Bus Inf Syst Eng 66(4):507–515 (2024)
https://doi.org/10.1007/s12599-024-00857-8

CATCHWORD

Data-Centric Artificial Intelligence

Johannes Jakubik · Michael Vössing · Niklas Kühl · Jannis Walk · Gerhard Satzger

Published online: 5 March 2024

Abstract Data-centric artificial intelligence (data-centric AI) represents an emerging paradigm that emphasizes the importance of enhancing data systematically and at scale to build effective and efficient AI-based systems. The novel paradigm complements recent model-centric AI, which focuses on improving the performance of AI-based systems based on changes in the model using a fixed set of data. The objective of this article is to introduce practitioners and researchers from the field of Business and Information Systems Engineering (BISE) to data-centric AI. The paper defines relevant terms, provides key characteristics to contrast the paradigm of data-centric AI with the model-centric one, and introduces a framework to illustrate the different dimensions of data-centric AI. In addition, an overview of available tools for data-centric AI is presented and this novel paradigm is differenciated from related concepts. Finally, the paper discusses the longer-term implications of data-centric AI for the BISE community.

Keywords Data-centric artificial intelligence · Data quality · Data work

Accepted after 2 revisions by Christine Legner.

J. Jakubik (✉) · M. Vössing · J. Walk · G. Satzger
Karlsruhe Institute of Technology, Kaiserstraße 12,
76131 Karlsruhe, Germany
e-mail: johannes.jakubik@kit.edu

N. Kühl
University of Bayreuth, Universitätsstraße 30, 95447 Bayreuth,
Germany

1 Introduction

Over the past decades, researchers and practitioners in artificial intelligence (AI) have focused on improving ML models in AI-based systems (model-centric AI paradigm). However, the provision and selection of suitable data also impact model effectiveness (e.g., performance) and efficiency (e.g., costs for labeling or for training computation). Despite a long history of research on data (Legner et al. 2020; Otto 2011; Zhang et al. 2019), the impact of data quantity and quality on AI-based systems is still often overlooked in both AI research (Parmiggiani et al. 2022) and AI practice (Sambasivan et al. 2021). Propagated by Andrew Ng and promoted in a series of workshops (Ng et al. 2021, 2022), data-centric AI emphasizes the development and application of methods, tools, and best practices for systematically designing datasets and for engineering data quality and quantity to improve the performance of AI-based systems (Strickland 2022). In particular, the new paradigm is not calling for simply acquiring more data but more appropriate data. While many facets of data-centric AI have previously been studied independently, this paradigm unites researchers from different fields (e.g., machine learning and data science, data engineering, and information systems) with the goal of improving machine learning approaches in real-world settings. This has far-reaching implications for the way AI-based systems are developed.

The objective of this article is to introduce practitioners and researchers in Business and Information System Engineering to data-centric AI as a complementary and mutually beneficial paradigm to model-centric AI. We define relevant terms, introduce key characteristics to contrast both paradigms, and introduce a framework for data-centric AI. We distinguish data-centric AI from

related concepts and, in particular, discuss potential contributions of and implications for the BISE community.

2 Model-Centric and Data-Centric AI

In previous years, research on ML has mainly focused on the development of model types, architectures, and the definition of suitable hyperparameters to improve performance. For example, the ML community often benchmarks different ML approaches based on fixed datasets – both in practical competitions (Kaggle 2023) as well as in academic research (e.g., Ronneberger et al. 2015). Utilizing publicly available benchmark datasets allows for valuable and scientific sound comparisons across approaches and has facilitated a significant acceleration in the performance of ML models. In addition, these benchmark datasets can be employed to ensure the reproducibility of proposed models. Overall, this led to an increasing maturity of model types, architectures, and hyperparameter selection.

Definition ***Model-Centric Artificial Intelligence*** is the paradigm focusing on the choice of the suitable model type, architecture, and hyperparameters from a wide range of possibilities for building effective and efficient AI-based systems.

However, in recent years, this strategy (i.e., solely optimizing models) has plateaued for many datasets with regard to the model performance. Similarly, with regard to real-world datasets, a focus on improving (complex) models does not necessarily lead to significant performance increases (e.g., Baesens et al. 2021). Furthermore, practitioners often want to use ML to solve unique problems for which neither public datasets nor suitable pre-trained models are available. For this reason, the focus of practitioners and researchers has gradually been shifting towards data, the second, somewhat neglected ingredient for the development of AI-based systems. In particular, researchers and practitioners recognize the need for more systematic data work as a means to improve the data used to train ML models. In fact, data is a crucial lever for an ML model to generate knowledge (Gröger 2021). Consequently, data quantity (e.g., the number of instances) and data quality (e.g., data relevance and label quality) largely influence the performance of AI-based systems (Gudivada et al. 2017). Data-centric artificial intelligence (data-centric AI) represents a paradigm that reflects this.

Definition ***Data-Centric Artificial Intelligence*** is the paradigm emphasizing that the systematic design and engineering of data are essential for building effective and efficient AI-based systems.

Data-centric AI differs from model-centric AI in terms of the general focus, the importance of domain knowledge, and the understanding of data quality:

- *Focus* Data-centric AI generally holds the ML model fixed instead of the dataset. Performance increases are achieved by improving the quality and quantity of the data given a fixed model.
- *Data Work and Domain Knowledge* Domain-specific data work is an integral component of data-centric AI. Data work is supplemented by the development of methods and semi-automated tools to accelerate the development of successful AI-based systems.
- *Perspective on Data Quality* Data-centric AI generates performance improvements based on more appropriate data. This implies that changes in ML model performance metrics also indicate the effectiveness of adjustments in the data. This results in a novel perspective on data quality that can be approximated by changes in metrics from the field of machine learning.

Despite these differences between model-centric and data-centric AI, the two paradigms are inherently complementary, as the development of AI-based systems should ultimately incorporate both paradigms. A high-level overview depicting this relationship is displayed in Fig. 1.

While the data-centric paradigm emerges from the ML community – and most academic endeavors dealing with it do focus on machine learning – , the term "data-centric AI" has also intruded the computer science and BISE communities. However, in fact, data-centric *machine learning* might have been a more appropriate term (Kühl et al. 2022). ML research generally focuses on designing methods that leverage data to increase the performance on a range of tasks (i.e., learn) with computational resources (Alpaydin 2020). Artificial intelligence, in contrast,

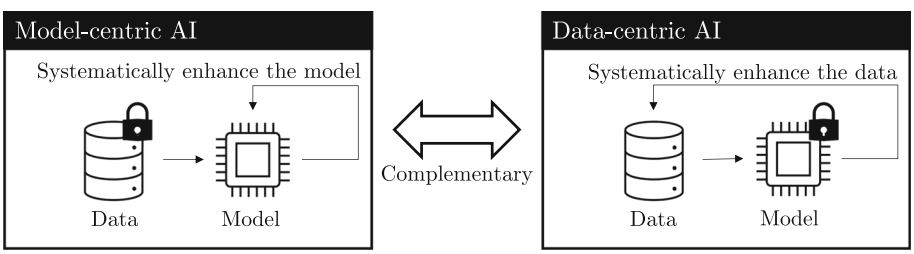

Fig. 1 Data-centric AI as an emerging, complementary paradigm for the development of AI-based systems

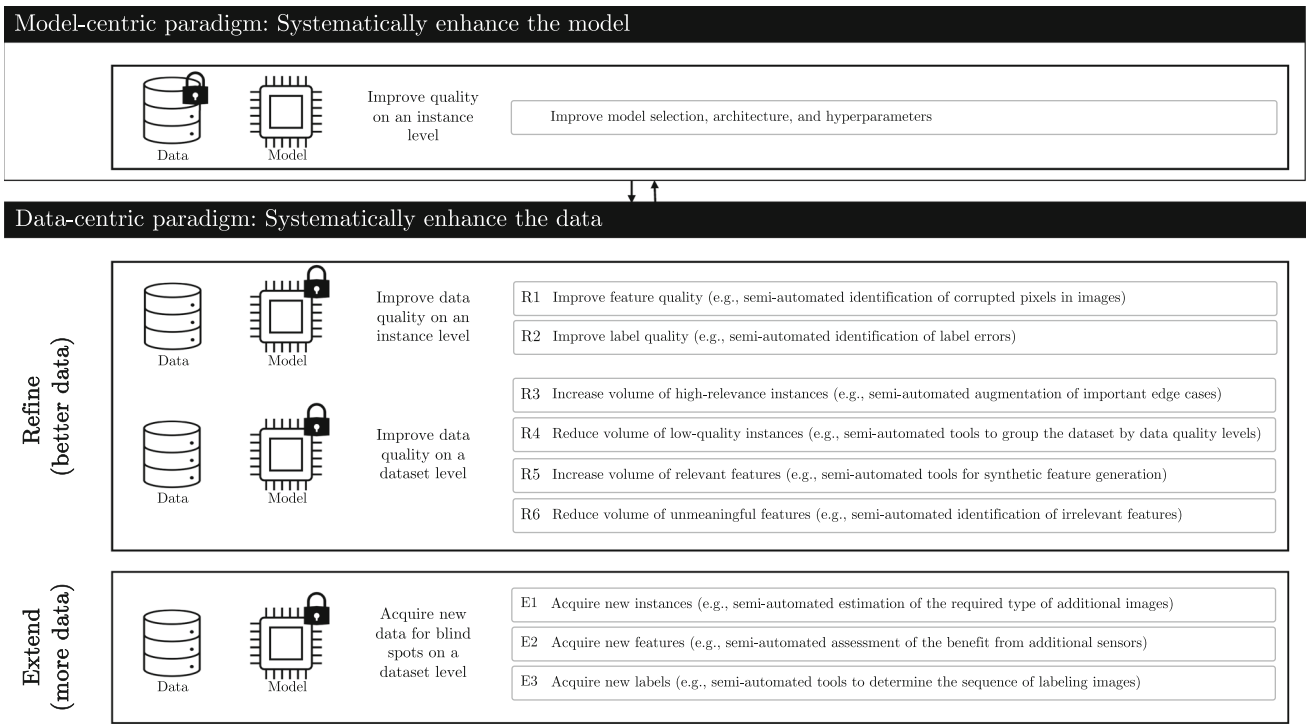

Fig. 2 Framework for the systematic design and engineering of data for data-centric AI (note that we illustrate a high-level representation of model-centric AI only. For more details on model-centric AI, we refer to Alpaydin (2020)

includes ML but also comprises a broader set of methods – e.g., logical programming or probabilistic methods – that allow an agent to interact with its environment.

3 Dimensions of Data-Centric AI

The framework for data-centric AI in Fig. 2 illustrates the different dimensions for the systematic design and engineering of data. While data-centric AI is also applicable to unsupervised (Amrani 2021) or reinforcement (Lin et al. 2022) learning, this summary focuses on supervised ML as the most prevalent real-world application of ML (Jordan and Mitchell 2015). Overall, we identify two major dimensions for data-centric AI – that is, the *refinement* of existing data (i.e., "better data") and the *extension* of this data by acquiring additional data (i.e., "more data").

The *refinement* of data refers to systematically improving the quality of existing data – measured with performance metrics of ML models. First, enhancing the quality of individual instances can be achieved by improving the quality of features or target labels (R1, R2). Regarding the representativeness of the data, the quality of data can be enhanced by increasing the number of high-relevance instances that strongly influence the learning process of ML models (R3). Such underrepresented but relevant instances need to be particularly taken into account for augmentations. Moreover, low-quality instances with, for example, incorrect labels or inaccurate feature values need to be identified and removed from the data (e.g., the semi-automatic identification of label errors in R4; see Northcutt et al. 2021, 2021). Thus, it is essential for data-centric AI to build semi-automated tools to better differentiate outliers (that should be removed from the dataset) from edge cases (that should be augmented to enhance the representativeness of the data). On a feature level, data refinement means increasing the volume of relevant features while excluding unmeaningful or even unfair ones. While these actions to refine datasets have partly been leveraged in the past, this work has mostly been performed manually.

The *extension* of data refers to systematically acquiring additional data for "blind spots" in the dataset. Extending the data becomes necessary when the existing data does not allow to sufficiently address the business problem. Additional data may help to develop an accurate ML model. There are two major motivations for extending the data: First, extending data helps to achieve an initial ML model performance that meets the requirements of the business problem. Second, extending the dataset helps to respond to shifts in the data distribution to maintain this performance over time. Hence, acquiring new data is crucial for both achieving high performance of the model and maintaining the performance of the AI-based system. Overall, we identify three dimensions along which data can be extended: First, new instances may be acquired, whereby each instance represents an observation (E1). Second,

additional, new features for each of the instances may be collected, e.g., by employing additional sensors (E2). Finally, collecting additional data may also refer to the retrieval of target labels for existing or new unlabeled instances (E3). Overall, extending the data by acquiring new features, observations, or target labels (E1–E3) primarily impacts the quality of data based on increasing data quantity (more data). In contrast, refining existing data targets the improvement of data quality based on operation with existing data (better data). Overall, both dimensions of the framework are heavily influenced by major IS topics, including, for example, data governance, data management, and AI governance. We discuss the relation between these topics and data-centric AI in the last section and use our framework to better link the importance of these topics to data-centric AI.

For widespread adoption of data-centric AI in research and practice, methods and appropriate tools are required. While some methods and corresponding measures already exist, e.g., for the identification of special instances in the data, there is a particular void of methods to systematically design and engineer the data. This includes, among others, methods for data versioning (e.g., Biewald 2020), methods that support the labeling process in terms of efficiency and performance (e.g., R2, E3; see Fiedler et al. 2019), methods for data exploration and visualization (e.g., R1, R2; see McInnes et al. 2018), and methods to identify special data instances (e.g., R1–R6; see Northcutt et al. 2021). We provide a summary on data-centric tools, including examples of commercial applications, in Table 1, where meta tools refer to tools that are of general importance for data-centric AI (i.e., across the dimensions of the data-centric AI framework). In general, methods from the field of transfer learning support the development of data-efficient AI-based systems across the dimensions of data-centric AI (R1–R6 and E1–E3) as pretrained models require a reduced amount of high-quality data. From a system's perspective, methods for semi-automated data exploration are particularly important, as such methods may eventually contribute to an enhanced data understanding, which is essential for improving data quality on an instance level (R1, R2), on a dataset level (R3–R6), and for extending the data (E1–E3). Overall, all required methods need to be supported by corresponding tools.

4 Delimitations of Data-Centric AI from Related Concepts

The data-centric AI paradigm relates to a number of concepts that have been studied in the BISE community over the past several decades, in particular Big Data, MLOps, and data-driven methods. In the following, we delineate data-centric AI from these closely related concepts.

The paradigms of *big data* and data-centric AI both focus on gathering more data to improve analytics and predictive tools. While the two fields are closely related, significant differences between the two exist: Big data generally refers to the collection, storage, and processing of large amounts of data (e.g., E1). However, there is less focus on what kind of data is stored (Chen et al. 2012). The general assumption is that more data is always better. In contrast, data-centric AI aims to improve the performance of AI systems by *systematically* acquiring more but also better data or even by removing deficient or irrelevant data (R1–R6). This is especially relevant in specialized domains lacking the option to collect large amounts of data. ML models are particularly sensitive to noise in the data when data volume is small (Baesens et al. 2021). In these cases, systematically designing and engineering datasets is crucial for the adoption and usage of AI. Moreover, data-centric AI includes additional operations on the data, such as the extension of data based on data collection in new contexts.

Machine Learning Operations (MLOps) is another research field closely linked with the data-centric AI paradigm. MLOps is concerned with putting AI projects into production and avoiding a multitude of pitfalls associated with this process. To address this gap, MLOps (Renggli et al. 2021) – oftentimes used interchangeably with Artificial Intelligence Operations (AIOps) – is required. MLOps is an engineering practice dealing with the application of tools, frameworks, and best practices to increase the number of AI projects that are brought to production. While data does play an important role, a major part of the MLOps paradigm focuses on engineering principles like continuous development, orchestration, monitoring, reproducibility, and versioning (Renggli et al. 2021). However, so far, there is very little focus on monitoring and versioning the datasets that would be adapted by methods from the field of data-centric AI (both in R1–R6 and E1–E3). Tracking different versions of data and the corresponding impact on ML models is essential to efficiently progress towards better data to increase the performance of ML models. Therefore, tools, frameworks, and best practices are required to facilitate data work and make modifications to the data an iterative part of AI projects instead of only preprocessing the data initially while iterating the search for the optimal model.

Finally, we discuss the difference between data-centric AI and *data-driven* methods due to the similarity in the terminology. Data-driven methods focus on processing data into information in order to present the derived information to decision-makers. Model-driven methods instead focus on mathematical models like optimization or simulation models. ML models are typically considered

Table 1 Examples of data-centric tools and meta tools categorized by the dimensions of the framework for data-centric AI

Tools purpose	Commercial examples	R1	R2	R3	R4	R5	R6	E1	E2	E3
Label error identification	cleanlab.ai		•							
Labeling support	prodi.gy									•
Synthetic data generation	gretel.ai			•		•		•	•	
Anomaly detection	Microsoft Azure	•			•		•			
Data gathering efficiency	–							•	•	
Edge case identification	iMerit edge case			•				•		
Visual data exploration	tableau.com	•	•	•	•	•	•			
...										
Meta tools										
Data versioning	wandb.ai	•	•	•	•	•	•	•	•	•
Personalized data work	–	•	•	•	•	•	•	•	•	•
Federated data work	lifebit.ai	•	•	•	•	•	•	•	•	•
Data verification	Google cloud	•	•	•	•	•	•	•	•	•
Data quality measurement	precisely.com	•	•	•	•	•	•	•	•	•
...										

both data-driven and model-driven since mathematical models are fed with a large amount of data (e.g., Turban 2011). The distinction between model-centric and data-centric is on a lower level of abstraction, though – it differentiates two paradigms for developing an ML model. This means that the type of method to generate information from data (data-driven vs. model-driven) does not necessarily determine the paradigm to develop the underlying model (data-centric vs. model-centric).

5 Implications for BISE Research

Until now, data-centric AI has largely been explored by researchers from the field of computer science. However, data-centric AI has the potential to fundamentally improve AI-based systems by complementing the paradigm of model-centric AI and thereby offering a more holistic development of AI-based systems. While data-centric AI promises to support the BISE community in the design of more effective information systems, BISE researchers are also well-positioned to advance data-centric AI. Figure 3 provides proposed areas with respect to BISE research on an individual, organizational, and cross-organizational level that are to be detailed in the following section. We group the proposed areas into advancements for dealing with data as such and with their incorporation within AI-based systems.

5.1 Individual Level

Data-centric AI emphasizes leveraging high-quality datasets, which frequently requires to accurately select a relevant, high-quality subset of large-scale datasets. This demonstrates the significance of a nuanced data understanding for data-centric AI. Recent semi-automated methods and tools propose to generate metadata in order to improve the understanding of high-dimensional datasets (Holstein et al. 2023). With a unified set of metadata, such approaches can not only foster an understanding *within* high-dimensional datasets but also comparisons *across* datasets. Advancing data understanding relies on research in information systems, which explores data visualization and interpretability methods to help researchers and practitioners gain insights into complex data patterns (Toreini et al. 2022). Interactive dashboards can enhance data understanding over time by allowing users to explore and analyze data from different perspectives, providing real-time insights (e.g., R1–R6). Building interactive data exploration tools enables users to engage with the data, facilitating ad-hoc analyses and supporting data profiling and data quality assessment. Additionally, information system research focuses on developing data profiling techniques that automatically extract statistical summaries, data distributions, and potential data quality issues (Abedjan et al. 2022), which help identifying missing values, outliers, and inconsistencies (R4) that may impact data-centric AI development. Understanding data in context, particularly within specific domains or industries, further enables data-centric AI solutions to address real-world challenges, while integrating domain-specific knowledge.

Data-centric AI emphasizes the importance of utilizing domain knowledge to refine and extend data towards increasing performances of AI-based systems. Thus, data-centric AI embraces both social and technical aspects by definition; and the underlying data work for data-centric AI, including annotation, curation, and preprocessing, is

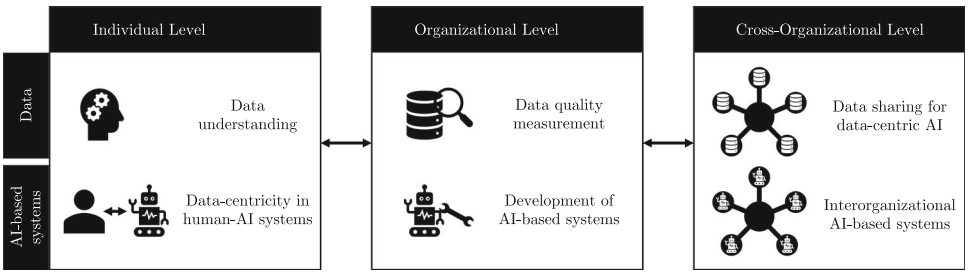

Fig. 3 Proposed areas of BISE research for the advancement of data-centric AI

inherently human-centered (Jarrahi et al. 2023). This is, among other examples, reflected in the utilization of semi-automated tools to improve the data work of human annotators and domain experts. As a consequence, data-centric AI requires one or multiple human(s) in the loop of the AI to guarantee access to domain knowledge. Therefore, the efficiency of data work is mainly driven by efficient and sustainable interactions between AI and domain experts in human-in-the-loop systems (e.g., by balancing the trade-off between the value of additional data for the AI and the time and cost of data work). Efficient interaction is especially important as the number of manual reviews to refine data is typically constrained due to the limited availability of human experts (e.g., physicians; see Holzinger 2016). For example, instead of asking physicians to select and adjust incorrect labels of images depicting specific diseases, semi-automated tools are required that preselect potentially mislabeled images (e.g., as part of R2), potentially important edge cases (e.g., as part of R3), or the most informative data instances for future labeling [e.g., E3, see also active learning (Hemmer et al. 2022)]. A physician can then review this subset in a fraction of the time. The design of semi-automated tools to facilitate data work requires social and technical considerations, opening various avenues for BISE research (e.g., personalized tools for data work, federated data work, etc.). The BISE community has a long history of analyzing the behavior and interactions of humans and AI, which now includes an additional facet in terms of downstream implications of human-in-the-loop systems for the improvement of data work as part of data-centric AI.

5.2 Organizational Level

Monitoring data quality is a critical organizational task in the context of data-centric AI (Schneider et al. 2023). Recent research has demonstrated that the performance of ML models is specifically affected by incomplete data, as well as low feature and low label accuracy (Budach et al. 2022). Monitoring the effect of data completeness, feature accuracy, and label accuracy on ML models during the enhancement of data quality in real-world datasets is essential to better understand promising ways of data enhancement (Aramburu et al. 2023). Ensuring appropriate data quality further requires data verification and validation, especially when dealing with real-time or sensor-generated data (Whang et al. 2023; Abbasi et al. 2016). Thus, there is a necessity for methods and tools to continuously verify and validate data (e.g., R1, R2) and provide feedback to data providers, enabling them to improve data quality. The shift towards the data-centric paradigm further changes the approach to measuring data quality, emphasizing continuous monitoring throughout the iterative data work process (Sambasivan et al. 2021). Quantifying data quality using ML model performance facilitates the assessment of the impact of data modifications on AI system performance across this process. Data-centric AI will benefit from a refinement in the understanding of data quality for AI-based systems and from the development of tools to continuously measure data quality in AI projects. Additionally, guidance from BISE researchers is needed to investigate innovative ways to seamlessly integrate diverse, heterogeneous data sources (e.g., E2), overcoming the limitation of current data integration approaches and enabling comprehensive data-centric AI applications (Grover et al. 2018). Through these efforts, BISE research can help to empower data-centric AI to harness high-quality data, improving the performance of AI-based systems across domains and industries.

Hitherto, the development of AI and associated systems not only requires high-quality data but has largely benefitted from established processes, such as CRISP-DM (Shearer 2000), which delineate and interconnect relevant stages in the development of data mining projects. Within the CRISP-DM framework, model-centric AI primarily focusses on the modeling stage. Conversely, data-centric AI accentuates data work, including data understanding and data preparation, which requires domain knowledge. We expect three major modifications in the development of AI-based systems based on the emergence of data-centric AI, which we visualize in the context of CRISP-DM in Fig. 4. First, data-centric AI enforces an iterative process between understanding data and preparing data for subsequent modeling. In this iterative process, data versioning can help to keep track of changes in the dynamically adjusted and augmented data. Second, during modeling,

data-centric AI advocates for selecting the most appropriate model based on the data understanding and the domain knowledge. Initial tests of different methods during the modeling stage require revisiting data understanding (e.g., is data quantity sufficient for a specific method? See E1–E3). Third, continuous model improvement is a central aspect of data-centric AI. This acknowledges that data is dynamic and ever-changing, and AI models need to be continuously updated and refined to maintain accuracy and relevance (Baier et al. 2021). This requires continuous data work and an adjusting data understanding over time. In the past, data work processes were barely routinized or standardized. The shift towards data-centric AI underlines that companies need to actively manage the day-to-day activities of data work to standardize processes, methods, and tools. BISE research is ideally suited to guide the process towards augmenting and implementing standard processes for data work and the development of AI-based systems in general including both theoretical and practical considerations.

5.3 Cross-Organizational Level

Often, relevant data is scattered across various organizations, so that data sharing and interorganizational cooperation represent important considerations in data-centric AI. Previous research, though, has identified identified a range of barriers that prevent data sharing (Fassnacht et al. 2023). BISE research can play a pivotal role in facilitating effective data sharing practices (Otto and Jarke 2019). Researchers can support the design and implementation of data sharing platforms and infrastructures that enable data owners to publish and share datasets with others in a controlled and collaborative manner (E1–E3). In this context, implementing data sharing standards, such as data marketplaces with predefined quality standards, becomes crucial for optimizing data and generating business value. These platforms should address important aspects such as data versioning, data licensing, and data citation to ensure proper data governance. Additionally, exploring incentive mechanisms for data sharing can encourage organizations to share their data by providing appropriate rewards, such as data credits, collaboration opportunities, or shared benefits, fostering a culture of data sharing. Developing trust and reputation systems for data sharing is equally important (Fassnacht et al. 2023), as they can help assess the reliability and credibility of data sources, allowing data-centric AI projects to identify high-quality and trustworthy datasets. Guidance from the BISE community to design those cross-disciplinary approaches is essential in addressing these complex questions and advancing the data sharing landscape for data-centric AI.

Cross-organizational usage of AI requires either data sharing or sharing locally trained models as part of federated learning (Hirt et al. 2023). In recent years, research and practice have started sharing ML models and utilizing federated learning to mitigate a lack of data. However, when sharing models instead of data, ensuring high data quality for each individual model across organizational entities is challenging (Deng et al. 2021). For example, without access to the entire set of data, it is difficult to assess the label quality (e.g., R2) or the relevance of instances (e.g., R3). This results in a need for methods and semi-automated tools to facilitate data work across organizational entities and distributed datasets. Overall, data-centric AI across organizations requires distributed data understanding and data preparation across different organizational entities. For cross-organizational data preparation, organizations need to agree on common standards for the processing of data. This also includes questions around the ownership of code pipelines for data preparation. From a technical perspective, common data processing requires

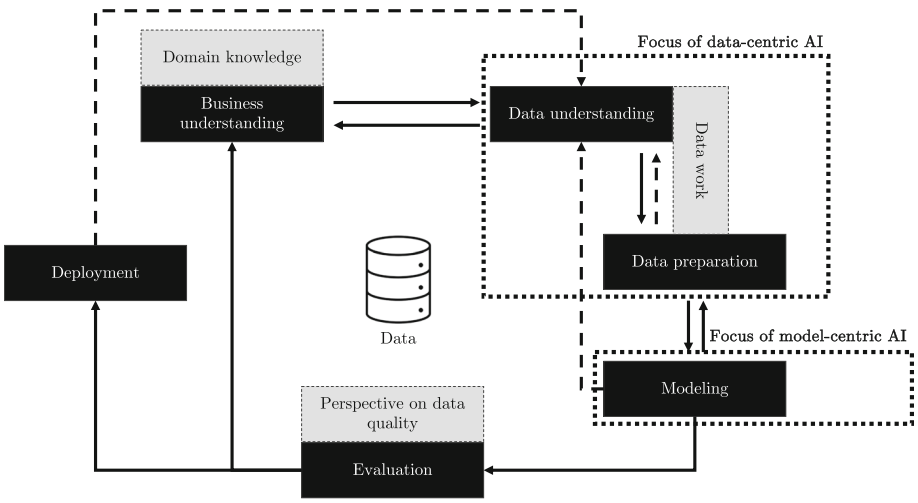

Fig. 4 Extending the Cross Industry Standard Processes for Data Mining (CRISP-DM) based on considerations from data-centric AI (black boxes and solid arrows indicate components of CRISP-DM, while the remaining components emerge from data-centric AI. Gray boxes refer to the key characteristics defined in Sect. 2)

aligned coding environments as well as versioning of code. The deployment of the resulting federated model may then have different implications for each organization's business understanding, domain knowledge, and data understanding (see Fig. 4). The BISE community has significant experience in cross-organizational research and, therefore, is well-positioned to inform data-centric AI across organizations.

6 Conclusion

Data is an indispensable component of any AI-based system. Data-centric AI and the corresponding focus on data work in the development of AI-based systems have significant implications for BISE researchers and practitioners. In this work, we introduced data-centric AI as an emerging paradigm, contrasted data-centric AI with related concepts, and highlighted a range of existing gaps in the literature that will benefit from guidance from the BISE community. The paradigm of data-centric AI has the potential to significantly improve the performance of AI-based systems in research and practice making it a promising field to study for BISE research.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Abbasi A, Sarker S, Chiang RH (2016) Big data research in information systems: toward an inclusive research agenda. J Assoc Inf Syst 17(2):1–32

Abedjan Z, Golab L, Naumann F, Papenbrock T (2022) Data profiling. Springer, Heidelberg

Alpaydin E (2020) Introduction to machine learning. MIT Press, Cambridge

Amrani H (2021) Model-centric and data-centric AI for personalization in human activity recognition. Ph.D. thesis, University of Milano-Bicocca

Aramburu MJ, Berlanga R, Lanza-Cruz I (2023) A data quality multidimensional model for social media analysis. Bus Inf Syst Eng 1–23

Baesens B, Höppner S, Verdonck T (2021) Data engineering for fraud detection. Decis Support Syst 150(113):492

Baier L, Kellner V, Kühl N, Satzger G (2021) Switching scheme: a novel approach for handling incremental concept drift in real-world data sets. In: Proceedings of the Hawaii international conference on systems sciences, pp 990–1000

Biewald L (2020) Experiment tracking with weights and biases. https://www.wandb.com/. Accessed 02 Dec 2022

Budach L, Feuerpfeil M, Ihde N, Nathansen A, Noack N, Patzlaff H, Naumann F, Harmouch H (2022) The effects of data quality on machine learning performance. arXiv:2207.14529

Chen H, Chiang RH, Storey VC (2012) Business intelligence and analytics: from big data to big impact. MIS Q 36(4):1165–1188

Deng Y, Lyu F, Ren J, Chen YC, Yang P, Zhou Y, Zhang Y (2021) Fair: quality-aware federated learning with precise user incentive and model aggregation. In: Proceedings of IEEE conference on computer communications. IEEE, pp 1–10

Fassnacht M, Benz C, Heinz D, Leimstoll J, Satzger G (2023) Barriers to data sharing among private sector organizations. In: Proceedings of the Hawaii international conference on system sciences (HICSS), pp 3695–3705

Fiedler N, Bestmann M, Hendrich N (2019) Imagetagger: an open source online platform for collaborative image labeling. In: Proceedings of RoboCup 2018: robot world cup XXII. Springer, Heidelberg, pp 162–169

Gröger C (2021) There is no AI without data. Commun ACM 64(11):98–108

Grover V, Chiang RH, Liang TP, Zhang D (2018) Creating strategic business value from big data analytics: a research framework. J Manag Inf Syst 35(2):388–423

Gudivada V, Apon A, Ding J (2017) Data quality considerations for big data and machine learning: going beyond data cleaning and transformations. Int J Adv Softw 10(1):1–20

Hemmer P, Kühl N, Schöffer J (2022) DEAL: deep evidential active learning for image classification. Deep Learn Appl 3:171–192

Hirt R, Kühl N, Martin D, Satzger G (2023) Enabling inter-organizational analytics in business networks through meta machine learning. Inf Technol Manag (forthcoming)

Holstein J, Schemmer M, Jakubik J, Vössing M, Satzger G (2023) Sanitizing data for analysis: designing systems for data understanding. Electron Market 33(1):1–18

Holzinger A (2016) Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Inform 3(2):119–131

Jarrahi MH, Memariani A, Guha S (2023) The principles of data-centric AI. Commun ACM 66(8):84–92

Jordan MI, Mitchell TM (2015) Machine learning: trends, perspectives, and prospects. Science 349(6245):255–260

Kaggle (2023) Kaggle competitions. https://www.kaggle.com/competitions. Accessed 05 Jul 2023

Kühl N, Schemmer M, Goutier M, Satzger G (2022) Artificial intelligence and machine learning. Electron Market 32(4):2235–2244

Legner C, Pentek T, Otto B (2020) Accumulating design knowledge with reference models: insights from 12 years' research into data management. J Assoc Inf Syst 21(3):735–770

Lin Q, Ye G, Wang J, Liu H (2022) RoboFlow: a data-centric workflow management system for developing AI-enhanced robots. In: Proceedings of the conference on robot learning. PMLR, pp 1789–1794

McInnes L, Healy J, Melville J (2018) UMAP: uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426

Ng A, Aroyo L, Coleman C, Diamos G, Reddi V, Vanschoren J, Wu C, S Z (2021) Data-centric AI workshop. https://datacentricai.org/neurips21/. Accessed 12 Feb 2022

Ng A, Laird D, He L (2022) Data-centric AI competition. https://https-deeplearning-ai.github.io/data-centriccomp/. Accessed 04 Dec 2022

Northcutt CG, Athalye A, Mueller J (2021) Pervasive label errors in test sets destabilize machine learning benchmarks. arXiv:2103.14749

Otto B (2011) Organizing data governance: findings from the telecommunications industry and consequences for large service providers. Commun Assoc Inf Syst 29(1):45–66

Otto B, Jarke M (2019) Designing a multi-sided data platform: findings from the international data spaces case. Electron Market 29(4):561–580

Parmiggiani E, Østerlie T, Almklov PG (2022) In the backrooms of data science. J Assoc Inf Syst 23(1):139–164

Renggli C, Rimanic L, Gürel NM, Karlas B, Wu W, Zhang C (2021) A data quality-driven view of MLOps. IEEE Data Eng Bull 44(1):11–23

Ronneberger O, Fischer P, Brox T (2015) U-Net: convolutional networks for biomedical image segmentation. In: Proceedings of the international conference on medical image computing and computer-assisted intervention, pp 234–241

Sambasivan N, Kapania S, Highfill H, Akrong D, Paritosh P, Aroyo LM (2021) "Everyone wants to do the model work, not the data work": data cascades in high-stakes AI. In: Proceedings of the CHI conference on human factors in computing systems, pp 1–15

Schneider J, Abraham R, Meske C, Vom Brocke J (2023) Artificial intelligence governance for businesses. Inf Syst Manag 40(3):229–249

Shearer C (2000) The CRISP-DM model: the new blueprint for data mining. J Data Warehous 5(4):13–22

Strickland E (2022) Andrew Ng: unbiggen AI. https://spectrum.ieee.org/andrew-ng-data-centric-ai. Accessed 12 Dec 2022

Toreini P, Langner M, Maedche A, Morana S, Vogel T (2022) Designing attentive information dashboards. J Assoc Inf Syst 23(2):521–552

Turban E (2011) Decision support and business intelligence systems. Pearson Education India

Whang SE, Roh Y, Song H, Lee JG (2023) Data collection and quality challenges in deep learning: a data-centric AI perspective. VLDB J 32(4):791–813

Zhang R, Indulska M, Sadiq S (2019) Discovering data quality problems: the case of repurposed data. Bus Inf Syst Eng 61:575–593