International Journal of Communication 8 (2014), 1727–1744 1932–8036/20140005

IJoC Big Data, Big Questions

A Dozen Ways to Get Lost in Translation: Inherent Challenges in Large-Scale Data Sets

LAWRENCE BUSCH
Michigan State University, USA

As noted by the late Susan Leigh Star, technoscientific research always involves simplification and standardization. In recent years, the collection and analysis of large-scale data sets (LSDS) have become the norm. These are often convenience samples analyzed by data mining techniques. Moreover, these data are often used as the basis for public and private policy and action. At the same time, the term "large-scale" suggests completeness, while ease of collection and analysis suggest that little else need be done. Both tend to crowd out other interpretations; hence understanding their limits should be of the utmost concern. This article discusses a number of the issues of concern that arise out of the necessary but potentially problematic simplifications/ standardizations found in LSDS.

Lost in Translation: Problems of Large-Scale Data Sets

> Textbooks generally pass over in silence the formation of cases; and philosophers do likewise when they come to admit the "rose" is an immediate datum, almost a natural fact. The truth is that without starting from the formation of cases, there can be no induction: here begins the creation of the uniformity of nature by the human mind, from which are produced the structures of every factual regularity. (Boldrini, 1972, p. 203)

The collection and analysis of large-scale data sets (LSDS) has become the norm over the last several decades. Fields as diverse as marketing, economics, particle physics, sociology, molecular biology, and logistics all employ both large-scale data collection and analysis. The National Science Foundation and the National Institutes of Health have made analysis of LSDS a program focus, as have many other federal agencies (Mervis, 2012). In addition to the obvious generators of LSDS, General Electric has announced that it will use this approach to design smarter aircraft engines (Vance, 2012) as well as to improve its other businesses and catch up with IBM (Catts, 2012). The Society of Human Resource Management is using LSDS to develop a single measure of the quality of a company's workforce, which could be used to compare companies (Green, 2012). Universities are using LSDS collected from students to quantify their behavior (Parry, 2012). Hospitals, now in the process of digitizing their records, are working with the

Lawrence Busch: lbusch@msu.edu
Date submitted: 2013–04–06

major information technology companies to mine their databases and improve care (Robertson, 2012). In fact, one physician has suggested that a huge integrated database be created that would include "all medical knowledge" (Shaywitz, 2012). In short, LSDS are now ubiquitous, making a better understanding of their promise and limitations necessary.

Although it is difficult to define precisely what a large-scale data set is, we may at least approximate a definition is several ways. First, we may argue that LSDS are sufficiently large that they would have been useless before the advent of digital computers.[1] Second, LSDS *require* digital storage; they are in that sense a creature of the massive increase in digital storage space (and concomitant decline in cost) that has occurred over the last 30 years. Finally, LSDS are sufficiently large that they permit (and perhaps even encourage) what has become known as "data mining."[2]

Importantly, many (though not all) of these data are neither populations nor samples in the classic sense understood in elementary statistics texts. They often involve convenience samples—people who bought a certain product, families that are part of a given government program, sensors that detect contaminants in a watershed, particles that pass through a certain target, books that Google has scanned, proteins from a given plant. They are often analyzed not by hypothesis testing, but by data mining techniques, which involve sifting through vast quantities of data to find the proverbial needle in the haystack.[3]

Arguably, the best-documented early case where the potential problems of convenience samples emerged was the 1948 *Reader's Digest* election poll. The poll erroneously predicted that Thomas Dewey would be elected president. However, nearly all the ex post facto critiques focused on sampling problems and the need to avoid convenience samples (e.g., Zetterberg, 2004). At the time, little or no attention was paid either to problems associated with the formation of cases or simplification (despite the fact that the use of convenience samples can be seen as a form of simplification).

In contrast, statistician Marcello Boldrini, quoted in the epigraph above, emphasized that the uniformities that allow us to engage in statistical analysis are *constructed* through the formation of cases; these categories into which we organize things have consequences for our statistical analyses. More recently, the late Susan Leigh Star (1983) noted that simplification is a common, even obligatory, part of

[1] Until digital records became commonplace, the U.S. Department of Agriculture arguably contained the premier statistical laboratory in the United States. As late as 1958, a manual explained how a room full of statistical clerks, each using a desk calculator, could be used to do a multiple regression analysis (Foote, 1958). Few other labs had the financing to engage in such activities.

[2] As Strasser (2012) notes, even as the size and scope of LSDS have increased markedly, the problem of having more information than the means available to analyze it was already present in the natural history museums of several centuries ago.

[3] This approach appears to have its roots in two unrelated fields a bit more than a century ago. On the one hand, particle physicists needed to trace the trajectories of subatomic particles in cloud chambers. On the other hand, actuaries had to calculate the probability of adverse events based on data about particular people or ships.

technoscientific work, and that simplification necessarily requires that certain things be left out of theories, empirical analyses, and published and unpublished papers. As she put it, “If you never sort the chaos of the world, it never becomes sensible” (Star, 1983, p. 205). She noted several ways in which simplification is manifested. Among them (a) ill-structured problems are broken into pieces and worked on as if they were well-structured; (b) replications are never precisely that because work at each site is organized somewhat differently; (c) limited resources require taking shortcuts; (d) outside agencies always demand particular formats, editing, and other rules; (e) there is always too much data so some must be discarded; (f) analysts are always accountable to someone and must summarize their findings; and (g) data are frequently “cleaned” by deleting anomalous cases. Indeed, all technoscientific work—this article included—requires simplification. Problems arise when, even as we simplify so as to make problems tractable, we forget such simplification and treat the results of our analyses as brute facts.

In this article I focus on the now commonplace creation, use, and analysis of large-scale data sets (LSDS). As with all quantitative analyses, LSDS require attention to both the formation of cases and to simplifications. Yet LSDS are particularly problematic because *more and more they are used to make public and private policy decisions* (Davidian & Louis, 2012). For example, military drones use Geographical Positioning System (GPS) data to determine what is (and is not) a target (Singer, 2009). Climate change models use aggregations of weather station measurements over time and space to make predictions about future weather conditions and policy recommendations about remediation (Busch, 2011a). Large-scale financial data sets are commonly used to make predictions of future growth and to guide investment decisions, including those leading to the recent crash. Moreover, the very size of LSDS appears to imply that the realities they display are in some sense(s) universal.

Below, I first examine the two complementary and increasingly intertwined roles played by LSDS—aggregation and drilling down. Then, I examine 12 challenges to the use of LSDS. The first seven focus largely on what goes into LSDS, while the latter five emphasize what emerges from them. I argue that although these challenges include the invention of new mathematical and statistical procedures for data collection and analysis, they also pose far deeper ontological and epistemological problems that require—per Boldrini and Star—constant reflection. I conclude by arguing that far more attention needs to be paid to these issues if we are make the best use of these new forms of data collection and analysis.

Aggregation vs. Drilling Down

In times past, the few LSDS that existed were such that users were either forced to work with aggregates (e.g., census data), or they consisted of massive tomes (e.g., telephone directories) or huge files (e.g., real property or medical case records) from which individual records could be retrieved. However, with few exceptions, those LSDS designed for aggregate analysis were such that drilling down was impossible, while those designed for drilling down were such that aggregation was all but impossible. All were both expensive and time consuming to collect, and even more expensive to analyze.

Today, many (perhaps most) LSDS can be designed/used either to aggregate—thereby determining properties of, that is, making visible, the aggregate phenomenon—or to drill down, locating characteristics of specific cases. That said, some data sets are still designed for aggregation, in that drilling down is deliberately blocked, while others are designed to permit drilling down, with little interest in—and even difficulty in creating—aggregations.

Of equal importance is that many LSDS were and remain designed to aid in decision making. Thus census data could be used to design government programs, to market goods to particular populations, or to identify particular social problems such as high crime rates. Similarly, real property records were and are used to identify owners of particular pieces of real estate and to make decisions about purchase or sale. Let us examine these two approaches more carefully.

Aggregation. Foucault (2007, 2008) has noted how in the 18th and 19th centuries all sorts of collectivities were given statistical reality as a result of various censuses of the population. As Foucault suggested, these statistical realities took on a life of their own. By virtue of their construction through the aggregation of individual events, they were examined, like Boldrini's rose noted in the epigraph, as brute data. Indeed, one of the founders of modern sociology—Emile Durkheim (1997 [1897])—began the slow task of understanding how to use aggregate statistics to demonstrate the presence of (or perhaps to create) otherwise invisible social phenomena such as suicide rates. Similarly, macroeconomic phenomena were given "substance" by the government-sponsored development of LSDS starting in the late 1930s (Mitchell, 2008).[4] In the wake of the Great Depression, government officials wished to better grasp the scope and nature of the problems facing them. The solution was to begin to collect and aggregate data about everything from wheat production to the lack of work. After several decades, it became possible to talk not merely about economic matters, but about "the economy," as if it were a separate entity distinct from politics or society.

More recently, vast quantities of meteorological data have been organized, collected, and analyzed in an effort to determine whether average temperatures are rising and the degree to which climate change is caused by human action. Initially, analysis of these data sets was limited by cost, scale, and lack of statistical tools. But today, any researcher in fields as diverse as sociology, genomics, marketing, literature, or particle physics, can access and analyze LSDS of one sort or another.

Importantly, the revelation/construction of new phenomena by virtue of aggregation/drilling down in LSDS is linked to an increasingly complex division of labor. With little exaggeration, it may be argued that those who determine the categories to be used in collecting data, the procedures for handling missing data, the specific subjects of data collection, the nature of the sampling methods used, and the means by which to construct and aggregate the data are often not those persons who engage in its analysis. Hence, most economists today build their models based on statistical data collected for other purposes by government agencies and private firms. Computational biologists search for patterns in data usually collected by others, often in multiple labs with different protocols; indeed, they may never actually set foot in any laboratory. And, marketing research firms frequently separate data collection from data analytics. As a consequence, the simplifications made in defining the population of interest, in data

[4] John Maynard Keynes (1939), despite or perhaps because of his solid mathematical skills, was particularly skeptical of using LSDS to create economic models.

collection, in sorting data into predefined categories, and in eliminating outliers may be invisible to those engaged in analysis; to them the data are the brute facts waiting to be analyzed.

Drilling down. Conversely, some LSDS are designed for drilling down. Hence, the various Internet search engines are designed to retrieve specific relevant results from a search. Similarly, LSDS created from cyclotrons are designed to reveal the location, speed, or other characteristics of subatomic particles; many genomic data sets are designed to isolate particular gene constructs and functions. And, of course, "personalized" marketing is all about drilling down, finding out as much as possible about a given consumer so as to target him/her with particularly attractive messages (Floridi, 2012).

But drilling down is also a part of the rapidly growing "surveillance society" in which many of us live (Lyon, 2003). Data are constantly collected on CCTV cameras, at toll booths on highways and bridges, at entrances and exits to rapid transit systems, on supermarket loyalty cards, at airport check-in counters, and in medical settings, among others. The recent revelations by Edward Snowden about U.S. National Security Agency data collection reveal that LSDS consisting of telephone records, consumer purchases, and other data are now available to government analysts. These LSDS are generally designed to allow the user to drill down to specific cases—the car that went through the toll plaza without stopping, the patient with a rare disease, the shopper who buys a particular breakfast cereal, the call to a certain person of interest, the purchase of a particular item, and so on. Such data may be focused on identifying "suspicious activities," various biometric characteristics (e.g., fingerprints, DNA, gait, facial features, ethnicity), patterns of behavior of particular persons, or unusual behavior of particular things (e.g., machines, geological formations, air currents).

Although data may be collected for a single purpose, both aggregation and drilling down are (now) possible at least in principle with all LSDS. Hence, one can use Google to determine how many Web pages cite a given term such as "aggregation" (28,600,000) or use electronic transit cards to determine the number of riders who embark at a particular stop. Similarly, one can combine census data with that from supermarket loyalty cards to determine (with more or less accuracy) the income, age, and other characteristics of shoppers. Once can drill down through global climate change data in an attempt to advise Ugandan farmers on how to mitigate the effects of drought or high temperatures attributed (circularly?) to climate change (Marx, Weber, Orlove, Leiserowitz, Krantz, Roncoli, & Phillips, 2007). One can combine data from as many as 100,000 sources to produce statistical profiles of particular individuals for law enforcement and other agencies (Woolner, 2011). In all instances, LSDS may produce knowledge previously inaccessible, but they may also lead us astray either in the way they aggregate or in the way they drill down, or both. They may lead us to erroneously believe that the bits of data collected are somehow like natural objects or to forget that the patterns we find may only appear when we simplify and standardize both the data and the procedures for collecting it.

Challenges for LSDS

Irrespective of their intended use, there are (at least) 12 (perhaps overlapping) challenges common to such data. All involve problems of either case formation, simplification, or both. It is not clear just how much each of these challenges is relevant to any given LSDS. Likely, each data set poses its own challenges and opportunities. Each is worthy of further study both to ensure that truth claims are appropriate and to minimize erroneous policy decisions.

1. *Lossiness.* LSDS may suffer from what engineers call "lossiness."[5] That is, data collection and/or analysis may involve aggregation, case construction, or standardization in such a way that certain aspects of the phenomena of interest are lost. Lossiness may occur by virtue of the simplification required for a large-scale project. For example, recently animal welfare standards were developed by the European Welfare Quality (2009) project. Researchers take 30–50 measures, reduce them to 12 criteria, further reduce them to 4 principles, and finally to an overall numerical assessment given to each farm. While one may use interpolation techniques to estimate the values of that which is lost, there is no way to return to the original phenomena of interest. Such information is irretrievable. Therefore, lossiness makes both checking for accuracy and replication difficult or impossible, as it eliminates the link between the objects of interest and the data collected.

But lossiness can also be connected to insufficient attention to the formation of cases. For example, lossiness poses a problem when replication is central to a given project. Proteomics researchers spent more than a decade collecting data on proteins found in DNA before discovering the lossiness in their protocols. Researchers assumed that they were using the same protocols in collecting those data, and they published them in the journals central to the emerging field. However, since proteins are not ever directly observable but must be inferred from laboratory instruments, each laboratory commonly uses a mass spectrometer to identify the proteins in question. Not surprisingly, spectrometers from different companies use slightly different methods and software to provide the data published in hundreds of papers. When those researchers attempted to replicate experiments performed on different spectrometers, they found replication impossible. Ultimately, they were only able to solve the problem by requiring all submitted papers to translate their data into a form that could be replicated on any properly functioning spectrometer. However, much of the first decade of data collection had to be discarded because of lossiness (Mackenzie, Waterton, Ellis, Frow, McNally, Busch, & Wynne, 2013). As Boldrini might have put it, inadequate attention had been paid to the formation of cases.

2. *Drift.* Directly related to lossiness is drift. Many LSDS are collected over long periods of time during which the phenomena of interest to researchers may change. For example, some years ago molecular biologists referred to "junk DNA" to describe seemingly irrelevant sequences; today those sequences have given rise to a wide range of "omic" sciences. Similarly, the methods and instruments used in data collection may be modified. Hence, in contemporary molecular biology, high throughput devices allow analysis of vast numbers of sequences in just a few days, whereas a half century ago, it took years to analyze relatively few sequences. Does this changing instrumentation reflect "the same" interests of biologists?

[5] The concept of lossiness appears to have originated in the problem of electricity transmission across long distances. More recently, it has been used to describe the problems associated with JPEG compression—namely that once compressed certain data in the picture are irretrievably lost. This is why a JPEG becomes blurred when its size is increased. I thank Ruth McNally for bringing this to my attention.

Furthermore, the training received by those analyzing the data may drift. Not too long ago, laboratory biologists learned to use information technologies in a rather ad hoc manner; today, numerous textbooks and formal programs promote computational biology. Laboratory biologists learned to use information technologies while working in the lab with organisms. At least in principle, it is possible for today's computational biologists to engage in their work with minimal contact with actual organisms.

Similarly, data collected 50 or 100 years ago on weather, watersheds, soils, and markets used different instruments than those used today. Although one may be able to "fit" the older numbers into contemporary data sets, their meanings, reliability, accuracy, and other properties may drift. Finally, researchers in neighboring fields may view the same data in considerably different ways. In sum, decisions about case formation and simplification made years ago may, in light of new knowledge, no longer be adequate for the problems at hand. Yet it is all too easy to collectively forget how these decisions were made as well as their import.

3*. Distancing.* LSDS are in many ways the logical conclusion of Leon Battista Alberti's 15th-century handbook on perspective drawing (itself based on a somewhat earlier invention by Brunelleschi [Edgerton, 2006]) and its later diffusion by Albrecht Dürer (Romanyshyn, 1989). Importantly, Alberti's followers believed that their "instrumental" approach revealed aspects of the world that were invisible to artists using older techniques, a position later taken up by Galileo in his distinction between primary (i.e., measurable, quantitative) and secondary (i.e., experienced) qualities.

LSDS often allow one to gain clarity only by distancing oneself from the phenomena of interest. For Alberti, this meant viewing the object of interest through a device and from a fixed point (later labeled Alberti's Window), as shown in Figure 1. This distancing described by Alberti allowed a form of objectivity based on mathematical relationships, much as does the use of LSDS. At the same time, it simplifies the problem at hand and makes possible the formation of cases.

Furthermore, both Alberti and Galileo claimed that their approach revealed a primary reality by distancing the observer from the observed through instrumentation. Moreover, in at least some instances, it appears that the instrumental observations *replace* the reality from which they were drawn. One amusing example involved the driver of a large truck who followed his GPS unit instructions while ignoring both road signs and people trying to stop him. He wedged his truck between two buildings in a small English village (Bell, 2008). Something similar might be said of the recent Google books analysis (Michel, Shen, Aiden, Veres, Gray, Team, et al., 2011). While the methods used allowed researchers to gaze at more than 5 million books from a fixed perspective provided by their computers, they surely did not examine more than a small fraction of those volumes directly. In addition, most of the volumes are in English, and the boundaries of the collection are more a matter of organizational permissions (those libraries that decided to take part in a digitization project) and of the limitations imposed by copyright law than they are of any scientific or statistical rationale.

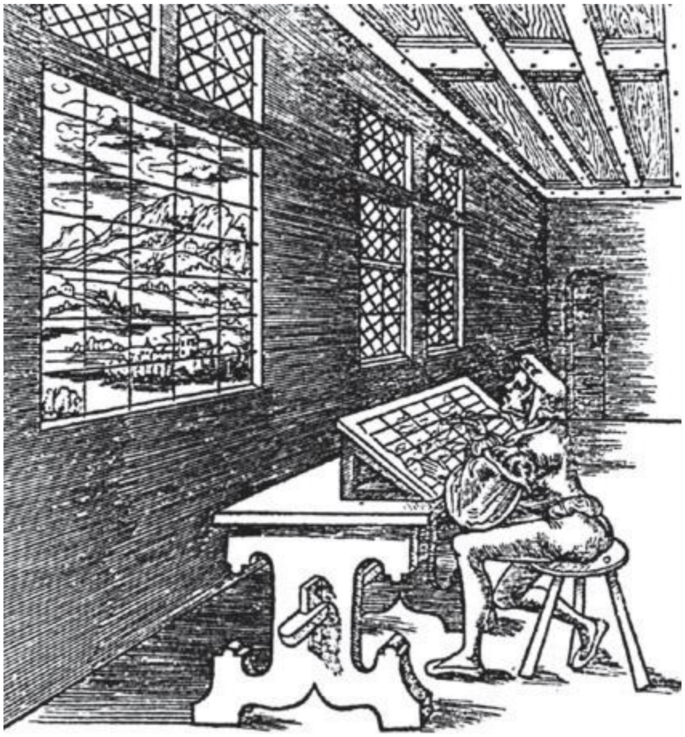

Figure 1. Alberti's Window (Edgerton, 2006, p. 162).

Moreover, their opening lines are revealing: "Reading small collections of carefully chosen works enables scholars to make powerful inferences about trends in human thought. However, this approach rarely enables precise measurement of the *underlying phenomena*" (Michel et al., 2011, p. 176, emphasis added). They claim to have revealed some "underlying phenomena," and indeed perhaps they have, but what might also be obscured by this?[6] Put differently, although there is little doubt that a statistical analysis of the contents of 5 million books provides a *different* understanding of the written word than does a careful reading of just a few, it is unclear as to whether, or in what sense, the larger sample necessarily produces a better answer to the research questions at hand. Indeed, it is not even clear that the research questions remain the same. What is important to recognize here is that there are complex tradeoffs between the size and apparent precision of the data set gained by distancing and the loss of far less mediated knowledge obtained by reading the books in question.

4. *Layering.* Nearly all studies involving LSDS employ a "layering" approach to phenomena of interest (Mol, 2012). But as Annemarie Mol argues, one may also take a "versional" approach to knowledge. Layering involves assuming that there are underlying relations that hold true uniformly (perhaps with some random error term) across time and space, while versions assume that the situatedness of things produces certain kinds of outcomes. For example, the U.S. policy toward the Soviet

[6] This issue is hardly new. Barfield (1965) already noted it nearly half a century ago.

Union during the Cold War took a layering approach, assuming that the Soviet Union's military saw the situation exactly as their American counterparts did. Hence, every time that intelligence reports revealed evidence that countered the layering approach, it was discarded by intelligence agencies. As revealed after the collapse of the Soviet Union, the Soviet view was considerably different, and American estimates were actually misleading.

Moreover, as Mol (2002) argued in her earlier work, the very practices of data collection have not merely epistemological but *ontological* consequences built into them. The examples she provides revolve around arteriosclerosis. For the patient, arteriosclerosis might appear as pain, for the surgeon as a blocked artery, and for the pathologist as a stain on a slide. Considerable work is required to piece together these differing ontologies of arteriosclerosis into a coherent narrative.[7]

The widespread use of LSDS clearly tips the balance in favor of layering. After all, in almost every instance, those who construct LSDS necessarily must remove the situatedness of the item of interest—for example, a particular protein, a survey response, a purchase of a particular consumer product—and reduce it to a set of variables. Recent work by Theodore Porter (2012) illustrates this problem in a somewhat different way. Porter argues that the dominance of facts, data, and statistics involves a "thinning" of the world. Specifically, in our collective efforts to produce greater precision, clarity, and objectivity, those aspects of things that are not amenable to numerical or statistical analysis—that situate particular phenomena—are systematically downgraded or removed from consideration. Such thinning involves the making of universal claims by discarding ambiguities, subtleties, and situatedness. Moreover, Porter quotes the "law" developed by statistician Donald Campbell:

> The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor. (Porter, 2012, p. 225)

In short, such indicators transform social settings as persons and institutions adapt their behavior to them, as their behavior becomes more "layered." For example, elementary and secondary teachers tend to teach to the test when that test is to be used as an indicator of their effectiveness. Since so many LSDS are used in public and private policy making, Campbell's law is perhaps applied more often than is readily apparent.

The work of both Mol and Porter suggest that the necessary decisions about simplification, case formation, and subsequent data analysis lead to the creation of a particular ontology. One might well ask: Is the ontology produced by a given approach to data collection and analysis the one we want to produce? Does it allow for other competing ontologies? Or does it crowd out competing ontologies that might in certain (other) circumstances be valued?

5. *Errors*. One must ask what is done with "errors." Of course, errors only exist within the context of rules of case formation and of simplification. LSDS are filled with all sorts of errors, but they are not

[7] For an example applied to animal welfare, see Busch (2011b).

handled in the same way across (and likely even within) fields of science. As Karen Knorr-Cetina (2003) noted in her comparison of high energy physics and molecular biology, high energy physicists spend considerable time and effort analyzing errors (attempting to sort out random from systematic error) in the data they generate, while molecular biologists consign such errors to the trash heap. Most social scientists and marketers as well spend little time analyzing errors, although a few engaged in survey analysis do raise those questions.

In the case of random samples, even a 95% confidence interval means that 5% of the reported results will be erroneous; in the case of convenience samples where the confidence interval is inherently unknowable, the error rate may be even higher. One recent editorial by a genomicist notes that "hunting for biological surprises without due caution can easily yield a rich crop of biases and experimental artefacts, and lead to high-impact papers built on nothing more than systematic experimental 'noise'" (MacArthur, 2012, p. 427). He goes on to note that high-throughput technologies often contain "systematic biases" that may conceal errors even from experienced researchers (see also Kaiser, 2012). And, surely, the drive for the prestige of being first to reveal an exciting finding encourages erroneous findings to be portrayed as fact.

Moreover, most public and private officials lack the competencies and resources needed to recollect or reanalyze LSDS. They may therefore be unaware of the errors they conceal. Furthermore, since their authority may be based on LSDS, they may strongly resist questioning the validity of such analyses (Porter, 1995). Moreover, the apparent objectivity of large numbers may suggest that no further hermeneutic work is needed.

6. *Standards.* Likely overlapping all of these issues is the role of standards. LSDS require herculean efforts of standardization—in data collection, analysis, and interpretation. As the epigraph to this article suggests, without standards, themselves based on conventions and categories that often appear self-evident, no analyses are possible, since all phenomena must be considered unique (Strasser, 2012). The process of creating uniformity through standardization, even as it is necessary, may obfuscate phenomena of considerable importance (Busch, 2011c). For example, Hulme, Dessai, Lorenzoni, and Nelson (2009) show that choosing different periods for baseline "normals" for temperature produces considerably different historical trends. It is by no means obvious which of these statistical groupings is likely to lead to the best climate policies.

Moreover, although standards are often considered neutral, purely technical objects, whose standards prevail is often a matter of prestige, power, and profit. The Chinese government fully understands this and has been trying to make its standards the international standards in a variety of fields. Most recently, it has invested heavily in the Internet of Things, a field that is likely to generate many LSDS. The government hopes to make Chinese standards *the* global standards in this domain (Voigt, 2012).

7. *Disproportionality.* A common property of LSDS is disproportionality. As Nowak, Bowen, and Cabot (2006) suggest, extreme outliers can have a very significant impact on whatever variable is of interest:

> Disproportionality has been examined in three broad realms of social science. First, research on environmental justice has investigated how negative environmental conditions (e.g., air quality, land use, water quality) disproportionally impact certain groups, classes, or minorities. . . .[8] Second, the concept of disproportionality is found in social science research that investigates patterns of social sanctions and interaction. This is an extensive area of research that examines how certain minority groups may be disproportionately subjected to state sanctions. . . . The third general area of disproportionality research in the social sciences examines the difference in the environmental impact associated with different forms of social groups. (pp. 155–156)

As part of the necessary simplifications made in the process of doing research, it is not uncommon for such outliers to be dropped as simply "errors." For example, a government-sponsored survey of residents of a given city or region might ask about the quality of social services and find that 95% of residents are "very satisfied" while 3% are "very dissatisfied." As a result, they might well conclude that the outliers (the 3%) are not sufficiently important to warrant further consideration. However, it might well be that those who are "very dissatisfied" are mostly of the same class, race, or ethnicity. Or perhaps most live in the same neighborhood. Or perhaps their concerns all relate to the same service.

Moreover, as Nowak et al. note, biophysical variables often suffer from the same problem. This is particularly the case for indicators of environmental degradation; a seemingly random event may be the tip of the proverbial iceberg. Doubtless, the same type of problems can be found in other fields of research. Yet the extreme outliers may be either discarded (as they are seen as merely errors or exceptions), or passed over if not found to be correlated with other variables currently of interest. Of course, one might argue that a "good" researcher would have pursued such questions. In all research, it is possible to probe deeper into the phenomenon of interest. However, there are always constraints of time, expertise, and budget limitations that *require* that certain questions remain unaddressed.

8. *Amplification/Reduction.* Both Don Ihde (1979) and Bruno Latour (1999) note the dilemma posed by the re-presentation of phenomena by scientific (and statistical) processes. On the one hand, those aspects of phenomena that can be calculated and standardized are amplified, while on the other hand those aspects of the phenomena that resist calculation and standardization are reduced. What is lost in this re-presentation? Is it of consequence? Note that these questions relate back to Mol's and Porter's. At the very least, re-presentation of phenomena such that they can be calculated and standardized recasts phenomena as calculable, standardized objects that are de-situated.[9]

This is of particular concern where LSDS are the result of audits of various sorts. Such audits almost always focus on those aspects of phenomena that are easily measurable (in part to keep audit

[8] Similarly, using EPA data, the late Bill Freudenberg (2006) showed how a small number of manufacturing plants produce the vast majority of the pollution.

[9] This is already a major issue in medical testing. Test results have in some fields replaced interaction with patients, but it is unclear whether this results in better care.

costs under control), but in so doing may erroneously conflate the audit results with the phenomenon of interest (Power, 1997). Several examples will help to clarify this point. In Britain, higher education audits collect vast quantities of data on so-called key performance indicators (KPIs). The assumption is that KPIs are at least adequate measures of scholarship across all disciplines.[10] Similarly, health care facilities may be ranked based on some set of mortality and morbidity rates, ignoring or glossing over the differences in the populations served by those institutions. Indeed, as philosopher Sandra Harding (2006) has suggested, there may well be persons or groups whose interests are served by the systematic ignorance produced in this way. This suggests that processes of simplification and case formation are themselves not neutral processes, but often serve (consciously or unconsciously) to promote various values.

9. *Narratives.* We must ask what kind of story is told by LSDS. While some might disagree, data hardly speak for themselves. Instead, as Star (1999) noted some years ago, "Many information systems employ what literary theorists would call a master narrative, or a single voice that does not problematize diversity. This voice speaks unconsciously from the presumed center of things" (p. 384). Yet even the most apparently obvious results require (1) a degree of interpretation (in the formation of cases, in data collection, and in analysis), and (2) the weaving of a master narrative around the data. Yet, despite this, the sheer size of LSDS tends to conceal other (perhaps equally or even more important) stories. Put differently, LSDS do not eliminate the problems usually framed as hermeneutics, but rather displace them.

An example is the recent tendency of investment banks to replace traders with computer algorithms—mathematical narratives. After all, the algorithms cost less than the annual salary of a trader and, it is argued, will work both more accurately and more consistently than a trader (Childs, 2012). But the algorithms do not do away with interpretation. Instead, they transfer the traders' *visible* hermeneutic work to those who construct and use the algorithms. There that work is often concealed. Perhaps more ominously, Poon's (2009) work suggests it was the shift of hermeneutic work that ultimately led to the collapse of the housing market. Mortgage companies shifted away from the careful weighing of the ability of particular borrowers to repay bankers, relying instead on credit bureau scores. An algorithm was used to impose a particular and apparently consistent interpretation of the world upon the data collected, which was then defined as the "correct" interpretation. But in so doing, the range of stories that could be told about mortgages was limited to those described in the algorithm.

Moreover, many analyses of LSDS use proprietary algorithms. In such instances, the formulation of the algorithm is not known to the user and is protected with intellectual property rights. This is the case, for example, with many ranking scores; the U.S. News and World Report (2013) rankings of everything from universities to law firms are a case in point. Hence, only the "official" interpretation is made available to others. If a faulty algorithm is used to analyze product data, the problem will likely become evident as the product is used. However, when the faulty algorithm is used to analyze complex public or private policies, its flaws may never be revealed.

[10] According to Head (2011) this idea was borrowed from Harvard management specialists.

Are displacements of this sort all of a kind? Or do they differ across types of data sets? And, are there certain aspects of phenomena that are systematically included or always left out of these mathematically or statistically defined narratives?[11] We simply don't know.

10. *Risk.* One might also ask how LSDS change the ways in which risk is depicted and acted upon. We have already seen a shift from risk as a voluntary undertaking (I engage in risky behavior) to risk as statistical probability of harm (mortality rate from driving) (Thompson & Dean, 1996). This is central to Beck's (1992) description of the "risk society." However, the use of LSDS, and (especially?) the graphical depiction of risk (or lack thereof) using LSDS have now become commonplace as well. In such instances, measurements (cases) are aggregated to produce (simplified) graphics. For example, with real-time weather maps one can "see" an approaching storm. Military personnel can "watch" events on the ground using remote screens that aggregate massive amounts of data from many sources (Shanker & Richtel, 2011). Developers can show digital drawings of neighborhoods with new buildings added in.

Of particular note is that such risk depictions and subsequent actions are always designed by persons or groups that have particular objectives in mind. In some instances, this might be the minimization of risk and in others it is maximization. For example, without doubt, soldiers monitoring events from a computer screen far removed from actual warfare will perceive risk differently from those actually on the battlefield. Perhaps more importantly, unless risks for specific subpopulations or individuals are taken into account, LSDS may push risks toward those least able to object. In short, there is now a wide range of new hermeneutical devices that build on the ubiquity of LSDS. How do these change perceptions of risk and risk-averse actions that people take?

11. *Experimentality.* As with all research, LSDS pose the problem of experimentality. Experiments rarely end in the laboratory; they require widespread use for us to become confident of their results. We are all familiar with recall notices for automobiles and pharmaceutical products. Similarly, investigations of air crashes often reveal problems that were unknown when the aircraft was initially tested (Downer, 2011). With the exception of a few cases of outright fraud, these are the result of problems that only appeared long after products were in use. Usually, the trials suggested that the product was robust; only after innumerable uses, perhaps under conditions not envisioned by developers, did the problem emerge. Statistically speaking, one might say that the initial trial sample was insufficiently large to reveal critical but rare problems many standard deviations from the mean.

Moreover, like physical products, one can only be confident about the use of software for the analysis of LSDS when it is in widespread use. As noted some years ago,

> As a rule software systems do not work well until they have been used, and have failed repeatedly, in real applications. Generally, many uses and many failures are required before a product is considered reliable. Software products, including those that have become relatively reliable, behave like other products of evolution-like processes; they

[11] Admittedly, all data sets are used to tell stories. But the hermeneutics of LSDS is rarely explored.

> often fail, even years after they were built, when the operating conditions change. (Parnas, van Schouwen, & Kwan, 1990, p. 636)

In short, no matter how thorough the research may be, the tacit and explicit simplifications about how technologies will be used and how policies will be carried out cannot possibly take all situations into account. LSDS are no exception to this rule.

12. *Ethics.* Last but not least, LSDS raise considerable ethical concerns. Indeed, one could easily devote an entire article or book solely to these issues. Ethical concerns include (a) balancing privacy with access in light of the vast and growing scope of LSDS (Pedreschi, Calders, Custers, Domingo-Ferrer, Finocchiaro, Giannotti et al., 2011); (b) avoiding using LSDS to discriminate against certain persons (e.g., those with genetic predispositions to certain diseases), groups (e.g., those least likely to have the means to complain about adverse effects of large-scale infrastructure projects), or even nations (e.g., by developing standards that make certain nations into standards takers, even if those standards serve them poorly); (c) use of LSDS to enhance social surveillance and social sorting (Lyon, 2003); (d) determining who owns what data that is located in LSDS (e.g., consider the recent flap over Instagram's attempt to use photos posted by users without getting their permission or providing compensation [Wortham, 2012]), and (e) undermining of democratic governance (e.g., use of gerrymandering based on LSDS so as to make the outcomes of elections nearly certain before elections are held). These ethical concerns raise numerous questions about LSDS including: How is privacy to be defined and protected? How can one know when discrimination is present? How can we be confident of results of analyses of LSDS when the algorithms used are trade secrets? What kinds of disclosures are necessary to ensure that LSDS are not used in an unethical manner? How can we protect ourselves against unnecessary surveillance and social sorting? Although ameliorating these ethical issues will require a great deal of debate and deliberation, some initial responses have been suggested (e.g., Pedreschi et al., 2011; von Schomberg, 2011).

Conclusions

This list is not meant to be all-inclusive. Doubtless, there are other issues in need of examination surrounding LSDS in various aspects of science and society, as well as models and algorithms based on them. As Susan Leigh Star might have noted, the consequences of the processes of simplification common across all technoscientific fields are hardly simple.

To date, most of these issues have not been adequately explored. Yet, for better and for worse, LSDS are with us to stay. We need to better understand how they are shaping and reshaping our world, such that they do not dissolve into something resembling the collages of Dadaism, thereby emphasizing the meaninglessness of daily life. As Davidian and Louis (2012) put it, "The future demands that scientists, policy-makers, and the public be able to interpret increasingly complex information and recognize both the benefits and pitfalls of statistical analysis" (p. 12). Doing so will be a considerable challenge.

References

Barfield, O. (1965). *Saving the appearances*. New York, NY: Harcourt, Brace, and World.

Beck, U. (1992). *Risk society: Towards a new modernity* (M. Ritter, Trans.). London, UK: SAGE Publications.

Bell, A. (2008, October 30). I'm stuck in a jam. *Manchester Evening News,* p. 6.

Boldrini, M. (1972). *Scientific truth and statistical method*. New York, NY: Hafner.

Busch, L. (2011a). Climate Change: How debates over standards shape the biophysical, social, political, and economic climate. *International Journal of Sociology of Agriculture and Food, 18*,167–180.

Busch, L. (2011b). How animal welfare standards create and justify realities. *Animal Welfare 20*, 21–27.

Busch, L. (2011c). *Standards: Recipes for reality*. Cambridge, MA: MIT Press.

Catts, T. (2012, April 30–May 6). GE heads west with $1 billion to spend. *BusinessWeek, 4277,* 38–39.

Childs, M. (2012). Computers elbow traders aside*. BusinessWeek, 4305*, p. 48.

Davidian, M., & Louis, T. A. (2012). Why statistics? *Science, 336*, 12.

Downer, J. (2011). "737-Cabriolet": The limits of knowledge and the sociology of inevitable failure. *American Journal of Sociology, 117*(3), 725–762.

Durkheim, E. (1997 [1897]). *Suicide: A study in sociology*. New York, NY: Free Press.

Edgerton, S. Y. (2006). Brunelleschi's mirror, Alberti's window, and Galileo's "perspective tube." *História, Ciências, Saúde-Manguinhos, 13*, 151–179.

Floridi, L. (2012). Big data and their epistemological challenge. *Philosophy and Technology, 25*(4), 435–437.

Foote, R. J. (1958). *Analytical tools for studying demand and price structures* (Agriculture Handbook No. 146).Washington, DC: USDA.

Foucault, M. (2007). *Security, territory, population: Lectures at the Collège de France, 1977–1978*. New York, NY: Palgrave Macmillan.

Foucault, M. (2008). *The birth of biopolitics: Lectures at the Collège de France, 1978–79*. New York, NY: Palgrave Macmillan.

Freudenburg, W. R. (2006). Environmental degradation, disproportionality, and the double diversion: Reaching out, reaching ahead, and reaching beyond. *Rural Sociology, 71*(1), 3–32.

Green, P. S. (2012, July 23–29). HR group creates workforce metrics. *BusinessWeek,* 44–46.

Harding, S. (2006). Two influential theories of ignorance and philosophy's interests in ignoring them. *Hypatia, 21*(3), 20–36.

Head, S. (2011). The grim threat to British universities. *New York Review of Books, 58*(1).

Hulme, M., Dessai, S., Lorenzoni, I., & Nelson, D. R. (2009). Unstable climates: Exploring the statistical and social constructions of 'normal' climate. *Geoforum, 40*(2), 197–206.

Ihde, D. (1979). *Technics and Praxis*. Dordrecht, Netherlands: D. Reidel.

Kaiser, J. (2012). Biomarker tests need closer scrutiny, IOM concludes. *Science, 335*, 1554.

Keynes, J. M. (1939). Professor Tinbergen's method, [Review of] *A Method and its application to investment activity* by J. Tinbergen. Geneva, Switzerland: League of Nations, 1939. *Economic Journal, 49*, 558–568.

Knorr Cetina, K. (2003). *Epistemic cultures: How the sciences make knowledge*. Cambridge, MA: Harvard University Press.

Latour, B. (1999). *Pandora's hope: Essays on the reality of science studies*. Cambridge, MA: Harvard University Press.

Lyon, D. (Ed.). (2003). *Surveillance as social sorting: Privacy, risk, and digital discrimination*. London,UK: Routledge.

MacArthur, D. (2012). Methods: Face up to false positives. *Nature, 487*(7408), 427–428.

Mackenzie, A., Waterton, C., Ellis, R., Frow, E. K., McNally, R., Busch, L., & Wynne, B. (2013). Classifying, constructing, and identifying life: Standards as transformations of "the biological." *Science, Technology & Human Values*, *38*(5), 701–722.

Marx, S. M., Weber, E. U., Orlove, B. S., Leiserowitz, A., Krantz, D. H., Roncoli, C., & Phillips, J. (2007). Communication and mental processes: Experiential and analytic processing of uncertain climate information. *Global Environmental Change, 17*(1), 47–58.

Mervis, J. (2012). Agencies rally to tackle big data. *Science, 336*, 22.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Team, T. G. B. et al. (2011). Quantitative analysis of culture using millions of digitized Books. *Science, 331*, 176–182.

Mitchell, T. (2008). Rethinking economy. *Geoforum, 39*, 1116–1121.

Mol, A. (2002). *The body multiple*. Durham, NC: Duke University Press.

Mol, A. (2012). Layers or versions? Human bodies and the love of bitterness. In B. Turner (Ed.), *Routledge Handbook of Body Studies* (pp. 119–129). New York, NY: Routledge.

Nowak, P., Bowen, S., & Cabot, P. E. (2006). Disproportionality as a framework for linking social and biophysical systems. *Society & Natural Resources, 19*(2), 153–173.

Parnas, D. L., van Schouwen, A. J., & Kwan, S. P. (1990). Evaluation of safety-critical software. *Communications of the ACM, 33*(6), 636–648.

Parry, M. (2012, August 3). Degrees, designed by the numbers. *The Chronicle of Higher Education, 58,* A1, A4–A8.

Pedreschi, D., Calders, T., Custers, B., Domingo-Ferrer, J., Finocchiaro, G., Giannotti, F. et al. (2011). Big data mining, fairness and privacy. *The Privacy Observatory Magazine*. Retrieved from http://privacyobservatory.org/current/40-big-data-mining-fairness-and-privacy

Poon, M. (2009). From new deal institutions to capital markets: Commercial consumer risk scores and the making of subprime mortgage finance. *Accounting, Organizations and Society, 34*(5), 654–674. doi:http://dx.doi.org/10.1016/j.aos.2009.02.003

Porter, T. M. (1995). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton, NJ: Princeton University Press.

Porter, T. M. (2012). Thin description: Surface and depth in science and science studies. *Osiris, 27*(1), 209–226.

Power, M. (1997). *The audit society: Rituals of verification*. Oxford, UK: Oxford University Press.

Robertson, J. (2012). Crunch two data sets, call me in the morning. *BusinessWeek, 4280*, 40–41.

Romanyshyn, R. D. (1989). *Technology as symptom and dream*. New York, NY: Routledge.

Shanker, T., & Richtel, M. (2011, January 17). In new military, data overload can be deadly. *The New York Times,* p. A1.

Shaywitz, D. A. (2012). A database of all medical knowledge: Why not? *The Atlantic*. Retrieved from http://www.theatlantic.com/health/archive/2012/07/a-database-of-all-medical-knowledge-why-not/260313

Singer, P. W. (2009). Military robots and the laws of war. *The New Atlantis: A Journal of Technology and Society, 26*, 27–47.

Star, S. L. (1983). Simplification in scientific work: An example from neuroscience research. *Social Studies of Science, 13*(2), 205–228.

Star, S. L. (1999). The ethnography of infrastructure. *The American Behavioral Scientist, 43*(3), 377–391.

Strasser, B. J. (2012). Data-driven sciences: From wonder cabinets to electronic databases. *Studies in History and Philosophy of Biological and Biomedical Sciences, 43*, 85–87.

Thompson, P. B., & Dean, W. (1996, Fall). Competing conceptions of risk. *Risk: Health, Safety and Environment, 7*, 361–384.

U.S. News and World Report. (2013). Rankings. Retrieved from http://www.usnews.com/rankings

Vance, A. (2012, December 10–16). GE tries to make its machines cool and connected *BusinessWeek, 4308,* 44–46.

Voigt, K. (2012, December 3). China looks to lead the Internet of things. Retrieved from http://edition.cnn.com/2012/11/28/business/china-internet-of-things/index.html?hpt=hp_c2

von Schomberg, R. (Ed.). (2011). *Towards responsible research and innovation in the information and communication technologies and security technologies fields*. Brussels: European Union.

Welfare Quality. (2009). *Welfare quality assessment standards for cattle*. Lelystad, Netherlands: Welfare Quality Consortium.

Woolner, A. (2011). You can run, but it’s hard to hide from TLO. *BusinessWeek, 4246*, 44–45.

Wortham, J. (2012, December 19). Facebook addresses ire over Instagram changes. *The New York Times,* p. B1.

Zetterberg, H. L. (2004). *US election 1948: The first great controversy about polls, media, and social Science*. Paper presented at the WAPOR regional conference on Elections, News Media and Public Opinion, Pamplona, Spain. Retrieved from http://zetterberg.org/Lectures/l041115.htm