CHAPTER 5
STRAIGHTWASHING: THE CLEANING AND ANALYSIS OF QUEER DATA

I arrived late and was quickly ushered through the university museum and into the historic library – the grand location for a drinks reception to celebrate the conclusion of an international conference on gender equality. It was late August and unusually humid. Makeup was running, shirts were starting to show early signs of sweat and coiffed hair was slipping under the evening heat. Waiting staff dotted the room with trays of canapés. Glasses were quickly refilled with wine as soon as they showed signs of emptying. To address the heat, a row of air conditioning units pushed a stream of cold air from one end of the room to the other. Attendees had also taken matters into their own hands and refashioned conference programmes as make-shift fans, fluttering like butterflies in the grand setting. Glasses clinked and the principal of the university, hosting the conference, took to the floor. Flanked on either side by government ministers and local dignitaries, the principal's speech focused on the university's reduction of its gender pay gap (the percentage difference in average pay between male and female staff). The principal outlined the most recent data, the current mean and median gap, the previous mean and median gap, and the change in percentage points. The numbers were presented with passion and, for myself and others in the room, appeared to take on a life of their own as monuments of gender equality.

The speech concluded and the crowd broke into chatter. Attendees seemed impressed with the principal's account of gender pay gap data at the university. The numbers certainly sounded positive. But I felt something was missing and, as others returned to the food and drink, I was left with more questions than answers. I was unsure what the data *really* told me about gender equality at the university. What were the experiences of female staff and how did they compare to male and non-binary staff? How did this intersect with the experiences of Black staff, disabled staff and LGBTQ staff? I also wanted to know more about how the data was analysed. Who counted as staff? How were part-time staff or those with multiple roles counted? What were the differences between mean and median averages? Data analysis shapes the stories we tell about EDI. Analysis of data can provide a cover.

Statements that focus on what was uncovered, rather than how this was uncovered, provide data with a long leash that can depart from the original intentions as to why it was collected. The transformation of numbers into 'good data' can become an objective in its own right. As resources are always limited, there is the risk that the collection and analysis of data drain energy from other initiatives intended to address inequality. It is therefore vital that, as with the collection of data, a queer approach critically investigates what happens during the analysis of gender, sex and sexuality data.

* * *

The cleaning of data can remove its queerness: paper surveys where respondents score out the response options 'female' and 'male' and write their own answer, interview recordings where participants flip the focus and ask questions of the researcher, census returns where LGBTQ couples identify themselves as 'married' even when governments do not recognize same-sex marriage. These examples demonstrate how collection methods can fail to restrict how participants share data about their lives and experiences. Although the move from paper to online surveys has made it harder for participants to select multiple response options, write-in comments or answer questions not intended for them, data analysis offers an additional opportunity to ensure that results align with the expectations of those who designed the study and the social world they wish to bring into being. In this chapter, I examine the preparation, analysis and usability of gender, sex and sexuality data. This process begins with cleaning, which involves the removal of data that breaks established rules (for example, the deletion of respondents that started a survey but did not answer any questions). Tamraparni Dasu and Theodore Johnson have estimated that 80 per cent of data analysts' time is spent on the cleaning and preparation of data – yet this stage in data's journey often remains undiscussed.[1] Cleaning data also involves the use of available information, such as responses to other questions in a survey, to add missing values or change answers provided. After cleaning, analysts are then required to use their judgement to splice or split categories, a practice known as aggregation and disaggregation. The assumption, often misplaced, that the number of people in a population who identify as LGBTQ is small can halt research before it even begins and materializes as the 'I have no problem

[1] Tamraparni Dasu and Theodore Johnson, *Exploratory Data Mining and Data Cleaning*, Wiley Series in Probability and Statistics (New York: Wiley, 2003).

with LGBTQ people but they don't work here' argument. The aggregation of LGBTQ groups offers a response to when 'small numbers' is used as an excuse for inaction. However, if research is conducted into the experiences of LGBTQ people, and the number of cases is smaller than anticipated, organizations might also use data to halt initiatives or cut funding. Small numbers therefore present multiple dangers for the analysis of data about LGBTQ people.

Related to analysis is the topic of big data, where datasets are huge in volume, collected in or near real time and diverse in scope. The speed and scale of big data are ideal for algorithmic decision-making, in which machines execute complex instructions that shape people's everyday lives. However, big data's ambition for total knowledge often overlooks the effects of homophobia, biphobia and transphobia on historical and contemporary data practices and is ill-suited to qualitative methods, which tend to predominate studies of LGBTQ people. In response, I highlight the merits of small data and queer data's potential to uncover new insights when research is undertaken that cuts across both big and small approaches.

Part II is a bridge between the collection of queer data and its uses for action. Data collection can result in researchers facing a messy assortment of transcripts, datasets and documents that, on their own, are hard to fathom. Analysis is the process of making sense of what this data tells us. Yet, whether working with qualitative or quantitative data, decisions made about analysis are never value-neutral. Biases and assumptions that inform the meanings of gender, sex and sexuality data do not stop when the researcher closes the online survey or turns off their dictaphone. Decisions made during analysis equally shape the findings of a research project and ultimately determine data's potential to impact the lives and experiences of LGBTQ people. This chapter therefore concludes with a discussion of Browne's account of running a survey for Pride in Brighton and Hove and the challenge of conducting data analysis that authentically represents the lives of LGBTQ people yet also produces results that are useable for future action. I frame this example within a wider debate about Gayatri Chakravorty Spivak's concept of strategic essentialism and the use of analytical methods that misrepresent LGBTQ people as a means to advance social and political change.

Cleaning, disaggregation and aggregation

Cleaning implies there is something wrong with the data collected. However, for many LGBTQ people, providing the 'wrong' answer is not a mistake but

an attempt to subvert cis/heteronormative assumptions that influenced the design of the collection instrument. For respondents who do not feel represented by the options offered, Ryan highlights 'a political obligation to alert the system of their presence by refusing to simply check one of the limited options with which they feel they are presented'.[2] Ryan tells the story of Jamie, who identifies as neither male nor female, who explains that 'for surveys where it's not going to be some big media frenzy thing I always write in my own thing and add things like intersex, other, transgender, and then circle transgender'.[3] Attempts to 'clean' LGBTQ transgressions during data analysis are evident in the recent history of the US census. In the 1990 census, for example, if two male respondents in the same household described their relationship as 'husband/wife' the census bureau would leave the relationship intact but change the sex of one of the male respondents to female.[4] As discussed in Chapter 4, Velte describes these practices as 'straightwashing' census data.[5] Although the option of 'unmarried partner' was added to the 1990 census, and several states provided legal recognition for same-sex couples during the period 2000 to 2010, the transformation of responses related to same-sex couples continued, to some extent, until the 2020 census. Michael Brown and Larry Knopp, commenting on the UK's 2001 censuses, have also noted that 'data-cleaning techniques and the lack of familiarity with same-sex couples' meant that 'census officials have had to closet some same-sex couples by recoding them, in order to maintain a consistent definition of household relation'.[6] These examples make clear the need for transparency about analytical approaches so that decisions, which reconfigure the meanings of data, are deliberated among the people about whom changes affect, rather than being presented as a natural part of data's journey from collection to use.

The idea of 'splitters' and 'lumpers' is another way to think about decisions made during data preparation.[7] Discussed by James Weinrich

[2] Ryan, 'Expressing Identity', 356.
[3] Ibid.
[4] Frisch, 'A Queer Reading of the United States Census', 65.
[5] Velte, 'Straightwashing the Census', 85.
[6] Michael Brown and Larry Knopp, 'Places or Polygons? Governmentality, Scale, and the Census in the Gay and Lesbian Atlas', *Population, Space and Place* 12, no. 4 (July 2006): 225.
[7] James D. Weinrich et al., 'A Factor Analysis of the Klein Sexual Orientation Grid in Two Disparate Samples', *Archives of Sexual Behavior* 22, no. 2 (1 April 1993): 157–68.

et al. in relation to the measurement of sexual identities, 'splitters' and 'lumpers' describe two opposing ways to bridge the gap between individual-level characteristics and population-level identity categories. For 'splitters', the grouping of identity characteristics should reflect individual nuances and complexities. The consideration of granular details means that it is likely that many categories will exist, some with very small numbers. 'Lumpers', on the other hand, favour an approach to classification that involves the smallest number of possible categories. As a result, bigger groups are created and there exists greater diversity within each group. There is no obligation to pick a side as both approaches are useful, depending on the research question under investigation. For example, imagine a survey of 100 people that provides twenty possible response options for sexual orientation. If respondents were equally spread across these twenty groups that would equate to only five people per sexual orientation category. The granular level of detail would make it hard for analysts to assess differences between sexual orientations and might ultimately mean that analysis by sexual orientation is abandoned altogether. Some aggregation is required. However, it is equally unnecessary to jump from twenty categories to two. Depending on the spread of responses, analysis can find a middle ground between the crude binary of heterosexual/homosexual and a potentially unwieldy list of twenty response options.

It is at this point in the analytical process where decisions can erase the experiences of bisexual people. Several scholars have commented on the dearth of studies that explore the topic of bisexuality within the wider field of queer studies, an omission that is often continued in practices related to gender, sex and sexuality data.[8] Efforts to respond to the 'straightwashing' of data can perpetuate the invisibility of bisexual respondents when they are aggregated into a wider LGB group and data is reported as representative of lesbian and gay respondents only (even when, as is often the case, bisexual

[8] April S. Callis, 'Playing with Butler and Foucault: Bisexuality and Queer Theory', *Journal of Bisexuality* 9, no. 3–4 (13 November 2009): 214; Surya Monro, Sally Hines, and Antony Osborne, 'Is Bisexuality Invisible? A Review of Sexualities Scholarship 1970–2015', *The Sociological Review* 65, no. 4 (1 November 2017): 663; Robin Rose Breetveld, 'Forms of Bisexual Injustice: Bi, Being, and Becoming a Knower', in *Bisexuality in Europe: Sexual Citizenship, Romantic Relationships, and Bi+ Identities*, ed. Emiel Maliepaard and Renate Baumgartner (Routledge, 2020), 156.

respondents constitute the largest proportion of those who identify as LGB).[9] Erasure of unique insights from bisexual respondents means that when data is published it seems as if bisexual people are missing.[10] The decision to 'split' or 'lump' must therefore consider whether this will minimize or maximize the future utility of the data. However, at the same time, excessive 'splitting' potentially dilutes the impact of data to change the social world and, in some cases, might give analysts a reason to abandon analysis of LGBTQ experiences altogether.

A similar challenge is common in research that aggregates minority racial and ethnic groups to form the acronym BAME. Particularly in light of the Black Lives Matter movement, researchers, practitioners and activists have critically examined whether the aggregation of racial and ethnic categories does more harm than good.[11] BAME is not a real-world reflection of an individual's identity but collates a mixture of histories, geographies, cultures, races and ethnicities. Foluke Ifejola Adebisi has argued, 'By putting everyone not white under the BAME umbrella, we describe no-one' and that 'BAME, like the porous idea of liberal "diversity" pretends all marginalized people are interchangeable'.[12] The BAME acronym also fails to underscore that, in the UK, those who identify as a mixed identity, which can cut across multiple categories including white, have a far younger age profile and are therefore likely to increase as a proportion of the population in the future.[13] Ọrẹ Ogunbiyi and Chelsea Kwakye, in their book *Taking up Space: The Black Girl's Manifesto to Change*, note how lazy uses of the BAME acronym silence frank conversations about the unique ways in which Black people, in particular, experience racism.[14] In addition, the acronym excuses those

[9] For example, in the 2018/19 academic year in UK higher education institutions that returned data to the Higher Education Statistics Agency, 3.1 per cent of students identified as bisexual whereas just 1.2 per cent of students identified as a gay man and 0.7 per cent identified as a gay woman/lesbian, in Advance HE, 'Equality in Higher Education', 309.

[10] Westbrook and Saperstein, 'New Categories Are Not Enough', 548.

[11] For discussion, see Cecilia Macaulay and Nora Fakim, '"Don't Call Me BAME": Why Some People Are Rejecting the Term', *BBC News*, 30 June 2020.

[12] Foluke Ifejola Adebisi, 'The Only Accurate Part of "BAME" Is the "and" … ', *Foluke's African Skies*, 8 July 2019.

[13] Data from the 2011 English and Welsh census shows that the average age of people who identified as a mixed ethnic group was eighteen whereas the average age of those who identified as a white ethnic group was forty-one, in Office for National Statistics, 'Age Groups', Ethnicity Facts and Figures, 22 August 2018.

[14] Chelsea Kwakye and Ọrẹ Ogunbiyi, *Taking up Space: The Black Girl's Manifesto for Change* (London: #Merky Books, 2019), sec. The B in B(A)ME.

in positions of power as they can claim they are helping 'BAME people' when overall data for the aggregated BAME group improves, even if they are doing nothing to improve the lives of people who experience anti-Black discrimination.

I have struggled with aggregation challenges in my research with university staff and students. In general staff or student surveys not focused on a particular identity group, the majority of respondents tend to identify as a white British ethnicity with other respondents spread across a diverse mix of ethnic groups. Even when data collection methods enable respondents to identify themselves as a specific ethnic group, the level of detail provided is often aggregated during analysis to form new groups that contain a larger number of respondents (such as Black, Asian or BAME). Gillborn et al. have described how the provision of *too few* ethnic categories produces meaningless results but that the provision of *too many* categories can be almost as bad.[15] They discuss a study of educational attainment in a school that used more than seventy separate ethnic categories to analyse the data. As so many categories were analysed, many with only one or two pupils, it was impossible to have any confidence in the results and the school reported no significant differences in attainment between ethnic groups. In this situation, analysts faced two options: aggregate ethnic groups into larger categories – using labels such as Asian, Black, mixed and white – or do not undertake any analysis due to the discrepancy in the size of the groups and the risk of over-analysing responses from a handful of participants. In my work, I have erred on the side that any disaggregated analysis, even at the very crude level of BAME versus white, is better than nothing at all.[16] I hope that in drawing attention to differences between high-level groups, the normalcy of whiteness is exposed and other researchers are encouraged to disaggregate their data, as much as possible, in future investigations.

When aggregating data related to race, ethnicity, sexual orientation and trans/gender identity, analysts also face the question of what to do with participants who select 'other' or 'prefer not to say'. Browne highlights an

[15] Gillborn, Warmington, and Demack, 'QuantCrit', 172.

[16] Critics of the term 'BAME', which intensified following the raised profile of the Black Lives Matter movement in the UK in 2020, highlight the need to disaggregate all data on race and ethnicity so that it describes the experiences of Black respondents. In particular, specific injustices related to anti-Black racism in areas such as education, health and policing are often lost when subsumed by the wider category of BAME.

assumption, among researchers, that participants who answer 'other' or 'prefer not to say' to a question about sexual orientation are most likely not heterosexual/straight.[17] Contrary to this assumption, Browne notes that if a participant wished to hide their sexual orientation they would assumedly select the response option least likely to invite attention (in other words, 'heterosexual/straight').[18] Heather Ridolfo et al. describe four types of participant that most frequently disclose as 'other' in a question about sexual identity: those who do not identify as LGBTQ but take issue with terms 'heterosexual' or 'straight'; those who reject traditional terms used to describe sexual identity; trans respondents; and those experimenting with or questioning their sexual identity.[19] The many reasons why someone might identify as 'other' highlights the error of assuming that responses should be aggregated into a broader category of LGBTQ rather than heterosexual/straight. In their account of a sexual orientation and gender identity survey conducted in the Japanese city of Osaka, Daiki Hiramori and Saori Kamano describe the use of additional response options 'Don't want to decide, haven't decided' and 'I do not understand the question' to further assess what is lost under the umbrella options of 'Other' and 'Prefer not to say'. Their study found that a sizeable proportion (5.2 per cent) of respondents selected the option 'Don't want to decide, haven't decided'.[20] Decisions made about how to aggregate data, deliberated among analysts rather than the participants about whom the data relates, remind us of the many judgements made during data analysis that influence the results produced.

Small numbers

Decisions made during analysis are particularly impactful when data about LGBTQ groups involve working with small numbers. The Gender Identity in US Surveillance Group, convened by the Williams Institute (a research centre based at the University of California Los Angeles), has reported hesitancy among US national data collection agencies to gather data on groups that would likely comprise less than 0.5 per cent of the

[17] Browne, 'Queer Quantification or Queer(y)Ing Quantification', 242.
[18] See Peter Betts, Amanda Wilmot, and Tamara Taylor, 'Developing Survey Questions on Sexual Identity: Exploratory Focus Groups' (Office for National Statistics, August 2008).
[19] Ridolfo, Miller, and Maitland, 'Measuring Sexual Identity Using Survey Questionnaires', 122.
[20] Hiramori and Kamano, 'Asking about Sexual Orientation and Gender Identity in Social Surveys in Japan', 451, 455, 460.

total population.[21] With an estimate that around 0.3 per cent of the adult population in the United States identify as trans, this arbitrary benchmark would therefore rule out inclusion in national counts.[22] However, even when counted, the issue of small numbers can create several problems for LGBTQ people in how this data is analysed and subsequently used. Ridolfo et al. describe the asymmetrical size of minority and majority sexual and gender groups in nationally representative surveys, which means that 'the slightest degree of error can dramatically impact estimates'.[23] As described in the previous chapter's account of same-sex couple data in the 2000 and 2010 US censuses, errors can render *all* data collected about LGB people suspect whether the count seems too high or too low. When those responsible for data collection fail to meaningfully engage and instil confidence among communities covered by the count, design questions that are exclusionary and/or fail to fully acknowledge the diversity and complexity of gender, sex and sexual identities, small numbers can become even smaller than originally anticipated. Browne has warned that if, for whatever reason, counts fail to match assumptions about the size of LGB populations this might weaken arguments for LGB equality and justify further inaction.[24]

A partial solution to the challenge of small numbers is the quantification of qualitative data collected via open-text boxes in surveys. In their article 'What Sexual and Gender Minority People Want Researchers to Know About Sexual Orientation and Gender Identity Questions: A Qualitative Study', Leslie W. Suen et al. highlight a strong desire for write-in answer choices for questions on sexual orientation and gender identity among the seventy-four people who participated in their focus groups and cognitive interviews, particularly among participants of colour.[25] As data collected from open-text boxes do not usually map to items on an existing coding framework (for example, where female equals one and male equals two), analysis is potentially difficult and time-consuming, particularly in large-scale exercises

[21] Discussed in Currah and Stryker, 'Introduction', 6.
[22] Gary J. Gates and Jody L. Herman, 'Beyond Academia: Strategies for Using LGBT Research to Influence Public Policy', in *Other, Please Specify*: *Queer Methods in Sociology*, ed. D'Lane Compton, Tey Meadow, and Kristen Schilt, (Berkeley: University of California Press, 2018), 81–2.
[23] Ridolfo, Miller, and Maitland, 'Measuring Sexual Identity Using Survey Questionnaires', 114.
[24] Browne, 'Queer Quantification or Queer(y)Ing Quantification', 247.
[25] Suen et al., 'What Sexual and Gender Minority People Want Researchers to Know about Sexual Orientation and Gender Identity Questions', 2310.

such as a national census. The view that open-text data is hard to analyse has likely dissuaded researchers from including this type of question in surveys and diversity monitoring forms.[26] In defence of open-text questions, Gloria Fraser et al. have noted how this approach 'is rarely used as the sole measure of gender' and 'may represent a missed opportunity for quantitative researchers'.[27] Fraser highlights how an open-text question 'encourages participants to self-identify using as many terms as they wish' and points to studies that demonstrate efficient and accurate methods to analyse the data collected.[28] Approaches include the manual coding of a small sample of open-text data, for example 5 per cent of all responses received, followed by the use of analytical software to automatically code the remainder of the data (with a researcher required to address instances where the computer is unable to determine a match).[29] For example, Anna Lindqvist et al. tested the use of an open-text box for gender in a survey of 794 people and found that 98.99 per cent of responses were easy to code.[30]

The move to digital collection methods has also enabled researchers to utilize technology to assist with analysis. For example, NRS proposed using technology in the online version of Scotland's 2022 census that would predict and auto-populate a response option for people who started typing text in the write-in box for questions on sexual orientation, religion, nationality and ethnicity. NRS explained that the approach would improve the respondent experience (it would be easier to complete the census), data quality (auto-populated options would be matched to a coding list, reducing the risk of error from manually matching open-text data to a coding list) and efficiencies in the coding of the data (more of this work could be automated).[31] Respondents that provide an open-text answer would have the option of accepting the auto-populated response or writing-in something

[26] Guidance from the Gender Identity in US Surveillance Group noted that 'fill-in-the-blank questions do not work for surveys that include tens of thousands of people', in Gender Identity in US Surveillance Group, 'Best Practices for Asking Questions to Identify Transgender and Other Gender Minority Respondents on Population-Based Surveys', xv.
[27] Gloria Fraser et al., 'Coding Responses to an Open-Ended Gender Measure in a New Zealand National Sample', *The Journal of Sex Research* 57, no. 8 (November 2019): 980.
[28] Gloria Fraser, 'Evaluating Inclusive Gender Identity Measures for Use in Quantitative Psychological Research', *Psychology & Sexuality* 9, no. 4 (2 October 2018): 343–57.
[29] Lara M. Greaves et al., 'The Diversity and Prevalence of Sexual Orientation Self-Labels in a New Zealand National Sample', *Archives of Sexual Behavior* 46, no. 5 (2016): 1325–36.
[30] Lindqvist, Sendén, and Renström, 'What Is Gender, Anyway', 6.
[31] National Records of Scotland, 'Letter to Culture, Tourism, Europe and External Affairs Committee', 25 October 2019.

different. NRS noted that 'use of predictive text minimises errors such as spelling mistakes and abbreviations, which means clean codeable data is collected'.[32] As a result, the technology would maximize the usability of data collected on 'other' identity characteristics and potentially help address the issue of 'small numbers'.

Following news of NRS's plan to use predictive-text technology, media attention focused on the draft list of twenty-one response options that would appear in the open-text box for the sexual orientation question.[33] A list of sexual orientations was drafted with input from LGBTQ organizations and included options such as asexual, bicurious, pansexual and queer.[34] Organizations such as the Christian Institute and Catholic Church expressed opposition, while LGB Alliance (a UK trans-exclusionary campaign group) argued the proposal 'would suggest that other sexual orientations exist beyond attraction to the opposite sex, same sex or both sexes' and asked that the census not include the term 'Other sexual orientation' as a response option.[35] In April 2020, NRS wrote to the Scottish Parliament committee with oversight of the census to announce that it had decided not to use predictive-text technology for the sexual orientation question.[36] NRS's letter also explained that the decision did not apply to other census questions on identity characteristics and that NRS would continue to use predictive-text technology for questions on religion, national identity and ethnicity. Like the proposed list of sexual orientations, the work-in-progress options were based on previous censuses, other surveys, desk-based research and engagement with stakeholders, and included 116 religions, 274 national identities and 241 ethnic groups.[37]

Considering the complexity of categorizing religious, national and ethnic identities, what does this example from Scotland's census tell us about small

[32] National Records of Scotland, 'Letter to Culture, Tourism, Europe and External Affairs Committee', 18 December 2019.
[33] See Chris Musson and Ben Archibald, 'Scots Face List of 21 Sexualities to Choose from in 2021 Census Such as Gynephilic', *The Scottish Sun*, 29 October 2019; Gina Davidson, 'Scotland's 2021 Census to Have 21 Sexual Orientation Choices for Adults', *The Scotsman*, 31 October 2019.
[34] National Records of Scotland, 'Letter to Culture, Tourism, Europe and External Affairs Committee', 18 December 2019.
[35] LGB Alliance, 'Letter to Culture, Tourism, Europe and External Affairs Committee', 26 November 2019.
[36] National Records of Scotland, 'Letter to Culture, Tourism, Europe and External Affairs Committee', 2 April 2020.
[37] Ibid., 18 December 2019.

numbers and the use of technology to improve the accuracy and efficiency of data collected about sexual orientation? As discussed in Chapter 3, the census can bring into being a population that 'makes sense' to the cis/heteronormative majority and design-out LGBTQ lives and experiences that fail to match these ideals. In Scotland, the lack of concern about the use of predictive-text for questions on religion, nationality and ethnicity suggests that opposition expressed had less to do with the technology deployed and more to do with efforts to shore up a definition of sexual orientation linked to a fixed, binary and trans-exclusionary understanding of sex. There exists huge potential in the use of open-text data to expand opportunities for participants to self-identify and ensure that those who write in answers are meaningfully coded and counted. Furthermore, the automated coding of open-text data and use of technologies, such as predictive-text, means that the provision of 'more response options' does not exacerbate the problem of small numbers nor risk diluting the uses of LGBTQ data for action.

Big data

Kitchin describes big data as huge in volume (exceeding 1,000 gigabytes), high in velocity (collected in or near real-time) and diverse in variety (collected from multiple sources).[38] Big data can take many forms, from the information generated by the Large Hadron Collider to content shared on social media platforms, and differs from projects that make statistical inferences about a population based on a random or representative sample. For example, rather than infer how people use the London Underground from a sample of travellers, big data collects and analyses information from the travel cards of *all* five million passengers who use the system each day.[39] The volume, velocity and diversity of information involved in a big data project necessitates an approach to analysis that can operate at scale and is therefore associated with the use of automated computer algorithms. As a means to make sense of larger, more complex datasets (rather than a particular approach to analysis) an algorithm is a sequence of instructions designed to perform a specific task. Algorithms can run a range of analytical approaches, such as predictive analytics where machine-learning techniques

[38] Kitchin, *The Data Revolution*, 68.
[39] Ibid., 72.

are used to identify the likelihood of future outcomes based on historical data (for example, the use of past data to predict how people will travel on the London Underground next week).

In their description of critical data studies, Craig Dalton and Jim Thatcher observe that big data 'is never a neutral tool' but 'always shapes and is shaped by a contested cultural landscape in both creation and interpretation'.[40] Gillborn et al. also note how those enamoured by the powers of big data imagine an approach to analysis driven by machines where 'theories and human reasoning are rendered obsolete because the "numbers speak for themselves"'.[41] As Gillborn et al. highlight, in reference to data on race and education, the illusion that 'numbers speak for themselves' is harmful. Ideas and assumptions about big data therefore have implications for the analysis of gender, sex and sexuality data. Jen Jack Gieseking picks up this question in their examination of how big data relates to lesbians and queer women.[42] Gieseking describes the Lesbian Herstory Archive in New York, one of the largest collections of information about lesbian lives and activities in the world, and how the archive is too small and too qualitative to meet the demands of big data. For Gieseking, the Lesbian Herstory Archive exemplifies the mismatch between LGBTQ data projects and big data, as well as the risk that 'society's obsession with big data further oppresses the marginalized by creating a false norm to which they are never able to measure up'.[43] Failure to recognize the impacts of homophobia, biphobia and transphobia on the historical and contemporary practices of amassing data about LGBTQ people means that extant datasets are unlikely to match the size or complexity required of big data projects. As with the collection of data, the analysis and use of big data can present an impartial sheen that masks biases that disadvantage LGBTQ people. At a time where the attention of governments is increasingly focused on potential insights from big data and funding bodies channel limited resources to projects that foreground data on a grand scale, there is a risk that data projects about LGBTQ lives and experiences might lose out.[44]

In contrast, small data projects are best understood as the approach to data collection and analysis before the advent of big data. Small data studies

[40] Craig Dalton and Jim Thatcher, 'What Does a Critical Data Studies Look Like, and Why Do We Care?', Society & Space, 12 May 2014.
[41] Gillborn, Warmington, and Demack, 'QuantCrit', 167.
[42] Gieseking, 'Size Matters to Lesbians, Too', 150.
[43] Ibid.
[44] Kitchin, *The Data Revolution*, 28.

investigate a specific phenomenon, which might involve the collection of data about a sample to make inferences about a larger population. Kitchin notes, 'Small data studies seek to mine gold from working a narrow seam, whereas big data studies seek to extract nuggets through open-pit mining scooping up and sieving huge tracts of land.'[45] For example, a small data study of online dating among university students might invite a representative sample of 100 students to complete a survey where they describe their dating activities during the past month. In contrast, a big data project might capture data from dating apps on the smartphones of all students, in real-time, and cross-reference this information with data deduced about users' gender, sex and sexuality based on their online browsing habits. Giesking explains how big data's ambition for total knowledge (in this example, *all* dating activity data about *all* students) is antithetical to a queer feminist approach as it denies the situated nature of knowledge and the oppressive factors that shape data's collection and organization.[46] For these reasons, data about LGBTQ people and big data projects are not common bedfellows. D'Ignazio and Klein describe the framing of big data versus small data as a false binary and instead raise the question, 'How we can scale up data for co-liberation in ways that remain careful, community-based, and complex?'[47] There is no requirement to choose between big and small data, and future studies of LGBTQ people need to utilize the benefits of both approaches. An uncritical embrace of big data puts the collection and analysis of gender, sex and sexuality data at risk, with LGBTQ people engaged as participants in a game where the design of the rules means they are destined to lose. Yet, when done right, the use of big and small data creates opportunities to scale up small data studies, combine data from multiple projects and utilize techniques that automate practices and offer new analytical insights from gender, sex and sexuality data.

Useable findings

Analysis is the bridge between data collection and the use of data for action. How analysis is approached will depend on the research question under investigation and the intended use of results from the study. For example, if

[45] Ibid., 29.
[46] Gieseking, 'Size Matters to Lesbians, Too', 154.
[47] D'Ignazio and Klein, *Data Feminism*.

research into the experiences of LGBTQ young people in school is intended to change the public's attitudes to education then punchy, policy-focused findings will likely have a greater impact than a detailed and dense report. An outcome-oriented approach requires researchers to look into the future, imagine the results of their study and determine how analysis can maximize the impact of their project. Queer data projects need to identify an end goal, which positively impacts the lives of whom the data relates, and work towards it. This poses a troublesome question: are analytical decisions that smudge or even misrepresent the identities of participants justified if researchers are confident this will maximize a project's potential to positively impact lives and experiences? Browne provides an example of this dilemma in her account of undertaking a questionnaire of attendees at the 2004 Pride festival in Brighton and Hove, England. Browne developed the project with trustees for the Pride festival, who understood the research as a means to collect data that would ensure the festival's longevity. The survey was completed by 7,210 people and captured data about the multiple dimensions of respondents' sexual orientation, including identity, behaviour and relationships. The nuanced approach meant it was possible to analyse different dimensions of sexual orientation and how they related to the economic, social and cultural impacts of the Pride festival. However, this comprehensive analysis was not undertaken. Browne explains, 'The drive to generate usable findings meant that […] there was pressure to re-establish the institutional discourses of sexualities which have validity when seeking support from sponsors, local authorities and grant awarding agencies'.[48] Even with the available data and a researcher conscious of not making decisions that flatten LGBTQ lives and experiences, the data's purpose (as a means to attract future funding and support) influenced the approach to analysis.

Browne's account of the Brighton and Hove Pride questionnaire highlights a tension of queer data: the push-and-pull between the analysis of data in ways that positively impact LGBTQ lives and the purity of methods used to achieve these outcomes. To conclude this chapter I want to position the queer tension between methods and outcomes within a wider body of scholarship, in particular Spivak's concept of strategic essentialism.[49] Spivak

[48] Kath Browne, 'Selling My Queer Soul or Queerying Quantitative Research?', *Sociological Research Online* 13, no. 1 (January 2008): 200–14.
[49] Gayatri Chakravorty Spivak, 'Subaltern Studies: Deconstructing Historiography [1985]', in *The Spivak Reader: Selected Works of Gayatri Chakravorty Spivak*, ed. Donna Landry and Gerald M. MacLean (New York: Routledge, 1996), 214.

describes how strategic essentialism involves the temporary presentation of an identity group as possessing fixed, intrinsic and innate qualities (or essences) as a means to advance political goals. Temporarily overlooking differences, and fixing in time and space the characteristics of an identity group, can provide a platform to mobilize action and make rights-based claims. Strategic essentialism has informed the contemporary field of identity politics, in which individuals forge political constituencies based on a diversity of shared characteristics. As an approach to the analysis of LGBTQ lives and experiences, strategic essentialism invites both benefits and dangers. The construction of constituencies based on identity relies on reductive stereotypes, an erasure of differences and inaccurate accounts of homogeneity that are sometimes hard to extinguish once unleashed (particularly as essentialist traits tend to benefit the least marginalized within groups).[50] As data makes the journey from collection to its use for action, any dilution of diversity and complexity likely favours individuals already within touching distance to the ideal of equality and where sexual orientation is the only characteristic that excludes them from full inclusion.[51] Moya Lloyd approaches the topic of strategic essentialism from an alternative direction and, while arguing that identities do not possess essential characteristics beyond what is constructed in the social world, challenges the view that essentialist and anti-essentialist positions are oppositional.[52] For Lloyd, essentialism and anti-essentialism are intertwined and a product of political systems based on the advocacy of political representatives and activists who speak on behalf of others, an arrangement that requires actors to couch demands in terms of defined and demarcated constituencies. If we adopt Lloyd's account, the use of gender, sex and sexuality categories to advance political goals cannot escape the pitfalls of strategic essentialism. Those working with queer data therefore need to tread carefully to ensure that analysis of gender, sex and sexuality data, which reduces and homogenizes difference, is reversible and ultimately improves the lives of LGBTQ people.

* * *

[50] Marie Moran, '(Un)Troubling Identity Politics: A Cultural Materialist Intervention', *European Journal of Social Theory* 23, no. 2 (May 2020): 265–6.
[51] Spade, *Normal Life*, 44.
[52] Moya Lloyd, *Beyond Identity Politics: Feminism, Power & Politics* (London: SAGE Publications, 2005).

The analysis of gender, sex and sexuality data marks another moment where the fingerprints of people, about whom the data does not directly relate, are evident. The cleaning of data suggests something is amiss with the accuracy of what was collected. In many cases, cleaning fixes unconscious errors on the part of participants – typing the wrong digit or answering a survey question not intended for them. However, among these errors, there also exist conscious attempts to subvert the rules and expectations of data collection practices. For LGBTQ people, *queering* data collection methods can offer a response to not being counted or being forced to identify in ways that fail to reflect an authentic account of an individual's life or experiences. The cleaning, disaggregation and aggregation of data therefore offer analysts a backstop and further opportunity to ensure data presents an account of the social world that those behind the project wished to bring into being. In projects that seek to use data to positively improve LGBTQ lives, aggregation and disaggregation techniques can maximize the usability of findings and boost the potential for impact. For example, the strategic aggregation of identity characteristics into groups, to refute the claim that some identities are 'too small' for meaningful analysis, or temporary overlooking of difference to form larger constituencies and strengthen a basis for political action. Analytical techniques are further supported by advances in qualitative approaches and technologies. For example, the development of new methods has made it more feasible to let respondents describe their identities, in their own words, in diversity-monitoring forms.

Debates about cleaning, aggregation and disaggregation take place against the expansive backdrop of big data and algorithms, which simultaneously analyse data and execute decisions based on the results. However, as Gieseking observes, big data does not generally accommodate information about LGBTQ people, which is often small in scope, qualitative and peppered with gaps and absences. The historical particularities of LGBTQ lives and experiences, and the methods used to collect data, means that a broad-brush approach to big data is ill-suited to many LGBTQ data projects. Data has a history. For big data to authentically represent LGBTQ lives and experiences it needs to accommodate these differences. The navigation of difference is equally core to discussions about the usability of data and the role of strategic essentialism to maximize gains in existing political systems. Again we return to an examination of who is included and excluded, and who makes these decisions. As with data collection, the analysis of gender, sex and sexuality data demonstrates how seemingly neutral data practices can actually entrench existing biases, assumptions and inequalities.