Data-related Practices for Creating Artificial Intelligence Systems in K-12

Viktoriya Olari
viktoriya.olari@fu-berlin.de
Freie Universität Berlin
Berlin, Germany

Ralf Romeike
ralf.romeike@fu-berlin.de
Freie Universität Berlin
Berlin, Germany

Abstract

Computer science curricula have started to include competencies related to artificial intelligence (AI) in K–12 education. However, before introducing a new topic into the classroom and suggesting competencies, it is essential to identify the central practices of the discipline. In the following research, we focus on identifying practices related to data, as current school curricula significantly underestimate the role of data, and understanding how data is processed is a key to understanding how AI systems function. We examine the theoretical literature on practices applied to data when creating AI systems, map the practices in a process model, validate the results of the mapping with domain experts, and contrast the results with current AI curricula for school students. The contribution of this work is a process model that summarizes data-related practices for AI systems built with machine learning, is comprehensively domain-embedded, and is aligned with K–12 education. Computer science educators can use it as a blueprint for defining competencies and designing learning arrangements that aim to enable students to create and understand AI systems.

CCS Concepts

• **General and reference** → *Surveys and overviews*; • **Computing methodologies** → **Artificial intelligence**; • **Social and professional topics** → **K-12 education**.

Keywords

AI Education, Data Education, Computer Science School Education

ACM Reference Format:
Viktoriya Olari and Ralf Romeike. 2024. Data-related Practices for Creating Artificial Intelligence Systems in K-12. In *The 19th WiPSCE Conference on Primary and Secondary Computing Education Research (WiPSCE '24), September 16–18, 2024, Munich, Germany.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3677619.3678115

1 INTRODUCTION

Artificial intelligence (AI) technologies are increasingly becoming part of information technology systems and enabling them to forecast, classify, detect, and extract information, reason, plan, search, and generate content, and more. A key driver for this progress is the availability of data. However, poor data quality creates a bottleneck. One may often get state-of-the-art results with high-quality data and simple algorithms, but rarely with low-quality data and the best algorithm [50]. Insufficient data quality is responsible for significant problems in AI systems, including amplifying bias, sexism, racism, and other forms of discrimination [17]. Therefore, the demand for a workforce that can responsibly leverage data for AI is huge. Promoting an understanding of AI systems is also critical for society as a whole and should start in schools [17, 32]. School students are already actively working with AI technologies; hence, they need an awareness and understanding of the data underlying these systems. In addition, they need awareness with respect to their role as data producers when they are interacting with AI and sharing their data.

A key to understanding AI systems is understanding how they are created and function [32, 49, 88]. In this context, a significant amount of work has been done to uncover the functionality of AI algorithms for school students using embodied interaction [38], robotics [53, 88], unplugged [42, 53], and computer-based approaches [36]. However, prior research indicates a significant gap. The role of data in AI is still significantly underappreciated in the context of AI education in schools. Current teaching approaches only scratch the surface of working with data. AI frameworks and curricula for school education devote little attention to the data-related skills required to build AI applications [54].

This gap holds unrealized learning potential as it offers students unique learning opportunities to better understand how AI systems function, including sources of bias, the reliability of AI systems, and the limits of the use of AI systems – all of which are considered essential for K–12 education [44, 49, 75]. For example, a common practice for students learning about AI is working with ready-to-use datasets without being involved in the data collection and preparation process [8, 26, 86]. In the real world, however, data is rarely immediately ready for use. Practitioners invest considerable effort in preparing datasets, iterating on the problem statement, and multiple rounds of data collection [18, 43, 91]. This process is prone to error but provides an opportunity for students to learn how biases are introduced into the dataset, such as measurement bias introduced during data collection [48]. Without this, students may even be misguided to think that all data is already perfect and accurately represents real-life phenomena, which would amplify the common misconception among school students that all data can be used by AI [39]. Furthermore, instead of passing ready-to-use data to an AI algorithm, students can be guided to follow common practices for creating an AI system [4], including defining success criteria for the AI solution, exploring the structure of the dataset, cleaning the data, and engineering features. In the process, they

WiPSCE '24, September 16–18, 2024, Munich, Germany

ACM ISBN 979-8-4007-1005-6/24/09
https://doi.org/10.1145/3677619.3678115

can learn that data must be improved before it can be used by an AI system and that, besides the AI algorithm, it is the interplay between the problem, the success criteria, and the type and quality of the data that determines whether the use of AI techniques adds value to solving a problem.

Closing this gap reveals unique opportunities for developing new approaches to teaching. For example, a narrative such as "the more data passed to an AI system, the better the result will be" evokes the impression that the creation of real AI systems is not possible in a school setting – students do not have the time to work with large amounts of data and schools do not have the necessary computational resources. In reality, although the volume of data is important, the quality of the data is of utmost importance [34, 35, 91, 92]. In fact, for many real-world problems practitioners do not have much data and can only solve these problems by amending existing data through expert knowledge [62]. Making these processes visible could foster the development of new teaching approaches in computer science education.

To expand knowledge about the role of data in AI education, we conducted a comprehensive theoretical analysis of the AI field, corroborating the results with experts, contrasting the findings with AI school curricula, and distilling the outcome into a process model that can be used by lesson and curriculum developers to formulate competencies. During the research process, we followed the approach and goal of computer science education – focusing on the key ideas of a field rather than on fleeting application knowledge – and were guided by the following overarching research question (RQ) with two sub-questions:

RQ: What are the key data-related practices that are applied to data when creating AI systems, and how should they be aligned to be used in K–12 education?

(1) What are the typical practices that practitioners apply to data when creating AI systems?
(2) How can the identified practices be aligned with AI education in K–12 and presented in a way that is useful for lesson and curriculum development?

The paper is organized as follows. In Sections 2 and 3, we discuss the role of data in the development of AI systems and previous theoretical work from the AI education research field on the incorporation of data in AI education. The methodology that we followed is described in Section 4. The model is outlined in Section 5. In Sections 6 and 7, we discuss threats to the validity of the model and give an outlook for future work.

2 THE ROLE OF DATA IN AI SYSTEMS

Data is a core component of information technology systems using AI techniques. AI techniques incorporate machine learning (ML) approaches such as supervised, unsupervised, and reinforcement learning and knowledge-based approaches, among others [19]. Supervised learning techniques learn patterns from labeled data with the goal of generalizing the patterns to unseen data. Unsupervised learning techniques work with unlabeled data to separate it into groups that share common characteristics [18]. In reinforcement learning, an agent produces data through interaction with the environment and learns from this data to better perform actions. In systems using knowledge-based approaches, data is manually handcrafted with the goal of representing it as knowledge, processing new data using this knowledge, and deriving new facts [63].

At the machine level, data used by AI systems is digitally stored on a device in binary values. It comes from different sources, such as humans, sensors, or machines, and is represented in different modalities, such as tables, images, text, audio, geospatial data, time series, and graphs [57]. The process of data collection, data modeling, and processing by an AI system is highly iterative. A common technology-agnostic model that practitioners use as a guide when developing an AI system is the CRISP-DM model [18], which was initially developed for data mining.

Depending on the concrete AI technology, the sub-processes may differ. For example, during data preparation and modeling, data cleaning and feature transformation are common across different ML approaches, while data labeling is inherent to supervised learning [34, 71, 91]. When developing knowledge-based systems, data preparation and modeling include eliciting expert knowledge from primary sources into knowledge protocols, interpreting it in a structured model, and formalizing it [70].

For any AI technology, careful data engineering is key to a reliable system. In the context of knowledge-based AI systems, the importance of the inclusion of multiple data sources and experts during the knowledge acquisition phase to omit bias [43] and the difficulties and limits of knowledge engineering have been known for many years [43, 63]. In the context of ML, the research direction of data-centric AI emerged recently [34, 71, 91]. It emphasizes that the careful selection and curation of data is key for ML systems. Rather than identifying more effective models to improve the performance of a system while leaving the data unchanged, a more accurate and reliable system can be achieved by keeping the model unchanged and continuing to improve the data [92].

Working with data continues even after a model has been deployed in production. Practitioners evaluate the model by measuring its performance on new data and interpreting its decisions. For both ML systems and knowledge-based systems, it is typical that performance will not be optimal after the first iteration [63]. Practitioners, therefore, move between all stages until the model accurately represents the data and meets the success criteria set in the understanding phase. When an AI system is deployed, it starts processing new data. Thus, it is critical to set up monitoring pipelines to understand the structure of the incoming data and observe the model's performance.

3 DATA PRACTICES IN AI EDUCATION

In order to understand how AI systems function, it is of paramount importance for novices to gain an understanding of how data is processed during AI system development [50]. This idea is reflected in the learning objectives of several AI curricula for school education [44, 49, 73, 75, 81]. For instance, when students learn about AI, they need to understand that computers learn from data and that ML is about statistical inference that finds patterns in data [75]. When building ML systems, students are expected to learn practices such as data selection, data filtering, and data splitting [67]. An AI-literate person should also be data-literate, that is, be able to collect,

manage, evaluate, and apply data in a critical manner [44, 59]. However, a recent literature review of the AI education field shows that current teaching approaches scratch only the surface of working with data [54].

A data literacy competency model developed by Grillenberger and Romeike [25] has been used as a guide to educate school students on data in AI systems. It emerged from a comprehensive analysis of the data management field and combines the data lifecycle with key high-level concepts in data management [24, 25]. As such, it offers a foundation from which to start working with data in the context of AI. However, it lacks processes immanent to AI technologies such as data labeling, data augmentation, feature engineering, data splitting, and the monitoring of new incoming data, among others. For these reasons and following the argumentation provided in the introductory section of this paper, we see a need for additional theoretical research that, compared to research done by Grillenberger and Romeike [25], focuses on the practices inherent in systems built with AI technologies and elaborates on these practices in detail.

4 METHOD

To identify key data-related practices, we investigated the AI field and structured the findings into a model. A similar procedure has been used in previous computing education research when structuring the field of data management for school education [24]. To align the model with K–12 education and avoid developing it in isolation from school practice, we looked at international AI curricula for K–12 with the goal of understanding what data-related practices are already expected to be understood by students. In the process, we discovered additional, and sometimes more general, practices, which we checked against the theoretical literature in the third step. Subsequently, we updated the model, evaluated it with the help of experts from the AI field, and informally discussed its usefulness with computer science teachers for secondary schools and computer science education researchers. After several iterations, we arrived at the final process model, which organizes practices in the data lifecycle that teachers and curriculum developers can use to design lessons and define competencies. Figure 1 visualizes the procedure. In the following subsections, we outline details of the three research stages.

4.1 Investigation of the Subject Field

The aim of the review of the AI field was to come up with an initial list of practices that practitioners apply to data when developing an AI system. To select the literature, we followed a purposeful sampling strategy. The purposeful strategy aimed to find information-rich studies that would provide an answer to the posed RQ [31] – in our case, to the first sub-question of the RQ. From the different strategies that purposeful sampling offers, we mostly used criterion sampling and snowballing, starting with reviewing standard textbooks used in education for AI and data science [4, 63, 69]. We additionally conducted searches of the ACM, SpringerLink, and ScienceDirect databases with the keywords "data lifecycle," "data-centric AI," "data processing," "artificial intelligence," "machine learning," and "knowledge-based" to find academic papers describing data processing in systems built with ML and knowledge-based techniques [2, 5, 10, 15, 24, 27, 33–35, 40, 70, 89–91]. In order to include a practical perspective, we also reviewed gray literature recommended by AI practitioners [6, 18, 50, 66, 74, 92, 93].

While reading the sources, we manually extracted practices that practitioners follow when developing a system that uses AI models along with their descriptions, ending up with a document of 111 pages that included descriptions of over 84 processes. For instance, from the following text passage describing the data science cycle, "a typical process starts with question or problem formulation, then goes through data collection, wrangling, cleaning, modeling, and finally representation, evaluation, and interpretation of the results" [4], we extracted the practices problem formulation, data collection, data wrangling, data cleaning, and data modeling. To determine redundancies and similarities, we mapped the practices in a hierarchical order. Because the mapping represented a process that is similar to the CRISP-DM reference model [18], we iteratively structured the practices in a process model using the CRISP-DM model as a guide while adding some additional stages. In order to make the process model focused and consistent, we excluded practices related to knowledge-based AI and hybrid AI. Thus, the final model contained *practices relevant to ML that are performed on data during one of the following stages of the data lifecycle: understanding the task, collecting data, understanding data, preparing data, implementing solutions, evaluating performance, deploying and monitoring the system, and sharing, deleting, and archiving data.*

Subsequently, the model was validated for correctness, representativeness, and relevance with the help of one practitioner who worked as a data scientist in a software company, one practitioner who had professional experience in data science and founded a company in the data science field, one researcher who had professional experience as a data scientist and worked in the field of responsible and explainable AI, and one researcher who had professional experience in developing curriculum and content for computing programs focusing on AI and working on the explainability of AI-based applications. The experts were recruited through the researchers' personal networks. Three experts gave written feedback, and one gave feedback during an informal interview. All the experts welcomed the process model, agreed with its general correctness, and gave suggestions to make it more comprehensive. Before following the advice to extend the model, the suggestions were checked against additional theoretical literature. If a suggestion was a data-related practice, as described above, the model was updated accordingly.

4.2 Investigation of the K–12 Education Field

To align the model with K-12 education and avoid developing it in isolation from school practice, we looked at AI school curricula guided by the second sub-question of the RQ. We focused on curricula because they provide a strong body of accumulated knowledge and are a useful starting point for understanding what researchers, teachers, and practitioners consider to be important for school students. To collect AI curricula, we used a snowballing strategy [31], starting with a review of papers identified in a recent systematic literature review on AI education in schools [60]. We included papers for further processing if they described what and how students

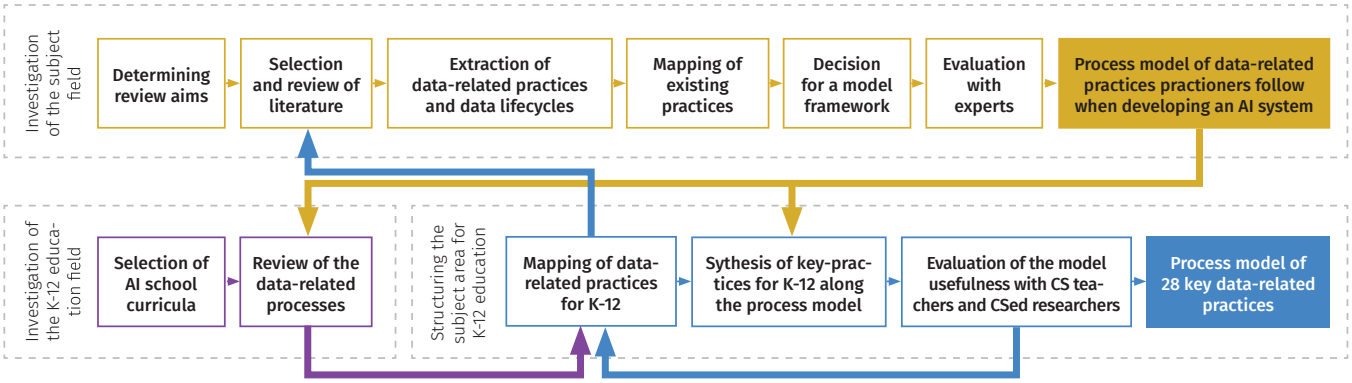

Figure 1: Overview of the analysis process.

should learn [13]. We also included the two most recent AI curricula that we were aware of from our previous research. After retrieving and closely reading these papers, we excluded 53 due to missing curricula, leaving a total of 49 papers. We then extracted a curriculum from each work, resulting in a large text corpus. Using the MAXQDA software, we iteratively searched for text passages that included keywords related to data that we defined based on the prior theoretical research (keyword list: data, information, input, file, image, picture, foto, photo, figure, text, digit, word, message, post, sound, recording, audio, music, tone, speech, song, video, graph, time series, time, date, spatial, table, number, numerical, survey, content, char, string, integer, boolean, float, array, list, map, dictionary, tuple, vector, matrix, binary, feature, category, class, object, pixel, N-Gram, tf–idf, DNA, variable, output, prediction, classification, recommendation, clustering, categorization, sequence, population, sample, observation, instance, point). If a sentence contained a data-related word, we auto-coded it as data-related.

4.3 Structuring the Subject Area for K–12 Education

To check the model for completeness and to identify any missing parts important for AI education, we manually reviewed all coded sentences, and if they contained references to a process, we added them to our mapping. In the process, we discovered additional, sometimes more general, practices. We then did an additional literature search and updated the model. Subsequently, we evaluated the model informally with three computer science teachers and several computer science education researchers, who were recruited through our personal network. The teachers expressed the need to make the model more suitable for classroom use and provide examples for every practice. We updated the model iteratively, adding examples of concrete sub-practices and general learning goals.

The final process model is comprehensively domain-embedded and aligned with K–12 education, as it includes strategies from the literature and practice and was evaluated for correctness and relevance by experts in the AI field. It is accompanied by examples and can be used by teachers and curriculum developers to plan lessons. In what follows, we describe the process model in detail.

5 THE PROCESS MODEL

The final process model consisted of 28 central data-related practices related to data processing during the development of an ML system. The practices are allocated to one of eight overarching stages, as illustrated in Figure 2. The stages represented by the same color are highly linked to each other. The arrows in the model emphasize that the process is highly iterative, and students may move back and forth between the stages. In the following, we provide core learning objectives for each stage, present practices, and give examples of relevant activities.

5.1 Understand the Task

In this stage, students can learn to analyze a real-world problem in order to determine whether ML technologies can add value to the solution and to define criteria describing when the problem has been successfully solved. To do this, students may engage with stakeholders, analyze the task and nature of data, and define success criteria for the solution [15, 18].

Understand the needs of the stakeholders. Understanding the needs of stakeholders is a critical process that must occur at the beginning of any project that employs ML technologies. Depending on their roles, stakeholders are concerned with different perspectives on data and needs [57].

For instance, data users want to be aware of how data is used by a given AI system, and data agents need data to be in a specific format before they can use it to create AI applications. Prior AI curricula often refer to students as data agents [9, 12, 45, 52, 61, 78, 80, 84, 85]. As data agents, students may engage with stakeholders who will use the solution and interact with the developed applications (data users) and those who contribute to dataset creation (data producers) [57].

Analyze the task of the project. ML technologies are employed to solve a wide variety of tasks. These tasks specify the problem to be solved [93]. Examples of tasks found in the AI curricula include prediction [9, 46, 77, 78, 81, 82], classification [9, 13, 41, 52, 56, 58, 64, 67, 72, 77, 78, 81, 82], text and speech generation [77, 82], image stylization [51], clustering [78], and recommendation [9, 20, 80]. Depending on the task, students as practitioners can decide what data is needed, how to prepare it, and whether past projects using similar data have used ML technologies to solve the task.

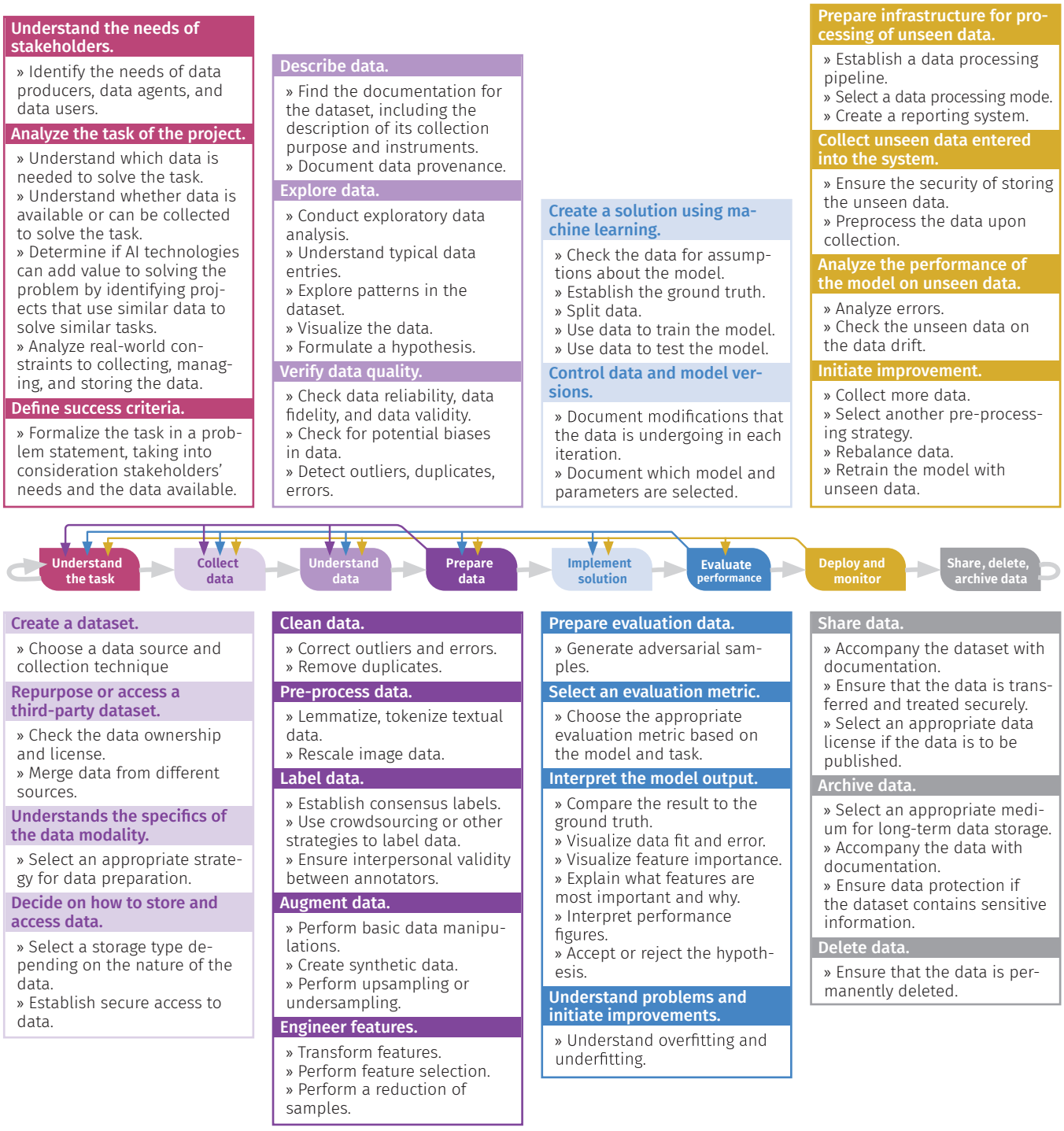

Figure 2: An eight-stage process model outlining the 28 data-related practices for AI education in K–12. The practices are approaches applied to data during the development of a ML system.

Define success criteria. A task is formalized in a problem statement, and success criteria are defined. According to experts, the goal is not to suggest a concrete technology that will solve the task but to describe how ML can add value to the project and what the success criteria are. The description may contain a preliminary plan for the solution. Although AI curricula do not include references to

this practice, theoretical literature [4, 18] and experts emphasize the importance of this step, as, otherwise, it is not clear when a project has successfully come to an end.

5.2 Collect Data

In this stage, students can learn strategies for collecting data and their associated difficulties, as these will influence the quality, reliability, and fairness of an ML system. They can also learn that, depending on the data modality, data must be pre-processed in different ways before it can effectively be used by an ML system.

Many AI curricula expect students to engage with or create systems based on supervised and unsupervised ML [9, 12, 45, 52, 61, 67, 78, 80, 82, 84, 85]. To do this, the students, as practitioners, need initial sets of data [22, 70, 92]. AI curricula emphasize that students should know how to collect data [81]. Students, as practitioners, have two options to collect data: creating a dataset or repurposing a dataset created by a third party.

Create a dataset. A dataset is created using a variety of techniques, such as surveying, crowdsourcing, scrapping, crawling the data from the internet, or using sensors. Some curricula refer to students creating their own datasets [78, 81]. During this step, students can learn that during the data sampling process, it is important to make sure that the data is reliable, valid, and fidelitous [5].

Repurpose or access third-party data. Collecting primary data can be time-consuming, so repurposing data [57] by accessing a third-party dataset could be more convenient. However, repurposing data might cause problems because the data could be outdated and not adapted to the requirements of the project. Existing AI curricula often refer to students using third-party datasets [44, 76, 78]. During this process, for students as practitioners, it is important to check the documentation of a dataset and the legal permissions for its use [57].

Understand the specifics of the data modality. Data used by ML algorithms is represented in different modalities. AI curricula suggest students work with images [14, 20, 37, 41, 47, 51, 72, 76, 78, 81], texts [3, 16, 41, 52, 58, 72, 77, 79, 81], tabular data [20], audio [76], video [12], and graph data [77]. Another common data modality is geospatial and time series data [57]. When selecting an appropriate ML algorithm, students need to understand the specifics of the data modality, as this influences data cleaning, modeling, and preparation methods.

Decide how to store and access data. Once data is collected or a dataset is chosen, how to access the data is decided. AI curricula suggest that school students be familiar with data storage and management in simple databases [68, 81], relational databases, local data files, and cloud storage [81]. However, ML systems might have specific requirements for storing data as they work with semi-structured and unstructured data. Knowledge of vector and graph databases is needed to store data with high-dimensional characteristics [29].

5.3 Understand Data

In this stage, the students can learn what data looks like and that, through careful observation, it can reveal a large amount of information about a real-world problem. They can also learn about difficulties that might influence the quality and reliability of the AI system.

Describe data. Describing the origins and processing of data is an essential, though not sufficient, prerequisite for the creation of a fair, accountable, transparent, and explainable ML system [87]. AI curricula expect students to know where data comes from [81, 84]. From the subject perspective, this might include describing data provenance (such as updates, release, licenses, authoring of the dataset, and documentation on data collection and pre-processing techniques, including demographics of the processes, such as who gathered the data) [23].

Explore data. The goal of exploration is to give an intuition about potential problems that need to be considered when using data in ML system development [21]. The exploration step is supported by creating visualizations, producing summary statistics (e.g., feature distributions), and creating data reports communicating the findings and relations between the features (e.g., through calculating correlations between features) [7]. We found that some AI curricula suggest students conduct a dataset analysis [76], understand data trends [81], visualize textual and numerical data [55, 65, 81], and communicate data [81] – all typical practices of data exploration.

Verify data quality. Verifying data quality means ensuring that the data fits the purpose [15] and is of high quality, which is vital for robust ML systems [35, 50]. Data is of high quality when it "accurately represents a phenomenon, and … exhibits empirical and explanatory power [5]." It must be reliable (e.g., consistent), valid, and have high fidelity (i.e., accurately represent the reality) [5]. AI curricula suggest students be familiar with the ideas of messy data [44, 81], general dataset quality [64], data representativity [78], data diversity [78], data homogeneity [83], data authenticity, and data accuracy [13, 84]. Therefore, to make a data quality assessment, students and practitioners should check the relevant dataset for outliers, duplicates, errors, and biases. It is important that students understand statistical fundamentals at this stage, including observation, population, sample size, and data distribution [21].

5.4 Prepare Data

During the data preparation phase, students can learn how to overcome difficulties identified in the data understanding stage and that different data cleaning, pre-processing, and engineering strategies will have a direct impact on the quality of an ML system. The data collected in the data collection step is often raw, meaning it is not ready to be used by an AI system due to noise such as erroneous attribute values, missing or incomplete values, or unnecessary information [93]. These issues might lead to an inaccurate and biased solution [92] and, therefore, must be corrected.

Clean data. Data cleaning and data pre-processing are common processes for all types of data and ML techniques [91, 92]. Data cleaning typically involves correcting errors and removing outliers and duplicates. For example, imputing missing values using the mean and predicting missing values using regression models are suitable practices for tabular data [92]. Existing AI curricula suggest that students understand how to correct a dataset [30].

Pre-process data. Data pre-processing differs depending on the nature and quality of the data as well as the specific ML technique

that a student aims to use. For instance, textual data needs to be lemmatized and tokenized. Image data needs to be rescaled. We did not find any references to pre-processing data in the AI curricula we reviewed.

Label data. Data labeling is inherent to supervised ML. It refers to the process of assigning one or more descriptive tags or labels to the data entries in a dataset and is considered to be a time-consuming and resource-intensive process [92]. Labels can be created by non-expert annotators in a crowdsourcing process or by annotators in cooperation with semi-supervised techniques [92]. To avoid labeling errors and ensure label quality, it is helpful to establish consensus labels, a labeling strategy, and interpersonal validity between the annotators. Existing AI curricula propose that students understand data labeling practices [32, 67, 77, 78].

Augment data. Data augmentation is a practice for increasing the size and diversity of data by artificially creating variations of existing data [92]. This strategy is helpful if there is limited data available or if creating a large annotated dataset is costly [92]. Common approaches for data augmentation include basic manipulations (e.g., for image data, scaling, blurring, rotating), augmentation data synthesis (generating synthetic data that closely resembles the existing data), and upsampling [92] and undersampling. We did not find references to data augmentation in the AI curricula that we reviewed.

Engineer features. Feature engineering refers to the process of formulating the most appropriate features given the data, the ML algorithm, and the task [93] and transforming raw data into the feature format usable by an ML algorithm [66]. This includes feature selection, extraction, transformation, and reduction. Feature selection is the process of obtaining a subset of features from an original feature set [11]. Feature extraction refers to the transformation of the original data into features with strong pattern recognition potential [11]. Feature reduction is the process of reducing the complexity of a dataset by reducing its feature size or sample size while retaining its essential information [92]. The success of ML models often depends on the success of feature engineering rather than the selection of an algorithm or a model [66]. In some AI curricula, we found references to practices of feature engineering, such as feature design [85], feature selection, feature extraction [67, 76, 78], feature transformation, and feature encoding [78, 81].

5.5 Implement the Solution

During the implementation stage, students can learn to use an ML algorithm to learn from data (supervised ML) and separate data into groups (unsupervised ML).

Create a solution using machine learning. During this process, practitioners check the data for assumptions about the selected technique, establish the ground truth if possible, and create visualizations to understand the learning process. For supervised ML, the data is split into training, validation, and testing data. Many AI curricula suggest students create solutions using ML techniques, including the practice of splitting datasets into training and testing datasets [9, 12, 45, 52, 61, 78, 80, 84, 85] as well as validation dataset [67, 78, 82].

Control data and model versions. During all stages, the responsible handling of data is of utmost importance and includes documenting modifications that the data is undergoing and issues discovered along the process [57]. We did not find any references to data control and versioning in AI curricula. However, experts and practitioners suggest that using a data version control system helps track the highly iterative process [35], find the best combination of data and models, and achieve reproducible results.

5.6 Evaluate Performance

At this stage, the students can learn to evaluate a model's performance from the perspective of the success criteria. They can also learn to compare the performance of models, observe how problems identified at the understanding stage influenced the model, and initiate improvements if the task is not yet satisfactorily solved.

Prepare evaluation data. Evaluation data helps to assess how well the model fits the data. For instance, a test dataset is used to test a supervised ML model. It must contain data pre-processed in the same way as the data on which the model was trained. The data must come from the same distribution. An evaluation dataset can be created to test the boundaries of the model. This consists of adversarial samples or data from another distribution [92]. We did not find references to the preparation of evaluation data in existing AI curricula.

Select an evaluation metric. The evaluation metric might depend on the model and the nature of the data. Frequently used metrics are accuracy, precision, recall, F1 score, root-mean-squared error, purity, and entropy [15]. An effective evaluation metric should be accurate, robust, scalable, and interpretable [15]. Typically, AI curricula refer to the accuracy metric [9, 45, 52, 78, 80–82].

Interpret the model output. This process includes practices for creating and interpreting performance figures [91, 92] and visualizing and interpreting the model output [15]. We did not find references to this practice in existing AI curricula. Here, students might use heatmaps to understand feature importance [94] or a confusion matrix to interpret performance for classification problems.

Understand problems and initiate improvements. If the performance of the model is not satisfactory (e.g., the system does not meet success criteria), iterations on prior stages are needed to augment the data, clean the data in a different way, add more data samples, interview more experts, select another dataset, or choose another ML model [15, 70]. We did not find references to this practice in existing AI curricula.

5.7 Deploy and Monitor

At this stage, students can learn how to use the ML model in production to solve the task identified at the beginning and share the solution with others. They can also learn that as soon as the solution is deployed and unseen data comes in, there will be a need for additional iterations for improvement.

We did not find any references to deployment and monitoring in AI curricula. However, it is possible for students to deploy ML systems in school contexts using tools such as App Inventor [72].

Prepare infrastructure for the processing of unseen data. After the successful performance evaluation, the solution is deployed with the unseen data [1]. Students can decide between real-time or batch-mode processing of the data [15]. As practitioners, students might create a reporting system about the incoming data.

Collect unseen data entered into the system. The unseen data that comes into the system is raw and needs to be pre-processed. Establishing data pre-processing procedures upon data collection is a crucial practice at this step. If the data is not pre-processed, the model will output an incorrect result. Ensuring the security of storing the unseen data is also important at this stage.

Analyze the performance of the model on unseen data. Since the structure of the unseen data may change over time, students as practitioners might need to check the data for drift. Data drift refers to changes in the distribution of input data [35] leading to inaccurate model performance. The output of the model can be stored along with the results of error analysis and unseen data for monitoring purposes.

Initiate improvement. As the world changes, the unseen data might be used to improve the ML model in the next development cycle. The students, as practitioners, might want to select a new strategy for pre-processing the data, conduct data rebalancing, and initiate model recreation.

5.8 Share, Archive, and Delete Data

Existing AI curricula does not explicitly reference the share, delete, and archive stages. However, working with practices at this stage may allow students to learn how to complete these steps safely and responsibly.

Share. When sharing data, students, as practitioners, must accompany the dataset with documentation. This typically includes information about stakeholders, the original purpose of data collection, a description of real-world constraints during the data collection, measurement instruments, data manipulations such as data cleaning, and issues discovered during the data understanding [57]. They also must ensure that the personally identifiable information is deleted and the data is transferred and treated safely.

Archive. When archiving the data, an appropriate medium for long-term data storage must be selected. The archived dataset must be accompanied by documentation.

Delete. When deleting data, the students, as practitioners, should ensure that it is permanently deleted.

6 LIMITATIONS

This work is based on an extensive theoretical analysis of the literature. However, depending on the specific ML algorithm or task, more specific processes are relevant, as shown by the sub-practices we added for teachers to make the concepts more tangible. The results of the studies that structure the field can be influenced by the individuals performing the procedures. To address these limitations and to ensure the high validity and representativeness of the model, we evaluated the model with several domain experts. Furthermore, curriculum developers should be aware that the model presents the upper limit of what is possible in upper-secondary computer science school education as it was created additively and does not include practices related to building AI systems using knowledge-based and hybrid AI approaches. When planning lessons, high-level practices can be introduced in the lower classes, and sub-practices can be introduced in higher grades.

7 OUTLOOK AND FUTURE WORK

In this work, we presented a process model that explores the role of data in AI for K–12 education. The process model is comprehensively domain-embedded and aligned with K–12 education. To increase the model's validity and ensure that it is correct, complete, and relevant for K–12, it was evaluated by experts from the AI field and cross-checked for usefulness by computer science teachers and computer science education researchers.

The results of this work are of interest to a wide audience. The model can be used in schools to teach socio-cultural perspectives on AI. Going through all eight stages and working with corresponding practices can give students the chance to gain important insights into the creation of AI systems and thus understand them better. The model also offers unique opportunities for developing new approaches to teaching AI in schools.

Researchers and curriculum developers can benefit from this work because the identified key practices can serve as a basis for defining skills – learned, observable, and performed acts – that students should master [28]. Educators can use the results as a blueprint for creating data-centered AI lessons to help students understand how AI systems work. The public can benefit from this work by learning about the role of data in the context of AI through the description of key practices. In the future, we will continue working closely with teachers to use the model as a blueprint for data-centered AI courses.

References

[1] Samuel Ackerman, Eitan Farchi, Orna Raz, Marcel Zalmanovici, and Parijat Dube. 2020. Detection of Data Drift and Outliers Affecting Machine Learning Model Performance over Time. (2020). https://doi.org/10.48550/ARXIV.2012.09258

[2] Mehwish Alam, Paul Groth, Pascal Hitzler, Heiko Paulheim, Harald Sack, and Volker Tresp. 2020. CSSA'20: Workshop on Combining Symbolic and Sub-Symbolic Methods and Their Applications. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. ACM, Virtual Event Ireland, 3523–3524. https://doi.org/10.1145/3340531.3414072

[3] Safinah Ali, Daniella DiPaola, Irene Lee, Jenna Hong, and Cynthia Breazeal. 2021. Exploring Generative Models with Middle School Students. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3411764.3445226

[4] Cecilia Rodriguez Aragon, Shion Guha, Marina Kogan, Michael Muller, and Gina Neff. 2022. *Human-Centered Data Science: An Introduction.* The MIT Press, Cambridge, Massachusetts.

[5] Lora Aroyo, Matthew Lease, Praveen Paritosh, and Mike Schaekermann. 2022. Data Excellence for AI: Why Should You Care? *Interactions* 29, 2 (March 2022), 66–69. https://doi.org/10.1145/3517337

[6] Aurek Chattopadhyaya, Matthew Van Dorenb, Reese Johnsonb, and Nan Niua. 2021. On the Role of Data Engineering Decisions in AI-Based Applications. (Feb. 2021). https://doi.org/10.5281/ZENODO.4818970

[7] Valentina Bellini, Marco Cascella, Franco Cutugno, Michele Russo, Roberto Lanza, Christian Compagnone, and Elena Giovanna Bignami. 2022. Understanding Basic Principles of Artificial Intelligence: A Practical Guide for Intensivists: Basic Principles of Artificial Intelligence. *Acta Biomedica Atenei Parmensis* 93, 5 (Oct. 2022), e2022297. https://doi.org/10.23750/abm.v93i5.13626

[8] Rolf Biehler and Yannik Fleischer. 2021. Introducing Students to Machine Learning with Decision Trees Using CODAP and Jupyter Notebooks. *Teaching Statistics* 43 (2021), S133–S142.

[9] H. Payne Blakeley and Cynthia Breazeal. 2019. *An Ethics of Artificial Intelligence: Curriculum for Middle School Students.* MIT Media Lab.

[10] Anna Bobasheva, Fabien Gandon, and Frederic Precioso. 2022. Learning and Reasoning for Cultural Metadata Quality: Coupling Symbolic AI and Machine Learning over a Semantic Web Knowledge Graph to Support Museum Curators in Improving the Quality of Cultural Metadata and Information Retrieval. *Journal on Computing and Cultural Heritage* 15, 3 (Sept. 2022), 1–23. https://doi.org/10.1145/3485844

[11] Jie Cai, Jiawei Luo, Shulin Wang, and Sheng Yang. 2018. Feature Selection in Machine Learning: A New Perspective. *Neurocomputing* 300 (July 2018), 70–79. https://doi.org/10.1016/j.neucom.2017.11.077

[12] Thomas K. F. Chiu. 2021. A Holistic Approach to the Design of Artificial Intelligence (AI) Education for K-12 Schools. *TechTrends* 65, 5 (Sept. 2021), 796–807. https://doi.org/10.1007/s11528-021-00637-1

[13] Thomas K. F. Chiu, Helen Meng, Ching-Sing Chai, Irwin King, Savio Wong, and Yeung Yam. 2022. Creation and Evaluation of a Pre-tertiary Artificial Intelligence (AI) Curriculum. 65 (2022), 30–39. https://doi.org/10.1109/TE.2021.3085878

[14] Beverly Clarke. 2019. *Artificial Intelligence - Alternate Curriculum Unit.* Exploring Computer Science, University of Oregon.

[15] Daswin De Silva and Damminda Alahakoon. 2022. An Artificial Intelligence Life Cycle: From Conception to Production. *Patterns* 3, 6 (June 2022), 100489. https://doi.org/10.1016/j.patter.2022.100489

[16] Stefania Druga and Amy J Ko. 2021. How Do Children's Perceptions of Machine Intelligence Change When Training and Coding Smart Programs?. In *Interaction Design and Children*. ACM, Athens Greece, 49–61. https://doi.org/10.1145/3459990.3460712

[17] Stefania Druga, Nancy Otero, and Amy J. Ko. 2022. The Landscape of Teaching Resources for AI Education. In *Proceedings of the 27th ACM Conference on on Innovation and Technology in Computer Science Education Vol. 1*. ACM, Dublin Ireland, 96–102. https://doi.org/10.1145/3502718.3524782

[18] Wayne W Eckerson, Nancy Hanlon, and Ramon Barquin. 2000. The CRISP-DM Model: The New Blueprint for Data Mining. 5, 4 (2000).

[19] European Commission. [n. d.]. Content and Technology, Proposal for a Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (AI Act).

[20] Carmen Fernández-Martínez, Isidoro Hernán-Losada, and Alberto Fernández. 2021. Early Introduction of AI in Spanish Middle Schools. A Motivational Study. *KI - Künstliche Intelligenz* 35, 2 (June 2021), 163–170. https://doi.org/10.1007/s13218-021-00735-5

[21] Sarah Friedrich, Gerd Antes, Sigrid Behr, Harald Binder, Werner Brannath, Florian Dumpert, Katja Ickstadt, Hans A. Kestler, Johannes Lederer, Heinz Leitgöb, Markus Pauly, Ansgar Steland, Adalbert Wilhelm, and Tim Friede. 2022. Is There a Role for Statistics in Artificial Intelligence? *Advances in Data Analysis and Classification* 16, 4 (Dec. 2022), 823–846. https://doi.org/10.1007/s11634-021-00455-6

[22] Zoubin Ghahramani. 2004. Unsupervised Learning. In *Advanced Lectures on Machine Learning*, O. Bousquet, G. Raetsch, and U. von Luxburg (Eds.). Springer-Verlag.

[23] Joan Giner-Miguelez, Abel Gómez, and Jordi Cabot. 2022. DescribeML: A Tool for Describing Machine Learning Datasets. In *Proceedings of the 25th International Conference on Model Driven Engineering Languages and Systems: Companion Proceedings*. ACM, Montreal Quebec Canada, 22–26. https://doi.org/10.1145/3550356.3559087

[24] Andreas Grillenberger and Ralf Romeike. 2017. Key Concepts of Data Management: An Empirical Approach. In *Proceedings of the 17th Koli Calling International Conference on Computing Education Research (Koli Calling '17)*. Association for Computing Machinery, New York, NY, USA, 30–39. https://doi.org/10.1145/3141880.3141886

[25] Andreas Grillenberger and Ralf Romeike. 2018. Developing a Theoretically Founded Data Literacy Competency Model. In *Proceedings of the 13th Workshop in Primary and Secondary Computing Education*. ACM, Potsdam Germany, 1–10. https://doi.org/10.1145/3265757.3265766

[26] Andreas Grillenberger and Ralf Romeike. 2019. About Classes and Trees: Introducing Secondary School Students to Aspects of Data Mining. In *Informatics in Schools. New Ideas in School Informatics*, Sergei N. Pozdniakov and Valentina Dagien\.e (Eds.). Springer International Publishing, Cham, 147–158.

[27] Mark Haakman, Luís Cruz, Hennie Huijgens, and Arie Van Deursen. 2021. AI Lifecycle Models Need to Be Revised: An Exploratory Study in Fintech. *Empirical Software Engineering* 26, 5 (Sept. 2021), 95. https://doi.org/10.1007/s10664-021-09993-1

[28] Thomas M. Haladyna. 2004. *Developing and Validating Multiple-choice Test Items* (0 ed.). Routledge. https://doi.org/10.4324/9780203825945

[29] Yikun Han, Chunjiang Liu, and Pengfei Wang. 2023. A Comprehensive Survey on Vector Database: Storage and Retrieval Technique, Challenge. arXiv:2310.11703 [cs]

[30] Julie Henry, Alyson Hernalesteen, and Anne-Sophie Collard. 2021. Teaching Artificial Intelligence to K-12 Through a Role-Playing Game Questioning the Intelligence Concept. *KI - Künstliche Intelligenz* 35, 2 (June 2021), 171–179. https://doi.org/10.1007/s13218-021-00733-7

[31] Mieke Heyvaert, Karin Hannes, and Patrick Onghena. 2016. *Using Mixed Methods Research Synthesis for Literature Reviews*. SAGE, Los Angeles.

[32] Tom Hitron, Yoav Orlev, Iddo Wald, Ariel Shamir, Hadas Erel, and Oren Zuckerman. 2019. Can Children Understand Machine Learning Concepts? The Effect of Uncovering Black Boxes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300645

[33] Robert Hoehndorf and Núria Queralt-Rosinach. 2017. Data Science and Symbolic AI: Synergies, Challenges and Opportunities. *Data Science* 1, 1-2 (Dec. 2017), 27–38. https://doi.org/10.3233/DS-170004

[34] Johannes Jakubik, Michael Vössing, Niklas Kühl, Jannis Walk, and Gerhard Satzger. 2024. Data-Centric Artificial Intelligence. *Business & Information Systems Engineering* (March 2024). https://doi.org/10.1007/s12599-024-00857-8

[35] Mohammad Hossein Jarrahi, Ali Memariani, and Shion Guha. 2023. The Principles of Data-Centric AI (DCAI). *Commun. ACM* 66, 8 (Aug. 2023), 84–92. https://doi.org/10.1145/3571724

[36] Sven Jatzlau, Tilman Michaeli, Stefan Seegerer, and Ralf Romeike. 2019. It s Not Magic After All@ Machine Learning in Snap! Using Reinforcement Learning. *2019 IEEE Blocks and Beyond Workshop (B&B)* (2019), 37–41.

[37] Ken Kahn, R Megasari, E Piantari, and E Junaeti. 2018. AI Programming by Children Using Snap! Block Programming in a Developing Country, Vol. 11082. Springer.

[38] Martin Kandlhofer, Gerald Steinbauer, Sabine Hirschmugl-Gaisch, and Petra Huber. 2016. Artificial Intelligence and Computer Science in Education: From Kindergarten to University. In *2016 IEEE Frontiers in Education Conference (FIE)*. 1–9. https://doi.org/10.1109/FIE.2016.7757570

[39] Keunjae Kim, Kyungbin Kwon, Anne Ottenbreit-Leftwich, Haesol Bae, and Krista Glazewski. 2023. Exploring Middle School Students' Common Naive Conceptions of Artificial Intelligence Concepts, and the Evolution of These Ideas. *Education and Information Technologies* (Jan. 2023). https://doi.org/10.1007/s10639-023-11600-3

[40] Damian Kutzias, Claudia Dukino, Falko Kötter, and Holger Kett. 2023. Comparative Analysis of Process Models for Data Science Projects:. In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence*. SCITEPRESS - Science and Technology Publications, Lisbon, Portugal, 1052–1062. https://doi.org/10.5220/0011895200003393

[41] Irene Lee, Safinah Ali, Helen Zhang, Daniella DiPaola, and Cynthia Breazeal. 2021. Developing Middle School Students' AI Literacy. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*. Association for Computing Machinery, Virtual Event, USA, 191–197. https://doi.org/10.1145/3408877.3432513

[42] Annabel Lindner and Stefan Seegerer. 2019. *AI Unplugged - Unplugging Artificial Intelligence - Activities and Teaching Materialon Artificial Intelligence.* Friedrich-Alexander-Universität Erlangen-Nürnberg.

[43] Yihwa Irene Liou. 1990. Knowledge Acquisition: Issues, Techniques, and Methodology. In *Proceedings of the 1990 ACM SIGBDP Conference on Trends and Directions in Expert Systems - SIGBDP '90*. ACM Press, Orlando, Florida, United States, 212–236. https://doi.org/10.1145/97709.97726

[44] Duri Long and Brian Magerko. 2020. What Is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3313831.3376727

[45] Zhuoyue Lyu, Safinah Ali, and Cynthia Breazeal. 2022. Introducing Variational Autoencoders to High School Students. 36 (2022), 12801–12809. https://doi.org/10.1609/aaai.v36i11.21559

[46] Katie Makar and Andee Rubin. 2022. A Framework for Thinking about Informal Statistical Inference. *Statistics Education Research Journal* 8, 1 (April 2022), 82–105. https://doi.org/10.52041/serj.v8i1.457

[47] Radu Mariescu-Istodor and Ilkka Jormanainen. 2019. Machine Learning for High School Students. *Proceedings of the 19th Koli Calling International Conference on Computing Education Research* (2019).

[48] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2022. A Survey on Bias and Fairness in Machine Learning. *Comput. Surveys* 54, 6 (July 2022), 1–35. https://doi.org/10.1145/3457607

[49] Tilman Michaeli, Ralf Romeike, and Stefan Seegerer. 2022. What Students Can Learn About Artificial Intelligence – Recommendations for K-12 Computing Education. In *Towards a Collaborative Society Through Creative Learning*, Keane, Therese, Lewin, Cathy, Brinda, Torsten, and Bottino, Rosa (Eds.). Springer Nature Switzerland, Cham, 196–208. https://doi.org/10.1007/978-3-031-43393-1_19

[50] Robert Monarch and Christopher D. Manning. 2021. *Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-Centered AI.* Sherlter Island, NY.

[51] Tsz Kit Ng and Kai Wa Chu. 2021. Motivating Students to Learn AI Through Social Networking Sites: A Case Study in Hong Kong. *Online Learning* 25, 1 (March 2021). https://doi.org/10.24059/olj.v25i1.2454

[52] Narges Norouzi, Snigdha Chaturvedi, and Matthew Rutledge. 2020. Lessons Learned from Teaching Machine Learning and Natural Language Processing to High School Students. *Proceedings of the AAAI Conference on Artificial Intelligence*

34, 09 (April 2020), 13397–13403. https://doi.org/10.1609/aaai.v34i09.7063
[53] Viktoriya Olari, Kostadin Cvejoski, and Øyvind Eide. 2021. Introduction to Machine Learning with Robots and Playful Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 17 (2021), 15630–15639.
[54] Viktoriya Olari, Kamilla Tenório, and Ralf Romeike. 2023. Introducing Artificial Intelligence Literacy in Schools: A Review of Competence Areas, Pedagogical Approaches, Contexts and Formats. In *Towards a Collaborative Society Through Creative Learning*, Therese Keane, Cathy Lewin, Torsten Brinda, and Rosa Bottino (Eds.). Vol. 685. Springer Nature Switzerland, Cham, 221–232. https://doi.org/10.1007/978-3-031-43393-1_21
[55] Luci Pangrazio and Neil Selwyn. 2019. 'Personal Data Literacies': A Critical Literacies Approach to Enhancing Understandings of Personal Digital Data. *New Media & Society* 21, 2 (2019), 419–437.
[56] Shruti Priya, Shubhankar Bhadra, Sridhar Chimalakonda, and Akhila Sri Manasa Venigalla. 2022. *ML-Quest*: A Game for Introducing Machine Learning Concepts to K-12 Students. *Interactive Learning Environments* (June 2022), 1–16. https://doi.org/10.1080/10494820.2022.2084115
[57] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 1776–1826. https://doi.org/10.1145/3531146.3533231
[58] Tejal Reddy, Randi Williams, and Cynthia Breazeal. 2021. Text Classification for AI Education. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*. Association for Computing Machinery, 1381. https://doi.org/10.1145/3408877.3439689
[59] Chantel Ridsdale, James Rothwell, Mike Smit, Michael Bliemel, Dean Irvine, Daniel Kelley, Stan Matwin, Brad Wuetherick, and Hossam Ali-Hassan. 2015. Strategies and Best Practices for Data Literacy Education Knowledge Synthesis Report. (2015). https://doi.org/10.13140/RG.2.1.1922.5044
[60] Saman Rizvi, Jane Waite, and Sue Sentance. 2023. Artificial Intelligence Teaching and Learning in K-12 from 2019 to 2022: A Systematic Literature Review. *Computers and Education: Artificial Intelligence* (June 2023), 100145. https://doi.org/10.1016/j.caeai.2023.100145
[61] Juan David Rodríguez-García, Jesús Moreno-León, Marcos Román-González, and Gregorio Robles. 2021. Evaluation of an Online Intervention to Teach Artificial Intelligence with LearningML to 10-16-Year-Old Students. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*. Association for Computing Machinery, Virtual Event, USA, 177–183. https://doi.org/10.1145/3408877.3432393
[62] David Rolnick, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, Alexandra Sasha Luccioni, Tegan Maharaj, Evan D. Sherwin, S. Karthik Mukkavilli, Konrad P. Kording, Carla P. Gomes, Andrew Y. Ng, Demis Hassabis, John C. Platt, Felix Creutzig, Jennifer Chayes, and Yoshua Bengio. 2023. Tackling Climate Change with Machine Learning. *Comput. Surveys* 55, 2 (Feb. 2023), 1–96. https://doi.org/10.1145/3485128
[63] Stuart J. Russell, Peter Norvig, Ming-wei Chang, Jacob Devlin, Anca Dragan, David Forsyth, Ian Goodfellow, Jitendra Malik, Vikas Mansinghka, Judea Pearl, and Michael J. Wooldridge. 2022. *Artificial Intelligence: A Modern Approach* (fourth edition, global edition ed.). Pearson, Harlow.
[64] Alpay Sabuncuoglu. 2020. Designing One Year Curriculum to Teach Artificial Intelligence for Middle School. In *Proceedings of the 2020 ACM Conference on Innovation and Technology in Computer Science Education*. ACM, Trondheim Norway, 96–102. https://doi.org/10.1145/3341525.3387364
[65] Jasmine B. Sami, Zachary Stein, Krystin Sinclair, and Larry Medsker. 2020. Data Science Outreach Educational Program for High School Students Focused in Agriculture. *Journal of STEM Education: Innovations and Research* 21, 1 (June 2020).
[66] Rachel Schutt and Cathy O'Neil. 2013. *Doing Data Science* (first edition ed.). O'Reilly Media, Beijing ; Sebastopol.
[67] Gilad Shamir and Ilya Levin. 2022. Teaching Machine Learning in Elementary School. *International Journal of Child-Computer Interaction* 31 (March 2022), 100415. https://doi.org/10.1016/j.ijcci.2021.100415
[68] Aaron Sloman. 2009. Teaching AI and Philosophy at School ?
[69] Alfred Z. Spector, Norvig, Peter, Wiggins, Chris, and Wing, Jeannette M. 2022. *Data Science in Context: Foundations, Challenges, Opportunities* (1st ed ed.). Cambridge University Press, United Kingdom, USA, Australia, India, Singapore.
[70] Rudi Studer, V.Richard Benjamins, and Dieter Fensel. 1998. Knowledge Engineering: Principles and Methods. *Data & Knowledge Engineering* 25, 1-2 (March 1998), 161–197. https://doi.org/10.1016/S0169-023X(97)00056-6
[71] Thordis Sveinsdottir, Pinelopi Troullinou, Stergios Aidlinis, Alexandra Delipalta, Rachel Finn, Panagiotis Loukinas, Julia Muraszkiewicz, Ryan O'Connor, Katrina Petersen, Michael Rovatsos, Nicole Santiago, Diana Sisu, Mistale Taylor, and Peter Wieltschnig. 2020. *The Role of Data in AI.* Technical Report. Zenodo. https://doi.org/10.5281/ZENODO.4312907
[72] Danny Tang. 2019. *Empowering Novices to Understand and Use Machine Learning with Personalized Image Classification Models, Intuitive Analysis Tools, and MIT App Inventor.* Ph. D. Dissertation. Massachusetts Institute of Technology.
[73] Matti Tedre, Peter Denning, and Tapani Toivonen. 2021. CT 2.0. In *Proceedings of the 21st Koli Calling International Conference on Computing Education Research (Koli Calling '21)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3488042.3488053
[74] Alexander Thamm, Michael Gramlich, and Alexander Borek. 2020. *The Ultimate Data and AI Guide: 150 FAQs about Artificial Intelligence, Machine Learning and Data.* Data AI Press, München.
[75] David Touretzky, Christina Gardner-McCune, Fred Martin, and Deborah Seehorn. 2019. Envisioning AI for K-12: What Should Every Child Know about AI?. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9795–9799. https://doi.org/10.1609/aaai.v33i01.33019795
[76] Touretzky, David. 2024. *Big Idea 1 – Perception.* Technical Report.
[77] Touretzky, David. 2024. *Big Idea 2 – Representation & Reasoning.* Technical Report.
[78] Touretzky, David. 2024. *Big Idea 3 – Learning.* Technical Report.
[79] Touretzky, David. 2024. *Big Idea 4 – Natural Interaction.* Technical Report.
[80] Touretzky, David. 2024. *Big Idea 5 – Societal Impact.* Technical Report.
[81] UNESCO. 2022. *K-12 AI Curricula: A Mapping of Government-Endorsed AI Curricula.* Technical Report ED-2022/FLI-ICT/K-12. UNESCO, Paris. 60 pages.
[82] Jessica Van Brummelen. 2019. *The Popstar, the Poet, and the Grinch: Relating Artificial Intelligence to the Computational Thinking Framework with Block-based Coding.*
[83] Jessica Van Brummelen, Tommy Heng, and Viktoriya Tabunshchyk. 2021. Teaching Tech to Talk: K-12 Conversational Artificial Intelligence Literacy Curriculum and Development Tools. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 17 (May 2021), 15655–15663. https://doi.org/10.1609/aaai.v35i17.17844
[84] Henriikka Vartiainen, Tapani Toivonen, Ilkka Jormanainen, Juho Kahila, Matti Tedre, and Teemu Valtonen. 2020. Machine Learning for Middle-Schoolers: Children as Designers of Machine-Learning Apps. In *2020 IEEE Frontiers in Education Conference (FIE)*. IEEE, Uppsala, Sweden, 1–9. https://doi.org/10.1109/FIE44824.2020.9273981
[85] Henriikka Vartiainen, Tapani Toivonen, Ilkka Jormanainen, Juho Kahila, Matti Tedre, and Teemu Valtonen. 2021. Machine Learning for Middle Schoolers: Learning through Data-Driven Design. *International Journal of Child-Computer Interaction* 29 (Sept. 2021), 100281. https://doi.org/10.1016/j.ijcci.2021.100281
[86] Xiaoyu Wan, Xiaofei Zhou, Zaiqiao Ye, Chase K. Mortensen, and Zhengyan Bai. 2020. SmileyCluster: Supporting Accessible Machine Learning in K-12 Scientific Discovery. *Proceedings of the Interaction Design and Children Conference* (2020).
[87] Karl Werder, Balasubramaniam Ramesh, and Rongen (Sophia) Zhang. 2022. Establishing Data Provenance for Responsible Artificial Intelligence Systems. *ACM Transactions on Management Information Systems* 13, 2 (June 2022), 1–23. https://doi.org/10.1145/3503488
[88] Randi Williams, Hae Won Park, Lauren Oh, and Cynthia Breazeal. 2019. PopBots: Designing an Artificial Intelligence Curriculum for Early Childhood Education. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9729–9736. https://doi.org/10.1609/aaai.v33i01.33019729
[89] Jeannette M. Wing. 2020. Ten Research Challenge Areas in Data Science. *Harvard Data Science Review* 2, 3 (Sept. 2020). https://doi.org/10.1162/99608f92.c6577b1f
[90] Yuanhao Xie, Luís Cruz, Petra Heck, and Jan S. Rellermeyer. 2021. Systematic Mapping Study on the Machine Learning Lifecycle. In *2021 IEEE/ACM 1st Workshop on AI Engineering - Software Engineering for AI (WAIN)*. IEEE, Madrid, Spain, 70–73. https://doi.org/10.1109/WAIN52551.2021.00017
[91] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, and Xia Hu. 2023. Data-Centric AI: Perspectives and Challenges. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, Shashi Shekhar, Zhi-Hua Zhou, Yao-Yi Chiang, and Gregor Stiglic (Eds.). Society for Industrial and Applied Mathematics, Philadelphia, PA, 945–948. https://doi.org/10.1137/1.9781611977653
[92] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. 2023. Data-Centric Artificial Intelligence: A Survey. arXiv:2303.10158 [cs]
[93] Alice Zheng and Amanda Casari. 2018. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists.* O'Reilly, Beijing Boston Farnham Sebastopol Tokyo.
[94] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning Deep Features for Discriminative Localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2921–2929. https://doi.org/10.1109/CVPR.2016.319