Voluntas
https://doi.org/10.1007/s11266-022-00479-2

RESEARCH PAPERS

The Wells-Du Bois Protocol for Machine Learning Bias: Building Critical Quantitative Foundations for Third Sector Scholarship

Thema Monroe-White[1] · **Jesse Lecy**[2]

Accepted: 2 March 2022
© International Society for Third-Sector Research 2022

Abstract This paper charts the rapid rise of data science methodologies in manuscripts published in top journals for third sector scholarship, indicating their growing importance to research in the field. We draw on critical quantitative theory (QuantCrit) to challenge the assumed neutrality of data science insights that are especially prone to misrepresentation and unbalanced treatment of subgroups (i.e., those marginalized and minoritized because of their race, gender, etc.). We summarize a set of challenges that result in biases within machine learning methods that are increasingly deployed in scientific inquiry. As a means of proactively addressing these concerns, we introduce the "Wells-Du Bois Protocol," a tool that scholars can use to determine if their research achieves a baseline level of bias mitigation. Ultimately, this work aims to facilitate the diffusion of key insights from the field of QuantCrit by showing how new computational methodologies can be improved by coupling quantitative work with humanistic and reflexive approaches to inquiry. The protocol ultimately aims to help safeguard third sector scholarship from systematic biases that can be introduced through the adoption of machine learning methods.

Keywords Algorithmic bias · Critical quantitative methods · Machine learning · Data science · Third sector

✉ Thema Monroe-White
tmonroewhite@berry.edu

Jesse Lecy
jdlecy@asu.edu

[1] Department of Technology, Entrepreneurship, and Data Analytics, Campbell School of Business, Berry College, 2277 Martha Berry Highway, Mount Berry, GA 30149-5024, USA

[2] Arizona State University, Tempe, USA

Introduction

Third sector scholarship embraces core values of diversity, equity, and inclusion. It is not surprising, then, that methodological pluralism is viewed as an appropriate extension of these core values. This approach has virtues: mixed method triangulation, the combination of statistical and qualitative approaches to social inquiry, has been touted as the new gold standard in social scientific research (Battaglio & Hall, 2018).

Pluralism also has costs. The universe of research methodologies has expanded over time, forcing journal editors and reviewers to gain familiarity with a larger set of tools to assess the quality of research during the peer review process. Fields like sociology, political science, and economics that have historically relied on traditional social science research methods are increasingly dominated by "sexy" new methods like machine learning (ML)—computational approaches to prediction or classification that learn from data and can improve with experience (Ayodele, 2010). Although many of these techniques are new to academic fields (Conte et al., 2012; Davenport & Patil, 2012), third sector scholars may increasingly experience pressures to embrace these techniques to maintain relevance (Hardwick et al., 2015).

The current movement to capitalize on the digital revolution appears to be distinct from previous academic phases of methodological diversification because it does not consist of paradigms diffusing across disciplines like game theory spilling over into economics or econometrics into political science. Rather, it has been driven largely by events outside academics—previously unimaginable amounts of data produced by online transactions, mobile devices, and social network platforms. These data capture human behavior in novel ways unprecedented for their

Published online: 05 April 2022

breadth, depth, and scale. Data availability in turn has created demand for computational tools that can mine these massive datasets for insight. Open-source software has democratized access to sophisticated analytical tools, and collaborative platforms have nurtured global communities of data specialists. This confluence of data, software, computing power and human expertise has catalyzed a rapid growth in new computational methods of inquiry.

The emerging field of data science—the systematic collection, management, analysis, visualization, explanation, and preservation of structured and unstructured data (Marshall & Geier, 2019)—sits at the intersection of social science, computer science, statistics, and information systems (Lazer, 2009). It leverages machine learning capabilities to extract new and exciting insights from plethoric data reservoirs, thus expanding the methodological toolkit available for social inquiry (Lazer, 2009; Chang et al., 2014). These new powers, however, are accompanied by a host of challenges and responsibilities that will be important to the third sector scholarship.

This work calls attention to the potential dangers associated with the *uncritical* adoption of data science approaches in third sector research by articulating a critical theory perspective on these trends in the popularity of new research methodologies and their potential impact on the field. We begin by cataloging scholarship that utilizes data science approaches in third sector research. We then introduce the critical quantitative theoretical framework, discussing how it can prepare scholars for the new wave of data-driven research. Finally, we propose the Wells-Du Bois protocol, the equivalent of a "Bechdel Test" for research produced with data science methodologies, to determine whether studies meet a minimal threshold of bias mitigation.

Theoretical Justification

Empirical Uses of Data Science

Big data methodologies are exciting because of the three V's: large amounts of data (volume) of varied types (variety) generated at a faster pace (velocity) than has previously been possible (LaValle et al., 2011), offering a rich canvas for the study of human behavior. Data science methodologies include a large and diverse collection of computational models used to identify associations (i.e., co-occurrence of items), descriptive analysis of dense data (i.e., visualizations), prediction (i.e., regression models, support vector machines), classifications (i.e., decision trees and neural networks), and clusters (i.e., k-means), which in turn can generate novel patterns and insights. These approaches can be applied to structured data (i.e., numeric spreadsheets) or unstructured data (i.e., text, video, audio files) furthering their applicability across contexts, disciplines, and data types.

The broad availability of open data and falling costs of proprietary data, the expansion of storage options and processing power, and the growth of powerful and free open-source software platforms has democratized the field of data science. The ability to leverage these resources to generate new insight is limited by two primary factors: expertise needed to transform, link, and wrangle existing data into a database that serves a specific purpose and analytical expertise. Data science continues to gain momentum across academic disciplines including economics, psychology, education across the humanities, sciences, and in medicine as technical capacity increases (Liao et al., 2018; Mullainathan & Spiess, 2017; Treaster, 2017).

An integral feature of third sector research is a community of scholars armed with theoretical tools and social systems expertise that allows them to identify and deconstruct inequalities and injustices that other fields may miss or ignore. With the rise of these now powerful research tools, the field is grappling with how to embrace methods that can enhance the rigor, scale, and ambition of existing research without compromising the values, diversity, and integrity of current scholarship.

The current crisis of reproducibility offers lessons about ways in which new methods can inadvertently distort research. Publication bias became a problem when inferential statistics became widely adopted in the social sciences. Methods that were supposed to make social science more objective made research less robust in fields that emphasized statistical significance over effect size. This misapplication of inferential techniques allowed editors to favor studies with splashy results instead of those that followed sound scientific approaches but produced banal findings that failed to overturn existing paradigms. As a result, some fields are facing a severe credibility crisis because the most impactful research published in top journals cannot be replicated (Baker, 2016). Data science offers a similarly seductive sirens call, inviting scholars to deploy new tools that promise insight, but only if they can navigate the perilous obstacles that accompany the journey.

The Rise of Data Science in Third Sector Scholarship

Bibliometric analysis is helpful for identifying trends in academic fields (Ellegaard & Wallin, 2015). It is deployed here to demonstrate a steady rise in data science methodologies and discourse in third sector scholarship. Evidence for this claim comes from publications taken from a convenience sample of high-impact third sector research over a

Table 1 Ten years of Computational Social Science Studies in Top Third Sector Research journals

Voluntas, NVSQ and PAR	2011–2012	2013–2014	2015–2016	2017–2018	2019–2020	Theoretical	Empirical
Categories						$n = 12$	$n = 13$
Artificial intelligence	–	–	–	–	6	4	2
Big data	–	–	1	3	2	3	3
Data mining	–	1	1	2	3	5	1
Data science	–	–	1	2	1	3	1
Machine learning	–	–	–	4	5	4	5
Text mining	–	–	1	4	6	4	7

Search term categories were collapsed for ease of interpretation

ten-year period to capture scholarship with a target audience of non-profit or non-governmental researchers and professionals. We deploy an exploratory analysis similar to Hodgkinson and Painter's (2003) examination of changes in the disciplinary content and breadth in *Voluntas* and ISTR conference proceedings from 1990 to 2002 to examine growth in the mention and use of data science methodologies over time. These trends demonstrate the growing popularity of these approaches in third sector research and the need for proactive reflection on what a rapid influx of new research methods might mean for the peer review process.

Table 1 summarizes the years in which data science papers (both theoretical and empirical) appeared in the flagship journals of three of the major academic associations engaged in third sector research (ARNOVA, ISTR, ASPA) indicating the growing trend of scholarly attention being given to these phenomena. Purely theoretical or descriptive research articles were separated from the empirical papers which employed computational / data science approaches in their methods or analytical analyses. For example, an article that discussed the use of artificial intelligence and automated systems in government decision making, but which did not itself use artificial intelligence or another computational approach to construct a model was considered "Theoretical." Notably, all but two PAR articles were theoretical. See Appendix A for a discussion of the sampling methodology to identify the scholarship described below.

Table 1 illustrates the relative distribution of data science concepts in third sector scholarship. Theoretical papers shared a common goal advising caution with respect to the wide scale deployment of AI, big data and machine learning approaches. Their advice, however, was primarily targeting practitioners, as opposed to their fellow researchers, and few articles contained any recommendations or reviews of the methodological impact of these tools on academic scholarship.

This cursory review of recent scholarship demonstrates the growing utilization of data science tools and frameworks to further scholarship in this domain. As the supply of manuscripts grow, we expect increasing demands being placed on the peer review process to assess the adequacy and appropriateness of these methods for supporting meaningful advances in knowledge. We next discuss some key challenges of assessing research that is produced using data science methodologies and a feasible approach for third sector scholars to identify and communicate potential for racial bias in the work.

Critical Lenses in Data Science

Critical quantitative theory (aka QuantCrit) is a lens that was developed to challenge the notion that quantitative research is inherently objective or neutral.[1] It offers third sector scholars a framework for evaluating quantitative methodologies across dimensions including but not limited to race, class, gender, and their intersections. It aims to reveal systematic inequalities in quantitative research by offering up models, measures and analytical practices that situate institutions and individuals within their broader social context (Wells & Stage, 2015). QuantCrit scholarship is similar to third sector research in that it is inter- or trans-disciplinary at its core (Sablan, 2019) and third sector research often draws on the theories and approaches from many social sciences (Corry, 2010). It has primarily been

[1] According to Gillborn et al. (2018), the tenants of critical quantitative theory are *"(1) the centrality of racism as a complex and deeply rooted aspect of society that is not readily amenable to quantification; (2) numbers are not neutral and should be interrogated for their role in promoting deficit analyses that serve White racial interests; (3) categories are neither 'natural' nor given and so the units and forms of analysis must be critically evaluated; (4) voice and insight are vital: data cannot 'speak for itself' and critical analyses should be informed by the experiential knowledge of marginalized groups; (5) statistical analyses have no inherent value but can play a role in struggles for social justice."*.

applied to sociology and education research (Garcia, et al., 2018) and is gaining traction in other social sciences.

The primary contention of QuantCrit is that quantitative data is just as socially constructed as any other form of data. Faircloth et al. (2015) put it well: "[i]f, as researchers, we impose our own thoughts and beliefs onto the data without giving careful consideration to alternative perspectives and approaches, a cultural mismatch will result between what we purport to measure and what is actually measured in multiple ways." Thus, QuantCrit challenges dominant frameworks whose very nature can be classified as employing White-centered logic or White supremacist ideological origins (Zuberi and Bonilla-Silva, 2008) holding central to their work a desire to advance social justice and the empowerment and legitimization of the unique perspectives and worldviews of marginalized communities (Kincheloe & McLaren, 1994).

The instructive lens of QuantCrit helps us anticipate and respond to some of the ways in which the emerging data science landscape can impact the third sector scholarship. A field that desires to balance methodological plurality with a commitment to social welfare, civil society, and the public good must adopt a pragmatic and principled approach to embracing change.

A Taxonomy of Harmful Data Science Practices

Data science leverages approaches in statistics and computer science/information systems to create generalizable knowledge from data (Dhar, 2013). However, data scientists collect and utilize data *from*, *by*, *about* and *affecting* people and society so in effect data are people. Decision's scholars tell us that the appropriateness of data, tools, and the data science workflow directly impact our understanding of behavior and social institutions (D'Ignazio & Klein, 2020). Recent examples of this impact have led to a variety of public harms resulting from the deployment of data science algorithms, including:

Unrepresentative Data Harms

Buolamwini and Gebru (2018) identified public harms caused by the deployment of facial recognition software that was calibrated with image databases that consisted of predominantly light-skinned males and fewer women and people with dark skin. As a result, the misclassification rate for darker-skinned female faces (34.7% error rate) was 43 times higher than lighter-skinned males (0.8% error rate). Poor performance that originates from lack of representation in training data disproportionately affects people of color.

Overly-Representative Data Harms

Training that realistically captures human discretion will often produce models that generate biased predictions because of the biases present in the underlying data. For example, policing data might contain arrest rates that are 4 to 10 times higher for Black citizens when underlying rates of criminal behavior are the same in White and Black segments of the population (Beck, 2018). As a result, predictive recidivism risk models have proven unreliable and biased against Black defendants (Dressel & Farid, 2018).

Identity Proxies

In many data science contexts laws forbid the explicit use of race as a criterion for denial of a product or service such as a loan or credit card. As a result, race variables are excluded from predictive models that inform these sorts of decisions. Race does not need to be explicitly present in a model, however, to impact the results (O'Neil, 2016). Geographic indicators like zip code and economic indicators can serve as proxies for race, resulting in de facto legal racial discrimination in predictive algorithms.

Harms of Ignorance

In many cases, algorithms perform amazingly well on their intended tasks, but deployment does not consider racial contexts nor potential harm. Noble (2018) documents how Google's keyword search algorithms returned pornographic and profane images as top results when "Black girls" or "Black women" were used as search terms. The algorithms performed well in most contexts, but when deployed without safeguards in contexts where exploitation has been historically prevalent they proved to be harmful.

Harms of Subpopulation Difference

Algorithms that perform well on average can have poor performance within subgroups of the population. For example, it is possible for a new drug to have a net positive population effect but when data is disaggregated it can be shown to harm a subpopulation. While FDA approval processes now require analysis of subpopulation differences, there is no equivalent standards to hold companies accountable for harm resulting from proprietary algorithms. These examples of racialized bias are part of a larger conversation within the data science and ML ethics community surrounding persistent public value failures and algorithmic fairness (Monroe-White & Marshall, 2019).

A Protocol for Bias in Data Science Research Methodologies

The mission of the International Society for Third Sector Research (ISTR) is to "foster research in civil society, its relationships with state and market sector engagement and philanthropy that brings about positive social change and informs public policy." (ISTR, 2019) Therefore, the relative infancy of data science as a discipline coupled within its growing popularity and potential to profoundly influence social institutions highlights the concern that "[q]uantitative data is often used to shut down, silence, and belittle equity work" (Gillborn et al., 2018). The intentionality deployed while integrating data science methods into third sector scholarship should be of central importance to the field.

Critical quantitative theory (QuantCrit) is a useful lens, but also more of a discourse than a systematic approach that can be used to probe a specific study for evidence of data science harm. The fundamental ontology of critical quantitative theory is to help disrupt conceptualizations of quantitative or computational methods as neutral or objective, and instead locate them as part of the broader socially constructed scientific process.

Embracing data science techniques while simultaneously employing a critical lens enables third sector scholars to better "bring about positive social change" and empower socially marginalized and minoritized groups through their work. Thus, as championed by Viterna et al. (2015), third sector scholars must "narrow the gap between the actors they study and the theoretical construct that they are supposed to represent."

We draw upon QuantCrit literature and the success of other bias-prevention tools such as the Bechdel-Wallace test in the literature and the Checklist Manifesto in medicine (Gawande, 2009). These two examples have achieved significant impact as parsimonious instruments that are easy to apply yet effective at identifying bias in representation or blind-spots in decision making. We propose here a similar heuristic designed to detect bias endemic in new computational research methods. The Wells-Du Bois protocol is a tractable approach for scholars, journal editors, and manuscript reviewers to determine whether research that uses machine learning promotes bias mitigating practices. The protocol is comprised of a checklist of items that authors can review before submitting work for publication, or a disclosure that a journal could require from authors.

The protocol is named after Ida Wells, a pioneering data journalist, and W.E.B. Du Bois, a brilliant social scientist, and data visualization pioneer. They devoted much of their intellectual and professional lives to developing novel empirical techniques for unmasking biases in human institutions and advancing new research methods for uncovering unfair treatment of minoritized populations. Their use as a namesake celebrates their rare ability to combine systematic data collection and analysis with humanistic and interpretative lenses to produce rigorous bodies of evidence that challenge the status quo.

The Wells-Du Bois Protocol

Ida B Wells-Barnett (1862–1931) was a pioneering data journalist, an influential member of Black literary circles, and a staunch antilynching advocate after the murder of three close associates outside of Memphis, Tennessee. She was editor of the *Free Speech* periodical in which she openly decried white-owned newspapers slanderous claims that the hanging, shooting, and burning alive of over 1100 Black men, women, and children were acts of righteousness in the defense of white women (Wells-Barnett, 1895). To counter these claims, she analyzed lynching records published in the *Chicago Tribune* (the leading journalistic outlet at that time) which included the victim's name (where applicable), location (city/town and state), date of lynching (month, day, and year), and accusation. She found that the accusations placed against the victims of these extra-judicial killings included "murder," "rape" alongside "stealing," "writing a letter to white woman," "enticing servant away," and "unknown." However, according to Wells-Barnett "the same crimes committed by white men against Negro [men,] women, [boys] and girls, is never punished by mob or the law." (Wells-Barnett, 1895, p. 74).

W.E.B. Du Bois (1868–1963) was the first Black person to receive his PhD from Harvard University and among his many accomplishments is his "Exhibit of American Negroes" which won the grand prize at the Exposition Universelle of 1900 in Paris (Smith, 2000). Referred to as the "visual theorist of race," Du Bois, a social scientist, and activist, collected and visualized novel data and statistics on, by and for Black people to draw attention to the double standard of justice in the US. The exacting yet hand-drawn charts, maps, and graphs, developed by Du Bois and his team at the Atlanta University Center (now Clark Atlanta University, a Historically Black University) told the story of Black people's population and economic growth over time, and literacy rates with timeless beauty and convincing visual precision (Lewis & Willis, 2010; Van Winkle, 2022) (See Fig. 1).

Wells-Barnett and Du Bois mobilized data as a tool to expose the hypocrisy of US institutions which sought to minimize the suffering of Black people and invalidate their human experience in service of the status quo. In other words, not only was their work critical in its origins and purpose, it was 'emancipatory', freeing members of the Black community from the persistent onslaught of

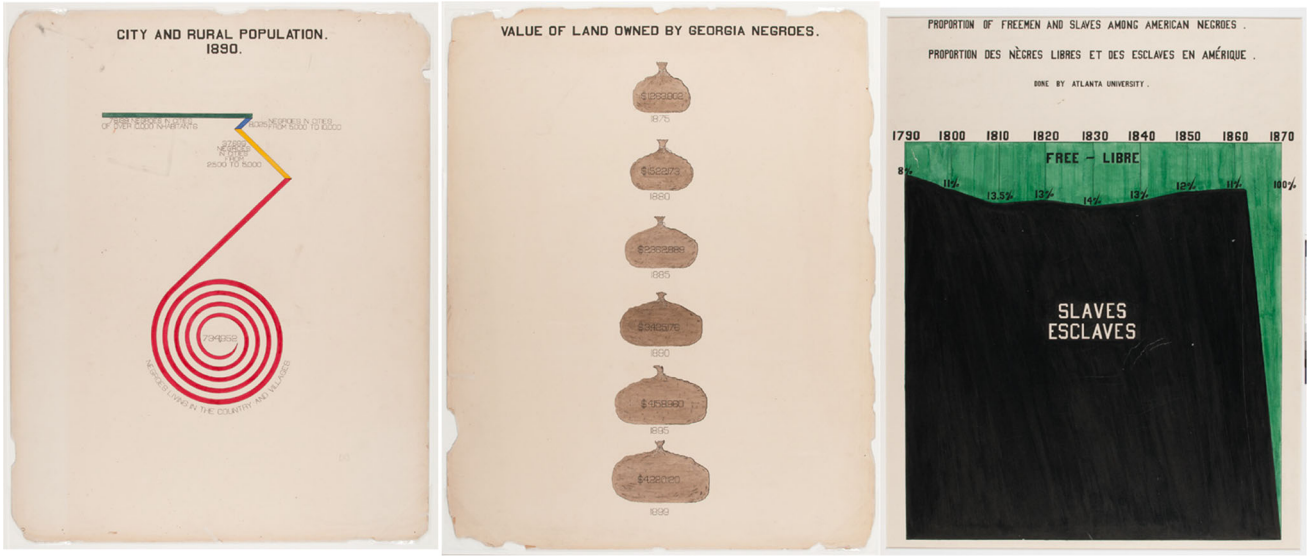

Fig. 1 **a** City and Rural Population; **b** Value of Land Owned by Georgia Negroes; and **c** Proportion of Freemen and Slaves among American Negroes (*Source*: W.E.B. Du Bois, 1900)

dehumanizing data framings and eugenicist narratives (Monroe-White, 2021). Their pioneering empirical work serves as a shining example of the non-neutral nature of data collection and data-driven discourse. Reflecting this lineage, the Wells-Du Bois protocol is designed as a feasible approach to determining whether research that uses data science methodologies has taken adequate steps to mitigate racial bias and other sub-group bias in the analysis. The seven items on the Wells-Du Bois protocol are:

Bad Data

(1) *Inadequate Data:* Does the data systematically omit or miscode a subpopulation? For example, when gender coding US names, does gender classification disproportionately fail on unfamiliar (e.g., non-Western) names?
(2) *Tendentious Data*: Was the data generated by, or does it represent subjective decisions made by other humans (e.g., judges granting parole)? If so, can their biases bias your results?

Algorithmic Bias

(3) *Harms of Identity Proxy:* Is there any way that models might systematically treat one race, gender, or class differently? E.g., unsupervised learning reproduces status quo groups like racially segregated neighborhoods). Alternatively, when you believe you are excluding race, can race still be constructed from other variables? E.g., the model may not explicitly include race; however, zip codes may act as proxies for race.
(4) *Harms of Subpopulation Difference:* Does the algorithm have differential performance across subgroups? For example, predicting race/gender from names is more accurate for majority populations than subpopulations.
(5) *Harms of Misfit Models:* If the models are predictive, have you examined their accuracy by subpopulation to ensure performance is not significantly different? Specifically, are what is your value-orientation and what are the public/social implications of this work?

Human Intent

(6) *Do No Harm:* Are you being transparent about the goals of your work? For example, gerrymandering voting districts to disempower voters and intentionally marginalize a segment of the population?
(7) *Harms of Ignorance:* Have you considered the unintended consequences of your research? That is, could the results be easily misappropriated to target disadvantaged populations?

Like the Bechdel test, which is notorious for its simplicity (for a film to pass the test, two women must talk to each other about something other than a man), this checklist employs parsimony as a means of taming complexity during the review of methodological rigor. The eight questions help authors to assess and disclose whether they have sufficiently attempted to mitigate bias. As pointed out by Gawande's work (2009), checklists are an

effective way to focus attention on and avoid errors caused by complexity and information overload.

The full Wells-Du Bois protocol is included as Appendix B and contains the full checklist of items with examples of each type of harm drawn from the literature and good-faith rules for mitigation bias associated with each item.

Discussion

Public sector and third sector scholars have detailed problems associated with big data and machine learning under a variety of conditions including e-governance, performance measurement in K-12 education, and public institutions (Lavertu, 2016; Mergel et al., 2016). They emphasize that success in machine learning approaches is highly dependent on human factors including question definition, proper data collection, data literacy, and analysis (Carnochan, et al., 2014). Lack of capacity to navigate changes caused by widespread adoption of these tools can lead to privacy loss, demographic and economic disparities in access, biased AI systems, and a lack of big data preparedness among public officials (Agarwal, 2018; Allard et al., 2018; Mergel et al., 2016).

Similarly, this paper contends that the rise of new computational methodologies presents predictable challenges to scholarly fields. Anyone advocating for the uncritical adoption of data-intensive approaches because of their "objectivity," "transparency," and "efficiency" should be reminded of the devastating impact that eugenics, the data-driven pseudoscience, still has on present-day welfare state, conceptualizations of social class, and institutions broadly (Roll-Hansen & Broberg, 2005; Jones, 2019; Wilkerson, 2020). Being data-driven is not sufficient to be objective and arguments constructed with data are not neutral. Mitigating bias in machine learning will require intentionality around their use.

The logic of data, how we identify, collect, prepare, analyze, and interpret it—whether big, small, unstructured, structured, supervised, or unsupervised—must be guided by a critical lens. The plentitude of failures in machine learning should remind researchers to be modest in expectations regarding immediate returns to big data and computational techniques. As new methods become increasingly commonplace in academia, we are called to practice humility with our claims (Weizenbaum, 1972), consider coupling quantitative outputs with humanistic and reflexive approaches (Harris, 2001), and to safeguard public values (Bozeman, 2002).

Scholars have raised red flags over governance and management challenges arising from large-scale deployment of "fourth revolution" technologies in practice (AI, facial recognition, social bots, etc.), but relatively scant attention has been paid to the ways in which these methodologies might overtime shape scholarship. Many of these ethical considerations have not found adequate operationalization in the empirical literature. There is limited yet promising evidence that third sector scholars are adopting data science methodologies with an eye toward critique, but without a strong framework in place, these efforts will be sporadic and unsystematic. While the methodological training in PhD programs will eventually evolve to encompass these novel computational approaches, in the meantime, current scholars may lack the sophistication necessary to rigorously evaluate research claims during the peer review process.

This manuscript leverages insights from the field of critical quantitative theory and examples of successful bias-reduction protocols to articulate a taxonomy of harmful data science practices and propose a checklist that can help authors embrace bias mitigation practices. The proposed Wells-Du Bois protocol is an example of a feasible stop-gap approach to addressing limited machine learning expertise in the field using a checklist of bias mitigating assessments completed by authors. Although imperfect, it does provide an actionable approach to an important and growing problem in the field.

Given that machine learning and other computational approaches are still relatively nascent in third sector research, the field has an opportunity to embrace the revolution with intention. Scholars can begin to leverage these powerful new methods to enhance research while remaining cognoscente of the significant potential for systematic racial or sub-group bias. Journals and reviewers can push for simple disclosure protocols that help avoid scientific errors that commonly arise when new computational methodologies are deployed in scholarship.

Appendix A: Literature Review Methodology

Impact criteria of journals reviewed were based on publicly available journal and/or society descriptions, aim and/or scope narratives and their impact factor.[2] Accordingly, the flagship journals of three of the major academic associations (ARNOVA, ISTR, ASPA) engaged in third sector research were selected: two top nonprofit journals and one public administration/nonprofit hybrid: Nonprofit and Voluntary Sector Quarterly (NVSQ), Voluntas: International Journal of Voluntary and Nonprofit Organizations (Voluntas) and Public Administration Review (PAR).

[2] Impact factor scores are measured as the total citations made in the journal citation report year to content published in a journal in the prior two or five years, divided by the number of articles and reviews published by the journal in the prior two or five years.

We identified papers containing data science approaches through four levels of review: journal selection, search, elimination of duplicates, and theoretical or methodological relevance. Next, within each journal, the initial search function was used to identify articles that mention computational social science or data science methodologies anywhere in the manuscript text, including: "artificial intelligence," "big data", "data mining," "data science," "machine learning," natural language processing" or "NLP," "neural net*," "text mining," and "sentiment analysis" without restriction by year. This search resulted in 88 articles (NVSQ: 12; PAR: 61; Voluntas: 15).

The next round of preprocessing required eliminating duplicates. Duplicates were frequently occurring in this dataset given that a single paper could contain as few as one or as many as six keywords. This resulted in the elimination of a total of 36 papers, leaving 52 articles (NVSQ: 10; PAR: 34; Voluntas: 8). Next, papers were evaluated based on the type of article that they represented. Original research articles were retained, while other types of manuscripts including book reviews, perspectives, and editorials were excluded. Likewise, articles that made mention of the search terms described above, but whose focus on them was cursory were also removed. This process resulted in the removal of 27 additional articles, for a final count of 25 articles pertaining directly to the research topic (NVSQ: 6; PAR: 13; Voluntas: 6). Finally, articles were evaluated on the use of methodologies as opposed to the discussion of their importance, impact, or use.

Among the empirical papers (n = 13), six used machine learning techniques to analyze unstructured text data, four used coding software to gain access to and/or create datasets for further analysis using traditional statistical approaches; two papers used automated approaches to prepare and clean data (i.e., transcription or harmonization of datasets) and two studies combined traditional statistical analyses with machine learning analytical approaches (Table 2).

One study explicitly opted not to use automated text mining approaches to analyze their corpus of data, because, in their words, "[w]hile computational sentiment analysis is very useful for analyzing a large corpus of documents, manual coders provide better accuracy than computational sentiment analysis" (Wasif, 2020). Lastly, two articles combined qualitative or statistical analyses with computational methods. Interestingly, no studies used data science methods or analytical tools to exclusively analyze non-text data.

The methodological and analytical priorities emphasized within these empirical studies included issues of data quality and the comparative advantage of manual vs. fully automated and semi-automated approaches. Among the studies using machine learning approaches for analysis, half explicitly discussed the value of human-generated content or the inadequacy of fully automated machine learning approaches for analyzing text data (see Fyall et al., 2018; Litofcenko et al., 2020; Scurlock, et al., 2020).

Other studies focused on the objectivity and efficiency of text mining approaches compared to manual coding procedures, emphasizing the value of transparency, and replicability (Walker, et al., 2019). Few studies, however, explicitly provided readers access to datasets or code (i.e., GitHub pages, R or Python code) which would facilitate the achievement of those values (see Lecy & Thornton, 2016; Guo & Saxton, 2018) and just one study incorporated model performance evaluation scores (Litofcenko, 2020).

Finally, most of the studies simply focused on describing the features of the computational approach or tool with

Table 2 Qualitative content analysis of empirical uses of data science methods, analytical approaches, and tools

Data science methods and analyses	Subcategories	References
Data analysis	Text data: automated or semi-automated text/content analysis, latent Dirichlet allocation (LDA), sentiment analysis, decision tree models	Chen and Nakazawa (2017); Fyall et al. (2018); Zhu and et al. (2019); Walker and et al. (2019); Litofcenko and et al. (2020); Santos and et al. (2020)
	Non-text data: *none*	*n/a*
	Combining machine learning (e.g., k-modes clustering) with traditional qualitative (e.g., LDA) or quantitative (e.g., multiple regression) approaches	Norris-Tirrell et al. (2018); Zhu and et al. (2019)
Data acquisition	Data acquisition: web scraping, online and social media APIs (Twitter)	Zhu et al. (2019); Guo and Saxton (2018); Nwakpuda (2020); Scurlock et al. (2020)
Data preprocessing and preparation	Data harmonization and cleaning: fuzzy logic matching algorithms; AI-enabled transcription services	Lecy and Thornton (2016); Nwakpuda (2020); Williamson et al. (2020)

minimal or no discussion of its limitations (e.g., Chen & Nakazawa, 2017; Norris-Tirrell et al., 2018; Santos et al., 2020; Walker et al., 2019; Zhu et al., 2019).

Appendix B: *The Wells-Du Bois Protocol: Mitigating Biased Practices in Data Science*

Data science leverages approaches in statistics and computer science/information systems to create generalizable knowledge (Dhar, 2013). Myriad recent examples of bias in machine learning models or failed AI platforms have highlighted instances of public harms resulting from the unenlightened deployment of data science applications (O'Neil, 2016). Social scientists have a vested interest in these important considerations since they utilize data *from*, *by*, and *about* people, so algorithmic bias or failure can cause material harm to real people and impact lives.

The Wells-Du Bois protocol is an actionable approach to avoiding harms in data science application. It is a tool inspired by the simplicity and efficacy of similar parsimonious instruments such as the Bechdel test for gender bias (Agarwal, 2018), the Apgar Score for newborn infant health (Gawande et al., 2007; Regenbogen et al., 2009), and myriad examples presented in Gawande's *Checklist Manifesto* (Gawande, 2011). The core insight is that these heuristics are not perfect measures of subtle latent constructs like gender bias or health, but they are approximately as accurate as more sophisticated instruments using only a couple of questions that are simple to apply.

Data science applications can cause harm when predictions are bad, model performance varies across groups, or engineers have failed to consider nefarious uses of the tools they are building. It is a challenge to determine whether complex machine learning models are likely to fail in these ways because it requires time and resources to "stress-test" models and validate them in the real world. Developing mathematical or algorithmic tests for bias requires advanced computational and statistical expertise and a sufficient understanding of algorithmic fairness (ethics for data scientists). Solutions require the calibration of model goals more than model parameters since algorithmic fairness requires trade-offs in performance between subgroups in the data, not optimization of a global model performance metric. These tasks require expertise beyond reach for most applied data science teams or social science collaborations.

The Wells-Du Bois (WDB) protocol is a stop-gap approach to a lack of broad data science capacity in a discipline or field. It uses a simple protocol to identify common harmful practices instead of explicit tests by experts or an attempt to be exhaustive. It is designed as a checklist approach to harm reduction undertaken by scholars or system engineers before manuscript submission or model deployment.

The goal is to capture the least onerous process that will prevent the most serious instances of data science harm. Users review the protocol before project deployment and assess whether there is a high likelihood of each type of data harm in their application, then decide whether they want to fix the issue or document and disclose the risk.

Whereas social science journals have developed nuanced ways to report regression results to avoid bias from omitted variables or under-specified models, there is no equivalent reporting protocol for research generated with a large class of machine learning applications that are calibrated using predictive fit metrics. In the academic context, journals could also consider requiring authors to submit the protocol along with manuscripts that utilize machine learning algorithms as part of the peer-review process. In such a context, it would be near impossible for reviewers to detect harmful practices based on the type of information usually disclosed in a methodology section in quantitative studies.

Assumedly social science publishing practices will evolve as research methodologies evolve. As such, the WDB protocol is designed to seed discussions that will generate insights that move this conversation forward. There will ultimately be more durable solutions to these problems. In the meantime, it is an actionable step that requires a minimal amount of time and expertise to implement.

The Protocol

The Wells-Du Bois protocol is a process by which authors or engineers can assess a project to identify potential sources of harm. The protocol does not ask authors or engineers to verify that the application is free of any problems. Rather, it suggests mitigation strategies for each type of potential harm when possible and promotes transparency when mitigation is not possible. In this way, it is a good faith threshold for mitigating biased research practices.

Bad Data

(1) Inadequate Data

Buolamwini and Gebru (2018) identified public harms caused by the deployment of facial recognition software that was calibrated with image databases that consisted of a larger number of light-skinned males than women and people with darker skin tones. As a result, the misclassification rate for darker-skinned female faces (34.7% error rate) was 43 times higher than lighter-skinned males (0.8%

error rate). Poor performance that originates from lack of representation in training data disproportionately affects women of color.

> *Inadequate Data Mitigation:* Representative data in most cases, but minoritized and marginalized groups, might need to be over-represented in some instances to achieve necessary sample sizes for neutral algorithms. Authors that use predictive algorithms can demonstrate harm-reduction strategies by (1) reporting sample sizes and descriptive statistics in training datasets separated out by socially constructed group identities like race or gender, and (2) reporting performance by in this way will help to demonstrate that algorithm performance is consistent across groups. More nuanced treatments would also include considerations of intersectionality—interactions of subgroup status like race and gender—since the lived experiences of individuals in these populations are shaped by the combined effects of these and other socially constructed identities.

(2) Tendentious Data

Tendentious data are data generated by or accurately capturing human behavior in the real world. As a result, human tendencies of imperfect decision making and implicit or explicit bias are baked into the data. It will accurately reflect socially constructed realities and capture contours of human action with high fidelity. As a result, when it is used as training data for machine learning models it will generate predictions that mirror biases present in society.

For example, Amazon built an artificial intelligence system to automate the review of thousands of applications they get for engineering positions. The system performed amazingly well at identifying highly qualified male engineers, but because it was trained using resumes from previous hiring processes and they had historically hired more male engineers than female engineers the algorithm penalized women and filtered out many highly qualified female candidates (Christian, 2020).

The problem was not in the construction of a training dataset, and it was not an algorithm that performs poorly. It is that real-world data were used to train a system and the algorithm accurately reproduced human bias that was encoded in the data. The tendentious data predisposed the machine learning models toward unfair predictions. The crux of the problem is that the algorithm is accurately predicting how Amazon recruiters behaved in the past (which is what it was trained to do), not whether a candidate can do the job well (which is what we want). Tendentious data produce models that are predisposed to the same cognitive inconsistencies or biases as the humans that generated the training data.

In a more subtle example, the over-policing of Black neighborhoods will result in higher arrest rates for Black citizens. As a result, predictive recidivism risk models will over-estimate their propensity of Black defendants to commit future crimes, deeming them as a higher risk than they are in reality (Dressel & Farid, 2018). In their study of more than more than 10,000 criminal defendants, ProPublica investigative journalists found that Black defendants were twice as likely to be misclassified as higher risk when compared to white defendants (Angwin et al., 2016).

The challenge is that arrest data accurately captures the number of arrests made by police, but it will over-represent the latent criminality of Black people because of the legacy of discrimination in the US criminal justice system. Thus, it both accurately represents the phenomenon of arrest but is biased in estimating the likelihood a Black defendant will commit a future crime or act of violence. Thus, it will generate tendentious predictive models that perform poorly when measured against objective reality.

> *Tendentious Data Mitigation:* Disclose whether outcomes used in a training dataset are socially constructed latent constructs (i.e., intelligence, status, effectiveness, performance) that will reflect subjective human judgements or consist of administrative or archival data produced by human actors. For example, typing speed is a fairly objective and robust performance measure, but anything subjective like teacher evaluations will be prone to human bias.

It is usually not possible to fix the data. Once bias gets baked into the data, it reflects the realities of the world and models can only be fit to the subjective interpretations (human tendencies) that are present in the data. Awareness and disclosure of the issue is a starting point. Better, less subjective data is a more durable longer-term solution, but that requires time and resources needed to create calibrated outcomes (e.g., technical assessments for job candidates instead of only resumes).

Some bias-mitigation strategies might also be possible. In the policing case, for example, train the models using subsets of crimes that are more objective in nature like armed robberies. Compare those models to similar ones trained on crimes that are more subjective in nature, like traffic stops. That may help to quantify the amount of bias in the model. Predictions made from the models could then be corrected by reducing risk scores from that category of race by the level of bias estimated from the

differences observed in risk associated with subjective versus objective outcomes. These techniques are commonly called "instrumental variables" in econometrics and are imperfect solutions. But beyond simply identifying and disclosing the problem while waiting for better data, they are one feasible approach to some level of bias mitigation.

Algorithmic Bias

(3) Harms of Identity Proxy

In many contexts where machine learning models enhance or replace human decision making, such as approval for a loan or credit card, laws forbid the explicit use of protected classes as a criterion for being denied a service (race, ethnicity, gender, language, and disability status). As a result, these variables are excluded from models to promote fairness and avoid lawsuits. However, they do not need to be explicit in models to be present. The role they serve in making algorithms biased or unfair can be fulfilled by identity proxies—composites of other variables in the model that predict the protected class with high accuracy.

The previous example of Amazon's attempt to build an AI system to identify the best job candidates from the mountains of resumes they receive demonstrates the challenge of identity proxies as well. They trained their algorithms using resumes from past positions that were filled and calibrating the models based on who was actually hired from the candidate pools. Historic bias in hiring decisions led to algorithms that systematically excluded women from the list of top candidates in the current list of resumes. Once engineers recognized this problem they were horrified so they removed gender from variables considered by the model, expecting that to resolve the issue. What they quickly discovered, however, is that resumes are full of gender proxies.

The algorithms could predict the gender of candidates based on things like their first name, which sports they listed (softball vs baseball), hobbies or interests (anything that mentioned "women's"), or which schools they attended (all-women's colleges were penalized). These examples seem somewhat obvious and could be mitigated by editing the algorithm once identified. The problem was that gender was deeply embedded in resumes in subtle ways that are hard to identify, such as the vocabulary used by candidates. The algorithms awarded points for the appearance of words like "executed" and "captured," which are more prevalent on male resumes. It proved to be impossible to identify and neutralize all gender proxies, so Amazon scrapped the project (Christian, 2020). Race and other subgroup proxies manifest in similar ways in data.

How pernicious is the problem? Some recent studies have shown how powerful machine learning algorithms have become at identifying patterns that are imperceptible to humans but can serve as signatures of group membership. For example, Bansal et al. (2012) used a support vector machine to predict the gender of a patient from an image of the iris, a result that surprised medical professionals since male and female irises look the same to humans. Similarly, Banerjee et al. (2021) showed AI can predict a patient's race from an X-ray despite the fact that physicians were unaware of any known correlations between features on medical images and race. They show that "detection is not due to trivial proxies or imaging-related surrogate covariates for race, such as underlying disease distribution" and that "performance persists over all anatomical regions and frequency spectrum of the images suggesting that mitigation efforts will be challenging."

These examples demonstrate how much information is unintentionally encoded in data. Amazon did not anticipate that gender would permeate all aspects of a candidate's resume. Bansal and Banerjee's teams were not designing medical imaging techniques that optimize detection of gender or race. They simply used archives of eye images or X-rays to demonstrate how powerful machine learning algorithms have become at detecting patterns that are invisible to experts. The lesson they demonstrate is that it is an unrealistic assumption that algorithms are not making predictions based on protected classes simply because variables like race, gender, or disability status are not explicitly present in models. Identity proxies are ubiquitous.

Harms of Identity Proxy Mitigation: Instead of dropping a category like race from the model, assuming it solves the problem, test for the presence of identity proxies. Make race the dependent variable in the model while retaining the other covariates. If you can accurately predict race or another protected class with the remaining covariates in your model, identity proxies are present. Using predictive models when identity proxies are present violates the spirit of protections afforded to historically disadvantaged classes. When relevant to the research, report whether identity proxies can be detected in the data.

(4) Harms of Subpopulation Difference

Harms of subpopulation difference arise when algorithm performance varies by subgroup within the data. For example, it is possible for a new drug to have a net positive population effect but when data are disaggregated, it can be shown to harm a subpopulation. While federal drug approval processes now account for subpopulation

differences, there are no standards to hold corporations accountable for subpopulation harm resulting from algorithms.

The extent of this problem was only recently uncovered. In 2016 ProPublica conducted a high-profile review of the COMPAS algorithm that predicts the risk of an incarcerated criminal committing a future crime if released. They found that algorithm to be biased against Black prisoners, assigning risk scores that were higher than justified by their actual observed behavior once released (Angwin et al., 2016). Stated differently, if a white prisoner and a black prisoner were assigned the same risk score, the white prisoner was more likely to recidivate. Black prisoners were being refused parole at higher rates as a result.

ProPublica's in-depth investigative reporting served as a lightning rod of sorts for machine learning scholars with expertise in algorithmic fairness, catalyzing a flurry of studies on the topic and an important breakthrough. The fact that the COMPAS algorithm was generating risk scores that were inconsistent across race was not a bug in their specific machine learning algorithm, but a feature of all machine learning models. Chouldechova (2017) and Pleiss et al. (2017) were able to prove that it is mathematically impossible to build machine learning models that perform consistently across subgroups in the data when baseline rates of the outcome differ by subgroup. In other words, the issue was not a poorly specified model in inadequate data inputs. Rather, it is a fundamental feature of machine learning models when baseline rates of the outcome vary by subgroups within the data. These examples of racialized bias are just one example of a larger conversation within the data science and ML ethics community surrounding persistent public value failures surrounding machine learning applications in the public domain (Monroe-White & Marshall, 2019).

> *Harms of Subpopulation Difference Mitigation:* There is no easy fix to the problem of machine learning models that perform differently for each subpopulation within the data. Mathematical proofs have shown that it is impossible to fix the issue using a different model specification or better data. Mitigation is less about solving the issue and more about recognizing the issue before you build or deploy your algorithm.

COMPAS failed because the algorithm designers defined fairness by calibrating risk scores using false positives alone (Angwin, et al., 2016). The risk scores produced by the models did in fact accurately predict whether an individual was likely to be re-arrested after release, regardless of race. But performance varied wildly on the other criteria–false negatives, whether an individual remained locked up when they were unlikely to commit a future crime. The algorithm deemed blacks much riskier than was supported by the data, so risk scores were much higher than they should have been. This result in a scenario where if you put inmates into risk categories, within each category white inmates were given lower risk scores and were much more likely to receive parole than Black inmates that had a similar likelihood of recidivism.

Publications that use machine learning classifiers should examine outcomes across various protected subgroup classes in the data like race and gender to determine whether baseline rates differ by group. If so, recognize that it is impossible to build a model that will perform similarly for all subgroups. With just two subgroups like data that only includes individuals from White and Black race categories, the models can achieve balanced false positive rates but have disparate rates of false negatives, or achieve balanced false negative rates and have disparate rates of false positives. They cannot achieve balance in both (Kleinberg et al., 2016). The more subgroups there are present, the harder it is to achieve balance.

Mitigation starts with awareness or the issues that arise when subgroups have different baseline performances in the data and disclosure of baseline rates, as well as reporting model performance statistics separately for subgroups to quantify the extent of the difference. That would meet a minimal requirement of fairness operationalized as transparency.

(5) Harms of Misfit Models

Models become misfits when they are optimized by minimizing the wrong type of error or preferencing the wrong outcome. Minimizing model error instead of maximizing model benefit. One of the lessons from the field of machine learning is that less accurate models are generally more useful. This relationship arises because of two model features: (1) model accuracy can be improved by overfitting data and (2) the most powerful machine learning models (e.g., neural networks) generate highly accurate results that no one understands and consequently are harder to translate into practice.

> *Harms of Misfit Models Mitigation:* The ML community has developed many practices like cross-fold validation to avoid over-fitting and has evolved an ethos of valuing simple models or assemblages of simpler models that offer insights to experts and decision-makers over black-box oracles.

Human Intent

(6) Do no Harm

Most data science failures are cases where models are biased in subtle ways or performance deviates from

expectations. The first item in the checklist asks, however, is the intent of the deployment to cause harm? For example, Cambridge Analytica used sophisticated data analytics to target marginalized and minoritized populations with misinformation to create a sense of frustration with the political process and make it less likely for the targeted individuals to vote (Schneble et al., 2018). Similarly, spatial analytics can be used to gerrymander voting districts to distort representation. In both cases, the goals are to suppress votes and undermine the democratic process, which one can argue are nefarious goals. In these examples there are no algorithmic failures or undetected bias. The models perform perfectly well, but their intent is to disadvantage or harm a subpopulation.

Do no Harm Mitigation: Can you confirm that your true intent matches the intent that is described in your manuscript or project documentation? Is the intent to benefit populations equally, or to advance objectives of fairness or justice?

(7) Harms of Ignorance

In many cases, algorithms perform amazingly well on their intended tasks, but deployment does not consider potential harm, especially in sensitive racial or gender contexts. For example, Noble (2018) documents how Google's keyword search algorithms returned pornographic and profane images as top results when "Black girls" or "Black women" were used as search terms. The search algorithm generally performed well in most contexts, but there was no proactive consideration for what might happen when deployed without safeguards in contexts where exploitation has been historically prevalent. The most pernicious examples of harmful practices are more likely unintended consequences that are difficult to anticipate when new technologies are deployed in unfamiliar or complex environments. Harms of ignorance are more aptly characterized as uncontemplated consequences—a sort of willful ignorance that results when scholars or engineers (composed primarily of While and Asian males) have not considered possible harms to minoritized groups (i.e., Asian women, Black males, etc.). Unintended consequences are hard to anticipate, whereas uncontemplated consequences could have been anticipated through a reasonable level of rumination.

Harms of Ignorance Mitigation: Heath and Heath (2013) promote the use of pre-mortem analysis to avoid mistakes in business. These exercises occur prior to the launch of a new product or service, and require teams to sit down together and write a hypothetical obituary for the project. The exercise forces them ahead of time to determine the most likely reasons that a project will fail. This sort of brainstorming then allows the team to protect against the most likely threats to success as they begin to implement the project. Similarly, to mitigate harms of ignorance scholars or engineers should entertain the question, what's the worst possible outcome that could result from this work? Are there specific groups that might be vulnerable to disproportionate harm by a failure of the algorithm? In the domain of research scholars should ask, could the data collected for this research, or the tools developed through the work be used for nefarious purposes?

Funding The authors did not receive support from any organization for the submitted work. No funding was received to assist with the preparation of this manuscript. No funding was received for conducting this study. No funds, grants, or other support was received.

Declarations

Conflict of interest The authors has no potential conflicts of interest to disclose.

Human and Animal participants This research did not involve human participants subjects and/or animals.

Informed Consent The research did not require use of informed consent as no human subjects or animal research was involved.

References

Agarwal, P. K. (2018). Public administration challenges in the world of AI and Bots. *Public Administration Review, 78*(6), 917–921.

Allard, S. W., Wiegand, E. R., Schlecht, C., Datta, A. R., Goerge, R. M., & Weigensberg, E. (2018). State agencies' use of administrative data for improved practice: Needs, challenges, and opportunities. *Public Administration Review, 78*(2), 240–250.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *ProPublica, May*, *23*(2016), 139–159.

Ayodele, T. O. (2010). Machine learning overview. *New Advances in Machine Learning*, 2.

Baker, M. (2016). Reproducibility crisis. *Nature, 533*(26), 353–366.

Banerjee, I., Bhimireddy, A. R., Burns, J. L., Celi, L. A., Chen, L. C., Correa, R., Gichoya, J. W. (2021). Reading race: AI recognises patient's racial identity in medical images. arXiv preprint https://arxiv.org/abs/2107.10356.

Bansal, A., Agarwal, R., & Sharma, R. K. (2012, November). SVM based gender classification using iris images. In: *2012 fourth international conference on computational intelligence and communication networks* (pp. 425–429). IEEE.

Battaglio, R. P., & Hall, J. L. (2018). Trinity is still my name: Renewed appreciation for triangulation and methodological diversity in public administration. *Public Administration Review, 78*(6), 825–827.

Beck, A. J. (2018). Race and ethnicity of violent crime offenders and arrestees, 2018. *Bureau of Justice Statistics*.

Bozeman, B. (2002). Public-value failure: When efficient markets may not do. *Public Administration Review, 62*(2), 145–161.

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender

classification. In: *Conference on fairness, accountability and transparency* (pp. 77–91).
Carnochan, S., Samples, M., Myers, M., & Austin, M. J. (2014). Performance measurement challenges in nonprofit human service organizations. *Nonprofit and Voluntary Sector Quarterly, 43*(6), 1014–1032.
Chang, R. M., Kauffman, R. J., & Kwon, Y. (2014). Understanding the paradigm shift to computational social science in the presence of big data. *Decision Support Systems, 63*, 67–80.
Chen, Y. W., & Nakazawa, M. (2017). Emotions and Pan-Asian organizing in the US Southwest: Analyzing interview discourses via sentiment analysis. *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, *28*(6), 2785–2806.
Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153–163.
Christian, B. (2020). *The alignment problem: Machine learning and human values*. WW Norton & Company.
Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertesz, J., & Nowak, A. (2012). Manifesto of computational social science. *The European Physical Journal Special Topics, 214*(1), 325–346.
Corry, O. (2010). Defining and theorizing the third sector. In *Third sector research* (pp. 11–20). Springer, New York, NY.
Davenport, T. H., & Patil, D. J. (2012). Data scientist: The sexiest job of the 21st century. *Harvard Business Review, 90*(5), 70–76.
Dhar, V. (2013). Data science and prediction. *Communications of the ACM, 56*(12), 64–73.
D'Ignazio, C., & Klein, L. F. (2020). *Data feminism*. MIT Press.
Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, *4*(1), eaao5580.
Du Bois,W. E. B. (1900). African American photographs assembled for 1900 Paris Exposition. Library of Congress,Washington, DC. Retrieved from http://www.loc.gov/pictures/collection/anedub/dubois.html.
Ellegaard, O., & Wallin, J. A. (2015). The bibliometric analysis of scholarly production: How great is the impact? *Scientometrics, 105*(3), 1809–1831.
Faircloth, S. C., Alcantar, C. M., & Stage, F. K. (2015). Use of large-scale data sets to study educational pathways of American Indian and Alaska Native students. *New Directions for Institutional Research, 2014*(163), 5–24.
Fyall, R., Moore, M. K., & Gugerty, M. K. (2018). Beyond NTEE codes: Opportunities to understand nonprofit activity through mission statement content coding. *Nonprofit and Voluntary Sector Quarterly, 47*(4), 677–701.
Garcia, N. M., Lopez, N., & Velez, V. N. (2018). QuantCrit: Rectifying quantitative methods through critical race theory. *Race Ethnicity and Education, 21*(2), 149–157. https://doi.org/10.1080/13613324.2017.1377675
Gawande, A. (2009). *The checklist manifesto: How to Get things right*. Metropolitan Books, New York.
Gawande, A. (2011). The checklist manifesto: How to get things right. *Journal of Nursing Regulation, 1*(4), 64.
Gawande, A. A., Kwaan, M. R., Regenbogen, S. E., Lipsitz, S. A., & Zinner, M. J. (2007). An Apgar score for surgery. *Journal of the American College of Surgeons, 204*(2), 201–208.
Gillborn, D., Warmington, P., & Demack, S. (2018). QuantCrit: Education, policy, Big Data and principles for a critical race theory of statistics. *Race Ethnicity and Education, 21*(2), 158–179.
Guo, C., & Saxton, G. D. (2018). Speaking and being heard: How nonprofit advocacy organizations gain attention on social media. *Nonprofit and Voluntary Sector Quarterly, 47*(1), 5–26.
Hardwick, R., Anderson, R., & Cooper, C. (2015). How do third sector organisations use research and other knowledge? A systematic scoping review. *Implementation Science, 10*(1), 84.
Harris, M. (2001). The place of self and reflexivity in third sector scholarship: An exploration. *Nonprofit and Voluntary Sector Quarterly, 30*(4), 747–760.
Heath, C., & Heath, D. (2013). *Decisive: How to make better choices in life and work*. Random House.
Hodgkinson, V., & Painter, A. (2003). Third sector research in international perspective: The role of ISTR. *Voluntas: International Journal of Voluntary and Nonprofit Organizations*, *14*(1), 1–14.
ISTR. (2019). International society for third-sector research strategic plan 2019–2024. https://cdn.ymaws.com/www.istr.org/resource/resmgr/docs/istr_strategic_plan_sept_201.pdf
Jones, R. (2019). Eugenics in education policy and the impact on African American students.
Kincheloe, J., & McLaren, P. (1994). Rethinking critical thinking and qualitative research. *Handbook of qualitative research. Thousand Oaks, CA: Sage Publications*
Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. arXiv preprint https://arxiv.org/abs/1609.05807
LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. *MIT Sloan Management Review, 52*(2), 21–32.
Lavertu, S. (2016). We all need help: "Big data" and the mismeasure of public administration. *Public Administration Review, 76*(6), 864–872.
Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., & Jebara, T. (2009). Life in the network: The coming age of computational social science. *Science (new York, NY), 323*(5915), 721.
Lecy, J., & Thornton, J. (2016). What big data can tell us about government awards to the nonprofit sector: Using the FAADS. *Nonprofit and Voluntary Sector Quarterly, 45*(5), 1052–1069.
Lewis, D. L., & Willis, D. (2010). *A small nation of people: WEB Du Bois and African American portraits of progress*. Zondervan.
Liao, H., Tang, M., Luo, L., Li, C., Chiclana, F., & Zeng, X. J. (2018). A bibliometric analysis and visualization of medical big data research. *Sustainability, 10*(1), 166.
Litofcenko, J., Karner, D., & Maier, F. (2020). Methods for classifying nonprofit organizations according to their field of activity: A report on semi-automated methods based on text. *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, *31*(1), 227–237.
Marshall B., and Geier S., (2019) Targeted curricular innovations in data science, In: *Proceedings of the IEEE Frontiers in Education Conference*.
Mergel, I., Rethemeyer, R. K., & Isett, K. (2016). Big data in public affairs. *Public Administration Review, 76*(6), 928–937.
Monroe-White, T. (2021). Emancipatory data science: a liberatory framework for mitigating data harms and fostering social transformation. In Proceedings of the 2021 on Computers and People Research Conference (pp. 23-30).
Monroe-White, T. and Marshall, B. (2019). Data science intelligence: Mitigating public value failures using PAIR principles. *Proceedings of the 2019 Pre-ICIS SIGDSA Symposium*. 4._https://aisel.aisnet.org/sigdsa2019/4
Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives, 31*(2), 87–106.
Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press.

Norris-Tirrell, D., Rinella, J., & Pham, X. (2018). Examining the career trajectories of nonprofit executive leaders. *Nonprofit and Voluntary Sector Quarterly, 47*(1), 146–164.

Nwakpuda, E. I. (2020). Major donors and higher education: Are STEM donors different from other donors?. *Nonprofit and Voluntary Sector Quarterly*, 0899764020907153.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. *arXiv preprint* https://arxiv.org/abs/1709.02012.

Regenbogen, S. E., Ehrenfeld, J. M., Lipsitz, S. R., Greenberg, C. C., Hutter, M. M., & Gawande, A. A. (2009). Utility of the surgical apgar score: Validation in 4119 patients. *Archives of Surgery, 144*(1), 30–36.

Roll-Hansen, N., & Broberg, G. (Eds.). (2005). *Eugenics and the welfare state: Sterilization policy in Denmark, Sweden, Norway, and Finland*. Michigan State University Press.

Sablan, J. R. (2019). Can you really measure that? Combining critical race theory and quantitative methods. *American Educational Research Journal, 56*(1), 178–203.

Santos, M. R., Laureano, R. M., & Moro, S. (2020). Unveiling research trends for organizational reputation in the nonprofit sector, *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, *31*(1), 56–70

Schneble, C. O., Elger, B. S., & Shaw, D. (2018). The cambridge analytica affair and Internet-mediated research. *EMBO Reports, 19*(8), e46579.

Scurlock, R., Dolsak, N., & Prakash, A. (2020) Recovering from scandals: Twitter coverage of oxfam and save the children scandals, *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, *31*(1), 94–110.

Smith, S. M. (2000). " Looking at one's self through the eyes of others": WEB Du Bois's photographs for the 1900 paris exposition. *African American Review, 34*(4), 581–599.

Treaster, J. B. (2017). Will you graduate? Ask big data. *The New York Times*, *2*.

Van Winkle, K. (2022). Above all made by themselves: The visual rhetoric of WEB Du Bois's data visualizations. *Technical Communication Quarterly, 31*(1), 17–32.

Viterna, J., Clough, E., & Clarke, K. (2015). Reclaiming the third sector from civil society a new agenda for development studies. *Sociology of Development, 1*(1), 173–207.

Walker, R. M., Chandra, Y., Zhang, J., & van Witteloostuijn, A. (2019). Topic modeling the research-practice gap in public administration. *Public Administration Review, 79*(6), 931–937.

Wasif, R. (2020). Does the media's anti-western bias affect its portrayal of NGOs in the muslim world? Assessing newspapers in Pakistan. *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, 1–16.

Weizenbaum, J. (1972). On the impact of the computer on society. *Science, 176*(4035), 609–614.

Wells, I. B. (1895). *A red record. Tabulated statistics and alleged causes of lynchings in the United States, 1892–1893–1894. Respectfully submitted to the nineteenth century civilization in the land of the free and the home of the brave.*

Wells, R. S., & Stage, F. K. (2015). Past, present, and future of critical quantitative research in higher education. *New Directions for Institutional Research, 2014*(163), 103–112.

Wilkerson, I. (2020). *Caste: The Origins of Our Discontents*. Random House.

Williamson, A. K., Luke, B., Furneaux, C. (2020). Ties that bind: Public foundations in dyadic partnerships. *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, 1–13.

Zhu, J., Huang, H., & Zhang, D. (2019). Big Tigers, Big Data: Learning social reactions to China's anticorruption campaign through online feedback. *Public Administration Review, 79*(4), 500–513.

Zuberi, T., & Bonilla-Silva, E. (Eds.). (2008). *White logic, white methods: Racism and methodology*. Rowman & Littlefield Publishers.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.