Current Data Policy in Europe and the United States

Michael Mattioli*

Data Policy in the United States: New Challenges

I. Introduction

Scientific, economic, and cultural progress have long relied upon our ability to represent the world abstractly. The development of the number zero in ancient civilizations enabled practical calculations of the passage of time;[1] physicists in the 17th Century discovered that the motion of objects in the heavens and on Earth could be predicted by elegant equations;[2] by digitizing genetic information, researchers in the late 20th Century succeeded in mapping the human genome.[3] As our ability to represent the world has leapt ahead, so too has our understanding of it. According to experts in the growing field of data science, we are now on the cusp of yet another leap ahead. Thanks to a confluence of factors – most notably, cheap and accessible computing power, new computational techniques, and the proliferation of computers and sensors in the developed world – researchers and scientists believe that they may soon gain new insights that were out of reach just a few years ago. These practices, commonly called "Big Data," may have the potential to deliver better treatments for diseases, reduce the rate of crime, improve automobile safety, enable the development of artificially intelligent assistants, and countless other improvements to the human condition.[4]

---

* Associate Professor of Law, Indiana University Maurer School of Law (Bloomington, IN). I wish to express my deep thanks to Maximilian Becker for inviting me to participate in a conference he organized on this subject, held in Düsseldorf in February, 2017 ("Rights in Data: Industry 4.0 and the IP Rights of Tomorrow"). This essay is a loosely edited version of a talk I presented at that event.

1 Charles Seife, Zero: The Biography of a Dangerous Idea 17 (Penguin, 2000) (explaining the development of the number zero first as a placeholder and later as an integer with a numerical value appearing in ancient Mayan calendars); Robert Kaplan, The Nothing That Is: A Natural History of Zero 80–87 (Oxford, 2000) (discussing the role that zero played in the Mayan reckoning of time).

2 Brian Greene, The Elegant Universe: Superstrings, Hidden Dimensions, and the Quest for The Ultimate Theory 55 (Vintage, 2005) (Newton … wrote down equations that quantitatively describe the strength of the gravitational force between two objects … This 'law of gravity' can be used to predict the motion of planets and comets around the sun, the moon about the earth … as well as more earthbound applications.").

3 The Human Genome Project Completion, http://www.genome.gov/11006943 (citing June 26, 2000 as the initial date of completion) (last visited June, 2017). See also, Press Release, International Consortium Completes Human Genome Project, http://www.genome.gov/11006929 (last visited June, 2017).

4 See generally, Sugimoto et al., Big Data is Not a Monolith (MIT Press, 2016).

ZGE / IPJ 9 (2017), 299–316 DOI: 10.1628/186723717X15069451170883
ISSN 1867-237X

Whether these possibilities will become reality will depend not only upon the work of engineers and scientists, but also the decisions of policymakers. This is because data is a product of human work and oversight and its collection, maintenance, quality, and exchange are affected by diverse areas of law and policy. This essay provides a brief overview of the primary legal mechanisms in the United States that influence how data is used privately, and then introduces two new challenges that have recently emerged against this backdrop. This is a broad subject to cover in a relatively short essay. Instead of comprehensively cataloging every law and regulation that could affect the use of data, then, the discussion is organized around a theme: how effectively data is being exchanged and put to new and useful purposes under the current policy framework. In doing so, this essay aims to introduce readers unfamiliar with American law to some of the most important areas of law and policy that affect data gathering and exchange. Areas of US law not directly related to these transactional issues are outside the scope of this essay.

This essay begins by describing two broad categories of law and policy that affect data reuse and exchange: mechanisms that can protect the economic value of data, and rules that seek to protect private information held within data. This background helps to inform the discussion that follows, which outlines two interrelated problems. The first problem is that many of the companies and institutions that hold useful data have few incentives (and some disincentives) to carefully document and disclose information about its provenance and pedigree. This is problematic because often, to make good use of data – i. e., to reuse it – researchers must know how it has been collected, organized, manipulated, and so forth over time. Separately, recent evidence suggests that cooperative problems are preventing the useful exchange of data among multiple data-holders in some industries. This presents the new and troubling potential for a world in which the only firms capable of exploiting Big Data technologies are large, vertically-integrated corporations that possess vast internal troves of information (and as a result, do not need to rely upon exchanges with other companies to do useful things).

By understanding how these problems relate to the current legal framework in the US, it is possible to explore new laws and policies designed to improve the situation. If readers can take just one idea away from this essay, it is that Big Data presents policy challenges that are distinct in some important respects from those that dominated legal discourse in earlier decades. Both problems discussed here, for instance, stem from a unique hallmark of Big Data: its ability to reveal new and unexpected insights in old data that was gathered for no specific purpose.

II. Definitional Challenges

Because there is no single body of law pertaining to data in the United States, there is, inconveniently, no single legally meaningful definition of what data is. The patchwork of state and federal laws and regulations that affect data use are tailored to address specific problems and behaviors, rather than a monolithic type of subject matter. As a result, it is helpful to frame any discussion of "Data Policy" under US law with a statement of what sort of data is being discussed.

In this essay, the term "data" refers to information that is stored in a digital form, such that it can be retrieved and its meaning can be interpreted at some later time. Data originates from myriad sources: it may be generated and gathered by machines that contain sensors, such as smartphones or personal health devices; alternatively, it may be gathered by individuals making qualitative assessments, such as doctors or marketers. A database is "a collection, assembly, or compilation, of data and related works arranged in a systematic or methodical way."[5] Data is typically stored in binary form (i.e., 1's and 0's), and it represents higher-level information pertaining to nearly anything under the sun, from scientific measurements, to linguistic information, to pictures or sounds.

Colloquially, the word "data" has become synonymous with "fact," but the formal definition of data is far more expansive. A brief digression into semantics explains this point: As historian Daniel Rosenberg has observed, the word "data" is the plural form of the Latin noun, *datum*, which describes a given, or a premise from which assumptions can be drawn.[6] The word "fact," by contrast, comes to English from the Latin verb *facere*, meaning "to do."[7] Hence, facts are things that *have been done* in the past, while data are starting points for future inquiry.[8] Although Big Data practices sometime use factual data, they may also include estimates, approximations, predictions, abstractions, and even demonstrably false information, as examples discussed later in this essay explain.

An unfortunately imprecise term, "Big Data" can best be understood through several hallmarks. First, it describes a set of processes, and not simply a large volume of data as the term implies; Second, these processes enable scientists and engineers to draw new insights from large volumes of preexisting data; Third, these insights are often unexpected and unforeseeable by those collecting the data initially. As a corollary, the data that fuels these new processes is often col-

5 Borrowed from a US bill that was never passed into law. Database Investment and Intellectual Property Antipiracy Act of 1996, 1996 H. R. 3531 (1996).

6 Daniel Rosenberg, *Data Before the Fact*, *in* "Raw Data" Is An Oxymoron 15–40 (Lisa Gitelman, ed., 2013).

7 *Id.*

8 *Id.* See the discussion of the Feist case, infra, for a discussion of U.S. Copyright Law's application to factual material. Feist Publications, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340, 111 S. Ct. 1282, 113 L. Ed. 2d 358 (1991).

lected with no specific end-use in mind. Because these practices rely heavily on vast computing power and diverse data, Big Data is a relatively recent development. Despite this newness, however, experts believe this phenomenon has the potential to bring about innovation and improvements on a scale similar to the Industrial Revolution of the 19th Century.

III. Protecting the Economic Value of Data

Economists have long understood that information that is costly to gather and cheap to copy is often subject to underproduction. The Nobel-prize winning economist, Kenneth Arrow, famously identified a paradox related to this problem: it is sometimes impossible to negotiate a sale for information without disclosing the information itself, thus lowering its value to zero, negating the need for a sale, and in the long term, suppressing incentives to gather the information in the first place. To address this problem, some countries have enacted *sui generis* intellectual property protection designed to protect the economic investment necessary to assemble databases. The European Union's Database Directive is a widely-known example.[9] As the following paragraphs explain, no such law exists in the United States. Formal intellectual property protection for data, meanwhile, is thin. As a result, companies and institutions that gather and organize useful data must rely on other mechanisms to prevent unwanted disclosures.

The US Supreme Court first addressed this problem in the venerable case of International News Service v. Associated Press. The dispute concerned a type of information that has been valued long before the advent of Big Data: news. Since the late 19th century, the United States has occupied four time zones – "pacific," "mountain," "central," and "eastern." In 1918, the six-hour time difference between cities on the east coast (eastern) and those on the west coast (pacific) gained special importance for newspaper publishers. This was because of where the news of greatest concern was originating: The European battlefront. The Associated Press (AP) and International News Service (INS) were competing American news agencies that employed journalists to collect the news, and then shared those reports with affiliate local newspapers around the US. AP reporters transmitted news of important events to the agency's central office, which would in turn send news stories to its 950 affiliate newspapers around the country.

INS seized upon an opportunity: Instead of paying journalists to gather news from Europe, the company waited for AP affiliates to publish stories on the East Coast, and then wrote original news reports based on the same underlying facts, which they telegraphed to INS affiliates elsewhere in the country. Thanks to the time difference, customers in California who purchased newspapers published

9 Directive 96/9/EC.

by INS affiliates could learn about the same events that customers of AP affiliates received. The Associated Press sued.

The Supreme Court recognized that this dispute presented the classic free-rider dilemma mentioned earlier – i. e., information that is costly to gather and cheap to reproduce will tend not to be gathered unless the information-gatherer has a means to prevent unwanted copying. Inconveniently for the Associated Press, although copyright protects expressions of facts, such as the specific words used in a news story, it does not apply to what the INS took without permission: the underlying facts. "The information respecting current events is not the creation of the writer," the Court explained, "but is a report of matters that ordinarily are *publici juris*; it is the history of the day."

As the Court saw it, the problem couldn't be resolved so simply, however. The court reasoned that news is valuable to the public and unlikely to be gathered robustly unless agencies like the Associated Press had some way to draw a return from their investment. Because the US had no law in place to protect the economic investment in newsgathering, the Court conceived of one: the Court declared that a "quasi-property" interest exists in freshly-gathered news. Weaving together concepts from intellectual property law and unfair competition, the Court explained that this "Hot News" doctrine worked as follows: news-gatherers such as the Associated Press would be entitled to prevent their competitors from copying the facts in a news story for a limited time. That period, the Court explained, would begin the moment a news story was gathered and would extend to the time at which it was no longer commercially valuable. In the sense that it offered exclusivity, it was something like an IP-like right, but one that could be asserted only against certain defendants (competitors) and the boundaries of which would be defined by market factors. Although the Hot News doctrine presents interesting and important theoretical questions, it has been invoked by courts only rarely since 1918.[10]

A second helpful stepping stone to understanding how US copyright law relates to data is the venerable 1991 decision of *Feist Publications v. Rural Telephone Service*.[11] The dispute arose when Feist, a telephone book publisher, copied a set of phone listings originally collected and published by Rural, a small telephone company located in Kansas. Rural sued Feist for copyright infringement. The US Supreme Court began its decision by noting – as it did in *INS* – that facts alone are not copyrightable.

Rural argued, however, that although the phone numbers Feist had copied were not individually copyrightable, Rural's overall section and arrangement of the phone numbers was. The Court recognized that copyright protection could,

10 Shyamkrishna Balganesh, *'Hot News': The Enduring Myth of Property in News*, 111 Colum. L. Rev. 419 (2011).

11 *Feist Publications, Inc., v. Rural Telephone Service Co.*, 499 U. S. 340 (1991).

in theory, extend to databases. "Factual compilations," it wrote, "may possess the requisite originality. [...] These choices as to selection and arrangement, so long as they are made independently by the compiler and entail a minimal degree of creativity [...]," can merit copyright protection.[12] Ultimately, however, the Court was unconvinced that Rural had exercised this sort of judgment in selection or arrangement. The directory was a complete listing of phone numbers listed in alphabetical order, based upon last names. The Court called this arrangement "obvious," and held that it did not possess sufficient originality – often called the *sine qua non* of copyrightability – to qualify for protection. "[C]opyright assures authors the right to their original expression," the Court wrote, "but encourages others to build freely upon the ideas and information conveyed by a work. This principle, known as the idea/expression or fact/expression dichotomy, applies to all works of authorship."[13]

Just two years later, the case of *Key Publications v. Chinatown Today* revealed that the bar for copyright protection in a compilation of information is quite low.[14] The defendant in that dispute had copied a phone directory of businesses in New York City's Chinatown without permission, and the Court of Appeals for the Second Circuit held that the original publisher's copyright in the selection and arrangement of the phone numbers had been infringed. Unlike Rural's phone directory, however, the telephone directory in *Key Publications* had not been merely an alphabetical listing of telephone numbers. Rather, it was a collection of telephone numbers organized by business type and selected according to the publisher's judgments. The court wrote, "selection implies the exercise of judgment in choosing which facts from the given body of data to include in a compilation."[15]

Along with original compilations, estimates constitute a distinct form of factual subject matter that has received copyright protection. In the 1994 decision of *CCC Information Services v. Maclean Hunter*, the Supreme Court determined that individual estimates of used car prices published by the plaintiff were "neither reports of historical prices nor mechanical derivations of historical prices," but rather, were "original creations" for purposes of copyright.[16] Based on this, the Court regarded them as copyrightable expressions. The Court based this conclusion on its finding that the "predictions were based not only on a multi-

12 *Id.*

13 *Id.* For a deeper discussion of these concepts, see Alan L. Durham, *Speaking of the World: Fact, Opinion and the Originality Standard of Copyright*, 33 Ariz. St. L. J. 791, 848 (2001).

14 945 F.2d 509 (2d Cir. 1991).

15 *Id.*

16 *CCC Info. Servs., Inc. v. Maclean Hunter Mkt. Reports, Inc.*, 44 F.3d 61, 67 (2d Cir. 1994). The merger doctrine, discussed later in this essay, was considered by the court in light of the alleged copying. Id. at 72. ("In this circuit, consideration of the merger doctrine takes place in light of the alleged copying to determine if infringement has occurred, rather than in analyzing the copyrightability of the original work.").

tude of data sources, but also on professional judgment and expertise." Five years later, the Court consistently ruled that a published set of collectible coin prices was imbued "with sufficient creativity and originality to make [it] copyrightable" because the estimated prices were derived from "considerable expertise and judgment."[17]

An important potential barrier for copyright protection over data is a longstanding rule called the "Merger Doctrine." This doctrine renders ineligible for copyright protection any otherwise original work that expresses an idea that can be expressed only in a limited number of possible ways. The rule has been applied in some noteworthy disputes concerning data in the United States. In the 2007 decision of *New York Mercantile Exchange v. IntercontinentalExchange*, for instance, the Court of Appeals for the Second Circuit applied the merger doctrine to bar copyright protection to a listing of market values for commodities contracts.[18] The court explained that the market values at issue were "economic fact[s] about the world," rather than estimates or predictions.

The forgoing discussion offers a few important insights into the level of exclusivity that US copyright law affords companies that gather and publish data. To the extent that an individual piece of data (a datum) is factual in nature – i.e., either true or held out as true – it is not protectable under copyright law.[19] Copyright protection *can* apply, however, to sets of data that result from originality in the act of selection or arrangement. As a practical matter though, such protection is unlikely to be very useful to many data publishers: electronic data is usually organized in a systematic (unoriginal) manner, and even when it is not, the scope of protection copyright provides is thin – i.e., unlicensed reproductions of the original data can avoid infringement by filtering or arranging the same data in different ways. Stated differently and as I have written elsewhere, it is sometimes relatively easy "to steal the tiles without copying the entire mosaic."[20]

Because of the relatively thin protection copyright provides, data gatherers often rely on alternative means of preventing copying of their data. The most common options are: trade secrecy (codified in state and federal statutes), the law of contracts (primarily operating at the state-level), and practical measures,

17 *CDN Inc. v. Kapes*, 197 F.3d 1256, 1260 (9th Cir. 1999) ("The evidence indicates that the plaintiff uses its considerable expertise and judgment to determine how a multitude of variable factors impact upon available bid and ask price data … As such, the Court finds that these prices were created, not discovered.").

18 *New York Mercantile Exchange v. IntercontinenalExchange*, 497 F.3d 109 (2d. Cir. 2007).

19 U.S. Courts seldom delve into precisely defining what facts are, but there is a line of cases that distinguish "hard facts" (statements or information that represent verifiable truths about the world) from "soft facts" (information that resides more in the realm of interpretation). See, e.g., Speaking of the World: Fact, Opinion and the Originality Standard of Copyright *in* Intellectual Property Protection of Fact-based Works: Copyright and Its Alternatives (Robert F. Brauneis ed., 2009).

20 Michael Mattioli, *Disclosing Big Data*, 99 Minnesota Law Review 525 (2014).

such as encryption. In addition, tort law (primarily common law operating at the state-level) and/or criminal law (primarily codified in state or federal statutes) may deter the theft or vandalizing of physical devices that contain data.

Until very recently, trade secret law in the United States has existed in the form of statutes passed by states, rather than by the federal government. Efforts to harmonize trade secret laws among the states have resulted in a uniform law of trade secrecy that has been adopted by 47 of the 50 states. In 2016, a second level of protection came in the form of the Defend Trade Secrets Act, a federal law that provides a cause of action for misappropriation of trade secrets.[21] At the state or federal level, the threshold requirements for a claim for misappropriation of trade secrets are fairly straight forward: the plaintiff must show that the defendant has disclosed or used information that is valuable because it is secret, and for which the trade secret holder has taken reasonable measures to protect. Significantly, trade secrecy may attach to information of all kinds – a far broader scope of protection than copyright law offers. For firms that gather or use valuable data, this might include not only the data itself but also related methods of data gathering or preparation.

Contracts are a second important mechanism that data holders use to prevent unwanted disclosure and copying of valuable information. Here, the seminal case of *ProCD v. Zeidenberg*, which was decided by the United States Court of Appeals for the Ninth Circuit in 1996, is a helpful starting point.[22] Perhaps coincidentally, this seminal decision also concerned telephone numbers: *ProCD* had collected and compiled telephone listing information from thousands of phone directories into a computer database. The company sold sets of CDs containing this database to two tiers of customers: an inexpensive version directed toward individual users, and an expensive version intended for business users. A contract contained inside of the box in which the CDs were sold (a so-called "shrink-wrap license") required purchasers of the "individual-use" version of the product to promise not to use their copy for business purposes. Directly disobeying this provision, *Zeidenberg* copied the information on an "individual-use" version of the product into an online database that he charged subscribers access to. He argued, however, that the contract should not be enforceable because he had no opportunity to read it before purchasing the product: although the box indicated that the product was intended for personal use, the contract itself was enclosed in the box. The Supreme Court disagreed, holding that shrink-wrap licenses like the one *ProCD* included in their box are enforceable contracts, even when the party accepting them has not read them. The decision reflected a policy determination that the efficiencies that can come from attaching licensing terms to

21 S. Rep. No. 114–220, at 3 (2016), available at https://www.congress.gov/congressional-report/114th-congress/senate-report/220/1).

22 *ProCD, Inc. v. Zeidenberg*, 86 F.3d 1447 (7th Cir. 1996).

products ("form licenses") tend to outweigh the potential harm that such contracts might do to consumers.[23] From the perspective of database publishers, the lesson of *Zeidenberg* was clear: shrink-wrap licenses can discourage certain uses of data that copyright law fails to discourage. Unlike formal intellectual property protection, however, breach of contract claims can only reach parties to a contract and not the world at large.[24]

It was against this backdrop that, in the 1990s, a push developed in some quarters for *sui generis* data protection. (This was part of a broader trend in intellectual property law at the time, under which circuit designs, plant varieties, and other specific subject matter was granted intellectual property-like protection through legislation.) Although a parade of bills for *sui generis* data protection came before Congress – at least seven geared toward protecting databases – none garnered the necessary political support to the passed into law.[25] Database publishers lobbied for these bills for the simple reason that they wished for more robust means of protecting the large investments they made in collecting and distributing valuable data. Prominent American academics voiced strong concerns, however, that such laws would limit scientific research and limit follow-on uses of data.[26] There were also constitutional concerns related to intellectual property and specific concerns with statutory language that had been presented. (The definition for the term database in at least one bill was confusingly broad, for instance.) During this period, unsuccessful alternate proposals were suggested as well, some of which would have provided liability-like penalties for the unpermitted use data. Readers familiar with data protection law in Europe, especially in light of recent developments in Big Data, may be surprised to learn that sui generis data protection is not a central theme of policy work or legal discourse in the US today. Prominent scholars in the United States are, of course, conducting important scholarship related to Big Data and the industries that this phenomenon connects with – e.g., so-called "Internet of Things," machine learning, etc. Presently, however, the greatest focus in this area relates to privacy and cybersecurity.

---

23 As a practical matter, consumers who are victims of certain harmful practices may have alternative avenues of recourse outside of claims for breach of contract.

24 This difference is sometimes referred to "relative rights" vs. "absolute rights" or the relativity of contract law.

25 See, e.g., 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). Database Investment and Intellectual Property Antipiracy Act of 1996, H.R. 3531, 104th Cong. § 7 (1996).

26 See, e.g., J.H. Reichman & Pamela Samuelson, *Intellectual Property Rights in Data?*, 50 Vand. L. Rev. 51 (1997).

IV. Protecting Privacy

A second branch of laws that affect the reuse of data stem from privacy. The notion of privacy as a fundamental right is relatively new in the United States, having first been proposed by Samuel Warren and Louis Brandeis in an 1890 article published in the Harvard Law Review.[27] Although there remains no monolithic data privacy law, a constellation of state and federal laws that relate to privacy affect how data may be collected, exchanged, and reused in myriad ways. Apart from legislation, in numerous decisions, the Supreme Court has also held that the Constitution protects citizens from certain invasions of privacy committed by the government. In addition, several federal agencies regulate behavior harmful to privacy. The Federal Trade Commission (FTC), for instance, protects consumers from deceptive or unfair practices of companies, including the wrongful collection and use of data.[28] One layer down, individual states have enacted specific privacy laws as well.[29]

A recent episode offers a helpful view of how the FTC limits the collection and use of data. In a recent dispute, the commission sued the television manufacturer Vizio in District Court, alleging that Vizio had been secretly monitoring television viewers' habits without properly notifying them. Vizio accomplished this, the FTC charged, by analyzing the digital signal originating from cable boxes and entering their televisions' "HDMI" video input ports. As the complaint alleged, "on a second-by-second basis, Vizio collected a selection of pixels on the screen that it matched to a database of TV, movie and Commercial contents." This fine-grained viewing information,[30] which had collected in over 11 million televisions since 2010, was highly valuable to advertisers. In February of 2017, the district court that decided this dispute devised a specific remedy under which Vizio was required to "Prominently disclose to the consumer, separate and apart from any privacy policy, terms of use page, or other similar document: (1) the types of viewing data that will be collected and used, (2) the types of viewing data that will be shared with third parties; (3) the identity or specific categories of such third parties; and (4) all purposes for defendants' sharing of such information." In the end, Vizio agreed to pay 2.2 million dollars to settle lawsuit.

Although a comprehensive study of all state and federal statutes that apply to privacy is beyond the scope of this essay, it is helpful to briefly mention some of the most common forms such laws take. Many states require companies that hold personal data to report privacy violations to the individuals affected. California law requires, for example, that "any person or business that conducts business in

27 Samuel Warren & Louis Brandeis, The Right to Privacy, 4 Harvard L. R. 193 (1890).

28 FTC Act, 15 U.S. C. § 41 et seq.

29 Cal. Govt. Code § 6267.

30 Maximillian Becker's insightful argument for "data-avoiding products" suggests a new solution to this problem that consumers and industry might prefer, see in this issue p. 371.

California, and that owns or licenses computerized data that includes personal information, shall disclose a breach of the security of the system following discovery or notification of the breach."[31] Some states have enacted laws that set minimum technical standards that must be met by companies that hold personal information relating to residents of the state. In Massachusetts, for instance, computer systems that hold such data must possess: "(1) Secure user authentication protocols [such as secure methods of selecting passwords and the use of token devices]; [...] (3) Encryption fall transmitted records and files containing personal information that will travel across public networks [...] (7) Reasonably up-to-date versions of system security agent software which must include malware protection."[32]

At the federal level, privacy legislation pertaining to data follows a similar piecemeal pattern, applying to specific uses of certain data or specific relationships – e. g., between citizens and the government. The Privacy Act of 1974, for instance, regulates the US Government's collection, maintenance, use, and disclosure of personally identifiable information about individuals. This includes "any records contained in a system of records by any means of communication many person, or to any other agency, except to request by, or with the prior written consent of, the individual to whom the record pertains." The law contains a number of exceptions, including provisions that permit the US government to use data in order to analyze census information, to help assemble archives of historical value, or to aid law enforcement.[33]

Another important federal law affecting individual privacy is the Health Insurance Portability and Accountability Act of 1996 (HIPAA). This law governs how healthcare providers and other entities may use and disclose personally identifying information. Under this law, such information falls into specific, narrow categories: patient names, zip codes, treatment dates, and other pieces of potentially identifying information, for instance.

Other laws that relate to data privacy include the Electronic Communications Privacy Act of 1996, which sets out procedures (e. g., search warrants) the government must follow when gathering certain types of data on individuals; The Fair Credit Reporting Act of 1970, which relates to the privacy of certain data maintained by consumer reporting agencies; The Children's Online Privacy Protection Rule, which sets out requirements for how online services may collect and use certain information relating to children; The Financial Modernization Act of 1999 (also known as Graham-Leach-Bliley Act), which requires certain financial

31 Cal. Civ. Code § 1798.82.

32 Mass. Gen. Laws Ann. ch. 93H, § 1; Mass. Regs. Code tit. 201, §§ 17.01 et seq., Standards for The Protection of Personal Information of Residents of the Commonwealth.

33 Cite to 2017 Whitehouse executive order: "In January of 2017 the White House issued an executive order quote enhancing Public Safety in the interior of the United States and quote section 14."

service providers to explain how they collect and use consumer data. These federal laws may sometimes preempt similar state-level legislation.[34]

Finally, it is important to appreciate that form contracts attached to or included within products and services may shrink the effective reach of some laws relating to privacy. Online services widely require new users to accept end-user license agreements which often include provisions that address private information. These provisions may include terms that permit a website or app to use data in certain ways that would otherwise result in liability. As mentioned in the discussion of the *Zeidenberg* case earlier, form contracts are generally enforceable. Even so, an aggrieved class of consumers may succeed in challenging the enforceability of such terms based on a theory of, for instance, unconscionability.

V. New Challenges

New challenges for data reuse and data exchange have emerged against the patchwork of laws that pertain to data. To appreciate these problems, it is helpful to consider how the data publishing landscape has changed since the 1990s. In those days, before the proliferation of personal computers or the internet, the business of selling and licensing data was relatively straightforward: data publishers gathered useful information, which they sold or licensed access to customers. The economic value in data today is far more complex. As mentioned earlier in this essay, Big Data researchers use data initially gathered indiscriminately and for no specific purpose. Relatedly, experts believe that the greatest promise of Big Data requires, as a precondition, the pooling of heterogeneous data from multiple sources.

This new economic picture presents some special problems. To intelligently use data that has been gathered automatically and indiscriminately, it is often important to understand where the data initially came from, how it was gathered, any changes that were made to the data along the way, and so forth. Considering the many sources of data that might be useful to researchers, and the many circumstances under which such data has been collected and organized, it is not surprising that there is no universal approach or even widely accepted best practices for documenting and disclosing such metadata. It seems unlikely that market forces will address this problem: much of the data being used by contemporary data scientists has not been gathered as an asset at the outset. Rather, such information is often gathered automatically (e. g., by sensors) and only later does its value come to light. This connects loosely to the second problem: powerful disincentives discourage companies from sharing useful data with one another.

34 See, e. g., Ohio Rev. Code Ann. § 1349.19 ("This section does not apply to any person or entity that is a covered entity as defined in 45 C. F. R. 160.103, as amended.").

Some of these forces are rooted in law, while others are woven into culture and appear to be deeply contextual.

1. The Data Disclosure Problem

Big Data is powerful, but it does not speak for itself. Because leading Big Data sources such as online searches, social media posts, credit card transactions, and mobile phone locations are initially unstructured, they must sometimes be carefully filtered, organized, and sometimes even altered by experts before they can be used.[35] These methods of data organization are far more important to the new breed of Big Data companies than to publishers of conventional compilations of facts and statistics. As the author of a recently published book on the subject explains, "[t]raditional 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."[36] For these reasons, commentators from the fields of Computer Science and Informatics have cautioned against blind reliance on unstructured data. Danah Boyd and Kate Crawford, leading Big Data scholars have warned that "[l]arge data sets from Internet sources are often unreliable, prone to outages and losses, and these errors and gaps are magnified when multiple data sets are used together."[37]

The need for experts to give Big Data meaning and form raises two pressing questions: first, whether inadequate disclosure of information about data provenance and pedigree significantly limits its reuse, and by extension, the development of new and important Big Data applications; second, whether policymakers should encourage more robust disclosure of such information.

In a recent publication, I analyzed these questions by conducting a set of ethnographic case studies into how data is used by companies at the vanguard of the Big Data phenomenon.[38] A central theme that emerged from the investigation is that data is often creatively manipulated prior to publication.[39] These manipula-

---

35 Mayer-Schönberger & Cukier at 32 (explaining that big data is inherently unstructured); Bill Franks, Taming The Big Data Tidal Wave 20 (Wiley, 2012) ("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.").

36 *Id.* at 21.

37 Danah Boyd & Kate Crawford, Critical Questions for Big Data, 15 Info. Comm. & Soc'y 662, 668 (2012).

38 See above Fn. 20.

39 This observation has been made by other scholars as well. See, e.g., Alan L. Durham, *Speaking of the World: Fact, Opinion and the Originality Standard of Copyright*, 33 Ariz. St. L.J. 791, 839 (2001) ("This is not to suggest that the reported population is nothing more than the census taker's fantasy, or that one figure cannot be more accurate than another, but any census data is, at least, the product of objective reality and subjective decisions rendered by the census taker.").

tions can be grouped into four broad categories: the filtering of useless noise from databases, finding and fixing gaps in the data by way of interferences, hiding or "masking" private information, and classifying data.

An overarching theme of my studyis that data is often imbued with myriad subjective judgments. One company I examined, for instance, defined, identified, and excised spam and other unwanted commercial content from Twitter datafeeds on an ad hoc basis. Two other companies employed data scientists who designed methods of selecting social media posts that customers might find helpful. One data scientist described how datasets may often include guesses: when he noticed that the biological sex of certain hospital patent records in a database had been omitted, he entered his guesses about such information by, for instance, looking to factors such as patient height, weight, and ailments. In a similar way, data scientists explained how the common practice of manipulating data to preserve consumer and patient privacy in accordance with federal laws.

The fact that data is often manipulated and infused with human judgment is not necessarily troubling in itself. To the contrary, the data scientists and companies I interviewed very carefully considered how their judgments affected the data. What *is* problematic, however, is the fact that the grand vision of Big Data espoused by its supporters involves a world in which data is used and reused, combined and disaggregated, endlessly offering new insights. Such fluidity is impossible, however, when information relating to the limitations of the underlying data has not been disclosed. A researcher who wishes to study, generally, the overall prevalence of cancer in the United Statesmay not care that the biological sex of some of the patients in her database is incorrect; a researcher who wishes to study the prevalence of cancer *in men* in the United States, however, would be deeply concerned by the possibility of the same error. The limits of Big Data, it seems, are defined by the limits of disclosure.

2. The Data-Pooling Problem

A second problem relates to data exchange. Economists and scholars of industrial organizations have long recognized that innovation can arise from pools of industrial and technical knowledge from diverse areas. According to experts, the greatest promise of Big Data lies, similarly, in the aggregation of data from multiple sources. There are many ways that data can be aggregated: a scientist might seek to license data from multiple publishers or data holders; a corporation could acquire or merge with another company that holds valuable data. According to experts in several industries increasingly drawing upon data, however, the most effective way to bring about innovation in the field of Big Data is through the creation of pools – institutions that aggregate and license-out sets of data initially gathered by member companies and institutions. This approach is already gained widespread support from experts in the world of cancer treatment.

Prompted by recent concerns that such "data pools" are not forming quickly or broadly enough, I recently conducted a study examining efforts to pool privately held data related to cancer treatment.[40] Through a set of interviews with experts from academia, the drug research industry, and the government, I identified several barriers to the aggregation of such data. Some of these barriers were unsurprising: many hospitals and doctors, for instance, explained that laws designed to protect patient privacy – most notably HIPAA – created an upfront liability risk for any healthcare provider that wished to share data. Offsetting this risk requires hiring experts to "mask" data that might reveal patient information. Research subjects reported that because such work requires deep expertise and cannot be completely automated, it is often expensive. Similarly, experts cited the lack of standards for medical data and a proliferation of proprietary formats as problems for data aggregation. Before contributing data to a pool, a healthcare provider would need to overcome these barriers – a potentially costly endeavor with unclear benefits.

The study also revealed some surprising barriers to the aggregation of data, at least in the realm of cancer research. Hospitals expressed the concern that sharing patient data with a pool could lead to the widespread disclosure of negative information about the hospital's track record for care. Pharmaceutical companies, meanwhile, explained that sharing too much information relating to clinical trials might give their competitors too much information about research methods, business plans, or information they would prefer to maintain as trade secrets. In a similar vein, academic researchers explained that the disclosure of valuable research data could jeopardize their individual opportunities to receive grants and promotions in the future.

3. Future Directions for Policy work

Together, the data disclosure problem and the data-pooling problem reveal some of the limitations of current US law and policy. The disclosure of information concerning data provenance and pedigree is not inherently "an intellectual property problem," of course, but these are problems that may be of concern to a body of law traditionally concerned with disclosure of technological information. Much like data, information describing how data has been collected, manipulated, and organized, is factual in nature. As a result, such information not qualify for copyright protection under US law. These steps entail processes and methods, however, which might prompt the question of whether they are patentable (and by extension, whether patent law might encourage their disclosure). Based on the accounts of engineers and experts that I have interviewed in the course of my work, the answer to this question is "often, no." Many of the techniques that data

40 Michael Mattioli, *The Data-Pooling Problem*, Berkeley Tech. L. J. (forthcoming, 2017).

scientists use to manipulate data are widely known and documented and as such, would not cross US Patent Law's threshold requirements for novelty and non-obviousness. Other techniques are entirely subjective and applied on an ad hoc basis. Finally, although software is patentable, recent Supreme Court jurisprudence has cast a degree of doubt upon the patentability of algorithms.[41] As such, they would be unlikely to meet US Patent Law's utility requirements. It appears that the most robust protection for information concerning data provenance and pedigree is the law of trade secrets. This helps to explain why intellectual property law does not encourage the disclosure of such information.

The data-pooling problem helps to highlight a different set of limitations. Perhaps most starkly, the problem reveals how laws related to privacy (e. g., HIPAA) can discourage and limit exchanges of data that might hold great social value. This is not to suggest that privacy laws should be weakened in any way. It is helpful to consider, however, that most US laws designed to protect personal privacy were enacted before the age of the Big Data. As such, these laws do not reflect a policy judgment that the social value of Big Data may, in some cases, outweigh countervailing concerns regarding privacy.

This essay is primarily descriptive in nature and does not aim to offer prescriptions. It is helpful to briefly consider if now could be a helpful time for policymakers to reconsider the reach of the privacy law as it applies to data-gatherers geared toward human health and safety. Perhaps in certain limited circumstances such as cancer research, the benefits of robust privacy protections do not outweigh the benefits that could come from greater exchanges of data. At the same time, it would be helpful to consider other areas of law and policy unrelated to privacy that might address the data-pooling problem. The government could, for instance, encourage the use of standards – i. e., related to data collection methods and storage formats. This step alone would likely reduce some of the upfront costs of aggregating data into pools. The US government could also offer myriad incentives (e. g., tax incentives, vouchers for expedited FDA review of drugs and medical devices) designed to encourage companies and institutions to exchange data. I explore these proposals and others in my recent article, The Data-Pooling Problem.

VI. Conclusions

Some of the most interesting challenges surrounding Big Data stem from one of its hallmarks: the ability to reveal new, unexpected, and even unforeseeable insights from data that was initially generated and gathered automatically and indiscriminately.

41 See *Alice Corp. Pty. v. CLS Bank Int'l*, 134 S. Ct. 2347, 189 L. Ed. 2d 296 (2014) (instructing that software claims that cover only abstract ideas and do not include any "inventive concept" are unpatentable).

For data gatherers, this setup presents upfront costs distinct from the costs that have dominated data and information policy debates in the US in the past. The core concern when *INS v. AP* was decided, for instance, was whether adequate incentives existed for publishers to do the hard work of *collecting* useful data. Eighty years later, the *Feist* case presented essentially the same core question. Today, the upfront costs of data-gathering seem to raise fewer concerns. In part, this is because data scientists who specialized in Big Data techniques often draw upon data that is generated automatically – electronic logs of online searches and transactions, data produced by doctors in the normal course of caring for them, electronic logs created by the multitudes of small, cheap, sensors in medical devices, smartphones, automobiles, utility systems, and so forth. Collecting useful data no longer seems to be the central problem.

The fact that data is generated automatically does not mean that it is useful, however. As mentioned, a hallmark of Big Data is that it often involves probing old data for new insights. Data scientists cannot intelligently ask new questions of old data without first understanding the characteristics of that data, including the limits of what it can describe. Contrary to the way it is often discussed, data is not a natural resource just waiting to be extracted from the world; it is, rather, often inevitably infused with subjective human judgments – e.g., decisions concerning data filtering, classifying, masking, or cleaning. This subjectivity can exist even when data is gathered by sensors or other automated systems: a thermometer, for instance, may produce accurate data that future data scientists would nevertheless be unable to use without some other piece of metadata – the specific characteristics of the device, whether it was situated indoors or outdoors, and so forth. Ironically, this helps explain why market demand does not appear to be encouraging the disclosure of such information. In a world where the potential future uses of data are largely speculative, companies and institutions that have ability to disclose useful information about data provenance and pedigree have relatively weak incentives to do so. The immediate costs are clear and the potential gains are not.

The unforeseeable nature of Big Data also seems to discourage its exchange. Unsure of what data they share might reveal in the future, many companies and institutions appear to feel rationally concerned that a dataset that seems innocuous today will be damaging in unexpected ways tomorrow. This may explain why data pools – institutions that some experts believe would ideal platforms for future innovation – are not taking form in great numbers or in many industries. There is no reason to guess that this problem is limited to the area of patient treatment data.

Because the constellation of laws and policies that relate to data disclosure and exchange in the United States took form before the unique problems raised by Big Data arose, now seems a good time to assess how well the legal framework is working, and if more should be done. Until then, the companies that stand to

benefit from Big Data may be only those that are large enough to assemble their own proprietary data silos.

Zusammenfassung

Dieser Aufsatz gibt einen kurzen Überblick darüber, wie die rechtlichen Rahmenbedingungen in den Vereinigten Staaten von Amerika den privaten Umgang mit Daten beeinflussen. Darauf aufbauend wird auf zwei aktuelle Probleme in Bezug auf Big Data eingegangen: Das erste Problem besteht darin, dass für viele Unternehmen und Institutionen, die Zugriff auf nützliche Daten haben, nur wenig positive, zum Teil sogar negative Anreize existieren, die Herkunft derartiger Daten offenzulegen. Das zweite Problem ist, dass kooperative Hemmnisse zwischen zahlreichen „data-holders" eine sinnvolle Aggregation von Daten (z. B. pooling) verhindern. Sofern die politischen Entscheidungsträger diese Probleme und deren Zusammenhänge mit den derzeitigen Rahmenbedingungen des US-amerikanischen Rechts verstehen, können möglicherweise neue politische Lösungen für diese Probleme gefunden werden.