12. Towards a Reflexive Digital Data Analysis

Karin van Es, Nicolás López Coombs & Thomas Boeschoten

Introduction

As mentioned in the introduction to this book, in April 2010 flights across Europe were grounded due to the prospect of an enormous spreading ash cloud caused by the eruption of the Icelandic volcano Eyjafjallajökull. Computer simulations depicted how the volcanic ash would likely disperse. The mathematical models used in these simulations, however, were not supplemented with actual samples of the ash concentrations in the region. The EU administration received widespread criticism for their blind trust in the images produced by the simulations, which not only lacked empirical validation but were based on controversial mathematical models (Gelernter 2010; Schäfer & Kessler 2013).

Although data analysis is hardly a novel practice, the amount of data available to us has drastically increased in recent years. This explosion of data has provided researchers with exciting resources to explore social practices and relationships and has made new approaches to cultural analysis possible (Berry 2011: 1). The current wealth of data can tempt the humanities researcher to adopt new forms of empiricism. However, data are not natural phenomena, but always exist within a particular social context. It is all too easy to lose sight of the fact that 'all data provide oligoptic views of the world: views from certain vantage points, using particular tools, rather than an all-seeing, infallible God's eye view' (Kitchin 2014b: 4). When it comes to sifting and analysing this data, a critical attitude is therefore necessary. Scholars, indeed required by our datafied society to develop new literacies and competencies (see Uricchio in this volume; Rieder & Röhle in this volume and Montfort 2016), can also rely on the skills they already possess. Trained in critical inquiry, they are particularly well equipped, we would suggest, to consider the ways that data are 'cooked' (Bowker 2013). By raising questions at the various stages of digital data research, this chapter brings into focus how researchers and their tools shape data. In so doing it advocates for a reflexive form of data analysis and data visualization that can serve as a critical intervention to dispel the blind optimism and faith in the objective quantification of human behaviour and sociality through

Big Data (Van Dijck 2014). We seek to expose the limitations and biases of contemporary data analysis, which can result in rash, consequential and regrettable decision-making, as was the case with the Eyjafjallajökull ash cloud.

We begin with a brief review of two misconceptions concerning Big Data research that need to be addressed, since they conceal implicit choices made prior to starting research. We then consider questions that correspond to each phase of data analysis. These by no means exhaustive explorations represent an initial attempt to outline a reflexive, transparent procedure to guide digital data research. We recognize that the publication space for documenting methodologies is limited (Bruns 2013) and that choices need to be made in what to communicate to others. This does not, however, alleviate scholars from the obligation to consider and document their decision-making process. Although we draw primarily from our own experience of working with social media data, the reflections provoked by these questions are fruitful for data analysis more generally.

Two main misconceptions

Among the many misconceptions about Big Data (see, for example, boyd & Crawford 2013), there are two widespread assumptions that are arguably the most crucial to correct before embarking on data-driven research, as their implicit choices have serious consequences for the subsequent research process.

First, Big Data is presumed to have the inherent 'authority' to speak for itself – as when Chris Anderson (2008) notoriously declared that the current deluge of data has rendered the scientific method obsolete. Scholars such as Rob Kitchin (2014b) and Nick Couldry (2016) have criticized the idea that knowledge production is free from theory or human bias and interpretation. Kate Crawford concurs: 'Data and data sets are not objective; they are creations of human design. We give numbers their voice, draw inferences from them, and define their meaning through our interpretations. Hidden biases in both the collection and analysis stages present considerable risks' (2013: para. 2). We can add that algorithmic tools are, by nature, opaque to most researchers (see Paßmann & Boersma in this volume).

Because data are not natural resources existing *a priori* to be extracted, but rather cultural entities that are co-produced (Vis 2013), both Johanna

Drucker (2011) and Kitchin (2014a) make the distinction between *data* ('given' in Latin) and *capta* ('taken'), each preferring the latter term. Whichever term is used, data/capta are selected and shaped by humans and their technologies. We intend to reflect on this process through the questions we raise in this chapter.

Second, there is the common misconception that digital data analysis involves amassing large amounts of data and using calculations to detect underlying patterns. This view ties in with, or derives from, two other mistaken assumptions: that all data analysis is quantitative, and that any analysis involving calculations seeks solely to establish patterns. As in other academic fields, scholars within media studies use different methods to answer different types of research questions. Such an approach also applies (or should apply) when data are involved. Some researchers aim simply to ascertain how often something has happened (e.g. how often certain words are used); others, however, seek to discover *how* or *why* it has happened – requiring, in the latter case, a qualitative approach (Crawford 2013: para. 7). This observation is particularly relevant for the humanities. As Kitchin argues, quantitative approaches, while useful 'in regards to explaining and modelling instrumental systems', are limited when it comes to trying to understand human life (2014a: 145).

In the humanities, data analysis often combines quantitative and qualitative approaches. For example, Lev Manovich (2012), referring to Franco Moretti's notion of 'distant reading',[1] sees more benefit in an oscillation between the two than in simply sticking to one of these orientations. Likewise, Burdick et al. (2012) propose a digital humanities practice in which 'toggling' between both perspectives (and their attendant methods) would become the norm (30). However, we should not forget that even research that relies heavily on computational tools for the calculation of large amounts of data and the visualization of patterns still requires the researcher to interpret these patterns. As Manovich observes, '[w]hile computer-assisted examination of massive cultural data sets typically reveals new patterns in this data [...] a human is still needed to make sense of these patterns' (2012: 468–69). Making sense of such patterns, Kitchin stresses, 'requires social [or, we might add, cultural] theory and deep contextual knowledge' (2014: 144).

1 The identification of large-scale trends, patterns and relationships in large numbers of literary texts, as opposed to the 'close reading' of individual texts, a common endeavour in literary studies.

Doing Digital Data Analysis

Because digital data analysis involves many possible methods, each of which functions best in conjunction with a theoretical framework that invests the collected data with meaning, one must carefully reflect on the procedures for working with data. The following section discusses how to work with data in a reflexive fashion – that is, in which researchers consider their own role in the construction of the data. Moreover, this approach entails that researchers take responsibility to discern how the given tools work with and shape the data. To fully adopt such a reflexive approach, researchers must consider important questions that relate to each of the three stages of digital data analysis: acquiring, cleaning and analysing. These phases, inspired by the seven stages of visualization (acquiring, parsing, filtering, mining, representing, refining and interacting) that Ben Fry (2007) explores in *Visualizing Data*, are meant to elicit a critical review of the data-making process. The researcher should be able to answer each of the questions and consider which ones to highlight in the analysis; communicating these questions to others helps keep the research process transparent. The answers need not necessarily be addressed at length in the final research product; they can often be referenced as footnotes, summarized in an attachment or, in the case of visualizations, supplied via explanatory captions.

1. Acquiring: Selecting Sources and Obtaining Data

No matter what the goal of the analysis, the researcher must identify and gather the relevant data at an early stage in the process in order to answer the research question. In digital data analysis there are four principal ways to acquire such data sets. First, researchers can create their own data – through surveys and interviews, the counting of phenomena, or the tracking of uses and practices (for example, by using A/B testing or analytics software). Second, they can download (open) data made available by governments or institutions such as WikiLeaks or the Pew Research Center. Third, they can extract data from the application programming interfaces (APIs) of popular platforms such as Google Maps, Twitter and Flickr through the writing of code or the use of readymade data extraction applications that enable researchers to retrieve data from the company's database in standard file formats. Finally, they can purchase access to data through social media API aggregation companies such as Gnip, Topsy and

DataSift. However, researchers should be aware that the existence of such commercial resellers limits free access to social media data (Manovich 2012). Indeed, each form of data gathering carries its own limitations and biases. APIs, for instance, not only provide data but are themselves 'data makers' as well: they construct and provide access to certain (meta)data (Vis 2013). This raises further questions about the reliability and validity of the data, as well as how representative it is (see Cornelius Puschmann and Julian Ausserhofer in this volume). Moreover, biases already exist in data sets in that data collection privileges certain social groups (see Leurs and Shepherd in this volume).

Although online data are readily available, such accessibility does not necessarily mean that it is ethical for researchers to use them (boyd & Crawford 2013; Zimmer 2010). Despite such data being 'public', people have expectations as to how this information will be presented and employed (see Markham & Buchanan 2012). Undertaking large-scale online research thus prompts a series of ethical questions. As a result of the significant changes in the scale and scope of data, traditional ethical guidelines relating to 'informed consent' and privacy, as well as the definition of 'human subjects' and the concept of 'harm' in relation to participants, need to be revisited (see Van Schie, Westra & Schäfer in this volume). Prior to embarking on any research project, one must consider research ethics (Markham & Buchanan discuss guidelines in their contribution to this volume) and assess *who* is doing the asking – and how that shapes research outcomes.

The researcher's first challenge is to define the research data. This process should be guided by theory, a research question or preliminary explorations. Note here that not all research requires 'big' data, and in some cases 'small data' (e.g. a focus on a single individual) can be more productive (boyd & Crawford 2013: 670). Small data afford different kinds of questions and methods than Big Data, and therefore yield different kinds of knowledge. The type of data collected depends on the research question (although the first analysis of a data set can also lead to the formulation of a research question). When acquiring data for research, whether big or small, the following questions should be considered:

- What ethical considerations have been taken into account when collecting the research data?
- What kind of data is being used?
- How was the data collected? Which tools or software were used, or who supplied the data?

– Which criteria were used to select the data set? Who is included or excluded from the data set?
– What are the limitations of these data-gathering methods? How reliable is the method of data collection?
– What metadata does the data set contain (for example, location, time, date of a tweet)?
– When combining data sets, what biases might result from the different contexts in which the data originated?

2. Cleaning: Parsing and Filtering Data

After reflecting on how the data was retrieved, the researcher needs to explain the decisions made to prepare for subsequent analysis, which also involves removing certain data from the data set. This part of the process concerns how the data has been organized into categories and which data has been retained. Researchers may find that a single data set is not sufficient to realize their objectives. This problem can often be addressed by combining data sets or enriching the data. One may include answers to the following questions in the analysis:

– What categories are used to organize the data?
– What do the categories assume about the meaning of the data to be measured and/or calculated?
– How has irrelevant data (that is, spam or 'noise') been dealt with?
– What is the 'quality' of the data (for example, were some data wrongly formatted and did they have to be restored)?
– How has the data(set) been enriched? For what purposes?

Here, it is important to recognize that when we organize data into categories (according to population, gender, nation, etc.), these categories tend to be treated as if they were discrete and fixed, when in fact they are interpretive expressions (Drucker 2011).

3. Analysing: Mining, Representing and Refining Data

To understand the data and discern underlying patterns, researchers will often use statistical and data-mining methods (Fry 2007: 5). Digital data analysis requires a basic knowledge of math and statistics (that is,

sampling and calculating mode/mean/median), so that researchers can assess whether patterns in the data are the result of chance and determine what biases are at play. Moreover, when studying social networks, one should also be aware that 'the degree distribution typically follows a power law distribution, i.e. most people have a few friends, while few people have [a great] many friends' (Tang et al. 2012: 5). In such instances it is futile to discuss averages.

Prior to finding correlations and making statistical claims, researchers should provide for the reader an assessment of the value and meaning of the metrics they are using. For example, in working with information gleaned from social media, it is generally taken for granted that 'shares', 'likes', 'follows' and 'retweets' are salient research material, although the *meaning* of each online gesture is not self-evident. Although it is easy to figure out which users have been retweeted most often, it is not clear what this means (why people have retweeted others' tweets); to discover the answer requires different, qualitative methods. The prestructured actions one finds on social media are not always similar and comparable (see the interview with Carolin Gerlitz in this volume). To put it simply, not all 'likes' are created equal.

Although large data sets are useful for detecting patterns and connections, Big Data research risks having its practitioners see correlations everywhere (Marcus & Davis 2014). The opposite of abductive reasoning, this tendency is called 'apophenia', defined by Dan Dixon as 'pattern recognition gone wrong, seeing only the pattern expected, no matter what data leads to it' (2012: 202). It is a particular pitfall in Big Data research since not all patterns and relationships found in the data are meaningful or truthful (Kitchin 2014: 13). We should recall here the commonplace of statistics that 'correlation does not imply causation' and the fallacy that data is self-evident.

In addition to statistical methods, there are numerous ways to visualize data and many tools for doing so. Visualizations 'may be used as analytical and interpretive tools – to reveal patterns or anomalies or concurrences – or they may be produced to illustrate findings or serve as the distillation of an argument' (Burdick et al. 2012: 43). They can also be used to tell stories in new ways, emphasizing different relations (see Venturini et al. in this volume). Instruments of visualization range from easy-to-use tools that provide WYSIWYG (what you see is what you get) interfaces for data taken directly from cloud services, databases, APIs (social-media-metrics providers such as Buzzcapture, Salesforce, OBI4wan, etc.) or imported spreadsheets such as Tableau and Gephi, to more sophisticated means that require programming (for example, R and D3) or designing in programs such as Illustrator.

Although each tool raises its own specific questions, the following will apply in most instances:

- How was the data prepared and combined for visualization (by filtering, transforming, calculating and enriching)?
- What purpose does the visualization serve?
- Why has this type of visualization been selected?
- How have the colours, sizes and shapes in the visualization been determined?
- What software has been used and why? What computational methods does the research employ?
- Which settings and algorithms were applied?
- How have the decisions related to the above-mentioned questions highlighted or downplayed aspects of the underlying data set?

Each graph and illustration should be provided with a number and a description, and in the analysis itself, it is important to differentiate between description (for example, an explanation of the content, type of graph, sample size, etc.) and the interpretation of what is shown.

Conclusions

The Icelandic ash cloud debacle caused cancellations of some 100,000 flights, stranded 10 million people and collectively cost airlines and airports over 2 billion dollars (*The Telegraph* 2011). For many EU citizens, it was also their most memorable exposure to data research, understandably triggering academic criticism over the objective appearance of the visuals (Schäfer & Kessler 2013). To overcome this significant blot on the field of digital data analysis, we need to engage critically and transparently with data. It is crucial for researchers to reveal how they and the tools they use have shaped their research, and how the data they employ has been influenced by the platforms they originated on. When undertaking data analysis, therefore, the researcher must reflect on the following considerations:

- where the data came from;
- who produced the data and for what purposes;
- what data are selected and how they relate to the larger data set;
- which tools were used for collection and analysis;
- why certain data and metrics were used for the research.

These considerations are mirrored in the more detailed questions relating to the phases of the research process; they can help create awareness of the choices made by researchers during their research and debunk the common misconception that data and data visualizations are neutral and objective. They represent a first attempt to cast light on the inner workings of Big Data research and join the plea that we as researchers have to be more transparent in our procedures in working with data. For consumers of data, these efforts will hopefully contribute to an increased awareness of the stages involved in the production of data and the adoption of a critical stance towards the data they interpret and make sense of. As we come to live in an increasingly datafied society, these aims seem more relevant than ever.

Acknowledgments

The authors would like to thank Eef Masson, Fernando van der Vlist and Mirko Tobias Schäfer for their critical remarks and suggestions to earlier drafts of this article.

References

Anderson, Chris. 2008. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." www.wired.com/2008/06/pb-theory/.

Berry, David M. 2011. "The Computational Turn: Thinking About the Digital Humanities." *Culture Machine* 12: 1–22. www.culturemachine.net/index.php/cm/article/viewArticle/440.

Bowker, Geoffrey C. 2013. "Data Flakes: An Afterword to *'Raw Data' Is an Oxymoron*." In Lisa Gitelman, ed., *"Raw Data" Is an Oxymoron*. Cambridge, MA: The MIT Press, 167–72.

boyd, danah, & Crawford, Kate. 2012. "Critical Questions for Big Data: Provocations for a Cultural, Technological, And Scholarly Phenomenon." *Information, Communication & Society* 15 (5): 662-679.

Bruns, Axel. 2013. "Faster Than The Speed of Print: Reconciling 'Big Data' Social Media Analysis and Academic Scholarship." *First Monday* 18 (10). http://firstmonday.org/ojs/index.php/fm/article/view/4879/3756.

Burdick, Anne, Johanna Drucker, Peter Lunenfeld, Todd Presner & Jeffrey Schnapp. 2012. *Digital_Humanities*. Cambridge, MA: The MIT Press.

Couldry, Nick. 2014. "Inaugural: A Necessary Disenchantment: Myth, Agency and Injustice in a Digital World." *Sociological Review* 62 (4): 880–97.

Crawford, Kate. 2013. "The Hidden Biases in Big Data." *HBR Blog Network*, *1*. https://hbr.org/2013/04/the-hidden-biases-in-big-data/.

Dijck, José van. 2014. "Datafication, Dataism and Dataveillance: Big Data Between Scientific Paradigm and Ideology." *Surveillance & Society* 12 (2): 197–208.

Dixon, Dan. 2012. "Analysis Tool or Research Methodology: Is There an Epistemology for Patterns?" In *Understanding Digital Humanities*, ed. David M. Berry. New York: Palgrave Macmillan, 191–209.

Drucker, Johanna. 2011. "Humanities Approaches to Graphical Display." *Digital Humanities Quarterly* 5 (1): 1–21. www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html.

Fry, Ben. 2007. "The Seven Stages of Visualizing Data." In *Visualizing Data*. Sebastapol, CA: O'Reilly Media, 1–18.

Gelernter, David. 2010. "Gefahren der Softwaregläubigkeit: Die Aschewolke aus Antiwissen." *Frankfurter Allgemeine Zeitung*, April 26. www.faz.net/aktuell/feuilleton/debatten/digitales-denken/gefahren-der-softwareglaeubigkeit-die-aschewolke-aus-antiwissen-1606375.html.

Kitchin, Rob. 2014a. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. London: Sage.

—. 2014b. "Big Data, New Epistemologies and Paradigm Shifts." *Big Data & Society* 1 (1): n.p.

Manovich, Lev. 2012. "Trending: The Promises and the Challenges of Big Social Data." In M.K. Gold, ed., *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press, 460–75.

Marcus, Gary & Ernest Davis. 2014. "Eight (No, Nine!) Problems with Big Data." *New York Times*, April 6, Op-Ed pages. www.nytimes.com/2014/04/07/opinion/eight-no-nine-problems-with-big-data.html?_r=1.

Markham, Annette & Elizabeth Buchanan. 2012. "Ethical Decision-making and Internet Research 2.0: Recommendations from the Association of Internet Researchers Ethics working committee." (Downloaded from http://aoir.org/reports/ethics2.pdf).

Schäfer, Mirko Tobias & Frank Kessler. 2013. "Trust in Technical Images." *Mtschaefer.net*. http://mtschaefer.net/entry/trust-technical-images/.

Tang, Jiliang, Xufei Wang & Huan Liu. 2012. "Integrating Social Media Data for Community Detection." In Martin Atzmueller, Alvin Chin, Denis Helic & Andreas Hotho (eds.), *Modelling and Mining Ubiquitous Social Media*. Berlin: Springer, 1–20.

The Telegraph. 2011. "How the 2010 Ash Cloud Caused Chaos: Facts and Figures," 24 May, sec. Finance. www.telegraph.co.uk/finance/newsbysector/transport/8531152/How-the-2010-ash-cloud-caused-chaos-facts-and-figures.html.

Vis, Farida. 2013. "A Critical Reflection on Big Data: Considering APIs, Researchers and Tools as Data Makers." *First Monday* 18 (10). http://firstmonday.org/ojs/index.php/fm/article/view/4878/3755.

Zimmer, Michael. 2010. "'But the Data Is Already Public': On the Ethics of Research in Facebook." *Ethics and Information Technology* 12 (4): 312–25.