Locating subjectivity in Data Science

Dakota Murray

Department of Informatics, Indiana University Bloomington, Bloomington, United States;
dakmurra@iu.edu

Abstract

The field of Data Science has grown to prominence and has impacted virtually every scientific domain. Part of the appeal of Data Science has been the mystique of objectivity that surrounds the field and its techniques. However, recent ethical and methodological scandals have begun to erode the allure of Data Science, exposing it an inherently human practice, just like all methodology. In this essay I aim to further dispel the mystique of objectivity that still surrounds Data Science by locating *subjectivity* throughout the entire Data Science Process. I demonstrate how the demographics of data science shape the kind of knowledge the field produces, and continue to demonstrate the subjectivity inherent in data collection, wrangling, and analysis. In doing so, I draw on literature from Critical Data Studies, Statistics, and Data Science itself. I argue that dispelling the mystique of objectivity is necessary to create a better and more mature Data Science. I close the essay by discussing five potential paths towards this improved science: diversity, methodological reflexivity, methodological improvements, openness, and a focus values.

1 Qualifying Exam Question #2

My original research research question is quoted below. In writing this essay, I tried to keep maintain the spirit of the original question. However, due to constraints over available time and length, I was forced to make several modifications from the original material. Moreover, as I gained a deeper understanding of the issues related to Data Science, I updated the essay to constitute an argument that forms more naturally out of literature.

> The recent explosion of Data Science and Big Data, and their growing importance in science, has lead to an increasing number of discussions on the limitations and implications of these methods. In technical spheres, these discussions tend to focus on epistemic issues, namely, what are the limits on what these tools allow us to know, and how valid is the knowledge they produce? However, in more social and humanistic fields, the discussion has instead centered on the ethical implications of applying these methods to real-world situations, with little discussion on their validity. There is little cross-talk between technical and socially-oriented scholars. However, in a world where scientific knowledge (and knowledge more generally) is used in ways that impact day-to-day life, validity can have direct ethical implications—knowledge built on an inproper premise can lead to failed technologies, misinformed governance, and bad decision-making. Data epistemology and ethics would benefit from a synthesis between these two areas of study. While there is no standardized procedure for data science, the general data analysis process can be divided into a series of stages, including question or problem formulation, data collection, pre-processing, inference, and "application"—the last being up to you to define. For each of these stages, Briefly summarize some of the tools, techniques, and procedures that constitute this stage
>
> 1. Summarize some of the important discussions relevant to this stage from both the technical and STS literature, with a particular focus on epistemic issues.
> 2. Summarize a case study in order to examine an epistemic issue in the production of scientific knowledge using data-intensive methods,
> 3. Your answer should include a short essay summarizing how technical, epistemic, and ethical issues interact broadly across data science as it relates to scientific knowledge.

In the version of the essay submitted here, I consider four major stages of the data analysis process, considering data analysis, data wrangling (cleaning/processing), data collection, and who gets to ask questions using data. Given my broad research interests, I chose to take a broad approach towards understanding each of these stages, drawing from a wide set of literature from different fields. I use these literature to discuss how *subjectivity* is part of the the Data Science process, and what different fields can tell us about this subjectivity.

I believe that this version of the essay easily satisfies the first and third stipulation in my question, below. In attempting to address the second stipulation, I found it difficult to identify in-depth case studies that deserved more than a sentence of mention in the essay; instead of relying on case studies, I ensured that many smaller examples were included throughout the essay; many of these examples, most of which are drawn form news headlines or issues I have encountered in my own work.

Also included in this essay is a section not stipulated in my original question: the final major section, *Paths Forward*. I came to include this section after discovering the great deal of literature discussing various kinds of improvements in Statistics and Data Science. This section is also in the spirit of my work—I don't seek to tear down quantitative methods or Data Science; rather, I hope that by acknowledging the issues, that a better methodology can be created. As such, a discussion on potential paths for improvement serves as a suitable conclusion for this essay.

Dakota Murray

Introduction

In his now (in)famous 2008 article in Wired Magazine, Chris Anderson claimed that Big Data would make scientific theory obsolete—correlation alone would suffice [1]. Similar evangelizing of Big Data and Data Science continued in many spaces since. Some have praised the transformative power of the so-called "Big Data Revolution" [2]. Others have spoken of the new "fourth paradigm" of scientific research in which knowledge is be gleaned directly from data, rather than from theory or experimentation [3]. Claims of revolutions and paradigms can feel overly Utopian or exaggerated; however, the real-world impacts of Big Data are hard to dismiss. From public sector data analytics to arcane and opaque Machine Learning algorithms, nearly every corner of society has been touched, in some way, by data collection, machine learning, statistics, or some other data-driven technology. Both within and outside of academia, "data" has become a metaphor that pervades discourse and frames conceptions of knowledge and evidence [4, 5]. The allure or big data builds upon the trust placed in the tools of quantification [6]; with the new "datification" [7], numbers are multiplied to bigger and more complex forms, which surely, it could be believed, makes them even more trustworthy. Viewed in terms of the "Law of Amplification" [8], the Big Data Revolution has brought with it technologies that amplify our abilities of dataficiation and knowledge production to dizzying new heights.

The discipline of Data Science has emerged to leverage the potential of Big Data. Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" [9] in order to understand phenomenon using data [10]. Data Science was imparted with the same mystique as Big Data, and became an industry and scientific buzzword after famously being termed the "sexiest job of the 21st century" [11]. The particular glamour of Data Science comes from its supposed potential to extract knowledge and novel insights directly from the data—data which is increasingly massive, heterogeneous, and messy. That this discipline has achieved status in the commercial sector, where consumer data has become a kind of capital [12] is not surprising; however, the techniques and tools of data science have also had a profound impact on scientific research. Data Science, like the field of Cybernetics[1] before it, has situated itself as a universal science thanks to its wide appeal and powerful metaphors [15]. Like Cybernetics, Data Science also promises to revolutionize every area of science. Whether in the natural sciences, social sciences, or humanities, the collection and use of data has become an increasingly important component of research [16]. Many disciplines have undergone major re-configurations to adapt to this new data-driven paradigm (e.g., [17, 18], resulting in new disciplinary titles such as *Digital Humanities* [19] and *Computational Social Science* [20]. "Data" has even become a powerful metaphor in areas of scholarship that have historically avoided the framing of quantification [5]. Whether the data is big or small, complex or simple, Data Science offers a new way that it can be understood.

Data science presents a range of opportunities for scientific research—it is hard to imagine modern scientific achievements like globally-integrated climate modelling [21] or the creation of the first ever image of a black hole [22] without the data science methodology and technologies that underpin them. However, just as Data Science can amplify the best qualities of scientific research, so too can it amplify limitations and ethical risks. These challenges are most apparent in applications of Data Science in the commercial and public sectors. For example, with Data Science and Big Data also came the so-called "Weapons of Math Destruction"—algorithmic systems which due to their scale, speed, and opaqueness, are responsible for exacerbating social ills [23]. These and similar systems have punished the poor [24], perpetuated harmful stereotypes and oppression [25], resulted in targeted policing [26, 27] and has lead to failures in public services [24, 28]; those that face the ill effects of these systems are disproportionately the underprivileged. The use of Data Science tools in scientific research has also resulted in ethical scandals. Facebook's attempt to measure emotional contagion by manipulating user's new feeds [29] stoked concerns over the place of human subjects, often unaware of their involvement, in Big Data research [30]. Research using Data Science has also violated ethical standards concerning the privacy and protection of research

[1] Cybernetics, like Data Science, exhibited broad appeal to many disciplines and was involved in acts of "legitimacy exchange" in which other disciplines would use Cybernetics methodology in order to benefit from its legitimacy; in turn, the use of Cybernetics methodology and metaphors by other disciplines would bolster the appeal and legitimacy of Cybernetics [13, 14]. It does not seem like a stretch to argue that Data Science is involved in similar forms of legitimacy exchange.

subjects. Examples include cases when researchers combined publicly available data to identify the pseudonymous artist Banksy [31], scraped many thousands of profiles from the dating site OKCupid [32], and used the images of unconsenting and vulnerable people to evaluates facial recognition algorithms [33]. Abuses of technology and ethical scandals in scientific research are cause for concern, but they are not new in and of themselves; what is new, however, is the scale and ubiquity of data-driven technologies that amplify the consequences of ethical risks to potentially millions of people.

Data Science has stoked not only ethical, but also epistemic and methodological dilemmas [34]. For example, Google Flu Trends, a Big Data tool attempting to predict Flu outbreaks using user's internet searches, demonstrated how "data hubris"—overconfidence in a result because of the size of the data—can lead to misleading results [35]. Bigger data doesn't necessarily mean better data [17, 18, 36]—sampling biases, noise, embedded assumptions, and confounding factors still impact data, no matter their size. Nor do correlations in bigger data automatically entail causal relationships [37]. For many problems, small and carefully-collected data might be better [16, 17]. However, even in statistical and experimental sciences operating with smaller data, the allure of easy statistical procedures, rather than sound statistical judgment, has contributed to the ongoing replication crisis [38, 39]. Despite these methodological limitations the allure of data persists. Polling during the 2016 presidential election demonstrated that even when data-driven predictions were wrong, the outcome could still be rationalized in the framework of the data—the allure of numbers endured [40]. Data Science offers clear methodological opportunities; the danger, however, stems from subtle, unclear, and unknown limitations. Despite claims to objectivity, Data Science has made clear the need for an understanding of the contingent nature of quantitative claims to knowledge.

In response to the ethical, methodological, and epistemic issues in Data Science, many fields have come forward offering new visions of the field. *Critical Data Studies* has emerged to address the need for a more thorough understanding of data. Critical Data Studies investigates the social, cultural, and political aspects of Big Data and Data Science [41]. Building on the rich theoretical frameworks developed by *Science and Technology Studies* and with extensive ethnographic fieldwork, Critical Data Studies has made strides in contextualizing data in terms of ethics, history, contemporary culture, and individual data practices (e.g., [42–44]). These scholars, however, are not the only ones discussing the implications of Data Science; other voices come from within the Data Science community itself. With a goal of establishing cross-disciplinary conversations about algorithmic issues, ACM's Conference on Fairness, Accountability, and Transparency has brought together a broad range of researchers. Within machine learning there has also been an increasing focus on enforcing mathematical definitions of fairness in classification (e.g., [45, 46]). In Statistics, conversations have centered on solutions to the ongoing replication crisis, advocating changes in the use of statistical significance [47] and encouraging reproducible research practices [48]. In other areas of Data Science, researchers have begun to grapple with the implicit ethical issues in behavioral Big Data research (e.g., [36, 49]). Each of these discussions contributes to an overall view of a better, more ethical, and more mature Data Science.

In this essay I aim to dispel the mystical quality of Data Science as a tool for generating objective knowledge; my intent is to locate the human factors, or the *subjectivity*, that permeate the practices of Data Science. Subjectivity, as I conceive of it here, is not meant as a diminutive term; I simply use it to refer to points of choice in which a data scientist must express personal judgment. My goal is not to tear down the authority of Data Science, but instead to demonstrate that Data Science, while powerful, is inherently a human practice. In doing so, my hope is that Data Science can mature as a discipline, and develop a path forward that leverages the field's potential while also mitigating its risks. To reflect the disciplinary heterogeneity of Data Science, I bring together voices from Critical Data Studies, Statistics, and Machine Learning, and even Data Science itself. These disciplines have long existed in separate silos with little chance for useful conversation; however, a better future for Data Science will require input from them all.

I organize this essay according to loosely-defined *stages* of data science. Partitioning the Data Science process into distinct stages suffers from two distinct difficulties. For one, the heterogeneity[2] of Data

[2]Many processes exist, perhaps two of the most prominent being CRISP-DM [50] and modifications of the Agile development methodology [51]. Crisp DM considers six major stages: Business understanding, Data understanding, Data

Science processes and the lack of widely accepted standards [53] means that any set of stages will fail to represent the work of of all, or likely even a majority, of practicing Data Scientists. The second difficulty is that stages of Data Science practice have no clear borders. Processes are also often iterative or cyclical, with researchers returning to earlier stages and updating past work. I acknowledge these limitations and choose to focus on only four broadly-defined stages that constitute major and relatively uncontroversial steps of the Data Science process. I first examine the kinds of people who can participate in data science and discuss how this might effect the kinds of research that is done; I center issues of gender, race, class, and geography. In the second section I investigate the choices made during the collection and reuse of data. The third section details the choices made during data wrangling, which consists of processing, cleaning, enriching, or otherwise preparing data for analysis. In the fourth section, I discuss the data analysis process itself, considering both subjectivity inherent to statistical methods and Big Data analysis more generally. Finally, I conclude by discussing several paths towards a better version of Data Science that are drawn from a wide range of disciplinary perspectives.

Who participates and who asks questions

Before analysis, the application of an algorithm, or even the processing of the data, the shape of Data Science is driven by the kinds of people who do the work. Feminist scholars like Donna Haraway were some of the first to note that different people, men and women for example, have different experiences and perspectives; therefore, a science and engineering dominated by men would lead to knowledge and technologies shaped by and suiting only men [54]. Data Science emerged primarily out of computing, a discipline that, especially after the introduction of personal computers, became associated with masculinity and has been dominated y men [55–57]. Whereas Data Science emerged out of computing, the techniques of the field have been widely adopted across science; however, in spite of ideals of universalism [58], scientific research as a whole remains dominated by men. Women continue to be underrepresented in high-impact scientific journals [59], on the editorial boards [60–64], and globally across scientific research [65]. The consequences of this underrepresentation have clear effects in the kind of knowledge being produced. For example, in medical research male authors are less likely than women to report on sex differences in medical trials [66], even though sex differences have been identified in the expression of illness [67] and drug responses [68, 69]. This speaks to a wider issue of scientific research persistently disregarding women or otherwise making wrong, misleading, and often harmful assumption [70]. Issues of gender participation in Science and Engineering are further complicated by *Technofeminism*—the perspective that gender and technology are mutually constructive, and that whereas gender dynamics shape technology and knowledge, so does technology shape gender [71]. This can be seen in household technologies which often freed men from household labour while reinforcing narrow gender roles for women [72]. In another example, an algorithmic technology for making automated hiring decisions was found to perpetuate sexist hiring practice that existed in historical data [73], a project that, fortunately, was abandoned. Without the participation of more women in Data Science, the kinds of knowledge it produces will necessarily be limited by the perspectives of only men.

The politics of data science extent beyond gender, also being influenced by race, class, and geography. Science and computing have been associated with a particular brand of white masculinity [56]; this is reflected in the underrepresentation of racial and ethic minorities among Science and Engineering degree holders in the workforce [74]. This racial disparity is likely one reason why ethnic minorities have often bore the brunt of ethical malfeasance in data science. For example, Safiya Noble [25] found that the design of search engine rankings can amplify harmful cultural stereotypes about Black, Asian, and Hispanic women, even though companies claim their rankings to be neutral. Related to this, many recent headlines have demonstrated examples of racist image classification [75] and failures of facial

preparation, Modelling, Evaluation, and Deployment, but highlighting the cyclic nature of these stages. Agile/Scrum methodology was first used in software development, and prioritizes iterative development and continuous delivery to a client. Both of these approaches are focused on business domains, and so do not readily map to scientific research. I do not use consider these methodologies in the present paper. Instead, the stages I consider were informed by the popular introductory text *Think Like a Data Scientist* which contains a more flexible and broadly-defined Data Science process [52].

recognition technology for Black [76] and east-Asian users [77,78]. In one recent and concerning study, researchers found that object detection of the kind used in self-driving cars had differential success rates between individuals with light and dark skin [79]. Similar issues persist when Data Science is used for the deployment of public services; for example, predictive policing promises improvements to the use of police resources, but often simply focuses resources on the policing of underprivileged and typically Black communities [26,27]; in other cases, when government assistance programs replace human employees with automated systems, it was often Black communities whom suffered [24]. Race is one of many determinants of who can participate in Data Science, and given the dominance of White data scientists, it is little surprise that the needs and experiences of ethnic minorities tend not not to be reflected in the data analysis process.

Socioeconomic class, often correlated with race, also dictates who is able to participate in data science. Just as men are more likely to have the skills and social networks necessary for data science work [17], socioeconomic status determines who is able to seek education, especially elite education. For example, those from low socioeconomic status are underrepresented among degree holders with a bachelors or higher [80]. Socioeconomic status has also been found to be associated with the destination for both graduate and postgraduate education [81], whereas prestige of doctoral education is associated with the access to elite professorships [82]. Advantages incurred by by socioeconomic privilege and academic prestige can snowball into further advantages, resulting in Matthew Effects that favor the already successful [83]. An ideal of academic research is that it is a meritocracy in which skill, not prestige, is the latent factor for researcher's success; however, this idea was recently contested in a recent study that found that work environment and access to resources, not raw talent or doctoral prestige, was associated with faculty productivity [84]. Education has long been one means by which social status is perpetuated, but in research this process also has the effect of shaping the kinds of questions that are asked. For example, recent studies have found that faculty hiring facilitates the spreading of ideas [85], and so a favoring of elite schools in faculty hiring [82] may amplify the ideas emerging from these elite institutions. The kinds of questions that are asked, and the kinds of research that are considered valuable, are in part determined by prestige and, in turn, socioeconomic status. Data Scientists and those using their methods will tend to be those coming from privileged socioeconomic backgrounds, and thus the perspectives of these people will shape the field.

Where one is born can determine a great deal about their life including access to food, education, public services, healthcare, and whether or not they will participate in knowledge production. At a basic level, geography determines what sorts of objects of research are available to ask questions about [86]; for example, Oceanography is unlikely to develop as a strong scientific discipline in a landlocked country. Given that a handful of wealthy and economically prosperous countries produce the vast majority of scientific publications [87,88], the kind of research that gets done is shaped by the needs and interests of these nations. For example, the "10/90 gap" describes how 90 percent of global medical research spending is targeted towards 10 percent of the world's population [89]; this leads to medical research that mostly benefits the peoples of the developed, "western" world [90]. In other cases, individuals from developing countries may struggle to contribute to science because they lack access to existing scientific literature which often is not open-access and sits behind paywalls requiring expensive library subscriptions [91]. Issues of geography are also influenced by issues of language. English has become the dominant language of scientific research [92]; non-native English speakers therefore face additional challenges when entering and contributing to global science [92–95]. The role of geography in determining participation in knowledge production is further complicated by its correlation with other factors such as race and class, correlation which is especially strong in countries like the United States and South Africa which have strong ethnic segregation. Those participating in Data Science will tend to come from only a handful of countries, and even then likely only a few regions within these countries; the questions that are asked and answered with the tools of Data Science will be skewed towards the needs and interests of these regions.

Data science, despite the ideal of constituting a "view from nowhere" [96], must be performed by someone in some place. The characteristics of the person doing Data Science are important because they can impact not only what questions are asked, but also the choices made throughout the Data

Science process. The view of Data Science, more often then not, comes from someone who is white, male, wealthy, and from a developed country. The domination of only a small number of perspectives in Data Science limits the potential of scientific research and can result in harms to the women, indigenous peoples, underprivileged groups, and those developing countries, all of whom have been excluded from the Data Science work. Locating the subjectivity in Data Science requires first understanding who gets to participate, whose eyes are on the data, and who is *not* involved in the analysis.

Data Collection

Data collection is inherently a human process, whether it be measurements of behavior in a psychology experiment, surveying a population about political beliefs, taking river samples for ecological research, or scraping information from websites. From the design of a data collection process to the choice of measurements and execution, choices have to be made, and so subjectivity becomes an implicit part of data collection. For example, in experimental study a researcher must design a study, choose instruments for measurement, decide how much data collect, and what subjects to exclude—all of which have the potential to impact the experiment's findings [97]. In the case of survey studies, researchers must choose a sampling strategy, decide how to write and present survey questions, and determine whether the survey will be administered by the researcher, or through services such as Amazon Mechanical Turk or Qualtrics. Perhaps the most important decision when designing a study is deciding *what* to measure and *how*. Measurement is the means of quantifying the world, and so the choice of measurement is fundamental to the kinds of knowledge that can be produced from a study. However, human judgment is necessary even in cases even when measurement is as seemingly straightforward as counting. For example, controversies erupted in early attempts to count the number of chromosomes in DNA; in spite of the ability to view samples of DNA under a microscope, samples were not obviously separated and so researchers would have exercise judgment about which chromosomes to include in a counting [91]. Another example of the issues of counting emerge from Bibliometrics—a discipline that seeks to study scientific research through publications, citations, and relations between them. A common topic in this field is the aggregation of citation and publication counts for individuals, institutions, and nations. This task at first seems straightforward, but actually requires judgment over how to distribute research credit in the case that publications have multiple co-authors (e.g., [98]), what time window to use when compiling citation counts (e.g., [99, 100]), and how to calculate aggregate indicators [101, 102]. Choices over what data to collect and how can have consequences throughout the entire research process. For example, recent experiments attempting to measure the ability of individual's "gaydar" in fact created a scenario so divorced from social context that measurements became meaningless, and the findings misleading [103]. Before Data Science or any kind of analysis can take place, data needs to be created. The creation of data requires choices that have consequences for the remainder of the Data Science process.

In many cases, a Data Scientist won't be involved in the collection of data, and instead will only happen upon a dataset after it is fully formed, created for some other purpose by some corporation, government, or research project[3]. This does not mean, however, that the judgment of the Data Scientist does not affect the data. Data do not speak for themselves, and so even when coming upon an already-collected dataset, the data scientist must interpret this data in light of their unique experiences and perspectives. Interpretation is necessary even in the most quantitative of disciplines. For example, some astronomical research requires calculations using historical records of solar eclipses; however, often ancient texts mentioning solar eclipses can be interpreted as either as historical record or poetic fiction, and researchers must make judgments about whether they include these records in their calculations or not [4]. Personal background can also influence how an individual sees and contextualizes data allowing its meaning and significance to be re-framed. Consider for example the history of Sarah and Angelina

[3]For example, the central role of consumer data in the "digital economy" [104] has transformed data into a form of capital, something to be collected for its own sake, from every source, and to the furthest extent [12]; through "datification", data has become a kind of currency that citizens pay in order to access corporate services [7]. Governments have also joined, releasing datasets on public repositories such as Data.gov and Data.gov.uk and Data.Taipei. Similarly, for scientific research there has been much discussion about the benefits of sharing data, making is available for replication and reuse [16, 105, 106].

Grimké, abolitionists operating within in the southern United States in the years before the civil war. These two women compiled thousands of local news stories including the descriptions of runaway slaves in order to paint a thorough and harrowing image of the scale and violence of slavery—a move that succeeded in shifting public opinion at a critical historical moment [4]. Similarly, the ability of individuals to re-frame existing data and direct it towards new purposes has been a hallmark of the Big Data movement. Examples of this re-framing include the reuse of social media data to measure political polarity [107] and to predict changes in the stock market [108]. What all of these examples demonstrate is that data never speak for themselves; even when a researcher comes to a dataset after its creation, they bring with them a set of assumptions that change how they think about and approach that data.

Data collection, like all of Data Science, is a subjective process. Data is never raw, instead it always comes "cooked" with theory and assumptions [4]. If collecting data, a data scientist will need to make choices about study design, what to measure, and how. If coming upon data only after its been collected, the data scientist still must interpret the data in light of their research goals, their background, and their unique perspective. Whether creating or reusing data, the data scientist must make use of their judgment and make choices, each of which may affect the findings of the project, and each of which might have been different.

Data Wrangling

Data almost never comes ready for analysis out of the box, instead it requires effort to be made fit for the analytical tools and goals of research. The term I use to describe this process is "data wrangling" [52], a popular term coined to refer to the variety or cleaning, pre-processing, enriching, and manipulations necessary prior to analysis. Data wrangling involves a series of choices about what processing steps are necessary and how these steps should be executed. Statisticians have described this flexibility in the data preparation phase as "researcher degrees of freedom" [97, 109]. Each researcher degree of freedom constitutes a choice made by a researcher, a choice which might seem straightforward but often one in which an alternative choice would have been equally justifiable. One example of a researcher degree of freedom is how to handle missing values; depending on the circumstances of their missingness, a data scientist might decide to exclude the observation with the missing value, code it with some default value, or use statistical imputation to replace it with a reasonable guess. Another common degree of freedom involves decisions over whether or not to normalize a skewed variable distribution in order to make them more amenable for statistical analysis. In other cases (e.g., Principal Components Analysis), variables may be normalized to occupy a similar range of values such that the effect of especially large values don't dominate the analysis analysis. Other times, variables are simplified to make analysis easier or more interpretable. For example, a continuous variable of yearly income might be discretized to a categorical variable of "high income" and "low income". Similarly, complex categorical variables may be aggregated or simplified: surveys of employment characteristics, for example, will often have large and complex sets of occupational codes which, depending on the research goal, may need to be simplified into categories like "STEM" vs "Not STEM", or "White Collar" vs. "Blue Collar". Each of these examples demonstrates a point of flexibility in the process of data wrangling, a researcher degree of freedom which might have been different if another data scientist performed the cleaning. These choices can have consequences for analysis and findings, so much so that some statisticians have argued that the ongoing replication crisis [110] can be attributed, in part, to variance in researcher's choices during the analysis process [111]. The analytical techniques of Data Science will rarely work on data out of the box—especially big and messy data. Instead, a data scientist will need to use their judgment to "wrangle" the data into an appropriate form fit for analysis; each choice introduces a new degree of freedom—a point of flexibility—into the analysis with the potential to impact the results.

Another common data wrangling task involves combining multiple datasets, a process that again requires the data scientist's judgment and can amplify researcher degrees of freedom. The ability to combine datasets from multiple choices has been lauded as one of the strengths of Big Data [2]. However, the process of merging multiple datasets is rarely, if ever, simple or straightforward; instead, bringing datasets requires work to overcome "data friction" [21] such that they can be "made commensurate" [4].

In some cases, this work seems straightforward. For example, if conducting an analysis of global economic indicators, datasets might need to be merged at the level of countries, and this will often require that idiosyncratic country names be made commensurate, including such actions as mapping "UK" and "United Kingdom", or "Vietnam" and "View Nam" to the same entities. However, in other cases this is less straightforward. For example, if one dataset provides indicators for "England", "Scotland", "Wales", and "Northern Ireland", whereas another includes only the "United Kingdom", then judgment is necessary to merge these data in a way that makes sense. Similarly, the status of countries can also shift over time; for example, economic data collected in 2008 would list the country "Sudan", however data collected in 2014 might list both "Sudan" and "South Sudan"[4], yet in order to make these datasets commensurate some mapping must be made. These examples demonstrate that there are situations when making datasets commensurate for which there is no clear choice, and for which many alternatives are equally justifiable. Sometimes a researcher may need to enforce or integrate some kind of "standard" [112] across datasets. For example, the American Community Survey is a popular source of survey data used in social sciences that employs a unique set of occupational codes; if performing an analysis of occupational salaries between countries, then these occupation codes may need to be mapped to others, a task that will likely be unclear and require further judgment. In other cases, data might need to be made commensurate at an even more fundamental or technical level. For example, integrating multiple corpora of images might require that they be made to share a common size or file format. This kind of data wrangling can persist even within the same project. For example, research into climate or ecological modelling will often require that the researchers account for variation and changes in sensory equipment or local and non-relevant changes to the environment around the sensor [4, 21, 113]. Just as data are rarely ready for analysis out of the box, they also are rarely ready to work with one another without significant labor. The choices made when combining datasets introduce additional flexibility into the Data Science process.

If data is always already "cooked", then it only becomes more so once the data scientist begins the work of data wrangling. When making data ready for analysis, choices have to be made such as what observations to keep, which to transform, and how data should be made to work together. These "researcher degrees of freedom" constitute points of flexibility in which other, often equally-justifiable, choices might have been made. The choices that any one data scientist makes during data wrangling will be informed by their experiences, assumptions, and motivations, all of which will differ from person to person. In this way, subjectivity continues to imbue Data Science.

Analysis—turning data into knowledge

Subjectivity continues to be a part of Data Science even until the stage of analysis. Choices over which analytical technique to apply, what covariates to include, and what significance level to use are additional researcher degrees of freedom that affect analysis [97, 109]. These analytical degrees of freedom, along with those introduced during data collection and wrangling, can have dramatic consequences on analysis. For example, a recent study crowd-sourced analysis of a dataset by providing the data to 29 Data Science teams and asking them the same research question; there was great variance in conclusions with teams reporting the presence positive, negative, *and* null relationships [114]. Researchers using statistical techniques might intentionally exploit these degrees of freedom in order to obtain statistically significant and thus publishable [110, 115] results, a practice that has been called *p-hacking* [116] and related to other similar practices such as *data dredging* [117] and *opportunistic bias* [118]. This exploitation has been posited as one explanation of the replication crisis [38], however others have argued that the problem stems not from exploitation, but instead from blind adherence to misunderstood statistical rituals [119]. For example, much scientific research has been defined by Null Hypothesis Statistical Testing—an approach to science that uses the *p-value* developed by Ronald Fisher in order to compare distributions of interest.

[4] South Sudan declared its independence in 2011. Another similar case would be how to calculate country metrics, such as population growth, following the annexation of Crimea in 2014, in which the national status of a large population is in question. In other cases, researchers may have to make choices about what to include as a country; for example, in some analyses, Hong Kong may be treated either as an independent state or as a city of China. Nation-level analyses present many cases in which researcher judgment is necessary.

The p-value was intended to measure the strength of evidence against some null hypothesis [120]. At the time, Fisher suggested a p-value threshold of 0.05[5] for rejecting the null hypothesis, but also argued that researchers should exercise judgment when interpreting statistical results. However, in the misguided pursuit of making scientific research more objective the 0.05 threshold became in a widely accepted binary rule for determining the presence of an effect. This development, many argue, has lead to a rash or poor statistical judgment and failed replications [119, 122]. Even in ideal situations, when a researcher is trained in statistics, sets a hypothesis and test condition beforehand, and refrains form p-hacking, the subtleties of data analysis can still result in subjective or contingent outputs, partially as a result of so-called "forking paths"[6] [123]. Sound statistical analysis requires sound judgment, though even in ideal cases, this judgment will vary between individual data scientists.

With Big Data, classical statistical inference was said to be obsolete [1]; instead, truth would emerge through correlation via the sheer force of millions or even billions or data points. However, recent developments to Big Data Analysis have demonstrated this to be false; truth does not naturally emerge from Big Data anymore than it does for small data. In fact, many aspects of Big Data make it *more* open to interpretation than for smaller data. Whereas a data scientist can interact with an excel spreadsheet in a way resembling physicality, Big Data presents a situation in which the data is too large to conceive of, let alone interact with in any physically meaningful way [124]. Instead, algorithmic tools must be used to process and work with and visualize these massive datasets. For example, clustering algorithms are a means of automatically grouping similar data points together in a way that can simplify the complexity of a dataset and identify interesting clusters of data points. However, many clustering algorithms have inherent flexibility. K-means, for example, is a popular clustering algorithm that is inherently stochastic, meaning that running the algorithm on the same data multiple times will likely identify different clusters of data. Moreover, k-means requires that the data scientist selects a value, k, which is the number of clusters to identify. Specific choices of k, while sometimes justifiable, are usually arbitrary and result from a process of trial and error and the data scientist's own interpretation. Visualization of Big Data also involves subjectivity. Whereas the structure of small datasets can be exhaustively explored, Big Datasets, often with millions of data points and highly complex structure, presents both technical and aesthetic challenges. Rather than plotting the data directly, *translation work* is necessary to map the complex and numerous data to some sort of visual component [125]; these visualizations, in order to shrink or simplify the data, will necessarily accentuate certain aspects while obfuscating others. This translation work is never neutral and will impact the way that the data is interpreted. In other cases, large and complex high-dimensional data can be *embedded* into lower-dimensional space, making it more amenable to analysis and visualization. However, depending on the structure of the data, there are many embedding algorithms to choose from, each of which require that the data scientist define some sort of measure of distance between data objects. Working with or even *looking at* Big Data is not a straightforward process—it requires instrumentation, translation work, and simplification. Even when Big Data is finally analyzed, it still does not automatically reveal truthful relationships. Recent works have shown that many correlations in Big Data are likely spurious, existing simply due to the data's size or by chance [126, 127]. Scientists have long had to contend with a confounding of correlation and causation, but the allure of Big Data can instill a sense of *data hubris*—overconfidence in results due to the sheer size of the data [35]. The very bigness of Big Data means that understanding it requires new approaches to analysis and visualization, approaches that often require researcher choice and judgment.

At no point in the process of Data Science is subjectivity absent—it pervades every decision and consideration from the choice of analytical tools to the interpretation of the results. Even under rigorous conditions, different people can come to different conclusions [114]. The analytical tools of Data Science present new ways of looking at and managing ever growing datasets, but they do not let the data speak for

[5] Ronald Fisher suggested using a 0.05 significance level, but others have recommended different levels, including 0.005 [121], but these levels are all to some degree arbitrary.

[6] This is the idea, credited to Gelman & Loken [123] demonstrating why issues of multiple comparisons can be a problem even when a hypothesis is set ahead of time and when no p-hacking is being done. The issue stems from the making of analytical decisions that depend on what is seen in the data—if the data were different, then the decision would likely to have also been different. These decisions, while usually justifiable, will often lead further down the path of statistical significance, even for spurious correlations.

themselves. The data scientist—a human with motivations, experiences, and a unique perspective,—will always be at the center of Data Science.

Paths Forward

So far I have argued that the objective mystique of Data Science is an illusion. A human, making human choices, lies at the center every stage of the Data Science process. The characteristics of this person including their gender, race, social class, country of origin, discipline, motivations, and more, all influence the choices they make while doing Data Science. That human subjectivity is central to Data Science dismantles the allure of objectivity that the discipline cultivates. As such, this essay may be viewed as an attack on Data Science and its claims to knowledge; however, this was not my intent. Subjectivity, despite the often-negative sentiment imparted on the term, is not a failing of Data Science. My intent was rather to argue that Data Science, like all disciplines and methodologies, is inherently a human endeavor in which human choice is central. Only after dismissing the mystical qualities of Data Science can the field mature into a more robust and ethically-conscious version of itself. In this section, I discuss five potential paths towards improvement for data science from a variety of disciplinary perspectives. These paths include a focus on diversity, methodological reflexively, development of improved methodology, a centering of openness, and a focus on values in research. Each of these paths acknowledges the inherent humanness of Data Science, but poses a distinct vision of its future. A new and better Data Science, one that leverages the opportunities of Big Data but also recognizes its subjectivity,will require progress towards all of these paths.

Diversity

One approach to improve Data Science focuses on making those doing the work more representative of the kinds of people their work impacts. Science, Engineering, and Computing have had a history of being white, male, and western [55, 56], a history that has passed on to Data Science. An analysis run by General Assembly—an online technical education company—found that women and racial minorities were underrepresented in their online data science courses [128]; these groups were also found to be underrepresented in Computer Science university programs [129] and science more generally [65, 74]. Many recent ethical scandals involving the application of Data Science have been attributed, in part, to the lack of diversity among those doing the analysis and designing the systems. By including people of different perspectives, experiences, and concerns into Data Science, the hope is that potential ethical and methodological issues will be considered early and remedied [130]. Several data-intensive organizations have already stated stances on improving diversity in data science, including Facebook [131], Microsoft [132], and the National Institute of Health [133]. In addition to gender and ethnic diversity, other forms of diversity have also been advocated for their potential for improving data science. For example, some have argued for methodological diversity in Data Science analyses in the hopes of making findings more robust [134]. Disciplinary diversity has also been advocated; for example, in the early days of Artificial Intelligence research Diana Forsythe [135] noted the challenges and opportunities of working collaboratively with engineers and computer scientists. More recently, Paul Dourish and Genevieve Bell echoed similar difficulties from their collaborations with Ubiquitous Computing researchers, though the also remark on the worry that some Computer Scientists have of being in service—becoming mere programmers—to other disciplines [136]. Scholars from Critical Data Studies have also reported insights and engagement when studying and working alongside practicing data scientists [137] and point towards a bright future of collaboration between the two fields [15, 42]. Diversity in team size also offers promise for improving Data Science as it has been found that large and small teams make distinct scientific contributions [138]. Conceptions of diversity are broad, each offering potential to making Data Science more equitable and more ethically and methodologically robust.

Calls for diversity, however, are not without critics; Google was the focus of two illustrative episodes. In 2017 the infamous "Damore" memo, written and circulated by an employee of Google, decried the company's inclusion initiatives that were aimed at assisting women [139]. In another case, Google

appointed to their new AI Ethics Board the president of the politically conservative "Heritage Foundation"; this move was met with public and internal backlash, quickly resulting in the dissolution of the ethics board [140]. At this political moment, "Diversity" is a polarizing term, and so attempts to implement diversity initiatives may be contentions; in other cases, discussions may center on what kinds of diversity are most desired. In spite of it sometimes being contested, diversity is one path that holds great potential for making Data Science better while also helping to heal historical injustice and bias.

Methodological Reflexivity

In ethnographic methodology, reflexivity refers to the acknowledgement that anything that is claimed or observed is "always a view from somewhere" [54], meaning that claims to knowledge are always situated within the social, political, and cultural circumstances of the person or group making the claim. Recent scholars in Critical Data Studies have suggested that data scientists approach their work in ways bearing striking similarities to ethnographic study [141]. For example, Elish & Boyd [142] found that data scientists would continuously interpret and re-interpret model outputs over the course of labelling, cleaning, and modelling, and each change to the model's outcome would lead the data scientist to rationalize new findings in the context of an internally-consistent worldview; this process, the authors argue, is akin to ethnographic study but without the corresponding reflexivity. Reflexivity encourages data scientists to consider their decisions in the context of their background and particular social circumstances, focusing not on eliminating biases but in recognizing them. Some scholars have found that data scientists often are already reflexive in their practice though with their own unique farmings. For example, one ethnographic study observed how researchers using Data Science methodology proceeded through a process of breakdown and repair; during periods of "breakdown", researchers were forced to engage with the origins and contexts of their data and in the process learn something new [124]. In another study, data scientist were found to critically engage with many aspects of their work, though often through the lens of efficiency [137]. In another recent study, data scientists were encouraged to engage with criticisms of their field, and were found to already be aware of many of these criticisms and considered them when making decisions [43]. Other disciplines offer alternative forms of reflexivity that may benefit Data Science; *Scientometrics* [98], *Metaknowledge* [143], and the *Science of Science* [144] each use quantitative methodology in order to gain macro-scale understandings of scientific research; approaches from these fields have already been used to understand the extent of bias in science [145] and the epistemic consequences of digital publishing [146].

Methodological reflexivity can take many forms, but all rest on a reflective engagement with the origins and consequences of one's own work. However, reflexivity requires the development of new disciplinary norms and practices, a development that may take years of effort by the Data Science community. Moreover, there is some evidence that scientists being open about their values can have at best mixed effects on public perceptions of their credibility [147]. As a path forward, reflexivity offers a sense of personal awareness that centers cultural and social concerns and can evolve according to discipline-specific needs, but it requires caution when engaging with the larger public and long-term and concerted effort to develop new norms and practices.

Methodological Improvements

Whereas methodological reflexivity suggests embracing subjectivity, quantitative researchers have instead advocated for additional protections against subjectivity through methodological and technical means. In statistics, for example, the impacts of researcher degrees of freedom might be mitigated by instituting more standardization in data processing, analysis, and reporting [97] and by implementing study pre-registration [148]. Others proposals have called for more rigorous statistical and methodological training for researchers so that they better understand and interpret statistical outcomes [48, 119]. Yet other statisticians have instead called for a turn away from a frequentest paradigm of statistics in favor of a Bayesian paradigm [38, 149]. Some instead advocate for abandoning the idea of statistical significance all together [150], or at least re-aligning its use [122, 151]. Others have instead suggested that researchers put less weight in p-values and focus instead on effect size [152]. One interesting proposal even advocated for

rotating scatter plots to make it harder to mistakenly infer causation [153]. Machine Learning researchers have also begun to consider mathematical means of mitigating bias and promoting fairness in classification systems (e.g., [154–156]).

Many of these methodological and technical changes offers clear and potentially useful methodological interventions, however none of them, in and of themselves, are sufficient to remove the subjectivity and ethical issues from Data Science. In Machine Learning, for example, mathematical solutions to bias fail to capture its social realities [157]. These, and other methodological fixes, could become "packaged intervention" [8]—politically expedient solutions that will often fail when transferred to a new context due in part to the "portability trap" [158]. An example of a packaged intervention can be found in calls to re-define statistical significance from a p-value threshold of 0.05 to one of 0.005 [121]; this intervention may help in some contexts, but will likely prove less effective when transferred adopted by disciplines that use large datasets and for which even negligible effects can appear highly significant [152, 159]. All of these proposals, while useful, will not remedy issues stemming from poor publication practices [115] or other biases across science [145]. Technical and methodological improvements are key to building a better Data Science; however, these improvements are by themselves insufficient. Technical solutions have to be coupled with new incentives, norms, and values.

Openness

Making the research process more open and transparent offers another path forward for Data Science, and one which builds on fundamental scientific ideals. Making data publicly available, for example, has the potential to improve the reproducibility of findings [106] and to facilitate the reuse and re-purposing data for further scientific discovery [105]. Policies for open data also have been shown to have strong support from the broad academic community [160–162]. Open code [163] and algorithmic transparency [164] have been proposed as ways of making scientific analyses as well as systems for automated decision making more transparent [23]. More extensive reporting or release of code for data processing and analysis also have the potential to reveal researcher degrees of freedom and make re-analysis and replication possible [48, 109, 165, 166]. An open research culture has the potential to improve awareness, reliability [167, 168].

Openness, while bringing benefits, also comes with challenges. For example, authors may not automatically comply to policies of open data [169–171] or open publications [172]. There are also practical challenges to openness. For example, making data and methods available requires resources, effort, and technical expertise [173] which is rarely formally incentivized [105, 174]. Moreover, making data open may put researchers at risks of having their research ideas "scooped" or losing control of their data [162]. There are also many practical and ethical limitations to transparency and accountability [175]. For example, even anonymous social data can cause harm to individuals [36, 49] or be re-identified using other public data (e.g., [31, 32, 176]. Transparency is also hindered by the very nature of algorithms used in data science, which often "black boxes" [177] that are unknowable even to their creators[7]. Openness is further complicated by issues surrounding informed consent. For example, a participant in a study may give permission for data to be used by the researcher, but may not imagine the potential for their data to be freely circulated and used for arbitrary purposes; in some cases, data sharing without informed consent can perpetuate histories of exploitation of underprivileged communities [180]. Open Data policies can also be weaponized by commercial and political interests in order to make research and decision-making on certain topics, such as climate change, more difficult [181]. Data scientists could learn from the open science movement by making their data and analysis publicly available and transparent; however, just as Data Science can amplify the benefits of open science, so too can it amplify the risks of openness.

[7]The common example are neural networks, algorithms that can be trained to classify data by adjusting internal numeric weights—in many cases thousands or millions of values—in response to training data. These weights are often difficult to interpret in any meaningful way. A consequence of this is that researchers are often unable to determine why an algorithms makes decisions. In this case of self-driving cars this has grown especially concerning because the reasons for the deaths of drivers and pedestrians are often unexplainable (e.g., [178, 179]). Eubanks [24] and O'Neil [23] explore other instance in which computational decisions have had deleterious effects in public and commercial decision making.

Values

Locating the source of an ethical failure in a Data Science project is incredibly difficult [173]. The practices of Data Science are highly heterogeneous , involving geographically distributed teams and stakeholders, often with no standardized work process [53]; because of this heterogeneity, generalizing ethical advice form any one project is difficult. The ethics of Data Science are further complicated by the tendency for Data Science teams to be international, distributed, and embedded within complex networks of people and institutions. The technical complexity of Data Science also mean that projects can have dire yet unintended consequences when applied to new contexts, even when the data scientists themselves were cautious and considerate [8, 23, 158]. Institutional Review Boards have existed to enforce ethical standards in research, but they are usually woefully unprepared for data-intensive research, especially that using behavioral big data [49] Given these difficulties, it can be difficult for a Data Scientists to know what the "ethical" choice is in any situation.

Duty-based ethical frameworks have emerged with the goal of providing strict ethical principals; these principals provide a set of guidelines for Data Scientists to follow in order to be ethical in their practice, even under conditions of uncertainty. For example, Cathy O'Neil's *Weapons of Math Destruction* [23] argued for a sort of "Hippocratic Oath" for data scientists. This oath would center on the potential harms of Data Science, avoiding known biases of technical systems, and would caution against the release of automated decision-making systems into social settings. This sentiment was echoed by Virginia Eubanks in *Automated Inequality* [182] who spoke about the efforts of the Data for Good Exchange to create such a code of ethics[8]. There are some signals that calls for increased ethics in Data Science are being heard and acted upon. For example, Kate Crawford, a prominent scholar at the forefront of ethical AI research, was the keynote speaker for NeurIPS[9] 2017 (The Artificial Intelligence Channel, 2017), one of the largest and most prestigious Machine Learning conferences. In another example, the Association for Computing Machinery has run, since 2018, the conference on Fairness, Accountability, and Transparency which seeks to address such issues as ethics in AI and Big Data. There have also been growing efforts to teach ethics to new Data Science students. MIT, for instance, plans to include computing ethics as a key component of the educational curriculum in their new *Stephen Schwarzman College of Computing* [184]. In other cases, calls for increased ethics come from within the Data Science community itself; recent publications to PLoS One [185] and PLoS Biology [36] outline concerns and possible concerns to issues in responsible and ethical Behavior Big Data research. Many people and institutions are approaching ethics from different angles; time will tell whether these result in a more mature, reflective, and ethically-concious data science process.

Conclusion

In this essay, I have attempted to dispel the mystique of objectivity that surrounds the practice of Data Science by arguing that subjectivity is an inherent part of its practice, as it is in all scientific methodologies. Every stage of the Data Science process requires that choices be made. Each choice constitutes a set of equally-justifiable "forking paths" which the scientists must choose between. The paths chosen by one data scientist may lead to a vastly different finding than would the paths chosen by another (e.g., [114]). Data scientists make their choices based on their personal background, their disciplinary culture, the infrastructure and tools they have available, the ethical norms of their institution, and seemingly inconsequential choices made in the heat of the moment. In nearly all cases, what appears as an objective or clear choice could have been different.

Just as data analysis has many paths, there are also many paths towards a better and more mature Data Science. In this essay I have outlined five of such paths. One involves making the practitioners of Data Science more representative of the people they impact with their analyses and automated

[8] Reporting on these efforts can be found on the blog of Virginia Eubanks herself [182] and was also reported on by Wired Magazine [183].

[9] This conference was originally titles the Conference on Neural Information Processing Systems and referred to using the acronym "NIPS"; however, due to this acronym's sexist connotations, the conference was re-branded in 2019 with the acronym "NeurIPS". I consider this a positive development, and so I use the new acronym here.

systems. Another path advocates for methodological reflexivity, a construct borrowed from ethnographic methodology that encourages the data scientist to consider how their own background and biases might impact their results. Some advocate to instead protect against bias by introducing methodological improvements and interventions to the Data Science practice. Openness promises to make data, analysis and algorithms more transparent in order to open them up to critique, verification, and replication. Finally, increased ethical training and the establishment of formal professional principals seek to center the social, cultural, and ethical implications of data science and to provide guidelines for ethical practice. Improving Data Science will likely require effort and progress to be made in each of these paths forward along with other progress along paths not explored here. By acknowledging the human factor in Data Science, rather than ignoring it, the community as a whole can work towards better, more mature, and methodologically robust version of Data Science.

Acknowledgments

I would like to thank my qualifying exam committee for their patience and understanding as I worked through my ideas for this proposal. I would also like to thank the Department of Informants and the many amazing professors and colleagues who have had an impact on my development as a research. I would especially like to thank Dr. Cassidy Sugimoto for her wonderful mentorship as I have progressed through my doctoral program.

References

1. Anderson C. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Wired. 2008;.
2. Mayer-Schönberger V, Cukier K. Big Data: A Revolution That Will Transform How We Live, Work, and Think. Reprint edition ed. Boston: Eamon Dolan/Mariner Books; 2014.
3. Hey T, Tansley S, Tolle K. The Fourth Paradigm: Data-Intensive Scientific Discovery. 1st ed. Redmond, Washington: Microsoft Research; 2009.
4. Gitelman L, editor. "Raw Data" Is an Oxymoron. Cambridge, Massachusetts ; London, England: The MIT Press; 2013.
5. Markham AN. Undermining 'data': A critical examination of a core term in scientific inquiry. First Monday. 2013;18(10). doi:10.5210/fm.v18i10.4868.
6. Porter TM. Trust in Numbers. Reprint edition ed. Princeton, N.J: Princeton University Press; 1996.
7. van Dijck J. Datafiction, dataism and dataveillance: Big Data between scientific paradigm and secular belief. Surveillance & Society. 2014;12.
8. Toyama K. Geek Heresy: Rescuing Social Change from the Cult of Technology. F first edition edition ed. New York: PublicAffairs; 2015.
9. Hayashi C, Yajima K, Bock HH, Ohsumi N, Tanaka Y, Baba Y, editors. Data Science, Classification, and Related Methods: Proceedings of the Fifth Conference of the International Federation of Classification Societies (IFCS-96), Kobe, Japan, March 27–30, 1996. Studies in Classification, Data Analysis, and Knowledge Organization. Springer Japan; 1998. Available from: https://www.springer.com/us/book/9784431702085.
10. Dhar V. Data Science and Prediction; 2013. Available from: https://cacm.acm.org/magazines/2013/12/169933-data-science-and-prediction/abstract.

11. Davenport TH, Patil DJ. Data Scientist: The Sexiest Job of the 21st Century. Harvard Business Review. 2012;(October 2012).

12. Sadowski J. When data is capital: Datafication, accumulation, and extraction. Big Data & Society. 2019;6(1):2053951718820549. doi:10.1177/2053951718820549.

13. Bowker G. How to Be Universal: Some Cybernetic Strategies, 1943-70. Social Studies of Science. 1993;23(1):107–127.

14. kline RE. The Cybernetics Moment; 2014. Available from: `https://jhupbooks.press.jhu.edu/content/cybernetics-moment`.

15. Ribes D. STS, Meet Data Science, Once Again. Science, Technology, & Human Values. 2018; p. 0162243918798899. doi:10.1177/0162243918798899.

16. Borgman CL. Big Data, Little Data, No Data: Scholarship in the Networked World. Cambridge, Massachusetts: The MIT Press; 2015.

17. boyd d, Crawford K. Critical Questions for Big Data. Information, Communication & Society. 2012;15(5):662–679. doi:10.1080/1369118X.2012.678878.

18. Kitchin R. Big Data, new epistemologies and paradigm shifts. Big Data & Society. 2014;1(1):2053951714528481. doi:10.1177/2053951714528481.

19. Burdick A, Drucker J, Lunenfeld P, Presner T, Schnapp J. Digital_Humanities. The MIT Press; 2016.

20. Salganik M. Bit by Bit: Social Research in the Digital Age. Princeton: Princeton University Press; 2017.

21. Edwards PN. A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming. Cambridge, Massachusetts London, England: MIT Press; 2013.

22. Overbye D. Darkness Visible, Finally: Astronomers Capture First Ever Image of a Black Hole. The New York Times. 2019;.

23. O'Neil C. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. 1st ed. New York: Crown; 2016.

24. Eubanks V. Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. New York, NY: St. Martin's Press; 2018.

25. Noble S. Algorithms of Oppression: How Search Engines Reinforce Racism. 1st ed. New York: NYU Press; 2018.

26. Lum K, Isaac W. To predict and serve? Significance. 2016;13(5):14–19. doi:10.1111/j.1740-9713.2016.00960.x.

27. Richardson R, Schultz J, Crawford K. Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice. Rochester, NY: Social Science Research Network; 2019. ID 3333423. Available from: `https://papers.ssrn.com/abstract=3333423`.

28. Chouldechova A, Benavides-Prado D, Fialko O, Vaithianathan R. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In: Conference on Fairness, Accountability and Transparency; 2018. p. 134–148. Available from: `http://proceedings.mlr.press/v81/chouldechova18a.html`.

29. Kramer ADI, Guillory JE, Hancock JT. Experimental evidence of massive-scale emotional contagion through social networks. Proceedings of the National Academy of Sciences. 2014;111(24):8788–8790. doi:10.1073/pnas.1320040111.

30. Chambers C. Facebook Fiasco: Was Cornell University's study of 'emotional contagion' a breach of ethics? The Guardian. 2014;.

31. Hauge MV, Stevenson MD, Rossmo DK, Comber SCL. Tagging Banksy: using geographic profiling to investigate a modern art mystery. Journal of Spatial Science. 2016;61(1):185–190. doi:10.1080/14498596.2016.1138246.

32. Cox J. 70,000 OkCupid Users Just Had Their Data Published; 2016. Available from: `https://motherboard.vice.com/en_us/article/8q88nx/70000-okcupid-users-just-had-their-data-published`.

33. Wernimont JOK Nikki Stevens. The Government Uses Images of Abused Children and Dead People to Test Facial Recognition Tech. Slate Magazine. 2019;.

34. Ekbia H, Mattioli M, Kouper I, Arave G, Ghazinejad A, Bowman T, et al. Big data, bigger dilemmas: A critical review. arXiv:150900909 [cs]. 2015;.

35. Lazer D, Kennedy R, King G, Vespignani A. The Parable of Google Flu: Traps in Big Data Analysis. Science. 2014;343(6176):1203–1205. doi:10.1126/science.1248506.

36. Zook M, Barocas S, Boyd D, Crawford K, Keller E, Gangadharan SP, et al. Ten simple rules for responsible big data research. PLOS Computational Biology. 2017;13(3):e1005399. doi:10.1371/journal.pcbi.1005399.

37. Canali S. Big Data, epistemology and causality: Knowledge in and knowledge out in EXPOsOMICS. Big Data & Society. 2016;3(2):2053951716669530. doi:10.1177/2053951716669530.

38. Colling LJ, Szűcs D. Statistical Inference and the Replication Crisis. Review of Philosophy and Psychology. 2018;doi:10.1007/s13164-018-0421-4.

39. Baker M. 1,500 scientists lift the lid on reproducibility. Nature News. 2016;533(7604):452. doi:10.1038/533452a.

40. Loukissas Y, Pollock A. After Big Data Failed: The Enduring Allure of Numbers in the Wake of the 2016 US Election. Engaging Science, Technology, and Society. 2017;3(0):16–20. doi:10.17351/ests2017.150.

41. Iliadis A, Russo F. Critical data studies: An introduction. Big Data & Society. 2016;3(2):2053951716674238. doi:10.1177/2053951716674238.

42. Neff G, Tanweer A, Fiore-Gartland B, Osburn L. Critique and Contribute: A Practice-Based Framework for Improving Critical Data Studies and Data Science. Big Data. 2017;5(2):85–97. doi:10.1089/big.2016.0050.

43. Moats D, Seaver N. "You Social Scientists Love Mind Games": Experimenting in the "divide" between data science and critical algorithm studies. Big Data & Society. 2019;6(1):2053951719833404. doi:10.1177/2053951719833404.

44. Metcalf J, Crawford K. Where are human subjects in Big Data research? The emerging ethics divide. Big Data & Society. 2016;3(1):2053951716650211. doi:10.1177/2053951716650211.

45. Kamiran F, Calders T. Data Preprocessing Techniques for Classification Without Discrimination. Knowl Inf Syst. 2012;33(1):1–33. doi:10.1007/s10115-011-0463-8.

46. Menon AK, Williamson RC. The cost of fairness in binary classification. In: Conference on Fairness, Accountability and Transparency; 2018. p. 107–118. Available from: http://proceedings.mlr.press/v81/menon18a.html.

47. McShane BB, Gal D. Blinding Us to the Obvious? The Effect of Statistical Training on the Evaluation of Evidence. Management Science. 2015;62(6):1707–1718. doi:10.1287/mnsc.2015.2212.

48. Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Percie du Sert N, et al. A manifesto for reproducible science. Nature Human Behaviour. 2017;1(1):0021. doi:10.1038/s41562-016-0021.

49. Shmueli G. Research Dilemmas with Behavioral Big Data. Big Data. 2017;5(2):98–119. doi:10.1089/big.2016.0043.

50. Shearer C. The CRISP-DM model: the new blueprint for data mining. J Data Warehouse. 2000;5:13–22.

51. Grady NW, Payne JA, Parker H. Agile big data analytics: AnalyticsOps for data science. In: 2017 IEEE International Conference on Big Data (Big Data). Boston, MA: IEEE; 2017. p. 2331–2339. Available from: http://ieeexplore.ieee.org/document/8258187/.

52. Godsey B. Think Like a Data Scientist: Tackle the data science process step-by-step. 1st ed. Shelter Island: Manning Publications; 2017.

53. Saltz J, Hotz N, Wild D, Stirling K. Exploring Project Management Methodologies Used Within Data Science Teams. AMCIS 2018 Proceedings. 2018;.

54. Haraway D. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. Feminist Studies. 1988;14(3):575–599. doi:10.2307/3178066.

55. Oldenziel R. Making Technology Masculine: Men, Women, and Modern Machines in America, 1870-1945. 1st ed. Amsterdam: Amsterdam University Press; 2004.

56. Turner F. From Counterculture to Cyberculture: Stewart Brand, the Whole Earth Network, and the Rise of Digital Utopianism. 60265th ed. Chicago: University of Chicago Press; 2008.

57. Ensmenger NL. The Computer Boys Take Over: Computers, Programmers, and the Politics of Technical Expertise. Aspray W, editor. Cambridge, Mass.: The MIT Press; 2012.

58. Merton RK. The Sociology of Science: Theoretical and Empirical Investigations. Storer NW, editor. Chicago: University of Chicago Press; 1979.

59. Shen YA, Webster JM, Shoda Y, Fine I. Persistent Underrepresentation of Women's Science in High Profile Journals. bioRxiv. 2018; p. 275362. doi:10.1101/275362.

60. Addis E, Villa P. The Editorial Boards of Italian Economics Journals: Women, Gender, and Social Networking. Feminist Economics. 2003;9(1):75–91. doi:10.1080/1354570032000057062.

61. Amrein K, Langmann A, Fahrleitner-Pammer A, Pieber TR, Zollner-Schwetz I. Women Underrepresented on Editorial Boards of 60 Major Medical Journals. Gender Medicine. 2011;8(6):378–387. doi:10.1016/j.genm.2011.10.007.

62. Cho AH, Johnson SA, Schuman CE, Adler JM, Gonzalez O, Graves SJ, et al. Women are underrepresented on the editorial boards of journals in environmental biology and natural resource management. PeerJ. 2014;2:e542. doi:10.7717/peerj.542.

63. Metz I, Harzing AW. Gender Diversity in Editorial Boards of Management Journals. Academy of Management Learning & Education. 2009;8(4):540–557. doi:10.5465/amle.8.4.zqr540.

64. Espin J, Palmas S, Carrasco-Rueda F, Riemer K, Allen PE, Berkebile N, et al. A persistent lack of international representation on editorial boards in environmental biology. PLOS Biology. 2017;15(12):e2002760. doi:10.1371/journal.pbio.2002760.

65. Larivière V, Ni C, Gingras Y, Cronin B, Sugimoto CR. Bibliometrics: Global gender disparities in science. Nature News. 2013;504(7479):211. doi:10.1038/504211a.

66. Sugimoto CR, Ahn YY, Smith E, Macaluso B, Larivière V. Factors affecting sex-related reporting in medical research: a cross-disciplinary bibliometric analysis. The Lancet. 2019;393(10171):550–559. doi:10.1016/S0140-6736(18)32995-7.

67. McCombe P, Greer J, Mackay I. Sexual Dimorphism in Autoimmune Disease. Current Molecular Medicine;9(9):1058–1079.

68. Franconi F, Brunelleschi S, Steardo L, Cuomo V. Gender differences in drug responses. Pharmacological Research. 2007;55(2):81–95. doi:10.1016/j.phrs.2006.11.001.

69. Jochmann N, Stangl K, Garbe E, Baumann G, Stangl V. Female-specific aspects in the pharmacotherapy of chronic cardiovascular diseases. European Heart Journal. 2005;26(16):1585–1595. doi:10.1093/eurheartj/ehi397.

70. Saini A. Inferior: How Science Got Women Wrong-and the New Research That's Rewriting the Story. Boston: Beacon Press; 2017.

71. Wajcman J. TechnoFeminism. 1st ed. Cambridge ; Malden, MA: Polity; 2004.

72. Cowan RS. More Work For Mother: The Ironies Of Household Technology From The Open Hearth To The Microwave. 0002nd ed. New York: Basic Books; 1985.

73. Dastin J. Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. 2018;.

74. Women, Minorities, and Persons with Disabilities in Science and Engineering: 2017. Arlington, VA: National Science Foundation, National Center for Science and Engineering Statistics; 2017. 17-310. Available from: `https://www.nsf.gov/statistics/2017/nsf17310/citation.cfm`.

75. Guynn J. Google Photos labeled black people 'gorillas'. USA TODAY. 2015;.

76. Breland A. How white engineers built racist code – and why it's dangerous for black people. The Guardian. 2017;.

77. Lee O. Camera Misses the Mark on Racial Sensitivity; 2009. Available from: `https://gizmodo.com/camera-misses-the-mark-on-racial-sensitivity-5256650`.

78. Zhao C. Is the iPhone X's facial recognition racist? Newsweek. 2017;.

79. Wilson B, Hoffman J, Morgenstern J. Predictive Inequity in Object Detection. arXiv:190211097 [cs, stat]. 2019;.

80. Musu-Gillette L. NCES Blog | Educational attainment differences by students' socioeconomic status; 2015. Available from: `https://nces.ed.gov/blogs/nces/post/educational-attainment-differences-by-students-socioeconomic-status`.

81. Jerrim J. Family Background and Access to "High Status" Universities. Sutton Trust; 2013.

82. Clauset A, Arbesman S, Larremore DB. Systematic inequality and hierarchy in faculty hiring networks. Science Advances. 2015;1(1):e1400005. doi:10.1126/sciadv.1400005.

83. Merton RK. The Matthew Effect in Science. Science. 1968;159(3810):56–63. doi:10.1126/science.159.3810.56.

84. Way SF, Morgan AC, Larremore DB, Clauset A. Productivity, prominence, and the effects of academic environment. Proceedings of the National Academy of Sciences. 2019; p. 201817431. doi:10.1073/pnas.1817431116.

85. Morgan AC, Economou DJ, Way SF, Clauset A. Prestige drives epistemic inequality in the diffusion of scientific ideas. EPJ Data Science. 2018;7(1):40. doi:10.1140/epjds/s13688-018-0166-4.

86. Livingstone DN. Putting Science in Its Place: Geographies of Scientific Knowledge. Chicago, Ill.: University of Chicago Press; 2003.

87. May RM. The Scientific Wealth of Nations. Science. 1997;275(5301):793–796. doi:10.1126/science.275.5301.793.

88. King DA. The scientific impact of nations. Nature. 2004;430:311.

89. Miranda JJ, Zaman J. "Exporting Failure": Why Research from Rich Countries may not Benefit the Developing World. Revista de saude publica. 2010;44(1):185–189.

90. Evans JA, Shim JM, Ioannidis JPA. Attention to Local Health Burden and the Global Disparity of Health Research. PLoS ONE. 2014;9(4). doi:10.1371/journal.pone.0090147.

91. Martin A, Lynch M. Counting Things and People: The Practices and Politics of Counting. Social Problems. 2009;56(2):243–266. doi:10.1525/sp.2009.56.2.243.

92. Gordin MD. Scientific Babel: How Science Was Done Before and After Global English. 1st ed. Chicago ; London: University of Chicago Press; 2015.

93. Duszak A, Lewkowicz J. Publishing academic texts in English: A Polish perspective. Journal of English for Academic Purposes. 2008;7(2):108–120. doi:10.1016/j.jeap.2008.03.001.

94. Flowerdew J. Scholarly writers who use English as an Additional Language: What can Goffman's "Stigma" tell us? Journal of English for Academic Purposes. 2008;7(2):77–86. doi:10.1016/j.jeap.2008.03.002.

95. Salager-Meyer F. Scientific publishing in developing countries: Challenges for the future. Journal of English for Academic Purposes. 2008;7(2):121–132. doi:10.1016/j.jeap.2008.03.009.

96. van House NA. Science and technology studies and information studies. Annual Review of Information Science and Technology. 2004;38(1):1–86. doi:10.1002/aris.1440380102.

97. Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science. 2011;22(11):1359–1366. doi:10.1177/0956797611417632.

98. Leydesdorff L, Milojević S. Scientometrics. 2012;.

99. Campanario JM. Providing impact: The distribution of JCR journals according to references they contribute to the 2-year and 5-year journal impact factors. Journal of Informetrics. 2015;9(2):398–407. doi:10.1016/j.joi.2015.01.005.

100. Nierop Ev. The introduction of the 5-year impact factor: does it benefit statistics journals? Statistica Neerlandica. 2010;64(1):71–76. doi:10.1111/j.1467-9574.2009.00448.x.

101. Lariviere V, Kiermer V, MacCallum CJ, McNutt M, Patterson M, Pulverer B, et al. A simple proposal for the publication of journal citation distributions. bioRxiv. 2016; p. 062109. doi:10.1101/062109.

102. Thelwall M. The precision of the arithmetic mean, geometric mean and percentiles for citation data: An experimental simulation modelling approach. Journal of Informetrics. 2016;10(1):110–123. doi:10.1016/j.joi.2015.12.001.

103. Gelman A, Greggor M, Simpson D. Gaydar and the Fallacy of Decontextualized Measurement. Sociological Science. 2018;5:270–280. doi:10.15195/v5.a12.

104. Poon M. Corporate Capitalism and the Growing Power of Big Data: Review Essay. Science, Technology, & Human Values. 2016;41(6):1088–1108. doi:10.1177/0162243916650491.

105. Borgman CL. The conundrum of sharing research data. Journal of the American Society for Information Science and Technology. 2012;63(6):1059–1078. doi:10.1002/asi.22634.

106. Gewin V. Data sharing: An open mind on open data. Nature. 2016;529(7584):117–119. doi:10.1038/nj7584-117a.

107. Conover MD, Ratkiewicz J, Francisco M, Goncalves B, Menczer F, Flammini A. Political Polarization on Twitter. In: Fifth International AAAI Conference on Weblogs and Social Media; 2011.Available from: `https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2847`.

108. Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market. Journal of Computational Science. 2011;2(1):1–8. doi:10.1016/j.jocs.2010.12.007.

109. Wicherts JM, Veldkamp CLS, Augusteijn HEM, Bakker M, van Aert RCM, van Assen MALM. Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. Frontiers in Psychology. 2016;7. doi:10.3389/fpsyg.2016.01832.

110. Ioannidis JPA. Why Most Published Research Findings Are False. PLoS Medicine. 2005;2(8). doi:10.1371/journal.pmed.0020124.

111. Carter EC, McCullough ME. Publication bias and the limited strength model of self-control: has the evidence for ego depletion been overestimated? Frontiers in Psychology. 2014;5. doi:10.3389/fpsyg.2014.00823.

112. Bowker G, Star SL. Sorting Things Out: Classification and Its Consequences. Revised edition ed. Cambridge, Massachusetts London, England: The MIT Press; 2000.

113. Ribes D. Notes on the Concept of Data Interoperability: Cases from an Ecology of AIDS Research Infrastructures. In: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. CSCW '17. New York, NY, USA: ACM; 2017. p. 1514–1526. Available from: `http://doi.acm.org/10.1145/2998181.2998344`.

114. Silberzahn R, Uhlmann EL, Martin DP, Anselmi P, Aust F, Awtrey E, et al. Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. Advances in Methods and Practices in Psychological Science. 2018;1(3):337–356. doi:10.1177/2515245917747646.

115. Young NS, Ioannidis JPA, Al-Ubaydli O. Why Current Publication Practices May Distort Science. PLoS Medicine. 2008;5(10). doi:10.1371/journal.pmed.0050201.

116. Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. The Extent and Consequences of P-Hacking in Science. PLoS Biology. 2015;13(3). doi:10.1371/journal.pbio.1002106.

117. Smith GD, Ebrahim S. Data dredging, bias, or confounding. BMJ : British Medical Journal. 2002;325(7378):1437–1438.

118. DeCoster J, Sparks EA, Sparks JC, Sparks GG, Sparks CW. Opportunistic biases: Their origins, effects, and an integrated solution. The American Psychologist. 2015;70(6):499–514. doi:10.1037/a0039191.

119. Gigerenzer G. Statistical Rituals: The Replication Delusion and How We Got There. Advances in Methods and Practices in Psychological Science. 2018;1(2):198–218. doi:10.1177/2515245918771329.

120. Sterne JAC, Smith GD. Sifting the evidence—what's wrong with significance tests? BMJ : British Medical Journal. 2001;322(7280):226–231.

121. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, et al. Redefine statistical significance. Nature Human Behaviour. 2018;2(1):6. doi:10.1038/s41562-017-0189-z.

122. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature. 2019;567(7748):305. doi:10.1038/d41586-019-00857-9.

123. Gelman A, Loken E. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "[U+FB01]shing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. 2013; p. 17.

124. Tanweer A, Fiore-Gartland B, Aragon C. Impediment to insight to innovation: understanding data assemblages through the breakdown–repair process. Information, Communication & Society. 2016;19(6):736–752. doi:10.1080/1369118X.2016.1153125.

125. Hohl M. From abstract to actual: art and designer[U+2010]like enquiries into data visualisation. Kybernetes. 2011;40(7/8):1038–1044. doi:10.1108/03684921111160278.

126. Taleb NN. Beware the Big Errors of 'Big Data'. Wired. 2013;.

127. Calude CS, Longo G. The Deluge of Spurious Correlations in Big Data. Foundations of Science. 2017;22(3):595–612. doi:10.1007/s10699-016-9489-4.

128. The Study of Data Science Lags in Gender and Racial Representation -. General Assembly; 2017. Available from: https://generalassemb.ly/blog/data-science-gender-race-disparity/.

129. Diversity Gaps in Computer Science:Exploring the Underrepresentation of Girls, Blacks and Hispanics. Gallup; 2016. Available from: http://services.google.com/fh/files/misc/diversity-gaps-in-computer-science-report.pdf.

130. Marr B. The Gender Diversity Crisis In Artificial Intelligence And Data Science – And How To Tackle It; 2019. Available from: https://www.linkedin.com/pulse/gender-diversity-crisis-artificial-intelligence-data-science-marr.

131. Hofleitner A. The value of diversity in data science research; 2017. Available from: https://research.fb.com/the-value-of-diversity-in-data-science-research/.

132. Hoffman J, Rao J. Diversity in data science: Microsoft Research's summer school aims high; 2015. Available from: https://www.microsoft.com/en-us/research/blog/diversity-in-data-science-microsoft-researchs-summer-school-aims-high/.

133. Valentine HA. Data Science: Meet Diversity | SWD at NIH; 2017. Available from: /blog/2017-05-26-data-science-meet-diversity.

134. Mann RP, Woolley-Meza O. Maintaining intellectual diversity in data science. Data Science. 2017;1(1-2):85–94. doi:10.3233/DS-170003.

135. Forsythe DE, Hess DJ. Studying Those Who Study Us: An Anthropologist in the World of Artificial Intelligence. 1st ed. Stanford, Calif: Stanford University Press; 2002.

136. Dourish P, Bell G. Divining a Digital Future: Mess and Mythology in Ubiquitous Computing. Reprint edition ed. The MIT Press; 2014.

137. Lowrie I. Algorithmic rationality: Epistemology and efficiency in the data sciences. Big Data & Society. 2017;4(1):2053951717700925. doi:10.1177/2053951717700925.

138. Wu L, Wang D, Evans JA. Large teams develop and small teams disrupt science and technology. Nature. 2019;566(7744):378. doi:10.1038/s41586-019-0941-9.

139. Wakabayashi D. Contentious Memo Strikes Nerve Inside Google and Out. The New York Times. 2018;.

140. Statt N. Google dissolves AI ethics board just one week after forming it. The Verge. 2019;.

141. Seaver N. Bastard Algebra. In: Data: Now Bigger and Better! The University of Chicago Press; 2015.Available from: https://www.press.uchicago.edu/ucp/books/book/distributed/D/bo20285526.html.

142. Elish MC, Boyd D. Situating Methods in the Magic of Big Data and Artificial Intelligence. Rochester, NY: Social Science Research Network; 2017. ID 3040201. Available from: https://papers.ssrn.com/abstract=3040201.

143. Evans JA, Foster JG. Metaknowledge. Science. 2011;331(6018):721–725. doi:10.1126/science.1201765.

144. Fortunato S, Bergstrom CT, Börner K, Evans JA, Helbing D, Milojević S, et al. Science of science. Science. 2018;359(6379):eaao0185. doi:10.1126/science.aao0185.

145. Fanelli D, Costas R, Ioannidis JPA. Meta-assessment of bias in science. Proceedings of the National Academy of Sciences of the United States of America. 2017;114(14):3714–3719. doi:10.1073/pnas.1618569114.

146. Evans JA. Electronic Publication and the Narrowing of Science and Scholarship. Science. 2008;321(5887):395–399. doi:10.1126/science.1150473.

147. Elliott KC, McCright AM, Allen S, Dietz T. Values in environmental research: Citizens' views of scientists who acknowledge values. PLOS ONE. 2017;12(10):e0186049. doi:10.1371/journal.pone.0186049.

148. Pain E. Register your study as a new publication option. Science | AAAS. 2015;doi:doi:10.1126/science.caredit.a1500282.

149. Gelman A, Shalizi CR. Philosophy and the practice of Bayesian statistics. The British journal of mathematical and statistical psychology. 2013;66(1):8–38. doi:10.1111/j.2044-8317.2011.02037.x.

150. McShane BB, Gal D, Gelman A, Robert C, Tackett JL. Abandon Statistical Significance. arXiv:170907588 [stat]. 2017;.

151. Gelman A, Stern H. The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant. The American Statistician. 2006;60(4):328–331. doi:10.1198/000313006X152649.

152. Sullivan GM, Feinn R. Using Effect Size—or Why the P Value Is Not Enough. Journal of Graduate Medical Education. 2012;4(3):279–282. doi:10.4300/JGME-D-12-00156.1.

153. Bergstrom CT, West JD. Why scatter plots suggest causality, and what we can do about it. 2018;.

154. Drosou M, Jagadish Hv, Pitoura E, Stoyanovich J. Diversity in Big Data: A Review. Big Data. 2017;5(2):73–84. doi:10.1089/big.2016.0054.

155. Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S. Certifying and Removing Disparate Impact. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '15. New York, NY, USA: ACM; 2015. p. 259–268. Available from: http://doi.acm.org/10.1145/2783258.2783311.

156. Nguyen L. A Proposal of Discovering User Interest by Support Vector Machine and Decision Tree on Document Classification. In: Proceedings of the 2009 International Conference on Computational Science and Engineering - Volume 04. CSE '09. Washington, DC, USA: IEEE Computer Society; 2009. p. 809–814. Available from: https://doi.org/10.1109/CSE.2009.112.

157. Binns R. Fairness in Machine Learning: Lessons from Political Philosophy. In: Conference on Fairness, Accountability and Transparency; 2018. p. 149–159. Available from: http://proceedings.mlr.press/v81/binns18a.html.

158. Selbst AD, Boyd D, Friedler S, Venkatasubramanian S, Vertesi J. Fairness and Abstraction in Sociotechnical Systems. Rochester, NY: Social Science Research Network; 2018. ID 3265913. Available from: https://papers.ssrn.com/abstract=3265913.

159. Khalilzadeh J, Tasci ADA. Large sample size, significance level, and the effect size: Solutions to perils of using big data for academic research. Tourism Management. 2017;62:89–96. doi:10.1016/j.tourman.2017.03.026.

160. Federer LM, Lu YL, Joubert DJ, Welsh J, Brandys B. Biomedical Data Sharing and Reuse: Attitudes and Practices of Clinical and Scientific Research Staff. PLOS ONE. 2015;10(6):e0129506. doi:10.1371/journal.pone.0129506.

161. Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, Read E, et al. Data Sharing by Scientists: Practices and Perceptions. PLOS ONE. 2011;6(6):e21101. doi:10.1371/journal.pone.0021101.

162. Tenopir C, Dalton ED, Allard S, Frame M, Pjesivac I, Birch B, et al. Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. PLOS ONE. 2015;10(8):e0134826. doi:10.1371/journal.pone.0134826.

163. Ince DC, Hatton L, Graham-Cumming J. The case for open computer programs. Nature. 2012;482(7386):485–488. doi:10.1038/nature10836.

164. Diakopoulos N, Koliska M. Algorithmic Transparency in the News Media. Digital Journalism. 2017;5(7):809–828. doi:10.1080/21670811.2016.1208053.

165. Peng RD. Reproducible Research in Computational Science. Science. 2011;334(6060):1226–1227. doi:10.1126/science.1213847.

166. Randles BM, Pasquetto IV, Golshan MS, Borgman CL. Using the Jupyter notebook as a tool for open science: an empirical study. IEEE Press; 2017. p. 338–339. Available from: http://dl.acm.org/citation.cfm?id=3200334.3200401.

167. Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, et al. Promoting an open research culture. Science. 2015;348(6242):1422–1425. doi:10.1126/science.aab2374.

168. Miguel E, Camerer C, Casey K, Cohen J, Esterling KM, Gerber A, et al. Promoting Transparency in Social Science Research. Science. 2014;343(6166):30–31. doi:10.1126/science.1245317.

169. Piwowar HA, Chapman WW. Public sharing of research datasets: a pilot study of associations. Journal of informetrics. 2010;4(2):148–156. doi:10.1016/j.joi.2009.11.010.

170. Federer LM, Belter CW, Joubert DJ, Livinski A, Lu YL, Snyders LN, et al. Data sharing in PLOS ONE: An analysis of Data Availability Statements. PLOS ONE. 2018;13(5):e0194768. doi:10.1371/journal.pone.0194768.

171. Savage CJ, Vickers AJ. Empirical Study of Data Sharing by Authors Publishing in PLoS Journals. PLOS ONE. 2009;4(9):e7078. doi:10.1371/journal.pone.0007078.

172. Larivière V, Sugimoto CR. Do authors comply when funders enforce open access to research? Nature. 2018;562(7728):483. doi:10.1038/d41586-018-07101-w.

173. Leonelli S. What difference does quantity make? On the epistemology of Big Data in biology. Big Data & Society. 2014;1(1):2053951714534395. doi:10.1177/2053951714534395.

174. Wallis JC, Rolando E, Borgman CL. If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. PLOS ONE. 2013;8(7):e67332. doi:10.1371/journal.pone.0067332.

175. Ananny M, Crawford K. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society. 2018;20(3):973–989. doi:10.1177/1461444816676645.

176. Pandurangan V. On Taxis and Rainbows; 2014. Available from: `https://tech.vijayp.ca/of-taxis-and-rainbows-f6bc289679a1`.

177. Latour B. Science in Action: How to Follow Scientists and Engineers Through Society. Reprint edition ed. Cambridge, Mass: Harvard University Press; 1988.

178. Knight W. Machine learning is making self-driving cars smarter, but it can also make their workings more mysterious. MIT Technology Review. 2016;.

179. Wakabayashi D. Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam. The New York Times. 2018;.

180. Radin J. "Digital Natives": How Medical and Indigenous Histories Matter for Big Data. Osiris. 2017;32(1):43–64. doi:10.1086/693853.

181. Levy KE, Johns DM. When open data is a Trojan Horse: The weaponization of transparency in science and governance. Big Data & Society. 2016;3(1):2053951715621568. doi:10.1177/2053951715621568.

182. Eubanks V. A Hippocratic Oath for Data Science; 2018. Available from: `https://virginia-eubanks.com/2018/02/21/a-hippocratic-oath-for-data-science/`.

183. Simonite T. Should Data Scientists Adhere to a Hippocratic Oath? Wired. 2018;.

184. Communications MS. Ethics, computing, and AI: Perspectives from MIT. MIT News. 2019;.

185. Buchanan E. Considering the ethics of big data research: A case of Twitter and ISIS/ISIL. PLOS ONE. 2017;12(12):e0187155. doi:10.1371/journal.pone.0187155.