MIS Quarterly Executive

Ethical Issues in the Big Data Industry

Big Data combines information from diverse sources to create knowledge, make better predictions and tailor services. This article analyzes Big Data as an industry, not a technology, and identifies the ethical issues it faces. These issues arise from reselling consumers' data to the secondary market for Big Data. Remedies for the issues are proposed, with the goal of fostering a sustainable Big Data Industry.[1,2]

Kirsten E. Martin
George Washington University
(U.S.)

The Big Data Industry

Big Data receives a lot of press and attention—and rightly so. Big Data, the combination of greater size and complexity of data with advanced analytics,[3] has been effective in improving national security, making marketing more effective, reducing credit risk, improving medical research and facilitating urban planning. In leveraging easily observable characteristics and events, Big Data combines information from diverse sources in new ways to create knowledge, make better predictions or tailor services. Governments serve their citizens better, hospitals are safer, firms extend credit to those previously excluded from the market, law enforcers catch more criminals and nations are safer.

Yet Big Data (also known in academic circles as "data analytics") has also been criticized as a breach of privacy, as potentially discriminatory, as distorting the power relationship and as just "creepy."[4] In generating large, complex data sets and using new predictions and generalizations, firms making use of Big Data have targeted individuals for products they did not know they needed, ignored citizens when repairing streets, informed friends and family that someone is pregnant or engaged, and charged consumers more based on their computer type. Table 1 summarizes examples of the beneficial and questionable uses of Big Data and illustrates the

1 Dorothy Leidner is the accepting senior editor for this article.

2 This work has been funded by National Science Foundation Grant #1311823 supporting a three-year study of privacy online. I wish to thank the participants at the American Statistical Association annual meeting (2014), American Association of Public Opinion Researchers (2014) and the Philosophy of Management conference (2014), as well as Mary Culnan, Chris Hoofnagle and Katie Shilton for their thoughtful comments on an earlier version of this article.

3 Both the size of the data set, due to the volume, variety and velocity of the data, as well as the advanced analytics, combine to create Big Data. Key to definitions of Big Data are that the amount of data and the software used to analyze it have changed and combine to support new insights and new uses. See also Ohm, P. "Fourth Amendment in a World without Privacy," *Mississippi. Law Journal* (81), 2011, pp. 1309-1356; Boyd, D. and Crawford, K. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon," *Information, Communication & Society* (15:5), 2012, pp. 662-679; Rubinstein, I. S. "Big Data: The End of Privacy or a New Beginning?," *International Data Privacy Law* (3:2), 2012, pp. 74-87; and Hartzog, W. and Selinger, E. "Big Data in Small Hands," *Stanford Law Review Online* (66), 2013, pp. 81-87.

4 Ur, B. et al. "Smart, Useful, Scary, Creepy: Perceptions of Online Behavioral Advertising," presented at the Symposium On Usable Privacy and Security, July 11-13, 2012, Washington, D.C. See also Barocas, S. and Selbst, A. D. "Big Data's Disparate Impact," 2015, draft available at SSRN 2477899; and Richards, N. M. and King, J. H. "Three Paradoxes of Big Data," *Stanford Law Review Online* (66), 2013 pp. 41-46.

Table 1: Examples of Beneficial and Questionable Uses of Big Data

	Beneficial Uses	Questionable Uses
By Technology		
License Plate Readers	Reading passing cars for tolls on highway; police locating stolen car	Used by private detectives; placed on trucks to gather license plate data broadly
Facial Recognition	Finding potential terrorists at large sporting events	Used by social networking sites to identify members in pictures
GPS	Location-based coupons; traffic predictions; directions on map	Location-based stalking; iPhone as a homing beacon
By Context		
Healthcare	Treatment of cancer; health of pregnancy; Google Flu Trends Insights into interaction between medications from search terms; insights into hospital spread of infections Identifying veterans' potential suicidal thoughts	Discrimination in healthcare and insurance; app knows how fit you are Development of a health score from purchase habits and from search terms
Education	Personalizing student instruction Accountability for performance by school Identifying students at risk of dropping out	Using data for possible admissions discrimination
Electricity	Turning on/off home electricity	Allowing criminals to know if you are home; smart homes hacked
Law Enforcement	Machine learning to identify burglar; accessing phone records to identify potential suspects in a mugging New York Fire Department using data mining to predict problems	Accessing smartphone without a warrant; identifying suspects by web browsing habits Individuals under scrutiny for not participating in tracking
Retail	Improving layout of store based on typical movements of customers Better coupons, suggested items WalMart's use of RetailLink to integrate suppliers with onsite supplier inventory	Tracking movements/shopping habits of spectators at a stadium using Verizon's Precision Marketing Insight program Price discrimination (e.g., Amazon, Orbitz) Target sending notice of pregnancy to unsuspecting teen's parents
Urban Planning	Traffic management; smart grid technology Use of popular app by competitive cyclists and runners for road planning Identifying areas for road improvement	Identifying who is listening to which radio station; EZ Pass responder tracked everywhere Possibility of hackers changing traffic lights and creating traffic jams Identifying areas for road improvement but focusing only on those with mobile apps

potential confusion on how Big Data fits in a community—if at all.

Part of the ambiguity in researching Big Data is choosing what to study. Big Data has been framed as: (1) the *ability* to process huge "treasure troves" of data and predict future outcomes, (2) a *process* that "leverages massive data sets and algorithmic analysis" to extract new information and meaning, (3) an *asset,* (4) a *moment* where the data volume, acquisition or velocity limits the use of traditional tools and (5) a *tactic* to operate at a large scale not possible at a smaller scale.[5]

Framing Big Data as an asset, ability or technique sterilizes an important ethical discussion. Big Data is mistakenly framed as morally neutral or having benefits that outweigh any costs. Grand statements such as "Big Data itself, like all technology, is ethically neutral"[6] are implicit in reports that focus on the strategic and operational challenges of Big Data, but which largely ignore the ethical and social implications.[7] The growing field of data analytics excludes ethical analysis in both practice and academia. Yet creating, aggregating and selling data can change relationships and business models and requires rethinking information governance strategies—including issues concerning ethics and privacy.[8]

I suggest Big Data should be analyzed as the Big Data Industry (BDI) in order to identify the systemic risks in current Big Data practices. Such an approach situates Big Data within a larger system of firms, organizations, processes and norms for analysis. The volume, variety and velocity[9] of the data, plus the novel analytics required to produce actionable information, renders Big Data a difference in kind rather than degree. To create and use these large data sets to maximum effect, many firms aggregate data to create a new "whole" and sell access to this new data set.

The separate and distinct firms in the Big Data Industry work through agreements to produce a product (Big Data) for customers—similar to any other industry.[10] In response, CIOs and CDOs (Chief Data Officers) are shifting to an outward, strategic focus in leveraging Big Data rather than the inward, service focus used for traditional data. At present, however, there are not yet any industry norms or supply chain best practices that can guide them.[11]

This article examines the ethical issues in the nascent Big Data Industry. Industries are the aggregate of firms involved in the production and distribution of a product—e.g., the software industry, the ERP industry, the automobile industry, etc. Importantly, if a market exists for a product, then a corresponding industry exists to meet that demand. And, as the market for Big Data continues to grow and be measured, the corresponding Big Data Industry, comprised of those firms involved in the production, analysis and use of Big Data, begins to coalesce around standard industry practices. (Note that this article focuses on privacy issues in the U.S. Big Data Industry; as described in the panel below, the privacy regulatory environments in the U.S. and Europe differ significantly.)

5 In order of reference: Barocas, S. and Selbst, A. D., op. cit., 2014; Hartzog, W. and Selinger, E., op. cit., 2013; *Big Data Management & Analytics,* Gartner, 2014, available at http://www.gartner.com/technology/topics/big-data.jsp; Mayer-Schönberger, V. and Cukier, K. *Big Data: A Revolution That Will Transform How We Live, Work, and Think,* Houghton Mifflin Harcourt, 2013; Richards, N. M. and King, J. H., op. cit., 2013.

6 Wen, H. "Big ethics for big data," *O'Reilly Radar,* June 11, 2012, available at http://radar.oreilly.com/2012/06/ethics-big-data-business-decisions.html.

7 For example, Gartner notes that there are three strategic and operational challenges: information strategy, data analytics and enterprise information management, but makes no mention of ethical challenges. See also *Big Data and Privacy: A Technological Perspective,* Report to the President, 2014, available at https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf; *Big Data Platform - Bringing Big Data to the Enterprise,* IBM ; and Manyika, J. et al. *Big Data: The next Frontier for Innovation, Competition, and Productivity,* McKinsey & Company, 2011, available at http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

8 The shift to creating value through monetizing data impacts relationships with stakeholders as well as policies internal to the organization—see Tallon, P. P., Short, J. E and Harkins, M. W. "The Evolution of Information Governance at Intel," *MIS Quarterly Executive* (12:4), 2013, pp. 189-198; Najjar, M. S. and Kettinger, W. J., "Data Monetization: Lessons from a Retailer's Journey," *MIS Quarterly Executive* (12:4), 2013, pp. 213-225.

9 The 3Vs of Big Data—volume, variety and velocity—were originally defined in a META/Gartner report but have subsequently been expanded with veracity, value, validity, variability and even visualization, leading to the term "V confusion"—see Grimes, S. "Big Data: Avoid 'Wanna V' Confusion," *InformationWeek,* August 7, 2013, available at http://www.informationweek.com/big-data/big-data-analytics/big-data-avoid-wanna-v-confusion/d/d-id/1111077?.

10 Firms monetizing the value of data require new tactics and strategies as well as, perhaps, accounting rules to capture the value (and risk) created in new transactions. See Monga, V. "The Big Mystery: What's Big Data Really Worth?," *Wall Street Journal,* October 13, 2014, available at http://blogs.wsj.com/cfo/2014/10/13/the-big-mystery-whats-big-data-really-worth/.

11 Lee, Y. et al., "A Cubic Framework for the Chief Data Officer: Succeeding in a World of Big Data," *MIS Quarterly Executive* (13:1), 2014, pp. 1-13.

Privacy: U.S. Versus EU

The use of Big Data in Europe faces a distinct set of regulatory constraints governed by the EU's Data Protection Directive (95/46/EC) and, for example, the United Kingdom's Data Protection Act 1998. Regulations require those using "personal data" to abide by the directive's requirements to being fair, to be clear as to the purpose of gathered information and, problematic for Big Data, to strive for minimization. See also the Bureau of National Affairs' World Data protection Report 14(9) as well as the U.K.'s Information Commissioner's Office Big Data and Data Protection (2014).

For example, Facebook recently was unable to comply with the stricter EU regulations because of a lack of adequate consent and control for users: Facebook users have no true opt-out mechanism, no valid consent for the transfer of data to third parties and a general lack of control over their data. In other words, Facebook's "take it or leave it" approach to choice is not sufficient for European law.[12] Generally, privacy is taken more seriously by regulators in the EU (and by U.S. companies doing business in Europe), with "data subjects" having a right to be forgotten, authentic user consent and a general leaning toward "opt-in" as the default.[13]

The article first examines the information supply chain within the Big Data Industry, including upstream sources of data and downstream uses of data. Next, it examines two crucial consumer-related ethical issues created by systemic norms and practices of the Big Data Industry: (1) the negative externality of surveillance and (2) destructive demand. Remedies for these potential issues are proposed, with the goal of fostering a sustainable Big Data Industry.

An industry-level analysis extends the examination of Big Data in three ways. First, framing Big Data as an industry highlights the participants, power relationships and systemic issues that arise within the production and use of Big Data, insights that are not available when Big Data is isolated as a technology. Second, an industry-level analysis captures pervasive industry practices that are missed when considering single uses of Big Data. These systemic issues can be resolved with the industry-specific measures described in the analysis. Finally, an industry-level analysis broadens the number of interested parties to all who have a stake in creating a sustainable Big Data Industry. All companies in controversial industries have their legitimacy questioned and have a vested interest in creating sustainable industry norms. In other words, the recognition that bad behavior may delegitimize the entire industry provides an incentive for industry leaders to curb such practices.[14] A brief overview of the leading firms in the Big Data Industry is given in the left panel on the next page.

The Big Data Industry's Supply Chain

Within the Big Data Industry, data, such as online consumer data or location data from an application, is passed from one firm to the next within an information supply chain, comparable to supply chains in traditional industries (see text panel on the next page). Within this supply chain, consumers provide information to firms, which then pass it to tracking companies, which may also pass it to data aggregators. Data aggregators

12 See also Lomas, N. "Facebook's Data Protection Practices Under Fresh Fire In Europe," *TechCrunch,* available at http://social.techcrunch.com/2015/02/23/facebook-ad-network/.

13 Scott, M. "Where Tech Giants Protect Privacy," *The New York Times,* December 13, 2014, available at http://www.nytimes.com/2014/12/14/sunday-review/where-tech-giants-protect-privacy.html.

14 The BDI requires not only information brokers to aggregate data, but also hardware, software and professional services firms to support the collection, storage and use of the data. Leaders include firms focused on analytics solutions (e.g., SAS, IBM, SAP) as well as industry specialists (e.g., Amazon Web Services) and service providers (Accenture). For more information, see Robb, D. "Top 20 Big Data Companies," Datamation, November 20, 2014, available at http://www.datamation.com/applications/top-20-big-data-companies-1.html. Importantly, many firms combine products and services that support the BDI—e.g., IBM (hardware, software and services), HP (cloud and storage), Dell (storage), SAP (analytics), Teradata and Oracle (hardware, software, services), SAS and Palantir (analytics and software) and Accenture (software and services). See Leopold, G. "Big Data Rankings: Leaders Generated $6B in Revenues," Datanami, December 4, 2014, available at http://www.datanami.com/2014/12/04/big-data-rankings-leaders-generated-6b-revenues/. While hardware, software, analytics and even technology consulting firms make most of the industry leader lists, missing are the data brokers and data aggregators that make up the information supply chain discussed in the next section.

Supply Chains

In a traditional business model, supply chains comprise a series of firms working together to deliver value by transforming raw material into a finished product. Trees are harvested in the forest, traded to the pulp manufacturer and eventually become the paper used to print an article; tomatoes are picked, packed, shipped and crushed into sauce to be used on a delivered pizza. The figure below illustrates a generic supply chain: each firm adds value to the product or service to transform the raw materials in one location and deliver a finished product to the end customer through value creation and trade.

All supply chains carry ethical issues both downstream and upstream. Software companies must ensure that their products are not eventually sold in Syria through a distribution center in Dubai; Apple is held accountable for the working conditions of its upstream suppliers, such as Foxconn. Supply chain researchers examine upstream sourcing issues, looking at how supplier selection takes account of, for example, the way forests are harvested in the paper industry or how apparel is manufactured overseas, as well as following products downstream through logistics and eventual sale and use.

act as distributors by holding consolidated information of many users across many contexts.

Data aggregators or data brokers may sell the information to researchers, government agencies or polling companies, or an ad network may use the information from an aggregator or broker to place an advertisement on a website when a user returns to browse or shop online. Survey firms, academic research teams, government agencies or private firms may also contract with a data broker directly to use data to supplement survey research, make employment decisions and investigate possible criminal activity. An information supply chain is thus created with multiple firms exchanging information and adding value to the data.

As with traditional supply chains, the information supply chain can be analyzed both by the downstream distribution and use of Big Data as well as by the upstream sourcing (see Figure 1).

The issues arising from the downstream use of Big Data and upstream sourcing of information are summarized in Figure 2 and described in detail below.

Issues with Downstream Customers and Uses of Big Data

As shown in Table 1, downstream uses of Big Data can be perceived as producing beneficial and questionable (often unethical and harmful) outcomes. However, the potential harm that can result from using Big Data should not detract from the benefits—from curing diseases to identifying fraud. Nonetheless, selling information increases the risk of secondary misuse of the data, with eventual harmful impacts on users. While the potential harm from *incorrect* information or false conclusions merits attention, harm downstream in the supply chain includes harm from the *correct* conclusions. For instance, Target famously correctly identified a pregnant teenager based on her purchase history and sent a congratulatory letter to her house, which was seen by her parents who were unaware that their daughter was pregnant.[15]

The harmful effects of using Big Data can be extended to include:

- *Value destruction* (rather than creation) for stakeholders
- *Diminished rights* (rather than realized) for stakeholders
- *Disrespectful* to someone involved in the process (rather than supporting them).

Such effects are not possible without information provided upstream, thereby linking

15 Duhigg, C. "How Companies Learn Your Secrets," *The New York Times,* February 16, 2012, available at http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html.

Figure 1: Example of Information Supply Chain Within the Big Data Industry

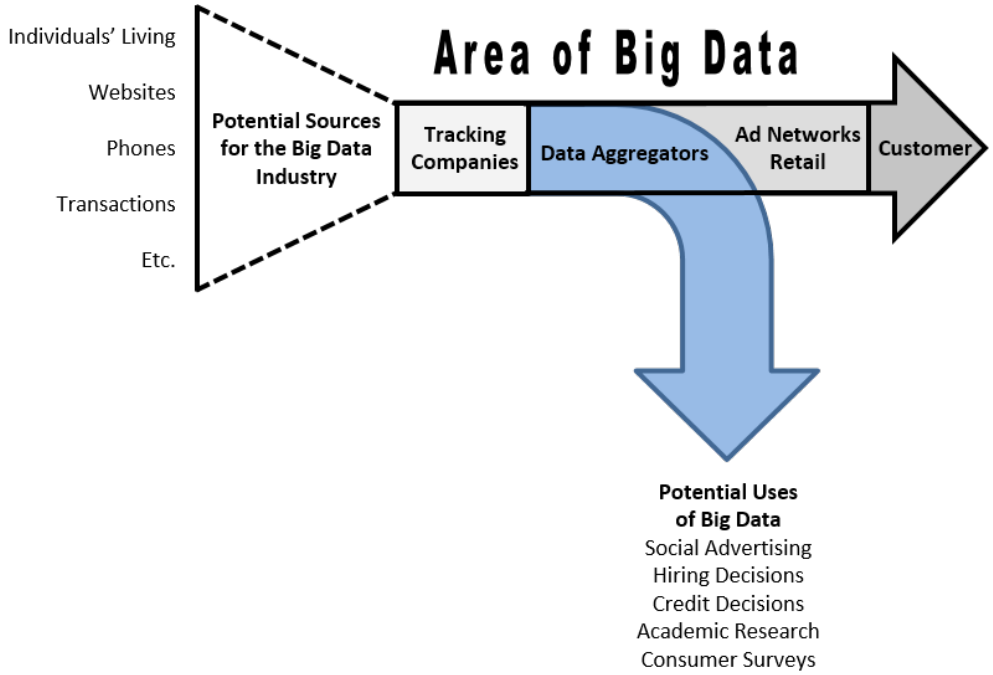
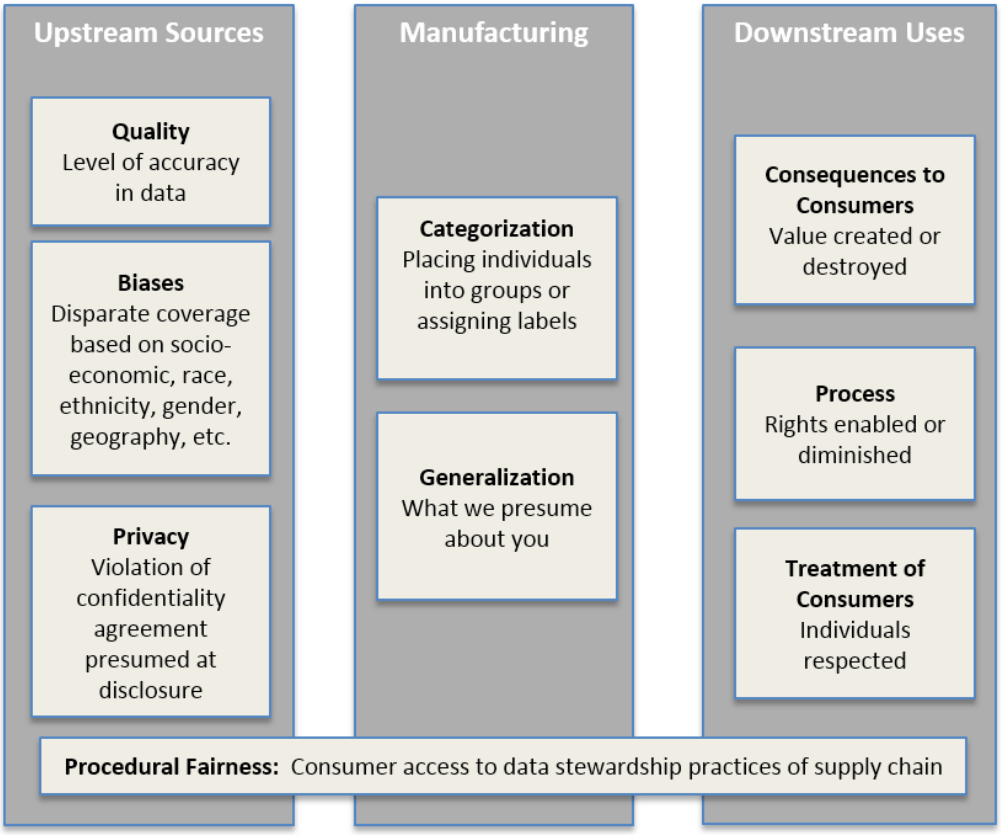


Figure 2: Issues within the BDI Supply Chain



all supply chain members to the eventual uses of information.[16]

First, data uses can be analyzed based on the *consequences* to the individual. More obvious adverse consequences include being denied credit, losing a job, having secrets outed to your family, paying more for insurance, etc. For example, information may be used downstream to modify insurance premiums or mortgage rates. However, there can also be positive consequences, as when downstream use identifies trends in demographics such as flu outbreaks, or prioritizes search results for a travel site.[17] Table 1 focuses on the consequences (both good and bad) from the use of Big Data.

A more egregious yet subtle consequence is what law scholar Ryan Calo conceptualizes as digital market manipulation. When firms know more information about consumers with an ever better ability to fine-tune the consumer experience, they are able to influence consumers at a personal level and to trigger vulnerability in consumers in their marketing.[18] Calo's argument suggests that Target, for example, would not only identify a consumer who is pregnant, but could also engineer food cravings in her through subtle triggers. As summarized by Calo, firms will increasingly be in the position to create 'suckers' rather than waiting for one to be born every minute.

The harm resulting from the use of Big Data can also be identified by asking not only how value is created or destroyed for individuals, but also whether individuals' *rights are being realized* in the process of using the data. Barocos and Selbst nicely illustrate the harm that can arise not only from the information supply chain, but also from the process followed in using Big Data. Big Data may develop learned prejudice algorithms based on pre-existing information. By basing predictive algorithms on previous data patterns, learned prejudice builds on previously institutionalized prejudice—for example, in areas such as college admissions or when a Google search on black-sounding names brings up arrest records. Such algorithms can produce objectionable outcomes, as with accidental or intentional discrimination.[19]

Finally, categorizing individuals under certain headings can be disrespectful to them—for example, the categorization of individuals based on their personal history, such as rape victim status, becomes an exercise in objectifying individuals as a mere category. Big Data aggregators have been known to list individuals by classifications such as alcoholics, erectile dysfunction sufferers and even as "daughter killed in car crash."[20] Even without value being destroyed, individuals can be disrespected through objectifying them as a mere category—particularly a category that overwhelms in significance, such as being the victim of a crime, struggling with an addiction or coping with a death.

Issues with Upstream Sources

In addition to the possible downstream harmful effects of using Big Data, firms in the information supply chain must also contend with issues concerned with upstream suppliers of data, in particular the possibility of partnering with bad suppliers. The ability to develop an ever-greater volume, velocity and variety of data

16 As stated by Bambauer, "estimating harm is a wearisome task"—see Bambauer, J. "Other People's Papers," draft paper, 2014, p. 15, available at http://masonlec.org/site/rte_uploads/files/Bambauer_Other_Peoples_Papers_GMU.pdf. Bambauer categorizes privacy harms as arising from collection, risk of misuse, aggregation, obstruction and hassle; Richards lists sorting, discrimination, persuasion and blackmail as potential harms—Richards, N. M. "The Dangers of Surveillance," *Harvard Law Review,* 2013, available at http://harvardlawreview.org/2013/05/the-dangers-of-surveillance/; Calo focuses more broadly on objective and subjective harms—Calo, M. R. "Boundaries of Privacy Harm," *Indiana Law Journal* (86), 2011, pp. 1131-1162.

17 For example, car insurance companies are moving toward usage-based premiums based on driving data collected in real time—see Boulton, C. "Auto Insurers Bank on Big Data to Drive New Business," *Wall Street Journal,* February 20, 2013, available at http://blogs.wsj.com/cio/2013/02/20/auto-insurers-bank-on-big-data-to-drive-new-business/. Similarly, health insurance companies can deny services and increase premiums through accessing data online—see Gittelson, K. "How Big Data Is Changing Insurance," *BBC News,* November 15, 2013, available at http://www.bbc.com/news/business-24941415.

18 Calo, M. R. "Digital Market Manipulation," *The George Washington Law Review* (82:4), 2013, pp. 995-1051.

19 For the concept of objectionable classification and biases, see Barocas, S. and Selbst, A. D., op. cit., 2015; Sweeney, L. "Discrimination in Online Ad Delivery," *acmqueue* (11:3), 2013, available at http://queue.acm.org/detail.cfm?id=2460278; and Cohen, J. E. "What Privacy Is for," *Harvard Law Review* (126), 2013, pp. 1904-1933.

20 For examples of objectionable categorizations, see Hill, K. "Data Broker Was Selling Lists Of Rape Victims, Alcoholics, and 'Erectile Dysfunction Sufferers'," *Forbes,* September 19, 2013, available at http://www.forbes.com/sites/kashmirhill/2013/12/19/data-broker-was-selling-lists-of-rape-alcoholism-and-erectile-dysfunction-sufferers/.

requires large, complex and distributed data sets from many sources. Sources of data within the Big Data Industry include consumers, products, location, machines and transactions (and all combinations of these). In fact, the variety of combined data differentiates Big Data from traditional data analysis: many data sources combine data types or use data in novel ways. This pooling of diverse, sometimes innocuous, pieces of data contributes to a greater potential for statistical significance or to making sense of new knowledge.[21]

Within the Big Data Industry, upstream sources may be undesirable because of the quality of information, biases in the data and privacy issues in the collection and sharing of information. Data quality may be an issue due to inaccuracies in the data or a lack of coverage.[22] Inaccuracies may arise from the manner in which the data was collected, the degree of imputed[23] data within the data source or from deliberate obfuscation by users.[24] Assessing the quality of upstream data is similar to assessing the quality of upstream sources in a manufacturing supply chain, where firms are free to specify the quality they desire for their products. However, firms using upstream information further down the information supply chain will be held accountable for the quality of that information.

Data may also have biases that skew it toward specific types of users, such as a particular race, ethnicity, gender, socioeconomic status or location. Using upstream data further down the supply chain requires an understanding of the level of bias in the data—skewed data will skew the results and limit the generalization of the findings. For example, location tracking can be beneficial to the community when used for transit scheduling; however, if one group is systematically ignored in the source data (e.g., groups with less access to mobile devices used to track location data), that group will not benefit from the improved transit system or may have traffic flow inaccurately predicted.[25]

Finally, and importantly for the ethical implications of the Big Data Industry, the firm supplying data should be assessed on how it respects privacy in the collection of information. Consumers disclose information within a set of privacy rules, and sharing that information with other firms in the supply chain may breach their privacy expectations. In other words, information always has "terms of use" or norms governing when, how, why and where it can be used.[26] For example, information shared with Orbitz, a travel website, has a distinct set of privacy expectations based on the individual's relationship with the website and the context of the interaction. Individuals may expect location information to be used to offer hotel or restaurant discounts for their destination, but they do not expect that information be passed to data aggregators and used a year later to make pricing decisions. Users disclose information with a purpose in mind and within an implicit confidentiality agreement.

Privacy law scholar Woodrow Hartzog suggests that this confidentiality agreement should be honored by firms that subsequently receive or gather the information within a concept of "chain link confidentiality."[27] The expectations present at initial disclosure—who should receive information, how it can be used, how long it will be stored—should persist throughout the online information supply chain.

21 Groves has previously categorized data sources as organic vs. designed—Groves, R. M. "Three Eras of Survey Research," *Public Opinion Quarterly* (75:5), 2011, pp. 861-871. Sources have also been categorized as analog vs. digital in *Big Data and Privacy: A Technological Perspective,* Report to the President, 2014. However, the differences in these categories are not always clear or meaningful in determining the appropriateness of the supplier.

22 For an analysis of quality and bias issues in Big Data sources, see Boyd, D. and Crawford, K. *Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon,* Microsoft Research, 2012; Lerman, J. "Big Data and Its Exclusions," *Stanford Law Review Online* (66), 2013, pp. 55-63; and Crawford, K. "The Hidden Biases in Big Data," *Harvard Business Review,* April 1, 2013.

23 Imputation is the process of replacing missing data with substituted values.

24 The role of obfuscation in protecting privacy is examined in Brunton, F. and Nissenbaum, H. "Vernacular resistance to data collection and analysis: A political theory of obfuscation," *First Monday* (16:5), 2011.

25 O'Leary, D. E. "Exploiting Big Data from Mobile Device Sensor-Based Apps: Challenges and Benefits," *MIS Quarterly Executive* (12:4), 2013, pp. 179-187.

26 Nissenbaum, H. *Privacy in Context: Technology, Policy, and the Integrity of Social Life,* Stanford University Press, 2009; Martin, K. "Understanding Privacy Online: Development of a Social Contract Approach to Privacy," *Journal of Business Ethics,* 2015, pp. 1-19; and Richards, N. M. and King, J. H. "Big Data Ethics," *Wake Forest Law Review* (23), 2014.

27 Hartzog, W. "Chain-Link Confidentiality," *Georgia Law Review* (46), 2011, pp. 657-704.

Role of Firms in the Information Supply Chain

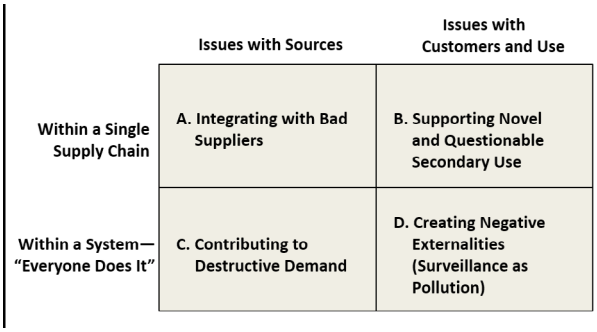
In conventional supply chains, upstream suppliers may have quality problems or unethical business practices that taint the final product. In the 1990s, for example, Wal-Mart and Nike infamously relied on overseas manufacturers that used child labor and unsafe working conditions. More recently, Apple has grappled with the reputational problems arising from using Foxconn, a supplier with harsh working conditions. Firms that willingly enter a supply chain have an obligation to ensure that the practices of other firms in the chain match their own. Similarly, organizations within the information supply chain are held responsible for the data stewardship practices of both upstream and downstream partners.

An organization's responsibility within a supply chain is derived from the benefits it receives from the practices of the supply chain. In accepting those benefits, the firm implicitly signs up to the practices of the supply chain—including potentially questionable practices of upstream sources. Nike benefits from the practices of its suppliers even though the working conditions of those suppliers leave a lot to be desired.

Each firm in the Big Data Industry contributes to, and benefits from, an information supply chain and willingly takes on part of the responsibility for actions and practices within that chain. For example, when Facebook seeks to use information from upstream data brokers such as Acxiom, Epsilon, Datalogix and BlueKai,[28] it must not only worry about its own collection methods, but also the upstream sources' data collection methods. Choosing and creating supply chains means that firms are responsible for the conduct and treatment of users throughout the chain. Thus Nike is held responsible for how its products are sourced, and coffee retailers are held responsible for how their coffee is farmed.

28 Hill, K. "Facebook Joins Forces With Data Brokers To Gather More Intel About Users For Ads," *Forbes,* February 27, 2013, available at http://www.forbes.com/sites/kashmirhill/2013/02/27/facebook-joins-forces-with-data-brokers-to-gather-more-intel-about-users-for-ads/.

Figure 3: Current Ethical Issues within the Big Data Industry



Systemic Issues in the Big Data Industry

As described above, the role of firms within their information supply chain should be analyzed, but the Big Data Industry includes firms that are developing generalized norms and practices. In effect, the *systemic* participation in the Big Data Industry gives rise to "everyone does it" ethical issues—where norms of practice are beginning to form across many firms and supply chains, as illustrated in Figure 3. Quadrants A and B capture the ethical issues within a single supply chain, as described above.

This section examines the ethical issues captured by Quadrants C and D, and links them to parallel, more traditional industries. The first issue is creating negative externalities (or surveillance as pollution), where surveillance is a byproduct of the systematic collection, aggregation and use of individual data (Quadrant D). The second is the growing problem of destructive demand within the Big Data Industry (Quadrant C), where the need for consumer data is pressuring consumer-facing firms to collect and sell increasing amounts of information with lower standards. Both sets of ethical issues stem from the *systemic* norms and practices within the industry. In addition, both are more consumer- or individual-focused and may apply to a particular subset of firms within the Big Data Industry.

The ethical issues that have to be faced at both the supply-chain level and the industry level are summarized in Table 2 (For comparison, the table provides corresponding examples from traditional industries; it also describes how CIOs and CDOs will have to deal with the issues.)

Table 2: Ethical Issues in the Big Data Industry

Ethical Issues	Big Data Industry Examples	Traditional Industry Examples	As Faced by CIOs and CDOs
Supply Chain Level			
Unfair or objectionable harms from using Big Data	Harms from downstream use, such as using Big Data to discriminate in consumer credit decisions or college admissions	Sale of computer systems in Iran or Syria; use of product in crime	How do downstream users of your data protect the consumer data or impact consumers?
Gathering of data as an intrusion or violation of privacy	Questionable upstream sourcing, such as purchasing location data surreptitiously gathered from mobile applications or using data from invisible web beacons unknown to user	Apple and Foxconn; Nike and sweatshops	What questions do you ask about using data from unknown or questionable sources?
Industry Level			
Harm to those not involved in the immediate decision or transaction caused by broad tracking of consumers and collection of information	Negative externality of surveillance, such as the hidden and systematic aggregation of data about individuals	Steel industry and pollution	How is your company possibly contributing to surveillance by participating in broad user tracking—or partnering within someone who does?
Focus on resale of consumer data; treating consumers simply as a means to supply the secondary market of information traders	Destructive demand, such as creating a flashlight application just to gather user contact or location data	Demand for residential mortgages created by the mortgage-backed securities industry; websites and applications used as bait	How is your company creating destructive demand by using data of questionable quality or that was collected by breaching privacy expectations?

Creating Negative Externalities (or Surveillance as Pollution)

In all markets, costs regularly accrue to parties not directly involved in an immediate decision or exchange. For example, a firm making steel can create harm to the community in the form of the pollution it produces. The steel company may contract with a customer—which does not feel the effects of pollution—without including the "cost" of pollution. This is an example of a *negative externality,* which exists when the harm done to others is not taken into account in the immediate transaction.[29]

There are also negative externalities in the Big Data Industry arising from the aggressive focus on collecting consumer data. The danger is that disclosing personal data can become the default,

29 Coase illustrated negative externalities with the example of a spark from a train that causes harm to farmers along the tracks—Coase, R. H. "Problem of Social Cost," *Journal of Law and Economics* (3), 1960, pp. 1-44. Importantly for Coase, negative externalities do not necessarily require government intervention, which carries its own cost, but may be resolved through private ordering between parties.

and individuals who choose *not* to disclose can be harmed. For example, individuals who attempt to opt out of aggressive data collection by using TOR[30] or other obfuscation technologies may be targeted by the National Security Agency as suspicious.[31] The harm to individuals who do not share their data is a result of the decisions of the majority who do share.

More complicated is when the harmful effect is compounded by many parties in an industry acting in a similar way. For example, a manufacturing firm may not take account of the harmful effects on the local community of the pollution it produces. However, the *aggregated* harm of pollution from manufacturers worldwide becomes a problem for society in general through global warming. Aggregated negative externalities are a consequence of "everyone does it"— the harm results from the fact that the practice is pervasive in an industry. The harm from aggregated actions across an industry is more than the sum of the harms caused by individual firms.

Firms within the Big Data Industry create an aggregated negative externality because they contribute to a larger system of surveillance through the breadth of information gathered and because firms that collect and aggregate data are invisible to users. In general, surveillance conflicts with the need of individuals to be unobserved as well as their need for uniqueness and a sense of self. An individual's personal space permits "unconstrained, unobserved physical and intellectual movement" to develop as an individual and to cultivate relationships.[32] Surveillance can cause harm by violating the personal space—both physical and metaphorical—that is important to develop as an individual and within relationships. Importantly, the fear of being watched and judged by others causes "spaces exposed by surveillance [to] function differently than spaces that are not so exposed" by changing how individuals behave and think.[33]

Surveillance works by affecting not only those who are being watched, but also those who are *not actually* being watched. In fact, the mere belief that someone is being watched is enough for individuals to act as though they are under surveillance. Prisons are designed so that only some of the prisoners are watched, but the prisoners do not know specifically who is being watched at any one time. Individuals do not need to know they are under surveillance to act as though they are under surveillance. Importantly for the Big Data Industry, the negative externality of surveillance means the industry can rely on those individuals not currently being watched to believe and act as though they are under surveillance.

Surveillance is particularly effective in changing behavior and thoughts when individuals (1) cannot avoid the gaze of the watcher and (2) cannot identify the watchers.[34] By aggregating data across disparate contexts online, the Big Data Industry contributes to the perception that surveillance is impossible to avoid yet also creates a data record that tells a richer, more personalized story than individual data points.[35] Broad data aggregators summarize highly diverse data (the "variety" in Big Data) so they can analyze individualized behavior. In addition, most data aggregators are invisible to the user and thereby aggravate the surveillance problem by being not only unknown, but also unreachable. Unknown and invisible firms that

30 TOR—The Onion Router—is a service to make accessing websites anonymous. Users' requests are routed among many other TOR users' requests and are bounced throughout the TOR network of client computers to remain hidden to outsiders. For more information, see https://www.torproject.org.

31 Zetter, K. "The NSA Is Targeting Users of Privacy Services, Leaked Code Shows," *WIRED,* July 3, 2014.

32 Fried, F. *An Anatomy of Values: Problems of Personal and Social Choice,* Harvard University Press, 1970; and Rachels, J. "Why Privacy Is Important," *Philosophy & Public Affairs,* 1975, pp. 323-333.

33 Cohen, J. E. "Privacy, Visibility, Transparency, and Exposure," *The University of Chicago Law Review* (75:1), 2008, pp. 181-201. The inability to escape online surveillance is illustrated in Brunton, F. and Nissenbaum, H., op. cit., 2011, and Strandburg, K. J. "Home, Home on the Web and Other Fourth Amendment Implications of Technosocial Change," *Maryland Law Review,* (70:3), 2011. In the words of Cohen, "Pervasive monitoring of every first move or false start will, at the margin, incline choices toward the bland and mainstream" thereby causing "a blunting and blurring of rough edges and sharp lines." —Cohen, J. E. "Examined lives: Informational privacy and the subject as object," *Stanford Law Review,* (52), 2000, pp. 1373-1438.

34 Cohen, J. E., op. cit., 2008.

35 The Mosaic Theory of privacy explains why privacy scholars are concerned with all elements of tracking, including transaction surveillance and purchasing behavior. This theory suggests that the whole of one's movements reveal far more than the individual movements—where the aggregation of small movements across contexts is a difference in kind and not in degree. See Kerr, O.S. "The Mosaic Theory of the Fourth Amendment," *Michigan Law Review* (111:3), 2012; and *United States v. Jones,* Supreme Court of United States, January 23, 2012, available at http://www.supremecourt.gov/opinions/11pdf/10-1259.pdf.

gather and store data contribute to the perception of omnipresent and omniscient surveillance and exacerbate the power imbalance between the watched and the watcher.[36]

Currently, the Big Data Industry does not consider or take account of the negative externality of surveillance. Firms that capture, aggregate or use Big Data create a cost to the larger community in the form of surveillance.

Contributing to Destructive Demand

In addition to the aggregate harm of surveillance, the Big Data Industry has the potential to foster *destructive demand* for consumer data when firms exert pressure on consumer-facing organizations to collect more information. As described below, consumers unknowingly can become suppliers to a secondary Big Data market.

The main source of information for the Big Data Industry is a byproduct of legitimate transactions with consumer-facing firms. Data is collected from a transaction in the primary market—e.g., checking the weather, buying groceries, using a phone, paying bills, etc.—and is then aggregated and merged to create a large robust data set. In effect, that data is seen as sitting inventory when a firm in the secondary Big Data market—such as a data broker or tracking company—creates value through the secondary use of the data. The consumer data from the initial transaction, such as buying books on Amazon or reading news on *The New York Times,* can be sold or repurposed in a secondary market without losing value. Examples of destructive demand created by secondary markets are described in the panel on the following page.

A tipping point exists where the product—whether residential mortgages as described in the panel or consumer information—is no longer *pushed* into the secondary market, but rather the secondary market becomes a *pull* for the product of the primary, consumer-targeted market. In this situation, the secondary market creates a destructive demand by exerting pressure on suppliers to adopt questionable or unethical practices to meet the demands of the secondary market. Primary market firms (e.g., residential mortgage originators) then treat customers as a mere means[37] to the secondary market (for mortgage-backed securities). The demand becomes particularly destructive when the service in the primary market serves as a lure (or bait) for the supply of the secondary market—as when mortgage originators became a lure to produce mortgages for the mortgage-backed securities market.

Within the Big Data Industry, websites and applications with trusted relationships with consumers can become the bait for Big Data, such as when a flashlight application tracks your location or when a website with numerous tracking beacons[38] stores consumer information. The primary market promises a customer-focused relationship (first-market relationship) when it is actually attempting to sell customers' information to a secondary market.

The attributes of the mortgage-backed securities market, and the destructive demand it created, provide a warning for the secondary market for consumer information in the Big Data Industry. The demand for the primary market becomes destructive:

1. *Where the secondary market becomes as or more lucrative than the primary market.* For example, the fee charged to consumers for mortgages was dwarfed by the profits from the sale of mortgages into the secondary market. Mortgage originators could lose money on a mortgage but still make a profit by selling the mortgage in the secondary market. Within the Big Data Industry, problems will arise when the sale of consumer information is more lucrative or, at minimum, equals the profits from the primary market activities, such as selling an application or providing a service.
2. *When the quality in the secondary market is less than in the primary market*—i.e., when the quality requirements of data brokers or data aggregators do not match the expectations of consumers who disclose information. For example, the mortgage-backed securities market was not concerned about the quality of the

36 Richards, N. M., op. cit., 2013.

37 The Mere Means Principle is an ethical principle that posits that you should never treat people merely as a means to your own ends.

38 A tracking is an often-transparent graphic image, usually no larger than 1 pixel x 1 pixel, that is placed on a website that is used to monitor the behavior of the user visiting the site.

Examples of Destructive Demand from Secondary Markets

Secondary markets can be beneficial. A secondary market for bicycles and cars can increase the life of the product. In fact, customers may be more willing to invest in a car in the primary "new car" market knowing that the robust secondary market for used cars exists to sell the car when necessary. Other secondary markets create value from items that would otherwise be thrown away—e.g., the byproduct from cattle ranching (wax) or from steel-making (scrap metal). The secondary market allows firms to capture value from seemingly waste products, such as ranchers selling the byproduct of cow fat used for candles.

However, secondary markets can apply perverse pressures to distort the demand, quality or price in the primary market. An example is the market for carbon credits. Firms who create HFC-23, a super greenhouse gas, as a byproduct of their manufacturing are paid to destroy it to prevent the gas causing environmental damage. However, the secondary market for HFC-23 became too lucrative: some firms had an incentive to create HFC-23 so they would be paid to destroy it. In fact, the World Bank paid $1 billion to two chemical factories in China to destroy HFC-23, and later evidence suggested these firms may have deliberately overproduced the gas so they could be paid to destroy it in the secondary market.

More problematic is when the secondary market begins to systematically distort the primary market, as in the well-known case of mortgage-backed securities and the residential mortgage market. The primary market for mortgages is between a lender and home-buyer. Financial institutions lend money to qualified individuals to buy a home at a rate that takes into account the potential risk of the individual defaulting on the loan.

A secondary market for residential mortgages uses consumer mortgages as the inventory for a new financial instrument: mortgage-backed securities (MBS). The MBS market increased dramatically between 2000 and 2008, and the associated demand for consumer mortgages to feed the MBS market led to lax sourcing in the primary mortgage market. Interestingly, the price did not change in the primary market; rates and interest rate spreads remained steady throughout the growth in the MBS market. However, the quality standards for consumer mortgages required in the primary market dropped to match the (lower) requirements in the secondary market. More mortgage originations and fewer denials led to a greater number of high-risk borrowers through lax sourcing for the MBS market.

This mismatch between the quality required in the secondary and primary markets proved particularly hazardous. The interests of firms in the secondary market did not align with those of consumers, and without a relationship with consumers there were higher default rates for the mortgages included in their MBS. However, when incentives of the secondary market were aligned with the primary market of the consumer, as in the case of affiliated investors, economists found no change in the mortgage default rates. The increase in private securitization by non-commercial bank financial firms, with lower requirements for quality, created a destructive demand for lower quality mortgages in the primary market.

residential mortgages they purchased from originators.

3. *When firms in the primary market have limited accountability to consumers for their transactions in the secondary market.* Primary market firms can hide their bad behavior when they sell into the secondary market because their activity in the secondary market is not visible or incorporated in the primary market. The term "moral hazard" refers to when individuals or institutions do not bear the full consequences of their actions, as in

the case of mortgage originators selling bad loans into the MBS secondary market. In the Big Data Industry, consumer-facing organizations are currently not held accountable for selling access to consumer data even by market forces, and their activities in the secondary market are invisible to the primary consumer market.

Guidelines for a Sustainable Big Data Industry

The Big Data Industry is currently in a unique, yet vulnerable, position, with identified systemic risks but without clear industry leaders to develop cooperative strategies. Moreover, the power of Big Data is generated by non-consumer-focused firms that aggregate and distribute the data, and regulating such firms has met with questionable success in the other industries.[39] However, all firms are tainted by the bad behavior and questionable practices of others in their industry and have a stake in a sustainable resolution. Three types of firms in the Big Data Industry are of particular importance in creating sustainable industry practices:

1. Possible leaders in the industry, which could emerge from their unique position as gatekeepers, such as consumer-facing companies, website operators and application providers. These companies control how information is initially gathered and how it is subsequently shared.
2. Organizations with unique influence and knowledge in the area of Big Data analytics, such as the American Statistical Association and the Census Bureau, as well as HHS and the National Research Council (which govern academics' Institutional Review Boards). These organizations have the stature and deep knowledge of research, data sets, analytics and confidentiality to begin to set standards of practice.
3. Providers of key products within the Big Data Industry, such as Palantir, Microsoft, SAP, IBM, etc. These companies have few competitors and unique knowledge of analytic products and services, and can offer advice to firms at a critical point to analyze and use Big Data.

As this article has shown, the ethical issues and problems facing the Big Data Industry are similar to those faced by other industries. Practical solutions to creating mutually beneficial and sustainable relationships within the industry include visible data stewardship practices, greater data due process internally and using the services of a data integrity professional. These solutions, which are summarized in Table 3 and Figure 4, directly address the issues identified in this article. (Table 3 also describes how CIOs and CDOs can address the problems.) Despite the potential to create harm, the Big Data Industry has the potential to be a force for good and the focus therefore should be on implementing the solutions described below to create value for all stakeholders.[40]

1. Identify and Communicate Data Stewardship Practices

Current information supply chains are not visible, putting consumers at a disadvantage in choosing preferred supply chains or holding a firm responsible for its decision to join a particular supply chain. Such information asymmetries could be minimized by clearly illustrating the upstream sourcing information and downstream use in order to report the data stewardship practices. Data stewardship includes the rules about internal treatment and external sharing of information for different types of data. Industry groups can develop data stewardship best practices for firms and, more importantly, coalesce around a format for communicating data stewardship practices.

Making the supply chain visible will clearly identify a firm's position in the chain and enable the firm to take responsibility for the upstream and downstream practices of others. A firm's different upstream sources of information, the type of information collected, its internal uses

39 For a comparison of regulating the credit reporting industry with regulating Big Data, see Hoofnagle, C. J. *How the Fair Credit Reporting Act Regulates Big Data,* paper presented at Future of Privacy Forum Workshop on Big Data and Privacy: Making Ends Meet, September 10, 2013, available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2432955.

40 For a balanced view on solutions that both optimize the use of technology and respect privacy and ethics, see Mayer, J. and Narayanan, A. "Privacy Substitutes," *Stanford Law Review Online* (66), 2013, pp. 89-96; and Bambauer, J., op. cit., 2014.

Table 3: Possible Solutions to the Big Data Industry's Ethical Issues

Type of Issue	Cause of Problem	Potential Solution	As Faced by CIOs and CDOs
Supply Chain Sourcing and Use Issues	**Data Stewardship**		
	Firms not accountable for conduct of upstream sources and downstream customers	Illustrate role of firm in larger supply chain Make machine readable notification of supply chain information available to policy makers, reports and privacy advocates	Identify and take ownership of upstream sources and downstream customers/uses of data Ensure information about data stewardship practices is available to experts and novices
	Supply chain not visible	Make data stewardship practices of supply chain visible	Do not enter into confidentiality agreements that preclude explaining your data partners, either upstream sources or downstream users
Surveillance as Negative Externality	**Data Due Process**		
	Harm to others not captured by firms collecting, storing or using personally identifiable information (PII)	Minimize surveillance	Make tracking visible to consumer
		Internalize cost of surveillance with increased data due process	(Industry) Require additional data due process for firms acquiring and retaining PII
Destructive Demand for Consumer Information	**Data Integrity**		
	Secondary market of data trading has lower quality requirements than primary consumer-focused market	Use a data integrity professional when handling or selling PII	(Industry) Institute data integrity professional or board for projects partnering with Big Data sources and customers
	Secondary market is not visible to primary market (consumers)	Make activity in secondary market visible to regulators and consumers	Account for and communicate additional risk from partnering in secondary market for Big Data through disclosure

and storage, and the firm's possible downstream customers and recipients are all important for understanding the entirety of the supply chain and the firm's data stewardship practices. An illustrative example is shown in Figure 5. The data sources, type of data and level of identifiability are important for understanding the upstream sourcing practices; the firm's primary use, secondary use and storage explains the purpose and vulnerability of the data; and the types of data, recipients and level of trust in the recipients explains the downstream uses of the data collected.

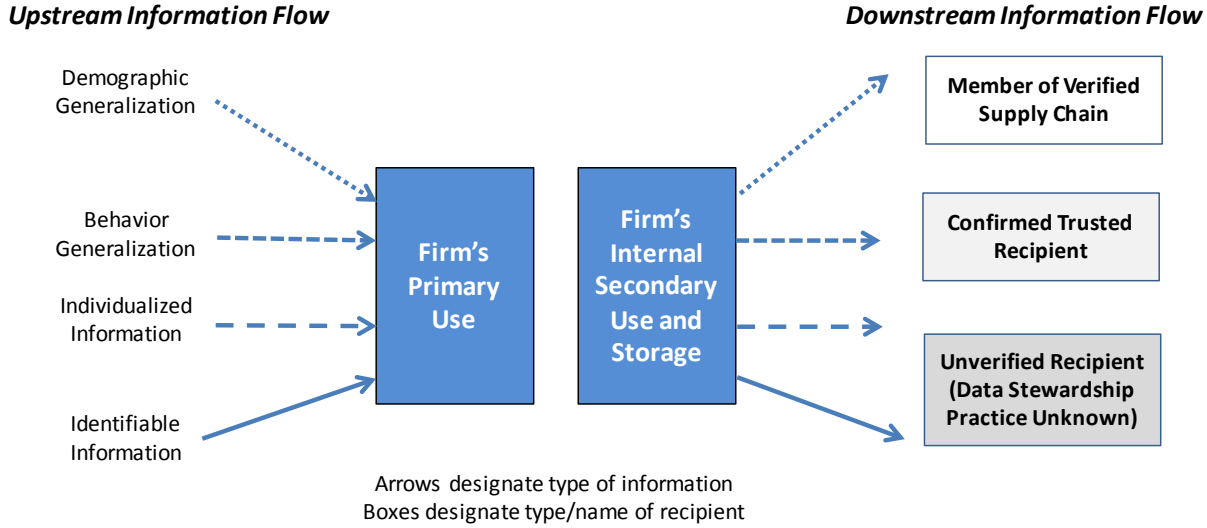
While the information supply chain may look complicated, a similar problem has been resolved in areas such as free-trade coffee, organic food and sustainable fishing: trusted supply chains are identified, certified and valued by customers and customer groups. The information supply chain of a particular firm should be similarly available to

Figure 4: Guidelines for a Sustainable Big Data Industry

Problems	Supply Chain Sourcing and Use Issues	Surveillance as Negative Externality	Destructive Demand for Consumer Information
General Solutions	*Make supply chain visible to technologists, researchers, consumers and regulators*	*Decrease surveillance harm and internalize costs when contributing to surveillance*	*Make secondary market for consumer data visible*
Guidelines for the Big Data Industry			
1. Identify and communicate data stewardship practices	Make data sources and uses of information visible and searchable (see Figure 2)		Clearly identify firms within an information supply chain
2. Differentiate data due process model for PII and non-PII*		Internalize surveillance cost: additional data due process required when retaining PII	
3. Quantify activity in secondary market for Big Data	Communicate to consumers, regulators and investors the value created and associated risk from activity in secondary market for information		Explain % of data sold and % of sales from selling information in secondary market for information
4. Institute data integrity professional or board for Big Data Analytics	Ensure adherence to and compliance with stewardship norms through professional data integrity	Internalize surveillance cost: require data integrity professional or board when using personal data (PII)	Make primary data collectors responsible for quality of information gathered

*The ability to fully differentiate between personally identifiable information (PII) and non-PII is debatable, as argued by Narayanan, A. and Shmatikov, V. "Myths and Fallacies of Personally Identifiable Information," *Communications of the ACM* (53:6), 2010), pp. 24-26.

Figure 5: Example of a Firm's Information Supply Chain Diagram



industry groups, customer groups and regulators that have the knowledge necessary to certify a level of data stewardship within the supply chain. Making information supply chains available in a machine-readable form would support the illustration in Figure 5, as has been developed and effectively called for by Cranor.[41] Users would then be able to identify and choose trusted and certified supply chains.

Providing information about a firm's larger supply chain and data stewardship practices in a uniform way is critical not only for helping users directly, but also for allowing researchers, reporters, technologists and academics to easily diagram and analyze the many different supply chains and provide an audit trail for the data.

2. Differentiate Data Due Process Requirements for Personal Data

Two approaches can be used to manage surveillance as a negative externality: (1) individual firms can reduce their role in contributing to surveillance and (2) the industry can implement policies to internalize the cost of surveillance for firms. First, surveillance is most effective (and therefore most harmful) when the watcher is hidden yet omnipresent.[42] Firms can reduce their role in consumer surveillance by becoming more visible to the consumers and by limiting data collection. The negative externality of surveillance suggests that firms that are invisible to users, such as data aggregators and data brokers, have a special role in the online surveillance system. Both data aggregators and data brokers are invisible to users while aggregating data across diverse sources. Making the tracking of individuals obvious at the time of data collection can diminish the harm of surveillance.

In addition to decreasing the effectiveness and related harm of surveillance, internalizing the cost of surveillance for firms is an effective tool to diminish this negative externality. For example, data brokers and aggregators that store and distribute information within the Big Data Industry could have additional data due process requirements imposed on them for collecting, retaining and distributing personally identifiable information (PII). While some have claimed that PII is not clearly distinguishable,[43] firms that retain information that can be linked back to an individual so it can be fused with other information about the same individual should have an additional obligation of data due process.

Citron and Pasquale outline three areas of data due process requirements, which are instructive moving forward: (1) identifying audit trails, (2) offering interactive modeling and (3) supporting user objections.[44] In addition to firms being required to provide an audit trail for how information is sourced, used and distributed similar to that shown in Figure 5, they could also be required to offer interactive modeling of the use of information and a process to enable individuals to examine and object to the information stored. These additional requirements would impose a cost on those that opt to retain personally identifiable information. The additional obligations would increase the cost of retaining the information, internalize the previously externalized harm (surveillance) and possibly dissuade some firms from using and retaining PII.

Requiring better internal oversight of the data stewardship practices and additional data due process procedures would increase the cost of holding individualized yet comprehensive data and internalize the cost of contributing to surveillance. Many negative externalities are beyond the scope of a single firm to rectify; the

41 Cranor, L. F. "Necessary but Not Sufficient: Standardized Mechanisms for Privacy Notice and Choice," *Journal on Telecommunications and High Technology Law* (10), 2012, pp. 273-307. Rather than focus on the type of information, the firm's storage, information use or third-party access to data would be highlighted if such tactics diverge from commonly accepted practices. Research demonstrates that users care most about the possible secondary use or third-party access to information both online and with mobile devices, as noted by Martin, K., op. cit., 2015; Shilton, K. and Martin, K. E. "Mobile Privacy Expectations in Context," *Telecommunications Policy Research Conference* (41), 2013; and Martin, K. E. "Privacy Notices as Tabula Rasa: An Empirical Investigation into How Complying with a Privacy Notice Is Related to Meeting Privacy Expectations Online," *Journal of Public Policy and Marketing,* 2015.

42 Cohen, J. E., op. cit., 2008.

43 Ohm, P. "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization," *UCLA Law Review* (57), 2009, pp. 1701-1777; and Narayanan, A. and Shmatikov, V. "Myths and Fallacies of Personally Identifiable Information," *Communications of the ACM* (53:6), 2010, pp. 24-26.

44 Citron, D. K. and Pasquale, F. "The Scored Society: Due Process for Automated Predictions," *Washington Law Review* (89), 2014, pp. 1-33; see also Crawford, K. and Schultz, J. "Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms," *Boston College Law Review* (55:1), 2014, pp. 93-129.

cost of reigning in surveillance is too much for one firm to bear and the effect of a single firm changing its data practices would be minimal. For those that wish to use large samples of personally identifiable information, better data governance together with the services of a data integrity professional—who is certified and held accountable for the data practices of the firm—would ensure that data stewardship practices and data due process are followed. Similarly, an internal consumer review board, as advocated by Calo and cited in the draft White House Consumer Privacy Bill of Rights Act of 2015, would similarly internalize the cost of storing and using personally identifiable data.[45]

3. Quantify Activity in the Secondary Market for Big Data

Destructive demand flourishes when the interests of the secondary market for consumer information are not aligned with the primary market and when the secondary market is not visible to the primary market. By linking all relevant firms through an information supply chain, firms in the secondary market have "skin in the game" and thus an incentive to align interests.[46] In other words, by framing themselves as members of a larger supply chain, firms have a vested interest in ensuring others in the chain uphold data stewardship and data due process practices. Otherwise, their reputation would be at risk.

In addition, by making the secondary market more visible to the primary market, the primary market can take into consideration secondary market firms' actions. Consumers may be more (or less) willing to divulge information to a firm in the primary market depending on its type of involvement in the secondary market for selling information. Importantly, the current approach, where the secondary market for Big Data is invisible to the primary consumer-facing market, does not allow for such feedback.

Aligning interests not only benefits the primary market; it can also benefit quality and trusted firms in the secondary market. For example, within the mortgage-backed securities market, unaffiliated financial companies, which did not have interests aligned with the primary market, were not able to sell their securities at the same rate as those companies that were affiliated. In other words, this secondary market recognized the inherent risk of trading with companies whose quality criteria did not align with the consumer market. For the Big Data Industry, history suggests there would be a market for quality data practices in the secondary market for Big Data.

4. Institute Data Integrity Professional or Board for Big Data Analytics

The practical implications of these guidelines call for renewed attention to the training and development of data integrity professionals. The focus of their training should be on incorporating an ethical analysis, which is consistent with FTC Commissioner Julie Brill's focus on the role of technologists in protecting privacy in the age of Big Data, as well as Mayer and Narayanan's call for engineers to develop privacy substitutes within their design.[47]

First, professional data scientists are needed to implement the solutions outlined above to curtail surveillance and destructive demand, as well as to ensure data stewardship practices. Currently, advice for Big Data professionals, including data scientists, data analytics specialists, and business intelligence and analytics specialists, focuses on the challenges in using Big Data, such as leadership, talent management, technology, decision making and company culture. There is little advice on ensuring data integrity.[48]

Second, consumer review boards, made up partly of professional data scientists, would oversee and authorize research on human beings within the commercial space. As Calo notes, academics are required to receive clearance to conduct research from their Institutional Review Board and undertake associated training, even when the research is for societal benefit. Yet private companies conduct research

45 Calo, R. "Consumer Subject Review Boards: A Thought Experiment," *Stanford Law Review Online* (66), 2013, pp. 97-102.

46 For the mortgage-backed securities market, skin in the game—and aligning interests—was effective to avoid losses—James, C.M. "Mortgage-Backed Securities: How Important Is 'Skin in the Game'?," *FRBSF Economic Letter,* December 13, 2010.

47 Brill, J. *A Call to Arms: The Role of Technologists in Protecting Privacy in the Age of Big Data,* Sloan Cyber Security Lecture by Commissioner Julie Brill, Polytechnic Institute of NYU, October 23, 2013; Mayer, J. and Narayanan, A., op. cit., 2013.

48 Chen, H., Chiang, R. H. L. and Storey, V. C. "Business Intelligence and Analytics: From Big Data to Big Impact," *MIS Quarterly* (36:4), 2012, pp. 1165-1188.

without oversight, even when at the expense of the consumer.[49] Revelations that OKCupid and Facebook[50] had conducted experiments on users without their knowledge only show how prescient Calo was in the call for consumer review boards; and effective consumer review boards would require data integrity professionals.

Finally, academic institutions continue to develop degree courses in business analytics, business intelligence and data analytics to train Big Data professionals—but they do not require students to take a course in ethics. A survey of the top 15 such programs shows the intense focus on technique, with little regard given to privacy, ethics or corporate and professional responsibility.[51] Accreditation for such programs should require them both to train data integrity professionals who graduate with a degree in data science, data analytics or business intelligence, and to support the solutions proposed in these guidelines.

Concluding Comments

This article has examined Big Data within the context of the Big Data Industry and identified persistent issues and points of weakness in current market practices. In doing so, it has examined the industry's information supply chain of upstream suppliers and downstream uses of data, the ethical issues arising from the negative externality of surveillance caused by persistent tracking, aggregation and the use of consumer-level data, and the potential destructive demand driven by the secondary market for consumer information. Importantly, the article has identified the Big Data Industry as having both economic and ethical issues at the individual firm, supply chain and general industry level and has suggested associated solutions to preserve sustainable industry practices.

49 McAfee, A. and Brynjolfsson, E. "Big Data: The Management Revolution," *Harvard Business Review,* October 2012.

50 Stampler, L. "Facebook Isn't the Only Website Running Experiments on Human Beings," *Time,* July 28, 2014, available at http://time.com/3047603/okcupid-oktrends-experiments/.

51 The survey includes both bachelor's and master's programs from across schools/programs such as business and engineering. See http://www.informationweek.com/big-data/big-data-analytics/big-data-analytics-masters-degrees-20-top-programs/d/d-id/1108042?page_number=3 and http://analytics.ncsu.edu/?page_id=4184. Programs reviewed include Bentley, Columbia, LSU, NYU, GWSB, Northwestern, Rutgers, CMU, Harvard, MIT, NCSU, Stanford, UT Austin and UC Berkeley. Both *Information Week*'s and NCSU's lists focus on U.S. universities.

About the Author

Kirsten E. Martin
Kirsten Martin (martink@email.gwu.edu) is an assistant professor of Strategic Management and Public Policy at the George Washington University School of Business. She is principle investigator on a three-year grant from the National Science Foundation to study online privacy. Martin is a member of the advisory board for the Future Privacy Forum and the Census Bureau's National Advisory Committee, and is a fellow at the Business Roundtable Institute for Corporate Ethics. Her research interests include online privacy, the ethics of Big Data, privacy, corporate responsibility and stakeholder theory.