From Bias to Repair: Error as a Site of Collaboration and Negotiation in Applied Data Science Work

CINDY KAIYING LIN, Pennsylvania State University, USA
STEVEN J. JACKSON, Cornell University, USA

Managing error has become an increasingly central and contested arena within data science work. While recent scholarship in artificial intelligence and machine learning has focused on limiting and eliminating error, practitioners have long used error as a site of collaboration and learning vis-à-vis labelers, domain experts, and the worlds data scientists seek to model and understand. Drawing from work in CSCW, STS, HCML, and repair studies, as well as from multi-sited ethnographic fieldwork within a government institution and a non-profit organization, we move beyond the notion of error as an edge case or anomaly to make three basic arguments. First, error discloses or calls to attention existing structures of collaboration unseen or underappreciated under 'working' systems. Second, error calls into being new forms and sites of collaboration (including, sometimes, new actors). Third, error redeploys old sites and actors in new ways, often through restructuring relations of hierarchy and expertise which recenter or devalue the position of different actors. We conclude by discussing how an artful living with error can better support the creative strategies of negotiation and adjustment which data scientists and their collaborators engage in when faced with disruption, breakdown, and friction in their work.

CCS Concepts: • **Human-centered computing** → **Collaborative and social computing** → **Collaborative and social computing design and evaluation methods** → **Ethnographic studies**

Additional Key Words and Phrases: Error; Data Science; Machine Learning; Critical Data Studies; Repair; AI ethics

ACM Reference format:
Cindy Kaiying Lin and Steven J. Jackson. 2023. From Bias to Repair: Error as a Site of Collaboration and Negotiation in Applied Data Science Work. *Proc. ACM Hum.-Comput. Interact*, 7, CSCW1, Article 131 (April 2023), 32 pages, https://doi.org/10.1145/3579607.

1 INTRODUCTION

”The Foundations of AI are riddled with error,” writes *WIRED* journalist Will Knight [57], noting that widely used artificial intelligence (AI) datasets such as ImageNet are erroneous because they either contain test datasets overrepresenting certain populations, objects, or languages; or low-quality training datasets produced by underpaid and disinterested crowd workers. Natural Language Processing (NLP) experts Emily Bender et al. have similarly critiqued the rise of large language models, noting that many are built on internet datasets which skew toward younger populations, white men, and the Global North [12]. Gordon et al. describe the source of low-quality training datasets as noise and note that one way to ameliorate noise is through a strategy

2573-0142/2023/04 – Art 131 $15.00

https://doi.org/10.1145/3579607

of aggregation that packages discrepant annotator perceptions into "a single ground truth label" [38]. They recommend that data science teams pay attention to the "messy realities of our lives" by embracing the "contestation, disagreement, and deliberation" of annotation practices that more fully represent real-world behavior.

There is much to embrace in this line of critique. A proper and public accounting of error is essential to a long overdue reckoning, now underway, with the fallibility and limits of AI systems. This accounting includes recognition of how errors may stack or layer: that there are *systematic* rather than randomly distributed arrangements of error which collectively produce patterns of unequal effect in the world; and that the accumulated weight of prior stackings of error has contributed significantly to the racialized, gendered, classed, etc. configurations of the world [19][32] [34] [82] [86] [98]. From this vantage, the much recognized but imprecisely defined problems of bias may be no more and no less than the patterned outcomes of an accumulated and unequal history of error. As Meredith Broussard has recently argued, errors are not just accidents [20]. Similarly, Ruha Benjamin asserts that we often think about racism "as an aberration, a glitch, an accident, an isolated incident, a bad apple in the backwoods and outdated" [13]. But a glitch is rarely 'just a glitch.'

One response to this situation, commonly adopted within the machine learning (ML) community, is to double down on the accuracy of labels, and find new and more careful ways of testing models which build a stronger correspondence between modeled outputs and their real-world counterparts. A recent *Science* article, for instance, documents how ML practitioners have attempted to address the failures of models by designing tougher benchmarks (e.g., using different source imagery for training and testing data) [79], and new evaluation metrics which reward a model's ability to perform well against several benchmarks, attuned variously to the "accuracy, speed, memory usage, fairness, and robustness" of models [65]. These methods focus on eliminating errors by deriving a "correct" label most able to mimic the world "out there".

These are important and necessary efforts. Attending to error is both essential for the refinement of existing systems (especially where and when they tip in systematic and socially consequential ways) as well as a much-needed acknowledgment of modesty and fallibility, built around a forthright accounting of what AI can and cannot do in the world. But to conceive of error solely as an obstacle or impediment to an error-free machine learning, or to focus only on the outcomes (including highly negative ones) of these processes, risks reifying the ideal of a perfect(ed) AI that is placeless, oddly people-less—an ideal we believe is false. This imagining also misses out on the *work* of error - the value of error as a site for ongoing collaboration and negotiation within ML work, and a host of concrete questions and challenges that turn out to be essential to how real-world ML systems function.

This paper addresses some specific problems of error at two distinct stages of the ML cycle: (1) data preparation and (2) model building and evaluation. In each stage, we consider two such errors: label error and generalization error. Label error arises when the assigned labels are different or discrepant from ground truth data [106]. For the purposes of this paper, we focus on how practitioners manage label errors during the manual labeling and inference of images. Machine learning researchers typically view label error as resulting from a lack of clarity in annotation tasks [33]; ambiguities of input or output data [33]; variations of judgment (as well as expertise); and the value systems and subjectivities of human annotators [81][106]. To address these perceived problems, they have developed noise-robust or tolerant algorithms to handle label errors [76]; created specific label requirements to avoid misinterpretation or manually checked for repeated errors in training datasets [15]; and have used confidence learning on label quality to search for errors and train on clean data [78], among other solutions.

Generalization error occurs when a model becomes overfitted, adapting to the conditions of its training data in a way that makes it unable to perform well with unseen data. For instance, a ML model trained to classify wheat crops in South Africa might not do well with smallholder wheat crops in Ethiopia because it is trained to recognize the specific features and location of commercial-scale crops instead. Generalization error impacts how a single machine learning model can be flexibly and adaptively used across different datasets, thus, practitioners have developed a variety of measures to address it. Practitioners use a variety of regularization techniques which discourage the model from becoming too embedded in the particularistic details of training sets, or train the model on more (and more varied) examples. More recently, advancements in deep learning have allowed practitioners to introduce adversarial data or noise into training data to make the model more robust to noise from new data sets [59][88].

The above measures show how ML practitioners aim to resolve errors by either developing technical measures to control and measure errors or embracing the collection of more data as a way to limit and forestall errors. As such, the focus of ML practitioners in both academia and industry has been to improve the quality and range of training data. This work follows a 'limit and eliminate' approach, in which the value of errors is entirely negative; *error functions first and foremost as an obstacle to be overcome, in pursuit of ML models that correspond more accurately and comprehensively to the conditions of a complex world.*

But this perspective misses other aspects of the important work that error does, especially in collaborative work settings. We now consider how errors provide opportunities for data science experts and their collaborators to do three things. First, errors reveal *existing* structures of collaboration that go unseen or unappreciated within supposedly working systems. Second, errors animate new forms and sites of collaboration (including, sometimes, new actors). Lastly, errors rework old sites and actors in new ways, by restructuring relations of hierarchy and expertise that may alternately recenter or devalue the position of different actors. Errors show us how actors come together to negotiate, collaborate, and repair inherently glitchy and uneven systems.

In this paper, we will show how errors in data science drive social interactions and collaborative networks that reconfigure and reposition actors and their knowledge practices. More specifically, we show *how* errors are identified (or not identified) and *whose* estimations of error are ultimately heeded and addressed. Furthermore, we show how errors identified during data preparation are inextricably tied to model building and evaluation. In doing so, we illustrate that the stages of ML development are never wholly separated. We ask the following questions: How is error understood in ML? How is it differently perceived and managed by modelers, domain experts, and annotators? How do groups and teams, near and far, *collaborate* around errors, determine which errors are acceptable, and decide which ones are to be fixed? Under what conditions are errors identified, negotiated, and repaired, and under what conditions are they simply accepted and lived with?

We draw from ethnographic examinations of two sites to make our case. The first is a government research institution in Indonesia, working with local mappers to integrate data science into national mapping efforts. The second is a North American non-profit organization working with data labelers in the Middle East to develop open-access machine learning datasets and models for the purpose of vaccine distribution. Through closely thinking with their work, we illuminate how organizations address errors, revealing the deeply sociotechnical nature of error management efforts [1].

In addition to thinking through how ML practitioners limit and eliminate error, we depict three alternative ways of conceptually engaging with error: errors as revealing existing collaborations, initiating new collaborative structures, and reworking collaborative relations and hierarchies. We argue that errors are navigated differently by, and have varied impacts on, the actors involved. When our interlocutors treat errors *only* as obstacles to be overcome, they reinforce epistemic, and workplace hierarchies between scientific professionals and "lowly" paid technicians or annotators. Approaches that instead view errors as sources of information or new patterns (as an 'occasion' for learning, in the educational philosophy of John Dewey [30] [31]) may support forms of collaboration across domain expertise and data scientists that allow for unexpected possibilities – both within data practice, and vis-à-vis the wider worlds that data practice and AI touches.

The sections that follow begin by reviewing technical literature on label error and generalization error, interweaving findings from social science literature on error in technoscience and ongoing CSCW and HCI literature on breakdown, maintenance, repair, and the emergent field of human-centered machine learning (HCML). We describe our research methods and provide two case studies that illustrate the multifarious ways organizations and experts respond to generalization and label errors. We conclude by discussing the implications of taking error seriously as a central site of data science practice, what an artful living with error would entail, and error's broader impact on CSCW research and practice.

2 LITERATURE REVIEW

2.1 Limit and Eliminate: Machine Learning and the Problem of Error

Machine learning (ML) practitioners have long been confronted with the problem of error: the myriad glitches, breakdowns, and failures which challenge the basic work practices of data scientists and limit the effectiveness of and confidence in their results. Errors can result from the misclassifications of target variables, such as a ML model mistaking trees for buildings. Errors can also prevent ML models from generalizing across a variety of datasets, either because the training data differs greatly from the "real-world" data, or because the data itself is dynamic and regularly updated (e.g., email content).

Within this set, certain kinds of errors have been called out for attention. The first is *label error.* Label error occurs when annotators produce "noisy labels" in training data. Practitioners point out that these noisy labels are common in computer vision, given that while there are "standardized techniques" to clean tabular data, they are "less suited for perception e.g., pixel of an image" [[55]: 1]. ML practitioners view label error as a result of disagreement and variance in classifications among labelers. There is widespread acknowledgment among practitioners that such noisy labels are present not only within training datasets, but also test datasets [15][78]. Various attempts have been made to reduce such noise, driven by the fear that noisy labels compromise the quality of the training dataset itself [44]. To improve the quality of data, ML practitioners have primarily adopted a majority vote to decide on a single label [25]. Here, experts labeling pixels vote for what their labels represent in the event of a disagreement. In our second case study, a group of remote sensing scientists and data scientists use a majority vote to decide whether a building roof blocked by trees ought to be included as a label. A majority vote would decide if a label belonged.

Majority vote is not foolproof. ML practitioners, both within HCI and the larger computing community, argue that the majority vote is used "to construct a ground truth without preserving information about label distributions." [[47]:4]. Various human-centered machine learning measures to retain the distribution of labels include relying on soft labels (probabilistic labels instead of a single definite label) [47] and ensuring intra-annotator consistency by paying attention to the dominant labels an annotator would provide when classifying the same item repeatedly [38]. Davani et al. have embraced *both* disagreement between and inconsistency within annotators because they argue that such variance provides more "flexibility" and "better estimates for uncertainty" to model the real-life behavior of annotations [25]. In sum, ML practitioners are not only aware of the need to control the perceptions of annotators by aggregating or predicting their disagreements but have embraced the notion that there are *multiple ground truths* which could yield better-performing models. Our work builds from such technically minded strategies by showing *how* ML practitioners harness disagreement to embrace uncertainty and multiplicity in perceptions of error.

A second, more commonly recognized, kind of error is *generalization error.* It appears when models are brought into production. When models are tested on "real-world" data, they might be unable to grapple with newly ingested data. Hullman et al. refer to the problem of generalizability as "methods working well under certain conditions but fail[ing] when applied to new problems or in the world" [47]. While generalization errors may be caused by multiple factors ranging from "inadequate feature representation" to "small and imbalanced datasets", we hone in on the phenomenon of *data drift.* Data drift occurs when there is a change in the relationship between input and output data. For instance, a data scientist might encounter a natural drift in data such as mean temperatures changing with the season [73], and must account for such shifts in order to increase the model's generalizability. As generalization error deals directly with the extent to which an algorithm can be deployed to do the same task on different datasets, practitioners have developed a wide array of measures to limit the generalization error rate.

Practitioners typically use a variety of regularization techniques to approach these problems. These techniques include ensuring that a model is not overtrained, or does not have too many features, factors which make it difficult for a model to adapt to unseen data. Another common regularization technique is data augmentation, which makes slight modifications to training datasets such as rotating or flipping images, in order to make them more complicated and suitable for complex models [83]. In effect, data augmentation increases the size of the training dataset, and is considered by practitioners as a necessary step for deep learning models in particular [94]. Augmentation would be followed by the reweighting of samples in training data so as to prioritize the newly added training data. Other options include using a variety of machine learning models for different segments of the training data, or even changing the prediction target [35].

The responses to both generalization and label errors regard poor ML model performance as an isolated problem, one remedied through tinkering with datasets and model architecture. This tinkering can include statistical and mathematical responses such as including new uncertainty and model evaluation metrics, or efforts to reduce model parameters to minimize or reduce overfitting. Yet such responses often frame errors as anomalous "edge" events, rather than as occasions collaboratively managed by data scientists, domain experts, and data annotators [89]. Domain experts, especially, are often portrayed as playing a subsidiary and passive 'informant' role in the design and operation of ML systems [33]. Furthermore, statistical and data science measures reify the "limit and eliminate" approach to error, rather than paying attention to the generative value error provides by transforming how we know a particular phenomenon, learn who to work with, and who to be instructed by.

2.2 Negotiating Error: From Elimination to Repair

A second approach to the problem of error may be found in a growing body of work in CSCW, HCI, science and technology studies (STS), and critical data studies. In human-centered machine learning (HCML) [4][5] [69] [70][71] [72], scholars have examined how data science has differential impacts on marginalized populations, drawing attention to the relational aspects of technology and society while decentering its deterministic effects in projects of modernity and development. Centrally, Stevie Chancellor has recommended embracing failure as part of a more balanced and holistic approach to data science. Instead of emphasizing "technical conceptions of performance through quantitative metrics, such as evaluating error rates and efficiency," a sociotechnical perspective on failure encourages technologists to grapple with who is impacted by the failures of AI systems and how they navigate such breakdowns [23]. For instance, it is found that medical professionals lose trust in AI systems that fail to return clinically-relevant results to provide a diagnosis on the fly, even if the same system is relevant for other clinical cases [[21] cited in [23]]. As such, we understand error and allied concepts such as breakdown and failure as a new analytic to understand how teams, groups, and organizations reorganize in the face of error, and thus repair what has stopped functioning as intended.

2.2.1 Error as a site of hierarchy and power dynamics. Recent work on algorithmic bias, data annotation, and AI failures have shown that the identification and disclosure of error rely on the power relations and hierarchies of a workplace. CSCW scholars Taylor et al. discuss how nurses who are well-equipped to identify patient safety errors are often disincentivized from speaking up due to asymmetric power relations and strict hierarchies between physicians and nurses. They found that nurses desire to displace the responsibility of error communication to intelligent technologies such as robots, seeing them as "neutral third part[ies]" [[100]: 221:3].

In other workplace settings, Miceli et al. argue that the failures of data-driven technologies have often been narrowed down to a problem of bias by annotators who are often employed on a precarious basis [68]. As such, bias is often resolved by scrutinizing the quality of training data, instead of taking account of the "labor conditions, institutional practices, infrastructures, and epistemological stances encoded into datasets". Nithya Sambasivan and Rajesh Veeraraghavan have shown how AI developers may moralize error (and disclaim responsibility), characterizing field workers collecting training data as lazy, corrupt, and careless, further justifying the use of surveillance, automated checking tools, and cross-verification to eliminate error in training data [91].

Error is also differently perceived across different kinds of experts and professionals, creating friction between AI systems and the practices and material context of domain experts. For instance, Jung et al. studied how craft brewers using a digitized brewing system found it difficult to understand abstract digital values, sought less optimal target outputs, and preferred working more flexibly with their materials [53]. CSCW scholars have also discussed how domain experts are included in the process of ML development to reinforce the myth of AI systems as functioning, seamless systems [54] [58] [75]. Raji et al. in particular have noted a global push to include a variety of domain experts in AI tool development for COVID-19 in order to democratize the development of AI, with little to no concern placed on the actual functionality of the AI tools themselves [84]. Hence, while domain experts ranging from data annotators to specialized professionals play an important role in identifying errors specific to a field or discipline, they are sometimes regarded as lowly skilled, non-technical, or mere participants in ML development. In sum, these CSCW and HCML readings of error point to how failures and breakdowns are sites to study relations of hierarchy and expertise in data science collaborations; they do the work of making known power dynamics and resistance, as well as varying definitions of precision and accuracy.

2.2.2 Error as a site of collective work. A second line of work in STS and the social sciences has shown how errors are sites for collaborative, collective, and coordination work [16][22] [37], [42][52] [45]. Historian of science Lorraine Daston argues that diagnosing, eliminating, or—at a minimum— "taming" errors has been a central feature of modern science, constructing vigilant scientific subjects [26]. This vigilance is trained not through instituting individual perceptions, but by "collective seeing and naming" made possible by standardized descriptions of natural phenomena [27].

The management of error is also distributed throughout a network of practitioners and technologies. Sociologist of science and technology Donald MacKenzie's groundbreaking study of U.S. missile guidance in mid-century America showed how errors in ballistic missile guidance technologies were addressed through conflict and collaboration between the laboratories and corporations variously headed and directed by technologists, military members, and political figures [63]. In later work on computer system failures, MacKenzie critiques the category of "computer-related accidental death" because it reduces technical failure into a self-evident category [64]. Instead, he reveals how system design often contributes to human error. Even in contexts where humans appear to be absent, such as software errors, MacKenzie argues that multi-causality is the rule rather than the exception.

In a more recent context, sociologist of science Adrian Mackenzie notes how backpropagation, an algorithm that optimizes the connections between nodes in a neural network based on the error rate obtained in the previous iteration, is reinscribed in the actions of practitioners in a data science competition who compare the error rates of their models and iteratively optimize their model, thus replicating the self-adjusting optimization process commonly found within the algorithm [62]. In other words, the practice of machine learning itself shapes and structures the world these practitioners live in.

These dynamics complicate not only who should be held to account for system breakdowns, but also who gets to speak on behalf of such failures and hence develop strategies for mitigating them. Historian of technology Rebecca Slayton has shown how a group of elite software engineers in the mid-twentieth century United States questioned whether the software necessary to control an anti-ballistic missile (ABM) system could ever be made to work without error.

These computer scientists, including MIT's Joseph Weizenbaum, argued that reliable software could not be produced without rounds of debugging, real-world testing, and revision, even in stable, well-understood situations. [97]. Slayton's analysis reveals how computer scientists and software engineers came to be perceived by the US public as capable of speaking about the risks of complex computer systems, instead of independent third parties or social scientists.

The dominance of computer engineers diagnosing and 'speaking for' error in AI persists. Recent studies of AI ethics have shown how technology companies occasionally used errors to explain why it is important to improve existing products and efforts to "build better", mostly behind closed doors [1][84]. This makes it important, as STS and communication scholar Mike Ananny argues, to view algorithmic errors as "public problems" [1]. If errors are sites to expose that something unjust has been done, there should be stronger efforts to "reveal communities of algorithmic error – people who see and diagnose errors similarly, who strive for fixes together" [2]. This proposal to study how communities diagnose, engage, and become the authority on errors aligns with our next point on how errors animate collaborative structures and repair.

2.2.3 Error as a site of collaboration and repair. These insights around the complexity of breakdown and error are echoed in a body of CSCW work in repair studies (see for e.g., [46][49] [50] [87]). As this work makes clear, breakdown is not the exception case to the normal workings of technical systems, but rather their *ordinary mode of operation* – that is, technical systems (like other instances of established order and stability in the world) are in practice regularly breaking down, and sustained only through complex and ongoing acts of maintenance and repair (which are nevertheless commonly rendered invisible by our predilection for other figures and moments – for example, design and designers, use and users [8]). This understanding places error (and responses thereto) at the very center of technical process, rather than as an occasional and unfortunate edge case. It also locates error as a key site of learning and innovation, and very often one around which multiple interests and kinds of expertise group.

The potentially creative and collaborative nature of error is perhaps best exemplified by Klemp et al.'s [56] remarkable study of the role of error in jazz improvisation. Drawing on pragmatist philosophy (in particular, the pedagogical theories of John Dewey [29]), the piece follows the consequences of a single 'wrong' note struck in a 1958 recording of 'In Walked Bud' by Thelonious Monk - and the complex work of repair by Monk and members of his quintet to make the wrong note 'right': that is, fit within a redefined pattern and imagination of the piece. This work is deeply collaborative in nature, relying on the careful attunement and listening between members (itself aided by a longer history of familiarity and collaboration). It is also essentially creative, and one of the principal engines by which new sounds and insights are brought into worlds of jazz. Responses to error in data science we believe hold a similar potential for producing new ways of collaborating with one another and addressing failures and ruptures in more open, creative, and effective ways.

At stake in all these debates is the status of error - both as an analytic matter (how are we to conceive and think about error in a theoretical sense?) but far more importantly as a practical one: how do *actors themselves* encounter, make sense of, and work around error amidst the ongoing glitchiness and uncertainty of data science work? How do they do this work collaboratively, within teams and arrangements of differently placed interests and expertise? How are repairs and responses to errors effected, and how are judgments of accuracy, reliability, or 'good enough-ness' arrived at through these processes? The following sections explore these questions in two separate cases: the effort to establish an error-free map amongst government-contracted surveyors in Indonesia and the negotiation of acceptable error rates in the processing and annotation of satellite imagery.

3 METHODS AND FIELDSITES

The two following case studies build on in-person and virtual ethnographic fieldwork and interviews conducted at an Indonesian government research institution (National Mapping Agency) in Jakarta and a North American-based non-profit organization (Starlight). It is derived from more than 4 years of ethnographic study conducted by the first author in three rounds: from the summers of 2016 to 2018, March 2019 to June 2020, and virtually from August 2021 to July 2022. Recruitment occurred based on the first author establishing herself as a research fellow in

Starlight in 2021. Starlight specializes in the development of open-access machine learning training datasets and models. Where consent was given, the first author recorded team meetings and meetings with external corporate partners and tech consultants. The first author has also worked with a member from Starlight to ensure that the names of their organization and partners are anonymized and the context of their partnerships is minimized to ensure privacy.

Prior to Starlight, the first author gained access as a research intern in several government research institutions in Indonesia, including the National Mapping Agency. She browsed government libraries, interviewed staff members, and sat in their meetings after gaining access with an Indonesian research permit. The first author conducted more than 100 semi-structured qualitative interviews in English and Bahasa Indonesia with remote sensing scientists, data scientists, geospatial data technicians and policymakers in Jakarta. All research interlocutors have been anonymized to protect their privacy. Before interviews and observation with members of these organizations, the first author shared a verbal informed consent form and interview questions.

Since data science adoption and implementation rolled out in Indonesia's government institutions over several years, the first author's rewriting of the interview protocol happened several times, centered around notions of machine learning and map accuracy and perceptions of ground truth among scientific and technical experts. Our interview protocol in Starlight addressed questions of bias in machine learning models, including challenges of how to identify them, fix them, as well as communicate about them to clients and other partner organizations.

The first author created memos which were then shared and discussed with the second author (along with fieldnotes and interview transcripts) to elaborate on emerging themes in fieldnotes and interview data. Our data analysis followed a grounded theory approach, tracing practices from how training data was made to how models were designed before arriving at the analytic and conceptual framework of error. Our initial themes in the memos included "social hierarchies," "fuzzy landscapes," "actual pixel boundary," "ground truth," and "resolution" to conceptualize moments when fieldwork participants were triaging to identify image classification mistakes and fix them. Looking through these themes, we recognized that errors were the common denominator holding them together. With this realization, we returned to the practitioners to interview them further. Our coding and analysis went through several additional rounds, where themes of "precision," "repair," and "efficiency" were developed (and sometimes discarded or demoted) before arriving at the themes at work in the present paper.

Across both large government institutions and small non-profit organizations, we wished to observe how organizational structures and norms of collaboration and professional hierarchy shaped data science practices. To navigate these different settings, the first author leveraged her privilege as an information scholar first, at a large R1 public university, and second, as an information postdoctoral fellow in an Ivy League university to access these typically tightly controlled organizations. On the one hand, the first author had to ensure that her access to their clients and partners was treated with ample care and consideration, given that some of the products they worked on had different extents of openness to public use and sharing. On the other hand, the privilege of the first author is part and parcel of the stacking and recognition of the hierarchy of expertise that shapes the environments in which errors are constructed and discussed by her interlocutors. In Indonesia, the first author's positionality as a Singaporean Chinese woman both provided as well as challenged her full access to government offices on many fronts (the former due to her esteemed educational background in Singapore and the U.S.

and the latter based on her Chinese identity given longstanding racial tensions between Chinese and non-Chinese Indonesians). In some ways, the first author's transnational mobility, flexibility, and expertise allowed her to navigate both worlds with ease, as much as it presents its perils and limitations.

While the United States and Indonesia are different countries with varying norms on culture, technological advancements, and demographics, we found it essential to understand how these two collaborations apply machine learning to solve problems of resource management and landscape recognition in the Global South. In Indonesia, the National Mapping Agency has turned to data science to speed up the production of national maps since 2016. This effort began shortly before the National Strategy for Artificial Intelligence was established in 2020, an initiative that signaled the growing entanglement between state knowledge production and data science research. Since then, government researchers have heavily promoted the use of data science and AI in a range of applications, from biotechnology to environmental sciences, driven in part by growing fears of redundancy in the face of data-driven sciences. In 2019, Indonesian President Joko Widodo threatened to replace civil servants with AI. While this did not happen, a merger of more than 33 government research agencies in 2022 resulted in the layoff of hundreds of researchers, technicians, and research assistants. It is in this context of job scarcity and economic instability that we attempt to understand how errors in machine learning become sites to construct and legitimize new (and old) methods of knowledge production.

Starlight, on the other hand, acts as an intermediary between a data annotation company we have named Tarik and their client, Greenworld, a US-based consultancy that supports development and sustainability projects. A United Nations agency has commissioned Greenworld to develop a building detection model that can automate the creation of high-resolution building maps in Dhaka, Bangladesh. In addition to Bangladesh, the data of other Global South countries would be included to train the same model. These high-resolution building maps were to be used by health professionals to plan and estimate the population size in any particular neighborhood, and place mobile COVID-19 and childhood vaccination clinics accordingly. Greenworld's accurate, high-resolution building detection models ensured that the UN agency could quickly produce updated maps that would serve as proxies for population sizes. The goal is for the agency to know where exactly to place mobile vaccination clinics and maximize the impact and access to vaccines.

4 FINDINGS

In the following sections, we analyze how errors in data science are sites to disclose and reorganize structures of collaboration between differently positioned experts and data annotators/technicians. Our understanding draws upon existing CSCW, STS, and HCI work on how errors cannot be viewed simply as an obstacle to knowledge production but as an avenue for collective work, the enactment of hierarchy and power dynamics, and the reorganization of collaborations across and within organizations. Our empirical study shows how different institutions work with one another to identify errors, make sense of them, and decide what is necessary to address. Together, these two cases will show how errors reveal existing collaborations critical for sociotechnical systems to work, initiate new networks of relation and collaboration, and restructure current arrangements to either make way or further devalue the labor and expertise of those with lesser power.

4.1 Building a Collective Map in Indonesia

Our first case follows the role of label error in an ambitious machine learning and remote sensing-driven national remapping project in Indonesia. It begins with Bayu, an earth scientist (geodesy) trained in machine learning during his graduate studies at ITC Delft in the Netherlands. In his office near the outskirts of Jakarta, hundreds of cartographers have been trained to map Indonesia since 1969. At the time we spoke in 2017, Bayu was prototyping a method for automating the detection and mapping of buildings in Indonesia. This method would include a standardized machine learning model for detecting Indonesia's buildings and assist in making the nation's first-ever complete and large-scale (1: 5000) topographic map.

That afternoon, earth scientists in the mapping agency told us that Bayu's method could radically speed up the production of maps - a promising contribution for a severely overstretched agency charged with producing accurate and effective maps across Indonesia's more than 17,000 islands. We were seated in one of the agency's meeting rooms to learn about Bayu's work, sharing our opinions on why machine learning mattered for the nation's topographic maps.

A senior official at the agency explained that as this was the first time Indonesia had made large-scale topographic maps, the workforce required had to be highly trained and large in number. Another earth scientist had heard from his colleague conducting quality control on maps that the subcontracted private geospatial data operators had little experience classifying features on high-resolution remote sensing data. With a scarcity of mappers trained to make these maps, Bayu believed a machine learning approach could help detect and classify buildings quicker and more accurately. Given this, Bayu and his research team found machine learning to be instrumental in easing the labor and time required to identify and classify buildings.

After describing why machine learning was important for the nation's maps, audience members watched Bayu pull up an image. It was an image of a network of data points overlaid on top of a satellite image. He said, "This is Light Detection and Ranging (LiDAR) data. The Agency has recently introduced LiDAR to data operators to help them map faster."

What does fast mean here? Asked one senior earth scientist. Bayu waited to answer the question. Instead, he explained that each LiDAR data point has three features: intensity values (i.e., the amount of light energy recorded), elevation values, and a return number (i.e., the total number of returns for a given laser pulse) recorded from a remote sensing device fitted to an airplane or satellite. These values, Bayu explained, were then compiled into a 3D data "point cloud." This point cloud was different from a 2D pixelated satellite image because it emphasized features that data technicians could not observe from plain sight. Bayu showed a LiDAR point cloud of Germany that he had extracted from satellite imagery, drawing the eye to the yellow roofs that now popped in contrast to the black background.

Still, not all the earth scientists at the table were satisfied. They probed into how such a network of points could guide an operator's classification work. One of Bayu's long-term collaborators in the agency, Faizah, guided the audience through how to read a point cloud. "Unlike a point cloud, which intuition will tell you to join together based on the proximity of dots, a single pixel gives too much information. Pixels make an operator imagine too much." Furthermore, with each data point, numerical information such as the elevation height of where the light pulse hits a particular landscape is provided to help an operator decide if data points should be joined. For Bayu and Faizah, the distance of data points and the information each point had assisted data operators in classifying features on a high-resolution satellite imagery. Pixels on a high-resolution imagery would have provided too much detail to operators and made them prone to label errors.

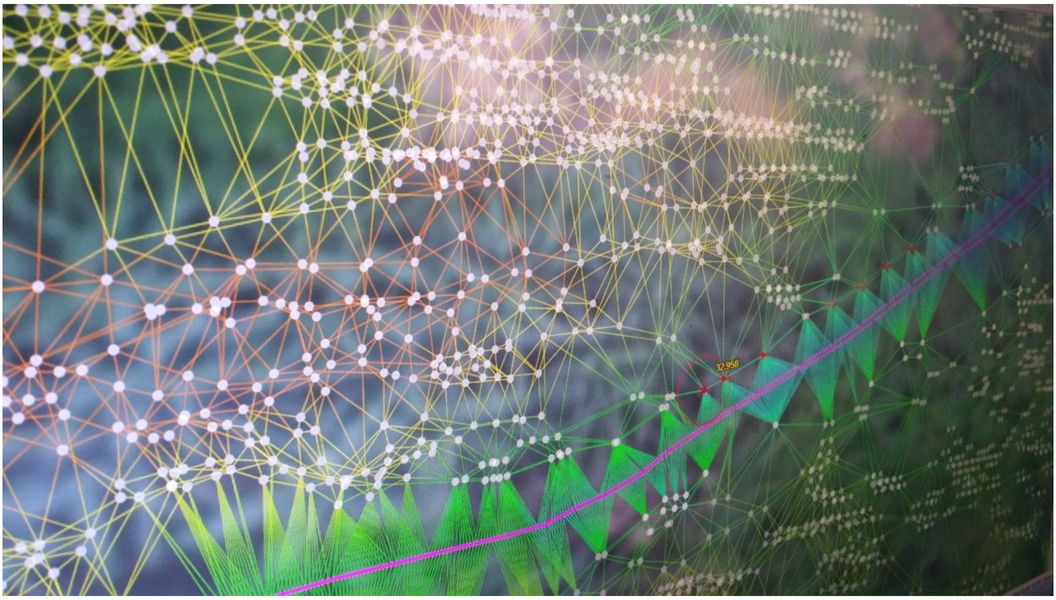

Fig. 1. This is an image of a LiDAR point cloud. Each shaded dot represents where a laser has hit the surface of a landscape. The different colors of the lines emanating from each dot indicate different elevation values. For example, the purple line represents a circle, and surrounding it are lines of a similar color because the elevation values of a riverbed are roughly identical (first author's photo).

4.1.1 Error initiates collaboration between senior and junior earth scientists.

To further illustrate the significance of using LiDAR, Bayu, and Faizah continued to guide the audience on how to read a point cloud in two ways. First, from aerial satellite imagery, Faizah showed how data operators could now "see through" forest cover. LiDAR penetrated the gaps of forest branches and leaves to spot the road, river, or building underneath. Second, operators could use the elevation value of each LiDAR point to classify features according to mapping standards. Since each LiDAR point had the elevation value of any object, operators could refer to the height of trees and confirm that a dense plot of trees they were looking at was indeed a forest. For instance, in Indonesia, forest cover is defined as an area of trees with >5 m height, >30% canopy cover, and excludes plantations, such as oil palm. True enough, when we were on site with data operators, operators often referred to LiDAR to confirm their vision and coordinate with others on hard-to-see objects.

At this point, there were questions for Bayu on whether LiDAR could only be used to classify buildings. Bayu chuckled, "I prefer using LiDAR point clouds for buildings...because the roofs are more regular than other landscapes". He clicked to the next slide and outlined a building's roof with his index finger. Underneath his finger was red LiDAR data points around a sea of blue LiDAR data points. No matter where you are, the roofs of buildings are fairly similar, Bayu added.

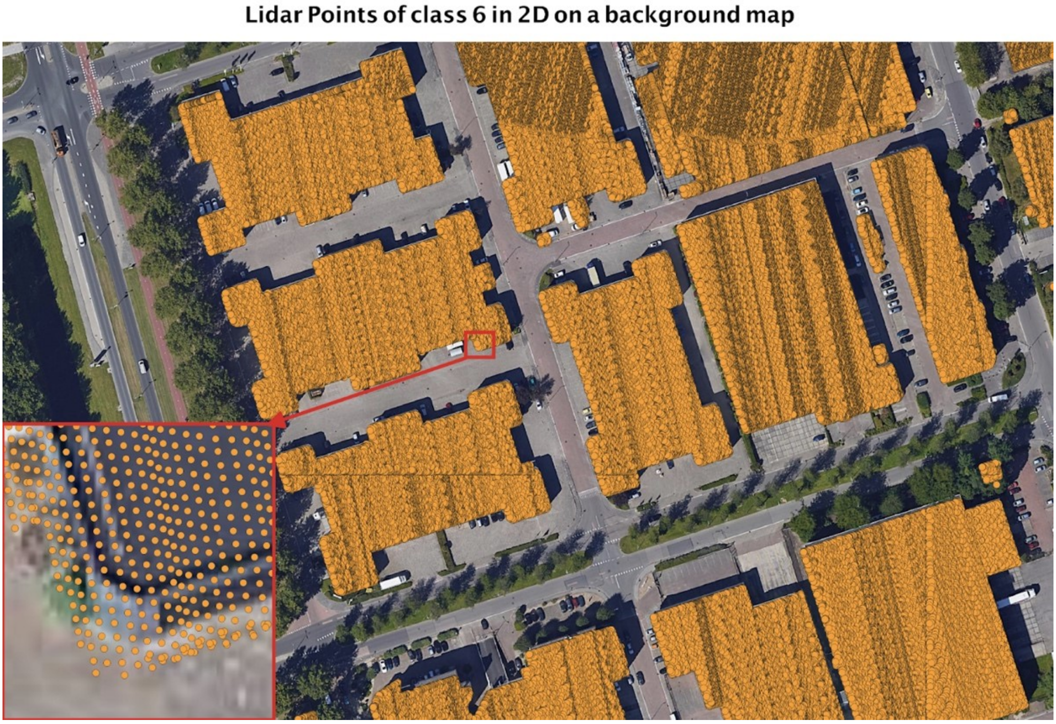

Fig. 2. This is not an image of Bayu's LiDAR data points, but it is used to illustrate what he regards as the edge data points of a building roof that would not have same elevation values as compared to the ground that sits next to it. Image credit to [10].

But that was only one type of building one senior earth scientist observed. Bayu pointed out that slums in Indonesia were challenging for his model to recognize. But he again emphasized that the elevation values of other building roofs would not differ much. At this point, a conversation ensued that brought out a larger set of questions. What exactly was a building? What about buildings that were blocked by vegetation? Or buildings that were too close to one another? When does the universality of buildings break down? And what had Bayu done to address this error?

Bayu showed a point cloud that was denser than the one shown earlier. It had surfaces so smooth that it felt like there were barely any gaps between points. Furthermore, the data points were color-coded such that points with similar elevation values possess the same color. As taught before by Bayu, points that are close in distance to one another and have identical elevation values can be regarded as belonging to the same group. He pointed at the edges of a building, and we looked at them again. He said, *I call these points, edge points.*

Bayu clicked on his slides again; this time, it showed that the space around buildings also possessed a different color. He called these points ground. Based on his chosen criteria that LiDAR points with similar elevation values belong to the same group, Bayu could now identify ground from non-ground data points, further allowing him to repair the problem of buildings that are too close to one another (see Fig. 2. above). When training a model, Bayu explained, even a slight space between points could become extremely large for machinic vision.

These images are what Bayu used to train a fully convolutional network (FCN) to classify point clouds into buildings. The model worked on two assumptions: First, a building's surface consisted of LIDAR data points with similar elevation values. That is, all roofs had a similar height. Second, a building's edges had points that did not have neighboring points on all of its sides. His model looped around these parameters such that it outlined the edge points of a building by itself. Distinguishing ground from non-ground at the level of every pixel, his model was considered by his superiors after the demo more precise than the data technician's technique of classifying pixels by "manual visual interpretation." Manual interpretation involves looking at the tone ("brightness," "contrast"), texture ("smoothness," "roughness"), pattern, shape, size, and association of different land features [66]. This type of classification hence produced a variety of outputs, instead of a single definite enumeration of what is perceived. The meeting concluded with no clear direction of whether Bayu's method was suitable for mapping out the whole of Indonesia, leaving the machine learning project for national mapping an incomplete one.

In a later interview, however, Bayu revealed that his FCN model had failed to classify features unique to Indonesia - a problem he attributed in part to the model's training data being generated from urban northern Europe. Here is where we can see how label errors also affected the generalizability of the model to other places. To correct this, he attempted to add more object classes to the ground truth image. These included "Uncleared Land" and "Waste," categories informed by his living in Jakarta, a metropolis well-known to be heavily polluted. Even though Bayu was initially adamant about automating the detection of buildings and ground, he had to add more classes to account for the complexity of "ground." As the model began to learn new classes, it became apparent that the collective seeing and discussion of Bayu and the senior earth scientists was central to making the model work.

Despite the presentation's emphasis on seamless and automated mapping, we also learned a year later that manual interpretation remained vital to Indonesia's topographic mapping projects and that Bayu's model was not fully adopted to map Indonesia. So what happened to LiDAR data purchased and processed by the National Mapping Agency? They were sent to data operators who were tasked with smoothing the elevation values of the LiDAR point cloud to make a good digital elevation model, a necessary component of large-scale maps. As "raw" elevation models made from LiDAR have minor variations and outliers, it was the job of technicians to smooth these out - a notably monotonous and laborious task. Seized by frustration, operators felt that they were completing menial tasks. "A machine needs to do my job!" one of them told me after tolerating a grueling night shift performing repetitive data cleaning and processing.

There are three important features to be gleaned from this first case study. The first is how Bayu shows the smoothness and efficacy of machine learning by carefully selecting the dataset it was trained on and providing the right examples. Building out of the commonly held principle to prevent label error held by machine learning practitioners - garbage in and garbage out [44], Bayu showed how clean and controlled LiDAR data could promise good results.

But simply using LiDAR data was not enough; he had to communicate how LiDAR data can tame aspects of label error through universalizing a definition of building and convincing senior earth scientists of the value of LiDAR data. In doing so, he initiated a new collaboration between junior and senior earth scientists, getting them to *collectively* see that most, if not, all, buildings had similar structures, using LiDAR data points. Given that accelerating the production of maps is an important selling point for Bayu's audience members, his demonstration of LiDAR data shows how machine learning may overcome the chronic staffing shortages of the agency and the magnitude of the mapping task at hand. In this way, we can view Bayu's careful handling of error as initiating a new way of collectively seeing and collaboratively training a machine learning model that connects the earth sciences and computing.

The second point is that even if datasets appear to disallow any form of contingency and ambiguity, potential errors were brought forth by heterogeneously placed experts, including those with more extended familiarity with the empirical worlds at hand. Consider how senior earth scientists were concerned about how different types of building might not be identified under the schema of Bayu's edge point theory. They were not convinced that Bayu's universal definition of building roofs could cover all kinds of buildings, a perspective informed by their training in remote sensing science. As such, from the classification of data technicians to the label error of Bayu, it revealed an existing collaborative structure that earth scientists with more experience working with remote sensing imagery could expand on potential mishaps or mistakes that were not yet identified.

The third and last point is how the collaboration worked—or failed to work—when label errors were regarded solely as a result of a data technician's work. Once a data technician's judgment was viewed as problematic, it allowed earth scientists to develop assumptions on how data technicians perceived satellite imagery. While domain scientists such as Bayu were regarded as "trained" in harnessing the value of satellite imagery such as LiDAR, data technicians trained in high school were regarded as incapable of doing the same. Even when Bayu's model did not work [84], and data technicians were still entrusted with the making of the map, it was Bayu's expertise that *added* value to datasets. By accepting this logic, Bayu's audience also accepted the premise that a LiDAR training dataset could overcome the error of subjective and uncoordinated human judgment. In other words, old actors such as geospatial data technicians were redeployed to center machine learning knowledge over manual annotation and labeling. When error becomes an obstacle, it is not only a site of conflict and tension across different sites and actors, but also a place for hierarchies of expertise to be made.

4.2 Quality Control of Bangladeshi Building Detection Models

In our second case study, we show how an interdisciplinary collaboration between private and non-profit organizations works to manage and correct imprecisely labeled data, revealing the *definitionally flexible* notion of "precision" in data science. Precision is an emic term that data scientists use to talk about label errors, especially regarding human annotation. We trace the evaluation of a data quality control standard that would enable a machine learning (ML) model to learn the specificity of Bangladesh's landscapes while ensuring that the same model could recognize landscapes elsewhere. It zooms in on a moment shared between a North American tech consulting firm and its subcontracted workers from a North American non-profit organization and a data annotation company based in the Middle East.

The collaboration we describe was oriented to nominally inclusive goals - notably the current lack of geodiversity in major open-source ML training datasets (creating an imbalance in the accuracy with which ML efforts can 'see' various environments - an analog, perhaps to AI's well-recognized bias problems along racial and other lines). Many machine learning training datasets for computer vision and image recognition tasks consist of places in the Global North. For instance, in Image Net, a commonly used image dataset for computer vision and machine learning applications, around 45% of the data is from the U.S., making it difficult for image classifiers to perform well on landscapes of the Global South [94]. To create more geodiverse training datasets, the chief data scientist of Starlight developed a long-term relationship with Tarik.

Founded in 2017, Tarik employs Middle Eastern youths and displaced refugees to become annotators. Tarik differentiates itself from the crowdsourced platform and freelancer work such as Mechanical Turk because employees work on a long-term and permanent basis with regular clients. Under this model, participating youths are hired to classify and provide a series of labels for a variety of clients, ranging from land cover class labels for non-profits like Starlight to detecting car trunks for automobile companies. Projects we observed throughout our fieldwork included, but were not exhaustive of, the labeling of clouds and annotation for building roofs in Bangladesh, Oman, Madagascar, the Philippines, and India.

Starlight and Tarik are contracted to Greenworld, a US-based tech consultancy. Commissioned by a United Nations Agency, Greenworld aims to create a building detection model for Bangladesh based on annotations from Starlight and Tarik. Their main goal is to plan for the placement of COVID-19 and childhood vaccination clinics, with buildings serving as proxies for population density in low-and-middle-income countries. Greenworld is the main client in this partnership, while Starlight and Tarik have been contracted for their annotation services. Given the stakes of the project, Greenworld—in charge of building the ML model—assigned Tarik more than 100,000 satellite images to be annotated. More than 80 workers from Tarik were employed three months before the project start date in order to annotate these images. Starlight primarily mediated communication between Greenworld and Tarik. When asked why the Chief Data Scientist of Starlight gave us two reasons. First, these two organizations have different points of expertise and technical training. Most annotators in Tarik were focused on creating labels for an image, while Greenworld members honed in on the details of ML modeling. Second, and perhaps more importantly, Tarik belonged to a working culture that the Chief Data Scientist from Starlight was more familiar with, and so they could act as a cultural mediator and negotiator between Greenworld (GW) and Tarik.

In a Zoom meeting in February 2022, three core team members in Starlight were on a call with Greenworld to review training data that would be used to train an open-source building detection model. After several weeks and two rounds of data annotation, Greenworld had called for the meeting to express concerns with the results achieved so far and advocated for a 98% accuracy rate for the data annotators.

For Starlight and Tarik, 98% was extremely high in their experience – close to 10% more than what is usually expected of them. A Greenworld staff member overseeing the overall implementation of the project, Thomas, expressed his empathy with Tarik, observing that Dhaka was one of the densest cities on the planet. This meant that buildings were likely to be close to one another and difficult to distinguish from overhead satellite imagery. But he emphasized that they nevertheless wanted an error rate of 2%. Thomas reasoned that the map would provide a direct reference for population density; a near 100% accuracy for classification is needed to achieve the UN agency's vision for health microplanning in Bangladesh. He pulled up a decision tree Greenworld had created to evaluate the labels and urged all reviewers and labelers to follow a systematic process in cases of uncertain attribution:

"Basically, we are saying that if there is any uncertainty about [labels], we would like the user to go down the chain here... I do want to let you know that we want this to be as collaborative as possible, and we know where Tarik is coming from, and we want to help to make their product as accurate and appealing as possible… [But] we are looking right now at 70% accepted and a 30%

unaccepted rate which, you know, puts us in a tough position on the contractual side... if we are looking at that across the hundreds of thousands of satellite images, that's a significant portion. As much as we can afford to, for the lack of better term, have a kumbaya moment here, we need to find out what's workable for us as well as what is the timeline for the Tarik team to improve so that from a few months, or even few weeks from now, we can get at the point in time."

For Thomas (GW), the decision tree Greenworld had developed would help remove uncertainty when all three organizations (Tarik, Greenworld, and Starlight) evaluated Tarik's annotations. Typically, a team of experienced annotators in Tarik reviewed all annotations. This would be followed by a random sample picked and reviewed by three core members of Starlight. Finally, Greenworld would evaluate the annotated dataset reviewed by both Tarik and Starlight. In the meeting, Thomas claimed that the errors Tarik made would delay timelines. The meeting revealed the unspoken hierarchy in this collaboration—only Greenworld could develop a decision tree to standardize the quality of a good label. According to their tree, Greenworld had evaluated that 30% of labels annotators provided were erroneous, i.e. did not successfully pass through their decision tree (figure below).

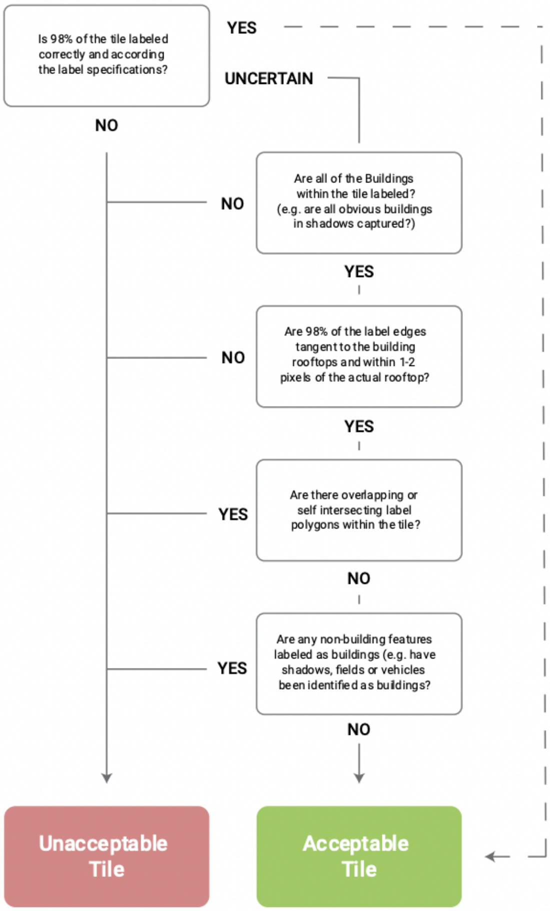
Is 98% of the tile labeled correctly and according the label specifications?
YES
UNCERTAIN
NO
NO
Are all of the Buildings within the tile labeled? (e.g. are all obvious buildings in shadows captured?)
YES
NO
Are 98% of the label edges tangent to the building rooftops and within 1-2 pixels of the actual rooftop?
YES
YES
Are there overlapping or self intersecting label polygons within the tile?
NO
YES
Are any non-building features labeled as buildings (e.g. have shadows, fields or vehicles been identified as buildings?
NO
Unacceptable Tile
Acceptable Tile

Fig.3. The decision tree evaluation criteria developed by Greenworld.

Labels were evaluated according to the following steps in the tree. The first step was to ensure that buildings were correctly classified according to the data annotation standards. If they were, the next step would be to evaluate if "all of the buildings within the chip (i.e. a cut out scene from a larger satellite image) have been labeled." The steps following this zoomed in upon smaller details of buildings and images, such as whether label edges were "within 1-2 pixels of the actual rooftop", whether "there are overlapping" labels, and whether shadows from buildings were included in a building label. Unfortunately, most annotators could not fulfill the last few requirements down the decision tree.

At this point, Starlight was befuddled. Why were there so many [label] errors? Hari, the Chief Data Scientist at Starlight, observed that they had gone through cases such as shadows. He asked if annotators in Tarik had instead committed a *systematic* error, given that Starlight had already trained annotators and was concerned that they were misinterpreting the classification standards, thus committing the same mistake repeatedly. In a later conversation, we learned from the founder of Tarik that apart from systematic errors, annotators sometimes "lose motivation" when labeling a single chip of dense buildings. The firm had agreed to a "pretty low (salary) rate" per chip, believing that the landscape would not be as dense as Bangladesh and hence, annotators could label more chips.

Annotators from Tarik also confessed that it was hard to be paid by chip, as there were many occasions where they only had one or two buildings to annotate. Greenworld quickly reassured Hari that the decision tree was not foolproof. As John, one of Greenworld's geospatial data science team members put it, a "98% accuracy is impossible to achieve, and in fact not common in machine learning". Greenworld acknowledged that error is bound to happen. But the kind of errors allowed to train an accurate machine learning model was important to know.

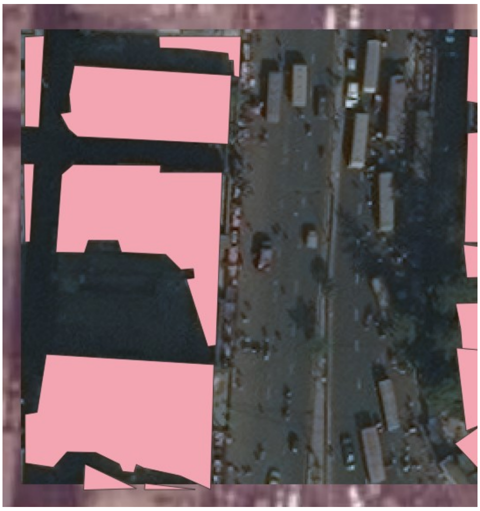

Fig. 4. Building roof blocked by shadows in the image of pink labels.

4.2.1 Error reworks relations of hierarchy.

To decide what label errors were *more* imprecise, John (GW) began to pull out labels he found difficult to accept or reject as training data. A discussion ensued. One of these labels showed a building roof blocked by shadows from surrounding buildings (Figure 4 above). He explained that the drawn line “captures a bit of the shaded edge of the building and stops arbitrarily where the real edge is, and that happens consistently.” Greenworld’s data scientist consultant, Mary, agreed, saying, “we are looking for [real] building edges to be [parallel] to the annotations of the labeler.” They scrolled up and down the image, and the Chief Data Scientist of Starlight spotted two buildings that were not labeled, making what John previously saw as uncertain a definite reject. Given that two buildings out of an estimated number of twenty were missing labels, the label was rejected according to the 98th percentile. This led to another discussion on buildings whose roofs were obstructed. Were trees blocking less than half of building roofs included in a label? How much of the tree could be included in a label?

John (GW) pulled up another image to help answer these questions. Mary, a team member of the data science team from Greenworld, asked to zoom into the image. On the top right corner of the image, a building’s edge was exposed but not annotated. The rest of the roof was completely covered in green forest, making it difficult to tell where the other end of the building ended. Thirty seconds passed. Both Starlight and Greenworld members were staring at the image, trying to infer the other end of the roof. Sean jumped in, “If we are spending too much time on it and we are looking *too long* for an error, we should accept [the chip]” even if this roof was not included.

Mary (GW) quickly added, “You can’t infer how far the building extends underneath the tree. I won’t *hallucinate* something that isn’t there”. Further cases demonstrated the importance of accepting labels that simply took too long to decide if there was a label error. It was now clear to all members in Greenworld and Starlight that the decision tree, whilst helpful in cases where multiple building roofs were unlabeled, tended to be less helpful in cases where building classification required extra inference and time. Any chances of including more guesswork from evaluators in Greenworld and Starlight such as a blocked roof was reduced to ensure that an open-source building detection model could be developed within a short period of time.

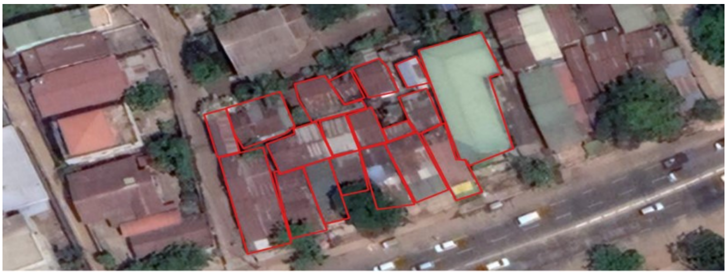

Fig. 5. An example provided by Greenworld on what a precise outlining of building roofs is.

There were cases where imprecision, whether in the form of guesswork or incorrect labels, were welcomed. These came to light when the meeting moved to another set of label errors annotators repeatedly committed: the accurate outlining of individual buildings (Figure 5 above). Because of Dhaka’s density, buildings side by side were often annotated as one building or had borders that overlapped with one another. Greenworld shared a few annotated images once more, showing cases where the lines that outlined buildings overlapped one another. At first sight, it was difficult for anyone to spot an overlap. But John (GW) zoomed right into the buildings, making the precision of the overlap clear. These overlaps, according to Mary (GW), would be fatal for the overall accuracy of the baseline model they were trying to tune to detect Bangladesh's buildings better. Already, computer code had been run to ensure that no overlaps occurred - but it couldn't detect some of these overlaps. At the same time, the labels for other buildings were fairly consistent—according to John (GW), they "were pretty good" and "captured the buildings pretty well". When comparing a single mistake against the rest of the perfectly annotated label, it was difficult for both Greenworld and Starlight to decide if an *entire* label should be rejected because of a single mistake. Thomas (GW) chimed in and said the team should accept the label, given that they have previously established that any more time spent on staring at a label "would drive them nuts".

Cory (GW), another member of the data science team under Mary (GW) added:

> "When things are ambiguous, to maintain efficiency, if we are thinking too hard we are leaning toward acceptance, if we are deliberating more than 20-30 secs, let's not overcomplicate it especially if it's uncertain again. We would just accept it."

The above quote shows how reviewers have to rely on annotators' perceptions when existing information is uncertain. Hence, potential imprecisions that may qualify as label errors were tolerated because both geospatial data analysts and data scientists in Greenworld could not judge within thirty seconds if a label is correct or not. That is, the issue of misclassification would not be solved simply by spending more time on a label. Team members had to decide in what contexts would a good enough label suffice for the issue at hand.

In cases where the edges of buildings are cut off at the borders of a satellite image (see Figure 6 below), annotators' guesswork was regarded as the main source of accuracy. Greenworld referred to these cut off images as a case of incomplete data. When annotators encountered incomplete data, they were trained by Hari from Starlight (SL) to toggle between a neighboring satellite image and the cut-off image they were labeling on their annotation platform. This way, they could tell if a building was on the other edge of a satellite image chip. Cory (GW) agreed with Hari (SL), citing that annotators were entrusted to have "studied" the images more fully and longer than reviewers. Here, we see how errors provide a site for the temporary reshuffling of roles: contracted technicians are viewed as experts above their clients as they had spent more time with the imagery.

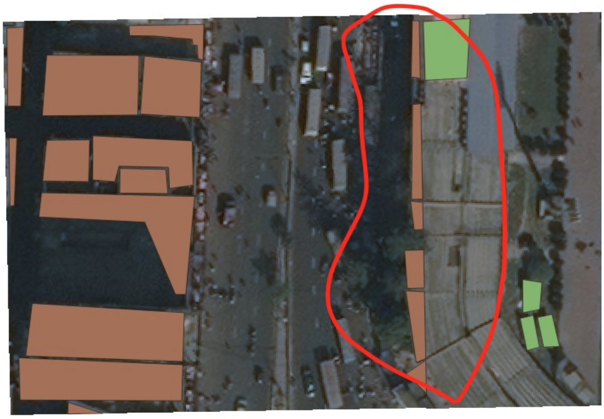

Fig. 6. An example provided by Greenworld on what incomplete information entails.

We learned later that the judgment of annotators however did not guarantee the success of a model all the time. By looking at additional information, annotators were developing training data based on materials the ML model would not have access to. This was an error of data-drift, which would potentially result in a generalization error. Cory (GW) explained shortly after a discussion on additional data, "…. We don't want to capture things that the model can't see without the [additional data] sources." A data science advisor to Greenworld from another firm further elaborated on why Cory's statement was important:

> "…Similar to a machine learning model trained on a certain dataset, the building concept is xyz and the model will try to look for the bracket of xyz in another region. These are new features you see, and it might fail but that is OK. Maybe you want to focus on these features and fine tune it. That's good enough…A model that can predict both buildings in Dhaka and another place is one that understands the concept of a building in Dhaka, not the exact building itself with all its specific features."

This quote shows how concepts of buildings can change across regions, even if they might belong to a similar group. A machine learning model doesn't need to know all the features of a building in one region. However, knowing such specificities prevents practitioners from using the model elsewhere. Between trusting the guesswork of annotators in cases of incomplete data and constraining their judgment from resulting in a generalization error, Greenworld and Starlight calibrated the extent of precision and discretion that annotators could exercise.

There are two lessons to be learned from this second case study. Conventionally, the reason for such accurate machine learning training data would be to retrain a building detection model that can generalize across regions. But another way to understand this emphasis on 98% accuracy and 2% error is to first consider how the problem of error in data annotation is neither absolute nor occurring in isolation. It must be understood in the context of work and organizational practices and priorities, where participants balance a set of competing demands. Here, priorities were shaped not only by reducing label and generalization errors, but also by the efficiency (cost minimization and productive time) of work. In a global setting where remote teams must coordinate across timezones, error redeploys old actors such as annotators in new ways by centering their labeling as a necessary prerequisite and nexus for efficient work.

This sets our cases apart from existing technical literature on ML strategies (and even HCML) for handling error, which has focused on questions of error in annotation in isolation while perhaps missing the tradeoffs, workarounds, and contradictory pressures that in fact shape real-world collaborative work settings [68]. For instance, in cases such as unclassified buildings, it was clear to reviewers that the label should be rejected. However, in more ambiguous cases, such as building roofs obscured by trees, evaluators were encouraged to accept labels that were less certain or secure in their attribution. Here, the question was no longer whether Tarik was precise in their classification. It was about ensuring that time and money was spent well, a responsibility and goal that was shared across - and as we have seen above, *negotiated between* - organizations.

Second, it also became clear as particular errors become tolerable in manual classification that it was challenging to convey to annotators which errors mattered more than others. For instance, if a missing building is considered more erroneous than outlining an overcast roof, annotators would be able to prioritize satellite chips that had more buildings that were hidden to ensure that their chip would be accepted and hence paid for. As annotators are not paid per building roof they outlined and were paid per chip instead, they often complained about annotating chips with many buildings, given that the added complexity would not factor into their overall salary.

Hence, the burden of achieving precise imprecision was most impactful on Tarik. As a result, in situations where it was difficult to decide at the outset what errors could be tolerated, existing structures of collaboration and partnerships fell into place. This impact of relying on older structures of collaboration was not borne silently by annotators. Starlight attempted to negotiate a higher per-chip rate for Tarik. They also tried to get Greenworld to pay annotators according to the time spent on annotation. While these requests were pending, Starlight filtered through a series of satellite images with less dense settings that would help Tarik annotators save labor time and earn more money.

4 DISCUSSION

In the cases above, we see how the identification and negotiation of error constitute a central aspect within the wider lives of machine learning and data science [2]. In ways that retrospective and summative ('end-of-pipeline') accounts regularly occlude, this work unfolds against a backdrop of deep uncertainty: things are rarely 'simply wrong,' even as they have come to seem so by the *end* of the process. Even if errors were anticipated—such as within the decision tree that Greenworld developed—they rarely remained the same throughout time. In real-world and applied data science work, judgments and responses to error are also profoundly *collaborative* in nature, involving the divergent interests and understandings of differently placed actors and institutions who must work together (in some fashion, whether through mutual accommodation, authority, coercion, or some complex mix of these) to arrive at a common-ish set of standards and expectations [2]. And the answers arrived at—in the two cases studied here, but we suspect in most others as well—are always situated and relative to the purposes at hand, making 'good enough' evaluations, and not abstract or context-free notions of accuracy, the central virtue of real-world data science practices.

These observations raise the stakes in taking error seriously and suggest the possibility of a richer and more generative encounter than the 'limit and eliminate' approach to error. This approach, we suggest has been the dominant way of thinking about and dealing practically with error in mainstream data science research to date. Our case studies instead suggest a potential shift in the basic imaginary of error, moving from a model of 'limit and eliminate', to one that foregrounds the *artful living with* error. What would it mean for CSCW, human-centered machine learning, and allied fields to take these propositions seriously? Can we imagine a more jazz-like (even Monk-like) relation to error in the contemporary practice and understanding of data science [56]? What difference (if any) would this difference make?

4.1 Error discloses existing structures of collaboration

First, by paying attention to breakdown and errors, we reveal existing structures of collaboration unseen or underappreciated in seemingly working systems. On the one hand, error is often used by powerful actors to make claims on who committed labeling mistakes in AI systems [68]. But at the same time, errors reveal how differently positioned experts can collectively reach agreement through different structures of collaboration – whether in more hierarchal and rigid arrangements, or more flexible and adaptive partnerships. As illustrated in the two case studies, errors are opportunities for collaborators to show how buildings break down and reveal the less-than-universal nature of man-made landscapes across regions. How these errors were disclosed and dealt with, however, depended on existing organizational cultures, norms, and hierarchies.

Greenworld, for instance, attempted to use a quality control standard or decision tree to evaluate the label errors of Tarik—in the process disclosing how they viewed themselves as an authority over their data annotators. This also reflected how they typically conducted their work to meet tight timelines and demands. Greenworld's vision of how annotation work should be performed was challenged by the intermediary Starlight, who introduced flexibility and openness in the collaboration with Tarik.

In contexts where organizational hierarchies were more rigid—such as within the Indonesian government—senior scientists could challenge Bayu, a junior earth scientist, on the universal application of LiDAR data and edge point schema to building structures in Indonesia. At the same time, by engaging them on their terms, Bayu could convince senior scientists that LiDAR datasets were able to navigate the problem of buildings being too close to one another.

We might consider how the disclosing properties of errors are a site for transparency and accountability – principles which are central to AI ethics (for critical views of AI ethics, see [18] [39] [74] [67] [85] [104]. Instead of treating errors as the inevitable result of ML practice, we conceive of errors as a practical site revealing the fallibility and fragility of ML, and exposing the collaborative structures of work that precede ML use and development. Errors are central for (listening to) and learning what has gone wrong, and provide avenues for ML practitioners and their collaborators to navigate the black box nature of ML (as well as organizations) through improvisational work [56].

Error is a property of an accountable AI ethics, given that it forces developers to exercise some responsibility for a system's breakdown, provides transparency into how error comes about, and reveals under what collaborative circumstances the system might be repaired. Instead of an artifact largely defined by technocrats [39] and compromised by corporate interests [105], an AI ethics centered on error broadens its vision to include the network of relations and collaborations necessary for rendering AI functional. Errors, whether a misclassification of LiDAR data points, or the failure of a building detection model to generalize across far-flung datasets of buildings, can provide a generative force for more purposive AI systems and worlds.

4.2 Error generates new forms and sites of collaboration

Second, errors can also serve as an organizing force and principle, bringing new forms and sites of collaboration into being. Errors can temporarily interrupt a sense of partnership and shared expectation within existing collaborations. They implode expectations, create breakdowns, and leave cracks within the seamless facade of (functioning) ML systems. But error also generates new collaborations through bringing in additional or alternative data, actors, and sites. In the first case study, for instance, Bayu and Faizah's ability to show that they could develop a machine learning method for mapping buildings rested on the assumption that data technicians were unable to accurately classify with high-resolution images. Label errors allowed Bayu and Faizah to introduce a new method of mapping that worked on LiDAR data points instead of pixels to classify buildings. Label errors allowed for the entry of new data and actors, reorganizing how geospatial data workflows were conventionally arranged for national mapping efforts.

In the second case study, we learned that the fear of label errors drove Greenworld to develop a new data quality control standard or decision tree. This decision tree brought in new actors and sites such as Starlight to evaluate the quality of Tarik's labels, bringing heterogeneity in how annotation pipelines are typically constructed between the primary client and the annotators.

This new collaboration brought in Starlight as an actor to negotiate the challenges of working across different time zones, expertise, and cultures. Evaluators from Greenworld and Starlight worked together to figure out if errors committed by annotators were necessarily bad. They coordinated their perceptions with visual inspection and decided what level of imprecision was allowable. In this way, there was some flexibility in determining how much and what kind of errors can be tolerated and reveals how errors invite change in collaborative structures.

Through their investigative and evaluative processes, Greenworld and Starlight came to realize what annotators of Tarik had long known through experience: precision was not a static entity [102]. Precision was instead remade through interactions between differently placed organizations and fields of study [93]. In particular, we can see how coercion and authority over annotators was less central than mutual accommodation between annotators and experts. This reveals how data science calls into being new forms of collaborative structures.

Noticing how error invites new actors, forms, and collaboration sites reaffirms an earlier point made by CSCW scholars Samir Passi and Steve Jackson on the rule-based nature of data science. By studying data science in action, one begins to understand how data analytics is not fully determined by abstract formulas and universal rules. Rather data science is a negotiated practice – rule-*based* rather than rule-*bound* – with limits to how practitioners develop and use data science techniques and models. A rule-based data science celebrates the work and "lived differences between theoretical reality, empirical richness and situated improvisations" rather than homogenize or flatten these potential frictions [[90]: 9]. We extend this further to show how the situated practice of dealing with error also meant bringing in new structures of collaboration. This is because AI rules, like the partnerships that sustain them, are not deterministic and absolute; its breakdown is met with flexibility and adaptation by its practitioners.

4.3 Error reworks existing collaborative relations and hierarchies

Third, error reworks or redeploys current collaborative structures, including in ways that reshape relations of hierarchy and authority, recentering and devaluing particular roles and positions. In our case studies, data science and its existing collaborative relations and hierarchies were mobilized by data scientists and domain experts to paradoxical ends. On one hand, data science can recenter old actors who have been marginalized in the field of machine intelligence. On the other hand, it can also be used to penalize those whose perceptions are viewed as prone to error. This ability to reshape and/or redeploy old structures for new ends is not necessarily because the implementation of data science displaces manual labor.

As we have shown in Case Study 1 there are existing hierarchies and power dynamics between remote sensing scientists and geospatial data technicians. Data technicians have long been regarded in the remote sensing community as performing menial and mundane mapping work in Indonesia [60]. Bayu, for example, strategically redeployed their old roles and practices to justify using ML in national mapping efforts [61]. This stands in contrast to CSCW scholarship that asserts that machine learning can replace traditional disciplines such as mapping [84].

At the same time, in Case Study 2, we show how annotators were gradually recognized for playing crucial roles in making quality ground truth data. While initially Tarik's labels were perceived as out of alignment with Greenworld's mapping standards, Greenworld later agreed that annotators had the ability to see images in ways that evaluators could not. When time is constrained, and reviewers lack supplementary data to judge what classifications can be considered accurate, annotators were entrusted with studying the images more than domain experts and data scientists. Error hence gives organizations an opportunity to redeploy old actors such as annotators within hierarchies that recenter the expertise and value of data annotation work. Error also shows how annotation work is recentered primarily in situations where timeline pressures mounted. Deferring to the judgment of data annotators only mattered when Greenworld's evaluators sensed that they were running short on time, and/or simply could not evaluate each chip and edge case with more given time. This intermittent recognition of annotators' judgment does pose the question of whether greater recognition of annotation work would be possible if timelines for such projects were looser and not pressed for time.

While Bayu's attempts to integrate data science in mapping devalued geospatial data technician labor in the first case study, the recentering of Tarik's technical expertise in the second case study shows us how the impacts of data science on relations of hierarchy and authority are not straightforward. As anthropologist Andrea Ballaestro argues, for professionals such as ML practitioners who regularly use or are familiar with "calculation grammar", technical practices are far less formulaic and "much more morally potent than they seem" [11].

Ballaestro describes how Costa Rican technocrats use a pricing algorithm to link social and technical concerns, providing a "distinctively technical place for ethics" that is nonetheless charged with the possibility of social transformation and change [8]. In the same way, we believe that errors provide domain experts and data scientists an ability to analyze what worlds data science affords more closely, and what worlds become compromised. In this way, it is the responsibility and risk of the ML practitioner and their collaborators to decide what steps should be taken in this double bind.

5 CONCLUSION

With increased attention to the harms and violence of AI systems in governance, employment, policing, and finance, there has been growing discussion both within CSCW and adjacent fields on the multiple failures and limits of AI systems, especially as they become entangled within complex fields of human and ecological practice. This has sometimes motivated a limit and eliminate approach, in which the fantasy of an error-free machine learning model seems just *that much* closer to attainable—a fantasy that centers the roles that technical ML practitioners play in this process.

Our paper has instead focused on the pragmatics and practice of error – how errors are currently managed in practice (by distributed, real-world remote actors); the complexity and skill of this work; how errors can offer new insight and discovery; and above all how errors can reveal, rework, and generate new and preexisting structures, relations, and form of collaboration in data science work (with complex and contingent effects on existing relations of expertise and authority). Shedding light on responses to error outside of "limiting and eliminating" reveals how sociotechnical errors are sites that social scientists and critical data studies and computing scholars should have a stake in further clarifying and elucidating.

We have provided two case studies of building detection models used for environmental and healthcare reasons to emphasize error's generative potential for collaboration, negotiation, and innovation. We acknowledge that label and generalization errors are not the only errors that plague machine learning practitioners and urge scholars to study how failures are being diagnosed and dealt with in ML communities. Already, critical conversations on the role of error in data science have begun in the fields of CSCW, FAACT, STS, and information science, with attention given to questions of who has the authority to determine errors [3] [4][32] [48] [84], how errors are defined according to normative standards of success and efficiency [2] and how perceived errors may hold consequences for the legitimacy of public data infrastructures [17].

We build on such work to think with the dynamic status of errors and how they can shape organizational structures and concomitant relations of expertise and authority within the environmental and geographical sciences. Our paper provides an empirical account of daily error management practices that demystify the all-encompassing powers that successful and progressive ML systems are depicted as having within popular discourse. More importantly, we wish to emphasize the artful ways that people collaborate and negotiate with error, and to illuminate how recognition for this work can lead us to more effective, more creative, more accountable, and more human-centered forms of machine learning and data science.

ACKNOWLEDGMENTS

First, we thank our fieldwork participants for their time, generosity, and meticulous knowledge sharing. Second, we are grateful for having the opportunity to present and receive feedback for parts of this paper at the University of Southern California's "Media as Sociotechnical Systems" initiative and an invited panel at the American Anthropological Association Annual Meeting 2022 entitled, "Simulations are Social Facts". And lastly, we would like to thank Padma Chirumamilla and other anonymized reviewers for their valuable feedback on earlier drafts. The Cornell Atkinson Center for Sustainability has partly supported this research.

REFERENCES

[1] Mark S. Ackerman. "The intellectual challenge of CSCW: the gap between social requirements and technical feasibility." *Human–Computer Interaction* 15, no. 2-3 (2000): 179-203.

[2] Mike Ananny. 2022. Seeing Like an Algorithmic Error: What are Algorithmic Mistakes, Why Do They Matter, How Might They Be Public Problems? In The Yale Information Society Project & Yale Journal Of Law And Technology White Paper Series. https://yjolt.org/sites/default/files/0_-_ananny_-_seeing_like_an_algorithmic_error.pdf

[3] Claudia Aradau and Tobias Blanke. 2021. Algorithmic Surveillance and the Political Life of Error. *Journal for the History of Knowledge* 2, no. 1: 10-10.

[4] Cecilia Aragon, Shion Guha, Marina Kogan, Michael Muller, and Gina Neff. Human-Centered Data Science: An Introduction. Cambridge, MA: MIT Press, 2022.

[5] Cecilia Aragon, Clayton Hutto, Andy Echenique, Brittany Fiore-Gartland, Yun Huang, Jinyoung Kim, Gina Nef, Wanli Xing, and Joseph Bayer. 2016. Developing a Research Agenda for Human-Centered Data Science. In Conference Companion Publication of the 2016 Conference on Computer Supported Cooperative Work and Social Computing. ACM Press, San Francisco, California, USA, 529–535. https: //doi.org/10.1145/2818052.2855518

[6] Atul Adya, Paramvir Bahl, Jitendra Padhye, Alec Wolman, and Lidong Zhou. 2004. A multi-radio unification protocol for IEEE 802.11 wireless networks. In Proceedings of the IEEE 1st International Conference on Broadnets Networks (BroadNets'04) . IEEE, Los Alamitos, CA, 210–217. https://doi.org/10.1109/BROADNETS.2004.8

[7] Sam Anzaroot and Andrew McCallum. 2013. UMass Citation Field Extraction Dataset. Retrieved May 27, 2019 from http://www.iesl.cs.umass.edu/data/data-umasscitationfield

[8] Seyram Avle and Silvia Lindtner. 2016. Design(ing) 'Here' and 'There': Tech Entrepreneurs, Global Markets, and Reflexivity in Design Processes. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16). Association for Computing Machinery, New York, NY, USA, 2233–2245. https://doi-org.proxy.library.cornell.edu/10.1145/2858036.2858509

[9] Gregory Bateson, Don D. Jackson, Jay Haley, and John Weakland. 1956. "Toward a theory of schizophrenia." Behavioral science 1, no. 4: 251-264

[10] Batran. 2021. A GIS Pipeline for LIDAR Point Cloud Feature Extraction. Towards Data Science. https://towardsdatascience.com/a-gis-pipeline-for-lidar-point-cloud-feature-extraction-8cd1c686468a

[11] Andrea Ballestero. 2015. The ethics of a formula: Calculating a financial–humanitarian price for water. *American Ethnologist* 42, no. 2: 262-278.

[12] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In Proceedings of the 2021 ACM Conference on

Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery, New York, NY, USA, 610–623. https://doi-org.proxy.library.cornell.edu/10.1145/3442188.3445922

[13] Ruha Benjamin. 2019. How Race and Technology 'Shape Each Other'. Emerson Today. https://today.emerson.edu/2019/10/18/ruha-benjamin-how-race-and-technology-shape-each-other/

[14] Mélanie Bernhardt, Daniel C. Castro, Ryutaro Tanno, Anton Schwaighofer, Kerem C. Tezcan, Miguel Monteiro, Shruthi Bannur et al. 2022. Active label cleaning for improved dataset quality under resource constraints. Nature communications 13, no. 1 (2022), 1-11.

[15] Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, Aäron van den Oord. 2020. Are we done with ImageNet? In Proceedings of Advances in Neural Information Processing Systems 2020. https://doi.org/10.48550/arXiv.2006.07159

[16] Dan Bouk. 2020. Error, Uncertainty, and the Shifting Ground of Census Data. *Harvard Data Science Review*, *2*(2). https://doi-org.proxy.library.cornell.edu/10.1162/99608f92.962cb309

[17] Dan Bouk and danah boyd. March 18, 2021. "Democracy's Data Infrastructure: The technopolitics of the U.S. census." Knight First Amendment Institute at Columbia University. https://knightcolumbia.org/content/democracys-data-infrastructure

[18] Solon Barocas, Andrew D Selbst, and Manish Raghavan. 2020. The hidden assumptions behind counterfactual explanations and principal reasons. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 80–89.

[19] Joy Buolamwini, Sorelle A Friedler, and Christo Wilson. [n.d.]. Gender shades: Intersectional accuracy disparities in commercial gender classification. http: //proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf. Accessed: 2022-1-12.

[20] Meredith Broussard. Forthcoming. More Than a Glitch: Confronting Race, Gender, and Ability Bias in Tech. Cambridge, MA: MIT Press.

[21] Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, and Michael Terry. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. Conference on Human Factors in Computing Systems - Proceedings: 1–14. https://doi.org/10.1145/3290605.3300234.

[22] Alexander Campolo. 2019. Steering by Sight: Data, Visualization, and the Birth of an Informational Worldview. PhD diss., New York University, 2019.

[23] Stevie Chancellor. 2022. Towards Practices for Human-Centered Machine Learning. *arXiv preprint arXiv:2203.00432* (2022).

[24] Edwin Chen. 2022. 30% of Google's Emotions Dataset is Mislabeled. Surge AI. https://www.surgehq.ai//blog/30-percent-of-googles-reddit-emotions-dataset-is-mislabeled

[25] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics* (2022) 10: 92–110.

[26] Lorraine Daston. 2005. Scientific error and the ethos of belief. *Social Research*: 1-28.

[27] Lorraine Daston. Cloud Physiognomy. Representations *135*(1), pp.45-71.

[28] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. Transactions of the Association for Computational Linguistics, 10:92–110.

[29] John Dewey. 1998 *The essential Dewey: Pragmatism, education, democracy.* Vol. 1. Bloomington, IN: Indiana University Press.

[30] John Dewey. 1986. Experience and education. In *The educational forum* (Vol. 50, No. 3, pp. 241-252). Taylor & Francis Group.

[31] John Dewey. 1938. Logic: The Theory of Inquiry. H. Holt and company, New York.

[32] Catherine D'Ignazio and Lauren F Klein. 2020. Data Feminism. MIT Press, Cambridge, MA.

[33] Anca Dumitrache, Lora Aroyo, and Chris Welty. 2015. CrowdTruth Measures for Language Ambiguity: The Case of Medical Relation Extraction. In In *Proc. of LD4IE Workshop, ISWC.* http://ceur-ws.org/Vol-1467/LD4IE2015_Dumitrache.pdf

[34] Virginia Eubanks. 2018. Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. St. Martin's Press, New York.

[35] Elena Samuylova and Emeli Dral. 2021. My data drifted. What's next?" How to handle ML model drift in production. Evidently AI. https://evidentlyai.com/blog/ml-monitoring-data-drift-how-to-handle

[36] Martin A. Fischler and Robert C. Bolles. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24, 6 (June 1981), 381–395. https://doi.org/10.1145/358669.358692

[37] Batya Friedman and Helen Nissenbaum. "Bias in computer systems." In *Computer Ethics*, pp. 215-232. Routledge, 2017.

[38] Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S. Bernstein. 2021. The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 388, 1–14. https://doi-org/10.1145/3411764.3445423

[39] Daniel Greene, Anna Lauren Hoffman, and Luke Stark. Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning. In Proceedings of the 52nd Hawaii International Conference on System Sciences, 2122-2131. https://hdl.handle.net/10125/59651

[40] Matthew Van Gundy, Davide Balzarotti, and Giovanni Vigna. 2007. Catch me, if you can: Evading network signatures with web-based polymorphic worms. In Proceedings of the first USENIX workshop on Offensive Technologies (WOOT '07) . USENIX Association, Berkley, CA, Article 7, 9 pages.

[41] James W. Demmel, Yozo Hida, William Kahan, Xiaoye S. Li, Soni Mukherjee, and Jason Riedy. 2005. Error Bounds from Extra Precise Iterative Refinement. Technical Report No. UCB/CSD-04-1344. University of California, Berkeley.

[42] Theodora Dryer. Designing Certainty: The Rise of Algorithmic Computing in an Age of Anxiety 1920-1970. University of California, San Diego, 2019.

[43] Melanie Feinberg. 2017. A design perspective on data. In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '17). ACM, New York, NY, 2952-2963. http://dx.doi.org/10.1145/3025453.3025837

[44] Clare Garvie. 2019. Garbage in, Garbage out. Face recognition on flawed data. Georgetown Law Center on Privacy & Technology (2019)

[45] Ian Hacking. 1990. The Taming of Chance. Cambridge University Press.

[46] Lara Houston, Steven J. Jackson, Daniela K. Rosner, Syed Ishtiaque Ahmed, Meg Young, and Laewoo Kang. 2016. Values in Repair. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16). Association for Computing Machinery, New York, NY, USA, 1403–1414. https://doi-org.proxy.library.cornell.edu/10.1145/2858036.2858470

[47] Jessica Hullman, Sayash Kapoor, Priyanka Nanayakkara, Andrew Gelman, and Arvind Narayanan. 2022. The worst of both worlds: A comparative analysis of errors in learning from data in psychology and machine learning. arXiv preprint arXiv:2203.06498 (2022).

[48] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery, New York, NY, USA, 560–575. https://doi-org.proxy.library.cornell.edu/10.1145/3442188.3445918

[49] Steven J. Jackson and Lara Houston. 2020. The Poetics and Political Economy of Repair. in Janet Wasko and Jeremy Schwartz, eds. Media: A Transdisciplinary Inquiry. Intellect Books, University of Chicago Press: Chicago.

[50] Steven Jackson. 2014. Rethinking Repair, in T. Gillespie, P. Boczkowski, and K. Foot, eds. Media Technologies: Essays on Communication, Materiality and Society. Cambridge, MA: MIT Press.

[51] Abigail Z Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (Virtual Event Canada). ACM, New York, NY, USA.

[52] Matthew Jones, 2018. How we became instrumentalists (again) data positivism since World War II. *Historical Studies in the Natural Sciences*, *48*(5), pp.673-684.

[53] Ju Yeon Jung, Tom Steinberger, John L. King, and Mark S. Ackerman. 2022. How Domain Experts Work with Data: Situating Data Science in the Practices and Settings of Craftwork. Proc. ACM Hum.-Comput. Interact. 6,

CSCW1, Article 58 (April 2022), 29 pages. https://doi-org/10.1145/3512905

[54] Frederike Kaltheuner, Abeba Birhane, Inioluwa Deborah Raji, Razvan Amironesei, Emily Denton, Alex Hanna, Hilary Nicole, Andrew Smart, Serena Dokuaa Oduro, James Vincent, Alexander Reben, Gemma Milne, Crofton Black, Adam Harvey, Andrew Strait, Tulsi Parida, Aparna Ashok, Fieke Jansen, Corinne Cath, and Aidan Peppin. 2021. Fake AI. Meatspace Press.

[55] Daniel Kang, Nikos Arechiga, Sudeep Pillai, Peter D. Bailis, and Matei Zaharia. 2022. Finding Label and Model Errors in Perception Data With Learned Observation Assertions. In Proceedings of the 2022 International Conference on Management of Data (SIGMOD '22). Association for Computing Machinery, New York, NY, USA, 496–505. https://doi-org.proxy.library.cornell.edu/10.1145/3514221.3517907

[56] Nathaniel Klemp, Ray McDermott, Jason Raley, Matthew Thibeault, Kimberly Powell, and Daniel J. Levitin. 2008. Plans, takes, and mis-takes. Outlines. Critical Practice Studies, 10(1), 4-21.

[57] Will Knight. March 31, 2021. The Foundations of AI are riddled with error. Wired Magazine. https://www.wired.com/story/foundations-ai-riddled-errors/#:~:text=The%20labels%20attached%20to%20images,driving%20cars%20and%20medical%20algorithms.

[58] P. M. Krafft, Meg Young, Michael Katell, Karen Huang, and Ghislain Bugingo. 2020. Defining AI in Policy versus Practice. Association for Computing Machinery, New York, NY, USA, 72–78. https://doi.org/10.1145/3375627.3375835

[59] Dongyue Li and Hongyang Zhang. 2021. Improved regularization and robustness for fine-tuning in neural networks." In 35th Conference on Neural Information Processing Systems (NeurIPS 2021): 27249-27262.

[60] Cindy Lin. 2020. How to make a forest. E-Flux. https://www.e-flux.com/architecture/at-the-border/325757/how-to-make-a-forest/

[61] Cindy Lin and Silvia Margot Lindtner. 2021. Techniques of Use: Confronting Value Systems of Productivity, Progress, and Usefulness in Computing and Design. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 595, 1–16. https://doi-org.proxy.library.cornell.edu/10.1145/3411764.3445237

[62] Adrian Mackenzie. 2017. Machine Learners: Archaeology of a data practice. Cambridge, MA: MIT Press.

[63] Donald MacKenzie. 1993. Inventing accuracy: A historical sociology of nuclear missile guidance. Cambridge, MA: MIT Press.

[64] Donald MacKenzie. 1994. Computer-related accidental death: an empirical exploration. *Science and Public Policy* 21, no. 4: 233-248.

[65] Zhiyi Ma, Kawin Ethayarajh, Tristan Thrush, Somya Jain, Ledell Wu, Robin Jia, Christopher Potts, Adina Williams, Douwe Kiela. 2021. Dynaboard: An Evaluation-As-A-Service Platform for Holistic Next-Generation Benchmarking. In Advances in Neural Information Processing Systems 34 (NeurIPS 2021). https://proceedings.neurips.cc/paper/2021/hash/55b1927fdafef39c48e5b73b5d61ea60-Abstract.html

[66] McWilliam, N., R. Teeuw, M. Whiteside, and P. Zukowskyj. 2005. Chapter 8: Image Interpretation and Processing GIS, GPS, and remote sensing. In *The Expedition Advisory Centre Royal Geographical Society 1 Kensington Gore*. https://www.rgs.org/CMSPages/GetFile.aspx?nodeguid=09c5b6e1-87f5-4ba9-9976-e03c383506ff&lang=en-GB

[67] Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. 2021.. Algorithmic impact assessments and accountability: The co-construction of impacts. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 735-746. 2021.

[68] Milagros Miceli, Julian Posada, and Tianling Yang. 2022. Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power? Proc. ACM Hum.-Comput. Interact. 6, GROUP, Article 34 (January 2022), 14 pages. https://doi-org/10.1145/3492853

[69] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q. Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). Association for Computing Machinery, New York, NY, USA, Paper 126, 1–15. https://doi-org.proxy.library.cornell.edu/10.1145/3290605.3300356

[70] Michael Muller, Melanie Feinberg, Timothy George, Steven J. Jackson, Bonnie E. John, Mary Beth Kery, and Samir Passi. 2019. Human-centered study of data science work practices. In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, Glasgow Scotland Uk, 1–8. https:

//doi.org/10.1145/3290607.3299018

[71] Michael Muller, Cecilia Aragon, Shion Guha, Marina Kogan, Gina Nef, Cathrine Seidelin, Katie Shilton, and Anissa Tanweer. 2020. Interrogating data science. In Conference Companion Publication of the 2020 Conference on Computer Supported Cooperative Work and Social Computing. ACM, Virtual Event USA, 467–473. https://doi.org/10.1145/3406865.3418584

[72] Michael Muller, Christine T. Wolf, Josh Andres, Michael Desmond, Narendra Nath Joshi, Zahra Ashktorab, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Evelyn Duesterwald, and Casey Dugan. 2021. Designing ground truth and the social life of labels. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3411764.3445402

[73] Microsoft Azure. https://docs.microsoft.com/en-us/azure/machine-learning/how-to-monitor-datasets?tabs=python

[74] Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. 2021. Assembling accountability: algorithmic impact assessment for the public interest. *Available at SSRN 3877437.*

[75] Arvind Narayanan. 2019. How to recognize AI snake oil. Arthur Miller Lecture on Science and Ethics.

[76] Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep K. Ravikumar, and Ambuj Tewari. 2013. "Learning with noisy labels. In Proceedings of Advances in Neural Information Processing Systems 26 (2013): 1-9.

[77] Gina Neff and Dawn Nafus. 2016. Self-tracking. MIT Press

[78] Curtis G. Northcutt, Lu Jiang, Issac L. Chuang. 2021. Confident Learning: Estimating Uncertainty in Dataset Labels. Journal of Artificial Intelligence Research 70 (2021): 1373-1411.

[79] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, Percy Liang. 2021. WILDS: A Benchmark of in-the-Wild Distribution Shifts. Proceedings of the 38th International Conference on Machine Learning, PMLR 139:5637-5664.

[80] Marina Kogan, Aaron Halfaker, Shion Guha, Cecilia Aragon, Michael Muller, and Stuart Geiger. 2020. Mapping out human-centered data science: Methods, approaches, and best practices. In Companion of the 2020 ACM International Conference on Supporting Group Work. ACM, Sanibel Island Florida USA, 151–156. https://doi.org/10.1145/3323994.3369898

[81] Desmond Patton, Philipp Blandfort, William Frey, Michael Gaskell, and Svebor Karaman. 2019. Annotating social media data from vulnerable populations: Evaluating disagreement between domain experts and graduate student annotators. In Proceedings of the 52nd Hawaii International Conference on System Sciences. https://hdl.handle.net/10125/59653

[82] Precarity Lab. *Technoprecarious.* Goldsmiths Press, 2020.

[83] Roberta Raileanu, Maxwell Goldstein, Denis Yarats, Ilya Kostrikov, and Rob Fergus. 2021. Automatic data augmentation for generalization in reinforcement learning. In Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021): 5402-5415.

[84] Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, Andrew D. Selbst. 2022. The Fallacy of AI Functionality. In 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Association for Computing Machinery, New York, NY, USA, 959–972. https://doi-org.proxy.library.cornell.edu/10.1145/3531146.3533158

[85] Inioluwa Deborah Raji and Jingying Yang. 2019. About ml: Annotation and benchmarking on understanding and transparency of machine learning lifecycles." *arXiv preprint arXiv:1912.06166.*

[86] Rashida Richardson, Jason Schultz, and Kate Crawford. 2019. Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice. (Feb. 2019).

[87] Daniela K. Rosner and Morgan G. Ames. 2014. "Designing for Repair? Infrastructures and Materialities of Breakdown." Proceedings of CSCW 2014, ACM Conference on Computer-Supported Cooperative Work and Social Computing. ACM Press, February 2014, 319-331.

[88] Hadi Salman, Greg Yang, Jerry Li, Pengchuan Zhang, Huan Zhang, Ilya Razenshteyn, Sébastien Bubeck. 2019. Provably Robust Deep Learning via Adversarially Trained Smoothed Classifiers. 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada: 1-12.

[89] Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. In Proceedings of the Conference on

Fairness, Accountability, and Transparency (Atlanta, GA, USA) (FAT* '19). Association for Computing Machinery, New York, NY, USA, 39–48.

[90] Samir Passi and Steven Jackson. 2017. Data Vision: Learning to See Through Algorithmic Abstraction. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17). Association for Computing Machinery, New York, NY, USA, 2436–2447. https://doi-org.proxy.library.cornell.edu/10.1145/2998181.2998331

[91] Nithya Sambasivan and Rajesh Veeraraghavan. 2022. The Deskilling of Domain Expertise in AI Development. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 587, 1–14. https://doi-org/10.1145/3491102.3517578

[92] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. Proc. ACM Hum.-Comput. Interact. 5, CSCW2, Article 317 (October 2021), 37 pages. https://doi-org.proxy.library.cornell.edu/10.1145/3476058

[93] Nick Seaver. 2021. Care and scale: decorrelative ethics in algorithmic recommendation. *Cultural Anthropology* 36, no. 3: 509-537.

[94] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, D. Sculley. 2017. No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World. In Proceedings of 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. arXiv:1711.08536v1

[95] Chirag Shah, Theresa Anderson, Loni Hagen, and Yin Zhang. 2021. An iSchool approach to data science: Human-centered, socially responsible, and context-driven. Journal of the Association for Information Science and Technology 72, 6 (2021), 793–796. https://doi.org/10.1002/asi.24444 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.24444.

[96] Manu Siddharta. 2019. Regularization Techniques in Deep Learning. https://www.kaggle.com/code/sid321axn/regularization-techniques-in-deep-learning/notebook

[97] Rebecca Slayton. 2013. Arguments that Count: Physics, Computing, and Missile Defense, 1949-2012. Cambridge, MA: MIT Press.

[98] Luke Stark and Jevan Hutson. 2022. Physiognomic Artificial Intelligence. forthcoming in Fordham Intellectual Property, Media & Entertainment Law Journal XXXII (2022). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3927300

[99] Anissa Tanweer, Cecilia R Aragon, Michael Muller, Shion Guha, Samir Passi, Gina Neff, and Marina Kogan. 2022. Interrogating Human-centered Data Science: Taking Stock of Opportunities and Limitations. In Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22). Association for Computing Machinery, New York, NY, USA, Article 99, 1–6. https://doi-org.proxy.library.cornell.edu/10.1145/3491101.3503740

[100] Angelique Taylor, Hee Rin Lee, Alyssa Kubota, and Laurel D. Riek. 2019. Coordinating Clinical Teams: Using Robots to Empower Nurses to Stop the Line. Proc. ACM Hum.-Comput. Interact. 3, CSCW, Article 221 (November 2019), 30 pages. https://doi.org/10.1145/3359323

[101] The Engine Room. 2022. AT THE CONFLUENCE OF DIGITAL RIGHTS & CLIMATE JUSTICE. https://www.theengineroom.org/new-report-at-the-confluence-of-digital-rights-climate-justice/

[102] Anna L. Tsing. (2012). On NonscalabilityThe Living World Is Not Amenable to Precision-Nested Scales. *Common knowledge*, *18*(3), 505-524.

[103] Pablo R. Velasco. 2019. Artificial Intelligibility and Proxy Error. spheres: Journal for Digital Cultures 5: 1-6.

[104] Richmond Y. Wong, Michael A. Madaio, and Nick Merrill. Seeing Like a Toolkit: How Toolkits Envision the Work of AI Ethics. *arXiv preprint arXiv:2202.08792* (2022).

[105] Meg Young, Michael Katell, and P.M. Krafft. 2022. Confronting Power and Corporate Capture at the FAccT Conference. In 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Association for Computing Machinery, New York, NY, USA, 1375–1386. https://doi-org.proxy.library.cornell.edu/10.1145/3531146.3533194

[106] Songzhu Zheng, Pengxiang Wu, Aman Goswami, Mayank Goswami, Dimitris Metaxas, Chao Chen. 2020. Error-Bounded Correction of Noisy Labels. In *International Conference on Machine Learning*, pp. 11447-11457. PMLR, 2020.

[107] Le Zhang, Ryutaro Tanno, Mou-Cheng Xu, Chen Jin, Joseph Jacob, Olga Ciccarelli, Frederik Barkhof, and Daniel C. Alexander. 2020. Disentangling Human Error from the Ground Truth in Segmentation of Medical Images. In Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada. https://proceedings.neurips.cc/paper/2020/file/b5d17ed2b502da15aa727af0d51508d6-Paper.pdf

Received July 2022; revised October 2022; accepted January 2023.