seminar.net
MEDIA, TECHNOLOGY & LIFELONG LEARNING
www.seminar.net

ISSN: 1504-4831 Vol 17, No 2 (2021) https://doi.org/10.7577/seminar.4397

Critical Data Studies and Data Science in Higher Education: An interdisciplinary and explorative approach towards a critical data literacy

Dan Verständig
Otto von Guericke University Magdeburg
Email: dan.verstaendig@ovgu.de

Abstract

This paper discusses an explorative approach on strengthening critical data literacy using data science methods and a theoretical framing intersecting educational science and media theory. The goal is to path a way from data-driven to data-discursive perspectives on data and datafication in higher education. Therefore, the paper focuses on a case study: a higher education course project in 2019 and 2020 on education and data science, based on problem-based learning. The paper closes with a discussion of challenges in strengthening data literacy in higher education, offering insights into data practices and the pitfalls of working with and reflecting on digital data.

Keywords: critical data literacy, data science, media education

Introduction

Algorithms, Big Data and Artificial Intelligence (AI) are the remarkable developments of recent years and have not only found their way into society as conceptual terms, they shape everyday actions to different degrees, and this has long since gone beyond basic recommendation systems for better online shopping experiences (Kitchin, 2021). Spam filters screen out unwelcome mails from our inboxes, voice assistance systems set appointments and remind us of open tasks, automated predictions in road traffic keep us free from traffic jam and facial recognition software in public video surveillance systems records patterns, identifies people, and highlights unusual or suspicious behavior (Lyon 2018).

All of the mentioned examples have in common that they rely on the basic principle of computation and pattern recognition. However, the idea of processing even big data sets is not new; also, the concept of AI is not an innovation of recent years. There is a remarkable historicizing on the efforts of AI (Buchanan, 2005). However, recent developments like the increase of computing power combined with networked technologies have unleashed a new quality of gathering, processing and working with Big Data. Additionally, new data practices have emerged and they are touching political, social and economic areas and shaping what is called 'the digital age'. There is a strong technological trend of turning almost every aspect of our daily life into data, be it social media activities, customer relation management, banking and health insurance or even smart cities. The digital age is characterized by *datafication*, as Cukier and Mayer-Schönberger (2013) framed it and as it has been further elaborated and discussed in the light of "data literacy" by Pangrazio and Sefton-Green (2020). In theory, processes of automation should make our lives easier and improve the individual quality of life as well as society in general. In practice, the processes of automation and computation not only increase complexity instead of reducing it, they also reproduce social inequalities and, thereby, punish the poor (Eubanks, 2017).

Why does all of this matter for education? On the one hand, it matters because digital technologies are changing the ways in which we see the world, think and learn. Therefore, knowledge about data, data practices and how data-driven models influence learning, our life in specific contexts and on a daily basis is an essential requirement in order to understand the digital world we live in. A confident, critical and responsible use of digital technologies is not only key to learning and working, but more generally for participation in society.

On the other hand, educational practice and research are challenged not only to incorporate digital technologies into curricular agendas in order to enhance learning and to address lifeworld problems. Educational research is called upon critical reflection on educational technology (ed-tech) (Selwyn 2014) and digital capitalist ideologies such as technological solutionism (Morozov, 2013). While discussions on the use of ed-tech are not

new, the Covid-19 pandemic and the social distancing that followed have affected education in many ways. In order to keep education running, educational institutions had to quickly adapt to the situation and, therefore, a strong push towards the pragmatic use of ed-tech occured in school and higher education. The situation has become a new market opportunity for commercial services and digital learning platforms (cf. Teräs et al., 2020). As Teräs et al. (2020) point out, choices made under these specific circumstances can potentially echo in the future as "new relations of power and control, new forms of student inequity and inequality, and other unpredictable effects" (ibid., p. 865). While these short-term actions and decisions taken by educational institutions might have long-term consequences (Selwyn, 2020), the pandemic highlights the problems that are being discussed for quite a while now (cf. Selwyn, 2019). These problems are related to the basic concept of education, the reduction of learning as an effective, computable process and the image of a learning person, shaped by arguments of enhancing learning through technological progress and "cruel optimism in ed-tech" (Macgilchrist, 2019). While the rhetoric on the use of digital tools are often human-centered, the methods utilized are not.

According to these outlined problems, the following question remains: How can education and media education contribute to ongoing debates about Critical Data Studies (CDS)? As Pangrazio and Sefton-Green (2020) suggest, data literacy might be the most appropriate educational response to the outlined problems and arising complexity in the digital age (ibid, p. 10). Still, the question remains how to foster data literacy and critical thinking?

This paper discusses an explorative approach on strengthening critical data literacy using data science methods and a theoretical framing intersecting educational science and media theory to focus on the sociotechnical "data assemblages" that make up Big Data (Kitchin and Lauriault, 2014). Following the inspirational book by Richterich (2018) on data ethics and Critical Data Studies, the goal is to path a way from data-driven to data-discursive perspectives of data and datafication in higher education.

The overall aim of this article is to reflect on the challenges of teaching data literacy in higher education. Therefore, the paper focuses on a case study, a higher education course project in 2019 and 2020 on education and data science, based on research-based learning. It closes with a discussion on the challenges on strengthening data literacy in higher education, offering insights into explorative data practices and the pitfalls of working with and reflecting on digital data.

Theoretical Framing: Data, Power and Algorithms

The perspective established in this contribution is based on some pre-assumptions and theoretical considerations that are necessary in order to critically assess digital data and set a specific perspective on the implications of datafication. The following points with respect to the practice-based framework improving CDS and data science established by

Neff et al. (2017) are shaping the approach: 1) data are a form of power; 2) with Big Data context matters even more; 3) data and code are fundamentally linked. These three aspects will be briefly unfolded before focusing on the actual project and findings.

Data and Power

Iliadis and Russo (2016) open up their introductory article on CDS with the statement that data are a form of power, since organizations "own vast quantities of user information and hold lucrative data capital" (ibid., p. 1). The power that private companies invoke with their user-centric business models can be framed as an economically driven act of mass data aggregation and user surveillance, which Zuboff (2019) defines as *surveillance capitalism*. In her theoretical concept towards the end of utopian rhetoric of the deliberating Internet, she describes how global tech companies such as *Google* and *Facebook* persuaded the users to give up their privacy for the sake of convenience and how the market shifted due to that development. The data gathered by these companies has been used not only to predict our behavior but also to influence and modify it. Zuboff (2019) analyzes how this actually has had disastrous consequences for democracy and freedom and whistleblowers such as Christopher Wylie, who released a cache of documents prompting the *Facebook/Cambridge Analytica* data scandal. This reminds us that social media targeting is not only a marketing hype but also an actual practice to influence the political process.

In addition to this, Couldry and Mejias (2019a, 2019b) developed the concept of data colonialism. Following the historically long developed and violently established practice or policy of control by one people or power over other people or areas in order to establish economic dominance, there is a shift of practices towards data and digital capitalism. With the digital quantification of the social fundamental patterns of colonialism still remain not only in the light of dataveillance or surveillance capitalism, but also in their very literal sense of offering a contract without transparency of the consequences. In data colonialism, data is appropriated through a new type of social relation, which they define as data relations. Data colonialism justifies what it does as an advance in scientific knowledge, personalized marketing, or rational management, just as historic colonialism claimed a civilizing mission. Data colonialism is a global phenomenon. However, *data* colonialism might be free of physical violence today, the prediction of behavior, the surveillance practices and the rhetoric by tech-leaders hiding the intention of profit has established a new form of power. A form of power in which citizens are intended to be fully transparent and private companies as well as governments hide their practices. This relation and shift of relational constraints is what Pasquale (2015) describes as a black box society. Transparency plays a vital role in algorithms, data and power structures; it is a topic that has been recently adopted by an increasing number of researchers in different perspectives highlighting data, processes, and stakeholder relations (O'Neil, 2016; Ananny and Crawford, 2016; Gillespie, 2016; Seaver, 2017).

There is still a lot of work to be done, not only for researchers but also for practitioners and

policy makers, in order to challenge bias in data and algorithms. Eubanks (2017), for example, systematically investigates the impacts of automation, data mining, policy algorithms, and predictive risk models on poor and working-class people in America, to elaborate how power applies within the use of algorithms and data. She establishes a rich perspective on how profiling, surveillance and containment relate to exclusion and punishment in the digital age. Instead of serving a better societal structure, digital technologies often worsen inequality and data can't provide what poor people actually need. Instead, they are treated like criminals on trial. For Eubanks (2017), we have forged a "*digital poorhouse* from databases, algorithms, and riskmodels" (ibid., p. 12). Drawing a historical line from county poorhouses of the nineteenth century towards the twentieth century, Eubanks argues, that the digital poorhouse "replaces the sometimes-biased decision-making of frontline social workers with the rational discrimination of high-tech tools" (ibid., p.192). Beside the problems that arise from automation, as described by Eubanks in detail, there is a strong relation between the understanding of power relations and discrimination. According to D'Ignazio and Klein (2020), power can be described as "[t]he current configuration of structural privilege and structural oppression, in which some groups experience unearned advantages [...] and other groups experience systematic disadvantages—because those same systems were not designed by them or with people like them in mind." (ibid., p. 24).

D'Ignazio and Klein (2020) who focus on an intersectional feminist approach to show how unequal power structures in the realm of data are being produced. In their work, they highlight several attempts to overcome these structural inequalities. They explain how, for example, an understanding of emotion can expand our ideas about effective data visualization (cf. p. 73 ff.), and how the concept of invisible labor can expose the significant human efforts required by the automated systems around us (cf. p. 178 ff.). Given these developments in emerging business models and personalized data, it is no great surprise that power relationships are also occurring in learning analytics. Broughan and Prinsloo (2019) note that interactions with data are strongly shaped by the question who is allowed to use and interact with data and analytics and in what roles, when they critically discuss the framing of learners as "data objects" rather than "data owners". Iliadis and Russo (2016) also remark that researchers "invoke data in the name of scientific objectivity" while often ignoring that data with reference to Gitelman (2013) are never "raw" but always "cooked" (ibid.). There is no neutral data, especially when working with digital technologies and doing research with digital data as in data science, a reflection on the social and cultural embedding and the narratives of the data is imperative.

Data are a form of power and the power relations are manifold. Although theoretical concepts and empirical studies reflect a high complexity and a rich discourse, some basic considerations can be condensed in order to utilize this assumption for educational purposes. In order to understand the complexity that comes with digital data, it might be rewarding to reflect on some basic questions, such as the characteristics of the data: Where

has the data been collected under what circumstances, and what was the intention? In addition, it might be important to consider by whom the data has been collected and how updating the data works. Considering the concept of data colonialism, it is also important to think about different stakeholders, therefore understanding who the users are and how they have been involved in the process of data collection or processing. This can become an important aspect considering the (lack of) transparency of an algorithmic system and the data involved.

Data and Context

As already established in the section on data and power context matters. Considering an educational setting, data and context can be understood in at least two ways. One the one hand, data are always embedded into social, cultural and historical contexts. Examining the contexts of data can be linked to critical thinking and reflection on the circumstances under which the data has been collected. On the other hand, context can be added, reconfigured or adjusted. This usually can be seen as a process of adding related information to data in order to work with it, give it a specific meaning or notion or even make use of it. This is not a trivial issue, since patterns and correlations often stand out against a background of context. Making sense of data is always framed in one way or another. While framing is necessary in order to find answers for specific questions, it might also be limiting, since one is probably leaving things out.

Context matters, but with occurrence of Big Data context matters even more. One of the specific characteristics of Big Data is that the data processing can be pointing to previously unrecognized connections. Big Data is not just more data, it describes a completely new quality that affects data assemblages, data collection analytics and challenges knowledge in particular. The latter can be observed in positions on new empiricism, as Kitchin (2014) critically assesses when discussing new epistemologies and paradigm shifts. However, data driven science must not be an end of theory, it rather requires even more awareness on the range and limits of the tools used. There is no objectivity created by machines. As Kitchin (2014) notes, “a piece of writing is not simply an order of letters and words; it is contextual and conveys meaning and has qualities that are ineffable” (ibid., p. 8). Therefore, the human factor of deciphering meaning or context is most important. However, there are some challenges when following this approach, since Big Data requires automated processing and analyzing of data. Additionally, the human factor does potentially not come without social determinism or human bias, as Eubanks (2017) showed in-depth. However, this depends on the context as well; it might be helpful to think of data processing in analogy to critical reading and writing at this point. In addition, establishing such a narrative can guide through a learning process on Big Data principles. In their paper on Big Data as a socio-technical phenomenon, boyd and Crawford (2012) argue that it is necessary to critically interrogate assumptions and biases concerning Big Data. Therefore, they formulate and discuss a set of critical questions regarding the implications of data

analytics and social formations. Retaining context remains a crucial factor and an ongoing challenge in the light of Big Data (cf. boyd and Crawford 2012, p. 671). In addition, boyd and Crawford (2012) highlight that there are questions of skills: "Wrangling APIs, scraping, and analyzing big swathes of data is a skill set generally restricted to those with a computational background" (ibid., p. 674). As already outlined with D'Ignazio and Klein (2020), it matters which persons are positioned in which settings and how to think of data science in the light of diversity.

With the rise of digital methods, it is possible to scrape data from many places and despite ethical concerns of scraping publicly available data. As boyd and Crawford (2012) note, that ethical questions need to be asked critically:

> Should someone be included as a part of a large aggregate of data? What if someone's 'public' blog post is taken out of context and analyzed in a way that the author never imagined? What does it mean for someone to be spotlighted or to be analyzed without knowing it? Who is responsible for making certain that individuals and communities are not hurt by the research process? What does informed consent look like? (ibid, p. 672)

As D'Ignazio and Klein (2017) highlight, it is not just on "the stages of data acquisition or data analysis that context matters. Context also comes into play in the framing and communication of results" (ibid., p.164). Therefore, critical data literacy is not only the ability to critically read data, but also to take into consideration how results can be framed. The numbers do not speak for themselves and data are always embedded into contexts. Being aware that certain framing or language in data presentation can be suggestive or misleading is an important aspect in communicating own results clearly.

The scale of data aggregation, the methods involved and the access to data are in power relations and always contextually framed. Therefore, it is important to discuss the implications that arise with a) the use of, b) the work and research with Big Data and c) the presentation of findings. An interdisciplinary approach on the implications briefly outlined here is necessary, since it allows a group of learners to shed light not only onto the mathematical challenges but also the political, social and economic dynamics that affect our lifeworld in various ways.

Data, Code and Algorithms

Digital data, especially Big Data, can barely be discussed without considering the computer code and the algorithms incorporating the mathematical models to process the data sets, visualize results and explore new contexts. Several researchers have highlighted the importance of considering not only data but also algorithms and their various implications on social relations and subjectivity (Manovich, 1999; Striphas, 2010; Gillespie, 2014; Pasquale, 2015; Eubanks, 2017). For Manovich (1999) the world is reduced to two kinds of software objects: data structures and algorithms. They are considered complementary to

each other:

> Algorithms and data structures have a symbiotic relationship. The more complex the data structure of a computer program, the simpler the algorithm needs to be, and vice versa. Together, data structures and algorithms are two halves of the ontology of the world according to a computer. (Manovich, 1999, p. 84)

Manovich examines the notion of database as a cultural form of its own and in the new way, it is structuring our experience of ourselves and the world. The analytical distinction between potentially messy data and algorithms that are processing and in a way cleaning the data can be made. Gillespie (2014) also thinks of algorithms as directly linked to databases in his article on the impact of algorithms on the public discourse:

> Algorithms are inert, meaningless machines until paired with databases on which to function. A sociological inquiry into an algorithm must always grapple with the databases to which it is wedded; failing to do so would be akin to studying what was said at a public protest, while failing to notice that some speakers had been stopped at the park gates. (Gillespie, 2014, p. 169)

Algorithms are designed to remain outside our grasp, and therefore unexpected encounters will remain. As there is no raw data and data sets are not neutral, but rather produced and embedded in various contexts, interdisciplinary perspectives from humanities remind us of the social value that is incorporated, echoed and reproduced in digital technologies. Although algorithms could be understood as abstract mathematical and computational tools, they are embedded into practices and they are produced by humans. Furthermore, data are not only a specific output of an algorithmic selection or process. The produced data can be seen as an input for further development and design of different algorithms. This describes a cycle in which data are output for given systems and models and input for future or other existing models and contexts. It also reveals some of the challenges considering the reflection of biased data. Since, bias in data is not only a technical or automated result, but rather a complex configuration of data input, modeling, processing and output. Data are generated in order to get specific insights, be it user information, shopping experiences or learning effectiveness, and it is the basis to reduce bias and optimize the algorithms we are engaging with. Therefore, the strong relation of data and the computational systems processing data is worth considering for a deeper understanding of the complexity within and the potential possibilities that can encounter one on a daily basis. Since algorithms often operate behind closed doors, the problem of accessing them remains. They might not be black boxes, but they are still opaque. However, most of the complex algorithms in the lived world rely on mathematical and computational models that can somehow be anticipated, be it algorithmic auditing or specific practices to re-contextualize data and playfully engage with the technologies. Even if algorithms may be too complex to explain in detail; even if efforts to elucidate the algorithms might require the use of data that violates regulations, there are tactics to

understand the general algorithmic principles to strengthen data literacy for empowerment in technical and non-technical audiences.

D'Ignazio (2017) advocates for "creative data literacy" and establishes the term offering five tactics to engage with data. They range from working with community centered data or the writing of "data biographies" to the production of learner-centered tools (ibid., p. 8ff.). Besides problem solving and critical thinking, creativity is one of the important transversal skills according to the digital education action plan (2021-2027) of the European Commission. A playful, explorative and creative way to address bias in data and algorithms can reach a wider group and include non-technical audience. For the course project that I will discuss in the next sections, I follow the work of D'Ignazio (2017) and the outlined theoretical and practical considerations on how to engage with data to strengthen creative data literacy.

Course Project: Education and Data Science

There is a variety of framings and understandings of education and learning and the discussion on technologies in education has a long tradition. Yet it holds the risk of emphasizing the technological dimension more than the challenges arising from pedagogical or educational work. Such tendencies and critical counter movements can be observed within the field of learning analytics. The broad availability of educational data has led to an interest in analyzing processes of learning and learning behavior. This often goes with the promise of a better learning experience, effective output measurements and a more personalized learning environment that enables a more student-centered mentoring and teaching.

However, education is a complex process and within the ideology of improvement and enhancement, there is little room for critical negotiation and negation. The latter is essential for critical thinking since it allows distancing from an artefact, situation or topic while reflecting about the experience one is confronted with. For educational settings, this is a crucial point and it requires an atmosphere of trust and openness. Therefore, understanding education as a process of re-configuring and continuously transforming the individual self and world relation, it might be rewarding not only to question circumstances focusing on the processes of one's individual positioning in a pluralized world, with different social norms and values, but also to consider structural transformations and societal challenges such as datafication. Therefore, the understanding of education established in this course project is rather inspired by the idea of the philosophy of education than an output-oriented perspective on learning. Of course, this has pragmatic limitations, since there is a goal for the students to complete the course and therefore tasks have to be accomplished, but considering education as a complex process of learning with a *perspective transformation*, it allows going beyond instructional learning and includes the moment of negation.

Didactical Approach and Structure

The course project covered a time period of one semester in which 18 students from the two study programmes computer science and media education (equally distributed), as well as one lecturer have been involved. The structure of the course combined theoretical considerations on data, media and data representation and practical implications with weekly tasks and exercises, hands on sessions and a joint course project in which all students participated. For the theoretical part, readings such as Eubanks (2017), D'Ignazio (2017), Iliadis and Russo (2016), Kitchin (2014), Gillespie (2014) and Gitelman (2013) have been considered as basic references and partly discussed in depth. The practical part required knowledge in working with big data sets, therefore the series "Making sense of data" by Myatt & Johnson (2009, 2012, 2014) has been introduced. The expected workload was about four to six hours a week considering the readings, programming assignments and the joint course project. This was quite high compared to a regular seminar or course in both study programmes.

The didactical concept, especially of the joint course project, has been established around problem-based and project-based learning (Mettas and Constantinou, 2007) following a project-led education model developed by Powell and Weenk (2003). While the roots of problem-based learning reach back to the 1970s. This method offers a good entry point on data science. Problem-based learning means to organize the curricular content around problem scenarios rather than disciplines. Problem-based learning is intended to promote the acquisition of knowledge that can be used flexibly, the development of transversal competences (e.g. problem solving, creativity, critical thinking) (cf. Duch et al., 2001). This approach also takes into account social competences and the ability to work in a team. Of course, problem-based learning cannot replace traditional teaching and learning methods such as a lecture, but it can complement them. Following Duch et al. (2001) any subject area can be adapted to problem-based learning with relative ease, considering some basic principles such as students motivation, encouraging discussions and reflections, integrating the problem into previous courses and prior knowledge of the students and the right level of complexity.

The course incorporated principles and methods of data science in order to offer an interdisciplinary introduction and to offer an entry point on how to actually work with Big Data under the condition of a research project. Data science is the field that is dedicated towards the research and analyzing of data in several ways. It invokes creativity and individual problem solving and it requires openness towards the data and their social, cultural and political embedding. Problem-based learning can be considered as an explorative way to foster a critical perspective on these phenomena. However, as a generic didactic approach, the framing is crucial and this is something the teaching staff should precisely take care of. This can be achieved by a set of measures, a selection of readings that highlight a critical perspective, by regularly encouraging discussions on the

implications of digital data and data practices and by highlighting potential challenges in applying data science methods on a basic level. Additionally, the choice of digital tools can be considered as an important factor establishing a critical perspective on digital data and data practices.

Choosing the right tools to work with is also an important aspect. We used several tools in order to address the topic of data science in that specific perspective. The university offers an eLearning management tool on the open source platform *Moodle*, enriched with a few add-ons, so writing announcements, providing small tasks and dropping material there served quite well. Additionally, we decided to use a project management tool in order to stay in contact and ensure social sharing. Therefore, we decided to work with *Slack*, which is a proprietary tool.

The course was structured in three parts, the first and introductory part was about theoretical positions of education, data science and societal challenges in general. Here, it was important to introduce not only the topic, but rather to enrich and stimulate discussions so the upcoming challenges and the complexity of the whole field in general could be anticipated. Therefore, assumptions such as the relation of data and power, context and meaning and the role of algorithms as well as algorithmic thinking have been widely discussed for a couple of weeks. In addition to the paper-driven discussions, we had some regular hands-on sessions, where we started working with the *Python* programming language and set up a common working environment in order to prospectively prevent long debugging sessions and error tracking procedures. In order to access data, we worked with *pandas*, which is a fast, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language. Going through the basic functionalities of pandas enabled the students to understand not only how this library works and how to utilize it, but also how Python functions in general. The coding part at this point was not only an enhancement or a variety of the course, it is fundamental for the overall goal. This part was a circular and tentative process in order to identify interesting problems the students want to engage with on a deeper level. Since the coding sessions have been set up as pair programming sessions of two students – a method derived from agile software development – it also served as a team building initiative.

The second part was about basic machine learning principles, mathematical models and their application on data (cf. Schutt and O'Neil, 2013; Finlay, 2014). At this point, we went into hands-on coding, using publicly accessible open data from *Kaggle*, which is a subsidiary of *Google*. Kaggle is an online community of data scientists and machine learning practitioners that allows users to find and publish data sets, explore and build models in a web-based data-science environment. As a community, the platform also enables users to work with other data scientists and machine-learning engineers, and enter competitions to solve data science challenges. The data sets derived from Kaggle served as an entry point to provide guided problem solving. We engaged with the *Titanic data set*,

which is a common and suitable starting point diving into machine learning (cf. Farag and Hassan, 2018). We used machine learning to create different models (Logistic Regression, K-Nearest Neighbor, Naïve Bayes) that predicted which passengers survived the Titanic shipwreck. The prediction and efficiency of the algorithms depend greatly on the predictive model. By solving this problem individually and in the group, we also critically discussed the implications of missing data, interpolation and gender in the group sessions, as well as in the joint Slack team. To some of the computer science students, this problem and the data set was not new, so we could also spend some time on critically reflecting on the overall structure of the data set and comparable tasks.

After these first impressions of how to engage with a data set, the third part of the course was dedicated to the joint group project, in which all of the students have been involved with different roles. The data science project, which has been determined together and discussed since the beginning of the second part of the course, was about winter, Christmas, cultures and globally shared emotions in the time of December 2019 to mid-January 2020. We wanted to visualize and understand at which time people post in social media about topics like #winter and #christmas and what the basic sentiments on these hashtags are. We discussed several assumptions, such as that winter offers a high variety of emotional engagements and that it is geographically depending how emotions are being articulated, we were thinking of a correlation from tweets and weather data. Since businesses on social media also use hashtags, we were discussing whether we could possibly find some patterns on official accounts from internationally recognized brands in various sectors, such as the food industry, service offerings and travel & tourism. Therefore, we decided to work with the microblogging service *Twitter*. The access to data is well structured and technically quite good to handle, since there is a well-documented API.

Although the actual joint group project has been introduced in the beginning of the course, it started only after setting the framing and basic assumptions. The work of step one and two took six weeks and it was necessary in order to align expectations, find a common ground and create an atmosphere of trust and openness within the group. After setting the topic and settling the idea, the joint group project can be divided into four steps:

1) Surveying existing open data sets: we asked the question, whether it is necessary at all to collect data or if there is already a publicly available data set that would benefit the interest of the group. Since we were looking for general usage patterns, an anonymized data set would have been a good choice. We found an open data set that met the requirements; the data set was fully anonymized and still offered some insights on the topic of winter and Christmas. However, the data set was missing geographical information and metadata, so we discussed it in the group and figured out that we were not able to make sense of the data, so we had to skip it. Nevertheless, this process was important, since data preparation and exploration is a vital factor in gaining a basic understanding of the data in general and for data science in particular. Exploring data

means not only to review the general structure of the data, it also serves to identify specific characteristics (features) with statistical analysis methods, for example by determining minima and maxima, the position measures and the data distribution as well as correlation measures. This process is often called exploratory data analysis.

2) Information retrieval and ethical considerations: the next step was to set up an environment for information retrieval and to collect the data. Twitter was not the first choice, but according to the given time frame, tools and the available resources, we found consent on using it. Again, access to data and data sources matters. Access to data in that sense is dependent on the business and their terms of use. This dependency on tech companies and their business models was a crucial point in discussing the ambiguity of getting insights into a topic and relying on others (individuals as well as businesses and their algorithmic selection) to get data at all. At this stage and according to the theoretical framing, it is mandatory to discuss and reflect the ethical implications of collecting digital data. As Boyd and Crawford (2012) note, "just because it is accessible does not make it ethical" (ibid., P. 671). As they elaborate, there is a gap between the individuals and institutions that have access to the data and the social media users that produce the data. Not all of the users active in public spheres would necessarily agree to have their data being re- or decontextualized. Therefore, we critically discussed the technological framework with regard to the ethical framework for publishing Twitter data as established by Williams et al. (2017) already before we designed the collection framework. We then assessed the potential field and user accounts and focused on businesses, users with a high follower count and a few user accounts of public figures, politicians and celebrities. This shaped our initial interest and narrowed down the project focus on a specific topic. The actual data collection was an automated process. This happened not as flawlessly as expected, since we had to reconfigure the automated process as we figured an error in the actual routine of storing the data properly. However, we got a data set to process and analyze.

3) Data analysis and processing: At this stage of the project, the challenge was not only to find patterns with regard to the initially determined aspects and retrieved data, but also to align the data with other factors. For example, when following the idea of high snowfall or unusual cold or warm temperatures in a specific geographic area, only the collected tweets with given geographical information could be considered as reliable data to align with. While inspecting the data the process evolved as looking for a needle in a haystack, but this actually was not the problem. It was the haystack itself. We might have asked the wrong questions and we might have translated our questions in a misleading way into computer code in order to retrieve the data. Nevertheless, at this point, this does not matter at all, for the purpose of critical data literacy and a deeper understanding of Big Data; it was a perfect situation, since it allowed reflection upon the way questions are directed towards digital data. While Big Data can reveal new insights or give answers to questions that have not been articulated, it can also be a hard task to beneficially analyze the data. In the case

of the joint group project, we literally ran out of time. This actually led to the decision not to further pursue the initial idea. It was demotivating for all participants of the course. However, we decided to go ahead with the setting and ended up with a web based app visualizing the tweets popping up around the globe and a basic adjustable user interface, such as a time slider. At this point, the project would have been a fail under real-world conditions. However, experiencing such a process under safe conditions, framed by discussions and reflections on the impact every single decision of this project had, it can still be considered as an exploration of the unexpected.

4) Evaluation and closing: We finished the project with an evaluation over the course of two weeks. The evaluation and conclusive discussion was about the actual project, its implications, learnings, and the course in general. Some of the students further elaborated the discussed topics by writing term papers on the course and the project; others attended hackathons and applied their (pre-existing and newly acquired) skills in other projects.

While we faced many challenges, the explorative joint project holds rich insights pathing a way towards critical data literacy in an interdisciplinary setting for higher education. Some of these insights, such as the importance of exceeding a specific disciplinary domain, teamwork, openness and the role of education will be outlined in the following sections.

Data Science and Interdisciplinary Teams

Data Science requires many skills that range from communication over technical expertise to flexible problem solving (cf. Schutt and O'Neil, 2013). These skills often exceed one domain and therefore it was the intention that students from different study programmes with different social, technological and cultural backgrounds come together in order to explore the work on digital data in a safe and guided educational setting.

The students developed a strong connection to the topic and the team itself. The course started with a basic quiz on data and information. This was a team building measure, since we defined our long-term goals, talked about individual expectations and gave our team a name, based on the answers of the quiz. During the project, the roles of the project members have developed and the students became active learners. Based on interest and skills they also switched roles between the different stages of the project. The pair programming tasks served well in order to strengthen the relationships between the students on an individual level. From a didactical point of view, it was beneficial to bring the students from both study programmes with a different level of expertise together in the pair programming sessions. Some guiding questions that appeared helpful and might inform and inspire similar project/course scenarios could be:

- What is the common ground for a team?
- What are the individual expectations?
- Which tools are necessary in order to work together?

- How to ensure equally distributed participation among the team members?

Errors and Openness

It is important to address the errors that have been made, not only in the sense of changing further parameters and rethinking decisions, but also in how to cope with errors in general. Since Data Science requires a lot of flexibility and experts with a well-equipped set of methods, it is important to address the errors that have been made. In the joint project, we have experienced many errors from simple coding errors and debugging sessions or finding a wrongly spelled character in the Python script, to more complex structural errors such as an inefficient or not working model or the (unintended) process of making data messy when comparing different data sets. Not paying enough attention to the possibilities for a solution could also lead to bad decisions. As the project provoked, relying on the data is not enough. It needs to be taken into consideration that there is always more than one possible answer to one particular question and each of the possibilities incorporates some level of uncertainty. Therefore, practicing tentative and problem-centered approaches is vital in order to become the ability of applying theoretical concepts in practice. This applies for computational skills as well as for pedagogical competencies.

As already outlined, openness is an important factor in order to stimulate the learning setting. This applies to the error management and decision-making as well as to the tools we used. We had a strong focus on using open data as well as open source software and we have discussed the topic in the beginning, as well as in the process of the joint group project. However, against this idea, we used a proprietary tool to organize our project management chats. An alternative to this is a self-hosted installation of an alternative software such as *Mattermost*, which has been installed recently on university servers. Mattermost would be the preferred choice for future courses. The discussions on openness can be summarized with the following guiding questions:

- What access do we have?
- How can we process the data?
- How is use and reuse of the data governed?
- What are the ethical implications of free and publicly available data?
- How to ensure that principles like access and openness can be communicated?

Education and Critical Data Literacy

As introduced in the theoretical pre-assumptions, it matters who is working on which data and how social implications can be addressed properly. Critical data literacy should go beyond being able to have mathematical and numerical skills. Speaking with Tygel and Kirsch (2015) it is “the set of abilities (data reading, data manipulation, data communication and data production) which allows one to use and produce data in a critical way” (ibid., p. 117). The course and especially the joint course project integrated

this methodology in practice and revealed some unexpected findings in a higher education setting.

As has been said before: skills matter. The course was open to students from different masters programmes, so the students were already able to draw on a wide range of methodological skills and knowledge from their individual learning experiences in their respective bachelors programmes. However, can this approach stand against other settings and be realized with non-technical learners or societal groups other than students who will prospectively work in the field? There are two distinctive answers to this question. The short answer would be optimistic and confidently opt for a yes, remarking that the course should not only be offered to students from media education and computer science, but rather for teacher education and social sciences as well. Addressing critical data literacy in educational settings does not only fulfill the idea of raising individual awareness on the ongoing challenges of datafication, but must be understood as a process of multiplication on critical awareness, therefore prospective teachers benefit at least as much as media educational practitioners from such a course. Second, problem-based learning might still be a key to easily adapt to a specific audience or group of participants. However, there are some limitations on scalability. The rather detailed description of the joint course project suggests that critical data literacy is a very complex topic and therefore it might be helpful to narrow the actual workshop down to a specific problem. This is depending on the given timeframe, prior knowledge and the specific goals. While it might be rewarding to discuss a specific data set – as a case study – in-depth with a group of experts, it might be demotivating to address technical issues within a non-technical audience. In order to raise awareness of the implications and the constellations of context and meaning in an educational setting, it might be helpful to rely on research findings that can be bound back to lifeworld problems, rather than discussing the significance of a graph. This means a didactical reduction in favor of explainability. Some of the key considerations on education and critical data literacy can be framed as in the following guiding questions:

- Are there aspects of critical data literacy that should be highlighted in particular?
- How to align an individual research project with institutional conditions?
- How can the educational setting raise awareness for diversity and data practices?

Conclusion

The goal of the course on education and data science was to establish a framework to address critical thinking on digital data with an interdisciplinary group of students of media education and computer science. An explorative project like the one presented here was highly depending on the motivation, engagement and endurance of the students. Hence, the educational setting within a climate of openness and trust has been of high significance. Even if the initial idea could not have been realized, the course offered rich insights into the complexity of getting in touch with Big Data.

Education is more and more shaped by increased worldwide discussions on the importance of teaching digital competencies and, of course, digital technology can support high quality and inclusive education. According to the European Commission and the digital education action plan (2021-2027) digital technology can meet the requirements of future skills such as digital and data literacy, but also creativity and analytical as well as critical thinking. Furthermore, it enables to work in teams and communicate beyond the borders of one's own domain. However, it requires a lot of flexibility to establish a creative and interdisciplinary setting as described within the course project. It requires openness towards the different languages that are being spoken. Higher education meets these requirements and allows a tentative approach on this topic. Nevertheless, efforts such as explorative projects on that topic are limited in their reach. Neither do they replace a comprehensive agenda on critical data literacy, nor do they fully anchor into existing study programmes. At the same time, the joint course project identified the exact need of addressing critical data literacy at the transdisciplinary intersection of computer science and media education. A deeper understanding of the social and cultural implications on data and knowledge of algorithmic processes can seed further engagements. Also, offering courses on data literacy is beneficial for students from computer science, since the skills developed in such projects are additional to what computer science curricula usually include. Therefore, it is even more important to establish a broad perspective on critical data literacy at the intersecting fields of data science and education.

References

Boyd, D. & Crawford, K. (2012). CRITICAL QUESTIONS FOR BIG DATA: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Broughan, C., & Prinsloo, P. (2020). (Re)centring students in learning analytics: In conversation with Paulo Freire. *Assessment & Evaluation in Higher Education*, *45*(4), 617–628. https://doi.org/10.1080/02602938.2019.1679716

Buchanan, B. G. (2005). A (Very) Brief History of Artificial Intelligence. *AI Magazine*, *26*(4), 53–60.

Costanza-Chock, S. (2020). *Design justice: Community-led practices to build the worlds we need*. The MIT Press. https://doi.org/10.7551/mitpress/12255.001.0001

Couldry, N., & Mejias, U. A. (2019a). *The Costs of Connection*. Stanford University Press.

Couldry, N., & Mejias, U. A. (2019b). Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject. *Television & New Media*, *20*(4), 336–349. https://doi.org/10.1177/1527476418796632

Cukier, K., & Mayer-Schönberger, V. (2013). The Rise of Big Data. How It's Changing the Way We Think About the World. Foreign Affairs, May/June.

D'Ignazio, C. (2017). Creative data literacy: Bridging the gap between the data-haves and data-have nots. *Information Design Journal*, *23*(1), 6–18. https://doi.org/10.1075/idj.23.1.03dig

D'Ignazio, C., & Klein, L. F. (2020). *Data feminism*. The MIT Press. https://doi.org/10.7551/mitpress/11805.001.0001

Duch, B. J., Groh, S. E., & Allen, D. E. (Eds.). (2001). *The power of problem-based learning: A practical "how to" for teaching undergraduate courses in any discipline* (1st ed). Stylus Pub.

Eubanks, V. (2017). *Automating inequality: How high-tech tools profile, police, and punish the poor* (First Edition). St. Martin's Press.

Farag, N., & Hassan, G. (2018). Predicting the Survivors of the Titanic Kaggle, Machine Learning From Disaster. *Proceedings of the 7th International Conference on Software and Information Engineering - ICSIE '18*, 32–37. https://doi.org/10.1145/3220267.3220282

Finlay, S. (2014). *Predictive Analytics, Data Mining and Big Data*. Palgrave Macmillan UK. https://doi.org/10.1057/9781137379283

Gillespie, T. (2014). The Relevance of Algorithms. In T. Gillespie, P. J. Boczkowski, & K. A. Foot (Eds.), *Media Technologies* (pp. 167–194). The MIT Press. https://doi.org/10.7551/mitpress/9780262525374.003.0009

Gitelman, L. (Ed.). (2013). *"Raw data" is an oxymoron*. The MIT Press. https://doi.org/10.7551/mitpress/9302.001.0001

Iliadis, A., & Russo, F. (2016). Critical data studies: An introduction. *Big Data & Society*, *3*(2), 205395171667423. https://doi.org/10.1177/2053951716674238

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 205395171452848. https://doi.org/10.1177/2053951714528481

Kitchin, R. (2021). *Data lives: How data are made and shape our world*.

Kitchin, R., & Lauriault, T. P. (2014). *Towards critical data studies: Charting and unpacking data assemblages and their work* (The Programmable City Working Paper 2 No. 2; p. 19).

Lyon, D. (2018). *The culture of surveillance: Watching as a way of life* (First published, reprinted). Polity.

Macgilchrist, F. (2019). Cruel optimism in edtech: When the digital data practices of educational technology providers inadvertently hinder educational equity. *Learning, Media and Technology*, *44*(1), 77–86. https://doi.org/10.1080/17439884.2018.1556217

Manovich, L. (1999). Database as Symbolic Form. *Convergence: The International Journal of Research into New Media Technologies*, *5*(2), 80–99. https://doi.org/10.1177/135485659900500206

Mettas, A. C., & Constantinou, C. C. (2007). The Technology Fair: A project-based learning approach for enhancing problem solving skills and interest in design and technology education. International Journal of Technology and Design Education, 18(1), 79–100. https://doi.org/10.1007/s10798-006-9011-3

Morozov, E. (2013). *To save everything, click here: The folly of technological solutionism* (First edition). PublicAffairs.

Myatt, G. J., & Johnson, W. P. (2009). *Making sense of data II: A practical guide to data visualization, advanced data mining methods, and applications*. Wiley.

Myatt, G. J., & Johnson, W. P. (2012). *Making sense of data III: A practical guide to designing interactive data visualizations*. Wiley.

Myatt, G. J., & Johnson, W. P. (2014). *Making sense of data I: A practical guide to exploratory data analysis and data mining* (Second edition). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118422007

Neff, G., Tanweer, A., Fiore-Gartland, B., & Osburn, L. (2017). Critique and Contribute: A Practice-Based Framework for Improving Critical Data Studies and Data Science. *Big Data*, *5*(2), 85–97. https://doi.org/10.1089/big.2016.0050

Pangrazio, L., & Sefton-Green, J. (2020). The social utility of 'data literacy.' *Learning, Media and Technology*, *45*(2), 208–220. https://doi.org/10.1080/17439884.2020.1707223

Powell, P. C., & Weenk, W. (2003). Project-led engineering education. Lemma.

Richterich, A. (2018). The Big Data Agenda: Data Ethics and Critical Data Studies. London: University of Westminster Press. DOI: https://doi.org/10.16997/book14

Schutt, R., & O'Neil, C. (2013). *Doing data science* (First edition). O'Reilly Media.

Selwyn, N. (2014). *Distrusting educational technology: Critical questions for changing times*. Routledge, Taylor & Francis Group.

Selwyn, N. (2019). What's the Problem with Learning Analytics? *Journal of Learning Analytics*, *6*(3). https://doi.org/10.18608/jla.2019.63.3

Selwyn. N. (2020). After COVID-19: The longer-term impacts of the coronavirus crisis on education. Melbourne: Monash University. https://educationfutures.monash.edu/all%2D%2D-present/after-covid-19

Striphas, T. (2015). Algorithmic culture. *European Journal of Cultural Studies*, *18*(4–5), 395–412. https://doi.org/10.1177/1367549415577392

Teräs, M., Suoranta, J., Teräs, H., & Curcher, M. (2020). Post-Covid-19 Education and Education Technology 'Solutionism': A Seller's Market. *Postdigital Science and Education*, *2*(3), 863–878. https://doi.org/10.1007/s42438-020-00164-x

Tygel, A. F., & Kirsch, R. (2016). Contributions of Paulo Freire for a Critical Data Literacy: A Popular Education Approach. *The Journal of Community Informatics*, *12*(3). https://doi.org/10.15353/joci.v12i3.3279

Williams, M. L., Burnap, P., & Sloan, L. (2017). Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation. *Sociology*, *51*(6), 1149–1168. https://doi.org/10.1177/0038038517708140

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for the future at the new frontier of power*. Profile Books.