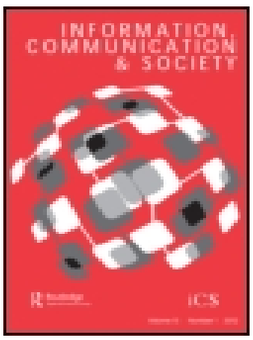

Routledge
Taylor & Francis Group

Information, Communication & Society

ISSN: 1369-118X (Print) 1468-4462 (Online) Journal homepage: http://www.tandfonline.com/loi/rics20

Impediment to insight to innovation: understanding data assemblages through the breakdown–repair process

Anissa Tanweer, Brittany Fiore-Gartland & Cecilia Aragon

To cite this article: Anissa Tanweer, Brittany Fiore-Gartland & Cecilia Aragon (2016): Impediment to insight to innovation: understanding data assemblages through the breakdown–repair process, Information, Communication & Society, DOI: 10.1080/1369118X.2016.1153125

To link to this article: http://dx.doi.org/10.1080/1369118X.2016.1153125

INFORMATION, COMMUNICATION & SOCIETY, 2016
http://dx.doi.org/10.1080/1369118X.2016.1153125

Routledge
Taylor & Francis Group

Impediment to insight to innovation: understanding data assemblages through the breakdown–repair process

Anissa Tanweer[a], Brittany Fiore-Gartland[b] and Cecilia Aragon[b]

[a]Department of Communication, University of Washington, Seattle, WA, USA; [b]Human Centered Design and Engineering, University of Washington, Seattle, WA, USA

ABSTRACT

As the era of 'big data' unfolds, researchers are increasingly engaging with large, complex data sets compiled from heterogeneous sources and distributed across networked technologies. The nature of these data sets makes it difficult to grasp and manipulate their materiality. We argue that moments of breakdown – points at which progress is stopped due to a material limitation – provide opportunities for researchers to develop new imaginations and configurations of their data sets' materiality, and serve as underappreciated resources for knowledge production. In our ethnographic study of data-intensive research in an academic setting, we emphasize the layers of repair work required to address breakdown, and highlight incremental innovations that stem from this work. We suggest that a focus on the breakdown–repair process can facilitate nuanced understandings of the relationships and labour involved in constituting data assemblages and constructing knowledge from them.

ARTICLE HISTORY
Received 1 September 2015
Accepted 8 February 2016

KEYWORDS
Big data; ethnography; data science; materiality; breakdown; repair

Introduction

As the era of 'big data' unfolds, researchers across myriad disciplines are increasingly engaging with large, complex data sets distributed across networked technologies (Kitchin, 2014). This emerging data-intensive mode of inquiry has been called the 'fourth paradigm' of scientific exploration (Hey, Tansley, & Tolle, 2009), inspiring research areas termed 'e-science' and more recently 'data science,' that call for the development of new knowledge infrastructures supporting new types of software ecologies (Borgman, 2007, 2015; Edwards et al., 2013). The pursuit of data-intensive scientific discovery in academia means that a wide array of digital technologies and networked infrastructures that produce, process, manage, store, and analyse data have become inextricable from the everyday work practices of a growing number of scholars.

What counts as a 'big' data set is relative to the available methods, tools, and practices for collecting, processing, and analysing it, and one of the primary challenges for researchers working with big data is that that its materiality can be partially obscured by its relative volume, variety, or velocity. Materiality is a term that has been conceptualized in a number

CONTACT Anissa Tanweer tanweer@uw.edu Department of Communication, University of Washington, Seattle, WA, USA

of ways in the humanities and social sciences (Sterne, 2014), ranging from observable consequences of intangible ideas to properties of physical objects. The way we approach materiality here is closer to the latter sense; essentially, we think of it as 'thingness' – or as Leah Lievrouw has put it, 'the physical character and existence of objects and artefacts that makes them useful and usable for certain purposes under particular conditions' (2014, p. 25). Materiality in this sense is often excluded from imaginations of digital data, a view that persists because 'a digital environment is an abstract projection supported and sustained by its capacity to propagate the illusion (or call it a working model) of immaterial behaviour' (Kirschenbaum, 2008, p. 11).

Kirschenbaum argues that the materiality of digital data has two forms: forensic materiality and formal materiality. Forensic materiality is the physical manipulation of matter for the inscribing of digital information; if it did not exist, he argues, data could not fill up a hard drive. Formal materiality is unique to digital data. Kirschenbaum (2008) characterizes it as 'the imposition of multiple relational computational states on a data set or digital object' (p. 12), a 'computationally specific phenomenon' (p. 9) involving 'the simulation or modelling of materiality via programmed software processes' (p. 9). It is this symbolic representation that is particularly difficult to comprehend with big data. The formal materiality of an Excel spreadsheet can be easily grasped and manipulated; one can open the document and scroll through its rows and columns, copy and paste, edit and delete. But when working with 90 terabytes of text or a trillion rows of data, as our research participants were doing, engaging with its formal materiality is a challenge.

Difficult to envision or engage with in their entirety, vast and distributed data sets demand new strategies for knowing, seeing, and communicating with data. We argue that encounters with breakdown – conceived of as points at which progress is stopped due to a material limitation – can lead to important insights into the materiality of digital data, representing essential sites of knowledge production for data science and any other big data analysis, and leading to incremental innovation. Our study highlights the quotidian nature of breakdown, as well as forms of repair labour that surface otherwise obscured relationships across the data assemblage.

Based on observations at a large university in a programme we identify by the pseudonym, the Data Science Collaboration (DSC), we outline a process by which researchers move from initially encountering breakdown, to generating insights about the materiality of their data, to enacting innovative computational strategies for repairing the breakdown. Our formulation of this process emerges through the study of three examples of researchers dealing with breakdown in their projects, each stemming from encounters with different kinds of material limitations. The articulation of the breakdown–repair process foregrounds the work that the researchers must do to move between each stage in the process. In doing this, we engage with and build upon several concepts articulated in previous literature. First, our work demonstrates that the process of seeing and knowing data is a representational act inextricable from their sociotechnical contexts of production. Second, we show how breakdown is an occasion for revealing the assemblage of human and non-human actors involved in a sociotechnical system. Third, we contribute to a growing focus on the role of quotidian articulation and repair work in producing those insights and leveraging them towards innovation.

Sociotechnical contexts of data

Big data are produced and used within particular sociotechnical contexts that shape expectations for what data can do, and influence interactions between people, data, and tools in important ways. Rob Kitchin refers to the 'assemblages' in which data are enmeshed: 'amalgams of systems of thought, forms of knowledge, finance, political economies, governmentalities and legalities, materialities and infrastructures, practices, organisations and institutions, subjectivities and communities, places, and marketplaces' (2014, p. 20). We adopt this perspective in thinking about the way things, people, and institutions are organized around data. One relevant point to be made about the 'co-determinous and mutually constituted' nature of data and their assemblages (Kitchin, p. 47) is that digital data are inextricable from the computational tools used to view and structure them. In their study of brain scans, for example, de Rijcke and Beaulieu show that scans are both digital and networked images that are 'increasingly part of suites of networked technologies rather than stand-alone outputs' (de Rijcke & Beaulieu, 2014, p. 132). This view foregrounds the mutual dependence of these different components, and the role of reconfigured instrumentality in shaping how they come to know their data. For these reasons, Carusi and Hoel call for 'a new ontological approach to technologically mediated vision, which takes into account the reciprocal and co-constitutive relations between vision, technologies, and objects' (Carusi & Hoel, 2014, pp. 215–216).

Another point that must be understood about the sociotechnical contexts in which data are embedded is that insights generated from data are neither reified truths nor are they self-evident, but rather are constructed and interpreted. Implicit in scientific data sets are processes of counting that can be understood as 'epistemic achievements that involve categorical judgments' (Martin & Lynch, 2009, p. 246). Every disciplinary institution and body of knowledge has 'its own norms and standards for the imagination of data,' and 'different data sets harbor the interpretive structures of their own imagining' (Gitelman & Jackson, 2013, p. 3). Often the big data that researchers work with are not just voluminous, they have been produced under varied circumstances or assembled from varied sources, which means that multiple contexts of production are implicated within a data set. The values and expectations across different contexts of data production and use are part of what constitutes what data mean and what data may do (Fiore-Gartland & Neff, 2015; Vertesi & Dourish, 2011).

Breakdown as revelatory disruption

Exploiting moments of breakdown as a theoretical and methodological probe into understanding otherwise invisible relationships among people and things is a common strategy within science and technology studies. Susan Leigh Star has demonstrated the ways in which obscured information infrastructure 'becomes visible upon breakdown' (Star, 1999, p. 382). Employing breakdown to make things visible is what Bowker would call an 'infrastructural inversion' – foregrounding the truly backstage elements of work practice, the boring things (Bowker, 1994).

An Actor Network Theory lens illuminates how technologies can be seen as black boxed, 'silent' intermediaries until occasions such as breakdown reveal them to be

mediators acting in visible and complex associations with one another (Latour, 2005), a phenomenon that Latour has referred to as depunctualization (Latour, 1999). Paying attention to moments when those technologies break down facilitates efforts 'to *make them talk*, that is, to offer descriptions of themselves, to produce *scripts* of what they are making others – human or non-human – do' (Latour, 2005, p. 79). This approach is easy enough to imagine in the example of a car breaking down, a moment that reveals the multitude of parts and interdependencies under the hood. But this 'opening the black box' (Latour, 1999) becomes more difficult to imagine when working with large scientific data sets that may be too massive to visualize in their entirety, too complex to see the parts and their relations at once, or too messy to readily ascertain relationships between their parts. In other words, the data do not necessarily 'offer descriptions of themselves' (Latour, 2005). What happens when breakdown does not reveal all the associations and relationships that make a difference in working with large data sets or the elements of the data assemblage? Breakdown in this context becomes more of an invitation for further investigation. In fact, our study demonstrates the tremendous amount of work that needs to be done to decipher the various components and relationships in the data that are 'revealed' through breakdown. Therefore, in this paper, we talk about depunctualization of 'black box' technologies not in terms of a passive phenomenon, but in terms of 'depunctualization *work*' that must be done to decipher those relationships. We also show how, once depunctualization work has led researchers to new understandings of their data, they must do more work to figure out how to leverage that insight to address the breakdown. The concept of articulation work (Gerson & Star, 1986) captures the kind of labour that this step entails:

> Articulation consists of all tasks involved in assembling, scheduling, monitoring and coordinating all of the steps necessary to complete a production task. This means carrying through a course of action despite local contingencies, unanticipated glitches, incommensurable opinions and beliefs or inadequate knowledge of local circumstances. (Gerson & Star, 1986, p. 266)

In the context of data science practice, these 'real-time adjustments' (Star, 1999) are part of a process of situated sense-making that responds iteratively to the ever-present challenges of making data flow in the appropriate form to the appropriate places at the appropriate times.

Breakdown as quotidian reality

In the previously discussed perspectives on infrastructure from scholars such as Bowker, Star, and Latour, breakdown is portrayed as a relatively uncommon, catastrophic failure that interrupts what is, under normal circumstances, a cohesively functioning entity – disruption that can sometimes lead to innovative leaps in design and practice (Petroski, 1985). But more recently, a subset of scholars are working towards reconceptualizing the place of breakdown in technology studies (see Graham & Thrift, 2007; Jackson, 2013). They draw attention to the idea that 'the world is always breaking' (Jackson, 2013, p. 223) and requires 'continuous efforts of repair and maintenance' (Graham & Thrift, 2007, p. 10). Jackson (2013) has proposed the work of repair as a 'facet or form of articulation work' (p. 223) and as a fruitful device for supporting 'broken world thinking.' He defines repair work as

> the subtle acts of care by which order and meaning in complex sociotechnical systems are maintained and transformed, human value is preserved and extended, and the complicated work of fitting to the varied circumstances of organizations, systems, and lives is accomplished. (Jackson, 2013, p. 222)

A focus on the quotidian nature of repair as a form of articulation work allows us to view incremental innovations that stem from that work:

> […] when things break down, new solutions may be invented. Indeed, there is some evidence to suggest that this kind of piece-by-piece adaptation is a leading cause of innovation, acting as a continuous feedback loop of experimentation which, through many small increments in practical knowledge, can produce large changes. (Graham & Thrift, 2007, p. 5)

The validity of re-centring the analytical lens on repair has been demonstrated through studies of e-waste recycling and cell phone repair ecologies in developing world settings (Burrell, 2012), where the expectation for breakdown is heightened, and breakdown becomes a site for renewed cultural and economic activity. These essential components of the technology landscape are often hidden from perspectives focused on moments of technology design and production, eliding their potential as sites for innovation.

Our study contributes to scholarly work framing technological breakdown as a continual and interminable phenomenon by doing two things: First, it demonstrates the quotidian work required to maintain and repair data, and second, it highlights the incremental nature of innovation that arises from this quotidian work. We adopt Jackson's 'broken world thinking' (2013) by taking breakdown as the starting point in thinking through the nature, use, and impact of technology. Just as Jackson considers broken world thinking to be 'both empirical and methodological' (p. 221), we employ breakdown as both an object of study, and as a lens for investigating data science practices in a manner that foregrounds the elusive materiality of big data.

Study design

This research is part of an ongoing ethnographic study of data science communities and collaborations in academia. Two of the authors embedded ourselves within a Data Science Collaboration (DSC) programme that takes place annually at a large public university. The DSC matches data science methodology experts with domain researchers from a range of disciplines – for example, astronomy, biology, and political science – to collaborate on data science projects throughout an academic term. For this period of time, the data science methodology experts serve as mentors to the domain researchers. A central feature of the programme is the co-location of the mentors and researchers two days per week in an effort to advance collaboration and productivity in a short period of time. Our analysis presented here is primarily based on observation within the space of co-location, where two ethnographers spent a total of approximately 50 hours conducting observations over the course of the ten-week academic term in which the DSC took place. The communication between collaborators in this setting made it an ideal site for observing the process that unfolds after researchers experience points of breakdown in their work. The six domain researchers and four mentors were constantly working side-by-side, discussing problems as they arose, and talking through solutions together. This field site allowed us to observe a rich set of interactions around data-intensive research, providing

invaluable insights that would have been difficult to obtain by watching researchers hack away on their computers in isolation, examining their digital traces, or merely asking them to retrospectively comment on the problem-solving process.

Our in-person observations were supplemented with archival analysis of project documentation and communication that occurred online. We also collected data in a series of 10 semi-structured interviews with every data science mentor and domain researcher participating in the DSC during the term in which we conducted observations. Additionally, after the DSC had ended and we began analysing our data, we conducted another round of five semi-structured interviews in order to validate our findings and further develop the analysis we present here. These interviews were conducted with three of the original DSC participants, as well as two other data scientists not involved with that iteration of the DSC, in order to ensure that our analysis had resonance outside the DSC setting. We asked former participants to update us on their research, and asked both groups to provide feedback on our preliminary categorization of breakdowns, insights that stem from those breakdowns, the strategies they use to repair breakdown, as well as the work they do to move between each of those stages in their research.

We used a modified grounded theory approach (Charmaz, 2014) in analysing our data, which began with open coding and constant comparison of our field notes and interview transcripts. This process allowed us to identify breakdown as an important theme across our data, and led us to more selectively code those situations to explore what encounters with breakdown could reveal. As we developed our preliminary analysis, we turned to literature on breakdown and repair to situate our findings, and then returned to our community of study to validate and further develop our analysis.

Findings

We did not enter the field intending to investigate breakdown in the course of data science work, but we quickly noticed how often things did not go according to plan for the participants in our study. The domain researchers identified problems and challenges they encountered day to day in the course of their research, and in some instances were encouraged to blog about and report back to the group on any 'blockers' they experienced, a term borrowed from Agile software development methods to mean anything that is blocking one's progress.[1]

These blockers are encountered so frequently that our respondents often describe data science as inherently being an exercise in problem solving. As an oceanographer named Rachel put it when asked about blockers in her work,

> that's the one constant. That's the one thing you can always count on happening … . it's very stop and start all the time. But there's a range of different issues from just really dumb technical stuff to more involved having to step back, and stop, and learn something new.

With this analysis, we are not attempting to exhaustively capture and explain the entire range of blockers Rachel mentions, but we *are* interested in those times she talks about 'having to step back, and stop, and learn something new.' Time and again, we saw people encounter problems that could only be resolved if they took the time to learn something new about their data set; for example, what it contained, how it was structured, or what kinds of dependencies existed between different elements of their data. We realized

that those material relationships were in part obscured by the fact that their digital data were so voluminous, variable, and complex. Furthermore, many of these relationships were baked into the computational tools for viewing, processing, and analysing the data, and required work to parse, extract, and interpret. And we saw that many times, moments of breakdown provided the occasion for the researchers to reimagine their data … again, to 'step back, and stop, and learn something new,' about it. These are the instances of breakdown that we explore in this paper.

Jackson refers to breakdown in two senses: as the inevitable decay of systems under the inescapable law of entropy, and as points of breakage resulting from 'bumping up against the limits of existing protocols and practices' (2013, p. 228). In the context of our fieldwork in the DSC, we understand breakdown in the latter sense, as *a point at which a material limitation prevents the use of current protocols and practices as expected*. The relevant protocols and practices we're discussing here – the ones that establish the limits the researchers bump up against – are their own research designs and plans. These researchers entered into their big data projects with a set of expectations for what their data sets contained, how much computational power and space they had at their disposal, and how much time they could reasonably take to execute their projects. In this paper, we discuss moments of breakdown in which researchers bumped up against those material limits, and the process that ensued. We present three cases from our ethnographic study that detail different kinds of breakdown and subsequent processes of repair.

Case 1: Breakdown through anomaly

Rachel is an oceanographer, and in the DSC, she was working with a data set compiled and synthesized from dozens of distinct oceanic expeditions. Rachel talks about two different kinds of anomalies researchers encounter when doing data science – 'the obvious ones which will break your script,' and 'the not obvious ones which will break your results.' Detecting the latter anomaly involves researchers comparing what they find against their informed expectations of what they *should* find, and is the sort of anomaly that plagued Rachel throughout the early stages of her project:

> Just to know that I would have to standardize the data – I didn't even realize at the beginning that that was a thing. And then I'd look at results and say 'hey, this is weird, why does everything suddenly shift at this point? Oh yeah, because of this. Oh, that's a problem.' So it was actually quite a painful and long startup period in a sense, where I just felt like I was constantly finding out things about the data that I'd wish I'd just known all in one go at the beginning.

In one instance, she noticed a very abrupt and sudden jump in her preliminary results. This anomaly indicated to her that *something* was wrong with her data, but did not in and of itself indicate *what* was wrong with her data. Before she could generate insight into the informational content of her data, she had to take a closer look at it. Rachel talks of the need to 'zoom' in and out of her data by looking at it at different scale – zooming out to see patterns in the data, zooming in to sort and sift and find the particularities that generate those patterns. In this case, she got closer to her data by sorting and sifting through the measurements collected on particular cruise expeditions. This depunctualization work of investigating the data content led her to realize that one cruise's measurements were drastically different from the others. She now understood something

new about the content of her data, but this new insight into variance across her data set was not enough to figure out how to repair the problem. First she had to figure out why the data was so anomalous. When Rachel tracked down another researcher who had been involved with the data collection for the cruise in question and asked why the measurements were so different from the others, it turned out that the instrument sensitivity had been adjusted on that cruise. 'There's a lot of that,' says Rachel. 'Struggling through, using other people's data, not having all of the information about where it came from, what happened, [when] there was some glitch.'

Once this step of articulation work had revealed important contextual information about the data's provenance, Rachel could work on developing a strategy for repairing the data and pushing past the point of breakdown. In the instance we've been describing, this meant calibrating the measurements that were collected using one level of sensitivity with the measurements that were collected using a different level of sensitivity. But Rachel also recognized that this sort of inconsistency in the data was something that could very well recur given the way data was being collected, so she initiated an innovative process of repair for avoiding future breakdowns. Rachel notes that in a traditional wet lab environment, things like the adjustment of the instrument's sensitivity would have been meticulously recorded in a lab book. 'If you were doing an experiment in the lab, you would just write a note [that said, for example], 'Oh I spilled something on my test tube,'' says Rachel. Looking for an analogous process in the collection of digital data, she worked with personnel in charge of instrumentation to make sure that changes to the instrument settings on cruises would be automatically recorded and time-stamped as the data are collected.

Case 2: Breakdown through size

Another case of breakdown occurred for Louis, an economist working on developing a counterfactual predictive pricing model in a particular commercial market sector. At the outset of his project, Louis' data set did not take up very much space on a disk, and he was planning on keeping his entire project contained on his laptop computer. Yet one day as he tried to run just one percent of his data in a statistical software package in the programming environment R, he turned to his data science mentor, Zach, and said, 'It's growing, it's about to die on 60 gigs!' as he maxed out his computer's RAM[2] capacity. 'That seems a little excessive,' said Zach. 'I'm all for giving machines more RAM, but if one [%] is taking 60 gigs, we might want to figure something else out.'

Just as in Rachel's case, after encountering this breakdown, Louis and Zach peered more closely into Louis' data to figure out what it contained, looking through specific fields in the data set to understand how it could be taking up so much memory. Upon looking through individual entries, Zach saw that Louis' data contained a lot of zeroes, and that the statistical package R was, by default, representing the data as a dense matrix in which zeroes are computed as a value instead of being skipped over as empty placeholders. Zach knew that data sets with a lot of zeroes could be compressed and represented as sparse matrices in which the zero values are excluded rather than computed in the same way that other values are processed. However, R was not designed to compute statistics on sparse matrices. Zach then set about figuring out how other people have worked around this problem, searching online for what Louis characterized as 'ad hoc solutions' that others had developed when facing similar issues.

Zach and Louis accumulated those ad hoc solutions into a single task that would allow R to run statistical analysis on a sparse matrix. This compression strategy was a novel innovation for the software package and significantly reduced RAM consumed by Louis's project.

Case 3: Breakdown through time

The third case details a project belonging to an astronomer named Sam, in which time was the primary constraint that led to breakdown. Sam was analysing images from a telescopic sky survey, a project that involved processing data from billions of pixels. Although in many of the other examples we observed, breakdown was encountered as an unexpected occurrence over the course of the DSC proceedings, in Sam's case, he began the DSC knowing that the amount of time it would take to execute his project rendered it intractable. He had not nailed down an accurate estimate of just how long it would take, but he knew that if he used the tools and techniques and database designs he was already familiar with, it would take many, many years. At one point he lamented that, 'it would take 100 years to actually do this.'

Whereas Rachel and Louis first had to sift through and peer into their data sets more closely to get a better understanding of what information and representations comprised them, Sam needed to figure out which elements of his data were the slowest to ingest into the database, and what aspects of his database structure were making his queries inefficient. But like Rachel and Louis, this also involved looking at his data on a different scale. Sam tried separating his data into more granular chunks, a process he referred to as 'atomization', to generate a different view of his data set and the relationships among its elements. Atomizing the ingest process meant uploading different components of his data separately in order to see how long each of them were taking. He also then ran test queries on a very small sub-sample of one-four-thousandth of his data to test its performance. From doing this depunctualization work, Sam realized that the pixel values were 'the largest bottleneck' in his work, as they were taking the longest time to ingest and they were slowing down his query time.

An important part of Sam's articulation work entailed consultation with database experts in the DSC who served as his mentors. In order to solve a problem like this, 'you have to be smarter, cleverer,' said Sam.

> That's where [the DSC mentors] really come into play, who have dealt with a lot of these database-type things before … . I've had really super useful discussions with [them] about details of databases, and building indices in databases, and things I had never thought about.

Based on the advice he received from his mentors in the DSC, Sam began redesigning his database representation and experimenting with various ways of optimizing his queries.

Sam and his mentors developed several strategies for optimizing the structural relationships of his database in order to drastically reduce the amount of time his project would take. First, they constructed a new representation of the database schema that excluded the onerous pixel data. By extracting the metadata they needed from the pixels and developing a plan to access the pixel data directly from their original file rather than including it in the database, they 'trimmed the fat' or eliminated the redundancy from their database and

ended up with a leaner, faster design. Second, they built a set of indices that would quickly direct a query to the appropriate values across the breadth of the database. Third, they rewrote their queries according to a logic that would run more efficiently on the new database structure.

Together these strategies reduced Sam's query time by more than a thousand fold, which was significant progress to make in a single academic term. The overall strategy still did not address the length of time it would eventually take to ingest the pixel data, however, and because they had eliminated the pixel values from the database, at some point they would also have to work on optimizing the process of ingesting the pixels from their original file. Still, Sam made significant, if incremental, progress in overcoming the limitation of time, and through this process gained insight into how the structural relationships of his database mattered for how he could work with his data.

Summary of findings

Distinguishable from other blockers encountered by researchers in the course of practicing data science (such as server crashes and bad code), each of these cases of breakdown stems from the formal materiality of the data set itself. In the cases that we describe here, the blocker is brought to the researchers' attention by an event indicating that they had reached a material limit established by their research protocols and practice. Because our participants could not visualize and understand the formal materiality of their data set in its entirety, these indicators of breakdown were important hooks and entrées into their data, providing occasion for researchers to learn about and manipulate the formal materiality of their data in order to work towards a resolution of the blocker.

These cases detail three different indicators of breakdown experienced by our participants: anomaly, size, and time. When researchers came across unexpected inconsistencies in their data that bumped against the limits of what they expected their data to contain, as Rachel did in the first case, we categorized this indicator of breakdown as an 'anomaly'. When we saw researchers such as Louis bump up against limits to the space and memory of the hardware that houses or processes their data, we refer to the indicator of this breakdown as 'size.' When we observed researchers unable to move forward because the duration required to execute a task computationally would exceed what the researcher considered to be tractable, we refer to the indicator of this breakdown as 'time.'

We summarize our findings in Figure 1 by bringing together the three cases discussed in this article, each representing parallel processes of breakdown and repair initiated by different indicators of breakdown. The summary chart outlines the indicator of breakdown, the type of depunctualization work the researcher did to understand that breakdown, the nature of the insight into the material form of the data, the type of articulation work the researcher did to leverage that insight to solve the problem, and finally, the techniques the researcher used to repair the breakdown. In the following discussion, we synthesize these cases into a typical process of repairing breakdown that stems from the formal materiality of digital data, and discuss how mapping breakdown–repair processes contributes to a richer understanding of data science practice and theories of materiality and big data.

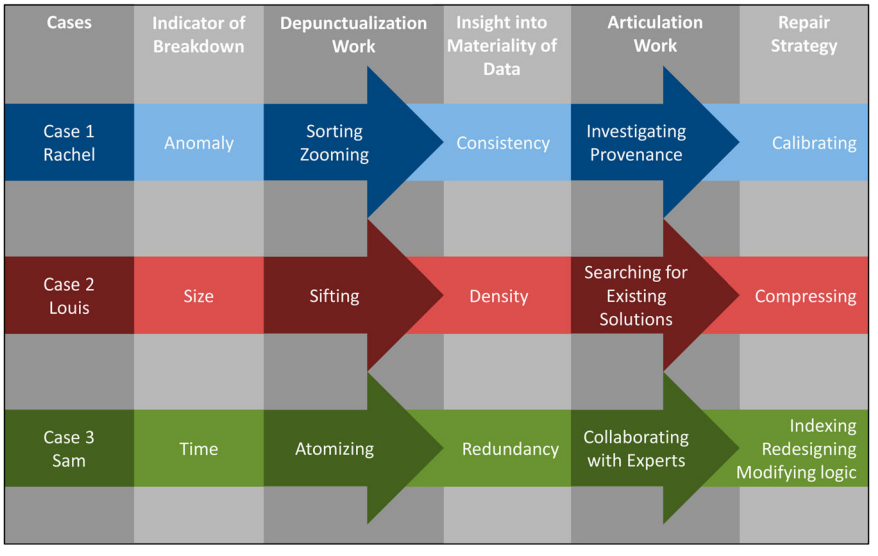

Figure 1. Summary of findings.

Discussion

Describing a typical breakdown–repair processes

Based on the cases of Rachel, Louis, Sam, and others, we describe a typical process by which data scientists move from breakdown to repair when that breakdown originates from limitations related to the formal materiality of their data. First, researchers have an initial encounter with an indicator of breakdown. This encounter opens the black box, so to speak, and partially reveals the interdependent parts contained within, in a process of depunctualization (Latour, 1999). However, we find the relationships between these revealed parts are not necessarily self-evident, and often require proactive investigation on the part of the data scientist in order to get a different perspective on their data set. This often means looking at their data on different scales, which can include zooming in by combing or sifting through individual data points, zooming out by creating visualizations that reveal broad patterns in their data, separating and organizing data into more granular chunks (a process our participants referred to as atomization) and sub-sampling, which refers to when a representative sample is taken from a larger sample of data.

This depunctualization work helps the researchers identify the source of breakdown and generate new insights by re-envisioning the material form of their data. With this new insight in mind, they engage in articulation work in order figure out what this new insight means, why the data are the way they are, and what can be done to stitch the constituent elements of the data assemblage together into a functioning whole. This often involves discussions with individuals who collected the data, searches for solutions used by others in response to similar problems, and dialogue with technical experts and other researchers on the conceptualization of research design. This articulation work helps them develop a strategy for innovative computational repair in which researchers

write or manipulate code to get past the point of breakdown and, in some cases, introduce sociotechnical innovations to prevent further encounters with breakdown.

In the DSC, when it was anomaly that stopped researchers in their tracks, that breakdown tended to yield insight into the informational content of the data with regard to its consistency, relevance, and accuracy. This in turn led to wrangling as a repair strategy (Kandel, Paepcke, Hellerstein, & Heer, 2012), which included techniques such as filtering, matching, and calibrating data. In our observations, breakdowns indicated by size yielded insight into the structural representation of the data in terms of its density and complexity. This insight led to re-representation as a repair strategy and techniques such as compression, which reduces the number of bits required to represent data, and dimensionality reduction, which refers to the elimination of certain variables. Breakdowns indicated by time yielded insights into the structural relationships between different elements of the data in terms of their redundancy and dependency upon one another. This led to optimization as a repair strategy, which incorporated techniques such as: parallelizing tasks to simultaneously run on multiple, distributed processing cores; indexing, or creating tables to quickly direct computational tasks to the location of data; caching, or creating a temporary location for storing information; redesigning the structural representation of the database; and modifying query logic to more efficiently access relevant data.

We found wrangling, re-representation, and optimization to be common computational strategies for overcoming breakdowns stemming from the material forms and consequences of working with big data. These strategies emerging from our cases are not intended to represent an exhaustive list of computational repair strategies involved in doing data science generally, but this preliminary typology helps us organize and attend to a particular type of breakdown associated with data science practice that stems from challenges in comprehending and manipulating the formal materiality of big data sets and the limitations of standard protocols and practices (Figure 2).

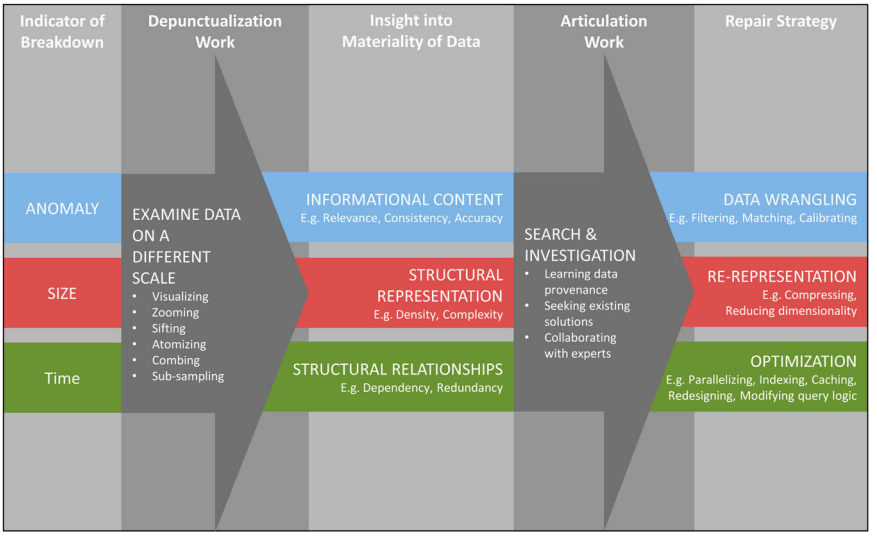

Figure 2. Typical breakdown–repair process.

Understanding data assemblages through the breakdown–repair process

Tracing this breakdown–repair process in the DSC does more than just delineate particular micro-processes involved in data science work; it serves as an analytical tool for surfacing insights into the labour, innovation, and lifecycle of data assemblages. For one thing, patterns in labour practices and relationships emerge through a mapping of the breakdown–repair process. Data science is often lauded for its development of sophisticated analytical algorithms, but our study illustrates that before analysis can be conducted on vast data sets, much of the work involved in data science entails contingent, improvised labour to overcome challenges in grasping and manipulating the elusive formal materiality of vast digital data sets. Although the depunctualization and articulation work our participants engaged in are iterative, ongoing, and central activities for data scientists, this labour is often overlooked and undervalued, as it tends to be performed by more junior positions in the academic hierarchy, such as graduate students and postdoctoral fellows. Acknowledging this integral labour of repair in data science is an important step in understanding the economic, political, and cultural transformations accompanying the 'data revolution' (Kitchin, 2014).

As other scholars have suggested, focusing on breakdown and repair supports a perspective on innovation that diverts us away from productivist narratives with exclusive focus on final outcomes. We find this to be true in our tracing of the breakdown–repair process, which provides the opportunity to foreground incremental innovation in particular. For example, in order to more effectively work with the material form of his data, Louis and his mentor had to customize the software tools at their disposal, a common necessity for researchers analysing large, heterogeneous data sets (Aragon, Bailey, Poon, Runge, & Thomas, 2008; Poon, Thomas, Aragon, & Lee, 2008; Vertesi & Dourish, 2011). This incremental innovation occurs even when work is unfinished and breakdown is left unfixed. In Sam's case, for example, his repair strategy did not fully resolve the breakdown; rather it demonstrates a more prolonged, ongoing, and incremental state of repair. This suggests that scholars of technology should leave room in our theorizations for an understanding of repair that does not assume whatever is broken can be fixed, and does not require that framing to identify repair work or innovation.

The breakdown–repair process also illuminates various iterations in the lifecycle of big data in academic research. For example, our analytical process allows us to understand the important and fraught role of data provenance in cases like Rachel's, in which data are being shared and repurposed to answer research questions they were not originally intended to address. Sharing and repurposing of data is becoming an increasingly common expectation in academia, and while it presents enormous potential for furthering academic research, the practice comes with a host of challenges (e.g. Borgman, 2007; Edwards et al., 2013; Trainer, Chaihirunkarn, Kalyanasundaram, & Herbsleb, 2015; Wallis, Rolando, & Borgman, 2013). When observing the way researchers deal with anomalies, we saw the articulation work of returning to the context of data production in order to understand, assess, and calibrate the quality and consistency of their data to be invaluable, and in many cases essential, to this process. Similar to what other scholars report (Faniel & Jacobsen, 2010; Rolland & Lee, 2013), we find that the use and reuse of research data required conversations and consultations with people involved directly in the data collection, documentation, and instrumentation associated with data production.

Mapping the breakdown–repair process allows us to see that data's formal materiality does not have straightforward and temporally consistent existence, but rather, is formed, reformed, and transformed through messy, iterative relationships in data assemblages.

Epistemic implications of the breakdown–repair process

Breakdowns have been long recognized in science and technology studies as disruptive occasions for revealing what is otherwise hidden and smoothly functioning infrastructure. But heeding a recent call for 'broken world thinking,' we highlight the inherently fragile, always-breaking nature of technology, the constant and quotidian labour that goes into its maintenance and repair, and the incremental innovation that arises from this labour. Steve Jackson asks us to ponder:

> Can repair sites and repair actors claim special insight or knowledge, by virtue of their positioning vis-à-vis the worlds of technology they engage? Can breakdown, maintenance, and repair confer special epistemic advantage in our thinking about technology? Can the fixer know and see different things – indeed, different worlds – than the better-known figures of "designer" or "user"? (Jackson, 2013, p. 229)

In drawing attention to the work of repair that goes into data science projects, we see the need to think more deeply about how the process of repairing data, of coming to terms with and manipulating its materiality, affect the ways we construct knowledge from it. Rachel's case shows us that certain data are rendered usable or not usable depending on whether anomalies in it can be repaired through detection and calibration; Louis' case shows us that certain techniques and methods can be applicable or not applicable depending on whether incompatibilities between data representations and tools can be repaired; Sam's case shows us that certain questions are rendered tractable or intractable depending on whether the structural relationships of the data set can be repaired through optimization. In other words, the work of repair is central to determining not just what is known, but what is *knowable*.

Conclusion

Researchers working with big data face challenges in coming to know and manipulate the formal materiality of their data because of its volume, complexity, and variance. Our ethnographic study highlights the quotidian nature of breakdown in data science practice, its importance as an occasion for gaining insight into the materiality of one's data, and the labour that goes into such repair. It is this labour that allows researchers to leverage breakdown as occasions for generating new ways of knowing, seeing, and working with data. Often dismissed as impediments that slow or derail a typical process of scientific inquiry, we argue that these encounters are underappreciated resources for knowledge production.

Using breakdown as an empirical object of study and methodological tool allowed us to further illuminate the micro practices of data science work and characterize the breakdown–repair process. Given the varied and voluminous nature of participants' data, moments of breakdown did not fully reveal a self-evident view of its inner workings. As participants engaged in depunctualization work associated with these breakdowns, they made visible otherwise obscured intermediaries within large scientific data sets, revealing

a complex process for making the material nature of the data knowable and workable. In order to build a more complete picture, researchers had to investigate their data by sifting through its content, sub-sampling, visualizing patterns in their results, and a number of other tasks. We demonstrate how the researchers engaged in articulation work to determine how to take new insights into the formal materiality of the data and leverage them towards a computational repair strategy that introduced incremental innovation within the assemblage of data and tools they were working with. These investigations, adaptations, adjustments, and repairs are generative sites of knowledge production that can be viewed as innovative practice emerging from the labour that goes into repairing different types of breakdown. Acknowledging the centrality of the breakdown–repair process has important implications for understanding the labour, innovation, and lifecycles of emerging data assemblages in the era of big data.

Notes

1. As part of the organization's effort to emulate a private sector start-up incubator model, they adopted certain terminology and procedures from the concepts of Agile software development, which is a set of software development methods aimed at fostering an adaptive project life cycle. In the DSC, aspects of some of these methods were imported into an academic context, including the stand-up meetings and the framing of problems as blockers.
2. RAM is the acronym for random access memory, which is the main type of computer memory that can be accessed randomly and can be quickly reached by the computer's processor.

Funding

This work was supported by the Gordon and Betty Moore Foundation and the Alfred P. Sloan Foundation.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Anissa Tanweer is a Ph.D. student in the Department of Communication at the University of Washington and a research assistant in the Human-Centered Data Science Lab. She is interested in the ways people organize with and around data, and how the availability of increasingly large, heterogeneous data sets is transforming the way we construct knowledge and make decisions. [email: tanweer@uw.edu]

Brittany Fiore-Gartland is a data science ethnographer in the eScience Institute and the Department of Human Centered Design and Engineering at the University of Washington. She is a Gordon and Betty Moore Foundation and Alfred P. Sloan Foundation Data Science Postdoctoral Fellow and a Washington Research Foundation Innovation Fellow. Her research focuses on the emerging practices and culture around data-intensive science. She is interested in the social and organizational dimensions of data-intensive transformations occurring across multiple sectors of work. [email: fioreb@uw.edu]

Cecilia Aragon is an Associate Professor in the Department of Human Centered Design & Engineering and a Senior Data Science Fellow at the eScience Institute at the University of Washington,

where she directs the Human-Centered Data Science Lab. She earned her Ph.D. in Computer Science from UC Berkeley in 2004. Her research focuses on human-centered data science, an emerging field at the intersection of computer-supported cooperative work and the statistical and computational techniques of data science. In 2008, she received the Presidential Early Career Award for Scientists and Engineers (PECASE) for her work in collaborative data-intensive science. Web: http://faculty.washington.edu/aragon/. [email: aragon@uw.edu]

ORCID

Brittany Fiore-Gartland http://orcid.org/0000-0003-3883-5874

References

Aragon, C. R., Bailey, S. J., Poon, S., Runge, K., & Thomas, R. C. (2008). Sunfall: A collaborative visual analytics system for astrophysics. *Journal of Physics: Conference Series*, *125*. doi:10.1088/1742-6596/125/1/012091

Borgman, C. L. (2007). *Scholarship in the digital age*. Cambridge, MA: MIT Press.

Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. Cambridge, MA: MIT Press.

Bowker, G. C. (1994). Information mythology: The world of/as information. In L. Bud-Frierman (Ed.), *Information acumen: The understanding and use of knowledge in modern business* (pp. 231–247). London: Routledge.

Burrell, J. (2012). *Invisible users: Youth in the Internet cafes of urban Ghana*. Cambridge, MA: MIT Press.

Carusi, A., & Hoel, A. S. (2014). Toward a new ontology of scientific vision. In C. Coopmans, J. Vertesi, M. Lynch, & S. Woolgar (Eds.), *Representation in scientific practice revisited* (pp. 201–222). Cambridge, MA: MIT Press.

Charmaz, K. (2014). *Constructing grounded theory*. London: Sage.

Edwards, P. N., Jackson, S. J., Chalmers, M. K., Bowker, G. C., Borgman, C. L., Ribes, D., … Calvert, S. (2013). *Knowledge infrastructures: Intellectual frameworks and research challenges*. Ann Arbor, MI: Deep Blue. Retrieved from http://hdl.handle.net/2027.42/97552

Faniel, I. M., & Jacobsen, T. E. (2010). Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. *Computer Supported Cooperative Work (CSCW)*, *19*(3–4), 355–375. doi:10.1007/s10606-010-9117-8

Fiore-Gartland, B., & Neff, G. (2015). Communication, mediation, and the expectations of data: Data valences across health and wellness communities. *International Journal of Communication, 9*, 1466-1484. Retrieved from http://ijoc.org/index.php/ijoc/article/view/2830

Gerson, E. M., & Star, S. L. (1986). Analyzing due process in the workplace. *ACM Transactions on Information Systems*, *4*(3), 257–270. doi:10.1145/214427.214431

Gitelman, L., & Jackson, V. (2013). *"Raw Data" is an oxymoron*. Cambridge, MA: MIT Press.

Graham, S., & Thrift, N. (2007). Out of order: Understanding repair and maintenance. *Theory, Culture & Society*, *24*(3), 1–25. doi:10.1177/0263276407075954

Hey, T., Tansley, S., & Tolle, K. (Eds.) (2009). *The fourth paradigm: Data-intensive scientific discovery*. Redmond, WA: Microsoft Research. Retrieved from http://research.microsoft.com/en-us/collaboration/fourthparadigm/

Jackson, S. J. (2013). Rethinking repair. In T. Gillespie, P. J. Boczkowski, & K. A. Foot (Eds.), *Media technologies: Essays on communication, materiality, and society* (pp. 221–239). Cambridge, MA: MIT Press.

Kandel, S., Paepcke, A., Hellerstein, J. M., & Heer, J. (2012). Enterprise data analysis and visualization: An interview study. *E Transactions on Visualization and Computer Graphics*, *18*(12), 2917–2926. doi:10.1109/TVCG.2012.219

Kirschenbaum, M. G. (2008). *Mechanisms: New media and the forensic imagination*. Cambridge, MA: MIT Press.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. London: Sage.

Latour, B. (2005). *Pandora's hope: Essays on the reality of science studies*. Cambridge, MA: Harvard University Press.

Latour, B. (2005). *Reassembling the social: An introduction to actor-network-theory*. Oxford: Oxford University Press.

Lievrouw, L. A. (2014). Materiality and media in communication and technology studies: An unfinished project. In T. Gillespie, P. J. Boczkowski, & K. A. Foot (Eds.), *Media technologies: Essays on communication, materiality, and society* (pp. 21–51). Cambridge, MA: MIT Press.

Martin, A., & Lynch, M. (2009). Counting things and people: The practices and politics of counting. *Social Problems*, *56*(2), 243–266. doi:10.1525/sp.2009.56.2.243

Petroski, H. (1985). *To engineer is human: The role of failure in successful design*. London: MacMillan.

Poon, S. S., Thomas, R. C., Aragon, C. R., & Lee, B. (2008). Context-linked virtual assistants for distributed teams: An astrophysics case study. In *Proceedings of the 2008 ACM conference on computer supported cooperative work (CSCW '08)* (pp. 361–370). New York, NY: ACM. doi:10.1145/1460563.1460623

de Rijcke, S., & Beaulieu, A. (2014). Networked neuroscience: Brain scans and visual knowing at the intersection of atlases and databases. In C. Coopmans, J. Vertesi, M. Lynch, & S. Woolgar (Eds.), *Representation in scientific practice revisited* (pp. 131–152). Cambridge, MA: MIT Press.

Rolland, B., & Lee, C. P. (2013). Beyond trust and reliability: Reusing data in collaborative cancer epidemiology research. In *Proceedings of the 2013 conference on computer supported cooperative work* (pp. 435–444). New York, NY: ACM. doi:10.1145/2441776.2441826

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

Sterne, J. (2014). "What do we want?" "Materiality!" "When do we want it?" "Now!" In T. Gillespie, P. J. Boczkowski, & K. A. Foot (Eds.), *Media technologies: Essays on communication, materiality, and society* (pp. 119–128). Cambridge, MA: MIT Press.

Trainer, E. H., Chaihirunkarn, C., Kalyanasundaram, A., & Herbsleb, J. D. (2015). From personal tool to community resources: What's the extra work and who will do it? In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing (CSCW '15)* (pp. 417–430). New York, NY: ACM. doi:10.1145/2675133.2675172

Vertesi, J., & Dourish, P. (2011). The value of data: Considering the context of production in data economies. In *Proceedings of the ACM 2011 conference on computer supported cooperative work (CSCW '11)* (pp. 533–542). New York, NY: ACM. doi:10.1145/1958824.1958906

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLoS ONE*, *8*(7). doi:10.1371/journal.pone.0067332