Original Research Article

Broken data: Conceptualising data in an emerging world

Sarah Pink[1], Minna Ruckenstein[2], Robert Willim[3] and
Melisa Duque[4]

Abstract

In this article, we introduce and demonstrate the concept-metaphor of broken data. In doing so, we advance critical discussions of digital data by accounting for how data might be in processes of decay, making, repair, re-making and growth, which are inextricable from the ongoing forms of creativity that stem from everyday contingencies and improvisatory human activity. We build and demonstrate our argument through three examples drawn from mundane everyday activity: the incompleteness, inaccuracy and dispersed nature of personal self-tracking data; the data cleaning and repair processes of Big Data analysis and how data can turn into noise and vice versa when they are transduced into sound within practices of music production and sound art. This, we argue is a necessary step for considering the meaning and implications of data as it is increasingly mobilised in ways that impact society and our everyday worlds.

Keywords

Digital data, broken world theory, repair, improvisation, ethnography

Introduction

In this article, we introduce the concept-metaphor of *broken data*. In doing so, we argue for and demonstrate an approach to data that attends to the often obscured or invisible elements of how it is disrupted, repaired, and always contingent on other processes of development around it. Re-thinking data in this way, we argue is an important and necessary step in accounting for what data can possibly mean in our everyday worlds. In doing so, we advance existing critical discussions of digital data through a perspective that mobilizes concepts relating to breakage, repair, and growth, and sees these processes as inextricable from the ongoing forms of creativity stemming from everyday contingencies and improvisatory human activity. By focusing ethnographic attention towards demystifying how digital data is constituted *in situ*, we argue that to better understand data futures, we need to account more fully for the incomplete, contingent and fractured character of digital data. We propose that to do so concept-metaphors (Moore, 1999, 2004) which are partial and perspectival framing devices, offer a means of describing the emergent qualities of materiality and environment.

In developing this discussion, we demonstrate how the insights drawn from ethnographic and autoethnographic encounters introduce a novel perspective to existing narratives about the imperfections of digital data.

Within existing literatures about data, there have already been several moves towards disrupting notions of data as objective, raw or pure. Scholars working in the interdisciplinary field of 'critical data studies' (Iliadis and Russo, 2016) have problematized the expectations and value assigned to data within the interplay between technological affordances and socio-cultural as well as political contexts (boyd and Crawford, 2012; Manovich, 2013; Markham, 2013). Others have drawn attention to the materialities and

[1]RMIT University, Australia; Halmstad University, Sweden
[2]University of Helsinki, Finland
[3]Lund University, Sweden
[4]RMIT University, Australia

Corresponding author:
Sarah Pink, RMIT University College of Design and Social Context Melbourne, Melbourne, Victoria 3000, Australia.
Email: Sarah.pink@rmit.edu.au

ontologies of digital data by focusing on the technological architecture of platforms, that make data 'platform ready' (Helmond, 2015), or by interrogating how digital data is represented, conceptualised, generated and responded to (Gitelman, 2013). A number of alternative concepts have been drawn on in order to re-think what data can be and do. For instance, Boellstorff (2013) has proposed theorizing Big Data through a concept of rotted data that 'reflects how data can be transformed in parahuman, complexly material, and temporally emergent ways that do not always follow a preordained, algorithmic "recipe"'; and Lupton (2016) has conceptualized data as 'lively', whereby correspondingly, when data stops working or being of use, it might be 'dying, dead, decaying, ageing, dirty, contaminated, worn out or sick'. This work opens up a questioning of the integrity and temporality of data and calls for ongoing exploration. It also alerts us to the need to acknowledge the material, social and environmental circumstances data is part of, and to the subsequent value of interrogating how data can be understood through modes of investigation, analysis and concepts that are usually applied to the analysis of everyday materialities.

In material culture studies, an intriguing body of scholarship has placed themes and practices of breakage, decay, repair, and displacement at the centre of an analytical narrative that advances 'an approach that takes seriously the seemingly banal fact that things are constantly falling out of place' (Domínguez Rubio, 2016: 60), whereby entangled processes of making and growing require our attention (Ingold and Hallam, 2014). Notions of 'breakdown' and 'repair labour' are also part of the academic lexicon in Science and Technology Studies (STS) for speaking of technology, 'data assemblages' and the work of data scientists (Jackson, 2014; Star, 1999; Tanweer et al., 2016). What Jackson (2014) has referred to as 'broken world' theory has been engaged for understanding digital data breakdown and repair (Tanweer et al., 2016). A focus on how data is made and broken highlights the collaboration between people and infrastructures: people's lives and work become entangled with data production (Berson, 2015). Such entanglements reveal themselves when we familiarize ourselves with data worlds. This familiarization can involve collaborations with the custodians of data, 'geeks and quants' (Bell, 2015: 25), or following how data becomes appropriated into the everyday, is valued and converted into forms of value (Fiore-Gartland and Neff, 2015; Ruckenstein, 2014).

In this article, we develop this discussion further by drawing on recent critical approaches to material culture and our own ethnographic insights. While anthropological accounts have long since recognized the processuality of the lives of material culture (Appadurai, 1988), a recent interdisciplinary focus on breakage, decay and repair offers new inspiration. Here, as the geographer Caitlin Desilvey (2006) proposes, we should 'accept that the artefact is not a discrete entity but a material form bound into continual cycles of articulation and disarticulation' (p. 333), whereby decay can be seen as 'generative of a different kind of knowledge' rather than simply a process of erasure (p. 323). Resonating with this Steyerl (2009) has written of the 'poor image', as it has emerged in a digital context, whereby degraded digital images no long refer back to authentic originals but relate to their 'swarm circulation, digital dispersion, fractured and flexible temporalities'. Likewise Jonathan Sterne's (2012) history of compression whereby digital sound degradation emerges with the advance of capitalism, sets out a background where digital imperfection has become part of and appropriated within popular culture. In what follows, acknowledging these existing moves, we first examine how the broken data concept-metaphor responds to this recent focus on breakage across social science and humanities research. We then show how concepts of repair, maintenance, and growth can be engaged to understand the materiality of data, and subsequently enable us to mobilise the concept-metaphor of broken data as a device through which to interpret data in ways that are not accommodated by existing data metaphors. We advance this discussion by attending ethnographically to the persistent materiality of digital data, considering how data is also part of (or at least inextricable from) an ongoingly emerging organic world of decay and growth (Ingold and Hallam, 2014). To develop our argument, we draw on three empirical examples, each of which highlights different instances of breakage as they emerge in relation to data, relating to self tracking, data analytics and music production, sound art and glitch art.

From broken data to repair work

As noted above, critical literatures in material culture studies have emerged around questions relating to breakage; Domínguez Rubio (2016) discusses art restoration and Dant (2010) and Desilvey (2006) have focused on material repair of everyday objects. This work has developed a critical perspective on earlier material culture studies approaches for seeing objects as 'self-evident and given' in their agency (Domínguez Rubio, 2016: 60) to call for attention to how 'objects are fragile and temporal realities' and 'that objects wear down and change, that they break, malfunction and have to be constantly mended, retrofitted and repurposed, or that they are routinely misused, misrecognized and disobeyed' (Domínguez Rubio, 2016: 60). The STS scholar Stephen Jackson (2014) has focused

on technology breakages to ask 'what happens when we take erosion, breakdown, and decay, rather than novelty, growth, and progress, as our starting points in thinking through the nature, use, and effects of information technology and new media?' (p. 174). Applying this thinking to digital data analytics, Tanweer et al. (2016) have taken an Actor Network Theory (ANT) approach that sees technological breakdown as 'continual' and 'interminable' (p. 740) rather than as a dramatic event. While we concur that an STS approach to digital data and materiality offers one useful starting point to understanding digital data (Pink et al., 2016), here we respond to these propositions to ask how they are further entangled with the organic, perceptual and experiential world (Pink and Fors, 2017), based in an ethnographic understanding of how data is part of the world we inhabit.

The broken data concept-metaphor calls for further situating the discussion in relation to repair and maintenance, addressing a dissatisfaction with concepts of innovation, and the idea that technological innovation alone can drive change. Russell and Vinsel's (2016) focus on what happens 'after innovation' encourages us to consider how maintenance and repair might have more to do with data work than technological innovation. They call on us to focus on 'the maintainers', who do the hidden type of mundane repair work, behind the scenes. When we translate this idea to apply it to our relationships with data, it might include 'domestic' or individual work with personal technologies and data, could concern working in an organisation with large data sets, or be part of the process of creative work with data. Here then, repair and maintenance is situated in a specific way in relation to narratives of technological innovation and digital data, in that it is conceptualised as preparatory work that might take place before any serious data analysis can take place. The broken data concept-metaphor calls for examining the production of data by paying attention to the mundane work that precedes data breakages or follows them. Repair sustains and enables things to carry on, beyond breakages and innovation, offering insights into the everyday materialities of data work.

The materiality of data

Recent work on digital materialities argues (Pink et al., 2016: 1) that the digital and the material are 'entangled elements of the same processes, activities and intentionalities'. The focus on repair and breakage in material culture studies and STS and Jackson's (2014) focus on the materiality of technology in an ongoing broken but generative and ongoingly reconstituted world (p. 175) invites us to consider how data is similarly breakable and repairable.

Discussions of the materiality of platforms, software, algorithms and data contribute to this trajectory. Dourish (2016) calls for 'recognition of the digital as material itself' (p. 31) by discussing programming for early computers, and demonstrates how the digital and material are inextricable through the example of emulation. Here, software becomes, as Dourish (2016) describes it, a 'tool for configuring a material arrangement of the delicately entwined digital and analog components that made up the original computer system – complete with flaws, mistakes, problems, undocumented features and unexpected idiosyncrasies' which need to be re-enacted in the new materiality of the host system (p. 43). Data can be understood as similarly material and flawed/incomplete. Tanweer et al. draw on Kirschenbaum's work (2008: 9) to suggest that digital data has a forensic materiality, which refers to its physical state, for instance in being able to 'fill up a hard drive', a formal materiality relating to computational processes constituted through software (Tanweer et al., 2016: 737). Understanding data as material means that we can think of it as an open, not 'discrete' entity (Desilvey, 2006: 333) that may be broken and repaired. Tanweer et al. also make the connection between the materiality of data and the possibility of repairing it, by turning to Jackson's notion of breakage and STS derived argument (following Star, 1999) that it is through breakdown or malfunction that infrastructures and things that are invisible to us in the everyday gain a new kind of visibility (Tanweer et al., 2016: 738). Through an empirical analysis of how data scientists in their research encountered breakdown or were held up due to material limitations in their work (Tanweer et al., 2016: 737), the analysis by Tanweer et al. (2016) emphasizes how the maintenance and repair of data, and the incremental innovations this entails, are integral to quotidian work of data scientists (p. 740). They see this focus on processes of breakdown and repair as a means to advance understandings of both the practices of data scientists and theoretical accounts of Big Data and materiality (Tanweer et al., 2016: 745).

These works play a corrective role in relation to assumptions that Big Data has objectively reliable accuracy or predictive qualities by showing how 'contingent, improvised labour' required to cope with the materiality of data (Tanweer et al., 2016: 748) are inseparable from the data work and the ways that different datasets are constituted. Design and futures anthropology approaches likewise emphasise how contingency (Irving, 2017), improvisation and creativity (Ingold, 2013) underpin the ways that our present and futures emerge. They have been applied to the labour of crafting and making (Ingold, 2013), and construction industry labour (Pink et al., 2017), specifically to demonstrate the flaws in predictive modes of

anticipatory governance and regulation. Here we draw attention to the labour involved in how different makers-users of data protagonize processes through which data comes to have meaning. The implications of this approach, as we take up towards the end of this article, are also to warn us against the predictive claims of Big Data, and related tendencies to data driven design and policy.

Data metaphors

The power of concept-metaphors lies in their ability to open spaces in which continuities and discontinuities across time, space and experience can be interrogated (Moore, 2004). Metaphors provoke ideas and act as a domain within which facts, connections and relationships are imagined. In the context of data studies, data is conceived as material and metaphors of breakage have been applied to its processes. Understanding data as being part of a digital materiality thus supports the application of a 'broken world' paradigm to data studies. Likewise, data metaphors can emphasise the mobility of the data. For instance, digital data can be seen as 'liquid', when 'data deluges', 'data tsunamis' and 'data flows' are referred to. Such liquidity or flow can also become 'blocked, stuck, leaking or frozen' (Lupton, 2015; Nafus, 2014; Pink, Ruckenstein et al., 2016) and can contribute to how data becomes 'felt' as lively (Lupton, 2016) through data visualisations that grant data liveliness (Ruckenstein, 2017). Such metaphorical understandings of digital data correspond with a concept of digital materiality which rejects 'an a priori definition about what is digital and what is material' (Pink et al., 2016: 10–11). Following Ingold's point that 'things are alive because they leak' (Ingold, 2008: 10) digital data made accessible through such metaphors as lively, liquid, or broken can be seen as always emergent, incomplete, ongoing, open leaky things (Pink et al., 2017). If we see data as not only breakable, but as emergent in such a way that they and other things can mutually leak into each other, we can also consider how digital-material data might grow as intertwined and inseparable from other things and processes.

Following this line of thought, things and materials – including technology – are *in and of the world* and are co-emergent with other (organic) things, bodies and persons (both sensorially and emotionally) rather than existing separately from an environment that they impact on (Pink and Fors, 2017). In this sense the idea that 'form, rather than being applied to the material, is emergent within the field of human relations' (Ingold and Hallam, 2014: 5) may also be extended to data. This process, Ingold and Hallam (2014) argue, involves 'neither making nor growing, but a kind of making-in-growing, or growing-in-making' (p. 5), and is akin to the concept of digital materiality (Pink et al., 2016: 12). A process of making-in-growing is also one in which data can become broken, and offers us a means to examine how (at least some kinds of) digital data emerges and transforms and to situate human action within this process.

In this section, we discussed a range of related metaphors that have been used to acknowledge how data is entangled in everyday environments, activity and experience with humans, while being simultaneously intimately bound up with and regulated by digital technologies and their software. We next interrogate further the utility of such concepts in empirical research through examples of three domains of data use, drawn from different studies, of personal self-tracking, data scientists and sound as data. In doing so our objective is to highlight the ongoing development and need to further our discussions of data along such lines, across diverse fields of research about data, and to reveal the gaps that need to be met by a data ethnography research agenda, rather than to present a singular and complete ethnographic study.

Self-tracking as making/growing data

In this section, we examine the experience of making, growing and breaking personal data through Sarah Pink's research with self-trackers. Through the experience of one participant, David, we emphasize the intricacies of how such processes emerge, from a sample of 18 participants in Australia and Sweden, with whom Pink and Vaike Fors undertook research, through video/audio/photographic interviews and re-enactments and auto-ethnography. Many participants demonstrated similarly idiosyncratic data activity.

One of the first things David showed Sarah was his broken wearable. While he told her it had been materially broken for some time, it had still functioned until 'it literally stopped working about three days ago'. He was nevertheless still wearing it (Figure 1).

As they talked, Sarah recalled the feel of her own wearable on her arm, which she had likewise continued to wear for weeks although it was not working. A series of digital material breakages created gaps in her data. Sometimes she forgot or had no time to charge it, breaking her data trace. She went overseas without a mobile data plan, breaking her connection. She subsequently recharged the technology and reconnected the app. But soon, after a smartphone software update, the wearable and app did not connect. In 2013, a market research survey found that 'more than half of U.S. consumers who have owned a modern activity tracker no longer use it. A third of U.S. consumers who have owned one stopped using the device within six

Figure 1. The broken wearable.

months of receiving it' (http://endeavourpartners.net/
assets/Endeavour-Partners-Wearables-and-the-Science-
of-Human-Behavior-Change-Part-1-January-20141.
pdf). Like many contemporary self-trackers Sarah
lasted less than six months. During that time she
repaired the technologies several times, and her data
was incomplete, spread over two apps that gave her
different readings. Yet each time she re-established
the routines associated with collecting and checking
personal data, she gained a sense of familiarity and
comfort that users associate with using these apps
(Pink and Fors, 2017).

David similarly told a story of his attempts at repair:

> **David**: How did that break off?...it fell off without me
> realizing at all that it was already gone, so I couldn't
> ever find it, and then it worked for a while with me
> poking it, maybe for about a week, and now, see it
> says on here [the smartphone app]...now it says
> 'alarm, trouble connecting' and I think it's dead.
> **Sarah**: So do you think there's anything you can do
> about it?
> **David**: Well according to this there are things I can do
> about it, but I don't believe they will work, I've looked,
> I mean it says you can try to re-synch it...I've heard
> other people as well say that they just break and that
> they only last for a year

David discussed possible techniques of maintenance
through which the technology could be repaired or
restored to adequate functionality, but concluded that
buying a new wearable would enable him to (imper-
fectly) repair his data and therefore continue with his
everyday data production.

In both examples, the digital materiality of data,
software and hardware became broken, since their
functioning (and the possibilities for their future
repair) was always interdependent with and contingent
on human, bodily, sensory, emotional, environmental
and other material circumstances that were not neces-
sarily predictable or reliable. Even when these technol-
ogies were functioning, the qualities of the data
produced were contingent on other elements, which
meant their data displays did not necessarily represent
software-prescribed activity categories for the user in
question. Generally for participants the accuracy of
their data in terms of actual steps or other activity
taken was not necessarily a priority (Pink and Fors,
2017) and tended to be idiosyncratically dispersed
across different platforms and apps, inaccurate and
often incomplete; its digital materiality was ongoingly
being damaged and repaired, and was never considered
to be 'perfect', finished or complete.

Self-trackers produce data within mundane daily
routines, like commuting, training, or sleeping
(Didžiokaitė et al., 2017; Pink et al., 2017). When rou-
tines are repeated they appear similar but are never
identically reproduced; a characteristic of routines is
the ongoing ways people innovate or improvise when
undertaking them, they are sites of incremental
(and sometimes more dramatic) change. Some such
activities can be defined as making or repair, while
others can be interpreted as entailing growth, particu-
larly when they involve human bodies or other organ-
isms. Therefore, self-tracking data, which is
contingent on the human body as part of the configur-
ation through which it is produced, cannot be under-
stood as separate from processes of growth.
Moreover processes of improvisation or repair of data
can be understood as happening within and as part of
growth.

As David and Sarah continued talking the signifi-
cance of how material and emotional elements config-
ured with his data and embodied experiences were
evident:

> **David:** when I first got it [the wearable] I was annoyed
> that it didn't pick up cycling, I was like, what you mean
> I just cycled all that and...so now I've started doing
> now is putting it in my pocket when I'm cycling, my
> jeans pocket, so when I'm cycling it measures the move-
> ment...yeah so you should try it
> **Sarah:** people said I should put it on my ankle

David: ... well [his partner] put hers on her ankle but it broke because it was too stretched, yeah so hers is broken too her first one yeah, she got second one. ... if I remember I put it in my pocket, and you don't get too many steps but, ... even though I know that I'm doing it, I still want it to know.

Sarah: you can tell it you're cycling can't you

David: You can but it doesn't add to your steps, ... so like I go swimming, you can tell it you do swimming and stuff but I just think it, ... you know that they collect that data, but I don't get any reward for that ... I've stopped telling it I was going for swims, because I thought, well why should I tell you, you don't care ...

As the contingencies around David's use of his wearable unfolded Sarah learned how his data-making was improvised, whereby gaps, or damage to his data trace were filled or repaired to create meaning for him. Here repair emerged in the form of the mundane work that he did to ensure that his data was coherent, for him. Where the data trace would otherwise be broken he focused on growing/making data that would repair this trace. These processes were inseparable from his bodily activity, partly because he intended to record it, but not because he wished do so accurately or in a way that corresponded with the categories offered by the app. Such data grows with the body as it inhabits digital material environments. Data-making understood as such involves users creatively honing ways of living with digital materialities whereby they feel comfortable and accomplished, thus data grows with people as they live their lives, in processual digital material environments. However, because such data making/growing does not necessarily subscribe to categories anticipated by apps or analytics it damages the types of knowledge such technologies seek to produce. It breaks categories and assumptions about what we could know through Big Data analytics using this data. There is a politics to this, both on a personal level as David's example illustrates, and concerning the responsibilities of data makers/growers and organisations that access personal data towards creating reliable data. As Sarah and David concluded:

Sarah: How useful do you think your data is to them if you're taking it [the wearable] off, then you go swimming, and not telling it when you're cycling?

David: ... I don't care. Like, I don't feel I have any responsibility to give them accurate data, ... Yeah that's a funny idea, but it's true if everybody's using it like I am then their data might just be completely compromised.

This research endorses critiques of assumptions that digital data and Big Data analytics effectively or objectively represent past or existing, or predict future, everyday worlds or activities. It further suggests that data is continuously damaged, repaired and made/grown through everyday improvisation and contingencies, and is emergent from specific configurations of processes and things of different affordances and qualities. Indeed in addition to breakdowns in its own materiality, such digital data cannot exist separately from other things and processes that involve breakage and growth or the repair activities that ensure that when things break the gaps are filled in.

Repairing social media data

In this section, we examine the implications of broken world theory for understanding data science and data analytics by exploring the data breakages in relation to social media data. As suggested above, these breakages might derive from the 'formal materiality' of the computational processes: the data set is too big to be analyzed with existing software tools, or the data analysis takes too much time or effort to be reasonable, requiring rethinking of research methods, or questions (Tanweer et al., 2016). Ruckenstein and her colleagues have encountered such breakages in relation to data work conducted in the 'Citizen Mindscapes' initiative, an interdisciplinary open data project that contextualizes and explores a Finnish-language social media data set ('Suomi24', or Finland24 in English), consisting of tens of millions of messages and covering social media over a time span of 15 years (see Lagus et al., 2016).

The Suomi24 data was generated by a media company, Aller: the data grew in the servers as an inseparable by-product of the company's practices. As many scholars have argued, data should not be seen in isolation, but it is made 'data' in the sociomaterial structures that support data work (Berson, 2015; Gitelman, 2013). In this case, the data that resided in the company servers for over a decade gained its 'liveliness' once the company decided to open the proprietary data for research purposes. For maintaining this liveliness a new infrastructure was needed for hosting and distributing the dataset. One such data infrastructure was already in place, the Language Bank of Finland, maintained by CSC (IT Centre for Science), which has been developed for acquiring, storing, offering and maintaining linguistic resources, tools and data sets for academic researchers. The Language Bank gave material structure to the Suomi24 data: it was repurposed as research data for linguistics.

The Korp tool, developed for the analysis of data sets stored in the Language Bank, allowed word searches, in relation to individual sentences, maintaining the life of the Suomi24 data as a resource for linguistic research. Yet, the sociomaterial arrangements
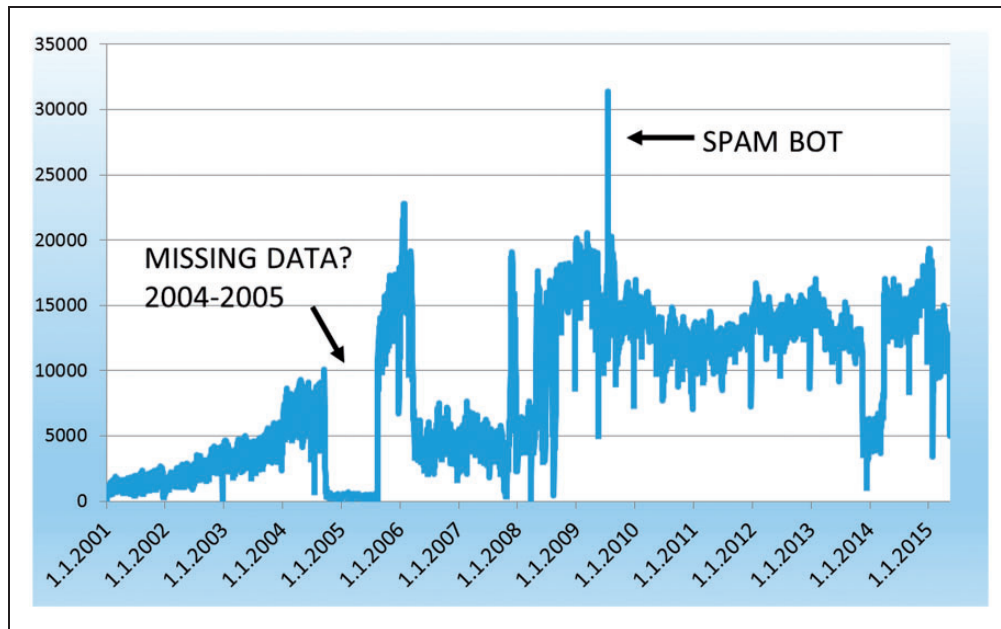
Figure 2. Identified gaps and anomalies in the Suomi24 data.

constrained other possible lives of the data – lives that were of interest to the Citizen Mindscapes research collective, aiming to work the data to accommodate the social science focus on topical patterns and emotional waves and rhythms characteristic of the social media. In the past two years, the collective, particularly those members experienced in working with large data sets, have been repairing and cleaning the data in order to make it ready for additional computational approaches.

The ongoing work has alerted us to breakages of data, raising more general questions about the origins and nature of broken data. Social media data, such as the Suomi24, is never an accurate, or complete representation of the society. From the societal perspective, the data is broken, offering discontinuous, partial, exaggerated or interrupted views to individual, social and societal aims. The preparation of data for research that takes societal brokenness seriously underlines the importance of understanding the limitations and biases in the production of the data, including insights into how the data might be broken. The first step towards this aim was a research report (Lagus et al., 2016) that evaluated and contextualized the Suomi24 data in a wide variety of ways. We paid attention to the readers and writers of the social media community as producers of the data; the moderation practices of the company were described to demonstrate how they shape the data set by excluding certain kinds of messages, for instance, advertisement, or those legally defined as advocating violence. The yearly volume and daily rhythms of the

data were calculated based on timestamps, and the topical hierarchies of the data were uncovered by attention to the platform features and conversational structures of the social media forum.

When our work identified gaps, errors and anomalies in the data, it revealed that data might be broken and discontinuous due to human or technological forces: infrastructure failures, trolling, or automated spam bots. With the information of gaps in the data (see Figure 2), we opened a conversation with the social media company's employees and learned that nobody could tell us about the 2004–2005 gap in the data. A crack in the organizational memory was revealed, reminding of the links between the temporality of data and human memory. In contrast, the anomaly in the data volume in July 2009 which we first suspected was a day when something dramatic happened that created a turmoil in the social media, turned out to be a spam bot, remembered very well in the company. Tanweer et al. (2016) emphasize the improvised labor needed to overcome challenges in working with large digital data; they introduce the obstacles and barriers that slow or derail the data science process as an important resource for knowledge production and innovation. In this case, the gaps and anomalies in the data led to significant conversations concerning the production of data, deepening our understanding of the human and material factors at play in processes of data generation.

For some researchers knowing all possible inconsistencies and breakages in the data is crucial. In the

field of statistics, for instance, research might require intimate knowledge of all possible anomalies of the data. What appears as incomplete, inconsistent and broken to some practitioners might be irrelevant for others, or a research opportunity. The role of the concept-metaphor of broken data is to open a space for discussion about these differences, maintaining them, rather than resolving them. It can highlight how data is seen as broken in different contexts and compare the breakages, then follow what happens after them, and focus on the repair and cleaning work.

In order to control breakages and unknown gaps, the Citizen Mindscape project is setting up a database to keep the data still: the data are transformed into 'frozen', or 'immutable data' that stops data growing. In other words, the aim of the database is to control the growth of data and freeze it into a digital materiality of a database. By analyzing and interpreting 'frozen data', or combining it with other data sets, the data can regain its liveliness in a manner that is both controlled – the data is known – and improvisatory.

With growing uses of secondary data, the ways in which data is broken might not be known beforehand, underlining the need to pay even more attention to brokenness and the work of repair. Given the increasing dependency in everyday life on data-driven applications, a better understanding of the consequences of data brokenness in the fields ranging from finance and insurance to health and education is called for (Ruckenstein and Schüll, 2017). Explorations that demonstrate who gives data life, and how it is kept alive and for what purposes, offer insights into data practices, paving the way for further exploration of data work and everyday data relations (Lupton, 2016; Ruckenstein, 2017). In the case of Suomi24 data, the data breakages suggest that we need to actively question data production and the diverse ways in which data is adapted for different ends by practitioners. As described above, the repurposed data requires an infrastructure, servers and cloud storage; the software and analytics tools enable certain perspectives and operations and disable others. Data is always inferred and interpreted in infrastructure and database design and by professionals, who see the data, and its possibilities, differently depending on their training. As Genevieve Bell (2015: 16) argues, the work of coding data and writing algorithms determines 'what kind of relationships there should be between data sets' and by doing so, data work promotes judgments about what data should speak to what other data. As our Citizen Mindscapes collaboration suggests, making 'data talk' to other data sets, or to interpreters of data, is permeated by moments of breakdown and repair that call for a richer understanding of everyday data practices.

Data to noise or noise to data in music production and glitch art

The example of the use of data in music and sound art offers us another way to consider how data might be broken or repaired, in ways that are contextual and part of its use. A sound file consists of data. It might be based on digitally encoded sonic variations captured by a microphone, other recording equipment, or generated by software or electronic circuits like in a synthesizer. To make a digital sound file audible it has to be decoded and output through equipment that reproduces sound, like speakers. In order for data to be experienced, it has to be transduced and turned into something that can be experienced sensorially. In anthropologist Stefan Helmreich's words transduction is 'the transmutation and conversion of signals across media that, when accomplished seamlessly, can produce a sense of effortless presence' (Helmreich, 2010: 10). The data of a sound file can be transduced and experienced as sound through speakers or headphones, but it can also be experienced as visual representations, for example waveforms on a screen.

Processes of transduction do not automatically differentiate between what is experienced as sound or as noise. Distorted parts of a sound file consist of data, and so do other parts of the file that we could experience as, for example, a voice uttering a sequence of words. What we hear as a distorted glitch becomes categorised as noise in relation to other parts of the file that we define as a meaningful and expected signal, or as appreciated content. It is only in the light of intended uses that some data might be considered broken, split, fractured, malfunctioning or noisy. Data should therefore be understood in relation to how it is intended to be processed.

The removal or filtering out of noise is an important part of data management. Noise abatement can be seen as a process of maintenance or repair. However, like all sorting and arrangement this maintenance, this filtering, has its own politics, aesthetics and peculiarities. If we assign the filtering to algorithms, someone or something has to decide how these algorithms sort out noise or smooth a data set or a data stream. When sound is recorded we get a data stream that is transduced into a file, or a file could also be generated by software. This raises questions concerning, how the file should be managed? Which parts of a sound file should be evened out, what should be removed and what should replace the removed data in order to keep the experience of the sound satisfying for a specific listener? For sound engineers or music producers there are no given and predefined general routines of how to deal with specific levels and characters of noise. Decisions on when and how to intervene in a data stream have to be made in relation to

norms, aesthetic appreciations and stylistic preferences. Imaginations about how a finalised processed sound should be affect choices during the management process. There is no predetermined definition of when data is broken or when unappreciated noise appears.

Once noise has been defined and framed, one can either filter it out or transform it into something valuable (Willim, 2014; cf. Krapp, 2011). Within the art world, and subsequently within digital culture, utilisations of the noisy and erroneous have become escape routes from predictably structured practices of creation. Within electronic music a plethora of artists started to experiment with glitches and the sounds of technological malfunction during the 1990s. Artists like *Oval*, *Pole* and *Ryoji Ikeda* created music based on sounds from skipping compact discs, broken electronic circuits and corrupted data streams. In 2000, Kim Cascone wrote about what he referred to as *The Aesthetics of Failure*:

> The 'post-digital' aesthetic was developed in part as a result of the immersive experience of working in environments suffused with digital technology: computer fans whirring, laser printers churning out documents, the sonification of user-interfaces, and the muffled noise of hard drives. But more specifically, it is from the 'failure' of digital technology that this new work has emerged: glitches, bugs, application errors, system crashes, clipping, aliasing, distortion, quantization noise, and even the noise floor of computer sound cards are the raw materials composers seek to incorporate into their music. (p. 12f)

Similar practices exist within video art, when distorted and grainy images were sought to create certain aesthetic effects. Broken media and data was also commodified and turned into products that could 'create glitches with ease'. For example, the plugin *Data Glitch* for Adobe's video software *Premiere* and *After Effects*, which was aimed at filmmakers and visual artists interested in prefabricated glitches and broken data. The plugin could be used to make glitches and to 'corrupt' moving images:

> Simulate Realistic Digital Image Glitches with Ease!
> Data Glitch is a native After Effects plugin that creates awesome realistic digital image glitches with total ease. Something you would see during a satellite transmission or a cable broadcast or from a damaged disk. Bad TV plugin is great for analog TV look, but this is 2010 and you hardly see anything that's analog anymore. This plugin simulates a realistic digital glitch effect.
> In real-life most of the glitches occur due to problems in encoding/decoding and sometimes data corruption. This plugin does exactly that. It encodes the data,

glitches the data and then decodes it similar to the real life situation. (Aeplugins 2010: http://aescripts.com/data-glitch/)

It was not surprising that this kind of software, which could create artistic effects 'with ease', met critique from artists and designers. As Menkman (2011) argued in her research on glitch art and software artists:

> There is an obvious critique here [from some artists]: to design a glitch means to domesticate it. When the glitch becomes domesticated into a desired process, controlled by a tool, or technology – essentially cultivated – it has lost the radical basis of its enchantment and becomes predictable. It is no longer a break from a flow within a technology, but instead a form of craft. For many critical artists, it is considered no longer a glitch, but a filter that consists of a preset and/or a default: what was once a glitch is now a new commodity. (p. 55)

Therefore, noise can be turned into something appreciated, and even something commodified. Broken data could be part of similar transformation or transduction processes. Glitches, noise and what is experienced as broken are relational and mutable.

The glitch art and music of recent decades had a number of predecessors, from the music of the Italian Futurist movement to the works of several experimental music composers. One composer and sound artist that might shed light on ideas about the brokenness of data is Alvin Lucier. In many of his works he has worked with how sound relates to different spaces. In 1969, he made *I am sitting in a room* (http://www.ubu.com/sound/lucier.html). The work connects ideas about sound, space, noise and human experience in an intriguing way, and shows how noise is relational and specific to contexts. While sitting in a room Lucier recorded his voice, the recording was played back through speakers in the room. The sound was recorded again, then played back and recorded in an iterative process. His voice became more and more distorted during the process, the resonances from the room and the technological mediation changed the sound. The sentences uttered by Lucier in the work describe the concept:

> I am sitting in a room different from the one you are in now. I am recording the sound of my speaking voice and I am going to play it back into the room again and again until the resonant frequencies of the room reinforce themselves so that any semblance of my speech, with perhaps the exception of rhythm, is destroyed. What you will hear, then, are the natural resonant frequencies of the room articulated by

speech. I regard this activity not so much as a demonstration of a physical fact, but more as a way to smooth out any irregularities my speech might have. (Kahn, 2009: 28)

The way the resonant frequencies of the room smoothed out Lucier's voice offers a perspective on what could be considered as noise and data. Here, there is no clear-cut demarcation between signal and noise in this context, no point at where data is obviously broken.

Discussion: Broken data and data futures

What does the study of the breakages, damage, forms of and improvisation and contingency in data making/ growth tell us about our everyday encounters with data and our possible data futures? As our three examples demonstrate, the production, analysis, reading and remaking of data involves complex processes of digital materiality, which like anything in everyday environments are relational, ongoingly changing and emergent. In each of these projects, concepts that were capable of acknowledging the brokenness and ongoingness of data were needed in order to explain the contingencies surrounding how data is used, improvised with and given meaning. Data is not necessarily accurate, complete or full aggregated representations of what individuals or societal groups have done, or able to predict what they will do. Indeed, our examples show that the metaphors of data being broken or grown might be seen as part of everyday data-making, integral to how data is experienced and prepared in data analysis, and inhabits processes of creative practice. Each example discussed has demonstrated different facets of breakage. The first shows how data is made in relation to the human body and sensory and emotional experience and its relationship to various technological materialities and software, everyday environments and socialities. If we situate personal data as part of this world, it cannot but be implicated in its processes of breakage and growth. In the second example, of data analytics, not only was the data used already potentially broken in the processes through which it was originally produced but was again broken through the techniques of cleaning, and processing required before being used for analysis. In the final example, any precise differentiation between data (as objective and clear) and noise was challenged. Processes of data breakage, repair and manipulation are integral to sound art practice, and have parallels in the examples of how data is broken in personal data and data analytics contexts. Indeed in each of these contexts, we can see how ongoing and not predetermined processes of human creativity underpin the improvisatory ways that people engage with data. Each example represents processes of making and creativity, in everyday life, workplace, and art. Moreover each example shows that even if we regard data as ongoingly being broken, it is only broken in ways that are contextual, and in relation to human intentionalities. Broken data is, as we have emphasised, a metaphor for understanding data, not an objective state.

Seeing data as crafted, made and growing in ways that are always creative, emergent and relational to the configurations of things and processes of which they are part has important implications for data uses and data policies. Our examples demonstrate ethnographically how the meanings of data are always contingent and that we cannot therefore regard data as having objectively reliable predictive qualities. While some allowances are already made for noisy data in data analytics methods, we suggest that more is needed. In order for digital data to become a part of processes of change, data practices need to be aligned with 'the generative processes of everyday life' (Pink and Fors, 2017). Studies that focus on data breakages and associated data work offer epistemological corrections and support for navigating a complex world in which policy-crafting as well as tech companies' proprietary software and data platforms have become participants alongside people in deciding how shared futures are promoted. Under these circumstances, we argue, the ways data is engaged as a technology to accompany us and guide us as we move on into unknown futures needs to be rethought.

By exploring ways in which data is broken in conjunction with other metaphors that bring it into a living and material world – such as its liveliness, growth, and decay – we can strengthen the understanding of how capacities of data technology might be harnessed to promote more responsible data futures. The approach we have introduced argues against technological solutionism, but does not deny the possible value of digital data in future making. Secondary data sources open possibilities for re-appropriating digital data and bringing it into wider ethical, political and social processes, not by protecting oneself against technological and communicative forces but by acting creatively with and within them to construct collective spaces. Our findings suggest that we need further interrogation of production, appropriation, and reappropriation of datasets, for instance by focusing on data sharing and participatory data-pooling (Delfanti and Iaconesi, 2016; Gregory and Bowker, 2016).

Summing up

In this article, we have developed and engaged the concept-metaphor of broken data in order to offer a mode

of interpreting how data is experienced, used and mobilised. In doing so we have examined the utility of concepts related to breakage and 'broken world' theories for understanding data in the present and its possible futures. As we have demonstrated a focus on breakage, repair and growth is an opportunity to learn about everyday data worlds, and to account for how these disrupt and break the linear, solutionist, and triumphant stories of Big Data. While such ideas have been posed to some extent in existing works, we have advanced this discussion by demonstrating how concepts related to breakage and that situate data as part of an emergent world can be mobilized in dialogue with ethnographic research in order to demonstrate the contingent and improvisatory modes through which data exists and the implications of this for the temporalities and imaginaries that data inhabits. More detailed ethnographic interrogation of how data is implicated in and implicates everyday life in realms from health to finance are needed. The specific implication of our discussion in this article is that particular questions that need to be followed through in future research relate to: (1) to recognise when data is broken and when and how these breakages are important for data work and generating meaning; (2) to acknowledge the processes of repair and maintenance that are part of the way data is produced, analysed and used; and (3) the need to nurture data to grow in transparent and ethical ways that are beneficial to all stakeholders.

The concept-metaphor of broken data therefore, we argue has a key role to play in future research and scholarship about and with digital data. However, as with any concept or metaphor, it needs to be situated within a suite of adjacent concepts and in relation to particular techniques of investigation that help us to comprehend the relevance of digital data. The concept-metaphor of broken data is, as we have shown in this article, usefully engaged as a critical response to uses of data that assume or imply that data presents pure, objective or complete information. However as an analytical device it is not simply critical but offers a mode through which to understand the ways that people engage with data and technologies as responses to breakage. As such this means that the concept of brokenness opens up analytical possibilities that focus on the detail of our engagements with data as it is produced or manipulated. This means that as an approach for understanding wider questions it would need to be combined with other techniques. For example a focus on broken data indeed has implications for how we consider the politics, governance and regulatory frameworks and generally power-relations of data and its use, but to comment on these more fully a focus on broken data would need to be attached to an analysis of wider discourses and processes of power as they are played out. A focus on broken data, because it necessitates attention to the often hidden detail of everyday activity is also aligned to particular research techniques, fruitfully, as shown in this article, involving ethnographic or auto-ethnographic encounters. This means that to research data through the metaphor of broken data a particular research design is required, which might combine ethnography with other techniques such as survey and interviewing methods, but that needs to rely on relatively small scale and in-depth studies in order to generate the insights needed.

As these final considerations reveal, the broken data concept-metaphor is not a single solution that will alone revolutionise how we think about Big Data. However, we argue that it is a vital element in any suite of methods and research design that seeks to understand what the actual implications of data in our present and futures might be.

Acknowledgments

Declaration of conflicting interests

Funding

References

Appadurai A (1988) *The Social Life of Things: Commodities in Cultural Perspective*. Cambridge: Cambridge University Press.

Bell G (2015) The secret life of Big Data. In: Boellstorf T and Maurer B (eds) *Data, Now Bigger and Better!*. Chicago: Prickly Paradigm Press, pp. 7–26.

Berson J (2015) *Computable Bodies: Instrumented Life and the Human Somatic Niche*. London: Bloomsbury.

Boellstorff T (2013) Making Big Data, in theory. *First Monday* 18(10). Available at: http://firstmonday.org/ojs/index.php/fm/article/view/4869/3750#56a (accessed 6th April 2017).

boyd d and Crawford K (2012) Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society* 15(5): 662–679.

Cascone K (2000) The Aesthetics of Failure: "Post-Digital" Tendencies in Contemporary Computer Music. *Computer Music Journal* 24(4): 12–18.

Dant T (2010) The work of repair: Gesture, emotion and sensual knowledge. *Sociological Research Online* 15(3): 7.

Delfanti A and Iaconesi S (2016) Open source cancer. Brain scans and the rituality of Biodigital Data sharing. In: Barney D, Coleman G, Ross C, et al. (eds) *The Participatory Condition in the Digital Age*. Minneapolis: University of Minnesota Press, pp. 123–143.

DeSilvey C (2006) Observed decay: Telling stories with mutable things. *Journal of Material Culture* 11(3): 318–338.

Didžiokaitė G, Saukko P and Greiffenhagen C (2017) The mundane experience of everyday calorie trackers: Beyond the metaphor of quantified self. *New Media & Society*, doi.org/10.1177/1461444817698478.

Domínguez Rubio F (2016) On the discrepancy between objects and things: An ecological approach. *Journal of Material Culture* 21(1): 59–86.

Dourish P (2016) Rematerializing the Platform: Emulation and the Digital Material. *Digital Materialities, Design and Anthropology*. Bloomsbury Publishing, pp. 29–44.

Fiore-Gartland B and Neff D (2015) Communication, mediation, and the expectations of data: Data valences across health and wellness communities. *International Journal of Communication* 9: 1466–1484.

Gitelman L (ed) (2013) *Raw Data is an Oxymoron*. Cambridge: MIT Press.

Gregory J and Bowker GC (2016) The data citizen, the quantified self, and personal genomics. In: Nafus D (ed.) *Quantified: Biosensing Technologies in Everyday Life*. Cambridge: MIT Press, pp. 211–222.

Helmond A (2015) The platformization of the web: Making web data platform ready. *Social Media + Society* 1(2): 2056305115603080.

Helmreich S (2010) Listening Against Soundscape. *Anthropology News*. December: 10.

Iliadis A and Russo F (2016) Critical data studies: An introduction. *Big Data & Society* 3: 1–7.

Ingold T (2008) Bringing Things to Life: Creative Entanglements in a World of Materials. *ESRC National Centre for Research Methods*, NCRM Working Paper Series, pp. 1–15.

Ingold T (2013) *Making*. Oxford: Routledge.

Ingold T and Hallam E (2014) Making and growing: An introduction. In: Hallam E and Ingold T (eds) *Making and Growing: Anthropological Studies of Organisms and Artefacts*. Farnham: Ashgate, pp. 1–24.

Irving A (2017) The art of turning left and right. In: Salazar J, Pink S, Irving A & Sjoberg J (eds) *Anthropologies and Futures*. London: Bloomsbury, pp. 23–42.

Jackson SJ (2014) Rethinking repair. In: Gillespie T, Boczkowski P and Foot K (eds) *Media Technologies: Essays on Communication, Materiality and Society*. Cambridge: MIT Press. DOI:10.7551/mitpress/9780262525374.003.0011.

Kahn D (2009) Alvin Lucier I am sitting in a room immersed and propagated. *OASE Journal for Architecture* 78: 24–37.

Kitchin R and McArdle G (2016) What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data and Society* 3: 1–10.

Krapp P (2011) *Noise Channels: Glitch and Error in Digital Culture*. Minneapolis: University of Minnesota Press.

Lagus KM, Pantzar M, Ruckenstein M, et al. (2016) *Suomi24: Muodonantoa aineistolle*. Helsinki: Faculty of Social Sciences, University of Helsinki.

Lupton D (2015) *Digital Sociology*. London: Routledge.

Lupton D (2016) *The Quantified Self: A Sociology of Self-Tracking*. Cambridge: Polity Press.

Manovich L (2013) *Software Takes Command*. New York: Bloomsbury.

Markham A (2013) Undermining 'data': A critical examination of a core term in scientific inquiry. *First Monday* 18(10). DOI: 10.5210/fm.v18i10.4868; http://firstmonday.org/ojs/index.php/fm/article/view/4868/3749 (accessed 28 December 2017).

Menkman R (2011) *The Glitch Moment(Um)*. *Network Notebooks*. Amsterdam: Institute of Network Cultures.

Moore H (1999) Anthropological theory at the turn of the century. In: Moore H (ed.) *Anthropological Theory Today*. Cambridge: Polity Press, pp. 1–23.

Moore HL (2004) Global anxieties: Concept-metaphors and pre-theoretical commitments in anthropology. *Anthropological Theory* 4(1): 71–88.

Nafus D (2014) Stuck data, dead data, and disloyal data: The stops and starts in making numbers into social practices. *Distinktion: Scandinavian Journal of Social Theory* 15: 208–222.

Pink S, Ardevol E and Lanzeni D (2016) Digital materiality: Configuring a field of anthropology/design? In: Pink S, Ardevol E and Lanzeni D (eds) *Digital Materialities: Anthropology and Design*. Oxford: Bloomsbury, pp. 1–26.

Pink S and Fors V (2017) Being in a mediated world: Self-tracking and the mind-body-environment. *Cultural Geographies* 24(3): 375–388.

Pink S, Ruckenstein M, Willim R, et al. (2016) DATA ETHNOGRAPHIES (5): Broken Data. Available at: https://dataethnographies.com/paper-v-broken-data/.

Pink S, Sumartojo S, Lupton D, et al. (2017) Mundane data: The routines, contingencies and accomplishments of digital living. *Big Data and Society*. doi.org/10.1177/2053951717700924.

Ruckenstein M (2014) Visualized and interacted life: Personal analytics and engagements with data doubles. *Societies* 4(1): 68–84.

Ruckenstein M (2017) Keeping data alive: Talking DTC genetic testing. *Information, Communication & Society* 20(1): 1024–1039.

Ruckenstein M and Schüll ND (2017) The datafication of health. *Annual Review of Anthropology* 46: 261–278.

Russell A and Vinsel L (2016) Hail the maintainers. *Aeon*. Available at: https://aeon.co/essays/innovation-is-overvalued-maintenance-often-matters-more (accessed 12 October 2017).

Star SL (1999) The ethnography of infrastructure. *American Behavioral Scientist* 43(3): 377–391. doi:10.1177/00027649921955326.

Sterne J (2012) *MP3: The Meaning of a Format*. Durham: Duke University Press.

Steyerl H (2009) In defense of the poor image. Available at: www.e-flux.com/journal/10/61362/in-defense-of-the-poor-image/ (accessed 23 October 2017).

Tanweer A, Fiore-Gartland B and Aragon C (2016) Impediment to insight to innovation: understanding data assemblages through the breakdownrepair process. *Information, Communication & Society* 19(6): 736–752.

Willim R (2014) Transmutations of noise. In: Czarniawska B and Löfgren O (eds) *Coping with Excess. How Organizations, Communities and Individuals Manage Overflows*. Cheltenham: Edward Elgar, pp. 118–132.