Teaching Data Concepts and Practices in Secondary School Education on Artificial Intelligence: Approaches, Mechanisms, and Emerging Local Theories

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

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

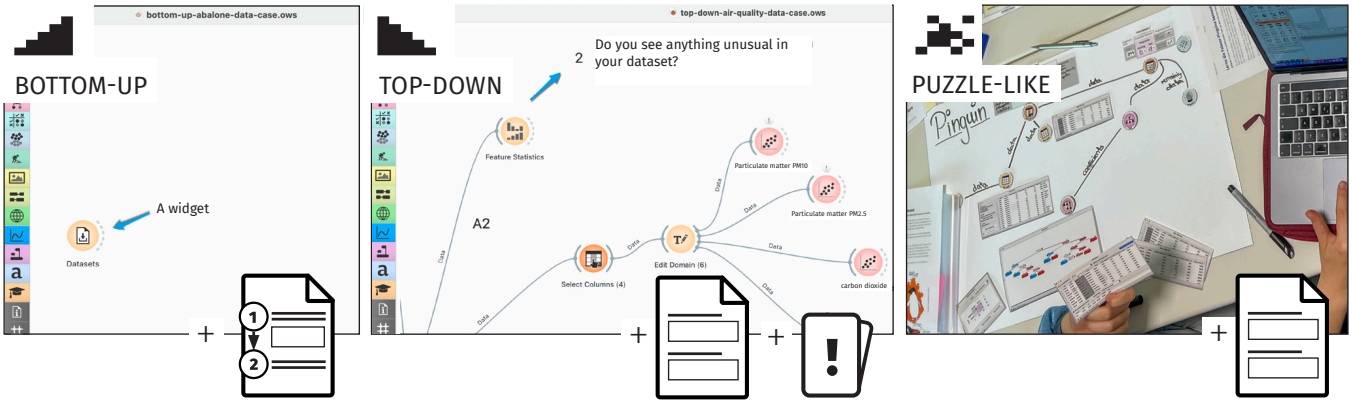

Figure 1: Three approaches to teaching about data with data cases: bottom-up, top-down, and puzzle-like.

Abstract

Data-related concepts and practices have been proposed to be a fundamental component of artificial intelligence (AI) school education. However, proposing concepts and practices is not enough. To enable teachers to introduce data concepts and practices in schools, it is necessary to understand the mechanisms that effectively support the learning of these under real school conditions. To this end, we designed and conducted a three-iteration design-based research study in collaboration with computer science and mathematics teachers, school students, and domain experts. In this paper, we present the results of the research process: the developed teaching approaches and the identified mechanisms that support learning of data concepts and practices following a conjecture-mapping approach. Based on the results, we explicate theoretically and empirically sound local instructional theories for teaching data concepts and practices in secondary school education on AI.

CCS Concepts

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

Koli Calling '25, Koli, Finland

ACM ISBN 979-8-4007-1599-0/25/11
https://doi.org/10.1145/3769994.3770021

Keywords

Data Education, AI education, K-12, Conjecture Mapping

ACM Reference Format:
Viktoriya Olari and Ralf Romeike. 2025. Teaching Data Concepts and Practices in Secondary School Education on Artificial Intelligence: Approaches, Mechanisms, and Emerging Local Theories. In *25th Koli Calling International Conference on Computing Education Research (Koli Calling '25), November 11–16, 2025, Koli, Finland*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3769994.3770021

1 Introduction

Artificial intelligence (AI) is increasingly being included as a topic in computer science (CS) school education worldwide [51]. Over the past decade, researchers, policymakers, industry representatives, and teachers have made significant efforts to define the competencies that school students need to develop in order to live and thrive in a world permeated with AI technologies [3, 6, 25, 29, 32, 35, 48–51]. One central competency mentioned by these frameworks is that primary and secondary school students should learn to design AI systems – not to develop commercial products or put them into service but rather to build the confidence and capacity to shape AI for human good by understanding the principles underpinning the design and behavior of AI (see, for example, [23, 32, 35, 37]).

From a subject-matter perspective, designing an AI system – especially a machine learning (ML) system – requires knowledge of the concepts and practices involved in collecting, storing, and pre-processing data. It also requires, based on the data available and the task, the ability to select an appropriate model architecture and a learning algorithm [2, 9, 15, 21, 24, 27, 40]. Understanding

the structure of the data and identifying any underlying issues is crucial because the behavior of an ML model is heavily influenced by its training dataset [24].

Despite its fundamental importance, the role of data in AI education has largely been underestimated thus far [38]. Several well-founded proposals have been made regarding the key data concepts and practices that students need to understand in order to design AI systems [36, 37]. However, current teaching approaches only scratch the surface of working with data [38]. Proposing concepts and principles is not enough. Since real school practice is impacted by school-specific requirements and challenges – such as rigid time constraints for teachers and students, the need to enable all students (not just those who are interested) to learn, and limited teacher knowledge, among other things – there is a need to develop effective teaching approaches.

To incorporate the teaching of data concepts and practices into AI education, we recognize two key needs. First, to support teachers, we must provide proven teaching approaches for real school settings. Second, in order to create these teaching approaches, we need to understand the mechanisms that effectively support the learning of data concepts and practices in real school settings.

To address these needs, we developed and evaluated a data case study method over the past two years in a three-iteration design-based research study. For this study, we, as CS education researchers, collaborated with CS and mathematics teachers and domain experts from the data science field. What follows is a report on two key research questions that we explored during the process:

(1) Which mechanisms support students' understanding of data concepts and practices, as well as their motivation to learn, and enable them to design simple ML systems?
(2) How should theoretically sound learning approaches be designed for use by teachers in real school practice?

The rest of the paper is organized as follows: First, we provide an overview of the theoretical background and related work on teaching data fundamentals in schools. Next, we detail our methodological approach, explaining the phases of design-based research and the conjecture-mapping approach. Lastly, we present the results of the research process: three teaching approaches and the learning mechanisms identified. Based on these findings, we propose five theoretically and empirically sound local instructional theories for teaching data fundamentals in secondary education.

2 Theoretical Background

To set the stage for identifying mechanisms and designing teaching approaches, we present prior work on data concepts and practices essential for school education on AI, the data case study as an established method for teaching about data, and known challenges from teaching data-related topics in schools.

2.1 Data Concepts and Practices

A large body of literature shows that working with data is an indispensable and time-consuming component of creating an AI system, particularly an ML system (see, for example, [1, 24, 42, 54]). Sculley et al. vividly illustrate in their work that only a small fraction of real-world ML systems are composed of ML algorithms. The necessary groundwork is extensive, especially for data collection, analysis, validation, and feature engineering [46]. Creating ML systems requires an understanding of key data concepts and mastery of key data practices [36, 37].

Knowing concepts is essential for communicating about ML systems. Just as describing a cell and its functions in biology lessons requires students to become familiar with a set of terms and understand their meanings (e.g., nucleus, Golgi apparatus, etc.), describing ML systems requires students, for example, to understand the types of data-based tasks possible (e.g., classification, regression), the data formats used by systems (e.g., tabular, time series, image), how and where data is stored (e.g., datasets and databases), the types of data problems (e.g., outliers, and missing data), how the data is transformed (e.g., cleaning and feature engineering), and how it is used to solve a task (e.g., data flow, training, validation, and testing data). Students must also understand how the results of models are tested and interpreted (e.g., performance metrics, underfitting, and overfitting). Olari et al. [36] provide further examples of data concepts, including a definition of what a data concept is.

Knowing and understanding the concepts is not sufficient to create an ML system; engaging in the data practices is essential. These practices are the actions applied to or carried out with data in mind during an ML project [37]. They include, for example, understanding the task; creating a dataset; deciding how to store the data; describing, exploring, and verifying the data quality; preprocessing the data (cleaning, labeling, and engineering features); using data to create models; preparing evaluation data; selecting an evaluation metrics; interpreting the results of the modeling process; and sharing, archiving, or deleting the data [12, 14, 55].

Designing an ML system requires the ability to communicate about ML systems using key data concepts and the ability to apply key data practices. However, a recent literature review showed that current ML education teaching approaches only scratch the surface of data practices, with some practices rarely or never being addressed [38]. For these reasons, we consider the development of teaching approaches that support understanding of data concepts and practices to be one of the central goals of our research study.

2.2 Teaching with the Data Case Study

In academic education in AI and data science, where teaching of data concepts and practices naturally occurs, an established teaching method is the *data case study* (also known as a "case study" or "lab"). As such, it is grounded in the tradition of constructivist, active, and situated learning [7, 13, 20, 22, 34, 53] and requires students to solve a *data case*, an authentic problematic situation accompanied by a dataset. In the process, students apply data concepts and practices taught in lectures, thereby developing data-based judgment and problem-solving skills [28]. While doing so, students internalize the fundamental importance of data for ML systems. Although several researchers in ML school education use data cases for teaching (e.g., [5]), the data case study has not been explicitly investigated as a teaching and learning method for AI school education thus far.

To enrich school education with the data case study method, school-specific challenges need to be considered. For instance, it is known that introducing data concepts and practices requires novices a great mental effort as they need to divide their attention between statistical concepts, CS, mathematics, and the domain

where the data comes from, while considering all aspects as part of a whole [10, 19]. The school students often do not have programming knowledge to work with data [30]. Time constraints imposed by schools make mastering a subject difficult. For teachers, besides lacking knowledge, generating and maintaining motivation and engagement among students is a challenge [14, 44].

In order to overcome the difficulties and enable active, constructivist learning, researchers report positive effects from letting students work with real-world datasets [18, 44] or letting them collect their own data, which enhances their sense of ownership over their learning [47]. They also report benefits from using low-code or no-code environments [30] to enable low-floor access for all students and from providing hands-on, module-based, iterative learning [44], as well as project-based learning [44]. Which contexts are appropriate for school education is a controversial issue [33]. Some research suggests personally relevant contexts [8], such as social media [17], sports [43], and food [30]. However, studies in the field of data literacy report that students were uninterested in everyday contexts [30]. Research on teaching other science-related subjects shows that students with low interest benefit most from daily-life and personal contextualization, while highly interested students benefit from unique contexts [16].

Reports on teaching data concepts and practices under real conditions are rare [30]. Evaluations of such interventions do not investigate which elements of the design cause learning. For this reason, we see the investigation of learning mechanisms that the data case study causes as another central goal of our research study.

3 Methodology

To further develop the data case study as a teaching approach that is sound from CS education, mathematics, and domain matter perspectives, suitable for secondary education, we needed expertise from the CS education research, school practice, and the domain providing the data for data cases. Therefore, we, as researchers in CS education, built and led a research team consisting of a CS and mathematics teacher with over 10 years of teaching experience, a mathematics and physics teacher with over 5 years of teaching experience, and a domain expert with over 35 years of experience working with data in the professional context of environmental science. The team met weekly for 1.5 years to prepare teaching approaches and reflect on mechanisms observed in the classroom.

3.1 The Design-Based Research Process

The research process for investigating mechanisms and developing teaching approaches was grounded in the design-based research approach following Prediger [41]. As such, it consisted of four closely interrelated steps, which were conducted in three cycles[1]: (1) specifying and structuring the learning objectives, (2) developing the teaching approach, (3) conducting and evaluating experiments with the developed approach, and (4) identifying mechanisms and developing local instructional theories about how design decisions in teaching approaches relate to learning outcomes.

[1]We refer to the first, second and third cycle as T1, T2 and T3, respectively. "T" stands for the word "trimester" – a period of three month. Every cycle lasted three months.

(1) The first phase of each cycle involved specifying the subject matter to be learned. To identify data concepts and practices relevant to teaching the topic of ML in schools in T1, we followed the model of data practices proposed by Olari et al. [37], who comprehensively analyzed the subject literature and CS school curricula on the topic of data in school education. We also drew on the collection of data concepts proposed by the same researchers [36]. From these collections, we selected the central concepts and practices that are commonly found in academic data cases (see, e.g., [53]).

(2) In the second phase of each cycle, we embodied our assumptions on how to support the learning of the selected data concepts and practices in concrete learning arrangements – the *data cases*. This process resulted in 18 data cases developed over a period of 1.5 years. The data cases and a detailed report on their development are presented in Olari et al. [39].

(3) The third phase of each cycle involved teaching with the data cases at a secondary school. The teachers from the research team conducted the lessons. In cases where lessons had to be canceled, a CS education researcher stepped in as a substitute teacher. A CS education researcher and the domain expert also participated in lessons to observe the mediating processes and collect data. They supported the teachers, particularly in T1, when the content and practices were still new; support decreased in T2, and by T3 the teachers had developed sufficient self-confidence to teach the topic independently. The study was coordinated with and approved by the Senate Department for Education, Youth and Family in Berlin and the ethics committee of Freie Universität Berlin. Before starting data collection, teachers, school students, and their guardians were provided with comprehensive information about the study to obtain informed consent for participation.

(4) In the fourth phase of each cycle, we evaluated the experiences and developed the theory. We reflected on how the design decisions supported the mediating processes observed under real school conditions and whether these processes led to the expected outcomes. As a result of this analysis, we formulated mechanisms that supported students' understanding of data concepts and practices and enabled them to design simple ML systems.

The design-based research postulates that discovery emerges through change [11, p. 145]. At the end of the research project, after we had changed the teaching approach three times and analyzed the mechanisms, we explicated five local instructional theories. These describe how the design of the teaching approaches supports the learning of data concepts and practices.

3.2 Conjecture Mapping for Articulating the Mechanisms

To make the design decisions and mechanisms explicit, we followed a conjecture-mapping approach described by Sandoval [45]. This approach helped us to specify the hypothesized, theoretically sound learning mechanism in each project cycle. Figure 2 demonstrates an excerpt from an initial conjecture map for T3.

Each *mechanism* comprised four elements: a high-level conjecture, an embodiment, mediating processes, and outcomes. The *high-level conjecture* stated how we aimed to support the learning of data concepts and practices under real school conditions. The *embodiment* (or *design*), instantiated the high-level conjecture in the form

of a task structure, a participant structure, tools, materials, and discursive practices. *Mediating processes* described the actions of students and the artifacts that we hypothesized would be observed in the classroom once the embodiment was implemented. *Outcomes* denoted the long-term learning goals students were expected to achieve. To hypothesize the *mechanisms* in the learning process, we explicitly connected elements of the conjecture map using arrows (see *design conjectures* and *theoretical conjectures*). For instance, as shown in Figure 2, we hypothesized that working with a puzzle-like data case and an unplugged version of Orange3 would lead students to reconstruct the data processing steps in a communicative manner. Repeated engagement in this process was expected to foster a conceptual understanding of data concepts such as data flow.

After trying out the design and analyzing the mediating processes that actually occurred as well as the learning outcomes of the students, we created a second conjecture map presenting the actual mechanism (i.e., what we actually achieved). The comparison of the initial and the actual conjecture maps served as a basis for reflection and for making changes in the next cycle of the project.

3.3 Data Collection and Analysis

To gain a comprehensive understanding of the mechanisms, we collected data on mediating processes observed in the classroom (weekly) and on outcomes (at the end of the course) using a mixed-method approach [52] [2]. Since we were interested in understanding the causal mechanisms at work in real school conditions in RQ1, i.e., conditions that are typically complex, temporary, and contextually variable, the theory was mostly formed by the qualitative analysis of data [31]. This approach helped us to understand the influence of contextual factors that cannot be statistically controlled and the unique processes at work in specific situations [31]. Miles and Huberman argue that "qualitative analysis, with its close-up look, can identify mechanisms, going beyond sheer association. It is unrelentingly local, and deals well with the complex network of events and processes in a situation" [31, p. 147]. To ensure validity, findings were regularly discussed within the research team and triangulated with quantitative results. In what follows, we describe the data collected and the data analysis procedure.

Mediating processes. The weekly data on mediating processes consisted of artifacts produced by the students, videotaped lessons, lesson observations (lesson observation protocol), pre-post tests on students' knowledge, post-evaluations of students' motivation (M2 instrument), and semi-structured interviews with the teachers (L2).

To identify the mediating processes, a CS education researcher qualitatively evaluated students' artifacts and manually summarized the results regarding correctness and common errors. The researcher also evaluated motivation surveys on how students perceived competency, pleasure, freedom of choice, and pressure (Example item for *pleasure*: "I found the activity very interesting." Responses ranged from 1 (completely disagree) to 5 (completely agree)). The students' interactions with the teaching materials and their engagement during the lessons were noted based on lesson observations and an initial review of the videotaped lessons [3].

[2]The instruments referenced in this paper can be found in the appendix.
[3]A detailed analysis of the videotaped data will follow in subsequent stages of the study.

An essential source for understanding the mediating processes was the teachers' perspective. Prior research has shown that expert teachers have a very nuanced understanding of the events happening in the classroom [26, p. 34]. To elicit this perspective from the interviews, the CS researcher coded the semi-structured interviews deductively according to the following categories: class average participation, individual participation, persistence to solve the task, distractions in the class, difficulties, students' use of support material, and demonstration of knowledge. Then, the researcher created summaries for each code and memos with illustrative examples. Text passages related to difficulties underwent additional inductive analysis. The most common difficulties are presented systematically in Olari et al. [39]. The researcher triangulated the results of the qualitative data analysis on mediating processes with the quantitative analysis of the pre-post knowledge tests.

Learning outcomes. The data on learning outcomes from the end of the course consisted of the project work submitted by the students and results from the post-course survey on students' perceptions of teaching conditions and motivation (M3 instrument).

To assess the learning outcomes, the CS education researcher qualitatively analyzed the students' final projects. Understanding of the data concepts was evaluated based on the conceptual correctness of the answers in the written reports. To evaluate students' ability to apply data practices to design simple ML systems (which we call *agency*), the researcher investigated the quantity and variety of widgets in the data flows, as well as their correctness and complexity. To assess teaching conditions and motivation, we calculated the group means for each of the following constructs: perceived content relevance, instructional quality, teacher interest, social integration, competence support, autonomy support, and error culture. An example item for *content relevance* was "The course made it clear that the subject matter is also important in everyday life." Students could select an answer on a scale from 1 to 4, where 1 meant "does not apply" and 4 meant "fully applies."

Learning groups. The student groups in different cycles may have differed substantially, limiting meaningful comparisons across T1, T2, and T3. To examine whether the groups were comparable, we administered a self-efficacy survey (E1 instrument) at the beginning of each course. Self-efficacy refers to a person's ability to persist and overcome difficulties while learning [4], which is relevant when acquiring data literacy (see Section 2.2). In addition, we asked students about their prior experiences with CS, working with data, and the broad topic of environmental science to which our data cases belong (SE instrument). To determine if there were any differences among the groups in each cycle, we compared the means across the groups using the one-way ANOVA. To determine among which groups the variances differed, we conducted a Tukey's range test. Information on each student's prior knowledge was manually extracted from observation notes and analyzed.

4 Results

In what follows, we first describe the study participants. Then, we report on the teaching approach in each project cycle, including findings on the mediating processes and outcomes. Based on the findings, we formulate theoretically and empirically sound mechanisms at the end of each section, thus answering RQ1.

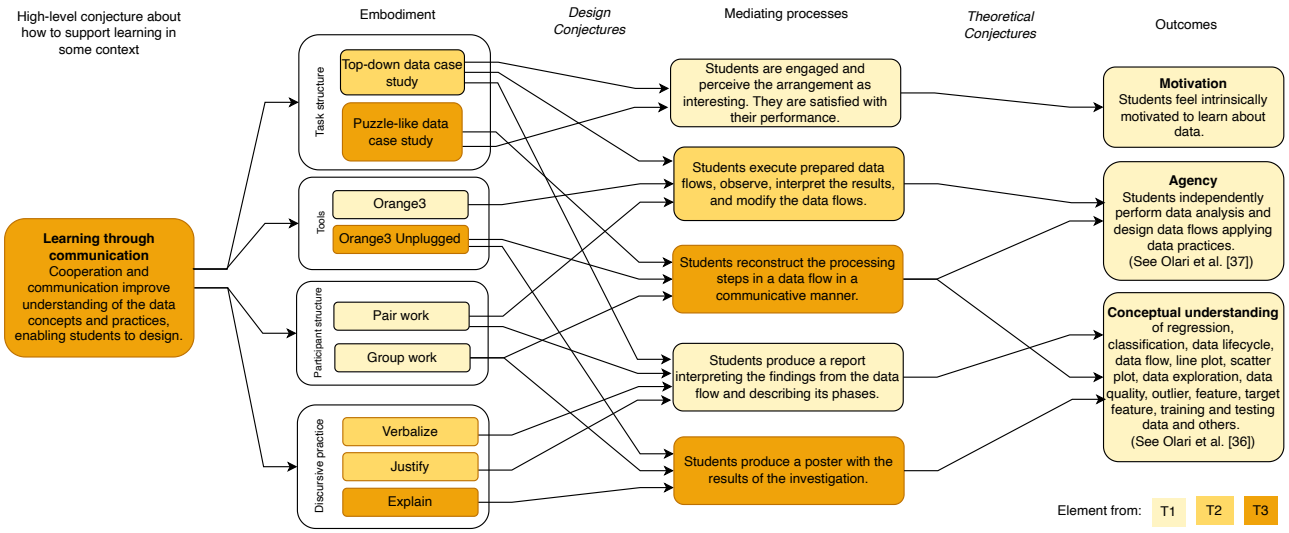

Figure 2: Excerpt from the initial conjecture map for T3. Different colors on the conjecture map indicate the iteration in which an element was introduced.

4.1 School and Participants

Experiments were conducted and evaluated with the developed approaches in a gymnasium, a school that prepares students for higher education at the university. The school was located in an urban area in Berlin and accommodated around 1,000 students. Courses, in which the experiments took place, were offered for the students in the fast-track classes, in which compulsory teaching content is learned more quickly than in regular classes and the students are offered such elective courses. Every course lasted three months and was offered in the "Computer Science and Mathematics" track. Courses in T1 and T2 lasted 48 school hours (48 × 45 minutes), and the course in T3 lasted 46 school hours. Since CS is not a compulsory subject in Berlin, the courses could be selected by students with and without prior knowledge in CS.

During the selection process, 44 fast-track students in grades 9 and 10, most of whom did not choose the course as their first preference, were randomly assigned to one of three groups. The resulting groups were relatively small, which is typical for CS classes in Germany (T1: n = 15; f = 5, m = 8, not specified = 2; mean age = 14.46 years; T2: n = 13; f = 4, m = 7; not specified = 2; mean age = 14.7 years; T3: n = 16; f = 4, m = 10, d = 1, not specified = 1; mean age = 14.72 years).

The self-efficacy levels among students prior to taking the course ranged from above average to rather high (T1: n = 15, mean = 3.12, SD = 0.39; T2: n = 11, mean = 2.66, SD = 0.42; T3: n = 15, mean=2.82, SD = 0.52). Tukey's test showed that students in T2 exhibited a lower level of self-efficacy than those in T1 and T3, which means that T2 students had a lower subjective belief in their ability to overcome difficult challenges through their own actions than T1 and T3 students. There was no difference between the T1 and T3 groups. As expected, the students had heterogeneous knowledge of CS. Nearly half of the students in each cycle had no prior programming experience, though most of them had experience working with spreadsheet programs. None of the students had in-depth knowledge of the domain area in which the data cases were contextualized.

4.2 The Bottom-Up Teaching Approach

In order to enable the students to master the data concepts and apply the data practices, the teaching approach in T1 was grounded in the academic data case study method. To make this approach workable in the school context, however, we adapted the method to address known difficulties, such as heterogeneous prior knowledge of the students in programming (challenge 1), difficulties in establishing and maintaining motivation, as identified in prior research (challenge 2), and time constraints and a lack of domain knowledge for both school students and teachers (challenge 3).

4.2.1 Embodiment. In the following, we outline and justify our design decisions for the teaching approach in T1.

Activity structure. Due to the established structure of the academic data cases (see [53] for examples), the school-specific data cases were initially constructed to be **bottom-up** (Figure 1, left). The bottom-up approach began with introducing students to a data case – a problematic situation and the dataset used to investigate it. Then, it guided the students step by step through a solution. In the process, the students built a data flow [4] and completed small, inquiry-based tasks after each step, thereby constructing an understanding of data concepts and practices through direct experiences

[4]Data flow in Orange3 is the complete path of data from the first to the last computing unit (called a "widget"). It begins with a widget to import the data, continues with widgets for data exploration and data pre-processing, and finishes with widgets using data to train an ML model and evaluate the results.

with them. We characterized this heavily situated, active learning approach in T1 in a high-level conjecture as "learning by imitation."

Task/materials. To appeal to as many students as possible, the data cases used the third-party, real-world datasets from familiar (e.g., average global temperatures) and uncommon (e.g., abalone, a marine mollusk, from Tasmania) contexts. Contrary to the academic data cases that expect university students to deeply familiarize themselves with the domain, each school-specific data case began with a one-page introduction to the context tailored to the students' everyday life experiences. To enable all students – not only those who had prior knowledge of CS – to apply data practices, we replaced the textual programming in computational notebooks such as Jupyter Notebook with visual programming in the flow-based data analysis environment Orange3, which is often used by school students without prior programming or data analysis experiences [14]. Since teaching data practices and concepts was new for the teachers, the school-specific data cases were organized into an all-in-one, multi-page workbook. Each student received a copy, which allowed for largely autonomous work and reduced the teachers' workload. The workbook contained tasks and spaces to fill in along with the information on data concepts and on the operation of "widgets," the computing units in Orange3.

Participant structure. Besides individual work, which is typical in academic data cases, we explicitly integrated partner and group work into the workbooks because meaningful communication among students is a central means and goal in school education, as was emphasized by the teachers in the research team.

Discursive practice. We facilitated discussions by asking the students to share their results from the small, inquiry-based tasks among each other.

4.2.2 Mediating Processes. In 24 lessons of T1, the students worked with bottom-up data cases prepared by the research team. When teachers reflected on mediating processes, they concluded that it was a *"challenge [...] to observe the learning in the students. (L1ADay2111, pos. 99)"*. It was difficult to determine what the students were learning and where they had difficulties (S1T1W3aud1tr, Pos. 10) (challenge 4). Teachers complained that the step-by-step instructions caused more important skills, such as reflection on the process, to be neglected (challenge 5).

In the students' artifacts, we observed that most students were able to compose data flows according to instructions and complete the inquiry-based tasks. However, the less persistent students often worked faster than the highly persistent students, delivering superficial answers and leaving out tasks (S1T1W3, Pos. 16), causing dissatisfaction among teachers (challenge 6).

From the lesson observations, we noticed that, despite explicit tasks for collaborative work, the students were uncomfortable engaging in group work and mostly worked alone (challenge 7). The motivational survey results indicated that the students found the learning with data cases rather interesting, as illustrated in Figure 3. They were mostly satisfied with their performance, felt able to choose from activities, and generally did not feel under pressure.

4.2.3 Outcomes. At the end of the course, the teachers were concerned about whether students would be able to find the problematic situation on their own and create data flows from scratch (challenge 8). Therefore, the research team heavily scaffolded the final projects. We provided a dataset, a context description, a list of questions to be answered, and a list of widgets to be used. Some students even received workable data flows.

Under these heavily scaffolded conditions, one-third (5 out of 15) of the students did not manage to create their own data flows and used workable data flows (challenge 9). Two-thirds (10 out of 15) of the students succeeded in creating their own data flows, though these were partially broken. Analysis of the project reports revealed that most students were able to operate with the data concepts that we covered in the course. However, the level of correctness varied greatly (challenge 10). For example, one group correctly used the concepts of model, correlation, feature, and target but incorrectly interpreted the R^2 as accuracy (*"The fourth model predicts the area of the forest fire best. Rain, temperature, wind, and RH were used as features because they are most strongly correlated with fires, and area was selected as the target. However, the model is still very poor, with an accuracy of just 13%" (Project 6).*)

Despite difficulties, the group tended to agree that the content covered in the course was relevant to their lives, the instructional quality was rather high, and the error culture was generally positive. The students felt more intrinsically than externally motivated, as can be seen in Figures 4 and 5.

4.2.4 Mechanisms for the Bottom-Up Teaching Approach. From observed mediating processes and outcomes, we conclude on three mechanisms in T1. In a course built around bottom-up cases, in which students compose data flows step by step according to instructions and interpret the results in context in small, inquiry-based tasks after each step, ...

Mechanism 1: ... the highly persistent students follow instructions introducing them to specific data practices and concepts, leading them by the end of the course to be able to create a data flow for a given context by correctly ordering the given widgets and articulating the results in context using data concepts. The less persistent students also follow the instructions but skip difficult tasks and, at the end of the course, can only articulate the results for a data flow given by the teachers, without being able to compose the data flows on their own.

Mechanism 2: ... the teachers have difficulties observing mediating processes in the classroom and, by the end of the course, trust their students to independently conduct a project only under heavily scaffolded conditions.

Mechanism 3: ... the motivation among the students is high. At the end of the course, the students perceive the teaching conditions as positive and feel intrinsically motivated to learn about data.

4.3 The Top-Down Teaching Approach

The main challenge from T1 that prompted changes of the teaching approach in T2 was the overemphasis on step-by-step instruction, which limited students' reflection on the data flows (challenge 5). The teachers argued that being able to reflect on and explain why a particular data concept or practice is used would help students transition to an abstract level of thinking. This would facilitate their conceptual understanding of data concepts and their ability to apply data practices in final independent projects. Additionally, the teachers emphasized the need to enable not only highly persistent students to learn but also those who are less persistent.

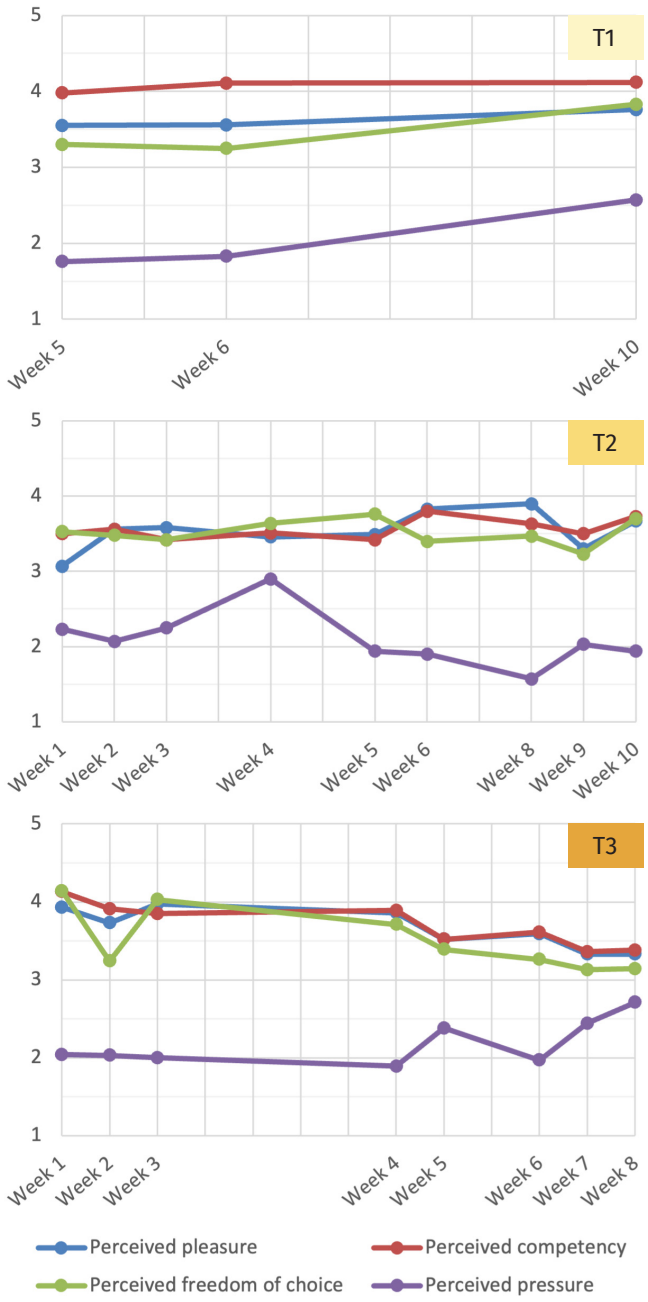

Figure 3: Development of students' motivation during T1 (above), T2 (middle) and T3 (below), where 1 means "completely disagree" and 5 is "completely agree".

4.3.1 Embodiment. In the following, we outline and justify our design decisions for the teaching approach in T2.

Activity structure. To enable students to reflect on the process, we transformed the bottom-up data case approach into a **top-down** approach (Figure 1, middle). Top-down means that after contextual introduction, a data case presented students with an elaborated data flow and asked them to complete a series of interpretation and reflection tasks (see the details of this approach in Olari et al. [39], Section 4.2.2). We characterized this situated, active learning approach in T2 in a high-level conjecture as "learning by reflection."

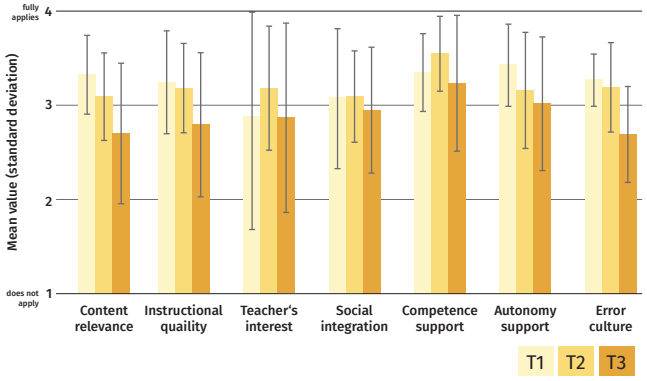

Figure 4: Perceived teaching conditions by the students at the end of T1, T2, and T3.

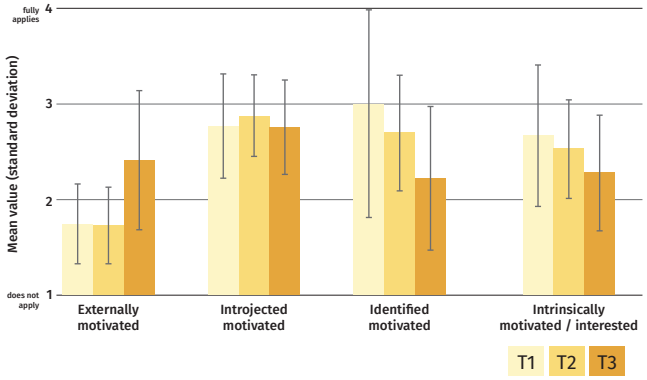

Figure 5: Quality of motivation to learn by the students at the end of T1, T2, and T3.

Task/materials. Since the students in T1 found the course content to be relevant, we continued using the real-world datasets in the top-down data cases. The flow-based data analysis environment Orange3 also remained unchanged, as all students in T1 were able to work with it. To help students understand where they were in the data flow, each data case began with an advanced organizer displaying the data practices and concepts to be covered. To make mediating processes and students' difficulties more observable (challenge 4) and enable the teacher to support the less persistent students, the all-in-one workbook containing all the necessary information was dismantled. Instead of a workbook, students received a one-page protocol and interpretation cards to help them interpret the widget results using the criteria, as well as widget cards to help them reconfigure the widget if needed. Teachers could now quickly assess how far the students had progressed and whether they were having difficulties by passing through the classroom.

Participant structure. The participant structure remained largely the same as in the first iteration.

Discursive practices. Since we recognized the need to reflect on the data flow, we explicitly added the instructions into tasks, forcing students to verbalize, interpret, and justify their answers.

4.3.2 Mediating Processes. In 26 lessons of T2, the students worked with the top-down data cases prepared by the research team. In the interviews, teachers reported being better able to observe learning and challenges than in T1. For instance, they realized that students had difficulties interpreting the results of a data flow in context (challenge 11). They also noticed that students did not understand how debug data flows if these get broken (*"They also had problems with doing this data discovery [...]. They lack debugging strategies."* (T2W5)) (challenge 12).

In the students' artifacts, we saw that the students were able to describe the results of leaf widgets on data exploration and to reflect on the steps in the data flow. However, most students struggled to explain why a particular ML model is suitable for a particular type of data and to interpret the model predictions in context (challenge 13).

From the lesson observations, we noticed that students could execute data flows and observe the results. However, the lack of communication (challenge 7) persisted. Similar to T1, we noticed that the students were uncomfortable working in groups and mostly talked only to their neighbors. In terms of motivation, the students were engaged in the work with top-down data cases and generally did not feel under pressure (see Figure 3).

4.3.3 Outcomes. At the end of T2, the teachers decided to give students a less scaffolded project than in T1. Instead of providing a detailed list of widgets for students to use, as in T1, we provided a more general description of the data practices that should be included in the data flow. Instead of writing a report, we asked students to explain the data flow and results of their investigations in a poster.

Under these conditions, all groups of students created simple data flows without being told which widgets to use. The use of widgets was rich and comparable to the use of widgets by students in T1. However, the data flows were much less comprehensive. Students in T2 used an average of 20 widgets per project, while students in T1 used an average of 98. In terms of data practices, the students demonstrated a general understanding of the stages of the data flow and adequately interpreted the results. They could articulate the results of data practices related to understanding and preparing data. However, four of five projects had errors in their data flows when applying data practices from the "evaluate performance" stage (challenge 14). We made similar observations when it came to the concepts: The students could operate with the data concepts that we covered related to data exploration and preparation but not with those related to using data for ML modeling and evaluation (challenge 15).

The perceived intrinsic motivation at the end of the course was around average, and extrinsic motivation was below average, as can be seen in Figure 5. This means that students with average self-efficacy levels in T2 felt similarly motivated to those with a higher level of self-efficacy in T1.

4.3.4 Mechanisms for the Top-Down Teaching Approach. From the observed mediating processes and outcomes, we conclude that in a course built around top-down data cases, with a focus on providing an overview of the process and articulating and interpreting the data practices and results in context,

Mechanism 4: ... students describe the results of the data flows and reflect on the steps in the data flow while having difficulties debugging the data flows. At the end of the course, students have an intuition for applying data practices and creating basic data flows for a given context without being explicitly told which widgets to use. They articulate the results of data exploration, although they make errors when training and evaluating simple ML models.

Mechanism 5: ... teachers are able to observe students' difficulties in the classroom. By the end of the course, they consider the students capable of applying data practices with guidance.

Mechanism 6: ... the motivation among students with an average level of self-efficacy is rather high. At the end of the course, students perceive the teaching conditions as positive and feel more intrinsically than extrinsically motivated to learn about data.

4.4 The Puzzle-Like Teaching Approach

The main reason for the changes in T3 was the lack of collaboration among students observed in T1 and T2 (challenge 7). To enable students to apply data practices, teachers emphasized the importance of collaboration between students, in which students practice discussing data concepts and practices to better understand them.

4.4.1 Third Embodiment. In the following, we outline and justify our design decisions for the teaching approach in T3. An excerpt of the conjecture map for T3 is illustrated in Figure 2.

Activity structure. To enable school students to practice data concepts and foster collaboration, we developed a **puzzle-like** teaching approach (Figure 1, right). Puzzle-like means that students received a set of widgets, tables, and a set of cards for widget configuration and were expected to work together without computers to reconstruct the underlying data flow. When ready, students could verify their solutions by constructing the same data flow in the Orange3 data mining environment. We characterized this active, collaborative learning approach in T3 in a high-level conjecture as "learning through communication."

Task/materials. The mix of the real-world datasets and flow-based data analysis environment remained unchanged in T3. In order to allow collaboration among students, we created Orange3-unplugged teaching materials. These enabled multiple students to work on the same data flow simultaneously. They could observe all the steps and results of the data flow, which is not easily possible in the Orange3 environment.

Participant structure. Although the general participant structure in T3 remained the same as in T2, as can be seen in Figure 2, we made significant changes to how students were expected to participate in tasks, including their responsibilities. For instance, we utilized the jigsaw teaching technique to allow students who missed previous lessons to catch up on the data concepts and practices from their more knowledgeable peers.

Discursive practice. We placed more emphasis on explanation and justification through the unplugged exercise. By requiring students to present their approaches to each other, we made it necessary for them to justify their ideas.

4.4.2 Mediating Processes. The students worked in 20 lessons of T3 with the puzzle-like data cases. In four lessons, they worked with a top-down data case. The teachers reported observing many mediating processes happening in the classroom. They noted lively interactions between students when reconstructing the data flows, peer teaching, and pair programming. Compared with T2, teachers reported being even better able to identify learning difficulties and reported feeling able to address them without the help of the research team. According to the teachers, the students showed endurance, and the class average actively participated: "*Well, once again, I had the impression that everyone was actually working and participating" (S1T3W2int2, Seg. 4).*

In the lesson recordings, we observed that students explained widgets and data flows to each other while reconstructing the data flows. Besides improved collaboration, we observed that during the reconstruction process, the students discovered implicit concepts, such as the input and output data of a widget, objects that flow between the widgets – elements that are essential for finding errors in the data flows. With this, the puzzle-like teaching approach addressed challenge 9, which was noticed by teachers in T2.

From the artifact analysis, we could clearly observe the mistakes that students made in the data flows. For instance, it can be seen in Figure 1 on the right that the students forgot to connect the widget *Tree* (decision tree model on a pink background) with the *Predictions* widget (crystal ball on a blue background). This type of error – failing to realize that making predictions requires not only the remaining data but also the trained model – was frequently observed in other student groups as well.

The weekly surveys on students' motivation showed that students perceived the work with puzzle-like data cases as rather interesting, as can be seen in Figure 3. The feeling of pressure increased during the final weeks of the course, during which the students prepared for the final project work.

4.4.3 Outcomes. In order to understand what each student knew and could do by the end of the course, the teachers decided to make the project work shorter and to be completed alone. The project consisted of two parts: theoretical and practical. For the theoretical part, the students were asked about their understanding of data concepts such as regression, classification, and data flow. In the practical part, the students received a dataset, a problematic situation, and two questions. Without guidance, they had to create a data flow that answered two questions and produce a short report on the outcomes.

The analysis of the data flows and reports showed that all students could perform tasks related to data exploration without guidance. However, most students made errors when training and/or testing regression and decision tree models (challenge 16). For example, some students connected the widgets correctly, but they selected the wrong target variable and predictors. Others tested the model with a training dataset or selected the wrong diagram to visualize the model training results (e.g., a confusion matrix for linear regression). Only one student could articulate and interpret the coefficients of linear regression. Most students had difficulties interpreting the decision tree model (challenge 17).

Despite them not being able to create working data flows, we observed students trying to debug the data flows. For example, one student tried to understand the problem by looking at the output of the widget in a data flow: *"I made a change to the Data Sampler, which is why there are 906 outputs again. However, the predictions and the confusion matrix are still not working properly" (Student 37).*

There was a noticeable drop in T3 students' perception of teaching conditions regarding content relevance and error culture in the classroom (challenge 18). Students perceived the practical relevance and course content as less significant for their personal advancement. The quality of motivation also changed, as illustrated in Figure 5. Amotivated learning was significantly higher at the end of T3 than in T1 and T2. Despite increased communication among students in T3, there were no noticeable changes in student social integration (challenge 19).

Since there were no differences in students' self-efficacy in T3 compared with T1, and since we observed a decline in motivation starting in week 5 of T3, unlike T2, we conclude that the mediating processes in T3 were less engaging than in T1 and T2. This finding corresponded with the teachers' views on student motivation. The teacher repeatedly mentioned that, since the students were learning and struggling, they perceived the lessons as less engaging.

4.4.4 Mechanisms for the Puzzle-Like Teaching Approach. From observed mediating processes and outcomes, we conclude that in a course built around puzzle-like data case studies, where the focus is on a deep understanding of data concepts and practices ...

Mechanism 7: ... the students examine the data flows in a criterion-based and communicative manner by reconstructing the data flows, verbalizing their steps and results, articulating errors, and verifying solutions in groups. At the end of the course, all students are able to explore and pre-process the data using the data practices covered in the course, while only some students are able to use data to create ML systems without errors.

Mechanism 8: ... the teachers clearly observe difficulties and misunderstandings about data concepts and practices among students and feel that they are able to address those in their future teaching. By the end of the course, they assess their students as being capable of creating their own data flows without guidance.

Mechanism 9: ... the motivation among students with an average to high level of self-efficacy is slightly above average, and at the end of the course, the students perceive the teaching conditions, such as content relevance, as neutral, while being more extrinsically than intrinsically motivated.

5 Discussion and Local Theories

In this study, we investigated mechanisms that help students understand data concepts and practices, as well as motivate and enable them to design simple ML systems. Based on the described approaches, challenges, and mechanisms, we explicate five local instructional theories (LIT) about which design features are suitable in which learning situations, thus answering RQ2.

After testing different approaches with students and analyzing the mediating processes and outcomes, we identified three approaches of the data case study method as being suitable for schools: top-down, bottom-up, and puzzle-like. In T1, students learned with the bottom-up data cases, in T2 with the top-down data cases, and in T3 with the top-down and puzzle-like data cases. All three approaches allow students to go through the data flow,

independently discover insights, and thus learn about data concepts and practices in a practical way in the classroom. Although the approaches were developed sequentially, the mechanisms that they cause show that the approaches are equally valid. One approach cannot overcome all challenges at once. Depending on the learning objective and class, the teacher can decide which approach to use in the classroom. Some challenges, such as a lack of programming skills (challenge 1), are addressed by all of them. Others, such as reflection on the process (challenge 5), are better served by one of the three approaches. **LIT1:** All three data case study approaches are appropriate for the classroom, provided that the groups of students are relatively small, have mixed prior knowledge of CS, an above average level of self-efficacy, some experience with spreadsheets, and a basic understanding of the domain that the data comes from.

The deeply contextualized, bottom-up data case approach guided students through data practices and concepts step by step. While the students found this approach interesting, the only mediating process we observed was their ability to compose data flows according to instructions and interpret the results in context. Observing this process is insufficient to determine what is difficult for students or what they do not understand. After engaging with the data cases multiple times, only highly persistent students were able to create their own data flows for a given context by correctly ordering the provided widgets under heavily scaffolded conditions. Students mostly worked quietly and alone. **LIT2:** If the goal is to motivate students to learn about data by deeply embedding content into the subject matter and providing a step-by-step introduction to a specific data concept or practice, and if the students are highly persistent and able to work independently, then the bottom-up approach is an appropriate choice for teachers. This approach requires teachers to proactively monitor students' progress, as working with bottom-up data cases does not provide much insight into the mediating processes, and students' difficulties are not immediately apparent (challenge 4).

The deeply contextualized, top-down data case approach provided students with an overview of a complete data flow, guiding them to reflect on and critically evaluate data practices. Students found working with top-down cases motivating. When students focused on critically reflecting on the data flow and getting an overview of the concepts and phases, they could create data flows for a given context by determining which widgets they needed based on the description of the sub-steps. **LIT3:** If the goal is to provide intuition for designing a simple ML system, from loading data to interpreting results deeply embedded in the domain, to reflect on the appropriateness of the data practices based on criteria, or to motivate students, and if the students have an average level of self-efficacy, then the top-down data case is an appropriate choice. There is less teacher involvement than in the bottom-up data case because the teacher can immediately monitor how far the students have progressed on the one-page data case and if they need help.

In the puzzle-like approach, students reconstructed a data flow from given elements in collaborative work, without the use of computers. Although this approach was not ranked highest in motivation surveys, it was the most useful for teachers to observe learning and identify students' difficulties. Teachers also generated ideas on how to address these issues. When students reconstructed the data flows, they could develop skills necessary for independently applying the data practices to construct the data flows. Despite these advantages, students perceived the content relevance of the puzzle-like data case as less significant and felt more externally than internally motivated. **LIT4:** If the goal is to enable students to practice articulating data concepts and practices, help teachers understand difficulties and misunderstandings that students have about them, teach a deep understanding of implicit data concepts explicitly, and teach skills for identifying errors in data flows – skills that are critical for the ability to apply data practices to create simple ML systems independently – and if the students have above-average self-efficacy, then a puzzle-like data case is an appropriate choice. The cost for teaching with this approach is students' superficial understanding of the context, lower perception of the content relevance, and lower intrinsic motivation (challenge 18).

In all three cycles, students were engaged and designed their own ML systems within a flow-based environment. This environment provided all our students with low-floor access to data practices and allowed them to create simple and complex ML systems, thus addressing the challenge of heterogeneous programming knowledge. The final projects submitted by the students were all different. **LIT5:** A flow-based programming environment, such as Orange3, is powerful enough to support many paths and styles. As a constructionist environment, it can serve as a foundation for learning data concepts and practices through design.

Based on these findings, if the goal is to motivate students to learn about data while fostering agency in a real school setting, enabling students to independently conduct data practices in a project and interpret the results of their work, then building a course around top-down and puzzle-like architectures in a data-flow environment, such as Orange3, is particularly promising. The approaches address the reported challenges of teaching data concepts and practices, help students develop problem-solving abilities, and adhere to school-specific learning goals, such as collaborative learning.

6 Conclusion

Investigating mechanisms is a time-consuming, iterative process that requires significant effort. To understand what works and why, one must change the conditions, observe the challenges, and reflect on them. In our study, we provided insights into how the investigation of mechanisms can be conducted using a conjecture-mapping approach. Based on the results, we outlined nine empirically and theoretically sound mechanisms for teaching data concepts and practices in secondary school AI education and explicated five local instructional theories. Future research should focus on understanding how combining the proposed teaching approaches can effectively contribute to students' ability to design ML systems using data practices and concepts.

Acknowledgments

We would like to thank the school students who participated in this study, as well as the teachers and the domain expert from the GEOMAR Helmholtz Centre for Ocean Research Kiel. We used DeepL Write and ChatGPT to support language refinement in this paper. We take full responsibility for its final form.

References

[1] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. 2023. Explainable Artificial Intelligence (XAI): What We Know and What Is Left to Attain Trustworthy Artificial Intelligence. *Information Fusion* 99 (Nov. 2023), 101805. doi:10.1016/j.inffus.2023.101805

[2] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. 2019. Software Engineering for Machine Learning: A Case Study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. IEEE, Montreal, QC, Canada, 291–300. doi:10.1109/ICSE-SEIP.2019.00042

[3] Ravinithesh Annapureddy, Alessandro Fornaroli, and Daniel Gatica-Perez. 2025. Generative AI Literacy: Twelve Defining Competencies. *Digit. Gov.: Res. Pract.* 6, 1, Article 13 (Feb. 2025). doi:10.1145/3685680

[4] Albert Bandura. 1977. Self-Efficacy: Toward a Unifying Theory of Behavioral Change. *Psychological Review* 84, 2 (1977), 191–215. doi:10.1037/0033-295X.84.2.191

[5] 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.

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

[7] Valentina Chkoniya. 2021. Success Factors for Using Case Method in Teaching Applied Data Science Education. *European Journal of Education* 4, 1 (April 2021), 77–86. doi:10.26417/236hbm84v

[8] Aayushi Dangol and Sayamindu Dasgupta. 2023. Constructionist Approaches to Critical Data Literacy: A Review. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*. ACM, Chicago IL USA, 112–123. doi:10.1145/3585088.3589367

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

[10] Sinem Demirci, University of California, Irvine, Mine Dogucu, University of Minnesota, Andrew Zieffler, University of Knoxville, and Joshua Rosenberg. 2024. Learning Difficulties of Introductory Data Science Students. In *Proceedings of the IASE 2023 Roundtable Conference - Fostering Learning of Statistics and Data Science*. International Association for Statistics Education, Toronto, Canada. doi:10.52041/iase2023.501

[11] Matthew W. Easterday, Daniel G. Rees Lewis, and Elizabeth M. Gerber. 2018. The Logic of Design Research. *Learning: Research and Practice* 4, 2 (July 2018), 131–160. doi:10.1080/23735082.2017.1286367

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

[13] T. Grandon Gill. 2011. *Informing with the Case Method: A Guide to Case Method Research, Writing, & Facilitation.* Informing Science Press, Santa Rosa.

[14] 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 Dagienė (Eds.). Springer International Publishing, Cham, 147–158.

[15] 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. doi:10.1007/s10664-021-09993-1

[16] Sebastian Habig, Janet Blankenburg, Helena Van Vorst, Sabine Fechner, Ilka Parchmann, and Elke Sumfleth. 2018. Context Characteristics and Their Effects on Students' Situational Interest in Chemistry. *International Journal of Science Education* 40, 10 (July 2018), 1154–1175. doi:10.1080/09500693.2018.1470349

[17] Anna Hartl, Elena Starke, Angelina Voggenreiter, Doris Holzberger, Tilman Michaeli, and Jürgen Pfeffer. 2024. Empowering Digital Natives: InstaClone - a Novel Approach to Data Literacy Education in the Age of Social Media. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1 (Sigcse 2024)*. Association for Computing Machinery, Portland, OR, USA and New York, NY, USA, 484–490. doi:10.1145/3626252.3630839

[18] Orit Hazzan and Koby Mike. 2024. *Guide to Teaching Data Science: An Interdisciplinary Approach.* Springer, Cham.

[19] Birte Heinemann, Simone Opel, Lea Budde, Carsten Schulte, Daniel Frischemeier, Rolf Biehler, Susanne Podworny, and Thomas Wassong. 2018. Drafting a Data Science Curriculum for Secondary Schools. In *Proceedings of the 18th Koli Calling International Conference on Computing Education Research*. ACM, Koli Finland, 1–5. doi:10.1145/3279720.3279737

[20] Clyde Freeman Herreid. 2011. Case Study Teaching. *New Directions for Teaching and Learning* 2011, 128 (Dec. 2011), 31–40. doi:10.1002/tl.466

[21] Frederick S. Hillier (Ed.). 2007. *Data Preparation in Neural Network Data Analysis.* Vol. 107. Springer US, Boston, MA, 39–62. doi:10.1007/978-0-387-71720-3_3

[22] Erin Rae Hoffer. 2020. Case-Based Teaching: Using Stories for Engagement and Inclusion. 2, 2 (2020), 75–80.

[23] Fiona M. Hollands, Daniella DiPaola, Cynthia Breazeal, and Safinah Ali. 2024. AI Mastery May Not Be for Everyone, but AI Literacy Should Be. In *Proceedings of the 2024 on ACM Virtual Global Computing Education Conference V. 1 (SIGCSE Virtual 2024)*. Association for Computing Machinery, Virtual Event, NC, USA and New York, NY, USA, 60–66. doi:10.1145/3649165.3690117

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

[25] Kaiyue Jia, Teresa H. M. Leung, Ngai Yan Irene Cheung, Yixun Li, and Junnan Yu. 2025. Developing a Holistic AI Literacy Framework for Children. *ACM Transactions on Computing Education* 25, 2, Article 21 (June 2025). doi:10.1145/3727986

[26] John D. Bransford, Ann L. Brown, and Rodney R. Cocking. 2000. *How People Learn: Brain, Mind, Experience, and School: Expanded Edition.* National Academies Press, Washington, D.C. 9853 pages. doi:10.17226/9853

[27] 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. doi:10.5220/0011895200003393

[28] Jana Lasser, Debsankha Manik, Alexander Silbersdorff, Benjamin Säfken, and Thomas Kneib. 2021. Introductory Data Science across Disciplines, Using Python, Case Studies, and Industry Consulting Projects. *Teaching Statistics* 43, S1 (July 2021). doi:10.1111/test.12243

[29] Duri Long, Aadarsh Padiyath, Anthony Teachey, and Brian Magerko. 2021. The Role of Collaboration, Creativity, and Embodiment in AI Learning Experiences. In *Creativity and Cognition (C&C '21)*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3450741.3465264

[30] Lilian Lopez, Zeyu Xiong, Kiara Chau, Gustavo Umbelino, Zihan Wu, and April Wang. 2025. datAR: A Situated Learning Approach for Data Literacy through Everyday Objects. In *Proceedings of the 30th ACM Conference on Innovation and Technology in Computer Science Education V. 1 (ITiCSE 2025)*. Association for Computing Machinery, Nijmegen, Netherlands and New York, NY, USA, 152–158. doi:10.1145/3724363.3729047

[31] Joseph A. Maxwell. 2004. Causal Explanation, Qualitative Research, and Scientific Inquiry in Education. *Educational Researcher* 33, 2 (March 2004), 3–11. doi:10.3102/0013189X033002003

[32] 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. doi:10.1007/978-3-031-43393-1_19

[33] Stephan Napierala. 2020. The Road to Finding Interesting Contexts for Teaching Data Literacy at School. In *Proceedings of the 15th Workshop on Primary and Secondary Computing Education*. ACM, Virtual Event Germany, 1–2. doi:10.1145/3421590.3421620

[34] Deborah Ann Nolan and Terry Speed. 2001. *Stat Labs: Mathematical Statistics through Applications* (corr. 2. print ed.). Springer, New York Berlin Heidelberg.

[35] OECD. 2025. *Empowering Learners for the Age of AI: An AI Literacy Framework for Primary and Secondary Education (Review Draft).* OECD, Paris.

[36] Viktoriya Olari and Ralf Romeike. 2024. Data-Related Concepts for Artificial Intelligence Education in K-12. *Computers and Education Open* 7 (Dec. 2024), 100196. doi:10.1016/j.caeo.2024.100196

[37] Viktoriya Olari and Ralf Romeike. 2024. Data-Related Practices for Creating Artificial Intelligence Systems in K-12. In *Proceedings of the 19th WiPSCE Conference on Primary and Secondary Computing Education Research*. Association for Computing Machinery, Munich, Germany.

[38] 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. doi:10.1007/978-3-031-43393-1_21

[39] Olari, Viktoriya and Romeike, Ralf. 2025. Data Case Study – A Teaching and Learning Method for Computer Science Education in Schools. In *3rd ACM Global Computing Education Conference (CompEd)*. ACM.

[40] Neoklis Polyzotis, Sudip Roy, Steven Euijong Whang, and Martin A. Zinkevich. 2018. Data Lifecycle Challenges in Production Machine Learning. *ACM SIGMOD Record* 47 (2018), 17–28.

[41] Susanne Prediger, Koeno Gravemeijer, and Jere Confrey. 2015. Design Research with a Focus on Learning Processes: An Overview on Achievements and Challenges. *ZDM* 47, 6 (Oct. 2015), 877–891. doi:10.1007/s11858-015-0722-3

[42] Gil Press. 2016. Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says. *Forbes* (March 2016).

[43] Ashley S Quiterio. 2025. Agency in Data Interactions: Teaching and Learning for Personal Data Literacy. In *Proceedings of the 24th Interaction Design and Children (Idc '25)*. Association for Computing Machinery, New York, NY, USA, 1192–1195. doi:10.1145/3713043.3731607

[44] 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). doi:10.13140/RG.2.1.1922.5044

[45] William Sandoval. 2014. Conjecture Mapping: An Approach to Systematic Educational Design Research. *Journal of the Learning Sciences* 23, 1 (Jan. 2014), 18–36. doi:10.1080/10508406.2013.778204

[46] D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. 2015. Hidden Technical Debt in Machine Learning Systems. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'15)*. MIT Press, Cambridge, MA, USA, 2503–2511.

[47] Ben Rydal Shapiro, Amanda Meng, Cody O'Donnell, Charlotte Lou, Edwin Zhao, Bianca Dankwa, and Andrew Hostetler. 2020. Re-Shape: A Method to Teach Data Ethics for Data Science Education. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–13. doi:10.1145/3313831.3376251

[48] 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. doi:10.1145/3488042.3488053

[49] Kamilla Tenório, Viktoriya Olari, Margarita Chikobava, and Ralf Romeike. 2023. Artificial Intelligence Literacy Research Field: A Bibliometric Analysis from 1989 to 2021. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*. ACM, Toronto ON Canada, 1083–1089. doi:10.1145/3545945.3569874

[50] 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. doi:10.1609/aaai.v33i01.33019795

[51] 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.

[52] Jan Van Den Akker. 1999. Principles and Methods of Development Research. In *Design Approaches and Tools in Education and Training*, Jan Van Den Akker, Robert Maribe Branch, Kent Gustafson, Nienke Nieveen, and Tjeerd Plomp (Eds.). Springer Netherlands, Dordrecht, 1–14. doi:10.1007/978-94-011-4255-7_1

[53] Carrie Wright, Qier Meng, Michael R. Breshock, Lyla Atta, Margaret A. Taub, Leah R. Jager, John Muschelli, and Stephanie C. Hicks. 2024. Open Case Studies: Statistics and Data Science Education through Real-World Applications. *Journal of Statistics and Data Science Education* (2024), 1–30. doi:10.1080/26939169.2024.2394541

[54] 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. doi:10.1137/1.9781611977653

[55] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. 2025. Data-Centric Artificial Intelligence: A Survey. *Comput. Surveys* 57, 5 (May 2025), 1–42. doi:10.1145/3711118