Developing a Theoretically Founded Data Literacy Competency Model

Andreas Grillenberger
Friedrich-Alexander-Universität Erlangen-Nürnberg
Computing Education Research Group
Erlangen, Germany
andreas.grillenberger@fau.de

Ralf Romeike
Friedrich-Alexander-Universität Erlangen-Nürnberg
Computing Education Research Group
Erlangen, Germany
ralf.romeike@fau.de

ABSTRACT

Today, data is everywhere: Our digitalized world depends on enormous amounts of data that are captured by and about everyone and considered a valuable resource. Not only in everyday life, but also in science, the relevance of data has clearly increased in recent years: Nowadays, data-driven research is often considered a new research paradigm. Thus, there is general agreement that basic competencies regarding gathering, storing, processing and visualizing data, often summarized under the term *data literacy*, are necessary for every scientist today. Moreover, data literacy is generally important for everyone, as it is essential for understanding how the modern world works. Yet, at the moment *data literacy* is hardly considered in CS teaching at schools. To allow deeper insight into this field and to structure related competencies, in this work we develop a competency model of data literacy by theoretically deriving central content and process areas of data literacy from existing empirical work, keeping a school education perspective in mind. The resulting competency model is contrasted to other approaches describing data literacy competencies from different perspectives. The practical value of this work is emphasized by giving insight into an exemplary lesson sequence fostering data literacy competencies.

KEYWORDS

data, data literacy, data science, data management, competency model, CS education

ACM Reference Format:
Andreas Grillenberger and Ralf Romeike. 2018. Developing a Theoretically Founded Data Literacy Competency Model. In *Proceedings of the 13th Workshop in Primary and Secondary Computing Education (WiPSCE '18), October 4–6, 2018, Potsdam, Germany*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3265757.3265766

1 INTRODUCTION

In recent years, the perception and use of has changed considerably: While in the past, data was a topic for computer scientists only, nowadays it becomes increasingly relevant in all scientific fields. Based on tremendous advances, especially in the emerging fields *data management* and *data science*, the awareness for data-driven technologies and methods has strongly increased. In particular, *data-intense scientific discovery* is nowadays even considered a new research paradigm (cf. [14]) alongside empirical, theoretical and computational/simulation approaches. But not only scientists come into contact with data regularly, instead data-driven technologies, results of data analyses and the task to store data appropriately can also be recognized in various situations throughout daily life. In recent years, data has become a topic of societal discourse, in particular focused on rather problematic aspects, such as the unauthorized disclosure and analysis of data (as recently done by Facebook and Cambridge Analytica[1]) or the influence of elections with the help of data analyses. Thus, knowing about the possibilities offered by data and data analysis plays an increasing role for developing an understanding of the world. As a result of an ACM workshop, Frank and Walker [11] summarize: *"As data, open, big, personal or in any other guise, becomes increasingly important, power will flow to those who are able to create, control and understand data. Those who cannot, will become powerless. Further, their ability to participate in society will be severely challenged as they lack the tools to engage with an important raw material of society."* Hence, to be able to cope with the new chances and challenges that arise, researchers, practitioners and generally everyone has to acquire some competencies and understand phenomena in the context of data, for instance how large volumes of data can lead to unexpectedly accurate predictions of information not obviously included in a data set. To foster basic knowledge and competencies in this area, for example Wolff and Koertuem [22] discussed simple aspects of data analysis and visualization with seventh and ninth grade students in the context of energy usage. Such competencies are more recently summarized under the term *data literacy*: In higher education, several approaches for teaching data literacy in interdisciplinary approaches and/or from a practical perspective have already been proposed and evaluated, but only few reports set a focus on competencies. Yet, the relevance of data literacy is not restricted to higher education: As a sound understanding of phenomena occurring in everyday life is based on knowledge about data analyses, predictions and their limits, data literacy is also a topic for CS teaching in schools. However, existing work can not be directly transferred to school education without further research, as higher education pursues different goals. In addition, so far no systematic review of competencies in this field was conducted, existing approaches are often merely based on best-practice examples. Hence, a data literacy competency model also suitable for primary

WiPSCE '18, October 4–6, 2018, Potsdam, Germany

ACM ISBN 978-1-4503-6588-8/18/10…$15.00
https://doi.org/10.1145/3265757.3265766

[1] https://www.nytimes.com/2018/04/04/technology/mark-zuckerberg-testify-congress.html

and secondary education cannot be anchored in these existing approaches, but needs to be built up systematically and anchored in school education, keeping didactic aspects in mind.

Thus, in this paper, after having created a theoretical foundation, we describe the development of a theoretically founded competency model of data literacy: In contrast to other work, we set our focus on a general knowledge perspective on the field. After developing the competency model, we discuss the resulting model, characterize it with exemplary competencies and contrast it to other approaches. To give an impression of its practical relevance, finally we conclude by outlining a data-literacy-oriented lesson sequence for secondary CS education.

2 DATA MANAGEMENT AND DATA SCIENCE AS FOUNDATION OF DATA LITERACY

Data literacy can be defined as the *"ability to collect, manage, evaluate, and apply data, in a critical manner"* [16] or, more extensively, as *"the knowledge of what data are, how they are collected, analyzed, visualized and shared, and [...] the understanding of how data are applied for benefit or detriment, within the cultural context of security and privacy."* [3] These are not the only approaches to defining this relatively new topical field, however, all definitions share an incorporation of various aspects related to handling data. From a CS perspective, most of the tasks described by the two definitions mentioned before originate from *data management* and *data science*: Data management, as well as the original field databases, focuses on rather static aspects related to data, in particular on how they are stored and accessed appropriately, while data science sets its focus on the rather dynamic aspects, such as data analysis and visualization. Hence, we assume that both fields give a clear impression of *data literacy* from a CS perspective and are a suitable basis for investigating this field in depth. Thus, central ideas of both fields need to be taken into account when developing a data literacy competency model.

Data management has already been thoroughly investigated from a CS education perspective, in particular its long-lasting key concepts were identified and structured in the *model of key concepts of data management* [13]. The core technologies, practices, design principles and mechanics of data management, as summarized in this model (cf. fig. 1), were derived empirically based on a qualitative content analysis of established textbooks from this field and structured by adopting the model of the Great Principles of Computing [10]. As data management and data literacy exhibit strong overlaps, we assume that this model is suitable for getting first insights into data literacy. Additional insights from a practical perspective are gained by investigating data life cycle models (e. g. fig. 2). Later, by comparing our resulting competency model with existing data literacy competency descriptions, this assumption is evaluated.

The second foundational field of data literacy, *data science*, was investigated in-depth particularly in the EDISON project[2]. As part of this project, a competency framework [9] and a body of knowledge [8] have been developed, along with a model curriculum and a professional framework. Especially the first two documents give important insight into this field: They were created based on the requirements set out, for example, in job advertisements for data analysts, which give a clear impression of what others expect from data scientists, but hardly consider a scientific perspective, which might set different focus points. Hence, as a basis for developing this competency model, in previous work [12] we also conducted a qualitative content analysis with the goal to describe the contents of data science with a focus on the scientific perspective: Documents describing several data science study programs were investigated with the goal to determine the content knowledge expected of the graduates. As this specific analysis is not the focus of this paper, but will function as an important basis for the competency model, key aspects are summarized below:

- From all study programs related to data science in Germany, we selected those which set a clear focus on data science ($N = 11$ in June 2018).
- By analyzing the respective module descriptions of mandatory courses, we identified central contents that every graduate should get to know.
- As a result, four central content areas of data science with various specific contents were identified: *data analysis and machine learning*, *big data*, *data privacy and data ethics* and *data storage*.

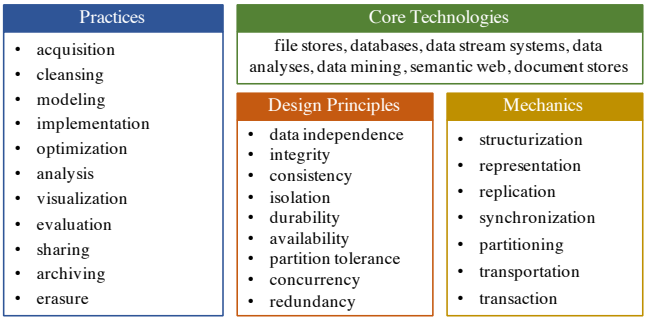

Figure 1: Model of key concepts of data management [13].

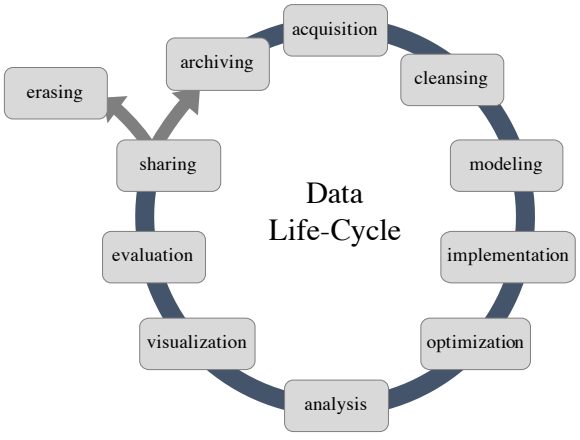

Figure 2: Data life cycle [13].

[2]EDISON is a research project trying to build a data science profession. http://edison-project.eu

- Although we set a focus on German study programs, we can expect a high international validity of our results, as we compared them with three study programs from the United States, which show similar focus points and only differ in details.

The existing work from both fields gives important insight into the respective fields and hence also into data literacy. Hence, following the same assumption as for data management, this work can serve as a basis for investigating data literacy from a CS education perspective.

3 DATA LITERACY IN (CS) EDUCATION

Although data literacy is not limited to higher education, it has hardly been considered from a general (CS) education perspective. For several years, even *data* in general was rarely discussed in CS education research, was instead focused on other aspects of CS. Despite not being in focus, there are indications that data literacy can be considered as a central part of general knowledge. For example, when adapting the concept of computational thinking (cf. [20]) for mathematics and science classrooms, as one of four central aspects Weintrop et al. [19] introduced *data practice*, described as *collecting, creating, manipulating, analyzing and visualizing data*: *"Data lie at the heart of scientific and mathematical pursuits. They serve many purposes, take many forms, and play a variety of roles in the conduct of scientific inquiry."* [19] Despite using a term other than data literacy, their work considers various aspects of data literacy as topics for high school education. Even from a CS education perspective, aspects of data literacy are not completely out-of-scope: For example, the CSTA/ISTE Computational Thinking Teacher Resources [5] involve several aspects of handling and analyzing data that are clearly related to data literacy, e. g. that grade 9 to 12 students should *"develop a survey and collect both qualitative and quantitative data to answer the question: 'Has global warming changed the quality of life?'"* and *"Use appropriate statistical methods that will best test the hypothesis: 'Global warming has not changed the quality of life'."* Another example are the 2017 CSTA K-12 Computer Science Standards [4], which also consider aspects related to data literacy, such as *"Identify and describe patterns in data visualizations, such as charts or graphs, to make predictions."* Yet, these approaches in general miss a technical foundation and do not consider data literacy in a systematic way.

Despite this clear relevance of data literacy for general knowledge, most work on this topic focuses on higher education: Inspired by the vision of Jim Gray [14], *data-intense scientific discovery* (also referenced to as *eScience*) is considered a new research paradigm based on processing and analyzing the immense amounts of observational research data. This new paradigm becomes important in almost every scientific discipline, hence there is a clear need to foster data literacy competencies in higher education. Following this need, in a study on strategies and best-practices for data literacy education, Ridsdale et al. [16] investigated articles about data literacy and related topics, but also gray literature such as reports, white papers and informal literature such as blog posts. Hence, they consider how data literacy is seen from different perspectives and identified 23 competencies (cf. fig. 3) and 64 tasks/skills of data literacy: For example, *"data discovery and collection"*, *"data manipulation"*

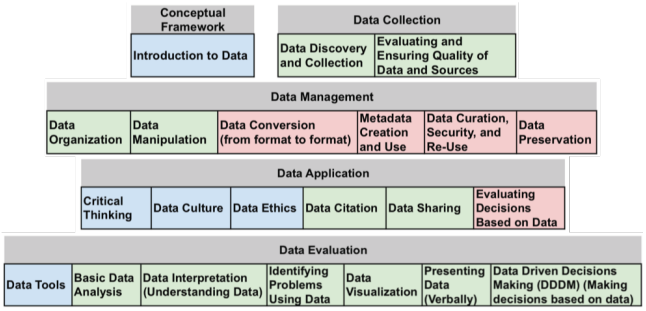

Figure 3: Data literacy competencies determined by Ridsdale et al. [16].

and *"data ethics"* are considered as data literacy competencies and *"identifies useful data"*, *"cleans data"* and *"applies and works with data in an ethical manner"* as exemplary knowledge/tasks. Yet, because of their different focus, directly adopting these results for school teaching is not possible.

4 DEVELOPING A DATA LITERACY COMPETENCY MODEL

The literature review has shown that, despite the relevance of data literacy, there is no competency model or description of data literacy available that is directly suitable for CS education in schools. Instead of adapting an existing approach, we decided to develop a completely new data literacy competency model based on a theoretical approach, keeping this larger target audience in mind throughout the development process. Therefore, we based our work on the assumption that the aforementioned work on data management and data science is a suitable basis for this purpose, which is evaluated afterwards by contrasting our model against the existing data literacy competency model by Ridsdale et al. [16]. In our approach, we set a focus on the scientific perspectives on the underlying fields. Following our basic assumptions, this allows to theoretically found and argue the resulting competencies with high validity.

In accordance with other competency models, in particular the one of the German educational standards for computer science in secondary schools [1] as well as the NCTM principles and standards for school mathematics [15], we decided to divide the model into two parts: *Content areas* reflect the CS content addressed by the competencies, while *process areas* emphasize the practical activities. This separation is promising for data literacy, as it considers two different perspectives on each data literacy competency by relating it to both a process area, which includes practices that reflect how people come into contact with data, how they handle and process them, but also to a content area which considers the theoretical background and the underlying scientific concepts that need to be understood. Due to their strong interconnection, these areas cannot be considered completely separate: For example, the potential process area *data analysis* represents an important practice in this field, as people mostly come into contact with data by reading about data analyses. Yet, for understanding how they work and for assessing their results it is not enough to know how to use software

to conduct data analyses. Instead, several concepts of data management and data science need to be understood, including aspects from the potential content areas *analysis methods*, *data storage*, *visualization* and *data ethics*. In CS lessons, depending on the desired educational goals, the focus can be shifted between content and process areas, but none of them can be left out completely. Hence, representing the competency model with two intertwined types of areas particularly emphasizes the wide variety of links between practically-oriented and content-oriented aspects.

With respect to the different natures of these areas, we first consider them separately: In the next two sections, we address the content areas and later the process areas and describe their origin, emergence and the resulting aspects. Afterwards, both types of areas are merged into a competency model and the resulting model is discussed and contrasted to existing approaches.

4.1 Deriving the Content Areas of Data Literacy

For deriving the content areas of data literacy, finding a basis that gives appropriate insight into the complete field is necessary. For this purpose, we use the aforementioned results of the study on contents of data science, but also refer to the content-oriented aspects of the model of key concepts of data management. Hence, seven aspects that will be used as the basis for deriving the content areas were identified: Coming from data science, there is *data analysis and machine learning*, *big data*, *data privacy and data ethics* and *data storage*; from data management, we get the *core technologies*, *practices* and *mechanisms* (cf. fig. 1). However, on closer examination, it becomes apparent that the aspects described there cannot be used as content areas without further discussion, because some of them have clear overlap (e. g. aspects of *big data* and *data storage* with parts of the *mechanisms* and *design principles*) and/or need to be clarified in detail as they are rather unspecific (such as *big data* in general). Also, some terms, for example *mechanics*, describe data management on a rather conceptual level, while others, such as *core technologies*, reflect a more abstract technological level.

Hence, for deriving the content areas of our data literacy model, these candidates were consolidated keeping in mind the target audience of the model, in particular secondary CS teachers. This led to the following criteria, which should be fulfilled by the final content areas:

The content areas…

- represent sets of strongly related concepts/ideas.
- focus on topics that are specifically relevant to data literacy, not only to CS in general.
- have as little overlap as possible.
- give clear insight into one part of data literacy.
- emphasize a content-related perspective on data literacy.

To find content areas that fulfill these criteria, we had to clarify the subject areas in particular by further characterizing these terms based on their respective definitions, but also by investigating their overlaps and differences. For this purpose, they were first narrowed down to a longer list of more specific topics. This allows for a more detailed insight into these areas while emphasizing their overlaps and similarities. Of course, this list cannot be complete, but only further characterizes the terms, which is appropriate for the targeted goal. This led to the following more detailed topics:

- *data analysis and machine learning:*
  - methods of data analysis, such as classification and clustering
  - predictions based on data
  - learning from data, in particular unsupervised and supervised learning
  - quality of data and analysis results
  - basic ideas of data analysis, such as data vs. information, information entropy, correlations vs. causalities
- *big data:*
  - correlation-based data analysis
  - techniques for managing large amounts of data
  - systems for storing large amounts of data
- *data privacy and data ethics:*
  - data ethics
  - basics of data security and safety
  - personal data
  - data privacy
- *data storage:*
  - systems for storing and managing data
  - function principles of data storage systems
- *core technologies:*
  - systems for storing and managing data
- *mechanics:*
  - function principles of data storage systems
  - representing data on a physical level
- *design principles:*
  - ways for accessing data
  - requirements on data stores and data storage

To eliminate overlaps and to (re-)combine similar aspects into one content area, we reorganized the topics and merged subordinate aspects under (partly new) superordinate terms, which resulted in four content areas described below. The links between the aforementioned topics and the content areas are shown in fig. 4: For example, *predictions based on data* became part of the content area *C3 (data analysis)*, while the basic ideas of data analysis cover aspects that are also related to *C3*, but also to *C1 (data and information)*.

(C1) *Data and information* was introduced as additional content area. It covers basic knowledge, such as the difference between information and data, ways for representing information as data, but also the difference between small and large amounts of data regarding their meaningfulness. Hence, this area contains aspects of the topical areas *data analysis / machine learning*, *big data*, *data storage* and *mechanics*.

(C2) *Data storage and access* is focused on aspects concerning the storage of and access to data and hence concepts that are particularly related to data management, but also considered relevant to data science. In particular, this area contains aspects such as replication or synchronization of data, representation of data on storage media, but also accessing data.

(C3) *Data analysis* is particularly focused on methods, algorithms and principles that are central to analyzing data, making predictions based on those and learning from data. With this focus, this content area is almost identical to the subject

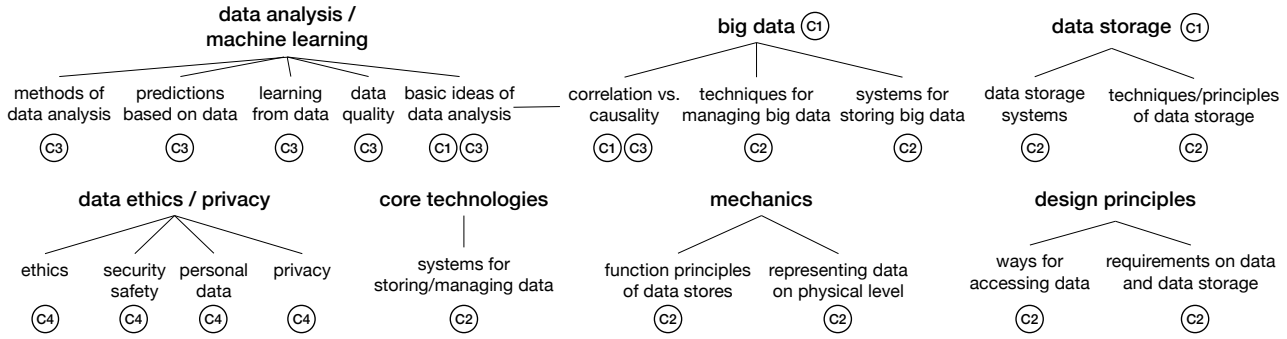

Figure 4: Division of the candidates for content areas into more specific parts and assignment to the final content areas (marked by C1-C4).

area *data analysis / machine learning*. Yet, in the name of the content area, the term *machine learning* was left out: As only aspects from this field are covered by data literacy, including this term could be misleading.

(C4) *Data ethics and protection* is directly derived from the subject area *data ethics and privacy*, yet by changing privacy to protection, the focus is expanded to also include aspects of data security and protecting data in general.

To summarize the above, these content areas represent various aspects of computer science: While the second area is obviously related to data management and the third to data science, their respective concepts are also related to other parts of computer science. For example, when accessing data, computer networks play an important role. But also when discussing the topic *data vs. information*, aspects of information theory and hence theoretical CS such as information entropy gain importance and algorithmic aspects play a central role. Hence, in addition to their relevance to data literacy, these content areas also emphasize strong roots of data literacy in and its links to various aspects of computer science that need to be taken into account in teaching.

4.2 Identifying Process Areas of Data Literacy based on the Data Life Cycle

As mentioned before, a data life cycle model was used to identify the process areas of data literacy. The processes mentioned in such a model represent important steps and tasks in this context and are reflected in data science and data management alike. In addition, they give a clear impression of how people come into contact with data and how they handle them. As a basis for this step, we use the data life cycle model which was developed based on the model of key concepts of data management [13] (cf. fig. 2). Although this model was created from a data management perspective, it is not specific to this field of CS. Instead, it shows high accordance with similar models, which differ in the terms used and in setting different emphases, but share the same meaning (cf. [13]). From this data life cycle, we derived eleven initial candidates for the process areas: *acquisition*, *cleansing*, *modeling*, *implementation*, *optimization*, *analysis*, *visualization*, *evaluation*, *sharing*, *archiving* and *erasure*.

From a scientific perspective, these process areas clearly describe the entire life cycle of data and cover all aspects mentioned in typical definitions of data literacy. Yet, a list of eleven process areas (and several content areas strongly connected to them) raises the legitimate question whether all of these processes are equally relevant for school education. In addition, this large list of terms makes a competency model relatively complex and extensive, and requires detailed knowledge to clearly distinguish the different aspects. Thus, in a next step, our goal was to evaluate these areas from a CS education perspective and in consequence to compress these areas to a more compact and comprehensible list. To achieve this goal, we discussed the list of candidates with teachers and researchers, some with, some without prior knowledge on data science and data management: With these participants, we considered the candidates as guidelines through fictional CS lesson sequences with the overarching goal to convey several aspects of data literacy. Keeping this goal in mind, in two subgroups concepts for lesson sequences with slightly different goals and foci were developed. During this development the participants identified several problems with the candidates for process areas, which were discussed afterwards:

- Some process areas can hardly be considered separately: Especially, *implementation* and *optimization* typically go hand in hand, but also *archiving* and *erasing* cannot be separated at all, as they are clear opposites to each other: a decision to archive data involves deciding against erasing those and vice versa. Also, data *acquistion* and *cleansing* are closely related and typically done at the same time. Hence, during consolidation these areas were merged, as trying to consider them separately raises problems for people using the competency model and is hardly reasonable because of their strong connections.
- The area *modeling* cannot be considered independently from several other areas: Modeling is already an essential task when deciding which aspects of the physical world to capture as data, but also when storing data, for example in a database, when planning data analyses. Hence, we consider two different types of modeling: data modeling and process modeling. While the latter, for example, is an inherent part of data analysis, the first one needs to be considered separately

and has a more specific value from a CS perspective. To emphasize data modeling instead of other types of modeling, we combined *modeling* with *data gathering* and *cleansing*, as data modeling particularly takes place in this part of the data life cycle.

- *Sharing* is a form of handling data which is similar to *archiving* and *erasing* in several aspects: it needs to consider data privacy aspects, methods how to give others access to data, needs to ensure that unauthorized access is prevented and decisions must be made with respect to ethical considerations. As all three processes are also not only applicable for the results of the analysis, but for all data throughout the whole process, merging them into one process area is reasonable.
- Finally, one area was considered missing: *interpreting* data and analysis results. In the prototypical process areas, this aspect was considered an inherent part of analyzing and visualizing. Yet, it is reasonable to emphasize this aspect more strongly, as interpretation is of tremendous importance for handling data and should never be left out. As *interpreting* typically goes hand in hand with *analyzing* and *visualizing* and as these aspects are also strongly related to each other, all three aspects were merged into one process area.

Based on these findings, we were able to condense the list of process areas by combining several areas. Also, with *interpretation*, an additional process area was introduced. As a result, we determined four process areas of data literacy:

(P1) **data gathering, modeling and cleansing**
This process area takes into account the early phases of handling data. It combines three aspects that cannot be considered separately: Data always needs to be structured in a way suitable for storage, access and use. This already provides two modeling aspects: Deciding for the part of the real world that should be captured as data and creating a suitable data model. Also, it is essential to detect and eliminate errors when gathering data and mistakes in the resulting data set as early as possible. With these aspects, this first process area addresses four questions: *Which attributes do I need to capture as data? How can I capture them? How can these data be stored in a way that I can later use them? Are the captured data usable for my purposes?*

(P2) **implementing and optimizing**
The implementation and optimization takes place on different levels: In particular, it includes the implementation of a data model in a suitable data storage system and storing the data in the system. But also in earlier phases, such as data gathering, and in later ones, like data analysis, (simple) algorithms may be implemented for fostering specific tasks. Accordingly, it can not only take place before the analysis, but also as part of it. Also, optimizing can pursue different goals related to improving data gathering, storage and analysis. Hence, the guiding questions of this process area are: *How can I practically realize data gathering, storage and analysis? How can I improve what has been achieved so far?*

(P3) **analyzing, visualizing and interpreting**
For analyzing data, several methods and principles, such as classification or clustering, may be used with the goal to extract new information from them. In addition, visualization is often important, as good visualizations support peoples' understanding, but even the analysis itself might be supported by visual methods. Hence, this area deals with three questions: *Which information can I extract from my data? How can I help people to easily grasp the essential? Which conclusions can I draw from my analysis results?*

(P4) **sharing, archiving and erasing** The last aspects in the data life cycle, sharing, archiving and erasing, are also essential from a data literacy perspective: In particular, sharing and archiving data include ideas such as consolidation, pseudonymization, and anonymization. Structural meta-data is used to find and organize data, while also being relevant for handling data on a daily basis. On the other hand, erasing data marks the end of the data life cycle and raises the challenge of securely deleting data. Along with sharing, it also clarifies the challenge that deleting data completely is typically not possible anymore if it has been shared with others. Hence, in this process area the following questions are raised: *Which data do I want to share with whom? Which data do I want to archive and how? How can I delete data appropriately?*

When considering the links of these process areas to the underlying fields of CS, it becomes evident that the second process area is particularly related to *data management* and its key concepts, while the third instead focuses on aspects of *data science*. In contrast, the first and last process areas are not related to specific areas of CS, but rather frame the others with generally relevant topics concerning handling and processing data. Hence, the process areas of data literacy emphasize both the static and dynamic aspects of data from a CS perspective, but also consider generally relevant topics concerning this field.

4.3 A Prototypical Data Literacy Competency Model

By combining the process and content areas as argued before, we can construct the competency model of data literacy shown in fig. 5.

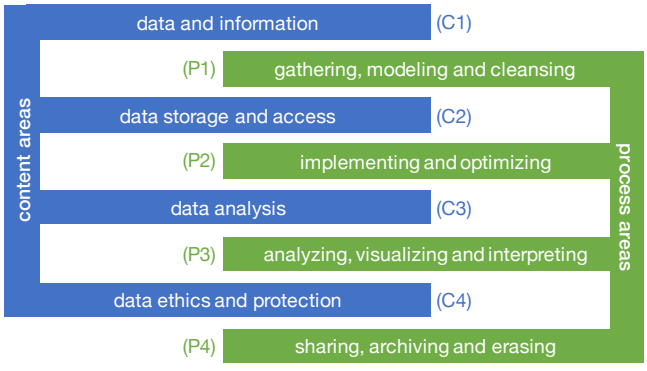

Figure 5: The developed data literacy competency model.

In this model, data literacy is shaped by aspects of data management, data science (including links to machine learning) and data

ethics on the content side and by aspects of handling and processing data on the practical side. This is consistent with popular definitions of data literacy: For example, Ridsdale et al. [16] define data literacy focusing on the practical aspects *"collect, manage, evaluate, and apply data"*, but also emphasize that these practices need to be applied *"in a critical manner"*, which makes a clear technical foundation of the competencies mandatory. Also others, e. g. Deahl [7] or Vahey et al. [17], set similar foci and describe data literacy mainly from a practical perspective, which is particularly covered by our process areas.

By design, the model emphasizes the strong link between content and process areas, as merely considering a single factor is not sufficient to clearly describe data literacy, give insight into this field and to develop appropriate competencies. On the contrary, both areas have to be considered closely intertwined in order to allow students to develop practical competencies that are technically sound by appropriate content knowledge. Hence, considering a process area without connecting it to any content area or vice versa is not intended by the model. While there are several obvious connections, for example between *P3 (analyzing, visualizing and interpreting)* and *C3 (data analysis)*, these are not the only links. Indeed, each process area has connections to all the content areas and vice versa. In table 1, we illustrate these connections by providing exemplary competencies for all combinations of process and content areas. As these competencies have not been evaluated or discussed further, they are not to be considered as a valid or complete list, instead giving an impression of the scope of the competency model.

As the exemplary competencies show, various links between process and content areas are possible. However, neither the model itself nor the abovementioned competencies distinguish different competency levels. Thus, in future work this model needs to be extended from a competency structure model to a competency level model by introducing a third dimension which considers different levels based on further research. However, as the model was developed based on a professional point-of-view on the field, these missing competency levels also suggest that the model is not restricted to school education, but may also fit for other educational levels.

In the presented form, the competency model offers many benefits for both research and practice: It allows to evaluate lessons and sequences regarding the acquisition of basic data literacy competencies. Also, it may be used as the basis for developing lessons and courses with a focus on fostering data literacy competencies and helps to technically substantiate them. In particular, the technical foundation of the model contributes to these possibilities: It was derived in a theoretically-argumentative way from two existing empirical studies, which take into account the two fields data science and data management that form a basis for data literacy. The origin of the developed model is also visualized in fig. 6.

5 COMPARISON WITH OTHER DATA LITERACY (COMPETENCY) MODELS

As mentioned before, the model developed in this work is not the only approach to characterize data literacy or its competencies. Starting from a description of the data inquiry process (*problem – plan – data – analysis – conclusions*), Wolff et al. [21] derived several competencies that were summarized into seven *foundational competencies* of data literacy (cf. fig. 7). These competencies also represent a data life cycle model similar to the one used for the development of our model. However, as our basis considers more details, it also gives better insight into the process areas, and in particular adds and explicates content areas that are not reflected in the model by Wolff et al. Yet, on an argumentative basis, they set a stronger focus on the area *problem – ask questions from data*. In our model, this aspect is not covered explicitly, as it rather provides the reasons for data analysis and does not cover CS related concepts/ideas that are not also part of analysis and evaluation. As we consider data literacy from a CS education perspective, we regard this aspect as being out-of-scope of the model, but without neglecting its importance.

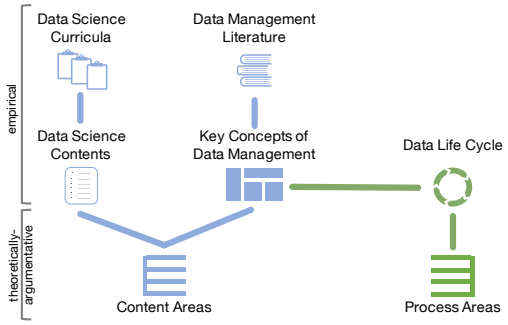

Figure 6: Visualization of the origin of the model.

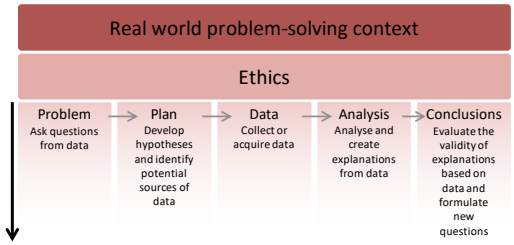

Figure 7: The space of data literacy skills by Wolff et al. [21].

The most popular study on data literacy, which was conducted by Ridsdale et al. [16], also results in a set of data literacy competencies, which are more detailed than the previously mentioned approach. As part of their work, Ridsdale et al. identified five knowledge areas with 22 competencies (cf. table 2). However, they use a different competency term, which in comparison to the commonly used competency definition by Weinert [18], seems to be on a more abstract level: For example, the competency *basic data analysis* mentioned by Ridsdale et al., following the common understanding, should rather be considered a competency area as it includes various different competencies. Yet, this difference does not interfere the comparison of the meaning behind both models: Although they are structured differently and originate from different methodological approaches, both models contain many similar aspects and show a strong overlap, in particular related to the practices and

Table 1: Matrix of exemplary competencies for the different combinations of process (P1–P4) and content areas (C1–C4).

	P1 gathering, modeling and cleansing	**P2 implementing and optimizing**	**P3 analyzing, visualizing and interpreting**	**P4 sharing, archiving and erasing**
C1 data and information	- choose suitable sensors for gathering the desired information as data - structure the gathered data in a suitable way for later analysis - evaluate if the captured data represents the original information correctly	- implement algorithms for gathering the desired data - implement simple algorithms to download data from web APIs - discuss optimizations and limits of data gathering	- combine data to gain new information - emphasize the desired information in visualizations - interpret data and analysis results to get new information	- decide whether to share original data - decide which of the original data to store to keep the required information - decide on an appropriate way to delete specific data
C2 data storage and access	- select a suitable data model - structure the gathered data in a suitable way for storage - visualize data models in a suitable way	- decide on a suitable data storage and store the data - use possibilities for enabling efficient access to data - increase storage efficiency using compression	- access the data in a suitable way for analysis - use suitable data formats for the data to analyze - store their analysis results appropriately	- decide whom to give access to the stored data - determine access rights for the data - discuss issues related to data validity when erasing data
C3 data analysis	- decide whether specific data influences results of analysis - structure data appropriately for analysis - connect data from different sources for analysis purposes	- implement simple analysis algorithms - determine adjustment screws for analysis - optimize data analyses in order to gain higher qualitaty results	- decide for appropriate analysis methods - visualize data and analysis results - interpret the results of analyses	- decide which analysis results to share with whom - reason whether storing the original data is necessary after analyzing them - decide whether it is reasonable to share information about the analysis process
C4 data ethics and protection	- reflect ethical issues when gathering information - decide whether combining different data sources is reasonable in specific contexts - discuss impacts on privacy when continuously capturing data	- discuss how to anonymize or pseudonymize data appropriately - exclude data from permanent storage based on ethical considerations - choose access rights to data based on privacy issues	- discuss the ethical impacts of the conducted data analyses and their results - decide whether analysis results are sufficiently anonymized - reflect whether analyzing specific data raises privacy issues	- reason whether storing data for further uses should be allowed from an ethical perspective - decide on appropriate ways to securely erase original data and analysis results - find ways for appropriately removing attributes that lead to privacy issues

process areas. The model by Ridsdale et al. contains all aspects that are also emphasized in our model, which leads to the assumption that our model does not add invalid aspects. However, Ridsdale et al. add some aspects not explicitly mentioned in our model, for instance *metadata creation and use* or *presenting data (verbally)*. These aspects are not completely out-of-scope for our model, but, with a focus on school education, it merely sets another emphasis and thus does not cover these areas equally: While *metadata* is of course an important topic, it is not considered as being on content area level, instead it is considered in *data and information (C1)*. Verbal presentation meanwhile is not in the focus of our model, as this is a competency which is not specific to data literacy. In general, the five knowledge areas presented by Ridsdale et al. are mostly equivalent to our content areas, yet they broaden the focus of the last content area from *data ethics and protection* (as we call it) to *data application* in general. However, data application is also covered in all other parts of our model, as the practices are oriented on data application in general.

Table 2: Knowledge areas and competencies of data literacy as identified by Ridsdale et al. [16]

knowledge area	competencies
conceptual framework	introduction to data
data management	data organization; data manipulation; data conversion; metadata creation and use; data curation, security, and re-use; data preservation
data evaluation	data tools; basic data analysis; data interpretation (understanding data); identifying problems using data; data visualization; presenting data (verbally); data driven decisions making (DDDM) (making decisions based on data)
data application	critical thinking; data culture; data ethics; data citation; data sharing; evaluating decisions based on data

In summary, the comparison to the two exemplarily chosen models shows that despite the different approach of our work, the results are similar to existing models: All three models generally cover the same parts and show no contradictions with each other. This also supports our fundamental assumption that existing works on data management and data science also give clear insight into data literacy, at least when a CS education perspective on the contents and practices described is taken. Correspondingly, by emphasizing the distinction of process and content areas, our model makes a clear contribution to research in data literacy education, as this model supports educators to keep both the contentual and practical perspectives in mind and to consider them appropriately. As a result of the clearly described and comprehensible approach, it is even possible for them to reconstruct the competencies with respect to a specific target audience. Additionally, this work contributes to theoretically founding discussions on data literacy. Also, the model becomes more understandable as the principles behind the competencies are well-described and as all competencies can be traced back to their origins.

6 USING THE COMPETENCY MODEL FOR PLANNING CS LESSONS

In order to give an example of how to use the developed competency model for CS education, in the following we will outline the development of a lesson sequence based on this model. In this example, the overall lesson goal is to raise students' awareness regarding the analysis of large amounts of data and predictions based on those. For determining more specific goals and targeted competencies, considering the process areas of data literacy was a helpful approach: As we do not want to give students strict "rules" for handling their data, but instead give them insight into the possibilities, limits and threats of this topic, conducting their own data analysis on real data in a context that affects them is a more suitable approach. Hence, at least the process area *P3 (analyzing, visualizing and interpreting)* has to be considered in this lesson sequence, but in order to include insight into the limits of such analyses, process area *P2 (optimizing and implementing)* needs to be taken into account as well. Yet, only giving practical insight without fostering technically sound knowledge is not appropriate to enable the transfer of knowledge to new situations and to allow recognition of general functions of such analyses. Instead, related content areas also need to be considered: In particular *C3 (data analysis)* gives the technical foundation for understanding how data analyses work and how to conduct them. As considering real-world problems is a central aspect of the planned lessons, also *C4 (data ethics and protection)* raises important concerns.

Regarding these four selected areas, we could for example strive for the following competencies in CS lessons:

- explain the function principle of a simple data analysis method
- explain how data can be predicted after learning from existing data
- interpret resulting/predicted data
- optimize a prediction model e. g. by modifying the amount of training data used
- discuss analysis results from an ethical perspective
- discuss the analysis approach and goals in terms of ethical and societal aspects

Although the related topics can become very complex, it is possible to achieve such competency goals in CS education: In a lesson sequence (three lessons of 90 minutes each) that was conducted with ninth grade students (about 15 years old), we were able to achieve these goals. We started with what most students know, an online shop intending to analyze purchases in order to predict what people will buy next. In an unplugged approach, the students were asked to determine rules in a fictitious data set on purchases in an online shop that was generated exactly for this task. In this context, classification trees were introduced and used to visualize the rules and to predict attributes. On this foundation, we started with the central task of our lesson: Based on attributes known about a set of students and the points in two previous examinations, the students in our class had to predict the points achieved by the other students in a third examination. For this purpose, real data were used: Students were given the chance to analyze real anonymized data about more than 600 Portuguese students that are published in the UCI Machine Learning Repository[3]. This data set includes various attributes of the students, their habits and their family situation as well as the points they scored in three examinations. Based on the process the students familiarized themselves with and carried out manually before, this task was performed with software assistance and automated, in order to show the high potential of such analyses and to allow students to adjust their analysis flexibly. For this purpose, we used the tool *Orange*[4] (cf. fig. 8), which enables data analysis without any programming knowledge by using a graphical interface to visualize and model the data flow. Using this tool, the students were able to conduct analyses and results that were fascinating for them: In particular, they were able to predict the third examination grade with a relatively high accuracy and experienced both the power of such analyses, but also their limits, e. g. when trying to optimize their results. With this simple approach, they were even able to reproduce and comprehend parts of a scientific study [2] and were able to understand a central principle of data analyses

[3] https://archive.ics.uci.edu/ml/datasets/student+performance
[4] https://orange.biolab.si

as they occur everywhere today. During classroom obersavtion, it was clearly visible that this task was raising the students' attention and is strongly connected to the students' lives. But they also became thoughtful as two main problems of such predictions became recognizable: For example, students who received bad grades twice or which come from problematic family situations are stigmatized when being graded this way, so that those who actually perform better would be misjudged. However, even students who regularly perform well and seem to have good family conditions might be judged wrong.

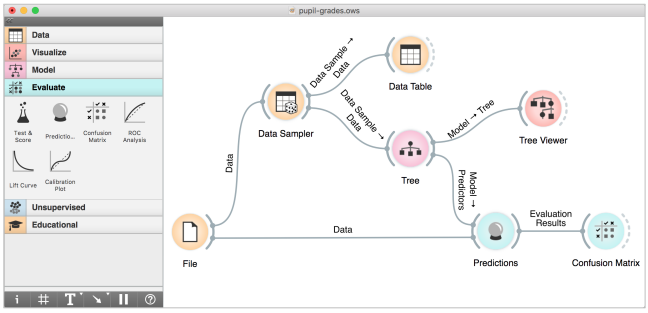

Figure 8: Data analysis and prediction in Orange.

7 CONCLUSIONS

In summary, the data literacy competency model developed and presented in this work describes central aspects and areas of this field and gives insight into the competencies that shape data literacy. It structures them in a way which makes them comprehensible and easy-to-grasp, in particular for educators. However, although we gave some exemplary competencies in table 1, in future work, the competency model needs to be filled with competencies that are further refined and elaborated.

As we have shown, the model developed in this work is in high accordance with already existing models that give insight into data literacy and structure this topic. Yet, our model can be used more flexibly and adapted for specific target audiences, as it allows for gaining detailed insight into its origin. By deriving it from empirical work in a theoretically-argumentative way, it is reasonable to assume a high validity of our results in particular from a scientific perspective on data literacy. Another defining aspect of our model is the separation of content and process areas, which emphasizes the equal relevance of both areas for adequate data literacy education. According to our experiences, this separation makes it easier to fill the model with appropriate competencies, as the combination of process with content areas gives a clear orientation that supports getting insight into this field. With our approach, we take another step towards a technical foundation of data literacy education: At an expert workshop on data literacy organized by the German Informatics Society (results published in [6]), the participants saw the creation of a standardized competency model and standardizing data literacy education as important steps towards improving data literacy education in general. For meeting these demands, the competency model developed in this work is a clear and important step forward.

In the same workshop, it was emphasized that starting in higher education is too late for data literacy education and that instead starting at school seems appropriate. As the outlined teaching example has shown, this is possible and reasonable: Several ideas of data literacy are relevant for general knowledge today and also can be discussed in the classroom without the need to acquire detailed knowledge on, for example, its rather complex mathematical foundations. Regarding everyday life in the digital age, considering data literacy in schools enables students to take advantage of the possibilities present today and to acquire knowledge and competencies necessary for understanding our world.

REFERENCES

[1] Torsten Brinda, Hermann Puhlmann, and Carsten Schulte. 2009. Bridging ICT and CS: Educational Standards for Computer Science in Lower Secondary Education. In *Proceedings of the 14th Annual ACM SIGCSE Conference on Innovation and Technology in Computer Science Education* (ITiCSE '09). ACM, Paris, France.

[2] Paulo Cortez and Alice Silva. 2008. Using Data Mining to Predict Secondary School Student Performance. In *Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008)*.

[3] David Crusoe. 2016. Data Literacy defined pro populo: To read this article, please provide a little information. *Journal of community informatics*, 12, 3.

[4] CSTA. 2017. K-12 Computer Science Standards, Revised 2017. https://drive.google.com/file/d/0B0TlX1G3mywqbXpydGdIVk00Y1U/view. (2017).

[5] CSTA and ISTE. 2011. Computational Thinking Teacher Resources. https://www.csteachers.org/resource/resmgr/472.11CTTeacherResources_2ed.pdf. (2011).

[6] 2018. *Data Literacy und Data Science Education: Digitale Kompetenzen in der Hochschulausbildung*. Gesellschaft für Informatik, Bonn.

[7] Erica Deahl. 2014. Better the Data You Know: Developing Youth Data Literacy in Schools and Informal Learning Environments. (2014).

[8] Yuri Demchenko, Adam Belloum, and Tomasz Wiktorski. EDISON Data Science Framework: Part 2. Data Science Body of Knowledge (DS-BoK) Release 2. (2017).

[9] Yuri Demchenko, Andrea Manieri, and Adam Belloum. EDISON Data Science Framework: Part 1. Data Science Competence Framework (CF-DS) Release 2. (2017).

[10] Peter J. Denning. 2003. Great Principles of Computing. *Commun. acm*, 46, 11, 15–20.

[11] Mark Frank and Johanna Walker. 2016. Some Key Challenges for Data Literacy. *Journal of community informatics*, 12, 3.

[12] Andreas Grillenberger and Ralf Romeike. 2018. Ermittlung der informatischen Inhalte der Data Science durch Analyse von Studienangeboten. In *Hochschuldidaktik der Informatik–HDI 2018*. Gesellschaft für Informatik, Bonn.

[13] Andreas Grillenberger and Ralf Romeike. 2017. Key Concepts of Data Management: An Empirical Approach. In *Proceedings of the 17th Koli Calling International Conference on Computing Education Research*. ACM, New York, 10 pages.

[14] Tony Hey, Stewart Tansley, and Tolle Kristin. 2015. Jim Gray on eScience: A Transformed Scientific Method. In *The Fourth Paradigm: Data-Intense Scientific Discovery*. Microsoft Research.

[15] 2000. *Principles and Standards for School Mathematics*. National Council of Teachers of Mathematics.

[16] Chantel Ridsdale, James Rothwell, Michael Smit, Hossam Ali-Hassan, Michael Bliemel, Dean Irvine, Daniel Kelley, Stan Matwin, and Bradley Wuetherick. 2015. Strategies and Best Practices for Data Literacy Education: Knowledge Synthesis Report. (2015).

[17] Phil Vahey, Louise Yarnall, Charles Patton, Daniel Zalles, and Karen Swan. 2006. Mathematizing middle school: Results from a cross-disciplinary study of data literacy. In *Annual Meeting of the American Educational Research Association*.

[18] Franz Emanuel Weinert. 2001. Vergleichende Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit. In *Leistungsmessungen in Schulen*. Beltz, 17–32.

[19] David Weintrop, Elham Beheshti, Michael Horn, Kai Orton, Kemi Jona, Laura Trouille, and Uri Wilensky. 2016. Defining computational thinking for mathematics and science classrooms. *J sci educ technol*, 25, 1, 127–147.

[20] Jeannette M Wing. 2006. Computational thinking. *Commun. acm*, 49, 3, 33–35.

[21] Annika Wolff, Daniel Gooch, Jose J Cavero Montaner, Umar Rashid, and Gerd Kortuem. 2017. Creating an understanding of data literacy for a data-driven society. *Journal of community informatics*, 12, 3.

[22] Annika Wolff and Gerd Kortuem. 2015. Visualising energy: teaching data literacy in schools. In *Sencity 2*.