The Evolution of Data Quality: Understanding the Transdisciplinary Origins of Data Quality Concepts and Approaches

Sallie Keller, Gizem Korkmaz, Mark Orr, Aaron Schroeder, and Stephanie Shipp

Social and Decision Analytics Laboratory, Biocomplexity Institute of Virginia Tech, Arlington, Virginia 22203; email: sallie41@vbi.vt.edu, gkorkmaz@vbi.vt.edu, morr9@vbi.vt.edu, aschroed@vbi.vt.edu, steph19@vbi.vt.edu

Annu. Rev. Stat. Appl. 2017. 4:85–108

First published online as a Review in Advance on January 6, 2017

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

This article's doi: 10.1146/annurev-statistics-060116-054114

Keywords

designed data, administrative data, opportunity data, reproducibility, total survey error, decision theoretic framework

Abstract

Data, and hence data quality, transcend all boundaries of science, commerce, engineering, medicine, public health, and policy. Data quality has historically been addressed by controlling the measurement processes, controlling the data collection processes, and through data ownership. For many data sources being leveraged into data science, this approach to data quality may be challenged. To understand that challenge, a historical and disciplinary perspective on data quality, highlighting the evolution and convergence of data concepts and applications, is presented.

1. INTRODUCTION

Data quality has historically been addressed by controlling the measurement and data collection processes and through data ownership. For many data sources being leveraged into data science, this approach to data quality may be challenged. Today, the data revolution is experiencing massive data acquisition and repurposing to support analyses of all kinds and across all of science, engineering, and business. This warrants a renewed look at data quality assessment.

This article presents a historical and disciplinary perspective on data quality, highlighting the evolution and convergence of data concepts and applications. The article begins with a brief consideration of the fundamental types of data that underlie today's data applications. Three types of data dominate current data applications:

- ***Designed data:*** data that have traditionally been used in scientific discovery. Designed data include statistically designed data collections, such as surveys or experiments, and intentional observational collections. Examples of intentional observational collections include data obtained from specially designed instruments such as telescopes, DNA sequencers, or sensors on an ocean buoy, and also data from systematically designed case studies such as health registries. Researchers have frequently devoted decades of systematic research to understanding and characterizing the properties of designed data collections.

- ***Administrative data:*** data collected for the administration of an organization, program, or service process. Examples of administrative data include Internal Revenue Service data for individuals and businesses, Social Security earnings records, Medicare and Medicaid health utilization data, 911 and Emergency Management Services, property tax data from local governments, credit data, banking and other financial data such as insurance coverage, production processes, and taxi trip data. When removed from their administrative function, the statistical properties of these data become problematic because they come with little to no documentation about coverage, representativeness, bias, and longitudinal gaps. In some cases, these statistical properties may be knowable, but simply have not been well-studied (NRC 2013a).

- ***Opportunity data:*** data generated on an ongoing basis as society moves through its daily paces. Opportunity data derive from a variety of sources such as GPS systems and embedded sensors, social media exchanges, mobile and wearable devices, and Internet entries. Captured through a variety of methods including direct flows, Internet searches, web crawling, and scraping, these data may exist in a variety of electronic and physical modalities. Though technological advances allow users to collect volumes of data opportunistically, these collections are likely to be the least understood and studied.

The three data types are highlighted throughout this article. They occur in different proportions in different disciplines, and each discipline takes a slightly different focus in managing the quality of its data. As a result, each discipline has contributed in different ways to the development and adoption of definition, methods, and approaches relevant to data quality. These are briefly summarized in **Table 1** and provide a guide for this review of the literature.

The remainder of this article describes data quality from the perspective of physical and biological sciences; engineering, computer science, and business; medicine and public health; social and behavioral sciences; and statistical sciences (Sections 2 through 6). Additional attention is given to the emerging role of opportunity data (Section 7). Section 8 presents conclusions and a forward vision.

Table 1 Contributions to evolution of data quality by discipline

Discipline	Contributions to data quality
Physical and biological sciences	■ Experimental methods ■ Data repositories/portals ■ Reproducibility and replication
Engineering, information technology, business	■ Pareto Principle (80% of a problem is triggered by 20% of the sources) ■ Fitness-for-use ■ Total data quality management ■ Data management ■ Standards
Medicine and public health	■ Clinical data standardization ■ Validity of self-reporting ■ Parsimony and respondent burden ■ Registries
Social and behavioral sciences	■ Total survey error (sampling and nonsampling error) ■ Randomized control trials, observational studies, and natural experiments
Statistics and official statistics	■ Decision theoretic approach ■ Statistical methods ■ Privacy and confidentiality ■ Quality improvement projects from staff within organization

2. THEMES FROM PHYSICAL AND BIOLOGICAL SCIENCES

2.1. Experimental Methods

The need to address data quality is a persistent one in the physical and biological sciences, where scientists often seek to understand subtle effects that leave minute traces in large volumes of data. The variety of designed data collections covers a broad range of disciplines from agronomy to zoology. Perhaps the most recognized starting point for data quality across the sciences using designed data collection was in the 1920s, with R.A. Fisher's revolutionary research in experimental design. Fisher's work introduced the theory, methods, and practice of randomization and replication (Fisher 1925). These concepts are critical for estimating error, understanding the bias and precision of data, and assessing the quality of data in general (Williams et al. 2006).

For most scientists, three factors motivate their work on data quality: first, the need to create a strong foundation of data from which to draw their own conclusions; second, the need to protect their data and conclusions from the criticisms of others; and third, the need to understand the potential flaws in data collected by others. The work of these scientists in data quality primarily concentrates on the design and execution of experiments, including in laboratory, field, and clinical settings. The key ingredients are measurement implementation, laboratory and experimental controls, documentation, analysis, and curation of data.

Data quality in the sciences centers on testing and refining theory, starting with prioritizing the parameters or variables to be used in the analyses. Implicit in this work and a tenet of data quality is that data quality cannot be thought of independently from the data user (Chapman 2005). Training in scientific methods provides practical solutions for theory development, such as methodologies, benchmarks for the most effective techniques, case studies, examples, and the importance of statistics (Chapman 2005, Milham 2012, Becker 2001, Keller et al. 2008).

2.2. Creation of Data Repositories

Scientists working with designed data made a major contribution to the data quality field with the creation of data repositories and portals that allow researchers access to important scientific data. Sharing data through repositories enhances both the quality and the value of the data through standardized processes for curation, analysis, and quality control (Contreras & Reichman 2015). By allowing broad access to data, these repositories encourage and support the use of previously collected data to test and extend previous results. Data repositories are quite common in science fields such as astronomy, genomics, and earth sciences (examples given below).

These repositories have accelerated discovery by expanding the reach of these data to scientists who are not involved in the initial data collection and experiments. Repositories address challenges that affect data quality through governance, interoperability across systems, and costs. Yet barriers remain, and these barriers vary by field. For example, sharing of data may not be possible due to competitiveness across organizations, the inability to share in-house software to analyze the data, the need to create data standards, or the fact that access to the data may violate privacy (Milham 2012).

One example of a successful data repository is the Sloan Digital Sky Survey (SDSS), a large-scale astronomy survey that has been in progress since 2000 (NRC 2014). SDSS was created in 1998 to use state-of-the-art instruments and software to conduct astronomical surveys at a level of detail not possible until then (**http://www.sdss.org/**). SDSS started with a commitment to create high-quality public datasets that would allow for collaborations around the world for all scientists, not just those with access to astronomy equipment. SDSS creates three-dimensional maps of the universe with multicolor images of 1/3 of the sky and spectra for more than 3 million astronomical objects. The data are calibrated, checked for quality, and made available on an annual basis to researchers through online databases. The availability of SDSS data has supported a vast range of scientific investigations by astronomers and other researchers around the world, demonstrating the role for repositories that ensure high data quality leading to reproducible experiments and findings. "Half of these achievements were among the original 'design goals' of the SDSS, but the other half were either entirely unanticipated or not expected to be nearly as exciting or powerful as they turned out to be" (Sloan Digital Sky Survey 2008, p. 7).

Another example is the sharing of cDNA microarray data through research consortia, which has led to a common set of standards and relatively homogeneous data classes (Becker 2001). There are many issues with the sharing of these data, which requires the transformation of biologic to numeric data. These issues may include loss of context, such as laboratory practices followed, and therefore lack of information about the quality of the data when they are transformed. To avoid this loss of information, the consortium ensures that documentation is comprehensive so that other researchers can assess the quality of the data and make comparisons with other studies using the same data (Becker 2001). The documentation also includes information on when incorrect assignments of sequence identity are made so that errors are not perpetuated in other studies.

A relatively new example of a data portal is the NSF-funded National Ecological Observatory Network (NEON). The NEON virtual laboratory components are connected using cyber-technology networks to create an integrated platform for regional- to continental-scale ecological research. Scientists and engineers use NEON to conduct real-time ecological studies across all levels of biology and all temporal and geographical scales. NEON's infrastructure enables hypothesis-driven basic biological and ecological research by providing raw and transformed data in close to real time (Keller et al. 2008). NEON is in the early stages of releasing data products, with the caveat that only limited quality control procedures, such as range checking and spike identification, have been implemented. Subsequent releases of provisional, and then science-grade, data

products are planned as part of their Science Design Plans for improving data quality. The plans include using and reporting statistical and scientific measures (e.g., uncertainty, quality assurance/quality control procedures, and sensor validations) and engineering measures (e.g., system verification and sensor calibrations; see **http://www.neonscience.org/**).

2.3. Reproducibility

Reproducibility is another facet of data quality that is particularly important in all of science and engineering, including the disciplines in the remaining sections. Many journals provide mechanisms to make reproducibility possible, including *PLoS*, *Nature*, and *Science* (McNutt 2014). This entails ensuring access to the computer code and datasets used to produce the results of a study. In contrast, replication of scientific findings involves research conducted by independent researchers using their own methods, data, and equipment that validate or confirm the findings of earlier studies. Replication is not always possible, however, so reproducibility is a minimum and necessary standard for confirming scientific findings (Peng 2009).

Reproducibility goes well beyond validating statistical results and includes empirical, computational, ethical, and statistical analyses (Madigan & Wasserstein 2014, Stodden 2015). For example, empirical reproducibility emphasizes documenting experiments in sufficient detail, computational reproducibility ensures that the same results are obtained from the data and code used in the original study, and statistical reproducibility focuses on statistical design and analysis to ensure replication of an experiment (Stodden 2015). There are also definitions of ethical reproducibility such as documenting the methods used in biomedical research or in social and behavioral science research so others can reproduce algorithms used in analysis (O'Neil 2016).

Many studies have been undertaken to understand reproducibility of scientific findings and have come to different conclusions about the findings. For example, one scientist argues that half of all scientific discoveries are false (Ioannidis 2005), others find that a large portion of the reproduced findings produce weaker evidence compared with the original findings (Nosek et al. 2015), and others find that more than 4/5 of the results are true positives (Jager & Leek 2013). Each of these studies used different methods to reach their conclusion. Despite this controversy, the premise underlying reproducibility is data quality in the form of good experimental design and execution, documentation, and making scientific inputs available for reproducing the scientific work.

The data revolution within the biological and physical science world is generating massive amounts of data from the research cited above as well as a wide range of other projects, such as those undertaken at the Large Hadron Collider and genomics-proteomics-metabolomics research. The methods and approaches developed to work with the data produced from this research have led to the development of repositories that curate the data, ensure data quality and comparability across studies, and, importantly, create trust when using the data. Curation of the data in these repositories anticipates data users' needs and perceptions, including the opportunity to reproduce the research (Stvilia et al. 2015).

3. THEMES FROM ENGINEERING, COMPUTER SCIENCE, AND BUSINESS

In the computer science, engineering, and business worlds, data quality management focuses largely on administrative data and is driven by the need to have accurate, reliable data for daily operations. The kinds of data traditionally discussed in this data quality literature are fundamental to the functioning of an organization—if the data are bad, firms will lose money (Lima et al. 2007), or defective products will be manufactured (Hazen et al. 2014).

The advent of data quality in the engineering and business worlds traces back to the 1940s and 1950s with Edward Deming (Deming & Geoffrey 1941, Deming 1950) and Joseph Juran (Juran 1951). Japanese companies embraced these methods and transformed their business practices using them. Deming's approach used statistical process control that focused on measuring inputs and processes and thus minimized product inspections after a product was built (Reilly 1994, Neave 2000). Juran (1964) integrated quantitative and qualitative approaches to quality issues through application of the Pareto Principle, that is, 80% of a problem is triggered by 20% of the sources. Although Deming's and Juran's methods were different, their questions were not. Their emphasis on data quality centered around the question, "What is the variation trying to tell us about a process, about the people in the process?" (Deming 1993). Taguchi (1992) later created statistical methods to eliminate variation by concentrating first on the design of the product and then on the manufacturing process.

Since those beginnings, there has been an abundance of work in the literature to define data quality and its dimensions. Redman (1992, 2004) defines high-quality data as those that are fit for their intended uses in operations, decision-making, and planning. An alternative statement of this is "data that are fit for use by data consumers" (Strong et al. 1997). These definitions imply that the data meet the specification requirements, which reflect the implied needs and the degree to which a set of data characteristics (e.g., completeness, validity, accuracy, consistency, availability, and timeliness) fulfill requirements (ISO 1992, Abate et al. 1998). Data quality is further defined from the perspective of the ease of use of the data (Wang & Strong 1996) with respect to the integrity, accuracy, interpretability, and value assessed by the data user (Strong et al. 1997) and other attributes that make the data valuable (Wang et al. 1995, O'Brien et al. 1999).

A slightly different approach is to view the information (data) from the perspective of eliminating errors to improve efficiency and quality of outputs, resulting in increasing consumer satisfaction. Extending the definitions to the operational realms, data quality is sometimes divided into three types: (*a*) operational quality, which examines strategic and tactical work, (*b*) behavioral quality, which focuses on the human aspects and everyday activities, and (*c*) process quality, which looks at techniques and methods including data control tools and information systems (Ballou et al. 1998, Lima et al. 2007).

In this vein, the term "information quality" was introduced in the late 1990s with the recognition that just as a physical product has quality associated with it, so does information (Wang 1998). It has been introduced as a multidimensional concept with respect to assessment, management, and context (Ge & Helfert 2007, Lee et al. 2002) At the earliest stage of the introduction of the term, the difference between data and information began to be debated. Most publications since the late 1990s use the terms data quality and information quality interchangeably. Data quality has become synonymous with fitness-for-use of the information provided to the consumers (Wang & Strong 1996), the quality of the content in an information system (Redman 1996), or the quality of any particular data source or product evaluated in light of its intended use (UNECE 2013). In this manuscript, the term data quality will be used to cover both data and information quality.

3.1. Total Data Quality Management

The engineering, computer science, and business fields were closely intertwined in the creation of the total data quality management (TDQM) approach (Redman 1992, Wang 1998). TDQM provides a general framework for understanding the improvement-through-management approach to data quality. TDQM defines data quality challenges in four areas: process, systems, policy and procedures, and data design (Mosley et al. 2010). The central idea is that to understand and change data quality, information systems should be considered analogous to manufacturing systems, with

Table 2 Hierarchical data quality dimensions

Level I Dimensions	Level II Dimensions
Intrinsic	Believability, accuracy, objectivity, reputation
Contextual	Value-added, relevancy, timeliness, completeness, appropriate amount of data
Representational	Interpretability, ease of understanding, representational consistency, concise representation
Accessibility	Accessibility, access security

Source: Wang & Strong (1996).

data as the raw material and data products as the output. This analogy afforded a natural springboard for the adoption of principles from total quality management (Juran & Godfrey 1999) into the data space, as evidenced by the TDQM movement adopting the International Organization for Standardizations' ISO 9000 (ISO 1992), which emphasizes the quality of any product.

TDQM puts equal emphasis on all aspects of quality management, from the technical, to administrative operations, to human resources, and it includes all stages of manufacturing from detecting issues in quality to final customer satisfaction and legal ramifications. The data quality management approach focuses on the process of collecting, recording, checking, editing, storing, and accessing data, and continually improving the process at each step. In the short run, errors are corrected. In the longer run, the ultimate goal is to prevent errors. Data quality is relative to time as processes improve. The analogy between ISO 9000 and data quality has proven useful, so much so that it serves as a basis for the TDQM framework ontology that includes the following: management responsibilities, operational and assurance costs, research and development, production, distribution, personnel management, and legal function (Wang & Strong 1996). This comprehensive nature of TDQM is one of its benefits.

A core feature of the TDQM process is its description of the dimensions of data quality, which provide a transparent basis for judging the quality of a data source. These dimensions capture multiple aspects of the systems generating the data products in a hierarchical fashion. **Table 2** presents one of the early examples of this approach (Wang & Strong 1996). Although there are several dimensional schemes (see Batini & Scannapieco 2006 and Batini et al. 2009 for a comprehensive review), **Table 2** is presented because of its historical significance for official statistics, which will be discussed in Section 5.

3.2. Data Management

The evolution of the field of data management overlaps significantly with data quality management and TDQM. The concept of data quality management within the world of data management developed in the 1980s in step with the technological ability to access stored data in a random fashion (**http://en.wikipedia.org/wiki/Data_management**). Specifically, as data management encoding process moved from the highly controlled and defined linear process of transcribing information to tape, to a system allowing for the random updating and transformation of data fields in a record, the need for a higher level of control over what exactly can go into a data field (including type, size, and values) became evident (Mosley et al. 2010). Two key data quality concepts came from these data management advances—ensuring data integrity and cleansing legacy data. Data integrity refers to the rules and processes put in place to maintain and assure the accuracy and consistency of a system that stores, processes, and retrieves data (Boritz 2005). Data cleaning refers to the identification of incomplete, incorrect, inaccurate, or irrelevant parts of the data and then replacing, modifying, or deleting this so-called dirty or coarse data (Wickham 2014, Wu 2013).

As the capability to store increasing amounts of data grew, so did the business motivation to improve the quality of administrative data and thereby improve decision making, reduce costs, and gain trust of customers (Redman 1992, 1998, 2001; Mandal 2004). High-quality data facilitates obtaining a competitive advantage by understanding customers and being able to deliver results quickly. Accurate, comprehensive, and timely data provide the foundation for achieving company goals (Redman 1996, 2001). Cleaning data is a short-term solution, and preventing errors is promoted as a permanent solution. The drawback to cleaning the data is that the process never ends, is costly, and may allow many errors to evade detection (Redman 1996, 2001).

The 1980s and 1990s presented the opportunity to combine managing business data for administrative purposes and accessing business data for building business analytics. Mandal (2004) defined business data as part of the business processes. In defining data quality in the context of statistical process control, he streamlines the Wang & Strong (1996) framework to include product features (e.g., price, color, and strength) and the ability of the data to provide useful information but not perceptions of quality (e.g., believability and reputation). Based on these criteria, Mandal defines eight classes of data—wrong, noisy, irrelevant, inadequate, hard, redundant, right, and rich data.

These characteristics correspond to the Wang & Strong (1996) categories in **Table 1**, except for the needed addition of redundancy, which Mandal (2004) feels is important for process calibration or to provide new information. He also points out the relation between factors that affect data quality, such as measurement bias, poor repeatability, correlation, and loss of some data, to data collection problems. Mandal recognizes that the data may be used for future applications and poses questions that should be asked when data are used for new purposes: "Will the data be wrong?" "Will the data be noisy?" and "Will the data be rich?"

4. THEMES FROM MEDICINE AND PUBLIC HEALTH

The use of data in the fields of medicine and public health is concerned with the broad notion of health information: all of the data related to an individual's or a population's health (Cabitza & Batini 2016). The uses of health information are broad and include, for example, clinical trials, direct facilitation of patient care, billing and insurance, scientific and epidemiologic research, and policy analysis.

The clinical patient health record is a longitudinal administrative record of an individual's health information. The health record is a set of nonstandardized data that spans multiple levels of aggregation, from a single measurement element (blood pressure) to collections of diagnoses and related clinical observations (Cabitza & Batini 2016). This complexity is compounded by the high degree of human interaction involved in the production of clinical records, including self-reported data, medical diagnosis, and other patient information. An electronic version of this record is an electronic health record (EHR). EHR systems are collections of EHRs and are being used to repurpose these data to support research for the personalization of health decisions (Richesson et al. 2013, FDA 2013).

Health information quality is concerned with highly dynamic, error-prone, and complex data. Data quality assessment of health information, historically concerned with accuracy and completeness (Arts et al. 2002, Weiskopf & Weng 2013), has moved toward developing a more comprehensive set of dimensions (Liaw et al. 2013, Weiskopf & Weng 2013) that align with those found in engineering, computer science, and business, as discussed in Section 3. However, there remains little consensus in terms of definitions and the broader conceptual framework (Cabitza & Batini 2016).

Two areas in which patient health records, including EHRs, are used to support research are patient registries and clinical trials. These align with the data repositories and experimental

methods as discussed in Section 2; however, because they are derived from patient health information, they have some special data quality challenges worth highlighting.

4.1. Registries

Health registries, such as Breast Cancer Family Registries, are designed observational data repositories used to follow specific patient populations for various purposes, such as tracking a disease natural history, clinical and cost-effectiveness of care and medical procedures, evaluation of patient safety, and policy value (**http://epi.grants.cancer.gov/CFR/about_breast.html**). Common patient populations are defined by a specific disease diagnosis or an exposure to medical procedures or treatments, such as a diagnosis of breast cancer. Some common health information data sources for a single registry could include patient-reported, clinician-reported, medical chart abstraction, EHRs, institutional administrative records, and vital records (Gliklich et al. 2014).

Data quality is driven by multiple dimensions such as clinical data standardization, the existence of common definitions of data fields, and the validity of self-reported patient conditions and outcomes. Recognized issues (Gliklich et al. 2014) include the definitions of data fields and their relational structure, the training of personnel related to data collection, data processing issues (data cleaning), and curation. Furthermore, adverse event detection is a fundamental driver of data quality. The final data structure for a registry balances parsimony (to reduce patient burden), validity (of all data elements) and use (to reach the specific goals of the registry).

4.2. Clinical Trials

Clinical trials are designed experiments based on the principle and concepts discussed in Section 2. Clinical trials are fundamentally dependent on data quality, which in this case includes complicated factors such as the trial design (randomization, blinding, sample size, baseline assessment, measures and outcomes, and recruitment), the patient population (e.g., variation in adherence by disease state), and the data entry methods, which can span EHRs, paper-based patient charts, third-party laboratory data, patient-reported outcomes, and direct entry by the study clinical staff (Soc. Clin. Data Manag. 2014).

The central dimensions of data quality for clinical trials are accuracy, completeness, and consistency, with particular attention to issues related to missing data in clinical populations. This can reflect the complex, nonrandom (censored) processes such as attrition or differential follow-up for highly marginalized populations. Some important data fields are prone to error, including dates and times. Fraudulent data entry is a serious concern for large multisite trials (George & Buyse 2015). A best practice for ensuring data quality is the formation of a formal data quality monitoring and surveillance system that is dynamic and responsive to potential errors and includes on-site visits of clinical sites for the monitoring of procedures (Friedman et al. 2015).

5. THEMES FROM SOCIAL AND BEHAVIORAL SCIENCES

Social and behavioral science researchers use a complementary mix of qualitative and quantitative approaches to collect data. The data collection often includes data that scientists collect themselves through designed data collection or observation. They also use secondary data, designed and administrative, from sources such as those collected by statistical agencies or other organizations. Data quality varies across topics or subfields of social and behavioral sciences. In some cases, the collection methodologies and measures, such as Gross Domestic Product and unemployment, are strictly defined to follow national or international reporting standards. In other cases, the data

sources are more idiosyncratic or problematic, such as datasets used to measure political issues on difficult-to-measure topics, such as human rights violations, corruption, and political institutions and regimes (**http://www.nsd.uib.no/macrodataguide/quality.html**). The data quality focus in this article is on quantitative methods, with special attention given to survey and experimental methods.

Similarly to the physical and biological sciences, the social and behavioral sciences have adopted the use of data repositories. One well-known example is the Inter-university Consortium for Political and Social Research that was established in 1962 (**http://www.icpsr.umich.edu/icpsrweb/landing.jsp**). Statistical organizations also provide repositories primarily of their own data collections, such as the Organization for Economic Development, EuroStat, the US Census Bureau, and other national and international statistical agencies. The many benefits of these repositories include providing a forum for building partnerships, supporting the use of multiple data sources in research, and providing access to data, especially historical data (Green & Gutmann 2007). The repositories have also supported the ability to reproduce research (LeVeque et al. 2012, Stodden et al. 2013, Stodden 2015).

When data are of unknown quality, social and behavioral scientists assess the validity of the data to understand the relationship between the theory and the data collected, use exploratory analysis to identify coverage or completeness of the data, and apply statistical methods to overcome data quality impacts on the analyses and subsequent conclusions. Although in many cases there may be little control over the original design and collection of the data, obtaining information about who collected the data and why can provide insights into the quality of the data.

5.1. Survey Data Quality

The advent of large-scale surveys in the 1930s provided a new source of scientific data for quantitative social and behavioral science research. These surveys moved the field from systematic observation to probability-based designed data collection (Groves 2011). Introducing probability-based surveys immediately improved data quality and the measurement of social and behavioral phenomena. The field of survey research has kept pace with the data quality methods from engineering, computer science, and business, including the incorporation of TDQM into survey processing during the early 1980s (Norwood 1990).

Survey research has concentrated on data quality from the perspective of the data collection process, starting with the research questions and the level of acceptable variability in the findings. Choices are made about the population of interest, the sampling approach (e.g., random, stratified, or cluster), and the sample size, subject to the available budget (Dippo 1997). These choices include how to carry out the processes as carefully as possible, how to document the processes, and how to identify the potential data quality issues that surround these processes, including both sampling and nonsampling errors. Data quality is improved by eliminating sources of error at each point in the data collection and review processes.

Nonsampling error accounts for the uncertainty about the quality of the survey or census data. This error is the sum of all nonsampling errors that occur when constructing the sampling frame, selecting the sample, collecting and processing the data, and estimating the data for analysis. Methods have evolved to understand the quality of survey and census data using a total survey error approach that applies statistical and qualitative analysis to provide a holistic view of survey error, especially nonsampling error (Biemer 2010). For example, Brooks & Bailer (1978) created error profiles for the Current Population Survey to inform both users and producers of statistics. Similarly, total survey error was examined in detail for the US Census Bureau's Survey of Income and Program Participation, a longitudinal survey with a complicated structure and response burden

resulting in high survey attrition (King et al. 1998). Total survey error methods maximize data quality within the constraints of budget and other resources (Biemer & Lyberg 2003).

5.2. Randomized Control Trials, Observational Studies, and Natural Experiments

Social and behavioral scientists use a range of experimental methods, from randomized control trials to observational studies to natural experiments. The emphasis here is on the data generated or acquired for these studies and the approaches to issues of data quality.

Randomized control trials, similar to the experimental methods in Section 2 and clinical trials in Section 4, are used for evaluating an intervention's effectiveness. Randomized control trials are used to answer a specific question and test this with treatment and control (comparison) groups that are assigned randomly by the researcher and in which the manipulation of the treatment is under the control of the researcher. Randomized control trials are common across many fields of study, such as for testing the effectiveness of a vaccine or a fertilizer in an agricultural field trial, or the efficacy of a health behavior intervention. They have been increasingly used in social and behavioral science research for evaluating policy and public programs (Orr 1999).

Data generated from randomized control trials falls into the category of designed data collection. The measurement processes range from survey instruments to behavioral and biometric measurement. Development of the measurement process, experimental design and implementation, data collection, and documentation form the backbone of the assessment of data quality. Rigorous standards for conducting and reporting randomized control trials have been established and are updated routinely (**http://www.consort-statement.org/consort-2010**).

Researchers are aware that randomized control trials are not always possible for ethical, safety, or other reasons. In addition, they do not explain why a policy works or what caused an improvement, because the environment cannot be controlled (Behn 2015). This has led to renewed interest in observational studies (Rosenbaum 2010, NRC 2012) and natural experiments (Levitt & List 2009) and suggests that a mixture of approaches is necessary to assess causation (Deaton & Cartwright 2016). In both cases, the researcher simply observes behavior by evaluating existing data. These studies are frequently augmented with additional designed data collection to capture ethnographic and socio-demographic characteristics.

Observational studies and natural experiments are similar to randomized control trials in that there are two groups (treatment and control) that are assigned randomly, referred to as "as-if random" (Rosenbaum 2010, Dunning 2012). However, the manipulation of the treatment is not under the control of the researcher, because the researcher is using already-collected data, usually from another source. Nonetheless, these experiments have the ability to advance empirical work if the assumption of randomness is credible, and if the data are used in a coherent model that incorporates behavior and additional data, such as technology or economic conditions (Rosenzweig & Wolpin 2000, Levitt & List 2007, Rosenbaum 2010, NRC 2012). It is frequently argued that these types of studies may be conducted at lower cost than randomized control trials because they take advantage of available data (Dunning 2012).

The principal difference between an observational study and a natural experiment rests with how the treatment is applied (Rosenbaum 2010, Dunning 2012). If external factors assign the treatment, for example, changes in social insurance benefits, weather events, or genetic difference in twins, it is considered a natural experiment. If the subjects could exercise choice in selecting the treatment, then it would be considered an observational study, for example, pursuing educational degree programs, taking a drug, or applying for a social benefit.

Natural experiments and observational studies provide additional insights into data quality and are especially relevant to this review article because they emphasize repurposing already-collected data, so the analyst has little to no ability to control the data collection or measurement processes. Understanding the sources of variation provides useful information about data quality (Meyer 1995). Meyer describes the need to acquire a "detailed knowledge of the theory, institutions, data collection, and other background relevant to a topic . . . to judge the importance of these problems for a given study" (p. 152). These dimensions can also be described using concepts from the evaluation literature in terms of threats to internal and external validity (Cook et al. 1979, Meyer 1995).

- Internal validity refers to whether one can draw inferences from the results of the study. Threats to internal validity include omitted variables and interactions, misspecification, mismeasurement, or endogeneity of explanatory variables, selection bias, and differential attrition of respondents from treatment or comparison groups.
- External validity involves assessing whether the effects found in an experiment are generalizable to other situations (individuals, contexts, or outcomes).
- Context validity refers to confusion over what is cause and what is effect; for example, are higher earnings due to educational credentials that signal ability or the actual learning that occurred in school?

In using laboratory experimental (e.g., randomized control trials) and naturally occurring data (e.g., natural experiments), Levitt & List (2009) argue that what is needed to describe the data-generating process is a model of behavior and its relationships to other contexts: "Theory is the tool that permits us to take results from one environment to predict in another, and generalizability of laboratory evidence should be no exception" (p. 170). They conclude that combining laboratory analysis with a decision-making model can be designed to reduce biases and provide useful information. In the same way as Dunning (2012), Deaton (2010), and others, Levitt & List (2009) advocate the benefits of combining the best of laboratory-generated and naturally occurring data in an empirical setting to measure social preferences.

Survey, observational, and experimental studies have traditionally used data consistent with the rubric surrounding the data revolution, such as the criteria described in **Table 3** used to assess data quality. They have also used statistical criteria to assess quality and to establish the credibility of

Table 3 Brackstone's data quality dimensions

Dimension	Definition
Relevance	The degree to which data inform issues of importance to the users of data. Although subjective and variable across current and potential data users, NSOs should strive to put in place a program to meet even potentially conflicting user needs. The degree to which this is possible, and the approach for implementation, is bounded by resource limitations.
Accuracy	The degree to which data represent the phenomena it was designed to measure. Standard statistical principles of bias, variance, and sources of error (coverage, nonresponse, etc.) are used to measure this dimension.
Timeliness	The temporal difference between the time at which the data were collected and when they become available.
Accessibility	From a user's perspective, the ease of accessing the data from an NSO. This incorporates the degree to which it is possible to ascertain the existence of the data, the processes for access, and, in some cases, the cost.
Interpretability	The degree to which the supplemental data and the metadata are useful for interpretation of the data and its uses. This includes conceptual issues, generation and classification of variables, and data collection methods.
Coherence	The extent to which data can be used with other data and over time. This captures conceptual standards, classifications, and methods that go beyond numerical consistency.

Abbreviation: NSO, national statistics organization
Source: These definitions are paraphrased from Brackstone's original 1999 article.

assumptions. These criteria can be directly applied today when repurposing large administrative datasets to study human behavior.

6. ADDITIONAL THEMES FROM STATISTICAL SCIENCES

Statistics clearly deserves its title as the science of uncertainty, especially in the field of data quality. Statistics has always played an important role in data quality management through statistical process control methods, visualization tools, and simulation experiments and is a fundamental tool for all sciences. From a statistical perspective, data quality has two key components. The first component is the design of data collection and measurement, and the second component is the uses of the data. Statisticians assert that data quality is a relative concept and therefore the metrics change depending on the data use (Spencer 1985).

One of the seminal works in the field originated with John Tukey in the application of exploratory data analysis to reveal patterns in data through graphical representations and multiple perspectives on data subsets (Tukey 1962, 1977). These patterns provide insight into the data quality as well as underlying trends and processes. Asking the right question is critical to data quality, as Tukey stressed in one of his famous quotes: "Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise" (Tukey 1962, p. 13).

Data quality has a direct impact on the ability to quantify uncertainty and the strength of the results. Statistics also provides ways to succinctly present data through summary tables, models, visualization methods, and forecasts, each intended to provide information but also to quantify uncertainty. Traditionally, the statistics discipline designs data collection to minimize bias and maximize information content, verifies the quality of the data after it is collected, and analyzes data in a way that produces insight or information to support decision-making (Madigan & Wasserstein 2014). This tradition is being challenged as repurposed administrative and opportunity data are being more commonly incorporated into statistical analyses.

6.1. Decision-Theoretic Approach

The decision-theoretic approach concerns the use of data in a decision framework and is the backbone of data quality assessment in statistics. The direct relationship between data quality and the consequences of actions derived from the use of the data was first highlighted in the literature by Morgenstern et al. (1963). There is a desire to ensure high data quality when the importance of the use of the data increases and the magnitude of the consequences of data error is high. Morgenstern et al. (1963, pp. 117–18) present these ideas as hypotheses:

- Hypothesis 1. The more rudimentary the use of the data, the less quality is needed.
- Hypothesis 2. The needed data quality increases as the magnitude of the consequences of data error increases.
- Hypothesis 3. As the probability that a decision uses data increases, the needed data quality increases.

These trade-offs can be analyzed in a decision-theoretic model, and the conditions under which they hold or fail can be examined (Spencer 1985).

Also of interest is the determination of data quality when data users do not make optimal choices. Perception of data quality is as important as actual data quality. This is particularly important in the context of repurposed data. Data that are perceived to be high quality are more likely to be used than data that are thought to be low quality (Spencer 1985). Benefit-cost analysis provides another mechanism to construct utility functions to represent the consequences of different levels

of data quality, although they cannot take into account all the potential uses of the data, only the immediate use being evaluated (Spencer 1985).

Karr et al. (2006) note that data quality is a "multi-disciplinary problem that brings together ideas from computer science, quality control, human factors research, and the statistical sciences; in any application, these [are] all linked by domain knowledge to the specific context of the problem" (p. 138). And the questions they pose are similar to survey research but this time are focused on achieving data quality (Karr et al. 2006):

- What cost is needed to achieve a specified level of data quality?
- What are the financial benefits of improved data quality?
- What are the costs of poor data quality?

For example, in agricultural field experiments, the researcher makes choices about the number of plots, what research questions are to be answered, and the number of treatments subject to the level of uncertainty that is acceptable, because accuracy and resources compete.

Decision-theoretic approaches bring together the underlying benefits of data quality with the outcome that the data can be used to inform and assess decisions. Data quality not only is an assessment of the data values but also takes into account factors such as accessibility, relevance, timeliness, metadata, documentation, user capabilities and expectations, cost, and context-specific domain knowledge. This is a theme across all disciplines. The fact that all aspects of data quality are important is evident when there is lack of data quality, as demonstrated by our inability to avert the September 11, 2001, terrorist attack because of data quality problems that prevented availability of prompt, accurate, and relevant data from federal databases (Karr et al. 2006).

It is this decision construct that drives all of statistics to find optimal strategies. The goal is to extrapolate this decision construct to using sources of data that were collected for other reasons and repurposing them for new research purposes. For example, how would one adopt concepts such as fitness-for-use, relative measures of data quality, and timeliness (Agafitei et al. 2015, Braaksma & Zeelenberg 2015, Couper 2013)? One could weigh the tradeoffs of lower quality data with increased timeliness in the short run and the reverse in the longer run. In a decision-theoretic construct, one can ask, "What am I willing to give up to know this now?" The answer to this question depends on the uses and decisions to be made. When repurposing data, these tradeoffs in data quality can be measured along the multiple TDQM dimensions, such as timeliness, accuracy, usefulness, and customer satisfaction.

This section ends with a cautionary note. An important aspect of data quality is the correct use and interpretation of statistics. An example is the use of the *p*-value, a frequently used statistical measure to assess the strength of scientific evidence. The American Statistical Association (Wasserstein & Lazar 2016) and Baker (2016) warn that the *p*-value should be only one of the many measures used to draw scientific conclusions or make policy and that the entirety of the methods and analysis should be described. The point is made that the binary nature of true or false associated with hypothesis testing can place too much emphasis on the *p*-value in making scientific inferences if other contextual factors are not included. These factors are exactly the data quality issues that have been repeated many times in this article including documentation of study design, validity of assumptions for data analysis, and rigor in data collection and measurement.

6.2. Official Statistics

The federal statistical system plays an important role in providing high-quality data to inform policymakers. Official statistics also have an important role in informing businesses and the public about economic conditions. Official statistics have evolved from the collection of the first US

Census of Population in 1790, to the collection of expenditure data in 1890, to development of sample design and error measurement in the 1930s. Later in the 1980s, business data quality practices were adopted, along with an additional focus on nonsampling error. Today, data quality must address the integration of survey and nonsurvey data (e.g., repurposed administrative data) into the creation of official data products. Each stage has improved data quality and introduced new dimensions.

Official statistics and data quality have always been closely related through two interrelated historical threads. The first thread focuses on the sources of measurement error (e.g., sampling and nonsampling error), the culmination of which are the contemporary approaches to total survey error (Biemer et al. 2014). The second thread emphasizes very broad notions of data quality that came to maturity not in official statistics but in information systems, as discussed in Section 3 (e.g., Deming 1993, Redman 1992).

These two historical threads intersected in the 1990s, a (re)union that continues to the present day to reflect fitness-for-use concepts from the management and industry practices [e.g., the work of Tayi & Ballou (1998), driven in part by Brackstone's (1999) seminal article]. Brackstone urges that official statistics should expand the notion of data quality from a mean-squared error approach to a more holistic approach, borrowing heavily from the TDQM process (e.g., Wang & Strong 1996). In principle, this holistic approach meant an expansion of both the dimensions on which to judge the quality of data and the processes and institutional structures to provide assurance of data quality. Official statistics is maturing into a comprehensive and modern approach to data quality assurance that incorporates both the total survey error approach and the data quality management approach (ESS 2015, Statistics Canada 2009, UK ONS 2013, Aust. Bur. Stat. 2009, Tam & Clark 2015).

Brackstone (1999) demonstrates the important influence of the TDQM approach, suggesting six dimensions of data quality for official statistics: relevance, accuracy, timeliness, accessibility, interpretability, and coherence. **Table 3** summarizes these definitions. In addition, Brackstone suggests a comprehensive set of institutional mechanisms for managing data quality in federal statistical organizations, and descriptions of the manner in which data quality dimensions might be affected by such mechanisms (e.g., corporate planning, user liaisons, dissemination). Today, Brackstone's work is strongly reflected in contemporary official statistical products both in terms of the dimensions of data quality and a more end-to-end business management model (Lee & Allen 2001; Biemer Lyberg 2003; Statistics Canada 2009; UK ONS 2013; Aust. Bur. Stat. 2009; ESS 2015; UNECE 2013, 2015).

The federal statistical system mandate is to produce statistics that are objective, relevant, accurate, and timely (NRC 2013b). But these characteristics compete with each other, and tradeoffs have to be made among them (Norwood 1990). The release of preliminary employment estimates that are subject to revision when the final estimates are released four months later is an example of these tradeoffs. Statistical design is the primary emphasis of statistical agencies, with a focus on reducing sampling and nonsampling errors. To support this, official statistics management processes in the United States adopted industry approaches to data quality during the 1980s with great success. This involves systematic identification of major sources of error, seeking out quality improvement projects from staff, and taking interdisciplinary approaches to examining quality issues (Norwood 1990).

In addressing sampling and nonsampling errors, Manski (2015) goes back to Morgenstern et al. (1963) to highlight Morgenstern's advocacy of providing measurement error (sampling and nonsampling error) in federal statistics. Manski notes that federal statistical agencies sometimes publish estimates without measures of error. Strategies for communicating the uncertainty in the estimates are similar to defining sampling and nonsampling errors, as in Biemer (2010). Manksi

defines nonsampling error (uncertainty) in federal statistics in three ways: (*a*) transitory statistical uncertainty takes into account the release of preliminary and revised estimates, such as the release of the Bureau of Labor Statistics preliminary monthly state employment estimates and revised estimates four months later, (*b*) permanent statistical uncertainty occurs as a result of nonresponse and inaccurate responses by respondents, and (*c*) conceptual uncertainty comes from not understanding the information derived from the official statistics or lack of clarity about the concepts themselves, such as how poverty is measured or how data are seasonally adjusted.

In the survey world, federal statistics have focused on automated editing, probabilistic record linkage, and implementing measurement error and survey methods to improve data quality. A fundamental premise is that the quality of the data should depend on the use of the data (as echoed throughout this entire article). Official statistics have adopted the data quality management approach in two ways—through the adoption of the data quality dimensions and attributes, and through the lens of total survey error that attempts to control sampling error and nonsampling error, the latter benefiting from the control processes found in businesses (Japec et al. 2015).

In the past decade, the use of external data for official statistics has garnered much interest, mostly in reference to administrative data (see UNECE 2015). In this context, the dimensions of data quality defined in **Table 3** are deemed useful for gauging the quality of administrative data. However, because administrative data are not in the control of a national statistical organization, other considerations have emerged (Verschaeren 2012). An influential approach that originated from Statistics Netherlands (Ossen et al. 2011, Daas et al. 2012, Statistics Netherlands 2012) introduced the following three hyperdimensions of data quality, specifically for administrative data: source, metadata, and data.

The source dimension reflects quality related to the data generator such as procedures for access; metadata refers to the existence of and quality of documentation and knowledge provided by the source; data refers to the quality of the data in terms of population coverage, nonresponse, and precision, among other more technical factors (see SN-MIAD 2013 for a review). An important research thrust has emphasized evaluation frameworks for judging the quality of external data, specifically administrative data, at the input stage to serve as a screening mechanism against extensive investment in external data sources of poor quality (EPA 2000, 2006; Daas et al. 2012; Ossen et al. 2011; Iwig et al. 2013; US Census Bureau 2015).

In the current data revolution, the situation is somewhat different. Current conditions require a wider set of data quality dimensions, including privacy, security, and complexity (UNECE 2015). In addition, the notion that data should be viewed in terms of potential new trade-offs (timeliness versus representativeness), as increasing efficiency when combining data sources, and as potentially generating new data products (Braaksma & Zeelenberg 2015) casts data quality as relational among all data sources. The use of repositories may help to ensure the quality and accessibility of these data to advance social and behavioral research (Petrakos et al. 2014). The creation of the Evidence-Based Policy Commission to recommend the creation of a clearinghouse for sharing administrative data at the US Census Bureau is also a historical step in this direction (OMB 2016).

7. OPPORTUNITY DATA

Although observational data have already been mentioned in previous sections, these new sources of data warrant additional discussion. The data streams that result from opportunity data, such as social media, open the opportunity to capture observations while individuals are in the act of behaving. Unlike studies using survey data, there is no researcher control over the process of data collection. These traditional measurement instruments gather activity data every ten years, as in

the US Census or through self-reported survey and interview data in daily, weekly, or monthly time frames. Opportunity data may give social scientists a look into levels of behavior that they have never before been able to observe (NRC 2012, Keller & Shipp 2017). Many of the data quality challenges identified in previous sections apply here, but because the data can be so massive and unstructured, the solutions may be different in both degree and kind.

The data quality issues associated with opportunity data include lack of transparency and reliability. For example, privately provided and owned data, such as Google searches, web portals, and sensor data, are not often saved at fixed intervals of time, and their underlying algorithms are not publicly available. Importantly from a statistical perspective, their representativeness and error structures are not known.

7.1. Social Media

Social media includes blogs, social networking websites, wikis, social bookmarking or folksonomies, and online media sharing. Social media is grouped into categories based on their functionalities: blogging, media sharing, micro blogging, social bookmarking, social friendship networks, social news, and wikis. Each category can be expected to provide a different form of data requiring different procedures to assure data quality. The enormity of the information that propagates through these communities presents an opportunity for harnessing the data into predictive analyses that touch on topics as diverse as movie box-office revenue (Asur & Huberman 2010), political elections (Tumasjan et al. 2010), civil unrest events (Chen & Neill, 2014, Ramakrishnan et al. 2014, Korkmaz et al. 2016), the stock market (Bollen et al. 2011, Zhang et al. 2011), flu trends (Lampos et al. 2010, Culotta 2010), housing market fluctuations (Wu & Brynjolfsson 2009), and even earthquakes (Sakaki et al. 2010). The use of social media also can inform difficult-to-measure topics such as unemployment rates (Ettredge et al. 2005), inflation (Guzman 2011), consumer sentiments (Choi & Varian 2012), consumer price indices (Cavallo 2015), and housing prices and sales (Wu 2013). Goel et al. (2010) provide a useful survey of work in this area and describe some of the limitations of web search data.

Data quality issues with social media arise for several reasons: Not everyone in a population of interest is present on social media, the degree of activity varies per user, not all users can be identified, and the collection of social media content created by an individual is selective. Unsophisticated collection and analysis of social media messages (e.g., without using additional sources of information) will be wrought with issues of representativeness and systematic error and/or bias (Daas et al. 2012).

Other data quality challenges associated with social media include spam, colloquial usage and intentional misspelling, lack of contextual relevance, and freshness of information. To address these challenges, especially to identify and filter spam content, supervised machine learning techniques with graph-centric (e.g., Kamaliha et al. 2008, Zhu et al. 2008) and content-centric approaches (e.g., Kolari et al. 2006, Ntoulas et al. 2006, Lin et al. 2008) have been developed.

Agarwal & Yiliyasi (2010) propose a TDQM approach to tackle the challenges in social media. Their approach maps data quality dimensions, social media categories, social media challenges, and data quality tools to bridge the gap between the data quality framework and its application in addressing data quality challenges. Their TDQM methodology is an iterative and continuous data quality monitoring and improvement process (Fisher et al. 2012). **Table 4** illustrates the relationship between social media challenges and data quality dimensions of the Wang & Strong (1996) framework given in **Table 2**.

Data quality in the context of social media encompasses many challenges due to the wide accessibility, performance, global audience, recentness, and ease of use of social media (Agarwal &

Table 4 Mapping social media challenges and data quality dimensions

Social Media Challenges	Data Quality Dimensions
Spam	Accuracy, believability, reputation, value-added, relevancy
Contextual relevance	Relevancy
Colloquial usage and intentional misspelling	Accuracy, value-added
Information overload	Amount of data, ease of understanding, manipulability, conciseness
Freshness of information	Accuracy, believability, reputation, timeliness

Source: Adapted from Agarwal & Yiliyasi (2010).

Yiliyasi 2010). Of fundamental importance is the lack of a common theoretical basis for addressing these challenges. However, there is some recent research to do this. Emamjome et al. (2013) propose a conceptual model for assessing data quality in the context of social media. Here, data quality is defined as the degree to which information is suitable for doing a specified task by a specific user, in a certain context. Their conceptual model builds directly off the TDQM concepts presented in Section 3. Emamjome (2014) further maps information dimensions in social media to three types of quality definitions: (*a*) manufacturing-based quality to define quality of stored information including syntax rules of media representation and language; (*b*) user-based quality to include conformance to users' cognitive and meaning system; and (*c*) value-based quality defined as the fitness of information to be used for a specific user, the efforts and costs to derive the information, and how much it contributes to organizations' decision-making.

Research based on opportunity data is gaining in popularity, and issues surrounding the use of these data are emerging. The initial data quality approaches are anchored on past concepts of data quality with significant attention to fitness-for-use.

8. CONCLUSIONS

Data quality transcends all the disciplinary boundaries of science, commerce, engineering, and governing activities. This review has revealed that, as an ever-evolving transdisciplinary undertaking, data quality has benefited from the contributions of all disciplines. Scientific fields created accessible repositories and methods to ensure careful documentation and preservation of data that anticipates future data uses. Engineering and business fields introduced both statistical and qualitative approaches and developed the concept of TDQM dimensions that other disciplines (including national statistical organizations) have adopted. Medicine and public health bridge the disciplines through the standardization of clinical data, refining validity measures, and minimizing respondent burden. Social and behavioral sciences introduced and refined survey quality through measures of sampling and nonsampling error, and embraced natural experiments and observational studies that use and repurpose large corpuses of data, including administrative data. The decision-theoretic approach that provides the underpinnings for sound statistical practices informs the data quality methods and approaches in each of the fields.

The ultimate data quality goal is to develop a disciplined process of identifying data sources, preparing the data for use, and assessing the value of these sources for the intended uses. Donoho (2015) provides a six-part framework that solidifies the future role of data quality and reproducibility of research. The first part includes the activities related to data exploration and preparation that involve learning about the data, identifying issues with the data, and addressing the issues for the selected variables. The second part encompasses the data representation and transformation

activities that address the range of formats and files types, such as text, images, sensors, and tabular data, as well as how the data are saved, and require the capability to work with all types of data. The third part incorporates the activities related to computing with data including utilization of multiple programming languages, such as R and Python, as well as data storage options, such as cluster and cloud computing. Fourth are the activities concerning data visualization. The fifth part concerns activities related to data modeling, including both generative modeling to infer properties of the underlying mechanism and predictive modeling through machine learning. The sixth part addresses the activities associated with creating a new discipline called the Science about Data Science, which focuses on identifying and creating a process to accelerate the science of learning from data.

Statistical sciences are central to Science about Data Science in understanding all types of data and the advances in their use. This may be more important than first imagined, and opportunities exist to amplify this role. Designed data will continue to provide a statistical baseline for accuracy, consistency, and representativeness, although at the expense of timeliness. Administrative and opportunity data allow for earlier and deeper insights into human behavior and organizational processes and the interactions of the two. What is different when using administrative and opportunity data is that we cannot change the collection of the data to improve its quality. What this means is that we must be explicit in accepting larger variation in results when using these data. Statisticians have many creative approaches for making these decisions about tradeoffs, and these tradeoffs between data quality and data usability require conversations between the statisticians and scientists, engineers, business people, administrators, and decision-makers across disciplines.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This research was partially supported by the US Census Bureau under a contract with the MITRE Corporation and was partially based on work supported by the National Science Foundation under Grant No. 1338491. We would like to thank Ted W. Allen and Daniel H. Weinberg for their review of the article.

LITERATURE CITED

Abate M, Diegert K, Allen H. 1998. A hierarchical approach to improving data quality. *Data Qual.* 4(1):365–69

Agafitei M, Gras F, Kloek W, Reis F, et al. 2015. Measuring output quality for multisource statistics in official statistics: some directions. *Stat. J. IAOS* 31(2):203–11

Agarwal N, Yiliyasi Y. 2010. Information quality challenges in social media. *Int. Conf. Inform. Q. (ICIQ).* **http://mitiq.mit.edu/ICIQ/Documents/IQ%20Conference%202010/Papers/3A1_IQChallengesInSocialMedia.pdf**

Arts DGT, De Keizer NF, Scheffer G. 2002. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. *J. Am. Med. Inform. Assoc.* 9(6):600–11

Asur S, Huberman BA. 2010. Predicting the future with social media, *2010 IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol. (WI-IAT)*, Vol. 1, pp. 492–99. Piscataway, NJ: IEEE

Aust. Bur. Stat. 2009. *The ABS Data Quality Framework*. Belconnen, Aust.: Aust. Bur. Stat. **http://www.abs.gov.au/ausstats/abs@.nsf/mf/1520.0**

Baker M. 2016. Statisticians issue warning over misuse of *p*-values. *Nat. News* 531:151

Ballou D, Wang R, Pazer H, Tayi GK. 1998. Modeling information manufacturing systems to determine information product quality. *Manag. Sci.* 44(4):462–84

Batini C, Cappiello C, Francalanci C, Maurino A. 2009. Methodologies for data quality assessment and improvement. *ACM Comput. Surv.* 41(3):16

Batini C, Scannapieco M. 2006. Introduction to data quality. In *Data Quality: Concepts, Methodologies and Techniques*, ed. C Batini, M Scannapieco, pp. 1–18. New York: Springer

Becker KG. 2001. The sharing of canal microarray data. *Nat. Rev. Neurosci.* 2(6):438–40

Behn R. 2015. The black box of randomized controlled trials. *Bob Behn's Perform. Leadersh. Rep.* 12(5):1

Biemer PP. 2010. Total survey error: design, implementation, and evaluation. *Public Opin. Q.* 74(5):817–48

Biemer PP, Lyberg LE. 2003. *Introduction to Survey Quality*. New York: Wiley

Biemer P, Trewin D, Bergdahl H, Japec L. 2014. A system for managing the quality of official statistics. *J. Off. Stat.* 30(3):381–415

Bollen J, Mao H, Zeng X. 2011. Twitter mood predicts the stock market. *J. Comput. Sci.* 2(1):1–8

Boritz JE. 2005. IS practitioners' views on core concepts of information integrity. *Int. J. Account. Inform. Syst.* 6(4):260–79

Braaksma B, Zeelenberg K. 2015. Re-make/re-model: should big data change the modelling paradigm in official statistics? *Stat. J. Int. Assoc. Off. Stat.* 31(2):193–202

Brackstone G. 1999. Managing data quality in a statistical agency. *Surv. Methodol.* 25(2):139–50

Brooks CA, Bailer BA. 1978. *An error profile: employment as measured by the current population survey*. Work. Pap. 3, Off. Fed. Stat. Policy Stand.

Cabitza F, Batini C. 2016. Information quality in healthcare. In *Data and Information Quality: Dimensions, Principles and Techniques*, ed. C Batini, M Scannapieco, pp. 21–51. London: Springer

Cavallo A. 2015. *Scraped data and sticky prices*. Work. Pap. 21490, Natl. Bur. Econ. Res.

Chapman AD. 2005. *Principles of data quality*. Rep., Glob. Biodivers. Inf. Facil., Copenhagen

Chen F, Neill DB. 2014. Non-parametric scan statistics for event detection and fore-casting in heterogeneous social media graphs. *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 1166–75. New York: ACM

Choi H, Varian H. 2012. Predicting the present with Google trends. *Econ. Rec.* 88:2–9

Contreras JL, Reichman JH. 2015. Sharing by design: data and decentralized commons. *Science* 350(6266):1312–14

Cook TD, Campbell DT, Day A. 1979. *Quasi-experimentation: Design Analysis Issues for Field Settings*. Boston: Houghton Mifflin

Couper M. 2013. Is the sky falling? New technology, changing media, and the future of surveys. *Surv. Res. Methods* 7:145–56

Culotta A. 2010. Towards detecting influenza epidemics by analyzing Twitter messages. *Proc. 1st Worksh. Soc. Media Anal.*, pp. 115–122. New York: ACM

Daas P, Roos M, Van de Ven M, Neroni J. 2012. *Twitter as a potential data source for statistics*. Work. Pap. 201221, Cent. Bur. Stat. **http://www.pietdaas.nl/beta/pubs/pubs/DiscPaper_Twitter.pdf**

Deaton A. 2010. Instruments, randomization, and learning about development. *J. Econ. Lit.* 48:424–55

Deaton A, Cartwright N. 2016. *Understanding and misunderstanding randomized controlled trials*. NBER Work. Pap. 22595. **http://www.princeton.edu/~deaton/downloads/Deaton_Cartwright_RCTs_with_ABSTRACT_August_25.pdf**

Deming WE. 1950. Lectures on Statistical Control of Quality. Tokyo: Nippon Kagaku Gijutsu Remmei

Deming WE. 1993. *The New Economics for Industry, Government, Education*. Cambridge, MA: MIT Press

Deming WE, Geoffrey L. 1941. On sample inspection in the processing of census returns. *J. Am. Stat. Assoc.* 36(215):351–60

Dippo CS. 1997. *Survey Measurement and Process Improvement: Concepts and Integration*. Hoboken, NJ: Wiley

Donoho D. 2015. *50 years of data science*. Presented at Tukey Centen. Worksh., Princeton, NJ, September 18. **http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf**

Dunning T. 2012. *Natural Experiments in the Social Sciences: A Design-based Approach*. Cambridge, UK: Cambridge Univ. Press

Emamjome F. 2014. A theoretical approach to conceptualize information quality in social media. *Proc. 25th Australas. Conf. Inf. Syst., Auckland, NZ.* **http://www.pacis-net.org/file/2013/PACIS2013-072.pdf**

Emamjome FF, Rabaa'i AA, Gable GG, Bandara W. 2013. Information quality in social media: A conceptual model. *Proc. Pac. Asia Conf. Inf. Syst. (PACIS 2013)* **http://www.pacis-net.org/file/2013/PACIS2013-072.pdf**

EPA (Environ. Prot. Agency). 2000. *Guidance for data quality assessment: practical methods for data analysis*. Tech. Rep. EPA QA/G-9, Environ. Prot. Agency, Washington, DC

EPA (Environ. Prot. Agency). 2006. *Data quality assessment: statistical methods for practitioners*. Tech. Rep. EPA QA/G-9S, Environ. Prot. Agency, Washington, DC

Ettredge M, Gerdes J, Karuga G. 2005. Using web-based search data to predict macroeconomic statistics. *Commun. ACM* 48(11):87–92

ESS (Eur. Stat. Syst.). 2015. Quality assurance framework of the European Statistical System, version 1.2. **http://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V1-2final.pdf/bbf5970c-1adf-46c8-afc3-58ce177a0646**

FDA (Food Drug Admin.). 2013. *Guidance for Industry: Electronic Source Data in Clinical Investigations*. Washington, DC: Dep. Health Hum. Serv. **http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm328691.pdf**

Fisher C, Lauría E, Chengalur-Smith S, Wang R. 2012. *Introduction to Information Quality*. Bloomington, IN: AuthorHouse

Fisher RA. 1925. Theory of statistical estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, Vol. 22, pp. 700–25. Cambridge, UK: Cambridge Univ. Press

Friedman LM, Furberg CD, DeMets DL, Reboussin DM, Granger CB. 2015. *Fundamentals of Clinical Trials*. New York: Springer. 5th ed.

Ge M, Helfert M. 2007. A review of information quality research—develop a research agenda. *Proc. 12th Int. Conf. Inf. Qual., Cambridge, MA, Nov. 9–11.* **http://mitiq.mit.edu/ICIQ/PDF/A%20REVIEW%20OF%20INFORMATION%20QUALITY%20RESEARCH.pdf**

George SL, Buyse M. 2015. Data fraud in clinical trials. *Clin. Investig.* 5(2):161–73

Gliklich R, Dreyer N, Leavy M. 2014. *Registries for Evaluating Patient Outcomes: A User's Guide*. Rockville, MD: Agency Healthc. Res. Qual. 3rd ed.

Goel S, Hofman JM, Lahaie S, Pennock DM, Watts DJ. 2010. Predicting consumer behavior with web search. *PNAS* 107(41):17486–90

Green AG, Gutmann MP. 2007. Building partnerships among social science researchers, institution-based repositories and domain specific data archives. *OCLC Syst. Serv. Int. Digit. Libr. Perspect.* 23(1):35–53

Groves RM. 2011. Three eras of survey research. *Public Opin. Q.* 75(5):861–71

Guzman G. 2011. Internet search behavior as an economic forecasting tool: the case of inflation expectations. *J. Econ. Soc. Meas.* 36(3):119–67

Hazen BT, Boone CA, Ezell JD, Jones-Farmer LA. 2014. Data quality for data science, predictive analytics, and big data in supply chain management: an introduction to the problem and suggestions for research and applications. *Int. J. Prod. Econ.* 154:72–80

Ioannidis JP. 2005. Why most published research findings are false. *PLOS Med.* 2(8):e124

ISO (Int. Stan. Organ.). 1992. ISO 9000—Quality Management. Geneva: ISO. **http://www.iso.org/iso/iso_9000**

Iwig W, Berning M, Marck P, Prell M. 2013. *Data quality assessment tool for administrative data*. Work. Pap. WP 46, Fed. Comm. Stat. Methodol.

Jager LR, Leek JT. 2013. An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics*. doi: 10.1093/biostatistics/kxt007

Japec L, Kreuter F, Berg M, Biemer P, Decker P, et al. 2015. Big data in survey research: AAPOR task force report. *Public Opin. Q.* 79:839–80

Juran JM. 1951. Directions for ASQC. *Ind. Qual. Control* 8(3):30–34

Juran JM. 1964. *Managerial Breakthrough: A New Concept of the Manager's Job*. New York: McGraw-Hill

Juran JM, Godfrey AB. 1999. *Juran's Quality Handbook*. New York: McGraw-Hill

Kamaliha E, Riahi F, Qazvinian V, Adibi J. 2008. Characterizing network motifs to identify spam comments. *Proc. 2008 IEEE Int. Conf. Data Min. Worksh. (ICDMW)*, pp. 919–28. Piscataway, NJ: IEEE

Karr AF, Sanil AP, Banks DL. 2006. Data quality: a statistical perspective. *Stat. Methodol.* 3(2):137–73

Keller M, Schimel DS, Hargrove WW, Hoffman FM. 2008. A continental strategy for the national ecological observatory network. *Front. Ecol. Environ.* 6(5):282–84

Keller S, Shipp S. 2017. Building resilient cities: harnessing the power of urban analytics. In *The Resilience Challenge: Looking at Resilience Through Multiple Lenses*, eds. J Bohland and P Knox, forthcoming. Springfield, IL: Charles C Thomas Ltd

King K, Petroni R, Singh R. 1998. *Quality profile for the survey of income and program participation*. Work. Pap. 230. US Census Bur. **https://www.census.gov/sipp/workpapr/wp30.pdf**

Kolari P, Java A, Finin T, Oates T, Joshi A. 2006. Detecting spam blogs: A machine learning approach. *Proc. Natl. Conf. Artif. Intelligence*, Vol. 2, pp. 1351–56. Palo Alto, CA: AAAI

Korkmaz G, Cadena J, Kuhlman CJ, Marathe A, Vullikanti A, Ramakrishnan N. 2016. Multi-source models for civil unrest forecasting. *Soc. Netw. Anal. Min.* 6:50

Lampos V, De Bie T, Cristianini N. 2010. Flu detector—tracking epidemics on Twitter. In *Machine Learning and Knowledge Discovery in Databases*, ed. W Daelemans, K Morik, pp. 599–602. New York: Springer

Lee G, Allen B. 2001. *Educated Use of Information about Data Quality*. Belconnen, Aust.: Aust. Bur. Stat.

Lee YW, Strong DM, Kahn BK, Wang RY. 2002. AIMQ: a methodology for information quality assessment. *Inf. Manag.* 40(2):133–46

LeVeque RJ, Mitchell IM, Stodden V. 2012. Reproducible research for scientific computing: tools and strategies for changing the culture. *Comput. Sci. Eng.* 14(4):13

Levitt SD, List JA. 2007. What do laboratory experiments measuring social preferences reveal about the real world? *J. Econ. Perspect.* 21:153–74

Levitt SD, List JA. 2009. Field experiments in economics: the past, the present, and the future. *Eur. Econ. Rev.* 53(1):1–18

Liaw ST, Rahimi A, Ray P, Taggart J, Dennis S, et al. 2013. Towards an ontology for data quality in integrated chronic disease management: a realist review of the literature. *Int. J. Med. Inform.* 82:10–24

Lima LFR, Macada ACG, Koufteros X. 2007. A model for information quality in the banking industry–the case of the public banks in Brazil. *Proc. 2007 Int. Conf. Inf. Qual.*

Lin YR, Sundaram H, Chi Y, Tatemura J, Tseng BL. 2008. Detecting splogs via temporal dynamics using self-similarity analysis. *ACM Trans. Web (TWEB)* 2(1):4

Madigan D, Wasserstein R. 2014. *Statistics and science: a report of the London workshop on the future of the statistical sciences*. Lond. Worksh. Future Stat. Sci. **http://www.worldofstatistics.org/wos/pdfs/Statistics&Science-TheLondonWorkshopReport.pdf**

Mandal P. 2004. Data quality in statistical process control. *Total Q. Manag. Bus. Excell.* 15(1):89–103

Manski CF. 2015. Communicating uncertainty in official economic statistics: an appraisal fifty years after Morgenstern. *J. Econ. Lit.* 53(3):631–53

McNutt M. 2014. Journals unite for reproducibility. *Science* 346(6210):679–79

Meyer BD. 1995. Natural and quasi-experiments in economics. *J. Bus. Econ. Stat.* 13(2):151–61

Milham MP. 2012. Open neuroscience solutions for the connectome-wide association era. *Neuron* 73(2):214–18

Morgenstern O. 1963. *On the Accuracy of Economic Observations*. Princeton, NJ: Princeton Univ. Press

Mosley M, Brackett MH, Earley S, Henderson D. 2010. *The DAMA Guide to the Data Management Body of Knowledge*. Bradley Beach, NJ: Technics

NRC (Natl. Res. Counc.). 2012. *Using Science as Evidence in Public Policy*. Washington, DC: Natl. Acad. Press

NRC (Natl. Res. Counc.). 2013a. *Frontiers in Massive Data Analysis*. Washington, DC: Natl. Acad. Press

NRC (Natl. Res. Counc.). 2013b. *Principles and Practices for a Federal Statistical Agency*. Washington, DC: Natl. Acad. Press

NRC (Natl. Res. Counc.). 2014. *Furthering America's Research Enterprise*. Washington, DC: Natl. Acad. Press

Neave HR. 2000. The Deming dimension: management for a better future. In *The Collection of the English Papers in the December 2006 Revision of the Deming Homepage*, pp. 69–78. Zumikon, Switzerland: Swiss Deming Institute. **https://www.skgep.gov.ae/docs/default-source/Articles/article2.pdf#page = 69**

Norwood JL. 1990. Distinguished lecture on economics in government: Data quality and public policy. *J. Econ. Perspect.* 4:3–12

Nosek BA, Aarts AA, Anderson JE, Anderson CJ, Attridge PR, et al. 2015. Estimating the reproducibility of psychological science. *Science* 349(6251):aac4716–aac4716. **http://science.sciencemag.org/content/sci/349/6251/aac4716.full.pdf**

Ntoulas A, Najork M, Manasse M, Fetterly D. 2006. Detecting spam web pages through content analysis. *Proc. 15th Int. Conf. World Wide Web*, pp. 83–92. New York: ACM

O'Brien JF, Bodenheimer RE Jr., Brostow GJ, Hodgins JK. 1999. *Automatic joint parameter estimation from magnetic motion capture data*. Tech. Rep., Georgia Inst. Technol., Atlanta, GA

OMB (Off. Manag. Budg.). 2016. Building the Capacity to Produce and Use Evidence. In *Analytical and Perspectives*, *Budget of the U.S. Government*, *Fiscal Year 2017*, pp. 69–77. Washington, DC: Off. Manag. Budg.

O'Neil C. 2016. The ethical data scientist: people have too much trust in numbers to be intrinsically objective. *Slate*, Feb. 4. **http://www.slate.com/articles/technology/future_tense/2016/02/how_to_bring_better_ethics_to_data_science.html**

Orr LL. 1999. *Social Experiments: Evaluating Public Programs with Experimental Methods*. Thousand Oaks, CA: Sage

Ossen SJ, Daas PJ, Tennekes M. 2011. Overall assessment of the quality of administrative data sources. *Proc. 58th World Statistical Congress*, *2011*, *Dublin*. The Hague, Neth.: Int. Stat. Inst.

Peng RD. 2009. Reproducible research and biostatistics. *Biostatistics* 10(3):405–8

Petrakos M, Santourian A, Farmakis G, Stavropoulos P, Oikonomopoulou G, et al. 2014. Analysis of the potential of selected big data repositories as data sources for official statistics. *Proc. 27th Panhellenic Stat. Conf.* Athens: Greek Stat. Inst.

Ramakrishnan N, Butler P, Muthiah S, Self N, Khandpur R, et al. 2014. 'Beating the news' with EMBERS: Forecasting civil unrest using open source indicators. *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min*, pp. 1799–808. New York: ACM

Redman TC. 1992. *Data Quality: Management and Technology*. New York: Bantam Books

Redman TC. 1998. The impact of poor data quality on the typical enterprise. *Commun. ACM* 41(2):79–82

Redman TC. 2001. *Data Quality: The Field Guide*. Boston: Digital Press

Redman TC, Blanton A. 1996. *Data Quality for the Information Age*. Norwood, MA: Artech House, Inc.

Redman TC. 2004. Data: an unfolding quality disaster. *DM Rev.* 14(8):21–23

Reilly NB. 1994. *Quality: What Makes It Happen?* New York: John Wiley and Sons

Ren GJ, Glissmann S. 2012. Identifying information assets for open data: the role of business architecture and information quality. *Proc. 2012 IEEE 14th Int. Conf. Commer. Enterp. Comput. (CEC)*, *pp. 94–100*. Piscataway, NJ: IEEE

Richesson RL, Hammond WE, Nahm M, Wixted D, Simon GE, et al. 2013. Electronic health records based phenotyping in next-generation clinical trials: A perspective from the NIH Health Care Systems Collaboratory. *J. Am. Med. Inform. Assoc.* 20:e226–31

Rosenbaum PR. 2010. *Design of Observational Studies*. New York: Springer

Rosenzweig MR, Wolpin KI. 2000. Natural "natural experiments" in economics. *J. Econ. Lit.* 38:827–74

Sakaki T, Okazaki M, Matsuo Y. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. *Proc. 19th Int. Conf. World Wide Web*, pp. 851–60. New York: ACM

Sloan Digital Sky Survey (SDSS). 2008. SDSS-III: massive spectroscopic surveys of the distant universe, the Milky Way galaxy, and extra-solar planetary systems. *SDSS*, Jan. 8. **https://www.sdss3.org/collaboration/description.pdf**

SN-MIAD. 2013. *Methodologies for integrated use of administrative data in the statistical process (MIAD)*. Stat. Netw. Tech. Rep., UNECE, Geneva. **https://ec.europa.eu/eurostat/cros/system/files/Preliminary%20report%20on%20Quality%20Assessment%20Framework_0.pdf**

Soc. Clin. Data Manag. 2014. *eSource implementation in clinical research: a data management perspective*. White Pap., Soc. Clin. Data Manag., McLean, VA. **http://www.clinicalink.com/wp-content/uploads/2014/06/SCDM-eSource-Implementation_061214.pdf**

Spencer BD. 1985. Optimal data quality. *J. Am. Stat. Assoc.* 80(391):564–73

Statistics Canada. 2009. *Statistics Canada Quality Guidelines*. Ottowa, Can.: Stat. Can. 5th Ed.
Statistics Netherlands. 2012. *49 Factors that Influence the Quality of Secondary Data Sources*. The Hague, Neth.: Stat. Neth.
Stodden V. 2015. Reproducing statistical results. *Annu. Rev. Stat. Appl.* 2:1–19
Stodden V, Borwein J, Bailey DH. 2013. Setting the default to reproducible. *Comput. Sci. Res. SIAM News* 46:4–6
Strong DM, Lee YW, Wang RY. 1997. Data quality in context. *Commun. ACM* 40(5):103–10
Stvilia B, Hinnant CC, Wu S, Worrall A, Lee DJ, et al. 2015. Research project tasks, data, and perceptions of data quality in a condensed matter physics community. *J. Assoc. Inform. Sci. Technol.* 66(2):246–63
Taguchi G. 1992. *Introduction to Quality Engineering: Designing Quality into Products and Processes*. Tokyo: Asian Product. Organ.
Tam S, Clarke F. 2015. *Big data, statistical inference and official statistics*. Res. pap. 1351.0.55.054, Aust. Bur. Stat., Canberra, Aust.
Tayi GK, Ballou DP. 1998. Examining data quality. *Commun. ACM* 41(2):54–57
Tukey JW. 1962. The future of data analysis. *Ann. Math. Stat.* 33:1–67
Tukey JW. 1977. *Exploratory Data Analysis*. New York: Pearson
Tumasjan A, Sprenger TO, Sandner PG, Welpe IM. 2010. Predicting elections with Twitter: what 140 characters reveal about political sentiment. *Proc. 4th Int. AAAI Conf. Weblogs Soc. Media, ICWSM 2010, Washington, DC, USA, May 23–26, 2010*
UK ONS (UK Off. Natl. Stat.). 2013. *Guidelines for measuring statistical output quality, version 4.1*. London: ONS. **http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/guide-method/method-quality/quality/guidelines-for-measuring-statistical-quality/index.html**
UNECE (UN Econ. Comm. Eur.). 2013. The generic statistical business process model (GSBPM). Version 5.0. **http://www1.unece.org/stat/platform/display/GSBPM/GSBPM+v5.0**
UNECE (UN Econ. Comm. Eur.). 2015. *Using Administrative and Secondary Sources for Official Statistics: A Handbook of Principles and Practices*. Geneva: UNECE
US Census Bur. 2015. *Review of administrative data sources relevant to the American Community Survey*. Work. Pap., US Dep. Commer., Washington, DC
Verschaeren F. 2012. *Checking the usefulness and initial quality of administrative data*. Presented at Meet. Am. Stat. Assoc. (ASA), 4th Int. Conf. Establishment Surv. **http://www.q2012.gr/articlefiles/sessions/20.2_Verschaeren_ESSnet%20Admin%20data.pdf**
Wang RY. 1998. A product perspective on total data quality management. *Commun. ACM* 41(2):58–65
Wang RY, Reddy MP, Kon HB. 1995. Toward quality data: An attribute-based approach. *Decis. Support Syst.* 13(3):349–72
Wang RY, Strong DM. 1996. Beyond accuracy: What data quality means to data consumers. *J. Manag. Inform. Syst.* 12:5–33
Wasserstein RL, Lazar NA. 2016. The ASA's statement on *p*-values: context, process, and purpose. *Am. Stat.* 70:129–33
Weiskopf NG, Weng C. 2013. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J. Am. Med. Inform. Assoc.* 20:144–51
Wickham H. 2014. Tidy data. *J. Stat. Softw.* 59(10):1–23
Williams RH, Zimmerman DW, Ross DC, Zumbo BD. 2006. *Twelve British Statisticians*. Raleigh, NC: Boson
Wu L, Brynjolfsson E. 2009. The future of prediction: how Google searches foreshadow housing prices and quantities. In *ICIS 2009 Proceedings*, paper 147
Wu S. 2013. A review on coarse warranty data and analysis. *Reliability Eng. Syst. Saf.* 114:1–11
Zhang X, Fuehres H, Gloor PA. 2011. Predicting stock market indicators through Twitter "I hope it is not as bad as I fear." *Procedia Soc. Behav. Sci.* 26:55–62
Zhu L, Sun A, Choi B. 2008. Online spam-blog detection through blog search. *Proc. 17th ACM Conf. Inform. Knowl. Manag*, pp. 1347–48. New York: ACM

Annual Review of Statistics and Its Application

Volume 4, 2017

Contents

Errata

An online log of corrections to *Annual Review of Statistics and Its Application* articles may be found at http://www.annualreviews.org/errata/statistics