Computers and Education Open 7 (2024) 100196

Contents lists available at ScienceDirect

Computers and Education Open

journal homepage: www.sciencedirect.com/journal/computers-and-education-open

Data-related concepts for artificial intelligence education in K-12

Viktoriya Olari [*], Ralf Romeike

Freie Universität Berlin, Computing Education Research Group, Königin-Luise-Str. 24-26, Berlin, 14195, Germany

ARTICLE INFO

Keywords:
Artificial Intelligence education
Computer Science education
K-12
Data
Data lifecycle
Key concepts

ABSTRACT

Due to advances in Artificial Intelligence (AI), computer science education has rapidly started to include topics related to AI along K-12 education. Although this development is timely and important, it is also concerning because the elaboration of the AI field for K-12 is still ongoing. Current efforts may significantly underestimate the role of data, the fundamental component of an AI system. If the goal is to enable students to understand how AI systems work, knowledge of key concepts related to data processing is a prerequisite, as data collection, preparation, and engineering are closely linked to the functionality of AI systems. To advance the field, the following research provides a comprehensive collection of key data-related concepts relevant to K-12 computer science education. These concepts were identified through a theoretical review of the AI field, aligned through a review of AI curricula for school education, evaluated through interviews with domain experts and teachers, and structured hierarchically according to the data lifecycle. Computer science educators can use the elaborated structure as a conceptual guide for designing learning arrangements that aim to enable students to understand how AI systems are created and function.

1. Introduction

Advances in Artificial Intelligence (AI) over the past few decades are increasingly changing the technology landscape. Software that works with data is gaining the ability to generate new content, predict future events, and make suggestions tailored to user profiles, among other things. To address these developments and prepare society to responsibly work with and shape AI technologies, AI is increasingly introduced as a topic in computer science education around the world. However, before introducing a new topic for teaching, it is essential to identify its central concepts [1].

In computing education research, there is a strong consensus that teaching should focus on key concepts of the subject rather than on short-lived technological developments. For this reason, catalogs of ideas, concepts, and principles of computer science and its subfields have been developed over the past decades. Well known are, for instance, Fundamental Ideas of Computer Science [2], Great Principles for Computing [3], Big Ideas in Computer Science for K-12 education [4], and Key Concepts of Data Management [1]. These catalogs provide insights into key aspects of the field and can be used to prepare topics for teaching or as a basis for developing computer science curricula [1].

For the area of AI, the conceptualization of the field for K-12 is still ongoing. Several catalogs of competencies, ideas, and design principles have been proposed [5–8]. However, systematization of the field with respect to data, the most fundamental component of AI systems, that educators can draw on when planning lessons, is still lacking [9]. When teaching students about the functionality and limitations of AI systems, the role of data is of significant importance [10–13].

In order to expand the knowledge of data-related concepts in AI for school education, we conducted a theoretical analysis of the AI field, corroborated the results with experts, and contrasted the findings with AI school curricula. The following overarching research question with two sub-questions guided us through the process:

RQ: What data-related concepts are essential when creating an AI system in the context of AI education for K-12?

- What are essential concepts related to data processing when creating an AI system?
- How can the identified concepts be aligned with AI education in K-12?

The reminder of the paper is organized as follows: First, we present the theoretical foundations of our work. In Section 2, we discuss the role of data in the development of AI systems. In Sections 3 and 4, we review previous theoretical work on data-related AI education for students and on characterizing a subject area through underlying concepts. In Section

[*] Corresponding author.
E-mail address: viktoriya.olari@fu-berlin.de (V. Olari).

https://doi.org/10.1016/j.caeo.2024.100196
Received 15 February 2024; Received in revised form 3 June 2024; Accepted 13 June 2024
Available online 1 July 2024

5, we outline the details of the methodology for identifying key concepts. In Section 6, we present and explain the key data-related concepts that were identified from the theoretical analysis of the field and aligned with previous AI curricula. In Section 7, we discuss our findings and suggest the directions for future research.

2. Data as a core component of AI systems

Data is a core component of software using AI techniques, which include machine learning approaches such as supervised, unsupervised and reinforcement learning, logic- and knowledge-based approaches among others [14]. Supervised learning techniques learn pattern in labeled data with a goal to generalize the pattern for the unseen data. Unsupervised learning techniques work with unlabeled data to separate it into groups that share common characteristics [15]. In reinforcement learning, an agent produces data though the interaction with the environment and learns from these data to perform better actions. In systems using logic- and knowledge-based approaches, the data is manually handcrafted with a goal to represent it as a knowledge, use this knowledge to process new data and derive new facts [16]. At machine level, data is stored digitally on a device in binary values. It comes from different sources, including sensors, machines or humans and at the application level, is represented in different modalities - such as text, image, audio, table or graph.

Understanding how data is processed is essential to understanding how AI systems work, and how reliable they are [10–13]. For logic- and knowledge-based AI systems, the importance of inclusion of multiple data sources and experts during the knowledge acquisition phase to omit bias [17] are known for many years [16,17]. In context of machine learning, research direction of data-centric AI emerged recently [10,11,18]. Compared to the model-centric machine learning, which focuses on identifying more effective models to improve performance of machine learning applications while leaving the data unchanged, data-centric researchers argue that systematic engineering of the data is a key to building an accurate machine learning system [12]. A spectrum of tasks of the data-centric machine learning includes data preparation, data augmentation, data quality assurance, error analysis [13], output monitoring and interpretation [19] among others.

The data processing steps during the development and deployment of AI systems have been described in life cycle models [19–23] with Cross-Industry Standard Process for Data Mining (CRISP-DM) model being one of the industry and academic baselines [19]. CRISP-DM model is an industry-, tool-, and application-neutral model that provides a blueprint consisting of six key stages: business understanding, data understanding, data preparation, modeling, evaluation, deployment [15]. Because data processing continues after the deployment of an AI system [21,24], in addition to these six stages, it is necessary to consider the inclusion of data collection, monitoring, sharing/archiving/deleting data as essential components of the data lifecycle.

3. Introducing AI as a topic in computer science school education

AI is increasingly being included as a topic in K-12 computer science curricula around the world. By 2021, AI curricula in school education have been endorsed by the governments of 11 UNESCO member states at various levels of school education [25]. Policymakers are also updating digital education recommendations. For example, the European Union recently published an update of the European Digital Competence Framework, DigComp 2.2. The document includes a list of more than 80 examples of knowledge, skills, and attitudes related to citizens interacting with AI systems [26]. In comparison, the first version from 2013 did not include any of these [27].

While these developments are timely and important, they are also concerning because the elaboration of the field of AI for K-12 is still ongoing. Since 2015, the body of research on AI education has been growing rapidly [28]. Researchers are developing and evaluating new tools [29], and experimenting with teaching the inner workings of AI algorithms [30–32]. However, elaborated competency models are just beginning to emerge [6,8,33]. Few studies target computer science students, who are expected to engage more deeply with the subject than students in other disciplines [34]. As AI technologies bring several fundamental changes to software development [7], there is still much work to be done. Currently, little research has focused on understanding the enduring key concepts of AI for K-12 [5], a step that is important before introducing new topics into computer science curricula.

4. Characterizing the AI field for K-12 through concepts

Characterizing a domain by concepts and underlying ideas has a long tradition in the sciences and has become a common approach in computer science [1,3,4]. In the sciences, concepts are described as systematic mental representations of the real world [35]. They can be *observable* (e.g., ”mammal”), *unobservable* (e.g., ”atom”), or they can be *related to processes* (e.g., ”photosynthesis”). In computer science, concepts have been described in terms of ideas and principles. For example, Schwill suggests that a fundamental idea in a domain is a schema for thinking, acting, describing, or explaining [2]. As such, it must be applicable or observable in multiple ways in different areas of the domain, can be demonstrated and taught at any intellectual level, can be clearly observed in the historical development of the domain, and is related to everyday language and thought. Denning emphasizes that besides focusing on lifeless and abstract concepts, it is essential to capture the *principles* - the mechanics of a discipline, the principles of design distilled from recurring patterns observed in practice (e.g., programming, engineering, innovating) [3]. Another important feature is that conceptual knowledge is *product independent* [36]. It enables students to understand a subject in a broader context and to transfer skills.

In AI education, several proposals have been made to capture the essence of the AI field for K-12. Touretzky et al. proposed a set of five big ideas [5]. The ideas focus on essential capabilities of AI systems such as perception, representation, reasoning, learning, natural interaction and societal impact. Tedre et al. elaborated on conceptual shifts in computational thinking for K-12 education and showed differences between traditional programming education and education focused on building machine learning models [7]. In addition to these systematizations of the field, a number of competency models for AI education have been proposed [6,8].

All these elaborations emphasize the role of data in AI systems. For example, the Big Idea ”Learning” encourages students to understand that computers learn from data and that machine learning is about statistical inference that finds patterns in data [5]. There is also an emerging line of research exploring the role of data literacy and data agency in learning about AI [37,38]. However, a recent literature review concluded that teaching approaches only scratch the surface of working with data, and competency models for data literacy, while providing a foundation for working with data in the context of AI, lack concepts inherent to AI technologies such as model development [9]. For these reasons, we see an urgent need to advance the field of AI education by systematizing it from the perspective of its fundamental component - data.

5. Methods

The objective of this work is to identify data-related concepts that are central to the creation of an AI system and that can be used by educators to structure AI education curricula and plan lessons. To achieve this goal, the following steps were necessary: (1) criteria-based specification of the construct ”data-related concept”, (2) systematic analysis of the domain and extraction of potential concepts from the literature, (3) alignment of the concepts identified in the literature with K-12 AI education, (4) evaluation of the aligned concepts with the domain experts and teachers. The process was iterative, as shown in Fig. 1. In the following, we report details on each of the stages.

5.1. Criteria-based specification of the construct "data-related concept"

For systematic identification of data-related concepts in literature, the construct ”data-related concept” must be characterized through criteria. Following the characterization of a concept in sciences, as outlined in Section 4, and the data lifecycle model that is used as a blueprint to process data in AI projects, as described in Section 2, *a data-related concept instantiates in a term that, in the specialist community of AI and data scientists, is used as a placeholder to describe a certain process (e.g., data augmentation), an observable entity (e.g., image) or an unobservable entity (e.g. data modality) which is related to data processing at one of the following stages of the data lifecycle: understand the task, collect data, understand data, prepare data, implement solution, evaluate performance, deploy and monitor, share/delete/archive data.* Following this specification, *IP Protocol* would not be regarded as a data-related concept as it is primarily used in the context of routing data packages across networks and does not directly relate to one of the stages of the data lifecycle [39]. *Data label* would be regarded as such because it is a critical component in the data preparation stage. During this stage, annotators provide meaningful labels to data when preparing it for use by an AI algorithm [23].

The concept is considered as central if it fulfills three criteria derived from prior research in computer science education, as described in Section 4:

1. **Product independency.** The concept must be independent of a product. For instance, *Apache Hadoop* [40] does not fulfill this criterion, as it stands for a concrete software library. However, *distributed data storage* fulfills it.
2. **Time stability.** The concept must be observable in the historical development of the domain. For instance, *prompting* does not fulfill this criterion, as it become popular recently [41] and is unclear whether it remains relevant over time. *Hypothesis testing* fulfills this criterion because it is a concept known for years from the field of statistics and still relevant in the AI and data science field [42].
3. **Conceptual clarity.** The concept must be universal and unambiguous. For instance, *Frankenstein dataset* does not fulfill this criterion because besides describing a dataset that combines data from apparently distinct sources while being from the same source [43], it can also describe a dataset made of synthetically generated data [44]. *Redundant data* and *synthetic data* [13,45] fulfill this criterion, as these are more universal and conceptually clear concepts.

5.2. Systematic analysis of the theoretical literature and extraction of potential concepts

Identifying concepts is challenging for several reasons. First, the concepts are not readily apparent because they can be found at different levels of abstraction, ranging from highly technical terms (e.g., *discriminative feature*) to more abstract theoretical ideas (e.g., *data protection*). Second, advances in AI technologies and data science are constantly introducing new concepts (e.g., *MLOps* [46]), which increases the complexity of identifying stable, long-lasting concepts. Therefore, deep immersion in the domain is required to ensure comprehensive understanding and identification of essential concepts. We began the process with an in-depth analysis of the theoretical sources on data processing in the development of AI systems. For selecting the literature, we followed the purposeful sampling strategy since our objective was neither to include all existing relevant studies (exhaustive strategy) nor to identify all relevant studies withing specified limits (selective strategy) [47]. Purposeful sampling strategy aims to find information-rich studies, in our case, studies with a high density on data-related concepts in AI system development. From the different strategies that purposeful sampling offers, we mostly used theory-based construct sampling, criterion sampling and snowballing, starting with reviewing standard textbooks used in the university education for AI and data science [16,23,48]. We additionally conducted a search at ACM, SpringerLink and ScienceDirect databases with a combination of the keywords ”data”, ”data lifecycle”, ”data-centric AI”, ”data processing”, ”Artificial Intelligence”, ”machine learning”, ”knowledge-based” to find academic papers describing data processing for systems built with both machine learning and logic- and knowledge-based techniques [1,10,11,13,19–22,24,49–53]. In order to include practical perspectives, we also reviewed grey literature that was recommended by AI practitioners [12,15,54–59].

While engaging with the sources, for each of the stages of the data lifecycle, we manually extracted an initial list of potential concepts along with their descriptions ending up with a document of 111 pages, including descriptions of 84 data-related processes, technologies and technical vocabulary among others.

To clarify the relationships between potential concepts and identify redundancies, we mapped the concepts in a hierarchical order considering the occurrences of concepts in the data lifecycle and grouping them into categories. This iterative process involved extensive consultation of additional theoretical literature, continuous revision, rearrangement and amendment of potential concepts, as well as the evaluation of terms that did not meet the criteria described in Section 5.1 for conceptual counterparts. For instance, *feature selection* and *feature extraction* were mapped as subordinate concepts of *feature engineering*. During the consultation of additional literature, we identified that additional subordinate concepts are relevant for feature

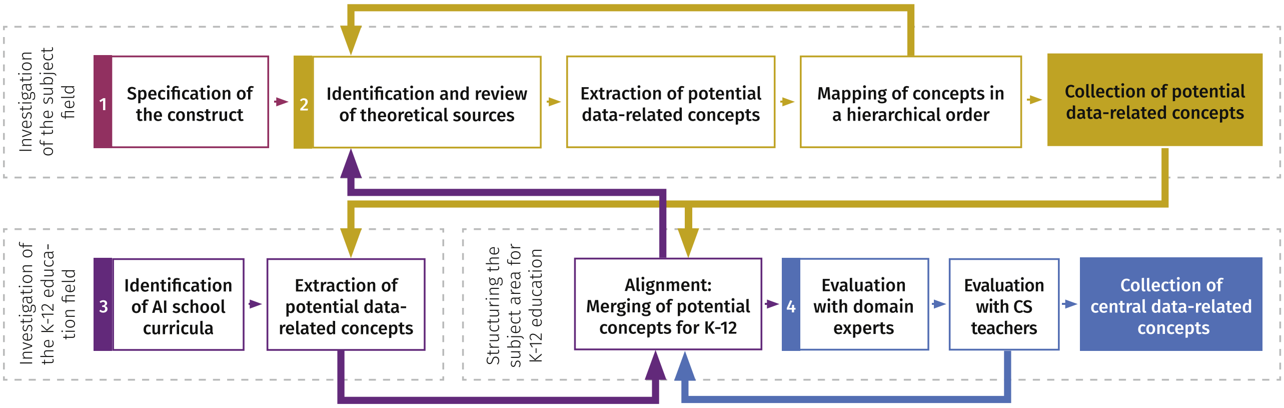
Investigation of the subject field
1
Specification of the construct
2
Identification and review of theoretical sources
Extraction of potential data-related concepts
Mapping of concepts in a hierarchical order
Collection of potential data-related concepts
Investigation of the K-12 education field
3
Identification of AI school curricula
Extraction of potential data-related concepts
Structuring the subject area for K-12 education
Alignment: Merging of potential concepts for K-12
4
Evaluation with domain experts
Evaluation with CS teachers
Collection of central data-related concepts

Fig. 1. Overview of the analysis process.

engineering such as *feature transformation* and *feature reduction* and included them into the collection. *General Data Protection Regulation GDPR* was replaced by *data-related legal regulation*. We continued to refine our collection, until we reached a point of theoretical saturation [60] and existing concepts began to recur. This saturation indicated that our collection of potential concepts was comprehensive and relevant for representing data processing in AI systems.

5.3. Alignment of concepts identified in the literature with AI education for K-12

To adapt the collection of potential concepts embodying a professional perspective on data processing for the use in K-12 education, an alignment was necessary. The alignment ensures that initial collection adequately includes data-related concepts already present in AI education for K-12 and does not miss any essential concept. It also highlights gaps if a concept is present in the collection but absent in AI education for K-12.

To identify data-related concepts in AI education for K-12, we focused on analyzing AI curricula for K-12 education because they provide a strong body of accumulated knowledge of what researchers, teachers, and practitioners consider to be important for school students. To conduct the alignment, four steps were required: (1) collection of AI curricula for K-12, (2) identifying of data-related sections in AI curricula, (3) extraction of data-related concepts from the data-related sections, (4) merging the data-related concepts from the literature with data-related concepts identified in the AI curricula.

To collect AI curricula for K-12, similar to identifying the subject literature, we followed the purposeful sampling strategy since our objective was not to include all existing AI curricula for K-12 [47]. Instead, we aimed to find information-rich curricula that included a substantial description of competencies and learning activities. From the strategies that the purposeful sampling offers, we mostly used the snowballing [47], starting with a review of papers identified in a recent comprehensive literature review on AI education for schools [61]. We included papers for further processing if they contained a description of what and how the students should learn [33]. We also included the two most recent AI curricula that we were aware of from our previous research. After retrieving and closely reading the papers, we excluded 53 papers due to missing curricula, leaving a total of 49 papers. From each paper, we extracted a curriculum, resulting in a large text corpus[1].

In order to identify data-related text sections, we iteratively searched the corpus for the inclusion of keywords used in the context of data (information, input, file, image, picture, foto, photo, photograph, 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 the sentence contained one or more of the keywords, we auto-coded the sentence as data-related by using the MAXQDA software.

To extract the concepts from the data-related text sections, we reviewed each sentence and labeled it with one of the stages of the data lifecycle. From the labeled sentences, we extracted the data-related concept following the criteria defined in Section 5.1 and structured the resulting collection hierarchically. Subsequently, we contrasted the results with our initial collection.

To merge the collections, if the initial collection did not contain a concept, we consulted the literature to understand the relationship between the newly found data-related concept and the concepts present in the initial collection. Our aligned collection of potential concepts included concepts found in both the theoretical literature and AI curricula, concepts found only in AI curricula, and concepts identified only in the theoretical literature.

5.4. Corroborating the results with domain experts and computer science teachers

Relying on the literature can be insufficient, and reducing complexity is a critical step in preparing content for teaching in K-12 computer science education. For these reasons, the collection of potential concepts had to be evaluated by experts for soundness and representativeness, and by computer science teachers for its suitability for classroom use.

For the soundness evaluation, we presented the collection to a data scientist, a domain expert in using AI for climate research, and a data literacy researcher. Two experts provided written feedback, and one provided feedback in an informal interview. All experts welcomed the collection, agreed with its general representativeness, and made additional suggestions to make it more comprehensive. The collection was subsequently updated.

In order to include the practical perspective as much as possible and to reduce the complexity of the collection, we involved computer science teachers. At several stages of the process, we demonstrated versions of the collection to four teachers to understand its general suitability for use in computer science classes and informally discussed their feedback. Teachers expressed the need to include more concrete concepts because the abstract concepts are difficult to teach. They also expressed the need for the collection to be accompanied by coherent examples of how these concepts can be incorporated into a series of lessons.

6. Results

The research resulted in a comprehensive collection of key data-related concepts for AI education in K-12 computer science classes, as illustrated in Fig. 2. The collection is described in detail in the sections that follow. Each section is organized similarly. First, we provide a context for the concepts, highlighting their significance in the overall process of developing AI systems. We then present selected data-related concepts and outline the relationships between them.

6.1. Concepts related to task understanding

Before working on any project that involves the use of AI techniques, it is critical to understand the problem that will be solved using AI. This includes understanding the nature of the task (concept *data-based task*), the specifics of the available data (concept *data modality*), the roles and needs of the stakeholders (concept *data stakeholder*) and defining when the task is complete (concept *success criterion*).

AI technologies can be used to solve a variety of *data-based tasks*. In AI curricula, we identified mentions of tasks such as regression [8,25,62–64], classification [25,33,62,63,65–74], detection and localization [64], segmentation [63], image generation [75], text generation [68,76], audio generation [73], machine translation [68,76], speech recognition [76], text summarization [76], conversational interaction [76], recommendation [69,77,78], reasoning [6,8,33,34,63,64,68,76,78,79], and others.

From the subject perspective, there are many more tasks that can be solved with AI technologies. Theoretical literature suggests that any *data-based task* [57] can be conceptualized in terms of an input space and an output space. For instance, in a *speech recognition task*, the audio signal in the input implies a text transcription in the output [80]. Different data-based tasks have unique data requirements [15,19,81,82]. For example, a *classification task* requires labeled data as the input.

When students learn about AI, they discover that AI systems can work with various types of data, including *image data* [25,63–66,75,77,83–85], *text data* [25,62,65–68,76,86,87], *tabular data* [77], *audio* [64],

[1] The text corpus is available upon request.

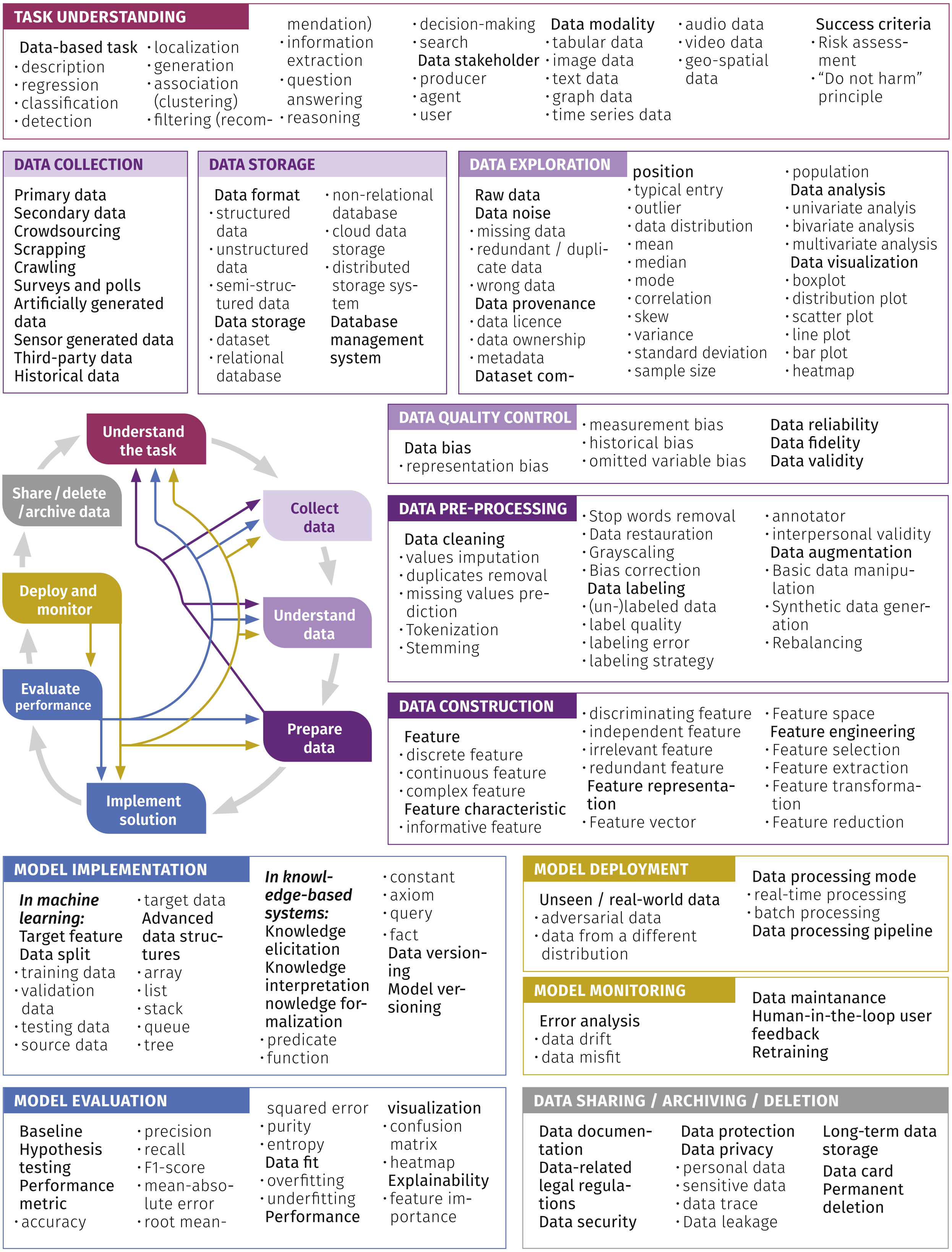

Fig. 2. Overview of the key data-related concepts for creating AI systems along the data lifecycle.

video [88], and *graph data* [68]. Other types of data common in practice are *time series* and *geo-spatial data* [89]. An essential idea suggested by practitioners is that the choice of *data modality*, defined as a particular way or mechanism of encoding information [90], influences the selection of data cleaning and pre-processing methods. For example, when working with image data, machine learning techniques require images to be rescaled and pixel values to be represented as vectors before the data can be processed by the algorithm.

While several AI curricula suggest that school students create and use datasets to develop their own AI applications [62,63,69,78,88,91–94], we found a lack of references on students analyzing stakeholder needs and formulating success criteria, as is common in practice and theoretical literature [55]. A *success criterion* states precisely and clearly when the task is completed. It must reflect the needs of *stakeholders*, including those who will use or be affected by the implemented solution (*users*), those who provide data to create the application (*producers*) and those who will leverage the data to create the application (*agents*) [89]. For instance, *users* of AI-based systems might be interested in how the data was used to create an AI system. They can require a set of explanations and controls grounded within product experiences.

6.2. Concepts related to data collection

Data collection is a fundamental step when creating an AI system. The data can be either collected for the specific purpose of the task (concept *primary data*) or it can be repurposed (concept *secondary data*). Common methods of data collection include automatic techniques (e.g., concept *scrapping*), human-involved approaches (e.g., concept *crowdsourcing*), generating data through simulations (e.g., concept *artificial data generation*).

In the AI curricula, we identified references to automatic data collection (students should explain how AI systems collect data via sensors [8]) and human-involved data collection (e.g., students should create labeled datasets [64]). Collection of *primary data*, while time-consuming, ensures that the data precisely meets the requirements of the specific task. Conversely, utilizing *secondary data*, which is pre-existing, can be more convenient but poses risks such as data being outdated or misaligned with the tasks specific requirements [59]. The chosen data collection techniques can have a significant impact on the quality of the data and, consequently, on the performance of the AI system. For example, *crowdsourcing* as a data collection method, involving many individuals [54] without a standardized strategy, can result in a dataset requiring extensive cleaning and preparation, thereby affecting the efficiency and accuracy of the subsequent AI application.

6.3. Concepts related to data storage

Data used for the development of any AI system, must be digitally stored (concept *data storage*). Because AI systems process data of different formats (concept *data format*), they have specific data storage requirements.

AI curricula suggest that students should be familiar with simple *databases* [25,79] and *relational databases* [25]. They also suggest students to understand advantages and disadvantages of cloud storage [25]. From the subject perspective, *relational databases* have several limitations for AI systems [56], as they primarily store tabular data (*structured data*), require definition of an exact scheme before storing any data and are not easy scalable if the data volume is too large to be stored on a single server. *Non-relational database* systems overcome these issues [56], as they handle *unstructured data* such as images and texts and *semi-structured data* such as XML files and are scalable through a *distributed storage system*. The data models in non-relational database systems are more flexible, so that additional data can be included without having to make changes to the overall schema of the database. The data stored in a database is accessed and manipulated through a corresponding *database management system*.

6.4. Concepts related to data exploration

The data collected in the data collection step is not ready to be used by an AI technique (concept *raw data*). Understanding the errors in the data (concept *data noise*), the origins of the data (concept *data provenance*) and the characteristics of the data (concept *data composition*) is essential to anticipate and resolve data-related problems early, which directly affects the quality and functionality of the final AI system. Data exploration requires knowledge of basic statistical concepts (concept *data analysis*) and various visualization techniques (concept *data visualization*). The result of data exploration is the knowledge about relationships in the data, its problems, and assumptions about the task [95].

In the AI curricula, we identified several references to the data exploration step. For example, school students should understand the concept of *messy data* [6,25], be aware of data origins [25,91], analyze datasets [64], understand data trends [25], find patterns in data and irrelevant correlations [63], create data visualizations [96] such as simple charts [25] and graphs [97].

From the subject perspective, data collected during data collection is *raw data*, meaning it is not yet ready to be used by an AI system. It is messy because it contains *noise* such as missing data (incomplete values), *wrong data* (erroneous values), *duplicate data and redundant data* [57]. All of these issues need to be addressed before the data is used by an AI system.

To avoid problems with AI systems being built on identical or overlapping datasets while assuming they come from distinct sources [43], it is important to understand the *data provenance* - the origin and previous processing of the data, including information about the *data license, ownership*, and *metadata*. Metadata might include information regarding the documentation of data collection and pre-processing techniques including demographics such as who collected the data and who funded it [98].

To understand the structure of the dataset, its *typical data entries, outliers*, and patterns within the data, and to gain an intuition about potential data-related problems [99], the *data composition* is explored using statistical methods [100]. Depending on the data modality, the *data analysis* and *data visualization* may differ. For example, when working with tabular data, data analysis involves calculating the correlations between variables (*bivariate and multivariate analysis*) [100] and visualizing the distribution of the variables to understand the potential skews in the data (univariate analysis).

6.5. Concepts related to data quality control

High quality data is essential for reliable AI systems [13,54]. Verifying data quality involves understanding the consistency of the data (concept *data reliability*), the representativity of the data (concept *data fidelity*), the accuracy of the data (concept *data validity*) and identifying any imbalances in the data (*data bias*).

AI frameworks mention some aspects of *data fidelity*, such as data representativity [63], dataset size [25,63,64,101], data reliability, such as homogeneity [102] and *data validity* such as quality, authenticity, and accuracy of training data [33,91]. Although many AI curricula mention bias [6,8,25,63–65,67,69,71,72,77,85,102], we did not find any occurrence of bias in the context of the data-related concepts.

From the subject perspective, data is high quality when it accurately represents a phenomenon, is collected, stored, and used responsibly, is maintainable over time, is reusable across applications, and has empirical and explanatory power [49]. In this context, *data fidelity* describes how well the dataset represents the reality. *Data reliability* illustrates data consistency, replicability and reproducibility of data. *Data validity* indicates how well the data explains things related to the phenomena captured by the data [49]. Imbalances in data are one of the sources of bias in AI systems. Therefore, when verifying data quality, data should be inspected for biases such as *representation bias*, which arises from how data is sampled from a population during the data collection process [103,104].

6.6. Concepts related to data preprocessing

In order to avoid creating distorted AI systems that can cause harm, issues identified in the earlier stages of the data lifecycle must be resolved during the data preprocessing step [12,105]. Data preprocessing includes several key operations, such as data correction (concept *data cleaning*), data labeling for tasks requiring labeled data (concept *data labeling*), increasing the size of the dataset if the original set is insufficient (concept *data augmentation*).

AI curricula suggest that students should be able to correct the dataset [106]. However, we did not find any further references to practices of data cleaning. From the subject perspective, the cleaning strategy depends on the data modality and the data-based task. For example, for tabular data, *values imputation* with mean or median, *prediction of missing values* using a regression model, *duplicates removal* are common [12]. For text data, typical cleaning strategies include splitting the documents into words, removing punctuation and symbols, making all words lowercase, removing stop words, and stemming words [107]. For image data, noise, blur, and distortion are removed using *image restoration*, which attempts to recover a degraded image by modeling the degradation with prior knowledge [95].

If the dataset size is small - which is an issue for many data-based tasks - data augmentation can be used [12]. We found one cursory reference to data augmentation in the prior AI curricula [72]. From the subject perspective, *data augmentation* is a technique to increase the size and diversity of data by artificially creating variations of the existing data. Common approaches for data augmentation are *basic manipulations* (e.g., scaling, smoothing, rotating, sharpening, contrast enhancement for image data [95]), *generating synthetic data* that closely resembles the existing data. At this stage, actions can be taken to handle *data bias* by, e.g., balancing the data distribution for the minority class (concept *rebalancing*, including *upsampling* and *undersampling*), as suggested by prior theoretical literature [12] and the experts from the AI field.

AI frameworks refer to the concept of *data labeling* [63,68,71,101]. We did not find any references to *label quality, labeling strategy*, and other essential concepts during this process. *Data labeling* refers to the process of assigning one or more descriptive tags (*labels*) to the data entries in a dataset [12], and is required for data-based tasks such as classification, regression, detection, and localization. Labels are created by *annotators* [12]. To avoid *labeling errors* and to ensure the *label quality*, a *labeling strategy* is needed. For example, when conducting labeling of images, the annotator needs to know whether an image label annotation is sufficient or whether a precise object shape annotation is required.

6.7. Concepts related to data construction

For data-based tasks using machine learning techniques, having clean and preprocessed data alone is insufficient for effective use by an AI algorithm. The data must be transformed into a format that the algorithm can process, a process known as feature engineering [58] (concept *feature engineering*). This involves a comprehensive understanding of what constitutes a feature (concept *feature*, concept *feature characteristic*), and the methods for representing the feature in a way that is compatible with the AI algorithm (concept *feature representation*).

AI curricula operate with the concept of *feature* in multiple contexts, such as feature vector [68], multi-dimensional feature space [68,108], feature selection [63,71], feature design [92], feature extraction [64,71], feature encoding [63]. From the subject perspective, a *feature* is a numeric representation of data [57] and can be *discrete, continuous*, or *complex* [109]. A single numeric feature is a *scalar*. An ordered list of features is known as a *feature vector*. Feature vectors sit within a *feature space*, which is a vector space. The input to a machine learning model is represented as a feature vector [57]. Features are closely tied to the model, as some types of models are more appropriate for some types of features than others [109]. In regression and classification tasks, a *target feature* is a feature that is to be predicted with a subset of other features, also called explanatory variables, dependent variables, or predictors [58].

The success of machine learning models depends on feature engineering [58,110]. *Feature engineering* refers to the process of formulating the most appropriate features given the data, the model, and the task [57], which includes *feature selection, extraction, transformation*, and *reduction*. Feature selection is the process of obtaining a subset of features from an original feature set [111]. The features must be *informative, discriminating*, and *independent. Irrelevant* and *redundant* features should be omitted. Feature *extraction* refers to the transformation of the original data to features with strong pattern recognition ability [111]. *Feature transformation* is the process of converting the original features into a new set of features using methods such as normalization or standardization [12]. *Feature reduction* is the process of reducing the complexity of a dataset by reducing the feature size or the sample size while retaining the essential information [12].

6.8. Concepts related to model implementation

During the model implementation, the data is used as an input for an AI algorithm. For data-based tasks employing machine learning, this process encompasses understanding how to partition the data (concept *data split*), how to model the data structures (concept *advanced data structures*) and how to manage versions of the data and model (concepts *data versioning* and *model versioning*). For data-based tasks utilizing logic- and knowledge-based approaches, the process includes manually handcrafting the data (concepts *knowledge elicitation, knowledge interpretation*, and *knowledge formalization*).

AI curricula operate with several data-related concepts in context of model implementation such as training data [63,69,78,88,91,92], testing data [62,63,93,94], evaluation data [71,73] and concepts such data split [71], composition of training data such as its quantity [62]. From the subject perspective, a process of splitting a dataset into training data and testing data is called *data splitt* [12,112]. Experts in the AI field emphasize that the data modality influences the splitting strategy (e.g., it would not be appropriate to randomly split a dataset for time-series data [112], as then the relationships between data points would be lost). *Training data* is used to train the model. Portion of the training set can be used as *validation data* to evaluate the model performance during the training [112]. *Testing data* is used to evaluate a trained model [12]. An important step in this context is a development of the effective testing dataset such as through controlling the distribution of data. In the context of transfer learning, the data is differentiated between *source and target data* [113,114]. The former refers to the data used for pre-training the model while the latter for fine-tuning the model [113]. AI curricula mention that working with data also requires knowledge of advanced *data structures* [25] such as *stacks, queues* and *trees* [34,115]. Working with other complex data structures such as *arrays* and *lists* is common.

AI curricula refer to data-based tasks in context of logic- and knowledge-based systems, such as *reasoning* (incl. logical deduction) [6,8,33,34,63,64,68,76,78,79] and *decision-making* [68,115]. However, we did not find any references to data processing during the model implementation. From the subject perspective, developing a knowledge-based system requires *knowledge elicitation* (acquiring and storing informal descriptions of the knowledge about the specific domain and the problem-solving process in knowledge-protocols), *knowledge interpretation* (representing the knowledge structures in a semi-formal way) [20] and *knowledge formalization* (expressing the natural language text in the formal specification language). Knowledge formalization includes understanding of components such as a *predicate, function, constant, fact* and *axiom* [16].

6.9. Concepts related to model evaluation

Model evaluation is essential for assessing the accuracy and robustness of a model on testing data. This process involves understanding how to measure model performance (concepts *performance metrics* and *baseline*), visualize the results (concept *performance visualization*), and interpret the outcomes (concepts *data fit* and *explainability*).

In AI curricula, we identified references to evaluating models based on *accuracy* [25,62,63,69,73,78,93]. However, accuracy is only one of many evaluation metrics, and other p*erformance metrics* may be more appropriate depending on the task according to the AI experts. Common metrics include *precision, recall, F1-score, root mean-squared error, purity*, and *entropy* [19]. AI curricula also discuss whether the data is effectively represented by the model, described by concepts such as *data fit* [13], *overfitting* and *underfitting* [43,85]. From the subject perspective, *overfitting* occurs when the model learns noise in addition to true regularities, whereas *underfitting* happens when the model fails to learn the true patterns in the data [116].

If a model performs poorly, improvements can be made by altering the data or the model itself. Improvements are compared to a *baseline* model, which is created with the simplest applicable AI algorithm. This comparison helps determine the effectiveness of the modifications. Visualizing the model's performance and understanding its decisions are also crucial steps during the model evaluation stage. One method of explaining the model decisions is *feature importance*, which identifies the most influential features in the model's predictions [117].

6.10. Concepts related to deployment and monitoring

After successful performance evaluation, the model is deployed to process unseen data (concept *real-world data*). To ensure accurate performance on unseen data, it is essential to preprocess the data before input into the model (concepts *data-processing pipeline* and *data processing mode*). Maintaining the model's accuracy over time requires rigorous documentation of model performance (concept *error analysis*), ongoing *data maintenance*, and incorporating human feedback [54,118] (concept *human-in-the-loop*).

We did not find references to data processing during the deployment and monitoring stages in AI curricula. However, from the subject perspective, it is necessary to preprocess the incoming, *real-world data* (e.g., cleaning, transforming) in the same way as preprocessing the original data used to create an AI model, which requires the establishment of *data processing pipelines* [56] and deciding between the *real-time* or *batch mode processing* of the data (e.g., once a day) [19]. Because the world changes, continuous monitoring of model and data performance is crucial. For instance, if *error analysis* indicates shifts in input data distribution leading to performance degradation (*data drift*), the model must be updated [13]. The process of maintaining the quality and reliability of data, such as providing infrastructures and data debugging possibilities, is called *data maintenance* [12].

6.11. Concepts related to sharing, deleting and archiving data

In the final step, data is either archived, deleted, or shared for further processing. If the data is archived or shared, documenting all prior steps of data processing is essential for reusing data in future AI systems [119] (concept *data documentation*).

We found some references to *data documentation* in prior AI frameworks [6,79]. From the subject perspective, when data is shared or archived, it must be accompanied by a comprehensive documentation describing data provenance, characteristics, composition, discovered issues, preprocessing steps, augmentation, and modeling, among other aspects [89]. Proper security measures must be implemented to prevent *data leakage* when sharing or archiving data. Key concepts at this stage include *data privacy*, which involves protecting individuals' privacy rights concerning personal data [56] and *data security*, which entails safeguarding data from destructive forces and unauthorized actions through authentication, access control, and encryption [56]. When deleting data, it is crucial to ensure *permanent removal* from all data storage [120].

7. Discussion

Through our analysis of the literature and AI curricula for K-12, we identified a collection of data-related concepts essential for the creation of AI systems within the context of AI education for K-12. Our work differs from other conceptualizations of the AI field in two major ways: (1) it focuses on identifying essential concepts related to data as the fundamental component of AI systems; (2) in addition to analyzing current AI curricula - as is common in other theoretical reviews of AI education [61,121,122] - it is based on an in-depth analysis of the theoretical literature, consultations with domain experts and computer science teachers. During this process, we identified many data-related concepts already present in the AI curricula. However, based on the theoretical work, we see more potential for including the data-related concepts into K-12 education. Here, we discuss the implications of what we observed during the process.

AI curricula contain a wide variety of data-based tasks. However, there is a limited guidance on what the success criteria for fulfilling the tasks are as well as how the data used to solve the task need to be prepared, constructed, and processed. Despite the importance of data collection in any AI project, AI curricula address corresponding concepts in a cursory manner. However, the collection of one's own data can serve to illustrate the difficulties inherent in the collection of representative data and facilitate a deeper understanding of the potential sources of bias. Prior research suggests that working with one's own data can enhance students' learning of AI [123,124].

A notable discrepancy exists between the theoretical work that underscores the significance of data quality control and the lack of references to concepts related to data quality control in AI curricula. Similarly, there is a discrepancy between the theoretical work that emphasizes the significance of data cleaning, data pre-processing, data augmentation and data construction and the lack of references to the corresponding concepts in AI curricula. Further research is necessary to elaborate how the concepts related to these processes can be introduced into AI education, as these are of paramount importance to the functionality and reliability of AI systems and might counteract one of the common naive conceptions of school students that all data can be used by AI [125]. Perhaps it is possible to draw upon previous research on data literacy, where data cleaning and preprocessing is a topic of interest [1,126].

Data labeling is already a component of AI education. However, we note that it appears to be insufficiently elaborated, as references to data labeling strategies and labeling errors are lacking in AI curricula, all of which are potential sources of data bias. Future research could elaborate on activities that teach intuition and difficulties in the labeling process.

The use of data to build machine learning models is covered in many AI curricula. However, because many AI curricula devote less attention to data cleaning, data pre-processing, and data construction, students may develop the erroneous belief that raw data is an accurate representation of reality, as prior research has shown [125]. Future research could elaborate on prioritizing the effective development of data before its use with AI algorithms, taking into account the specifics of data modalities.

We have observed that AI curricula emphasize the role of data in the context of tasks that can be solved with machine learning. However, many modern AI systems use a combination of machine learning methods and knowledge-based approaches [127]. Therefore, understanding how data is manually processed to build knowledge is essential for the design of AI systems. Although AI curricula refer to data-based tasks that are solved by means of logic- and knowledge-based systems, we did not find references to concepts on how to handcraft data when modeling such systems, which is another potential area for future research.

We identified a limited number of references to model evaluation. There is a clear need for future research to elaborate further on how students can be equipped with the means of evaluating model performance and explaining model outcomes. These skills are of critical importance not only in the context of creating AI systems but also in the responsible use of such systems.

The deployment, monitoring, sharing, deletion, and archiving of data is not prominent in AI curricula. However, prior research indicates that it is possible to deploy AI-based systems in a school context using tools such as App Inventor [66], suggesting that future research could elaborate on approaches to integrate essential concepts from monitoring, sharing, deleting, and archiving data into AI school education. This could raise awareness among students that even accurate AI systems often perform poorly on unseen data, or need to be updated over time to adapt to changes in the world.

8. Limitations

The results of the research are subject to several limitations. Prior research indicates that the creation of collections of concepts, as presented in this paper, is influenced by the individual performing the procedure. Additionally, there is no established criterion to prove the completeness of a collection [2]. To address these limitations and ensure the high validity and representativeness of the collection, we evaluated the collection with domain experts. Nevertheless, curriculum developers should be aware that, depending on the data-based task or data modality, the knowledge of additional data-related concepts is required. Furthermore, curriculum developers should be aware that the collection presents the upper limit of what is possible in upper-secondary computer science school education as it was created additive and in consultation with computer science teachers. When planning lessons, it is possible to imagine that the top concepts can be introduced in the lower classes, with the deeper concepts being introduced in higher grades. For example, data collection can be introduced in primary school, while data crawling may be a topic in upper secondary school.

9. Conclusion

The goal of this research was to systematize the AI education field for K-12 from the perspective of data, the most fundamental component of AI systems. The resulting collection of data-related concepts provides a rigorous framework that is of interest to a broad audience. Researchers and curriculum developers can benefit from this work as the identified key concepts can serve as a foundation for developing AI curricula that adequately consider the role of data. Computer science teachers can use the collection as a reference for terminology when developing data-focused lesson series on AI. For future work, we are collaborating closely with school students to evaluate the collection in a data-centered AI course and anticipate evolving it over time.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used DeepL Translate, DeepL Write, ChatGPT in order to check the spelling and grammar and improve the fluency of the text. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

CRediT authorship contribution statement

Viktoriya Olari: Writing – original draft, Writing – review & editing, Visualization, Project administration, Methodology, Formal analysis, Investigation, Data curation, Conceptualization. **Ralf Romeike:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Methodology, Validation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known conflict of interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Grillenberger A, Romeike R. Key concepts of data management: An empirical approach. Koli Calling '17: proceedings of the 17th koli calling international conference on computing education research. New York, NY, USA: Association for Computing Machinery; 2017. p. 30–9. https://doi.org/10.1145/3141880.3141886.ISBN 978-1-4503-5301-4

[2] Schwill A. Fundamental ideas of computer science. Bull European Assoc Theor Comput Sci 1994:53.

[3] Denning PJ. Great principles of computing. Commun ACM 2003;46(11):15–20. https://doi.org/10.1145/948383.948400.

[4] Bell T, Tymann P, Yehudai A. The big ideas in computer science for K-12 curricula. Bull EATCS 2018;124:36–46.

[5] Touretzky D, Gardner-McCune C, Martin F, Seehorn D. Envisioning AI for K-12: what should every child know about AI?. Proceedings of the AAAI conference on artificial intelligence. 33; 2019. p. 9795–9. https://doi.org/10.1609/aaai.v33i01.33019795.

[6] Long D, Magerko B. What is AI literacy? competencies and design considerations. CHI '20: Proceedings of the 2020 CHI conference on human factors in computing systems. New York, NY, USA: Association for Computing Machinery; 2020. p. 1–16. https://doi.org/10.1145/3313831.3376727.ISBN 978-1-4503-6708-0

[7] Tedre M, Denning P, Toivonen T. CT 2.0. Koli Calling '21: Proceedings of the 21st Koli Calling international conference on computing education research. New York, NY, USA: Association for Computing Machinery; 2021. p. 1–8. https://doi.org/10.1145/3488042.3488053.ISBN 978-1-4503-8488-9

[8] Michaeli T, Romeike R, Seegerer S. What students can learn about artificial intelligence – recommendations for K-12 computing education. In: Keane T, Lewin C, Brinda T, Bottino R, editors. Towards a Collaborative Society Through Creative Learning. WCCE 2022. IFIP Advances in Information and Communication Technology, 685. Cham: Springer; 2023. p. 196–208. https://doi.org/10.1007/978-3-031-43393-1_19.

[9] Olari V, Tenório K, Romeike R. Introducing artificial intelligence literacy in schools: a review of competence areas, pedagogical approaches, contexts and formats. In: Keane T, Lewin C, Brinda T, Bottino R, editors. Towards a collaborative society through creative learning. 685. Cham: Springer Nature Switzerland; 2023. p. 221–32. https://doi.org/10.1007/978-3-031-43393-1_21. ISBN 978-3-031-43392-4 978-3-031-43393-1

[10] Zha D, Bhat ZP, Lai K-H, Yang F, Hu X. Data-centric AI: perspectives and challenges. In: Shekhar S, Zhou Z-H, Chiang Y-Y, Stiglic G, editors. Proceedings of the 2023 SIAM international conference on data mining (SDM). Philadelphia, PA: Society for Industrial and Applied Mathematics; 2023. p. 945–8. https://doi.org/10.1137/1.9781611977653.ISBN 978-1-61197-765-3

[11] Jakubik J, Vössing M, Kühl N, Walk J, Satzger G. Data-centric artificial intelligence. Bus Inf Syst Eng 2024. https://doi.org/10.1007/s12599-024-00857-8.

[12] Zha D., Bhat Z.P., Lai K.-H., Yang F., Jiang Z., Zhong S., Hu X.. Data-centric artificial intelligence: a survey. 2023b. 2303.10158.

[13] Jarrahi MH, Memariani A, Guha S. The principles of data-centric AI (DCAI). Commun ACM 2023;66(8):84–92. https://doi.org/10.1145/3571724.

[14] European Commission. 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). Directorate-General for Communications Networks. 2021. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex:52021PC0206.

[15] Eckerson WW, Hanlon N, Barquin R. The CRISP-DM model: the new blueprint for data mining 2000;5(4).

[16] Russell SJ, Norvig P, Chang M-w, Devlin J, Dragan A, Forsyth D, Goodfellow I, Malik J, Mansinghka V, Pearl J, Wooldridge MJ. Artificial intelligence: a modern approach. Pearson series in artificial intelligence. fourth ed. global edition. Harlow: Pearson; 2022.ISBN 978-1-292-40113-3

[17] Liou YI. Knowledge acquisition: issues, techniques, and methodology. Proceedings of the 1990 ACM SIGBDP conference on trends and directions in expert systems - SIGBDP '90. Orlando, Florida, United States: ACM Press; 1990. p. 212–36. https://doi.org/10.1145/97709.97726.ISBN 978-0-89791-416-1

[18] Sveinsdottir T, Troullinou P, Aidlinis S, Delipalta A, Finn R, Loukinas P, Muraszkiewicz J, O'Connor R, Petersen K, Rovatsos M, Santiago N, Sisu D, Taylor M, Wieltschnig P. The role of data in AI. Tech. Rep. Zenodo; 2020. https://doi.org/10.5281/ZENODO.4312907.

[19] De Silva D, Alahakoon D. An artificial intelligence life cycle: from conception to production. Patterns 2022;3(6):100489. https://doi.org/10.1016/j.patter.2022.100489.

[20] Studer R, Benjamins V, Fensel D. Knowledge engineering: principles and methods. Data Knowl Eng 1998;25(1-2):161–97. https://doi.org/10.1016/S0169-023X(97)00056-6.

[21] Haakman M, Cruz L, Huijgens H, Van Deursen A. AI lifecycle models need to be revised: an exploratory study in Fintech. Empir Softw Eng 2021;26(5):95. https://doi.org/10.1007/s10664-021-09993-1.

[22] Kutzias D, Dukino C, Kötter F, Kett H. Comparative analysis of process models for data science projects:. Proceedings of the 15th international conference on agents and artificial intelligence. Lisbon, Portugal: SCITEPRESS - Science and Technology Publications; 2023. p. 1052–62. https://doi.org/10.5220/0011895200003393.ISBN 978-989-758-623-1

[23] Aragon CR, Guha S, Kogan M, Muller M, Neff G. Human-centered data science: an introduction. Cambridge, Massachusetts: The MIT Press; 2022.ISBN 978-0-262-54321-7

[24] Xie Y, Cruz L, Heck P, Rellermeyer JS. Systematic mapping study on the machine learning lifecycle. 2021 IEEE/ACM 1st workshop on AI engineering - software engineering for AI (WAIN). Madrid, Spain: IEEE; 2021. p. 70–3. https://doi.org/10.1109/WAIN52551.2021.00017.ISBN 978-1-66544-470-5

[25] UNESCO. K-12 AI curricula: a mapping of government-endorsed AI curricula. Tech. Rep. ED-2022/FLI-ICT/K-12. Paris: UNESCO; 2022.

[26] Vuorikari R, Kluzer S, Punie Y. DigComp 2.2, the digital competence framework for citizens: with new examples of knowledge, skills and attitudes. Luxembourg: Publications Office of the European Union; 2022.ISBN 978-92-76-48882-8

[27] DIGCOMP: a framework for developing and understanding digital competence in Europe. LU: Publications Office; 2013.European Commission. Joint Research Centre. Institute for Prospective Technological Studies.

[28] Tenório K, Olari V, Chikobava M, Romeike R. Artificial intelligence literacy research field: a bibliometric analysis from 1989 to 2021. Proceedings of the 54th acm technical symposium on computer science education v. 1. Toronto ON Canada: ACM; 2023. p. 1083–9. https://doi.org/10.1145/3545945.3569874. ISBN 978-1-4503-9431-4

[29] Yim IHY, Su J. Artificial intelligence (AI) learning tools in K-12 education: a scoping review. J Comput Educ 2024. https://doi.org/10.1007/s40692-023-00304-9.

[30] Jatzlau S, Michaeli T, Seegerer S, Romeike R. It's not magic after all – machine learning in snap! using reinforcement learning. 2019 IEEE Blocks and beyond workshop (B&B). 2019. p. 37–41.

[31] Biehler R, Fleischer Y. Introducing students to machine learning with decision trees using CODAP and Jupyter Notebooks. Teach Stat 2021;43:S133–42.

[32] Olari V, Cvejoski K, Eide Ø. Introduction to machine learning with robots and playful learning. Proc AAAI conf Artif Intell 2021;35:15630–9. https://doi.org/10.1609/aaai.v35i17.17841.

[33] Chiu TKF, Meng H, Chai CS, King I, Wong S, Yam Y. Creation and evaluation of a pre-tertiary artificial intelligence (AI) curriculum. IEEE Trans Educ 2022;65:30–9. https://doi.org/10.1109/TE.2021.3085878.

[34] Kandlhofer M, Steinbauer G, Hirschmugl-Gaisch S, Huber P. Artificial intelligence and computer science in education: from kindergarten to university. 2016 IEEE Frontiers in education conference (FIE). 2016. p. 1–9. https://doi.org/10.1109/FIE.2016.7757570.

[35] Kampourakis K. On the meaning of concepts in science education. Sci Educ 2018; 27(7-8):591–2. https://doi.org/10.1007/s11191-018-0004-x.

[36] Hartmann W, Näf M, Reichert R. Informatikunterricht planen und durchführen. 1. korrigierter nachdruck. Berlin Heidelberg: Springer, eXamen.press; 2007.ISBN 978-3-540-34484-1

[37] Olari V, Romeike R. Addressing AI and data literacy in teacher education: a review of existing educational frameworks. The 16th workshop in primary and secondary computing education. Virtual Event, Germany: Association for Computing Machinery; 2021Article17. https://doi.org/10.1145/3481312.3481351.

[38] Tedre M, Vartiainen H, Kahila J, Toivonen T, Jormanainen I, Valtonen T. Machine learning introduces new perspectives to data agency in K—12 computing education. 2020 IEEE Frontiers in education conference (FIE). 2020. p. 1–8. https://doi.org/10.1109/FIE44824.2020.9274138.

[39] Cerf V, Kahn R. A protocol for packet network intercommunication. IEEE Trans Commun 1974;22(5):637–48. https://doi.org/10.1109/TCOM.1974.1092259.

[40] Manikandan SG, Ravi S. Big data analysis using Apache Hadoop. 2014 International conference on IT convergence and security (ICITCS). Beijing, China: IEEE; 2014. p. 1–4. https://doi.org/10.1109/ICITCS.2014.7021746.ISBN 978-1-4799-6541-0

[41] Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput Surv 2023;55(9):1–35. https://doi.org/10.1145/3560815.

[42] Biau DJ, Jolles BM, Porcher R. P Value and the theory of hypothesis testing: an explanation for new researchers. Clin Orthop Relat Res 2010;468(3):885–92. https://doi.org/10.1007/s11999-009-1164-4.

[43] Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, Aviles-Rivero AI, Etmann C, McCague C, Beer L, Weir-McCall JR, Teng Z, Gkrania-Klotsas E, AIX-COVNET, Ruggiero A, Korhonen A, Jefferson E, Ako E, Langs G, Gozaliasl G, Yang G, Prosch H, Preller J, Stanczuk J, Tang J, Hofmanninger J, Babar J, Sánchez LE, Thillai M, Gonzalez PM, Teare P, Zhu X, Patel M, Cafolla C, Azadbakht H, Jacob J, Lowe J, Zhang K, Bradley K, Wassin M, Holzer M, Ji K, Ortet MD, Ai T, Walton N, Lio P, Stranks S, Shadbahr T, Lin W, Zha Y, Niu Z, Rudd JHF, Sala E, Schönlieb C-B. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell 2021;3(3):199–217. https://doi.org/10.1038/s42256-021-00307-0.

[44] Hu G, Peng X, Yang Y, Hospedales TM, Verbeek J. Frankenstein: learning deep face representations using small data. IEEE Trans Image Process 2018;27(1): 293–303. https://doi.org/10.1109/TIP.2017.2756450.

[45] Aldoseri A, Al-Khalifa KN, Hamouda AM. Re-thinking data strategy and integration for artificial intelligence: concepts, opportunities, and challenges. Appl Sci 2023;13(12):7082. https://doi.org/10.3390/app13127082.

[46] Kreuzberger D, Kühl N, Hirschl S. Machine learning operations (MLOps): overview, definition, and architecture. IEEE Access 2023;11:31866–79. https://doi.org/10.1109/ACCESS.2023.3262138.

[47] Heyvaert M, Hannes K, Onghena P. Using mixed methods research synthesis for literature reviews. Los Angeles: SAGE; 2016.ISBN 978-1-4833-5830-7

[48] Spector AZ, Norvig P, Wiggins C, Wing JM. Data science in context: foundations, challenges, opportunities. 1st ed. United Kingdom, USA, Australia, India, Singapore: Cambridge University Press; 2022.ISBN 978-1-00-927220-9 978-1-00-927223-0

[49] Aroyo L, Lease M, Paritosh P, Schaekermann M. Data excellence for AI: why should you care? Interactions 2022;29(2):66–9. https://doi.org/10.1145/3517337.

[50] Alam M, Groth P, Hitzler P, Paulheim H, Sack H, Tresp V. CSSA'20: Workshop on combining symbolic and sub-symbolic methods and their applications. Proceedings of the 29th ACM international conference on information & knowledge management. Virtual Event Ireland: ACM; 2020. p. 3523–4. https://doi.org/10.1145/3340531.3414072.ISBN 978-1-4503-6859-9

[51] Hoehndorf R, Queralt-Rosinach N. Data science and symbolic AI: synergies, challenges and opportunities. Data Sci 2017;1(1-2):27–38. https://doi.org/10.3233/DS-170004.

[52] Bobasheva A, Gandon F, Precioso F. 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. J Comput Cult Heritage 2022;15(3):1–23. https://doi.org/10.1145/3485844.

[53] Wing JM. Ten research challenge areas in data science. Harvard Data Sci Rev 2020;2(3). https://doi.org/10.1162/99608f92.c6577b1f.

[54] Monarch R., Manning C.D.. Human-in-the-loop machine learning: active learning and annotation for human-centered AI. 2021. Sherlter Island, NY ISBN 978-1-61729-674-1.

[55] Chattopadhyaya A., Van Dorenb M., Johnsonb R., Niua N.. On the role of data engineering decisions in AI-based applications2021; 10.5281/ZENODO.4818970.

[56] Thamm A, Gramlich M, Borek A. The ultimate data and AI guide: 150 FAQs about artificial intelligence, machine learning and data. München: Data AI Press; 2020. ISBN 978-3-9821737-0-2

[57] Zheng A, Casari A. Feature engineering for machine learning: principles and techniques for data scientists. Beijing Boston Farnham Sebastopol Tokyo: O'Reilly; 2018.ISBN 978-1-4919-5324-2

[58] Schutt R, O'Neil C. Doing data science. first ed. Beijing ; Sebastopol: O'Reilly Media; 2013.ISBN 978-1-4493-5865-5

[59] Statistische Beratungseinheit / fu:stat;. Statistik-Reader. Berlin: Freie Universität Berlin | Fachbereich Wirtschaftswissenschaft; 2023.

[60] Saunders B, Sim J, Kingstone T, Baker S, Waterfield J, Bartlam B, Burroughs H, Jinks C. Saturation in qualitative research: exploring its conceptualization and operationalization. Qual Quan 2018;52(4):1893–907. https://doi.org/10.1007/s11135-017-0574-8.

[61] Rizvi S, Waite J, Sentance S. Artificial intelligence teaching and learning in K-12 from 2019 to 2022: a systematic literature review. Comput Educ Artif Intell 2023: 100145. https://doi.org/10.1016/j.caeai.2023.100145.

[62] Norouzi N, Chaturvedi S, Rutledge M. Lessons learned from teaching machine learning and natural language processing to high school students 2020;34(09): 13397–403. https://doi.org/10.1609/aaai.v34i09.7063.

[63] Touretzky D. Big idea 3 – learning. https://ai4k12.org/big-idea-3-overview/; 2024.

[64] Touretzky D. Big idea 1 – perception. https://ai4k12.org/big-idea-1-overview/; 2024.

[65] Lee I, Ali S, Zhang H, DiPaola D, Breazeal C. Developing middle school students' AI literacy. Proceedings of the 52nd ACM technical symposium on computer science education. Virtual Event, USA: Association for Computing Machinery; 2021. p. 191–7. https://doi.org/10.1145/3408877.3432513.

[66] Tang D. Empowering novices to understand and use machine learning with personalized image classification models, intuitive analysis tools, and MIT App inventor. Massachusetts Institute of Technology; 2019. Ph.D. thesis.

[67] Reddy T, Williams R, Breazeal C. Text classification for AI education. Proceedings of the 52nd ACM technical symposium on computer science education. Association for Computing Machinery; 2021. p. 1381. https://doi.org/10.1145/3408877.3439689.ISBN 978-1-4503-8062-1

[68] Touretzky D. Big idea 2 – representation & reasoning. https://ai4k12.org/big-idea-2-overview/; 2024.

[69] Blakeley HP, Breazeal C. An ethics of artificial intelligence: curriculum for middle school students. MIT Media Lab; 2019.

[70] Sabuncuoglu A. Designing one year curriculum to teach artificial intelligence for middle school. Proceedings of the 2020 ACM conference on innovation and technology in computer science education. Trondheim Norway: ACM; 2020. p. 96–102. https://doi.org/10.1145/3341525.3387364.ISBN 978-1-4503-6874-2

[71] Shamir G, Levin I. Teaching machine learning in elementary school. Int J Child-Comput Interact 2022;31:100415. https://doi.org/10.1016/j.ijcci.2021.100415.

[72] Macar U. Castleman B. Mauchly N. Jiang M. Aouissi A. Aouissi S. Maayah X. Erdem K. Ravindranath R. Clark-Sevilla A. Salleb-Aouissi A. Teenagers and artificial intelligence: bootcamp experience and lessons learned. 2023. 10.48550/ARXIV.2312.10067.

[73] Van Brummelen J, Shen JH, Patton EW. The popstar, the poet, and the grinch: Relating artificial intelligence to the computational thinking framework with block-based coding. Proc Int Conf Comput Think Edu 2019;3.

[74] Priya S, Bhadra S, Chimalakonda S, Venigalla ASM. ML-Quest: a game for introducing machine learning concepts to K-12 students. Interact Learn Environ 2022:1–16. https://doi.org/10.1080/10494820.2022.2084115.

[75] Ng TK, Chu KW. Motivating students to learn AI through social networking sites: a case study in Hong Kong. Online Learn 2021;25(1). https://doi.org/10.24059/olj.v25i1.2454.

[76] Touretzky D. Big idea 4 – natural interaction. 2024. https://ai4k12.org/big-idea-4-natural-interaction/.

[77] Fernández-Martínez C, Hernán-Losada I, Fernández A. Early introduction of AI in Spanish middle schools. A motivational study. KI - Künstliche Intelligenz 2021;35 (2):163–70. https://doi.org/10.1007/s13218-021-00735-5.

[78] Touretzky D. Big idea 5 – societal impact. 2024. https://ai4k12.org/big-idea-5-societal-impact/.

[79] Sloman A. Teaching AI and philosophy at school ? School of Computer Science, The University of Birmingham; 2009.

[80] Paullada A, Raji ID, Bender EM, Denton E, Hanna A. Data and its (dis)contents: a survey of dataset development and use in machine learning research. Patterns 2021;2(11):100336. https://doi.org/10.1016/j.patter.2021.100336.

[81] Schopf T, Arabi K, Matthes F. Exploring the landscape of natural language processing research. Proceedings of the conference recent advances in natural language processing - large language models for natural language processings. INCOMA Ltd., Shoumen, BULGARIA; 2023. p. 1034–45. https://doi.org/10.26615/978-954-452-092-2_111.ISBN 978-954-452-092-2

[82] Bengesi S. El-Sayed H. Sarker M. K.Houkpati Y. Irungu J. Oladunni T. Advancements in generative AI: a comprehensive review of GANs, GPT, autoencoders, diffusion model and transformers. 2023. 10.48550/ARXIV.2311.10242.

[83] Mariescu-Istodor R, Jormanainen I. Machine learning for high school students. Proceedings of the 19th Koli Calling international conference on computing education research. 2019.

[84] Kahn K, Megasari R, Piantari E, Junaeti E. AI programming by children using Snap! block programming in a developing country11082. Springer; 2018.

[85] Clarke B. Artificial intelligence - alternate curriculum unit. University of Oregon: Exploring Computer Science; 2019.

[86] Druga S, Ko AJ. How do children's perceptions of machine intelligence change when training and coding smart programs?. Interaction design and children. Athens Greece: ACM; 2021. p. 49–61. https://doi.org/10.1145/3459990.3460712.ISBN 978-1-4503-8452-0

[87] Ali S, DiPaola D, Lee I, Hong J, Breazeal C. Exploring generative models with middle school students. CHI '21: Proceedings of the 2021 CHI conference on human factors in computing systems. New York, NY, USA: Association for Computing Machinery; 2021. p. 1–13. https://doi.org/10.1145/3411764.3445226.ISBN 978-1-4503-8096-6

[88] Chiu TKF. A holistic approach to the design of artificial intelligence (AI) education for K-12 schools. TechTrends 2021;65(5):796–807. https://doi.org/10.1007/s11528-021-00637-1.

[89] Pushkarna M, Zaldivar A, Kjartansson O. Data Cards: purposeful and transparent dataset documentation for responsible AI. 2022 ACM Conference on fairness, accountability, and transparency. Seoul Republic of Korea: ACM; 2022. p. 1776–826. https://doi.org/10.1145/3531146.3533231.ISBN 978-1-4503-9352-2

[90] Guo W, Wang J, Wang S. Deep multimodal representation learning: a survey. IEEE Access 2019;7:63373–94. https://doi.org/10.1109/ACCESS.2019.2916887.

[91] Vartiainen H, Toivonen T, Jormanainen I, Kahila J, Tedre M, Valtonen T. Machine learning for middle-schoolers: children as designers of machine-learning apps. 2020 IEEE Frontiers in education conference (FIE). Uppsala, Sweden: IEEE; 2020. p. 1–9. https://doi.org/10.1109/FIE44824.2020.9273981.ISBN 978-1-72818-961-1

[92] Vartiainen H, Toivonen T, Jormanainen I, Kahila J, Tedre M, Valtonen T. Machine learning for middle schoolers: Learning through data-driven design. Int J Child-Comput Interact 2021;29:100281. https://doi.org/10.1016/j.ijcci.2021.100281.

[93] Lyu Z, Ali S, Breazeal C. Introducing variational autoencoders to high school students. Proc AAAI Conf Artif Intell 2022;36(11):12801–9. https://doi.org/10.1609/aaai.v36i11.21559.

[94] Rodríguez-García JD, Moreno-León J, Román-González M, Robles G. Evaluation of an online intervention to teach artificial intelligence with LearningML to 10-16-year-old students. Proceedings of the 52nd ACM technical symposium on computer science education. Virtual Event, USA: Association for Computing Machinery; 2021. p. 177–83. https://doi.org/10.1145/3408877.3432393.

[95] Zhang Y, Allen TT, Rodriguez Buno R. Exploratory image data analysis for quality improvement hypothesis generation. Qual Eng 2024:1–20. https://doi.org/10.1080/08982112.2023.2285305.

[96] Sami JB, Stein Z, Sinclair K, Medsker L. Data science outreach educational program for high school students focused in agriculture. J STEM Educ Innov Res 2020;21(1).

[97] Pangrazio L, Selwyn N. 'Personal data literacies': a critical literacies approach to enhancing understandings of personal digital data. New Media Soc 2019;21(2): 419–37.

[98] Giner-Miguelez J, Gómez A, Cabot J. DescribeML: a tool for describing machine learning datasets. Proceedings of the 25th International conference on model driven engineering languages and systems: companion proceedings. Montreal Quebec Canada: ACM; 2022. p. 22–6. https://doi.org/10.1145/3550356.3559087.ISBN 978-1-4503-9467-3

[99] Friedrich S, Antes G, Behr S, Binder H, Brannath W, Dumpert F, Ickstadt K, Kestler HA, Lederer J, Leitgöb H, Pauly M, Steland A, Wilhelm A, Friede T. Is there a role for statistics in artificial intelligence? Adv Data Anal Classif 2022;16 (4):823–46. https://doi.org/10.1007/s11634-021-00455-6.

[100] Bellini V, Cascella M, Cutugno F, Russo M, Lanza R, Compagnone C, Bignami EG. Understanding basic principles of artificial intelligence: a practical guide for intensivists: basic principles of artificial intelligence. Acta Biomedica Atenei Parmensis 2022;93(5):e2022297. https://doi.org/10.23750/abm.v93i5.13626.

[101] Hitron T, Orlev Y, Wald I, Shamir A, Erel H, Zuckerman O. Can children understand machine learning concepts? The effect of uncovering black boxes. CHI '19: Proceedings of the 2019 CHI conference on human factors in computing systems. New York, NY, USA: Association for Computing Machinery; 2019. p. 1–11. https://doi.org/10.1145/3290605.3300645.ISBN 978-1-4503-5970-2

[102] Van Brummelen J, Heng T, Tabunshchyk V. Teaching tech to talk: K-12 conversational artificial intelligence literacy curriculum and development tools. Proceedings of the AAAI conference on artificial intelligence. 35; 2021. p. 15655–63. https://doi.org/10.1609/aaai.v35i17.17844.

[103] Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Comput Surv 2022;54(6):1–35. https://doi.org/10.1145/3457607.

[104] Fabbrizzi S, Papadopoulos S, Ntoutsi E, Kompatsiaris I. A survey on bias in visual datasets. Comput Vis Image Underst 2022;223:103552. https://doi.org/10.1016/j.cviu.2022.103552.

[105] Gupta S, Gupta A. Dealing with noise problem in machine learning data-sets: a systematic review. Procedia Comput Sci 2019;161:466–74. https://doi.org/10.1016/j.procs.2019.11.146.

[106] Henry J, Hernalesteen A, Collard A-S. Teaching artificial intelligence to K-12 through a role-playing game questioning the intelligence concept. KI - Künstliche Intelligenz 2021;35(2):171–9. https://doi.org/10.1007/s13218-021-00733-7.

[107] Allen TT, Sui Z, Akbari K. Exploratory text data analysis for quality hypothesis generation. Qual Eng 2018;30(4):701–12. https://doi.org/10.1080/08982112.2018.1481216.

[108] Wan X, Zhou X, Ye Z, Mortensen CK, Bai Z. SmileyCluster: supporting accessible machine learning in K-12 scientific discovery. Proceedings of the interaction design and children conference. 2020.

[109] Liu H, Motoda H. Feature selection for knowledge discovery and data mining. Boston, MA: Springer US; 1998. https://doi.org/10.1007/978-1-4615-5689-3. ISBN 978-1-4613-7604-0 978-1-4615-5689-3

[110] Verdonck T, Baesens B, Óskarsdóttir M, Vanden Broucke S. Special issue on feature engineering editorial. Mach Learn 2021:s10994–021–06042–2. https://doi.org/10.1007/s10994-021-06042-2.

[111] Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: a new perspective. Neurocomputing 2018;300:70–9. https://doi.org/10.1016/j.neucom.2017.11.077.

[112] Joseph VR, Vakayil A. SPlit: an optimal method for data splitting. Technometrics 2022;64(2):166–76. https://doi.org/10.1080/00401706.2021.1921037.

[113] Zhang A, Lipton ZC, Li M, Smola AJ. Dive into deep learning. Cambridge University Press; 2023.

[114] Hanneke S, Kpotufe S. On the value of target data in transfer learning. Proceedings of the 33rd International Conference on Neural Information Processing Systems. 885. Red Hook, NY, USA: Curran Associates Inc; 2019. p. 11. https://doi.org/10.48550/ARXIV.2002.04747.

[115] Burgsteiner H, Kandlhofer M, Steinbauer G. IRobot: teaching the basics of artificial intelligence in high schools. AAAI'16: Proceedings of the Thirtieth AAAI conference on artificial intelligence. 30. Phoenix, Arizona: AAAI Press; 2016. p. 4126–7. https://doi.org/10.1609/aaai.v30i1.9864.

[116] Sehra S, Flores D, Montanez GD. Undecidability of underfitting in learning algorithms. 2021 2nd International conference on computing and data science (CDS). Stanford, CA, USA: IEEE; 2021. p. 591–4. https://doi.org/10.1109/CDS52072.2021.00107.ISBN 978-1-66540-428-0

[117] Zhou B. Khosla A. Lapedriza A. Oliva A. Torralba A. Learning deep features for discriminative localization. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 2921-2929. 10.48550/ARXIV.1512.04150.

[118] Honeycutt DR, Nourani M, Ragan ED. Soliciting human-in-the-loop user feedback for interactive machine learning reduces user trust and impressions of model accuracy. Proc AAAI Conf Hum Comput Crowdsourcing 2020;8:63–72. https://doi.org/10.48550/ARXIV.2008.12735.

[119] Werder K, Ramesh B, Zhang RS. Establishing data provenance for responsible artificial intelligence systems. ACM Trans Manage Inf Syst 2022;13(2):1–23. https://doi.org/10.1145/3503488.

[120] Ginart AA, Guan MY, Valiant G, Zou J. Making AI forget you: data deletion in machine learning. Proceedings of the 33rd international conference on neural information processing systems. Red Hook, NY, USA: Curran Associates Inc.; 2019.

[121] Almatrafi O, Johri A, Lee H. A systematic review of AI literacy conceptualization, constructs, and implementation and assessment efforts (2019–2023). Comput Educ Open 2024;6:100173. https://doi.org/10.1016/j.caeo.2024.100173.

[122] Casal-Otero L, Catala A, Fernández-Morante C, Taboada M, Cebreiro B, Barro S. AI literacy in K-12: a systematic literature review. Int J STEM Educ 2023;10(1):29. https://doi.org/10.1186/s40594-023-00418-7.

[123] Srikant S, Aggarwal V. Introducing Data science to school kids. Proceedings of the 2017 ACM SIGCSE technical symposium on computer science education. Seattle Washington USA: ACM; 2017. p. 561–6. https://doi.org/10.1145/3017680.3017717.ISBN 978-1-4503-4698-6

[124] Register Y, Ko AJ. Learning machine learning with personal data helps stakeholders ground advocacy arguments in model mechanics. ICER '20: Proceedings of the 2020 ACM conference on international computing education

research. New York, NY, USA: Association for Computing Machinery; 2020. p. 67–78. https://doi.org/10.1145/3372782.3406252.ISBN 978-1-4503-7092-9

[125] Kim K, Kwon K, Ottenbreit-Leftwich A, Bae H, Glazewski K. Exploring middle school students’ common naive conceptions of Artificial Intelligence concepts, and the evolution of these ideas. Educ Inf Technol 2023. https://doi.org/10.1007/s10639-023-11600-3.

[126] Ridsdale C. Rothwell J. Smit M. Bliemel M. Irvine D. Kelley D. Matwin S. Wuetherick B. Ali-Hassan H. Strategies and best practices for data literacy education knowledge synthesis report. 2015. 10.13140/RG.2.1.1922.5044.

[127] van Bekkum M., de Boer M., van Harmelen F., Meyer-Vitali A., ten Teije A.. Modular design patterns for hybrid learning and reasoning systems: a taxonomy, patterns and use cases. 2021. 2102.11965.