EXPLORATION OF LOCATION DATA: REAL DATA IN THE CONTEXT OF INTERACTION WITH A CELLULAR NETWORK

Lukas Höper[1], Susanne Podworny[1], Carsten Schulte[1] and Daniel Frischemeier[2]
[1]Paderborn University, Germany
[2]University of Münster, Germany
lukas.hoeper@uni-paderborn.de

Students are not aware and have little understanding of collecting and processing personal data in their everyday contexts of interaction with data-driven digital artifacts. To be aware of where, how and why data are collected and processed is important to be self-determined. Therefore, we develop and evaluate a teaching sequence to provide reasoning about data as a fundamental aspect of statistical literacy. This teaching sequences deals with the context of interaction with a cellular network where location data are collected. Students get real location data from an unknown person which can be explored with the aim to characterize the person. Students gain different insights by using different basic filters and explain how they achieve these. The results of the exploratory study indicate that students learned to gain insights by exploring given location data and that these insights may describe the person with detailed aspects that may not necessarily be true.

INTRODUCTION

Students at school are certainly active users of data-driven services which often use machine learning methods. But students are often not aware of where, how, and why data about them are collected and processed (Pangrazio & Selwyn, 2019; Tedre et al., 2020). This is problematic because if the students are not aware about these processes about collection and processing of their personal data, they cannot interact with data-driven digital artifacts in a self-determined way and cannot be self-determined participants in the digital data-driven world (Engel, 2017). Therefore, we developed the concept *data awareness* which aims to foster an awareness and understanding of the collection and processing of personal data within an interaction between human and data-driven digital artifacts. The concept combines various contents and skills, including data literacy and statistical literacy and focusses on the interaction between a human (e.g., the students) and a data-driven digital artifact, where personal data are collected and processed. Schulte & Budde (2018) describe such an interaction system between a human and a digital artifact as a hybrid interaction system. In the context of data collection and processing, we are talking specially about data-driven digital artifacts. An example is the cellular network, with which a user interacts in everyday live when calling, texting or using the mobile internet. A lot of data is involved in these interactions between a user and a cellular network, including location data (Wicker, 2012). To enable students to evaluate a cellular network as data-driven digital artifact in this interaction, a perception of possible types of data involved and moreover, of the possible uses and implications of, for example, location data is necessary.

To foster such data awareness, we develop and evaluate a teaching sequence for data science lessons leading grade 8 to 10 for 14- to 16-year-old students, in which this particular context of interaction with a cellular network takes place and an understanding of possible processing steps of real location data are provided. Thereby, we address reasoning about data as a skill of data literacy and statistical literacy to finally provide an understanding of the collection and processing of location data in this interaction context. Thus, we aim to foster students' skills to get insights from real location data, so that the students will be able to generate ideas of possible uses of location data as personal data.

CONCEPT DATA AWARENESS

Data awareness is an educational concept which aims to foster awareness, attention and understanding of where, how, and why personal data arise and are processed during (and probably after) an interaction between a user and a data-driven digital artifact. In such an interaction system – consisting of a human, a data-driven digital artifact and the interaction between these actors – a lot of different data may be produced and collected. To foster data awareness, the students have to be able to recognize the collection and processing of personal data about themselves while interacting with a data-driven digital artifact. This leads to using data-driven digital artifacts in a more self-determined way.

In: R Helenius, E Falck (Eds.), *Statistics Education in the Era of Data Science*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
Aug-Sept 2021, Online conference. iase-web.org/Conference_Proceedings.php

Explicitly and implicitly collected data

The OECD's taxonomy (OECD, 2014) distinguishes between the following four types of data: *provided data* as data which are created aware and actively by the user; *observed data* as data which are gathered by observation and recording and of which the user is not necessarily aware of; *derived data* as data which are generated by processing of existing data; *inferred data* as data which are generated by probability-based analytically processing (OECD, 2014). A similar distinction between different kinds of data is made by Pangrazio & Selwyn (2019). Overall, on the one hand there are data which are collected explicitly when the user providing them while interacting with a data-driven digital artifact. On the other hand, data can be implicitly collected and generated. Therefore, in interactions between a human and data-driven digital artifact, we distinguish between explicitly and implicitly collected data. For example, while using the cellular network there are explicitly collected data like the telephone number of the person to be called and implicitly collected data like the location data which are tracked by the cellular network provider. To be aware of and recognizing implicitly collected data in such interactions certainly requires an understanding of the technology of a data-driven digital artifact which can be described with an analytical perspective of architecture (e.g., Schulte & Budde, 2018).

Primary and secondary purpose of using and processing data

Explicitly or implicitly collected data can be processed by aiming to generate more data or get information about the person, such as generating data models by using machine learning methods or producing a digital self of the person (e.g., Bode & Kristensen, 2015; Tedre et al., 2020). Zuboff (2019) describes in her explanations of surveillance capitalism that providers of data-driven digital artifacts track their users in order to perform profiling and to use the users' experiences for commercial purposes, such as predictions of users' future behavior. Therefore, the collected data are often transferred in a different context which changes the meaning and implications of the data and thus also of the data-driven digital artifact – this can be described as the interpretive perspective of relevance (Schulte & Budde, 2018) which is necessary for the evaluation of the data-driven digital artifact.

The purposes of using and processing data can essentially be divided into two areas: 1) data are used or processed to operate features of the data-driven digital artifact (primary purpose) or 2) to investigate developments of the data-driven digital artifact or to achieve additional (e.g., commercial) purposes (secondary purpose). In the example of the cellular network, this involves processing the location data to establish a connection between mobile phones, or more precisely, to operate the cellular network efficiently as a primary purpose – otherwise, the whole cellular network would have to be flooded for connection requests. Various secondary purposes of using and processing collected data are imaginable, such as getting general insights of the user. Wicker (2012) describes some insights which stem from location data. In summary, to evaluate collecting and processing of personal data while interacting with a data-driven digital artifact includes a) an understanding of primary purposes of using and processing the collected data by the data-driven digital artifact itself or by the provider, and b) the ability to create concrete ideas for secondary purposes.

Furthermore, the concept aims to empower students to make an informed evaluation of the data-driven digital artifact in a concrete everyday interaction and to decide whether to continue the interaction or to shape the interaction system. Therefore, students have to be able to recognize which data are collected and for which purposes they might be processed, including developing ideas for secondary purposes.

Reasoning about data and data moves

In order to foster data awareness, it is necessary to (basically) understand the architecture of data-driven digital artifacts, including data processing procedures: How can explicitly and implicitly collected data be used to gain insights about a user for primary and secondary purposes? The data is usually derived from a specific interaction context and has been collected explicitly or implicitly, but it is also considered in other contexts, especially for secondary purposes. For example, location data could be used to establish a connection between mobile phones, and also to investigate the movement behavior of users. This raises the question of what happens when this data is processed with further relevant knowledge for other purposes. Therefore, reasoning about data is an important skill of data literacy und statistical literacy (Biehler et al, 2018; Garfield & Ben-Zvi, 2008). In order to provide a basic understanding of data processing steps that can be used to gain insights from data, Erickson et al. (2018)

suggest a set of data moves. The authors recommend including them in data science education to make something that seems like data science: filtering, grouping, summarizing, calculating, merging, reorganizing. For example, by filtering, some insights about a person can quite easily be gained from location data from that person.

RESEARCH QUESTION

The developed teaching sequence takes the context of interacting with the cellular network and presents an example of the students' everyday life. While running the cellular network, location data are implicitly collected the whole time. In order to develop ideas about for which additional purposes location data could be used for, the students learn what location data are and what insights can be gained from them. Thereby they should be enabled to develop ideas about further processing purposes and apply filtering as a data move. Following the methodology of conjecture mapping (Sandoval, 2014), we investigate the design conjecture of whether the students are able to gain insights when exploring location data. When this occurs, they can achieve various aims of the concept data awareness. Therefore, in this paper we focus on the gained insights during the exploration phase. When students can gain concrete insights by exploring the location data which are visualized on an interactive map, they may develop an understanding of data processing steps. We investigate how students explore location data to determine whether exploration of location data in the setting of this teaching sequence can serve to foster an understanding of what insights can be gained from location data to help students develop ideas of secondary purposes of using and processing location data. This would enable students to evaluate the collection and processing of location data in everyday contexts, aiming for more self-determined interactions with data-driven digital artifacts that collect and process location data. For the purpose of this research interest, following research question is raised: *What insights do students gain from exploring real location data after filtering and visualizing them on an interactive map?*

METHODOLOGY

Teaching Sequence

We have designed and realized a four-lesson teaching sequence for grade 8 to 10 in the context of interacting with the cellular network. While such interactions between a user and cellular network, a lot of data are collected, for example, location data. The students explore given location data of a real person which are generated by using the cellular network and are published by a German newspaper (https://www.zeit.de/digital/datenschutz/2011-03/data-protection-malte-spitz). This dataset contains location data from August 2009 to February 2010.

First of all, the developed and evaluated teaching sequence is described by the key steps of the lessons. At the beginning, the functionality of the cellular network is introduced by explaining the question: How does a call between two mobile phones occur? In this first step, the students explore how a cellular network is structured and how it works. In doing so, students get a basic understanding about the functioning of a cellular network. Afterwards, the students are asked to answer the questions what data are collected by using a cellular network, why they are collected and for what purpose they are used. This involves that location data are necessary for running the cellular network, specifically for efficient routing of data packages when establishing a mobile connection. Without the collection and use of the location data the whole cellular network has to be flooded and searched to find the target. Through these first steps of the teaching sequence, students gain insights into the architecture and relevance of the data-driven digital artifact that is the mobile network. After that, the students discuss in class what location data are and in which other contexts they might be collected. To get an understanding of processing location data by using basic data moves and thus to get a perception of processing data in secondary purpose, students are then asked to explore a set of given location data. Therefore, the students get a prepared Jupyter Notebook with which they characterize the person they do not know by creating a profile about him/her. The Jupyter Notebook uses a self-developed package with prepared methods which can be used easily by the students and require almost no previous knowledge in programming with Python. For this, we prepared code for the exploration by filtering the data (filtering by range of time, a day of the week, a day of the month or a month) and visualizing them on an interactive map. In the Jupyter Notebook the filter-methods are described and illustrated by short examples. After that, in the prepared Jupyter Notebook contains a full worked example (e.g., Atkinson et al., 2000) of Python code cell and a markdown cell with a description and interpretation of the

potential home of the unknown person. Furthermore, the Jupyter Notebook provides a process for the exploration of the location data: 1) formulating a question, 2) setting the filters and visualizing the location data on an interactive map, 3) exploring the visualized location data, and 4) describing and interpreting the data and answering the previous question. Thereby, students can make and check various assumptions. The insights interpreted by the students about the person are then discussed and collected in class. Different interpretations may occur, such as different possible jobs. This is followed by a discussion of the findings in the whole class. Thereby, the students can reflect the teaching sequence on an individual and a societal level and evaluate the hybrid interaction system consisting of a person, the cellular network and the interaction between these actors.

Data Collection and Analysis

The teaching sequence was implemented in a 10th grade compulsory elective computer science class of secondary school in Germany within a total of 20 students – 3 female and 17 male students. The students had no prior programming experiences with Python or any other programming language and no experiences with Jupyter Notebooks. Because of the pandemic situation, the lessons were done online from students' homes. A few students worked in pairs on the same Jupyter Notebook, so that we collected 15 of the students' Jupyter Notebooks to investigate the research question. They got the task to explore the location data in the prepared Jupyter Notebook by filtering and examining the location data to get as many insights about the person as possible. By using the provided Jupyter Notebook, the students are supposed to explore about three more ideas. Therefore, they should write down the Python code for filtering in a code cell, describe the code and note the interpretation with the insight as a result of the exploration of the filtered location data within a text cell. We call a pair of these semantically connected cells (one code cell and one text cell) a 'exploration block' and use these blocks as analytical unit for the evaluation. These blocks with Python code and textual descriptions with the noted and justified insights are evaluated qualitatively to investigate the research question. Therefore, we used qualitative content analysis in an inductive way (Mayring, 2015) with which we analyze the exploration blocks in the students' Jupyter Notebooks. Different researchers from the authors of this article made an interpretation of the exploration blocks by coding them in an inductive way. Coding results were discussed until consensus has been reached.

RESULTS

The 15 students' Jupyter Notebooks contain on average 2.7 exploration blocks consisting of a code cell and a text cell. Overall, there are 40 complete or partially complete exploration blocks. 35 of these have correct (executable) Python code in the code cell. Of all the exploration blocks, 33 contain an interpretation with a formulated insight as a result of exploring the location data. In total, there are 31 complete exploration blocks which are structured like the worked example in the prepared Jupyter Notebook – executable code cell, description of the filtered location data and interpretation with a resulting insight.

Table 1 lists the categories with frequencies as a result of coding the 33 interpretations. Please note that several times an interpretation (in one cell) was allocated to several categories because of its length.

Table 1. Categories as result of coding the interpretations in students' Jupyter Notebooks

Categories	**frequencies**
Person is employed	7
Place where the person works	7
Profession of the person	2
Home of the person	5
Used mode of transport to work	3

Categories (continued)	**frequencies**
Home of family or friends	5
Vacation resort	3
Travel (for leisure or work)	2
Leisure activities	3
Family status	2

In total, students drew conclusions about the unknown person's work, about the home, about transport mode, about the home of family or friends and the other categories mentioned in table 1. For example, one student wrote "*The person probably works at this time [filtered from 11:00 to 12:00]. The*

person probably works in Berlin near the Federal Ministry for Economic Affairs and Energy, but also travels frequently." to infer the unknown person's working place. This was a typical interpretation in the students' Jupyter Notebooks.

Of the 35 executable code cells, multiple filters were applied five times in succession within one code cell. In total, 6 code cells contained filtering by month, 7 contained filtering by a day of the week, 2 contained filtering by a day of the month, and 25 contained filtering by range of time like in the worked example in the prepared Jupyter Notebook.

FINDINGS

Most of the students used the worked example as a guide for their explorations which can be seen on the one hand in the formulation of the exploration cells, which were phrased in a similar way as in the worked example, and on the other hand in the most frequently used filtering by time period, which was the only one used in the worked example. Few students applied multiple filters successfully within one code cell which possibly indicates a higher level of understanding of data processing using filtering as a data move (knowledge about the architecture).

The results in table 1 show that most students were able to gain some insights about the unknown person, with essentially no prior knowledge of Python or processing of data. They did this by applying a few simple filters to the given location data. Only a few students applied more than one filter in one code cell to get an insight, but most interpretations referred to the exploration of the location data to which only one filter was applied. Thus, students could certainly gain insights about the person with a few simple filters as a data move. Compared to the list of possible insights from the exploration of location data by Wicker (2012), the insights of the students (see table 1) can also be found there, but on the whole, they are less detailed and even remain rather superficial. The students did not make any findings, such as which doctor the person visits how often, or to which religion the person might belong (Wicker, 2012). Students may have underestimated the possibilities of what insights can be gained by exploring location data.

The students drew some conclusions that are quite contradictory. For example, some students mentioned that the person is a bus driver and others that the person works at a school. As a result, students learned that the insights are subjective and also not always accurate or correct, which was also brought up and discussed in class after the exploration when discussing the insights. This goes along with the interpretive perspectives of relevance on data, which are of a more subjective nature and thus very depending on the individual relevance knowledge of a student. While discussing the findings in class, it became apparent that the students were confident in their gained insights and tended to overestimate the truthfulness of their interpretations. This made the students aware that (automated) processing of their data could also be misinterpreted, which might have negative consequences.

DISCUSSION AND CONCLUSION

The teaching sequence shows how an interaction context taken from learners' everyday lives can be used to foster an awareness, as an example, of where, how and why personal data are collected and processed. In this context of interaction between users and the cellular network, data are collected and processed explicitly as well as implicitly. The students have learned that especially location data are processed in this context for a primary purpose – to establish a connection between two mobile phones. To evaluate the collection and processing of location data in this and similar contexts, it is also relevant to consider what secondary purposes this data could be used for. Therefore, the lessons show an example where the collected location data is explored in a further context – the characterization of the person. By exploring the location data, the students were able to gain various insights about the person using filtering methods as a basic data move, whereas in comparison to Wicker (2012), more comprehensive insights could have been gathered. Altogether, the findings indicate that the students learned steps about processing location data and using them to gain various insights about the person. Students showed an overestimation of the truth value of their insights but underestimated the possibilities of gaining detailed and concrete insights.

Statistical literacy should be initiated in school for example by exploring real data of real everyday interaction contexts, to foster a data awareness, thus the students develop an understanding and awareness of the collection and processing of personal data during interaction with data-driven digital artifacts. In order to address other everyday contexts within data processing that is often

automated, machine learning methods could also be used to foster a basic understanding of such automated processing of personal data, like competencies of AI literacy (e.g., Long & Magerko, 2020). In the context of this teaching sequence, the processing of location data with machine learning methods could be addressed in further lessons, for example, to operate personalized location-based services or to investigate the movement behavior of people. The teaching sequence was implemented in a computer science class, but could also be used in statistics education, possibly emphasizing the statistical method of filtering. In further work, we will revise the teaching unit with the new findings and continue to develop the concept data awareness. Therefore, we will investigate further conjectures oriented by the method conjecture mapping for a design-based research process. Also, we develop and evaluate different teaching sequences for various grades in secondary school education.

REFERENCES

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from Examples: Instructional Principles from the Worked Examples Research. *Review of Educational Research, 70*(2), 181–214. https://doi.org/10.3102/00346543070002181

Biehler, R., Frischemeier, D., Reading, C., & Shaughnessy, J. M. (2018). Reasoning About Data. In D. Ben-Zvi, K. Makar, & J. Garfield (Eds.), *International Handbook of Research in Statistics Education* (pp. 139–192). Springer International Publishing. https://doi.org/10.1007/978-3-319-66195-7_5

Bode, M., & Kristensen, D. (2015). The digital doppelgänger within. A study on self-tracking and the quantified self-movement. In R. Canniford & D. Bajde (Eds.), *Assembling Consumption. Researching actors, networks and markets* (pp. 119–134). Routledge. https://doi.org/10.4324/9781315743608

Engel, J. (2017). Statistical Literacy for acitve citizenship: A call for Data Science Education. *Statistics Education Research Journal, 16*(1), 44–49.

Erickson, T., Finzer, B., Reichsman, F., & Wilkerson, M. (2018). *Data Moves: One Key to Data Science at the School Level. Looking Back, Looking Forward.* Tenth International Conference on Teaching Statistics (ICOTS10, 2018), Kyoto, Japan.

Garfield, J. B., & Ben-Zvi, D. (2008). *Developing Students' Statistical Reasoning.* Springer Netherlands. https://doi.org/10.1007/978-1-4020-8383-9

Long, D., & Magerko, B. (2020). What is AI Literacy? Competencies and Design Considerations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10/ghbz2q

Mayring, P. (2015). Qualitative Content Analysis: Theoretical Background and Procedures. In A. Bikner-Ahsbahs, C. Knipping, & N. Presmeg (Eds.), *Approaches to Qualitative Research in Mathematics Education* (pp. 365–380). Springer Netherlands. https://doi.org/10.1007/978-94-017-9181-6_13

OECD. (2014). *Summary of the OECD Privacy Expert Roundtable, "Protecting Privacy in a Data-driven Economy: Taking Stock of Current Thinking*". DSTI/ICCP/- REG(2014)3.

Pangrazio, L., & Selwyn, N. (2019). 'Personal data literacies': A critical literacies approach to enhancing understandings of personal digital data. *New Media & Society, 21*(2), 419–437. https://doi.org/10.1177/1461444818799523

Sandoval, W. (2014). Conjecture Mapping: An Approach to Systematic Educational Design Research. *Journal of the Learning Sciences, 23*(1), 18–36. https://doi.org/10.1080/10508406.2013.778204

Schulte, C., & Budde, L. (2018). A Framework for Computing Education: Hybrid Interaction System: The need for a bigger picture in computing education. *18th Koli Calling International Conference on Computing Education Research* (Koli Calling '18), 18, 10.

Tedre, M., Vartiainen, H., Kahila, J., Toivonen, T., Jormanainen, I., & Valtonen, T. (2020). Machine Learning Introduces New Perspectives to Data Agency in K—12 Computing Education. *2020 IEEE Frontiers in Education Conference* (FIE), 1–8. https://doi.org/10.1109/fie44824.2020.9274138

Wicker, S. B. (2012). The loss of location privacy in the cellular age. *Communications of the ACM, 55*(8), 60–68. https://doi.org/10.1145/2240236.2240255

Zuboff, S. (2019*). The age of surveillance capitalism: The fight for a human future at the new frontier of power* (First edition). PublicAffairs.