6. Unlocking big data: at the crossroads of computer science and the social sciences

Oliver Posegga

1 INTRODUCTION

During the past two decades, we've experienced fundamental technological advances that are already having a profound impact on all levels of society – one that is likely to persist. A significant portion of social interaction has shifted from the physical to the digital world, as has the production, consumption, and dissemination of information. In the digital world, every interaction and information is represented by data, which have become one of the most valuable resources of the twenty-first century.

Unlocking full access to this resource promises progress in established fields of research, not exclusively but especially in the social sciences, and insights into novel phenomena that arise from the digital transformation of society. At the same time, the growing availability of methods to process and analyse such data, most notably from the domain of machine and deep learning, put the realisation of this promise within reach. In light of this potential, the early twenty-first century has been referred to as a 'golden age' for the social sciences, which is based on a 'measurement revolution' that significantly expands their methodological repertoire and empirical prowess (Kleinberg, 2008; Watts, 2007).

While there is little doubt that this vision has a bright outlook, it is also true that realising its potential may be more complicated than anticipated. Despite a considerable amount of progress in specific areas, the general verdict seems to be that, to date, we have learned little about the social mechanisms that underlie the phenomena these novel data sources allow us to observe (Lazer et al., 2021; Watts, 2007). This sobering assessment of the last 20 years of research raises a simple question: *Why?*

The answer is, of course, complex and the subject of multiple seminal papers (Hofman et al., 2021; Lazer et al., 2020, 2021; Ruths & Pfeffer, 2014; Wallach, 2018; Watts, 2007). Contributing to the discussion and working towards a meaningful response requires taking a closer look at the underlying problem and its components. The problem is that even with an abundance of *data*, a growing portfolio of appropriate *methods*, and a rich body of *theories* about relevant phenomena of interest to the social sciences, we often fail to develop *research designs* that address meaningful questions and lead to valid answers. Thus, we miss the opportunity to unlock the full potential of the novel data available for research.

The problem is not merely technical or methodological but also interdisciplinary and fundamentally related to the goals of the associated disciplines, especially the social sciences and computer science. Improving the state of the art requires understanding the type of data; the assumptions that underlie the methods used to collect, process, and analyse it; the role of the theories about the studied phenomena; and their interdisciplinary nature.

This chapter provides a brief introduction to the problem, establishes the role and nature of data in their context, and outlines the obstacles we face in addressing contemporary research questions at the intersection of computer science and social science. Before addressing the characteristics of digital data in the context of human behaviour, and particularly social behaviour, I will briefly establish their origins and importance in the following.

2 ON THE ORIGINS AND RELEVANCE OF NOVEL DATA SOURCES

One reason so many novel data sources have emerged is the digital transformation of nearly every aspect of society. Throughout this ongoing process, the nature of social interaction and the production, consumption, and dissemination of information has significantly changed. In essence, modern information and communication technologies such as the internet, social media, and smartphones provide ubiquitous and continuous access to networks of people and information. Moreover, such technologies simplify the production of multimedia content, which can be produced with a few taps on the screen of a smartphone. Information can be shared with the click of a button and uploaded from almost anywhere at any time to a constantly evolving ecosystem of social media platforms with billions of users across the globe. Once uploaded, information becomes content that can be consumed, shared, and reacted to by other users, who might stumble upon it in feeds curated by platform providers or by receiving it from members of their personal networks.

This continuous flow of information and the interaction between individuals gives rise to large-scale, digitally enabled social networks. Below their surface runs a complex stack of technologies that share a property in common: to operate properly and satisfy their providers' goals, they must log the activities and interactions of their users and store the information they share. As a result, huge amounts of data resembling traces of digital interaction and shared content accumulate in the databases run by platform providers, and become an integral component of their services. Almost as a by-product, the data produced in this context promise deep insights into the fabric and patterns of human behaviour and social interaction on an unprecedented scale (Kleinberg, 2008).

Such data promise to advance our understanding of phenomena that have been studied since long before the rise of social media and similar drivers of digitally enabled social networks. Studies of this type often revolve around whether the online world mirrors the offline world or if it is subject to novel dynamics that are typically not observed offline (Jungherr, 2018). Regardless of the outcome, the results of such studies are likely to offer ground for future research. For example, in research on small-world networks, Travers and Milgram (1977) conducted a comprehensive and laborious experiment. They found that the average person in North America was, on average, only six or fewer social connections – that is, in a chain of the proverbial 'friend of a friend' – away from any other person. These 'six degrees of separation' and other characteristics of small-world networks have been the subject of a plethora of research in online and offline contexts. With significantly less effort compared to Travers and Milgram, for example, researchers at Facebook found that the average member of their platform is connected to any other member by a distance of 4.7 steps (Ugander et al., 2011).

While confirming small-world properties, findings such as these raise further questions regarding the nature of social connections online. How, for example, do friendships on Facebook differ from the social relationships Travers and Milgram studied long before the emergence of social media, the internet, or even mobile phones? Following this thought, the interesting questions often emerge once we begin to ask how social behaviour might have changed in the presence of digitally enabled social networks and how online and offline networks are intertwined. Such questions, and the study of social behaviour online more generally, can lead to research on novel phenomena that emerge from the complex interplay between the online and offline worlds.

One prominent area of research in this context aims at understanding how the media system has transformed given the many ways in which the production, dissemination, and consumption of information, including news, has changed with the advent of social media and related technologies (Chadwick, 2017; Jungherr et al., 2019). For example, the growing importance of digital and digitally born media has altered the role of traditional mass media. The once almost exclusive purview of traditional media to promote topics, frameworks, and speakers to the public is now routinely challenged by novel actors and individual users empowered by the open nature of the contemporary media system (Jungherr et al., 2019). Digital media also plays a fundamental role in political campaigning and elections (see Jungherr, Chapter 25 in this volume).

Another area of research investigates phenomena arising from collective behaviour afforded, in particular, by social media platforms and technologies that allow individuals to organise and coordinate in various ways. Prominent examples include crisis management (Eismann et al., 2016, 2018, 2021; Reuter et al., 2018), riots (Bastos et al., 2015; Panagiotopoulos et al., 2014; Segerberg & Bennett, 2011), and even revolutions (Bennett & Segerberg, 2012; Wolfsfeld et al., 2013).

While the data described above are a by-product of individuals using the platforms and services offered by organisations, the same technologies driving them can be used to create controlled settings to elicit data for academic purposes. For example, smartphones – which function as mobile sensor platforms – can be used to track individuals' proximity and measure the frequency and strength of their social relationships while controlling for various contextual factors (Stopczynski et al., 2014). Widely available smartwatches can perform the same tasks, while also providing data on an array of individual vital signs. In so doing, they can be used to substitute more complex and sometimes more complicated devices, such as smart badges, which have been designed as measurement devices for social interaction in close proximity (Pentland, 2007). In addition, such devices can be paired with specifically designed mobile applications to prompt participants with surveys to elicit additional data (Miller, 2012). Thus, the pervasiveness of mobile technologies can be used to collect data on human behaviour while maintaining the necessary degree of control over the data collection, as required by some research designs (see also Struminskaya & Keusch, Chapter 5 in this volume).

Similarly, interactive websites can be used to distribute and conduct surveys or to run more complex online experiments. In both cases, the novelty of the approach does not necessarily lie within the technical ability to create a survey form or design a digital experiment, but it is the possibility of making it accessible to a broad audience using digital distribution channels, such as social media platforms. Moreover, providing the participants of a study with additional technologies, such as specific software (e.g., browser plugins or mobile applications), allows researchers to monitor their activities throughout a study, such as by recording their clicks and interactions with websites while consuming media content or by collecting data on the content they are served by algorithms on social media platforms (Christner et al., 2021).

Finally, crowdsourcing has become an established form of conducting and supporting research online. Scholars can extend the scope of studies significantly by distributing an open call to the members of online communities and asking them to participate in academic endeavours by identifying patterns in images and labelling data (Sorokin & Forsyth, 2008) or participating in experiments and surveys (Paolacci et al., 2010), among other approaches. In the extreme, this concept can be extended to citizen science, where the research process is opened to the public (Bonney et al., 2014; Silvertown, 2009).

In summary, the digital transformation of society, especially of social interaction and information behaviour, has led to a variety of interesting phenomena that bridge the online and offline worlds. At the same time, the technologies that drive this transformation offer novel opportunities for advanced research designs. Unravelling the former and fully utilising the latter requires a thorough understanding of the technologies involved and the nature of the data they produce. Before elaborating further on the challenges of working with this type of data, it is essential to discuss the fundamental data types resulting from the above.

3 DIFFERENT TYPES OF DATA

Many labels have been used in reference to the data described above. Quite a number of them are vague, and 'big data' is probably the most opaque among them. While 'big data' is commonly used to describe the tremendous volume of data generated thanks to the prevalence of social media platforms and similar data sources, the term lacks any conceptual clarity and is thus best avoided. To account for the specific characteristics of such data, the term *digital trace data* is a much better alternative. While the term pertains primarily to data produced as a by-product of technology use, it serves as a reference point for data elicited deliberately using the same or similar technologies in controlled settings.

Digital trace data are defined as 'records of activity (trace data) undertaken through an online information system (thus, digital)', where 'a trace is a mark left as a sign of passage; it is recorded evidence that something has occurred in the past' (Howison et al., 2011). This definition captures a wide variety of data created by human interactions with technology, such as undirected actions (e.g., logging in to a website or clicking a button), interactions directed at others (e.g., establishing a social relationship or uploading content), and interactions directed at information (e.g., liking a picture or sharing a link). Thus, digital traces can resemble simple log data (such as transaction data from online markets (see Przepiorka, Chapter 13 in this volume); or trace data from online dating (see Skopek, Chapter 12 in this volume)), which document atomistic actions, and complex data, representing aggregated information such as content shared on social media platforms (e.g., tweets, Wikipedia articles, or YouTube videos (see Breuer et al., Chapter 14 and Schwemmer et al., Chapter 15 in this volume)).

It is worth noting that trace data do not necessarily have to be digital (Howison et al., 2011). For example, observations and recordings of human activity offline, which document events that took place in the past, qualify as trace data, and are frequently used in some disciplines. Among other names, data of this type are referred to as process-generated data, that is, data produced as a by-product of human activity and *not* in response to a deliberate stimulus of an observer, such as newspaper articles or transcripts of political speeches (Baur, 2011; Johnson & Turner, 2003).

Digital trace data share some of the properties of these other data but due to their origin have three fundamental characteristics that make them unique.

First, digital trace data are found rather than reported, thus they are referred to as *found* data (Howison et al., 2011). This characteristic refers to the fact that they are a by-product of individuals using information and communication technologies without the intent of creating data for academic purposes and often without reacting to a stimulus designed by an observer to elicit data for research (other than proprietary research conducted by the platform operator). This distinguishes them from data collected via measurement instruments specifically designed to collect data for research, such as surveys or experiments. This has two implications for working with digital trace data: they are not biased by measurement instruments – they represent human activity recorded in the absence of the typical effects introduced by observers and instruments present in research settings; but they are subject to other biases, many of which might be unknown. Most notably, digital trace data are affected by the design of the platforms from whence they originate. While this can be accounted for in some cases, not all properties of the technologies involved are transparent to their users and outside observers. For example, the algorithms that recommend content to Twitter users or rank Google search results are not fully transparent and introduce biases to user behaviour observable through digital trace data. While digital trace data offer a high level of resolution and novel ways of observing human behaviour in digital spaces, it is vital to account for limitations introduced by the fact that they are created to operate the technologies and enable the services from whence they originate. In this sense, they follow the logic and goals of the organisations running the platforms rather than those of independent observers interested in using digital trace data for research.

Second, digital trace data are *event-based* rather than *summary data* (Howison et al., 2011). While they can differ in their degree of abstraction, digital trace data typically document events at a specific time. For example, a messenger service might produce data that log interactions between individuals, and each record might represent a message exchanged between two individuals. Even for a small population using the service with moderate frequency, such data will likely contain hundreds of interaction events between pairs of users. While such events are interesting for some research areas, the events themselves would rarely be the focal unit of analysis. In many instances, the data would be used to study the relationships unfolding between the service users, with the events serving as proxies for social relationships between users.

Analysing these relationships would require an abstraction from the events. A common way to perform such an abstraction would be to identify pairs of users who interact frequently and assume that theirs is a somewhat stable social relationship, thus converting event-based data to summary data. The result, a relationship between individuals inferred from event data, involves non-trivial assumptions about interpreting patterns emerging from event data. Compared to other means of data collection, such as surveys or interviews during which the individuals involved must characterise the nature of their relationships in response to precise questions, digital trace data and aggregates derived from them can be challenging to validate. At the same time, they provide insights into human behaviour that are not biased by such questions and can be used to study otherwise unobservable patterns of behaviour that might not be reported by individuals if prompted by an explicit stimulus.

Third, digital trace data are *longitudinal* by nature (Howison et al., 2011). As a product of technological artefacts, digital trace data comprise records of activity that are typically timestamped. Whether they represent an exchange via email or social media, digital content in the form of a tweet or Wikipedia article, an edit of a blog post or a ‘like’ of a video, the events recorded by the technological infrastructure that underlies the platforms on which the activity takes place logs that activity over time and thus produces time series data by default. While other sources of data collection, especially traditional instruments used in the social sciences, can be designed to collect longitudinal data, they are often used in cross-sectional designs, mainly when longitudinal designs are too labour intense or have little chance of producing reliable outcomes. While digital trace data thus provide a higher temporal resolution, working with them entails additional assumptions when temporal aggregations are required.

The properties described above result from digital trace data being a by-product of information systems online, that is, of the underlying data-generating process. They are, in some aspects, similar to what is usually referred to in academia as secondary data. In contrast to primary data, secondary data are not collected originally by the researcher conducting a study but by third parties (e.g., other researchers or organisations). When working with secondary data, one gives up control over the data generation and collection processes, including the dataset's characteristics. Similarly, when working with digital trace data, researchers rarely have control over the data-generating process, which is managed by the entity governing the technology that underlies the information system from which the data are extracted. This lack of control, which often imposes research limitations, can be compensated for in part by designing controlled research environments based on the same technologies that underlie proprietary platforms and services, or by establishing research collaborations with platform and service providers.

Consider, for example, technologies used in the advertising industry to monitor how individuals interact with websites. These can be employed to elicit browsing behaviour and clickstream data of individuals to understand patterns of online news consumption (Flaxman et al., 2016; Mukerjee et al., 2018). Other examples include studies in which participants are equipped with smartphones, which act as mobile sensor platforms under the researcher's control, and thus can track the behaviour of participants – with their knowledge and consent (Stopczynski et al., 2014). Further, cooperating operators of established platforms can aid in regaining control over relevant sections of the data generation and collection processes. Notable examples include studies that rely on established platforms' communities to participate in academic research. For example, scholars cooperated with the operator of Eve Online, a massively multiplayer online game, and asked the player base to identify anomalies in images in exchange for in-game rewards (Sullivan et al., 2018).

Similarly, users of MyHeritage and Geni.com participated in creating a crowdsourced genealogical dataset that comprises millions of genealogical records and is now available for research (Hsu et al., 2021; Kaplanis et al., 2018). While such approaches have significant advantages in terms of control over the resulting data, they often depend on the researchers' capabilities to establish cooperation or utilise novel technologies in often complex research designs. Thus, as with digital trace data, there are drawbacks. At the same time, the challenges imposed by collecting and using such data are related less to control over the involved processes and are, instead, more grounded in instrument and measurement design and related to challenges with which the academic community is more familiar.

To exploit data from either source fully, it is essential to account for the origin of the data and their characteristics. This requires understanding the data-generating processes unfolding between individuals, as well as the technologies they use, to achieve specific goals under or outside researchers' control.

4 CHALLENGES AT THE INTERSECTION BETWEEN SOCIAL AND COMPUTER SCIENCE

Many of the challenges involved in working with the types of data described above pertain to data access, data-generating processes, and the role of theory in explaining phenomena of interest. Research on predicting election outcomes based on digital trace data, specifically social media data collected from Twitter, illustrate some of these challenges well.

Ever since the inception of Twitter, its rich data have attracted scholars from a variety of disciplines, which has led to a substantial body of research evolving around the platform. A particular strand of Twitter-based research aims at using the signals issued by millions of Twitter users – in the form of tweets, retweets, mentions, hashtags, following relationships, and more – to predict events within and outside of the platform. Examples include the prediction of stock market trends (Oliveira et al., 2017; Pagolu et al., 2016), box office success (Arias et al., 2014), and the outcome of elections (Tumasjan et al., 2011).

Predicting elections is a welcome example, as there is ample research on the subject. The anatomy of a study in this category is straightforward: the independent variables are typically measures derived from user activity in the period leading up to the election (e.g., the number of tweets containing the hashtags associated with political parties or candidates, their relative share of the overall activity observed during that time, and the sentiment inferred from the tweets); and such measures are then used as an input for statistical models, which are trained on the data to predict the outcome of the election. After the election, the model's results are compared to the actual election outcomes.

While many of these studies present remarkably accurate models, they have been met with fierce criticism and suffer from several issues. Among the more prominent problems are a lack of reproducibility and generalisability of the findings, the often required intensive fine tuning of the presented models, and a general lack of a theoretical foundation that would allow for an appropriate discussion of the findings in the context of established theories (Gayo-Avello, 2013; Gayo-Avello et al., 2011; Jungherr et al., 2012, 2017).

In essence, despite the popularity of this line of research, it has taught us little about relevant social processes in the context of elections.

A complete list of issues with this line of research is beyond this chapter's scope. Nevertheless, it is interesting to explore some fundamental questions that underlie the idea of predicting election outcomes and raise some general questions about the use of digital trace data in this context.

4.1 Data Access

It is worth noting the prominent role Twitter has played in enabling this and related lines of research. While there are multiple reasons for this, one crucial aspect is its policy regarding *data access*. Early on, Twitter provided almost unrestricted access to the platform's data, including official interfaces to monitor the full live stream of tweets being published and a full history search interface to look up historical tweets. Complete archives of all tweets were created and shared online during that period. These archives were valuable resources for research, as they provided a complete picture of discussions and activities taking place on Twitter.

Over time, free access to this resource was reduced significantly with Twitter's decision to monetise most access to data. For example, rather than offering access to the entire stream of tweets published by the platform's users, Twitter began to restrict that access to a random sample. This posed a significant challenge for academic research, as control of the sampling logic was suddenly shifted to Twitter itself, outside the control of researchers (Morstatter et al., 2013). Other factors made it too complicated to control for characteristics of the data Twitter offered, such as unknown demographics and additional restrictions introduced for the look-up of historical data. Some of these restrictions have been lifted for academic research in recent years, but others – such as unknown characteristics of Twitter's population – persist as challenges to working with digital trace data collected from Twitter (see also Kashyap, Rinderknecht et al., Chapter 3 in this volume).

Twitter is, nevertheless, a positive example of providing managed access to data. Other platforms are almost inaccessible for academic research. At one end of the spectrum between closed and open access models, Facebook is known for being quite restrictive in providing access to its data (Hegelich, 2020). While the platform is accessible to some degree, Facebook's access model heavily favours business applications and is tailored towards advertisers that target specific audiences on the platform. This severely limits research opportunities based on Facebook data and leaves it to proprietary research conducted by Facebook, which cannot be replicated or evaluated without access to the data used by corporate research teams.

At the other end of the spectrum, platforms such as Reddit and Wikipedia provide open access to their data, and hence their data is widely available for research. Beyond access, however, it is important to note that data collected from these platforms may still be subject to technical limitations. In the case of Reddit, posts and comments published on the platform can be up and downvoted by Reddit users. While the messages themselves can be collected from the platform, the votes they receive are aggregated at the time of data collection. Thus, longitudinal data on up and downvotes is not available by default and needs to be created by repeatedly collecting data from the platform.

While there are many reasons to restrict access to digital trace data that are accumulated in the databases of platform providers (e.g., privacy concerns and regulations), researchers have argued for better and controlled access to them (Hegelich, 2020; Lazer et al., 2020). Nevertheless, data access remains one of the most significant challenges in academic research. The election example outlined earlier illustrates some of the issues that arise from this limited access. As previously mentioned, the demographics of Twitter's population are mostly unknown, and access to data has, for a long time, been restricted to a sample of the overall activity on Twitter. This leads to an interesting question: What part of the population that has participated in the election is represented by data collected from Twitter, and is it reasonable to assume that public discussions on Twitter generally reflect public opinion about the election?

While the question may not be focal for research that is merely interested in prediction, it becomes crucial in light of established research on elections, democratic processes, and public opinion. Few studies in this line of research actively address these sorts of questions and the issues of data and information quality that can challenge reproducibility and transparency. Related issues, such as limitations with respect to access to historical data, can introduce additional barriers to the replication of studies based on digital trace data.

Overcoming limitations to data access is difficult. On the one hand, academia must coordinate efforts to convince platform providers to cooperate. On the other hand, scholars must carefully consider which data sources to choose in conducting research, and triangulate their results using data from multiple platforms and instruments (e.g., by using online surveys to estimate relevant characteristics of the population from which they collect digital trace data). At the very least, it is vital to raise questions with respect to validity issues that may arise from limited access, and to embrace open data initiatives as a way to address reproducibility issues.

4.2 Data-Generating Processes

It is important to consider the *data-generating processes* that underlie digital trace data. In general, digital trace data result from individuals interacting with information systems online. Such interactions are subject to various influences that determine the information they contain. From a sociotechnical perspective, individual interactions with technological artefacts are actualised affordances – opportunities provided by a technological artefact, in a specific context, as perceived by individuals who decide to act upon those opportunities and use the technology in ways that help them achieve particular goals (Leonardi, 2012).

Twitter, for example, affords its users the opportunity to follow other users, publish tweets, and tag keywords or other users in those tweets. Data logged by Twitter merely resembles the result of the actualisation of an affordance, while it is silent on the individual's context and intent. An individual might, for example, follow the Twitter profile of a political candidate during an election, which would likely be part of a dataset that could be used to study social media use during the election. Whether the visible tie between the individual and the candidate would indicate political support or is simply the prerequisite for the individual to be notified about the candidate's posts depends on the individual's intention and goals. Both can typically only be inferred and derived from assumptions about the individual. At the same time, this missing information is vital in determining what is measured using digital trace data, a fundamental question of validity that pertains to every study working with this type of data.

Two other factors, context and design, can provide further insights into the use of technology. Context is a vague term that typically refers to the circumstances and the environment surrounding the interaction observed through digital trace data. In the context of election studies, an apparent contextual boundary is set by the timeframe and topic of a conversation taking place on Twitter. For example, it might be reasonable to assume that a tweet that mentions a political candidate in a conversation between multiple people during the campaign period is relevant to public discourse on the election. Likewise, an old tweet outside of this context in a different conversation might not be relevant.

A study investigating patterns of social relationships among a cohort of students provides another example outside the realm of social media. The students who participated in the study offered data on their communication and interaction patterns in multiple media types, their course schedules, and GPS data on their physical location during the study. The availability of GPS data and the course schedules allowed the researchers to control the context of their interactions and distinguish, for example, between interactions related to shared classes and those outside of the students' university schedules (Stopczynski et al., 2014).

While rich contextual information allows for more robust and refined analyses, they are difficult to obtain and sometimes altogether unobtainable. Nevertheless, efforts to control for and understand the context in which digital trace data have been produced are crucial to strengthening the foundation for the assumptions required to interpret those data.

Finally, it is essential to understand the design of a specific technology and how it affects user behaviour. Technology design rarely happens by accident; rather, it is intentional and follows a particular purpose, such as to promote features and the intuitive communication of intended use. Thus, a technology's design reflects the intentions and ideas of its designer, which must be accounted for when interpreting the results of interactions between users and a technological artefact.

Again, we return to the example of predicting elections based on Twitter data and a single tweet that is part of a conversation related to the election. One might ask whether the author of the tweet participated in the discussion. Twitter has a vast population, and millions of users discuss a plethora of topics on the platform at any time of the day. Somehow, a user must come across the conversation and decide to participate. Thus, the answer to our question may seem obvious for those familiar with Twitter: the topic or parts of the exchange must have ended up in the user's feed (the list of tweets a user sees each time the platform is visited). The follow-up question is why it ended up there, which leads to an interesting component of Twitter – the algorithm responsible for curating individualised feeds of content provided to each user.

With this knowledge that one of the primary mechanisms of delivering content to Twitter users is their private feed and that an algorithm is responsible for composing it, one might wonder about its inner workings. Unfortunately, that is quite elusive. The algorithm is not transparent; its inner workings are unknown to anyone but its designers. Thus, how and why a user might have decided to exhibit a specific behaviour on the platform often remains unclear. Perhaps the user follows someone who partook in the conversation, which led to it being part of their feed. It is equally likely that the user might have seen a trending topic, which leads to them exploring it and then finding the conversation and becoming a participant. In this particular example, as in studies related to the consumption and spread of information on Twitter, this design aspect of Twitter poses a challenge that requires researchers to make assumptions about platform use (Howison et al., 2011; Lazer & Radford, 2017).

It is also worth noting that design is subject to change. For example, Twitter increased the maximum length of tweets from 140 to 280 characters; Facebook now allows its users to react to content published on the platform with a variety of emoticons, whereas historically it restricted such reactions to 'liking' content; YouTube recently decided to limit its response features rather expand them, so while users can still like or dislike videos, the number of dislikes is no longer displayed. These changes can potentially expand or restrict user behaviour and change how they interact with a given platform. Thus, scholars must familiarise themselves with the design of and user behaviour on platforms from whence they obtain their data to strengthen the validity of their assumptions and increase the robustness and replicability of their results.

4.3 Role of Theory

Research based on digital trace data and the digital environments from which they originate requires particular attention to *the role of theory*. As mentioned earlier, the popularity of research based on digital trace data is due in part to the digital transformation of society and the surge of data that has resulted. At the very least, this transformation has shifted a significant fraction of social interactions and information exchange to digital spheres, where they are mediated by their enabling technologies. This has implications for established theories on social behaviour – the phenomena to which they pertain remain either unaffected or are subject to changes. In the former case, what has been studied 'offline' is either not subject to what happens 'online' or is mirrored in both domains; in the latter case, empirical evidence online contradicts empirical evidence offline. Both cases are interesting and can provide valuable contributions to the body of social science theories. However, research must engage in theorising to deliver such contributions actively. While this may seem trivial, the complex and diverse range of phenomena studied through the lens of digital trace data and their interdisciplinary nature often renders theorising a complicated task.

Consider, again, the simple case of predicting election outcomes, which illustrates this complication quite well. It is worth noting that the perspective taken here – framing the problem as a prediction task – is one closer to computer science, especially given the description of the basic architecture of a study aimed at predicting election outcomes based on digital trace data. A valid research goal in this domain is to devise a method to produce accurate and precise predictions as a way to demonstrate the predictive power of the data used in combination with the method. Social sciences studies, however, tend to focus far less on demonstrating anything about the data and method; their objective is far more often to build upon and thus extend the body of social science theories on a specific subject. Thus, in the social sciences, the question would not primarily be *if* Twitter data can predict election outcomes and *how* those predictions could be improved, but *why* digital trace data from Twitter might be indicators for political support or public opinion during an election, why this might lead to a voting intention, which might then affect the outcomes of an election (Gayo-Avello et al., 2011; Hofman et al., 2021; Jungherr et al., 2017). Framed this way, the goal shifts significantly from predicting the outcome to being able to offer a supportable explanation for the outcome. In this sense, a study that merely predicts the outcome of a social process without a valid link to a theoretical foundation that helps explain the prediction might be less valuable to the social sciences.

Even this simplified depiction of the slightly different goal structures of two disciplines that are crucial to unlocking the potential of digital trace data for the study of social phenomena helps us to understand some of the prevalent issues that are detrimental to advancing the field.

Predictions, of course, are not explanations, and vice versa – much like correlation and causation are related but different. Research based on digital trace data, particularly in its early days, was prone to issues resulting from this conflation. There has been no shortage of studies aimed at predicting all sorts of elections (Schoen et al., 2013), which indicates the popularity of this line of research. On the one hand, this popularity has helped to demonstrate the value and significance of digital trace data. It has rightfully drawn attention to where these data originate and the social processes that unfold in those places. Further, it has demonstrated the capability of innovative methods and research designs to identify relevant signals and interesting patterns of human behaviour with complex and large datasets comprising digital trace data. On the other hand, the success of predictive models has led to claims that require explanations not provided by the research designs used in such studies, which, in turn, has led to strong criticism (Hofman et al., 2021; Schoen et al., 2013).

Research often distinguishes between offline and online phenomena, which raises another issue related to the role of theory when working with digital trace data. The question of why Twitter data might be an indicator of political support is, undoubtedly, intriguing. Considering, however, that Twitter represents only a fraction of the population that might be eligible to vote in any particular election, and that Twitter users are people who are subject to a multitude of influences outside of the platform, it is reasonable to assume that there are more interesting questions. For example, one might ask about the role Twitter – and, by proxy, all the interactions and information exchanged on the platform – plays in the surrounding media system. Albeit a broad question, it is justified if the goal is to understand why messages exchanged on the platform are supposed to be related to individual and collective behaviour during elections. The question becomes even more relevant when one considers the long history of research on the intersection between media systems and elections. For example, the way we measure and understand public opinion and agenda setting during and outside of election periods is based on the idea of a media system that comprises traditional mass media (print, radio, and television). This media system, however, has changed significantly; it has been replaced by a hybrid system that is subject to various influences and feedback loops between media organisations and individuals (Jungherr et al., 2019).

Given the intertwined and complex nature of the system that underlies the phenomenon of interest, an exclusive focus on a single platform, or an emphasis on the relevance of what is happening online, seems to fall short of providing comprehensive and robust explanations. It is important to make clear that this does not disqualify studies that look at a particular phenomenon through the lens of a specific platform; rather, we need to address the big picture to provide more robust explanations for interesting phenomena at the intersection between the social sciences, computer science, and other sciences. This argument leads to another point: it is equally detrimental to ignore digital trace data and remain wedded only to that which is most familiar to researchers.

In addition to the challenges mentioned here, there is a long list of unresolved issues related to the subject. Notably, concerns related to privacy and ethics are absent from the discussion; nevertheless, they often inhibit access to data that could otherwise be made accessible for academic research with appropriate precautions. Similarly, methodological issues are another important subject that is only partially addressed. Several of the issues discussed above, especially those related to understanding the behaviour of users responsible for generating digital trace data, might be resolved by employing research designs based on multiple data sources and types of data. In particular, the combination of quantitative (digital trace) data and qualitative data (e.g., interviews or observations) offers promising avenues for future research (Grigoropoulou & Small, 2022).

5 CONCLUSION

In summary, novel data on human behaviour are a seemingly abundant resource of the twenty-first century. This remarkable quantity of data, however, often comes with equally remarkable challenges with respect to their quality. Further, for academic research based on such data, and digital trace data in particular, barriers to controlled access rank high among the most demanding challenges. To resolve these challenges requires coordinated and continuous effort; they can be circumvented with careful consideration and innovative research designs. Further, it is paramount to keep in mind that data-generating processes on digital platforms are often opaque and out of researchers' control; understanding such processes requires the study of how users behave on such platforms, the context in which they do so, and how the design of the technologies of those platforms shapes or affects their behaviour. Based on this foundation, engagement in theorising is another important key to unlocking the potential of digital trace data. This endeavour requires academia to address fundamental, interdisciplinary challenges, particularly at the intersection between the social sciences and computer science.

The emergence and popularity of fields dedicated to research at this intersection, first and foremost computational social science, signifies the relevance and potential of research in this direction. At the same time, the sobering conclusion that we seem far off from understanding the social processes that underlie the phenomena studied in this field should make us revisit and rethink the foundations of this research. This may require tremendous effort but overcoming the challenges we face in working with digital trace data is worth it. After all, no one ever said a measurement revolution would be easy.

REFERENCES

Arias, M., Arratia, A., & Xuriguera, R. (2014). Forecasting with Twitter data. *ACM Transactions on Intelligent Systems and Technology*, *5*(1), 1–24.

Bastos, M. T., Mercea, D., & Charpentier, A. (2015). Tents, tweets, and events: The interplay between ongoing protests and social media. *Journal of Communication*, *65*(2), 320–350.

Baur, N. (2011). Mixing process-generated data in market sociology. *Quality & Quantity*, *45*(6), 1233–1251.

Bennett, W. L., & Segerberg, A. (2012). The logic of connective action: Digital media and the personalization of contentious politics. *Information, Communication & Society*, *15*(5), 739–768.

Bonney, R., Shirk, J. L., Phillips, T. B., Wiggins, A., Ballard, H. L., Miller-Rushing, A. J., & Parrish, J. K. (2014). Next steps for citizen science. *Science*, *343*(6178), 1436–1437.

Chadwick, A. (2017). *The hybrid media system: Politics and power*. Oxford University Press.

Christner, C., Urman, A., Adam, S., & Maier, M. (2021). Automated tracking approaches for studying online media use: A critical review and recommendations. *Communication Methods and Measures*, 1–17.

Eismann, K., Posegga, O., & Fischbach, K. (2016). *Collective behaviour, social media, and disasters: A systematic literature review*. ECIS.

Eismann, K., Posegga, O., & Fischbach, K. (2018). *Decision making in emergency management: The role of social media*. ECIS.

Eismann, K., Posegga, O., & Fischbach, K. (2021). Opening organizational learning in crisis management: On the affordances of social media. *Journal of Strategic Information Systems*, *30*(4), 101692.

Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, *80*(S1), 298–320.

Gayo-Avello, D. (2013). A meta-analysis of state-of-the-art electoral prediction from Twitter data. *Social Science Computer Review*, *31*(6), 649–679.

Gayo-Avello, D., Metaxas, P., & Mustafaraj, E. (2011). Limits of electoral predictions using Twitter. *Proceedings of the International AAAI Conference on Web and Social Media.*

Grigoropoulou, N., & Small, M. L. (2022). The data revolution in social science needs qualitative research. *Nature Human Behaviour*, 1–3.

Hegelich, S. (2020). Facebook needs to share more with researchers. *Nature*, *579*(7800), 473–474.

Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., Margetts, H., Mullainathan, S., Salganik, M. J., & Vazire, S. (2021). Integrating explanation and prediction in computational social science. *Nature*, *595*(7866), 181–188.

Howison, J., Wiggins, A., & Crowston, K. (2011). Validity issues in the use of social network analysis with digital trace data. *Journal of the Association for Information Systems*, *12*(12), 2.

Hsu, C.-H., Posegga, O., Fischbach, K., & Engelhardt, H. (2021). Examining the trade-offs between human fertility and longevity over three centuries using crowdsourced genealogy data. *PloS One*, *16*(8), e0255528.

Johnson, B., & Turner, L. A. (2003). Data collection strategies in mixed methods research. In A. Tashakkori & C. Teddlie (Eds), *The SAGE handbook of mixed methods in social and behavioral research* (pp. 297–319). SAGE.

Jungherr, A. (2018). Normalizing digital trace data. In N. J. Stroud & S. McGregor (Eds), *Digital discussions: How big data informs political communication* (pp. 9–35). Routledge.

Jungherr, A., Jürgens, P., & Schoen, H. (2012). Why the pirate party won the German election of 2009 or the trouble with predictions: A response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe,

I. M. ‘predicting elections with Twitter: What 140 characters reveal about political sentiment’. *Social Science Computer Review*, *30*(2), 229–234.

Jungherr, A., Schoen, H., Posegga, O., & Jürgens, P. (2017). Digital trace data in the study of public opinion: An indicator of attention toward politics rather than political support. *Social Science Computer Review*, *35*(3), 336–356.

Jungherr, A., Posegga, O., & An, J. (2019). Discursive power in contemporary media systems: A comparative framework. *International Journal of Press/Politics*, *24*(4), 404–425.

Kaplanis, J., Gordon, A., Shor, T., Weissbrod, O., Geiger, D., Wahl, M., Gershovits, M., Markus, B., Sheikh, M., & Gymrek, M. (2018). Quantitative analysis of population-scale family trees with millions of relatives. *Science*, *360*(6385), 171–175.

Kleinberg, J. (2008). The convergence of social and technological networks. *Communications of the ACM*, *51*(11), 66–72.

Lazer, D., & Radford, J. (2017). Data ex machina: Introduction to big data. *Annual Review of Sociology*, *43*, 19–39.

Lazer, D., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., Freelon, D., Gonzalez-Bailon, S., King, G., & Margetts, H. (2020). Computational social science: Obstacles and opportunities. *Science*, *369*(6507), 1060–1062.

Lazer, D., Hargittai, E., Freelon, D., Gonzalez-Bailon, S., Munger, K., Ognyanova, K., & Radford, J. (2021). Meaningful measures of human society in the twenty-first century. *Nature*, *595*(7866), 189–196.

Leonardi, P. M. (2012). Materiality, sociomateriality, and socio-technical systems: What do these terms mean? How are they related? Do we need them? In P. M. Leonardi, B. A. Nardi, & J. Kallinikos (Eds), *Materiality and organizing: Social interaction in a technological world* (pp. 25–48). Oxford University Press.

Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science*, *7*(3), 221–237.

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. (2013). Is the sample good enough? Comparing data from Twitter’s streaming API with Twitter’s firehose. *Proceedings of the International AAAI Conference on Web and Social Media*.

Mukerjee, S., Majó-Vázquez, S., & González-Bailón, S. (2018). Networks of audience overlap in the consumption of digital news. *Journal of Communication*, *68*(1), 26–50.

Oliveira, N., Cortez, P., & Areal, N. (2017). The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems with Applications*, *73*, 125–144.

Pagolu, V. S., Reddy, K. N., Panda, G., & Majhi, B. (2016). Sentiment analysis of Twitter data for predicting stock market movements. *2016 International Conference on Signal Processing, Communication, Power and Embedded System*.

Panagiotopoulos, P., Bigdeli, A. Z., & Sams, S. (2014). Citizen–government collaboration on social media: The case of Twitter in the 2011 riots in England. *Government Information Quarterly*, *31*(3), 349–357.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*(5), 411–419.

Pentland, A. S. (2007). Automatic mapping and modeling of human networks. *Physica A: Statistical Mechanics and Its Applications*, *378*(1), 59–67.

Reuter, C., Hughes, A. L., & Kaufhold, M.-A. (2018). Social media in crisis management: An evaluation and analysis of crisis informatics research. *International Journal of Human–Computer Interaction*, *34*(4), 280–294.

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, *346*(6213), 1063–1064.

Schoen, H., Gayo-Avello, D., Metaxas, P. T., Mustafaraj, E., Strohmaier, M., & Gloor, P. (2013). The power of prediction with social media. *Internet Research*, *23*(5), 528–543.

Segerberg, A., & Bennett, W. L. (2011). Social media and the organization of collective action: Using Twitter to explore the ecologies of two climate change protests. *The Communication Review*, *14*(3), 197–215.

Silvertown, J. (2009). A new dawn for citizen science. *Trends in Ecology and Evolution*, *24*(9), 467–471.

Sorokin, A., & Forsyth, D. (2008). Utility data annotation with Amazon Mechanical Turk. *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*.

Stopczynski, A., Sekara, V., Sapiezynski, P., Cuttone, A., Madsen, M. M., Larsen, J. E., & Lehmann, S. (2014). Measuring large-scale social networks with high resolution. *PloS One*, *9*(4), e95978.

Sullivan, D. P., Winsnes, C. F., Åkesson, L., Hjelmare, M., Wiking, M., Schutten, R., Campbell, L., Leifsson, H., Rhodes, S., & Nordgren, A. (2018). Deep learning is combined with massive-scale citizen science to improve large-scale image classification. *Nature Biotechnology*, *36*(9), 820–828.

Travers, J., & Milgram, S. (1977). An experimental study of the small world problem. In M. Newman, A.-L. Barabási, & D. J. Watts (Eds), *The structure and dynamics of networks* (pp. 179–197). Elsevier.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2011). Election forecasts with Twitter: How 140 characters reflect the political landscape. *Social Science Computer Review*, *29*(4), 402–418.

Ugander, J., Karrer, B., Backstrom, L., & Marlow, C. (2011). The anatomy of the Facebook social graph. https://doi.org/10.48550/ARXIV.1111.4503

Wallach, H. (2018). Computational social science≠ computer science+ social data. *Communications of the ACM*, *61*(3), 42–44.

Watts, D. J. (2007). A twenty-first century science. *Nature*, *445*(7127), 489–489.

Wolfsfeld, G., Segev, E., & Sheafer, T. (2013). Social media and the Arab Spring: Politics comes first. *International Journal of Press/Politics*, *18*(2), 115–137.