Data-Centric AI - Challenges and Opportunities

Team Extreme - IDD MATHS 21124015, 21124031, 21124032, 21124034, 21124051

Abstract

Artificial intelligence (AI) systems are primarily algorithms that learn archetypal traits from massive data clouds to solve problems. There are two primary approaches for enhancing the AI systems perfomance: model-centric AI and data-centric AI. Model centric AI entails incrementally enhancing an algorithm or piece of code while maintaining the nature and volume of the data being collected. In data centric AI, the data quality is continually improved, keeping the model constant. Even while model-centric AI has dominated over the previous three decades, it has recently come under fire for being restricted to business and industries where platforms with millions of users may freely rely on generalized solutions. Instead of advocating either approach, this paper supports a 'both and' position. The alternative data-centric approach may well be necessary to overcome the accused limitations of model centric AI. It shouldn't, however, lead to a reduction in interest in model-centric AI. According to us, successful 'problem solving' requires both an analysis of how we act upon things (algorithm) and an understanding of their data/properties and states.

Index Terms

Data-Centric AI, Model-Centric AI, Data Engineering, Challenges

I. INTRODUCTION

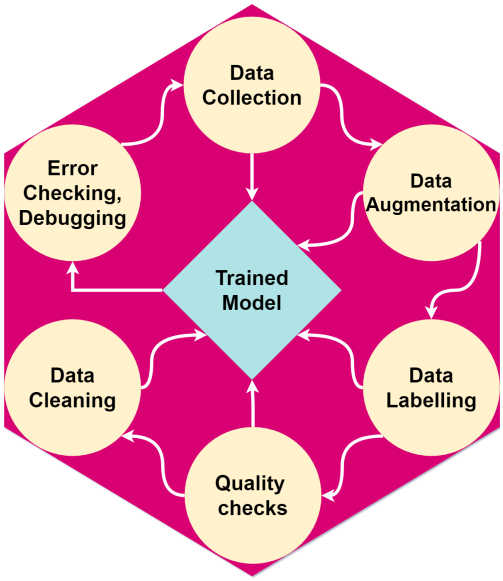

Fig. 1: Data - Centric AI Workflow

Machine learning and artificial intelligence models are used in nearly all industry sectors today, including those in the automotive, healthcare, financial, and semiconductor industries. Earlier, the core of AI was ML models. But when AI spread across many industries and grew in demand, ML models became nothing more than a commodity and no longer the core of AI. Instead, greater performance in ML models and Systems would be driven by the training data. Data is therefore the new oil. People frequently believe that data is static, which is a major factor in ML models becoming commonplace. The data literally means "that which is supplied," which is true. For the majority of individuals, machine learning entails downloading a pre-made dataset and creating a model. Afterward, work on enhancing the model's performance. What we refer to as data in real-time business is the result of the processes. When compared to actual model construction, maintaining and enhancing the deployed models accounts for a larger share of the cost and efficacy of many real-time ML systems. The main answer for these applications is to improve the quality of the data.

Everyone who uses machine learning (ML) will agree to the point that actual data cleansing and preparation account for 80% of ML activities. Therefore, a key duty for an ML engineer to take into consideration is assuring the data quality. Data fluctuates significantly for higher stakes ML/AI. This data flow has a longer duration and various unfavorable effects. From data collection through model deployment, data movement can occur anywhere. Conventional methods which are completely dependent on code are overshadowed by modern day ML/AI systems which takes into consideration both code and data. Practitioners strive to work upon the model/code to improve ML/AI system instead of focusing more on data part of the system. However, it is often preferable to improve the input data in actual applications than experimenting with various models. Through case studies in AI, Computer Vision, and other fields, Andrew Ng illustrated this concept and showed how enhancing the data will result in a better model. AI systems that are model-centric focus on how to alter the model (code) to enhance performance. Data Centric AI's prime focus is to amend the input-output labels in a systematic manner so as to achieve a model with better performance. Model centric and data centric methods need to be balanced well in order to provide a strong AI solution.

This paper brings together pros of both the approaches rather than discarding any of them completely. We contend that, in order to overcome the model-centric AI's shortcoming, additional consideration must be given to the alternative data-centric AI. This shouldn't, however, lead to a decline in interest in the model-centric strategy because it gives us the anchor, we need to evaluate an AI system's performance as

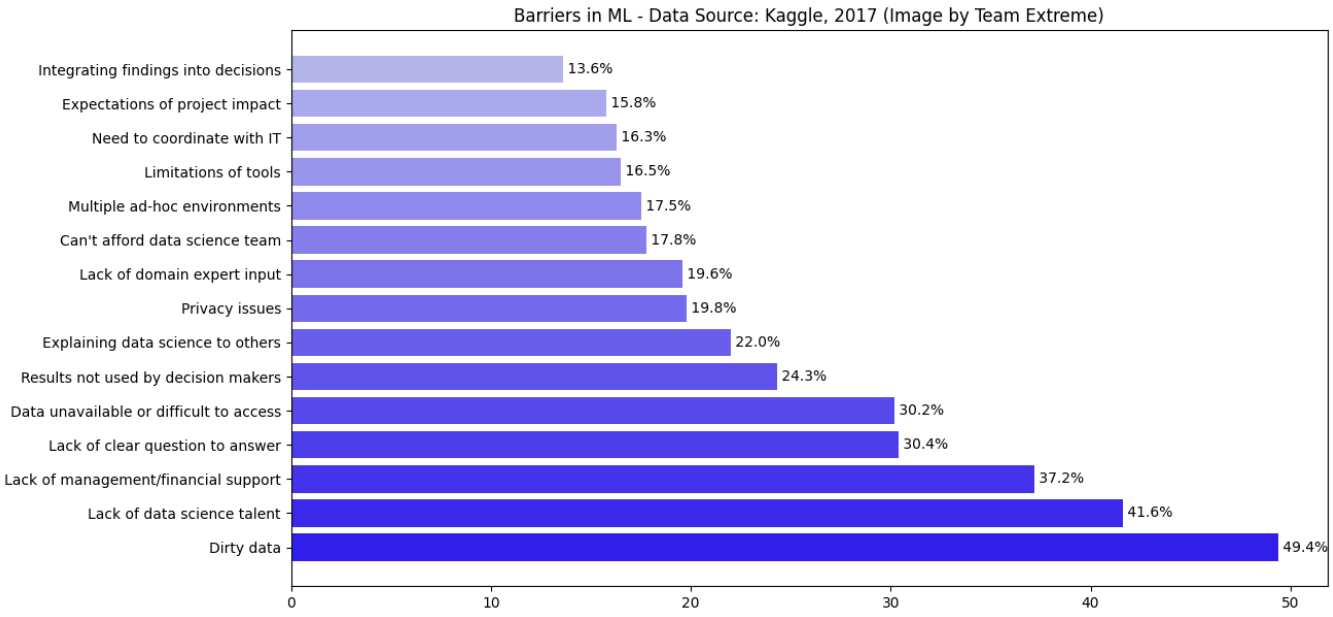

Fig. 2: Barriers faced in ML; Data Source - Kaggle Survey, 2017

the data set is upgraded. The structure of the essay is as follows. We present the idea of data-centric AI and explain why it is becoming more and more popular among ML practitioners in Section II. Section III offers a road map for achieving data-centric AI. We support our position with arguments in Section IV, where we also wrap up the essay.

II. AN OVERVIEW OF MODEL-CENTRIC AI

Model-Centric Approaches often follow a structured sequence during production that involves scoping the project, data collection and augmentation, data storage, data cleaning, data visualization, visual analytics, model construction and training, and finally, model evaluation. In these situations, the development and improvement of the model's algorithms typically receive the majority of the attention, with data engineering being a one-time assignment. Modern approaches also tend to analyze biases and fairness, focusing on loss functions such as Cross-Entropy, Errors in the mean square and mean absolute percentages, etc. In traditional ML production environments, a stronger emphasis is generally given to the code rather than the resultant trained model to improve its style, readability, and unambiguity. Since the models are to be trained with new data continuously, it necessitates in a broader sense for the distinctions between a data scientist and ML engineer to corrode gradually, with all collaborators developing general expertise on skills and best practices for all the domains involved. Even though model training is essential, the development of AutoML systems has resulted in the progressive decline in the requisite amount of human intervention. The model-centric design, also known as the application-centric or data-driven paradigm depending on the use case, typically necessitates minimal work on the side of the data engineer due to the conventional one-time nature of data acquisition and preprocessing. Suppose, however, data collection and model training was to become a continuous process where subsequent semi-supervised models learn from the same data they collect. In that instance, the workforce would be directed towards constantly labelling and augmentation of collected data, and a movement towards Data-Centric concepts would take place.

III. LIMITATIONS OF MODEL-CENTRIC AI

Model-centric AI is predicated on machine learning approaches that prioritize improving model architectures (algorithm/code) as well as the underlying hyper-parameters. In this method, data is created essentially just once and is maintained throughout the development of the AI system. Model-centric AI has been successful over the past few decades, yet it has an unfixable flaw. It particularly thrives in organizations and sectors when there are customer platforms with millions of users are free to depend on generic fixes. In these conditions, the majority of consumers would be satisfied by a single AI system, nonetheless, outliers would be practically useless. Examples of such organizations and sectors include the advertising sector, where firms like Google, Baidu, Amazon, and Facebook have access to vast amounts of data (sometimes in a standardized format), which they may use to build model-centric AI systems. Standardized solutions like those offered by a single AI system cannot be used in sectors like manufacturing, agriculture, or healthcare where customized solutions are preferred versus one-size-fits-all recipes. Instead, they should conceive their strategy to ensure that their model's (algorithm's) model learns what it needs to learn from having complete data that includes all crucial cases and is labelled consistently.

Due to the rapid pace of innovation in today's technology-driven world, AI models and the algorithms they are modelled

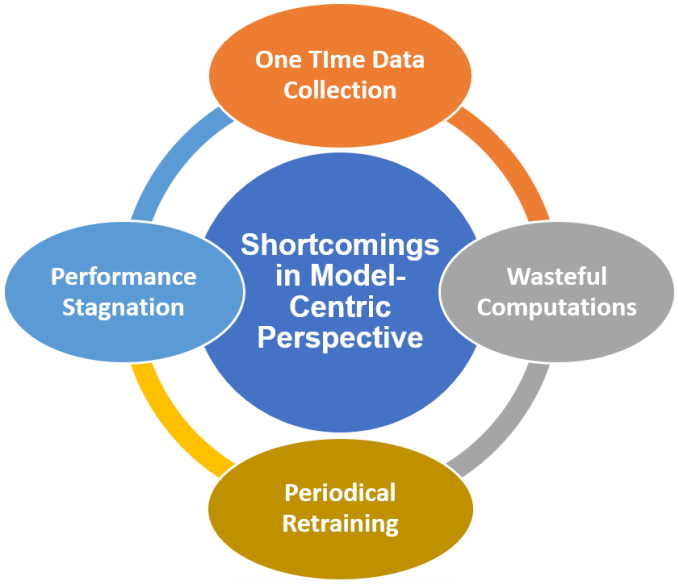

Fig. 3: Shortcomings in Model-Centric Systems

after may quickly become outdated and require retraining on new data, primarily because significant trends' unforeseen or unexpected behaviour may outpace the model's usefulness and relevance. Such model-centric AI's capabilities are severely constrained in fields where the model itself is created to support ongoing scientific study or as a component of a more complex model. As a result, training data is produced using a smaller model based on the restricted findings and hypotheses. Typically, data collection is not practicable in these situations, or work involves limited data from expensive datasets. In such cases, the shortcomings of Model-Centric technologies become apparent, since training a model-centred algorithm may result in biased findings, mainly if the initial model used to generate the data was built on incomplete knowledge or under a lack of domain expertise. With general trends in Deep-Learning requiring focus on large amounts of data, there is a general tendency to prefer reusing existing models to fit into our use cases with specific training data. Such a shift is perceived as a step towards the Data-Centric movement, where the model is fixed, with the only variable inputs being the training data and classifiers.

IV. AN OVERVIEW OF DATA-CENTRIC AI

Introduced to the mainstream by the co-founder of DeepLearning.AI and adjunct Professor at Stanford University, Andrew Ng, on March 24, 2021, data-centric AI is a data-engineering strategy that tries to improve an AI system's performance by methodically boosting the quality of the data used to train the underlying model, under mutually exclusive yet collectively exhaustive methodologies and categories. Data-centric AI can improve the performance of AI models and services through augmentation, extrapolation, and interpolation. Data-centric AI can assist in making AI services more accurate and dependable by expanding the data that is accessible to them and enabling them to use it more efficiently. With the help of training data from many sources, including synthetic data, public data sets, and private data sets, data-centric AI is created utilizing this innovative methodology. This strategy can lessen the time and effort needed to generate training data while also helping to increase the quality of the data. It can also help increase the effectiveness with which training data is used by AI services. Additionally, data-centric AI can process additional data sets because the data is personalized. This means that regardless of the magnitude of the data set, data-centric AI can analyze and learn from it and make decent predictions.

Furthermore, data-centric AI is not limited to a specific type of data. It can learn from text, images, audio, and video. In general, a data-centric AI strategy comprises using the appropriate labels and fixing any problems, getting rid of noisy data inconsistencies, data augmentation, feature engineering, analysis of errors employing domain experts to identify the accuracy or inaccuracy in data points, etc. Data-centric AI constantly evaluates the created AI model in conjunction with updated data. A developed AI model would typically only train on a data set once during the production stage before the software development process can be completed with the deployment of the model with required functionality. The underlying model will eventually come across edge-case instances of data points that are entirely different from those encountered during the training phase. This behaviour is expected as the workflow of data-centric AI is assumed to include the successive improvement of data, in particular, in businesses and industries that cannot afford to have enormous data points (e.g., manufacturing, agriculture, and healthcare). As a result, evaluating the model's quality would also happen more regularly than only once. A model would be able to recognize, judge and then answer back appropriately to variational data distributions owing to the production systems' capacity to offer rapid feedback. In fact, this ability gives data-centric approaches a competitive advantage over their model-centric counterparts.

The ever-arising challenges and problems in today's world require continuous optimization and tuning of the model, along with simultaneous collection, processing, augmentation, and labelling of high-volume data. In cases of filter-list-based blocking/ moderation of social media content, restriction of cyber-threats, fraud detection, and spam tracking, a shift from a data-driven or application-centric to a data-centric perspective focusing on labelling and cleaning of data is apparent, with the information being collected in high frequency, at an hourly basis or even faster. The boundaries between business and technology are vanishing, with tools and techniques like ML and DL requiring assistance from domain experts and consultants to modify inputs or generate better algorithms. DL or Deep-Learning based approaches have become popular owing to the enormous scope and capability for collecting, storing, and processing Big Data, mainly due to their excellent performance with big data and technological advances.

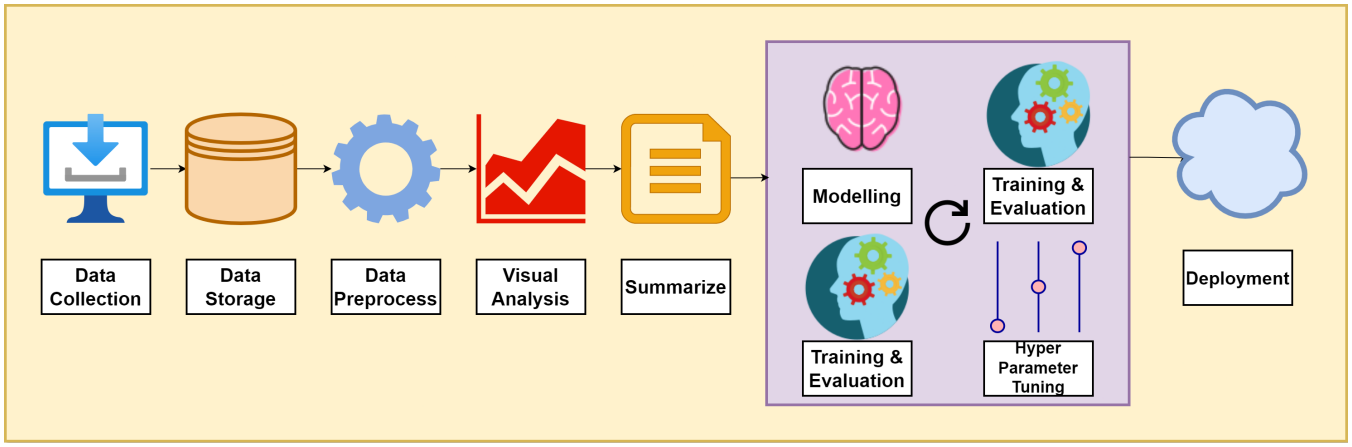

Fig. 4: Production Workflow in Model-Centric Systems

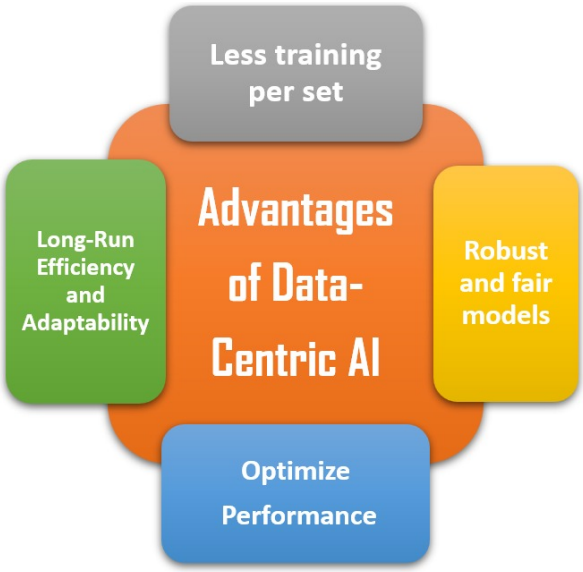

Fig. 5: Advantages in Data-Centric Systems

In Deep-Learning based applications, the distinguishing hierarchy and structure of features or parameters are learned from the data. At the same time, they are usually coded by a human domain expert in typical machine learning applications. Then, through algorithms such as gradient descent and backpropagation, the deep learning algorithm learns and fits itself for accuracy. This methodology allows for mimicking the human brain at a primitive level, allowing DL models to make predictions more precisely through a combination of weights, inputs, and biases. Since a massive volume of data is processed through multiple layers of neural networks, the aspect of clean, labelled information becomes vital, as the presence of dirty, repeated, inconsistent data can cause unnatural biases, failure in edge cases, erroneous predictions, the poor performance of the model, wasteful computations, etc. Most real-world datasets are noisy, unstructured, and unorganized, with several biases, outliers, missing values, repeated values, etc.

To address this issue from a data-centric standpoint when developing their AI systems, businesses and industries need to pay more attention to guaranteeing a consistent collection of high-quality data sets than treating this vital product development phase as a one-time task.

1) Limited Data: The crucial issue is the need for large data sets that are both type-intensive and quantity-comprehensive. Contrary to Internet corporations (like Google and Baidu), the amount of data that manufacturing industries have access to is frequently constrained. Typically, they use data sets comprising 102–103 pertinent data points to train their models. As a result, when using approaches designed for hundreds of millions of data points, a model trained on no more than 103 relevant cases to detect some flaw (or a rare disease) will struggle.

2) Solution Customization: Highly individualized solutions are also necessary. Think about a manufacturing company that sells a variety of goods. Since each manufactured product would require a uniquely trained ML system. The idea of a single universal AI system for flaw identification and classification across every item would not be effective.

V. CHALLENGES FACED BY DATA-CENTRIC AI TECHNOLOGIES

Although data-centric AI appears to be a perfect replacement for model-centric perspectives, several limitations exist, which necessitates that we use both views in tandem to employ better data-engineering practices and yet continue to optimize and tune the model to fit, following conventional methodologies. Data-Centric production faces challenges such as maintaining consistency, adequate volume retention after cleaning, quality of maintenance, the necessity of a proper data versioning system, etc. Amongst the various issues, one of the most significant is the loss of volume associated with cleaning and validating data. An AI model can only receive a massive amount of high-quality data if low-quality datasets are removed.

Therefore, a data-centric method frequently needs a more significant data volume than a model-centric one. This brings potential issues where cleaning might decrease the insufficient amount of collected data under technological or monetary constraints; for example, in scientific research, the model might be continuously tuned to generate and work on experimental data. Working in a constantly evolving setting can be challenging, primarily due to the changing algorithms and margins for errors. In many cases, rules-based algorithms result in a high rejection rate because the technology needs to differentiate between authentic defective components and acceptable levels of variation, especially in manufacturing and production-based industries. This forces a large percentage of human follow-up inspection, which raises costs and slows down production lines.

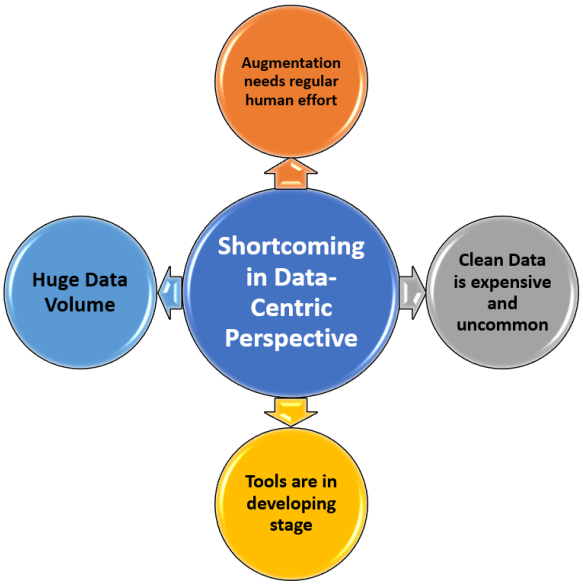

Fig. 6: Shortcomings in Data-Centric Systems

Robustness is the model's ability to preserve performance with a small amount of noise in the data, such that the training error rates are consistent with testing error rates. Fairness refers to the model's tendency to defend against unnatural biases such that the model does not inadvertently introduce any biases. In the case of data quality issues, a focus is given to robust training algorithms when validation of noisy data is insufficient and/or to fairness when appropriate pre-processing levels are insufficient to remove biases from data. An AI system should be trained on the same data type it will test and analyze, including any edge-case variances. Additionally, as part of quality control methods, properties of data records that are not causative features should be randomized during training. An AI model quickly loses accuracy without consistent data annotation. Unfortunately, it can be challenging to maintain a high level of consistency. However, this brings forth a significant challenge to data-centric AI, as it requires human annotation that is costlier than machine computation, especially in environments that emphasize maximum automation.

VI. DATA-CENTRIC AI - EXECUTION AND ANALYSIS

Data preparation is a notable example of tedious step of machine learning(ML) lifecycle. Since data quality directly affects a model's quality, it is also one of the most crucial processes. This section will discuss the significance of exploratory data analysis (EDA), data visualization, and other tools for preparing data for machine learning (ML) pipelines and identifying data quality problems. Data scientists examine and glean essential insights from the data using EDA techniques. Additionally, efficient EDA dramatically benefits from the talents and subject expertise of data scientists in this area. To encourage more statisticians, particularly academics, to research a wide range of fascinating difficulties,We present the traditional yet current subject of data quality from a statistical perspective. The data quality landscape is discussed along with the research underpinnings in computer science, overall quality management, and statistics.

The use of two case studies based on an EDA approach to data quality motivates a collection of research questions for statistics that cover theory, methodology, and software tools. Data visualization is a crucial EDA approach that uses visual elements like charts and graphs to make analysis simple and efficient. When it comes to data quality profiling, visual EDA is very pertinent. With visual features like charts and graphs, data visualization is a crucial EDA technique that simplifies and streamlines analysis.Visual EDA is especially pertinent in the context of data quality profiling. To investigate and summarise multiple data sets, data scientists use exploratory data analysis (EDA), which typically employs data visualisation tools. By figuring out how to alter data sources to best obtain the required answers, it makes it easier for data scientists to detect trends, spot anomalies, test hypotheses, or validate assumptions. EDA aids in comprehending the variables used in data collecting and how they relate. Typically, it is used to look into what information the data might reveal outside of the formal modelling or hypothesis testing assignment. It can also assist you in determining the suitability of the statistical methods you're considering using for data analysis. John Tukey, an American mathematician, developed EDA methods, which are still extensively used in the data discovery process. EDA's main objective is to help with data analysis before making any assumptions. It can help with identifying obvious errors, better comprehending data patterns, identifying outliers or unexpected events, and identifying fascinating relationships between the variables. Data scientists can make sure their findings are accurate and pertinent to any targeted business objectives by using exploratory analysis. EDA assists stakeholders by making sure they are asking the right questions. EDA can help in answering inquiries about standard deviations, categorical variables, and confidence intervals.

After it is finished and new insights have been discovered, EDA can be utilised for more sophisticated data analysis or

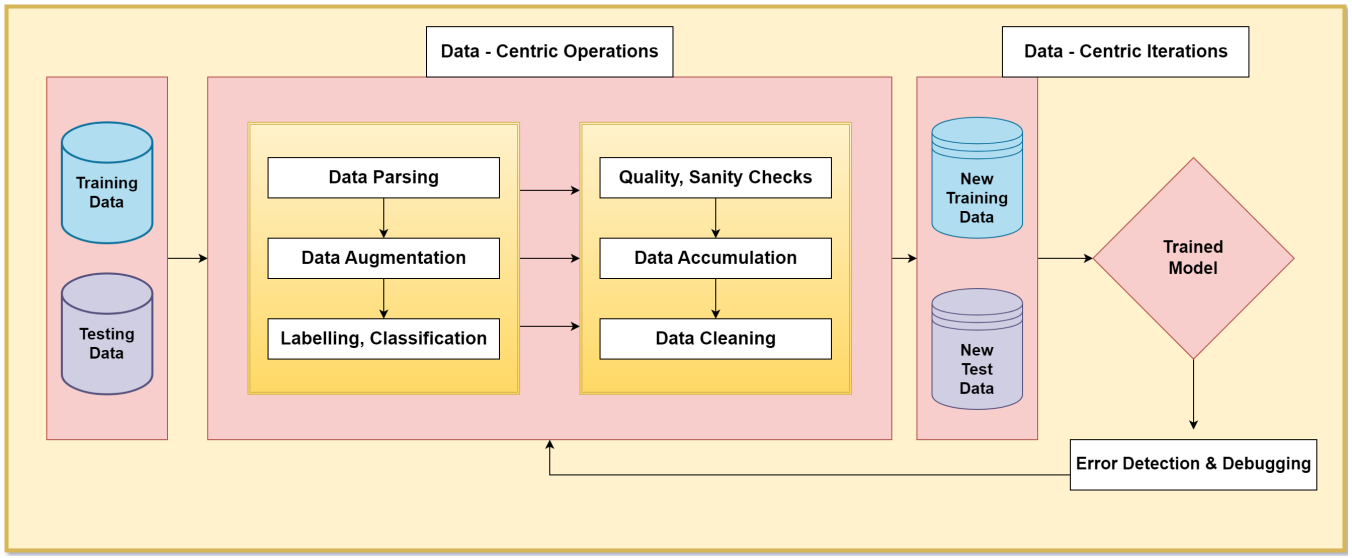

Fig. 7: Production Workflow in Data-Centric Systems

modelling, including machine learning. Consider data to be a shared resource, Applications are built on its framework.But when understood, processed, and stored within the application scope, the same data is handled differently in terms of description, access, and protection. Reusability will be made possible and is a crucial step in the transformation when a shared asset model for data. The data architecture complexicity rises due to data duplication brought on by data silos and an application-centric perspective, notably in data protection and provenance issues. A flexible, consistent data model is required to describe data. The system must have a shared understanding to share, interpret, and process data.

Reusability is restricted by the enormous number of ideas an engineer must be familiar with and comprehend if each application retains its data model. Data redundancy is increased by limited reusability, which causes more complex data governance. The total system complexity and the number of data concepts are directly correlated. A smooth flow can be achieved by maintaining the appropriate interfaces for data delivery. Interfaces that are well-defined and simple to use are critical for reuse and an essential technological asset for data management. There should be a set of guidelines and rules for data governance that are implemented. So, in the ideal case scenario, the technical implementation should be intimately related to the governance process. It is necessary to create and implement some guidelines for data maintenance.

The likelihood of reuse on a bigger scale increases with the degree of openness of the unified data model. Since brand-specific vehicle software architectures are common in the automobile industry, Original Equipment Manufacturers (OEMs) have been developing them to satisfy the requirements of a single brand or company. This sense has countless permutations of components, including transit buses and procedures. However, the data architecture itself is one area that might be changed without harming brand-specific solutions. Data is now the centre of digital transformation and the most crucial component of architecture. But this only works if everyone is aware of the information. To guarantee adaptability, scalability, and a decoupled infrastructure, standards must be established. Additionally, as more OEMs and members of the community get involved in the standards, the ongoing work to fix possible problems with the data models will result in a major improvement in data accuracy.

The adoption of a data-centric approach can be achieved by combining several steps. For AI systems to be able tolearn from the smaller data sets that are readily available in most businesses, research teams should first focus on assuring good quality data. This involves ensuring that the data collected adequately shows the ideas we want the AI to learn. Corporate domain professionals should handle data engineering together with AI specialists. By doing this, AI is made simpler and, as a result, more available to various sectors. Rather than investing time and energy in creating software, teams might use ML Ops platforms, with much of the scaffolding software required to make an AI system easier to produce. As a result, the time between proof of concept and production will drastically shorten to just a few weeks instead of taking years.

VII. DEVELOPMENTS IN RECENT TIMES

We have dealt with the unparalleled significance of Data-Centric AI over Model-Centric AI. In recent times, industries are witnessing some prominent shift in their core architecture and becoming more data oriented. Some firms are bringing more data-centric approach towards solving their daily tasks. A new division has developed a number of tools, including CleanLab, LandingLens, Snorkel, AutoAugment, HoloClean, Albumentations, etc., to support various production processes in response to identifying and resolving the stepwise challenges encountered in the current workflow methodologies, where data engineering and model engineering are split into

separate cycles. Such an advancement has been made possible by structured thinking and the division of tasks into mutually exclusive yet cumulatively exhaustive parts. Various benchmarks, such as dcbench, ImageNet, and MLPerf, have been developed to assess and validate the effectiveness and quality of said tools based on several parameters, which may be generalised based on training data size, budget-restricted data-cleaning, object-detection, computer-vision, etc.

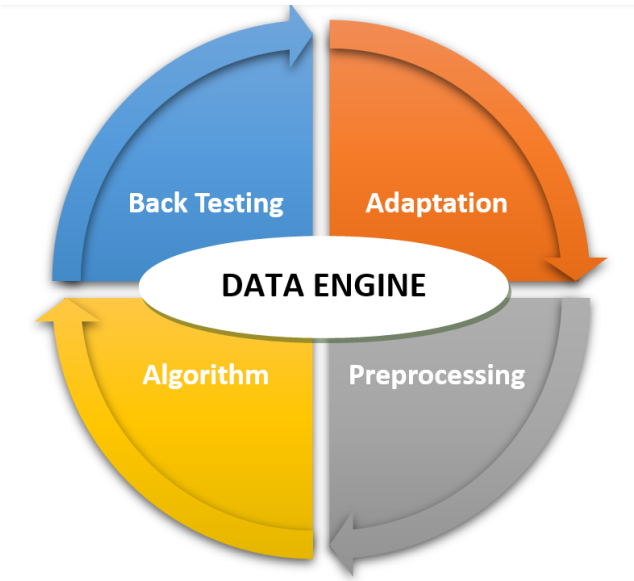

Fig. 8: Production Workflow in RoboFlow

The pipelines of creating AI-enhanced robots are being managed by RoboFlow, a cloud-based workflow management system founded in January 2020. RoboFlow is data-centric, in contrast to the majority of conventional robotic development techniques, which are largely process-centric. This remarkable trait makes it particularly ideal for creating data-driven AI-enhanced robots. RoboFlow breaks down the entire robotic development process into 4 building blocks, starting with data processing and moving on through algorithm development, back testing, and application adaptation.

A very broad interdisciplinary, Materials Science (MS), discipline that targets to detect and form new materials, is particularly affected by the transition from theory to practice—predicting structural features and discovering novel materials like perovskites, nanoparticles, and nanoclusters. By using Machine Learning, we can obtain new types of interatomic potentials—specifically, the ML potentials applied to a wide range of material systems with accuracy equivalent to that of first-principles calculations—has been a particularly active topic. Despite the advent of computing methods for modelling quantum mechanics in the 1970s, calculations were still constrained by the quantity of atoms and electrons, which results in ever longer run-times. Modern computational methods to electronic structures are based on solving the Kohn-Sham (KS) equation, which has made important advancements in the materials sciences. The KS total energy of anatase TiO2 nanoparticles (NPs) at various temperatures may be predicted quickly and correctly using only a little amount of theoretical data utilising a novel, data-centric machine learning methodology that we have mentioned earlier. Therefore, a shift to a more data-oriented approach improved the accuracy of the computations indicated above.

VIII. CONCLUSION

It would initially seem more sensible to think about changing the model (algorithm/code) rather than the data to increase the system's performance. The necessity of using data-centric techniques, however, becomes relevant under the limitations of model-centric AI solutions as discussed. Model-centric classification algorithms employ regularisation techniques to adjust the algorithm or code in order to enhance model performance. However, for the majority of real-time applications to work well, reliable data is essential. However, we believe that this should take into account the underlying model to the same extent. As a result, the model-centric and data-centric approaches should be viewed as two sides of the same coin, and the model-centric approach's drawbacks in particular organisations and industries shouldn't lead to its complete abandonment. This is the case for a number of pragmatic and obvious reasons.

First, our intuitive knowledge of problem solving needs to place a lot of focus on how we act upon things, in addition to knowing items' properties and information about them. In the context of AI development, this would suggest a preference for spending more time designing and optimising intelligent algorithms and more precisely changing the hyper-parameters of the underlying model as well as the code required to implement the designed algorithms. Fixing the data set is extremely important at this stage of the project since only with a defined data set can models be compared and grouped according to performance.

Second, starting the design of AI systems with a model-centric approach offers plenty of opportunity to amass experience in comprehending real-world problems and potential computational solutions, as opposed to the experience gains that could have been obtained had research and industry adopted the data-centric approach. In fact, when attempting to solve a problem, it is often in our nature to start by acting in order to make an impact before going further to understand the properties of the items in our environment. Reinforcement learning, a recently well-publicized machine learning (ML) technique, is based on the idea that intelligent agents learn through interacting with their environments.

Third, it would be difficult to connect the final models with the real-world problems since the produced models would have undergone a completely different development process. In the early stages of AI development, academia and industry would have favoured the data-centric strategy to its counterpart model-centric method.

ACKNOWLEDGEMENTS

It is only fair that we are thankful to our CSO211 course professor Dr. Sanjay Kumar Singh, and our Teaching Assistants for the valuable insights and guidance they provided us with. This paper would not have been possible without their mentorship and support.

REFERENCES

[1] Karan Goel, "Our Journey towards Data–Centric AI: A Retrospective", Available: http://ai.stanford.edu/blog/data-centric-ai-retrospective/ ,15thSeptember2021.

[2] Christopher Ré, Feng Niu, Pallavi Gudipati, Charles Srisuwananukorn, "Overton: A Data System for Monitoring and Improving Machine-Learned Products", proc. Of 10th Annual Conference on Innovative Data Systems Research (CIDR), 12th – 15th January 2020.

[3] Nithya Sambasivan et. al., "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI", proc. of CHI Conference on Human Factors in Computing Systems (CHI 2021), 8th – 13th, 2021. Available: https://doi.org/10.1145/3411764.3445518

[4] Andrew Ng, "MLOps: From Model – centric to Data – centric AI", 35th Conference on Neural Information Systems (NeurIPS), 14th December 2021.

[5] Andrew Ng, "Data-centric AI Development: A Critical Shift in Perspective", Available: https://read.deeplearning.ai/thebatch/data-centric-ai-development-part-2-a-critical-shift-inperspective/,2021.

[6] Fabiana Clemente, "From Model-centric to Data-centric: A new paradigm for AI Development", Available: https://towardsdatascience.com/from-model-centric-to-datacentric-4beb8ef50475,30thMarch2021.

[7] A. Ng. (2021) A chat with andrew on mlops: From modelcentric to data-centric ai. DeepLearningAI. [Online]. Available: https://www.youtube.com/watch?v=06-AZXmwHjot=1607s

[8] D. Berscheid. (2021) "Data-centric machine learning: Making customized ml solutions production-ready." Dida.Do. [Online]. Available: https://dida.do/blog/data-centric-machine-learning

[9] A. Ng. (2021) "AI doesn't have to be too complicated or expensive for your business." Harvard Business Review. [Online]. Available: https://hbr.org/2021/07/ai-doesnt-have-to-be-too-complicatedor-expensive-for-your-business

[10] R. S. Sutton and A. G. Barto, Reinforcement learning: An introduction. MIT press, 2018.

[11] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton et al., "Mastering the game of go without human knowledge," nature, vol. 550, no. 7676, pp. 354–359, 2017.

[12] O. H. Hamid and J. Braun, "Reinforcement learning and attractor neural network models of associative learning," in International Joint Conference on Computational Intelligence. Springer, 2019, pp. 327– 349.

[13] A. Nair, B. McGrew, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Overcoming exploration in reinforcement learning with demonstrations," in 2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2018, pp. 6292–6299.

[14] T. M. Moerland, J. Broekens, and C. M. Jonker, "Emotion in reinforcement learning agents and robots: a survey," Machine Learning, vol. 107, no. 2, pp. 443–480, 2018

[15] Neoklis Polyzotis, Matei Zaharia, "What can Data-Centric AI Learn from Data and ML Engineering?," arXiv preprint arXiv:2112.06439 (2021). Available: https://arxiv.org/pdf/2112.06439.pdf

[16] Steven Euijong Whang, Yuji Roh, Hwanjun Song, Jae-Gil Lee, "Data Collection and Quality Challenges in Deep Learning: A Data-Centric AI Perspective," arXiv preprint arXiv:2112.06409 (2021). Available: https://arxiv.org/pdf/2112.06409.pdf

[17] Mustafa Hajij, Ghada Zamzmi, Karthikeyan Natesan Ramamurthy, and Aldo Guzman, "Data-centric AI Requires Rethinking Data Notion," arXiv preprint arXiv:2110.02491 (2021). Available: https://arxiv.org/pdf/2110.02491.pdf

[18] Shen Wang, Mehdi Nikfar, Joshua C. Agar, and Yaling Liu, "Stacked Generative Machine Learning Models for Fast Approximations of Steady-State Navier-Stokes Equations," arXiv preprint arXiv:2112.06419 (2021). Available: https://arxiv.org/pdf/2112.06419.pdf

[19] Sabri Eyuboglu, Bojan Karlaš, Christopher Ré, Ce Zhang, and James Zou.2022. "dcbench: A Benchmark for Data-Centric AI Systems,". Available: https://dl.acm.org/doi/pdf/10.1145/3533028.3533310

[20] Mark Mazumder, Colby Banbury, Xiaozhe Yao. 20 July 2022. "Dataperf:Benchmarks for Data-Centric AI Development.". Available: https://arxiv.org/pdf/2207.10062.pdf.

[21] Hima Patel, Shanmukha Guttula, Ruhi Sharma Mittal, Naresh Manwani, Laure Berti-Equille, and Abhijit Manatkar. 2022. "Advances in Exploratory Data Analysis, Visualisation and Quality for Data Centric AI Systems,". Available: https://dl.acm.org/doi/10.1145/3534678.3542604

[22] LJ MIRANDA, Jul 30, 2021, "Towards Data-centric Machine Learning,". Available: https://ljvmiranda921.github.io/notebook/2021/07/30/data-centric-ml/.

[23] Dr. Chetana Hegde, "Anomaly Detection In Time Series Data Using Data- Centric Ai", 2022 IEEE International Conference on Electronics, Computing and Communication Technologie, Available: https://ieeexplore.ieee.org/document/9865824

[24] Oussama H. Hamid, "From Model-centric To Data-centric Ai: A Paradigm Shift Or Rather A Complementary Approach", 2022 8th International Conference on Information Technology Trends (ITT), Available: https://ieeexplore.ieee.org/document/9863935

[25] Irena Atova, Kwang-Cheng Chen, Ahmed E.Kamal, Malamati Louta, "Data Science And Artificial Intelligence", June 2020 IEEE Communications Magazine, Available: https://www.academia.edu/49186370/Data_Science_and_Artificial_ Intelligence

[26] Yiqi Zhong, Lei Wu, Xianming Liu, Junjun Jiang,"Exploiting The Potential Of Datasets: A Data-centric Approach For Model Robustness", 10 Mar 2022, Available: https://arxiv.org/abs/2203.05323

[27] Guo Ye, Jiayi Wang, Han Liu, 2021:"RoboFlow: a Data-centric Workflow Management System for Developing AI-enhanced Robots" Blue Sky Papers, 5th Conference on Robot Learning (CoRL 2021), London, UK. Available: https://proceedings.mlr.press/v164/lin22c/lin22c.pdf

[28] Anne Gerdes (2022) "A participatory data-centric approach to AI Ethics by Design", Applied Artificial Intelligence, 36:1, 2009222, Available: https://www.tandfonline.com/doi/pdf/10.1080/08839514.2021.2009222?needAccess=true

[29] Huan Hu, Yajie Cui, Zhaoxiang Liu, Shiguo Lian "Data-Centric AI Paradigm Based on Application-Driven Fine-Grained Dataset Design" Available: https://arxiv.org/ftp/arxiv/papers/2209/2209.09449.pdf

[30] Lora Aroyo, Matthew Lease, Praveen Paritosh, Mike Schaekermann "Data Excellence for AI: Why Should You Care" Available: https://arxiv.org/ftp/arxiv/papers/2111/2111.10391.pdf

[31] Mohammad Motamedi, Nikolay Sakharnykh, Tim Kaldewey, 29 Oct 2021: "A Data-Centric Approach for Training Deep Neural Networks with Less Data" Available: https://arxiv.org/pdf/2110.03613.pdf

[32] Daniel Alvarez-Coello, Daniel Wilms, Adnan Bekan, Jorge Marx Gómez. 2021: "Towards a Data-Centric Architecture in the Automotive Industry", Elsevier B.V. Available: https://reader.elsevier.com/reader/sd/pii/S1877050921002581

[33] J. Zico Kolter , 2019: "Provably robust deep learning" Available: https://www.cs.cmu.edu/~cliu6/16-883/robust_deep_learning.pdf

[34] Hasan Kurban, Mustafa Kurban, Mehmet M. Dalkilic, 24 August 2022: "Rapidly predicting Kohn–Sham total energy using data centric AI", Scientifc Reports Available: https://www.nature.com/articles/s41598-022-18366-7

[35] Nabeel Seedat, Fergus Imrie, Mihaela van der Schaar, 09 Nov 2022: "DC-Check: A Data-Centric AI checklist to guide the development of reliable machine learning systems" Available: https://arxiv.org/pdf/2211.05764.pdf