DOI:10.1145/3458723

Documentation to facilitate communication between dataset creators and consumers.

BY TIMNIT GEBRU, JAMIE MORGENSTERN, BRIANA VECCHIONE, JENNIFER WORTMAN VAUGHAN, HANNA WALLACH, HAL DAUMÉ III, AND KATE CRAWFORD

Datasheets for Datasets

DATA PLAYS A critical role in machine learning. Every machine learning model is trained and evaluated using data, quite often in the form of static datasets. The characteristics of these datasets fundamentally influence a model's behavior: a model is unlikely to perform well in the wild if its deployment context does not match its training or evaluation datasets, or if these datasets reflect unwanted societal biases. Mismatches like this can have especially severe consequences when machine learning models are used in high-stakes domains, such as criminal justice,[1,13,24] hiring,[19] critical infrastructure,[11,21] and finance.[18] Even in other domains, mismatches may lead to loss of revenue or public relations setbacks. Of particular concern are recent examples showing that machine learning models can reproduce or amplify unwanted societal biases reflected in training datasets.[4,5,12] For these and other reasons, the World Economic Forum suggests all entities should document the provenance, creation, and use of machine learning datasets to avoid discriminatory outcomes.[25]

Although data provenance has been studied extensively in the databases community,[3,8] it is rarely discussed in the machine learning community. Documenting the creation and use of datasets has received even less attention. Despite the importance of data to machine learning, there is currently no standardized process for documenting machine learning datasets.

To address this gap, we propose *datasheets for datasets*. In the electronics industry, every component, no matter how simple or complex, is accompanied with a datasheet describing its operating characteristics, test results, recommended usage, and other information. By analogy, we propose that every dataset be accompanied with a datasheet that documents its motivation, composition, collection process, recommended uses, and so on. Datasheets for datasets have the potential to increase transparency and accountability within the machine learning community, mitigate unwanted societal biases in machine learning models, facilitate greater reproducibility of machine learning results, and help researchers and practitioners to select more appropriate datasets for their chosen tasks.

After outlining our objectives, we describe the process by which we developed datasheets for datasets. We then provide a set of questions designed to elicit the information that a datasheet for a dataset might contain, as well as a workflow for dataset creators to use when answering these questions. We conclude with a summary of the impact to date of datasheets for datasets and a discussion of implementation challenges and avenues for future work.

Objectives. Datasheets for datasets are intended to address the needs of two key stakeholder groups: dataset creators and dataset consumers. For dataset creators, the primary objective is to encourage careful reflection on the process of creating, distributing, and maintaining a dataset, including any underlying assumptions, potential risks or harms, and implica-

tions of use. For dataset consumers, the primary objective is to ensure they have the information they need to make informed decisions about using a dataset. Transparency on the part of dataset creators is necessary for dataset consumers to be sufficiently well informed that they can select appropriate datasets for their chosen tasks and avoid unintentional misuse.[a]

Beyond these two key stakeholder groups, datasheets for datasets may be valuable to policy makers, consumer advocates, investigative journalists, individuals whose data is included in datasets, and individuals who may be impacted by models trained or evaluated using datasets. They also serve a secondary objective of facilitating greater reproducibility of machine learning results: researchers and practitioners without access to a dataset may be able to use the information in its datasheet to create alternative datasets with similar characteristics.

Although we provide a set of questions designed to elicit the information a datasheet for a dataset might contain, these questions are not intended to be prescriptive. Indeed, we expect that datasheets will necessarily vary depending on factors such as the domain or existing organizational infrastructure and workflows. For example, some the questions are appropriate for academic researchers publicly releasing datasets for the purpose of enabling future research, but less relevant for product teams

» key insights

- **There are currently no industry standards for documenting machine learning datasets.**
- **Datasheets address this gap by documenting the contexts and contents of datasets: from their motivation, composition, collection process, and recommended uses.**
- **Datasheets for datasets can increase transparency and accountability within the machine learning community, mitigate unwanted biases in machine learning models, facilitate greater reproducibility of machine learning results, and help researchers and practitioners to choose the right dataset.**
- **Datasheets enable dataset creators to be intentional throughout the dataset creation process.**
- **Iterating on the design of datasheets with practitioners and legal experts helped improve the questions.**
- **Datasheets and other forms of data documentation are increasingly commonly released along with datasets.**

IMAGE BY GOODSTUDIO

a We note that in some cases, the people creating a datasheet for a dataset may not be the dataset creators, as was the case with the example datasheets that we created as part of our development process.

creating internal datasets for training proprietary models. As another example, Bender and Friedman[2] outline a proposal similar to datasheets for datasets specifically intended for language-based datasets. Their questions may be naturally integrated into a datasheet for a language-based dataset as appropriate.

We emphasize that the process of creating a datasheet is not intended to be automated. Although automated documentation processes are convenient, they run counter to our objective of encouraging dataset creators to carefully reflect on the process of creating, distributing, and maintaining a dataset.

Development Process

Here, we refined the questions and workflow provided over a period of approximately two years, incorporating many rounds of feedback.

First, leveraging our own experiences as researchers with diverse backgrounds working in different domains and institutions, we drew on our knowledge of dataset characteristics, unintentional misuse, unwanted societal biases, and other issues to produce an initial set of questions designed to elicit information about these topics. We then "tested" these questions by creating example datasheets for two widely used datasets: Labeled Faces in the Wild[16] and Pang and Lee's polarity dataset.[22] We chose these datasets in large part because their creators provided exemplary documentation, allowing us to easily find the answers to many of the questions. While creating these example datasheets, we found gaps in the questions, as well as redundancies and lack of clarity. We therefore refined the questions and distributed them to product teams in two major U.S.-based technology companies, in some cases helping teams to create datasheets for their datasets and observing where the questions did not achieve their intended objectives. Contemporaneously, we circulated an initial draft of this article to colleagues through social media and on arXiv (draft posted Mar. 23, 2018). Via these channels we received extensive comments from dozens of researchers, practitioners, and policy makers. We also worked with a team of lawyers to review the questions from a legal perspective.

We incorporated this feedback to yield the questions and workflow provided in the next section: We added and removed questions, refined the content of the questions, and reordered the questions to better match the key stages of the dataset life cycle. Based on our experiences with product teams, we reworded the questions to discourage yes/no answers, added a section on "Uses," and deleted a section on "Legal and Ethical Considerations." We found that product teams were more likely to answer questions about legal and ethical considerations if they were integrated into sections about the relevant stages of the dataset lifecycle rather than grouped together. Finally, following feedback from the team of lawyers, we removed questions that explicitly asked about compliance with regulations, and introduced factual questions intended to elicit relevant information about compliance without requiring dataset creators to make legal judgments.

Questions and Workflow

In this section, we provide a set of questions designed to elicit the information that a datasheet for a dataset might contain, as well as a workflow for dataset creators to use when answering these questions. The questions are grouped into sections that approximately match the key stages of the dataset lifecycle: motivation, composition, collection process, preprocessing/cleaning/labeling, uses, distribution, and maintenance. This grouping encourages dataset creators to reflect on the process of creating, distributing, and maintaining a dataset, and even alter this process in response to their reflection. We note that not all questions will be applicable to all datasets; those that do not apply should be skipped.

To illustrate how these questions might be answered in practice, we produced an appendix that includes an example datasheet for Pang and Lee's polarity dataset.[22] (The appendix is available online at https://dl.acm.org/doi/10.1145/3458723.) We answered some of the questions with "Unknown to the authors of the datasheet." This is because we did not create the dataset ourselves and could not find the answers to these questions in the available documentation. For an example of a datasheet that was created by the creators of the corresponding dataset, please see that of Cao and Daumé.[6,b] We note that even dataset creators may be unable to answer all the questions provided here. We recommend answering as many questions as possible rather than skipping the datasheet creation process entirely.

Motivation. The following questions are primarily intended to encourage dataset creators to clearly articulate their reasons for creating the dataset and to promote transparency about funding interests. The latter may be particularly relevant for datasets created for research purposes.

1. **For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
2. **Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?**
3. **Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.
4. **Any other comments?**

Composition. Dataset creators should read through these questions prior to any data collection and then provide answers once data collection is complete. Most of the questions here are intended to provide dataset consumers with the information they need to make informed decisions about using the dataset for their chosen tasks. Some of the questions are designed to elicit information about compliance with the EU's General Data Protection Regulation (GDPR) or comparable regulations in other jurisdictions.

Questions that apply only to datasets that relate to people are grouped together at the end of the

b See https://github.com/TristaCao/into_inclusivecoref/blob/master/GICoref/datasheet-gicoref.md.

section. We recommend taking a broad interpretation of whether a dataset relates to people. For example, any dataset containing text that was written by people relates to people.

5. **What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)?** Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

6. **How many instances are there in total (of each type, if appropriate)?**

7. **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).

8. **What data does each instance consist of?** "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.

9. **Is there a label or target associated with each instance?** If so, please provide a description.

10. **Is any information missing from individual instances?** If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.

11. **Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)?** If so, please describe how these relationships are made explicit.

12. **Are there recommended data splits (for example, training, development/validation, testing)?** If so, please provide a description of these splits, explaining the rationale behind them.

13. **Are there any errors, sources of noise, or redundancies in the dataset?** If so, please provide a description.

Datasheets for datasets have the potential to increase transparency and accountability within the ML community, mitigate unwanted societal biases in ML models, facilitate greater reproducibility of ML results, and help researchers and practitioners select more appropriate datasets for their chosen tasks.

14. **Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)?** If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

15. **Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?** If so, please provide a description.

16. **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** If so, please describe why.

If the dataset does not relate to people, you may skip the remaining questions in this section.

17. **Does the dataset identify any subpopulations (for example, by age, gender)?** If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

18. **Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset?** If so, please describe how.

19. **Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?** If so, please provide a description.

20. **Any other comments?**

Collection process. As with the questions in the previous section, dataset creators should read through these questions prior to any data collection to flag potential issues and then provide answers once collection is complete. In addition to the goals outlined earlier, the following questions are designed to elicit information that may help researchers and practitioners to create alternative datasets with similar characteristics. Again, questions that apply only to datasets that relate to people are grouped together at the end of the section.

21. **How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)?** If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

22. **What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)?** How were these mechanisms or procedures validated?

23. **If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?**

24. **Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?**

25. **Over what timeframe was the data collected?** Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

26. **Were any ethical review processes conducted (for example, by an institutional review board)?** If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

If the dataset does not relate to people, you may skip the remaining questions in this section.

27. **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?**

28. **Were the individuals in question notified about the data collection?** If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

29. **Did the individuals in question consent to the collection and use of their data?** If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

30. **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

31. **Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted?** If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

32. **Any other comments?**

Preprocessing/cleaning/labeling. Dataset creators should read through these questions prior to any preprocessing, cleaning, or labeling and then provide answers once these tasks are complete. The questions in this section are intended to provide dataset consumers with the information they need to determine whether the "raw" data has been processed in ways that are compatible with their chosen tasks. For example, text that has been converted into a "bag-of-words" is not suitable for tasks involving word order.

33. **Was any preprocessing/clean-**

The process of creating a datasheet is not intended to be automated. Although automated documentation processes are convenient, they run counter to our objective of encouraging dataset creators to carefully reflect on the process of creating, distributing, and maintaining a dataset.

ing/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.

34. **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)?** If so, please provide a link or other access point to the "raw" data.

35. **Is the software that was used to preprocess/clean/label the data available?** If so, please provide a link or other access point.

36. **Any other comments?**

Uses. The following questions are intended to encourage dataset creators to reflect on the tasks for which the dataset should and should not be used. By explicitly highlighting these tasks, dataset creators can help dataset consumers to make informed decisions, thereby avoiding potential risks or harms.

37. **Has the dataset been used for any tasks already?** If so, please provide a description.

38. **Is there a repository that links to any or all papers or systems that use the dataset?** If so, please provide a link or other access point.

39. **What (other) tasks could the dataset be used for?**

40. **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

41. **Are there tasks for which the dataset should not be used?** If so, please provide a description.

42. **Any other comments?**

Distribution. Dataset creators should provide answers to these questions prior to distributing the dataset either internally within the entity on behalf of which the dataset was created or externally to third parties.

43. **Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created?** If so, please provide a description.

44. **How will the dataset be distributed (for example, tarball on website, API, GitHub)?** Does the dataset have a digital object identifier (DOI)?

45. **When will the dataset be distributed?**

46. **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

47. **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

48. **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

49. **Any other comments?**

Maintenance. As with the previous questions, dataset creators should provide answers to these questions prior to distributing the dataset. The questions in this section are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan to dataset consumers.

50. **Who will be supporting/hosting/maintaining the dataset?**

51. **How can the owner/curator/manager of the dataset be contacted (for example, email address)?**

52. **Is there an erratum?** If so, please provide a link or other access point.

53. **Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)?** If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?

54. **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)?** If so, please describe these limits and explain how they will be enforced.

55. **Will older versions of the dataset continue to be supported/hosted/maintained?** If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.

56. **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.

57. **Any other comments?**

Impact and Challenges

Since circulating an initial draft of this article in March 2018, datasheets for datasets have already gained traction in a number of settings. Academic researchers have adopted our proposal and released datasets with accompanying datasheets.[7,10,23,26] Microsoft, Google, and IBM have begun to pilot datasheets for datasets internally within product teams. Researchers at Google published follow-up work on *model cards* that document machine learning models[20] and released a *data card* (a lightweight version of a datasheet) along with the Open Images dataset.[17] Researchers at IBM proposed *factsheets*[14] that document various characteristics of AI services, including whether the datasets used to develop the services are accompanied with datasheets. The Data Nutrition Project incorporated some of the questions provided in the previous section into the latest release of their Dataset Nutrition Label.[9] Finally, the Partnership on AI, a multi-stakeholder organization focused on studying and formulating best practices for de-

veloping and deploying AI technologies, is working on industry-wide documentation guidance that builds on datasheets for datasets, model cards, and factsheets.[c]

These initial successes have also revealed implementation challenges that may need to be addressed to support wider adoption. Chief among them is the need for dataset creators to modify the questions and workflow provided earlier based on their existing organizational infrastructure and workflows. We also note that the questions and workflow may pose problems for dynamic datasets. If a dataset changes only infrequently, we recommend accompanying updated versions with updated datasheets.

Datasheets for datasets do not provide a complete solution to mitigating unwanted societal biases or potential risks or harms. Dataset creators cannot anticipate every possible use of a dataset, and identifying unwanted societal biases often requires additional labels indicating demographic information about individuals, which may not be available to dataset creators for reasons including those individuals' data protection and privacy.[15]

When creating datasets that relate to people, and hence their accompanying datasheets, it may be necessary for dataset creators to work with experts in other domains such as anthropology, sociology, and science and technology studies. There are complex and contextual social, historical, and geographical factors that influence how best to collect data from individuals in a manner that is respectful.

Finally, creating datasheets for datasets will necessarily impose overhead on dataset creators. Although datasheets may reduce the amount of time that dataset creators spend answering one-off questions about datasets, the process of creating a datasheet will always take time, and organizational infrastructure and workflows—not to mention incentives—will need to be modified to accommodate this investment.

Despite these implementation challenges, there are many benefits to creating datasheets for datasets. In addition to facilitating better communication between dataset creators and dataset consumers, datasheets provide an opportunity for dataset creators to distinguish themselves as prioritizing transparency and accountability. Ultimately, we believe that the benefits to the machine learning community outweigh the costs.

c https://www.partnershiponai.org/about-ml/

Acknowledgments

We thank P. Bailey, E. Bender, Y. Bengio, S. Bird, S. Brown, S. Bowles, J. Buolamwini, A. Casari, E. Charran, A. Couillault, L. Dauterman, L. Dodds, M. Dudík, M. Ekstrand, N. Elhadad, M. Golebiewski, N. Gonsalves, M. Hansen, A. Hickl, M. Hoffman, S. Hoogerwerf, E. Horvitz, M. Huang, S. Kallumadi, E. Kamar, K. Kenthapadi, E. Kiciman, J. Krones, E. Learned-Miller, L. Lee, J. Leidner, R. Mauceri, B. Mcfee, E. McReynolds, B. Micu, M. Mitchell, S. Mudnal, B. O'Connor, T. Padilla, B. Pang, A. Parikh, L. Peets, A. Perina, M. Philips, B. Place, S. Rao, J. Ren, D. Van Riper, A. Roth, C. Rudin, B. Shneiderman, B. Srivastava, A. Teredesai, R. Thomas, M. Tomko, P. Tziachris, M. Whittaker, H. Wolters, A. Yeo, L. Zhang, and the attendees of the Partnership on AI's April 2019 ABOUT ML workshop for valuable feedback.

References

1. Andrews, D., Bonta, J., and Wormith, J. The recent past and near future of risk and/or need assessment. *Crime & Delinquency 52*, 1 (2006), 7–27.
2. Bender, E. and Friedman, B. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Trans. of the Assoc. for Computational Linguistics 6* (2018), 587–604.
3. Bhardwaj, A. et al. DataHub: Collaborative data science & dataset version management at scale. *CoRR* abs/1409.0798 (2014).
4. Bolukbasi, T., Chang, K., Zou, J., Saligrama, V., and Kalai, A. Man is to computer programmer as woman is to homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems* (2016).
5. Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the Conf. on Fairness, Accountability, and Transparency* (2018). 77–91.
6. Cao, Y. and Daumé, H. Toward gender-inclusive coreference resolution. In *Proceedings of the Conf. of the Assoc. for Computational Linguistics* (2020). abs/1910.13913.
7. Cao, Y. and Daumé, H. Toward gender-inclusive coreference resolution. In *Proceedings of the Conf. of the Assoc. for Computational Linguistics* (2020).
8. Cheney, J., Chiticariu, L., and Tan, W. Provenance in databases: Why, how, and where. *Foundations and Trends in Databases 1*, 4 (2009), 379–474.
9. Chmielinski, K. et al. The dataset nutrition label (2nd Gen): Leveraging context to mitigate harms in artificial intelligence. In *NeurIPS Workshop on Dataset Curation and Security*, 2020.
10. Choi, E. et al. QuAC: Question answering in context. In *Proceedings of the 2018 Conf. on Empirical Methods in Natural Language Processing*.
11. Chui, G. Project will use AI to prevent or minimize electric grid failures, 2017.
12. Dastin, J. Amazon scraps secret AI recruiting tool that showed bias against women, 2018; https://reut.rs/3imOH4d.
13. Garvie, C., Bedoya, A., and Frankle, J. *The Perpetual Line-Up: Unregulated Police Face Recognition in America*. Georgetown Law, Center on Privacy & Technology, Washington, D.C., 2016.
14. Hind, M. et al. Varshney. Increasing trust in AI services through supplier's declarations of conformity. *CoRR* abs/1808.07261 (2018).
15. Holstein, K., Vaughan, J., Daumé, H, Dudík, M., and Wallach, H. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of 2019 ACM CHI Conf. on Human Factors in Computing Systems*.
16. Huang, G., Ramesh, M., Berg, T., and Learned-Miller, E. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Technical Report 07-49. University of Massachusetts Amherst, 2007.
17. Krasin, I. et al. OpenImages: A public dataset for large-scale multi-label and multi-class image classification, 2017.
18. Lin, T. The new investor. *UCLA Law Review 60* (2012), 678.
19. Mann, G. and O'Neil, C. Hiring Algorithms Are Not Neutral, 2016; https://hbr.org/2016/12/hiring-algorithms-are-not-neutral.
20. Mitchell, M. et al. Model cards for model reporting. In *Proceedings of the Conf. on Fairness, Accountability, and Transparency* (2019). 220–229.
21. O'Connor, M. How AI Could Smarten Up Our Water System, 2017.
22. Pang, B. and Lee, L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Assoc. for Computational Linguistics*. 2004, 271.
23. Seck, I., Dahmane, K., Duthon, P., and Loosli, G. Baselines and a datasheet for the Cerema AWP dataset. *CoRR* abs/1806.04016 (2018). http://arxiv.org/abs/1806.04016
24. Doha Supply Systems. Facial Recognition, 2017.
25. World Economic Forum Global Future Council on Human Rights 2016–2018. How to Prevent Discriminatory Outcomes in Machine Learning; 2018. https://www.weforum.org/whitepapers/how-to-prevent-discriminatory-outcomes-inmachine-learning.
26. Yagcioglu, S., Erdem, A., Erdem, E., and Ikizler-Cinbis, N. RecipeQA: A challenge dataset for multimodal comprehension of cooking recipes. In *Proceedings of the 2018 Conf. on Empirical Methods in Natural Language Processing*.

Timnit Gebru is founder of DAIR Institute, Palo Alto, CA, USA.

Jamie Morgenstern is an assistant professor at the University of Washington, Seattle, WA, USA.

Briana Vecchione is a Ph.D. student at Cornell University, Ithaca, NY, USA.

Jennifer Wortman Vaughan is Senior Principal Researcher at Microsoft Research, New York, NY, USA.

Hanna Wallach is Partner Research Manager at Microsoft Research, New York, NY, USA.

Hal Daumé III is Senior Principal Researcher at Microsoft Research and a professor at the University of Maryland, College Park, MD, USA.

Kate Crawford is Senior Principal Researcher at Microsoft Research, and Research Professor at USC Annenberg, CA, USA.

Watch the authors discuss this work in the exclusive *Communications* video. https://cacm.acm.org/videos/datasheets-for-datasets