Data and its (dis)contents: A survey of dataset development and use in machine learning research

Amandalynne Paullada
Department of Linguistics
University of Washington

Inioluwa Deborah Raji
Mozilla Foundation

Emily M. Bender
Department of Linguistics
University of Washington

Emily Denton
Google Research

Alex Hanna
Google Research

Abstract

Datasets have played a foundational role in the advancement of machine learning research. They form the basis for the models we design and deploy, as well as our primary medium for benchmarking and evaluation. Furthermore, the ways in which we collect, construct and share these datasets inform the kinds of problems the field pursues and the methods explored in algorithm development. However, recent work from a breadth of perspectives has revealed the limitations of predominant practices in dataset collection and use. In this paper, we survey the many concerns raised about the way we collect and use data in machine learning and advocate that a more cautious and thorough understanding of data is necessary to address several of the practical and ethical issues of the field.

1 Introduction

The importance of datasets for machine learning research cannot be overstated. Datasets have been seen as the limiting factor for algorithmic development and scientific progress [Halevy et al., 2009, Sun et al., 2017], and a select few benchmark datasets have shaped some of the most significant developments in the field. Benchmark datasets have also played a critical role in orienting the goals, values, and research agendas of the machine learning community [Dotan and Milli, 2020].

In recent years, machine learning systems have been reported to achieve 'super-human' performance when evaluated on benchmark datasets, such as the GLUE benchmark for English textual understanding [Wang et al., 2019]. However, recent work that has surfaced the shortcomings of such datasets as meaningful tests of human-like reasoning ability reveals how this appearance of progress may rest on faulty foundations.

As the machine learning field increasingly turned to data-driven approaches, the sort of skilled and methodical human annotation applied in dataset collection practices in earlier eras was spurned as 'slow and expensive to acquire', and a turn toward unfettered collection of increasingly large amounts of data from the Web, alongside increased reliance on non-expert crowdworkers, was seen as a boon to machine learning [Halevy et al., 2009, Deng et al., 2009]. These data practices tend to abstract away the human labor, subjective judgments and biases, and contingent contexts involved in dataset production. However, these details are important for assessing whether and how a dataset might be useful for a particular application, for enabling better, more systematic error analysis, and for acknowledging the significant difficulty required in constructing useful datasets. Enormous scale has been mythologized as beneficial to generality and objectivity, but all datasets have limitations and biases [boyd and Crawford, 2012].

NeurIPS 2020 Workshop: ML Retrospectives, Surveys & Meta-analyses (ML-RSA), Virtual.

The data-driven turn in AI research — which has placed large scale datasets at the center of model development and evaluation — makes a careful, critical review of the datasets and practices of dataset creation and use crucial for at least two reasons: first, as systems trained in this way are deployed in real-world contexts that affect the lives and livelihoods of real people, it is essential that researchers, advocacy groups and the public at large understand both the contents of the datasets and how they affect system performance. Second, as the culture of the field has focused on benchmarks as the primary tool for both measuring and driving research progress [Schlangen, 2020], understanding what they are measuring (and how well) becomes increasingly urgent.

We conduct a survey of the literature of recent issues pertaining to data in machine learning research, with our primary focus on work in computer vision and natural language processing. We structure our survey around three themes. The first (§3) deals with studies which critically review the design of the datasets used as benchmarks. This includes studies which audit existing datasets for bias (§3.1), those which examine existing datasets for spurious correlations which make the benchmarks gameable (§3.2), those which critically analyze the framing of tasks (§3.3), and work promoting better data collection and documentation practices (§3.4). Next, we review approaches to improving these aspects of datasets, while still maintaining the fundamental research paradigm (§4). In looking at approaches to filtering and augmenting data and modeling techniques aimed at mitigating the impact of bias in datasets, we see further critiques of the current state of datasets. However, we find that these approaches do not fully address the broader issues with data use. Finally, we survey work on dataset practices as a whole, including critiques of their use as performance targets (§5.1), perspectives on data management (§5.2) and reuse (§5.3), and papers raising legal issues pertaining to data collection and distribution (§5.4).

2 Definitions

We follow Schlangen [2020] in distinguishing between *benchmarks*, *tasks*, *capabilities*, and *datasets*. While his work focused on natural langauge processing, we broaden these defintions to include aspects of other machine learning applications. In this context, a *task* is constituted of an input space and output space and an expected mapping between them. Schlangen notes that there are typically both *intensional* and *extensional* definitions of tasks. An intensional definition describes the relationship between input and output (e.g. the output in automatic speech recognition is a transcription of the audio signal in the input), where an extensional definition is simply the set of input-output pairs in the dataset. Thus tasks are exemplified by *datasets*, i.e. sets of input-output pairs that conform, if valid, to the intensional definition of the task. Tasks can be of interest for two (not mutually exclusive) reasons: Either they map directly into a use case (e.g. automatic transcription of audio data) or they illustrate cognitive *capabilities*, typical of humans, that we are attempting to program into machines. In the former case, a task is suitable as a *benchmark* (for comparing competing systems to each other) if the task is well-aligned with its real-world use case and the dataset is sufficiently representative of the data the systems would encounter in production. In the latter case, establishing the value of the task as a benchmark is more involved: as Schlangen argues, success on the task has to be shown to rely on having some set of capabilities that are definable outside of the task itself and transferable to other tasks.

3 Dataset design and development

> "Every data set involving people implies subjects and objects, those who collect and those who make up the collected. It is imperative to remember that on both sides we have human beings."
>
> – Mimi Ọnụọha, 'The Point of Collection'[1]

In this section, we review papers which explore issues with the contents of datasets that arise due to the manner in which they were collected, the assumptions guiding the dataset construction process and the set of questions guiding their development.

[1] Ọnụọha [2016]

3.1 Representational concerns

In recent years there has been growing concern regarding the degree and manner of representation of different sociodemographic groups within prominent machine learning datasets. For example, a glaring under-representation of darker skinned subjects has been identified within prominent facial analysis datasets [Buolamwini and Gebru, 2018] and the images in object recognition datasets have been overwhelmingly sourced from Western countries [DeVries et al., 2019]. Zhao et al. [2018] found a stark underrepresentation of female pronouns in the commonly used OntoNotes dataset for English coreference resolution. Stereotype-aligned correlations have also been identified in both computer vision and natural language processing datasets. For example, word co-occurrences in natural language processing datasets frequently reflect social biases and stereotypes relating to race, gender, (dis)ability, and more [Garg et al., 2018, Hutchinson et al., 2020] and correlations between gender and activities depicted in computer vision datasets have been shown to reflect common gender stereotypes [Zhao et al., 2017, Burns et al., 2018, van Miltenburg, 2016]. Dixon et al. [2018] found that a dataset for toxicity classification contained a disproportionate association between words describing queer identities and text labeled as 'toxic.' In an examination of the person categories within the ImageNet dataset [Deng et al., 2009], Crawford and Paglen [2019] uncovered millions of images of people that had been labelled with offensive categories, including racial slurs and derogatory phrases. In a similar vein, Prabhu and Birhane [2020] examined a broader swath of image classification datasets that were constructed using the same categorical schema as ImageNet, finding a range of harmful and problematic representations, including non-consensual and pornographic imagery of women. In response to the work of Crawford and Paglen [2019], a large portion of the ImageNet dataset has been removed [Yang et al., 2020]. Similarly, Prabhu and Birhane's [2020] examination prompted the complete removal of the TinyImages dataset [Torralba et al., 2008].

3.2 Exposing spurious cues exploited by ML models

While deep learning models have seemed to achieve remarkable performance on challenging tasks in artificial intelligence, recent work has illustrated how these performance gains may be due largely to 'cheap tricks'[2] rather than human-like reasoning capabilities. Geirhos et al. [2020] illustrate how deep neural networks rely on *shortcuts*, or decision rules that do not extrapolate well to out-of-distribution data and are often based on incidental associations, for performance. Oftentimes, these shortcuts arise due to annotation artifacts in datasets that allow models to overfit to training data and to rely on nonsensical heuristics to 'solve' the task. Recent work has revealed the presence of shortcuts in commonly used datasets that had been conceived of as proving grounds for particular competencies, such as 'reading comprehension' and other 'language understanding' capabilities. Experiments that illuminate data artifacts, or 'dataset ablations' as Heinzerling [2019] calls them, involve simple or nonsensical baselines such as training models on incomplete inputs and comparing performance to models trained on full inputs. Much recent work in NLP has revealed how these simple baselines are competitive, and that models trained on incomplete inputs for argument reasoning, natural language inference, and reading comprehension — i.e., tasks structured such that no human could do much more than randomly guess the correct output — perform quite well [Niven and Kao, 2019, Gururangan et al., 2018, Poliak et al., 2018, Kaushik and Lipton, 2018][3]. Sugawara et al. [2020] show that models trained on instances from machine reading comprehension datasets with scrambled words can still get the 'right' answer. Many of these issues result from the assumptions made in task design and in the underspecification of instructions given to human data labelers, and can thus can be addressed by rethinking the format that dataset collection takes. In light of this, recent work has proposed critical approaches to designing annotation frameworks to leverage human 'common sense' [Srivastava et al., 2020] and more critical approaches to reading comprehension dataset creation and use [Gardner et al., 2019] to pre-empt spurious correlations.

3.3 How do datasets legitimize certain problems or goals?

As the previous sections have laid out, the mapping between inputs and 'gold' labels contained in datasets is not always a meaningful one, and the ways in which tasks are structured can lead models

[2] To borrow a term from Levesque [2014]

[3] Storks et al. [2019] and Schlegel et al. [2020] provide more comprehensive reviews of datasets and dataset ablations for natural language inference.

to rely on faulty heuristics for making predictions. The problems this raises aren't limited to misleading conclusions based on benchmarking studies: When machine learning models can leverage spurious cues to make predictions well enough to beat a baseline in the test data, the resulting systems can appear to legitimize spurious tasks that do not map to real world capabilities. Jacobsen et al. [2020] point out that shortcuts in deep learning, as described in Section 3.2, make ethically dubious questions seem answerable, and advise, 'When assessing whether a task is solvable, we first need to ask: should it be solved? And if so, should it be solved by AI?' Simply because a mapping can be learned does not mean it is meaningful, and as we review in Section 3.2, this mapping can rely on spurious correlations in the data.

Decisions about what to collect in the first place and the 'problematization' that guides data collection leads to the creation of datasets that formulate pseudoscientific, often unjust tasks. For example, several papers in recent years that attempt to predict attributes such as sexuality and other fluid, subjective personal traits from photos of human faces presuppose that these predictions are possible and worthwhile to make. But these datasets, like those discussed above, enable a reliance on meaningless shortcuts. These in turn support the apparent 'learnability' of the personal traits in question: an audit by Agüera y Arcas et al. [2018] found that a model trained on the 'gaydar' data was really learning to spot stereotypical choices in grooming and self-expression, which are by no means universal, while Gelman et al. [2018] discuss how such a study strips away context and implies the existence of an "essential homosexual nature." The task rests on a pseudoscientific essentialism of human traits. Another example, from NLP, is GermEval 2020 Task 1 [Johannßen et al., 2020], which asked systems to reproduce a ranking of students by IQ scores and grades using only German short answer texts produced by the students as input. By setting up this task as feasible (for machine models or otherwise), the task organizers suggested that short answer texts contain sufficient information to 'predict' IQ scores and furthermore that IQ scores are a valid and relevant thing to measure about a person [Bender, 2020]. Not only are these task formulations problematic, but as we describe in Section 5.3, once sensitive data has been collected, it can be misused.

3.4 Collection, annotation, and documentation practices

A host of concerns regarding the practices of dataset collection, annotation, and documentation have been raised within recent years. In combination, these concerns reflect what Jo and Gebru [2020] describe as a *laissez-faire* attitude regarding dataset development: rather than collecting and curating datasets with care and intentionality — as is more typical in other data-centric disciplines — machine learning practitioners have adopted an approach where anything goes. As one data scientist put it, "if it is available to us, we ingest it" [Holstein et al., 2019].

The common practices of scraping data from internet search engines, social media platforms, and other publicly available online sources faced significant backlash in recent years. For example, Prabhu and Birhane [2020] identified the presence of millions of non-consensual pornographic imagery within prominent computer vision datasets and facial analysis datasets have received pushback due to the inclusion of personal Flickr photos without data subject's knowledge [Solon, 2019]. In many instances, the legality of the data usage has come into question, as we discuss further in Section 5.4.

Dataset annotation practices have also come under great scrutiny within recent years. Much of this has focused on how subjective values, judgments, and biases of annotators contribute to undesirable or unintended dataset bias [Ghai et al., 2020, Hube et al., 2019, van Miltenburg, 2016, Misra et al., 2016, Sap et al., 2019]. More generally, several researchers have identified a widespread failure to recognize annotation work as *interpretive work*, which in turn can result in a conflation of *gold* labels in a collected dataset and *real-world* objects, for which there may be no single ground truth label [Miceli et al., 2020, Aroyo and Welty, 2015].

Recent work by Tsipras et al. [2020] has revealed that the annotation pipeline for ImageNet does not reflect the intention of its development for the purpose of object recognition in images. They note that ImageNet, constructed with the constraint of a single label per image, had its labels largely determined by crowdworkers indicating the visual presence of that object in the image. This has led to issues with how labels are applied, particularly to images with multiple objects, where the class of interest could include a background or obscured object that would be an unsuitable result for the image classification task of that particular photo. Furthermore, the nature of image retrieval for the annotation tasks biases the crowdworkers' response to the labeling prompt, making them much less effective at filtering out unsuitable examples for a class category. This is just one of several inconsistencies and biases in the data that hints at larger annotation patterns that mischaracterize the real world tasks these datasets are meant to abstractly represent, and the broader impact of data curation design choices in determining the quality of the final dataset.

Dataset documentation practices have also been a central focus, especially as dataset development processes are increasingly being recognized as a source of algorithmic unfairness. A recent study of publications that leverage Twitter data found data decisions were heavily under-specified and inconsistent across publications [Geiger et al., 2020]. Scheuerman et al. [2020] found a widespread under-specification of annotation processes relating to gender and racial categories within facial analysis datasets. Several dataset documentation frameworks have been proposed in recent years in an effort to address these concerns, with certain frameworks looking to not just capture characteristics of the output dataset but also report details of the procedure of dataset creation [Gebru et al., 2018, Bender and Friedman, 2018, Holland et al., 2018].

The lack of rigorous and standardized dataset documentation practices has contributed to reproducibility concerns. For example, recent work undertook the laborious task of reconstructing ImageNet, following the original documented dataset construction process in an effort to test the generalization capabilities of ImageNet classifiers [Recht et al., 2019]. Despite mirroring the original collection and annotation methods — including leveraging images from the same time period — the newly constructed dataset was found to have different distributional properties. The differences were largely localized to variations in constructing ground truth labels from multiple annotations. More specifically, different thresholds for inter-annotator agreement were found to produce vastly different datasets, indicating yet again that ground truth labels in datasets do not correspond to truth.

This section centers on issues with criticisms of data themselves, and how representational issues, spurious correlations, problem legitimization, and the haphazard collection, annotation, and documentation practices are endemic to ML datasets. In the next section, we review ML methods which have been developed to deal with these issues.

4 Filtering, augmenting, and other twists on datasets

> "[Sabotage is] the impossibly small difference between exceptional failures and business as usual, connected by the fact that the very same properties and tendencies enable either outcome."
> — Evan Calder Williams, 'Manual Override'[4]

Further insight into issues with dataset contents can be found in work that attempts to address these problems, still from within the same general paradigm. In this section, we survey recent work that proposes methods for exploring and adjusting datasets toward identifying and addressing some of the issues outlined in Section 3.

The massive sizes of contemporary machine learning datasets make it intractable to thoroughly scrutinize their contents, and thus it is hard to know where to begin looking for the kinds of representational and statistical biases outlined in the previous sections. While many of the biases discovered in datasets were found by using intuition and domain expertise to construct well-designed dataset ablations and audits, recent work has also proposed tools for using statistical properties of datasets to surface spurious cues and other issues with contents. The AFLITE algorithm proposed by Sakaguchi et al. [2020] provides a way to systematically identify dataset instances that are easily gamed by a model, but in ways that are not easily detected by humans. This algorithm is applied by Le Bras et al. [2020] to a variety of natural language processing datasets, and they find that training models on adversarially filtered data leads to better generalization to out-of-distribution data. Additionally, recent work proposes methods for performing exploratory data analyses based on training dynamics that reveal edge cases in the data, bringing to light labeling errors or ambiguous labels in datasets [Swayamdipta et al., 2020]. These methods crucially rely on statistical patterns in the data to surface problem instances; it is up to human judgment to make sense of the nature of these problematic instances, whether they represent logical inconsistencies with the task at hand, cases of injustice, or both. As Denton et al. [2020] propose in the 'data genealogy' paradigm, we can qualitatively assess the design choices with respect to data sources, motivations, and methods

[4] Williams [2016]

used for constructing datasets. Prabhu and Birhane [2020] and Pipkin [2020] show that meticulous manual audits of large datasets are compelling ways to discover the most surprising and disturbing contents therein; data-driven approaches may serve primarily as a map for where to begin looking. Pipkin spent hundreds of hours watching the entirety of MIT's 'Moments in Time' video dataset [Monfort et al., 2019]. They provocatively point out through in their artistic intervention "Lacework" that the curators of massive datasets may have less intimate familiarity with the contents of these datasets than those who are paid to look at and label individual instances.

In response to a proliferation of challenging perturbations derived from existing datasets to improve generalization capabilities and lessen the ability for models to learn shortcuts, Liu et al. [2019] propose 'inoculation by fine-tuning' as a method for interpreting what model failures on perturbed inputs reveal about weaknesses of training data (or models). Recent papers also outline methodologies for leveraging human insight in the manual construction of counterfactual examples that complement instances in natural language processing datasets to promote better generalization [Gardner et al., 2020, Kaushik et al., 2020].

The case of VQA-CP [Teney et al., 2020] provides a cautionary tale of when a perturbed version of a dataset is, itself, prone to spurious cues. This complement to the original VQA dataset that consisted of instances redistributed across train and test sets was found to be easy to 'solve' with randomly generated answers. Cleverly designed sabotages that are meant to strengthen models' ability to generalize may ultimately follow the same patterns as the original data, and are thus prone to the same kinds of artifacts. This has prompted attempts to make models more robust to any kind of dataset artifact, but also suggests that there is a broader view to be taken with respect to rethinking how we construct datasets for tasks overall.

Considering that datasets will always be imperfect representations of real-world tasks, recent work proposes methods of mitigating the impacts of biases in data. Teney et al. [2020] propose an auxiliary training objective using counterfactually labeled data to guide models toward better decision boundaries. He et al. [2019] propose the DRiFT algorithm for 'unlearning' dataset bias. Sometimes, noise in datasets is not symptomatic of statistical anomalies or labeling errors, but rather, a reflection of variability in human judgment. Pavlick and Kwiatkowski [2019] find that human judgment on natural language inference tasks is variable, and that machine evaluation on this task should reflect this variability.

We emphasize that these procedural dataset modifications and bias mitigation techniques are only useful insofar as the dataset in question itself represents a valid task. In making lemonade from lemons, we must ensure the lemons are not ill-gotten or poorly formed.

5 Dataset culture

> "Data travel widely, but wherever they go, that's where data are. For even when data escape their origins, they are always encountered within other significant local settings."
> – Yanni Loukissas, *Data Are Local*

A final layer of critiques looks at the culture around dataset use in machine learning. In this section, we review papers that ask: What are issues with the broader culture of dataset use? How do our dataset usage, storage, and re-usage practices wrench data away from their contexts of creation? Lastly, what are the legal and privacy implications of these datasets?

5.1 Benchmarking practices

Benchmark datasets play a critical role in orienting the goals of machine learning communities and tracking progress within the field [Dotan and Milli, 2020, Denton et al., 2020]. Yet, the near singular focus on improving benchmark metrics has been critiqued from a variety of perspectives in recent years. Geoff Hinton has critiqued the current benchmarking culture, saying it has the potential to stunt the development of new ideas [Simonite, 2018]. Natural language processing researchers have exhibited growing concern with the singular focus on benchmark metrics, with several calls to include more comprehensive evaluations — including reports of energy consumption, model size, fairness metrics, and more — in additional to standard top-line metrics [Ethayarajh and Jurafsky, 2020, Dodge et al., 2019, Schwartz et al., 2019]. Sculley et al. [2018] examine the incentive structures that encourage singular focus on benchmark metrics — often at the expense of empirical rigor — and offer a range of suggestions including incentivizing detailed empirical evaluations, including negative results, and sharing additional experimental details. From a fairness perspective, researchers have called for the inclusion of disaggregated evaluation metrics, in addition to standard top-line metrics, when reporting and documenting model performance [Mitchell et al., 2019].

The excitement surrounding leaderboards and challenges can also give rise to a misconstrual of what high performance on a benchmark actually entails. In response to the recent onslaught publications misrepresenting the capabilities of BERT language models, Bender and Koller [2020] encourage natural language processing researchers to be attentive to the limitations of tasks and include error analysis in addition to standard performance metrics.

5.2 Data management and distribution

Secure storage and appropriate dissemination of human-derived data is a key component of data ethics [Richards and King, 2014]. To have a culture of care for the subjects of the datasets we make use of requires us to prioritize the well being of the subjects in the dataset throughout collection, development *and* distribution. In order to do so systematically, the machine learning community still has much to learn from other disciplines with respect to how they handle the data of human subjects. Unlike in the social sciences or medicine, the machine learning field has yet to develop the data management practices required to store and transmit sensitive human data.

Metcalf and Crawford [2016] go so far as to suggest the re-framing of data science as human subjects research, indicating the need for institutional review boards and informed content as researchers make decisions about other people's personal information. Particularly in consideration of an international context, where privacy concerns may be less regulated in certain regions, the potential for the data exploitation is a real threat to the safety and well being of data subjects [Mohamed et al., 2020]. As a result, those that are the most vulnerable are at risk of losing control of the way in which their own personal information is handled. Without individual control of personal information, anyone who happens to be given the opportunity to access their unprotected data to can act with little oversight, potentially against the interests or wellbeing of data subjects. This can become especially problematic and dangerous in the most sensitive contexts of personal finance information, medical data or biometrics [Birhane, 2020].

However, machine learning researchers developing such datasets rarely pay attention to this necessary consideration. Researchers will regularly distribute biometric information — for example, face image data — without so much as a distribution request form, or required privacy policy in place. Furthermore, the images are often collected without any level of informed consent or participation [Harvey and LaPlace, 2019, Solon, 2019].

Even when these datasets are flagged for removal by the creators, researchers will still attempt to make use of that now illicit information through derivative versions and backchannels. For example, Peng [2020] finds that after certain problematic face datasets were removed, hundreds of researchers continued to cite and make use of copies of this dataset months later. Without any centralized structure of data governance for the research in the field, it becomes nearly impossible to take any kind of significant action to block or otherwise prevent the active dissemination of such harmful datasets.

5.3 Use and Reuse

Several scholars have written on the importance of reusable data and code for reproducibility and replicability in machine learning [Stodden and Miguez, 2014, Stodden, 2020]. While the reuse of scientific data is often seen as an unmitigated good for scientific reproducibility [Pasquetto et al., 2017], here, we want to consider the potential pitfalls of taking data which had been collected for one purpose and using it for one in which it was not intended, particularly when this data reuse is morally and ethically objectionable to the original curators. Science and technology scholars have considered the potential incompatibilities and reconstructions needed in using data from one domain in another [Edwards, 2013]. Indeed, Strasser and Edwards discuss several major questions for big data in science and engineering, asking critically "Who owns the data?" and "Who uses the data?" [2017, p. 341-343]. Although in Section 5.4 we discuss ownership in a legal sense, ownership also suggests an inquiry into who the data have come from, such as the "literal [. . .] DNA sequences" of individuals [Strasser and Edwards, 2017, p. 342] or other biometric information. In this case, considering data reuse becomes a pivotal issue of benchmark datasets.

Instances of data reuse in benchmarks are often seen in the scraping and mining context, especially when it comes to Flickr, Wikipedia, and other openly licensed data instances. Many of the instances in which machine learning datasets drawn from these and other sources which are serious privacy violations are well-documented by Harvey and LaPlace [2019].

The reuse of data from one context to the context of machine learning is exemplified well by historian of science Joanna Radin's exploration of the peculiar history of the Pima Indians Diabetes Dataset (PIDD) and its introduction into the UCI Machine Learning Repository [Radin, 2017]. The PIDD has been used thousands of times as a "toy" classification task and currently lives in the UCI repository, a major repository for machine learning datasets. The data were collected by the National Institutes of Health from the Indigenous community living at the Gila River Indian Community Reservation, which had been extensively studied and restudied for their high prevalence of diabetes. In her history of this dataset, Radin is attentive to the politics of the creation and processing of the data itself, rather than its deployment and use. The fact that "data was used to refine algorithms that had nothing to do with diabetes or even to do with bodies, is exemplary of the history of Big Data writ large." [2017, p. 45]. Moreover, the residents of the Reservation, who refer to themselves as the Akimel O'odham, had been the subject of intense anthropological and biomedical research, especially due to a high prevalence of diabetes, which in and of itself stemmed from a history of displacement and settler-colonialism. However, their participation in research had not yielded any significant decreases in obesity or diabetes amongst community members.

Another concerning example of data reuse occurs when derivative versions of an original dataset are distributed — beyond the control of its curators — without any actionable recourse for removal. The DukeMTMC (Duke Multi-Target, Multi-Camera) dataset was collected from surveillance video footage from eight cameras on the Duke campus in 2014, used without consent of the individuals in the images and distributed openly to researchers in the US, Europe, and China. After reporting in the *Financial Times* [Murgia, 2019] and research by Harvey and LaPlace, the dataset was taken down on June 2, 2019. However, Peng [2020] has recently highlighted how the dataset and its derivatives are still freely available for download and used in scientific publications. It is nearly impossible for researchers to maintain control of datasets once they are released openly or are not closely supervised by institutional data repositories.

5.4 Legal issues

A host of legal issues have been identified pertaining to the collection of benchmark datasets. Benchmarks are often mined from the internet, collecting data instances which have various levels of licensing attached and storing them into a single repository. Different legal issues arise at each stage in the data processing pipeline, from collection to annotation, from training to evaluation, from inference and the reuse of downstream representations such as word embeddings and convolutional features [Benjamin et al., 2019]. Legal issues also arise which impact a host of different people in the process, including dataset curators, AI researchers, copyright holders, data subjects (those people whose likenesses, utterances, or representations are in the data), and consumers (those who are not in the data but are impacted by the inferences of the AI system). Different areas of law can protect (and also possibly harm) each of the different actors in turn [Khan and Hanna, 2020].

Benchmark datasets are drawn from a number of different sources, each with a different configuration of copyright holders and permissions for their use in training and evaluation in machine learning models. For instance, ImageNet was collected through several image search engines where licensing/copyright restrictions on data instances in those images are unknown [Russakovsky et al., 2015]. The ImageNet project does not host the images on their website, and therefore sidestep the copyright question by claiming that they operate like a search engine [Levendowski, 2018, ftn. 36]. PASCAL VOC was collected via the Flickr API, meaning that the images were all held through the Creative Commons license [Everingham et al., 2010]. Open licenses like Creative Commons allow for training of machine learning models under fair use doctrine [Merkely, 2019]. Faces in the Wild and Labeled Faces in the Wild were collected through Yahoo News, and via an investigation of the captions on the images we can see that the major copyright holders of those images are news wire services, including the Associated Press and Reuters [Berg et al., 2004]. Other datasets are collected in studio environment, where images were taken by dataset curators and therefore are copyright holders, which avoids potential copyright issues.

US copyright law is not well-suited to cover the range of uses of benchmark datasets, and there is not much case law establishing precedent in this area. Legal scholars have defended the use of copyrighted material in data mining and training models by suggesting that this material's usage is protected by fair use, since it entails the non-expressive use of expressive materials [Sag, 2019]. Levendowski [2018] has argued that copyright is actually a useful tool for battling algorithmic bias by offering a larger pool of works from which machine learning practitioners can draw from. She argues that, given that pre-trained representations like `word2vec` and other word embeddings suffer from gender and racial bias [Caliskan et al., 2017, Packer et al., 2018], and other public domain datasets are older or obtained through means likely to result in amplified representation of stereotypes and other biases in the data (e.g. the Enron text dataset), that using copyright data can battle biased datasets and be used their use would fall under copyright's fair use exception.

Even in cases in which all data were collected legally from a copyright perspective — such as through open licenses like Creative Commons — many downstream questions remain, including issues about privacy, informed consent, and procedures of opt-out [Merkely, 2019]. Copyright guarantees are not sufficient protections for safeguarding privacy rights of individuals, as seen in the collection of images for the Diversity in Faces and MegaFace datasets [Solon, 2019, Murgia, 2019]. Potential privacy violations arise when datasets contain biometric information which can be used to identity individuals, including faces, fingerprints, gait, and voice amongst others. However, at least in the US, there is no national-level privacy law which deals with biometric privacy. A patchwork of laws exist in Illinois, California, and Virginia which have the potential to safeguard the privacy of data subjects and consumer. However, only the Illinois Biometric Privacy law requires corporate entities to provide notice to data subjects and obtain their written consent [Khan and Hanna, 2020].

The machine learning and AI research communities have responded to the crisis by attempting to outline alternatives to licensing which make sense for research and benchmarking practices more broadly. The Montreal Data License[5] outlines different contingencies for a particular dataset, including whether the dataset will be used in commercial versus non-commercial settings, whether representations will be generated from the dataset, whether users can annotate the label or use subsets of it, and more [Benjamin et al., 2019]. This is a step forward in clarifying the different ways in which the dataset can be used once it has been collected, and therefore is a clear boon for AI researchers who create their own data instances, such as photos developed in a studio or text or captions written by crowdworkers. However, this does not deal with the larger issue of the copyright status of data instances scraped from the web, nor the privacy implications of those data instances.

In this section, we've shed light on issues around benchmarking practices, dataset use and reuse, and the legal status of benchmark datasets. These issues are more about the peculiar practices of data in machine learning culture, rather than the technical challenges associated with benchmark datasets. In this way, we want to highlight how datasets work *as* culture — that is, "not [as] singular technical objects that enter into many different cultural interactions, but… rather [as] unstable objects, culturally enacted by the practices people use to engage with them" [Seaver, 2017]. Interrogating benchmark datasets from this view requires us to expand our frame from simply technical aspects of the system, to thinking how datasets intersect with communities of practice, communities of data subjects, and legal institutions [Selbst et al., 2019].

6 Conclusion

> "Not all speed is movement."
> – Toni Cade Bambara, 'On the Issue of Roles'[6]

In this paper, we have presented a survey of issues in dataset design and development, as well as reflections on the broader culture of dataset use in machine learning. A viewpoint internal to this culture values rapid and massive progress: ever larger training datasets, used to train ever larger models, which post ever higher scores on ever harder benchmark tasks. What emerges from the papers we survey, however, is a viewpoint, largely external to that current culture of dataset use,

[5] `https://montrealdatalicense.com/`

[6] Bambara [1970]

which reveals intertwined scientific and ethical concerns appealing to a more careful and detail-oriented strategy.

Critiques of dataset design and development, especially of datasets that achieve the requisite size via opportunistic scraping of web-accessible (but not necessarily freely reusable) resources, point up various different kinds of pitfalls: First there are pitfalls of representation wherein datasets are biased both in terms of which data subjects are predominantly included and whose gaze is represented. Second, we find pitfalls of artifacts in the data, which machine learning models can easily leverage to 'game' the tasks. Third, we find evidence of whole tasks which are spurious, where success is only possible given artifacts because the tasks themselves don't correspond to reasonable real-world correlations or capabilities. Finally, we find critiques of insufficiently careful data annotation and documentation practice, which erode the foundations of any scientific inquiry based on these datasets.

Attempts to rehabilitate datasets and/or models starting from the flawed datasets themselves further reinforce the problems outlined in the critiques of dataset design and development. The development of adversarial datasets or challenge sets, while possibly removing some spurious cues, doesn't address most of the other issues with either the original datasets or the research paradigm.

Critiques of the dataset culture itself focus on the overemphasis on benchmarking to the exclusion of other evaluation practices, data management and distribution, ethical issues of data reuse, and legal issues around data use. Hyper-focus on benchmarking pushes out work that connects models more carefully to their modeling domain and approaches not optimized for the available crop of benchmarks. The papers we surveyed suggest a need for work that takes a broader view than is afforded by the one-dimensional comparison of systems typical of benchmarks. Furthermore, critiques of data management and distribution show the need for growing a culture of care for the subjects of datasets in machine learning, i.e. to keep in mind that 'data are people' and behave appropriately towards the people from whom we collect data [Raji, 2020]. Reflections of issues of data reuse emphasize the connection between data and its context, and the risks of harm (to data subjects and others) that arise when data is disconnected from its context and carried to and recontextualized in new domains. Finally, we surveyed papers exploring the legal vulnerabilities inherent to current data collection and distribution practices in ML.

What paths forward are visible from this broader viewpoint? We argue that fixes that focus narrowly on improving datasets by making them more representative or more challenging might miss the more general point raised by these critiques, and we'll be trapped in a game of dataset whack-a-mole rather than making progress, so long as notions of 'progress' are largely defined by performance on datasets. At the same time, we wish to recognize and honor the liberatory potential of datasets, when carefully designed, to make visible patterns of injustice in the world such that they may be addressed (see, for example, the work of Data for Black Lives[7]). Recent work by Register and Ko [2020] illustrates how educational interventions that guide students through the process of collecting their own personal data and running it through machine learning pipelines can equip them with skills and technical literacy toward self-advocacy — a promising lesson for the next generation of machine learning practitioners and for those impacted by machine learning systems.

In closing, we advocate for a turn in the culture towards carefully collected datasets, rooted in their original contexts, distributed only in ways that respect the intellectual property and privacy rights of data creators and data subjects, and constructed in conversation with the relevant scientific and scholarly fields required to create datasets that faithfully model tasks and tasks which target relevant and realistic capabilities. Such datasets will undoubtedly be more expensive to create, in time, money and effort, and therefore smaller than today's most celebrated benchmarks. This, in turn, will encourage work on approaches to machine learning (and to artificial intelligence beyond machine learning) that go beyond the current paradigm of techniques idolizing scale. Should this come to pass, we predict that machine learning as a field will be better positioned to understand how its technology impacts people and to design solutions that work with fidelity and equity in their deployment contexts.

[7] `https://d4bl.org/`

References

[1] Blaise Agüera y Arcas, Alexander Todorov, and Margaret Mitchell. Do algorithms reveal sexual orientation or just expose our stereotypes? *medium. com*, 2018.

[2] Lora Aroyo and Chris Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24, Mar. 2015. doi: 10.1609/aimag.v36i1.2564. URL `https://aaai.org/ojs/index.php/aimagazine/article/view/2564`.

[3] Toni Cade Bambara. On the issue of roles. *The Black Woman: An Anthology*, pages 101–10, 1970.

[4] Emily M. Bender. Is there research that shouldn't be done? is there research that shouldn't be encouraged? *medium. com*, 2020.

[5] Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. doi: 10.1162/tacl_a_00041. URL `https://www.aclweb.org/anthology/Q18-1041`.

[6] Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463. URL `https://www.aclweb.org/anthology/2020.acl-main.463`.

[7] Misha Benjamin, Paul Gagnon, Negar Rostamzadeh, Chris Pal, Yoshua Bengio, and Alex Shee. Towards Standardization of Data Licenses: The Montreal Data License. *arXiv:1903.12262 [cs, stat]*, March 2019.

[8] T.L. Berg, A.C. Berg, J. Edwards, M. Maire, R. White, Yee-Whye Teh, E. Learned-Miller, and D.A. Forsyth. Names and faces in the news. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages 848–854, Washington, DC, USA, 2004. IEEE. ISBN 978-0-7695-2158-9. doi: 10.1109/CVPR.2004.1315253.

[9] Abeba Birhane. Algorithmic colonization of africa. *SCRIPTed*, 17:389, 2020.

[10] danah boyd and Kate Crawford. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5): 662–679, 2012.

[11] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of Machine Learning Research*, volume 81, pages 77–91, New York, NY, USA, 23–24 Feb 2018. PMLR.

[12] Kaylee Burns, Lisa Anne Hendricks, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, 2018.

[13] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, April 2017. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aal4230.

[14] Kate Crawford and Trevor Paglen. *Excavating AI: The Politics of Images in Machine Learning Training Sets*, 2019. URL `https://www.excavating.ai/`.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

[16] Emily L. Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Morgan Klaus Scheuerman. Bringing the people back in: Contesting benchmark machine learning datasets. In *Proceedings of the Participatory Approaches to Machine Learning Workshop*, 2020.

[17] Terrance DeVries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 52–59. Computer Vision Foundation / IEEE, 2019. URL http://openaccess.thecvf.com/content_CVPRW_2019/html/cv4gc/de_Vries_Does_Object_Recognition

[18] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73, 2018.

[19] Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1224. URL https://www.aclweb.org/anthology/D19-1224.

[20] Mimi Ọnụọha. The Point of Collection. *Points*, 2016. URL https://points.datasociety.net/the-point-of-collection-8ee44ad7c2fa.

[21] Ravit Dotan and Smitha Milli. Value-laden disciplinary shifts in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 294, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3373157. URL https://doi.org/10.1145/3351095.3373157.

[22] Paul N. Edwards. *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Infrastructures Series. The MIT Press, Cambridge, Massachusetts London, England, first paperback edition edition, 2013. ISBN 978-0-262-51863-5 978-0-262-01392-5.

[23] Kawin Ethayarajh and Dan Jurafsky. Utility is in the eye of the user: A critique of nlp leaderboards. In *arXiv:2009.13888*, 2020.

[24] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. ISSN 0920-5691, 1573-1405. doi: 10.1007/s11263-009-0275-4.

[25] Matt Gardner, Jonathan Berant, Hannaneh Hajishirzi, Alon Talmor, and Sewon Min. On making reading comprehension more comprehensive. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 105–112, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5815. URL https://www.aclweb.org/anthology/D19-5815.

[26] Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.findings-emnlp.117.

[27] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018.

[28] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018.

[29] R. Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 325–336, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372862. URL `https://doi.org/10.1145/3351095.3372862`.

[30] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *arXiv preprint arXiv:2004.07780*, 2020.

[31] Andrew Gelman, Greggor Mattson, and Daniel Simpson. Gaydar and the fallacy of decontextualized measurement. *Sociological Science*, 5(12):270–280, 2018. ISSN 2330-6696. doi: 10.15195/v5.a12. URL `http://dx.doi.org/10.15195/v5.a12`.

[32] Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, and Klaus Mueller. Measuring social biases of crowd workers using counterfactual queries. *arXiv preprint arXiv:2004.02028*, 2020.

[33] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[34] Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, 2009.

[35] Adam Harvey and Jules LaPlace. MegaPixels: Origins and endpoints of biometric datasets "In the Wild". https://megapixels.cc, 2019.

[36] He He, Sheng Zha, and Haohan Wang. Unlearn dataset bias in natural language inference by fitting the residual. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 132–142, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-6115. URL `https://www.aclweb.org/anthology/D19-6115`.

[37] Benjamin Heinzerling. Nlp's clever hans moment has arrived. *The Gradient*, 2019.

[38] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677*, 2018.

[39] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019.

[40] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300637. URL `https://doi.org/10.1145/3290605.3300637`.

[41] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Craig Denuyl. Social biases in nlp models as barriers for persons with disabilities. In *Proceedings of ACL 2020*, 2020.

[42] Jörn-Henrik Jacobsen, Robert Geirhos, and Claudio Michaelis. Shortcuts: Neural networks love to cheat. *The Gradient*, 2020.

[43] Eun Seo Jo and Timnit Gebru. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 306–316, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372829. URL `https://doi.org/10.1145/3351095.3372829`.

[44] Dirk Johannßen, Chris Biemann, Steffen Remus, Timo Baumann, and David Sheffer. Germeval 2020 task 1 on the classification and regression of cognitive and emotional style from text: Companion paper. In *Proceedings of the 5th SwissText & 16th KONVENS Joint Conference*, volume 2624, 2020.

[45] Divyansh Kaushik and Zachary C. Lipton. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1546. URL `https://www.aclweb.org/anthology/D18-1546`.

[46] Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *ICLR*, 2020.

[47] Mehtab Khan and Alex Hanna. The legality of computer vision datasets. *Under review*, 2020.

[48] Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E Peters, Ashish Sabharwal, and Yejin Choi. Adversarial filters of dataset biases. *International Conference on Machine Learning (ICML)*, 2020.

[49] Amanda Levendowski. How copyright law can fix artificial intelligence's implicit bias problem. *Wash. L. Rev.*, 93:579, 2018.

[50] Hector J Levesque. On our best behaviour. *Artificial Intelligence*, 212:27–35, 2014.

[51] Nelson F. Liu, Roy Schwartz, and Noah A. Smith. Inoculation by fine-tuning: A method for analyzing challenge datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2171–2179, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1225. URL `https://www.aclweb.org/anthology/N19-1225`.

[52] Ryan Merkely. Use and Fair Use: Statement on shared images in facial recognition AI, March 2019.

[53] Jacob Metcalf and Kate Crawford. Where are human subjects in big data research? the emerging ethics divide. *Big Data & Society*, 3(1):2053951716650211, 2016.

[54] Milagros Miceli, Martin Schuessler, and Tianling Yang. Between subjectivity and imposition: Power dynamics in data annotation for computer vision. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2), October 2020. doi: 10.1145/3415186. URL `https://doi.org/10.1145/3415186`.

[55] Ishan Misra, C. Zitnick, Margaret Mitchell, and Ross Girshick. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[56] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287596. URL `https://doi.org/10.1145/3287560.3287596`.

[57] Shakir Mohamed, Marie-Therese Png, and William Isaac. Decolonial ai: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, pages 1–26, 2020.

[58] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–8, 2019. ISSN 0162-8828. doi: 10.1109/TPAMI.2019.2901464.

[59] Madhumita Murgia. Who's using your face? The ugly truth about facial recognition. *Financial Times*, September 2019.

[60] Timothy Niven and Hung-Yu Kao. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1459. URL `https://www.aclweb.org/anthology/P19-1459`.

[61] Ben Packer, M. Mitchell, Mario Guajardo-Céspedes, and Yoni Halpern. Text embeddings contain bias. Here's why that matters. Technical report, Google, 2018.

[62] Irene V. Pasquetto, Bernadette M. Randles, and Christine L. Borgman. On the Reuse of Scientific Data. *Data Science Journal*, 16:8, March 2017. ISSN 1683-1470. doi: 10.5334/dsj-2017-008.

[63] Ellie Pavlick and Tom Kwiatkowski. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694, 2019.

[64] Kenny Peng. Facial recognition datasets are being widely used despite being taken down due to ethical concerns. Here's how., October 2020.

[65] Kenny Peng. Facial recognition datasets are being widely used despite being taken down due to ethical concerns. here's how. *Freedom to Tinker*, 2020.

[66] Everest Pipkin. On lacework: watching an entire machine-learning dataset. *unthinking.photography*, July 2020.

[67] Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-2023. URL `https://www.aclweb.org/anthology/S18-2023`.

[68] Vinay Uday Prabhu and Abeba Birhane. Large image datasets: A pyrrhic win for computer vision? *ArXiv*, abs/2006.16923, 2020.

[69] Joanna Radin. "Digital Natives": How Medical and Indigenous Histories Matter for Big Data. *Osiris*, 32(1):43–64, September 2017. ISSN 0369-7827, 1933-8287. doi: 10.1086/693853.

[70] Inioluwa Deborah Raji. The discomfort of death counts: Mourning through the distorted lens of reported covid-19 death data. *Patterns*, 1(4):100066, 2020.

[71] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of Machine Learning Research*, volume 97, pages 5389–5400, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[72] Yim Register and Amy J Ko. Learning machine learning with personal data helps stakeholders ground advocacy arguments in model mechanics. In *Proceedings of the 2020 ACM Conference on International Computing Education Research*, pages 67–78, 2020.

[73] Neil M Richards and Jonathan H King. Big data ethics. *Wake Forest L. Rev.*, 49:393, 2014.

[74] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, December 2015. ISSN 0920-5691, 1573-1405. doi: 10.1007/s11263-015-0816-y.

[75] Matthew Sag. The new legal landscape for text mining and machine learning. *Journal of the Copyright Society of the USA*, 66, 2019.

[76] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. In *AAAI*, 2020.

[77] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1163. URL `https://www.aclweb.org/anthology/P19-1163`.

[78] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R. Brubaker. How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1), May 2020. doi: 10.1145/3392866. URL `https://doi.org/10.1145/3392866`.

[79] David Schlangen. Targeting the benchmark: On methodology in current natural language processing research. *ArXiv*, abs/2007.04792, 2020.

[80] Viktor Schlegel, Goran Nenadic, and Riza Batista-Navarro. Beyond leaderboards: A survey of methods for revealing weaknesses in natural language inference data and models. *arXiv preprint arXiv:2005.14709*, 2020.

[81] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green AI. *CoRR*, abs/1907.10597, 2019. URL `http://arxiv.org/abs/1907.10597`.

[82] D. Sculley, Jasper Snoek, Alexander B. Wiltschko, and A. Rahimi. Winner's curse? on pace, progress, and empirical rigor. In *ICLR*, 2018.

[83] Nick Seaver. Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society*, 4(2):2053951717738104, 2017. doi: 10.1177/2053951717738104.

[84] Andrew D Selbst, danah boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 59–68, 2019.

[85] Tom Simonite. *Google's AI Guru Wants Computers to Think More Like Brains*, 2018. URL `https://www.wired.com/story/googles-ai-guru-computers-think-more-like-brains/`.

[86] Olivia Solon. Facial recognition's 'dirty little secret': Millions of online photos scraped without consent. In *NBC News*, 2019.

[87] Olivia Solon. *Facial recognition's 'dirty little secret': Millions of online photos scraped without consent*, 2019. URL `https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-onl`

[88] Megha Srivastava, Tatsunori Hashimoto, and Percy Liang. Robustness to spurious correlations via human annotations. *International Conference on Machine Learning (ICML)*, 2020.

[89] Victoria Stodden. The data science life cycle: A disciplined approach to advancing data science as a science. *Communications of the ACM*, 63(7):58–66, June 2020. ISSN 0001-0782, 1557-7317. doi: 10.1145/3360646.

[90] Victoria Stodden and Sheila Miguez. Best Practices for Computational Science: Software Infrastructure and Environments for Reproducible and Extensible Research. *Journal of Open Research Software*, 2(1):e21, July 2014. ISSN 2049-9647. doi: 10.5334/jors.ay.

[91] Shane Storks, Qiaozi Gao, and Joyce Y Chai. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*, 2019.

[92] Bruno J. Strasser and Paul N. Edwards. Big Data Is the Answer . . . But What Is the Question? *Osiris*, 32(1):328–345, September 2017. ISSN 0369-7827, 1933-8287. doi: 10.1086/694223.

[93] Saku Sugawara, Pontus Stenetorp, Kentaro Inui, and Akiko Aizawa. Assessing the benchmarking capacity of machine reading comprehension datasets. In *AAAI*, pages 8918–8927, 2020.

[94] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017.

[95] Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of EMNLP*, 2020. URL https://arxiv.org/abs/2009.10795.

[96] Damien Teney, Ehsan Abbasnedjad, and Anton van den Hengel. Learning what makes a difference from counterfactual examples and gradient supervision. In *European Conference on Computer Vision*, 2020.

[97] Damien Teney, Kushal Kafle, Robik Shrestha, Ehsan Abbasnejad, Christopher Kanan, and Anton van den Hengel. On the value of out-of-distribution testing: An example of goodhart's law. *arXiv preprint arXiv:2005.09241*, 2020.

[98] Antonio Torralba, Rob Fergus, and William T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(11):1958–1970, November 2008. ISSN 0162-8828. doi: 10.1109/TPAMI.2008.128. URL https://doi.org/10.1109/TPAMI.2008.128.

[99] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. From imagenet to image classification: Contextualizing progress on benchmarks. *International Conference on Machine Learning (ICML)*, 2020.

[100] Emiel van Miltenburg. Stereotyping and bias in the flickr30k dataset. In *Proceedings of Multimodal Corpora: Computer vision and language processing (MMC 2016)*, pages 1–4, 2016.

[101] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[102] Evan Calder Williams. Manual Override. *The New Inquiry*, March 21 2016. URL https://thenewinquiry.com/manual-override/.

[103] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 547–558, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367.

[104] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1323. URL https://www.aclweb.org/anthology/D17-1323.

[105] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2003. URL https://www.aclweb.org/anthology/N18-2003.