CHAPTER 5

Eliciting and Transforming Data

> We must be careful not to confuse data with the abstractions we use to analyze them.
>
> William James (cited in Rice, 2007, p. iii)

The word "data" comes from Latin, where *datum* means something that is given – this term is also, incidentally, the singular for data in English. This linguistic root has led to a simplified understanding of what data represent. For many researchers, especially following the positivist tradition (see Chapter 2), data are a nonproblematized category, which includes aspects of events that are recorded and ready for analysis. And yet, there are many gaps between what is given in experience (i.e., events), raw data (the records or traces of events), transformed data (raw data that are processed), and data analysis. In the earlier quote, William James reminds us that raw data should not be confused with data that are transformed into categories, concepts, and codes; moreover, raw data always have the potential to disrupt expectations that have been shaped by theory (see also Chapter 3). The data we end up working with are far from "what is given" as a separate, static, and finite outcome. Data are transformed through research and analysis – through action (and interaction) guided by theories, questions, and interests (see Chapters 3 and 4).

Data are produced in various ways – from experiments and interviews to corpus construction – and take many forms, including numeric, text, sound, and image. We define data as the traces of experiences researchers use to address a research question. Data are the cornerstone of empirical social research. This is because they can capture experiences of breakdown or stumbling in which our theories are shaken by unexpected empirical evidence and our analytical methods prove insufficient (for the distinction between data and disruptive data, see Chapter 3). However, not all data provoke a rupture in our understanding; otherwise, the term would be relatively narrow, depending on what researchers find surprising or

https://doi.org/10.1017/9781009031066.006 Published online by Cambridge University Press

thought-provoking. However, we maintain that all data have the *potential* for disruption and, when adopting a pragmatism stance, we are particularly interested in exploring this potential. To this end, we argue, we need to be sensitive to the gaps or differences between phenomena, data, and analysis and, particularly in the context of new forms of data available nowadays, the importance of data construction as a recurring – not given and once and for all – process. Hence the pragmatist proposition we advance in this chapter: *Data are always transformations*.

In this chapter, we first review the different roles data have in research – as reality, as construction, as disruption – and propose to conceptualize it as a process, whereby records of a phenomenon (raw data) are collected and transformed (becoming transformed data). Importantly, for the purpose of this chapter we do not distinguish between the event as it happened (the noumena) and the experience of the event (the phenomena) and designate both using the same term. For instance, if an interview is an event, raw data can be the audio recording while transformed data start from the transcription, which entails choices (e.g., the system of transcription), and continue with the extraction of excerpts for detailed analysis and include word frequencies, which transform the raw data into numeric values. Second, we develop a classification of existing data collection methods and data types. Third, we discuss how technology has profoundly impacted data production, making large naturally occurring datasets available, including big qualitative datasets. Finally, we examine the opportunities and challenges of big qualitative data and discuss how pragmatism can help achieve these potentials and avoid the pitfalls.

5.1 What Are Data?

The vocabulary of data in psychology and social research is expansive, and it includes notions such as variable, information, fact, statistic, input, sample, population, finding, theme, and meanings. But, most of all, it includes the pervasive distinction between "quantitative" and "qualitative" data (Henderson et al., 1999; Holton & Walsh, 2016). The main difference made is thus between numerical and nonnumerical data. For example, household earnings, frequencies of the use of pronouns in a political speech, or the count of how often children interact with each other are recorded as numbers and, as such, are considered quantitative. In contrast, the drawings made by an artist, the sounds produced by an orchestra, or the words uttered to convince a friend to go bowling are considered qualitative. Data taking the form of visuals (still or moving), sounds, smells,

https://doi.org/10.1017/9781009031066.006 Published online by Cambridge University Press

touch, and, most of all, text (oral or written) are all considered qualitative. However, we suggest that this division is primarily analytical and often does not hold upon closer scrutiny. This is because numerical data always originate in "qualitative" experience, and all nonnumerical data can be counted and thus used in statistical analysis. In our examples, political speeches and school interactions are grounded in text, sound, touch, and so on; yet various aspects of them are quantified. Conversely, musical scores are a formal language that can easily be translated into numerical terms, but the sounds made by an orchestra and, in particular, the experience of listening to the orchestra in question are not (unless it is a digital audio recording). Quantifying the qualitative and qualitizing the quantitative are common processes within research. For most researchers, these transformations are part of the treatment or preparation of data for analysis (i.e., "moving" data from one state to another). We propose that data can change "state" as part of this processing (e.g., the movement from raw to transformed data and back again, including back and forth between qualitative and quantitative forms) and always afford further transformations depending on the research purpose and question. Thus, processes of data creation and transformation take center stage instead of being mere "data collection."

To note, there is a continuum between data transformation and data analysis but there is also a qualitative difference between them. While processing data from raw to transformed still keeps the focus on the data themselves (the main outcome is the new and transformed data), analysis moves between data and findings that answer a research question (and so the main outcome is the finding). Of course, researchers can also draw findings or conclusions from the process and outcome of data transformation but this is not their primary aim when working with data in the sense used in this chapter.

Even though data are the cornerstone of empirical research, they are often undertheorized. Methodology books usually present typologies of data rather than discuss what data are for or problematize practices of collecting and working with data. This omission goes back partially to the implicit definition of data as "what is given" and a general focus on data analysis rather than data collecting and creating. When data collection is a topic, it is primarily discussed in terms of samples and sampling methods (e.g., Devers & Frankel, 2000; Faugier & Sargeant, 1997; Marshall, 1996), although there has been some questioning of what "collection" actually entails (e.g., Backett-Milburn et al., 1999; Smagorinsky, 1995). Within these critical reflections, the notion of collecting data is regarded with

https://doi.org/10.1017/9781009031066.006 Published online by Cambridge University Press

suspicion given that it seems to suggest that data preexist the process of research; similarly, the established article section of "findings" to report research results suggests that conclusions are found rather than created. In contrast, pragmatism leads us to consider both data collection and data analysis as constructive processes and, in essence, data as resulting from the constrained engagement of the researcher with the world. In order to situate this latter view, we review below some common understandings of data in social science that build toward a pragmatist view of data as a process.

5.1.1 Data as Reality

Collecting data with the assumption that one is collecting aspects of a transcendental and universal Reality is peculiar to the realist traditions (in a narrow sense; Daston, 1992). Within the positivist paradigm, data are judged primarily in terms of their accuracy and truthfulness. The ostensibly independent quality of "good" data implies that it reflects reality and can be used to study phenomena in a direct, unmediated, and universal manner. This view ignores the constructed nature of data and removes the role of the researcher and the broader social and cultural context in shaping the research. This is not to say that there are no such things as "facts" or that this notion has no place in the pragmatist approach. In our post-truth and postfact context (Berentson-Shaw, 2018), truth must remain an essential criterion in science and public debate. The problem, however, is that the "data as Reality" approach is static and reductionist. It focuses on a narrow correspondence between theory and world, data and events. At the extreme, this approach equates the data collected with the phenomenon under study.

5.1.2 Data as Construction

The idea that data are constructed through research is widespread within the social sciences, especially among qualitative researchers operating within more constructionist paradigms (see Chapters 1 to 3). This approach places the researcher back into the relationship between data and world and focuses on the researcher's role in data elicitation and transformation (Carolan, 2003; Hagues, 2021). This goes beyond discussions of prompted or unprompted data, covered later in this chapter, and starts from the very decision to call specific information "data" and consider it relevant for a given research question. All the choices made following this (e.g., sampling, collection, transcription, codification, and analysis) reveal that

https://doi.org/10.1017/9781009031066.006 Published online by Cambridge University Press

data do not simply exist "out there," like a carving waiting to be extracted from the marble, but are produced as part of a human activity – the activity of research. Like all activity, research is guided by human interests (see Chapter 4) and mediated by culture (Wertsch, 1998); as a consequence, its tools and products are necessarily cocreations in the triangular relationship between the researcher, the phenomenon, and culture (including theories, the literature, and commonsense). Holding the view that research constructs data might sound relativist but it does not have to be – it is not "everything goes" (e.g., everything is data, all data are equal); rather, it foregrounds the mediated relation between data and events in the world.

5.1.3 *Data as Disruption*

Pragmatism acknowledges the constructed nature of data and the facticity of data (not in a transcendental or Real sense but as a truth of human activity). It emphasizes the potential of data to disrupt our expectations. This disruption entails both object (data) and subject (expectation). This understanding draws on the view that reflective thinking, or executive function, begins when we encounter obstacles or problems (see Dewey, 1903, 1997). In George Herbert Mead's words, "analytical thought commences with the presence of problems and the conflict between different lines of activity" (1964b, p. 7). Data that trigger analytical – and creative – thought typically originate in a conflict between our theories/assumptions and the new data encountered. Data as disruption are the unsettling of old views and thus the seed of new interpretations, which is the basis of scientific progress. However, one problem with adopting this position is that it downplays "nondisruptive" data. Nondisruptive data are essential for research because there could be no exceptions, surprises, or disruptions without established patterns, theories, and assumptions. No data are intrinsically disruptive or nondisruptive; it all depends on the research question, the theory, and the broader research assumptions. Thus, from a pragmatist standpoint, research should be clear about its guiding theories and questions and remain attentive to the disruptive potential of data.

5.1.4 *Data as a Process*

The guiding pragmatist insight we develop in this chapter is conceptualizing data as a process. This is the idea that data are dynamic rather than static, as something crafted rather than given. The path from data to analysis entails transformations. It is not only the case that data emerge from an

https://doi.org/10.1017/9781009031066.006 Published online by Cambridge University Press

initial transformation of raw data (traces or recordings of human activity) into "something" that can be analyzed (excerpts, categories, numbers) – but our relationship with data also changes in the process of research. This process should not be understood exclusively in terms of preparing data for either quantitative or qualitative analysis (Manikandan, 2010; Rubin & Rubin, 2005). It should not be confused with conducting more than one analysis on the same piece of data (for instance, using two types of statistics or employing both thematic and discursive analyses). Data as a process involves a continuous reflection on the kinds of transformations available – to consider the same data through different lenses, especially lenses that cut across the quantitative and qualitative divide. This potential for transformation rests in the fact that all raw data are *perspectival* (i.e., they can always be approached, understood, and acted upon differently, including within the same study). An online comment or an internet meme, for example, can be part of a much larger sample and coded numerically to identify patterns while, at the same time, being used for in-depth semiotic analysis. The raw data remain the same yet their "collection" and treatment are no longer static; the raw data are "processed" in various ways to afford various analyses.

In order to unpack data as a process, we need to distinguish between four distinct levels of data: (1) *the events* – the object of interest in the world, in all its complexity and tangled with other events; (2) *the record or raw data* – the traces of the events, such as archives, memories, survey scores, audio recordings, and digital footprints; (3) *processed or transformed data* – transforming the raw data to enable certain types of analyses, such as selecting excerpts, wrangling numbers, categorizing types, and quantifying qualities; and (4) *the analysis* – finding patterns or explanations by examining the transformed data using various analytic procedures, such as content analysis, correlations, discursive analysis, and linear regression. The events are facts that exist in the past. The raw data are the traces of events in the present. While the raw data are unchangeable (any change would result in transformed data), data transformation can move freely back and forth between transformations (e.g., quantification, categorization, sampling, or aggregating) and the raw data.

There are specific processes connecting, on the one hand, the event with raw data and, on the other hand, transformed data with analysis. These transformations define what data "are," and these processes are often multiple and even open-ended. For example, several steps can be taken to move from the event (e.g., the experience of going through a war) to analysis (e.g., the themes that describe this experience). For instance, memories

https://doi.org/10.1017/9781009031066.006 Published online by Cambridge University Press

need to be expressed orally or in writing, and voices need to be recorded and transcribed leading to numerous selections, choices, and transformations (e.g., placing the data in a table, using time stamps, level of transcription detail, number of variables). The idea of data as a process foregrounds that there are many forking paths between events in the world and data used in analyses.

For most of the twentieth century, a key challenge for research was getting records of events of interest. In quantitative research, obtaining survey data or conducting experiments was generally more time-consuming than running the analyses on the resultant data. While the challenge of finding the right participants or alternative sources remains in place (Macnab et al., 2007), the processes of recording, transcribing, and doing descriptive analysis have been simplified nowadays by a series of technological developments, not least the invention of computers and the general availability of research software. Today we are likely to record or hold too much raw data (entering the infamous "data dungeons"; Bauer & Gaskell, 2000) while the range of analytical methods have expanded and started to include highly technical procedures (e.g., natural language processing). This reversal – the relative accessibility of data, including rise of big qualitative data, and the difficulty of analyzing it – is accentuated by another type of gap, that between data collection and data analysis.

Traditional methods like experiments and surveys and, to some extent, interviews imply (or are particularly suited to) specific analytic strategies, such as comparisons of means, correlations, and thematic analyses. These traditional methods tend to collect data that are prestructured for a specific type of analysis. What is recorded in experiments and surveys typically takes a numerical form. For example, complex actions and interactions are reduced to categorical outcomes. Interviews allow for broader analyses, but they are challenging to scale up given the resource demands of interviewing and detailed transcription. This immediate connection between record and analysis, both quantitative and qualitative, obscures the processual nature of data because the time spent between data collection and analysis is reduced. However, digitalization has changed this dynamic. There are few "ready-made" methods for analyzing naturally occurring big qualitative data and the often-rich associated data. For example, a social media post has metadata (e.g., time, location, user details) and response data (e.g., replies, upvotes, circulation). Digitization means that the "space" between records and analysis widened, records are increasingly abundant, but they are also messy. Naturally occurring traces have high ecological validity; however, they often require extra processing to become suited for

https://doi.org/10.1017/9781009031066.006 Published online by Cambridge University Press

research (e.g., sampling, cleaning, joining, enumerating, and wrangling). This new context lends itself to pragmatist approaches (Chapters 1 and 2), systemic theories (Chapter 3), abductive questions (Chapter 4), and, as we shall see next, more complex and creative forms of analysis (Chapters 6 and 7). Most of all, they demand a deeper reflection on data elicitation and data types, the two topics we move to next.

5.2 Data Elicitation and Data Types

Traditionally, social sciences research has understood data as something to be "collected," either in the field or in the lab. Beyond the fact that this distinction has lost a lot of its meaning with the advent of research done online – with experiments moving into people's homes and interviews taking place with researchers still in their lab – it raises more fundamental questions about what constitutes a realistic or artificial context. Indeed, while the lab is often presented as a place of increased control, where researchers can test hypotheses in a quasi-vacuum, it is also decried as an artificial situation, where events might fundamentally differ from what happens "in real life" – because no human behavior is ever in a vacuum. However, while experimental procedures in a lab might be an extreme case, most traditional data collection methods involve some degree of artificiality, in the sense that the situation in which the data are gathered is at least partially created by researchers to produce said data. In other words, researchers using traditional methods do not simply "collect" data; instead, they elicit or create data (Hood et al., 2012).

A common distinction is between *naturally occurring data* (Reader & Gillespie, 2022) – termed "unobtrusive" or "nonreactive" (Reader et al., 2020; Webb et al., 1966) – and *constructed data*. Naturally occurring data are produced outside the research process (i.e., exist independently of any instructions from the researcher). They are also part of ongoing chains of events that make up the world; they are consequential outside the research process and shape the world of tomorrow (e.g., people making plans, flying planes, contesting identities, giving feedback, making friendships, and debating points of view). For example, online posts are naturally occurring data that can become data for research even if they were not created for research (something that raises particular ethical concerns; see the final section of this chapter and Chapter 8).

Using naturally occurring data for research typically entails either corpus construction or observation. In corpus construction, researchers search for preexisting naturally occurring data that can address their research

https://doi.org/10.1017/9781009031066.006 Published online by Cambridge University Press

question. This could include personal diaries (Gillespie et al., 2007), formal complaints and incidents (Van Dael et al., 2021), social media posts (Whittaker & Gillespie, 2013), cockpit voice recordings (Noort et al., 2021a), and even FBI records (Gillespie, 2005b). In observational research, researchers choose a context, situation, or event and collect the data as it happens, such as during a protest (Power, 2018) or in the aftermath of a disaster (Cornish, 2021). In both cases, however, the route from data to analysis is complex. While "traces of events" might be "naturally occurring," what ends up being analyzed is necessarily a constructed subset of the actual events. The corpus construction method entails numerous choices about what is and what is not in the corpus. It requires a delicate crafting of the corpus to suit one or more research questions. Equally, the observation method filters what ends up being analyzed through the experience, questions, and concerns of the researcher (Mulhall, 2003). In both cases, events themselves are too abundant for direct analysis. Researchers have to select, simplify, or describe. To this end, research questions (i.e., researcher interests) are critical because they provide criteria for isolating, extracting, and even abstracting data.

When researchers talk about constructed data, they often distinguish "prompted data" (e.g., interviews) and "controlled data" (e.g., experiments). In interviews, the aim is to prompt answers that are guided to varying degrees by the researcher (i.e., structured, semistructured, and unstructured interviewing; Kvale & Brinkmann, 2008; Qu & Dumay, 2011). While for most interviews, there is an assumption that respondents can freely produce their views, the opinions generated necessarily bear the mark of the interactional context of the interview itself (it is an inter-view; Farr, 1984). Experiments entail controlled data since the researcher tries to standardize the setting and collect structured reactions quantitatively. It is no surprise that control and standardization are defining characteristics of the experimental method (Wolfle et al., 1949). Surveys are somewhere between prompted and controlled data, depending on how they are constructed. For example, surveys on misinformation tend to be quite controlled (e.g., controlling the stimuli, the sample, how accuracy is assessed), whereas surveys of opinions entail less control (they can be very narrow inventories of opinions, but they do not necessarily control much beyond the response format) – methodological differences that raise the problem of expressive responding or the deliberate production of insincere responses in misinformation surveys (see Schaffner & Luks, 2018).

One of the main limitations of the distinction between naturally occurring and constructed data is that it suggests that some data exist "out there"

https://doi.org/10.1017/9781009031066.006 Published online by Cambridge University Press

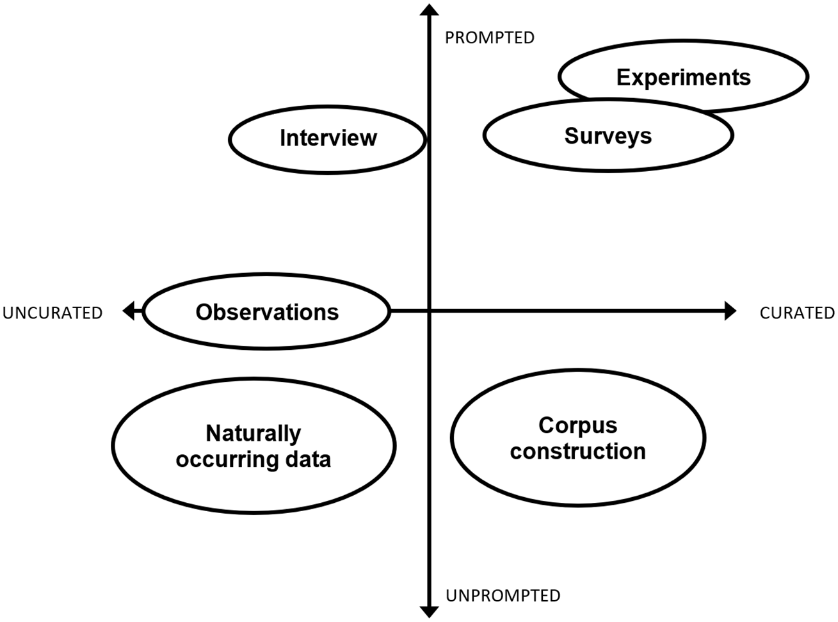

Figure 5.1 A typology of data elicitation

and researchers do little to influence them, while others are fabricated for research. However, we can think about these two categories in more subtle terms. For example, one of the most relevant criteria is related to the degree of control of the researcher. This can refer to control over the situation of data collection or control over the content of the data itself. Following this new distinction, we can talk about prompted or unprompted and curated or uncurated methods of data elicitation (see Figure 5.1 for details).

In Figure 5.1, *prompted or unprompted* refers to how much the situation is controlled and created by the researcher. Experiments offer the most control and construct more "artificial" situations for participants, followed by surveys and interviews. Observations are on the midline because, in traditional observation techniques, the researcher contributes to a certain extent to the situation. But, especially in covert observations, one could argue that the influence of the researcher is minimal. *Curated or uncurated* refers to the extent to which data are selected for inclusion into the dataset (e.g., how strict the criteria are about what constitutes data). Experiments and surveys are the most curated, as they collect only limited and predefined information. Constructed corpora are also heavily curated, as the

https://doi.org/10.1017/9781009031066.006 Published online by Cambridge University Press

data are collected based on clear criteria for inclusion/exclusion. On the other hand, interviews and video observations are the least curated, as they offer little control over what ends up in the transcript or video record.

One general observation when considering this typology is that most methodological attention tends to be focused on the opposing quadrants of *uncurated and unprompted* and *prompted and curated*. These correspond, largely, to the naturally occurring and constructed data categories mentioned earlier. And yet naturally occurring data does not necessarily have to be uncurated, as is the case for instance in corpus construction, where much effort is put in selecting relevant data to answer a research question. Some of the biggest expansions in data sources, discussed in more detail in the following section, concern what we call here (uncurated) "naturally occurring data." This is any source of naturally occurring data that is either taken as a self-contained dataset (e.g., analyzing a whole archive) or selected on arbitrary terms (e.g., analyzing the last three months of conversation on a forum). These can be closed-circuit television (CCTV) footage, interactions on social media, entire diaries available to the public, black box recordings, or public speeches. Because these data are both unprompted and uncurated, they land themselves particularly well to zooming in and out through multi-resolution research (see Chapter 7) without being the only data that can afford such analytical movements (as we shall see in Chapter 6).

In the end, when it comes to typologies, we need to move past simple dichotomies like qualitative and quantitative data and, in many ways, the field of methodology is doing so by operating increasingly with new concepts such as *structured and unstructured data* (Bolla & Anandan, 2018; Fong et al., 2015; Gillespie & Reader, 2022; Shabou et al., 2020). While structured data are generally quantitative and unstructured data are generally qualitative, these new notions shift the emphasis away from a binary toward the idea that there can be degrees of structure. They also shift focus from the implicit (and misleading) assumption that a specific type of data necessarily calls for a specific method of analysis. Instead, the focus is on "datafication" as the process of structuring records. In practical terms, structured means anything that can be neatly organized in a table or database, and unstructured is anything that cannot be put in a table without losing essential information. Of course, there is always a grey area because there are many degrees of structuring (arguably, all digital data, even when qualitative, are structured in 1s and 0s). For example, tweets can be put in a table along with related information (e.g., time sent, number of likes, retweets, responses), but their textual content remains unstructured (except in so far as it is digital). Our proposal is that the movement

https://doi.org/10.1017/9781009031066.006 Published online by Cambridge University Press

between structured and unstructured (what we call data transformation or data as a process) has enormous research potential. Often this move is from unstructured data to at least partially structured data through automatic or manual coding (quantitizing; Chapter 7). However, it can also be in the reverse direction, to recover the rich qualitative particulars underlying previously structured data (qualitizing; Chapter 7).

In summary, we need to rethink old terminology regarding data. In particular, we need a deeper reflection on the role of researchers in eliciting data and how much control they have over the content and context of data collection. Distinctions between prompted or unprompted and curated or uncurated add nuance and help us navigate where the abundance of new data, especially big qualitative data, is "coming from." The structured–unstructured terminology points us to processes of working with data, which is a significant advance. It is increasingly important to move beyond types of data and toward understanding how data are selected, reshaped, and used. Texts analyzed only for word frequencies are structured data. Images classified by humans or an algorithm are also structured data. But both texts and images are also unstructured qualitative data.

What is essential for our argument is that, in both cases, the unstructured (raw) form remains, *enabling the researcher to move between unstructured and structured* in the process of working with data. Both unstructured and structured data can be subjected to pluralistic forms of analysis, and this process is helped by the fact that unstructured data can always be structured and, in many cases, structured data can be unstructured. We need, however, more opportunities for the latter. For example, text can be analyzed systematically to reveal a high-level pattern while retaining the ability to zoom into individual quotes. But structured data such as survey ratings do not typically enable zooming down into the thoughts/utterances underlying the given rating except if the research is set up in such a way as to allow this (Rosenbaum & Valsiner, 2011). Luckily, as we shall discuss next, new types of data are extremely rich and can be structured in more than one manner, allowing for both quantification and exemplification and, as such, enabling us to gain both quantitative rigor and qualitative subtlety.

5.3 Big Qualitative Data

Quantities of data are increasing exponentially. Things, people, and organizations increasingly produce traces (potential raw data) as a byproduct of their activities. Digitization has made these data easy to store, search, and (security and privacy issues aside) share. This increase in data creates

https://doi.org/10.1017/9781009031066.006 Published online by Cambridge University Press

opportunities but also pitfalls (Salganik, 2019). From the use of smartphones to social media participation, most of us – but not everyone (Van Dijk, 2020) – leave numerous digital traces, and these records can easily be collected as data and analyzed with or without our consent. The ascendancy of big data (Yaqoob et al., 2016) and big data techniques (Foster et al., 2016), and their spread in the corporate domain, testifies to the digital data boom. These data and techniques will revolutionize traditional research methods (e.g., surveys, experiments, and interviews). Previously, numerical data collected on a piece of paper or typed in a computer were the only record available for analysis from an experiment (bar the existence of fieldnotes from the experimenter). Nowadays, there may be high-quality audio and video footage of participants during the experiment (Ross & Vallée-Tourangeau, 2021). This abundance of new data means that the gap between raw data and analysis is increasing because we increasingly need to decide how to select and structure the data for analysis. The traditional record, meant for a particular analysis, is becoming rare. We increasingly face a multitude of choices, both in terms of what we consider raw data and in terms of how we transform this into analyzable data, which is both exciting and challenging.

In practical terms, increasing digitization has three consequences for social research: new types of data, more behavioral data, and increasing quantities of data.

5.3.1 Increased Access to More Types of Data

The new types of data include social media data, video footage, live interactional data, and digital archives. While some of these have been around for decades, they have become increasingly important for researchers, and there is an upsurge in digital data and digital research (Brynjolfsson & McAfee, 2011; González-Bailón, 2013; Mayer-Schönberger & Ramge, 2022).

Social media data, including text and images (e.g., memes and emojis), conversations or posts, and a wide range of online behaviors (e.g., liking, linking, forwarding, following), are a rapidly growing data type. Whereas today we are taking for granted the diversity of social media platforms, we need to keep in mind how recent many of these platforms are and how they are transforming both individual lives and society. For instance, Facebook was founded on February 4, 2004, the first tweet dates from March 21, 2006, the first Instagram post was released on July 16, 2010, and TikTok was launched in China in September 2016. At the time of writing, the new social media apps include Supernova, Locket, Sunroom,

https://doi.org/10.1017/9781009031066.006 Published online by Cambridge University Press

and PearPop. No doubt, many of these new platforms will fail, and new platforms will arise – which goes to show how rapidly these technologies and social spaces are developing. Our point is that researchers are increasingly presented not only with social data but new social phenomena that did not exist before the early 2000s (e.g., de Saint Laurent et al., 2020; Stahl & Literat, 2022).

Video footage is a type of data coming from people sharing videos online and from organizations (e.g., broadcasts, CCTV, training, teaching, and news). With video footage, we can gain insight into the mundane aspects of life (whereas in the past, only special events were filmed) and gain a broader range of perspectives on a single event (e.g., videos of key events like the US Capitol Hill riot; Tynes, 2021). Since people are increasingly used to being filmed, they react less to it (e.g., CCTV), and thus it is increasingly likely that there will be video footage of critical events. Natural disasters, social uprisings, unethical behavior, and whistleblowing increasingly produce digital traces. Thus, retrospective verbal accounts will increasingly give way to rich audio-visual data that provide insight into events as they happened.

Live data come out of people answering requests to record reactions or fill up surveys while events are taking place or at specific moments during the day, but they can also include a direct record of activity (e.g., collecting browser data, real-time messaging, tweeting, recording movement on the GPS). This "live" aspect of the data, when combined with computing, can lead to real-time analysis of behavior in ways that feed back into the phenomenon itself (e.g., monitoring hate speech online; Paschalides et al., 2020). This is in stark contrast with traditional research practice, which takes months or years to get from data collection to results. In the case of live data, the research participant does the observation/data collection more or less willingly (e.g., participants install an app for recording their web activity, online interactions, or personal experiences; Christensen et al., 2003).

Digital archives are not new. From the invention of the first computer in the 1940s to the creation of the Internet in the 1980s, we have been accumulating digital archives. What differs today is the ease of access and the analytical possibilities opened by natural language processing and object recognition techniques. This has enabled archives to continue growing while remaining almost instantly searchable. Archival data are also becoming richer and more multifaceted with extra data such as time stamps, document history, and email communications and messages about documents (Beer & Burrows, 2013; Falk, 2003). Thus, digital archives are not

https://doi.org/10.1017/9781009031066.006 Published online by Cambridge University Press

only becoming bigger but they are also increasingly combining and stitching together multiple types of data (both structured and unstructured).

5.3.2 Increased Access to Behavioral Data

Another opportunity for researchers dealing with new data, particularly in its digital forms, is that it allows them more direct access to behavioral information. While traditional methods such as surveys and interviews are usually based on self-report (people reporting behavior), and experiments construct behavior in somewhat artificial environments, online spaces offer easy access to behavioral data. This is not limited to online behavior. The video footage, live data, and digital archives increasingly include data on offline behavior (e.g., purchases, video footage of daily life, images of important events, and reports of medical error). Although there is a debate about the relationship between online and offline behaviors (Kim et al., 2017), it is increasingly recognized that there is no relation. The digital realm increasingly contains traces of our nondigital behavior. And, although it is imperfect behavioral data, it must be compared to the alternatives, such as recollections of behavior and declarations of behavioral intent.

5.3.3 Increased Large Quantities of Unstructured Data

Big qualitative data also challenge the old assumption that quantitative (structured) data are relatively easy to accumulate in larger quantities and thus offer breadth, while qualitative (unstructured) data take more time to gather and thus offer depth. Contemporary researchers often have vast amounts of unstructured and unanalyzed data, creating new opportunities and challenges. Key questions include: How does one preserve some depth when the data are too vast to analyze manually? How does one best simplify one's data by structuring and quantifying them? How does one select what information to keep and disregard? Big data, especially of the unstructured type, make new methods possible (e.g., natural language processing; Hirschberg & Manning, 2015) while, at the same time, rendering other methods dated (e.g., the statistics taught traditionally to psychologists are not suited to big data).

The aforementioned three questions, taken together, force us to rethink our research practices. And yet most methods, books, and courses trail behind these challenges and generally stay at the level of generic tools like observations, experiments, surveys, and interviews that are presented as the

https://doi.org/10.1017/9781009031066.006 Published online by Cambridge University Press

methodological canon. Moreover, these books and courses also implicitly, if not explicitly, work on the assumption that some methods are better than others. For example, in the social science domains that try to emulate the natural sciences, experiments are considered the gold standard for obtaining causal information and making predictions (two key attributes of positivism; see Chapter 2). Besides the fact that the value of a method always depends on the goal we are trying to achieve with it, the reality that we now can access large amounts of data about actual behaviors should make us rethink our hierarchies, preferences, and assumptions.

From a pragmatist standpoint, the abundance of new data, especially naturally occurring big qualitative data, provides a valuable opportunity. This is because these big qualitative datasets are more likely to lead to useful knowledge. Not only are these data "big" and high-powered but they are also high in validity. These big qualitative datasets will have a huge impact on social science research because they are unprecedented in their quantitative scale, rich in their qualitative details, and have high validity due to their proximity to human behavior. These data are part of human life (i.e., naturally occurring rather than induced or artificial); thus, by analyzing them, researchers can get close to (and contribute to) what actually happens. The pragmatist point is this: To create useful knowledge for humans, it is recommended to start with what humans are actually doing.

5.4 Accessing Data and Ethical Challenges

The opportunities of abundant new data, particularly of the unstructured and digital kind, need to be understood in the context of several constraints, especially access and ethics. In this subsection, we take a closer look at the main ethical challenges surrounding big qualitative data and, connected to this, the question of how data are accessed and by whom.

5.4.1 Accessing Data

To understand issues surrounding access, we should first review some key types of data sources. *Repositories*, for example, are existing datasets that have been curated by other researchers and are made available for secondary research (Pinfield et al., 2014). Online *archives* are websites where data, usually naturally occurring and not gathered for research, are shared (e.g., parliamentary debate transcripts on government websites; de Saint Laurent, 2014). In both cases, existing data are made available. *Websites and platforms* offer ways of collecting data online where the medium is the data

https://doi.org/10.1017/9781009031066.006 Published online by Cambridge University Press

(e.g., social media, collaborative platforms; Kordzadeh & Warren, 2013). Finally, moving toward explicitly constructed data, we have *apps and programs* aimed at collecting data with the participants' consent (Zydney & Warner, 2016). For these, the data are actively collected by the researchers. While information from many of these sources can be freely and relatively easily accessible, this is not always the case. The optimism surrounding big data and the enthusiasm for big data research are tempered by the fact that most social media platforms restrict researchers' access to downloading and processing their content. Some of these restrictions are a response to past unethical practices (see, for instance, Facebook's reaction to the Cambridge Analytica scandal; Brown, 2020). But access is also restricted for corporate reasons. Under the banner of protecting users' privacy, companies have effectively been given private control over the conversations that make up the public sphere – and researchers who want to examine what is occurring in these online public spheres risk being locked out.

There are two main dimensions when it comes to accessing data in general and online data in particular: *extraction* and *authorization*. For online data, extraction can take several forms (see also Edwards et al., 2020). First, manual extraction can be done by downloading files manually or copying text and images from a social media platform into open documents. Second, Application Programming Interfaces (APIs) can be used to query databases directly, enabling downloading data at high speed and in a structured form. Third, web scraping can be used to automatically collect data from websites where there is no API access. This entails computer algorithms that simulate being a user of the website, opening pages, and then programmatically extracting data. Many platforms try to prevent web scraping, but it is widely used (e.g., it is how search engines construct their databases of the Internet).

Another issue for data access is authorization (Asghar et al., 2017). Sometimes the data themselves are open access, which means that everyone can have access to them (e.g., Reddit, Wikipedia). Most of the time, however, some form of authentication is needed (i.e., the researcher must register or apply for access). Some platforms use a mixture of both (e.g., the download is slower without authentication). Other times, researchers are asked to pay to access data, especially if downloading at scale. However, charging researchers for noncommercial access is a questionable practice because researchers can claim fair use (i.e., analyzing the data is in the public interest and not for commercial purposes).

As a result of these constraints, data access is often limited in practice to those researchers who have the technical skills and/or financial means to

https://doi.org/10.1017/9781009031066.006 Published online by Cambridge University Press

access it. A lot of studies are done using manual data collection, which limits the amount of data collected and, in addition, many platforms are deliberately difficult to copy and paste from. Consequently, less representative platforms get overstudied and third parties' profit from selling data to those who can afford to buy it (e.g., one can pay to have a dataset curated for you from social media platforms). For the latter, the main clients are companies who want to analyze markets or reputations, so the tools are oriented toward market research and the prices are often prohibitive for researchers.

5.4.2 *Ethical Issues*

Social scientists have an ethical responsibility to study big qualitative datasets. First, it is important to study these data because online interactions are increasingly central to people's lives. Online interactions can have significant consequences for individuals (e.g., support groups, conspiracy theories) and society (e.g., misinformation and politics; Kyza et al., 2020). At the same time, we should refrain from blindly assuming that studies about what happens offline apply to online behaviors and vice versa. Second, we should access this kind of data in order to propose a critical and pluralistic approach to it. Most digital data are currently analyzed by data scientists/computer scientists and/or for commercial purposes. Their focus is also primarily on structured data. Psychologists and social scientists have a lot to offer this field of research in terms of theoretical, methodological, and ethical reflections. Most of all, social scientists can contextualize the data themselves and the research practice they are part of in social, political, and historical terms, which is too often missing in big data investigations. For instance, social scientists have questioned the use of algorithms assumed to be neutral when, in reality, they are trained on data that are never "neutral" (Martin, 2019; O'Neil, 2016; Stinson, 2022). Third, we should access this kind of data to hold large social media platforms accountable. For example, there is considerable research interest in misinformation on social media but, because Facebook's algorithms are not accessible, researchers cannot independently verify Facebook's claims about the success of their practices in this area.

However, there is a myriad of ethical challenges for social scientists using big qualitative data. The specific issues vary depending on the details of the study. Nevertheless, key questions include: How private are the data (e.g., do people share their innermost thoughts or rate washing machines)?

https://doi.org/10.1017/9781009031066.006 Published online by Cambridge University Press

Who else has access to the data? Are the data to be shared? If they are "stolen" data, have they been made widely accessible by others? How relevant are the data (e.g., are they private but about a pressing issue)? How much personal information do they contain (e.g., can people be identified, even indirectly)? What will be done with the data? Will quotes and extracts be shared in publications? Will the data be made available?

Admittedly, many of these questions do not have clear-cut answers and, as such, require considerable moral deliberation (see Chapter 8). But they should be asked by researchers before they engage in collecting and analyzing online data. As with any other type of data, safeguarding practices include *consent*, *privacy*, and rules around *data sharing* (Smith et al., 2016). Yet online data add extra concerns to each category. First, researchers often cannot ask for consent from the participants but can be asked to make the existence of the project public so that participants can opt out (which is quite difficult in practice). Good practice in this regard is to ask oneself how reasonable it is for the participants to expect their data to be publicly accessed and analyzed (e.g., is it an open platform that people know to be "watched," like Twitter, or a small forum that is assumed to be private?). When it comes to privacy, removing identifying characteristics is often not enough. If the researcher publishes quotes and extracts, he or she should consider whether it may be advisable to modify the quotes sufficiently so that they cannot be searched for online. Finally, on data sharing, if the data collected were originally public, it makes sense to share the curated dataset (which also enhances replicability). But one should ask: How "curated" is the dataset? For example, even if the data are public, if your dataset looked at the behavior of a few isolated platform users over ten years, it might be questionable to share it. Also, one should ask: What are the legal requirements? For example, in Switzerland (in 2022), data obtained by scraping open websites cannot legally be shared; only the code used to obtain them may be made public.

For these reasons, ethics committees often struggle when researchers work on "new," big, and/or digital data (Ferretti et al., 2021) because there are few specific policies in place or they are country-specific. In general, informed consent does not readily apply, underaged participants cannot be reliably screened out, data may be stolen, and public platforms have different types of users and create different expectations of privacy not easily known by nonspecialist ethics committees (e.g., users of Twitter and Reddit usually know that their data are public, Facebook users may be less aware of the extent to which their data are public).

https://doi.org/10.1017/9781009031066.006 Published online by Cambridge University Press

5.5 The Potentials of Data

In this chapter, we argued that data entail a record of real-world phenomena collected, stored, and transformed to enable analysis. Instead of static instances of one "kind" or another, a pragmatist conceptualization focuses on data as a process of construction taking place between researcher and world, a process that can disrupt established theoretical views or empirical patterns. Instead of the traditional dichotomy between quantitative and qualitative research, there is a new focus on the role of the researcher (captured by whether data are prompted or unprompted, curated or uncurated) and the "movements" between structured and unstructured data. This idea of not just transforming data but retransforming them (moving back and forth between quantities and qualities) is particularly suitable for dealing with the abundance of "new" forms of data and, in particular, the digital big data boom that is currently shaping psychology and the social sciences. While these data come with ethical and access challenges, they represent great opportunities for zooming in and out of datasets that are rich, multidimensional, and often surprising. This is not meant to say that multiple forms of analysis are applicable only to online or big data. As we will see in Chapter 6, what the current data context mainly brings to the fore is the widening gap between data and analysis. Instead of predetermined and linear relationships between the kind of data being collected and their processing, we are left, due to current advances, more aware of our methodological choices in transforming data. Key among them is the possibility of *combining* structured and unstructured data (Chapter 6) and, finally, of considering in research *the same piece of data* as (potentially) structured *and* (potentially) unstructured (Chapter 7). These practices call, naturally, for mixing methods of analysis, a call that is highly congruent with a pragmatist approach.

https://doi.org/10.1017/9781009031066.006 Published online by Cambridge University Press