Published in final edited form as:
Account Res. 2017 ; 24(6): 344–358. doi:10.1080/08989621.2017.1327813.

Data-Intensive Science and Research Integrity

David B. Resnik, JD, PhD[*],
National Institute for Environmental Health Ethics, National Institutes of Health, 111 Alexander Drive, Research Triangle Park, NC, 27709, USA. resnikd@niehs.nih.gov

Kevin C. Elliott,
Lyman Briggs College, Department of Fisheries and Wildlife, and Department of Philosophy, Michigan State University

Patricia A. Soranno, PhD, and
Department of Fisheries and Wildlife, Michigan State University

Elise M. Smith, PhD
National Institute for Environmental Health Ethics, National Institutes of Health

Abstract

In this commentary, we consider questions related to research integrity in data-intensive science and argue that there is no need to create a distinct category of misconduct that applies to deception related to processing, analyzing, or interpreting data. The best way to promote integrity in data-intensive science is to maintain a firm commitment to epistemological and ethical values, such as honesty, openness, transparency, and objectivity, which apply to all types of research, and to promote education, policy development, and scholarly debate concerning appropriate uses of statistics.

Keywords

data-intensive science; research integrity; misconduct; ethics; education; deception; transparency

The growing importance of data-intensive science (a.k.a. “big data”) is one of the most significant trends in contemporary research (Bell et al 2009, Miller 2010, Elliott et al 2016, Sigumoto et al 2016). Data-intensive science has been described as research in which the capture, curation, and analysis of large volumes of data is central to the scientific question. While efforts to handle such large quantities of data often challenge traditional approaches (Hey et al 2009), historians have pointed out that scientists have faced challenges in collecting and analyzing large volumes of data for centuries (Laudan 1981, Muller-Wille and Charmantier 2012). Nevertheless, contemporary data-intensive science is distinctive because it incorporates new computational methods and technologies and tends to involve large, interdisciplinary scientific teams (Leonelli 2014; Strasser 2012). Data-intensive research now plays an important role in many disciplines, including particle physics, astronomy, cancer therapy, drug discovery, radiology, ecology, climatology, geology, molecular biology,

[*]corresponding author.

molecular biology, systems biology, economics, and social science (Elliott et al 2016, Sugimoto et al 2016).

In data-intensive science, investigators often use statistical methods encoded in computer algorithms to search for patterns and associations inone or many largedatasets, which may contain billions of terabytes of information, millions of data-points, and hundreds of variables (Ekbia et al 2015). For example, GenBank (2016) contains data from 200 million DNA sequences, PubMed (2016) includes 26 million citations from the biomedical literature, and the National Energy Research Scientific Computing Centerincludes 2.8 petabytes of information related to nuclear physics (Compare Business Products 2016).Some examples of data-intensive science include: genome-wide association studies, which examine thousands of genetic variants to determine whether some of these are associated with diseases or health-related phenotypes (National Genome Research Institute 2015); social network analyses, which examine thousands of social relationships to understand how these are related to behaviors (Scott 2013); and models of climate change, which are based on thousands of measurements involving numerous variables, such as surface temperatures, ocean temperatures, plankton growth,and cloud cover (Intergovernmental Panel on Climate Change 2013). Because data-intensive science often involves collaborators from different disciplines who may not be able to criticize each other’s contributions, trust among members of research teams is paramount.

Data-intensive science is often contrasted with hypothesis-driven science, with data-intensive methods focusing on identifying statistical associations rather than testing hypotheses concerning causal relationships (Prensky 2009, Steadman 2013). We think it is better to characterize scientific research as an iterative process in which scientists move back and forth between different epistemic activities (Elliott 2012, O’Malley et al 2009). For example, when studying a new or poorly understood phenomenon, scientists may begin with loosely formulated questions that guide exploratory research and the development of new concepts, instruments, and techniques. As the research progresses, new questions are likely to arise, and in some cases hypotheses can be formulated and tested, which in turn may generate new questions (O’Malley et al 2010). In other cases, research which is initially designed to confirm or falsify a hypothesis may yield large quantities of data that may be reused for more data-driven research of an exploratory nature, leading to thediscovery of unexpected patterns or associations. On this view, data-intensive and hypothesis-driven modes of research both play important roles in scientific investigation(Elliott et al. 2016).

Some have criticized data-intensive science by claiming that it does not provide rigorous tests of hypotheses and thus tells us nothing about causal relationships or mechanisms. Critics have also pointed out that it is relatively easy to find statistically significant but meaningless associations in large datasets (see Fan et al 2014 and Ekbia et al 2015 for discussion of these critiques). We contend that many of these criticisms can be alleviated by understanding the limitations of data-intensive research methods and using them appropriately. For example, computational researchers have reverse engineered computer-generated algorithms to peer inside the “black box” of computer-mediated statistical inference. Data-intensive methods may also be treated as components of broader research programs in which statistical associations are subjected to further investigation (Kitchin

2014, Leonelli 2012).Elliott et al (2016) argue that instead of privileging one form of inquiry over another, scientists should focus on employing the best methods for addressing knowledge gaps and creating research teams with the necessary expertise to employ those methods successfully.

In this commentary, we will focus on questions related to research integrity in data-intensive science. While many scientists and scholars have written about other ethics topics related to data-intensive science, such as sharing and owning data and protecting privacy (van Wel and Royakkers 2004, Raman and Ramos 2013), few have discussed research integrity issues. We shall argue that there is no need to create a distinct category of misconduct that applies to deception related to processing, analyzing, or interpreting data and that the best way to promote integrity in data-intensive science is to maintain a firm commitment to epistemological and ethical values, such as honesty, openness, transparency, and objectivity,whichapply to all types of research, and to promote education, policy development, and scholarly debate concerning appropriate uses of statistics (Resnik 2000).

Research Integrity

Research integrity can be understood as adhering to commonly accepted ethical and professional norms in the conduct of research and making responsible decisions when faced with ethical dilemmas (National Academy of Science Sciences 1992, Steneck 2006, De Winter and Kosolosky 2013, Shamoo and Resnik 2015). Research integrity is important to 1) advance the goals of science, 2) promote trust among scientists, 3) foster the public’s support for research and hold science accountable to the public; and 4) ensure that science conforms to moral standards (Shamoo and Resnik 2015). Some widely accepted ethical norms in science include: honesty (e.g., not faking or fudging data), objectivity (e.g.,minimizing bias and self-deception), carefulness (e.g., keeping good records, avoiding sloppiness) openness (e.g.,sharing of data, methods, etc.), fair allocation of credit (e.g., appropriate authorship), intellectual freedom (e.g. absence of censorship), respect for colleagues, respect of intellectual property, promotion of animal welfare, protection of the rights and welfare of human subjects in research, and social responsibility (Macrina 2013, Shamoo and Resnik 2015).

Misconduct is one of the central concepts related to research integrity. Misconduct can be understood, in a very general sense, as an egregious violation of science’s widely accepted ethical norms (Shamoo and Resnik 2015). If we think of scientific behavior as falling on a spectrum, misconduct lies at the unethical end and research integrity lies at the ethical end, with questionable research practices (such as not disclosing a conflict of interest or questionable authorship attribution) falling in the middle (Shamoo and Resnik 2015).

However, misconduct also has alegal meaning in the United States and many other countries. Government agencies, academic institutions, and scientific journals have adopted their own definitions of misconduct for compliance and oversight purposes.For example, if the U.S. Office of Research Integrity (ORI) finds that an investigator has committed misconduct inresearch funded by the National Institutes of Health (NIH), it may bar the investigator from receiving federal research funds for a period of time. If a university determines that one of its employees has committed research misconduct, it may terminate his or her employment. If the editors of a journal determine that published paper includes fabricated or falsified data or plagiarized text, they may retract the paper (Resnik et al 2015c).To enhance our understanding of current definitions of research misconduct it will be useful to place them in historical context.

A Brief History of Research Misconduct Definitions

One of the first published accounts of unethical practices in science appeared in a book titled *The Decline of Science in England*, written by nineteenth century British mathematician, scientist, and philosopher Charles Babbage (1830). Babbage chastised some of his colleagues for engaging in research practices that he regarded as deceptive and unethical, including forging (making up data), trimming (excluding data that contradict one's hypothesis), cooking (designing an experiment to obtain a predetermined result, not to test a hypothesis) (Babbage 1830). Dishonesty is a common thread in these three unethical behaviors (Resnik 1998). While Babbage's ideas were never adopted as official government policies, they haveinfluencedhow scientists and policymakers have thought about honesty in science (Broad and Wade 1993).

Although famous episodes of misconduct occurred in science long before and after Babbage's time (such as the Piltdown Man hoax), governments did not mount a significant response to the problem of misconduct until the 1980s, when highly-publicized cases of fraudulentresearch caught the attention of the public and U.S. Congress. In 1981, Representative Al Gore Jr. held hearings on fraud in federally-funded biomedical research and discovered that neither government agencies nor universities had effective policies for dealing with such problems (Steneck 1999). The prevailing wisdom at the time was that government rules were unnecessary because science was self-regulating. Little had changed when Representative John Dingell also held hearings on the same topic four years later. Dingell's hearings prompted Congress to pass legislation requiring the Public Health Service (PHS), which funds NIH research, to adopt misconduct regulations (Steneck 1999).The PHS complied with this mandate, and in 1989 also began requiring students and trainees supported by its funds to receive education in responsible conduct of research (RCR). The congressional legislation also created the Commission on Research Integrity, chaired by Harvard University scientist Paul Ryan, which held hearings from 1992 to 1995. The Commission made recommendations concerning federal policies for defining, reporting, investigating, and preventing misconduct. During the remainder of the decade, U.S. government agencies, including the NIH, the National Science Foundation, and the Department of Education, revised and harmonized their misconduct policies (Steneck 1999).

In 2000, the Office of Science and Technology Policy (OSTP) adopted a research misconduct policy that applies to all federal agencies. Institutions that receive federal funds for research are required to follow this policy (Shamoo and Resnik 2015). The policy defines research misconduct as:

> [F]abrication, falsification, or plagiarism in proposing, performing, or reviewing research, or in reporting research results…Misconduct does not include honest

> error or difference of scientific opinion…(Office of Science and Technology Policy 2000: 76262).

The policy also states that to make a finding of misconduct an agency or an institution must determine that:

> There be a significant departure from accepted practices of the relevant research community; and the misconduct be committed intentionally, or knowingly, or recklessly; and the allegation be proven by a preponderance of evidence(Office of Science and Technology Policy 2000: 76262).

The federal policy characterizes fabrication as “making up data or results and recording or reporting them,” falsification as “manipulating research materials, equipment, or processes, or changing or omitting data or results such that the research is not accurately represented in research record,” and plagiarism as “the appropriation of another person’s ideas, processes, results, or words without giving appropriate credit (OSTP 2000: 76262).”The federal policy focused on three unethical behaviors—fabrication, falsification, and plagiarism (FFP)—and eliminated a category previously used by the PHS and NSF, “other serious deviations from accepted research practices”(Dooley and Kerch 2000).

OSTP officials decided to eliminate the “other serious deviations” category from the definition because they viewed it as too vague and open-ended (National Academy of Sciences 1992, Resnik 2003). They were concerned that including this category in the definition of misconductmight 1) encourageallegations which have little to do withtruthfulness in research, such as accusations of sexual harassment or financial mismanagement, and 2) deter scientists from conducting important research that uses innovative or unorthodox methods or concepts because they would fear that they could be accused of seriously deviating from accepted scientific practices (National Academy of Sciences 1992, Dooley and Kerch 2000). Historically, ground-breaking scientific advances have come from methodological and conceptual innovationsthat deviate from the*status quo*(Kuhn 1970, Laudan 1981). For example, Isaac Newton developed differential and integral calculusto calculate changes in velocity and acceleration of objects in motion (Kline 1982). [1]Founders of quantum mechanics proposed statistical models of the behavior of subatomic matter whichchallenged traditional notions of causation, determinism, and observation in science (Hughes 1989).

Though the U.S.has taken the lead in addressing research misconduct, other countries, research institutions, and journal associations have also adopted definitions of research misconduct. While most of these include FFP, some encompass other categories, such asserious deviations, inappropriate authorship, violating confidentiality of peer review, or failing to disclose conflicts of interest (Resnik et al 2015b). Some also list forms of deception in research other than FFP. For example, Australia defines misconduct as “fabrication, falsification, plagiarism, or deception in proposing, carrying out or reporting the results of research (Australian Government 2007). The U.K.’s definition includes FFP as well as misrepresentation, which is defined as “suppression or deliberate or negligent

[1] Wilhelm Gottfried Leibniz independently developed calculus (Kline 1982).

misrepresentation of findings and/or data (United Kingdom Research Council 2012). Many U.S. research institutions also have definitions of misconduct that include categories of behavior that go beyond FFP, such as serious deviations and significant violations of human or animal research regulations(Resnik et al 2015a).

Deception in Data-Intensive Science

When the U.S. began developing its policies on research misconduct, data-intensive practices played a less prominent role in scientific research. Though data-intensive science was occurring (for example, the Human Genome Project began in 1990), it was on asmaller scale than it is today, and the technological capacity and computational expertise required to process and analyze large amounts of information had not yet been developed. Consequently, there may be forms of deception in data-intensive science that researchers and policymakers could not adequately consider when they were defining misconduct as FFP.Clearly, a scientist analyzing a small dataset could commit misconduct by fabricating or falsifying data, but scientists could engage in other deceptive practices involving data.

The case of former Duke University genomics researcher Anil Potti illustrates some scientific and ethical issues that can arise in potentially deceptive practices involving the statistical analysis of large datasets. Potti and his collaborators were working on a method to discernstatistical relationships between publicly available DNA microarray data for tumors andtumor cell-line responses to various chemotherapy agents.Suspicions concerning the research began to surface when two biostatisticians, Keith Baggerly and Kevin Combes, were unable to reproduce the results of a paper Potti and coauthors published in *Nature Medicine* (Potti et al 2006, Coombes et al 2007). Duke University's institutional review board (IRB) suspended clinical trials of the method in response to these concerns, but then restarted them again when it determined that there were no problems with the research that would place human subjects unduly at risk. Third year medical student Bradford Perez, who was working with Potti, discovered that the method had not been independently validated and the computer program was unstable. He brought these concerns to Duke University officials, who suggested thathis criticism of Potti's work amounted to a difference of scientific opinion and did not meet the definition of research misconduct (Goldberg 2015).However, Duke University decided to launch a misconduct investigation after learning that Potti had falsely claimed he was a Rhodes Scholar on grant applications. The IRB suspended the clinical trials upon learning that Potti was being investigated for misconduct, and the University agreed to a legal settlement with patients who claimed they were injured in these studies(Goldberg 2015).

In 2015, Duke University and ORI (2015) found that Potti had committed misconduct. However, their findings were based on evidence that Potti had fabricated and falsified data used in his research, not on concerns relating to the legitimacy of his research methods. If Duke and ORI had not determined that Potti fabricated and falsified data the case probably would have been dropped, since the other scientifically questionable acts which Potti allegedly committed do not fit the FFP definition of misconduct used by the federal government and Duke University (2007). This counterfactual scenario raises the issue of whether the definition of misconduct should be expanded to include deceptive practices related to data processing, analysis, or interpretation which do not involve or result in FFP. While these forms of deception are not unique to data-intensive science, they become more difficult to accurately detect when working with large volumes of data. Therefore, the increasing prevalence of data-intensive research in recent years makes it important to consider whether the definition of misconduct should be updated.

Deception in data-intensive science could occur in multiple ways at various stages in the research. First, because it can be difficult to distinguish between meaningful patterns in the data and random noise, data often must undergo several stages of processing prior to analysis, including cleaning, editing, and coding (Miller 2010, Ekbia et al 2015).Researchers could deceptivelymodify or exclude data from further analysis without proper justification (Resnik 2000, Fan et al 2014). This type of deception might go unnoticed, since most studies have valid data processing criteria and procedures which can be difficult to decipher.

Second, during the data-analysis stage researchers must makenumerous choices among different statistical tests, inferential modes and practices, statistical evaluation criteria (e.g.,Bayesian versus frequentist inference, significance levels or p-values), as well as whether to use existing and emerging techniques developed by computer scientists for mining data (Hand 1998). Deceptive data-analysis could occur if researchers deliberately use an inappropriate statistical test or data-mining technique to boost support for a preferred conclusion, if they overstate the statistical significance of their results (Resnik 2000, Fan et al 2014), or if they leave out evidence or results that conflict with parts of the presented analysis.

Third, during the data-interpretation stage researchers make variousassumptions based on interpretive frameworks, analytical tools, concepts, and background theories (Elliott et al 2016, Longino 1990). Deception could occur if researchers overstate the scientific or practical significance of their findings or claim to have demonstrated causal connections when they have only produced evidence of statistical associations (Resnik 2000, Master and Resnik 2013).

Additionally, as with any form of scientific approach, failure to properly describe and disclose any of the steps related to data processing, analysis, or interpretation could entail deception and lead to irreproducibility, bias, or erroneous inference (Fan et al 2014).However, because it can be difficult to trace every step and decision in data-intensive science, adequate record-keeping procedures must be used. Discussions of best practices pertaining to transparency concerning methods, databases, computational algorithms, and supplementary material are ongoing (Fan et al 2014).

Clearly, deliberately concealing, obfuscating, or misusing methods used to process, analyze, or interpret data would be dishonest and unethical,but should such behavior be regarded as misconduct? The line between “misconduct” and “scientific disagreement” could be difficult to draw if we consider deceptive data processing, analysis, or interpretation to be a type of misconduct (Resnik and Stewart 2012). For example, in 1994 Charles Herrnstein and Thomas Murray published *The Bell Curve*, a highly controversial book that claimed to demonstrate statistical associations between race, intelligence, income, and educational achievement (Murray and Herrnstein 1994). Most of the book's critics claimed that the authors posited racist assumptions and drew racist conclusions, but some arguedthat the authors had applied statistical methods inappropriately (Devlin et al 1997). For example, Fischer et al (1996) reanalyzed Herrnstein and Murray's data and claimed that the effects of race were an artifact of how the authors weighted variables used in factor analysis and that income and other socioeconomic factors become more important than race when one uses a different weighting. Clearly, *The Bell Curve* raises important scientific and ethical questions pertaining to appropriate applications of statistical methods, but we do not think that the best way to deal with these issues is to treat them as misconduct allegations. The best way to resolve statistical issues like those raised by*The Bell Curve* is through critical reflection and scholarly debate, not through burdensome and costly legal procedures.

If deceptive uses of statistics which do not involve FFP were classified as misconduct, scientists could face misconduct allegations from critics or competitors who disagree with their methods, a situation which could have damaging consequences for scientific research. First, institutional officials, funding agencies, and journals could be inundated with misconduct allegations related to deceptive uses of statistics and might not have enough time or resources to deal with important cases related to FFP. Expanding the definition of misconduct to include deceptive uses of statistics could distract institutions, agencies, and journals from focusing on more important research integrity concerns.

Second, researchers might refrain from developing or implementing innovative and useful methods for processing, analyzing, or interpreting data to protect themselves againstmisconduct allegations. This concern may be especially important in data-intensive science, where methods and techniques continue to evolve in response to increases in the size, heterogeneity, and complexity of datasets (Fan et al 2014). The importance of promoting progress and innovation in data-intensive science supports the conclusion that deceptive data processing, analysis, or interpretation(which does not involve or lead to fabrication, falsification, or plagiarism)should not be classified as misconduct, at least for legal purposes (Resnik 2000). While these deceptive practices would qualify as "serious deviations," one could argue that government agencies, research institutions, and journals should not treat them as research misconduct to avoid creating a chilling effect on innovative uses of statistics.

However, as also noted above, some countries and institutions include the "serious deviations" category in their misconduct definitions and some include forms of deception other than fabrication or falsification. So, it would not be out of the question to treat deceptive data processing, analysis, or interpretation as misconduct. The key would be to propose a clear definition of deception related to processing, analyzing, or interpreting data which does not stifle progress and innovation.What might such a definition look like?

An analogy with deceptive manipulation of digital images may lend some insight into this question. Computer programs, such as Photoshop, have made it possible to manipulate digital images of proteins, cells, and other structures to enhance the clarity and quality of the image (Rossner and Yamada 2000). However, one can also use these programs to manipulate animage dishonestly and deceptively. For example, one could decrease the brightness or contrast of a gel image to make a band disappear, splice together gel images to make bands appear, or make copies of an image of a cell to make it appear that one has repeated the experiment (Rossner and Yamada 2000). Since 2000, the incidence of cases involving digital image manipulation handled by the ORI has increased dramatically. Prior to 2000, less than 5 cases per year that went to ORI involved image manipulation; after 2000, the ORI has dealt with approximately 25 such cases per year. From 1994 to 2000, the ORI made only 4 findings of misconduct related to image manipulation; from 2001 to 2007, it made 15 of these findings (Parrish and Noonan 2009).

Some journals have developed policies that distinguish between appropriate and inappropriate image manipulation (Bosch et al 2012). Many of these require authors to submit original images and include technical guidelines concerning gel electrophoresis, Western blotting, and microscopy (e.g. Nature Publishing Group 2016). The point of such policies is to ensure that the manipulated image published by the journal is an accurate representation of data (Rossner and Yamada 2004). As with the manipulation of images, it is likely to be difficult to distinguish appropriate and inappropriate forms of data management, but the formulation of explicit policies could provide greater clarity.

Integrity in Data-Intensive Science

Scientific journals, funding agencies, and professional societies could promote integrity in data-intensive science by developing guidelines concerning ethical data processing, analysis, and interpretation and promoting scholarly debate concerning appropriate uses of statistics.The goal of such policies should be to promote honesty, openness, objectivity, and transparency inthese scientific practices(Resnik 2000, Stodden et al 2016).For example, policies could require authors tofully describe and disclose methods used to acquire, edit, code, clean, select, audit, and analyze data; name computer programs used to process or analyze data and share programs or codesthat are not widely available; discuss and describe frameworks, theories, or other assumptions used to interpret data; and provide reviewers and the public with access to primary datathat is not considered proprietary or protected by confidentiality rules pertaining to human subjects (Soranno et al 2015, Stodden et al 2016).Many of these recommendations are included in guidelines developed by the Center for Open Science (2016). The American Statistical Association (2016) has developed guidelines for ethical statistical practice that could also inform policies concerning data manipulation. Journals and scholarly societies could play a supportive role both by developing and promulgating data-management policies and by providing venues for publishing datasets and code to make them publicly available along withappropriate supporting information (Elliott et al 2016).

Additional educationin ethical issues in data processing, analysis, and interpretation could also play an important role in promoting integrity in data-intensive science. Though many RCR courses include material on collecting, storing, analyzing, and sharing data, most do not include material on processing or interpreting data (Dubois and Duecker 2009).Research methods courses could also include material on ethical issues in processing,analyzing, and interpretingdata. Researchers could also provide students and trainees with mentoring on appropriate data practices when planning and implementing research projects and writing papers.It would also be valuable for education and mentoring programs toinclude training on workflow management/documentation and successful team science, which could help to promote more open and responsible management of large datasets.Similarly, collaborators with different areas of disciplinary expertise are often needed to engage in successful data-intensive science, which means that trainees should be prepared to work successfully and ethically in large interdisciplinary teams (Cheruvelil et al 2014).

While most people would agree that education and policy development concerning ethical data practices could help to promote integrity in data-intensive science, the more controversial issue remains: should funding agencies, journals, or research institutions develop a distinct category of deception related to data processing, analysis, or interpretation to include in the definition of research misconduct?Because we are concerned that a separate category of deceptive data practices could distract institutions, funding agencies, and journals from focusing on FFP cases, deter useful innovation in data-intensive science, and give the misleading impression that the improper handling of data isprevalent in data-intensive research, we do not favor development of such a category, especially if it has legal implications. We think the best way to move forward is to find ways to promote adherence to the acceptedethical and epistemological valuespresent in both traditional and data-intensive science and tackle the new challenges that can arise when working with very large datasets.Critical reflection and scholarly debate on best practices and transparency guidelines for data-intensive science is a crucial step in this direction (Sugimoto et al 2016).

Acknowledgments:

This research is funded, in part, by the Intramural Program of the National Institute of Environmental Health Science (NIEHS), National Institutes of Health (NIH); and the National Science Foundation grants EF-1065786 to PAS and SES-1449466 to KCE and PAS. It does not represent the view of the NIEHS, NIH, NSF, or U.S. government.We are grateful to Kendra Cheruvelil and Georgina Montgomery for helpful discussions.

References

American Statistical Association. (2016). Ethical guidelines for statistical practice. Available at:http://www.amstat.org/ASA/Your-Career/Ethical-Guidelines-for-Statistical-Practice.aspx. Accessed: November 8, 2016.

Australian Government. National Health and Medical Research Council. Australian Research Council. (2007). Australian Code for the Responsible. Conduct of Research. Available at: https://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/r39.pdf. Accessed: October 28, 2016.

Babbage C (1830). Reflections on the Decline of Science in England, and on Some of its Causes. Available at: http://www.gutenberg.org/files/1216/1216-h/1216-h.htm. Accessed: October 27, 2016.

Bell G, Hey T, Szalay A. (2009). Beyond the data deluge. Science 323(5919):1297–129819265007

Bosch X, Hernández C, Pericas JM, Doti P, Marušić A. (2012). Misconduct policies in high-impact biomedical journals. PLoS One 7(12):e51928.23284820

Broad W, Wade N. (1993). Betrayers of Truth, 2nd ed. New York: Simon and Schuster. Center for Open Science 2016.

The Transparency and Openness Promotion Guidelines.Available at: https://cos.io/top/. Accessed: December 14, 2016.

Cheruvelil KS, Soranno PA, Weathers KC, Hanson PC, Goring SJ, Filstrup CT, Read EK. (2014). Creating and maintaining high-performing collaborative research teams: the importance of diversity and interpersonal skills. Frontiers in Ecology and Environment 12(1):31–38.

Compare Business Products. 2016 Top 10 largest databases in the world. Available at:http://www.comparebusinessproducts.com/fyi/10-largest-databases-in-the-world. Accessed: November 8, 2016.

Coombes KR, Wang J, Baggerly KA. (2007). Microarrays: retracing steps. Nature Medicine 13(11): 1276–1277.

Devlin B, Feinberg SE, Resnick DP, Roeder K (eds.). (1997). Intelligence, Genes, and Success: Scientists Respond to The Bell Curve. New York: Springer-Verlag.

De Winter J, Kosolosky L. (2013). The epistemic integrity of scientific research. Science and Engineering Ethics 19(3):757–774.23054672

Dooley JJ, Kerch HM. (2000). Evolving research misconduct policies and their significance for physical scientists. Science and Engineering Ethics 6(1):109–121.11273428

Dubois JM, Dueker JM.(2009). Teaching and assessing the responsible conduct of research: a Delphi consensus panel report. Journal of Research Administration40(1):49–70.22500145

Duke University. (2007). Duke University policy and procedures governing research misconduct. http://provost.duke.edu/wp-content/uploads/FHB_App_P.pdf#page=33. Accessed: November 11, 2016.

Ekbia H, Mattioli M, Kouper I, Arave G, Ghazinejad A, Bowman T, Suri VR, Tsou A, Weingart S, Sugimoto C. (2015). Big data, bigger dilemmas: a critical review. Journal of the Association for Information Science and Technology 66(8):1523–1545.

Elliott KC. (2012). Epistemic and methodological iteration in scientific research. Studies in History and Philosophy of Science 43:376–382.

Elliott KC, Cheruvelil KS, Montgomery GM, Soranno PA. (2016). Conceptions of good science in our data-rich world. Bioscience [published online October 1, 2016].

Fan J, Han F, Liu H. (2014). Challenges of big data analysis. Natural Science Review 1: 293–314.

Fischer CS, Hout M, Jankowski MS, Lucas SR, Swidler A, Vos K. (1996). Inequality by Design: Cracking the Bell Curve Myth. Princeton, NJ: Princeton University Press.

GenBank. (2016). GenBank and WGS statistics. Available at: https://www.ncbi.nlm.nih.gov/genbank/statistics/. Accessed: November 6, 2016.

Goldberg P (2015). Duke officials silenced med student who Reported trouble in Anil Potti's lab. The Cancer Letter, January 9, 2015. Available at: http://cancerletter.com/articles/20150109_1/. Accessed: November 6, 2016.

Hand DJ. (1998). Data mining: statistics and more? The American Statistician 52(2):112–118.

Herrnstein R, Murray T. (1994). The Bell Curve: Intelligence and Class Structure in American Life. New York: Free Press.

Hey T, Tansley S, Tolle K. (2009). The Fourth Paradigm: Data-Intensive Scientific Discovery. Redmond, WA: Microsoft Research.

Hughes RIG. (1989). The Structure and Interpretation of Quantum Mechanics. Cambridge, MA: Harvard University Press.

Intergovernmental Panel on Climate Change. (2013). Climate Change 2013: The Physical Basis. Available at: http://www.ipcc.ch/report/ar5/wg1/. Accessed: October 12, 2016.

Kitchin R (2014). The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences. London: Sage.

Kline M (1982). Mathematics: The Loss of Certainty. Oxford, UK: Oxford University Press.

Kuhn TS. (1970). The Structure of Scientific Revolutions, 2nd ed. Chicago: University of Chicago Press.

Laudan L (1981). Science and Hypothesis: Historical Essays on Scientific Methodology. Dordrecht, Netherlands: Reidel.

Leonelli S (2014). What difference does quantity make? On the epistemology of Big Data in biology. Big Data & Society, Apr-Jun, 2014: 1–11.

Longino H (1990). Science as Social Knowledge. Princeton: Princeton University Press.

Macrina F (ed.). (2013). Scientific Integrity, 4th ed. Washington, DC: American Society for Microbiology Press.

Marx X (2013). Biology: the challenges of big data. Nature 498(7453):255–260.23765498

Master Z, Resnik DB.(2013). Hype and public trust in science.Scienceand Engineering Ethics19(2): 321–335.

Miller HJ. (2010). The data avalanche is here: should we be digging? Journal of Regional Science 50(1):181–201.

Muller-Wille S, Charmantier I. (2012). Natural history and information overload: the case of Linnaeus. Studies in the History and Philosophy of Biological and Biomedical Sciences43: 4–15.

National Academy of Sciences. (1992). Responsible Science: Ensuring the Integrity of the Research Process. Washington, DC: National Academy of Sciences.

National Genome Research Institute. (2015). Genome-wide association studies. Available at: https://www.genome.gov/20019523/. Accessed: October 27, 2016.

Nature Publishing Group. (2016). Image integrity and standards. Available at: http://www.nature.com/authors/policies/image.html. Accessed: November 8, 2016.

Office of Research Integrity. (2015). Case Summary: Potti, Anil. Available at: http://ori.hhs.gov/content/case-summary-potti-anil. Accessed: November 6, 2016.

Office of Science and Technology Policy. (2000). Federal misconduct policy. Federal Register 65(235): 76260–76264.

O’Malley M, Elliott KC, Burian R. (2010). From genetic to genomic regulation: Iterative methods in miRNA research. Studies in the History and Philosophy of Biological and Biomedical Sciences41:407–417.

O’Malley M, Elliott KC, Haufe C, Burian R. (2009). Philosophies of funding. Cell 138:611–615.19703386

Parrish D, Noonan B.(2009). Image manipulation as research misconduct. Science and Engineering Ethics15(2):161–167.19125357

Prensky MH (2009). Sapiens digital: from digital immigrants and digital natives to digital wisdom. Innovate 2009; 5 Available at:http://www.innovateonline.info/index.php?view=article&id=705. Accessed: December 14, 2016.

Popper K (1959). The Logic of Scientific Discovery. London: Routledge.

Potti A, Dressman HK, Bild A, Riedel RF, Chan G, Sayer R, Cragun J, Cottrill H, Kelley MJ, Petersen R, Harpole D, Marks J, Berchuck A, Ginsburg GS, Febbo P, Lancaster J, Nevins JR. (2006). Genomic signatures to guide the use of chemotherapeutics. Nature Medicine12(11):1294–1300.

PubMed. (2016). Home. Available at: https://www.ncbi.nlm.nih.gov/pubmed. Accessed: November 6, 2016.

Raman H, Ramos I (eds.). (2013). Ethical Data Mining Applications for Socio-Economic Development. Association for Computing Machinery Available at: http://www.igi-global.com/book/ethical-data-mining-applications-socio/73551?f=e-book. Accessed: October 31, 2016.

Resnik DB. (1998). The Ethics of Science. New York: Routledge.

Resnik DB. (2000). Statistics, ethics, and research: an agenda for education and reform. Accountability in Research 8(1): 163–188.

Resnik DB. (2003). From Baltimore to Bell Labs: reflections on two decades of debate about scientific misconduct. Accountability in Research 10(2):123–135.14577424

Resnik DB, Neal T, Raymond A, Kissling GE. (2015a). Research misconduct definitions adopted by U.S. research institutions. Accountability in Research22(1):14–21.25275621

Resnik DB, Rasmussen LM, Kissling GE. (2015b). An international study of research misconduct policies. Accountability in Research 22(5):249–66.25928177

Resnik DB, Wager E, Kissling GE. (2015c). Retraction policies of top scientific journals ranked by impact factor. Journal of Medical Librarian Association 103(3):136–139.

Resnik DB, Stewart CN. (2012). Misconduct versus honest error and scientific disagreement. Accountability in Research 19(1): 56–63.22268506

Rossner M, Yamada K. (2004). What’s in a picture? The temptation of image manipulation. The Journal of Cell Biology 166(1):11–15.15240566

Scott J (2013). Social Network Analysis. London, UK: Sage.

Shamoo AE, Resnik DB. (2015). Responsible Conduct of Research, 3rd ed. New York: Oxford University Press.

Soranno PA, Cheruvelil KS, Elliott KC, Montgomery G. (2015). It’s good to share: Why environmental scientists’ ethics are out of date. BioScience 65:69–73.26955073

Steadman I (2013). Big data and the death of the theorist. Wired, January 25, 2013. Available at:http://www.wired.co.uk/news/archive/2013-01/25/big-data-end-of-theory. Accessed: December 14, 2016.

Steneck NH. (1999). Confronting misconduct in science in the 1980s and 1990s: what has and has not been accomplished? Science and Engineering Ethics 5(2):161–176.11657853

Steneck NH. (2006). ORI Introduction to Responsible Conduct of Research. Washington, DC: Office of Research Integrity.

Stodden V, McNutt M, Bailey DH, Deelman E, Gil Y, Hanson B, Heroux MA, Ioannidis JP, Taufer M. (2016). Enhancing reproducibility for computational methods. Science 354(6317):1240–1241.27940837

Strasser BJ. (2012). Data-driven sciences: from wonder cabinets to electronic databases. Studies in the History and Philosophy of Biological and Biomedical Sciences 43:85–87.

Sugimoto CR, Ekbia HR, Mattioli M (eds.). (2016). Big Data is Not a Monolith. Cambridge, MA: MIT Press.

United Kingdom Research Council. (2012). The Research Ethics Guidebook. The RCUK (Research Council UK) Code of Conduct. Available at http://www.ethicsguidebook.ac.uk/Research-Council-funding-122. Accessed: October 28, 2016.

van Wel L, Royakkers L. (2004). Ethical issues in web data mining. Ethics and Information Technology 6(2):129–140.