



Data Sprint Learning. Exercising Proximity to Data in Teaching Situations

Aprendizaje a partir de *data sprints*.
Ejercitar la proximidad a los datos en contextos docentes

 **Laura Kocksch**
laurak@ikl.aau.dk
Technoanthropology Lab, Aalborg University

 **Mace Ojala**
maco@itu.dk
IT University of Copenhagen

 **Katharina Kinder-Kurlanda**
Katharina.Kinder-Kurlanda@aau.at
University of Klagenfurt

Received: 23/12/2021

Accepted: 20/06/2022

ABSTRACT

This paper reports on a data sprint conducted as part of a PhD course on digital methods and data critique at the University of Klagenfurt. We reflect on how our data sprint contributed to this higher educational setting, and point to ways in which the data sprint method can be developed further based on our experience. The paper discusses how the sprint fabricated a moment of “critical proximity” for students that were mainly working with qualitative social science methods. The data sprint allowed them to put their critique on “big data” into practice by working with selected sets of data from Twitter and Scopus. We reflect on our collective experience and draw conclusions on the use of data sprints in teaching. Data sprints encourage us to engage with feelings of being underwhelmed and overwhelmed by data that provoke our social science way of critique. Our data sprint tangibly demonstrates that data work is in fact “messy”: transgressing ideals of good data management, biased, ambiguous and open-ended. But instead of turning away from this “wildness”, we urge to make use of it in teaching settings. This wildness allows to step out of conventional modes of critique, and into modes of action. We conclude with a protocol as a practical guide for everyone who wants to introduce data sprints in their teaching.

KEYWORDS

data sprint, data work, data analysis, pedagogy, critical proximity

How to cite this article:

Kocksch,L.; Ojala,M.; & Kinder-Kurlanda,K. (2022). Data Sprint Learning. Exercising Proximity to Data in Teaching Situations. *Dígitos. Revista de Comunicación Digital*, 8: 31-50. DOI: 10.7203/drdcd.v1i8.232

RESUMEN

Este artículo presenta los resultados de un 'data sprint' realizado como parte de un curso de doctorado sobre métodos digitales y crítica de datos en la Universidad de Klagenfurt. Con este artículo, reflexionamos sobre cómo nuestro *data sprint* contribuyó a este entorno educativo superior y señalamos las formas en las que el método de *data sprint* se puede desarrollar, aún más, a través de nuestras experiencias. El documento analiza cómo el *sprint* fabricó un momento de "proximidad crítica" en estudiantes que trabajaban principalmente con métodos cualitativos en el campo de las ciencias sociales. El *data sprint* permitió a los estudiantes poner en práctica críticas sobre la noción de *Big Data* al trabajar con conjuntos de datos seleccionados de Twitter y Scopus. A partir de una reflexión sobre nuestras experiencias colectivas, compartimos conclusiones sobre el uso de *data sprints* en la enseñanza. El método *data sprint* nos anima a interactuar con el sentimiento de estar decepcionados o abrumados por los datos, que provocan un modo de crítica perteneciente a las ciencias sociales. Nuestro *data sprint* demuestra, de manera tangible, que el trabajo de datos es, de hecho, "desordenado"; ya que transgrede los ideales de una buena gestión de datos, al ser sesgado, ambiguo y sin límites fijos. Pero en vez de alejarnos de este "estado salvaje", nosotros invitamos su uso en la enseñanza. Este "estado salvaje" permite exceder modos convencionales de crítica y entrar en modos de acción. De este modo concluimos ofreciendo un protocolo, a modo de guía práctica, para todo aquel que quiera introducir los *data sprint* en su docencia.

PALABRAS CLAVE

sprint de datos, trabajo con datos, análisis de datos, pedagogía, proximidad crítica



Data Sprint Learning. Exercising Proximity to Data in Teaching Situations

1. Introduction

This article reports on a one-day data sprint we conducted online as part of a PhD course in digital methods and data critique at the University of Klagenfurt. We address the educational potential of data sprints, particularly their ability to bring students into "critical proximity" with digital data (Latour, 2005: 253; Birkbak *et al.*, 2015) Our data sprint induced feelings of being overwhelmed in students, when they looked at a digital data set for the first time; encountering its sheer volume or being 'wooded' by colourful visualisations. On the other hand, students also expressed feeling underwhelmed by a lack of relevant questions the data could answer, feeling suspicious of the data and its reductionist nature. This prompted a discussion of the partiality of data with the students which was not tailored towards a distant critique, but turned into a set of analytic categories that imbued the subsequent analytic process. The data sprint facilitated a 'wildness' of working with data that was both disconcerting to students (being overwhelmed and underwhelmed at the same time), but also engendered playfulness and experimentation. Students reacted to the data sprint with both excitement and defiance; feeling at the same time intrigued to acquire technical skills, while also being forced to work with data that they found inherently lacking and biased. These reactions are possible only from a position of closeness to data, where its potential and risks are

out in the open. We describe this disconcerting during the sprint in relation to how it fostered a close-up and productive reflection of data - rather than abstract critique. We argue that it is worth facilitating under- and overwhelmedness, disconcerting and bewilderment in educational data sprints, and we show how we crafted them in the course of our exercise.

A data sprint is a limited-duration event where people are invited to come together to sort, interpret, and visualise a prepared dataset (Venturini *et al.* 2018b; Jensen *et al.*, 2021; Munk *et al.* 2019a). Compared, on the one hand, to workshops, data sprints are more open ended, non-committal and exploratory in nature. On the other hand, compared to hackathons, data sprints are focused on a data set or data sets selected and curated by the organisers prior to the event, fostering a more thorough analysis and conversation during the sprint. Data sprints are interdisciplinary, and can even be seen as non-disciplinary in the sense that what counts as success is left implicit, open for interpretation, and might not adhere to standards of any established field (Munk *et al.*, 2019b). Participants are invited to “mess around” with the data brought in by the facilitators, encouraging also those not proficient in data analysis, statistics or digital methods to get involved. The data thus serve as a boundary object (Star and Griesemer, 1989) during the sprint; while they are pre-configured they also allow flexibility. Issue experts, designers and software developers are invited to facilitate, provoke and guide the process. Controversy might arise as different parties attempt to gain authority in the interpretation of data or use it to support their specific claims about the world. In this regard, data sprints are not solely meetings where insights are extracted from data, but settings in which specific (differences in) epistemic politics become observable.

Data sprints generate extraordinary situations in which participants and organisers who often bring qualitative social science backgrounds can get into proximity with digital data. Sprints thus overcome common themes of critique towards data, and urge participants to “get their hands dirty” (Munk, 2021). This brings together the tradition of controversy mapping with that of participatory design (Venturini *et al.*, 2018b; Munk *et al.*, 2019; Jensen *et al.*, 2021; Venturini & Munk, 2021). In this approach, data sprints aim more at crafting a situation in which stakeholders come together and unfold their specific points of view based on the data at hand. In the process both participants and the organisers step on uncharted territory, eliciting often surprising and unanticipated insights in data (Jensen, 2020; Jensen *et al.*, 2021). Our sprint, however, drawing on this realisation, had pedagogic goals; the sprint familiarised students with a selection of digital methods while also giving them the chance to dip a toe into data work. This forged a contrast to the rest of the PhD course that was based on generic principles of data management and data ethics and scholarly critique of data-driven surveillance and big tech.

Although data sprints have been employed in a variety of domains (see for example Sanderhoff, 2014; Berry *et al.*, 2015) and including a range of stakeholders, or “issue experts” (Munk *et al.*, 2019a) (e.g., scientists, activists, public officials, educators, etc.), so far their potential as teaching situations has not been discussed explicitly. Most

commonly, data sprints are seen as ways to investigate a topic or “issue” (Marres, 2015) while also interfering in and reflecting upon the affordances of the platform and method in the process (Omena *et al.*, 2020; Pearce *et al.*, 2020). As data sprints have proven useful in prompting hands-on reflexivity, this paper aims at exploring their potential in educational settings. We argue that data sprints combine the useful role of “toy problems” and “real world” data, allowing students to try and test data before moving to their own research projects. In the case of PhD education, data sprints offer situations of extraordinary encounters with data, allowing “playing around” with data without committing to include results in a publication or thesis (which is usually the main concern of PhD students). Data sprints allow competent play with data instead of *testing* them for their utility in a thesis or publication. The data sprint’s format also allows reflection of the data analysis process as a whole in a very short time - which may be accomplished over years in the actual PhD project and thus much more difficult to oversee. To investigate the additive value of data sprints in educational settings, we ask: how would a data sprint trigger “critical proximity” in ways a theoretical course would not?

For our data sprint, we chose a topic that aligned with the topic of the digital methods and data critique focused PhD course: Open Data and Open Science. We anticipated students to have limited knowledge of Open Data and Open Science, but as digital humanities and social science researchers to be equipped with the conceptual tools to identify recurring themes or specific problematizations.

The choice of topic and desired outcome of the sprint was threefold. Firstly, we designed the sprint to act as a teaching situation where students would learn about Open Data as an ongoing controversy in science (Levin & Leonelli, 2016; Prainsack & Leonelli, 2018). Secondly, Open Data and sharing of research data was a topic that we, the organisers, were interested in and have published on (e.g., Kinder-Kurlanda *et al.*, 2017; Kinder-Kurlanda & Weller, 2020; Sørensen & Kocksch, 2021). Coming into the data sprint with our own research questions as scholars in science and technology studies, the data sprint allowed us to work on our specific questions towards Open Data controversies, e.g, what similarities and differences manifest in the published research literature and discourse on social media, how openness and concealment are discussed, what methods are (not) widely acknowledged etc. Lastly, it was our goal as educators to exercise the method of data sprints itself, familiarise ourselves with the nitty-gritty of putting together events like this and reflect upon their productiveness in educational settings. Whereas the three goals intertwine, this article specifically aims at a discussion of the last. In synthesis, data sprint learning provided learning for students and educators alike.

In section 2, we outline our data sprint procedure. The section stands in the place of a methodology chapter you might expect in a traditional article. However, it does not include a description of our participant observation data collection or close description of our analysis because we did not follow a strict methodological procedure as prescribed by e.g. ethnographic methods. We report instead as organisers, educators and participants of the data sprint, putting our past experiences in teaching and impressions

during and after the data sprint in conversation. We report how the situation was set up and how it encouraged specific forms of action. We thereby lean on description of teaching situations in science and technology studies (STS) that emphasise how learning is composed by specific socio-material assemblages in teaching situations (Sørensen, 2009). Rather than being interested in analysis or evaluation of the sprint, we aim at reflecting the data sprint method as a way to forge a specific learning setting.

Section 3 is more conventional to readers of ethnographic papers: we present three episodes from the sprint which we in the concluding section 5, emphasise as particularly relevant for understanding how the data sprint created an extraordinary teaching situation.

However leading up to the conclusion, we offer a protocol as a condensed description of our data sprint in the form of a set of instructions in section 4. The “protocol” fulfils a task different from that of the Digital Methods Initiative community where protocols are step-by-step descriptions of tools and techniques which can be reproduced to collect and analyse comparable data sets (Rogers, 2019; Venturini *et al.*, 2018a). Instead our notion of protocol points to what has been suggested by Ballesterio and Winthereik (2021:11) in the form of an “analytic protocol”, that is a “practical guide [...] to set up conditions to create analytic timespace” for ethnographic analysis. Protocols in that sense “invoke a sense of organised reflection” (Ballesterio & Winthereik, 2021:11) that can aid theorising and foster creativity in ethnographic practice. We find that our protocol engenders such systematic reflection of data sprints as moments of getting close to data while remaining reflexive. While Ballesterio and Winthereik argue that protocols should be used to prolong moments of analysis in ethnography and possibly invite others into the process, our protocol relies on a short-term method (quite literally “a sprint”), that fosters fast and messy decision making. As we point out, the wildness of the data sprint method has potential for ethnographic encounters with data. Specifically, we reflect how working with data close up can generate analytically productive *disconcertment* and *bewildering*; reactions that are cultivated by employing our protocol. It is intended to be re-used and developed in other teaching settings where data forms a matter of concern (Latour, 2004; Puig de la Bellacasa, 2017).

2. Method and structure of the data sprint

The data sprint was a 6-hour event conducted over a video-conferencing platform in June 2021. The data sprint was organised in two sessions. A morning session included introductions of research topics and interests by the participants and an introduction to the data by the organisers, and developing research questions. In an afternoon session students worked in groups of 2-4 that were interested in a specific question developed in the first.

Figure 1 shows the structure of the data sprint, including introductory sessions and open working sessions. Section 1 introduced the data sprint and gave an introduction into the data collection process for both data sets. Section 2 introduced students to network visualisations, specifically Gephi. Sections 3, 4 and 5 were designed as group working time with occasional prompts to “freeze” and report. The presentation deck

made use of the “sprint” metaphor (with both the background and the “freezes”). The exact time slots were adopted during the sprint.



Agenda	
10-10.40	Introduction to data sprint, Data collection 1 (research references) and 2 (social media)
10.40-10.50	Break
10.50-11.20	Intro to networks
11.20-11.30	Break
11.30-12.30	Activity 1+2 (inkl. Freezes): Question formulation
12.30-13.30	Lunch Break
13.30-14.30	Activity 3+4 (inkl. Freezes): Network analysis
15-15.20	Break
15.20-16	Activity 5+6: Finalizing prototypes and wrapup

Figure 1. Structure of the data sprint (as shown to participants, original slide deck)

The groups were supported by the organisers in handling data graphs and tweaking visualisations with Gephi (Bastian *et al.*, 2009). The second session closed with a discussion of interesting findings and further questions. The entire event was facilitated by and documented on Miro – a canvas tool for online collaborative work. Throughout the event we made various efforts to account for the fact that we as participants were all coming together online, while also sitting alone in very different dorm rooms, patios and offices. We tried to create elements of commonality (e.g. everyone was asked to bring a mouse in addition to the laptop to facilitate better navigation of the graphics) and different possibilities to engage with each other, by voice, in chats, on boards etc. to allow for different needs and wants regarding interaction.

Ten people gathered for the data sprint, including three organisers. The participating PhD students were already advanced in their projects. In the sessions leading up to the data sprint they had been provided with literature and presentations on a variety of issues around methodological, epistemological and ethical concerns of Big Data and their treatment in the more social science-oriented literature. For example, “platform politics” had been discussed as well as the issue of researchers becoming complicit in platform power and user data commodification. The sprint thus provoked the critical suspicion towards data that they had trained in the rest of the course. The event started with a round of introductions where participants stated their backgrounds and familiarity with data work, Digital Methods as advanced by Rogers (2013; 2019) and others, and the topic Open Data/Open Science. Half of the participants had previously worked with or considered working with digital methods, mainly with social media data and digital archives. The other half reported they were not familiar with digital methods and expressed doubts how valuable such methods could be for their research. Some

articulated that in order to be critical about digital methods, they would want to know more about their “inner” functioning.

We requested the participants to prepare by installing the network analysis tool Gephi (Bastian *et al.*, 2009) before the event and watch two video tutorials. We had prepared two independent data sets. The first was a set of academic references, sourced from Elsevier's bibliographic database Scopus. The second was a set of social media posts from Twitter, collected using 4CAT (2021) via the Twitter API v2 Academic Research Track, a data product the company made available in January 2021 (Parack, 2021). Prior to the sprint, both data sets were pruned and cleaned to allow easier handling on students' own computers. This entailed filtering out entries in the Scopus set which lacked semantic data, i.e., an abstract or summary, and narrowing the Twitter sample down to span no more than one week, and transforming from JSON to CSV. We intentionally left the data sets *semi-cleaned*, e.g., leaving in entries that were misaligned (information on authors in the wrong column in the Scopus set) or hard to discern (extensive Twitter links). We also left files in formats that we anticipated might turn out to be non-trivial to open with Excel. While we tested turning CSV files into the specialised network format using the Table 2 Net tool, we did not provide ready-to-go Gephi files. Data file formatting was not intended to be an essential part of the workshop, however, in hindsight we recognize it productive to leave data in a rough shape such that students experience some of the “underwhelming” chores of data work. The following section reflects in detail on how students interacted with the data sets.

3. Data Sprint Learning. Being under- and overwhelmed, disconcerted and bewildered

This section describes three aspects of what characterised the process of learning during the data sprint: manoeuvring feelings of being underwhelmed and overwhelmed disconcertment over biases, and a negotiation of both wildness and bewilderment.

3.1. Being both under- and overwhelmed

Discussion during the sprint and participant feedback collected afterwards expressed sensations of being both under- and overwhelmed by the data, the tools and the sprint format itself. Students reported that their first impression was that of too many things to install, learn to operate and understand. For most of them it was not only the first time working with social media data or a corpus of literature references and the first introduction to network analysis and Gephi, but also the first introduction to analysing and visualising digital data at all. In the foreground of much of the work during the data sprint was therefore familiarising students with how to encounter such data and tools in the first place: how to open a CSV file? What re-formatting is needed for it to open correctly (with the right separator)? How to spot that it is opened incorrectly?

These ‘alien encounters’ with data resulted in students being overwhelmed by the means and potential that lay before them. They were ‘wooded’ by some of the visualisations that they produced with our guidance, putting them in a state of amazement that initially

prohibited any questioning or critique.

But on the other hand students expressed being underwhelmed; they pointed out that the data was full of gaps, absences, biases and that much of data work seemed almost mundane. This became most apparent in the first session where we instructed students to open a CSV file with Excel or LibreOffice Calc. Most of them encountered a separator issue (between the “,” used in the generated data sets, and the locale setting of “;” in their program). Translating the data back and forth from Calc into Excel and Google Spreadsheet seemed random to students and a tedious process to overcome. The mundanity of data work stood in conflict with how they had imagined it: a thorough analytic process. Their underwhelming was added to when they had managed to open the CSV file as a table which still did not allow a lot of conclusions. Students lacked familiarity with *reading* a CSV file, leaving them with a table without any interesting insights.

For the students competent in qualitative methods these struggles led to being underwhelmed by the perceived reductionist and positivist nature of the methods, and the banality of the data (which was even deepened with the data changing seemingly randomly from a table into text and back). In other words, students were not impressed with the data and data work, even though they were amazed and also overwhelmed by the techniques in use and visualisations we showed them as a possible product of the sprint process.

3.2. Disconcertment over data bias

In order to provoke their social science interest more, we explored with them how the data had been collected and discussed what biases and data gaps were introduced in the process; or how absences “haunted” the data (Meldgaard Kjær *et al.*, 2021). Our intention was to make them see the data as carefully crafted rather than “raw” (Bowker, 2009; Gitelman & Jackson, 2013: 2), and therefore capture their interest in the data as the result of a specific selection process. We then asked them to identify traces of that process and additional gaps and biases in the data, hoping to train them in situating the data set.

In preparation for the event, we had compiled two lists of biases to seed discussion, one for each data set. For the first dataset from Scopus, we listed:

- Scopus contains mainly English language publications in their final stages
- My search prioritizes highest citation counts (last years may only be published next year due to scientists’ citation practices)
- Scientific Publications only

For the second dataset sourced from Twitter, we listed the following potential biases:

- Twitter contains intentional as well as casual conversation

- High volume limits the timespan for our query
- What does #opendata OR “open data” OR #openscience OR “open science” not match?
- How is the experience of using Twitter via the app different from using it via a spreadsheet program?
- Is all content human-made? Does this matter?

Stating that data is incomplete and potentially biased is not enough. It does not come as much of a surprise to any sociologist, critical data studies or science and technology studies scholar. It also risks being defeating. Rather, we took it as a starting point and stated more precisely how *the* particular data set at hand was crafted in a specific way and hence haunted by specific absences (Meldgaard Kjær *et al.*, 2021). During the sprint, data gaps familiarised the students with the process of collecting data and composing simplified data sets that would allow students easy access to methods and data. We did not bring the biases up to check them off our list - to make our encounter with data more “pure” as we have moved the necessary tedious data composition out of the way. Rather, we fostered a conversation about biases to reach “disconcertment” in students - a critical, yet, productive, reflexive analytic mode to discern between different epistemic systems (Verran, 2001).

We asked students to open the data sets on their computers and look at them in standard spreadsheet software. Students were prompted to analyse the data and make notes on: “What data is missing? Whose point of view is rendered absent? What would be nice to have?”

Students came back with various additions to our data, e.g., that they did not contain any information about gender, age and ethnicity of authors neither in Scopus nor on Twitter, which does not allow us to investigate further whether certain points of view are over-/underrepresented. Neither did the data contain any geographical information or institutions which could have helped to make transparent whether the conversation on open data was dominated by Western ideals of science, disregarding any precarities of opening indigenous knowledge, for example. Students also noted that Twitter data lacked real names and that, or that in comparison, Twitter data contained more controversy than the Scopus data set. See figure 2 for some of the notes taken during this session of the data sprint.

The exercise had a two-fold effect: it brought students closer to the data sets, letting them read through and think about what objects the data refer to. As students were not familiar with the dataset or the topic, this moment contributed to their feeling of overwhelmedness and underwhelmedness. Letting them go through the data with the sole goal to point out what was missing appeared arbitrary. It was left open to them what exactly to identify, leaving them space to employ any line of critique about digital data that they were familiar with. For example, one student suggested applying typical lines of critique from her social science background, e.g., biases based on gender, race, class or age as a strategy to identify biases in the data set.

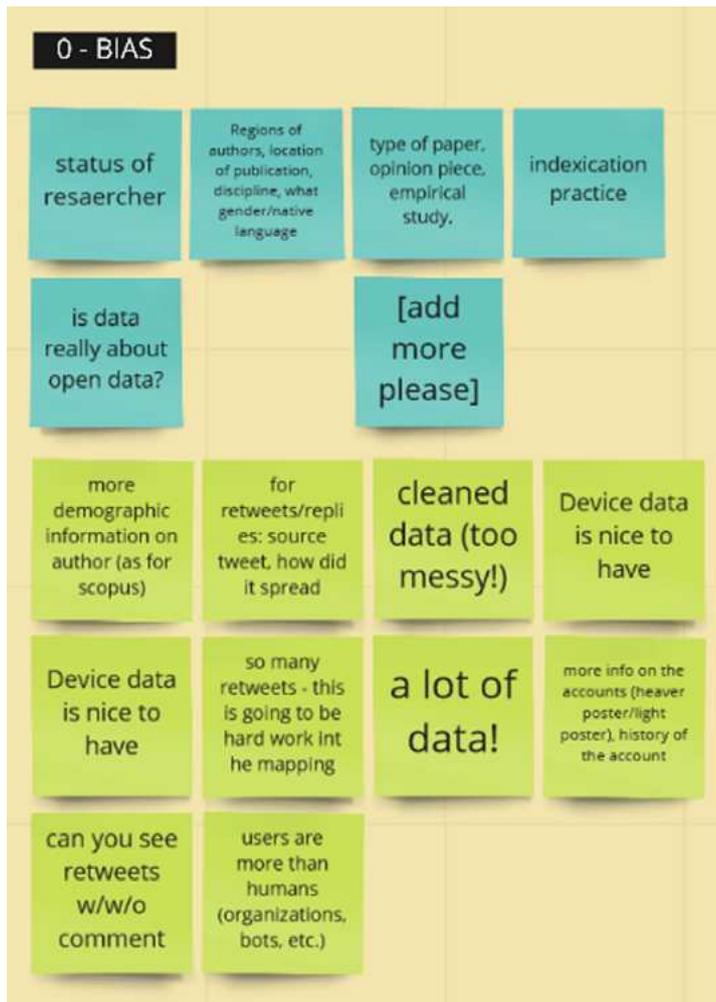


Figure 2. Collaborative note-taking work in progress

Secondly, letting students point to biases trained their critical gaze: asking what and how knowledge is imbued with a specific local and historical gaze is a key strategy in feminist STS (Haraway, 1988), something that we exercised here in a very pragmatic and concrete form. Students were asked to localise their “disconcertment” with data in the data set, turning it into an analytic strategy. Disconcertment allows us to notice moments of suspicion and scepticism and turn them into an analytic strategy. Localising “biases” was a task of critiquing up close.

Thirdly, we could have concluded the work here, and continued a conversation about dramatic effects of biases and gaps in all data (which would have certainly been worthwhile, see Criado Pérez, 2019). But we did not. We did not want either ourselves or the students to lean back and ruminate about biases and their consequences. Rather

the sprint format prompted us all to move on and stay with what we had, however disconcerting.

In the next step, we asked students how to deal with the biases and gaps they identified. We started from four responses to biased data: “(1) Ignore them, (2) Don’t use the data, (3) Reflect upon them in a later analysis or (4) Contextualise the data employing the biases.” See figure 3 top.

Subsequently we asked students to come up with more ways of addressing biases in the specific data set. Together we compiled a new list, with more perspectives. The extended list in figure 3 included (5) throwing statistics at the biases (quantifying whether they are statistically significant) (6) consulting with interdisciplinary colleagues, non-academics, minorities or domain experts (qualifying whether they are relevant or not and what harm/benefit the data set could do nevertheless) (7) searching for existing societal biases like race, gender, class, etc. (linking the data to larger societal problems making it an exemplary result of power structures) or (8) throwing more data at it (adding more data to present a more “complete” picture).

The feeling of disconcertment was later re-examined when students were analysing



Figure 3. Suggestions for how to deal with biases in data.

the data sets by themselves: some decided to take out “big” nodes in their Gephi visualisations to see how that would change the relations between other nodes – accounting for making hierarchies in data appear in a different light. Others suggested it needed more data to make their initial findings more sound, or one suggested that it needed domain experts to interpret the data they had visualised.

So, how did the students experience data gaps and biases differently through the proximity with data during the sprint? Firstly, they were prompted to localise biases and data gaps in the concrete data sets at hand. This shifted generalised and abstract critique of data bias to very concrete questions of authorship, platform politics, collection methods and who was present.

Secondly, broad strokes critique of data bias was not a point of closure, but a disconcertment turned into a starting point for future queries. Besides quantifying the biases or de-biasing the data with more complete data, others suggested to rather make biases a strength and employ them analytically, e.g. by inviting affected publics in to comment and utilise the data. When visualising the data with Gephi, some proposed to delete some of the data to see other relations more clearly. These wild ideas demonstrate that the data sprint urged students to exercise agency and materially engage with the partiality and composition of data sets pragmatically and productively. The sprint format afforded balancing material work with a feeling of disconcertment. We stayed with the biased data, making biases part of the pragmatic data work. Biases became nothing to evade, circumvent or be defeated by, but something that is part of the analytic process. Rather than imagining a clean – complete, unbiased, absolute – data set, students made use of their disconcertments surrounding data’s reductionism and partiality. Due to - not despite - their partiality, data became a relevant object of inquiry to students (Jacomy, 2021; Haraway, 1988).

3.3 Wildness and bewilderment

Rather than being well-disciplined, accountable and tidy, in other words tame and domesticated, data sprints are a rowdy mess. This is their virtue. Sprints focus on whichever doings seem most plausible for short-term goals. As the name suggests, sprints are relatively short, intense and single-minded events relaxed from the concerns of what will follow afterwards. Isolated from other data work, a sprint is thus a safe place to experiment. A data sprint typically avoids having pre-defined and well delineated success criteria. As a research situation the research questions, if formulated at all, are typically exploratory. A data sprint does not aim at, and cannot reach robust findings, consideration of nuance and edge cases, or final solution to all the biases. Hence, data sprints are especially appropriate when interfacing with participants (Munk *et al.*, 2019a; Jensen *et al.*, 2021), or for generating ideas at the early stages of a research.

Rather than, for example, teaching and disciplining the participants to the more well-defined issues of traditional data management the impetus is to find new questions and to allow for surprise and *bewilderment* to happen without becoming incapable of acting. A simple but illustrative example of how a certain bewildered wildness manifested in our data sprint, observe some of the data filenames used during the sprint:

- 4CAT Dataset/convert-csv-excel-70212a65460708be9088c35e73f72b36.csv
- 4CAT Dataset/convert-csv-excel-70212a65460708be9088c35e73f72b36.actually-c-separated.csv
- 4CAT Dataset/opendata-or-open-data-a884822c87a999a5f95d8a521846c7c5.ndjson
- Scopus Dataset/Scopus-2/scopus-2.csv
- Scopus Dataset/Scopus-2/scopus2-nurAutJahCit.csv
- Folders Scopus Dataset/new, Scopus Dataset/medium and Scopus Dataset/waste

Some of the above shown filenames index the data source or the device of collection (*Scopus**, *4CAT**), some document the procedures taken (*convert-csv-excel**; multilingual **-nurAutJahCit**, **.actually-c-separated**), some document versioning (**-2**; because this was improvised, there is no **-1**), some recycle unique identifiers such as *70212a65460708be9088c35e73f72b36* from the data sets.

The filenames are idiosyncratic, improvised and do not follow an established naming scheme as good data management would expect – they are more wild than that. However, they are neither random nor without any management; for instance, we observed the application of practices of no-space, CamelCase and generally conveying meaning in the filenames, but the practices are unsystematically applied. This wildness of filenames generalises to other decisions done while conducting loosely governed data work without the constraints of documentation and specification requirements, provenance, metadata management and security requirements of a more serious

data management regime of a more sustainable research situation, especially in collaborations.

The data sprint did not encounter wildness as it may be common in technology design - where research "in the wild" refers to research or design conducted outside of experimental settings (Hutchins, 1995; Rogers, 2011). The wildness composed during our sprint was that of an experiment; a setting of contrived and liberating unruliness. Instead of getting wildness into the design process, the sprint aimed at introducing wildness as a moment of becoming familiar with and uninhibited by data work while reflecting about it.

4. A checklist to data sprint learning

In this article we argue that facilitating moments of underwhelmedness, overwhelmedness, and bewilderment is productive for learning situations. The data sprint forged encounters with data that affected students; both exciting them to work with data in the future and confronting them with the nitty-gritty of data work. We were able to exercise in the data sprint these moments because our group (both teachers and students) was composed of scholars who were used to taking a critically reflective stance as a default. In participants from very different academic backgrounds the data sprint may have had a different effect, possibly engendering less, rather than more reflection and critique. The elusive state that we wanted to capture in this exercise was the experience of familiar notions and habits being challenged and possibly being turned upside down. Engendering similar moments of bewilderment in, say, most engineers would probably have required a completely different sprint setting, for example a reading and writing of philosophy texts or the generation of a theory. Our data sprint reached such an encounter with data by following some principles we will summarise in the following.

The article leaves you with lessons learnt from our data sprint: five dos and two don'ts. We ask you to take it as a device to slow down judgement during your own data sprints in teaching contexts; a device to prompt revision, critique, stumbling; a physical object you can toss around, cut up, rework, leave coffee stains on and discard if necessary.

Five dos

- *Choose a very particular case*, grounded in research. This allows to mitigate overwhelming and underwhelming, as students are capable of responding to and articulating data in regard to a concrete issue at hand.
- *Curate a small data set* that is comprehensive to students when looking at it without any visualisation techniques (in order to be able to point to gaps and biases).
- *Clean up data beforehand* - there will still be plenty that is new and bewildering to students.
- *Assign preparatory homework*, in moderation. Get participants to install and try out tools beforehand. When conducting a course on data management, think

about how to present the “ideal” case and how the data sprint can productively counternarrate formal data management rules.

- *Prepare students* that the data sprint will be exploratory and they are not being *tested*. We observed some defiance and self-deprecating tendencies that could be mitigated by making clear that the data sprint is about bringing forth an individual’s expertise rather than testing their competencies in a fixed set of technical skills. We had some trouble enacting that; while the organisers had more technical expertise, the goal was to have students select and design visualisations that were useful to them (enacting their own expertise in their field of study).

Two don’ts (some of which we failed at)

- *Don’t prioritise explanation time* over workshop time; be aware that both overlap but the sprint should give enough time to explore data independently. Let students know the sprint is a playground and they should not mistake it as a test of their abilities (forming an utterly different academic encounter than they are used to). We see this point related to the last point stated above: With extensive explanations of the data collection method, students saw organisers as experts on the topic - which is not the case. We missed the goal to have students become experts of the data themselves by using it for their individual research projects.
- *Don’t hesitate to set up dynamic, multimodal prompts* such as “chat storms” (asking for written answers in the chat), background music, and different camera perspectives. In an online sprint, processes of group formation and a sense of commonality can be aided by these methods. They help to loosen the atmosphere and enhance the feeling of a playground.

5. Conclusion and implications

Data work in contemporary technoscience requires dealing with ambiguity, vagueness, inconclusiveness, effects of being overwhelmed and underwhelmed by data, and its untameable biases. We have argued that our data sprint did not only teach students about such messiness in a matter-of-fact or distanced critical mode, but let them experience it first-hand. Furthermore, it prompted them to put some aspects of the received critique on platforms, methods or quantitative data to work. This was reached through the nature of “wildness” that the temporary, extraordinary data sprints events embody: students don’t have to be inhibited by what they read – concepts, best practices, principles, methods or critique – nor do they need to cite literature, maintain metadata or be precise about their terminology. They are allowed, enabled and indeed expected to leave inhibitions aside to explore, do and *tinker* (Mol *et al.*, 2010; Law, 2011), and generate preliminary, affective, contradictory or undertheorized insights.

Pedagogically, we argue, *wildness* of a data sprint is valuable. Firstly ignoring the time-consuming 80% of wrangling of the so-called “80/20 rule”, and jumping straight into the 20% of analysis allows getting further along the data processing “pipeline”. But the data sprint also showed that the practices of cleaning and analysis can not be neatly

separated and that, in fact, leaving some roughness in the data urged students to ponder about the composition of data. This point has been made in relation to local traces, such as the separator (Loukissas, 2019). Students found themselves confronted with the tedious side of data work (for example translating formats, making local copies, etc) which fostered analytic overwhelming and underwhelming as well as disconcertment over the banality and politics of data at the same time.

Secondly, by doing so, wild data management enables first-hand experience of engaging with the struggles of data work, rather than disables that experience like *a-priori* critical distance tends to do. Proximity invites the material to present some of its own, inherent troubles. Facing the temptations, passions and excitements is part of engaging the Zuckerbergian “move fast and break things” experience in a critical proximity (Latour, 2005: 253; Birkbak *et al.*, 2015; Herbrechter, 2017). Learning to manoeuvre between proximity and distance is necessary for reflection, and zooming in and zooming out is productive for critique and understanding not only data as such, but also its conditions. Establishing a habit for moving between proximity and distance can, in longer term, move an individual researcher or a research project toward *critical technical practice* (Agre, 1997) – and given that most contemporary academic work is deeply involved with technology (Latour, 1987) – what we might call a *critical technoscientific practice*.

Thirdly, wildness levels the playing field by decentering data work expertise and opens the process to wider participation and people with a diversity of skill sets. Students were able to apply their critique without staying back, turning into analytic strategies *with* the data.

If the bewildering outputs of a data sprint – provisional insights, sketches, visualisations, provocations, hypotheses, personal contacts, data versions and scripts – are to be integrated into a more structured research project with the usual research accountability needs this mess needs to be tamed. If left undomesticated, the sprinting will not be able to contribute to moving the research forward in productive and sustainable ways. As such, wild data management is a teachable moment of taking “technical debt” to be paid later, for the benefit of gaining momentum to achieve short-term goals (Cunningham, 1992). In the educational context of the PhD course, balancing this tradeoff was safely contained in the isolated experimental setting, with the data sprint as a “lab”, rather than directly imposed on the students’ own PhD research projects. The reflection and insights gained in the data sprint could be taken up by students in different ways and settings later on - and the ‘debt’ be paid there - rather than, e.g. be addressed in a longer data sprint.

As authors, we recognize that encouraging wild data management is itself a risky proposition; it suggests that throwing out good practices under pressure, “in a hurry”, is fine. This is a hazardous lesson since all research work is under pressure. Not only is research intrinsically hard epistemic work, but also the conditions of research are a source of pressure. The sum of the funding schemes (including enrollment in a PhD programme), the toxic ‘publish or perish’ doctrine and other hybrid entanglements within which these PhD students will conduct their work now and in the foreseeable

future seem to suggest that there are always reasons to cut corners. Perpetual sprinting and hurrying seems not only a rational idea, but under these conditions an imperative! This is a reasonable, but unfortunate conclusion. We by no means suggest that data work in research should replace its ideals for stringent standards with wildness typical of data sprints, or valorize accumulating “technical debt” without strategies to pay it back. We argue however that gaining the experience first-hand in critical proximity (Latour, 2005: 253; Herbrechter, 2017; Birkbak *et al.*, 2015) connects a researcher to the temptations, excitements and doubts of data work, and helps scholars develop a critical technical practice instead of, or in addition to, categorical rejection of data work. Facilitating a stance of critical technical practice in the different phases of ‘sprinting’ was surely the biggest challenge for us as teachers and brought its own moments of over- and underwhelmedness with the intended, perceived and/or achieved results.

The data sprint urged us all to continue working with data despite its biases, and thus stay with the trouble for a while in critical proximity. The result of surfacing specific biases during the sprint was not that data was less biased afterwards, nor solely that we can “tick the box” by saying we reflected on biases and thus legitimise using the data nonetheless. Rather than surrendering to biases and sensations of overwhelmedness and underwhelmedness, sprint participants stayed with the vagueness and slipperiness of data through modes of tinkering with the composition of data, trying out different “views” and swapping data out for each other. The data sprint’s material orientation as well as playful wildness facilitated students not simply to think about or expose biases in the data as an empirical finding, but turned biases into pragmatic work.

The mode of critique elicited through sprinting was that of engagement and productiveness, of *thinking-with* (Puig de la Bellacasa, 2017: 71), proposing an analytic tool to ask fruitful questions rather than establish closure. The sprint presented data not as an explanation for anything but much more as something that requires suspicion, that fosters disconcertment, halting and questioning. Student’s sensations of being underwhelmed and not inclined at first sight was immensely valuable; they did not fall into the sheer promises of data work, but pointed to gaps and biases.

We emphasise that data spring learning was not limited to the students learning digital methods, but also allowed us, as organisers and authors to learn crafting situations in which data operates as a device to foster critical questioning and slows down judgement.

Acknowledgements

The data sprint was funded by the Center for Advanced Internet Research (CAIS NRW). We owe our gratitude to the group of PhD fellows and colleagues from D!ARC (Digital Age Research Center) and RUSTlab (Ruhr University Science and Technology Studies Lab) who participated in the data sprint; for without their engagement, expressions of bewildering and provoking reflections this paper would not exist. We are also indebted to our anonymous reviewers and the editors who significantly improved the structure and argument of this manuscript. Further, we are grateful to Barbara Nino Carreras who generously translated our abstract into Spanish.

References

- 4CAT Capture and Analysis Toolkit [Computer software]. (2021). OILab and Digital Methods Initiative at the University of Amsterdam. Retrieved from <https://github.com/digitalmethodsinitiative/4cat>
- Agre, P., Toward a Critical Technical Practice: Lessons Learned in Trying to Reform AI. In Bowker, Geoffrey, ed. (1997). *Social Science, Technical Systems and Cooperative Work: Beyond The Great Divide*. USA: L. Erlbaum Associates Inc. ISBN 978-0-8058-2403-2.
- Ballestero, A., Winthereik B. (2021). *Experimenting with Ethnography. A Companion to Analysis*. Duke University Press.
- Bastian M., Heymann S., Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*.
- Berry, D. M., Borra, E., Helmond, A., Plantin, J. C., & Rettberg, J. W. (2015). The data sprint approach: exploring the field of Digital Humanities through Amazon's application programming interface. *Digital Humanities Quarterly*, 9(4).
- Birkbak, A., Petersen, M. K., & Elgaard Jensen, T. (2015). Critical proximity as a methodological move in techno-anthropology. *Techné: Research in Philosophy and Technology*, 19(2), 266-290.
- Borgman, C. L., Scharnhorst, A., & Golshan, M. S. (2019). Digital data archives as knowledge infrastructures: Mediating data sharing and reuse. *Journal of the Association for Information Science and Technology*, 70(8), 888-904.
- Bowker, G. C. (2009). *Memory Practices in the Sciences*. MIT Press
- Corti, L., Van den Eynden, V., Bishop, L., & Woollard, M. (2019). *Managing and sharing research data: a guide to good practice*. Sage.
- Criado Perez, C. (2019) *Invisible Women. Exposing Data Bias in World Designed for Men*. Vintage Books.
- Cunningham, W. (1992). The WyCash Portfolio Management System. OOPSLA '92. <http://c2.com/doc/oopsla92.html>. Accessed 1 November 2021.
- Gitelman, L. and Jackson, V. Introduction. In Gitelman, L. (ed.) (2013). *"Raw Data" is an Oxymoron*. MIT Press.
- Godfrey-Smith, P. (2003). *Theory and Reality. An Introduction to the Philosophy of Science*. Chicago: University of Chicago Press.
- Haraway, D. (1988). Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, 14(3), 575–599. <https://doi.org/10.2307/3178066>
- Haraway, D. (2016). *Staying with the trouble: Making Kin in the Chthulucene*. Duke University Press.
- Herbrechter, S. (2017) Critical proximity, *Journal for Cultural Research*, 21:4, 323-336, DOI: [10.1080/14797585.2017.1370485](https://doi.org/10.1080/14797585.2017.1370485)

- Hutchins, E. (1995). *Cognition in the Wild*. MIT Press.
- Jacomy, M. (2021). *Situating Visual Network Analysis*. Aalborg Universitetsforlag, Aalborg Universitet. Det Humanistiske Fakultet. Ph.D.-Serien
- Jensen, T. E. (2020). Exploring the Trading Zones of Digital STS. STS Encounters - DASTS working paper series, 11(1), 89-116. https://www.dasts.dk/wp-content/uploads/4_Trading_FV_1.pdf
- Jensen, T. E., Birkebæk, A., Madsen, A. K., & Munk, A. K. (2021). Participatory Data Design: Acting in a digital world. In *Making and Doing STS*. MIT Press.
- Kinder-Kurlanda, K., Weller, K., Zenk-Möltgen, W., Pfeffer, J., & Morstatter, F. (2017). Archiving information from geotagged tweets to promote reproducibility and comparability in social media research. *Big Data & Society*. <https://doi.org/10.1177/2053951717736336>
- Kinder-Kurlanda, K. E., & Weller, K. (2020). Research Ethics Practices in a Changing Social Media Landscape. *AoIR Selected Papers of Internet Research, 2020*. <https://doi.org/10.5210/spir.v2020i0.11251>
- Latour, B. (1987). *Science in Action: How to Follow Scientists and Engineers Through Society*. Harvard University Press. ISBN 0-674-79291-2
- Latour, B. (2004). Why has critique run out of steam? From matters of fact to matters of concern. *Critical Inquiry*, 30(2), 225-248.
- Latour, B. (2005). *Reassembling the Social. An Introduction to Actor-Network Theory*. Oxford University Press.
- Law, J. (2011). Heterogeneous Engineering and Tinkering. <http://www.heterogeneities.net/publications/Law2011HeterogeneousEngineeringAndTinkering.pdf>. Accessed 1 November 2021.
- Levin, N., & Leonelli, S. (2017). How Does One “Open” Science? Questions of Value in Biological Research. *Science, Technology & Human Values*, 42(2), 280–305. <https://doi.org/10.1177/0162243916672071>
- Loukissas, Y. A. (2019). *All data are local: Thinking critically in a data-driven society*. MIT Press.
- Marres N. (2015). Why Map Issues? On Controversy Analysis as a Digital Method. *Science, Technology, & Human Values*, 40(5):655-686. doi:10.1177/0162243915574602
- Meldgaard Kjær, K., Ojala, M., Henriksen, L. (2021). Absent Data: Engagements with Absence in a Twitter Collection Process. *Catalyst*, 7(2). <https://doi.org/10.28968/cftt.v7i2.34563>
- Mol, A., Moser, I. and Pols, J. (eds) (2010). *Care in Practice: on Tinkering in Clinics, Homes and Farms*. Transcript verlag, Bielefeld.
- Munk, A. (2001). The digital minced meat. *EASST Review*. 40(1). <https://easst.net/article/the-digital-minced-meat/>
- Munk, A. K., Tommaso, V., & Meunier, A. (2019a). Data Sprints: A Collaborative Format in Digital Controversy Mapping. In J. Vertesi, & D. Ribes (red.), *Digital STS: A Field*

Guide for Science & Technology Studies (s. 472-496). Princeton University Press.
<https://doi.org/10.2307/j.ctvc77mp9.34>

Munk, A. K., Madsen, A. K., & Jacomy, M. (2019b). Thinking Through The Databody: Sprints as Experimental Situations. In Å. Mäkitalo, T. Nicewonger, & M. Elam (Eds.), *Designs for Experimentation and Inquiry: Approaching Learning and Knowing in Digital Transformation* (1 ed., pp. 110-128). Routledge. <https://doi.org/10.4324/9780429489839>

Omena, J. J., Rabello, E. T., & Mintz, A. G. (2020). Digital Methods for Hashtag Engagement Research. *Social Media and Society*, 6(3), 1-18. <https://doi.org/10.1177/2056305120940697>

Parack, S. (2021). Introducing the new Academic Research product track. *Twitter Developer forum*. Retrieved from <https://twittercommunity.com/t/introducing-the-new-academic-research-product-track/148632>.

Pearce, W., Özkula, S. M., Greene, A. K., Teeling, L., Bansard, J. S., Omena, J. J., & Rabello, E. T. (2020). Visual cross-platform analysis: digital methods to research social media images. *Information Communication and Society*, 23(2), 161-180. <https://doi.org/10.1080/1369118X.2018.1486871>

Prainsack, B., & Leonelli, S. (2018). "Responsibility". In *Science and the politics of openness*. Manchester, England: Manchester University Press. Retrieved Dec 22, 2021, from <https://www.manchesteropenhive.com/view/9781526106476/9781526106476.00013.xml>

Puig de la Bellacasa, M. (2017). *Matters of care: Speculative ethics in more than human worlds*. University of Minnesota Press.

Rogers, R. (2013). *Digital Methods*. MIT Press.

Rogers, R. (2019). *Doing Digital Methods*. Sage.

Rogers, Y. (2011). Interaction design gone wild: striving for wild theory. *Interactions*, 18(4) <https://doi.org/10.1145/1978822.1978834>

Sanderhoff, M. (ed.). (2014). *Sharing is Caring. Openness and Sharing in the Cultural Sector*. Statens Museum for Kunst. Copenhagen, Denmark. <https://www.smk.dk/en/article/the-sharing-is-caring-anthology/>

Star, S., Griesemer, J. (1989). Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science*, 19 (3): 387-420. DOI: <https://doi.org/10.1177/030631289019003001>

Sørensen, E. (2009). *The Materiality of Learning: Technology and Knowledge in Educational Practice* (Learning in Doing: Social, Cognitive and Computational Perspectives). Cambridge: Cambridge University Press. DOI:<http://doi.org/10.1017/CBO9780511576362>

Sørensen, E. and Kocksch, L. (2021). Data Durabilities: Towards Conceptualizations of Scientific Long-Term Data Storage. *Engaging Science, Technology, and Society*, 7(1): 12-32. DOI: <https://doi.org/10.17351/ests2021.777>

- Venturini, T., Bounegru, L., Gray, J., & Rogers, R. (2018a). A reality check(list) for digital methods. *New Media & Society*. 20(11): 4195-4217. DOI: <http://doi.org/10.1177/1461444818769236>
- Venturini, T., Munk, A., & Meunier, A. (2018b). Data-Sprinting: a Public Approach to Digital Research. In: Celia Lury; Rachel Fensham; Alexandra Heller-Nicholas. *Routledge handbook of interdisciplinary research methods*, Routledge, Routledge international handbooks, 978-1-138-88687-2. ff10.4324/9781315714523-24ff. Ffhal-01672288f
- Venturini, T., & Munk, A. K. (2021). *Controversy Mapping: A Field Guide*. John Wiley & Sons.
- Verran, H. (2001). *Science and an African logic*. University of Chicago Press.