Investigating Explanations that Target Training Data

Ariful Islam Anik and Andrea Bunt

Department of Computer Science, University of Manitoba, Winnipeg, Manitoba, Canada

Abstract

To promote transparency in black-box machine learning systems, different explanation approaches have been developed and discussed in the literature. However, training dataset information is rarely communicated in these explanations despite the utmost importance of training data to a system trained with machine learning techniques. We investigated explanations that focus on communicating training dataset information to end-users in our work. In this position paper, we discuss our prototype explanations and highlight findings from our user studies. We also discuss open questions and interesting directions for future research.

Keywords 1

Explanations, Training Data, Machine Learning Systems, Transparency.

1. Introduction

While machine learning (ML) and artificial intelligence (AI) are being increasingly used in a range of automated systems, a lack of transparency in these black-box systems can be a barrier for end-users to interpret the systems' outcomes [28,32]. This lack of transparency can also negatively impact end-users' trust and acceptance of the systems [13,36].

To increase system transparency, prior work has investigated a range of explanation approaches for machine learning systems [2,7,9,14,36,37]. These explanations provide the users with information about the systems and their decisions by mostly focusing on explaining the decision factors, the criteria, and the properties of the outcomes [2,7,9,14,36,37]. While evaluations of these approaches [4,7,9,16,23,35] have shown them to be valuable, previously studied explanations rarely communicate information about training data or how the system was trained. Since machine learning algorithms look at the underlying patterns and characteristics of the training data to decide on the outcomes, training data and training procedures can have a fundamental impact on the performance of machine learning systems [8]. For example, biased training data can lead to systematic discriminations by the systems [5,6,22].

Our work focuses on designing and studying *data-centric* explanations that provide end-users with information on the data used to train the system [1]. In this position paper, we first summarize how we designed and evaluated data-centric explanations that communicate information on the training data to end-users. We also discuss interesting and important future research directions that have arisen from our work.

2. Related Work

With the goal of increasing transparency in machine learning systems, prior work has investigated a range of explanation approaches that explain the outcomes and/or a system's rationale behind the outcomes. These explanations can be categorized into different groups based on the focus of the provided information. For example, *input-influence* explanations [4,14] describe the degree of influence of the inputs to the system output. In contrast, *sensitivity-based* explanations [4,36] describe how much the value of an input has to

Joint Proceedings of the ACM IUI 2021 Workshops, April 13-17, 2021, College Station, USA
EMAIL: aianik@cs.umanitoba.ca (A. 1); bunt@cs.umanitoba.ca (A. 2)

differ to change the output. Other popular explanation approaches include *demographic-based* explanations [2,4], which describe the aggregate statistics on the outcome classes for different demographic categories (e.g., gender, race), while *case-based* explanations [4,7] use example instances from the training data to explain the outcome. Prior work also explored *white-box* explanations [9] that explain the internal workings of an algorithm, and *visual* explanations [25,39] that explain the outcomes or the model through a visual analytics interface. Most of these approaches either focus on the decision process or the factors in the decision process.

Prior work has also investigated the impact of different explanation approaches on end-users' perception of machine learning systems [4,7,9,16,23,35]. While increased transparency through explanations tends to universally increase users' acceptance of the systems [13,21,24], the impacts on trust have been mixed [9,13,23,26,30,33,34]. Prior work has also studied the impact of explanations on end-users' sense of fairness, finding that certain explanation styles impact fairness judgments more than the others [4,16].

Given that training data is fundamental to the performance of machine learning systems, Gebru et al. advocated the concept of documenting important information (e.g., motivation, creation, compositions, intended use, distribution) about datasets before releasing them, proposing a standard dataset documentation sheet for this purpose [17]. This documentation approach is receiving attention in the machine learning community [10,40] and in some organizations [3,31]. Our research focuses on investigating how such information could be communicated to *end-users* and how it might impact their perceptions of machine learning systems.

3. Data-centric Explanations

In this section, we present a high-level description of our approach to explanations that communicate the underlying training data. We also summarize our key evaluation results to date. A more detailed discussion of our work can be found in [1].

Our data-centric explanations focus on providing end-users with information on the training data used in machine learning systems. We leveraged Gebru et al.'s datasheets for datasets [17] as a starting point to design data-centric explanations, using an iterative process to transform this information into forms that were meaningful and understandable to end-users. Figure 1 provides an overview of one of our prototype data-centric explanations. Our iterative design and evaluation led us to include five different categories of training data information (Figure 1: Left). Within each category, the prototype explains dataset information using a question-and-answer format (example is given in Figure 1: A).

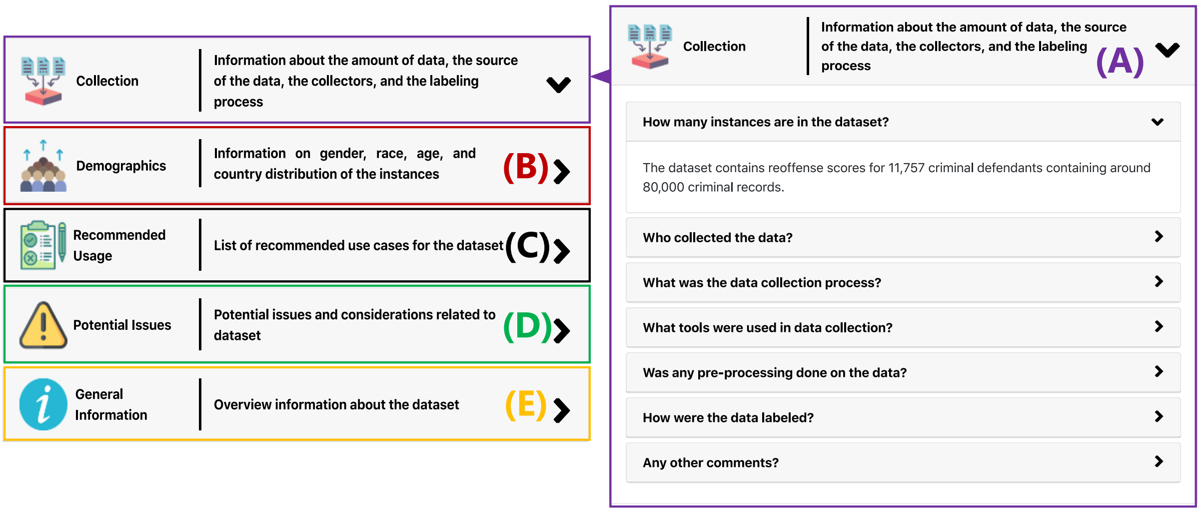

Figure 1: Overview of data-centric explanations as described in [1]. On the left, we can see the main screen with the five categories of information provided in the explanations. On the right (A), we see the expanded version of the collection category. (B), (C), (D), and (E) refer to the other categories (demographics, recommended usage, potential issues, and general information) respectively.

We evaluated our prototype explanations in a mixed-method user study with 27 participants to assess their potential to impact end-users' perceptions of machine learning systems. Our evaluation used a scenario-based approach, where we presented participants with a set of scenarios describing potential real-world systems along with the accompanying explanations. The scenarios varied in the perceived stakes of the systems (high stakes vs low stakes) and the characteristics of the training data revealed in the accompanying explanations (balanced training data vs training data with red flags). Our study also included a semi-structured interview session with each participant where we probed on issues surrounding trust, fairness, and characteristics of the system scenarios and training data.

We found in our evaluation that the data-centric explanations impacted participants' perceived level of trust in and the sense of fairness of the machine learning systems. We found that participants had more trust in the system and thought the system was fair when the explanations revealed a balanced training dataset with no errors compared to when explanations pointed out issues in the training data. Our study also provided qualitative insights into the value end-users see in having training-data information available. For example, participants liked having access to the demographics information as they felt it helped them identify biases. We also noticed initial indications of participant expertise affecting attitudes towards the explanations. Machine learning experts expected other users to have difficulty understanding explanations; however, we did not such concerns expressed by participants with less prior knowledge of machine learning. In fact, almost all participants reported that the explanations were easy to understand and expressed interest in having them available.

4. Opportunities and Challenges with Data-centric Explanations

Our initial evaluation of the data-centric explanation prototypes suggested that end-users are capable of and interested in understanding information about training datasets. Our results also point to interesting future research directions that we discuss in this section.

While our study findings suggest that participants positively receive data-centric explanations, some participants also wanted additional information about the systems and the decision factors, particularly to judge fairness. A significant body of research has investigated explanations that focus on the factors of a decision and the decision process (i.e., process-centric information) [9,14,25,36,39]. While each of the explanation approache has its own benefits, it would be interesting to explore ways to combine explanations of training data with process-centric explanations. Doing so would also allow us to investigate how end-users might prioritize the different types of explanations, as well as how the different approaches might complement each other.

We also see opportunities for the community to study and discuss different evaluation methods. For example, a common method for evaluating explanations of machine learning systems is to use fictional system scenarios (which we also used in our study with data-centric explanations) [4,19,29,38,41]. A downside of this method is that it requires participants to role-play rather than experience the systems directly, which in turn impacts the ecological validity of the study findings. There are a number of challenges with moving towards evaluations with real-life systems. For example, before we can evaluate our explanations in a real setting, we need more documented datasets available for real-world systems and we need more machine learning specialists to buy into the idea of data-centric explanations and be more open to incorporating data-centric explanations in real-life systems.

One of the goals for explanations, in general, is to ensure fairness in machine learning systems by revealing more details about the systems and their decision process. However, measuring users' perceptions of fairness is a challenging task. While a common approach is to adapt and use prior scales proposed for organizational justice [4,12,16] (which we also use in our study), these scales do not necessarily capture the fact that fairness is multi-dimensional and context-dependent [18,19]. A first necessary step in developing more robust study instruments is to develop a common definition of "fairness". There is existing work in this direction that we can build upon [11,20]. A second key evaluation challenge is having objective measures to complement the commonly collected questionnaire data (e.g., self-reported Likert scale values [4,7,9,16,19,29]). Developing such measures, particularly ones that can be feasibility collected, is an important area of future work.

Finally, we are interested in how explanations such as ours might influence the perceptions of stakeholders other than potential end-users, who are often the target pool in evaluations [4,7,16,23,35]. For example, for explanations of training data, one interesting audience could be companies and organizations that want to purchase machine learning systems to see whether data-centric explanations might impact on their purchasing decisions. Another potential audience for the data-centric explanations are journalists, who play an important role in reporting black-box systems and communicating them to the general public [15]. We know from prior work that journalists have criticized machine learning systems for their black-box nature [27].

5. Summary

Explaining the training data of machine learning systems has the potential to provide a range of benefits to end-users and other stakeholders in terms of increased transparency of the systems. Our study with data-centric explanations found some evidence that such explanations can impact people's trust in and fairness judgment of machine learning systems. We discussed some important directions for future work, which we hope will encourage discussion with researchers working on a variety of explanation styles and approaches.

6. References

[1] Ariful Islam Anik and Andrea Bunt. 2021. Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, (To appear).

[2] Liliana Ardissono, Anna Goy, Giovanna Petrone, Marino Segnan, and Pietro Torasso. 2003. Intrigue : Personalized Recommendation of Tourist Attractions. *Applied Artificial Intelligence: Special Issue on Artificial Intelligence for Cultural Heritage and Digital Libraries* 17, 8–9: 687–714.

[3] M. Arnold, D. Piorkowski, D. Reimer, J. Richards, J. Tsay, K. R. Varshney, R. K.E. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilovic, R. Nair, K. Natesan Ramamurthy, and A. Olteanu. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* 63, 4–5. https://doi.org/10.1147/JRD.2019.2942288

[4] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. "It's reducing a human being to a percentage"; perceptions of justice in algorithmic decisions. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* 2018-April: 1–14. https://doi.org/10.1145/3173574.3173951

[5] Tolga Bolukbasi, Kai Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*, 4356–4364.

[6] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (Proceedings of Machine Learning Research), 77–91. Retrieved from http://proceedings.mlr.press/v81/buolamwini18a.html

[7] Carrie J. Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. *Proceedings of the 24th International Conference on Intelligent User Interfaces*: 258–262. https://doi.org/10.1145/3301275.3302289

[8] Toon Calders and Indrė Žliobaitė. 2013. Why unbiased computational processes can lead to discriminative decision procedures. *Studies in Applied Philosophy, Epistemology and Rational Ethics* 3: 43–57. https://doi.org/10.1007/978-3-642-30487-3_3

[9] Hao Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray,

F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*: 1–12. https://doi.org/10.1145/3290605.3300789

[10] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen Tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2020. QUAC: Question answering in context. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*: 2174–2184. https://doi.org/10.18653/v1/d18-1241

[11] Alexandra Chouldechova and Aaron Roth. 2018. The Frontiers of Fairness in Machine Learning. 1–13. Retrieved from http://arxiv.org/abs/1810.08810

[12] Jason A Colquitt and Jessica B Rodell. 2015. Measuring justice and fairness. In *The Oxford handbook of justice in the workplace.* Oxford University Press, New York, NY, US, 187–202. https://doi.org/10.1093/oxfordhb/9780199981410.013.8

[13] Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten Van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. In *User Modeling and User-Adapted Interaction*, 455–496. https://doi.org/10.1007/s11257-008-9051-3

[14] Anupam Datta, Shayak Sen, and Yair Zick. 2017. Algorithmic Transparency via Quantitative Input Influence. *Transparent Data Mining for Big and Small Data*: 71–94. https://doi.org/10.1007/978-3-319-54024-5_4

[15] Nicholas Diakopoulos. 2015. Algorithmic Accountability: Journalistic investigation of computational power structures. *Digital Journalism* 3, 3: 398–415. https://doi.org/10.1080/21670811.2014.976411

[16] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K.E. Bellamy, and Casey Dugan. 2019. Explaining models: An empirical study of how explanations impact fairness judgment. *Proceedings of the 24th International Conference on Intelligent User Interfaces*: 275–285. https://doi.org/10.1145/3301275.3302310

[17] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2018. Datasheets for datasets. In *5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 1–27. Retrieved from http://arxiv.org/abs/1803.09010

[18] Ben Green and Lily Hu. 2018. The Myth in the Methodology: Towards a Recontextualization of Fairness in Machine Learning. *Proceedings of the machine learning: the debates workshop*.

[19] Nina Grgic-Hlaca, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. 2018. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. *The Web Conference 2018 - Proceedings of the World Wide Web Conference, WWW 2018*: 903–912. https://doi.org/10.1145/3178876.3186138

[20] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*: 3323–3331.

[21] J. L. Herlocker, J. A. Konstan, and J. Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 241–250. https://doi.org/10.1145/358916.358995

[22] Lauren Kirchner, Surya Mattu, Jeff Larson, and Julia Angwin. 2016. Machine Bias. *Propublica* 23: 1–26. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[23] Rene F. Kizilcec. 2016. How much information? Effects of transparency on trust in an algorithmic interface. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*: 2390–2395. https://doi.org/10.1145/2858036.2858402

[24] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will you accept an imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*: 1–14. https://doi.org/10.1145/3290605.3300641

[25] Josua Krause, Adam Perer, and Kenney Ng. 2016. Interacting with predictions: Visual inspection of black-box machine learning models. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*: 5686–5697. https://doi.org/10.1145/2858036.2858529

[26] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more? the effects of mental model soundness on personalizing an intelligent agent. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*: 1–10. https://doi.org/10.1145/2207676.2207678

[27] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2020. How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica*. Retrieved from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[28] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. 2018. Fair, Transparent, and Accountable Algorithmic Decision-making Processes. *Philosophy & Technology* 31, 4: 611–627. https://doi.org/10.1007/s13347-017-0279-x

[29] Brian Y. Lim and Anind K. Dey. 2009. Assessing demand for intelligibility in context-aware applications. *UbiComp 2009: Ubiquitous Computing*: 195. https://doi.org/10.1145/1620545.1620576

[30] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the 27th international conference on Human factors in computing systems - CHI 09*, 2119. https://doi.org/10.1145/1518701.1519023

[31] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, Figure 2: 220–229. https://doi.org/10.1145/3287560.3287596

[32] Frank Pasquale. 2015. *The Black Box Society*. Harvard University Press. https://doi.org/10.4159/harvard.9780674736061

[33] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and Measuring Model Interpretability. Retrieved from http://arxiv.org/abs/1802.07810

[34] Pearl Pu and Li Chen. 2006. Trust building with explanation interfaces. *Proceedings of the 11th International Conference on Intelligent User Interfaces* 2006: 93–100. https://doi.org/10.1145/1111449.1111475

[35] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as Mechanisms for Supporting Algorithmic Transparency. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*: 1–13. https://doi.org/10.1145/3173574.3173677

[36] Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 97–101. https://doi.org/10.18653/v1/n16-3020

[37] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus Robert Müller. 2017. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems* 28, 11: 2660–2673. https://doi.org/10.1109/TNNLS.2016.2599820

[38] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical notions vs. Human perception of fairness: A descriptive approach to fairness for machine learning. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: 2459–2468. https://doi.org/10.1145/3292500.3330664

[39] Paolo Tamagnini, Josua Krause, Aritra Dasgupta, and Enrico Bertini. 2017. Interpreting black-box classifiers using instance-level visual explanations. *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics, HILDA 2017*: 1–6. https://doi.org/10.1145/3077257.3077260

[40] Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2020. RecipeQA: A challenge dataset for multimodal comprehension of cooking recipes. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*: 1358–1368. https://doi.org/10.18653/v1/d18-1166

[41] Yunfeng Zhang, Q. Vera Liao, and Rachel K.E. Bellamy. 2020. Efect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*: 295–305. https://doi.org/10.1145/3351095.3372852