Navigating Data Errors in Machine Learning Pipelines: Identify, Debug, and Learn

Bojan Karlaš
Harvard University
bkarlas@mgh.harvard.edu

Babak Salimi
University of California, San Diego
bsalimi@ucsd.edu

Sebastian Schelter
BIFOLD & TU Berlin
schelter@tu-berlin.de

***Abstract*—Addressing data errors such as wrong, missing, noisy, biased, or out-of-distribution values has become a crucial part of the machine learning (ML) development lifecycle. Unfortunately, traditional approaches have relied either on treating the symptom by refining the model architecture, or on improving data quality by repairing incorrect values regardless of their significance to the downstream model. Both strategies end up tackling the problem in isolation, and disregard the structure of modern ML pipelines, which involve a series of steps for data preprocessing, model training, and prediction processing. Consequently, they miss the opportunity to consider how different data errors propagate through the pipeline and how they impact its ability to perform downstream tasks. In recent years, the research community has made significant strides towards more holistic approaches for identifying the most harmful data errors, performing the most beneficial repairs, and ensuring reliable performance even if some data errors remain present. This tutorial will survey some prominent work published in this space and showcase several tools that have been developed. By combining theoretical foundations with practical demonstrations, attendees will gain actionable strategies to diagnose and mitigate data quality issues, improving the reliability, fairness, and transparency of ML systems in real-world settings.**

I. Introduction

ML systems are increasingly deployed in high-stakes domains such as healthcare, finance, and law enforcement, where their decisions profoundly impact individuals and communities. To ensure trust in these systems, it is essential to focus on their accuracy, fairness, robustness, and reliability [9], [25], [46], [73]. Developing trustworthy ML systems involves navigating a complex multi-stage pipeline—spanning data preparation, model training, predictive queries, and evaluation—where failures at any stage can lead to significant performance degradation. A key observation is that many such failures are directly caused by data errors, including missing, incorrect, invalid, biased, and out-of-distribution values. These errors propagate through ML pipelines, compromising outcomes even when models and algorithms are otherwise well-refined. Debugging and mitigating these errors consumes considerable time and effort for developers, highlighting the need to systematically address data quality issues.

Existing methods for addressing bias in ML models primarily focus on algorithm-specific solutions, which often treat the symptoms of poor data quality rather than tackling its root causes [4], [47]. Traditional data cleaning techniques [34], [37], [38], [44], [64] produce a single "best-guess" version of cleaned data but offer no guarantees of unbiasedness or representativeness. Similarly, explainability methods, while valuable for interpreting model predictions, often analyze models in isolation and overlook the broader ML pipeline, where errors from upstream stages can propagate and amplify. In real-world ML applications, training data is typically derived from multiple heterogeneous source datasets through complex ML pipelines [5], [80]. These pipelines integrate, transform, and encode raw data into features, exacerbating data quality challenges such as inconsistencies, noise, and bias. To overcome these limitations, a holistic, data-centric approach that scrutinizes and refines the entire ML pipeline is necessary to ensure robust, fair, and reliable systems in deployment.

Addressing these challenges requires a shift in perspective: understanding ML pipelines as interconnected workflows where data quality issues must be tackled holistically across all stages. This tutorial focuses on recent work that highlights the importance of quantifying task-specific data contributions, such as Shapley values for identifying problematic data points [19], [29], reasoning about end-to-end pipelines to trace error propagation [21], [22], [68], and providing quality guarantees to enable robust learning under uncertainty, incomplete data, and distributional inconsistencies [51], [58], [89]. These techniques collectively provide the foundation for understanding how data errors propagate through ML pipelines and how targeted interventions can mitigate their downstream impact.

This tutorial equips participants with actionable tools to identify, debug, and reason about data quality challenges, while enhancing accuracy, fairness, reliability, and robustness in ML systems. Structured into two 90-minute parts, the first session will introduce methods for identifying data errors, debugging ML pipelines, and learning from imperfect data. The second hands-on session will allow attendees to apply practical tools to real-world tasks, such as prioritizing problematic data points, tracing error propagation, and implementing robust learning strategies. By combining these perspectives, participants will gain a holistic understanding of how to address data quality issues, enabling them to build robust, transparent, and trustworthy ML systems for real-world challenges.

II. Outline of the Survey Part

In the first part of the tutorial, we will present a survey of relevant works covering the notion of data importance as a framework for identifying data errors, the application of

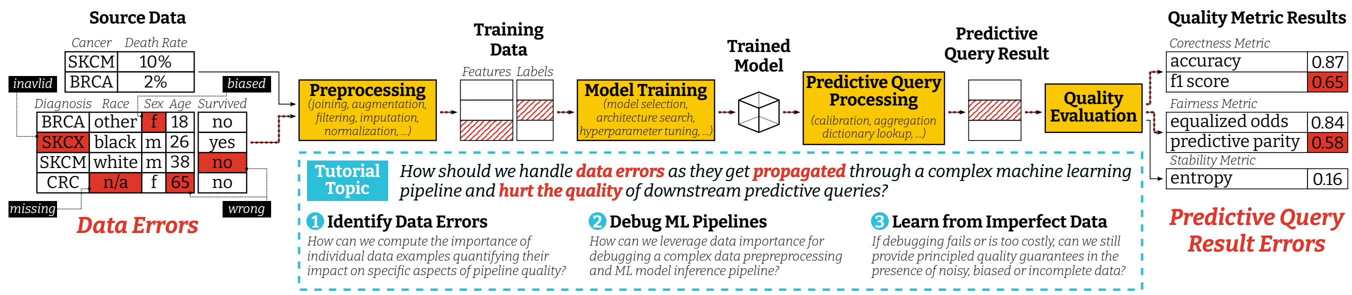

Fig. 1: Data errors are often the root cause of many quality issues in modern ML pipelines. Handling such errors as they traverse complex pipeline steps is a key challenge for practitioners. This tutorial covers some significant recent developments, presents novel tools, and explores opportunities for future work in this space.

these methods for debugging end-to-end ML pipelines, and the approaches for providing quality guarantees despite the presence of data errors.

A. Data Importance for Data Error Detection

Repairing data is often a very costly process requiring a lot of human effort. Therefore, identifying data with the most significant negative impact on the downstream ML model could allow practitioners to prioritize their efforts. The general strategy pursued by recent approaches is to define some method of measuring the importance of individual units of data with respect to their impact on the downstream ML model.

Quantifying Data Importance. We will start by covering a simple way to measure data importance using the *leave-one-out (LOO)* score. We will cover several generalizations of this approach including the Shapley value [19], [29], Banzhaf value [76], Beta Shapley [39], and others. We will also cover gradient-based methods [36], [37], as well as some uncertainty-based methods [55], [59]. Finally, we will cover methods geared toward specific aspects of model quality such as fairness [62], as well as methods specialized for large language models (LLMs) [43].

Takeaway: Attendees will get acquainted with the notion of data importance in the context of data debugging, as well as various methods for quantifying it. They will also develop a sense of the strengths and weaknesses of various methods, allowing them to reason about selecting the best method for their own data debugging scenario.

Overcoming Computational Challenges. Despite their effectiveness, many methods for quantifying data importance come with enormous computational costs, limiting their applicability in real-world data debugging tasks [30]. For example, the Shapley value involves a sum over exponentially many subsets, making it intractable in practical settings. In this part, we will explore approaches for speeding up the computation, including Monte Carlo methods [19], using the K-nearest neighbors as a proxy model [28], and model-based estimation [13].

Takeaway: This part will introduce the attendees to certain algorithmic approaches for making data importance computationally efficient by leveraging various tools well-known to the data management community. We hope that our overview will inspire some of the attendees to make future contributions.

B. Data Debugging in ML Pipelines

The discussed techniques are designed for a static, preprocessed training dataset of a given model. However, this assumption ignores the circumstances in real-world ML applications, where models are trained on data that is preprocessed as part of an *ML pipeline* [5], [33], [40], [56], [63], [69], [71], [79]–[81]. Such a pipeline typically accesses several heterogeneous input datasets, integrates them and encodes them into features to produce the actual training data for the model (Figure 1). This raises challenges for debugging – data errors should be identified in the *source data* of a pipeline, while existing debugging methods are designed for already preprocessed *training data* (the output of the preprocessing step). The second part of our survey will bring these ML pipelines into play.

Libraries and Systems for ML pipelines. We will start by giving an overview of the implementation of ML pipelines, which typically combine several systems and libraries. Examples from the open source space include combinations of Pandas [77] with scikit-learn [57], [70], Spark [82] with SparkML [48], Tensorflow Transform [5] with Apache Beam [2], and systems such as Apache SystemDS [8], MLflow Recipes [14] from Databricks or Metaflow [49] from Netflix. Furthermore, we will cover proprietary pipeline abstractions in commercial cloud services such as Amazon SageMaker [3], Microsoft Azure ML [52], or Google's Vertex AI [20].

Takeaway: Attendees will learn about shared design patterns and abstractions across various libraries, as well as their shortcomings which make debugging more difficult.

Characteristics of Real-World ML Pipelines. Next, we will summarize two large-scale empirical studies on the characteristics of real-world ML pipelines encountered in large companies, code repositories and cloud platforms. In particular, we will focus on a study of thousands of production ML pipelines at Google [80], and on insights from the analysis of millions of GitHub notebooks and ML.Net pipelines from Microsoft [63].

Takeaway: Attendees will learn about detailed usage statistics for common libraries and operators in these pipelines, and will be confronted with findings that that contradict conventional wisdom, e.g., that data ingestion, data preprocessing and model analysis account for higher compute costs than model training, or that a large proportion of pipelines train traditional non-neural ML models.

Inspecting and Debugging Data in ML pipelines. Finally, we will discuss techniques to inspect pipelines and debug their input and output data. We will start by reviewing techniques for the basic analysis of ML pipelines [22], [53], [60]. Subsequently, we will dive into work on provenance-based data debugging of ML pipelines and discuss approaches which leverage fine-grained provenance information [24] to reason about the input and output data of a pipeline. Examples include a continuous integration system to screen pipelines for issues like data leakage and label errors [68], techniques to efficiently compute data importance over pipelines of various shapes [33], or the identification of training data points whose removal would fix user complaints in prediction queries [18], [79].

Takeaway: Attendees will learn details about different pipeline representations, the efficient computation of fine-grained data provenance for a pipeline, and how this enables the adaptation of existing data debugging techniques to the pipeline's source data. We will furthermore highlight the connection to related areas such as incremental view maintenance.

C. Learning from Uncertain and Incomplete Data

Uncertain and incomplete data, arising from data errors, missing values, and biases, are pervasive in real-world ML applications. These imperfections distort the underlying data distribution, compromising the fairness, accuracy, and generalizability of ML models. Traditional approaches, such as fairness-aware learning, data cleaning techniques [10], [32], [44], [64], [83], [84], and methods addressing selection bias [12], [15], [26], [41], [65] or labeling errors [31], [85], [87], often fail to recover a representative dataset, limiting their effectiveness in practice. Furthermore, robust model learning methods aim to ensure resilience against adversarial perturbations [27], [66], [72], [86] and distributional shifts [6], [54], but rely on restrictive assumptions that rarely hold. These challenges are exacerbated in complex ML pipelines, where data imperfections propagate and interact, amplifying performance degradation.

Quantifying and Handling Incomplete Data. Recent advancements have shifted focus from perfecting data to reasoning under its inherent uncertainty and incompleteness. Early contributions extended nearest neighbor classifiers to handle incomplete information, ensuring predictions align with the most reliable available data [35]. The dataset multiplicity problem formalized the challenges posed by unreliable or conflicting datasets, emphasizing the need to quantify uncertainty and its impact on predictions [51]. To address noisy and incomplete inputs, frameworks have emerged that provide statistical guarantees by constructing certain and approximately certain models, ensuring robust performance in tasks like linear regression and support vector machines [88]. Another critical development is the possible worlds framework, which trains models across multiple plausible interpretations of uncertain data, enhancing resilience in ambiguous settings [89]. Further, methods addressing fairness concerns, such as consistent range approximation, certify that predictive models remain unbiased despite biases in training data [90]. Robustness to programmable data biases has also been explored, where decision trees are evaluated over biased datasets to ensure consistent predictions and fairness [50].

Takeaway. As part of this overview, attendees will learn about the limitations of traditional approaches for cleaning, debiasing, and learning robust models. The tutorial will highlight recent progress in learning from incomplete, uncertain, and inconsistent data, as well as methods for propagating the impact of these imperfections on ML models. We will also cover applications for quantifying robustness and improving data cleaning. As a result, participants will gain practical insights into building resilient models that can effectively learn and perform under real-world data imperfections.

D. Open Challenges & Conclusion

To conclude the survey part, we will highlight some problems that we believe to be relevant to this space but have received little attention from the research community. Specifically, in the era of AI-assisted programming, there is a question of how the discipline of data debugging will be impacted [7], [17], [67], [75]. Additionally, there is the question of how data debugging as a part of the broader AI development lifecycle will be impacted by the recently introduced AI-regulation such as the EU AI Act, GDPR, SCPA, etc. [1], [11], [16], [74]

Takeaway: Attendees will get an overview of opportunities for future research directions in the space of data debugging.

III. Outline of the Hands-On Session

The hands-on session is divided into two parts, each of which is implemented in a separate Google Colab notebook.

A. Tools for Identifying Data Errors, Computing Pipeline Provenance, and Quantifying Uncertainty in Model Training and Predictions

Structure. The hands-on session will start with a one-hour introduction to various tools for identifying data errors, computing pipeline provenance, and quantifying uncertainty in model training and predictions. This part will leverage synthetically generated data from a hiring scenario, in particular a set of recommendation letters together with multiple tables of side data such as demographic information and social media details of the applicants. The corresponding ML use case will be to train a classifier to predict the sentiment of a recommendation letter. We will walk attendees through various examples of software tools for methods discussed in the survey such as kNN-Shapley [28], Gopher [62], [91], mlinspect [22], [23], [68], Datascope [33], and Zorro [89]. We will introduce the APIs of these tools, provide code snippets with usage examples, as well as a set short 5-minute programmming tasks for attendees, to encourage them to explore the tools themselves.

Content. We will start by showcasing how to identify and "recover" from data errors via data debugging. The data for this part consists of a single preprocessed table as training data without any complex features. We will inject synthetic noise such as label errors into the data and show how this negatively impacts the downstream quality metrics of the classifier. We will apply tools from Section II-A to identify impactful tuples with data errors, provide them to an "oracle" cleaning function and show how such prioritized cleaning improves quality metrics. Next, we will introduce ML pipelines for data preprocessing into the scenario (as discussed in Section II-B), which include additional side tables, and preprocess the data with a complex operations such as (fuzzy) joins, filters, projections with user-defined functions, as well as costly feature encoders. We will show how to compute fine-grained data provenance for the pipeline outputs and identify previously injected data errors in the source data of the pipeline based on the provenance information and the previously introduced tools. We conclude our tool introduction with scenarios discussed in Section II-C, where data quality issues cannot be fully resolved through cleaning. Here, we demonstrate how to reason about and quantify uncertainty in model training and predictions. Using a subset of the data, we inject synthetic missing attributes and uncertain labels to simulate real-world imperfections. We will focus on Zorro [89], a framework that symbolically propagates uncertainty due to missing values through the training process, allowing us to compute prediction ranges for model outputs. Attendees will observe how incomplete and uncertain data impact prediction reliability and robustness and will visualize the resulting uncertainty ranges for specific test points.

B. Data Debugging Challenge

In the final half hour of the hands-on session, we will present attendees with a challenging data cleaning task, inspired by recent benchmarks for data-centric AI development [45]. The attendees will be given access to a prepared training dataset with data errors unknown to them, and access to classifier with a validation set. Moreover, they will be given an "oracle" function, which allows them to specify a limited set of training tuples to clean (by supplying their identifiers). This oracle function will then evaluate the classifier (retrained on the partially cleaned data) on a hidden test set, and report the metric on this test set to the attendee. This will allow attendees to test their previously acquired knowledge about the data debugging tools in a challenging example scenario. We additionally plan to implement a live "leaderboard" to showcase which submissions introduced the highest improvements.

IV. Prerequisites & Context

Prerequisites. The target audience for our survey are researchers and practitioners with an interest in the intersection of data management, machine learning, and data quality. We intend to cover both theoretical aspects as well as practical aspects related to the design and deployment of real-world ML applications to appeal to a large audience. The survey assumes a very basic understanding of machine learning. For the hands-on session, attendees will need a laptop with internet access, and basic Python and data wrangling skills. We plan to implement the tutorial and tasks in Google Colab notebooks to avoid needing any local software installation or data downloads.

Difference with Previous Tutorials. This tutorial will be held for the first time and is custom-designed for the ICDE audience. However, this tutorial shares some overlap with several tutorials presented in recent years at data management venues, with key differences that we highlight here. The tutorial on "Data Collection and Quality Challenges for Deep Learning" at VLDB'20 [78] covered a topic similar to ours. However, the field has produced substantial developments over the past years in terms of methods and tools which we will present in our tutorial. The tutorial "Explainable AI: Foundations, Applications, Opportunities for Data Management Research" [61] at SIGMOD'22 also covered data importance methods, among other topics, but they focused on the context of model interpretability, as opposed to data debugging. Similarly, the tutorial "Applications and Computation of the Shapley Value in Databases and Machine Learning" at SIGMOD'24 [42] focuses entirely on the Shapley value and its various applications, while in our case this is presented as one of the various tools available for identifying data errors in ML pipelines.

V. Presenters

Bojan Karlaš is a postdoctoral research fellow at Harvard University. He currently works on developing machine learning pipelines for processing biomedical data and extracting clinically meaningful insights. Previously, he did his Ph.D. at the Systems Group of ETH Zürich working with Ce Zhang where he was developing systems for managing the ML development lifecycle with a specific focus on data debugging.

Babak Salimi is an Assistant Professor at the Halıcıoğlu Data Science Institute and is affiliated with the Department of Computer Science and Engineering at UC San Diego. His research focuses on the intersection of data management, machine learning, and responsible data science. His contributions have been recognized with the NSF CAREER Award (2024), SIGMOD Research Highlight Award (2020), SIGMOD Best Paper Award (2019), and VLDB Best Demo Award (2018).

Sebastian Schelter is a professor at the Berlin Institute for the Foundations of Learning and Data (BIFOLD). His research focuses on the intersection of data management and machine learning, with an aim on real-world applications. He was previously at the University of Amsterdam, New York University, and Amazon. His contributions have been recognized with a SIGMOD Systems Award, a SIGMOD Best Demo Runner-Up Award, and a Best Paper Runner-Up Award from the TRL workshop at NeurIPS.

References

[1] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). 2016. [Online; accessed 17-Feb-2019].

[2] Tyler Akidau, Robert Bradshaw, Craig Chambers, Slava Chernyak, Rafael J Fernández-Moctezuma, Reuven Lax, Sam McVeety, Daniel Mills, Frances Perry, Eric Schmidt, et al. The dataflow model: a practical approach to balancing correctness, latency, and cost in massive-scale, unbounded, out-of-order data processing. *Proceedings of the VLDB Endowment*, 8(12):1792–1803, 2015.

[3] Amazon Web Services. SageMaker Pipelines. https://aws.amazon.com/sagemaker/pipelines/, 2020.

[4] Agathe Balayn, Christoph Lofi, and Geert-Jan Houben. Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems. *The VLDB Journal*, 30(5):739–768, 2021.

[5] Denis Baylor, Eric Breck, Heng-Tze Cheng, Noah Fiedel, Chuan Yu Foo, Zakaria Haque, Salem Haykal, Mustafa Ispir, Vihan Jain, Levent Koc, et al. Tfx: A tensorflow-based production-scale machine learning platform. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1387–1395, 2017.

[6] Aharon Ben-Tal, Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Advanced Risk & Portfolio Management® Research Paper Series*, 2011.

[7] Sahil Bhatia, Jie Qiu, Niranjan Hasabnis, Sanjit A Seshia, and Alvin Cheung. Verified code transpilation with llms. *NeurIPS*, 2024.

[8] Matthias Boehm et al. Systemds: A declarative machine learning system for the end-to-end data science lifecycle. *CIDR*, 2020.

[9] Lukas Budach, Moritz Feuerpfeil, Nina Ihde, Andrea Nathansen, Nele Noack, Hendrik Patzlaff, Felix Naumann, and Hazar Harmouch. The effects of data quality on machine learning performance. *arXiv preprint arXiv:2207.14529*, 2022.

[10] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

[11] California Consumer Privacy Act (CCPA). Requests to Delete. https://oag.ca.gov/privacy/ccpa#sectiond, 2024.

[12] Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. Sample selection bias correction theory. In *International conference on algorithmic learning theory*, pages 38–53. Springer, 2008.

[13] Ian Covert, Chanwoo Kim, Su-In Lee, James Zou, and Tatsunori Hashimoto. Stochastic amortization: A unified approach to accelerate feature and data attribution. *NeurIPS*, 2024.

[14] Databricks. Mlflow Recipes. https://mlflow.org/docs/latest/recipes.html, 2022.

[15] Wei Du and Xintao Wu. Robust fairness-aware learning under sample selection bias. *arXiv preprint arXiv:2105.11570*, 2021.

[16] EU AI Act. Article 10: Data and Data Governance. https://artificialintelligenceact.eu/article/10/, 2024.

[17] Raul Castro Fernandez et al. How large language models will disrupt data management. *VLDB*, 2023.

[18] Lampros Flokas, Weiyuan Wu, Yejia Liu, Jiannan Wang, Nakul Verma, and Eugene Wu. Complaint-driven training data debugging at interactive speeds. In *Proceedings of the 2022 International Conference on Management of Data*, pages 369–383, 2022.

[19] Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, pages 2242–2251. PMLR, 2019.

[20] Google. Vertex AI Pipelines. https://cloud.google.com/vertex-ai/docs/pipelines, 2021.

[21] Stefan Grafberger, Paul Groth, and Sebastian Schelter. Automating and optimizing data-centric what-if analyses on native machine learning pipelines. *Proceedings of the ACM on Management of Data*, 1(2):1–26, 2023.

[22] Stefan Grafberger, Paul Groth, Julia Stoyanovich, and Sebastian Schelter. Data distribution debugging in machine learning pipelines. *The VLDB Journal*, 31(5):1103–1126, 2022.

[23] Stefan Grafberger, Shubha Guha, Julia Stoyanovich, and Sebastian Schelter. Mlinspect: A data distribution debugger for machine learning pipelines. In *Proceedings of the 2021 International Conference on Management of Data*, pages 2736–2739, 2021.

[24] Todd J Green, Grigoris Karvounarakis, and Val Tannen. Provenance semirings. In *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 31–40, 2007.

[25] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, 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*, pages 1–16, 2019.

[26] Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19, 2006.

[27] Jinyuan Jia, Xiaoyu Cao, and Neil Zhenqiang Gong. Intrinsic certified robustness of bagging against data poisoning attacks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 7961–7969, 2021.

[28] Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nezihe Merve Gurel, Bo Li, Ce Zhang, Costas J Spanos, and Dawn Song. Efficient task-specific data valuation for nearest neighbor algorithms. *arXiv preprint arXiv:1908.08619*, 2019.

[29] Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J Spanos. Towards efficient data valuation based on the shapley value. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1167–1176. PMLR, 2019.

[30] Ruoxi Jia, Fan Wu, Xuehui Sun, Jiacen Xu, David Dao, Bhavya Kailkhura, Ce Zhang, Bo Li, and Dawn Song. Scalability vs. utility: Do we have to sacrifice one for the other in data importance quantification? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8239–8247, 2021.

[31] Heinrich Jiang and Ofir Nachum. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*, pages 702–712. PMLR, 2020.

[32] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012.

[33] Bojan Karlaš, David Dao, Matteo Interlandi, Sebastian Schelter, Wentao Wu, and Ce Zhang. Data debugging with shapley importance over machine learning pipelines. In *The Twelfth International Conference on Learning Representations*, 2023.

[34] Bojan Karlas, Peng Li, Renzhi Wu, Nezihe Merve Gürel, Xu Chu, Wentao Wu, and Ce Zhang. Nearest neighbor classifiers over incomplete information: From certain answers to certain predictions. *Proc. VLDB Endow.*, 14(3):255–267, 2020.

[35] Bojan Karlaš, Peng Li, Renzhi Wu, Nezihe Merve Gürel, Xu Chu, Wentao Wu, and Ce Zhang. Nearest neighbor classifiers over incomplete information: From certain answers to certain predictions. *VLDB*, 2020.

[36] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pages 1885–1894. PMLR, 2017.

[37] Sanjay Krishnan, Jiannan Wang, Eugene Wu, Michael J Franklin, and Ken Goldberg. Activeclean: Interactive data cleaning for statistical modeling. *Proceedings of the VLDB Endowment*, 9(12):948–959, 2016.

[38] Sebastian Kruse, Thorsten Papenbrock, Hazar Harmouch, and Felix Naumann. Data anamnesis: Admitting raw data into an organization. *IEEE Data Eng. Bull.*, 39(2):8–20, 2016.

[39] Yongchan Kwon and James Zou. Beta shapley: a unified and noise-reduced data valuation framework for machine learning. *arXiv preprint arXiv:2110.14049*, 2021.

[40] Peng Li, Zhiyi Chen, Xu Chu, and Kexin Rong. Diffprep: Differentiable data preprocessing pipeline search for learning over tabular data. *Proceedings of the ACM on Management of Data*, 1(2):1–26, 2023.

[41] Anqi Liu and Brian Ziebart. Robust classification under sample selection bias. *Advances in neural information processing systems*, 27, 2014.

[42] Xuan Luo and Jian Pei. Applications and computation of the shapley value in databases and machine learning. In *Companion of the 2024 International Conference on Management of Data*, SIGMOD/PODS '24, page 630–635, New York, NY, USA, 2024. Association for Computing Machinery.

[43] Xiaozhong Lyu, Stefan Grafberger, Samantha Biegel, Shaopeng Wei, Meng Cao, Sebastian Schelter, and Ce Zhang. Improving retrieval-augmented large language models via data importance learning. *arXiv preprint arXiv:2307.03027*, 2023.

[44] Mohammad Mahdavi and Ziawasch Abedjan. Baran: Effective error correction via a unified context representation and transfer learning. *Proceedings of the VLDB Endowment*, 13(12):1948–1961, 2020.

[45] Mark Mazumder, Colby Banbury, Xiaozhe Yao, Bojan Karlaš, William Gaviria Rojas, Sudnya Diamos, Greg Diamos, Lynn He, Alicia Parrish, Hannah Rose Kirk, et al. Dataperf: Benchmarks for data-centric ai development. *Advances in Neural Information Processing Systems*, 36, 2024.

[46] Sean McGregor. Preventing repeated real world ai failures by cataloging incidents: The ai incident database. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 15458–15463, 2021.

[47] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[48] Xiangrui Meng, Joseph Bradley, Burak Yavuz, Evan Sparks, Shivaram Venkataraman, Davies Liu, Jeremy Freeman, DB Tsai, Manish Amde, Sean Owen, et al. Mllib: Machine learning in apache spark. *Journal of Machine Learning Research*, 17(34):1–7, 2016.

[49] Metaflow.org. A framework for real-life ML, AI, and data science. https://metaflow.org, 2024.

[50] Anna Meyer, Aws Albarghouthi, and Loris D'Antoni. Certifying robustness to programmable data bias in decision trees. *Advances in Neural Information Processing Systems*, 34, 2021.

[51] Anna P Meyer, Aws Albarghouthi, and Loris D'Antoni. The dataset multiplicity problem: How unreliable data impacts predictions. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 193–204, 2023.

[52] Microsoft. Azure Machine Learning Pipelines. https://learn.microsoft.com/en-us/azure/machine-learning/concept-ml-pipelines, 2018.

[53] Mohammad Hossein Namaki, Avrilia Floratou, Fotis Psallidas, Subru Krishnan, Ashvin Agrawal, Yinghui Wu, Yiwen Zhu, and Markus Weimer. Vamsa: Automated provenance tracking in data science scripts. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1542–1551, 2020.

[54] Hongseok Namkoong and John C Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[55] Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021.

[56] Kwanghyun Park, Karla Saur, Dalitso Banda, Rathijit Sen, Matteo Interlandi, and Konstantinos Karanasos. End-to-end optimization of machine learning prediction queries. In *Proceedings of the 2022 International Conference on Management of Data*, pages 587–601, 2022.

[57] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[58] Alireza Pirhadi, Mohammad Hossein Moslemi, Alexander Cloninger, Mostafa Milani, and Babak Salimi. Otclean: Data cleaning for conditional independence violations using optimal transport. *Proceedings of the ACM on Management of Data*, 2(3):1–26, 2024.

[59] Geoff Pleiss, Tianyi Zhang, Ethan Elenberg, and Kilian Q Weinberger. Identifying mislabeled data using the area under the margin ranking. *Advances in Neural Information Processing Systems*, 33:17044–17056, 2020.

[60] Neoklis Polyzotis, Martin Zinkevich, Sudip Roy, Eric Breck, and Steven Whang. Data validation for machine learning. *Proceedings of machine learning and systems*, 1:334–347, 2019.

[61] Romila Pradhan, Aditya Lahiri, Sainyam Galhotra, and Babak Salimi. Explainable ai: Foundations, applications, opportunities for data management research. In *Proceedings of the 2022 International Conference on Management of Data*, SIGMOD/PODS '22, pages 2452–2457, 2022.

[62] Romila Pradhan, Jiongli Zhu, Boris Glavic, and Babak Salimi. Interpretable data-based explanations for fairness debugging. In *Proceedings of the 2022 International Conference on Management of Data*, pages 247–261, 2022.

[63] Fotis Psallidas, Yiwen Zhu, Bojan Karlas, Jordan Henkel, Matteo Interlandi, Subru Krishnan, Brian Kroth, Venkatesh Emani, Wentao Wu, Ce Zhang, et al. Data science through the looking glass: Analysis of millions of github notebooks and ml. net pipelines. *ACM SIGMOD Record*, 51(2):30–37, 2022.

[64] Theodoros Rekatsinas, Xu Chu, Ihab F. Ilyas, and Christopher Ré. Holoclean: Holistic data repairs with probabilistic inference. *Proc. VLDB Endow.*, 10(11):1190–1201, 2017.

[65] Ashkan Rezaei, Anqi Liu, Omid Memarrast, and Brian Ziebart. Robust fairness under covariate shift. *arXiv preprint arXiv:2010.05166*, 2020.

[66] Elan Rosenfeld, Ezra Winston, Pradeep Ravikumar, and Zico Kolter. Certified robustness to label-flipping attacks via randomized smoothing. In *International Conference on Machine Learning*, pages 8230–8241. PMLR, 2020.

[67] Sebastian Schelter and Stefan Grafberger. Messy code makes managing ml pipelines difficult? just let llms rewrite the code!, 2024.

[68] Sebastian Schelter, Stefan Grafberger, Shubha Guha, Bojan Karlas, and Ce Zhang. Proactively screening machine learning pipelines with arguseyes. In *Companion of the 2023 International Conference on Management of Data*, pages 91–94, 2023.

[69] Maximilian E Schüle, Luca Scalerandi, Alfons Kemper, and Thomas Neumann. Blue elephants inspecting pandas: Inspection and execution of machine learning pipelines in sql. In *EDBT*, pages 40–52, 2023.

[70] Scikit-learn. Pipelines and composite estimators. https://scikit-learn.org/stable/modules/compose.html, 2024.

[71] Shafaq Siddiqi, Roman Kern, and Matthias Boehm. Saga: A scalable framework for optimizing data cleaning pipelines for machine learning applications. *Proceedings of the ACM on Management of Data*, 1(3):1–26, 2023.

[72] Jacob Steinhardt, Pang Wei Koh, and Percy Liang. Certified defenses for data poisoning attacks. In *NeurIPS*, 2017.

[73] Julia Stoyanovich, Serge Abiteboul, Bill Howe, HV Jagadish, and Sebastian Schelter. Responsible data management. *Communications of the ACM*, 65(6):64–74, 2022.

[74] DigiChina Stanford University. Internet information service algorithmic recommendation management provisions.

[75] Matthias Urban and Carsten Binnig. Caesura: Language models as multi-modal query planners. *CIDR*, 2024.

[76] Jiachen T Wang and Ruoxi Jia. Data banzhaf: A robust data valuation framework for machine learning. In *International Conference on Artificial Intelligence and Statistics*, pages 6388–6421. PMLR, 2023.

[77] Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010.

[78] Steven Euijong Whang and Jae-Gil Lee. Data collection and quality challenges for deep learning. *Proceedings of the VLDB Endowment*, 13(12):3429–3432, 2020.

[79] Weiyuan Wu, Lampros Flokas, Eugene Wu, and Jiannan Wang. Complaint-driven training data debugging for query 2.0. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pages 1317–1334, 2020.

[80] Doris Xin, Hui Miao, Aditya Parameswaran, and Neoklis Polyzotis. Production machine learning pipelines: Empirical analysis and optimization opportunities. In *Proceedings of the 2021 international conference on management of data*, pages 2639–2652, 2021.

[81] Gyeong-In Yu, Saeed Amizadeh, Sehoon Kim, Artidoro Pagnoni, Ce Zhang, Byung-Gon Chun, Markus Weimer, and Matteo Interlandi. Windtunnel: towards differentiable ml pipelines beyond a single model. *Proceedings of the VLDB Endowment*, 15(1):11–20, 2021.

[82] Matei Zaharia, Mosharaf Chowdhury, Michael J Franklin, Scott Shenker, and Ion Stoica. Spark: Cluster computing with working sets. In *2nd USENIX workshop on hot topics in cloud computing (HotCloud 10)*, 2010.

[83] Richard S. Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. Learning fair representations. In *ICML (3)*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 325–333. JMLR.org, 2013.

[84] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

[85] Hantian Zhang, Ki Hyun Tae, Jaeyoung Park, Xu Chu, and Steven Euijong Whang. iflipper: Label flipping for individual fairness. *Proceedings of the ACM on Management of Data*, 1(1):1–26, 2023.

[86] Xuezhou Zhang, Xiaojin Zhu, and Stephen Wright. Training set debugging using trusted items. In *AAAI*, 2018.

[87] Yixuan Zhang, Boyu Li, Zenan Ling, and Feng Zhou. Mitigating label bias in machine learning: Fairness through confident learning. *arXiv preprint arXiv:2312.08749*, 2023.

[88] Cheng Zhen, Nischal Aryal, Arash Termehchy, and Amandeep Singh Chabada. Certain and approximately certain models for statistical learning. *Proceedings of the ACM on Management of Data*, 2(3):1–25, 2024.

[89] Jiongli Zhu, Su Feng, Boris Glavic, and Babak Salimi. Learning from uncertain data: From possible worlds to possible models. *NeurIPS*, 2024.

[90] Jiongli Zhu, Sainyam Galhotra, Nazanin Sabri, and Babak Salimi. Consistent range approximation for fair predictive modeling. *VLDB*, 2023.

[91] Jiongli Zhu, Romila Pradhan, Boris Glavic, and Babak Salimi. Generating interpretable data-based explanations for fairness debugging using gopher. In *Proceedings of the 2022 International Conference on Management of Data*, pages 2433–2436, 2022.