Journal of Information Technology and Computing
https://jitc.sabapub.com
ISSN: 2709-5916

2023 Volume 4, Issue 1 : 46 – 58
DOI : 10.48185/jitc.v4i1.725

Quality challenges in Deep Learning Data Collection in perspective of Artificial Intelligence

N Gowri Vidhya[1,*], D.Nirmala[2], T. Manju[1]

[1]First Department of Information Technology, St. Joseph's Institute of Technology, Chennai, India
[2]Department of Electronics and communication Engineering, Prince Shri Venkateshwara Padmavathy Engineering College, Chennai, India

Received: 01.03.2023 • Accepted: 12.06.2023 • Published: 27.06.2023 • Final Version: 27.06.2023

Abstract: With reinforcement learning powered by big data and computer infrastructure, data-centric AI is driving a fundamental shift in the way software is developed. To treat data as a first-class citizen on par with code, software engineering must be rethought in this situation. One surprise finding is how much time is spent on data preparation throughout the machine learning process. Even the most powerful machine learning algorithms will struggle to perform adequately in the absence of high-quality data. Advanced technologies that are data-centric are being used more frequently as a result. Unfortunately, a lot of real-world datasets are small, unclean, biased, and occasionally even tainted. In this study, we focus on the scientific community for data collecting and data quality for deep learning applications. Data collection is essential since modern algorithms for deep learning rely mostly on large-scale data collecting than classification techniques. To enhance data quality, we investigate data validation, cleaning, and integration techniques. Even if the data cannot be completely cleaned, robust model training strategies enable us to work with imperfect data during training the model. Furthermore, despite the fact that that these issues have gotten less attention in conventional data management studies, bias and fairness are significant themes in modern application of machine learning. In order to prevent injustice, we investigate controls for fairness and strategies for doing so before, during, and after model training. We believe the information management community is in a good position to address these problems.

Keywords: AI, Data centric, data cleansing, validation and integration

1. Introduction

Deep learning is often used to glean knowledge from enormous volumes of data. Natural language processing, healthcare, and self-driving cars are just a few of the many applications. Deep learning has become so well-liked because of its exceptional performance when combined with the availability of vast amounts of data and robust computer infrastructure. According to IDC [1], the entire amount of data will have rapidly expanded to 175 zettabytes by 2025. Additionally, software is capable executing a wide range of tasks at superhuman levels thanks to powerful GPUs and TPUs. Machine learning is replacing software in software engineering, which is a fundamental paradigm shift [2]. Traditional software engineering includes all three phases of creating, implementing, and debugging code. In contrast, machine learning starts with data and trains a function on it. Particularly, the time spent acquiring, cleaning, and preparing information for machine learning training accounts for 45% [3] or even 80–90% [4] of the entire time. A machine learning platform's high level code also needs a lot fewer lines of code than conventional software does. Finally, to keep

* Corresponding Author: vighyagowri@gmail.com

the training model improving, hyper - parameter tweaking may be needed. This entire process—from data preparation to model deployment—has been actively developed by businesses and is widely acknowledged as a new software development paradigm.

2. Overview of the Study

Data-centric AI [5], whose primary goal is to enhance data pre-processing for greater model accuracy rather than the model training algorithm, has gained prominence more lately. These trends compel us to investigate difficulties related to deep learning data gathering and quality from a data-centric AI standpoint. Figure 1 depicts a streamlined process from beginning to end, from data collection to model deployment.

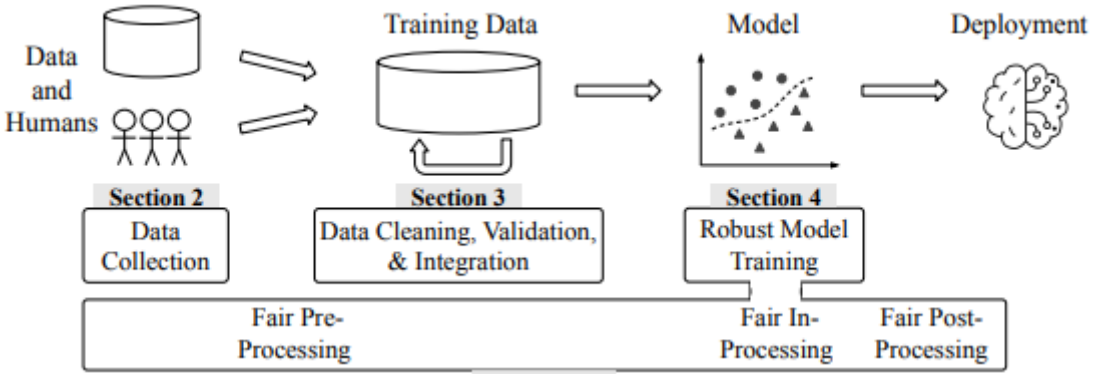

Figure 1. Deep learning challenges in AI perspective

Since deep learning systems are much more sophisticated [6], we only go over the most important steps here. Our talk will start with data collection. Deep learning needs more training data than traditional machine learning because feature engineering is not as problematic. Sadly, a lack of data and the challenge of articulating the learned models inhibits many businesses from implementing deep learning. The second topic is data cleansing and validation. Although there is a plethora of knowledge on data cleaning, not all techniques directly improve deep learning performance [7–10]. A recent deep learning issue called data poisoning is another issue that has to be addressed, especially by the data management community. Data poisoning is becoming a bigger issue because attackers create data with the malicious intent to reduce the model accuracy of AI systems. In response, the goal of the field of study known as data sanitization is to defend against such assaults [11–14]. Nevertheless, in addition to cleaning and verifying data to improve model accuracy, it is also increasingly important for responsible AI to show that justice is opposed to biassed data. In fact, an increasing number of studies on data validation have recognised that advancing AI ethics, particularly justice, is an important subject for future research [15]. Fairness measurements and unfairness mitigation are the core subjects of model fairness research [16,17], whether it be before, during, or after model training. Recent studies are now addressing model fairness and robustness together due to their close linkage, where data bias and noise may affect each other within the same training data [18,20,25]. While the issues covered in this survey are diverse, we believe that in order to enhance data-centric AI, it is essential to have a thorough awareness of the data difficulties that arise during the deep learning process [19]. Each subtopic is not only significant but is also the subject of extensive community research. Research on data collection, cleaning, and validation has usually been done by the data management community [21–23]. Robust model training is highly valued by the machine learning and security communities, whereas fair model training is highly

valued by the machine learning and fairness communities. Due to their close ties to the input data, the data management industry is currently conducting extensive research on issues like fairness and robustness [24].

3. Data Collection

The description of data collection has been modified and shortened in light of a presentation [25] and a survey [26] conducted by two of the authors. There are three fundamental ways to collect data. Data acquisition is the process of locating, upgrading, or creating new datasets. The process of classifying data in a manner that is instructional so that a machine-learning model may learn from it is the second problem. Given the high cost of labeling, other strategies can be used, such as crowd sourcing, shoddy supervision, and semi-supervised learning. Furthermore, if data and models already exist, they can be improved instead of starting from scratch with data collection or categorization.

3.1. Data Acquisition

Data acquisition, or the process of locating datasets suitable for use in machine learning model training, is the first step to be taken in the situation of insufficient data. In this survey, we explore three techniques: data generation, data augmentation, and data discovery. Data discovery is the process of indexing and searching databases [27]. To construct fabricated instances, data augmentation manipulates or combines tagged samples. If there isn't enough data, the last option is to create datasets on one's own via crowdsourcing or synthetic data production techniques.

3.1.1 Data Discovery

Data discovery refers to the problem of indexing and searching datasets that are either existing in corporate data lakes [28] or on the Web [29]. One example is the Goods system [30], which searches tens of billions of datasets in Google's data lake. Goods uses a post-hoc methodology to crawl datasets from diverse sources and extract information without the help of the dataset owners in order to create a single dataset library. Each entry in the catalogue contains details about a dataset, such as its size, provenance, who developed it, who read it, and its schema. Goods also provides dataset annotations, monitoring, and search. Google Dataset Search, a public version of Goods [31], supports science dataset searches. Recently, these data discovery tools have developed to become more interactive. An interactive data management and search application built on top of the Jupyter Notebook data science platform is Juneau [32], which serves as a suitable example. The key technological challenge in this scenario is locating the pertinent tables. Juneau employs similarity metrics that logically capture the objective of each data set's creation in order to compare records, schemas, and provenance information. Finding tables that can be joined or unioned effectively is essential when using data lakes, and LSH-based algorithms have been developed to perform set overlap search or unionable attribute retrieval on tables [33].

3.1.3 Data Generation

Another way to collect or acquire fresh data is through generating it. It's usual to use crowd sourcing platforms like Amazon Mechanical Turk [34], where one can define tasks and pay people to generate or locate data. For instance, a task can direct workers to find face images of a particular demographic on public websites [35]. For some domains, such as those involving mobility data and driving data, a simulator or generator can also be employed. Two examples are Hermoupolis [36] and Crash to Not Crash [92]. Domain randomization [37] is a potent technique for generating a variety of realistic

data from a simulator by altering its parameters. We can see that GANs also generate new data, but they require a sufficient amount of real data for training.

3.2. Data Labeling

The next step is to label the instances if there are sufficient datasets. We discuss data labeling strategies for making use of pre-existing labels as well as for manually or automatically labeling data without labels [38].

3.2.1 Utilize Existing Labels

The typical labelling technique is called semi-supervised learning [39], and it aims to anticipate future labels based on those that have already been assigned. You can use the machine learning benchmarks that already exist [40], which provide labelled data for a variety of tasks. The most fundamental form is self-training [41], in which a model is trained on the easily available labelled data before being applied to the unlabeled data. Following acceptance, the forecasts with the highest confidence values are added to the training set. Although other approaches, including Tritraining [42], Co-learning [43], and Co-training [34], do not make this assumption, this strategy is predicated on the notion that we may trust the high confidence.

3.2.2 Manual Labeling from No Labels

If there are no labeling to commence with but the company has the funds to pay workers, a frequent tactic is to use crowd sourcing services like Amazon Mechanical Turk to perform labelling. Given how important labelling is, there are services designed specifically for it, such as Google Cloud Labeling [45] as well as Amazon Contributory factor Ground truth [44]. Choosing labelling tasks, hiring labelers, and giving them the resources and assistance they require to classify the data are all possible with Sagemaker. Because the workers don't always have the required expertise, crowd source may not be feasible. As a result, because it could be expensive, subject-matter experts should only be consulted as a last resort.

3.2.3 Automatic Labeling from No Labels

Weak supervision, which tries to (semi-)automatically create imperfect labels (henceforth referred to as "weak" labels), has gained favour in recent years. Weak supervision operates at a scale where another higher volume may makes upward for the lower labeling quality. Weak supervision is beneficial in applications when there are few or no labels to begin with. Early approaches include crowdsourcing and distant supervision [46], which labels training data using external knowledge sets. More recently, data programming has improved on these techniques by creating and combining many labelling algorithms to produce weak labels.

Improving Existing Data

One can enhance the quality of current data and models in addition to searching and classifying datasets. This method works well in a variety of situations. Let's say the target application is innovative or complex, with no external datasets that are relevant, or where gathering more data no longer improves the model's accuracy due to its poor quality. A preferable choice in this case could be to enhance the available data. Relabeling is one efficient method of label improvement. Kristy choi et a l[47] .'s illustration of the significance of label improvement is done by comparing the model's accuracy trends to more training instances for datasets of various characteristics. Even if

more data are used, the model's accuracy plateaus as the quality of the data deteriorates rather than increasing.

4. Data Validation, Cleaning, and Integration

Various errors are typically present in the training data. Through the use of data visualisation and schema construction techniques, data validation [48] capabilities in machine learning platforms like TensorFlow Extended (TFX) [49] enable the early discovery of such data errors. Data cleaning can be used to correct the data, and a wealth of literature [50] has been written about various integrity requirements.

4.1. Data Validation

Data visualisation is frequently used for data validation for machine learning and is very effective [45]. A human may perform quick but critical sanity checks on the data using visualisation, which is more effective than traditional data cleaning and helps prevent later, more serious errors. A sample opensource programme called Facets [8] presents a variety of statistics and dataset contents that can be used for data sanity checks to prevent more serious issues in the future. In addition to manual visualisation, research has been done on the automatic production of new visuals that can be used for validation [52]. Interesting visualisations are frequently produced by an innovative system called SeeDB [51]. To gauge interest, SeeDB uses a utility metric with a deviation component.

4.1.1 Schema-based validation

Schema-based validation [27,] is often used in real life. Tensorflow Data Validation (TFDV) [36] establishes this assumption on the assumption of a continuous training environment in which input data is regularly provided. TFDV builds a data schema using previous data sets to evaluate incoming data sets and alert users to data anomalies.

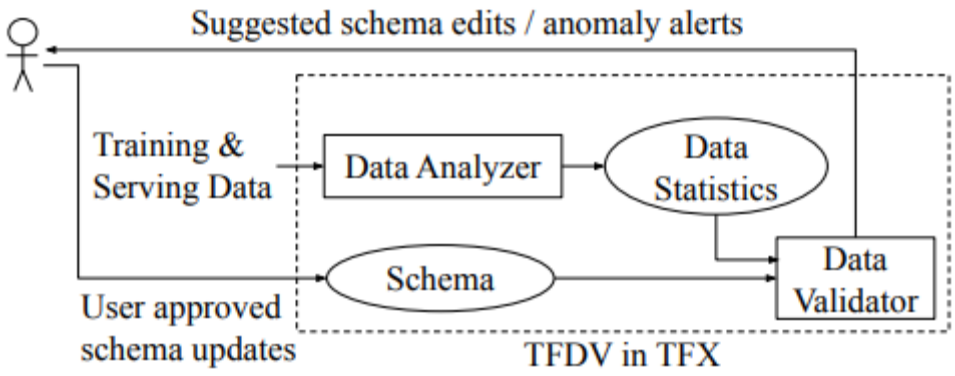

Figure 2. Tensor flow validation

In order to potentially address the underlying cause of the each abnormality, TFDV provides concrete action items. This architecture does not offer a summary of the metrics for the characteristics, unlike a typical database schema, which does. The user must then choose whether to modify the schema or correct the data if a new dataset differs from the preexisting schema.

5. Data Cleaning

Data cleaning has a long tradition of removing various well-defined flaws by meeting integrity criteria including critical obstacles, domain restrictions, referential constraints, & ability. For an introduction, consult the book Information Extraction [53]. There has also been a recent poll on data cleansing methods for computer vision and vice versa [54]. Let first describe one of the newest current statistics cleaning techniques to show how sophisticated these processes have become. William et al [55] uses probabilistic inference to fix data by satisfying three essential requirements: multiple consistency criteria, independent dictionaries for value verification, and application of quantitative statistics.

6. Robust Model Training

Even when the correct data has been gathered and cleansed, data quality issues may still arise during model training. Real-world datasets are considered to be imprecise and erroneous despite the process of data cleaning. The problem of data poisoning has sparked a great deal of interest in the machine learning community because it has been studied in concept (i.e., robust statistics) and reality for more than 50 years [56]. Can the machine-learning model learn from the data and make predictions as if it were clean? The primary inquiry that need to be addressed. In cases where we are unable to recover all of the clean data, it aims to develop machine learning algorithms that are immune against the most severe data corruptions. It focuses on data feature corruptions.

6.1. Noisy Features

Noisy features are commonly introduced via adversarial attacks. We focus on the arsenic attack, also known as contaminating of the training data, in order to adhere to the core theme of our study. Even during training phase of a machine-learning model, an attacker attempts to contaminate the training data by adding purposefully produced data to fool the training process. External conditions like colour noise and image blurring that might not be removed by data cleaning can also contribute to noisy features in addition to adversarial noise.

6.2. Missing Features

Considering missing data can reduce statistical significance and provide skewed predictions, data imputation has been a contentious issue in both statistics and machine learning. Any form of data can have functionalities, but because of the large current rate and recurrent sensor failure that create missing values, researchers particularly concentrate on multivariate time information in this work.

7. Fair Model Training

Now that model fairness is in focus, biased data may result in a model that is discriminating and, consequently, unjust. The goal of robust training the model, where this problem is closely related, is to address bias instead of disturbance in the learning algorithm. One well-known example is the Northpointe COMPAS tool, which predicts a defendant's chance of committing another crime. According to a ProPublica investigation [57], white defendants are considerably less likely to be labeled as high risk than black defendants, which turns out to be false in practice. Other well-known examples include an AI-based adopting this approach that excludes prospective employees based on their gender [3], an AI-based picture viewer that mistakenly classifies people as belonging to a particular race [10], and more. The study of algorithmic fairness was created as a result of these

events. Different factors can be at play in COMPAS' prejudice. The training data may be biassed in cases where there is more information available for a certain group. It's possible that factors outside of race contributed more to crime than race itself. Even the fairness metric can be questioned if it does not accurately reflect reality. Fairness analysis is typically a very complicated subject that includes factors not seen in the data. Since fairness and ethics are covered in detail in the current fair ML book [26], here we merely focus on fairness concerns with technical solutions. We go over how to assess fairness and how to minimize unfairness in particular detail.

8. Convergence With Robustness Techniques

Methods for fairness and robustness have recently begun to converge. This direction is inevitable because both approaches deal with data challenges, but neither one supersedes the other. Fair training just focuses on removing the bias from the data and presumes that it is unadulterated. However, the sensitive feature itself could be hazy or even nonexistent. However, robust training places more emphasis on raising accuracy overall and ignores differences in performance amongst different sensitive groups. Fairness and strength are typically not antithetical principles. For instance, if the data is already biased due to the removal of an excessive amount of data from an underrepresented group, discarding noisy data for robust training may exacerbate bias [58]. Figure 3 depicts these processes.

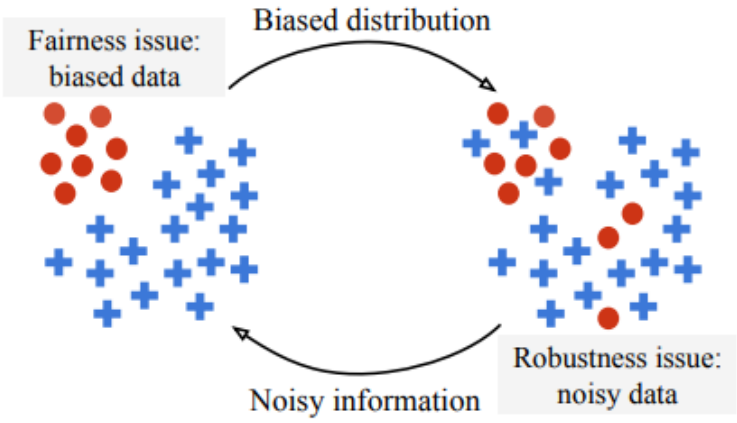

Figure 3. Fairness and robustness issue

The convergence can happen in three ways: by improving the robustness of fair procedures (fairness-oriented), by improving the fairness of robust techniques (robust-oriented), and by combining equally fair and robust training methods. We give a summary of recent research for each of the three tactics.

Fairness-oriented Approaches

The first path towards convergence is to increase the dependability of fair training. Currently, there are two methods for doing this research: when the sensitive group information is garbled or lacking altogether. If some users actively disregard or hide their group affiliations, the first scenario might occur. An analysis of fair training results on noisy sensitive group information shows that the true fairness violation on a clean sensitive group can be constrained by the distance between this group and its noisy variant [59]. With the intention of assessing the fairness of the real data distribution by changing the unfairness tolerance, noise-tolerant fair training algorithms [60] have also been proposed. In the second case, the sensitive attribute is completely missing. The data collection

procedure in this situation occasionally fails to obtain related data due to a number of circumstances, including legal restrictions. Distributional Robust Optimization (DRO) [62] has been used to improve the model performance for minority sensitive groups without using the group information from [61]. The objective is to roughly minimize the worst-case (latent) group loss by identifying the worst-performing samples (Figure 4) and assigning them more weight. Adversarial reweighted learning for impartiality [63] executes antagonistic learning between a classified as well as an opponent that finds less accurate clustered regions and sets greater weights on such regions, assuming that unobserved sensitive traits are linked with the characteristics and labels. Robustness-centered Techniques In order to improve a model's general accuracy, robust training is planned, yet it may discriminate.

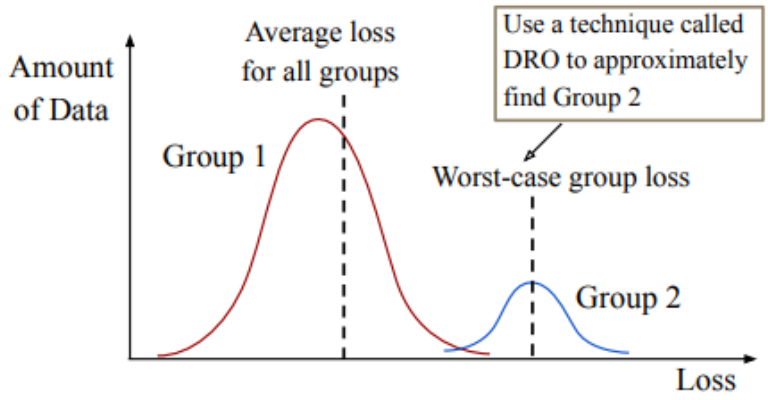

Figure 4. DRO based fair algorithm

The data collection procedure in this situation occasionally fails to obtain related data due to a number of circumstances, including legal restrictions. Distributionally Robust Optimization (DRO) [62] has been used to improve the model performance for minority sensitive groups without using the group information from [61]. The objective is to roughly minimize the worst-case (latent) group loss by identifying the worst-performing samples (Figure 4) and assigning them more weight. Adversarially reweighted learning for impartiality [63] executes antagonistic learning between a classifier as well as an opponent that finds less accurate clustered regions and sets greater weights on such regions, assuming that unobserved sensitive traits are linked with the characteristics and labels. Robustness-centered Techniques In order to improve a model's general accuracy, robust training is planned, yet it may discriminate.

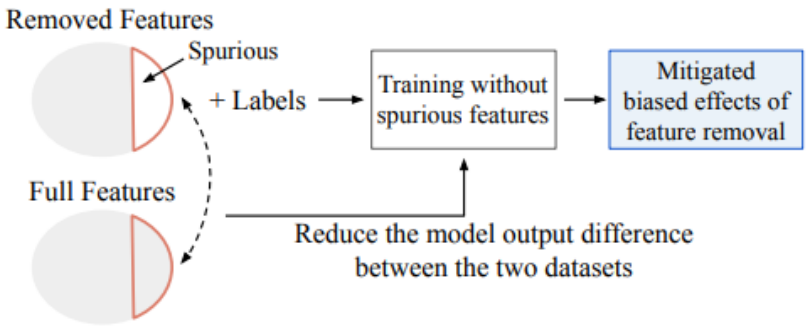

Figure 5. Self training technique

Similar Fusions Equitable training that is both effective and fair is possible. One goal is to simultaneously make the model learning fair and reliable. In the mutual information-based accessing

basic as FR-Train [67], a classifier, a discriminative model for fairness, and a classification algorithm for robustness compete to make the classifiers fair and robust (Figure 6). A modern sample selection method [68] adaptively selects training samples for trustworthy and equitable training the model (Figure 7). This approach doesn't demand that the model be changed or that more recent data be used. A benevolence ERM architecture [69] has been proposed in light of the discovery that group-dependent labeling noises may reduce model accuracy and fairness.

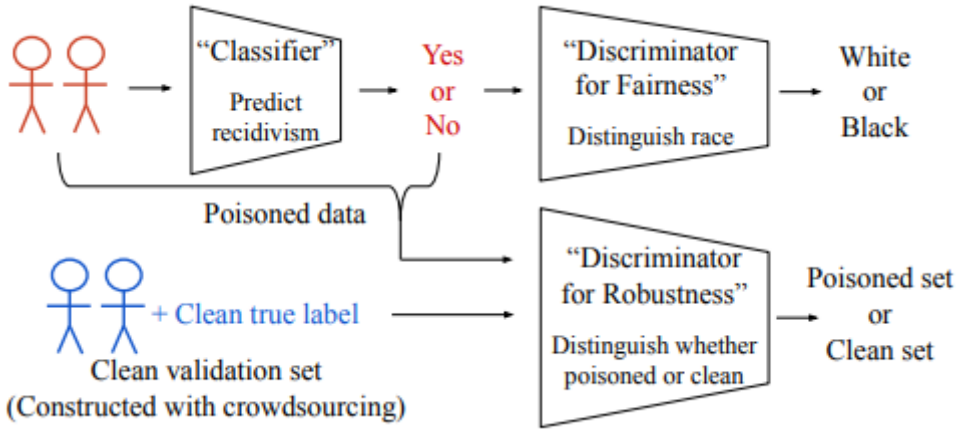

Figure 6. FR Training

Because the surrogate loss more closely resembles the actual loss, group-dependent label sounds are less harmful. Playing the role of an enemy and coming up with attacks that hurt both accuracy and fairness is another method for assuring accurate and equitable training. A gradient-based attack method is proposed in fairness-targeted poisoning attacks [70] that select the best attack sites with the greatest fairness-reducing impact.

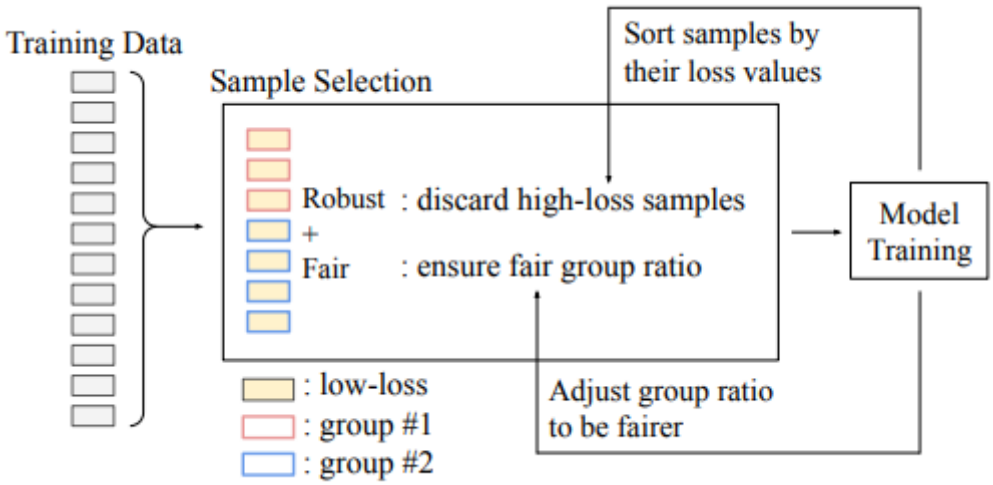

Figure 7. Adaptive sample selection

9. Overall Findings and Future Directions

Our findings are listed. In Section 3, we discussed the three phases of data collection strategies: data collection, data labeling, and data and model enhancement. Some of the strategies have been studied by the machine learning community, while others have been studied by the data management community. In Section 4, we covered the key methods for data integration, cleaning, sanitization, and validation. Data validation can be done using visualizations and schema knowledge. Although

modern strategies, as detailed in section 5, are primarily focused on improving model accuracy, data cleaning has received substantial investigation. Data sanitization has a special flavor in that it can defend against attacks using poison. Data integration is a hurdle when dealing with multimodal data. In Section 6, we discussed how noisy or missing labels lead to poor generalization on test data. Research on noisy labeling is now limited by accumulated noise or only looks at training data in part. Hybrid and semi-supervised techniques can achieve very high accuracy even with noisy training data. Self- and semi-supervised methods are actively being developed to benefit from massive volumes of unlabeled data. We covered convergence with ro-bustness processes, unfairness mitigation techniques, and fairness evaluations in Section 8. The mitigation may be done prior to, during, or after the model training. Pre-processing is advantageous when training data can be changed. In-processing can be useful when the training algorithm can be altered. Post-processing may be used when we are unable to change the data or model training. Fair and robust types of convergence of robustness techniques can be distinguished.

Concluding Remark

In the future of data-centric AI, deep learning will only grow more crucial as the importance of data collecting and quality improvement rises. The four key topics we discussed were data collection, data filtering, validation, and integration, robust model construction, and fair model training. Although many communities have looked into these topics, they must be used in conjunction. Our poll is intended to act as a catalyst in the development of data-centric AI, where we believe all data approaches will eventually merge with effective and fair training procedures.

References

[1] Amazon Mechanical Turk. https://www.mturk.com/. Accessed July 13th, 2022.
[2] Amazon SageMaker Ground Truth. https: //aws.amazon.com/sagemaker/groundtruth/. Accessed July 13th, 2022.
[3] Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/ article/us-amazon-com-jobs-automation-insightidUSKCN1MK08G. Accessed July 13th, 2022.
[4] CrowdFlower Data Science Report. https: //visit.figure-eight.com/rs/416-ZBE-142/images/ CrowdFlower DataScienceReport 2016.pdf.
[5] Data age 2025. https://www.seagate.com/our-story/ data-age-2025/.
[6] Data-centric AI resource hub. https:// datacentricai.org/.
[7] Data prep still dominates data scientists' time, survey finds. https://www.datanami.com/2020/07/06/dataprep-still-dominates-data-scientists-time-surveyfinds/.
[8] Facets – visualization for ML datasets. https://paircode.github.io/facets/. Accessed July 13th, 2022.
[9] GCP AI platform data labeling service. https:// cloud.google.com/ai-platform/data-labeling/docs. Accessed July 13th, 2022.
[10] Google apologises for Photos app's racist blunder. https://www.bbc.com/news/technology-33347866. Accessed July 13th, 2022.
[11] Kaggle. https://www.kaggle.com. Data Collection and Quality Challenges in Deep Learning: A Data-Centric AI Perspective 21
[12] Principles for AI ethics. https://research.samsung.com/ artificial-intelligence. Accessed July 13th, 2022.
[13] Responsible AI practices. https://ai.google/ responsibilities/responsible-ai-practices. Accessed July 13th, 2022.
[14] Responsible AI principles from Microsoft. https:// www.microsoft.com/en-us/ai/responsible-ai. Accessed July 13th, 2022.
[15] Software 2.0. https://medium.com/@karpathy/software2-0-a64152b37c35.

[16] South Korean AI chatbot pulled from Facebook after hate speech towards minorities. https://www.theguardian.com/world/2021/jan/14/ time-to-properly-socialise-hate-speech-ai-chatbotpulled-from-facebook. Accessed July 13th, 2022.
[17] Trusting AI. https://www.research.ibm.com/artificialintelligence/trusted-ai/. Accessed July 13th, 2022.
[18] Alekh Agarwal, Alina Beygelzimer, Miroslav Dud´ık, John Langford, and Hanna M. Wallach. A reductions approach to fair classification. In ICML, pages 60–69, 2018.
[19] Pulkit Agrawal, Rajat Arya, Aanchal Bindal, Sandeep Bhatia, Anupriya Gagneja, Joseph Godlewski, Yucheng Low, Timothy Muss, Mudit Manu Paliwal, Sethu Raman, Vishrut Shah, Bochao Shen, Laura Sugden, Kaiyu Zhao, and Ming-Chuan Wu. Data platform for machine learning. In SIGMOD, pages 1803–1816, 2019.
[20] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald C. Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. Software engineering for machine learning: a case study. In ICSE, pages 291–300, 2019.
[21] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias: There's software used across the country to predict future criminals. And its biased against blacks., 2016.
[22] Saeed Anwar and Nick Barnes. Real image denoising with feature attention. In CVPR, pages 3155–3164, 2019.
[23] Abolfazl Asudeh, Zhongjun Jin, and H. V. Jagadish. Assessing and remedying coverage for a given dataset. In ICDE, pages 554–565, 2019.
[24] Stephen H. Bach, Daniel Rodriguez, Yintao Liu, Chong Luo, Haidong Shao, Cassandra Xia, Souvik Sen, Alexander Ratner, Braden Hancock, Houman Alborzi, Rahul Kuchhal, Christopher R´e, and Rob Malkin. Snorkel drybell: A case study in deploying weak supervision at industrial scale. In SIGMOD, pages 362–375, 2019.
[25] Tadas Baltrusaitis, Chaitanya Ahuja, and LouisPhilippe Morency. Multimodal machine learning: A survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell., 41(2):423–443, 2019.
[26] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness and Machine Learning. fairmlbook.org, 2019. http://www.fairmlbook.org.
[27] Denis Baylor, Eric Breck, Heng-Tze Cheng, Noah Fiedel, Chuan Yu Foo, Zakaria Haque, Salem Haykal, Mustafa Ispir, Vihan Jain, Levent Koc, Chiu Yuen Koo, Lukasz Lew, Clemens Mewald, Akshay Naresh Modi, Neoklis Polyzotis, Sukriti Ramesh, Sudip Roy, Steven Euijong Whang, Martin Wicke, Jarek Wilkiewicz, Xin Zhang, and Martin Zinkevich. TFX: A tensorflow-based production-scale machine learning platform. In KDD, pages 1387–1395, 2017.
[28] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, et al. AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development, 2019.
[29] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art, 2017.
[30] David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In ICLR, 2020.
[31] David Berthelot, Nicholas Carlini, Ian J. Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In NeurIPS, pages 5050–5060, 2019.
[32] Felix Biessmann, Jacek Golebiowski, Tammo Rukat, Dustin Lange, and Philipp Schmidt. Automated data validation in machine learning systems. IEEE Data Eng. Bull., 44(1):51–65, 2021.
[33] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Srndic, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In ECML PKDD, pages 387–402. Springer, 2013.
[34] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In COLT, pages 92– 100, New York, NY, USA, 1998. ACM.
[35] Matthias Boehm, Iulian Antonov, Sebastian Baunsgaard, Mark Dokter, Robert Ginth¨or, Kevin Innerebner, Florijan Klezin, Stefanie N. Lindstaedt, Arnab Phani, Benjamin Rath, Berthold Reinwald, Shafaq Siddiqui, and Sebastian Benjamin Wrede. Systemds: A declarative machine learning system for the end-to-end data science lifecycle. In CIDR, 2020.
[36] Eric Breck, Martin Zinkevich, Neoklis Polyzotis, Steven Whang, and Sudip Roy. Data validation for machine learning. In MLSys, 2019.

[37] Dan Brickley, Matthew Burgess, and Natasha F. Noy. Google dataset search: Building a search engine for datasets in an open web ecosystem. In WWW, pages 1365–1375, 2019.
[38] Michael J. Cafarella, Alon Y. Halevy, Hongrae Lee, Jayant Madhavan, Cong Yu, Daisy Zhe Wang, and Eugene Wu. Ten years of webtables. PVLDB, 11(12):2140– 2149, 2018.
[39] Jos´e Cambronero, John K. Feser, Micah J. Smith, and Samuel Madden. Query optimization for dynamic imputation. Proc. VLDB Endow., 10(11):1310–1321, 2017.
[40] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey. CoRR, abs/1810.00069, 2018.
[41] Haw-Shiuan Chang, Erik G. Learned-Miller, and Andrew McCallum. Active bias: Training more accurate neural networks by emphasizing high variance samples. In NeurIPS, pages 1002–1012, 2017.
[42] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. Nature Scientific Reports, 8(1):6085, 2018.
[43] Andrew Chen, Andy Chow, Aaron Davidson, Arjun DCunha, Ali Ghodsi, Sue Ann Hong, Andy Konwinski, Clemens Mewald, Siddharth Murching, Tomas Nykodym, Paul Ogilvie, Mani Parkhe, Avesh Singh, Fen Xie, Matei Zaharia, Richard Zang, Juntai Zheng, and Corey Zumar. Developments in mlflow: A system to accelerate the machine learning lifecycle. In DEEM@SIGMOD, pages 5:1–5:4, 2020.
[44] Irene Y. Chen, Fredrik D. Johansson, and David A. Sontag. Why is my classifier discriminatory? In NeurIPS, pages 3543–3554, 2018.
[45] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In KDD, pages 785–794, 2016.
[46] Yu Cheng, Ilias Diakonikolas, and Rong Ge. Highdimensional robust mean estimation in nearly-linear time. In SIAM, pages 2755–2771, 2019.
[47] Kristy Choi, Aditya Grover, Trisha Singh, Rui Shu, and Stefano Ermon. Fair generative modeling via weak supervision. In ICML, pages 1887–1898, 2020.
[48] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2):153–163, 2017.
[49] Alexandra Chouldechova and Aaron Roth. A snapshot of the frontiers of fairness in machine learning. Commun. ACM, 63(5):82–89, 2020.
[50] Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Leveraging labeled and unlabeled data for consistent fair binary classification. In NeurIPS, pages 12739–12750, 2019.
[51] Andrew Cotter, Heinrich Jiang, and Karthik Sridharan. Two-player games for efficient non-convex constrained optimization. In ALT, pages 300–332, 2019.
[52] Gabriela F. Cretu, Angelos Stavrou, Michael E. Locasto, Salvatore J. Stolfo, and Angelos D. Keromytis. Casting out demons: Sanitizing training data for anomaly sensors. In IEEE S&P, pages 81–95, 2008.
[53] Ekin D. Cubuk, Barret Zoph, Dandelion Man´e, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation strategies from data. In CVPR, pages 113–123, 2019.
[54] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. SIAM Journal on Computing, 48(2):742–864, 2019.
[55] William Dieterich, Christina Mendoza, and Tim Brennan. Compas risk scales: Demonstrating accuracy equity and predictive parity. Technical report, Northpoint Inc, 2016.
[56] AnHai Doan, Alon Y. Halevy, and Zachary G. Ives. Principles of Data Integration. Morgan Kaufmann, 2012.
[57] Mohamad Dolatshah, Mathew Teoh, Jiannan Wang, and Jian Pei. Cleaning crowdsourced labels using oracles for statistical classification. PVLDB, 12(4):376–389, 2018.
[58] Xin Luna Dong and Theodoros Rekatsinas. Data integration and machine learning: A natural synergy. In KDD, pages 3193–3194, 2019.
[59] Mike Dreves, Gene Huang, Zhuo Peng, Neoklis Polyzotis, Evan Rosen, and Paul Suganthan G. C. Validating data and models in continuous ML pipelines. IEEE Data Eng. Bull., 44(1):42–50, 2021.
[60] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.
[61] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. In ITCS, pages 214–226, 2012.
[62] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In KDD, pages 259–268, 2015.

[63] Raul Castro Fernandez, Ziawasch Abedjan, Famien Koko, Gina Yuan, Samuel Madden, and Michael Stonebraker. Aurum: A data discovery system. In ICDE, pages 1001–1012, 2018.
[64] Dean P. Foster and Robert A. Stine. Alpha-investing: a procedure for sequential control of expected false discoveries. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70(2):429–444, 2008.
[65] Karan Goel, Albert Gu, Yixuan Li, and Christopher R´e. Model patching: Closing the subgroup performance gap with data augmentation. In ICLR, 2021.
[66] Ian J. Goodfellow. NIPS 2016 tutorial: Generative adversarial networks. CoRR, abs/1701.00160, 2017.
[67] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In NIPS, pages 2672–2680, 2014.
[68] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In ICLR, 2015.
[69] J. Gordon. Introducing tensorflow hub: A library for reusable machine learning modules in tensorflow., 2018.
[70] Stefan Grafberger, Julia Stoyanovich, and Sebastian Schelter. Lightweight inspection of data preprocessing in native machine learning pipelines. In CIDR, 2021.