OPEN DATA LITERATURE REVIEW

Emmie Tran and Ginny Scholtes[†]

Open data describes large datasets that governments at all levels release online and free of charge for analysis by anyone for any purpose. Entrepreneurs may use open data to create new products and services, and citizens may use it to gain insight into the government. A plethora of time saving and other useful applications have emerged from open data feeds, including more accurate traffic information, real-time arrival of public transportation, and information about crimes in neighborhoods. But data held by the government is implicitly or explicitly about individuals. While open government is often presented as an unqualified good, sometimes open data can identify individuals or groups, leading to invasions of privacy and disparate impact on vulnerable populations.

This review provides background to parties interested in open data, specifically for those attending the 19th Annual BCLT/BTLJ Symposium on open data. Part I defines open data, focusing on the origins of the open data movement and the types of data subject to government retention and public access. Part II discusses how open data can benefit society, and Part III delves into the many challenges and dangers of open data. Part IV addresses these challenges, looking at how the United States and other countries have implemented open data regimes, and considering some of the proposed measures to mitigate the dangers of open data.

† J.D. Candidates, 2016, University of California, Berkeley School of Law. Thank you to Chris Hoofnagle, Lecturer in Residence, Berkeley Law, for your indispensible help in producing this literature review. We also express gratitude to the Fall 2014 Berkeley Technology Law Journal Symposium Team for their contributions: Stephanie Cheng, Gabrielle Dorais, Jitesh Dudani, Anna Gassot, Jackie Kimble, Jenna Le, Maria Paula Souza, Michael Poon, Ariel Rogers, Monsura Sirajee, and Yu Tanebe. Finally, thank you to Ravi Antani and Sydney Ryan for their help editing this literature review.

I. WHAT IS OPEN DATA?

"Open data" is information that is accessible to everyone, machine readable, offered online at zero cost, and has no limits on reuse and redistribution.[1] Advances in collecting, processing, disseminating, and preserving information have resulted in the proliferation of data from a wide variety of sources. Governments collect a wide variety of information from their citizens, compile that information into de-identified datasets, and release those datasets to the public as open data. While the modern open data movement—emphasizing online availability—is relatively recent, the principles behind today's open data have been operating for more than a century.

A. ORIGINS OF THE MODERN OPEN DATA MOVEMENT

The push for access to government data has its historical underpinnings in the "open government" movement, which began in the 1800s in Britain with calls for proactive publication of laws and parliamentary minutes. In the United States, the open government movement was solidified by the passage of the Freedom of Information Act in 1966,[2] which allows for the full or partial disclosure of previously unreleased information and documents controlled by the United States government.[3]

Most government records came on paper or were otherwise "closed."[4] Over time, transparency concerns transformed into calls for closer collaboration between government and

---

1 U.S. DEP'T OF COMMERCE, FOSTERING INNOVATION, CREATING JOBS, DRIVING BETTER DECISIONS: THE VALUE OF GOVERNMENT DATA 28 (2014).

2 The Freedom of Information Act was enacted in 1966 and was subsequently amended in 1974 with a series of changes more commonly known as the Freedom of Information Act (FOIA) Amendments of 1974.

3 Albert Meijer et al., *Understanding the Dynamics of Open Data: From Sweeping Statements to Complex Contextual Interactions*, *in* OPEN GOVERNMENT: OPPORTUNITIES AND CHALLENGES FOR PUBLIC GOVERNANCE 101, 103 (Mila Gascó-Hernández ed., 2014); Harlan Yu & David G. Robinson, *The New Ambiguity of "Open Government,"* 59 UCLA L. REV. DISC. 178 (2012).

4 JONATHAN GRAY & HELEN DARBISHIRE, BEYOND ACCESS: OPEN GOVERNMENT DATA AND THE RIGHT TO (RE)USE PUBLIC INFORMATION 21 (Creative Commons, 2011). "Government information was traditionally held as paper documents but is increasingly held as electronic documents and in databases."

society, with demand soaring for access to public sector data in formats that facilitate reuse.[5] Eventually, the Electronic Freedom of Information Act Amendment of 1996 was enacted with the intent of bringing the law "into the electronic age;" it made clear that electronic records are "agency records" covered by FOIA and required agencies to provide information "in any form or format requested," including in electronic form, "if the record is readily reproducible by the agency in that form or format."[6]

In addition to open access, the open government movement sought to enhance transparency through the proactive release of "open by default" government data.[7] The convergence of "open data" and "open government" principles is now reflected in public policy, with the United States government implementing an open data policy through the Office of Management and Budget's Open Government Directive (OGD), which called for agencies throughout the executive branch to take steps to promote transparency, participation, and collaboration in the publication and use of government data.[8]

B. TYPES OF DATA SUBJECT TO GOVERNMENT RETENTION AND PUBLIC ACCESS

Governments collect a wide range of data from their citizens. Government data exists across diverse categories, including but not limited to information published by agencies for statistical purposes, information in administrative records on individuals and businesses, and physical measurements about natural phenomena.[9] For example: the Census Bureau collects confidential data

5 Jeremy Weinstein & Joshua Goldstein, *The Benefits of a Big Tent: Opening Up Government in Developing Countries*, 60 UCLA L. REV. DISC. 38 (2012).
6 5 U.S.C. § 552(a) (1996).
7 Weinstein & Goldstein, *supra* note 5, at 38.
8 Jeffrey Alan Johnson, *From Open Data to Information Justice*, 16 ETHICS & INFO. TECH. 263, 264(2014).
9 U.S. DEP'T OF COMMERCE, *supra* note 1, at 9.

on age, sex, race and Hispanic origin on a decennial basis;[10] the Census Bureau also conducts an Equal Employment Opportunity Tabulation which collects race, ethnicity and sex data across certain geographies and job categories;[11] and every taxpayer submits personal financial information to the Internal Revenue Service each year.[12] Furthermore, public health data from the Center for Disease Control ("CDC") was a benchmark for predicting the outbreak of influenza in flu algorithms, and postal codes are "fixed categories" for direct marketers to predict demographics and socioeconomic clusters.[13] Access to certain types of data, such as court records, provides citizens with an essential window into the functioning of government.[14] Many types of data governments collect about individuals are subject to public access in an open data regime.

The public can download all of this information in electronic format and, more recently, via Application Programming Interfaces ("APIs"). APIs allow programmers to access data directly and feed data into applications, which makes complex analysis of government datasets much easier.[15] In this way, open data can feed big data, where a diverse and unstructured digital collection of information is mined to discover patterns.[16] Given that "algorithms for big data are often built on reliable, established, and authoritative data sources," data from public sector organizations can provide stability for these models.[17]

---

10 U.S. DEP'T OF COMMERCE, *supra* note 1, at 18.
11 U.S. DEP'T OF COMMERCE, *supra* note 1, at 18.
12 U.S. DEP'T OF COMMERCE, *supra* note 1, at 16.
13 Anne L. Washington, *Government Information Policy in the Era of Big Data*, 31 REV. POL'Y RES. 319 (2014).
14 Barbara Ubaldi, *Open Government Data: Towards Analysis of Open Government Data Initiatives* 4 (OECD Publishing, Working Paper on Public Governance No. 22, 2013); Amanda Conley, Anupam Datta, Helen Nissenbaum & Divya Sharma, *Sustaining Privacy and Open Justice in the Transition to Online Court Records: A Multidisciplinary Inquiry*, 71 MD. L. REV. 772, 774 (2012).
15 U.S. DEP'T OF COMMERCE, *supra* note 1, at 28.
16 Washington, *supra* note 13, at 319.
17 Washington, *supra* note 13, at 320.

II. HOW OPEN DATA CAN BENEFIT SOCIETY

Open data presents many opportunities to benefit society. Placing these possible benefits in three general categories: open data can promote economic development,[18] foster effective governance through information-based policies,[19] and increase civic engagement and democratic accountability.[20] Not all these potential benefits have come to fruition, possibly due to the challenges discussed in Part III.

A. ECONOMIC DEVELOPMENT

Open data policies can encourage innovation by lowering barriers to entry and increasing competition.[21] In a 2014 report on the value of government statistics,[22] the U.S. Department of Commerce updated the nation on its progress on its new strategic plan for economic development,[23] which aims—among other things—to "redefine how [the federal government] manage[s], optimize[s], and enable[s] public access to [its] treasure trove of data."[24] The report made four broad findings regarding government "data published by the Principal Federal Statistical Agencies within the Executive Branch."[25]

First, every year government data potentially "guides trillions of dollars of investments,"[26] although this figure may exaggerate open data's economic impact as the Department of Commerce

18 *See* U.S. DEP'T OF COMMERCE, *supra* note 1; Barbara Ubaldi, *Open Government Data: Towards Analysis of Open Government Data Initiatives* 4 (OECD Publishing, Working Paper on Public Governance No. 22, 2013) (arguing that open data is an important source of economic growth, new forms of entrepreneurships, and social innovation).
19 *See infra* Part II.B.
20 *See infra* Part II.C.
21 Tim O'Reilly, *Government as a Platform*, *in* OPEN GOVERNMENT: COLLABORATION, TRANSPARENCY, AND PARTICIPATION IN PRACTICE 11, 15–17 (Daniel Lathrop & Laurel Ruma eds., 2010).
22 U.S. DEP'T OF COMMERCE, *supra* note 1.
23 *See generally* U.S. DEP'T OF COMMERCE, AMERICA IS OPEN FOR BUSINESS: STRATEGIC PLAN, FISCAL YEARS 2014–2018 (2014).
24 U.S. DEP'T OF COMMERCE, *supra* note 1, at 1.
25 U.S. DEP'T OF COMMERCE, *supra* note 1, at 10.
26 U.S. DEP'T OF COMMERCE, *supra* note 1, at 3.

came to this conclusion by looking at the GDP and estimating the investments that could rely on government data. Businesses, organizations, individuals, and all levels of government rely on open data to make economic decisions, and the report includes many examples of private sector decisions based on open data.[27]

Second, the Department of Commerce report found that the cost of providing data is "small" relative to the "potential benefits."[28] Average annual spending, including the decennial census, by the Principal Federal Statistical Agencies is $3.7 billion.[29] While the report did not provide an exact estimate of the potential benefits provided by this data, it did note that the data allows most actors in the nation's nearly $17 trillion economy to make more informed decisions that optimize resource use and allow businesses to stay competitive.[30]

Third, the report found that government data is "uniquely comprehensive, consistent, confidential, credible, relevant, and accessible."[31] By collecting and providing useful data, the

27 U.S. DEP'T OF COMMERCE, *supra* note 1, at 14. For example: retail companies, such as Target, use American Community Survey data to "tailor [their] merchandise offerings for customers in its neighborhoods," U.S. DEP'T OF COMMERCE, *supra* note 1, at 19; businesses, such as PETCO, use geographic information systems based on demographic census data to decide where to open new locations, U.S. DEP'T OF COMMERCE, *supra* note 1, at 34; businesses use the Bureau of Labor Statistics' monthly data on prices received by domestic producers of goods and services to help set contract prices U.S. DEP'T OF COMMERCE, *supra* note 1, at 25; a community organization "formed to keep Philadelphia's downtown clean, safe, beautiful, and fun" uses census data on where people live and work to optimize its budget (*id.* at 14); various religious groups use census data to decide where to build houses of worship U.S. DEP'T OF COMMERCE, *supra* note 1, at 26; high school and college graduates and adults considering a career change use the Bureau of Labor Statistics' Occupational Outlook Handbook to make informed decisions about what career to pursue U.S. DEP'T OF COMMERCE, *supra* note 1, at 22; individual home buyers and real estate investors can use census data on pricing, household income, population growth, unemployment rate, and rental rates U.S. DEP'T OF COMMERCE, *supra* note 1, at 36; Michigan's Department of Community Health used census data to build hospitals where they were most needed U.S. DEP'T OF COMMERCE, *supra* note 1, at 23); and Fontana Unified School District in California used census data to optimize construction of new schools after rapid population growth U.S. DEP'T OF COMMERCE, *supra* note 1, at 24.

28 U.S. DEP'T OF COMMERCE, *supra* note 1, at 3.

29 U.S. DEP'T OF COMMERCE, *supra* note 1, at 11.

30 U.S. DEP'T OF COMMERCE, *supra* note 1, at 3.

31 U.S. DEP'T OF COMMERCE, *supra* note 1, at 3, 22–29.

government remedies a failure of the market to produce an optimal level of information.[32] The Confidential Information Protection and Statistical Efficiency Act of 2002 is one of many legal provisions designed to protect the privacy of survey respondents and regulate agencies that maintain databases with confidential information.[33] Motivated in part by a May 9, 2013 Executive Order promoting open data, the federal government now provides online access to a wide variety of data free of charge, including statistical reports and scientific and economic information that influences policies set in place by federal agencies, such as NASA, the Department of the Interior, and the IRS.[34] The federal government also supports private entities in making data more available and easier to use, a boon to the public as a whole, but especially to research communities.[35]

Finally, the report found that government data is "commercially valuable,"[36] although the methods used to calculate this value are highly speculative. The report estimated the annual revenues of firms that "engage in [go]vernment data-intensive business activities" such as Bloomberg, Gallup, Nielsen, and Zillow.[37] Using those annual revenues, the report estimated the value of government

32 Because information generally does not share the characteristics of products that can be produced efficiently by the market (rivalrous, excludable, has recognizable benefits, exhibits constant returns to scale, generates no externalities), the market would likely produce less than the socially optimal level of information. U.S. DEP'T OF COMMERCE, *supra* note 1, at 19–21. For further discussion of information as a market failure, see J. Bradford Delong & A. Michael Froomkin, *Speculative Microeconomics for Tomorrow's Economy*, FIRST MONDAY(Feb. 7, 2000), http://firstmonday.org/ojs/index.php/fm/article/view/726; Beth Allen, *Information as an Economic Commodity*, 80 AM. ECON. REV. 268, 271 (1990); Benjamin J. Bates, *Information as an Economic Good: Sources of Individual and Social Value*, *in* THE POLITICAL ECONOMY OF INFORMATION 76–94 (Vincent Mosco & Janet Wasco eds., 1988).

33 *See* 44 U.S.C. § 3501 (2002); *see also* titles 13 and 22 for confidentiality provisions specific to the Census Bureau and the Bureau of Economic Analysis, respectively.

34 U.S. DEP'T OF COMMERCE, *supra* note 1, at 28–29; *see also* THE WHITE HOUSE, U.S. OPEN DATA ACTION PLAN (May 9, 2014), *available at* https://www.whitehouse.gov/sites/default/files/microsites/ostp/ us_open_data_action_plan.pdf.

35 U.S. DEP'T OF COMMERCE, *supra* note 1, at 28–29.

36 U.S. DEP'T OF COMMERCE, *supra* note 1, at 4.

37 U.S. DEP'T OF COMMERCE, *supra* note 1, at 31–38. These activities may be grouped into four categories: (1) value-added re-packagers that "aggregate Government data from many different Federal Government agencies, private sector firms, other governments, and international organizations within a single application," for example Geolytics, Haver Analytics, and Socrata, (2) benchmarkers that "create statistics or other data products using non-Government data, but use Government data as a reference and standard with which to adjust, weight, or test the validity of their products," for example Nielsen and Gallup, (3) analysts that "use Government data (either directly or indirectly through re-packagers)

data to be between $24 billion per year and $221 billion per year.[38] Despite the large range, the estimates demonstrate that government data may support an "important sector of the economy."[39]

B. EFFECTIVE GOVERNANCE THROUGH INFORMATION-BASED POLICIES

Open data allows state, local, and all levels of federal government to create, promote, and execute information-based policies. Because it aims to foster innovation, efficiency, and effectiveness in government services, open data can help foster collaboration across and within public agencies and departments. [40] It also provides a platform for innovative and improved service delivery. [41] For example: local law enforcement maps crimes using census demographic and housing data;[42] state and city governments enact rent stabilization laws that use census information as benchmarks;[43] the CDC relies on census data to decide where to target outreach and cancer screening services;[44] the U.S. Department of Defense used the epidemic computational model GLEAM to track the spread of infectious diseases;[45] local governments use spatial data sharing to manage traffic accidents and address data;[46] and several U.S. cities, including Seattle, San Francisco,

to create products for other firms or to generate research based on the data," for example CNBC, The New York Times Company, and the News Corporation, and (4) data brokers that "compile Government data (and sometimes proprietary private sector data) from around the world and add value by aggregating and integrating the data," for example Acxiom, Bloomberg, Esri, Experian, IHS Global Insight, and Thomson Reuters. U.S. DEP'T OF COMMERCE, *supra* note 1, at 32–33.

38 *See* U.S. DEP'T OF COMMERCE, *supra* note 1, at 31–42.

39 U.S. DEP'T OF COMMERCE, *supra* note 1, at 42.

40 Ubaldi, *supra* note 18, at 14.

41 Ubaldi, *supra* note 18, at 14.

42 U.S. DEP'T OF COMMERCE, *supra* note 1, at 25.

43 U.S. DEP'T OF COMMERCE, *supra* note 1, at 33.

44 U.S. DEP'T OF COMMERCE, *supra* note 1, at 16.

45 Gianluca Misuraca et al., *Policy-Making 2.0: Unleashing the Power of Big Data for Public Governance, in* OPEN GOVERNMENT: OPPORTUNITIES AND CHALLENGES FOR PUBLIC GOVERNANCE 180 (Mila Gasco-Hernandez ed., 2014).

46 Glenn Vancauwenberghe, Ezra Dessers, Joep Crompvoets & Danny Vandenbroucke, *Realizing Data Sharing: The Role of Spatial Data Infrastructures, in* OPEN GOVERNMENT: OPPORTUNITIES AND CHALLENGES FOR PUBLIC GOVERNANCE 155, 158 (Mila Gasco-Hernandez ed., 2014) (studying how the amount of regulations in spatial data sharing affected the accessibility of spatial data sharing in four processes: the development of zoning plans, the management of traffic accident registrations, the management of address data, and the mapping of flood areas).

and Detroit, use the software-based demographic and development modeling tool UrbanSim for urban development.[47]

Through aggregation and publication of data sets, government can interact with many different stakeholders that have diverging interests, such as individual citizens, interest groups, and companies, to engage in complex decision-making.[48] Open data can also attract new entrants, such as technologists, into debates around different policies, ideally resulting in evidence-based government action that satisfies the needs of the affected parties.[49] Also, public officials may have political incentives to initiate open data policies because using information and communication technologies to improve government efficiency can enhance their reputation for innovation.[50]

C. CIVIC ENGAGEMENT AND PUBLIC ACCOUNTABILITY

Open data brings information that government policies are based on to the fingertips of any citizen with access to the Internet. With such ready access to this information, citizens may engage in the process of governance more effectively than when such information was available only by request, in hardcopy, or in person.[51] For example, some scholars advocate "citizensourcing," or

47 Misuraca, *supra* note 45, at 181.

48 *See* Albert Meijer, Josta de Hoog, Mark van Twist, Martijn van der Steen & Jorren Scherpenisse, *Understanding the Dynamics of Open Data: From Sweeping Statements to Complex Contextual Interactions*, *in* OPEN GOVERNMENT: OPPORTUNITIES AND CHALLENGES FOR PUBLIC GOVERNANCE 101, 107 (Mila Gascó-Hernández ed., 2014); *see also* Tim O'Reilly, *Government as a Platform*, *in* OPEN GOVERNMENT: COLLABORATION, TRANSPARENCY, AND PARTICIPATION IN PRACTICE 11, 34–35 (Daniel Lathrop & Laurel Ruma eds., 2010) (engaging health care data might enable the government to create a dynamic pricing system for Medicare reimbursements based on outcomes).

49 *See* Weinstein & Goldstein, *supra* note 5, at 41.

50 Weinstein & Goldstein, *supra* note 5, at 43.

51 *See* David G. Robinson et al., *Enabling Innovation for Civic Engagement, in* OPEN GOVERNMENT: COLLABORATION, TRANSPARENCY, AND PARTICIPATION IN PRACTICE 85 (Daniel Lathrop & Laurel Ruma eds., 2010); *see also* Tim O'Reilly, *Government As a Platform*, *in* OPEN GOVERNMENT: COLLABORATION, TRANSPARENCY, AND PARTICIPATION IN PRACTICE 11, 12 (Daniel Lathrop & Laurel Ruma, eds., 2010) (explaining how open data policies allows the government to act as a platform, enabling problem solving by citizens); Sarah Schacht, *Democracy, Under Everything, in* OPEN GOVERNMENT: COLLABORATION, TRANSPARENCY, AND PARTICIPATION IN PRACTICE 155, 157 (Daniel Lathrop & Laurel Ruma eds., 2010) (arguing that the incompetent communication of legislative information is a major gap between full citizenry engagement in the legislative process).

"taking a task that is traditionally performed by a designed public agent (usually a civil servant) and outsourcing it to an undefined, generally large group of people in the form of an 'open call.'"[52] With open data policies in place, the government can provide reliable, useful open data that individuals, research institutions, and interest groups can use to evaluate the effectiveness of government policies and develop alternative policies.[53]

As open data policies aim to increase transparency, they have the potential to allow the public to evaluate policies and hold government accountable.[54] For the public to use and interact with the information, it must be free over the Internet in open, structured, machine-readable formats that are downloadable in bulk.[55] Government websites containing open data can motivate users to interact with that data by offering advanced search options, standard web feed formats, and links to other sources.[56] It is good practice for the government to ensure that users can authenticate all published data, for example through National Institute of Standards and Technology (NIST) standard "digital signatures."[57]

52 Dennis Hilgers & Christoph Ihl, *Citizensourcing: Applying the Concept of Open Innovation to the Public Sector*, 4 INT'L J. PUB. PARTICIPATION 67, 72–73 (2010) (presenting a framework for citizen-engaged governance with three dimensions: "Citizen Ideation and Innovation," "Collaborative Administration," and "Collaborative Democracy").
53 *See* Weinstein & Goldstein, *supra* note 5, at 44–46 (arguing that open data can be especially beneficial for emerging democracies, where transparent government performance is critical as information on service delivery allows political oversight, citizen monitoring, and feedback via the electoral process).
54 *See* Weinstein & Goldstein, *supra* note 5, at 39 (arguing that an "open data campaign can accelerate demand for information and generate a public conversation about what kind of data matter for accountability"); Maxat Kassen, *A Promising Phenomenon of Open Data: A Case Study of the Chicago Open Data Project*, 30 GOV'T INFO. Q. 508 (2013) (explaining that a goal of open data is to technology to usher in a new era in participatory democracy); Ubaldi, *supra* note 18, at 4 (explaining how open data can (1) help the public better understand what the government does and how well it performs, (2) hold government accountable for wrongdoing or unachieved results, and (3) provide a basis for public participation and collaboration in the creation of innovative, value-added services).
55 Robinson et al., *supra* note 51, at 87.
56 Robinson et al., *supra* note 51, 87–89.
57 Robinson et al., *supra* note 51, 87–89.

Thus, open data has the potential to promote civic engagement and effective governance.[58] A commitment to open data involves reorienting the production of public information towards adaptability, and necessarily makes bureaucracies more citizen-facing.[59] This reorientation may lead to changes in broader norms about openness, and may also accelerate cultural changes inside governments.[60]

III. THE CHALLENGES AND DANGERS OF OPEN DATA

While public access to government data can promote economic development, information-based policies, civic engagement, and democratic accountability, open data regimes pose multiple challenges. The main challenges and dangers of open data are conceptualization, disparate impact and civil rights violations, invasion of privacy and lack of consent, the failure of de-identification, security breaches, and transparency as an end itself.

A. CONCEPTUALIZATION

Research into the implications, legal and otherwise, of open data cannot happen in a vacuum; researchers must familiarize themselves with the social context of data collection and distribution. Primary collectors of data are in unique, privileged positions to know the nuances of data, including subtle nuances of their strengths and weaknesses. But open data relies on the idea that datasets can be reused by third parties for differential purposes; the failure to understand the full context could lead to problematic interpretations of the open data regime. To ensure that data sharing can be efficient, effective, accurate, and unbiased, researchers must first conceptualize and understand how data is collected, why data is collected, and who benefits from open data. Doing so would address

58 *See* Weinstein & Goldstein, *supra* note 5, at 44–46; *see also* O'Reilly, *supra* note 51; Schacht, *supra* note 51.
59 Weinstein & Goldstein, *supra* note 5, at 43.
60 Weinstein & Goldstein, *supra* note 5, at 43.

some existing legal challenges such as defining the scope of the right of access to information, clarifying and consistently applying the legal exceptions to openness, clarifying who owns government information, and "solving the compliance minefield that webmasters encounter."[61]

Some scholars have attributed the lack of conceptualizing with the failure of the Open Government Directive, which sought to provide the public with unprecedented access to government datasets on www.data.gov. Agencies put up new datasets but the directive did not adequately spell out what constituted a dataset and many agencies "sliced an diced" their numbers to look compliant.[62]

Additionally, the discourse of open data is often infused with the idea that data are objective when they are actually defined in a "normative, political, and ethical" process "that is often contested and has consequences for subsequent analysis, interpretation and action."[63] Failing to critically examining how data is assembled raises the risk that researchers use data without "knowing the politics of why and how such databases are constructed, the technical aspects of their generation, or having personal familiarity with the phenomena captured."[64]

B. DISPARATE IMPACT AND CIVIL RIGHTS VIOLATIONS

Analysis and use of open data can lead to possible civil rights violations through disparate impact on vulnerable populations, even—or maybe especially—when performed on a massive scale by experts with highly sophisticated software. Data mining "attempts to locate statistical

61 Ubaldi, *supra* note 14, at 37–38. In the United Sates, an online compliance checklist for designers of federal websites contains approximately twenty-four different regulatory regimes with which all public federal websites must comply. These range from privacy and usability to the Freedom of Information Act (FOIA) to the Paperwork Reduction Act.
62 Alon Peled, *When Transparency and Collaboration Collide: The USA Open Data Program*, 61 J. AM. SOC'Y SCI. & TECH. 2085, 2086 (2011); *see also infra* Part IV.A.
63 Rob Kitchin, DATA REVOLUTION: BIG DATA, OPEN DATA, DATA INFRASTRUCTURES AND THEIR CONSEQUENCES 19 (2014).
64 *Id.* at 22.

relationships in a dataset" and "automate the process of discovering useful patterns, revealing regularities upon which subsequent decision-making can rely."[65] There are multiple points in this process where bias can unintentionally enter a system designed to be fair. For example, a computer program used to sort medical school applicants on the basis of previous admission decisions was discovered to have relied upon biased data,[66] and "Google queries for black-sounding names were more likely to return contextual (i.e., key-word triggered) advertisements for arrest records than those for white-sounding names."[67]

First, data mining defines the "target variable," or the outcome of interest, and "class labels," or the different classes of data that the program must distinguish, in order to convert an amorphous problem into an algorithm that computers can use to parse a dataset.[68] Defining the target variable is a subjective process, leaving room for the data miner's subconscious bias to enter the analysis.[69] Second, data mining algorithms learn by example, and in order to teach the algorithms what to look for, data miners use "training data."[70] If the training data contains bias, either through subtly mislabeled data examples that reflect societal prejudice or through incomplete data collection that over or underrepresents certain classes, then data mining's results will be biased.[71] Third, data mining involves "feature selection," where data miners pick input variables to further differentiate the data.[72] Feature selection can lead to biased results because data miners may fail to include enough features that make pertinent distinctions between members of a protected class, and the features that are

65 Solon Barocas & Andrew Selbst, BIG DATA'S DISPARATE IMPACT 7 (Sept. 14, 2013) (unpublished manuscript) (on file with authors).
66 *Id.* at 11.
67 *Id.* at 12.
68 *Id.* 6–7.
69 Teresa Scassa, *Privacy and Open Government*, 6 FUTURE INTERNET 397, 399 (2014).
70 Barocas & Selbst, *supra* note 65, at 10.
71 Barocas & Selbst, *supra* note 65, at 10.
72 Barocas & Selbst, *supra* note 65, at 17.

included may contain baked-in prejudices.[73] Therefore, the collection and aggregation of big data may produce a high margin of error through the corruption of collected data and flaws in data entry.[74]

Criteria that sort data to achieve the desired outcome (for example, factors that accurately predict which applicant will excel at the job) may also serve as "proxies" for membership in a protected class, teaching the data mining algorithm that there is a relationship between the desired outcome and those who are not members of a protected class.[75] So, even systems that do not explicitly take into account factors such as race, gender, or socioeconomic status can unintentionally discriminate against vulnerable populations. This effect is especially concerning when the data unintentionally teaching systems to discriminate comes from open government initiatives. The diverse sources of data and methods of data entry make it challenging for data miners to control for every known risk or bias, even when the source of the data is the government.[76] Sometimes, government data can be misleading or even inaccurate.[77]

Cumulatively, open data has the potential to exacerbate societal inequities because most citizens cannot access, interpret, and translate this data into action. On the other hand, the most empowered users are already located in data companies, and have greater skills and access to proprietary data to

---

73 Barocas & Selbst, *supra* note 65, at 17.
74 K. Krasnow Waterman & Paula J. Bruening, *Big Data Analytics: Risks and Responsibilities*, 4 INT'L DATA PRIVACY L. 89, 91 (2014).
75 Barocas & Selbst, *supra* note 65, at 20.
76 Waterman & Bruening, *supra* note 74, at 91.
77 Bill Allison, *My Data Can't Tell You That*, *in* OPEN GOVERNMENT: COLLABORATION, TRANSPARENCY, AND PARTICIPATION IN PRACTICE 263, 265 (Daniel Lathrop & Laurel Ruma eds., 2010). When examining records maintained on the buildup of spent fuel assemblies at power plants, authors Donald L. Bartlett and James B. Steele found inconsistent and illogical figures. This is despite that fact that "the Nuclear Regulatory Commission and the nuclear industry had developed an elaborate system, one in which each fuel assembly was assigned its own serial number."

enhance open data.[78] Heavy reliance on big data may marginalize groups that are not part of the data since they are "exclude[d] from the kinds of interactions that produce data and . . . their viewpoints [are] invisible to those who collect the data."[79]

One study conducted by Solly Benjamin and his colleagues in Bangalore showed that the digitization of land records in Bangalore was primarily being used by middle and upper income individuals and by corporations to gain ownership of land from the marginalized and poor.[80] The newly digitized records were the "basis for instructions to land surveyors and lawyers and others to challenge titles, exploit gaps in titles, take advantage of mistakes in documentation, [and] identify opportunities and targets for bribery, among others."[81]

In addition to the dangers of perpetuating biases vis-à-vis omission of entire datasets, data systems permit the identification of vulnerable subpopulations that have been subject to a number of human rights abuses in the past, including forced migration, internment, genocide, and crimes against humanity.[82] For example, special registration systems covering the Jewish and Gypsy populations of the Netherlands played a role in the detention of Dutch Jews and Gypsies prior to their eventual deportation and death during World War II.[83] More broadly, population data systems played a role in some of the worst human rights abuses of the past two centuries, such as the Nazi Holocaust, the internment of Japanese Americans, and the Rwandan genocide.[84]

78 Michael Gurstein, *Open Data: Empowering the Empowered or Effective Data Use for Everyone?*, FIRST MONDAY (Feb. 7, 2011), http://journals.uic.edu/ojs/index.php/fm/article/view/3316/2764.
79 Johnson, *supra* note 8.
80 Gurstein, *supra* note 78.
81 Gurstein, *supra* note 78.
82 William Seltzer & Margo Anderson, *The Dark Side of Numbers: The Role of Population Data Systems in Human Rights Abuses*, 68 SOC. RES., 482, 483 (2001).
83 *Id.* at 486.
84 *Id.* at 484.

At the same time, privileged members of society may opt out of public processes in order to avoid triggering public records. Transparency may drive a reluctance to vindicate rights in a court[85] and drive individuals to arbitration. More affluent people can use mechanisms such as the land trust to obscure their presence in putatively open records, such as property databases.[86]

C. INVASION OF PRIVACY AND LACK OF CONSENT

Open government raises broad privacy challenges. The first is balancing individual discretion over personal information with government transparency and accountability.[87] Large swathes of personal information collected by governments are deemed "public" according to various laws or regulations.[88] However, in many cases, decisions about their publicity were made in an era before the Internet when public records were practically obscure.[89] Promoting citizen confidence in government is the goal of open data model governance, yet this may occasionally clash with the protection of citizen privacy when "public personal information" is reused in ways that subject individuals did not envision.[90]

Digitized information can be rapidly copied, mined, matched, and used for a broad range of purposes that many would consider invasive of privacy.[91] For example, in the aftermath of the tragic school shooting in Newtown, Connecticut, a newspaper used public registry data to show the names

85 Peter A. Winn, *Online Court Records: Balancing Judicial Accountability and Privacy in an Age of Electronic Information*, 79 WASH. L. REV. 307, 308 (2004).
86 Matt Richtel, *For Tech Titans, Sharing Has Its Limits*, N.Y. TIMES, Mar. 14, 2015, at BU 4. Tech industry moguls – the champions of transparency – are using non-disclosure actions to prevent publicity firms from being given their home addresses.
87 Scassa, *supra* note 69, at 402. The government has the power to coerce individuals to reveal truthful information with one prime example being the U.S. Census. Even though the census was created with strong confidentiality guarantees, newer extractions of personal information may have no privacy guarantees at all.
88 Scassa, *supra* note 69, at 403.
89 Scassa, *supra* note 69, at 403.
90 Scassa, *supra* note 69, at 402–3. Other "public personal information" include "political campaign contributions, public servant salary information, building or renovation permits, land titles information, and so on."
91 Scassa, *supra* note 69, at 402–3.

and addresses of all registered gun owners in two New York counties.[92] Even though that information had been acceptably public when its access was limited to a government office, the public suddenly considered the gun registry data too invasive when it was represented on an interactive map and posted on the Internet.[93]

In another example, gay rights advocates used campaign finance disclosure statements to expose the names and donations of businesses and individuals who donated to Proposition 8, a proposed amendment to the California Constitution that sought to ban gay marriage.[94] From there, it was not difficult for opponents of the proposition to create an online interactive map that matched the name of each donor who supported Proposition 8, along with the amount of their donation, to the street address.[95] In addition to the major physical safety concerns associated with releasing this information, these campaign finance disclosure statements and related disclosure requirements are replete with the potential for repurposing by commercial entities,[96] home invasions, or identity theft.[97]

A second privacy challenge relates to the blurring of private and public sector information. Governments tend to be held to a stricter standard than private actors with respect to personal information collected from individuals.[98] However, governments are increasingly interacting with

92 Scassa, *supra* note 69, at 404.
93 Scassa, *supra* note 69, at 404.
94 Washington, *supra* note 13, at 324. This information is often times complete and accurate because they are official documents people often times use to claim tax dedications.
95 Washington, *supra* note 13, at 324.
96 Thomas P. Keenan, *Are They Making Our Privates Public? – Emerging Risks of Governmental Open Data Initiatives,* 375 IFIP ADVANCES IN INFO. & COMMC'N TECH. 1, 2 (2012). Genealogy websites like Ancestry.com, which receive the majority of their data from government sources, profit from their unbridled use. With a swift click of a button, Ancestry.com users can receive birth, death, marriage, immigration, travel and military service records whose subjects most likely never consented to their release. Commercial misuse of this data is ripe for abuse, with the most commonly cited example being insurance companies using an applicant's familial death records to infer that applicant's longevity and liability.
97 *Id.* at 7.
98 Scassa, *supra* note 69, at 405.

citizens through social networking platforms such as Facebook, Twitter, and Google Plus.[99] In some cases, private sector companies also provide the digital infrastructure for online feedback or complaints mechanisms, with the companies acting as a client and service provider of the government.[100] This form of citizen engagement may permit governments to gain access to an additional layer of personal information that individuals may not expect and under security measures that are less than desirable.[101]

Finally, the concerns highlighted above all run counter to a basic tenet of privacy research and scholarship: consent. Citizens usually receive basic and vague notices of the collection and future uses of their personal information, but those notices likely do not cover the many ways and possible extent to which government and private entities would use their data.[102] The collection and release of this information may also give rise to the practice of governmental data analysis; such actions by the government may constitute unreasonable searches since the data mining is carried out without sufficient judicial approval and within the public-private digital space.[103] On a broader level, the parsing of this data might have a "chilling effect" on many important activities and behaviors since citizens may fear additional governmental scrutiny into their daily lives.[104]

D. THE FAILURE OF DE-IDENTIFICATION

Open data may also contribute to the big data environment in which governments and private entities monitor citizens. Some scholars have argued that the existing privacy protection paradigm,

---

99 Scassa, *supra* note 69, at 405.
100 Scassa, *supra* note 69, at 405.
101 Scassa, *supra* note 69, at 406. Government agencies tacitly employ and enforce the privacy, security and other policies employed by these social media providers when they adopt the use of their tools, even if the tools and the policies governing them do not meet acceptable security standards.
102 Tal Z. Zarsky, *Governmental Data Mining and its Alternatives,* 116 PENN ST. L. REV. 285, 296 (2011).
103 *Id.*
104 *Id.* at 297.

based on de-identifying personally identifiable information, is an increasingly inadequate privacy safeguard as the amount of publicly available information about individuals grows.[105] De-identified public records, when combined with other publically available digitized information, can be re-identified and used for a broad range of privacy invasive purposes by both government and non-government actors.[106] A supposedly anonymous unique identifier can serve as the basis for recognizing an individual in and across multiple databases, and is more of a "unique persistent identifier" than a truly anonymous identifier.[107]

A report released by the United States General Accounting Office examined the practice of "combining existing person-specific data with additional data that refer to the same persons, their family and friends, school or employment, area of residence or geographic environment (hereinafter 'record linkage')."[108] Although the Privacy Act generally prohibits agencies and staff from disclosing or releasing identified data, this prohibition merely requires that the identity of individuals cannot be "reasonably deduced."[109] At least one agency (National Center for Health Statistics) warns users against attempting to re-identify persons in its public-use dataset, but the agency does not specify penalties for doing so.[110] Concerns about re-identification and record linkage have grown because of the proliferation of "data snoopers" and "data detectives," yet there is a lack of guidance and regulation in the area of re-identification.[111]

105 Arvind Narayanan & Vitaly Shmatikov, *Privacy and Security: Myths and Fallacies of "Personally Identifiable Information"*, 53 COMM. OF THE ACM 24, 26 (June 2010).
106 Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701 (2010).
107 Barocas & Selbst, *supra* note 65, at 53.
108 U.S. GOV'T ACCOUNTABILITY OFFICE, GAO-01-126SP, Record Linkage and Privacy: Issues in Creating New Federal Research and Statistical Information 10 (2001).
109 *Id.* at 69.
110 *Id.* at 70.
111 *Id.* at 69.

This lack of guidance is resoundingly apparent when analyzing Toronto's database on calls to the city's complaint and service request line. Although the database is generally de-identified to show only a partial postal code of the caller, there are numerous cases where the field contains a precise intersection.[112] Using databases like Google Maps and Google Street View, it is possible that "data detectives" could trace the calls back to the individual property and hence, to the owner.[113]

E. SECURITY

Aside from the threat that data the government purposely releases is reused in ways that invade individual privacy, there is also potential for unintentional releases of government information. As more and more court records are becoming digitized and publicly accessible online, sensitive information may slip through established filters and censors.[114] Court records contain a lot of personal information and potentially could include the entire case file.[115] Even if it does not appear at first glance that the information is that private, "when combined with other publicly available data . . . it may provide ample information for identity thieves."[116] Usually in state courts, the lawyers and their clients redact sensitive information like Social Security numbers, but some items may slip through, and sensitive information will be even harder to locate and remove when documents are scanned into PDF format.[117] While there are judicial rules and state statutes to help decide what is excluded and included, often in practice those considerations are on the backburner given the custom and convenience of the rules.[118] A concern is that "when they are in electronic format, court

112 Keenan, *supra* note 96, at 4.
113 Keenan, *supra* note 96, at 4.
114 Amanda Conley, Anupam Datta, Helen Nissenbaum & Divya Sharma, *Sustaining Privacy and Open Justice in the Transition to Online Court Records: A Multidisciplinary Inquiry*, 71 MD. L. REV. 772, 776 (2012).
115 *Id.*
116 *Id.* at 782.
117 *Id.*
118 *Id.* at 796.

clerks may in some instances, without oversight, decide to simply place all the records online to avoid having to complete paper requests at the courthouse and to provide greater accessibility to the interested parties."[119]

The New York City DataMine project was unveiled on October 6, 2009, touting a freely available collection of 103 municipal datasets.[120] However, "discussion on the public Sunlight Labs Google Group revealed that one XLS table listing the city's more than 1,100 women's organizations contained not only the personal email address of the group's contact with the New York City Commission on Women's Issues (CWI), but what he or she was using as the Secret Question and Secret Answer" in their user profiles.[121]

F. TRANSPARENCY AS AN END IN ITSELF

Support for open data presupposes transparency as a method of promoting good government, but taken to extremes, transparency can undermine government or upset other values. Some critics see open data as representing a confusion of ends and means, where transparency has become an end in itself rather than an instrument of government accountability. For instance, Slee warns of the "data doppelgänger:" the "shadow of commercial interests that follow civic hackers wherever they go; the new markets that spring up inevitably from the ruins of the old."[122] According to Slee, while one generally beneficial effect of open data is undermining the power of those who previously controlled access to the data, a possibly harmful result of that effect is the proliferation of new markets and new businesses.[123] These new open data-based markets "are likely to consist of a few,

119 *Id.* at 797.
120 *Id.*
121 *Id.* at 2.
122 Tom Slee, *Seeing Like a Geek*, CROOKED TIMBER (June 25, 2012), http://crookedtimber.org/2012/06/25/seeing-like-a-geek/.
123 *Id.*

big firms, each with significant market power."[124] So while governments release open data intending to democratize and diversify markets by giving data free of charge, open data can actually have the opposite effect and lead to more monopolistic markets.[125] Bates also cautions against the optimistic dominant narrative of open data, pointing out its potential to worsen economic disparity and "critically contextualiz[ing] [open data] within contemporary capitalist processes."[126] Finally, MacNeil emphasizes concerns over the actual uses of data.[127] Secondary users of data could be anyone and have no duties to data subjects or to the public.[128] According to MacNeil, academic freedom and utilitarianism cannot justify broad access to archival data for research purposes.[129] Rather, ethical duties of data archivists require a rights-based approach, with expert review of requests for access to protect privacy and vulnerable members of society.[130]

A 2003 article highlighted how open government and data brokers were beginning to systemically undermine civil liberties.[131] In it, Chris Hoofnagle observed that big data companies, many of which hailed from the direct marketing field, were scooping up personal information in public records and reselling it to law enforcement and intelligence agencies. In turn, government agencies were monitoring privacy laws, because such restrictions would hamper the ability of data brokers to sell data back to the government. Data brokers had become big supporters of open

---

124 *Id.*
125 *See id.*
126 Jo Bates, *"This Is What Modern Deregulation Looks Like:" Co-optation and Contestation in the Shaping of UK's Open Government Data Initiative*, JOURNAL OF COMMUNITY INFORMATICS (2012), http://ci-journal.net/index.php/ciej/article/view/845/916/.
127 HEATHER MACNEIL, WITHOUT CONSENT: THE ETHICS OF DISCLOSING PERSONAL INFORMATION IN PUBLIC ARCHIVES 82–83 (1992).
128 *Id.*
129 *Id.*
130 *Id.*
131 Chris Jay Hoofnagle, *Big Brother's Little Helpers: How ChoicePoint and Other Commercial Data Brokers Collect and Package Your Data for Law Enforcement*, 29 N.C. J. INT'L L. & COM. REG. 595 (2003).

government and transparency laws—they borrowed the transparency agenda in order to promote transparent citizenry using the rhetoric of transparent government.

Many repercussions result from radical transparency agendas. First, openness in court records could lead to less cooperation with government because individuals may fear that their affairs will become grist for the media.[132] Second, a loss of control over data with subsequent aggregations can cause less transparency and more inaccuracy.[133] Third, disclosed data can distort reality, misleading the public and providing a false appearance of public accountability.[134]

Advocates of transparency often invoke Louis D. Brandeis' famous quote, "[s]unlight is said to be the best of disinfectants."[135] Evgeny Morozov observes that when transparency causes privacy invasions or begins to conflict with other liberal values, "disinfectants, alas, are of little use to sunburn victims."[136]

IV. ADDRESSING THE CHALLENGES AND DANGERS OF OPEN DATA

Given the dangers open data poses, a functional open data system requires measures to protect the privacy, civil rights, and security of the public. This Part reviews the open data systems currently in use, both in the United States and in other countries, then reviews some of the proposed ex ante and ex post measures to mitigate the dangers of open data. Then, this Part discusses a

132 *See, e.g.*, EVGENY MOROZOV, TO SAVE EVERYTHING, CLICK HERE: THE FOLLY OF TECHNOLOGICAL SOLUTIONISM 63–99 (2013). The article concludes that information "needs to be collected and distributed in full awareness of the institutional and cultural complexity of the institutional environment in which it is gathered. Sometimes preserving the social relations that enable that environment to exist—for example, to make policing of crimes possible—might require producing data that is only half transparent or half accessible . . . . The tyranny of openness—the result of our infatuation with Internet-centrism—must be resisted." *Id.* at 99.

133 *Id.*

134 *Id.*

135 Louis Brandeis, OTHER PEOPLE'S MONEY AND HOW THE BANKERS USE IT 62 (1914).

136 Morozov, *supra* note 132, at 63 (showing how the liberal value of transparency can be illiberal, especially when openness becomes an end in itself rather than as a means to accountability and arguing that transparency advocates, perhaps unwittingly, buy into a rational choice theory paradigm of problems when calling for openness).

common critique of open data as purporting to support liberal values when it actually leads to illiberal ends.

A. OPEN DATA SYSTEMS CURRENTLY IN USE

1. Open Data in the United States

President Obama entered office aiming to establish "an unprecedented level of openness in Government" and set about prioritizing open-government initiatives.[137] The Open Government Directive (OGD) in December 2009 ordered all agencies to "make as much information as possible available online" through data.gov, but few agencies embraced the directive, with only 5 out of 169 participating agencies accounting for 99.37% of all datasets and applications in May 2011.[138] Both practical reasons—preparing data sets for public use takes resources—and political reasons—getting blame for bad data and no reward for providing good data—explain agencies' reluctance to participate in OGD.[139] Even though the White House hailed data.gov a success, and governments around the world embraced open data principles, the program had been criticized for its vague definition, impractical goal of maximizing transparency, emphasis on technology as a marker of transparency, and failure to establish standards for openness, prioritization and maintenance of

137 Alon Peled, *When Transparency and Collaboration Collide: The USA Open Data Program,* 61 J. AM. SOC'Y SCI. & TECH. 2085, 2086 (2011). On May 21, 2009 the White House launched www.data.gov with aspirations to make this site "the premier web-publishing location for all important federal databases." *Id.*

138 *Id.* at 2086–88. In Peled's study, the research team downloaded weekly "snapshots in time" of the performance data and concluded that most agencies met the minimal OGD requirements and did "virtually nothing" afterwards. *Id.* In 2011 when this study was conducted, almost 30% of all agencies had done nothing in data.gov for a full year or even more. *Id.*

139 *Id.* at 290–91. Gathering and analyzing data is expensive, and agencies use datasets as bargaining chips with other agencies either for money or information exchanges *Id.* at 2090. The article summarizes the issue by saying, "it is unfair and politically unfeasible to demand federal agencies to give for free the information products they labored hard to create." *Id.* The federal agencies did not want to give the Open Data Program any sort of datasets, including less valuable ones that they could not trade with other agencies, because they would "post it forever and delete nothing." *Id.* Also, because the Open Data Program would post information without mechanism for deletion, agencies would not be able to change bad data, leading to confusion and public blame. *Id.* "In other words, Open Data architects offered federal agencies a bad 'trade': Open Data would receive fame and glory for 'freeing government data,' [but] federal agencies would be robbed of their ability to trade data." *Id.* at 2091. Looking at the future of the Open Data Program, the article notes that federal agencies will continue adopting a "passive-aggressive stance" by giving only the least amount of useless data, but keeping their valuable datasets close their chests. *Id.*

data.[140] The new data.gov, unveiled in 2012, has also been criticized as suffering from flaws similar to those in the original version.[141] One recommendation to improve data.gov is to establish a formal internal Federal Information Marketplace that would to improve data-sharing via trade so that information can be controlled and sent to the right people in real time.[142]

States and local governments have also implemented open data policies. For example, many states have electronic filing requirements for campaign finance information, and the ethics commissions and campaign finance disclosure agencies of all fifty states have web presences.[143] A case study found that the City of Chicago's open data project, which aims to increase transparency and encourage citizens to access the data, successfully promoted democratic government because it hosted several data sets across sectors, collaborated with a highly active civic base, and solicited the help of various foundations and academic institutes.[144]

---

140 Alon Peled, *Re-Designing Open Data 2.0,* PROCEEDINGS OF THE CEDEM13 CONFERENCE 243, 246–47; *see also* Micah L. Sifry, *"You can be the Eyes and Ears": Barack Obama and the Wisdom of Crowds*, *in* OPEN GOVERNMENT: COLLABORATION, TRANSPARENCY, AND PARTICIPATION IN PRACTICE 117, 121 (Daniel Lathrop & Laurel Ruma eds., 2010). Suggested research avenues to improve data.gov include: (1) how agencies change the data to appear as if they are complying with the transparency policies and (2) what role IT plays in the "cat and mouse game between politicians seeking to control bureaucrats and bureaucrats developing new means to evade such control." Peled, *supra* note 137, at 2092.

141 Peled, *supra* note 140, at 248–50 (arguing that to be successful, Open Data 2.0 must become a component of a broader transparency program that delivers high-quality data for a low data-integration cost, manage cost by keeping technology reasonable, and establish measures that decrease the gap between the "data haves" and the "data have-nots").

142 Peled, *supra* note 137, at 2092.

143 Edwin Bender, *Case Study: FollowTheMoney.org*, *in* OPEN GOVERNMENT: COLLABORATION, TRANSPARENCY, AND PARTICIPATION IN PRACTICE 216 (Daniel Lathrop & Laurel Ruma eds., 2010). However, a lack of political will plagues the movement for greater transparency. *Id.* at 219. On the state level, this lack of political will manifests in a variety of ways, from poor quality of data posted online in Missouri to Utah's campaign finance website being offline for an entire legislative session. *Id.* Even in states that require electronic records, these records are often available only in formats that do not allow for analysis in a database, as is the case in Montana and Connecticut. *Id.* at 220. South Carolina's efforts to require online filing have been plagued by incomplete data and lack of funding to implement the electronic filing system.

144 Kassen, *supra* note 54, at 509–11. According to the study, the platform readily provides user-friendly access to data published by the city with a variety of search tools and solicits feedback from citizen users, staying true to the goal of transparency. *Id.* at 510. The platform's functionality and features have been expanded upon through collaboration with non-profit organizations and independent developers. *Id.* This expansion has resulted in related projects focused on matters such as lobbying, public works, and access to public resources. *Id.* at 511.

Several organizations have recently launched projects to make open data more accessible and usable to the public. For example, the National Institute on Money in State Politics launched FollowTheMoney.org to help citizens track the influence of campaign contributions in all fifty states;[145] MAPLight.org aggregates data on money and politics from US Congress, California State Legislature, and Los Angeles officials;[146] and GovTrack gathers and analyzes publically available legislative information to provide users with a dynamic view of Congress.[147] While these projects aimed to increase accountability by making such information public, transparency has largely failed as a way to regulate the activity of public officials.

2. Open Data Outside the United States

Regions and countries outside the United States have also instituted open data policies. The Open Government Partnership has grown from eight member countries at its inception in 2011 to sixty-five participating countries committed to goals of transparency, responsiveness, accountability, and effectiveness.[148] The EU has also been a leader in open data systems; it currently maintains the EU Open Data Portal[149] and has several directives pertaining to open data. In an effort to expand European residents' right to knowledge, the EU adopted the Public Sector Information ("PSI") Directive in 2003, which regulates the obligations of public sector bodies in the EU when re-using

145 *Id.* at 216. The Legislative Committee Analysis Tool groups members of Congress based on their committee assignments and primary campaign contributors, showing which donors with economic interests have targeted donations to members of specific committees. *Id.* at 217. Project Vote Smart combines candidate biographies and voting information with donor lists, and makes this data available to other websites via APIs. *Id.* at 217–18.

146 Daniel Newman, *Case Study: MAPLight.org*, *in* OPEN GOVERNMENT: COLLABORATION, TRANSPARENCY, AND PARTICIPATION IN PRACTICE 225, 227–233 (Daniel Lathrop & Laurel Ruma eds., 2010). The tools on their website allow users to compare campaign money, votes, and special interest positions on bills. *Id.* at 227.

147 Joshua Tauberer, *Case Study: GovTrack.us*, *in* OPEN GOVERNMENT: COLLABORATION, TRANSPARENCY, AND PARTICIPATION IN PRACTICE 203, 205–07 (Daniel Lathrop & Laurel Ruma eds., 2010) (explaining that GovTracks allows users to track changes to legislative bills, personalize their view into Congress by creating customized newsfeeds, and engage with other users through a question and answer forum).

148 *Participating Countries*, OPEN GOVERNMENT PARTNERSHIP, http://www.opengovpartnership.org (last visited Mar. 7, 2015).

149 *See* EUROPEAN UNION OPEN DATA PORTAL, https://open-data.europa.eu/en/data/ (last visited Mar. 11, 2015).

public sector data.[150] The EU's Data Protection Directive protects individuals' fundamental right to privacy regarding personal data usage, broadly defined to include personal information encompassed in public sector data.[151] These two interests—the right to reuse PSI and the protection of personal data—are often considered as being incompatible and several cases "discussed before some European Data Protection Authorities have shown that conditions for the reuse of PSI may raise issues of data protection."[152]

Scholars and policymakers suggest frameworks for striking a balance between the reuse of PSI and the protection of personal data in accordance with the suggestions of the European data protection authorities, both the European Data Protection Supervisor ("EDPS") and the Data Protection Working Party Article 29 ("WP29").[153] The WP29 recommends making data anonymous to avoid disclosure of personal information and a "case by case" approach in connection with the enforcement to protect data subjects' rights,[154] whereas the EDPS recommends a "proactive approach."[155] While the European Commission is a strong advocate of the release of public sector

150 Hans Graux, *Open government data: reconciling PSI re-use rights and privacy concerns*, EUROPEAN PUBLIC SECTOR INFORMATION PLATFORM, TOPIC REPORT NO. 2011/3 at 4.
151 *Id.* at 5.
152 Ugo Pagallo & Eleonora Bassi, *Open Data Protection: Challenges, Perspective and Tools for the Reuse of PSI*, DIGITAL ENLIGHTENMENT YEARBOOK, 179, 179 (2013). For example, the Fair-Play Alliance, a Slovakian NGO advocating government transparency, built an application to help citizens detect potential corruption. *Id.* at 8. But data in the application, which originated from the government, raised privacy concerns and a possible Data Protection Directive violation because it allowed the identification of specific individuals. *Id.* at 9–10. In another example, the detailed crime maps published by UKCrimeStats, while providing a useful service to citizens, also risked invading individuals' privacy because a user could link crimes on the map with identifiable individuals in the neighborhood, posing a difficult question of what qualifies as personal data under the Data Protection Directive. *Id.* at 11–14.
153 Pagallo & Bassi, *supra* note 152, at 183–84.
154 Pagallo & Bassi, *supra* note 152, at 184. However, the authors argue that the "case by case" approach could lead to a greater heterogeneity of practices between Public Sector Bodies and between Member States.
155 Pagallo & Bassi, *supra* note 152, at 184 ("Here, 'a proactive approach means that institutions assess and subsequently make clear to data subjects - before or at least at the moment they collect their data – the extent to which the processing of such data includes or might include its public disclosure.") (citations omitted). The authors consider that, while this approach can be very useful in some case, it does not fit all the personal data gathered by Public Sector Bodies over the past decades. Pagallo & Bassi, *supra* note 152, at 184.

data and the harmonization towards coherent national policies, open data implementation activities at the national and local levels often lack coordination with supranational strategies.[156]

Many countries have adopted individual open data initiatives to some extent. For example, New Zealand,[157] Kenya,[158] Austria,[159] the UK,[160] the Netherlands,[161] and many other EU member countries[162] have open data initiatives that aim to make government information available online for free to the public.

B. PROPOSED MEASURES TO MITIGATE THE DANGERS OF OPEN DATA

Open data's potential for privacy, civil rights, and security violations has become more apparent and more studied over the past several years. In response, scholars, advocates, and policy-makers have proposed measures to mitigate the dangers of open data.

---

156 Isabell Egger-Peitler & Tobias Polzer, *Open Data: European Ambitions and Local Efforts. Experiences from Austria*, *in* OPEN GOVERNMENT: OPPORTUNITIES AND CHALLENGES FOR PUBLIC GOVERNANCE 137, 138–140 (M. Gascó-Hernandéz ed., Springer 2014).

157 New Zealand launched its modern open government initiative in 2011. *Declaration on Open and Transparent Government*, ICT, http://ict.govt.nz/guidance-and-resources/open-government/declaration-open-and-transparent-government/ (last visited Mar. 11, 2015); *see also Open Government and Data Information Programme*, ICT, http://ict.govt.nz/programmes-and-initiatives/open-and-transparent-government/open-government-information-and-data-work-programm/ (last visited Mar. 11, 2015). New Zealand also launched a "Policing-Act-Wiki," a collaborative democracy platform to allow the public to contribute to the drafting of a new law. Hilgers, *supra* note 52, at 80.

158 *See* KENYA OPEN DATA, https://opendata.go.ke (last visited Mar. 11, 2015); Weinstein & Goldstein, *supra* note 49, at 44–46. The Kenya Open Data Initiative started in July 2011; by January 2012, the open data portal had released over four hundred datasets. Weinstein & Goldstein, *supra* note 49, at 44–46. Citizen groups are also promoting open data, and organizations like Uwezo and Twaweza in East Africa are prioritizing access to data on government performance. *Id.*

159 *See* OFFENE DATEN ÖSTRREICHS, https://www.data.gv.at; *see also* Isabell Egger-Peitler & Tobias Polzer, *Open Data: European Ambitions and Local Efforts. Experiences from Austria*, *in* OPEN GOVERNMENT: OPPORTUNITIES AND CHALLENGES FOR PUBLIC GOVERNANCE 137, 143–44 (Mila Gascó-Hernández ed., 2014) (explaining that in Austria, municipalities are the drivers of opening public data; the federal level is regarded as "dispassionate," reserved, and concentrated on a primarily coordinating role).

160 *See* DATA.GOV.UK: OPENING UP GOVERNMENT, http://data.gov.uk; *see also* Bates, *supra* note 126 (arguing that the "shaping of OGD is open to significant contestation," focusing on "the situation in the UK where the OGD initiative intersects with the UK government's programme of forced 'austerity' and marketisation of public services.").

161 *See* DATA.OVERHEID.NL: HET OPENDATAPORTAAL VAN DE NEDERLANDSE OVERHEID, https://data.overheid.nl (last visited Mar. 11, 2015); *see also* Meijer et al., *supra* note 48, at 108–110 (analyzing case studies on open data regarding public transportation and regarding policing information in the Netherlands).

162 EU member states are also expected to formulate national open data policies. Egger-Peitler & Tobias Polzer, *supra* note 159, at 141.

1. Ex-ante measures

Many scholars advocate for ex ante measures, or restrictions on open data before the data is published online and made available to the public. These measures fall into four general categories. First, transparency about the process of government gathering and distributing data can help protect against privacy invasions, prevent perpetuation of social inequities, and foster trust between individuals and institutions that use open data.[163] A review of the UK's transparency program concluded that privacy and transparency are compatible if they are treated together and with the knowledge that transparency programs can lead to invasions of privacy.[164] The review also made fourteen recommendations to the UK government to protect privacy while promoting transparency, most of which deal with disclosing the methods of data collection and types of data collected, raising awareness about existing data protection schemes, and ensuring that the public's privacy interests are sufficiently represented in government agencies collecting and distributing data.[165]

163 *See* Kieron O'Hara, *Are They Making Transparent Government, Not Transparent Citizens: A Report on Privacy And Transparency for the Cabinet Office*, REVIEW OF PRIVACY AND TRANSPARENCY 58–75 (2011); Peled, *supra* note 140, at 250 (arguing that a successful open data program must be a component of a broader transparency program that delivers high-quality, contextualized data for a low data-integration cost); Bhuvaneswari Raman, *The Rhetoric and Reality of Transparency: Transparent Information, Opaque City Spaces and the Empowerment Question*, 8 J. COMMUNITY INFORMATICS 9, 9 (2012) (arguing that without holistic spatial information transparency systems that take into account the wider political economy of information and land, open data can adversely affect the land claims of poor groups in society); Neil M. Richards & Johnathan H. King, *Big Data Ethics*, 46 WAKE FOREST L. REV. 393, 419 (2014) (arguing that transparency is necessary for the regulation of big data because it fosters trust between individuals and the institutions that use big data); *see also* Natali Helberger, *Form Matters: Informing Consumers Effectively,* 1 IVIR 51, 46 (2013) (arguing that mandated transparency can regulate the current market if the information presented is streamlined, digestible, and personalized enough for consumers to make informed decisions about agreeing to privacy and security policies).

164 *See* O'Hara, *supra* note 163, at 27.

165 *See* O'Hara, *supra* note 163, at 58–75 (recommending that the UK (1) represent privacy interests on the Transparency Board, (2) use disclosure, query and access controls selectively, (3) focus on and increase awareness of the technological paradigm of information science, (4) move toward a demand-driven regime by allowing information entrepreneurs ask for the datasets they need, (5) release a register of government data assets listing datasets, their contents and whether they are publicly available or not, (6) create transparency panels that would consider requests for data and whether there is a prima facie threat to privacy, (7) create for procedure for pre-release screening of data to ensure respect for privacy, (8) extend the research base and maintain an accurate threat model, (9) aggregate and disseminate current transparency research in a digestible format, (10) ensure that the rules for retaining information remain adequate when the default is to release rather than retain information, (11) maintain existing procedures for identifying harms and remedies, (12) use

Transparency practices would also apply to private entities that re-package open government data. In a May 2014 report, the FTC recommended that Congress consider enacting legislation to make data broker practices more visible to consumers and to give consumers greater control over the immense amounts of personal information about them collected and shared by data brokers.[166] As discussed in Part III, however, many commentators on open records and transparency criticize reliance on transparency as a solution to the challenges of open data.[167]

Second, the government could regulate the type of data it can collect, either through statutes passed by Congress, rules passed at an agency level to govern agency conduct, or executive order or policy guidance from the White House. Avoiding collecting certain types of data can safeguard against civil rights abuses by preventing governments (or parties that use data released by the government) from targeting certain groups, and several countries with a history of misuses of data have consciously decided not to gather or save certain data that permits associating an individual with a potentially vulnerable group.[168]

Third, the government could regulate the type of data it can make available to the public. Given the danger that anonymized public records could be re-matched with other mined data and used for privacy invasive purposes, scholars argue that open model governments should reassess

---

data.gov.uk to raise awareness of data protection responsibilities, (13) investigate the vulnerability of anonymised databases, and (14) be transparent about the use of anonymisation techniques).

166 FED. TRADE COMM'N, DATA BROKERS: A CALL FOR TRANSPARENCY AND ACCOUNTABILITY VIII 50 (May 2014), *available at* http://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf.

167 *See, e.g.*, MOROZOV, *supra* note 132 at 63–99. The article concludes that information "needs to be collected and distributed in full awareness of the institutional and cultural complexity of the institutional environment in which it is gathered. Sometimes preserving the social relations that enable that environment to exist—for example, to make policing of crimes possible—might require producing data that is only half transparent or half accessible . . . . The tyranny of openness—the result of our infatuation with Internet-centrism—must be resisted." MOROZOV, *supra* note 132, at 99.

168 *See* Seltzer & Anderson, *supra* note 82, at 495 (explaining (1) how population data systems permit the identification of vulnerable subpopulations, which historically, have contributed to a number of human rights abuses, including forced migration, internment, genocide, and crimes against humanity and (2) how regulating data collection can help prevent such abuses).

whether it is appropriate or necessary to disclose certain public records, even in an anonymized format.[169]

Fourth, the government could instate methodological and technical safeguards on how data is collected, analyzed, and presented. For example, scholars have proposed that agencies could require data systems to be based on sample rather than full-count data gathering or deliberately introducing errors into the data set in order to reduce the level of detail.[170]

2. Ex-post measures

Interested parties have also proposed a variety of ex-post measures, or policies to mitigate the dangers of open data after the data have been released to the public. First, the government could regulate, through a variety of statutory or administrative mechanisms, the use of open data after it has been published by the government. Rules and regulations governing personal information flows could resolve the essential privacy problem posed by the use of open data by preventing use of information that leads to privacy violations.[171] The FTC recommended that Congress consider legislation requiring data brokers to provide consumers access to, some control over, and a better understanding of their data.[172] Also, government could require entities that use open data as part of

169 Teresa Scassa, *Privacy and Open Government*, 6 FUTURE INTERNET 397, 408 (2014).
170 Seltzer & Anderson, *supra* note 82, at 495–98.
171 Richards & King, *supra* note 163, at 409–13 (explaining that because "virtually all information exists in intermediate states between completely public and completely private," shared private information can be confidential).
172 FED. TRADE COMM'N, *supra* note 166, at viii. The FTC report made four recommendations for this type of legislation. First, Congress should seek to enable consumers to easily identify data brokers. FED. TRADE COMM'N, *supra* note 166, at viii. The Commission suggests the creation of a centralized mechanism, such as an Internet portal, where data brokers can provide information about themselves, including links to opt-outs. *Id.* Second, Congress should require data brokers to clearly disclose what kind of data they are obtaining. FED. TRADE COMM'N, *supra* note 166, at viii. Third, data brokers should be required to disclose the sources of data, so that consumers are better able to determine if they need to correct their data. FED. TRADE COMM'N, *supra* note 166, at viii. Finally, Congress should require consumer-facing entities to provide prominent notice to consumers that they share consumer data with data brokers and provide consumers with options to opt out or access their information. FED. TRADE COMM'N, *supra* note 166, at viii.

data analysis software to undergo audits aimed to evaluate the entity's adherence to privacy, security and civil rights protection requirements.[173]

Second, private entities that use open data published by the government could self-regulate by adopting practices to safeguard privacy when they use open data. The FTC recommended that the data broker industry adopt several best practices, including implementing privacy-by-design, refraining from collecting information from minors, and adopting reasonable precautions to ensure that downstream users of their data do not use it for unlawful discriminatory purposes.[174] Similarly, scholars have called on statisticians and other professionals involved in the development and dissemination of population data to be sensitive to the ethical implications of their work, for example by implementing a set of standard ethical sensibilities concerning information technologies.[175]

Third, government or industries could develop technological safeguards to help prevent privacy, security, or civil rights violations through the use of published open data. For example, limiting the types of predictions and inferences entities can make based on open data can help prevent privacy invasions and unintentional discrimination.[176] Reforms could target how data brokers select the data that trains their data analysis software, for example by prohibiting the use of training data tainted by prejudice.[177]

---

173 OPEN TECHNOLOGY INSTITUTE, NEW AMERICA FOUNDATION, DATA AND DISCRIMINATION: COLLECTED ESSAYS 17–18 (Seeta Peña Gangadharan, Virginia Eubanks & Solon Barocas eds., 2014).
174 FED. TRADE COMM'N, *supra* note 166, at ix.
175 *See, e.g.*, Seltzer & Anderson, *supra* note 82, at 499; Richards & King, *supra* note 163, at 409 (arguing that the use of open data and bid data should be governed by an ethical code governed by the four normative values of privacy, confidentiality, transparency and identity).
176 *See* Richards & King, *supra* note 163, at 422–23 (arguing that considerations for identity are necessary to preserve human autonomy).
177 Barocas & Selbst, *supra* note 65, at 47–50.

Fourth, our legal system will have to develop a more nuanced approach to dealing with a new age of privacy, security and civil rights violations aided by open data. Current anti-discrimination law, namely Title VII, is not well equipped to deal with the discriminatory features of data mining.[178] For example, legal issues such as "How can you prove a discrimination case against a computer?" and "Can due process be violated if an automated decision-making system is simply running code?" may halt cases involving open data's civil rights violations.[179]

V. CONCLUSION

Open data presents opportunities to develop the economy, increase government effectiveness through information-based policies, and promote civic engagement and democratic accountability. Despite these benefits, widespread use of open data poses privacy, security, and civil rights challenges. The United States and other countries have implemented measures to deal with some of those issues, but most of the possible dangers of open data remain unmitigated.

178 Barocas & Selbst, *supra* note 65, at 23–42. Also, because the political climate and many members of the Supreme Court have moved away from supporting the elimination of status-based inequality as a purpose for antidiscrimination law, any reforms to data mining that specify a protected class, especially race, may fail a challenge to its constitutionality. Barocas & Selbst, *supra* note 65, at 54–56.

179 OPEN TECHNOLOGY INSTITUTE, *supra* note 173, at 7, 51.