About Classes and Trees: Introducing Secondary School Students to Aspects of Data Mining

This is an author draft of the paper

Grillenberger A., Romeike R. (2019) About Classes and Trees: Introducing Secondary School Students to Aspects of Data Mining. In: Pozdniakov S., Dagienė V. (eds) Informatics in Schools. New Ideas in School Informatics. ISSEP 2019. Lecture Notes in Computer Science, vol 11913. Springer, Cham

The original publication is available at: https://doi.org/10.1007/978-3-030-33759-9_12

Andreas Grillenberger[0000−0003−1760−2051] and Ralf Romeike

Freie Universität Berlin, Computing Education Research Group
Königin-Luise-Str. 24/26, 14195 Berlin, Germany
{andreas.grillenberger,ralf.romeike}@fu-berlin.de

Abstract. Today, data is no longer just important to computer science. Instead, basic competencies in managing, processing and using data are necessary in almost all other sciences and even in everyday life. Such competencies empower students to handle their own and others' data adequately and allow them to use data-related technologies and tools in a critically-reflected way. Although aspects of this topic are typically already part of computer science curricula for secondary schools, particularly fostering data-related competencies is often not the focus, so that large parts of this exciting topic have not arrived in the classroom yet.
In this paper, we investigate the exemplary topic *data analysis and predictions* from a secondary education perspective. After summarizing the technical and didactic foundations, we describe a theoretically sound teaching concept which aims to foster the acquisition of basic competencies in this field and to contribute to a better understanding of these important aspects of the digital world. Besides presenting the teaching concept, the paper discusses the methodical structure as well as the software tool used. In addition, the mostly positive results and impressions of an evaluation with ninth-grade students are presented.

Keywords: Data · Data Literacy · Data Mining · Data Analysis · Prediction · Teaching Concept · Secondary Education · Evaluation

1 Data in the Digital World

In today's world, data is an important basis of manifold developments which are often subsumed under the term *digitalization*. However, although everyone continuously generates and stores various data, when using those we typically take a rather passive role: Evaluating and processing all the data is mostly left to companies, whose products and services we use. Even more important is the limited or often lacking understanding of how such analyses work and hence also for their accompanying phenomena and impact: Despite an extensive discussion in the social discourse, it is difficult for large parts of the population to assess the power, possibilities and dangers of data analysis. Hence, they hardly have the opportunity to position themselves accordingly. This is particularly important as today various decisions related to using data-driven services have to be made

taking into account the personal cost-benefit ratio and the effects on society, for example when motor vehicle insurances desire to record driving behaviour. The ability to reflect such developments in a critically-reflective manner is particularly important when data-driven approaches are used in the background and/or without allowing people to decide for or against participating in this system: for example, rating persons based on data is not only carried out by well-known credit agencies, but increasingly also by government institutions. In China this development is already so far advanced that all inhabitants will soon be scored positively or negatively as part of a *social credit system*, with the aim to educate them to a desired behaviour [1]. The consequences associated with such developments can hardly be evaluated without a sound basic knowledge of how data are handled and used, otherwise the extent and possibilities remain hidden. In order to prepare for a self-determined and mature life in the digital society, school— and in this case particularly computer science teaching—has to provide insight into such developments and empower students to reflect them in a critically-reflected manner. Although different approaches for fostering basic competencies regarding handling and usage of data are already recognizable in computer science lessons, there is still a large gap [5] which indicates that computer science teaching in this area still has to develop further.

In this paper, we present a theoretically sound teaching concept focusing on *data analysis* and *prediction*. In the course of the lessons, students are not only given the chance to gain insight into the function of data analyses and predictions, but they are also empowered to carry out their own analyses based on real data sets and thus to fathom both the power and limits of automated data analyses. The teaching concept promotes a critical examination of everyday handling of data, but also enables students to deepen their experiences independently. In the following, we first summarize the state of research in this area and give an overview of CS teaching in this context. Then, before the teaching concept is presented in section 3, the focus of the concept on aspects of data mining is discussed, relevant technical contents are outlined and the selection of the tool used is discussed. Finally, in section 4 we describe the predominantly positive experiences gained during an evaluation of the teaching concept at school.

2 Current State of Teaching and Research

From a scientific perspective, the field *data* is particularly important: Not only is it an important basis for all developments summarized under terms such as *big data*, *data mining* and *data science*, but also for example in *machine learning*. But such developments are also triggering changes in other subjects, in particular related to research: For example, Hey et al. [10] emphasize the relevance of data-oriented research as a new research paradigm. In this context, also the need for fostering data competencies for everyone is stressed. The basic competencies everyone needs in this context are often summarized under the term *data literacy*. Ridsdale et al. [14] describe it as "the ability to collect, manage, evaluate, and apply data, in a critical manner" and, in a summative research approach, also

describe several key competencies related to this field. According to the common understanding, data literacy competencies have to be differentiated from such that are related to data science: Data science requires deeper competencies and focuses on professionally oriented aspects of the field *data*.

In computer science curricula, at the moment the extensive field *data* is mainly taken up with a focus on *databases* [5][1]. In this context, central basics are considered: in particular by introducing data models, the importance of defined data structures and of data types becomes evident, while at the same time key concepts are introduced, such as *redundancy*, *consistency* and *durability*. Beyond *databases*, however, data have rather little importance in current CS teaching: Although they play a certain role in programming and are also indispensable in topics such as *structure of the Internet*, the focus of these topics is usually different, so that data and related concepts are only considered marginally. Thus, the current role of this topic in the classroom hardly reflects its general importance. Accoringly, only few teaching concepts could be found which go a step further and take a broader look at the complete field. However, there are at least approaches that take up aspects of the topic which are typically less relevant in school teaching. For example, in a German simulation game [3], the idea of data protection is introduced. In another example, a teacher implemented a platform called InstaHub that allows students to build their own social network as well as a related teaching concept which allows students to get valuable insight into such platforms [4]. These and similar concepts deal with central topics of the field data. Yet, for several increasingly important areas such as data analysis and prediction, corresponding ideas are not yet to be found.

In computer science education research, however, progress has been made in recent years, particularly concerning the foundation of data-related aspects from an educational perspective: On the one hand, from a rather technical point-of-view, the entire subject area *data management* was investigated with the goal to identify key concepts and practices characterizing this field and to also consider recent developments [7]. Yet, as not only the technical aspects play an important role for teaching, also the perspectives of students, teachers and society as a whole have been investigated [8]. As part of this work, also a competency model of *data literacy* was developed, which, in contrast to the one by Ridsdale et al. takes the perspective of CS education and hence sets different foci [6], but in general both models are consistent. This competency model consists of four content and process areas each (cf. fig. 1). Thus, it is not only focused on the concepts or technical content of this subject area, but also emphasizes the practical perspective on the topic. However, this project is not the only approach to consider rather modern data-oriented aspects in school: For example, in a joint project with mathematics education, currently Heimann et al. [9] develop a data science curriculum for secondary schools, which sets similar foci as the competency model described before. Despite these different approaches to inves-

[1] Although the study [5] is about five years old, there were only few changes in CS curricula in the last years.

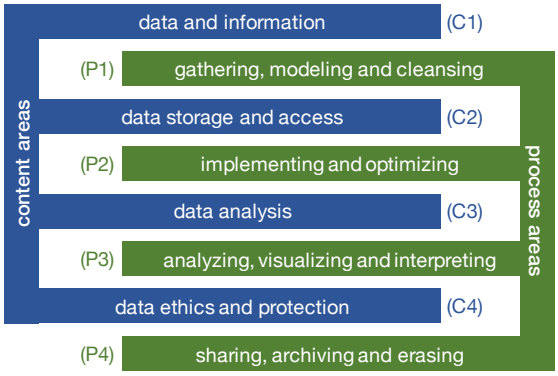

Fig. 1. The data literacy competency model used as basis for the lesson sequence [6].

tigate the topic from a CS education perspective, the topic has so far remained a marginal topic in computer science education research and teaching.

3 Presentation of the Teaching Concept

In order to address the previously characterized gap in current computer science teaching, the lesson sequence presented aims to foster a basic understanding and the acquisition of competencies related to data analysis and predictions. The planned lesson sequence is designed for only four lessons of 90 minutes each, so that it can be flexibly included into teaching. However, of course it can be adapted individually if necessary, since a much more in-depth or somewhat more superficial examination is possible at various points. In general, we cannot expect students to have basic knowledge in how data analysis or predictions work, therefore the lesson sequence was designed in such a way that no previous knowledge needs to be built upon, which also makes it suitable for both lower and upper secondary level.

In order to decide which competencies are emphasized, we used the data literacy competency model mentioned before (cf. fig. 1). This model is not only well-founded in CS education research but, by distinguishing content and process areas, it also helps to focus on both, content and practice, at the same time. In order to achieve the desired goals, from a content perspective the focus was set on *data analysis (C3)*, while also general aspects of *data and information (C1)* have to be taken into account, as we do not expect any prior knowledge. In addition, because of the relevance of this topic in society, also data *ethics and protection (C4)* cannot be left out, so that from a content perspective, we take up parts of all content areas except *data stores and data storage (C2)*. As we have designed this teaching concept so that it can be taught in only four lessons, we also could not include aspects from all process areas. Thus, we focus on the analysis and prediction itself, so that *P4 (analyzing, visualizing and interpreting)* is emphasized while the other process areas are at most considered marginally.

Hence, in the teaching sequence we particularly aimed at fostering the following data literacy competencies:

- explain why and how (possibly new) information can be obtained from stored data (C1/P3)
- characterize the difference between correlation- and causality-based relations in data as well as their respective meaningfulness (C1/P3, C4/P3)
- sketch the process of a (correlation-based) data analysis (C3/P3)
- characterize a typical analysis method and explain the underlying principle using a suitable example (C3/P3)
- perform a simple data analysis using a common method, manually as well as using a suitable software tool (C3/P3)
- predict missing attributes of a data set using a self-conducted data analysis (C3/P3)
- evaluate the outcome of the prediction and explain ideas for improvement (C3/P3)
- reflect the results taking into account ethical and social implications (C4/P3)

As we did not expect any prior knowledge, we decided to divide the course into two blocks: Initially, an introduction to the data analysis process is given in order to allow students to understand how the analysis process works. For this purpose, it is useful to focus on specific methods and on getting an overview of the analysis process. Afterwards, we introduce the students to some basics of data analysis and predictions without using digital analysis tools, as for this purpose we do not need to use larger amounts of data, so that analysis can be carried out manually in order to understand the important principles. Afterwards, in order to be able to estimate the potential and risks of automated data analysis and to get more valid and realistic analysis results, a software tool is used for analyzing data. At this step, we also switch to a larger data set that leads to more interesting results. While in the beginning we focus on a fictive data set (from the context of online shopping), the data set selected for the second block even more directly affects students, so that it leads to a critical discussion of the results and the analysis itself.

3.1 Basics: Data Mining, Classification and Prediction

Before we describe the lesson sequence in detail, we need to focus on the relevant basics of data mining and data analyses. While classical data analyses often pursue the goal to structure and summarize existing data, particularly by using aggregate functions in order to describe the data by minimum, maximum and average values, *data mining* follows a different approach: It focuses on discovering new information and often on predicting unknown attributes of a data set based on other data. The term *data mining* can be understood as an analogy to *gold mining*: It describes digging for valuable information in a large mountain of data. Different methods are used for this purpose, of which most can be traced back to the basic principles *classification*, *clustering* and *association*:

– *Classification* refers to dividing a data set into several classes. Often, the goal of a classification is to predict unknown attributes of one instance of the data set by looking at all the other instances of the same class. However, as using classification, the classes cannot be inferred from the data, the existing classes need to be known: For example, for classifying students by their performance in school, the classes may be derived from the grades, for classifying them by how far they live from school, it must be decided on which classes are introduced (such as less than $5\,km$, more than $5\,km$).
– *Clustering* adresses the limits of classification: It pursues the same goal as classification, but in this case the clusters are not predetermined but instead determined inductively from the data. Often, the rules for assigning instances of a data set to a certain cluster are an important result of clustering. So, for example, groups of people (i. e. clusters) knowing each other might be determined in social networks.
– *Association analysis* focuses on discover rules that describe a data set that either explain causal relationships or can be based on correlations. They are particularly useful for predicting unknown values. However, particularly in large data sets, finding associations is a complex task.

In the lesson sequence, the focus is on the methods *classification* and *association*, which are easy to understand, but also allow getting insight into how data analyses work. In this case, classification is used to find similarities in the data, to structure them accordingly and to derive findings from it, which are elaborated into rules as part of an association analysis. Thus, both methods go hand in hand and show how predictions work: By learning rules from an already classified data set, an automated classification of further data can take place, which allows for predicting unknown attributes of these data. In order to automate such analyses, a multitude of different classification algorithms exists, which often become very complex and are therefore not discussed in detail in the classroom. Instead, we focus on a basic method, the *classification tree*. These trees can be used as an intuitive approach to gaining an overview on association rules, as these rules are visualized as a decision tree, whose nodes represent decisions and whose leaves are used for class allocation. Hence, based on such a tree, predictions can easily be made, just by looking at a specific data set and following the tree's nodes from the root to a leave.

Besides the analysis methods, also the analysis process leading to a prediction is an important part of the teaching sequence: For making valid high-quality predictions, it is particularly important to consider the whole analysis process, as the quality is i. a. influenced by the selection of the sample on which the associations are determined. Hence, in the teaching sequence, we also give an overview on this process and relate the methods discussed in school to the overall process to give students an orientation during the analysis process (cf. fig. 2).

3.2 Tool Selection: the Data Mining Tool *Orange*

While no software tool is necessary for the first part of the lesson series, in the second part the aim is to make the power and potential of data analysis

Fig. 2. Analysis process discussed during the lesson sequence.

visible for the students by enabling them to conduct their own analysis with real data. For this purpose, a suitable tool is needed. Again, the selection of the tool is particularly led by the criterion, that it should not require any prior knowledge and be intuitively usable. When selecting the tool, we also took the criteria into account that Resnick et al. established for tools that support creative thinking [13]: These tools should "*make it easy for novices to get started (low threshold)*"[2] [13, 12], make it possible "*for experts to work on increasingly sophisticated projects (high ceiling)*" [13, 12] and "*support and suggest a wide range of explorations*" [13] (*wide walls*). Accordingly, the use of a classical programming language such as Python, which is very common for professional data analyses, is hardly reasonable for this lesson series. Instead, graphically oriented analysis tools are particularly suitable, especially tools in which users describe the analysis as a data flow model, for example *rapidminer*[3] and *Orange*[4]: these tools allow students to directly transfer the knowledge on the analysis process to the automated data analysis. Both tools provide all the functionalities that we need for the specific lesson sequence, but they also offer many additional possibilities, so that they could be used also for more sophisticated analyses. We finally decided to use *Orange* in the classroom for two reasons: First, it was possible to further reduce the complexity of the tool since it was an open source program in which unneeded modules could be hidden. Furthermore, *Orange* can be used and distributed without any license to be required, while *rapidminer* requires both teachers and students to apply for an academic license, which is a barrier for using the tool. Some impressions of *Orange* are given later in figs. 3 to 5.

3.3 Description of the Lesson Sequence

In the following, we give an overview of the course design. For a detailed description, refer to the overall concept published on the project website[5], which contains all the work materials and a detailed description for teachers.

In the first 90 minute lesson, the central aspects of the analysis process are emphasized. For motivating the importance of the topic, at the beginning of the lesson sequence a newspaper article is presented, that describes the attempt of a US retailer to recognize whether its customers are pregnant in order to send

[2] Later, *low threshold* was also refered to as *low floor*.
[3] `https://rapidminer.com/educational-program/`
[4] `https://orange.biolab.si`
[5] `https://dataliteracy.education`

them targeted advertisments [11]. This article encourages students to discuss how the retailer can determine that a customer is pregnant, which attributes about its customers it probably collects and how these attributes could lead to assuming a pregnancy. Thus, this discussion immediately draws students into the topic and also gives the teacher an impression of what students know about data analyses and how they think these work. Afterwards, based on other examples, the value and usefulness of data for different purposes, companies and business models is discussed and students can provide own examples they know from their daily lives. The teacher directs these discussions towards introducing the terms *causality*, *correlation* and *prediction*, which are relevant to be known for the complete lesson sequence. Starting with the students' ideas, afterwards a model of data analysis processes is being created.

Based on this introduction, in the next lesson, the process from a data set to a prediction is carried out manually, so that different principles of these analyses become recognizable, particularly the importance of a sufficient data sample, the selection of valid rules/associations and the creation of the data model. For this lesson, students work on a given simple and fictive data set (in the example from the context of online retail) and examine it for relations within the data, which are then formulated as rules (i. e. associations). In order to make these rules easier to grasp, to give a better overview of them and to simplify applying them to data, a (non-binary) decision tree is introduced as a form of representation. Afterwards, using this tree, the rules are applied to another data set with unknown attribute values. The given data set was designed in such a way that not all possible rules apply to all instances of the data set. Thus, students need to decide whether they consider an association as valid that is only valid for about 80% of the data. Hence, they also need to reflect about the targeted analysis quality and about problems resulting from unfavorably chosen rules. At the end of the lesson, the central task for the next lesson is introduced: to predict students' school grades based on a real data set.

In the third lesson, a freely accessible data set with (anonymised) data about Portuguese students [2] is analysed. This data set contains various personal data about more than 600 students (e. g. jobs of the parents, amount of spare time) as well as their grades in three exams. Using the tool *Orange*, which is not introduced in detail to the students, the aim is to generate a decision tree and hence a prediction model, which is then used for predicting the third grade from all the remaining data. As Orange not only allows viewing the actual results of the analysis, but also getting insight into the intermediate steps, students can for example have a look at the generated classification tree (fig. 4), but also on the data sample that was selected randomly. By discussing the teacher's fictional goal of significantly reducing the correction effort by using analysis and predictions, students get into questioning and evaluating how good the analysis quality can get in comparison with the teacher's effort (size of the data sample). This is particularly interesting because of the high analysis quality, which students can for example observe by examining a confusion matrix (fig. 5). Thus, the direct

involvement of the students holds a high potential for critically discussing the possibilities and threats of data analyses and predictions.

The last lesson focuses on additional use cases and on a critical reflection of those: For this purpose, it is planned to transfer the competencies acquired so far to other contexts and thus, for example, to question the use of data in medicine, by insurance companies and banks and to discuss legal, ethical and moral aspects of these analyses within the framework of a jigsaw puzzle.

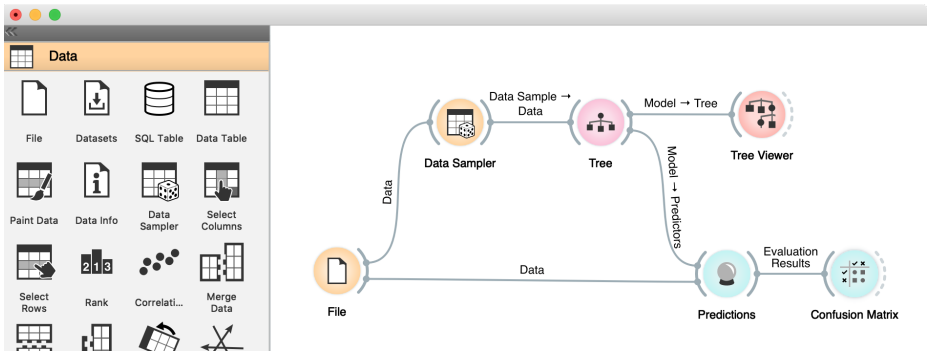

Fig. 3. Model of a data analysis in the analysis tool *Orange*.

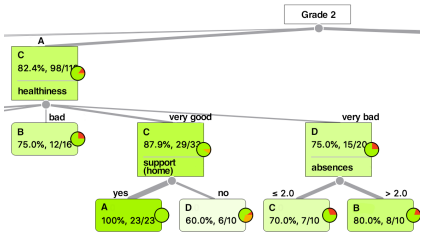

Fig. 4. Part of a classification tree in *Orange*.

actual \ predicted	1.0	2.0	3.0	4.0	5.0	6.0
1.0	34	10	2	0	0	0
2.0	1	52	31	1	0	0
3.0	0	1	182	34	0	0
4.0	0	0	5	212	19	0
5.0	0	0	0	9	40	0
6.0	0	0	0	9	7	0

Fig. 5. Confusion matrix for examining the analysis quality.

4 Evaluation in School

The lesson sequence was evaluated in two ninth grade classes of a German secondary school with 15 (three female) and 12 (no female) students respectively. Both classes were taught by the same teacher and had no pre-knowledge from teaching regarding data in general, databases or even programming. For organisational reasons, only three lessons (90 minutes each) were available, hence we

had to drop the fourth lesson and included some of the discussion aspects in the other lessons, so that those were very packed. The first author of this paper observed the lessons using observation sheets and conducted a guided interview with the teacher. Also, the students were interviewed with questionnaires at the end of the last lesson in order to cover different perspectives.

In general, the observations were rather positive: Although both courses were very different concerning students' interest in the topic and their motivation to participate in discussions, in general most students were taking part after a while, which was also surprising for the teacher, who did confirm this observation. Generally spoken, there was a large overlap between the researcher's and teacher's observations. In contrast, the students' perspective on the topic seemed a bit different in the questionnaires, but there was a high variation in their answers in almost all questions asked. As the questionnaire was handed to them as the last task of the lesson and hence also filled in by them very quickly, these results have to be considered carefully. Yet, taking together the three methods, there are some central results that can be deduced:

Interest and Motivation: Although the evaluation took place in two rather difficult, both inattentive and poorly motivated classes, both the classroom observation and the perception of the teacher showed a high interest and a motivated participation. According to the teacher, both the interest in the topic and the motivation seemed to be higher than for other topics. In addition, the frequent and intensive discussions were highly noticeable, which was unusual in particular for one of the classes. However, the student survey partly contradicted these observations, especially with regard to the perceived interest in the topic, which was rated as rather low.

Prior Knowledge and Experience: In the course of the lessons it became apparent several times that the examples used and the entire topic have strong references to the daily life of the students and to their experiences: In many cases, they additionaly brought their own examples and were able to find intuitive explanations for questions that arose. Based on the students' ideas, even the data analysis process could be deduced well. Thus, it became clearly recognizable that data and data analysis play an important role in students' life and that they have some ideas on how these work.

Comprehensibility: In general, the topics considered in the lesson sequence seemed to be comprehensible and understandable for the students. This was particularly evident in the discussion phases, where they often included aspects in their argumentations that had been considered earlier. Only the distinction between correlation- and causality-based data analysis occasionally led to difficulties, so that more time should be devoted to this aspect. Also, the students stated in the questionnaire that the tasks were easy to solve for them and that they now feel to have an understanding of what can be done with data. So, in general, at least the basics of the subject area are comprehensible for the students.

Structure and Tool Selection: The structure of the lesson sequence was appropriate from both the teacher's and observer's point of view. However, the

time should be distributed differently: The first block of manual data analysis was a bit too long, hence that it took considerable time before the students were allowed to conduct practical data analysis on the computer. This particularly seemed to lower their motivation as from previous teaching the students were used to working at the computer in every lesson, so this led to displeasure in the class. Therefore, a stronger integration of manual and automated data analysis has to be considered depending on what the students' are used to in class. Also, the lessons were a bit too packed, so that taking more time and adding at least a fourth lesson, as planned originally, seems necessary.

5 Summary

Summarizing, the developed lesson sequence could bring some important aspects of data analysis to school teaching. Particularly, it allows students to deal with this important topic and understand some of the underlying concepts. But it also helps them to develop some basic skills that are needed for conducting data analysis and for making predictions by themselves. It has turned out to be particularly important not to stop at the data analysis step, but instead to make predictions based on it, as this helped students to get into this topic, because they for example know that people are rated based on data in different contexts. Thus, the lesson sequence enabled students to understand aspects that are fascinating for them because of their "magical" effects, but can be explained with basic knowledge.

In general, the experiences of both, teacher and researcher, were very positive as this topic was not only important for the students, but as they also coped with the topic very well and developed a basic understanding. This particularly enabled them to discuss about this topic in a sound way and to conduct simple analyses by themselves. The active participation in the lessons confirmed the interest in the topic and the relevance for the learners. The choice of the software tool also seemed appropriate, as the students were not confronted with any challenges when using it and were able to master it intuitively.

However, the evaluation revealed aspects that should be taken into account in the future: In particular, the first phase with only manual data analyses was too long, as this contradicted the usual teaching. Accordingly, for motivational reasons, stronger intertwining the manual and automated data analyses will be sought in the future. In addition, it turned out that including further contexts, which was planned for the fourth lesson, was clearly lacking in the end, as the learners still found it difficult to relate the acquired competencies to new examples. Hence, shortening the topic to only three 90 minute lessons seems not advisable.

Overall, the developed teaching concept and its implementation and evaluation reveal the potential of the topic for school and also confirms the assumption that this topic appears important for the students and is important for their everyday lifes. In particular, we were also able to show that this topic, which is often regarded as complex, can be reduced for and addressed in teaching even

at lower secondary level, without having to abstract too much from the central aspects. Hence, the developed lesson sequence can be considered as a first step to bringing more data-oriented aspects to school teaching. However, when looking at the data literacy competency model used as a basis, there are many more aspects in this field that should clearly be adresses in teaching.

References

1. Botswan, R.: Big data meets Big Brother as China moves to rate its citizens, (2018). https://www.wired.co.uk/article/chinese-government-social-credit-score-privacy-invasion (visited on 09/06/2019)
2. Cortez, P., and Silva, A.: Using Data Mining to Predict Secondary School Student Performance. In: Proceedings of 5th Annual Future Business Technology Conference, Porto, 2008, pp. 5–12. EUROSIS-ETI (2008)
3. Dietz, A., and Oppermann, F.: Planspiel "Datenschutz 2.0". LOG IN (2011)
4. Dorn, J.: InstaHub, (2018). https://instahub.org (visited on 09/06/2019)
5. Grillenberger, A., and Romeike, R.: A Comparison of the Field Data Management and its Representation in Secondary CS Curricula. In: Proceedings of the 9th Workshop in Primary and Secondary Computing Education. ACM (2014)
6. Grillenberger, A., and Romeike, R.: Developing a Theoretically Founded Data Literacy Competency Model. In: Proceedings of the 13th Workshop in Primary and Secondary Computing Education. WiPSCE '18, 9:1–9:10. ACM (2018)
7. Grillenberger, A., and Romeike, R.: Key Concepts of Data Management: An Empirical Approach. In: Proceedings of the 17th Koli Calling International Conference on Computing Education Research. Koli Calling '17, pp. 30–39. ACM (2017)
8. Grillenberger, A., and Romeike, R.: What Teachers and Students Know about Data Management. In: Tomorrow's learning: Involving Everyone. Learning with and about Technologies and Computing - the 11th IFIP TC 3 World Conference on Computerr in Education, WCCE 2017, Dublin, Ireland, July 3-6, 2017, Revised Selected Papers. IFIP AICT, pp. 557–566. Springer, Heidelberg (2018)
9. Heinemann, B.: Drafting a Data Science Curriculum for Secondary Schools. In: Proceedings of the 18th Koli Calling International Conference on Computing Education Research. ACM, New York, NY, USA (2018)
10. Hey, T., Tansley, S., and Tolle, K.: The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, Redmond, Washington (2009)
11. Hill, K.: How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did, (2012). https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did (visited on 09/06/2019)
12. Myers, B.A., Hudson, S.E., and Pausch, R.: Past, Present and Future of User Interface Software Tools. ACM Transactions on Computer Human Interaction 7(1) (2000)
13. Resnick, M., Myers, B., Nakakoji, K., Shneiderman, B., Pausch, R., Selker, T., and Eisenberg, M.: Design Principles for Tools to Support Creative Thinking. National Science Foundation workshop on Creativity Support Tools (2005)
14. Ridsdale, C., Rothwell, J., Smit, M., Ali-Hassan, H., Bliemel, M., Irvine, D., Kelley, D., Matwin, S., and Wuetherick, B.: Strategies and Best Practices for Data Literacy Education: Knowledge Synthesis Report, Dalhousie University (2015)