Accepted manuscript version of: Gray, J. & Bounegru, L. (2019) "What a Difference a Dataset Makes? Data Journalism And/As Data Activism". In *Data in Society: Challenging Statistics in an Age of Globalisation*, J. Evans, S. Ruane and H. Southall (eds). Bristol: The Policy Press.

What a difference a dataset makes?
Data journalism and/as data activism

Jonathan Gray and Liliana Bounegru

Introduction

How and when might data journalism be viewed as a form of "data activism"? Data activism has been conceptualised as a set of practices which "interrogate the fundamental paradigm shift brought about by datafication", including through resisting surveillance and mobilising data to denounce injustice and advocate for change (Milan and van der Velden, 2016). In this chapter we examine three ways in which data journalism can serve not just to reinforce and reify dominant regimes of datafication – or ways of rendering life into data (van Dijck, 2014) – but also to interrogate them and make space for public involvement and intervention around data infrastructures.

Researchers and practitioners contend that the boundaries between journalism and activism may become porous, particularly when it comes to emerging digital technologies and media practices (Russell, 2017). A case in point is the Panama Papers, which has been described as both the "biggest leak in the history of data journalism" (Snowden, 2016), winning a top prize at the international "Data Journalism Awards", as well as the "coming-of-age of leaktivism", an emerging form of "social protest" (White, 2016). This is not the first time that data journalism has been associated with mega-leaks. Wikileaks was said to play a key role in obtaining broader recognition for data journalism as an emerging class of news work (Rogers, 2011, 2012). Other studies indicate the entanglements between data journalism and fields such as civic hacking and open data advocacy (Baack, 2017).

Data journalism has been broadly defined as "journalism done with data" incorporating a variety of different practices (Gray et al., 2012). Researchers and practitioners alike have discussed the relationship between data journalism and the promotion of facts, scientific norms and cultures of objectivity (Bounegru, 2012; Anderson, 2015; Gray et al., 2016). Many data journalism projects may be argued to embody a form of what Desrosières calls "proof in use realism", or an attitude that "'reality' is nothing more than the database to which they

have access" (Desrosières, 2001). According to such a view, data visualisations, data stories, data interactives may treat datasets and databases as self-evident collections of facts which designate different aspects of the world.

However, despite the abundance of "fact talk" amongst practitioners (e.g. Rogers, 2013; Silver, 2014), journalists may use data to do more than simply represent facts. Overemphasising the representational capacities of data, may lead us to underemphasise what else it can do (cf. Espeland and Stevens, 2008; Verran, 2015). There are many other things that data can do and ways in which it can become a "matter of concern" (Latour, 2004) in journalism, whether as an event to be covered, a cause to be advocated, a good to be protected, a topic to be editorialised, a set of skills to be adopted, a product to be monetised, a source to be substantiated or an object to be investigated.

In this chapter we explore how data journalism can serve not just to affirm and amplify "modes of authorized seeing" (Jasanoff, 2017) through the representation of facts, but rather to assemble actors to challenge, investigate, reimagine and explore alternatives to dominant forms of datafication. It might thus be considered as a potential site of experimentation and participation in emerging forms of data politics, data culture and "data worlds" (Gray, 2018; Gray et al, 2018). In order to examine this, in the following sections we look at three ways of doing things with data in the context of journalism: (1) assembling data publics; (2) making data differently; and (3) investigating datafication.

Assembling data publics

Journalists do not only use data to tell stories. They also use datasets and data infrastructures in order to facilitate various forms of interactivity, engagement, participation and collaboration. These efforts to assemble what Ruppert calls "data publics" (2015) can be understood in relation to broader social practices of participation organised using the internet and digital technologies (Fish et al., 2011).

This includes *invitations to explore, re-use and tell stories with data*. One of the defining features of the Guardian Datablog was its prominent links to "download the data" or "get the data", along with questions such as "What can you do with it?" and a recurring open invitation to "Please post your visualisations and mash-ups on our Flickr group" (Guardian, 2011). Rather than focusing on a single story or visualisation, data is thus considered the basis for different ways of seeing and knowing which readers are encouraged to explore.

Databases are also viewed as journalistic outputs in themselves, which can be used not only to provide information, but also to *mobilise action and facilitate interaction*. ProPublica's "Dollars for Docs" project provides a database of payments from pharmaceutical companies to doctors and hospitals (ProPublica, 2016). The project invites visitors to print out and ask their doctors about payments they have received. In a similar vein, the ICIJ have released a dataset from the Panama Papers so that "regulators and ordinary citizens from around the globe [can] probe the newly available data and find new connections that may have escaped reporters" (ICIJ, 2016). Like the petition or the hashtag, databases can thus be considered as devices to assemble and mobilise publics around issues.

Databases can be used as *crowdsourcing mechanisms* to enrich existing data or generate new data. They may thus serve as devices to structure interaction and participation in specific ways in order to advance data journalistic reporting. The Guardian's MP's Expenses encouraged readers to classify and comment on documents about UK parliamentary expense claims (Gray et al., 2012: 36, 137-139). *La Nacion* in Argentina used their VozData collaboration platform to review 40,000 leaked audio recordings related to the death of a government lawyer (La Nacion, 2017). In such cases interaction and participation is highly conventionised in order to enable distributed collaboration at larger scales.

Data journalists also use various tools to *open up their data work and collaborate with others*. Many data journalists use the GitHub platform to develop, collaborate around and publish datasets, analysis and code associated with their work (Bounegru, 2015). An investigation from BuzzFeed News into racial divisions in St. Louis County, US, used Github to publish data as well an open-source notebook showing the code for their analysis (BuzzFeed News, 2014). Others use online services such as Google Docs and Sheets to

organise distributed collaboration around data investigations, such as The Bureau of Investigative Journalism's collaborative investigation of spending cuts to services for children (TBIJ, 2018a). For projects such as the Panama Papers, databases can become what Susan Leigh Star calls "boundary objects" (Bowker et al., 2015), infrastructures which assemble and hold together different communities with diverse interests.

These arrangements to facilitate involvement around data journalism projects illustrate different practices for "making data public". They may organise and enrol data publics in a wide range of roles including as *contributors, collaborators, co-investigators, coders, designers, innovators, auditors, witnesses* and *activists*. In such cases datasets may not only provide factful representations of the world, but also facilitate the gathering of publics and the coordination and conventionalisation of data work, collective inquiry and other forms of social action.

Making data differently

When data journalists are not able to identify data around the issues and objects that they would like to report on, they may attempt to *change how it is made* or *make it for themselves*. As alluded to above, one way of collecting data is through *crowdsourcing mechanisms*. When other avenues have failed, data journalists have used such mechanisms to request input from users, such as crowdsourcing data on water bills in France in order to investigate unfair pricing practices (Gray et al., 2012: 106-107).

Many data journalists have sought to *create structured databases on the basis of official documents which can be scattered in different locations*. A network of journalists associated with the FarmSubsidy project sought to create a single database of EU farm subsidy data by extracting information provided through freedom of information requests and PDF documents, so that they could look at how much large beneficiaries received across multiple countries (Gray et al., 2012: 121-122). A similar approach was used by the Financial Times to investigate the EU's structural funds (Gray et al., 2012: 64-66) as well as by The Bureau of Investigative Journalism to map and report on local funding cuts (Mair et al., 2017).

Data journalists may also *assemble data themselves using their own methods, techniques and devices*. In the "Migrants' Files" project a network of European journalists gathered data about deaths of migrants en route to Europe through a combination of Non-Governmental Organisation data, lists from journalists and media monitoring (Gray, Lämmerhirt, & Bounegru, 2016). The Guardian's "The Counted" project used a similar combination of sources and strategies, as well as reader submissions, social media monitoring, Google Alerts and community-building efforts in order to compile a database of police killings in the US (Gray, Lämmerhirt, & Bounegru, 2016). The Bureau of Investigative Journalism's *Dying Homeless* project sought to "count those that die homeless on UK streets" organising a collaborative investigation with using a combination of online forms, chat channels and the #makethemcount hashtag on Twitter (TBIJ, 2018b). Beyond the screen, data journalists have also sought to gather structured data using sensors, drones and other devices (Pitt, 2014).

In many of these cases journalists were responding to the lack of data from other sources. When infrastructures involved in the creation of data do not address or align with their interests or concerns, they may *inventively repurpose* information from other sources or *create their own infrastructures for making data* (Gray et al., 2018). This is not simply a case of "filling the gaps" in existing regimes of quantisation or datafication, but rather of *rendering* different aspects of collective life into data through fieldwork (whether through on-site reporting, sensors or drones), screen work or collaborations with others.

Investigating datafication

In some circumstances, data may become not just a "matter of fact", but a "matter of concern" for journalists (Latour, 2004), leading them to consider it an object of investigation: how and by whom it was generated, how it is used, what it shows and does not show, how it may be manipulated, and the different kinds of biases, inequalities and injustices that it may give rise to. As such journalists may not just take data for granted as "given", but also may consider its "scenography" (Latour, 2008), the conditions and settings in which it is created, used and shared.

A huge amount of digital data is generated as a result of interactions with online platforms, apps and digital devices. While journalists can often tell stories *with* such digital data, more or less unproblematically, they may also tell stories about the production of digital data through *investigations into platforms, algorithms and digital datafication*. ProPublica's "Machine Bias" series reporters investigate "algorithmic injustice and the formulas that influence our lives", including price discrimination by online platforms and insurance companies; bias in criminal risk scores; platforms allowing advertisers to exclude users by race and how artificial intelligence engines are trained to be racist (ProPublica, 2018). Techniques for these investigations include *scraping data from platforms*, *obtaining data from advertising programmes* and *comparing predictions to outcomes*.

Sam Lavigne's "Infinite Campaign" for The New Inquiry playfully explores "the bizarre rubrics Twitter uses to render its users legible" by scraping data from Twitter's ad creation page and creating "a taxonomy of human beings according to Twitter and its data brokers" (New Inquiry, 2017). This is then used as the basis for video sequences combining phrases from this taxonomy with stock footage to create an endlessly scrolling array of clips displaying "the fantasies by which Twitter understands us" (Figure 1).

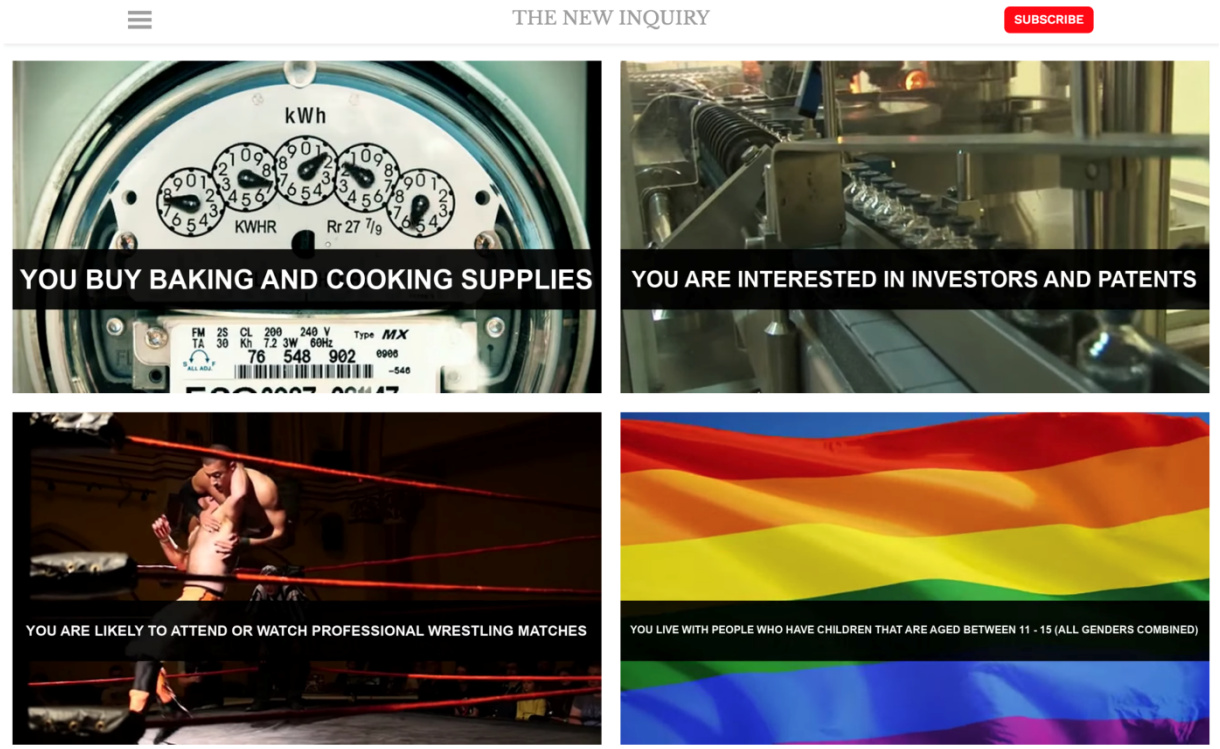

Figure 1: New Inquiry's "The Infinite Campaign".

In these cases *data becomes problematised* for journalists. Rather than being taken as a resource to be straightforwardly utilised, data becomes an object to be interrogated. As with the previous section on "making data differently" in these cases the social and technical conditions of creation become a "matter of concern". The aspects of data journalism discussed above can hence be mobilised in the service of what has been called "algorithmic accountability" (Diakopoulos, 2015), providing investigative reports on the operations of platforms, algorithms and other agents of digital datafication. While algorithmic accountability reporting often emphasises analytical operations and decision-making processes such as "prioritization, classification, association, and filtering" (400), it is worth noting that journalists may also attend to ways in which data is shaped through platforms and infrastructures, which make these algorithmic operations possible (Gray et al., 2018).

Conclusion

The cases we have examined suggest ways in which data journalists may facilitate broader public engagement and debate around datafication, rendering visible what is involved in making and using data. Although the question of when data journalism can fruitfully be considered as data activism requires situational analysis (looking at settings and relations, rather than features of projects), nevertheless we hope to have illustrated how the study of data journalism practices may enrich and complement research on the repertoires of data activism. Attending to such practices may also inform experimentation around how data journalism projects may serve not just to reproduce and communicate established facts and dominant forms of datafication, but also how they may constitute a site of collective inquiry, intervention and involvement in cultures and politics of data.

References

Anderson, C. W. (2015). Between the Unique and the Pattern. *Digital Journalism*, 3(3), 349–363.

Baack, S. (2017). Practically Engaged. *Digital Journalism*. 6(2), 673–692.

Bounegru, L. (2012). Data Journalism in Perspective. In J. Gray, L. Bounegru, & L. Chambers (Eds.), *The Data Journalism Handbook*. Sebastopol, CA: O'Reilly Media. Retrieved from http://datajournalismhandbook.org/

Bounegru, L. (2015) GitHub as Transparency Device in Data Journalism, Open Data and Data Activism. Retrieved from: http://lilianabounegru.org/2015/07/08/github-as-transparency-device-in-data-journalism-open-data-and-data-activism/

BuzzFeed News (2014) Analysis and data notes for the August 20, 2014 BuzzFeed News article, "The Ferguson Area Is Even More Segregated Than You Probably Guessed". Retrieved from: https://github.com/BuzzFeedNews/2014-08-st-louis-county-segregation

Desrosières, A. (2001). How Real Are Statistics? Four Possible Attitudes. *Social Research*, 68(2), 339–355.

Diakopoulos, N. (2015). Algorithmic Accountability. *Digital Journalism*, 3(3), 398–415.

Espeland, W. N., & Stevens, M. L. (2008). A Sociology of Quantification. European Journal of Sociology / Archives Européennes de Sociologie, 49(3), 401–436.

Fish, A., Murillo, L. F. R., Nguyen, L., Panofsky, A., & Kelty, C. M. (2011). Birds of the Internet. *Journal of Cultural Economy*, 4(2), 157–187.

Gray, J. (2018). Three Aspects of Data Worlds. *Krisis: Journal for Contemporary Philosophy*. Issue 1, 4–17.

Gray, J., Bounegru, L., & Chambers, L. (Eds.). (2012). *The Data Journalism Handbook*. Sebastopol, CA: O'Reilly Media. Retrieved from http://datajournalismhandbook.org/

Gray, J., Bounegru, L., Milan, S., & Ciuccarelli, P. (2016). Ways of Seeing Data: Toward a Critical Literacy for Data Visualizations as Research Objects and Research Devices. In

Kubitschko, S & Kaun, A (Eds.), *Innovative Methods in Media and Communication Research*, London: Palgrave Macmillan, pp. 227–251.

Gray, J., Lämmerhirt, D., & Bounegru, L. (2016). Changing What Counts: How Can Citizen-Generated and Civil Society Data Be Used as an Advocacy Tool to Change Official Data Collection? CIVICUS and Open Knowledge. http://papers.ssrn.com/abstract=2742871

Gray, J., Gerlitz, C., & Bounegru, L. (2018). Data Infrastructure Literacy. *Big Data and Society*, 5(2), 1–13.

Guardian (2011). All of our data journalism in one spreadsheet. *The Guardian* January 27

ICIJ (2016). ICIJ Releases Panama Papers Offshore Company Data. Retrieved from: https://www.icij.org/blog/2016/05/icij-releases-panama-papers-offshore-company-data/

Jasanoff, S. (2017). Virtual, Visible, and Actionable: Data Assemblages and the Sightlines of Justice. Big Data & Society, 4(2), 1–15.

La Nacion (2017). Dos años de análisis de las escuchas de Nisman. Retrieved from: https://www.lanacion.com.ar/1976325-dos-anos-de-analisis-de-las-escuchas-de-nisman

Latour, B. (2004). Why Has Critique Run out of Steam? From Matters of Fact to Matters of Concern. *Critical Inquiry*, 30(2), 225–248.

Latour, B. (2008). What Is the Style of Matters of Concern? Amsterdam: University of Amsterdam.

Mair, J., Keeble, R. L., & Lucero, M. (Eds.). (2017). *Data Journalism: Past, Present and Future*. London: Abramis.

Milan, S., & van der Velden, L. (2016). The Alternative Epistemologies of Data Activism. *Digital Culture & Society*, 2(2), 57–74.

New Inquiry (2017). Taxonomy of Humans According to Twitter. Retrieved from: https://thenewinquiry.com/taxonomy-of-humans-according-to-twitter/

Pitt, F. (2014). Sensors and Journalism. New York: Tow Center for Digital Journalism. Retrieved from https://towcenter.org/research/sensors-and-journalism/

ProPublica (2016). Dollars to Docs. Retrieved from: https://projects.propublica.org/docdollars/

ProPublica (2018). Machine Bias series. Retrieved from: https://www.propublica.org/series/machine-bias/

Rogers, S. (2011). Wikileaks Data Journalism: How We Handled the Data. The Guardian. January 31.

Rogers, S. (2012). Behind the Scenes at the Guardian Datablog. In J. Gray, L. Bounegru, & L. Chambers (Eds.), *Data Journalism Handbook*. Sebastopol, CA: O’Reilly Media. Retrieved from http://datajournalismhandbook.org/1.0/en/in_the_newsroom_3.html

Rogers, S. (2013). *Facts are Sacred*. London: Faber and Faber.

Russell, A. (2017). *Journalism as Activism: Recoding Media Power*, Chichester: John Wiley & Sons.

Silver, N. (2014, March 17). What the Fox Knows. Retrieved January 30, 2018, from https://fivethirtyeight.com/features/what-the-fox-knows/

Snowden (2016, April 3). Biggest leak in the history of data journalism just went live, and it's about corruption. http://panamapapers.sueddeutsche.de/en/ [Twitter post]. Retrieved from: https://twitter.com/Snowden/status/716683740903247873

Bowker, G. C., Timmermans, S., Clarke, A. E., & Balka, E. (Eds.). (2015). *Boundary Objects and Beyond: Working with Leigh Star*. Cambridge, MA: The MIT Press.

TBIJ (2018a). Councils "Dangerously Close to Brink" as Half Plan to Cut Spending on Vulnerable Children. Retrieved from: https://www.thebureauinvestigates.com/stories/2018-02-07/childrens-services-perilous-as-councils-struggle-to-balance-their-budgets

TBIJ (2018b) Dying Homeless: Counting the Deaths on UK Streets. Retrieved from: https://www.thebureauinvestigates.com/stories/2018-04-23/dying-homeless

van Dijck, J. (2014). Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology, *Surveillance & Society*, *12*(2), 197–208.

Verran, H. (2015). Enumerated Entities in Public Policy and Governance. In Mathematics, Substance and Surmise (pp. 365–379). New York: Springer, Cham.

White, M. (2016). The Panama Papers: Leaktivism's Coming of Age. *The Guardian*, April 5.