www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 09, September 2025, pp : 897-905

e-ISSN : 2583-1062
Impact Factor : 7.001

# DATA QUALITY CHALLENGES FOR MACHINE LEARNING MODEL

**Padakanti Mahesh**[1]

[1]M.Sc, Department Of Mathematics & Computer Science, Osmania University, Telangana, India.

## 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. In this paper, we focus on the scientific community for data collection and data quality for deep learning applications. Data collection is essential since modern algorithms for deep learning rely more on large-scale data collection 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 these issues have gotten less attention in conventional data management studies, bias and fairness are significant themes in modern applications of machine learning. In order to prevent injustice, we investigate controls for fairness and strategies for doing so before, during, and after model training.

**Keywords**: Artificial Intelligence, Data Cleaning, Data Validation, Robustness Techniques, Machine Learning.

## 1. INTRODUCTION

Deep learning frequently extracts knowledge from vast amounts of data. Natural language processing, healthcare, and self-driving cars are just a few of 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. IDC [1] says that by 2025, the total amount of data will have grown very quickly to 175 zetta bytes. Software can also do a lot of different tasks at superhuman speeds thanks to GPUs and TPUs that are very powerful. Machine learning is replacing software in software engineering, which is a fundamental paradigm shift [2]. Traditional software engineering includes all three phases — creating, implementing, and debugging code. In contrast, machine learning starts with data and trains a function on it. Specifically, the acquisition, cleaning, and preparation of information for machine learning training consumes 45% [3] or even 80–90% [4] of the total time. Additionally, the high-level code of a machine learning platform necessitates significantly fewer lines of code than that of conventional software. Finally, to keep the training model improving, hyperparameter tweaking may be needed. Businesses have actively developed this entire process—from data preparation to model deployment—and widely acknowledge it as a new software development paradigm.

## 2. OVERVIEW OF THE STUDY

Recently, data-centric AI [5] has become more popular. Its main goal is to improve data pre-processing for more accurate models, not the model training algorithm. Because of these trends, we need to look into problems with gathering and quality data for deep learning from the point of view of data-centric AI. From gathering data to deploying the model, Figure 1 shows a streamlined process from start to finish. 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 learnt models discourage many businesses from implementing deep learning. The second topic is data cleaning and validation. Even though a lot is known about cleaning data, not all methods directly improve the performance of deep learning [7–10]. The data management community, in particular, must address a recent deep learning issue known as data poisoning. It's getting worse when attackers create data on purpose to make AI systems less accurate. This is called "data poisoning." In response, the goal of the field of study known as data sanitisation is to defend against such assaults [11–14]. Still, it is important for responsible AI to show that justice is against biassed data, in addition to cleaning and verifying data to make models more accurate. In fact, an increasing number of studies on data validation have acknowledged that advancing AI ethics, particularly justice, is an important subject for future research [15]. How to measure fairness and how to reduce unfairness are the main topics of model fairness research [16, 17], whether it's done before, during, or after training. Model fairness and robustness are now being studied together because they are closely connected. For example, data

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 09, September 2025, pp : 897-905
www.ijprems.com
editor@ijprems.com
e-ISSN : 2583-1062
Impact Factor : 7.001

bias and noise can affect each other in the same training data [18, 20, 25]. While the issues covered in this survey are diverse, we believe that 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. The data management community typically conducts research on data collection, cleaning, and validation [21–23]. The machine learning and security communities value robust model training a lot, while the machine learning and fairness communities value fair model training a lot. Because they work with a lot of data, the data management industry is doing a lot of research on topics like fairness and robustness [24].

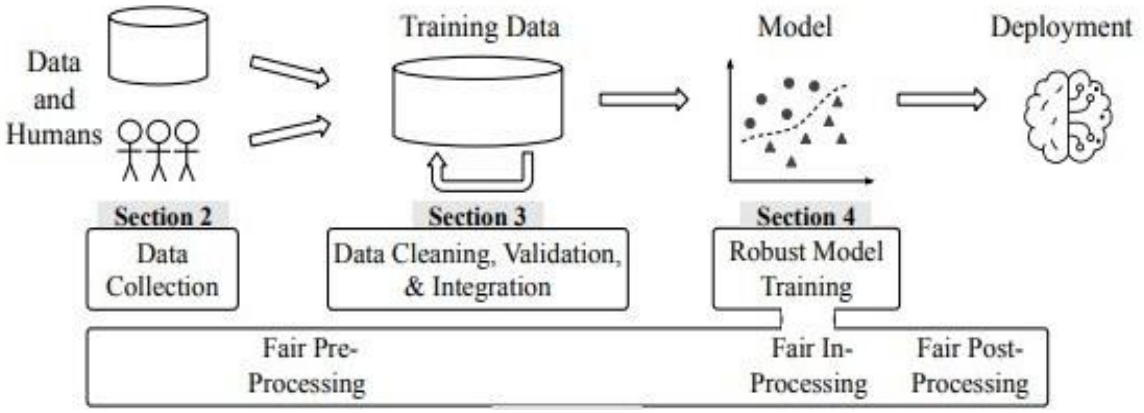


**Figure 1:** Deep learning challenges in AI perspective

## 3. DATA COLLECTION

Because of a presentation [25] and a survey [26] that two of the authors did, the description of how the data was collected has been changed and shortened. There are three fundamental ways to collect data. Data acquisition is the process of locating, upgrading, or creating new datasets. The second problem is classifying data in an instructional manner so that a machine-learning model can learn from it. Because labelling costs a lot, other methods can be used, like crowdsourcing, poor supervision, and learning with only some supervision. Furthermore, we can improve existing data and models rather than starting from scratch with data collection or categorisation.

### 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

The problem of indexing and searching datasets, either existing in corporate data lakes [28] or on the Web [29], is known as data discovery. One example is the Goods system [30], which searches tens of billions of datasets in Google's data lake. Goods use a post-hoc methodology to crawl datasets from diverse sources and extract information without the help of the dataset owners 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 provide dataset annotations, monitoring, and search. Google Dataset Search, a public version of Goods [31], supports science dataset searches. Recently, these data discovery tools have become more interactive. An interactive data management and search application built on top of the Jupyter Notebook platform for data science 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 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].

#### 3.1.2. Data Generation

Another way to collect or acquire fresh data is through generating it. It's usual to use crowdsourcing 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]. Some domains, like those involving mobility data and driving data, also benefit from the use of a simulator or generator. Two examples

are Hermoupolis [36] and Crash to Not Crash [92]. Domain randomisation [37] is a potent technique for generating various 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 labelling strategies for making use of pre-existing labels as well as for manually or automatically labelling data without labels [38].

#### 3.2.1. Utilize Existing Labels

The most common way to label things is called semi-supervised learning [39], and it tries to guess what labels will be given next by looking at the labels that have already been given. You can use the machine learning benchmarks that already exist [40], which provide labelled data for various tasks. Self-training [41] is the simplest type. In this method, a model is trained on the easy-to-find labelled data before being used on the unlabelled data. Once accepted, we add the forecasts with the highest confidence values to the training set. Although other approaches, including Tri-Training [42], Co-Learning [43], and Co-Training [34], do not make this assumption, this strategy is predicated on the notion that we may trust it with high confidence.

#### 3.2.2. Manual Labeling from No Labels

If the company doesn't have any labelling to do but has the money to pay workers, they will often use crowdsourcing services like Amazon Mechanical Turk to do it. Given how important labelling is, there are services designed specifically for it, such as Google Cloud Labelling [45] as well as Amazon Contributory Factor Ground Truth [44]. Choosing labelling tasks, hiring labellers, 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, crowdsourcing may not be feasible. Therefore, consult subject-matter experts only as a last resort due to their potential expense.

#### 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 recently. Weak supervision operates at a scale where another higher volume may make up for the lower labelling quality. Weak supervision is beneficial in applications when there are few or no labels to begin with. Early methods include crowdsourcing and distant supervision [46], which use outside knowledge sets to label training data. More recently, data programming has improved on these techniques by creating and combining many labelling algorithms that produce weak labels.

#### 3.2.4. Improving Existing Data

Along with searching and sorting datasets, one can improve the quality of existing data and models. This method works well in various 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. Relabelling is one efficient method of label improvement. Kristy Choi et al. [47] show how important it is to improve labels by looking at the model's accuracy over time with more training examples from datasets with different features. 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

Machine learning frequently uses data visualisation for data validation and finds it to be very effective [45]. A human may perform quick, 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 open-source programme called Facets [8] presents various statistics and dataset contents that can be used for data sanity checks to prevent more serious issues in the future. To gauge interest, See DB uses a utility metric with a deviation component.

#### 4.1.1. Schema-Based Validation

Real life often uses schema-based validation [27]. TensorFlow Data Validation (TFDV) [36] bases its assumption on a continuous training environment that regularly provides input data. TFDV uses old data sets to build a data schema that it can use to look at new data sets and let users know about any problems with the data.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 09, September 2025, pp : 897-905
www.ijprems.com
editor@ijprems.com
e-ISSN : 2583-1062
Impact Factor : 7.001

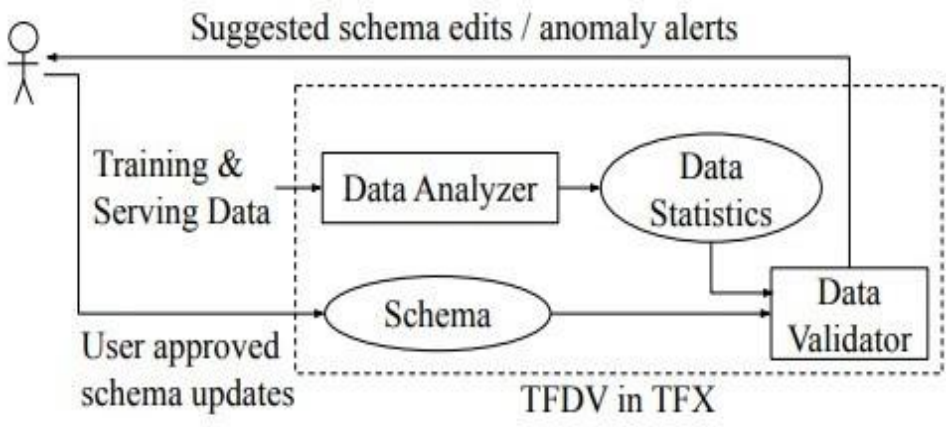


**Figure 2:** Tensorflow validation

To potentially address the underlying cause of each abnormality, TFDV provides concrete action items. There is no summary of the metrics for the characteristics in this architecture, but there is one in a typical database schema. If a new dataset is different from the old schema, the user must then decide whether to change the schema or fix the data.

## 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, and abilities.

## 6. ROBUST MODEL TRAINING

Even after properly cleaning and collecting the data, errors can still occur. This can happen during model training. Real-world datasets are considered imprecise and erroneous, despite the process of data cleaning. Can the machine-learning model learn from the data and make predictions as if it were clean? We need to address the primary inquiry. In the event that we can't get back all of the clean data, the goal is to create machine learning algorithms that can handle even the worst data corruption. It focuses on data feature corruptions.

### 6.1. Noisy Features

Adversarial attacks commonly introduce noisy features. We focus on the arsenic attack, also known as contaminating the training data, to adhere to the core theme of our study. An attacker tries to taint the training data of a machine-learning model by adding data that was purposely made to trick the training process, even while the model is still being trained. 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

Data imputation has been a controversial topic in both statistics and machine learning because missing data can make results less significant and predictions less accurate. Any kind of data can be useful, but researchers are focussing on multivariate time information in this work because of the high current rate and frequent sensor failures that leave values blank.

## 7. FAIR MODEL TRAINING

Now that the focus is on model fairness, biassed data can lead to a discriminating and consequently unjust model. Robust training of the model, a closely related problem, aims to address bias rather than disrupt the learning algorithm. One well-known example is the Northpointe COMPAS tool, which predicts a defendant's chance of committing another crime. Other well-known examples include an AI-based approach that excludes prospective employees based on their gender [3], an AI-based picture viewer that incorrectly classifies people as belonging to a particular race [10], and more. These events gave rise to the study of algorithmic fairness. Different factors can be at play in COMPAS' prejudice. The training data may be biased 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. We can question even the fairness metric if it fails to accurately reflect reality. Fairness analysis is typically a very complicated subject that includes factors not seen in the data. The current fair ML book [26] covers fairness and ethics in detail, so here we only concentrate on fairness concerns and their technical solutions. We go over how to assess fairness and minimise unfairness in particular detail.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 09, September 2025, pp : 897-905
www.ijprems.com
editor@ijprems.com
e-ISSN : 2583-1062
Impact Factor : 7.001

## 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. The sensitive feature itself could be unclear or even absent. However, robust training places more emphasis on raising accuracy overall and ignores differences in performance among different sensitive groups. Fairness and strength are typically not antithetical principles. Figure 3 depicts these processes.

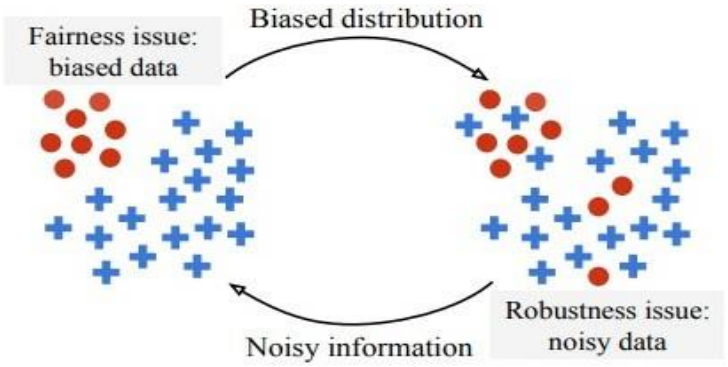


**Figure 3:** Fairness and robustness issue

There are three ways for convergence to happen: by making fair procedures more robust (fairness-oriented), by making robust techniques more fair (robust-oriented), or by combining training methods that are both fair and robust. We summarise the most recent research for each of the three tactics.

### 8.1. Fairness-oriented Approaches

The first step towards convergence is to increase the dependability of fair training. Currently, we employ two methods for conducting this research: either when the sensitive group information is unclear or completely absent. If some users actively disregard or hide their group affiliations, the first scenario might occur. This was shown by looking at the results of fair training on noisy sensitive group information. In the second case, the sensitive attribute has been completely removed. The data collection procedure in this situation occasionally fails to obtain related data due to several circumstances, including legal restrictions. The objective is to roughly minimise the worst-case (latent) group loss by identifying the worst-performing samples (Figure 4) and assigning them a greater weight.

This is done by assuming that sensitive traits that aren't seen are connected to the characteristics and labels. Robustness-Centred Techniques, we plan robust training to improve a model's general accuracy,but it may discriminate.The data collection procedure in this situation occasionally fails to obtain related data due to several circumstances, including legal restrictions. The objective is to roughly minimise the worst-case (latent) group loss by identifying the worst-performing samples (Figure 4) and assigning them a greater weight. This is done by assuming that sensitive traits that aren't seen are connected to the characteristics and labels. Robustness-Centred Techniques We plan robust training to improve a model's general accuracy, but it may discriminate.

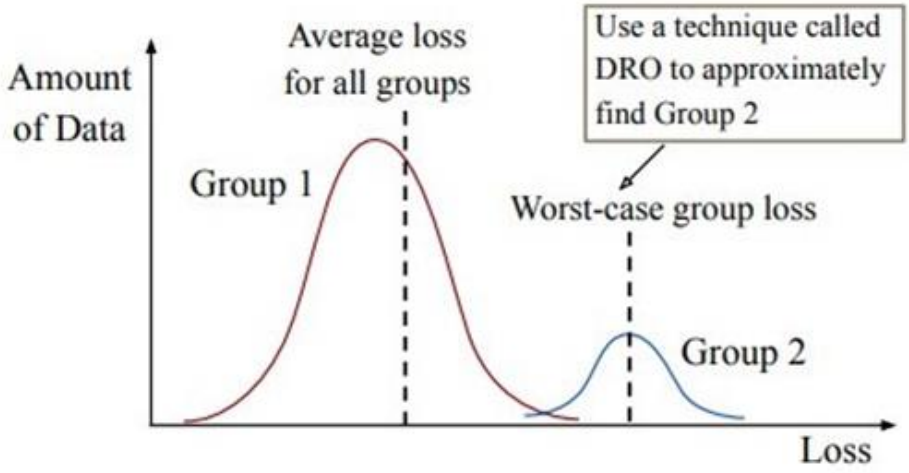


**Figure 4:** DRO based fair Algorithm

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 09, September 2025, pp : 897-905
www.ijprems.com
editor@ijprems.com
e-ISSN : 2583-1062
Impact Factor : 7.001

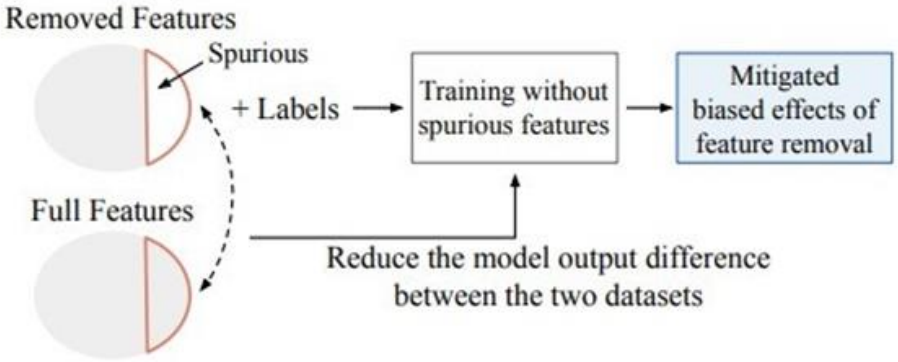


**Figure 5:** Self training techniques

Similar Fusions Equitable training that is both effective and fair is possible. One goal is to simultaneously make the model learning fair and reliable. According to FR-Train, there is a competition between a classifier, a discriminative model for fairness, and a classification algorithm for robustness to make the classifiers fair and robust (Figure 6). A new sample selection method picks training samples that are fair and reliable so that the model can be trained correctly (Figure 7). This approach does not require changing the model or using more recent data.

Label sounds that depend on groups are less harmful because the fake loss is more like the real loss. 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.

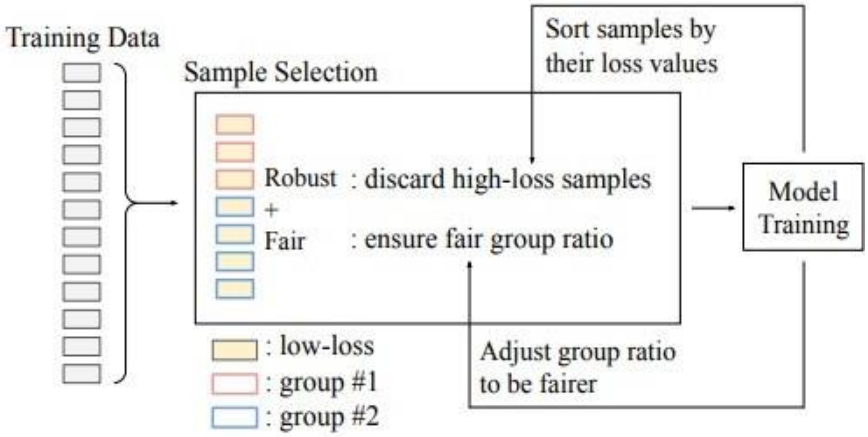


**Figure 6:** FR Training

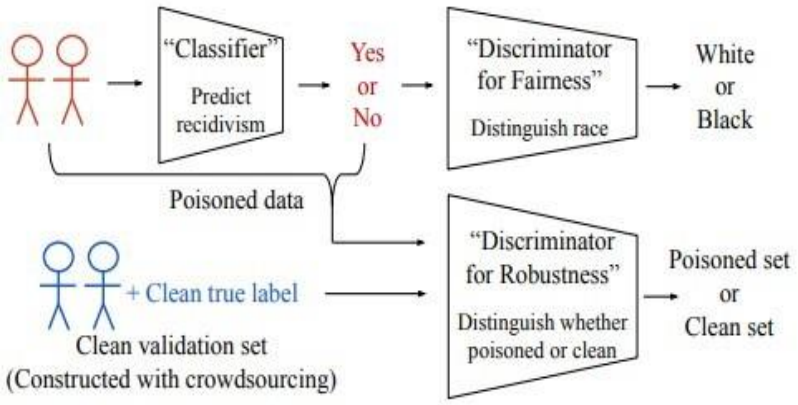


**Figure 7:** Adaptive sample selection

## 9. OVERALL FINDINGS AND FUTURE DIRECTIONS

We have listed our findings. We discussed the three phases of data collection strategies: data collection, data labelling, and data and model enhancement. The machine learning community has studied some of the strategies, while the data management community has studied others. We covered the key methods for data integration, cleaning, sanitisation, and validation. Visualisations and schema knowledge can facilitate data validation. Primarily focus on improving

model accuracy, but data cleaning has received substantial investigation. Data sanitisation has a special flavour in that it can defend against attacks with poison. Data integration is a hurdle when dealing with multimodal data. We discussed how noisy or missing labels lead to poor generalisations based on test data. Accumulated noise or partial focus on training data now limits research on noisy labelling. Hybrid and semi-supervised techniques can achieve very high accuracy even with noisy training data. Researchers are actively developing self- and semi-supervised methods to leverage massive volumes of unlabelled data. We talked about convergence with robustness processes, techniques for reducing unfairness, and fairness evaluations. We can perform the mitigation either before, during, or after the model training. When it is possible to change the training data, pre-processing becomes advantageous. Processing becomes beneficial when we have the ability to modify the training algorithm. When we cannot alter the data or the model training, we may resort to post-processing. We can distinguish between fair and robust types of convergence in robustness techniques.

## 10. CONCLUSION

Deep learning will become even more important in the future of data-centric AI as it becomes more important to collect data and make it better. The four key topics we discussed were data collection, data filtering, validation and integration, robust model construction, and fair model training. Many communities have explored these topics, but they need to work together. Our poll is meant to help the growth of data-centric AI. Eventually, we think that all data approaches will come together with fair and effective training methods.

## 11. 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 minorities. https://www.theguardian.com/world/2021/jan/14/ time-to-properly-socialise-hate-speech-ai-chatbotpulled-from-facebook. Accessed July 13th, 2022. Towards
[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.