2014

Disclosing Big Data

Michael Mattioli

Article

Disclosing Big Data

Michael Mattioli[†]

INTRODUCTION

This Article investigates whether intellectual property law sufficiently encourages "big data" producers to disclose how they collect, organize, and transform valuable sources of data.[1] Today, a lattice of technologies mediates our interactions with the world, automatically recording what we buy, where we go, details of our health, what we say, and to whom.[2] Left un-

[†] Associate Professor of Law, Indiana University Maurer School of Law. This Article benefitted from comments offered by Mark Janis, Marshall Leaffer, Gideon Parchomovsky, Rebecca Eisenberg, Jessica Litman, Mark McKenna, Sean Seymore, Tim Holbrook, Justin Hughes, Michael Madison, Katherine Strandburg, Brett Fischmann, Peter Lee, Jorge Contreras, Chris Seaman, Suzan Frankel, Miriam Bitton, Jason Du Mont, Lea Shaver, Jeffrey Stake, David Delaney, Jason Rantanen, Cassidy Sugimoto, Hamid Ekbia, Inna Kouper, and Brad Greenberg. This Article also owes thanks to Microsoft, Facebook, Google, DataSift, TrueLens, Treato, CancerLinQ, and the individuals at these organizations who consented to be interviewed. Copyright © 2014 by Michael Mattioli.

[1] *See generally* IAN AYRES, SUPER CRUNCHERS 60–63 (2007) (identifying this phenomenon years before the term "big data" came into vogue); STEPHEN BAKER, THE NUMERATI 98–99 (2008) (discussing the necessity of computers for gathering wide swaths of information); BILL FRANKS, TAMING THE BIG DATA TIDAL WAVE 20 (2012) (discussing big data practices from a technology-oriented perspective);VIKTOR MAYER-SCHÖNBERGER & KENNETH CUKIER, BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK (2013) (canvassing the big data phenomenon and identifying specific big data practices); ERIC SIEGEL, PREDICTIVE ANALYTICS: THE POWER TO PREDICT WHO WILL CLICK, BUY, LIE, OR DIE 2–3 (2013) (exploring the societal impact of the big data phenomenon). *See infra* note 12 and accompanying text (listing a selection of the many newspaper and magazine articles discussing the topic of big data published between 2010 and 2013).

[2] *See, e.g.*, Oswaldo Trelles et al., *Big Data, But Are We Ready?*, 12 NATURE REV. GENETICS 224, 224 (2011) (discussing big data in the context of biological research); Patrick Tucker, *Has Big Data Made Anonymity Impossible?*, MIT TECH. REV. (May 7, 2013), *available at* http://www.technologyreview.com/news/514351/has-big-data-made-anonymity-impossible (citing movie

touched, these records are valueless. Through innovative techniques of data reuse, however, experts are beginning to draw value from this raw data.[3] This relatively new phenomenon is commonly referred to as "big data," and many experts believe that it will soon lead the way to new frontiers in science and innovation.[4]

Among the many challenges that big data raises, one of the most urgent relates to data reuse. Leading commentators in the fields of informatics and computer science argue that the data fueling big data practices in many settings is inadequately documented and disclosed.[5] The nondisclosure of data's provenance and pedigree, they argue, impedes data reuse, which in turn can prevent innovative applications of the big data method.[6]

choices, locational data generated by mobile phones, and even recordings made by surveillance cameras as sources of big data); Ken Terry, *Big Data Analytics*, INFORMATIONWEEK, Mar. 1, 2013, at 8 (describing a number of big data projects designed to investigate the link between genetics and disease, including one run by Kaiser Permanente supported by a $25 million grant from the National Institutes of Health).

[3] *See* FRANKS, *supra* note 1, at 20 ("The biggest challenge with big data may not be the analytics you do with it, but the . . . processes you have to build to get it ready for analysis."); JAMES MANYIKA ET AL., MCKINSEY GLOBAL INST., BIG DATA: THE NEXT FRONTIER FOR INNOVATION, COMPETITION, AND PRODUCTIVITY 11 (2011), *available at* http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation ("Demand for deep analytical talent in the United States could be 50 to 60 percent greater than its projected supply by 2018."); *see also* Press Release, Office of Sci. & Tech. Policy, Exec. Office of the President, Obama Administration Unveils "Big Data" Initiative: Announces $200 Million In New R&D Investments 1 (Mar. 29, 2012) [hereinafter 2013 White House Press Release] *available at* http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf (announcing a substantial government investment in big data research).

[4] *See supra* note 3; *see also infra* Part I (providing background discussion on big data).

[5] *See infra* Part I.B; *see also* Christine L. Borgman, *The Conundrum of Sharing Research Data*, 63 J. AM. SOC'Y FOR INFO. SCI. & TECH. 1059, 1059–60 (2012) (discussing the fact that not much data sharing is actually taking place).

[6] *See, e.g.*, Declan Butler, *When Google Got Flu Wrong*, 494 NATURE 155, 155–56 (2013), *available at* http://www.nature.com/polopoly_fs/1.12413!/menu/main/topColumns/topLeftColumn/pdf/494155a.pdf (describing how Google Flu Trends, a leading source of flu-related information that is fueled by big data practices, has provided misleading information due to undetected biases in their practices); Quentin Hardy, *Why Big Data Is Not Truth*, N.Y. TIMES BITS BLOG (June 1, 2013, 8:00 AM), http://bits.blogs.nytimes.com/2013/06/01/why-big-data-is-not-truth ("[M]ost data sets, particularly where people are concerned, need references to the context in which they were created."); Ari Zoldan, *More Data, More Problems: Is Big Data Always Right?*,

This problem is subtle and thus requires some clarifying. Leading Computer Science and Informatics commentators are concerned with a problem beyond whether data *itself* is sufficiently disclosed, or whether big data practitioners are disclosing their methods of analyzing data.[7] The problem of most pressing concern to some commentators, rather, is the fact that, in many settings, insufficient information is made available concerning how data is initially collected and prepared.[8] Understanding where data comes from, and how it has been organized and manipulated by its stewards can be critical to its downstream reuse—the very essence of the big data method. Some commentators believe the problem of inadequate data disclosure threatens the very future of big data itself.[9] New policies geared toward encouraging the disclosure of big data practices thus appear to be normatively desirable.

Although the big data disclosure problem is not inherently an "intellectual property problem," it raises familiar concerns for intellectual property law, a primary goal of which is to encourage technological disclosure in order to speed innovation.[10] Through legal analysis and an original set of industry case

WIRED (May 10, 2013, 12:49 PM), http://www.wired.com/2013/05/more-data-more-problems-is-big-data-always-right (offering an example of how biases in data collection and preparation practices can distort research findings); *cf.* Borgman, *supra* note 5, at 1067–69 (critiquing reuse as a justification for data sharing, but recognizing the importance of data sharing for the process).

[7] *See* Omer Tene & Jules Polonetsky, *Big Data for All: Privacy and User Control in the Age of Analytics*, 11 NW. J. TECH. & INTELL. PROP. 239, 270–71 (2013) (finding that it is not always the algorithms or the accuracy of the data that requires scrutiny, but rather the factors considered and the inferences drawn from the data).

[8] *See id.*; *see also* NAT'L ACAD. OF SCI. ET AL., ENSURING THE INTEGRITY, ACCESSIBILITY, AND STEWARDSHIP OF RESEARCHING DATA IN THE DIGITAL AGE 41, 63 (2009) (discussing the importance of disclosing the steps used to generate data as well as conclusions drawn from the data, and finding that, despite the benefits of disclosure, there are instances when access to data is limited). The widespread disclosure and availability of data itself is arguably of great importance. The Author has reserved an empirical examination of this question for a future publication.

[9] *See infra* Part I.B (explaining how insufficient disclosure impedes reuse, which is considered a significant value of big data).

[10] *See* Mark A. Lemley, *The Surprising Virtues of Treating Trade Secrets As IP Rights*, 61 STAN. L. REV. 311, 332 (2008) ("Patent and copyright law do not exist solely to encourage invention, however. A second purpose——some argue the main one——is to ensure that the public receives the benefit of those inventions." (footnote omitted)); *id*. at 333 ("There is decent evidence to support the idea that at least one function of an IP right is not just to encourage new invention, but to encourage the *dissemination* of those new ideas.").

studies, this Article explains why big data practices do not fit neatly into the traditional intellectual property paradigms of patent or copyright. As a result, existing intellectual property policy does little to meaningfully encourage the disclosure of these practices. Simultaneously, a variety of forces, both legal and economic, are powerfully pushing data producers toward nondisclosure.

These conclusions prompt an inquiry: whether, as a body of law traditionally concerned with encouraging technological disclosure, intellectual property should be amended to address big data's disclosure problem.

To explore this question, this Article presents a hypothetical intellectual property based solution to big data's disclosure problem. The plan would seek to promote the disclosure of big data practices by providing data producers with a limited exclusive right in a closely-related asset—data itself. This new intellectual property construct (dubbed herein a "dataright" for convenience) would be conditioned on a data producer's full and complete disclosure of its data preparation practices. Importantly, this right would entitle data producers to block downstream *use* of data, but not reproduction or distribution. These limitations and unique aspects of the big data phenomenon distinguish this proposal from a set of database protection bills Congress has considered since the 1990s.[11] As this Article shows, however, this solution would possess significant drawbacks, suggesting that perhaps intellectual property is not the best framework to solve big data's disclosure problem. More discussion and debate are necessary.

This Article is divided into three Parts: Part I provides a primer on big data practices, and situates this new methodology within intellectual property law. This background discussion explains important characteristics of the big data phenomenon that have not been discussed in legal scholarship. Part II presents a series of original case studies gathered from interviews with experts working at the vanguard of this new field. Part III examines how intellectual property law influences the disclosure of big data practices and asks, critically, whether intellectual property offers a helpful model solution to big data's disclosure problem. By presenting a intellectual property based solution as an exploratory device rather than a formal legisla-

[11] *Infra* note 201 (listing relevant bills considered by Congress).

tive proposal, this Article aims to initiate a much-needed policy debate. A brief conclusion follows.

I. SITUATING BIG DATA WITHIN INTELLECTUAL PROPERTY LAW

Part practice and part philosophy, big data has been the subject of myriad American newspaper articles, op-eds, magazine features, and books published since 2010.[12] Despite its ever-growing popularity, however, the big data phenomenon is widely misunderstood.[13] This Part defines big data and explains why this emerging phenomenon raises important questions for intellectual property policy—a relationship that commentators have not yet explored. This background discussion frames a pressing policy question: does intellectual property law adequately encourage the disclosure of big data practices?

A. DEFINING BIG DATA

The term, "big data," refers to a new method of empirical inquiry.[14] This method consists of certain practices that become more useful as electronic data recorded from devices and ser-

[12] Based on a LexisNexis Academic search, in the year 2013, leading U.S. newspapers including *The New York Times*, *The Wall Street Journal*, and *USA Today* published 637 articles and opinion pieces on the subject of big data. *See, e.g.*, L. Gordon Crovitz, *Why 'Big Data' Is a Big Deal*, WALL ST. J., Mar. 25, 2013, at A15; Chuck Raasch, *'It Powers My Life,'* USA TODAY, Dec. 13, 2012, at 1A; Alexandra Stevenson, *Big Data Fund*, N.Y. TIMES, Oct. 18, 2013, at B5. Likewise, leading U.S. periodicals have published in-depth cover stories on the big data phenomenon. *See, e.g.*, *Data, Data Everywhere*, THE ECONOMIST, Feb. 27, 2010, at 3 (canvassing the broad promise and potential of big data); Alissa Quart, Cover Story, *The Body-Data Craze*, NEWSWEEK, June 26, 2013 (exploring personal fitness tracking devices as a rapidly growing source of big data); Michael Specter, *Climate by Numbers*, THE NEW YORKER, Nov. 11, 2013, at 38 (describing power of big data in the agricultural industry).

[13] *See, e.g.*, Luciano Floridi, *Big Data and Their Epistemological Challenge*, 25 PHIL. & TECH. 435, 436 (2012) *available at* http://link.springer.com/content/pdf/10.1007%2Fs13347-012-0093-4.pdf (reporting that the term "big data" is poorly defined); Karen E.C. Levy, *Relational Big Data*, 66 STAN. L. REV. ONLINE 73, 73 n.3 (Sept. 3, 2013), http://www.stanfordlawreview.org/sites/default/files/online/topics/66_StanLRevOnline_73_Levy.pdf (noting the "slipperiness" of the term's meaning, and explaining that big data describes a phenomenon that entails a set of practices performed on data resources); MAYER-SCHÖNBERGER & CUKIER, *supra* note 1, at 6 ("There is no rigorous definition of big data.").

[14] *See, e.g.*, MAYER-SCHÖNBERGER & CUKIER, *supra* note 1, at 6 (explaining that big data refers to a method of "extract[ing] new insights").

vices grows.[15] Today, experts in academia, government, and private industry are using the big data method to improve the quality of medical treatment, to cultivate more robust crops, to increase the efficiency of the national electrical grid, to improve the flow of traffic on highways, and to predict the flow of financial transactions across the globe.[16] Popular wisdom in technology circles holds that no avenue of human endeavor will not soon be touched and transformed by this new technique.[17]

To understand the big data method in practical terms, it is helpful to consider a brief example: In 2010, researchers at Stanford, Columbia, and Microsoft Corporation developed a new way to predict harmful interactions between pharmaceuticals.[18] In a break from traditional methods of predicting the interplay between drugs (e.g., studying chemical interactions and human physiology),[19] the group relied on an unlikely resource: the Internet. In cooperation with Microsoft, the researchers analyzed logs of millions of online searches made by consenting users of the Google, Bing, and Yahoo! search engines.[20] Using statistical techniques, they observed that users who searched for the names of two drugs—Paxil and Pravastatin—were like-

[15] *See id*. ("One way to think about the issue . . . is this: big data refers to things one can do at a large scale that cannot be done at a smaller one . . . .").

[16] *See, e.g.*, Tene & Polonetsky, *supra* note 7, at 243–51 (describing a number of domains in which the big data method is being used, including healthcare, mobile communications, energy, traffic management, retail, and online commerce).

[17] MAYER-SCHÖNBERGER & CUKIER, *supra* note 1, at 6.

[18] Ryen W. White et al., *Web-Scale Pharmacovigilance: Listening to Signals from the Crowd*, 20 J. AM. MED. INFORMATICS ASS'N 404 (2013), *available at* http://jamia.bmj.com/content/20/3/404.full.pdf; *see also* Stanford Ctr. for Internet & Soc'y, *The Privacy Paradox—Health and Medical Privacy*, YOUTUBE (Feb. 27, 2012, 00:32:24), http://www.youtube.com/watch?v=ntL4WMGkiXo [hereinafter Altman] (depicting Dr. Altman describing his process).

[19] *See, e.g.*, Nicholas P. Tatonetti et al., *Data-Driven Prediction of Drug Effects and Interactions*, 4 SCI. TRANSLATIONAL MED., Mar. 14, 2012, at 1, 1–3, *available at* http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3382018/pdf/nihms-373483.pdf (describing existing methods of predicting drug-drug interactions through the study of protein structure and chemical composition); *see also* CTR. FOR DRUG EVALUATION & RESEARCH, FOOD & DRUG ADMIN., GUIDANCE FOR INDUSTRY: DRUG INTERACTION STUDIES—STUDY DESIGN, DATA ANALYSIS, IMPLICATIONS FOR DOSING, AND LABELING RECOMMENDATIONS 2 (Feb. 2012), *available at* http://www.fda.gov/downloads/Drugs/GuidanceCompliance RegulatoryInformation/Guidances/ucm292362.pdf (recommending in vitro testing followed by clinical trials to test drug-drug interactions).

[20] Altman, *supra* note 18; White et al., *supra* note 18, at 1.

ly to also enter search terms related to hypoglycemia.[21] This correlation led the researchers to hypothesize, and later to experimentally confirm, that Paxil and Pravastatin can cause adverse side effects when taken together.[22]

The Stanford drug study has been widely cited by commentators because it demonstrates a characteristic that sets big data apart from traditional methods of empirical study: big data draws insights from records gathered automatically and indiscriminately a priori.[23] Since the dawn of the scientific method, researchers have typically studied the world by first articulating questions and hypotheses and only *later* collecting empirical evidence.[24] The big data method turns this process on its head by asking new questions of old data.[25] This new kind of empiricism is made possible by the vast tapestry of electronic devices and services that automatically record information about daily life in the developed world.[26] Internet search histo-

[21] Altman, *supra* note 18; White et al., *supra* note 18, at 1. Specifics on the statistical methods used are described in a recent publication. Bethany Percha et al., *Discovery and Explanation of Drug-Drug Interactions Via Text Mining*, 17 PAC. SYMP. ON BIOCOMPUTING 410, 411–13 (2012).

[22] Altman, *supra* note 18; White et al., *supra* note 18, at 1.

[23] *See, e.g.*, FRANKS, *supra* note 1, at 20–21 ("Traditional structured data doesn't require as much effort in these areas since it is specified, understood, and standardized in advance. With big data, it is necessary to specify, understand, and standardize it as part of the analysis process in many cases."); *id.* at 209 ("The fact is that data is never, ever as clean as they want it to be, and it is often not as clean as it really needs to be."); MAYER-SCHÖNBERGER & CUKIER, *supra* note 1, at 45 ("Conventional, so-called relational, databases are designed for a world . . . in which the questions one wants to answer using the data have to be clear at the outset, so that the database is designed to answer them—and only them—efficiently."). In their book, *"Raw Data" Is an Oxymoron*, Lisa Gitelman and Virginia Jackson similarly observed that "data are always already 'cooked' and never entirely 'raw.'" Lisa Gitelman & Virginia Jackson, *Introduction*, *in* "RAW DATA" IS AN OXYMORON 2 (Lisa Gitelman ed., 2013).

[24] *See, e.g.*, HUGH G. GAUCH, JR., SCIENTIFIC METHOD IN BRIEF 57 (2012) ("Observation is always selective. It needs a chosen object, a definite task . . . a point of view, a problem.").

[25] *See, e.g.*, MAYER-SCHÖNBERGER & CUKIER, *supra* note 1, at 44–45 ("[T]he questions we want to ask [about big data] often emerge only when we collect and work with the data we have."). *But see, e.g.*, GAUCH, *supra* note 24, at 57.

[26] *See* Tucker, *supra* note 2 (citing movie choices, locational data generated by mobile phones, and even recordings made by surveillance cameras as sources of big data); Martin White, *Big Data—Big Challenges*, ECONTENT (Nov. 9, 2011), http://www.econtentmag.com/Articles/Column/Eureka/Big-Data-Big-Challenges-78530.htm ("Big Data extends beyond structured data and

ries, social media connections and posts, and credit card records are chief sources, as are network-connected sensors in smartphones, personal health devices, automobiles, and home appliances.[27] Added to the mix is a flood of clinical and genetic data generated by healthcare providers.[28]

Because many applications of the big data method draw upon information that describes intimate details of our lives, it is not surprising that legal commentary on the subject has, to date, focused on the theme of privacy. In a landmark publication on the subject, Paul Ohm explored the troubling fact that in settings where data about individuals can be aggregated from multiple sources, anonymity can never be completely guaranteed.[29] Ohm discussed the big data privacy problem further in a 2013 essay, in which he noted that big data could allow governments and corporations to more easily spy on, and possibly even discriminate against private individuals.[30] Not all privacy scholars are as concerned, however: Omer Tene and Jules Polonetsky have advocated a loosening of privacy regulations in order to unleash the full power of big data for economic and social growth.[31]

As legal scholars continue to debate the appropriate policy responses to big data's privacy implications, they appear to agree that big data has a profound potential to foster innovation. Tene and Polonetsky identify a set of industries likely to benefit from big data: healthcare, electrical power distribution, mobile communications, traffic management, retail, payments,

includes unstructured data of all varieties: text, audio, video, click streams, log files, and more.").

[27] Tucker, *supra* note 2.

[28] *See* Terry, *supra* note 2, at 8 (describing big data projects designed to investigate the link between genetics and disease, including one run by Kaiser Permanente supported by a $25 million grant from the National Institutes of Health); Trelles et al., *supra* note 2, at 224 (discussing big data in the context of genetic research).

[29] Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. Rev. 1701, 1703–05 (2010).

[30] Paul Ohm, Response, *The Underwhelming Benefits of Big Data*, 161 U. PA. L. REV. ONLINE 339, 340 (2013), http://www.pennlawreview.com/online/161-U-Pa-L-Rev-Online-339.pdf.

[31] Tene & Polonetsky, *supra* note 7, at 264; Omer Tene & Jules Polonetsky, *Privacy in the Age of Big Data: A Time for Big Decisions*, 64 STAN. L. REV. ONLINE 63, 64–65 (2012), http://www.stanfordlawreview.org/sites/default/files/online/topics/64-SLRO-63_1.pdf.

and myriad online products and services.[32] Even Ohm, who argues that enthusiasm for big data should be tempered, unequivocally asserts that big data will deliver important technological benefits: "Whether applied to crises in medicine, in climate, in food safety, or in some other arena," Ohm writes, "Big Data techniques will lead to significant, new, life-enhancing (even life-saving) benefits that we would be ill advised to electively forego."[33] Leading commentators from the fields of computer science and informatics share the view that big data will (perhaps inevitably) spur important new innovations.[34]

The Obama Administration has also recognized big data's potential to stimulate technological progress. In March, 2012, the Administration announced that six federal departments and agencies would commit over $200 million to advance the state of the art in the field.[35] These commitments included research grants offered by the National Science Foundation, and a variety of new initiatives within the Department of Defense (autonomous robotics), the National Institutes of Health (genetic data studies), and the Department of Energy (data visualizations).[36] On May 1, 2014, the Executive Office of the President published a report presenting a detailed picture of how big data has already influenced society, how it will likely steer future innovation, and the policy challenges it presents. Echoing legal commentary on the subject, the report concluded, "Big data technologies are driving enormous innovation while raising novel privacy implications."[37]

Big data is a powerful new method of understanding the world around us. Like earlier technologies that have shed light on the human experience, such as photography, it holds a high potential to promote technological change and also a conspicuous set of concerns for individual privacy. The primacy of these privacy concerns in legal discourse has overshadowed a second

[32] Tene & Polonetsky, *supra* note 16, at 243–51; Tene & Polonetsky, *supra* note 31, at 64–65.

[33] Ohm, *supra* note 30, at 339–40.

[34] *See, e.g.*, MAYER-SCHÖNBERGER & CUKIER, *supra* note 1, at 5 (discussing how data can be reused to promote innovation).

[35] 2013 White House Press Release, *supra* note 3, at 1.

[36] *Id.* at 2–3.

[37] EXEC. OFFICE OF THE PRESIDENT, BIG DATA: SEIZING OPPORTUNITIES, PRESERVING VALUES 61 (May 2014), *available at* http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_5.1.14_final_print.pdf.

and equally important challenge, however: fostering an industrial and scientific landscape in which our society's creators can put the big data method to innovative new uses that enhance social welfare. As Part I.B explains, this goal is barred by a roadblock that perhaps only policymakers can remove.

B. THE BIG DATA DISCLOSURE PROBLEM

Despite the bold expectations surrounding big data, the phenomenon's full potential remains largely unrealized. Books, magazine articles, and academic journals frequently cite a small set of anecdotes that demonstrate the phenomenon's power—including the Stanford drug study mentioned earlier—but these examples are isolated experiments rather than evidence of widespread industrial and scientific activity.[38] The frequency with which the same anecdotes are repeated in the literature seems to underline this conclusion. This raises a puzzling question: Why has big data not yet delivered the big innovations that commentators predict?

According to technology experts, the answer lies in the challenges of data reuse.[39] Much of the rhetoric describing big data's potential for innovation assumes that data can be easily and meaningfully reused and recombined in order to examine new questions.[40] As Christine Borgman of UCLA explains, "If

[38] *See* White et al., *supra* note 18. Other examples of frequently cited big data anecdotes include: using airline data to predict airfare and flight arrival times, *cf.* Oren Etzioni et al., *To Buy or Not To Buy: Mining Airfare Data To Minimize Ticket Purchase Price*, 9 ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY & DATA MINING 119–28 (2003), *available at* http://www.cis.temple.edu/~yates/papers/hamlet-kdd03.pdf (using airline data to predict airfare), an anecdote describing Target's use of customer data to impute when a customer may be pregnant, *cf.* Charles Duhigg, *How Companies Learn Your Secrets*, N.Y. TIMES (Feb. 16, 2012), http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html (describing Target's use of customer data to impute when a customer may be pregnant), Google's process of tracking influenza, *cf. Flu Trends: How Does This Work*, GOOGLE.ORG, http://www.google.org/flutrends/about/how.html (last visited Oct. 29, 2014) (tracking influenza based on linguistic data found in user searches), and predictive policing, *cf. Predictive Policing*, NAT'L INST. JUST., http://www.nij.gov/topics/law-enforcement/strategies/predictive-policing/pages/welcome.aspx (last modified June 9, 2014) (drawing on crime-related data to predict when and where crimes are most likely to occur).

[39] *See* Borgman, *supra* note 5, at 1059 ("The 'dirty little secret' behind the promotion of data sharing is that not much sharing may be taking place.").

[40] Data reuse is a central theme at big data conferences and symposia. *See, e.g.*, *Programme with Presentations*, DIGITAL CURATION CONF., http://

the rewards of the data deluge are to be reaped, then researchers who produce those data must share them . . . in such a way that the data are interpretable and reusable by others."[41] In a similar vein, Limor Peer, Ann Green, and Libbie Stephenson write, "The idea that the data will be used by unspecified people, in unspecified ways, at unspecified time[s] . . . is thought to have broad benefits."[42] In their 2013 book surveying the big data phenomenon, Viktor Mayer-Schöenberger and Kenneth Cukier explain that the potential for data reuse is the central source of value in the big data method. "In a big-data world," they write, "[d]ata's value shifts from its primary use to its potential future uses."[43]

In reality, however, substantial impediments prevent data from being easily reused. One set of challenges is purely technical: because data is often recorded and published in a wide variety of formats, researchers have difficulty aggregating data from multiple sources.[44] This problem will likely be overcome in time. The federal government and a number of international standard-setting organizations are already developing and encouraging the use of standard formats for data in order to enable big data aggregation. The U.S. National Institute of Standards and Technology (NIST), for instance, assembled a working group on big data in 2013 that aims to develop a common set of

www.dcc.ac.uk/events/idcc14/programme-presentations (last visited Oct. 29, 2014).

[41] Borgman, *supra* note 5, at 1059.

[42] Limor Peer, *Mind the Gap in Data Reuse: Sharing Data Is Necessary But Not Sufficient for Future Reuse*, LONDON SCH. ECON. & POLI. SCI. (Mar. 28, 2014), http://blogs.lse.ac.uk/impactofsocialsciences/2014/03/28/mind-the-gap-in-data-reuse.

[43] MAYER-SCHÖNBERGER & CUKIER, *supra* note 1, at 99; *see also id.* at 147–48 ("More likely, we'll see the advent of new firms that pool data from many consumers, provide an easy way to license it, and automate the transactions."); AYRES, *supra* note 1, at 61–62 ("Businesses realize that information has value. Your databases not only help you make better decisions, database information is a commodity that can be sold to others. So it's natural that firms are keeping better track of what they and their customers are doing.").

[44] *See* Michael J. Madison, *Commons at the Intersection of Peer Productions, Citizen Science, and Big Data: Galaxy Zoo*, *in* GOVERNING KNOWLEDGE COMMONS 209 (Brett M. Frischmann et al. eds. 2014); *see also* Borgman, *supra* note 5, at 1070 ("Indeed, the greatest advantages of data sharing may be in the combination of data from multiple sources, compared or 'mashed up' in innovative ways." (citing D. Butler, *Mashups Mix Data Into Global Service: Is This the Future for Scientific Analysis?*, 439 NATURE 6 (2006))).

big data definitions, taxonomies, and reference architectures.[45] The International Standards Organization and the W3 have assembled similar groups to explore the adoption of standard formats.[46] In addition to the development of standards, machine learning systems such as IBM's "Watson" are becoming ever more adept at extracting useful data from unstructured sources of information, such as medical journal articles.[47]

A second barrier to widespread data reuse is at once more subtle and more challenging. Data is often deeply infused with the subjective judgments of those who collect and organize it.[48] As Danah Boyd, a leading big data commentator, explains, "[W]orking with Big Data is . . . subjective, and what it quantifies does not necessarily have a closer claim on objective truth . . . ."[49] Kate Crawford, another leading voice in this emerging field, recently wrote, "Hidden biases in . . . [the] analysis stages present considerable risks, and are as important to the big-data equation as the numbers themselves."[50]

These commonly embedded judgments present a problem for data reuse.[51] As Christine Borgman explains, "Reusers of data may not know, or be able to know, what prior actors did to

[45] *See NIST Big Data*, NAT'L INST. STANDARDS & TECH. (June 7, 2013), http://bigdatawg.nist.gov.

[46] *See* Keith Hare, *Report of Study Group on Next Generation Analytics and Big Data*, FARANCE INC. (June 5, 2013), http://www.jtc1sc32.org/doc/N2351-2400/32N2388b-report_SG_big_data_analytics.pdf; *Customer Experience Digital Data Community Group*, W3C, http://www.w3.org/community/custexpdata (last visited Oct. 29, 2014). The "W3C" is the main international standards-setting organization for the World Wide Web.

[47] *See, e.g.*, Ajay Royyuru, *IBM's Watson Takes on Brain Cancer*, IBM RES., http://www.research.ibm.com/articles/genomics.shtml (last visited Oct. 29, 2014).

[48] *See, e.g.*, NAT'L ACAD. OF SCI. ET AL., *supra* note 8, at 34 ("Because digital data can be manipulated more easily than can other forms of data, digital data are particularly susceptible to distortion. Researchers—and others—may be tempted to distort data in a misguided effort to clarify results. In the worst cases, they may even falsify or fabricate data.").

[49] Danah Boyd & Kate Crawford, *Six Provocations for Big Data* 4 (Sept. 21, 2011), http://ssrn.com/abstract=1926431; *see also* Hardy, *supra* note 6 (discussing Crawford's views further).

[50] Kate Crawford, *The Hidden Biases of Big Data*, HARV. BUS. REV. BLOG (Apr. 1, 2013, 2:00 PM), http://blogs.hbr.org/2013/04/the-hidden-biases-in-big-data.

[51] *See, e.g.*, Nick Bilton, *Disruptions: Data Without Context Tells a Misleading Story*, N.Y. TIMES BITS BLOG (Feb. 24, 2013, 11:00 AM), http://bits.blogs.nytimes.com/2013/02/24/disruptions-google-flu-trends-shows-problems-of-big-data-without-context.

the data. Each step in . . . processing data requires judgments, few of which may be fully documented. Later interpretations thus may depend upon multilevel inferences that are statistically problematic."[52] Peer similarly comments, "[I]t is often difficult to interpret and make use of the data . . . when you don't understand how the data were generated."[53] A senior technical specialist at Microsoft Corporation interviewed for this Article echoed these statements. "It's essential to be transparent about not only the source of the data," he explained, "but [also] the method used to gather it, any changes to it, and the basis for any decisions made about the source. In fact, without that, I wouldn't trust the data at all."[54]

Speaking at a conference in 2013, Kate Crawford presented a vivid example of how data devoid of context cannot be meaningfully reused or put to new purposes. When a powerful hurricane struck the East Coast in 2012, the largest number of status updates published online originated from urban areas with high numbers of social media users, rather than from locations where the storm had actually struck.[55] Crawford described how a hypothetical database could be created that included every online update that mentioned the name of the hurricane. A future researcher relying on this database alone to study the storm's progress could incorrectly guess that it hit regions exclusively populated by technology-savvy professionals.[56] If the same researcher knew the method by which the database had been built, however—a search for every message published online that referred to the storm's name—then she might be able to avoid this faulty conclusion. Data devoid of context can also be devoid of meaning.

In some academic research settings, institutional norms mandate disclosure of how data has been collected and prepared. Leading scientific and economic journals, for instance,

[52] Borgman, *supra* note 5, at 1067.

[53] Peer, *supra* note 42.

[54] E-mail from Buck Woody, Senior Technical Specialist, Microsoft Corp., to author (July 7, 2014) (on file with author).

[55] *See* Hardy, *supra* note 6 (concluding from this episode as relayed by Crawford that "most data sets, particularly where people are concerned, need references to the context in which they were created").

[56] *See id.* (quoting Crawford's comment that these were "privileged urban stories"); *see also* Zoldan, *supra* note 6 ("The majority of the tweets originated from Manhattan, largely because of the high concentration of smartphone and Twitter usage.").

require authors to submit information about their data sources and detailed descriptions of specific techniques they used to prepare the data for study.[57] Federal agencies that fund scientific research such as the National Institutes of Health similarly require grant recipients to disclose their sources of data, and their data-preparation practices.[58]

One might wonder why market forces should not be expected to encourage similar disclosures of industrial and commercial data. If a data producer consistently releases undocumented data, after all, one might expect that the company would develop a poor reputation and that consumers would turn to more reliable publishers. This view misunderstands the commercial context in which big data has developed. The devices and services fueling this phenomenon are provided by companies for whom data *itself* is typically a byproduct, rather than a source, of business.[59] Providers of search engines, mobile phones, health devices, public utilities, and other primary big data sources have little or no impetus to disclose their methods of data collection and preparation because there is not, as yet, a commercial market for such abstract information.[60] Big data represents a secondary, and largely speculative, public value

[57] *See, e.g.*, Editorial, *Social Software*, 4 NATURE METHODS 189 (2007), *available at* http://www.nature.com/nmeth/journal/v4/n3/full/nmeth0307-189.html (last visited Oct. 29, 2014) (requiring that authors submit all algorithms, software, and related data); *Availability of Data and Materials*, NATURE.COM, http://www.nature.com/authors/policies/availability.html (last visited Oct. 29, 2014) ("[A] condition of publication in a Nature journal is that authors are required to make materials, data and associated protocols promptly available to readers without undue qualifications."); *The* American Economic Review*: Data Availability Policy*, AM. ECON. ASS'N, http://www.aeaweb.org/aer/data.php (last visited Oct. 29, 2014) (requiring that authors include all datasets and descriptions of how intermediate datasets were made, as well as citing software used).

[58] *See NIH Data Sharing Policy and Implementation Guidance*, NAT'L INST. HEALTH, https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm (last updated Mar. 5, 2003) ("Documentation provides information about the methodology and procedures used to collect the data, details about codes, definitions of variables . . . and the like.").

[59] *See* AYRES, *supra* note 1, at 60 ("All too often information management is limited to historical data, to recent and not-so-recent information about past transactions. Business is now very good at tracking these kinds of data, but businesses as a group still have not gone far enough in proactively creating useful new data.").

[60] *Cf.* Boyd & Crawford, *supra* note 49, at 6–7 (discussing what little data Twitter releases to researchers and the problems resulting from lack of disclosure of storage methods).

that resides far downstream from the commercial exchanges that take place between data producers and their customers.[61]

Beyond the lack of any affirmative economic incentives to disclose their practices, big data producers may face strong *disincentives* to disclose. Privacy regulations, for instance, might discourage institutions that collect and transmit medical records from conveying information about their anonymization practices that could be used to identify patients.[62] One can imagine that competitive concerns might also discourage disclosure of data preparation methods. A health device manufacturer, for instance, might not want its customers or competitors to learn of shortcomings or errors in the data that its devices produce. Likewise, it is unlikely that big data producers would want to disclose information that reveals weaknesses in their methodologies—i.e., low quality data. Finally, some publishers of big data may view their methods of data preparation as valuable trade secrets that provide a competitive advantage.

The grand vision of big data as an engine for innovation relies on the assumption that data can be reused, combined, and repurposed. As this Part has explained, however, technical, commercial, and epistemological roadblocks render this assumption faulty. Most significantly, big data's producers tend to infuse their products with subjective judgments that, when left undisclosed, limit the data's potential for future reuse.

C. HOW INTELLECTUAL PROPERTY LAW INFLUENCES DISCLOSURE

A central goal of American intellectual property law is to spur innovation by encouraging technological disclosures.[63] Big data's disclosure problem suggests that intellectual property law is not meeting this goal in an important new technological field. To assess this hypothesis, it is first necessary to consider how the law applies to big data practices. Thankfully, although big data is new, it is not *so* new that it cannot be situated within the existing intellectual property framework. In fact, longstanding intellectual property debates pertaining to soft-

[61] *See* Borgman, *supra* note 5, at 1071 ("This . . . rationale, to enable others to ask new questions of extant data, benefits prospective users more than producers of data."). Thanks to Lea Shaver for helping me phrase this explanation.

[62] *See, e.g.*, *id.* at 1072 (citing similar factors).

[63] *See* Lemley, *supra* note 10, at 333.

ware, algorithms, and databases are directly relevant to big data. The following overview of the law's relationship to such subject matter lays the groundwork for examining this Article's original case studies.[64]

Vendors of information-based products have long secured exclusivity in their processes and knowhow through the law of trade secrets.[65] The Uniform Trade Secrets Act (UTSA), which has been adopted by most states, defines trade secrets as "information" that is (i) valuable, and (ii) reasonably protected.[66] The definition of "information" under the UTSA is expansive, covering technical and non-technical information, including methods, knowhow, and even ideas.[67] Importantly, information need not be *absolutely* secret to merit trade secret protection; it must only be the subject of reasonable efforts to prevent disclosure.[68] Remedies for trade secret misappropriation can include a range of monetary damages as well as injunctive relief.[69]

Information-based processes that are not readily perceived by consumers are particularly well suited for trade secret protection. Google's well-known "PageRank" algorithm, and the al-

[64] This Part explores how intellectual property law might apply to the *data preparation methods* at the root of big data's disclosure problem. The law's relationships to data itself or to methods that draw meaning from data (i.e. analytics) are ancillary to, and for the most part, outside the scope of this discussion.

[65] *See* Peter S. Menell, *The Challenges of Reforming Intellectual Property Protection for Computer Software*, 94 COLUM. L. REV. 2644, 2652 (1994) ("The [software] industry had developed principally through trade secret protection."); Mark A. Lemley & David W. O'Brien, *Encouraging Software Reuse*, 49 STAN. L. REV. 255, 258 (1997) ("Trade secret law remained the dominant form of legal protection of software through the mid-1970s.").

[66] *See* UNIF. TRADE SECRETS ACT § 1(4), 14 U.L.A 538 (2005); *see also* RESTATEMENT (THIRD) OF UNFAIR COMPETITION § 39 (1995) ("A trade secret is any information that can be used in the operation of a business or other enterprise and that is sufficiently valuable and secret to afford an actual or potential economic advantage over others.").

[67] Vincent Chiappetta, *Myth, Chameleon or Intellectual Property Olympian? A Normative Framework Supporting Trade Secret Law*, 8 GEO. MASON L. REV. 69, 76 (1999) ("Trade secret law . . . extends to technical and non-technical information, expression, ideas and facts, embracing such things as customer and supplier lists, financial information, methods of doing business, future marketing, sales and product plans and even employee names, job responsibilities and phone numbers.").

[68] *See, e.g.*, Lemley, *supra* note 10, at 317 (discussing the requirement that "the holder of the trade secret, took reasonable precautions under the circumstances to prevent its disclosure").

[69] *See* RESTATEMENT (THIRD) OF UNFAIR COMPETITION §§ 44–45.

gorithms used by high-speed electronic trading firms are two well-documented examples.[70] Source code—the instructions that software developers compose and which consumers cannot view—is also commonly protected through trade secrecy.[71] Pamela Samuelson has observed that trade secrecy in the software industry may also extend to "industrial techniques of a practical nature that [are] often the fruit of . . . experience and trial and error."[72]

Trade secret law was at the heart of an academic debate concerning software in the 1990s that has bearing on the big data disclosure problem. At that time, leading intellectual property scholars argued that, by discouraging the disclosure of source code and related practices, trade secret law would slow the pace of software innovation. Robert G. Bone, for instance, cautioned that trade secrecy would lead to wasteful duplicative efforts among software engineers working at different firms.[73] Pamela Samuelson cited a second significant cost: secrets are sometimes expensive to keep.[74] Drawing on these insights,

[70] *See* VAN LINDBERG, INTELLECTUAL PROPERTY AND OPEN SOURCE: A PRACTICAL GUIDE TO PROTECTING CODE 130–31 (2008) (discussing Google's use of trade secrecy); Indictment at 1–4, United States v. Aleynikov, 10 Crim. 96 (S.D.N.Y. Feb. 11, 2010), 2010 WL 4000356 (describing steps that Goldman Sachs & Co. used to maintain trade secret rights in their high-speed trading algorithms).

[71] *See* Data Gen. Corp. v. Digital Computer Controls, Inc., 357 A.2d 105, 112–13 (Del. Ch. 1975) (holding that the contents of a computer program distributed only in object code format were protectable trade secrets); *see also* Lemley, *supra* note 10, at 325 ("They are free to market products incorporating the secret, and to disclose the secret itself to others in the service of making money."); Wendy Seltzer, *Software Patents and/or Software Development*, 78 BROOK. L. REV. 929, 981 (2013) (discussing the advantages of software companies using trade secrecy to protect their inventions and technology).

[72] Pamela Samuelson et al., *A Manifesto Concerning the Legal Protection of Computer Programs*, 94 COLUM. L. REV. 2308, 2329 (1994) (observing that trade secrecy in the software industry extends to "the totality of unpatented knowledge utilized in industry" (citation omitted)).

[73] *See* Robert G. Bone, *A New Look at Trade Secret Law: Doctrine in Search of Justification*, 86 CAL. L. REV. 241, 266–67 (1998) ("[B]ecause trade secret law permits independent invention—and even gives the second inventor protection—firms will continue to seek the same invention, thereby wastefully duplicating the efforts of the first inventor."); *see also* Chiappetta, *supra* note 67, at 89–90 ("[T]here are significant reasons to suspect that the incremental encouragement offered by trade secret law does not outweigh its costs.").

[74] *See* Samuelson et al., *supra* note 72, at 2409 ("Substantial societal costs are incurred when program know-how is kept as a trade secret. Some of these arise from the costs of maintaining secrecy; others derive from the ex-

commentators during this period warned that widespread trade secrecy would reduce the rate of cumulative innovation in the software industry.[75]

Some scholars saw a silver lining, however, in the fact that software methods can sometimes be reverse-engineered. Jerome H. Reichman argued that trade secrecy was not an absolute bar to the dissemination of knowhow in the software industry because reverse engineering is permitted by the law, and often easy to perform on object code.[76] Mark Lemley identified a second potential benefit of trade secrecy: the availability of trade secret protection in the software industry, Lemley argued, could *encourage* the dissemination of information by compelling innovators to invest less in building physical barriers—e.g., encryption—to their secrets.[77]

The holders of industrial secrets that are *particularly* easy to keep might, of course, elect to forgo legal protection altogether and instead simply not document or disclose their methods. This strategy might, in some cases, be preferable to taking the affirmative (and more costly) steps necessary to maintain trade secret protection.

Like algorithms, many big data practices likely fit within trade secret law's expansive definition of "information."[78] Because such practices are typically implemented through software, a big data producer could also obtain trade secret protec-

penditures directed at reverse engineering or engaging in other efforts to duplicate or independently recreate the know-how.").

[75] *See, e.g.*, Chiappetta, *supra* note 67, at 89 ("Finally, there is no requirement of public disclosure, meaning no education of competitors . . . and no affirmative dedication to the public."); Lemley & O'Brien, *supra* note 65, at 276 ("Progress in computer science, the useful arts, and programming, as in other fields, depends on the ability of innovators and researchers to build up on earlier advances.").

[76] *See* J.H. Reichman, *Computer Programs as Applied Scientific Know-How: Implications of Copyright Protection for Commercialized University Research*, 42 VAND. L. REV. 639, 701 (1989) ("A would-be competitor who is denied access to the originator's source code may nonetheless reconstruct a skeletal version of it by using special computer programs to decompile and reverse engineer the object code . . . .").

[77] *See* Lemley, *supra* note 10, at 333–34 ("Paradoxically, however, trade secret law actually encourages broader disclosure and use of information, not secrecy. It does so in two ways. First, the legal protection trade secret law provides serves as a *substitute* for investments in physical secrecy that companies might otherwise make.").

[78] UNIF. TRADE SECRETS ACT § 1, 14 U.L.A. 538 (1985) ("'Trade secret' means information, including a formula, pattern, compilation, program, device, method, technique, or process.").

tion over the code that assists experts in carrying out these practices.[79] Moreover, from a practical perspective, secrecy over such information may be even easier to maintain than secrecy over software methods. The recent commentary describing big data's disclosure problem indicates that, unlike software, big data practices cannot be reverse-engineered.[80] That is, an expert cannot decipher just how a set of data was assembled with nothing more to work from than the data itself. As a result, the academic arguments that trade secrecy may sometimes promote disclosure of software methods seem inapplicable to big data practices.

Theoretically, patent law might push the developers of some big data practices toward public disclosure.[81] All patent applications must contain a detailed written description on the invention claimed, which The United States Patent and Trademark Office (USPTO) publishes eighteen months after the date that an application is filed.[82] In return for granting their knowledge to the public, patentees receive a far more robust form of protection than trade secret holders enjoy: the ability to enjoin any unauthorized use, manufacture, sale, or importation of their innovations for twenty years.[83]

Despite its advantages, patent protection extends to a narrower set of processes and methods than trade secrecy. Algorithms that amount to abstract ideas, for instance, do not meet

[79] For information on trade secret protection, see *id. See also* Wellogix, Inc. v. Accenture, LLP, 716 F.3d 867 (5th Cir. 2013) (finding software developer's source code contained trade secrets).

[80] *See* Ohm, *supra* note 29, at 1711 (mentioning and citing to a number of legal scholars placing faith in the power of anonymization through big data processes). *But see id.* at 1716–27 (noting the potential for reversing data anonymization techniques).

[81] *See generally*, Patent Act, 35 U.S.C. § 101 (2012) ("Whoever invents or discovers any new and useful process . . . may obtain a patent therefor, subject to the conditions and requirements of this title."); State St. Bank & Trust Co. v. Signature Fin. Grp., Inc., 149 F.3d 1368 (Fed. Cir. 1998) (action against assignee of patent for computerized accounting system used to manage mutual fund investment structure), *abrogated by In re* Bilski, 545 F.3d 943 (Fed. Cir. 2008) (patent application for method of hedging risk in field of commodities trading); CLS Bank Int'l v. Alice Corp., 717 F.3d 1269 (Fed. Cir. 2013) (suit concerning infringement and validity of patents generally directed to methods or systems that help lessen settlement risk of trades of financial instruments using a computer system).

[82] Patent Act, 35 U.S.C. § 112 (2012) (requiring disclosure sufficient to permit an individual skilled in the art to make and use the invention).

[83] *Id.*

the threshold eligibility requirements for patent protection.[84] Only processes that are novel, non-obvious, and useful may be eligible for patent protection.[85] A number of statutory bars, such as the sale or prior public use of an invention long before the date a patent is applied for, may also lead the USPTO to reject a patent application.[86] A final limitation on patentability possibly relevant to big data is patent law's requirement of definiteness. Patent claims—the so-called "metes and bounds" of patent protection—must be written in sufficiently definite terms.[87] This rule has led the Federal Circuit to invalidate patents claiming processes that rely on subjective judgments.[88] In the 2005 decision of *Datamize LLC v. Plumtree Software Inc.*, for example, the court determined that patent claims that relied on the subjective opinion of a person performing the claimed invention failed for indefiniteness.[89] Claim terms that involve, but do not rely entirely upon, subjective judgment may be sufficiently definite, however.[90]

While big data practices would presumably overcome the utility bar, it is unclear whether they are sufficiently novel and non-obvious to merit patent protection. In addition, as the *Datamize* decision instructs, patent protection would probably

[84] *See* Alice Corp. v. CLS Bank Int'l, 134 S. Ct. 2347 (2014) (holding that adding a computer to perform a set of functions that are otherwise abstract ideas does not confer patentability).

[85] Patent Act, 35 U.S.C. §§ 102–103.

[86] *Id.* § 102(b).

[87] *Id.* § 112 (requiring that claims have a definite meaning that individuals skilled in the art can understand).

[88] *In re* Musgrave, 431 F.2d 882, 893 (C.C.P.A. 1970); Datamize LLC v. Plumtree Software, Inc., 417 F.3d 1342, 1350 (Fed. Cir. 2005), *abrogated by* Nautilus, Inc. v. Biosig Instruments, Inc., 134 S. Ct. 2120 (2014). The definiteness test used in *Datamize* has since been refined by the Supreme Court such that "a patent is invalid for indefiniteness if its claims, read in light of the specification delineating the patent, and the prosecution history, fail to inform, with reasonable certainty, those skilled in the art about the scope of the invention." *Nautilus*, 134 S. Ct. at 2124. However, it is not clear that the Court's holding in *Nautlius* would change the outcome reached by the court in *Datamize*.

[89] *Datamize*, 417 F.3d at 1350 ("The scope of claim language cannot depend solely on the unrestrained, subjective opinion of a particular individual purportedly practicing the invention.").

[90] *See* Exxon Research & Eng'g Co. v. United States, 265 F.3d 1371, 1375 (Fed. Cir. 2001) ("If the meaning of the claim is discernible, even though the task may be formidable and the conclusion may be one over which reasonable persons will disagree, we have held the claim sufficiently clear to avoid invalidity on indefiniteness grounds.").

be unavailable to any practices that rely *entirely* upon subjective judgments. Finally, it is possible that some such methods would be merely abstract ideas ineligible for patent protection. The study presented in Part II probes these open questions.

Even when patent protection is available to information processing methods, trade secrecy may nevertheless be preferable. In a landmark publication on the economics of trade secrecy, David Friedman, William Landes, and Richard Posner identified two situations in which trade secrecy is preferable to patent protection: when patent protection seems too costly relative to the value of an invention, or when patent protection would provide a reward substantially lower than the value of an invention—for example, if an invention could easily be kept secret for a period of time longer than it would take other inventors to come up with the idea on their own.[91] Thus, the perceived cost of obtaining patent protection and the perceived value of secrecy could direct a big data producer toward secrecy even when patent protection might be available.

This overview of the relationship between intellectual property and big data would be incomplete without a brief look at copyright.[92] Unlike patent law and trade secrecy, copyright protection does not provide exclusivity in processes or methods.[93] Copyright may in some cases, however, protect the *products* of such practices. Originality, the sine qua non of copyrightability, has been found in data estimates, classifications, and in compilations (selections and arrangements) assembled through practices that rely upon subjective human

[91] David Friedman et al., *Some Economics of Trade Secret Law*, 5 J. ECON. PERSP., Winter 1991, at 61, 64.

[92] Copyright is discussed at greater length in Part III in connection with a new policy proposal.

[93] Copyright Act, 17 U.S.C. § 102(b) (2012) ("In no case does copyright protection for an original work of authorship extend to any idea, procedure, process, system, method of operation, concept, principle, or discovery, regardless of the form in which it is described, explained, illustrated, or embodied in such work."); *see also* Baker v. Selden, 101 U.S. 99, 103 (1879) ("The copyright of a book on perspective, no matter how many drawings and illustrations it may contain, gives no exclusive right to the modes of drawing described . . . ."); Morrissey v. Procter & Gamble Co., 379 F.2d 675, 678 (1st Cir. 1967) ("To permit copyrighting would mean that a party or parties, by copyrighting a mere handful of forms, could exhaust all possibilities of future use of the substance."). An important barrier to copyright protection for data is the merger doctrine—a venerable legal rule that bars copyright to works expressing ideas that can only be articulated in a limited number of ways. *Morrissey*, 379 F.2d at 678–79.

judgments.[94] Curiously, however, copyright does not require the authors of compilations to disclose their methods of assembly ex ante. Only if the copyrightability of a compilation is challenged in court is such information disclosed. Copyright is thus unlikely to promote the disclosure of big data practices.

The foregoing discussion can be reduced to several key insights: Many big data practices can probably be maintained as trade secrets, or even more simply, as undocumented procedures. Recent commentary on big data's disclosure problem indicates that big data practices are difficult, and perhaps even impossible to reverse-engineer.[95] Theory suggests that this makes secrecy an attractive option for big data producers.

Patent law presents a murkier picture. It is unclear, for instance, whether many big data practices would be sufficiently novel and non-obvious to merit patent protection. Moreover, the widely discussed subjectivity of big data practices suggests that perhaps many such methods could not be claimed in a manner definite enough to capture a meaningful scope of protection or any protection at all, for that matter.[96] Finally, even if a particular big data practice was patentable, theory instructs that trade secrecy is still preferable if the cost of obtaining a patent seems higher than any value one might expect to draw from the practice. These observations provide a helpful framework for examining the case studies presented in the next Part.

II. INDUSTRY PRACTICES

This Article asks whether intellectual property law adequately encourages the disclosure of big data practices. Because the big data phenomenon is relatively new, however, objective indicators such as patent filing behavior are of limited descrip-

[94] *See, e.g.*, CCC Info. Servs., Inc. v. Maclean Hunter Mkt. Reports, Inc., 44 F.3d 61, 67 (2d Cir. 1994) (stating that individual estimates of used car prices published by plaintiff were "original creations" for purposes of copyright); Am. Dental Ass'n v. Delta Dental Plans Ass'n, 126 F.3d 977, 979 (7th Cir. 1997) (holding short numerical codes copyrightable subject matter). The Copyright Act explicitly protects compilations "selected, coordinated, or arranged in such a way that the resulting work as a whole constitutes an original work of authorship." Copyright Act, 17 U.S.C. § 101 (2012); *see also* Feist Publ'n, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340, 362 (1991) (holding an obvious arrangement ineligible for copyright protection).

[95] *See supra* note 80 and accompanying text.

[96] In theory, highly subjective practices would also be difficult to disclose in a written description that would meet patent law's enablement requirement.

tive value. This question can be examined, however, by surveying the characteristics of big data practices—specifically, the possible patentability of these practices, the difficulty of uncovering these practices through reverse engineering, and the contexts in which these practices are being deployed.

This study draws on a set of interviews and surveys that I conducted with informaticists, data scientists, lawyers, and business professionals working at the vanguard of big data across different industries. In order to present a deep and varied portrait of the big data phenomenon, I assembled a listing of all big data companies, initiatives, and projects described in national newspapers, books, journals, and online press published since the phenomenon was first widely reported in late 2009.[97] I then interviewed individuals at these organizations by telephone.[98] All interviews lasted at least forty-five minutes, and some lasted hours and delved into subjects as technical and arcane as "probability-based methods of data masking."[99]

Two themes that emerged from this investigation are that big data practices are highly subjective, and difficult to uncover through reverse engineering. As a result, big data practices lend themselves toward secrecy. In addition, a number of disincentives to disclose, both economic and legal, further push toward secrecy. These findings varied, however, across different types of big data practices. To present these differences, this study is organized around four primary big data practices: filtering non-relevant data (i.e., "noise") from large datasets, identifying and correcting errors based on estimates or guesses, "masking" data in order to preserve the anonymity of individuals, and classifying data.

A. SEARCHING THE HAYSTACKS

Locating useful information within a large corpus of data is, in a sense, the ultimate search for a needle in a haystack.

[97] *See* MANYIKA ET AL., *supra* note 3, at 1 (outlining the explosion of data volume, collection, and use by various entities).

[98] All interviews were semi-structured. Only a subset of the many corporations and individuals whom I contacted agreed to be interviewed, which may have introduced sampling bias into this study. If any such bias exists, however, it is difficult to know how, or whether, it may have impacted the results of this study.

[99] This arcane-sounding practice involves obscuring only those portions of a dataset that could be used as "keys" to discover the identity of individuals whose identities a publisher wishes to keep private.

Online sources used by big data providers, such as social networks and online forums, span a vast array of topics and are often littered with "noise" in the form of spam (i.e., unwanted commercial messages). As a result, data culled from these sources must often be sifted and sorted before it can be put to good use.

Several technology startups boast special expertise in sifting data. One, aptly named DataSift, provides its customers with specialized streams of data culled from the hundreds of millions of daily posts made to social networks such as Twitter.[100] These data streams include the content of written messages, as well as related "metadata" describing, for instance, when online posts were written, or the gender, age, and geographic locations of authors.[101] The company aims to deliver data streams that provide helpful insights into the public's opinion of brands, news events, and even political candidates.[102]

Commenting for this Article, a vice president at the company described how the service sifts relevant data from the "firehose" of posts flowing from Twitter. "For every Tweet we receive," he explained, "we filter and enrich the content, by turning the 140 characters of each tweet into up to 400 fields of metadata."[103] The precise way that DataSift accomplishes this, he explained, depends deeply on a "human element."[104] For instance, the company routinely "encounters anomalies and data that are not 100% complete" from Twitter.[105] Such problematic data can be identified and sifted away by human reviewers.

The vice president shared a helpful example[106]: Recently, one of the company's clients requested a list of the twenty most popular athletes in America. The client, a clothing manufacturer, planned to use this list to decide which players' names to include on a new line of athletic jerseys. To find the answer, DataSift scoured the Internet to see which players on various sports teams were mentioned most often. The raw number of times a player was mentioned didn't reflect popularity alone,

[100] *See* DATASIFT, http://www.datasift.com (last visited Oct. 29, 2014).

[101] *See Data Enrichments*, DATASIFT, http://www.datasift.com/platform/data-enrichments (last visited Oct. 29, 2014).

[102] *See Data Sources*, DATASIFT, http://www.datasift.com/platform/datasources (last visited Oct. 29, 2014).

[103] Telephone Interview with Patrick Morrisey, DataSift (June 6, 2013).

[104] *Id*.

[105] *Id.*

[106] *Id.*

however: a player might be mentioned for positive or negative reasons. For this reason, the company relied on subjective human judgments to help determine the sentiment behind the online posts that it uncovered.[107]

In contrast to DataSift, which assembles information on a vast number of topics, other big data companies focus on a single subject. One example is Treato, formed in 2007 by an Israeli computer scientist named Roee Sa'adon.[108] The company's website describes its operation succinctly: "Treato automatically collects . . . the massive amount of content patients . . . generate online to extract relevant information, connect the dots and create the big picture of what they are saying about their personal treatment- and condition-related experiences."[109]

Like DataSift, Treato sifts commercial messages out of its dataset and often relies on subjective human judgments to do so. Sa'adon (now the company's Vice President of Technology) explained that selecting "high quality" information sources (i.e., sources that are relatively free of spam) is an important first step in this process.[110] In addition to being selective about its sources of data, Treato also carefully combs through its archive for commercial messages that should be excised. "A rigorous filtering process is necessary," Sa'adon explained, "and human judgment is often needed."[111] Treato employs full-time "data editors" who examine the online posts the company collects and ensure that any messages that seem commercial are removed.[112]

Treato also relies on experts to review the accuracy and quality of the non-commercial messages it encounters. According to Sa'adon, this stage in the process relies on the judgment of physicians hired by the company to review drug-related in-

[107] *See* Krystal Peak, *DataSift Debuts a Way To Find the Tweets You Need*, VATORNEWS (Nov. 16, 2011), http://vator.tv/news/2011-11-16-datasift-debuts-a-way-to-find-the-tweets-you-need.

[108] *See Our Story*, TREATO, http://www.corp.treato.com/story.html (last visited Oct. 29, 2014).

[109] *Id.*

[110] Telephone Interview with Roee Sa'adon, Vice President of Tech., Treato (Jun. 3, 2013); *see also* Roee Sa'adon, *Is Twitter a Good Source for Health Insights?*, TREATO BLOG (Mar. 18, 2013), http://www.blog.treato.com/is-twitter-a-good-source-for-health-insights/ (explaining the challenge of culling valuable health-related information from Twitter).

[111] Telephone Interview with Roee Sa'adon, *supra* note 110.

[112] *Id.*

formation to see if statements made online are consistent with general knowledge in the medical community.[113] The final set of approved posts that remain is compiled into a database, which as of this writing contains about 1.6 billion posts pertaining to 26,000 drugs and conditions.[114] The company grants users of its website access to this processed dataset.[115]

Yet another industry where big data selections are being compiled and sold is directed advertising. With the rise of social networks, a wealth of new and more detailed data pertaining to brand preferences, shopping habits, and even personal hobbies has become available to advertisers. TrueLens, a Boston-based firm that operates in this sphere, offered comments for this Article. The company's director of product marketing offered the following anecdote to explain the high level of subjectivity in its practices.[116] Suppose that an airline decides to launch two new routes from both Boston and San Francisco to Denver. The airline has a list of its past customers, but it does not know which of these customers are likely to be interested in the Boston-Denver route versus the San Francisco-Denver route. This is where big data sifting steps in. By analyzing publicly available information about the airline's customers (e.g., information that customers opted to share publicly on their social media profiles, publicly posted photos, check-ins and comments, etc.) the company is able to identify which of the airline's past customers are more likely to be interested in one particular route over the other.[117]

Significant human judgment goes into assembling this data, TrueLens's marketing director explained. Data scientists at the company might have a hunch, for example, that customers most interested in the airline's new route are those who live in major cities and who also enjoy skiing.[118] Relying on this hunch, they will create a selection of customers who match these criteria. With the benefit of this information, the airline can direct advertisements and promotional offers only to customers who

[113] *Id.*

[114] TREATO, http://www.treato.com (last visited Oct. 29, 2014).

[115] *Id.*

[116] Telephone Interview with Anish Kattukaran, TrueLens (July 10, 2013); *see also* TRUELENS, http://www.truelens.com (last visited Oct. 29, 2014) (helping marketers grow customer relations through social behavioral data and predictive analytics).

[117] Telephone Interview with Anish Kattukaran, *supra* note 116.

[118] *Id.*

are most likely to be interested.[119] Together, the examples in this Part reveal that the practice of sifting data often relies upon highly subjective judgments.

B. CLEANSING

The raw datasets that big data practitioners work with often contain errors. In part, this may be a consequence of their sheer size: unprecedented volumes of data imply unprecedented numbers of errors. A less obvious source of errors is the automatic and indiscriminate information-gathering that is a hallmark of the big data method. Even more subtly, some data errors manifest when error-free data from different sources is merged. In practice, identifying and correcting such errors is as much an exercise in aesthetics as statistics.

An informaticist interviewed for this Article offered the following example to describe the subjectivity of data cleaning in the healthcare industry. A cancer research project headed by the U.S. government recently requested a limited dataset of patient records from a Catholic health system.[120] The project's organizers required the sex and gender of every patient to be included in the dataset. Motivated by the religious beliefs of its leaders, however, the Catholic health system had long been identifying transgendered and transsexual patients as being of "UNKNOWN" sex and gender.[121] In order to deliver accurate data on the biological sex of the patients, the health system employed informaticists who imputed or inferred the sex of all patients who were labeled "UNKNOWN" based on related available data, such as height and weight. Deciding which information mattered was key to this process: "A diagnosis of prostate cancer would lead us to decide that an individual was male, regardless of data that suggested otherwise, such as a petite body size," the informaticist explained.[122] Thus, the final listing of patients delivered to the government was in part a product of professional judgment.

Because data cleaning is often highly subjective, different informaticists could easily produce different final products. While one expert might impute sex from certain discrete values

[119] *Id.*

[120] Telephone Interview with Josh Mann, Assistant Dir. of Oncology Tech. Solutions, at Am. Soc'y of Clinical Oncologists (Oct. 8, 2013).

[121] *Id.*

[122] *Id.*

such as diagnosis, height and weight, another might look through a doctor's notes to see textual references to gender, such as "he" or "she."[123] These two approaches could easily lead to different results. The informaticist who provided this example opined, "Cleaning big data is sometimes fairly subjective. Different professionals can dream up different data points to interpolate from."[124]

A data expert and economist from a prominent social network offered another helpful hypothetical example of how data cleaning works.[125] Suppose a big data analyst working for an online business wishes to collect data on how long visitors stay on her employer's website. When the analyst collects relevant data from the company's web server, she finds that most visitors appear to stay on the website for 2–5 minutes. Some of the data doesn't make sense, however: the server reports many visits lasting "0 minutes" in length, some visits lasting several days in length, and a few inscrutable results such as "infinity" and "not a number."

Faced with these anomalous results, the analyst might first try to find the sources of the errors. She may guess, for instance, that the records of visits lasting "0 minutes" were generated by automated software agents known as "bots." The visits apparently lasting for days, meanwhile, were probably generated by users who walked away from their computers without closing their web browsers. Lastly, she surmises that a bug in the web server's software caused the reports of "infinity" and "not a number."

After identifying the sources of these errors, the analyst "will probably clean data differently for different exercises," the expert interviewed for this article explained.[126] For instance, if she wishes to learn how all visitors interact with the website (including inactive users), she may decide to delete only the entries reporting "0 minutes" to correct for software bots. If the analyst's goal is to learn how long users stay on the website before clicking on links that take them to other websites, however, she may also delete all entries greater than 10 minutes to correct for inactive browsers. Ultimately, the final cleaned da-

[123] *Id.*

[124] *Id.*

[125] *See* E-mail Exchange with Michael Bailey, Econ. Research Manager, Facebook (July 2014) (on file with author).

[126] *Id.*

taset will reflect the analyst's judgments about the sources of error and her specific goals.

A simplified example provides a detailed picture of how data cleaning works in practice. The following table contains demographic information about four fictional medical patients:

Table 1: Dataset A[127]

Name	Age	D.O.B.	Address	Diagnosis	SSN
Jim Smith	05	11/22/1963	123 Main St.	Arthritis	123456
Smith, Kris	44	30/13/1970	123 Mane St.	Fracture	123456
C.J. Craig	121	1/1/1993	Munick, Germ.	B.	N/A
Sue Jordan	74	1/13/1940	Georgetown	DVT	4921923

Suppose that a big data publisher received this data from a doctor's office and wanted to identify and, where possible, correct all errors before sharing it with customers and partners. A few of the errors in this example are so obvious that they could be identified automatically by software. The date of birth of Jim Smith in the first row, for instance, does not correspond with the patient's age. Likewise, the date of birth in the second row contains an invalid month entry of "30." Software performing a statistical analysis of the four patient's' ages would notice that the age of "C.J. Craig" in row three, which was entered as "121," is improbably high—a statistical outlier that is likely an error.

Some of the remaining errors in Table 1 might require human judgment to correct. Common sense may be required to deduce, for instance, that a town probably would not contain two streets named "Main" and "Mane." The abbreviation "DVT"

[127] This hypothetical was reviewed and developed with the help of Michael Bailey of Facebook. *See id.* I also wish to credit Paul Ohm, who illustrated data de-identification in a similar format in a 2010 article on big data and privacy. *See* Ohm, *supra* note 29.

in row 4 presents an even deeper ambiguity: a doctor might be called upon to explain that the abbreviation could refer to either "deep vein thrombosis" or "diverticulitis." Without more information, however, it could be difficult to guess which is correct. Setting this ambiguity aside, an informaticist could assemble the data into the following intermediate form:

Table 2: Dataset A Cleaned

Name	Age	D.O.B.	Address	Diagnosis	SSN
Jim Smith	50	11/22/1963	123 Main St.	Arthritis	123456
Kris Smith	44	03/13/1970	123 Main St.	Fracture	N/A
C.J. Craig	21	1/1/1993	Munich, Germ.	B.	NONE
Sue Jordan	74	1/13/1940	Georgetown	DVT	492192

At this stage, the table still contains some ambiguous and missing information, but less than before. Now suppose that the following second database is shared by a local hospital:

Table 3: Dataset B

Last Name	First Name	Gender	Street	City	Complaint
Smithe	James	M	Main	Anytown	Joint pain
Jones	Deb.	F	Maple	Shellbyville	Foot
Jordan	Suzanne	F	Pine	Washington	Leg pain

The records in Dataset B are obviously formatted differently from those in Dataset A. As a result, an informaticist would need to conform or "normalize" the two sets before merging them—a step requiring a subjective judgment about how the data should be organized. But deeper subjective judgments shape the final product. Judging by street addresses and the type of injury, for example, it is likely that "James Smithe" is the same "Jim Smith" listed in Dataset A. In other words, the first row of Table 3 contains duplicative information. A researcher who did not make this guess would conclude that there are two patients who suffer from arthritis when, in all likelihood, there is only one.

Following similar logic, the informaticist guesses that "Suzanne Jordan" in Dataset B is "Sue Jordan" in Dataset A. Going further, she deduces that Ms. Jordan's complaint of leg pain in Dataset A implies deep vein thrombosis rather than diverticulitis. Ultimately, the final cleaned and merged datasets could appear as follows:

Table 4: Final Cleaned and Merged Dataset

Name	Age	Sex	D.O.B.	Address	Problem	SSN
James Smith	50	M	11/22/1963	123 Main St.	Arthritis	123456
Kris Smith	44	F	3/13/1970	123 Main St.	Fracture	N/A
C.J. Craig	21	F	1/1/1993	Munich, Germany	Fracture	N/A
Suzanne Jordan	74	F	1/13/1940	Pine Street, Washington	Deep Vein Thrombosis	492192
Deb Jones	47	F	1/1/1966	Maple Street, Shelbyville	Foot	N/A

Note that the age of "Deb Jones" in the final row is an average of the ages of the other participants. Although this value is probably not Deb Jones' true age, a big data publisher might insert it in the dataset because it would permit the patient's condition to be included without significantly disrupting the other valid age-related data.[128] As the expert consulted on this example explained, however, such a decision, like so many aspects of the data cleaning process, would rely on the subjective judgments of the person preparing the data.[129]

C. MASKING AND SUPPRESSION

Many big data producers obfuscate or "mask" personally identifying information contained in the raw data they begin with. In some industries, the law mandates this practice. Under the Health Insurance Portability and Accountability Act of 1996 (HIPAA), for instance, personal health records cannot be shared between institutions unless names, zip codes, treatment dates, and other specific identifiers are removed.[130] Even in the absence of a legal mandate, market forces have pushed some big data producers to mask personal data. Like data selection and data cleaning, data masking is a mix of science and art, the product of which is often infused with subjective judgments.

The simplest way to anonymize a dataset is to completely strip it of information that can be used to identify individuals, such as names, addresses, and phone numbers. Although this approach is often effective at ensuring anonymity, it also tends to destroy useful information. Completely removing personally identifying information makes it impossible to analyze data longitudinally, for example.[131]

A less destructive alternative is to systematically replace personally identifying information with dummy values. This approach permits the identification of the same individuals over time.[132] A big data marketing firm director interviewed for this Article explained that prior to sharing customer shopping habits with outside analysts, some companies replace every customer's name with a unique "hash"—a random string of letters and numbers.[133] By studying the behavior of a particular

[128] E-mail Exchange with Michael Bailey, *supra* note 125.

[129] *Id.*

[130] 45 C.F.R. § 164.514(e) (2013).

[131] *See, e.g.*, INFORMATION COMMISSIONER'S OFFICE, ANONYMISATION: MANAGING DATA PROTECTION RISK CODE OF PRACTICE 83–84 (Nov. 2012), [hereinafter CODE OF PRACTICE] *available at* http://ico.org.uk/for_organisations/data_protection/topic_guides/anonymisation.

[132] *See supra* note 18 and accompanying text (describing an example of this form of anonymization recently used to mask online search queries).

[133] Telephone Interview with Anonymous Source #1 (July 10, 2013) (although most individuals interviewed for this Article consented to being identi-

hash over time, analysts can then understand an individual customer's habits without learning the customer's name.[134]

A mobile phone operator recently followed this approach when MIT researchers asked it for data describing its customers' GPS locations in order to study Bay Area and Boston-area traffic patterns.[135] The phone company generated a dataset in which every phone number was replaced with a randomized value that was used consistently through all of its records.[136] By doing so, the phone operator made it possible for individual phones to be studied over time without disclosing any real phone numbers.[137]

Data masking sometimes involves techniques far more complex than replacing names with dummy values. An example was offered by experts at CancerLinQ, a project organized by the American Society of Clinical Oncologists in 2012.[138] CancerLinQ aggregates clinical information from hospitals around the country relating to cancer treatment. Such information includes, for instance, lab tests and doctors' notes.[139] The system then culls this data and correlates the successfulness of treatments with patient characteristics in order to provide treatment suggestions.[140]

Commenting for this Article, an informaticist and a lawyer working on CancerLinQ explained that preserving patient privacy often requires significant ingenuity. "Simply mechanically stripping HIPAA's eighteen restricted identifiers from our dataset would erase valuable information, such as dates of key care events and demographics," they described.[141] Instead of deleting information entirely, experts working on CancerLinQ turned to a software firm that specializes in de-identifying pa-

fied by name, several individuals commented only on condition of anonymity).

[134] *Id.*

[135] Pu Wang et al., *Understanding Road Usage Patterns in Urban Areas*, SCI. REP., (Dec. 20, 2012), http://www.nature.com/srep/2012/121220/srep01001/pdf/srep01001.pdf.

[136] *See* Pu Wang et al., *Understanding Road Usage Patterns in Urban Areas: Supplementary Information*, SCI. REP. 4 (Dec. 20, 2012), http://www.nature.com/srep/2012/121220/srep01001/extref/srep01001-s1.pdf.

[137] *Id.*

[138] *CancerLinQ*, AM. SOC'Y OF CLINICAL ONCOLOGY, http://www.asco.org/quality-guidelines/cancerlinq (last visited Oct. 29, 2014).

[139] *Id.*

[140] *Id.*

[141] Telephone Interview with Am. Soc'y of Clinical Oncologists (Oct. 7, 2013).

tient data.[142] The software allows the users of the system to prioritize the preservation of key information as well as permitted permutations, such as shifting all treatment dates equally to preserve a longitudinal record of the length of a particular patient's treatment without reporting actual dates of treatment. The software can slightly alter ages, geographic locations, treatment dates, and the like in order to meet the HIPAA mandate without rendering the data useless.[143]

One might guess that de-identification software entirely automates the process of data masking, but an informaticist interviewed explained that human judgment plays an important role in the process:

> At every step of the way, there [are] a lot of subjective questions. For example, the person using the software must be able to say how much they trust the recipient of the data or whether they think the data might be publicly exposed. The entire process of statistical de-identification is filled with subjective questions.[144]

The foregoing examples show that data masking sometimes entails subjective judgments. This practice is still heavily anchored to objective criteria, however.

D. CLASSIFYING

Yet another technique of altering data prior to publication is classification. Like the well-known optical illusion that portrays either a young woman or an old woman,[145] the picture that big data draws is often in the eye of its beholder. Nowhere is the fundamental subjectivity of perception more apparent than in the classifications and taxonomies that big data practitioners impose upon the data they work with.

[142] *Id.* The software firm is named "Privacy Analytics." *Id.*

[143] *Id.*

[144] Telephone Interview with Anonymous Source #2 (Oct. 7, 2013). This observation is amply supported by data practice guides published by the UK's Information Commissioner's Office. *See* CODE OF PRACTICE, *supra* note 131, at 81 ("[T]he choice of a particular method of anonymisation will depend on many factors, including an understanding of the potential risk of exposing personal data inappropriately, the sensitivities of the data, and the amount of control that the data controller has over the uses to which the anonymised data will be put. . . . *Hence the choice of an anonymisation technique should always be a matter for the data controller's judgment, based on the context of data sharing or use*." (emphasis added)).

[145] A reproduction of this well-known image appeared in Edwin G. Boring, *A New Ambiguous Figure*, 42 AM. J. PSYCHOL. 444, 444 (1930).

Classification is particularly useful in big data applications that cull linguistic data for insights (e.g., the analysis of online social media posts in order to gauge consumer opinions). A number of startup companies now specialize in this practice of so-called "sentiment analysis."[146] In a recently published interview, a chief scientist at one such company offered an example of the unavoidable subjectivity in classifying such data: "If a laptop is big, it's negative. But if a hard drive is big, it's good."[147] Although sophisticated software might be able to make some such classifications on its own, experts who commented for this article explained that human judgment is often required to make accurate and useful classifications of linguistic data.[148]

Classification also plays an important role in big data applications that relate to consumer shopping habits. For example, companies that aim to predict what shoppers will buy in the future purchase big data sets in which consumers are "coded" according to categories. By analyzing a set of purchasing data, for instance, an informaticist might cluster customers into unexpected categories, such as people who buy "Brussels sprouts and sugared cereal."[149] A director at one such company explained that there is no simple formula for creating these types of classifications; rather, what is needed is a deep knowledge of the subject matter and the ability to find patterns in the data.[150]

Classifying data often requires an appreciation for context that only a human can judge. In 2011, Dr. Monica Stephens of Humboldt State University gathered and presented scores of online Twitter posts in a map of the United States that identifies where hateful speech is most prevalent.[151] In carrying out this project, Dr. Stephens realized that identifying "hate" is more difficult than simply searching for certain words. Depending on context, some derogatory terms can take on a positive or negative connotation. To address this problem, Dr. Stephens

[146] BAKER, *supra* note 1, at 99, 121 (discussing the practice of divining human sentiment from big data sources).

[147] *Id.* at 114 (quoting Nicolas Nicolov of Umbria Communications).

[148] Telephone Interview with Anonymous Source #1, *supra* note 133.

[149] BAKER, *supra* note 1, at 43–65 (discussing the practice of grouping consumers into such "buckets").

[150] Telephone Interview with Anish Kattukaran, *supra* note 116.

[151] Monica Stephens, *FAQ: Geography of Hate*, FLOATING SHEEP (May 10, 2013, 10:40 PM), http://www.floatingsheep.org/2013/05/hatemap.html.

had her assistants manually review each post, remove those not derogatory in nature, and then classify the speech in the posts that remained.[152] Thus, the final processed dataset reflects subjective classifications that were made by Dr. Stephens and her research team.[153] Classifying data to facilitate analysis is a key big data practice, and like data sifting, it appears to sometimes *entirely* rely upon subjective human judgments.

III. IMPLICATIONS & RECOMMENDATIONS

This study's chief empirical finding is that big data practices—the ways in which raw data are transformed into useful datasets—frequently entail subjective judgments. Because these judgments are typically performed in an ad hoc fashion in response to the unique circumstances in which a given set of data is initially gathered, they are a mystery to downstream users. Big data practices are easy secrets to keep.

The foregoing study of how big data practices work enables an examination of whether intellectual property law does anything to encourage their disclosure. As this Part explains, the answer is, for the most part, "no." Big data practices do not fit neatly within the existing intellectual property paradigms of patent or copyright law. At the same time, the fact that these practices are not self-disclosing (i.e., they cannot be easily reverse-engineered) lends them well to trade secret status, or to mere nondisclosure. These conclusions point toward the need for new policies designed to encourage the disclosure of big data practices. To address this need and to stimulate further discourse, this Part outlines a hypothetical intellectual property based prescriptive measure.

A. PATENT LAW IS UNLIKELY TO ENCOURAGE BIG DATA DISCLOSURES

Many of the big data practices uncovered by this study appear either unlikely to meet patent law's threshold eligibility requirements, or potentially eligible but nevertheless unlikely to garner a meaningful scope of patent protection. Thus, patent law does not appear to meaningfully encourage the disclosure of big data practices.[154]

[152] *Id.*

[153] *See id.*

[154] *See supra*, Part I.C.

As explained earlier in this Article, the Federal Circuit has instructed that patent process claims must not rely on subjective judgments.[155] Such claims fail patent law's requirement of definiteness, the court has explained, because they do "not notify the public of the patentee's right to exclude since the meaning of the claim language would depend on the unpredictable vagaries of any one person's opinion."[156] Process claims may, however, involve some degree of human judgment.[157]

Some big data practices appear to rely *entirely* on subjective judgments made in an ad hoc fashion. Services that rely on humans to sift useful information from a large dataset are good examples. Treato calls upon staff physicians to select useful patient information culled from online forums, for instance.[158] Likewise, the marketing firms surveyed rely on human judges to determine which information should be included in their final products.[159] Data classification also sometimes relies entirely on subjective judgments. Statistical techniques can be used to help identify clusters of customer behavior and traits, but ultimately, subjective judgment is necessary to create useful classifications.[160] These big data practices would thus appear to be ineligible for patent protection because they cannot be claimed with sufficient definiteness.

Other big data practices rely only partially on subjective judgments, however, and could probably be claimed with sufficient definiteness. Consider the case of data cleaning that was performed on health records provided by a Catholic health system.[161] Inferring the sex of individuals who were not accurately coded into the system required ingenuity, but it was nevertheless based upon objective criteria—namely, physical factors that indicated sex, such as height and weight.[162] Likewise, data masking practices, such as replacing identifying information with dummy values, may involve subjective assessments of

[155] *See supra* note 88 and accompanying text.

[156] Datamize, LLC v. Plumtree Software, Inc., 417 F.3d 1342, 1350 (Fed. Cir. 2005), *abrogated by* Nautilus, Inc. v. Biosig Instruments, Inc., 134 S. Ct. 2120 (2014).

[157] *See supra* note 90 and accompanying text.

[158] *See supra* Part II.A.

[159] *See supra* Part II.A.

[160] *See supra* note 149 and accompanying text.

[161] *See supra* note 120 and accompanying text.

[162] *See supra* note 120 and accompanying text.

risk, but they are objectively anchored.[163] Even more complex methods, like the statistical practices used at CancerLinQ to mask personally identifying information,[164] could likely be expressed in claim terms that are sufficiently definite to receive patent protection.

Although big data practices that are objectively anchored could probably be claimed with sufficient definiteness, other barriers to patent protection may nevertheless stand in the way, of course. A failure to show sufficient novelty or nonobviousness, for instance, could lead to a rejection or a later invalidation.[165] Likewise, various statutory bars to patent protection may apply.[166] Importantly, methods of preparing data that are merely abstract ideas would be denied patent protection as ineligible subject matter.[167]

Even when patent protection is available to such practices, however, big data producers may nevertheless prefer the path of nondisclosure. As explained earlier in this Article, trade secrecy is preferable to patenting when an invention can easily be kept secret for a period of time longer than it would take other inventors to come up with the idea on their own.[168] Many big data practices fall squarely into this category. Like Google's Pagerank and the algorithms used by high-speed trading companies, big data practices yield commercially valuable products and services while remaining entirely out of view.[169] A legal expert on big data at Microsoft supported this conclusion, stating that "[i]f [big data practices] are going to be used almost entirely internally, behind a firewall, then the company may not need or want patent protection and the disclosure it requires."[170] This would seem to make trade secret protection, or mere casual nondisclosure, even more attractive to big data producers than

[163] *See supra* Part II.C.

[164] *See supra* Part II.C.

[165] *See supra* Part I.C.

[166] *See supra* Part I.C.

[167] *See* Alice Corp. v. CLS Bank Int'l, 134 S. Ct. 2347 (2014) (holding that adding a computer to perform a set of functions that are otherwise abstract ideas does not confer patentability).

[168] *See supra* note 91 and accompanying text.

[169] Individuals interviewed for this Article confirmed that big data practices typically render datasets very difficult and sometimes impossible to reverse-engineer. As Paul Ohm has discovered, however, it is sometimes possible to "re-identify" data that has been masked. Ohm, *supra* note 29.

[170] E-mail from Microsoft Source to author (Feb. 10, 2014) (on file with author).

it has long been to software producers. Unlike software object code, most big data products cannot be reverse-engineered to reveal the processes that went into their creation.[171]

A second scenario where nondisclosure is economically preferable to patent protection is when patent protection seems too costly relative to the value of an invention.[172] Like the software industry, the world of big data is fast-paced. Big data producers may often view the economic value of their practices as relatively short-lived, and as a result, not worth the time and trouble of obtaining patent protection. In such cases, data producers may simply neglect to document and disclose their practices. One of the experts interviewed for this article corroborated this view while discussing his time working at a big data producer. "We rarely slowed down to go through the burdensome process of patent filing," he stated, "not to mention that [applying for a patent] is expensive and time consuming."[173] Patent protection may simply not be worth the candle.

Beyond the lack of legal incentives to disclose big data outlined above, there appear to be a number of meaningful disincentives to disclosure. Privacy regulations, for instance, would discourage a publisher of medical records from disclosing its method of data suppression. (Such disclosures would likely facilitate unwanted re-identification.) Likewise, data producers may sometimes feel that disclosing their methods would reveal flaws in their methodologies or weaknesses in their underlying data. In short, intellectual property law may often not be the only reason why data producers choose not to disclose their methods.

B. COPYRIGHT LAW PROVIDES THIN PROTECTION FOR BIG DATA CORPORA

Copyright law offers surprisingly thin protection for corpora of big data. The following paragraphs explain why this is so, and also lay the theoretical foundation for this Article's central policy discussion.

As explained in Part I, copyright law can protect original expression found in compilations of data. Like ceramic frag-

[171] This reality is due to the simple fact that it is usually impossible to guess the various techniques and judgments that go into processing a dataset.

[172] *See supra* note 91 and accompanying text.

[173] E-mail from Google source to author (Feb. 13, 2014) (on file with author).

ments composed into an intricate mosaic, some data are individually unremarkable but collectively capture an original expression. The Copyright Act thus protects compilations "selected, coordinated, or arranged in such a way that the resulting work as a whole constitutes an original work of authorship."[174] Courts have deemed compilations of data copyrightable when the process of selecting or arranging the data required an exercise of subjective judgment. The Second Circuit has explained that "[s]election implies the exercise of judgment in choosing which facts from a given body of data to include in a compilation."[175] Likewise, arrangement "refers to the ordering or grouping of data into lists or categories that go beyond the mere mechanical grouping of data as such, for example, the alphabetical, chronological, or sequential listings of data."[176]

The forms of data sifting described in this Article clearly meet the originality bar as forms of selection. Some companies rely entirely on human judges to identify and sift-away unwanted commercial content.[177] But even more nuanced acts of selection are also being performed: big data companies that select social media posts that they believe customers will find the most helpful, or that select customers from a list who seem the most likely to buy a product, are both examples of exercises in judgment that would lead to a copyrightable selection.

As a practical matter, however, such protection is unlikely to effectively curtail unwanted copying. This is because copyists could, in theory, easily appropriate individual datums without copying their specific arrangement or selection within the database.[178] To invoke the metaphor used earlier, it is often possi-

[174] Copyright Act, 17 U.S.C. § 101 (2012). An important limitation on this form of copyright, however, is that it "extends only to the material contributed by the author of such work . . . and does not imply any exclusive right in the preexisting material." *Id.* § 103.

[175] Key Publ'ns, Inc. v. Chinatown Today Publ'g Enters., Inc., 945 F.2d 509, 513 (2d Cir. 1991).

[176] *Id.* (quoting U.S. COPYRIGHT OFFICE, GUIDELINES FOR REGISTRATION OF FACT BASED COMPILATIONS 1 (1989)).

[177] *See supra* Part II.A.

[178] *See* Warren Publ'g, Inc. v. Microdos Data Corp., 115 F.3d 1509, 1520 (11th Cir. 1997) (holding no infringement on plaintiff's directory because defendant's selection and arrangement varied from plaintiff's); BellSouth Adver. & Publ'g Corp. v. Donnelley Info. Publ'g, Inc., 999 F.2d 1436, 1441–42 (11th Cir. 1993) (holding no infringement, even though defendant copied a substantial amount of material from plaintiff's directory); Triangle Publ'ns, Inc. v. Sports Eye, Inc., 415 F. Supp. 682, 684–86 (E.D. Pa. 1976) (holding that data

ble to steal the tiles without copying the entire mosaic. As a result, the scope of copyright protection that corpora of big data enjoy is likely thin.[179]

Big data corpora that contain classifications also seem to meet copyright's originality requirement. In the 2007 case of *American Dental Association v. Delta Dental Plans Association*, the Second Circuit held that individual six-digit codes for dental procedures were copyrightable works of authorship that met Copyright's originality threshold.[180] Judge Easterbrook, who decided the case, explained that the plaintiff's selection of "six digits rather than five" reflected a judgment that more dental procedures would be added to the catalog over time, for instance.[181] Judge Easterbrook also found that the plaintiff's placement of related procedures in similar numerical series (e.g., the 2500 series or the 4200 series) was an expression of judgment that met Copyright's threshold for originality.[182] "Classification is a creative endeavor," Easterbrook concluded.[183]

Unfortunately for big data producers, other circuit courts have explicitly refused to grant copyright protection to data

compiled and published by plaintiff pertaining to races could be used by defendant in a competing publication because only the form of expressing the data, and not the data itself, is copyrightable); Jason R. Boyarski, *The Heist of* Feist*: Protection for Collections of Information and the Possible Federalization of "Hot News*,*"* 21 CARDOZO L. REV. 871, 904 (1999) ("Since courts have generally found comprehensive takings from copyrightable compilations to be noninfringing, collectors of information have been unable to obtain relief for damage to their investments as a result of substantial, competitive copying.").

[179] *See* David E. Shipley, *Thin but Not Anorexic: Copyright Protection for Compilations and Other Fact Works*, 15 J. INTELL. PROP. L. 91, 141 (2007) ("The protection copyright grants to a compilation may not be anorexic, but it certainly remains very lean."); Julie Wald, Note, *Legislating the Golden Rule: Achieving Comparable Protection Under the European Union Database Directive*, 25 FORDHAM INT'L L.J. 987, 1016–17 (2002) ("Following *Feist*, the U.S. appellate courts consistently demonstrated that copyright protection given to databases is extremely limited . . . . In the post-*Feist* era, it is increasingly difficult to prevent a competitor from taking substantial amounts of factual material from copyrighted collections of information and using it in a competing product.").

[180] Am. Dental Ass'n v. Delta Dental Plans Ass'n, 126 F.3d 977, 979 (7th Cir. 1997) (holding the short numerical codes to be copyrightable subject matter).

[181] *Id.*

[182] *Id.*

[183] *Id.*

that represents classifications.[184] Writing for the Third Circuit in the 2004 case of *Southco, Inc. v. Kanebridge Corporation*, Justice (then Judge) Alito explained that offering such protection would violate Copyright's longstanding tenet that protection may not extend to words or short phrases.[185] Judge Alito reasoned that extending copyright protection to a single number—say, "46,873"—would potentially lead anyone who used that number to become an infringer.[186]

Some big data producers discussed in this Article publish classifications that reflect subjective judgments, such as types of consumers, or the sentiment behind language. Generalizing numbers into numerical ranges in order to hide personally-identifying information could also arguably constitute a form of classification. Although the *Delta Dental* decision might lead one to think that copyright covers such subject matter, the overwhelming body of case law on this subject points in the opposite direction. Big data producers cannot rely on copyright to prevent unwanted copying of data classifications.

The foregoing analysis is by no means comprehensive—it serves only to show that, although the subjectivity of big data may imbue corpora of big data with a degree of copyrightable expression, such protection is unlikely to be robust.

[184] *E.g.*, ATC Distribution Grp., Inc. v. Whatever It Takes Transmissions & Parts, Inc., 402 F.3d 700, 707–08 (6th Cir. 2005).

[185] Southco, Inc. v. Kanebridge Corp., 390 F.3d 276, 285–87 (3d Cir. 2004) (en banc). An additional basis for denying protection was that, unlike the dental classifications in *Delta Dental*, the screw fastener numbers were arbitrarily selected and as a result, "totally unoriginal." *Id.* at 289 (Becker, J., concurring).

[186] *Id.* at 286. A survey of case law indicates that the bar on short words and phrases is not absolute and is typically applied with sensitivity to the specific words and phrases that are used. In a 2012 decision, for instance, the First Circuit wrote, "[A]pplicability of this law very much turns on the specific short phrases at issue, as not all short phrases will automatically be deemed uncopyrightable." Soc'y of Holy Transfiguration Monastery, Inc. v. Gregory, 689 F.3d 29, 52 (1st Cir. 2012); *see also* Health Grades, Inc. v. Robert Wood Johnson Univ. Hosp., Inc., 634 F. Supp. 2d 1226, 1238 (D. Colo. 2009) ("Accordingly, 'it does not make sense to state categorically that no combination of numbers or words short enough to be deemed a 'phrase' can possess 'at least some minimal degree of creativity' as required for copyright protection . . . ." (quoting *Southco*, 390 F.3d at 298 (Roth, J., dissenting) (citation omitted))). The regulation, the court wrote, should be viewed as only "a rough starting point for an originality analysis." *Soc'y of Holy Transfiguration Monastary*, 689 F.3d at 52.

C. ENCOURAGING DISCLOSURE OF BIG DATA PRACTICES

Big data is widely viewed as an engine for innovations that could enhance social and economic welfare.[187] This study indicates that such innovations may never come to light, however, if data producers do not document and disclose their practices. Troublingly, a variety of economic and legal forces discourage disclosure. As such, policies designed to encourage the disclosure of big data practices would seem to be normatively desirable. In order to spur discussion, this Part presents a policy model rooted in intellectual property law. This model is not offered as a formal legislative proposal, but rather, as an exploratory device intended to spur much-needed discussion and debate.

Big data's disclosure problem is not, of course, inherently an "intellectual property problem." Rather, intellectual property law is concerned with problems of technological disclosure, and thus it is potentially a relevant and helpful policy tool in this context.[188] The purpose of this discussion is to explore whether an intellectual property-based solution would be helpful. If not, different solutions might be developed far outside the precincts of intellectual property. The Federal Trade Commission (FTC), for instance, already investigates how big data practices affect consumers and could, in theory, enact rules that would encourage greater disclosures.[189] Another potential avenue for policymaking could be new limits on the availability of trade secret protection with respect to big data practices. One can easily also envision that the FDA might mandate new data disclosure rules pertaining to the technologies within its purview. These possibilities and others like them are valuable topics for future study and debate. This discussion is solely in-

[187] *See supra* Part I.

[188] As Brett Frischmann has noted, "Intellectual property laws are a prominent but by no means exclusive means of addressing the supply-side problem where free riding is a concern and appropriating benefits through market exchange of the intellectual resource or some derivative product is relevant to investment decisions." BRETT FRISCHMANN, INFRASTRUCTURE: THE SOCIAL VALUE OF SHARED RESOURCES 263 (2012).

[189] *See* FED. TRADE COMM'N, PROTECTING CONSUMER PRIVACY IN AN ERA OF RAPID CHANGE (2012), *available at* http://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-report-protecting-consumer-privacy-era-rapid-change-recommendations/120326privacyreport.pdf (discussing the effect of big data on consumer privacy and providing policy recommendations).

terested in the viability of an intellectual property-based solution.

What might an intellectual property-based solution to the big data disclosure problem look like? A carefully tailored form of sui generis intellectual property protection is one possibility.[190] Specifically, one might imagine a new legal entitlement—termed herein a "dataright" for convenience—that would be available to applicants who disclose clear and complete descriptions of their data collection and preparation methods alongside the data shaped by those methods.[191]

This new legal construct would be defined by three characteristics found in nearly all forms of intellectual property: (1) subject matter covered by the right; (2) exclusive rights conferred to publishers of this subject matter; and (3) a set of acquisition rules upon which exclusivity is conditioned. Concerning subject matter, a dataright could protect any data that has been collected or manipulated according to one or more methods not readily apparent to a person of ordinary skill in the art.[192] Protection might extend to individual datums as well as corpora of data. In this respect, a dataright could protect a finer-grained set of subject matter than copyright, for instance, which typically extends only to entire compilations of data. As a result, potential downstream users would be unable to skirt around the right in the way that copyright permits.[193]

[190] Because a central goal of intellectual property law is to encourage technological disclosures, the law of intellectual property seems eminently suitable for addressing big data disclosures. Moreover, the limitations of patent law and copyright law discussed earlier in this Part suggest that a new "sui generis" form of protection could be appropriate in this context. Readers should not conclude that encouraging big data disclosure is inherently an "intellectual property problem," however; there are likely many ways to encourage big data disclosures that do not rely on granting monopoly-like rights. This Article's overarching goal is to direct discourse toward the development of such policies.

[191] For a discussion of possible Constitutional limitations on such a law, see Yochai Benkler, C*onstitutional Bounds of Database Protection: The Role of Judicial Review in the Creation and Definition of Private Rights in Information*, 15 BERKELEY TECH. L.J. 535, 543–45 (2000).

[192] This standard borrows from patent law, which invokes the "person of ordinary skill" to resolve issues pertaining to initial protection. Patent Act, 35 U.S.C. § 103 (2012). Part II of this Article provides many examples of what sorts of methods would qualify for protection.

[193] *See supra* Part III.B (describing the limitations of copyright protection).

Turning to the subject of exclusivity, we might wish for dataright holders to be entitled to sue unauthorized *users* of their data for injunctive relief for some limited period of time. An exemplary "use" of data would be, for instance, applying a dataset to analysis in order to study a new problem or phenomenon. We might wish for dataright holders to *not* be entitled to prevent third parties from reproducing or distributing descriptions of the subject matter itself. Thus, underlying data could be freely reproduced and distributed barring any additional restrictions imposed by publishers through, for instance, contracts.[194] This limited exclusive entitlement would aim to balance data producers' desire to control downstream *use* against the public's interest in having widespread access to data.[195]

Turning to acquisition rules, dataright protection under this hypothetical plan would be available only to publishers who disclose all data collection and organization practices relevant to each piece of data they seek to protect.[196] This disclosure requirement is analogous to patent law's requirement that applicants disclose their inventions in formalized applications and, to a lesser degree, to copyright's requirement that authors seeking protection fix their works in tangible media.[197] The acquisition rules of dataright would be unique, however, in the

[194] Data producers have long relied upon contracts to curtail unwanted copying. U.S. COPYRIGHT OFFICE, REPORT ON LEGAL PROTECTION FOR DATABASES 22 (1997), *available at* http://www.copyright.gov/reports/db4.pdf ("For many database producers, contracts provide a major source of protection, either complementing copyright law or picking up the thread where it falls short."). This method of "self-help" in the data publishing industry may prevent some unwanted copying, but publishers have long lamented that contracts alone are far weaker than intellectual property protection because they avail only against licensees and not against unlicensed downstream copyists. *See The Consumer and Investor Access to Information Act of 1999: Hearing on H.R. 1858 Before the Subcomm. on Telecomms., Trade, & Consumer Prot. of the H. Comm. on Commerce*, 106th Cong. 67–68 (1999) [hereinafter *Hearing*] (statement of Lynn O. Henderson, President, Doane Agricultural Services, on behalf of the Agricultural Publishing Association) (disagreeing with the assertion that contracts provide adequate protection).

[195] Patent law strikes a similar balance: patent holders can seek to enjoin third parties from making, using, selling, or distributing their inventions, but third parties are free to reproduce and distribute descriptions of the inventions themselves. Patent Act, 35 U.S.C. § 154(a).

[196] This plan is inspired by the notion of a semi-patent, which Gideon Parchomovsky and I previously introduced. Gideon Parchomovsky & Michael Mattioli, *Partial Patents*, 111 COLUM. L. REV. 207 (2011).

[197] Patent Act, 35 U.S.C. §§ 112, 114.

respect that the subject matter they protect—data—would be different from the subject matter they disclose—methods.[198]

This hypothetical form of sui generis protection might be effective at encouraging some big data disclosures that would otherwise not be made. Data publishers have long demonstrated a desire for sui generis protection that would grant them greater control over downstream uses of their data.[199] Economic theory presented earlier in this Article indicates that in settings where publishers value such exclusivity more than they value exclusivity in their practices of data collection and organization, they would likely prefer dataright protection over trade secrecy.[200] By the same token, there would also be situations where this proposal would be unlikely to encourage new or valuable disclosures. Because this proposal offers publishers only an economic incentive, for instance, it would be ill-suited to encourage disclosure in settings in which privacy or strong commercial interests push toward secrecy.

In addition to its limitations, this proposal could face significant political challenges. Since the 1990s, Congress has regularly considered bills designed to provide sui generis protection for electronic databases.[201] Most of these proposals entailed a cause of action that database publishers could assert to prevent unauthorized copying.[202] The chief policy rationale behind these proposals was that because data is costly to gather

[198] In this respect, this proposal is very much like one that Gideon Parchomovsky and I dubbed the "semi-patent": a form of patent protection that would hinge on the publication of all research results that went into the development of the technology. Parchomovsky & Mattioli, *supra* note 196, at 208. The reach of intellectual protection is never perfectly coextensive with the degree of disclosure required. A patent may disclose a limited number of embodiments of an invention, for instance, and yet effectively capture a much wider range of subject matter. Similarly, a copyright covering a specific musical work may also cover similar works that embody similar themes.

[199] *Cf. Hearing*, *supra* note 194, at 67–68 (discussing how contracts provide inadequate protection that impacts downstream use of data).

[200] *See supra* Part I.C.

[201] Consumer Access to Information Act of 2004, H.R. 3872, 108th Cong. (2004); Database and Collections of Information Misappropriation Act, H.R. 3261, 108th Cong. (2003); Collections of Information Antipiracy Act, H.R. 354, 106th Cong. (1999); Consumer and Investor Access to Information Act, H.R. 1858, 106th Cong. (1999); Collections of Information Antipiracy Act, H.R. 2652, 105th Cong. (1998); Database Investment and Intellectual Property Antipiracy Act of 1996, H.R. 3531, 104th Cong. (1996).

[202] *E.g.*, Database Investment and Intellectual Property Antipiracy Act of 1996, H.R. 3531, 104th Cong. § 7 (1996) (providing injured database owners remedies for unauthorized copying by others).

and easy to copy, its collection requires the incentive of intellectual property-like protections.[203]

Although Congress considered at least six such proposals since 1996, none succeeded in garnering the necessary political support to become law. Leading commentators have cogently argued that poorly conceived sui generis database laws could chill socially and economically valuable uses of data. Most notably, Pamela Samuelson and Jerome H. Reichman have argued that proposals considered by Congress in the 1990s set a "new milestone for mischief" by overreaching the protection offered by traditional intellectual property laws.[204] By limiting scientific access to valuable data, they argued, such proposals threatened to "undermine the competitive ethos on which market economies depend."[205] Electronic database protection remains a contentious subject in intellectual property policy discourse.[206] Beyond these data-specific problems, sui generis proposals of all kinds arguably raise problems. As Mark Janis and Stephen Smith have explained, specialized forms of intellectual property protection designed around specific technologies tend to be inherently inflexible and can reduce the consistency and predictability of our system of intellectual property as a whole.[207]

The dataright described in this Part would fundamentally differ from earlier sui generis data protection proposals in important respects, however. A dataright would not entitle a data publisher to halt copying or distribution of its data. Instead, this right would be squarely aimed at unauthorized use of data. As such, this proposal would not limit the public's access to da-

[203] *See, e.g.*, 142 CONG. REC. 12,483 (1996) (statement of Hon. Carlos J. Moorhead) ("Information companies must dedicate massive resources to gathering and verifying factual material, presenting it in a user-friendly way, and keeping it current and useful to customers.").

[204] J.H. Reichman & Pamela Samuelson, *Intellectual Property Rights in Data?*, 50 VAND. L. REV. 51, 164 (1997).

[205] *Id.* at 163.

[206] Paula Baron, *Back to the Future: Learning from the Past in the Database Debate*, 62 OHIO ST. L.J. 879, 879 (2001) ("The appropriate form of legal protection for databases has been increasingly contentious since the early 1990s."); *see* Reichman & Samuelson, *supra* note 204, at 75–76 (describing "how radically the world intellectual property policymaking arena has changed").

[207] Mark D. Janis & Stephen Smith, *Technological Change and the Design of Plant Variety Protection Regimes*, 82 CHI.-KENT L. REV. 1557, 1560 (2007).

ta—the main source of "mischief" commentators cited in earlier data protection bills. Rather, the proposal would permit a data publisher to control downstream *use* (i.e., analysis) of its data for a limited time. Moreover, like other forms of intellectual property, a dataright would demand a valuable disclosure from its publisher. Data protection bills considered by Congress in the past entailed no such quid pro quo.[208] In this way, a dataright represents a novel balance between the necessary evil of exclusivity and the social benefits that can come from disclosure.

Setting political challenges aside, this proposal would entail some practical hurdles. The most significant would involve the risk of selective nondisclosure by data producers. Simply stated, data producers might elect to publish vague, incomplete, or inaccurate descriptions of their practices in order to receive protection. Although this risk is real, intellectual property law has long dealt with similar problems by imposing high penalties on rights holders. Under patent law, for instance, the doctrine of inequitable conduct provides that applicants who make factual misrepresentations to the USPTO during the prosecution process may have their patents invalidated.[209] A similar penalty of unenforceability would be appropriate under this plan.

Finally, alongside the benefits it could bring, this proposal would introduce new costs. Most likely, this plan would open the door to new litigation focused on two issues: whether a purported "use" of data constitutes infringement of a dataright, and whether a particular disclosure is sufficient to merit exclusivity in associated data. Although these challenges would be somewhat new for courts, patent law offers close parallels to these challenges, the resolution of which would involve similar questions of fact and of law. And of course, the provisioning of datarights would need to be overseen by a government institution with the expertise and competency to determine whether applications have adequately disclosed their methods.

This Part has explored what a special form of intellectual property protection adapted to the special challenges of big data would look like. Whether this proposal or one like it should be adopted into law is a conclusion that could only be drawn af-

[208] *See* bills cited *supra* note 201.

[209] U.S. PATENT & TRADEMARK OFFICE, MANUAL OF PATENT EXAMINING PROCEDURE § 2016 (9th ed. 2014).

ter careful discussion and debate among academics, lawmakers, data experts, and others who hold a stake in the exciting new frontier of big data. Ultimately, there may be no place for intellectual property based solutions to the big data disclosure problem. On the other hand, including data within the pantheon of protectable subject matter could yield economic and social benefits that outweigh the significant costs that such a step would necessarily entail.

CONCLUSION

This Article reveals that intellectual property law is not meaningfully encouraging producers of big data to disclose some of their most valuable practices to the public. This conclusion calls attention to a pressing policy problem. If big data practices remain undisclosed, innovation in this important field could languish.

This Article seeks to direct policy discourse toward the need to encourage greater disclosure of big data practices. There may be many ways to further this goal, such as new rulemaking within federal agencies, or perhaps a legislative change to intellectual property law. This Article explores this latter possibility by presenting a dataright as an exploratory device. By offering big data producers something new and valuable—an exclusive right to limit downstream use of their data—this new intellectual property right could encourage valuable technological disclosures that would otherwise remain shrouded in secrecy. This solution would carry substantial drawbacks however: it would encourage disclosure only in settings where data producers value data exclusivity more than they value secrecy in their methods. Moreover, this solution would entail significant new costs. Whether these costs would be outweighed by the plan's benefits would be a productive starting point for future discussion.

In light of big data's growing economic and social importance, policymakers and the public should be concerned with how our legal system will influence the production and use of this valuable new resource. Currently, our intellectual property system is not well configured to meet its goal of encouraging technological disclosures in this new frontier. Now is the time for policymakers and legal experts to explore solutions that will help us reap the full rewards of big data—for today, and for the vast and as yet undefined future.