an open access journal

Citation: Geiger, R. S., Cope, D., Ip, J., Lotosh, M., Shah, A., Weng, J., & Tang, R. (2021). "Garbage in, garbage out" revisited: What do machine learning application papers report about human-labeled training data? *Quantitative Science Studies*, *2*(3), 795–827. https://doi.org/10.1162/qss_a_00144

DOI:
https://doi.org/10.1162/qss_a_00144

Peer Review:
https://publons.com/publon/10.1162/qss_a_00144

Supporting Information:
https://doi.org/10.1162/qss_a_00144

Received: 18 February 2021
Accepted: 8 June 2021

Corresponding Author:
R. Stuart Geiger
stuart@stuartgeiger.com

Handling Editor:
Ludo Waltman

The MIT Press

RESEARCH ARTICLE

"Garbage in, garbage out" revisited: What do machine learning application papers report about human-labeled training data?

R. Stuart Geiger[1*], Dominique Cope[2], Jamie Ip[2], Marsha Lotosh[3*], Aayush Shah[2], Jenny Weng[2], and Rebekah Tang[1*]

[1]University of California, San Diego
[2]University of California, Berkeley
[3]Webster Pacific
*The majority of the work on this paper was conducted when this author was affiliated with the University of California, Berkeley.

Keywords: bias, data reporting, documentation, machine learning, research reporting, training data

ABSTRACT

Supervised machine learning, in which models are automatically derived from labeled training data, is only as good as the quality of that data. This study builds on prior work that investigated to what extent "best practices" around labeling training data were followed in applied ML publications within a single domain (social media platforms). In this paper, we expand by studying publications that apply supervised ML in a far broader spectrum of disciplines, focusing on human-labeled data. We report to what extent a random sample of ML application papers across disciplines give specific details about whether best practices were followed, while acknowledging that a greater range of application fields necessarily produces greater diversity of labeling and annotation methods. Because much of machine learning research and education only focuses on what is done once a "ground truth" or "gold standard" of training data is available, it is especially relevant to discuss issues around the equally important aspect of whether such data is reliable in the first place. This determination becomes increasingly complex when applied to a variety of specialized fields, as labeling can range from a task requiring little-to-no background knowledge to one that must be performed by someone with career expertise.

1. INTRODUCTION

Supervised machine learning (ML) is now widely used in many fields to produce models and classifiers from training data, which allows for automation of tasks such as diagnosing medical conditions (Shipp, Ross et al., 2002; Ye, Qin et al., 2003), identifying astronomical phenomena (Ball & Brunner, 2010; Fluke & Jacobs, 2020), classifying environmental zones (Lary, Alavi et al., 2016; Ma, Li et al., 2017), or distinguishing positive versus negative sentiment in documents (Prabowo & Thelwall, 2009; Ravi & Ravi, 2015; Thelwall, Buckley et al., 2010). Applying supervised ML requires labeled training data for a set of entities with known properties (called a "ground truth" or "gold standard"), which is used to create a classifier that will make predictions about new entities of the same type.

"Garbage in, garbage out" is a classic saying in computing about how problematic input data or instructions will produce problematic outputs (Babbage, 1864; Mellin, 1957), which is especially relevant in ML. Yet data quality is often less of a concern in ML research and education, with these issues often passed over in major textbooks (e.g., Friedman, Hastie, & Tibshirani, 2009; Goodfellow, Bengio, & Courville, 2016; James, Witten et al., 2013). Instead, the focus is typically on the domain-independent mathematical foundations of ML, with ML education and research often using clean, tidy, and prelabeled "toy" data sets. While this may be useful for theoretically oriented basic ML research, those applying ML in any given domain must also understand how low-quality or biased training data threatens the validity of the model (Buolamwini & Gebru, 2018; Dastin, 2018; Geiger, Yu et al., 2020; Obermeyer, Powers et al., 2019).

In this paper, we empirically investigate and discuss a wide range of issues and concerns around the production and use of training data in applied ML. Our team of seven labelers systematically examined published papers that applied supervised ML to a particular domain, sampling from three sets of academic fields: life and biomedical sciences; physical and environmental sciences; and social sciences and humanities. For each paper, we asked up to 15 questions about how the authors reported using supervised ML and how they reported obtaining the labeled training data used to produce the model or classifier. We particularly focus on human-labeled or human-annotated training data, in which one or more individuals make discrete assessments of items. Given that many issues and biases can emerge around human labeling, we examine whether papers reported following best practices in human labeling.

Our project is based on the methodology of structured content analysis, which seeks to systematically turn qualitative phenomena into categorical and quantitative data (Riff, Lacy, & Fico, 2013). We draw on and situate our study within the growing efforts to bridge the fields of qualitative and quantitative science studies (Leydesdorff, Ràfols, & Milojević, 2020; Bowker, 2020; Cambrosio, Cointet, & Abdo, 2020; Kang & Evans, 2020). Quantitative science studies often examines the outputs of science, such as analyzing bibliometrics and other already quantitative trace data to understand how scientists' final products have been received within science and other institutions. In contrast, qualitative science studies often examines the research process "in action" (Latour, 1987) to investigate how science is produced, such as using more ethnographic or historical methods. This project is in between these two traditions: Our method involves systematically quantifying information from qualitative sources, rather than using already-quantitative trace data; we examined a broad set of publications from across domains, rather than more in-depth case studies; and we analyzed and quantified information about research practices, rather than how publications are cited.

As our research project was a human-labeling project studying other human-labeling projects, we took care in our own practices. Before the research project began, we detailed all questions and valid responses, developed instructions with examples, and had a discussion-based process of reconciling disagreements. Another key issue in data labeling is that of construct validity and operationalization (Jacobs & Wallach, 2019): Is the labeling process actually capturing the theoretical construct that the authors are claiming to capture? In our study, we only have access to the paper reporting about the study and not the actual study or data set itself. This means that our fundamental unit of analysis must be what the papers report, even though our broader intent is to understand what the study's authors and labelers actually did. Many papers either do not discuss such details at all or lack sufficient detail to make a determination. For example, many papers did note that the study involved the creation of an original human-labeled data set, but did not specify who labeled it. For some of our items, one of the most common labels we gave was "no information." This is a concerning issue, given how crucial such information is in understanding the validity of the training data set, and by extension, the validity of the classifier.

2. LITERATURE REVIEW AND MOTIVATION

2.1. The Problem with Low-Quality and Biased Training Data

Curating high-quality training data sets for ML involves skill, expertise, and care, especially when items are individually labeled by humans. There can be disastrous results if training data sets are taken as a gold standard when they should not be. Supervised ML models are typically evaluated exclusively using a held-out subset of the original training data set, making systematic flaws in a training data set difficult to identify or audit within the traditional paradigm of ML. These concerns are particularly pressing when ML is used for deeply subjective and politicized decisions, such as in finance, hiring, welfare, and criminal justice. Many ML training data sets have been found to be systematically biased along various axes, including race and gender, which impacts the accuracy of those ML models (e.g., Buolamwini & Gebru, 2018). In other cases, more subtle issues arise around labels, such as a paper claiming to have produced an ML classifier distinguishing criminals from noncriminals using only facial images, with allegedly overwhelmingly high accuracy (Wu & Zhang, 2016). As Bergstrom and West (2020) critique, their labels were problematically derived from the source of the photos: Criminals were taken from prison mug shots, while noncriminals were taken from professional social network profiles. Because people generally do not smile in mug shots but do smile in profile photos, Bergstrom and West argue that the original team effectively built a smile classifier, but claimed it was a criminality classifier.

In another domain, an exposé (Dastin, 2018) reported that Amazon built an internal ML system for hiring that was later scrapped after it was determined to have substantial gender biases. The training data set used was based on hiring managers' past decisions, where resumes from those hired were given one label and those not hired were given another label. The classifier was thus trained to approximate years of past decisions, and given that Amazon has had significant gender gaps in their workforce (like many tech companies), this meant that such systematic biases were reinforced and rationalized through ML. This is the case even though gender was intentionally excluded as a feature in the model, as the classifier used other features that were a proxy for gender to more closely approximate the biases in the training data. Had the training data been a new data set labeled by a diverse team of trained HR professionals tasked with evaluating resumes with a focus on nondiscrimination, this might have produced a quite different classifier.

ML in the field of medicine is poised for explosive growth, although critics raise similar concerns about training data. Medical privacy risks arise for patients whose health care records may be used in formulating a training data set (Vayena, Blasimme, & Cohen, 2018). Furthermore, there is evidence of biases in health care applications of ML, and in some instances, the consequences of biases may directly impact patients' survival. One study in the United States labeled patients' medical records with their severity of illness, using a proxy variable that ostensibly required little human judgment: the cost of the patient's health care. Yet when this data was used to train a classifier, it caused significant bias against African American patients, who historically have had differential access to medical care (Obermeyer et al., 2019). The medical field itself is encountering new questions surrounding human labeling and annotation. For example, one widely used application is the interpretation of medical imaging. The human who labels MRI images as cancerous or not-cancerous must have specific expertise compared someone who labels product reviews as positive or negative. Meta-research in radiology has found practicing radiologists have about a 3–5% error rate (Brady, 2016), which raises the question of whether radiology training data sets should be independently labeled by multiple experts to ensure data quality. Finally, as with many fields, the introduction of ML using pre-existing data from a particular environment and setting has the distinct potential to reproduce and perpetuate existing systemic biases, especially when that classifier is deployed to a different environment and setting (DeCamp & Lindvall, 2020).

2.2. "Garbage in, Garbage out" Version 1

This project is heavily based on a prior study (Geiger et al., 2020), which similarly had a team of labelers examine issues around training data in a random sample of published papers. That study examined a narrow subset of peer-reviewed and preprint papers in a specific field: applied ML papers trained on Twitter data. They looked for 13 pieces of information in each paper, which they argued were important to understanding the validity of the training data labeling process. This included if the data was human or machine labeled, who the labelers were, how many labelers rated each item, and rates of interrater reliability (if multiple labelers rated each item). The study found a wide divergence both in the level of information reported and in adherence to best practices in human labeling. For example, of papers reporting a new human-labeled training data set, about 75% gave some information about who the labelers were, 55% specified the number of labelers, 11% released the training data set itself, and 0% reported how much crowdworkers were paid for their work.

We expanded on Geiger et al.'s study, drawing heavily from their published questions and protocols. We followed the same general process of having labelers rate each item independently, then reconciled disagreements through a discussion led by the team leader. We made some small modifications and extensions to the questions, which were recommended by the original authors for future work or were better suited to the expanded scope. We added questions about the field/domain of the paper and about the reconciliation process when multiple labelers labeled each item. We also rewrote some of the labeling instructions, label categories, and provided examples, often to clarify ambiguities.

2.3. Best Practices in Human Labeling of Training Data

Geiger et al. (2020) give a substantial review of existing work around human labeling of training data, including an extensive discussion of best practices in this work. They argue that much of the labeling work for ML is a form of structured content analysis, which is a methodology long used in the humanities and social sciences to turn qualitative or unstructured data into categorical or quantitative data. This involves teams of "coders" (also called "annotators," "labelers," or "reviewers") who "code," "annotate," or "label" items individually. (Note that we use "label/labeler" in this paper, although we began with using "annotate/annotator," which is still present in some of our data and protocols.) One textbook describes content analysis as a "systematic and replicable" (Riff et al., 2013, p. 19) method with established best practices, as Geiger et al. summarize:

> A "coding scheme" is defined, which is a set of labels, annotations, or codes that items in the corpus may have. Schemes include formal definitions or procedures, and often include examples, particularly for borderline cases. Next, coders are trained with the coding scheme, which typically involves interactive feedback. Training sometimes results in changes to the coding scheme, in which the first round becomes a pilot test.

> Then, labelers independently review at least a portion of the same items throughout the entire process, with a calculation of "inter-rater reliability" (IRR) or "inter-annotator agreement" (IAA). Finally, there is a process of "reconciliation" for disagreements, which is sometimes by majority vote without discussion and other times discussion-based. (Geiger et al., 2020, p. 2–3)

Structured content analysis is a difficult task, requiring both domain-specific expertise about the phenomenon to be labeled and domain-independent expertise to manage teams of labelers. Historically, undergraduate students have often performed such work for academic researchers. With the rise of crowdwork platforms such as Amazon Mechanical Turk, crowdworkers are often used for data labeling tasks. New software platforms have been developed to support more microlevel labeling and annotation or labeling at scale, including in citizen science (Bontcheva, Cunningham et al., 2013; Chang, Amershi, & Kamar, 2017; Nakayama, Kubo et al., 2018; Pérez-Pérez, Glez-Peña et al., 2015). For example, the Zooniverse (Simpson, Page, & De Roure, 2014) provides a common platform for citizen science projects across domains, where volunteers label data under scientists' direction.

2.4. Meta-Research and Methods Papers in Linguistics and NLP

We also draw inspiration from metaresearch and standardization efforts in Linguistics and Natural Language Processing (NLP) (Bender & Friedman, 2018; McDonald, Schoenebeck, & Forte, 2019). These fields have developed extensive literatures on standardization and reliability of linguistic labels, including best practices for corpus annotation (e.g., Doddington, Mitchell et al., 2004; Hovy & Lavid, 2010; Linguistic Data Consortium, 2008). In Geiger et al.'s (2020) study, the publisher with the highest information score was the Association for Computational Linguistics. There has been much work in linguistics and NLP around these issues, such as Sap et al.'s study of racial bias among labelers (Sap, Card et al., 2019). Blodgett et al. conducted a content analysis of how 146 NLP researchers discuss "bias" and found that while this has become a prominent topic in NLP, papers' discussions of motivations and methods around bias "are often vague, inconsistent, and lacking in normative reasoning" (Blodgett, Barocas et al., 2020, p. 5,454). There is also related work in methods papers focused on identifying or preventing "low-effort" responses from crowdworkers (Mozetič, Grčar, & Smailović, 2016; Raykar & Yu, 2012; Soberón, Aroyo et al., 2013), which raise issues around fair labor practices and compensation (Silberman, Tomlinson et al., 2018).

2.5. The Open Science, Reproducibility, and Research Integrity Movements

Two related movements in computationally supported knowledge production have surfaced issues around documentation. First, open science is focused on broader availability to the products of research and research infrastructure, including open access to publications, software tools, data sets, and analysis code (Fecher & Friesike, 2014). The related reproducibility movement calls for researchers to make protocols, data sets, and analysis code public, often focusing on what others need to replicate the original study (Kitzes, Turek, & Deniz, 2018; Wilson, Bryan et al., 2017). Such requirements have long been voluntary, with few incentives to be a first mover, but funding agencies and publications are increasingly establishing such requirements (Gil, David et al., 2016; Goodman, Pepe et al., 2014).

One notable effort is around formally specifying what each author of a paper actually did, which has long been standard in medical journals (Rennie, Flanagin, & Yank, 2000). Author role documentation has gained popularity with the more recent Contributor Roles Taxonomy

Project (or CRediT) (Brand, Allen et al., 2015). CRediT declarations are increasingly required by journals, which has led to novel quantitative science studies research (Larivière, Pontille, & Sugimoto, 2020). We also draw inspiration from work about capturing information in ML data fows and supply chains (Gharibi, Walunj et al., 2019; Schelter, Böse et al., 2017; Singh, Cobbe, & Norval, 2019) and developing tools to support data cleaning (Krishnan, Franklin et al., 2016; Schelter, Lange et al., 2018). We note that this work has long been part of library and information science, particularly in Research Data Management (Borgman, 2012; Medeiros & Ball, 2017; Sallans & Donnelly, 2012; Schreier, Wilson, & Resnik, 2006). There is much more work to be done on quantitatively studying issues around research integrity (Silberman et al., 2018; Zuckerman, 2020), which institutionally has often been limited to more egregious and blatant cases of plagiarism and fabrication.

2.6. Fairness, Accountability, and Transparency in ML

Within the field of ML, there is a growing movement in the Fairness, Accountability, and Transparency (or FAccT) subfield, with many recent papers proposing training data documentation in the context of ML. Various approaches and metaphors have been taken in this area, including "datasheets for datasets" (Gebru, Morgenstern et al., 2018), "model cards" (Mitchell, Wu et al., 2019), "data statements" (Bender & Friedman, 2018), "nutrition labels" (Holland, Hosny et al., 2018), a "bill of materials" (Barclay, Preece et al., 2019), "data labels" (Beretta, Vetrò et al., 2018), and "supplier declarations of conformity" (Hind, Mehta et al., 2018). Many go far beyond the concerns we have raised around human-labeled training data, as some are also (or primarily) concerned with documenting other forms of training data, model performance and accuracy, bias, considerations of ethics and potential impacts, and more. Our work is strongly aligned with this movement, as we seek to include data labeling within these areas of concern. However, as we discuss in our conclusion, a single one-size-fits-all standard may be necessary but not sufficient to address concerns of fairness and bias.

We also call attention to those developing methods for "de-biasing" ML, which is a fast-moving and contentious research area (for surveys and comparative work, see Mehrabi, Morstatter et al., 2019 and Friedler, Scheidegger et al., 2019). Much of this work is in developing domain-independent fairness metrics for evaluating trained models (e.g., Hardt, Price, & Srebro, 2016; Zafar, Valera et al., 2017), which are used to modify trained models or predictions (e.g., Amini, Soleimany et al., 2019; Karimi Mahabadi, Belinkov, & Henderson, 2020). However, other work has approached these issues more as a problem of data set preprocessing (Calmon, Wei et al., 2017) or database repair (Salimi, Howe, & Suciu, 2020). Critics note that domain-independent approaches may fall into what Selbst, Boyd et al. (2019, p. 60) identify as "abstraction traps," such as failing to account for the particularities of different kinds and qualities of discrimination in a given social context—a critique Hanna, Denton et al. (2020) make of fairness research that treats race as a single fixed attribute. We did not ask any questions about how papers discuss de-biasing or data cleaning due to the large number of questions we were already asking and the novelty of such approaches, but these concerns are deeply related.

3. DATA AND METHODS

3.1. Data: ML Papers Performing Classification Tasks

Our goal was to find a corpus of papers using supervised ML across disciplines and application domains, including papers producing an original labeled data set using human labeling. We used the Scopus bibliographic database (Baas, Schotten et al., 2020), which contains about

Table 1. Summary of sampling across all three corpora

Corpus	Papers in corpus	# randomly sampled	% sampled
Social Sciences & Humanities	5,346	70	1.30
Life & Biomedical Sciences	9,507	60	0.63
Physical & Environmental Sciences	11,030	70	0.63
Total	25,883	200	0.77

40,000 publications that a review board has verified for various qualities, including being peer reviewed, regularly published for at least 2 years, and governed by a named editorial board of experts. We searched for journal articles and conference proceedings from 2013 to 2018 where the title, abstract, or keywords included “machine learning” and either “classif*” or “supervi*” (case-insensitive). We ran three stratified samples across Scopus’s Subject Area classifications[1]: Physical Sciences (which includes engineering and earth/ecological sciences); Social Sciences & Humanities (a single category); and Life Sciences & Health Sciences (two categories, which we combined). Table 1 describes our sampling. More details about the corpora are in the appendix, which is available as supplementary materials and in our data repository (see Section 3.4).

3.2. Labeling Team, Training, and Workflow

Our labeling team included one research scientist who led the project (RSG) and undergraduate research assistants, who worked 6–10 hours per week for course credit as part of a university-sponsored research experience program (DC, JI, ML, AS, JW, and RT). The project began with six students for one semester, five of whom continued on the project for the second semester. All students had some coursework in computer science and/or data science, with a range of prior experience in ML in both a classroom and applied setting. Students’ majors and minors included Electrical Engineering & Computer Science, Data Science, Statistics, Economics, Linguistics, and Biology. For the first four weeks, the team leader trained the students in both broader ML concepts and the specific questions to be answered for this project. The team first labeled and discussed a practice set of 40 papers sampled from across the three corpora, which were not included in the final data set. In these initial weeks, the team learned the coding schema and the reconciliation process, which were further refined.

Following this training, the labeling workfow was that each week, a set of papers were randomly sampled from one corpus, typically between 10–15 papers. The students independently reviewed and labeled the same papers, using different web-based spreadsheets to record labels. The team leader synthesized labels and identified disagreement. The team met in person or by videochat to discuss the week’s cases of disagreement. The team leader explained various issues in question and built a consensus about the proper label (as opposed to purely majority vote). The team leader had the final say when a consensus could not be reached.

All 200 papers were labeled by at least four labelers; one labeled 137 items and another labeled 100 items. Following the first round of labeling and reconciliation, we conducted a

[1] https://web.archive.org/web/20200812203800/https://service.elsevier.com/app/answers/detail/a_id/15181/supporthub/scopus/

second round of verification. Where there was any initial disagreement on labels in the first round, each paper was re-examined and discussed by at least two labelers and the team leader. The second round began multiple months after the first round, meaning that there was at least one month between when each paper was examined and re-examined. If there was still disagreement, the final decision was made by the team leader. The team leader did a final check to review every label for all 200 papers.

3.3. IRR and Labeled Data Quality

All human labeling projects that involve multiple labelers should evaluate the intersubjective reliability of the labeling process (Tinsley & Weiss, 1975). We present IRR metrics using three metrics. For all metrics, we recoded "unsure" and blank responses to both be blank (NaN), but treated "N/A" answers as a distinct judgment. First, we calculated mean total agreement, or the proportion of items where all labelers initially gave the same label before reconciliation, but not counting blank and unsure responses. As Table 2 shows, this is a more stringent metric: All nonblank/unsure responses must be the same for an item to have a 1 score, otherwise the score is 0. Second, we present the mean percentage correct rate, which is the proportion of labelers who initially gave the same label arrived after discussion and reconciliation, but also not counting blank and unsure responses. As Table 2 shows, this is a more forgiving metric: If five out of six labelers give the same final correct label, the score is 0.83 for that item. For these two metrics, we calculated per-question scores by taking the mean of all scores for an item.

We also present the widely used Krippendorff's alpha (Krippendorff, 1970) metric, although we strongly advise against relying on it. Our data does not meet the statistical assumptions for both Fleiss's kappa and Krippendorf's alpha, which are popular because they support missing labels for 3+ labelers and take into account the possibilities that raters made decisions based on random chance. However, this requires assuming a uniform prior possibility of such a random distribution, which generally only applies if each possible response by raters is equally likely. Rates can be dramatically lower when there is a highly skewed distribution of response categories (Quarfoot & Levine, 2016; Oleinik, Popova et al., 2014). Our data set has highly skewed distributions, especially for many of the more specialized questions, which lead to minuscule scores for some questions with especially skewed distributions (e.g., prescreening for crowdwork; reported IRR).

Table 3 presents both our custom metrics and Krippendorff's alpha for all questions. Mean total agreement rates ranged from 37.5% to 66%, with an average of 48.0% across all questions. Mean percentage correct rates ranged from 65.4% to 85.8%, with an average of 73.1% across all questions. Some questions that had lower rates (especially for mean total agreement)

Table 2. Example of IRR calculations for sample rows

#	Labeler #1	Labeler #2	Labeler #3	Labeler #4	Labeler #5	Labeler #6	Final/correct label	Total agreement	Mean percentage correct
1	yes	unsure	yes	yes	yes	[blank]	yes	1	1
2	yes	yes	yes	yes	no	yes	yes	0	0.83
3	no	n/a	yes	no	yes	[blank]	yes	0	0.4
4	yes	no	yes	[blank]	yes	unsure	no	0	0.25

Table 3. Interrater reliability metrics per question

Question	Mean total agreement	Mean percentage correct	Krippendorff's alpha
Original classification task	66.0%	84.8%	0.670
Classifier area/domain	34.5%	65.4%	0.520
Labels from human judgment	37.5%	68.2%	0.517
Human labeling for training data	46.5%	77.3%	0.517
Used original human labeling	43.5%	71.0%	0.498
Original human labeling source	43.5%	71.1%	0.330
Prescreening for crowdwork	58.5%	84.2%	0.097
Labeler compensation	46.0%	68.0%	0.343
Training for human labelers	48.0%	70.0%	0.364
Formal instructions	47.5%	66.8%	0.337
Multiple labeler overlap	48.5%	69.3%	0.370
Synthesis of labeler overlap	53.0%	83.4%	0.146
Reported interrater reliability	55.5%	85.8%	0.121
Total number of human labelers	50.5%	69.3%	0.281
Median number of labelers per item	48.5%	69.3%	0.261
Link to data set available	41.0%	66.1%	0.322
Average across all questions	48.0%	73.1%	0.356
Median across all questions	48.0%	70.0%	0.343

were due to a labeler making an incorrect assessment on an earlier question, which determines whether they answer subsequent questions or mark them as "N/A."

In interpreting these metrics, we note that the standard approach of human labeling checked by IRR metrics treats individual humans as scientific instruments that turn complex phenomena into discrete structured data. If there is a high degree of IRR, then reconciliation can easily take place through a majority vote process involving no discussion, or if rates are quite high, then many researchers assume they can use just one of those human labelers per item in future work. These rates were not high enough for us to have confidence that we could have a purely quantitative/majority-vote reconciliation process, much less a process of only using one labeler per item. However, these rates are sufficient to show there is enough agreement to proceed to a discussion-based reconciliation process and a final check of all items by the team leader. As McDonald et al. (2019) discuss, standardized IRR metrics such as Krippendorf's alpha are useful in highly structured labeling projects that do not have a discussion-based reconciliation process, as they only evaluate the agreement of independent initial labels. Such metrics would be more essential to the validity of our study if we were conducting a quantitative, majority-rule reconciliation process or if only a subset of items were reviewed by multiple labelers. We included mean percentage correct rates to partially include the reconciliation and verification process.

Furthermore, our approach was largely focused on identifying the presence or absence of various kinds of information within long-form publications. This is a different kind of human judgment than is typically involved in common tasks using human labeling for ML (e.g., labeling a single social media post for positive/negative sentiment) or traditional social science and humanities content analysis (e.g., categorizing newspaper articles by topic). Our items were full research publications with many pages of detail, which followed many different field-specific conventions and genres. Our labelers were looking for up to 15 different kinds of information per paper, each of which could be found anywhere in the paper. We reflected that in our reconciliation process, most of the time when labelers disagreed, it was because some had caught a piece of information in the paper that others had not seen. Once that information was brought to the group, it was most often the case that some labelers said that they had missed that information and changed their response. It was less common for our team to have disagreements arising from two labelers differently interpreting the same text, especially after the first few weeks. For such reasons, we are relatively confident that if, after our process, no individual member of our team has identified the presence of such information, then it is quite likely not present in the paper.

3.4. Software, Data Sets, and Research Materials

We used Google Sheets to enter labels. For computational analysis and scripting for corpus collection, data management, and data analysis, we used Python 3.7 (van Rossum, 1995), using the following libraries: Pandas dataframes (McKinney, 2010) for data parsing and transformation; SciPy (Jones, Oliphant et al., 2001) and NumPy (van der Walt, Colbert, & Varoquaux, 2011) for quantitative computations; Matplotlib (Hunter, 2007) and Seaborn (Waskom, Botvinnik et al., 2018) for visualization; and SimpleDorff (Perry, 2020) for IRR calculations. Analysis was conducted in Jupyter Notebooks (Kluyver, Ragan-Kelley et al., 2016) using the IPython (Pérez & Granger, 2007) kernel.

Data sets, analysis scripts, labeling instructions, and other supplementary information can be downloaded from GitHub[2] and Zenodo[3]. Data sets include all labels from all labelers for the first round of independent labeling and the consolidated set of final labels and scores for all items. Paper URLs/DOIs have been anonymized with a unique salted hash. Analysis scripts are in Jupyter Notebooks and can be explored and modified in any modern web browser using the cloud-based MyBinder.org (Project Jupyter, Bussonnier et al., 2018)[4].

4. FINDINGS

Figure 1 shows a summary of results. For this figure, we recoded (or consolidated) some questions with many answers to reflect whether the paper reported an answer to that question. For example, for "original human labeling source," any answer that specified a source is "yes," while "no information" is "no." This is also how we calculated paper information scores in Section 5. Figure 1 illustrates how we asked more detailed questions for papers based on answers to prior questions. For example, 103 papers used labels from human judgment—either

[2] https://github.com/staeiou/gigo_qss_2021

[3] https://doi.org/10.5281/zenodo.4906636

[4] https://mybinder.org/v2/gh/staeiou/gigo_qss_2021/HEAD

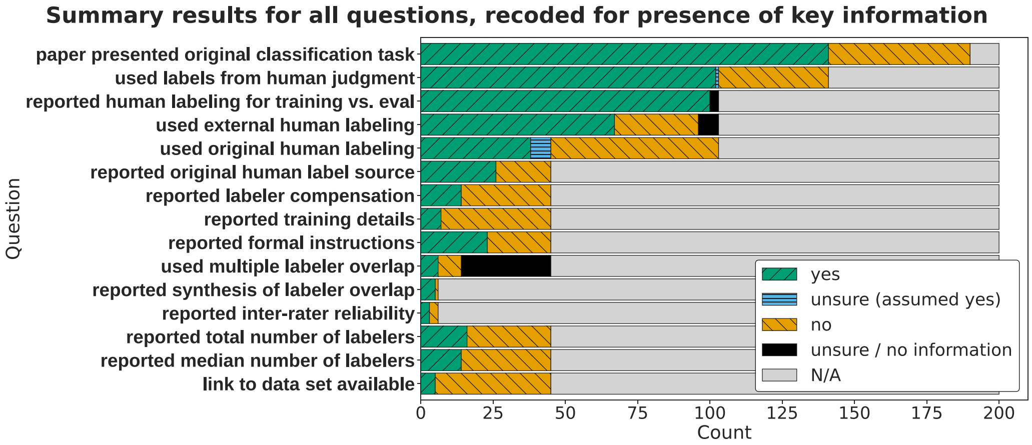

Figure 1. Summary of results. Note that some questions have been recoded to show the presence or absence of information.

″yes″ or ″unsure (assumed yes)″—and the next three questions were answered for those 103 papers. The remaining 10 questions were answered for the 45 papers that could be assumed to use original human labeling, with two of those questions only answered for the six papers involving multiple labeler overlap.

4.1. Original ML Classification Task

The first question was whether the paper was conducting an original classification task using supervised ML. Our keyword-based process of generating the corpus included some papers that used ML keywords but were not actually presenting a new ML classifier. However, defining the boundaries of supervised ML and classification tasks is difficult, particularly for papers that are long, complex, and ambiguously worded. We defined ML broadly: any automated process that does not exclusively rely on explicit rules, in which the performance of a task increases with additional data (Mitchell, 1997, p. 2). We decided that this can include simple linear regressions, although there is much debate about if and when simple linear regressions are a form of ML. However, as we were also looking for classification tasks, linear regressions were only included if they were used to make a prediction in a set of defined classes. We defined an "original" classifier to mean a classifier that the authors made based on new or old data, which excludes the exclusive use of pretrained classifiers or models. We found that some papers claimed to be using ML, but when we examined the details, these did not fall under our definition.

As Table 4 shows, the majority of papers in our data set were involved in an original classification task. We placed 10 papers in the "N/A" category—meaning they did not give enough detail for us to determine, were not in English, were not able to be accessed, or were complex boundary cases.

4.2. Classifier Area/Domain

The next question categorized the paper into one of eight fields/areas of study. We had sampled three broad disciplinary categories (Social Sciences & Humanities, Biomedical & Life Sciences, and Physical & Environmental Sciences), which are determined by Scopus on a per-journal/conference level. We made these area/domain determinations based on the paper's content, without consulting the Scopus-provided category. As Table 5 shows, our data

Table 4. Is the paper presenting an original/newly created ML classifier?

	Count	Proportion
Yes	141	70.50%
No	49	24.50%
N/A (paper ineligible or inaccessible)	10	5.00%
Total	200	100.00%

set contained a wide variety of ML application fields. Medical papers had the plurality of responses, followed by Linguistic; then papers from Biological, Physical, Soft/hardware, and Geo/ecological had similar sizes.

4.3. Labels from Human Judgment

While all approaches to curating training data involve some kind of human judgment, this question focused on cases where humans made discrete judgments about a set of specific items, which were then turned into labels for training data. More than a quarter of the papers in our corpora used some form of automation, scripting, or quantitative thresholds to label items. For example, one boundary case used medical records to label patients with or without high blood pressure (hypertension). We decided that if a medical practitioner made a diagnosis that researchers used as the label, it was human labeled. If the researchers set a quantitative threshold for high blood pressure, then parsed medical records for blood pressure readings with a script, it was not human labeled. In addition, individual human labeling could be done for all of the paper's training data (the typical case) or only a portion. For example, some authors reported using scripts or thresholds to label some items (e.g., the "easy" cases) then labeled the remaining items manually.

Table 5. Classifier area/domain

	Count	Proportion
Medical	43	30.50%
Linguistic	24	17.02%
Biological (nonmedical)	17	12.06%
Physical	14	9.93%
Soft/hardware	14	9.93%
Geo/ecological	13	9.22%
Activities and actions	7	4.96%
Demographic	5	3.55%
Other	4	2.84%
Total of applicable papers (presenting original ML classifier)	141	100.00%
Nonapplicable papers	59	–

Table 6. Were labels derived from humans making discrete judgments of items?

	Count	Proportion
No/Machine-labeled	38	26.95%
Yes for all items	53	37.59%
Yes for some items	10	7.09%
No information (implicit yes)	39	27.66%
Unsure (but assumed yes)	1	0.71%
Subtotal: papers assumed to use human labeled-data	103	73.05%
Total of applicable papers (presenting original ML classifier)	141	100.00%
Nonapplicable papers	59	–

In some instances, we determined that answer could be an "implicit yes" if ample evidence indicated a particular labeling method that most likely used humans at some point, but it was not explicitly stated by the authors. For example, many medical papers reported using diagnoses from a patient's medical records as labels. Some of these papers gave substantial detail about who originally made the diagnosis and even what diagnostic criteria were used, while others generated labels based on medical records and did not explicitly state that a human (e.g., a medical practitioner) made the diagnosis. If we could reasonably assume a human was involved in the original diagnosis, we generally labeled the second type of papers as "no information (implicit yes)." One paper was far less clear about the source of the data than other "implicit yes" papers, such that we labeled it "Unsure." However, we included the paper in subsequent questions because we felt we could answer subsequent questions about it, which reused externally obtained data for labeling.

As Table 6 shows, the second highest response are papers that do not clearly state whether their labeling was performed by a human or a machine, but contained enough contextual details for us to be reasonably confident in assuming that human labeling was used. Note that this question was originally titled "Labels from human annotation" throughout the labeling and reconciliation process, but was renamed in the analysis stage to better reflect the instructions.

4.4. Human Labeling for Training Versus Evaluation

This question and all subsequent questions were only applicable to papers that involved human labeling, which had "yes" or "implicit" designations to the previous question. This allowed for further specification of human labeled data usage within each publication. As Table 7 shows, human labeling for training data is the typical case, where labels are created and then used to train the classifier. Often part of this data is held out as a test set to evaluate the classifier. Human labeling for evaluation only is when the authors of the paper train the classifier using nonhuman-labeled data, but use humans to evaluate the validity of either that data set or the classifier. The overwhelming majority of papers took the more standard approach of using labels as training data, but a few did have human evaluation of classifiers trained with machine-labeled data. This question had lower rates of "unsure," where the paper did not give enough information to make a determination.

Table 7. Was human-labeled data used for training data or to evaluate a classifier trained on nonhuman-labeled training data?

	Count	Proportion
Human labeling for training data	94	91.26%
Human labeling for evaluation only	6	5.83%
Unsure	3	2.91%
Total of applicable papers (assumed to use human-labeled training data)	103	100.00%
Nonapplicable papers	97	–

4.5. Original and/or External Human-Labeled Data

Our next question was about whether papers that used human labeling used original human labeling, which we defined as a process in which the paper's authors obtained new labels from human judgments for items. This is in contrast to externally obtained data, which involves reusing existing private or public data sets of human judgments. Table 8 shows that most of the papers in our corpus that used labels from human judgment were reusing externally labeled data. Our assumption behind this question is that papers that rely on existing data sets may have less of a burden to discuss the details around the labeling process in the paper itself, as readers could review the cited paper for such details. In some cases, external and original human labeling were combined, such as if authors reused a existing labeled data set and then further labeled it for additional information.

Like the prior question, this question had lower rates of "unsure/no information" where the paper did not give enough information to make a determination. We note that for all of the papers we labeled as ″unsure/no information″ we had enough contextual or implicit information to assume that it was not a reused/externally labeled data set. This means that the total number of papers we assume to include at least some original human labeling is 45.

4.6. Summary of ML Papers' Approaches to Training Data

We synthesized responses to the prior questions to summarize the general breakdown of applied ML publications' approach to their data. Out of the 141 papers in our sample that

Table 8. Did authors reuse an existing human-labeled data set (external), create a new human-labeled data set (original), or both?

	Count	Proportion
Only external	58	56.31%
Only original	29	28.16%
Original and external	9	8.74%
Unsure/no information (but can assume original)	7	6.80%
Subtotal: assumed to include some original human labeling	45	43.69%
Total of applicable papers (assumed to use human-labeled training data)	103	100.00%
Nonapplicable papers	97	–

Table 9. Approach to training data by corpus: count (proportion). Totals may not equal 100% due to rounding.

	Life Sciences & Biomedical	Physical & Environmental Sciences	Social Sciences & Humanities	All corpora
Original human-labeled data	12 (26.7%)	13 (25.0%)	13 (29.5%)	**38 (26.95%)**
External human-labeled data	20 (44.4%)	20 (38.5%)	18 (40.9%)	**58 (41.1%)**
Machine-labeled data	12 (26.7%)	15 (28.8%)	11 (25.0%)	**38 (26.95%)**
Unsure	1 (2.2%)	4 (7.7%)	2 (4.5%)	**7 (5.0%)**
Subtotal: ML classifier papers	*45 (100%)*	*52 (100%)*	*34 (100%)*	***141 (100%)***
(No ML classifier/NA)	*15*	*18*	*26*	***59***
Grand total	**60**	**70**	**60**	**200**

presented an original ML classifier, 27% used machine-labeled data (either by the authors or from a reused data set), 41% used an existing human-labeled data set, 27% produced a novel human labeled data set, and 5% did not provide enough information for us to answer. Table 9 and Figure 2 present these results by corpus, which show few differences at this level.

4.7. Original Human Labeling Source

Our next question asked who the labelers were for the 45 papers that used original human labeling. As Table 10 shows, we found a diversity of approaches to the recruitment of human labelers. The plurality of papers gave no information about who performed their labeling task. The "survey/self-reported" category refers to papers that have individuals label data they generated, which included surveys as well as studies such as those using motion tracking, where subjects recorded performing different physical gestures. In contrast to Geiger et al.'s prior findings about papers that used Twitter data, none of the papers in our data set reported using

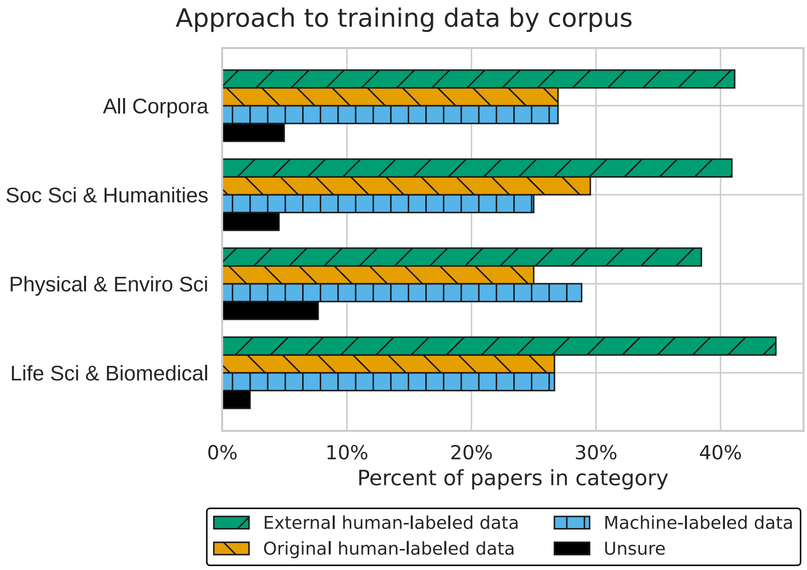

Figure 2. Approach to training data by corpus, excluding ineligible papers.

Table 10. Who were the humans doing the labeling work?

	Count	Proportion
Paper's authors	10	22.22%
No information	19	42.22%
Other with claim of expertise	9	20.00%
Other no claim of expertise	2	4.44%
Survey/self-reported	5	11.11%
Total of applicable papers (involving original human labeling)	45	100.00%
Nonapplicable papers	155	–

crowdworking platforms. We did not consider volunteer citizen science crowdsourcing platforms to be crowdworking.

4.8. Labeler Compensation

The next question inquired as to if and what type of compensation was offered to labelers for their work. Our labels for compensation included money or gift cards, class credit, paper authorship, other compensation, explicitly stating no compensation was given (or volunteers), and no information. As Table 11 shows, we observed that most publications did not provide this information, and therefore the label of "no information" was given to the majority of papers for this question.

4.9. Training for Human Labelers and Formal Instructions

The next two questions (see Tables 12 and 13) focused on how labelers were prepared for their work. We defined training as practicing the labeling task with interactive feedback (e.g., being told what they got right or wrong, or being able to ask questions) prior to starting the main labeling work for the study. Formal instructions are documents or videos containing guidelines, definitions, and examples that the labelers could reference as an aid. In two cases, the paper gave enough detail for us to know that no definitions or instructions were given to labelers beyond the text of the question, but about half of papers did not give enough information to make a determination.

Table 11. How were labelers compensated, if at all?

	Count	Proportion
Paper authorship	10	22.22%
Volunteer/explicit no compensation	4	8.89%
Other compensation specified	0	0.00%
No information	31	68.89%
Total of applicable papers (involving original human labeling)	45	100.00%
Nonapplicable papers	155	–

Table 12. Were any details specified about how labelers were trained?

	Count	Proportion
Some training details	7	15.56%
No information	38	84.44%
Total of applicable papers (involving original human labeling)	45	100.00%
Nonapplicable papers	155	–

4.10. Multiple Labeler Overlap

Our next three questions were all about using multiple labelers to review the same items. Having multiple independent labelers is typically a foundational best practice in structured content analysis, so that the integrity of the labels and the schema can be evaluated (although see McDonald et al., 2019). For multiple labeler overlap, our definitions required that papers state whether all or some of the items were labeled by multiple labelers, otherwise "no information" was recorded. We can reasonably assume that papers that did not mention whether multiple labelers were used for each item did not engage in this more intensive process, although we cannot be certain. As Table 14 shows, very few papers mentioned using multiple labelers per item, with the overwhelming majority not giving any indication.

4.11. Synthesis of Labeler Overlap and Reported IRR

The next two questions (see Tables 15 and 16) built on the previous question, which were only answered if the paper had been given the label of "yes for all items" or "yes for some items." For these papers that had multiple labeler overlap, we examined the method by which labeler disagreement was reconciled and whether any IRR or IAA metric was reported. We did not record what kind of IRR/IAA metric was used, such as Cohen's kappa or Krippendorff's alpha, but many different metrics were used. We also did not record what the exact statistic was, although we did notice a wide variation in what was considered an acceptable score.

4.12. Total and Median Number of Human Labelers

We then asked two final questions regarding how many individuals completed a paper's labeling task. Because this information can be presented differently based on the labeling process, we divided this into two. The total number of human labelers referred to all human

Table 13. What kind of formal instructions and/or examples were given to labelers?

	Count	Proportion
Instructions with formal definitions or examples	21	46.67%
No instructions beyond question text	2	4.44%
No information	22	48.89%
Total of applicable papers (involving original human labeling)	45	100.00%
Nonapplicable papers	155	–

Table 14. Multiple labeler overlap

	Count	Proportion
No	8	17.78%
Yes for all items	6	13.33%
Yes for some items	0	0.00%
No information	31	68.89%
Total of applicable papers (involving original human labeling)	45	100.00%
Nonapplicable papers	155	–

labelers involved in the project at any time (see Table 17). The median number of human labelers per item referred to how many labelers evaluated each item in a publication’s data set, which were greater than one in the case of papers that had multiple labelers per item (see Table 18). Eight papers specifed that there was only one labeler per item, which matches with the data in the first question about multiple labeler overlap. The majority of the papers did not provide enough information to answer the question.

4.13. Link to Data Set Available

Our final question was about whether the paper contained a link to the data set containing the original human-labeled training data set. Note that this question was only answered for papers involving some kind of original or novel human labeling, and papers that were exclusively re-using an existing open or public data set were left blank to avoid double-counting. We did not follow such links or verify that such data was actually available. As Table 19 shows, the overwhelming majority of papers did not include such a link, with five papers (11.11%) using original human-labeled training data sets linking to such data. Given the time, labor, expertise, and funding in creating original human labeled data sets, authors may be hesitant to release such data until they feel they have published as many papers as they can, especially junior scholars. Data sharing also requires specific expertise in data formats, documentation, and platforms, which may not be equally distributed across academic disciplines.

5. PAPER INFORMATION SCORES

After finalizing the labels, we quantifed the information that each paper provided about training data, based on how many questions we could answer for each paper. We developed a

Table 15. How were disagreements between labelers reconciled?

	Count	Proportion
Qualitative/discussion	3	50.00%
Quantitative/no discussion	2	33.33%
No information	1	16.67%
Total of applicable papers (involving multiple overlap)	6	100.00%
Nonapplicable papers	194	–

Table 16. Did the paper report an interrater reliability metric?

	Count	Proportion
Yes	3	50.00%
No	3	50.00%
Total of applicable papers (involving multiple overlap)	6	100.00%
Nonapplicable papers	194	–

Table 17. Total number of labelers in the project

	Count	Proportion
1	2	4.44%
2	6	13.33%
3	2	4.44%
5	1	2.22%
7	1	2.22%
10	1	2.22%
30	2	4.44%
659	1	2.22%
???	29	64.44%
Total of applicable papers (involving original human labeling)	45	100.00%
Nonapplicable papers	155	–

Table 18. Median number of labelers per item

	Count	Proportion
1	8	17.78%
2	5	11.11%
3	1	2.22%
??	31	68.89%
Total of applicable papers (involving original human labeling)	45	100.00%
Nonapplicable papers	155	–

Table 19. Link to data set available

	Count	Proportion
No	40	88.89%
Yes	5	11.11%
Total of applicable papers (involving original human labeling)	45	100.00%
Nonapplicable papers	155	–

total and normalized information score, as different studies demanded different levels of information. For example, our questions about whether IRR metrics and reconciliation methods were reported are only applicable for papers involving multiple labelers per item. However, all other questions are relevant for any project involving original human labeling. As such, papers involving original human labeling without multiple labelers per item had a maximum of 11 points, while those with multiple labelers per item had a maximum of 13 points. The normalized score is the total score divided by the maximum score possible.

5.1. Overall Distributions of Information Scores

Figure 3 shows histograms for total and normalized information scores, which show that scores varied substantially. As Geiger et al. (2020) also found, this roughly suggests two overlapping distributions and thus populations of publications: one centered around total scores of 3–5 and normalized scores of 0.3 and another centered around total scores of 9 and normalized scores of 0.7. The normalized information score ranged from 0 to 1, with one paper having a normalized score of 0 and three papers with a full score of 1. The total information score ranged from 0 to 11, with no paper receiving a full score of 13, which would have required a study involving multiple labeler overlap that gave answers to all questions, including IRR metrics and reconciliation method. Overall, the mean total score was 5.4, with a median of 5 and a standard deviation of 3.2. The mean normalized information score was 0.472, with a median of 0.455 and a standard deviation of 0.268. This is quite similar to the findings by Geiger et al. (2020) for their normalized scores, which had a mean of 0.441, a median of 0.429, and a standard deviation of 0.261.

5.2. Information Scores by Corpus and Application Areas

We analyzed information scores by corpus for all papers using original human labeling. Figure 4 is a boxplot illustrating the distribution of normalized information scores by corpus[5]. There was a lower median score (red lines in boxplots) for Social Science & Humanities papers (0.364) than Life Science & Biomedical papers (0.455) and Physical & Environmental Science papers (0.455). However, when examining means between groups ($\overline{\boldsymbol{X}}$ in boxplots), the Physical & Environmental Science papers had a lower mean (0.428) than Social Science & Humanities papers (0.482) and Life Science & Biomedical papers (0.519). We ran a one-way analysis of variance (ANOVA) of normalized information scores by corpus. No statistically significant difference was found ($p = 0.65$, $F = 0.43$). Because we run three statistical tests in

[5] For this and all other boxplots in this paper: The main box is the interquartile range (IQR), or the 25th and 75th percentiles. The middle red line is the median, the black $\overline{\boldsymbol{X}}$ is the mean. The outer whiskers are the highest and lowest data points in a range of 1.5 times the IQR from the median. Grey diamonds are outliers beyond 1.5 times the IQR from the median.

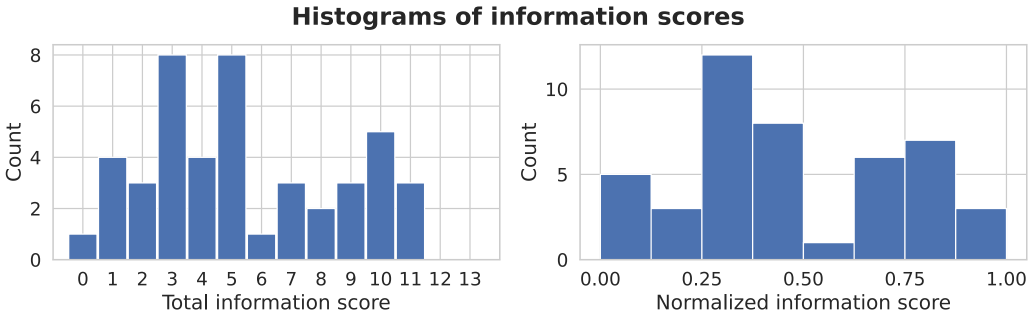

Figure 3. Histograms of total and normalized information scores for all papers involving original human labeling.

this paper, we apply a Bonferroni correction to address the multiple comparisons problem (Dunn, 1961), moving our p-value target from 0.05 to 0.0166.

Next, we conducted a similar analysis by the classification area/domain. A boxplot showing the distribution of normalized information scores is shown in Figure 5. These were not stratified random samples, and we ended up with far more papers in some categories than others, with only 1 item for Physical and Other. The small sample size makes formal statistical tests difficult to interpret, and the assumption of homoscedasticity is not fulfilled due to the wide range in standard deviations between these groups (e.g., 0.13 for Geo/ecological to 0.39 for Demographic). We recommend against making generalizable statistical tests or generalizations based on this analysis, but we report these scores to inform future work. Most groups' mean and median scores were between 0.4 and 0.6, with papers in the Linguistic category having lower medians (0.318). The most common categories—Linguistic, Medical, and Biological—also had much wider distributions and IQRs, but similar means. Activities & actions was the highest scoring category in terms of the mean, median, and upper and lower IQR. In these studies, it is generally the case that the data are recordings of a person performing an activity, and each label is the activity they are asked to perform. This research design may lead authors to more concretely detail such methods.

5.3. Normalized Information Scores by Document Type

For the 45 papers using original human labeling, 33 were journal articles and 12 were conference papers. We conducted an analysis of normalized information scores by document type, which showed larger differences. As Figure 6 shows, articles have a higher mean (0.53 vs. 0.31) and median (0.45 vs. 0.27). We ran a two-tailed Welch's unequal variances t test (Welch, 1947) (variances differed by 0.024) and found a statistically significant difference ($p = 0.0086$, $t = 2.86$). We applied a Bonferroni correction to the p-value threshold to address

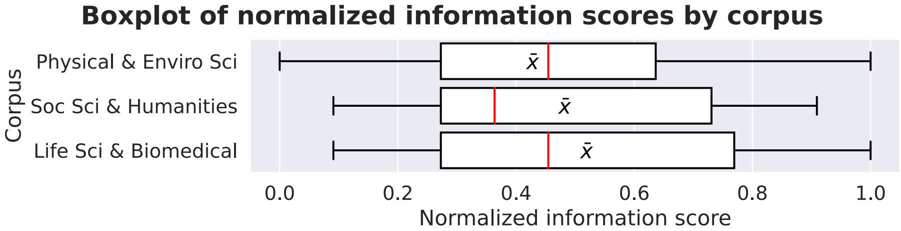

Figure 4. Boxplots of normalized information scores for papers using original human labeling, by corpus.

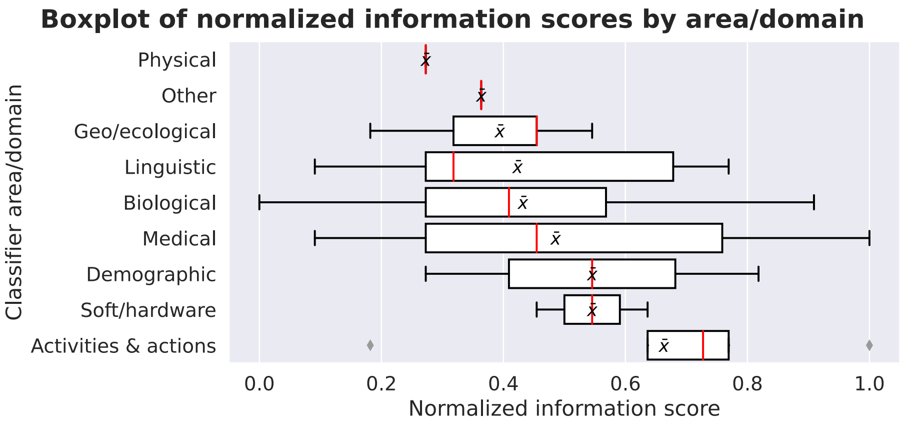

Figure 5. Boxplots of normalized information scores for all papers involving original human labeling by application area/domain. Physical and Other only had one paper. Activities & actions do not have whiskers because no items had scores from 1 to 1.5 * IQR, but did have two outliers outside the 1.5 * IQR range.

the problem of multiple comparisons, but the p-value is well below our adjusted target of 0.0166. This means that in our sample, we can assume that articles generally provide more information about training data than conference papers.

5.4. Label Source Information Scores

Finally, because of the relatively small number of papers involving original human labeling (n = 45) that lead to low statistical power for paper information scores, we examined all papers that presented an original ML classifier (n = 141) based on whether they gave information sufficient to determine if their data set's labels were derived from original human labeling. As discussed in Section 4.3, we gave many papers the answer "no information (implicit yes)," which means we could reasonably assume that labels were made by humans, but the paper never explicitly said humans were involved. Papers with answers "Yes for all items," "Yes for some items," and "No/machine-labeled" were scored 1. Papers with answers "No information (implicit yes)" and "Unsure (but assumed yes)" were scored 0. N/A papers that did not present an original classifier were excluded.

Figure 7 shows the label source reporting rates by corpus, which shows strikingly similar rates. Social Science & Humanities papers had a rate of 72.7%, compared to rates of 71.1% for the other two corpora. Figure 8 shows the label source reporting rates by application area, which shows a much wider range. Activities &actions also has the highest rate at 100% (likely for the same reasons hypothesized earlier), with the lowest rate being Gee/ecological at 46.1%. We also note the differences in these results and the overall paper information scores,

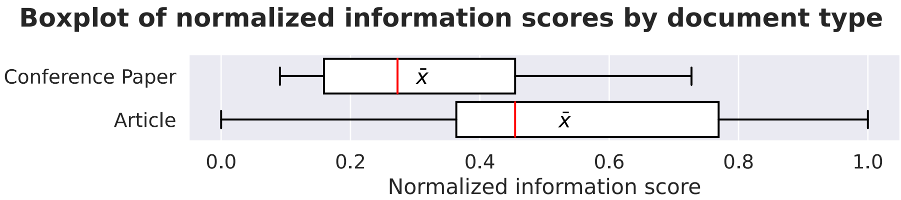

Figure 6. Boxplot of normalized information scores for papers using original labeling, by document type.

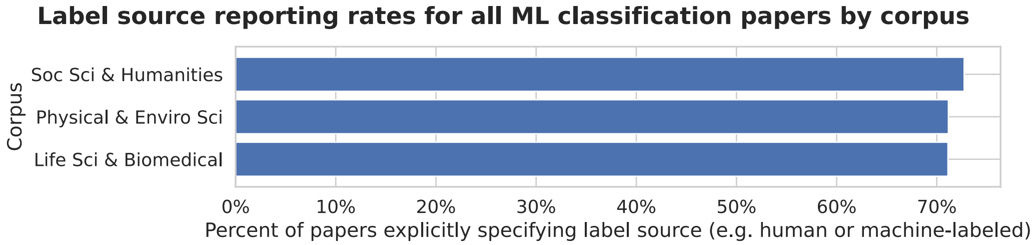

Figure 7. Label source reporting rates for papers presenting an original classifier, by corpus.

which were inversely ranked for the larger categories of Linguistic, Medical, and Biological. While Linguistic papers had lower median information scores, they had far higher rates of label source reporting (79.2%), compared to Medical (69.8%) and especially Biological (58.8%) papers.

Figure 9 shows label source reporting rates by journal article versus conference paper, which shows a much higher rate for articles. We ran a two-tailed Welch's unequal variances *t* test (variances differed by 0.082) and a statistically significant difference was not found (p = 0.038, t = 2.35). We must apply a Bonferroni correction to the *p*-value threshold to address the problem of multiple comparisons, and the *p*-value is above our adjusted target of 0.0166.

5.5. Conclusion to Information Score Results

In conclusion, our quantitative metrics show quite varying ranges and distributions of information scores, which does give evidence for the claim that there is substantial and wide variation in the practices around human labeling, training data curation, and research documentation. The ranges of the boxplots of normalized information scores are substantial, for both IQRs (25th and 75th percentile) and the whiskers at 1.5 * IQRs. Ranges are larger when sampling by corpus, but still substantial for the application areas with more papers (e.g., Medical, Biological, Linguistic).

We specifically call for more investigation into applied ML geo/ecological research, which often classifies land use from aerial photos or photos of geological samples. These had the lowest rates of label source specification and the lowest mean normalized information scores (excluding the categories that only had one paper). However, from our experience, some

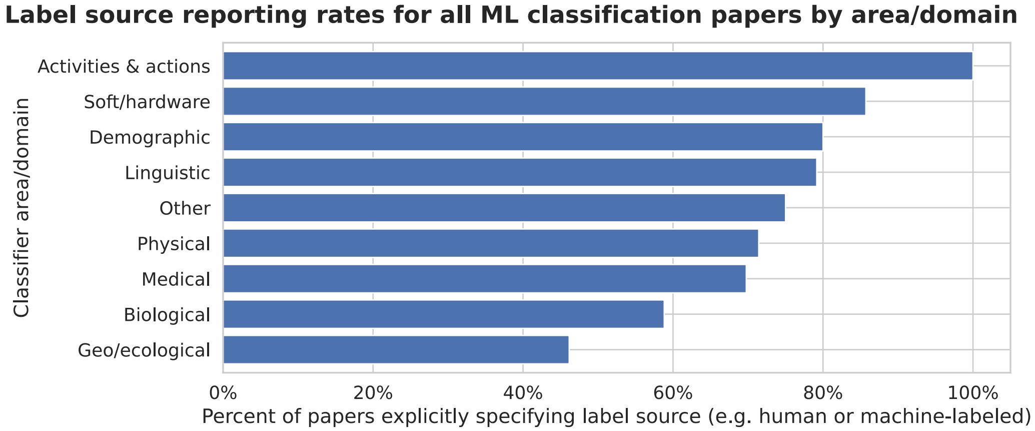

Figure 8. Label source reporting rates for papers presenting an original classifier, by area.

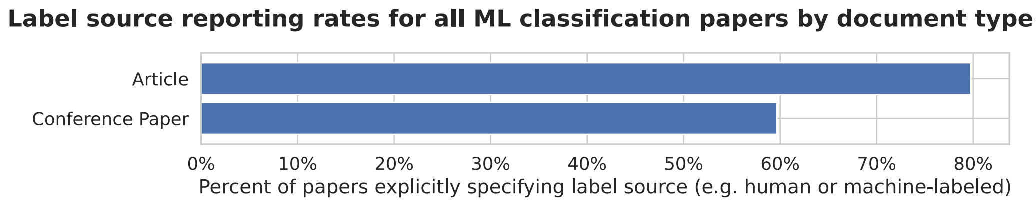

Figure 9. Label source reporting rates for papers presenting an original classifier, by document type.

papers with lower scores did give excellent levels of detail about how they were using an existing land use classification schema they cited (e.g., the widely used USGS guide by Anderson, Hardy et al. [1976]), but did not give any details about who applied that schema to the aerial photos. We can also hypothesize that in fields with widely established and shared methodological standards, researchers could have far higher rates of adherence to methodological best practices around data labeling, but have lower rates of reporting that they actually followed those practices in papers.

Finally, we draw attention to the different rates when we grouped by corpus versus application area. In our sampling, the corpus was the Scopus-provided metadata field, which is determined at the publication level when a journal is added to Scopus[6]. At this level, we saw fewer differences in quantitative scores. In contrast, our application area field is determined for each paper based on the content, independent of the journal or venue in which it was published. Scores varied far more when broken out by application area, which is likely due in part to noise in the smaller samples. However, this may also indicate that methodological reporting rates vary even more within subfields or types of research objects. For future work, we recommend that researchers pay specific attention to differences between fields or objects of study, rather than solely group papers in the high-level way we did with our three corpora.

6. CONCLUDING DISCUSSION

6.1. Findings

First, our study shows that contemporary applications of supervised ML across disciplines often rely on training data sets in ways that either reuse existing human-labeled data sets or label items with some kind of automated process. Of the papers in our data set that presented an original ML classifier, only 26.7% produced a new human-labeled data set as part of their study—a rate that did not substantially vary among our three corpora from the biomedical & life sciences, the physical & environmental sciences, and the humanities & social sciences. Second, of the applied ML publications that did produce a new human-labeled training data set, there was significant divergence in reporting methodological details and following best practices in human labeling. A small number of publications received top information scores, but approximately two-thirds of publications involving original human labeling did not provide enough information for us to answer more than half of the subsequent questions we asked about the labeling process.

This cross-disciplinary trend is cause for concern, given that high-quality training data is essential to the validity of ML classifiers and human judgment is notoriously difficult to standardize. When comparing across our three broad corpora of social science & humanities,

[6] https://web.archive.org/web/20210531200329/https://service.elsevier.com/app/answers/detail/a_id/14882/supporthub/

biomedical & life sciences, and physical & environmental sciences papers, we only see marginal differences in the level of information papers provide. We do see more robust evidence that journal articles have higher rates of reporting information about training data than conference papers, which may relate to conference papers being shorter and only involving a single cycle of peer-review[7].

6.2. Implications

6.2.1. *The black-boxing of training data*

ML is increasingly used across disciplines and application domains, but the quality of supervised ML classifiers is only as good as the data that is used to train it. Based on our findings, we argue for more attention to be placed on the specific details of how that training data is labeled. There is a recent wave of work that interrogates ML models once they are trained, as well as considerations about "automation bias" (Skitka, Mosier, & Burdick, 1999)—that people often treat trained models as a "black box," with their outputs unquestioned and taken as given. These concerns must also extend to the labeling and curation of training data sets, some of which become widely reused without being examined.

For example, Crawford and Paglen (2019) have called attention to problematic racial labels of images in the popular ImageNet training data set, which has been a standard benchmark data set in image recognition for over a decade. Birhane and Prabhu (2021) found thousands of images in the 80 Million Tiny Images data set that were labeled with offensive racial and gender-based slurs. The careful curation of data sets has long been a central tenet in the institutions of science, although standards and practices can change dramatically over time and across contexts. Historians of science such as Bowker (2005) and Gitelman (2013) remind us that data is never "raw," as data always is produced and used within a messy assemblage of partially overlapping human institutions, each of which have their own practices, values, and assumptions. To this end, we call for applied ML researchers and practitioners who are reusing human-labeled data sets to exercise as much caution and care around the decisions to reuse a labeled data set as they would if they were labeling the data themselves.

Finally, we have not asked any questions about how papers discuss data cleaning, but we encourage more investigation and consideration of how the often-backgrounded work of data cleaning is performed, managed, and documented. We could have asked another dozen questions about how papers did or did not discuss how they cleaned their data. For future work, we would encourage researchers to study what applied ML papers report about how they cleaned and preprocessed their data. We also see much future work in studying to what extent applied ML papers report efforts at de-biasing data sets and models.

6.2.2. *Institutional change around data documentation*

We call on the institutions of science—publications, funders, disciplinary societies, and educators—to play a major role in working out solutions to these issues of data quality and research documentation. We see this work as part of the open science and reproducibility movement, specifically the movement for open access to research data sets, materials, protocols, and analysis code. However, even advocates of this movement have long discussed how individual researchers do not have incentives to be first-movers in being more open than usual about the messiness in all research, because it leaves their work more open to rebuttal (Ali-Khan, Harris, & Gold, 2017; Smaldino, 2016; Zimring, 2019). In our own experience,

[7] Not all conference papers are peer reviewed, but all conference proceedings indexed by Scopus are peer reviewed.

we have certainly felt the temptation to not report certain details that would lead others to have less confidence in our study, such as our IRR metrics.

In looking towards solutions, we see a parallel with issues in open access to publications, which often requires individual researchers to choose if they want to pay for open access out of their own funding. While some first-movers paid for this out of their own budgets, open access is currently being far more effectively tackled at the institutional level in ways that will not require individuals (and especially first-movers) to pay the costs. So too do we see institutional solutions to the issue of methodological detail, where a common floor could be established that is equally applicable to all researchers. We also see resonance with the various proposed efforts at standardizing documentation about ML models and data sets (Barclay et al., 2019; Bender & Friedman, 2018; Beretta et al., 2018; Gebru et al., 2018; Hind et al., 2018; Holland et al., 2018; Mitchell et al., 2019; Raji & Yang, 2019) and urge that human labeling details be included in such efforts.

On the publication process, we note that research publications are limited by length restrictions, which can leave little space for details. We can hypothesize that having a dedicated and visible space for methodology and data set documentation would make these concerns more central for authors, reviewers, editors, and readers, although we can only speculate as to the best way for this to be implemented. For example, *Nature* has far shorter word limits for a main research article (2,000 to 2,500 words), which means methodological and data set documentation is often fully detailed in appendices, which can be of any length. Does this approach more easily lead to readers and/or reviewers ignoring such details and focusing more on results? However, *Nature* also requires that authors fill out a peer-reviewed checklist form that asks general and domain-specific questions about statistical details (e.g., "a description of all covariates tested") and about the data set (e.g., for behavioral science, "State the research sample … provide relevant demographic information … and indicate whether the sample is representative")[8]. Do these kinds of mandatory structured disclosure forms make these concerns more central to authors and reviewers, even if they are not as accessible to readers?

We also note that peer reviewers and editors play a major role in deciding what details are considered extraneous. First, we urge reviewers to make space for what some may see as "boring" methodological details. More importantly, we call on editorial boards to openly signal in author and reviewer guidelines that they invite or even require extended discussion of methodological details. To this end, one recent trend is the growth of multistakeholder groups that have collectively released formal guidelines or best practices statements on research reporting, such as the CONSORT guidelines for reporting randomized clinical trials (Schulz, Altman et al., 2010), the COREQ guidelines for reporting qualitative research (Tong, Sainsbury, & Craig, 2007), or the PRISMA guidelines on reporting meta-analyses and systematic reviews (Moher, Liberati et al., 2009).

For example, PRISMA guidelines on reporting meta-analyses and systematic reviews have been mandated in the author guidelines of many journals (including *The Lancet*[9], *PLOS ONE*[10], and *Systematic Reviews*[11]), which require authors to fill out the 27-item PRISMA checklist[12]. One interesting trend with such multistakeholder best practices statements in

[8] https://www.nature.com/documents/nrreporting-summary-flat.pdf

[9] https://els-jbs-prod-cdn.jbs.elsevierhealth.com/pb/assets/raw/Lancet/authors/tl-info-for-authors.pdf

[10] https://journals.plos.org/plosone/s/submission-guidelines#loc-systematic-reviews-and-meta-analyses

[11] https://systematicreviewsjournal.biomedcentral.com/submission-guidelines/preparing-your-manuscript/research

[12] https://prisma-statement.org/documents/PRISMA%202009%20checklist.pdf

medicine is the proliferation of subdomain-specific "extensions" that further specify methodological reporting standards. For example, the EQUATOR network tracks 32 extensions to the CONSORT guidelines[13], including guidelines for reporting randomized clinical trials in pain management (Gewandter, Eisenach et al., 2019), orthodontics (Pandis, Fleming et al., 2015), and psycho-social interventions (Montgomery, Grant et al., 2018).

However, there have been disagreements over the impact and efficacy of these more structured approaches. Page and Moher's[14] meta-analysis of 57 papers studying uptake of the PRISMA guidelines (Page & Moher, 2017) found that while more papers are reporting details in the PRISMA guidelines after it was released in 2009, some details remain low even for papers claiming to adhere to the guidelines. For example, for nine of the PRISMA items, fewer than 67% of papers actually reported the information in question. Fleming, Koletsi, and Pandis (2014) found that following the widespread uptake of the PRISMA guidelines by certain publications, more meta-analysis articles reported methodological details, but disproportionately those in the PRISMA guidelines. The authors of that study raise concerns that PRISMA has overdetermined the peer review process: Authors who are fully compliant with PRISMA are no longer reporting other methodological details that Fleming et al. claim are also relevant in such work and were in other competing meta-analysis guidelines that ultimately lost to PRISMA.

6.2.3. *Are there universal best practices for the labeling of training data?*

The efforts around methodological standards in medicine raise an important question about the wisdom of seeking a single one-size-fits-all set of best practices for any application of supervised ML. However, contemporary efforts around "fairness" or "transparency" in ML often work towards more universal or domain-independent approaches, which are applied to a wide range of application areas (e.g., finance, social services, policing, hiring, medicine). Yet in our work examining publications from quite different academic fields, we found ourselves needing to pay close attention to the various kinds of specialized expertise that are required to label a training data set for a particular purpose. As Bowker and Star (1999) and Goodwin (1994) discuss, all classification systems rely on a shared cultural context, which can be exceedingly difficult to formally specify and often falls apart at the edges. It can be difficult to know beforehand what level of shared cultural context and expertise will be involved.

Some of the papers we analyzed described in great detail how the people who labeled their data set were chosen for their expertise, from seasoned medical practitioners diagnosing diseases to youth familiar with social media slang in multiple languages. That said, not all labeling tasks require years of specialized expertise, such as more straightforward tasks we saw, including distinguishing positive versus negative business reviews or identifying different hand gestures. Even projects in the same domain can require different levels of expertise, such as a data set of animal photos labeled just for the presence of cats and dogs, versus labeling the same photos for the specific breed of cats and dogs. Furthermore, we found that some labeling tasks are well suited to semiautomated labeling where labelers are assisted with rule-based approaches, while others are not. Finally, even the more seemingly straightforward classification tasks can still have substantial room for ambiguity and error for the inevitable edge cases, which require training and verification processes to ensure a standardized data set.

[13] https://www.equator-network.org/?post_type=eq_guidelines&eq_guidelines_clinical_specialty=0&s=+CONSORT+extension

[14] Moher is the lead author of the PRISMA statement.

The labeling protocol and schema we developed and used in this paper—which is based on and extends prior work (Geiger et al., 2020)—is an effort at creating a cross-disciplinary standard for any given research project that uses human-labeled training data. While we believe that any peer reviewer or reader can ask these same questions of any ML application paper, they are only a starting point. Issues of validity, consistency, reliability, reproducibility, and accountability require further investigation. The kind of domain-independent criteria we used should be seen as necessary but not suficient criteria for having confidence in a labeled data set. We do not advocate for a single, universal, one-size-fits-all solution, but instead seek to spur conversations within and across disciplines about better approaches to bring the work of data labeling into the foreground. We see a role for the classic principle of reproducibility, but for data labeling: Does the paper provide enough detail so that another researcher could hypothetically recruit a similar team of labelers, give them the same instructions and training, reconcile disagreements similarly, and have them produce a similarly labeled data set?

Data publications could also play a major role in this issue, which are standalone peer-reviewed publications that do not answer a research question, but instead spend the entire paper describing the creation of a data set in rich detail (Candela, Castelli et al., 2015; Chavan & Penev, 2011; Costello, 2009; Smith, 2009). In seeking to bring the work of data labeling from the background to the foreground, our work is also aligned with scholars who have focused on the often undercompensated labor of crowdworkers and have called for researchers to detail how much they pay for data labeling (Silberman et al., 2018).

6.3. Limitations

To conclude, we reflect that our study also has the same kinds of limitations that many human labeling projects have. For example, given the concerns we raise about domain-specific expertise, our team may have missed or misinterpreted crucial details when examining papers. The second issue is around the reliability and reproducibility of our team's labeling process. In conducting this study, we have become quite familiar with the difficulties of getting a medium-sized team to build a consensus around reducing complex objects into quantifable data. We specifically chose to have a more detailed and time-intensive process in which disagreements were discussed, which traded off with the total number of items we were able to label. We believe this trade-off was the right decision, given our focus on methodological rigor, but it does mean our samples are smaller than we would like. The lower sample size means that we have less confidence in the statistical generalizability of our sample to the population of all applied ML publications. However, we see a wide range of future work that can be done to extend these efforts, such as with teams of domain-specific experts that examine applied ML fields in their area of expertise.

Finally, we only have access to what each publication reported about the work they did, and not the research project itself, which means our unit of analysis is methodological reporting. For example, researchers could have far higher rates of following methodological best practices around data labeling, but have lower rates of reporting that they actually followed those practices in papers. We could even hypothesize an inverse relationship between a field's overall adherence to methodological best practices and researchers' rates of reporting adherence to those practices, if such practices become so routine and mundane that they are left implicit in publications. For these reasons, we strongly advise against interpreting our quantitative scores as an unproblematic proxy for methodological rigor, especially for the scores by discipline and area. However, given our interest in how labeling practices impact the validity of ML models and classifiers, future work could extend this work through other methods, such as surveys and ethnographic studies of ML researchers.

ACKNOWLEDGMENTS

We would like to acknowledge support for this project from the UC-Berkeley Undergraduate Research Apprenticeship Program (URAP) and to Stacey Dorton for administrative and logistical support. We thank the peer reviewers for their feedback and suggestions.

AUTHOR CONTRIBUTIONS

R. Stuart Geiger: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing—original draft, Writing—review & editing. Dominique Cope: Investigation, Validation, Writing—original draft, Writing—review & editing. Jamie Ip: Data curation, Investigation, Software, Validation. Marsha Lotosh: Investigation, Validation, Visualization. Aayush Shah: Investigation, Validation. Jenny Weng: Investigation, Validation, Writing—review & editing. Rebekah Tang: Investigation.

COMPETING INTERESTS

The authors have no competing interests.

FUNDING INFORMATION

This work was funded in part by the Gordon & Betty Moore Foundation (Grant GBMF3834) and the Alfred P. Sloan Foundation (Grant 2013-10-27), as part of the Moore-Sloan Data Science Environments grant to UC-Berkeley. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

DATA AVAILABILITY

All data sets, analysis scripts, protocols, labeling instructions, and other supplementary information required to replicate and reproduce these findings can be downloaded on GitHub[15] and Zenodo[16].

REFERENCES

Ali-Khan, S. E., Harris, L. W., & Gold, E. R. (2017). Motivating participation in open science by examining researcher incentives. *eLife*, *6*, e29319. https://doi.org/10.7554/eLife.29319, PubMed: 29082866

Amini, A., Soleimany, A. P., Schwarting, W., Bhatia, S. N., & Rus, D. (2019). Uncovering and mitigating algorithmic bias through learned latent structure. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES '19* (pp. 289–295). Association for Computing Machinery. https://doi.org/10.1145/3306618.3314243

Anderson, J. R., Hardy, E. E., Roach, J. T., & Witmer, R. E. (1976). *A land use and land cover classification system for use with remote sensor data*, volume 964. US Government Printing Office. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.189.3029&rep=rep1&type=pdf

Baas, J., Schotten, M., Plume, A., Côté, G., & Karimi, R. (2020). Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quantitative Science Studies*, *1*(1), 377–386. https://doi.org/10.1162/qss_a_00019

[15] https://github.com/staeiou/gigo_qss_2021

[16] https://doi.org/10.5281/zenodo.4906636

Babbage, C. (1864). *Passages from the life of a philosopher*. London: Longman, Green, Longman, Roberts, and Green.

Ball, N. M., & Brunner, R. J. (2010). Data mining and machine learning in astronomy. *International Journal of Modern Physics D*, *19*(7), 1049–1106. https://doi.org/10.1142/S0218271810017160

Barclay, I., Preece, A., Taylor, I., & Verma, D. (2019). Towards traceability in data ecosystems using a bill of materials model. *arXiv preprint arXiv:1904.04253*. https://arxiv.org/abs/1904.04253

Bender, E. M., & Friedman, B. (2018). Data statements for NLP: Toward mitigating system bias and enabling better science. *Transactions of the ACL*, *6*, 587–604. https://doi.org/10.1162/tacl_a_00041

Beretta, E., Vetrò, A., Lepri, B., & De Martin, J. C. (2018). Ethical and socially-aware data labels. In *Annual International Symposium on Information Management and Big Data* (pp. 320–327). Springer. https://doi.org/10.1007/978-3-030-11680-4_30

Bergstrom, C. T., & West, J. D. (2020). *Calling bullshit: The art of skepticism in a data-driven world*. London: Random House Publishing Group.

Birhane, A., & Prabhu, V. U. (2021). Large image datasets: A pyrrhic win for computer vision? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (pp. 1537–1547). https://arxiv.org/abs/2006.16923. https://doi.org/10.1109/WACV48630.2021.00158

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of “bias” in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5454–5476). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.485

Bontcheva, K., Cunningham, H., Roberts, I., Roberts, A., Tablan, V., … Gorrell, G. (2013). GATE Teamware: A web-based, collaborative text annotation framework. *Language Resources and Evaluation*, *47*(4), 1007–1029. https://doi.org/10.1007/s10579-013-9215-6

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, *63*(6), 1059–1078. https://doi.org/10.1002/asi.22634

Bowker, G. (2005). *Memory practices in the sciences*. Cambridge, MA: MIT Press.

Bowker, G. C. (2020). Numbers or no numbers in science studies. *Quantitative Science Studies*, *1*(3), 927–929. https://doi.org/10.1162/qss_a_00054

Bowker, G. C., & Star, S. L. (1999). *Sorting things out: Classification and its consequences*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/6352.001.0001

Brady, A. P. (2016). Error and discrepancy in radiology: Inevitable or avoidable? *Insights into Imaging*, *8*(1), 171–182. https://doi.org/10.1007/s13244-016-0534-1, PubMed: 27928712

Brand, A., Allen, L., Altman, M., Hlava, M., & Scott, J. (2015). Beyond authorship: Attribution, contribution, collaboration, and credit. *Learned Publishing*, *28*(2), 151–155. https://doi.org/10.1087/20150211

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *ACM Conference on Fairness, Accountability and Transparency* (pp. 77–91). https://proceedings.mlr.press/v81/buolamwini18a.html

Calmon, F. P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17* (pp. 3995–4004). Curran Associates Inc. https://proceedings.neurips.cc/paper/2017/file/9a49a25d845a483fae4be7e341368e36-Paper.pdf

Cambrosio, A., Cointet, J.-P., & Abdo, A. H. (2020). Beyond networks: Aligning qualitative and computational science studies. *Quantitative Science Studies*, *1*(3), 1017–1024. https://doi.org/10.1162/qss_a_00055

Candela, L., Castelli, D., Manghi, P., & Tani, A. (2015). Data journals: A survey. *Journal of the Association for Information Science and Technology*, *66*(9), 1747–1762. https://doi.org/10.1002/asi.23358

Chang, J. C., Amershi, S., & Kamar, E. (2017). Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, CHI ’17* (pp. 2334–2346). New York, NY, USA. https://doi.org/10.1145/3025453.3026044

Chavan, V., & Penev, L. (2011). The data paper: A mechanism to incentivize data publishing in biodiversity science. *BMC Bioinformatics*, *12*, S2. https://doi.org/10.1186/1471-2105-12-S15-S2, PubMed: 22373175

Costello, M. J. (2009). Motivating online publication of data. *BioScience*, *59*(5), 418–427. https://doi.org/10.1525/bio.2009.59.5.9

Crawford, K., & Paglen, T. (2019). Excavating AI: The politics of training sets for machine learning. https://excavating.ai

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

DeCamp, M., & Lindvall, C. (2020). Latent bias and the implementation of artificial intelligence in medicine. *Journal of the American Medical Informatics Association*, *27*(12), 2020–2023. https://doi.org/10.1093/jamia/ocaa094, PubMed: 32574353

Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S. M., & Weischedel, R. M. (2004). The Automatic Content Extraction (ACE) Program: Tasks, data, and evaluation. In *Proceedings of the 2004 4th International Conference on Language Resources and Evaluation*, Vol. 2 (pp. 837–840). Paris: European Language Resources Association. https://www.lrec-conf.org/proceedings/lrec2004/pdf/5.pdf

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, *56*(293), 52–64. https://doi.org/10.1080/01621459.1961.10482090

Fecher, B., & Friesike, S. (2014). Open science: One term, five schools of thought. In S. Bartling & S. Friesike (Eds.) *Opening science: The evolving guide on how the internet is changing research, collaboration and scholarly publishing* (pp. 17–47). Springer International. https://doi.org/10.1007/978-3-319-00026-8_2

Fleming, P. S., Koletsi, D., & Pandis, N. (2014). Blinded by PRISMA: Are systematic reviewers focusing on PRISMA and ignoring other guidelines? *PLOS ONE*, *9*(5), e96407. https://doi.org/10.1371/journal.pone.0096407, PubMed: 24788774

Fluke, C. J., & Jacobs, C. (2020). Surveying the reach and maturity of machine learning and artificial intelligence in astronomy. *WIREs Data Mining and Knowledge Discovery*, *10*(2), e1349. https://doi.org/10.1002/widm.1349

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* ’19* (pp. 329–338). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287589

Friedman, J., Hastie, T., & Tibshirani, R. (2009). *The elements of statistical learning: Data mining, inference, and prediction*, 2nd ed. Springer. https://doi.org/10.1007/978-0-387-84858-7

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. *arXiv preprint arXiv:1803.09010*. https://arxiv.org/abs/1803.09010

Geiger, R. S., Yu, K., Yang, Y., Dai, M., Qiu, J., … Huang, J. (2020). Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 325–336). https://doi.org/10.1145/3351095.3372862

Gewandter, J. S., Eisenach, J. C., Gross, R. A., Jensen, M. P., Keefe, F. J., … Turk, D. C. (2019). Checklist for the preparation and review of pain clinical trial publications: A pain-specific supplement to CONSORT. *Pain Reports*, *4*(3), e621. https://doi.org/10.1097/PR9.0000000000000621, PubMed: 28989992

Gharibi, G., Walunj, V., Alanazi, R., Rella, S., & Lee, Y. (2019). Automated management of deep learning experiments. In *Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning, DEEM’19* (pp. 8:1–8:4). New York, NY, USA. https://doi.org/10.1145/3329486.3329495

Gil, Y., David, C. H., Demir, I., Essawy, B. T., Fulweiler, R. W., … Yu, X. (2016). Toward the geoscience paper of the future: Best practices for documenting and sharing research from data to software to provenance. *Earth and Space Science*, *3*(10), 388–415. https://doi.org/10.1002/2015EA000136

Gitelman, L. (Ed.) (2013). *Raw data is an oxymoron*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/9302.001.0001

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, MA: MIT Press. https://www.deeplearningbook.org

Goodman, A., Pepe, A., Blocker, A. W., Borgman, C. L., Cranmer, K., … Slavkovic, A. (2014). Ten simple rules for the care and feeding of scientific data. *PLOS Computational Biology*, *10*(4), e1003542. https://doi.org/10.1371/journal.pcbi.1003542, PubMed: 24763340

Goodwin, C. (1994). Professional vision. *American Anthropologist*, *96*(3), 606–633. https://doi.org/10.1525/aa.1994.96.3.02a00100

Hanna, A., Denton, E., Smart, A., & Smith-Loud, J. (2020). Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* ’20* (pp. 501–512). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372826

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS’ 16* (pp. 3323–3331). Curran Associates Inc. https://doi.org/10.5555/3157382.3157469

Hind, M., Mehta, S., Mojsilovic, A., Nair, R., Ramamurthy, K. N., … Varshney, K. R. (2018). Increasing trust in AI services through supplier’s declarations of conformity. *arXiv preprint arXiv:1808.07261*. https://arxiv.org/pdf/1808.07261

Holland, S., Hosny, A., Newman, S., Joseph, J., & Chmielinski, K. (2018). The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677*. https://arxiv.org/abs/1805.03677

Hovy, E., & Lavid, J. (2010). Towards a ‘science’ of corpus annotation: A new methodological challenge for corpus linguistics. *International Journal of Translation*, *22*(1), 13–36.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Jacobs, A. Z., & Wallach, H. (2019). Measurement and fairness. *arXiv:1912.05511 [cs]*. https://arxiv.org/abs/1912.05511

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer. https://doi.org/10.1007/978-1-4614-7138-7

Jones, E., Oliphant, T., Peterson, P., et al. (2001). SciPy: Open source scientific tools for Python. https://www.scipy.org/

Kang, D., & Evans, J. (2020). Against method: Exploding the boundary between qualitative and quantitative studies of science. *Quantitative Science Studies*, *1*(3), 930–944. https://doi.org/10.1162/qss_a_00056

Karimi Mahabadi, R., Belinkov, Y., & Henderson, J. (2020). End-to-end bias mitigation by modelling biases in corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8706–8716). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.769

Kitzes, J., Turek, D., & Deniz, F. (2018). *The practice of reproducible research: Case studies and lessons from the data-intensive sciences*. University of California Press, Oakland. https://practicereproducibleresearch.org

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., … Willing, C. (2016). Jupyter Notebooks: A publishing format for reproducible computational workfows. In F. Loizides & B. Schmidt (Eds.), *Positioning and power in academic publishing: Players, agents and agendas* (pp. 87–90). IOS Press. https://doi.org/10.3233/978-1-61499-649-1-87

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, *30*(1), 61–70. https://doi.org/10.1177/001316447003000105

Krishnan, S., Franklin, M. J., Goldberg, K., Wang, J., & Wu, E. (2016). ActiveClean: An interactive data cleaning framework for modern machine learning. In *Proceedings of the 2016 International Conference on Management of Data, SIGMOD ’16* (pp. 2117–2120). New York, NY, USA. https://doi.org/10.1145/2882903.2899409

Larivière, V., Pontille, D., & Sugimoto, C. R. (2020). Investigating the division of scientific labor using the Contributor Roles Taxonomy (CRediT). *Quantitative Science Studies*, *2*(1), 111–128. https://doi.org/10.1162/qss_a_00097

Lary, D. J., Alavi, A. H., Gandomi, A. H., & Walker, A. L. (2016). Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, *7*(1), 3–10. https://doi.org/10.1016/j.gsf.2015.07.003

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. Cambridge, MA: Harvard University Press.

Leydesdorff, L., Ràfols, I., & Milojević, S. (2020). Bridging the divide between qualitative and quantitative science studies. *Quantitative Science Studies*, *1*(3), 918–926. https://doi.org/10.1162/qss_e_00061

Linguistic Data Consortium. (2008). ACE (Automatic Content Extraction) English annotation guidelines for entities version 6.6. https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v6.6.pdf

Ma, L., Li, M., Ma, X., Cheng, L., Du, P., & Liu, Y. (2017). A review of supervised object-based land-cover image classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *130*, 277–293. https://doi.org/10.1016/j.isprsjprs.2017.06.001

McDonald, N., Schoenebeck, S., & Forte, A. (2019). Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 72:1–72:3. https://doi.org/10.1145/3359174

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the*

9th Python in Science Conference (pp. 51–56). https://conference.scipy.org/proceedings/scipy2010/mckinney.html. https://doi.org/10.25080/Majora-92bf1922-00a

Medeiros, N., & Ball, R. (2017). Teaching integrity in empirical economics: The pedagogy of reproducible science in undergraduate education. In M. Hensley & S. Davis-Kahl (Eds.), *Undergraduate research and the academic librarian: Case studies and best practices*. Chicago: Association of College & Research Libraries. https://scholarship.haverford.edu/cgi/viewcontent.cgi?article=1189

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. https://arxiv.org/abs/1908.09635

Mellin, W. (1957). Work with new electronic 'brains' opens field for army math experts. *The Hammond Times*, *10*, 66.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., … Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). https://doi.org/10.1145/3287560.3287596

Mitchell, T. (1997). *Machine learning*. New York: MacGraw-Hill.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLOS Medicine*, *6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097, PubMed: 19621072

Montgomery, P., Grant, S., Mayo-Wilson, E., Macdonald, G., Michie, S., … Yaffe, J., and on behalf of the CONSORT-SPI Group. (2018). Reporting randomised trials of social and psychological interventions: The CONSORT-SPI 2018 Extension. *Trials*, *19*(1), 407. https://doi.org/10.1186/s13063-018-2733-1, PubMed: 30060754

Mozetič, I., Grčar, M., & Smailović, J. (2016). Multilingual Twitter sentiment classification: The role of human annotators. *PLOS ONE*, *11*(5), e0155036. https://doi.org/10.1371/journal.pone.0155036, PubMed: 27149621

Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., & Liang, X. (2018). doccano: Text annotation tool for human. https://github.com/doccano/doccano

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342, PubMed: 31649194

Oleinik, A., Popova, I., Kirdina, S., & Shatalova, T. (2014). On the choice of measures of reliability and validity in the content-analysis of texts. *Quality & Quantity*, *48*(5), 2703–2718. https://doi.org/10.1007/s11135-013-9919-0

Page, M. J., & Moher, D. (2017). Evaluations of the uptake and impact of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement and extensions: A scoping review. *Systematic Reviews*, *6*(1), 263. https://doi.org/10.1186/s13643-017-0663-8, PubMed: 29258593

Pandis, N., Fleming, P. S., Hopewell, S., & Altman, D. G. (2015). The CONSORT Statement: Application within and adaptations for orthodontic trials. *American Journal of Orthodontics and Dentofacial Orthopedics*, *147*(6), 663–679. https://doi.org/10.1016/j.ajodo.2015.03.014, PubMed: 26038070

Pérez, F., & Granger, B. E. (2007). IPython: A system for interactive scientific computing. *Computing in Science and Engineering*, *9*(3), 21–29. https://doi.org/10.1109/MCSE.2007.53

Pérez-Pérez, M., Glez-Peña, D., Fdez-Riverola, F., & Lourenço, A. (2015). Marky: A tool supporting annotation consistency in multi-user and iterative document annotation projects. *Computer Methods and Programs in Biomedicine*, *118*(2), 242–251. https://doi.org/10.1016/j.cmpb.2014.11.005, PubMed: 25480679

Perry, T. (2020). SimpleDorff—Calculate Krippendorff's Alpha on a DataFrame. https://github.com/LightTag/simpledorff

Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined approach. *Journal of Informetrics*, *3*(2), 143–157. https://doi.org/10.1016/j.joi.2009.01.003

Project Jupyter, Bussonnier, M., Forde, J., Freeman, J., Granger, B., Head, T., … Willing, C. (2018). Binder 2.0—Reproducible, interactive, sharable environments for science at scale. In F. Akici, D. Lippa, D. Niederhut, & M. Pacer (Eds.), *Proceedings of the 17th Python in Science Conference* (pp. 113–120). https://doi.org/10.25080/Majora-4af1f417-011

Quarfoot, D., & Levine, R. A. (2016). How robust are multirater interrater reliability indices to changes in frequency distribution? *The American Statistician*, *70*(4), 373–384. https://doi.org/10.1080/00031305.2016.1141708

Raji, I. D., & Yang, J. (2019). ABOUT ML: Annotation and benchmarking on understanding and transparency of machine learning lifecycles. *arXiv:1912.06166 [cs, stat]*. https://arxiv.org/abs/1912.06166

Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, *89*, 14–46. https://doi.org/10.1016/j.knosys.2015.06.015

Raykar, V. C., & Yu, S. (2012). Eliminating spammers and ranking annotators for crowdsourced labeling tasks. *Journal of Machine Learning Research*, *13*(Feb), 491–518. https://doi.org/10.5555/2188385.2188401

Rennie, D., Flanagin, A., & Yank, V. (2000). The contributions of authors. *JAMA*, *284*(1), 89–91. https://doi.org/10.1001/jama.284.1.89, PubMed: 10872020

Riff, D., Lacy, S., & Fico, F. (2013). *Analyzing media messages: Using quantitative content analysis in research*. New York: Routledge. https://doi.org/10.4324/9780203551691

Salimi, B., Howe, B., & Suciu, D. (2020). Database repair meets algorithmic fairness. *ACM SIGMOD Record*, *49*(1), 34–41. https://doi.org/10.1145/3422648.3422657

Sallans, A., & Donnelly, M. (2012). DMP Online and DMPTool: Different strategies towards a shared goal. *International Journal of Digital Curation*, *7*(2), 123–129. https://doi.org/10.2218/ijdc.v7i2.235

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1668–1678). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1163

Schelter, S., Böse, J.-H., Kirschnick, J., Klein, T., & Seufert, S. (2017). Automatically tracking metadata and provenance of machine learning experiments. In *Machine Learning Systems Workshop at NIPS*. https://learningsys.org/nips17/assets/papers/paper_13.pdf

Schelter, S., Lange, D., Schmidt, P., Celikel, M., Biessmann, F., & Grafberger, A. (2018). Automating large-scale data quality verification. *Proceedings of the VLDB Endowment*, *11*(12), 1781–1794. https://doi.org/10.14778/3229863.3229867

Schreier, A. A., Wilson, K., & Resnik, D. (2006). Academic research record-keeping: Best practices for individuals, group leaders, and institutions. *Academic Medicine: Journal of the Association of American Medical Colleges*, *81*(1), 42. https://doi.org/10.1097/00001888-200601000-00010, PubMed: 16377817

Schulz, K. F., Altman, D. G., Moher, D., & for the CONSORT Group. (2010). CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. *PLOS Medicine*, *7*(3), e1000251. https://doi.org/10.1371/journal.pmed.1000251, PubMed: 20352064

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19* (pp. 59–68). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287598

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., … Golub, T. R. (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine*, *8*(1), 68–74. https://doi.org/10.1038/nm0102-68, PubMed: 11786909

Silberman, M. S., Tomlinson, B., LaPlante, R., Ross, J., Irani, L., & Zaldivar, A. (2018). Responsible research with crowds: Pay crowdworkers at least minimum wage. *Communications of the ACM*, *61*(3), 39–41. https://doi.org/10.1145/3180492

Simpson, R., Page, K. R., & De Roure, D. (2014). Zooniverse: Observing the world's largest citizen science platform. In *Proceedings of the 23rd International Conference on World Wide Web, WWW '14 Companion* (pp. 1049–1054). New York, NY, USA. https://doi.org/10.1145/2567948.2579215

Singh, J., Cobbe, J., & Norval, C. (2019). Decision provenance: Harnessing data flow for accountable systems. *IEEE Access*, *7*, 6562–6574. https://doi.org/10.1109/ACCESS.2018.2887201

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, *51*(5), 991–1006. https://doi.org/10.1006/ijhc.1999.0252

Smaldino, P. (2016). Why isn't science better? Look at career incentives. *The Conversation*, https://theconversation.com/why-isnt-science-better-look-at-career-incentives-65619

Smith, V. S. (2009). Data publication: Towards a database of everything. *BMC Research Notes*, *2*, 113. https://doi.org/10.1186/1756-0500-2-113, PubMed: 19552813

Soberón, G., Aroyo, L., Welty, C., Inel, O., Lin, H., & Overmeen, M. (2013). Measuring crowd truth: Disagreement metrics combined with worker behavior filters. In *CrowdSem 2013 Workshop*. https://ceur-ws.org/Vol-1030/paper-07.pdf

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, *61*(12), 2544–2558. https://doi.org/10.1002/asi.21416

Tinsley, H. E., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology*, *22*(4), 358. https://doi.org/10.1037/h0076640

Tong, A., Sainsbury, P., & Craig, J. (2007). Consolidated criteria for reporting qualitative research (COREQ): A 32-item checklist for interviews and focus groups. *International Journal for Quality in Health Care*, *19*(6), 349–357. https://doi.org/10.1093/intqhc/mzm042, PubMed: 17872937

van der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. *Computing in Science Engineering*, *13*(2), 22–30. https://doi.org/10.1109/MCSE.2011.37

van Rossum, G. (1995). Python Library Reference. https://ir.cwi.nl/pub/5009/05009D.pdf

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLOS Medicine*, *15*(11), e1002689. https://doi.org/10.1371/journal.pmed.1002689, PubMed: 30399149

Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Ostblom, J., … Qalieh, A. (2018). Seaborn: Statistical data visualization using Matplotlib. https://doi.org/10.5281/zenodo.592845

Welch, B. L. (1947). The generalization of Student's problem when several different population variances are involved. *Biometrika*, *34*(1/2), 28–35. https://doi.org/10.2307/2332510

Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., & Teal, T. K. (2017). Good enough practices in scientific computing. *PLOS Computational Biology*, *13*(6), e1005510. https://doi.org/10.1371/journal.pcbi.1005510, PubMed: 28640806

Wu, X., & Zhang, X. (2016). Automated inference on criminality using face images. *arXiv:1611.04135 [cs]*. https://arxiv.org/abs/1611.04135

Ye, Q.-H., Qin, L.-X., Forgues, M., He, P., Kim, J. W., … Wang, X. W. (2003). Predicting hepatitis B virus–positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning. *Nature Medicine*, *9*(4), 416–423. https://doi.org/10.1038/nm843, PubMed: 12640447

Zafar, M. B., Valera, I., Rogriguez, M. G., & Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics* (pp. 962–970). PMLR. https://proceedings.mlr.press/v54/zafar17a.html

Zimring, J. (2019). We're incentivizing bad science. *Scientific American*, https://blogs.scientificamerican.com/observations/were-incentivizing-bad-science/

Zuckerman, H. (2020). Is "the time ripe" for quantitative research on misconduct in science? *Quantitative Science Studies*, *1*(3), 945–958. https://doi.org/10.1162/qss_a_00065