How Sustainable Is Big Data?

Charles J. Corbett
UCLA Anderson School of Management
110 Westwood Plaza, Box 951481
Los Angeles, CA 90095-1481
email: charles.corbett@anderson.ucla.edu
tel: +1-310-825-1651
fax: +1-310-206-3337

December 14, 2017

Abstract

The rapid growth of "big data" provides tremendous opportunities for making better decisions, where "better" can be defined using any combination of economic, environmental, or social metrics. This essay provides a few examples of how the use of big data can precipitate more sustainable decision making. However, as with any technology, the use of big data on a large scale will have some undesirable consequences. Some of these are foreseeable, while others are entirely unpredictable. This essay highlights some of the sustainability-related challenges posed by the use of big data. It does not intend to suggest that the advent of big data is an undesirable development. However, it is not too early to start asking what the unwanted repercussions of the big data revolution might be.

Keywords: big data, sustainability, energy, operations, life-cycle assessment (LCA)

Received: December 2017; accepted: December 2017 by Kal Singhal, with no revision.

Introduction

Big data is here to stay, but what are some of the environmental and social consequences of the big data revolution? How sustainable is big data? The advent of big data provides revolutionary new opportunities for increased understanding of the environmental and social impacts of supply chains, with the concomitant potential for improvement along those dimensions. Big data also gives rise to both known and unknown environmental and social challenges. The purpose of this essay is to highlight some of those challenges. My intention is not to argue that big data is a phenomenon to be resisted. However, any technological breakthrough, if adopted on a sufficiently wide scale, will have far-reaching externalities, both positive and negative.

I use the term *big data* according to the emerging consensus (e.g., Etzion and Aragon-Correa, 2016, p. 148), which holds that big data is not necessarily "big" but rather that it is differentiated from "traditional data" by any of the "4Vs": *volume*, *variety*, *velocity*, and *veracity*. Goes (2014) argues that the promise of big data is the ability to exploit various combinations of these 4Vs.

Define *sustainability* loosely as making decisions while simultaneously taking into account economic, environmental, and social considerations. When sustainability is defined this way, it becomes clear that sustainability is inherently intertwined with big data. When we seek to measure the environmental and social impact of our decisions, an explosion in both the volume and the variety of data naturally results. Environmental impacts could be on global climate, health of watersheds, human health, biodiversity, etc., while social impacts could affect workers, consumers, communities, societies, or value chain actors.

In the past, we may have received periodic updates on climatic conditions, or sporadic insights into the treatment of workers at vendor facilities. Now, however, the real-time monitoring of such phenomena at ever greater granularity results in a much greater *volume* and *velocity* of data. Now that such data can include anything from temperatures to satellite images to social media posts, the *variety* is widening too. The *veracity* of data also varies widely, depending on factors such as whether weather data are observed or extrapolated, or whether worker conditions are self-reported or independently verified.

Although the main purpose of this essay is to highlight sustainability-related challenges associated with big data, I do not want it to sound negative. So, I will first provide a few (unpresentative and unscientifically selected) examples of the exciting benefits and opportunities that big data already provides or promises.

Examples of big data and sustainable operations

This section offers a few examples of how big data is, or can be, used to enhance our understanding of the impacts supply chains have on environmental and social conditions, and vice versa. These examples are not intended to be comprehensive or representative. Instead, they are provided to illustrate the wide range of (potential) applications of big data to sustainable operations, and to serve as a counterweight to the subsequent section's emphasis on the emerging sustainability-related challenges associated with big data.

Many of the examples of firms successfully reducing their environmental footprint involve harnessing large amounts of timely data. The importance of performance measurement as a tool to improve environmental performance has long been known, as the examples in Corbett and Van Wassenhove (1993) illustrate. One good example is the energy management program at Walt Disney World (Allen, 2005), which was initiated in the late 1990s and already involved collecting hourly information on the consumption of electricity, water, and other resources at a highly localized level. By gathering and sharing this information with the appropriate stakeholders (e.g., the maintenance crew or executive

managers), Walt Disney World has been able to reduce its annual electricity usage by some 100 million kilowatt hours while earning a 53% internal rate of return for their efforts. Monthly report cards provide historical and benchmarking information; showing the managers of Epcot how they performed relative to the Animal Kingdom helps to generate healthy competition among them. Conversely, real-time monitoring of a building's HVAC system allows a repair crew to be notified immediately if a control for a door to an air-conditioned space malfunctions in the evening, rather than only finding out when the next utility bill arrives a month or two later. The velocity with which the data is collected, processed, and shared, is critical to ensuring the data has the intended impact on energy consumption.

In a different setting, Marr (2017) points out how Caterpillar's Marine Division uses shipboard sensors to monitor a wide range of systems. The resulting data provide new insights into optimal operating practices; in one instance, a customer discovered that running more generators at lower power is more efficient than running fewer generators at maximum output. Big data help firms design better materials (National Academy of Sciences 2014), such as new photovoltaic materials with higher efficiency (p. 17) or GM's new thermoelectric materials for higher fuel efficiency (p. 24).

It is well understood that a changing climate will have a wide range of consequences for all kinds of organizations and supply chains. The exact effects of climate change on the conditions in any given location are still not well understood, but the combination of increasingly comprehensive historical data and fine-grained climate simulation models allows for more tailored predictions of how different regions will be affected. Some activities, such as wine growing and ski tourism, depend heavily on highly local microclimates, so forecasts must be available at a much more detailed spatial scale than was previously possible. For instance, Jones et al. (2005) combine data on the ratings of wines from regions around the world with a widely used climate simulation model (HadCM3) to predict how the wines from each region will be affected, positively or negatively, by changes in local climatic conditions from 2000–2049. Ski tourism, which depends on the thickness and the persistence of snowpack on specific slopes, provides a similar example. Sun et al. (2016) used statistical downscaling to predict that various mountain locations in Southern California will be snow-free several weeks earlier by mid-century than is currently the case. Investors in winter sports facilities would do well to consider this kind of detailed forecast. In some instances, the activities of the supply chain itself cause changes in weather patterns: Thornton et al. (2017) combine data on shipping emissions and on 1.5 billion lightning strikes on a 10×10 kilometer grid to find that the number of lightning strikes was elevated by 20–100% along polluted shipping lanes. These are just a few examples; there are numerous similar studies.

This kind of data gathering is no longer constrained to earthbound monitoring stations, as various satellite-based systems provide more data with more detail and at higher frequencies. This trend will continue to accelerate because of the constant effort to miniaturize satellites. Woellert et al. (2011) outline a range of opportunities being opened by the use of CubeSats. These are small satellites that have a mass of around one kilogram and a volume of 10 cubic centimeters. CubeSats provide much more fine-grained monitoring of atmospheric conditions. These cheap satellites will also allow near real-time tracking of animal populations, and they can help disaster relief agencies allocate resources by providing images of earthquake damage. Satellite imagery is already being used to detect illegal logging after the fact, but Lynch et al. (2013) argue that daily observations are required if we are to take preventive action against illegal logging, instead of just observing it from a distance. Laurence and Balmford (2013) propose that satellite data could help prevent inappropriate road-building. This would help prevent ecological disasters long before they occur, because even a single road through a forest can wreak environmental damage far out of proportion to the physical footprint of the road itself. All these kinds of analyses will require a staggering amount of data. And, the attendant databases will rapidly become massive, because many studies of this kind require longitudinal data with very fine spatial and temporal resolution.

At the opposite end of the spectrum, firms are becoming more and more interested in the conditions experienced by workers in their supply chains. Traditionally, much work in this area depended on sending auditors, third-party or otherwise, to assess the extent to which factories implemented the environmental and social practices expected of them. With the advent of smartphones, workers in factories around the world can now directly and anonymously report any conditions or practices they are exposed to. For instance, Walmart uses such worker-generated data collected through LaborVoices (de Felice 2015, p. 553). LaborLink is a similar effort. Firms must navigate between great opportunities and serious challenges to make the best decisions about how to use this data, since it comes from individual workers, in real time, covering a range of dimensions, and with unknown accuracy.

It would be easy to provide countless other examples of how big data can enhance sustainability, whether by allowing firms to make better decisions about the operation of their supply chains, or by allowing regulators to exert tighter control over those supply chains. The opportunities are boundless and exciting. Nevertheless, the big data revolution is also rapidly generating sustainability-related challenges of its own, which is the subject of the next section.

Sustainability challenges related to big data

The advent of big data presents unbounded opportunities to improve decision-making and to ensure more sustainable outcomes. However, like any other technological advance, it also brings challenges that will become increasingly acute as the use of big data becomes more prevalent. Some of these challenges may seem far fetched today, but recall that the internal combustion engine was once considered an environmental breakthrough (Kirsch, 2000). It is unlikely that advocates of fossil fuel–powered vehicles in the early 1900s could have had an inkling of the dramatic effects this technology would have on global air quality and climate over the course of the subsequent century. As mentioned earlier, the intention of this essay is not to argue that the rise of big data is an undesirable trend; the intent is to stress that we should be cognizant of some of its concomitant downsides. Below, I will review some of these risks associated with using or managing big data.

Social and ethical consequences of using big data

When initially deciding where to roll out its Prime Free Same-Day Delivery service, Amazon aimed to serve as many people as it could, using its data to identify ZIP codes that contained a high concentration of Amazon Prime members. Maps of various cities produced by *Bloomberg BusinessWeek* (Ingold and Soper, 2016), show the areas within the city limits that initially received same-day delivery. In the case of Boston, it shows that the Roxbury neighborhood was excluded, while all surrounding neighborhoods did receive same-day delivery. The extent to which Roxbury was an anomaly is highlighted when one considers how far beyond the city limits the same-day delivery area stretched. The population of Roxbury is 59% Black. The *Bloomberg BusinessWeek* analysis found similar (though not quite as striking) effects in other major cities.

Without a doubt, Amazon did not set out to distinguish neighborhoods based on racial or ethnic composition. In the *Bloomberg BusinessWeek* article, an Amazon spokesperson stated, “Demographics play no role in it. Zero.” However, the result of their analysis was undesirable enough that Amazon rapidly backtracked. They added same-day service to initially excluded regions in Boston, New York and Chicago. Even though race played no direct role in Amazon’s analysis, the algorithm they used led to “apparent discrimination,” as defined in Galhotra et al. (2017).

Fairness in algorithms is now the subject of significant research efforts, as also mentioned in Cohen (2017, p. 25). For instance, Calders and Verwer (2010) define a *discrimination score* to measure the strength of group discrimination, and Galhotra et al. (2017) highlight some limitations of that score and generalize it to settings with more complex inputs. They also propose ways of testing software for fairness, something that Amazon presumably wishes it had done before rolling out the early phase of its same-day delivery program.

A different way in which virtually all of us have been negatively affected by big data is through the various hacks that have occurred over the years, exposing our personal data to unauthorized parties. Equifax revealed a particularly large data breach in September 2017. The company disclosed that sensitive personal information for some 146 million customers were stolen, including financial records and social security numbers. Many other large organizations around the world have been the subject of similar hacks.

The vast majority of individuals whose data was compromised are not directly financially affected, beyond perhaps the cost associated with additional identity protection services. Lai et al. (2012) observe that there appears to be relatively little research on the consequences for victims of the subsequent potential identity theft. They mention various studies that estimate the total costs to consumers on the order of $50 billion in 2008–2009, or $6,000 per victim (p. 353). The horror stories reported by victims such as Amy Krebs are more salient (Shin, 2014). Cohen (2017, pp. 21–22) mentions that the breach at Ashley Madison, a Canadian online dating site that specializes in extramarital affairs, had severe consequences for families affected, and may have led to several unconfirmed suicides. The social and emotional costs incurred by individuals following data breaches can be substantial, in turn causing reputational damage and increased regulatory scrutiny of the firms that were hacked. Even firms that were not hacked can suffer consequences when a competitor is breached: Experian issued a warning about the risks it may face due to the "increased legislative and regulatory activity" that followed the Equifax breach (*Financial Times*, 2017).

Discussions of various ethical aspects of big data are emerging. Zwitter (2014) observes that global big data is shifting the power balance between the various stakeholders. They argue that major consequences can follow from many small actions taken by many individuals online (such as re-tweets or Facebook likes), and these consequences require a different perspective on what constitutes ethical behavior. Even offline actions, such as an individual parking his or her car in front of his or her own house, can be used to predict information such as demographics and voting behavior, as Gebru et al. (2017) describe in their application of deep learning to images from Google Street View. They also note that this raises important ethical concerns (p. 5). Richards and King (2013) highlight three paradoxes related to the ethics of big data. First, they note that although big data supposedly enhances transparency, much of the process by which the data are collected and analyzed is invisible to the public. Second, they pinpoint that in many cases, even though big data is about large numbers of individuals, its purpose is often precisely to identify specific individuals. And third, they also caution that big data will fundamentally change existing power structures.

A particular ethical quandary associated with big data is how it is used to make technology as addictive as possible. As Alter (2017) notes, the people who design games, websites, and other interactive experiences run endless tests on millions of users to collect massive amounts of data on which features (fonts, sounds, swipes, incentives, etc.) maximize user engagement and keep users coming back time and again. In Alter's (2017) words: "As an experience evolves, it becomes an irresistible, weaponized version of the experience it once was. In 2004, Facebook was fun; in 2016, it's addictive." Tristan Harris, a former Google Design Ethicist, lists ten ways in which product designers hijack our psychological vulnerabilities in order to keep our attention focused on their creations; among others, he lays out how our phone and the collection of apps that reside on it are like carrying slot machines in our pockets (Harris 2016). He argues

that designers should use the data at their disposal to protect us from our addictive tendencies, rather than exploit them; we should protect our time just as we protect our privacy.

To summarize, the use of big data is generating a wide range of ethical challenges. Some are more obvious than others, but they must be addressed if society is to reap the many positive benefits that big data has to offer.

Big data may not be the right data

How do we prioritize all this data? We have ever-growing amounts of increasingly fine-grained data on variables related to climate, and we should use that data by all means. But, when world leaders adopted 17 Sustainable Development Goals (SDGs) at a United Nations summit in September 2015, climate change mitigation was Goal 13 of those 17[1]. It may be harder to measure progress on some of the other SDGs, such as "zero hunger" (Goal 2) or "peace, justice, and strong institutions" (Goal 16), but that difficulty does not mean those areas should be neglected.

Even within each goal, it is essential to first define the objective, and only then try to determine appropriate indicators, rather than the reverse. Hák et al. (2016) point out the danger of letting data availability drive priorities: "Operationalisation of the targets through indicators would be methodologically incorrect and might lead to distortions in the development of the policy agenda (such an approach might cause the false interpretation that only what can be measured is important to our lives)" (p. 568). This lack of data is not an idle threat; Sachs (2012) identified data shortcomings as "one of the biggest drawbacks" (p. 2210) of the Millennium Development Goals (the predecessor to the SDGs). With all the excitement about the vast amount of data becoming available for analysis, we must always ask what is not being captured.

Moreover, even when data appears to exist, it may not be correct. Veracity is always a concern. Firms and nations increasingly formulate quantitative targets related to sustainability: For example, firms set "science-based targets" and nations work together under the Paris agreement to reduce greenhouse gas emissions. The success of this kind of initiative inevitably hinges on the credibility of the associated emissions data. Ongoing debates, such as the discussion about China's CO_2 emissions, indicate that there is as much as 10% uncertainty about the magnitude of those emissions over the period 2000–2013, which could be a decisive factor in whether China's cumulative emissions will be consistent with a 2 °C warming target (Liu et al., 2015; Korsbakken et al., 2016). At the firm level, Melville and Whisnant (2014) document various kinds of errors two firms made in their greenhouse gas emissions reporting. Blanco et al. (2016) point out that firms' reports of supply chain emissions are even more difficult to interpret. Melville et al. (2017) find that firms that paid attention to the accuracy of their carbon emissions data also achieved lower emissions. In the context of big data, these errors would suggest that improving the veracity of sustainability-related data may be beneficial in itself.

Finally, even when the data is correct in a narrow technical sense, it can easily be abused. With data arriving ever more rapidly, it is easy to fall into the trap of churning out analyses and rankings without due consideration of the underlying phenomena or the impact of those rankings. There are multitudes of rankings of countries, states, or firms on all kinds of environmental or social metrics. These rankings may sometimes have some informative value, but it is rare that a device as simple as a ranking can capture the many nuances involved in social or environmental concerns. Delmas et al. (2013) provide an illustration of how combining multiple rankings of firms' corporate social responsibility performance can lead to substantially better insights than any single ranking by itself. Many readers of this essay will find the

[1] See http://www.un.org/sustainabledevelopment/development-agenda/, last accessed November 15, 2017.

concerns about business school rankings to be all too familiar. These ranking methods are dissected in Bachrach et al. (2017), who argue that the fundamental flaws in the methodologies used to rank business schools negatively impact those schools' ability to meet their social obligations. This danger of the misuse of big data is well examined in Lazer et al. (2014). They use the large error in predictions made by Google Flu Trends as an example of "big data hubris," which is the "often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis" (p. 1203). They close with the reminder that, despite the enormous opportunities provided by big data, much information is also contained in "small data" (p. 1205) that is not, or cannot be, contained in big data.

In short, despite the excitement associated with emerging big data related to environmental and social indicators, "small data" is still a critical component of environmental and social progress too.

Big data may not mean better decisions

Part of the implicit premise underlying the excitement around big data is the assumption that more data will lead to better decisions. There is a vast literature on biases and heuristics (e.g., Kahneman 2011). Muthulingam et al. (2013) provide one example in a sustainability-related context: they document that managers who are faced with well-structured information about energy-efficiency initiatives will disproportionately choose items that appear clower to the top of the list, even when other initiatives further down are economically and environmentally superior. In the face of more data, these biases are likely to persist, or possibly become even more acute, due to the additional cognitive burden associated with big data. The concern that more data might lead to worse decisions is not new, as illustrated by a well-known quote from Herbert Simon (1971, pp. 40-41): "In an information-rich world, the wealth of information means a dearth of something else: a scarcity of whatever it is that information consumes. What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources that might consume it."

Even when there is no obvious bias at work, individuals may not respond to big data in the intended manner. One well-documented instance is the "rebound effect" in energy conservation, in which individuals who adopt energy-efficient technologies sometimes partly or fully negate these energy savings by consuming more energy elsewhere. Asensio and Delmas (2016, p. 207) find mixed evidence on this effect using real-time energy consumption data from 118 households over 9 months, yielding 374 million observations. The households in the treatment group were given feedback on their energy use, either in terms of cost or environmental health. The health treatment generally led to more durable energy conservation, but the households that received the cost treatment, after an initial reduction, ended up increasing their energy use related to heating and cooling. Even more surprisingly, both treatment groups experienced an increase in refrigerator energy use, which the authors attribute to unintuitive design of the temperature controls in the appliances concerned, making it hard for the participants to know which way to adjust the knob. This goes to show that no amount of data can compensate for the simple issue of poor product design. (Although, without the large dataset produced during this study, the product design issue would not have become clear.)

Discussions about big data often focus on the increasing *volume* of data, but the increasing *variety* of data poses growing challenges for decision-making. Decisions that have environmental consequences often involve data about a range of impacts, such as neurotoxicity, carcinogenicity, biotoxicity, global warming potential, land use effects, and various social indicators that may also be relevant. Multi-criteria decision methods can help to inform policies or decisions that balance such a range of environmental and social

indicators in an appropriate manner. The use of these methods is growing, but still relatively nascent (Linkov and Moberg, 2012; Ch. 2), and these methods are not without their drawbacks.

Consider the case of Alternatives Analysis, an approach intended to identify safer chemicals while avoiding inadvertent substitution of toxic chemicals with even more undesirable substances. This approach is a response to chemical policies changing from a risk management focus to a prevention focus. Data on the human health and ecosystem health impacts of a chemical become more reliable as we gain more experience with that chemical. However, negative "experience" is precisely what policies such as the California Safer Consumer Products program and the European Union's Registration, Evaluation, Authorization, and Restriction of Chemicals (REACH) program seek to avoid. By the time we have big data on the ecological or human health impacts of a chemical, we are already well past the time for prevention and well into the time for risk management and mitigation. Malloy et al. (2016) describe some of the challenges involved with making early assessments about safer alternatives when the data on risks and impacts is severely incomplete and multidimensional. Linkov and Moberg (2012) provide an in-depth discussion of how multi-criteria decision analysis can be used in a variety of environmental decision contexts. A recent workshop we held at UCLA with some 15 participants involved in various aspects of alternatives analysis revealed that there is a great need for more systematic approaches to making these kinds of decisions, but also that application of existing multi-criteria decision methods to these questions is far from straightforward. In each of these settings, the main challenge is in how to deal with the *small* data, rather than the big data.

To summarize, big data can clearly help inform environmental and social decisions in some situations, but in other contexts, we must focus on making better decisions with little or no data rather than waiting for more information to arrive.

Big data can change the manufacturing landscape

Big data is not only a technological revolution in itself. It will also facilitate other potentially large shifts, including in the physical world. One such potential consequence is that the spread of big data may foster wider application of mass customization, which in turn would likely be linked to a broader use of 3D printing or additive manufacturing. Two other articles in this special issue, by Feng and Shanthikumar (2017, p. 17) and Guha and Kumar (2017, pp. 13–14), also point to this potential effect.

This raises the question whether additive manufacturing is more or less sustainable than conventional production. Several recent articles, including a special issue of the *Journal of Industrial Ecology*, investigate this question. The overall takeaway is that the answer is not obvious. Summarizing existing studies, Kellens et al. (2017) observe that the specific energy involved in additive manufacturing is "1 to 2 orders of magnitude" (p. S63) higher than for conventional manufacturing. However, they also note that additive manufacturing can lead to environmental benefits if a larger portion of a damaged part can be reused in a repair process, or if parts are redesigned appropriately. This need for appropriate redesign was also found by Mami et al. (2017) in the manufacture of aircraft doorstops. In addition, Walachowicz et al. (2017) find that additive manufacturing has lower impact along various dimensions than conventional processes for the repair of gas turbine burners. In the case of injection molding, Huang et al. (2017) estimate that additive manufacturing uses slightly less energy than conventional manufacturing. In the manufacture of eyeglasses, Cerdas et al. (2017) find that the comparison between additive and conventional manufacturing depends heavily on the material used. They caution that one of the potential benefits of 3D printing is to allow more distributed manufacturing, but that such a more dispersed production system is harder to regulate. Taking a life cycle perspective, Huang et al. (2016) predict major savings due to the weight reduction in aircraft components that 3D printing allows.

The point here is not to argue that 3D printing is the only innovation that is likely to be accelerated by the big data revolution, nor that 3D printing is more or less sustainable than conventional manufacturing. The point is to highlight that the consequences of big data on sustainability are likely to reach well beyond big data itself and even into the physical world. The example of CubeSats mentioned earlier is another instance of this linkage. The demand for the big data that CubeSats can provide will have real environmental consequences, when the satellites are launched but also when they become space junk.

Managing and storing big data

There is no question that in many instances, better data can help reduce energy or material consumption, but managing that data still requires physical processes, which consume energy. A widely cited analysis by Gartner (2007) claims that information and communication technology (ICT) accounted for about 2% of global CO_2 emissions in 2007, and this amount was comparable to the emissions associated with aviation. Although Malmodin et al. (2010) point out that the comparison is distorted, they confirm that the 2% estimate is about right. (They estimate the ICT portion of CO_2e emissions as 1.3%.) In other words, the energy consumption of ICT is not huge, but it is already significant, and it is growing.

Predicting how the impact of ICT will evolve over time involves balancing two counteracting factors. On the one hand, storing and transmitting data is becoming more efficient over time. Aslan et al. (2017) estimate that electricity intensity of data transmission has decreased by about 50% per year since 2000. Operators of some of the largest data centers increasingly rely on renewable energy. Apple's data centers operate on 100% renewable energy (Apple 2017, p. 41), and Google will reach 100% renewable energy for its operations in 2017 (Google 2017, p. 9); Facebook is similarly committed to powering its operations with 100% renewable energy (Facebook 2017). On the other hand, our data footprint is growing dramatically, and with it the energy required by data centers, data networks and connected devices. The International Energy Agency projects that data center electricity use will increase by 3% by 2020, despite a tripling in workload; its forecast for data networks ranges anywhere from a 70% increase to a 15% decrease by 2021 (IEA 2017, p. 103). An encouraging sign is that, as Khuntia et al. (2017) find, firms that invest in green IT not only achieve lower IT equipment energy consumption, but also earn higher profits.

Whether the energy used for storing and transmitting data is renewable or not, that energy has to be generated somehow, and even renewable energy comes with (significant) costs, in the form of land use, material use, noise and visual pollution, and more. A rapid increase in the energy demands associated with big data is therefore a concern which we need to confront. How can we translate figures such as "2% of global CO_2 emissions" to numbers on a scale that apply to individuals or companies? For instance, what are the greenhouse gas emissions associated with storing 1 TB of data? Estimates vary widely. An analysis by the Natural Resources Defense Council (2012) compares various technologies and data hosting scenarios, and the result was a range of 0.6 kg CO_2e per year (i.e., using best practices, public cloud storage) to 15.9 kg CO_2e per year (worst case scenario, on-premise with virtualization) (see Figure 2). The high end of this range translates to 15.9 metric tons of CO_2e per year per TB of data, which would be comparable to the annual emissions of several passenger vehicles. This seems unreasonably high. A report by Google (2011, p. 6) estimates that keeping email on a locally hosted server at a small business causes 103 kg of CO_2 per year per user. Because this still seems high, I asked an expert in this area, Professor Eric Masanet at the McCormick School of Engineering at Northwestern University. He provided some very helpful estimates that seem more believable (Masanet, 2017).

Taking data from Shehabi et al. (2016), a Lawrence Berkeley National Laboratory report he coauthored, he estimates US data centers stored about 300 million TB of data in 2016 (Figure 12, p. 14), which consumed around 8.3 billion kWh (Figure 16, p. 17), hence 27.7 kWh per TB of data. This is the energy directly used for storage by the IT equipment, which we must multiply by the power usage effectiveness

(PUE, or *total data center facility energy use ÷ IT equipment energy use*) to obtain the total consumption of the data centers, which includes the energy consumption of associated technology such as cooling and lighting. According to Figure 21 in Shehabi et al. (2016, p. 25), the total data center energy usage is about 72 billion kWh in 2016, of which 43 billion kWh was for IT equipment, yielding a PUE of 72 ÷ 43 = 1.67. The total energy consumption of data centers is then 1.67 × 27.7 kWh = 46.33 kWh per TB of data per year.

Converting this to CO_2e emissions is not straightforward, since energy consumption varies widely across data centers. Using an EPA (2017) estimate for the emission factor (the US national weighted average CO_2 marginal emission rate) of 0.744 kg of CO_2 per kWh, the carbon emissions associated with data storage would be 0.744 × 46.33, or approximately 35 kg of CO_2 per TB per year. Of course, this is a very rough estimate at best, and with changes in storage technology and energy mix, this footprint per TB will likely improve. At the same time, with the continuation of big data and the advent of augmented reality and virtual reality, our data footprint will likely increase.

The numbers so far have only focused on storage, but *transmitting* data also consumes energy. Coroama et al. (2013) estimate that a videoconference transmission between Switzerland and Japan in 2009 accounted for 200 kWh per TB, substantially above the 46.33 kWh per TB estimated above for storage. Weber et al. (2010) find that online delivery of music to consumers generally causes a lower carbon footprint than physical CDs, but Mayers et al. (2014) predict that downloading large games over the Internet will cause higher carbon emissions than physical delivery via Blu-Ray discs. Altogether, these (and other) studies illustrate that, while there is considerable uncertainty about the energy impacts of big data storage and transmission, they are large enough to take seriously.

What makes all these figures more disconcerting is that the vast majority of data stored can safely be considered as waste. This is sometimes referred to as "dark data," or "the information assets organizations collect, process and store during regular business activities, but generally fail to use for other purposes."[2] Again, reliable estimates of how much dark data exists are hard to come by, but the numbers are quite astounding. IBM estimates that 90% of all data stored is never used (Johnson, 2015). Paul Rogers, chief development officer at General Electric, stated in Wharton (2014) that "[o]nly about one-half of 1% of the world's data is being analyzed" (p. 2), which means that the other 99.5% is dark data. Cohen (2017, p. 3) cites another source that mentions the same figure of 0.5%. A Veritas survey suggests that 54% of data is dark, while 32% is ROT (*redundant*, *obsolete*, *trivial*), and only 14% is business-critical. This survey also estimates there will be almost $900 billion in avoidable costs of data storage by 2020 (Hernandez, 2015). The prevalence of dark data and the costs of data storage suggest that a sizable proportion of the energy use associated with big data is avoidable, and we must begin to consider the waste hidden in big data in the same way we think about physical waste. It is not just energy that is wasted: Shehabi et al. (2016, p. 28) estimate that data centers in the US were responsible for the consumption of 626 billion liters of water in 2014.

We need to start using tools such as value stream mapping for data flows, just as we already do to identify waste in physical flows. This would help uncover the vast amounts of unnecessary and obsolete copies of data currently being stored because we have not yet started treating it as actual waste with a real cost. The ISO/IEC 38505 series of standards for governance of data provide a helpful organizing framework for how to collect, store, report, decide, distribute, and dispose of data. The concepts of the circular economy that are already being applied to supply chains (e.g., Agrawal et al., 2017) might also help curtail the unbounded growth of data being transmitted or sitting in storage. As data become too big, it will become excessively costly to distribute it, so it will need to be analyzed locally at data repositories. Space missions may provide inspiration on how to do this: the explosion of data collected during a mission far

[2] https://www.gartner.com/it-glossary/dark-data, last accessed October 27, 2017.

exceeds the severe constraints on what can be transmitted back to earth, which means that local data reduction will be necessary (National Academy of Sciences, 2014; p. 12).

This is a good time to recall the analogy of the internal combustion engine. When that technology was in its infancy, people were certainly concerned about the visible air pollution it created, but the thought that something as small as a car could change our climate must have been inconceivable. Now, several billion cars later, we know better. The point I wish to make here is that, even though the environmental benefits of big data are often large and obvious, it is not too early to start measuring and minimizing the environmental costs, as some firms already do by investing in renewable energy to power their data centers. This applies to big data, but also to other complementary technologies that have surfaced earlier in this essay, such as additive manufacturing and CubeSats.

Conclusions

We have only begun to explore the potential of big data to improve decision making in many areas of life. These applications permeate the field of sustainability, broadly defined. I have described a few examples, and other articles in this special issue provide more. Devalkar et al. (2017) outline how big data can help agriculture in India. Swaminathan (2017) describes opportunities for using big data to assist in humanitarian operations. These opportunities are exciting and profoundly promising, and we should pursue them accordingly.

However, while doing so, we should not lose track of the fact that rushing to collect and exploit ever "bigger" data will inevitably have undesirable side effects. Some of these side effects have already surfaced, but others may arise in unexpected areas. We should not dampen the excitement deservedly attached to big data, but we should also be vigilant about potential side effects. I have catalogued some of these potential unwanted byproducts. Some of them may turn out to be irrelevant in the long term, but some others not mentioned here will surely emerge.

One often hears the slogan *Big data is the new oil*, but like all analogies, it only goes so far. It is true that fossil fuel was a critical driver of growth and change in the global economy during the twentieth century (*The Economist*, 2017), but as Thorp (2012) points out, information is inherently renewable, unlike fossil fuel. Thorp does draw a different parallel, arguing that oil has also been the cause of untold environmental and social devastation. He observes that "data spills" have already occurred, and he asks when we might encounter "dangerous data drilling practices" or suffer the long-term effects of "data pollution." In her discussion of ethical issues in big data, Martin (2015) also refers to the surveillance that results from the systematic way in which individual data is collected as pollution. She draws a number of parallels between traditional supply chains and information supply chains, with implications for how data should be managed throughout the supply chain to minimize the negative aspects of the growth of big data. For instance, just as manufacturers are concerned about ethical sourcing, firms in the big data industry should ensure that the data they rely on are obtained ethically.

One could argue that, in its earliest days, the fossil fuel revolution was mostly beneficial and relatively harmless. Its disastrous side effects were the result of the sheer breadth and depth of the penetration of fossil fuel–based products into every aspect of human life. Moreover, the collateral inertia associated with the vast investments made over the years have created the pronounced path dependency that has caused so much difficulty as we try to migrate away from fossil fuels.

We are now making important decisions about big data—decisions about issues such as technology platforms, governance mechanisms, ownership structures, and access rights. All these decisions could have pivotal implications for what options will be available to us later, when the costs of big data start

coming into focus. In order to ensure we use big data in a sustainable way, we must always be on the alert for potential repercussions, even repercussions that seem far fetched to us now. The big data revolution has opened a vast uncharted frontier, and we must explore this frontier with enthusiasm, but also with caution.

Acknowledgements

I am grateful to the Editor, Kal Singhal, for providing this opportunity; to Eric Masanet for his quick and detailed response to my questions about energy use of data centers and for other suggestions, and to Christian Blanco, Suresh Muthulingam and Vincente LeCornu for helpful comments on an earlier version of this essay. As usual, all errors are my own.

References

Agrawal, V.V., A. Atasu, L.N. Van Wassenhove. 2017. New Opportunities for Operations Management Research in Sustainability. *Manufacturing & Service Operations Management*. Forthcoming.

Allen, P. 2005. How Disney Saves Energy and Operating Costs. HPAC Engineering. Retrieved December 2, 2017, http://www.hpac.com/building-controls/how-disney-saves-energy-and-operating-costs.

Alter, A. 2017. Tech Bigwigs Know How Addictive Their Products Are. Why Don't The Rest Of Us? Wired, March 24, 2017. Retrieved December 13, 2017, https://www.wired.com/2017/03/irresistible-the-rise-of-addictive-technology-and-the-business-of-keeping-us-hooked/

Alter, A. 2017. *Irresistible: The Rise of Addictive Technology and the Business of Keeping Us Hooked*. Penguin Press. New York, NY.

Apple. 2017. Environmental Responsibility Report. 2017 Progress Report, Covering Fiscal Year 2016. Retrieved December 13, 2017. https://images.apple.com/environment/pdf/Apple_Environmental_Responsibility_Report_2017.pdf.

Asensio, O.I., M.A. Delmas, M.A. 2016. The dynamics of behavior change: Evidence from energy conservation. *Journal of Economic Behavior & Organization* **126** 196-212.

Bachrach, D.G., E. Bendoly, D. Beu Ammeter, R. Blackburn, K.G. Brown, G. Burke, T. Callahan, K.Y. Chen, V.H. Day, A.E. Ellstrand, O.H. Erekson. 2017. On Academic Rankings, Unacceptable Methods, and the Social Obligations of Business Schools. *Decision Sciences*. Forthcoming.

Blanco, C., F. Caro, C.J. Corbett. 2016. The state of supply chain carbon footprinting: analysis of CDP disclosures by US firms. *Journal of Cleaner Production* **135** 1189-1197.

Calders, T., S. Verwer, S. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* **21**(2) 277-292.

Cerdas, F., M. Juraschek, S. Thiede, C. Herrmann. 2017. Life cycle assessment of 3D printed products in a distributed manufacturing system. *Journal of Industrial Ecology* **21**(S1) S80-S93.

Cohen, M.C. 2017. Big Data and Service Operations. *Production and Operations Management* (this issue). Forthcoming.

Corbett, C.J., L.N. Van Wassenhove, L.N. The green fee: internalizing and operationalizing environmental issues. *California Management Review* **36**(1) 116-135.

Coroama, V.C., L.M. Hilty, E. Heiri, F.M. Horn. 2013. The direct energy demand of internet data flows. *Journal of Industrial Ecology* **17**(5) 680-688.

de Felice, D. 2015. Business and Human Rights Indicators to Measure the Corporate Responsibility to Respect Challenges and Opportunities. *Human Rights Quarterly* **37** 511-555.

Delmas, M.A., D. Etzion, N. Nairn-Birch. 2013. Triangulating environmental performance: What do corporate social responsibility ratings really capture? *The Academy of Management Perspectives* **27**(3) 255-267.

Devalkar, S.K., S. Seshadri, C. Ghosh, A. Mathias. 2017. Data Science Applications in Indian Agriculture. *Production and Operations Management* (this issue). Forthcoming.

The Economist. 2017. Fuel of the future: Data is giving rise to a new economy. May 6, 2017. Retrieved December 3, 2017, https://www.economist.com/news/briefing/21721634-how-it-shaping-up-data-giving-rise-new-economy.

EPA. 2017. Greenhouse Gases Equivalencies Calculator - Calculations and References. Retrieved November 16, 2017, https://www.epa.gov/energy/greenhouse-gases-equivalencies-calculator-calculations-and-references.

Etzion, D., J.A. Aragon-Correa. 2016. Big data, management, and sustainability: Strategic opportunities ahead. *Organization & Environment* **29**(2) 147-155.

Facebook. 2017. Retrieved December 13, 2017, https://sustainability.fb.com/clean-and-renewable-energy/.

Feng, Q., J.G. Shanthikumar. 2017. How Research in Production and Operations Management May Evolve in the Era of Big Data. *Production and Operations Management* (this issue). Forthcoming.

Financial Times. 2017. Experian warns of increased scrutiny after Equifax hack. November 15, 2017. Retrieved November 15, 2017, https://www.ft.com/content/ec01484c-c9e8-11e7-ab18-7a9fb7d6163e.

Galhotra, S., Y. Brun, A. Meliou. 2017. Fairness testing: testing software for discrimination. *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*. ACM, 498-510.

Gartner. 2007. Gartner Estimates ICT Industry Accounts for 2 Percent of Global CO2 Emissions. April 26, 2007. Retrieved November 16, 2017, https://www.gartner.com/newsroom/id/503867.

Gebru, T., J. Krause, Y. Wang, D. Chen, J. Deng, E.L. Aiden, L. Fei-Fei. 2017. Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. *Proceedings of the National Academy of Sciences*. Forthcoming.

Goes, P.B. 2014. Big Data and IS Research. *MIS Quarterly* **38(**3) iii-viii.

Google. 2011. Google's Green Computing: Efficiency at Scale. Retrieved October 30, 2017, https://static.googleusercontent.com/media/www.google.com/en//green/pdfs/google-green-computing.pdf.

Google. 2016. Enviromental Report. Retrieved December 13, 2017, https://static.googleusercontent.com/media/www.google.com/en//green/pdf/google-2016-environmental-report.pdf.

Guha, S., S. Kumar. 2017. Emergence of Big Data Research in Operations Management, Information Systems, and Healthcare: Past Contributions and Future Roadmap. *Production and Operations Management* (this issue). Forthcoming.

Hák, T., S. Janoušková, B. Moldan. 2016. Sustainable Development Goals: A need for relevant indicators. *Ecological Indicators* **60** 565-573.

Harris, T. 2016. How Technology is Hijacking Your Mind — from a Magician and Google Design Ethicist. Medium, May 18, 2016. Retrieved December 13, 2017, https://journal.thriveglobal.com/how-technology-hijacks-peoples-minds-from-a-magician-and-google-s-design-ethicist-56d62ef5edf3.

Hernandez, P. 2015. Enterprises are Hoarding 'Dark' Data: Veritas. Datamation. Retrieved October 30, 2017, https://www.datamation.com/storage/enterprises-are-hoarding-dark-data-veritas.html.

Huang, R., M.E. Riddle, D. Graziano, S. Das, S. Nimbalkar, J. Cresko, E. Masanet. 2017. Environmental and economic implications of distributed additive manufacturing: The case of injection mold tooling. *Journal of Industrial Ecology* **21**(S1) S130-S143.

IEA. 2017. Digitalization & Energy. Retrieved December 13, 2017, http://www.iea.org/publications/freepublications/publication/DigitalizationandEnergy3.pdf.

Ingold, D., S. Soper. 2016. Amazon doesn't consider the race of its customers. Should it? Bloomberg, April 21, 2016. Retrieved October 27, 2017. http://www.bloomberg.com/graphics/2016-amazon-same-day.

Johnson, H. 2015. Digging up dark data: What puts IBM at the forefront of insight economy. Silicon Angle. Retrieved October 27, 2017, https://siliconangle.com/blog/2015/10/30/ibm-is-at-the-forefront-of-insight-economy-ibminsight/.

Jones, G.V., M.A. White, O.R. Cooper, K. Storchmann. 2005. Climate change and global wine quality. *Climatic change* **73**(3) 319-343.

Kahneman, D. 2011. *Thinking, fast and slow*. Farrar, Straus and Giroux, New York.

Kellens, K., M. Baumers, T.G. Gutowski, W. Flanagan, R. Lifset, J.R. Duflou. 2017. Environmental dimensions of additive manufacturing: Mapping application domains and their environmental implications. *Journal of Industrial Ecology* **21**(S1) S49-S68.

Khuntia, J., T.J.V. Saldanha, S. Mithas, V. Sambamurthy. 2017. Information Technology and Sustainability: Evidence from an Emerging Economy. *Production and Operations Management*. Forthcoming.

Kirsch, D.A. 2000. *The Electric Vehicle and the Burden of History*. Rutgers University Press.

Korsbakken, J.I., G.P. Peters, R.M. Andrew. 2016. Uncertainties around reductions in China [rsquor] s coal use and CO2 emissions. *Nature Climate Change* **6**(7) 687-690.

Lai, F., D. Li, C.T. Hsieh. 2012. Fighting identity theft: The coping perspective. *Decision Support Systems* **52**(2) 353-363.

Laurance, W.F., A. Balmford. 2013. Land use: a global map for road building. *Nature* **495**(7441) 308-309.

Lazer, D., R. Kennedy, G. King, A. Vespignani. 2014. The parable of Google Flu: traps in big data analysis. *Science* **343**(6176) 1203-1205.

Linkov, I., E. Moberg. 2012. *Multi-Criteria Decision Analysis: Environmental Applications and Case Studies.* CRC Press. Boca Raton, FL.

Liu, Z., D. Guan, W. Wei, S.J. Davis, P. Ciais, J. Bai, S. Peng, Q. Zhang, K. Hubacek, G. Marland, R.J. Andres. 2015. Reduced carbon emission estimates from fossil fuel combustion and cement production in China. *Nature* **524**(7565) 335-338.

Lynch, J. 2013. Choose satellites to monitor deforestation. *Nature* **496**(April 18, 2013) 293-294.

Malloy, T.F., V.M. Zaunbrecher, C. Batteate, A. Blake, W.F. Carroll, C.J. Corbett, S.F. Hansen, R. Lempert, I. Linkov, R. McFadden, K.D. Moran. 2016. Advancing alternative analysis: integration of Decision Science. *Environmental Health Perspectives* **125**(6) 066001-1 – 12.

Malmodin, J., Å. Moberg, D. Lundén, G. Finnveden, N. Lövehagen. 2010. Greenhouse gas emissions and operational electricity use in the ICT and entertainment & media sectors. *Journal of Industrial Ecology* **14**(5) 770-790.

Mami, F., J.P. Revéret, S. Fallaha, M. Margni. 2017. Evaluating Eco-Efficiency of 3D Printing in the Aeronautic Industry. *Journal of Industrial Ecology* **21**(S1) S37-S48.

Marr, B. 2017. Forbes, February 7, 2017. IoT And Big Data At Caterpillar: How Predictive Maintenance Saves Millions Of Dollars. Retrieved December 11, 2017, https://www.forbes.com/sites/bernardmarr/2017/02/07/iot-and-big-data-at-caterpillar-how-predictive-maintenance-saves-millions-of-dollars/#203abf737240.

Martin, K.E. 2015. Ethical Issues in the Big Data Industry. *MIS Quarterly Executive* **14**(2) 67-85.

Masanet, E. 2017. Private communication.

Mayers, K., J. Koomey, R. Hall, M. Bauer, C. France, A. Webb. 2015. The carbon footprint of games distribution. *Journal of Industrial Ecology* **19**(3) 402-415.

Melville, N.P., R. Whisnant. 2014. Energy and carbon management systems. *Journal of Industrial Ecology* **18**(6) 920-930.

Melville, N.P., T.J. Saldanha, D.E. Rush. 2017. Systems enabling low-carbon operations: The salience of accuracy. *Journal of Cleaner Production* **166** 1074-1083.

Muthulingam, S., C.J. Corbett, S. Benartzi, B. Oppenheim, B., 2013. Energy efficiency in small and medium-sized manufacturing firms: order effects and the adoption of process improvement recommendations. *Manufacturing & Service Operations Management* **15**(4) 596-615.

National Academy of Sciences. 2014. *Big Data in Materials Research and Development: Summary of a Workshop.* The National Academies Press. Washington, DC.

Natural Resources Defense Council. 2012. The Carbon Emissions of Server Computing For Small- To Medium-Sized Organizations: A Performance Study of On-Premise vs. The Cloud. Retrieved October 27, 2017, https://www.nrdc.org/sites/default/files/NRDC_WSP_Cloud_Computing_White_Paper.pdf.

Richards, N.M., J.H. King. 2013. Three paradoxes of big data. *Stanford Law Review Online* **66**(41) 41-46.

Sachs, J.D., 2012. From millennium development goals to sustainable development goals. *The Lancet* **379**(9832) 2206-2211.

Shehabi, A., S. Smith, D. Sartor, R. Brown, M. Herrlin, J. Koomey, E. Masanet, N. Horner, I. Azevedo, W. Lintner. 2016. United States Data Center Energy Usage Report. Lawrence Berkeley National Laboratory, Berkeley, California. LBNL-1005775

Shin, L. 2014. 'Someone Had Taken Over My Life': An Identity Theft Victim's Story. Forbes. November 18, 2014. Retrieved November 15, 2017, https://www.forbes.com/sites/laurashin/2014/11/18/someone-had-taken-over-my-life-an-identity-theft-victims-story/#3e6b4a3f25be.

Simon, H. A. 1971. Designing Organizations for an Information-Rich World. M. Greenberger, ed. *Computers, Communication, and the Public Interest*. The Johns Hopkins Press. Baltimore, MD.

Sun, F., A. Hall, M. Schwartz, D.B. Walton, N. Berg. 2016. Twenty-First-Century Snowfall and Snowpack Changes over the Southern California Mountains. *Journal of Climate* **29**(1) 91-110.

Swaminathan, J.M. 2017. Big Data Analytics for Rapid, Impactful, Sustained and Efficient (RISE) Humanitarian Operations. *Production and Operations Management* (this issue). Forthcoming.

Thornton, J.A., K.S. Virts, R.H. Holzworth, T.P. Mitchell. 2017. Lightning enhancement over major oceanic shipping lanes. *Geophysical Research Letters* **44**(17) 9102-9111.

Thorp, J. 2012. Big Data Is Not the New Oil. *Harvard Business Review*. November 30, 2012. Retrieved December 3, 2017, https://hbr.org/2012/11/data-humans-and-the-new-oil.

Walachowicz, F., I. Bernsdorf, U. Papenfuss, C. Zeller, A. Graichen, V. Navrotsky, N. Rajvanshi, C. Kiener. 2017. Comparative energy, resource and recycling lifecycle analysis of the industrial repair process of gas turbine burners using conventional machining and additive manufacturing. *Journal of Industrial Ecology* **21**(S1) S203-S215.

Weber, C.L., J.G. Koomey, H.S. Matthews. 2010. The energy and climate change implications of different music delivery methods. *Journal of Industrial Ecology* **14**(5) 754-769.

Wharton. 2014. What Big Data Can Mean for Sustainability. September 12, 2014. Retrieved December 3, 2017, http://knowledge.wharton.upenn.edu/article/what-big-data-means-for-sustainability/.

Woellert, K., P. Ehrenfreund, A.J. Ricco, H. Hertzfeld. 2011. Cubesats: Cost-effective science and technology platforms for emerging and developing nations. *Advances in Space Research* **47**(4) 663-684.

Zwitter, A. 2014. Big data ethics. *Big Data & Society* **1**(2) 1-6.