DOI: *10.46991/BYSU:F/2022.13.2.028*

THE STATE OF DATA: REFLECTIONS ON USING "BIG" AND ADMINISTRATIVE DATA SOURCES IN SOCIAL RESEARCH

Scot Hunter https://orcid.org/0000-0002-6101-732X
Mr, MSc, University of Stirling, Research Fellow in Social Sciences, University of Stirling (UK), Email: s.p.hunter@stir.ac.uk
Marina Shapira https://orcid.org/0000-0002-8860-1841
Dr, PhD, University of Oxford, Associate Professor in Sociology, University of Stirling (UK), Email: marina.shapira@stir.ac.uk

Abstract: *Recent computing power and storage advancements have meant more data are being collected and stored. Referred to as 'Big data', these data sources offer researchers myriad opportunities to make observations about the social world. These data can be massive, provide insight into whole populations rather than just a sample, and be used to analyse social behaviour in real time. Administrative data, a subcategory under the big data umbrella, also offers researchers abundant opportunities to conduct highly relevant research in many areas, including sociology, social policy, education, health studies and many more. This paper offers reflections on social research during the digital age by examining different forms of data, both 'big' and 'small', and their associated advantages and disadvantages. The paper concludes by suggesting that although big data has some promising elements, it also comes with some limitations and will not replace 'traditional' social surveys. And yet, when used in conjunction with social surveys, appropriately and ethically, big data could offer the researchers additional valuable insights.*
Keywords: *sociology, methodology, big data, administrative data, quantitative data analysis.*

ՏՎՅԱԼՆԵՐԻ ՎԻՃԱԿԸ. ՄՏՈՐՈՒՄՆԵՐ ՍՈՑԻԱԼԱԿԱՆ ՀԵՏԱԶՈՏՈՒԹՅՈՒՆՆԵՐՈՒՄ ՄԵԾ ԵՎ ՎԱՐՉԱԿԱՆ ՏՎՅԱԼՆԵՐԻ ԱՂԲՅՈՒՐՆԵՐԻ ՕԳՏԱԳՈՐԾՄԱՆ ՎԵՐԱԲԵՐՅԱԼ

Սքոթ Հանթեր
Սթիրլինգի համալսարանի (Մեծ Բրիտանիա) մագիստրոս, Սթիրլինգի համալսարանի սոցիալական գիտությունների գիտաշխատող
Մարինա Շապիրա
փիլիսոփայական գիտությունների դոկտոր (Օքսֆորդի համալսարան), Սթիրլինգի համալսարանի (Մեծ Բրիտանիա) սոցիոլոգիայի ամբիոնի դոցենտ

Ամփոփում. *Տվյալների պահպանման ոլորտում ժամանակակից առաջընթացի շնորհիվ հնարավոր է դարձել ավելի ու ավելի մեծ ծավալների տվյալներ կուտակել և պահպանել: Տվյալների մշտապես կուտակվող և աճող հենքերը, որոնք կոչվում են «մեծ տվյալներ», առաջարկում են սոցիալական աշխարհը դիտարկելու բազմաթիվ նոր հնարավորություններ: Մեծ տվյալների միջոցով հնարավոր է պատկերացում կազմել ամբողջ բնակչության, և ոչ միայն, ընտրանքային համախմբության մասին, և դրանք օգտագործել իրական ժամանակում զանգվածային սոցիալական վարքի վերլուծության համար: Վարչական տվյալները՝ որպես մեծ տվյալների ենթատեսակ, նաև առաջարկում են համապատասխան հետազոտություններ իրականացնել բազմաթիվ, այդ թվում՝ սոցիոլոգիայի, սոցիալական քաղաքականության, կրթության, առողջապահության ոլորտներում: Սույն հոդվածն անդրադառնում է թվային դարաշրջանում սոցիալական հետազոտություններում տվյալների տարբեր ձևերի՝ ներառյալ մեծ և փոքր տվյալների կիրառմանը, դրանց առավելություններին և թերություններին: Ենթադրվում է, որ թեև մեծ տվյալները որոշակի խոստումնալից բնութագրիչներ ունեն, այնուամենայնիվ պարունակում են նաև որոշակի սահմանափակումներ և չեն փոխարինի սոցիալական ավանդական հարցումներին: Սոցիալական հարցումների հետ մեկտեղ՝ մեծ տվյալների ճիշտ օգտագործմամբ հետազոտողները կարող են լրացուցիչ արժեքավոր պատկերացումներ ձեռք բերել:*
Բանալի բառեր – *սոցիոլոգիա, մեթոդաբանություն, մեծ տվյալներ, վարչական տվյալներ, քանակական տվյալների վերլուծություն*

СОСТОЯНИЕ ДАННЫХ: РАЗМЫШЛЕНИЯ ОБ ИСПОЛЬЗОВАНИИ ИСТОЧНИКОВ «БОЛЬШИХ» И АДМИНИСТРАТИВНЫХ ДАННЫХ В СОЦИАЛЬНЫХ ИССЛЕДОВАНИЯХ

Скот Хантер
магистр наук. университет Стерлинга, научный сотрудник социальных наук, университет Стирлинга (Великобритания)
Марина Шапира
доктор философии, Оксфордский университет, доцент кафедры социологии Стерлингского университета (Великобритания)

Journal of Sociology: Bulletin of Yerevan University, Vol. 13 (2), Dec 2022, pp. 028-037
Received: 05.08.2022
Revised: 24.10.2022
Accepted: 31.10.2022

Аннотация: *Недавние достижения в области вычисления и хранения данных привели к тому, что все больше данных накапливается и хранится теперь. Источники данных, называемые «большими данными», предлагают исследователям множество возможностей для наблюдения за социальным миром. Эти данные могут быть массивными, давать представление о всей совокупности, а не только о выборке, и использоваться для анализа социального поведения в режиме реального времени. Административные данные как подкатегория больших данных, также предлагают исследователям широкие возможности для проведения актуальных исследований во многих областях, включая социологию, социальную политику, образование, исследования в области здравоохранения и т.д. В данной статье рассматриваются социальные исследования в эпоху цифровых технологий путем изучения различных форм данных, включая «большие» и «малые» данные, а также связанные с ними преимущества и недостатки. В заключение выдвигается предположение, что, хотя большие данные имеют некоторые многообещающие характеристики, они также имеют некоторые ограничения, и не заменят «традиционные» социальные опросы. И все же при правильном использовании больших данных в сочетании с социальными опросами исследователи могут получить дополнительную ценную информацию.*
Ключевые слова: *социология, методология, большие данные, административные данные, количественный анализ данных.*

Every century, a new technology – steam engine, electricity,
atomic energy or microprocessors –
has swept away the old world
with the vision of a new one.
Today it seems we are entering the era of 'Big data
(Michael Coren as cited in Ariba 2021: np).

The state of data: reflections on using 'big' and administrative data sources in social research

We are living in the 'era of big data' (Ariba, 2021: 1), a digital age in which social phenomena can now be counted and stored in multiple ways and on even larger scales than ever before possible (Jasanoff, 2017). The explosion of data, a by-product of the shift from analogue to digital and technological advances in computing power and information storage, has enabled researchers to make previously impossible hypotheses about how the social world works (Salagnik, 2018; Jasanoff, 2017). The driving powerhouse of this ‘data revolution’ has been the wide-scale production and use of big data, which has not only given researchers new opportunities but become an intrinsic feature of societal processes, influencing how business is conducted, how policy is designed, and how people are observed and managed (Kitchin, 2022: 3; Mills, 1956; Jasanoff, 2017). First coined in the mid-1990s, the term 'big data' was initially used when referring to datasets too big for computers to store at that time. Now, big data ‘encompasses a range of other qualities immanent in the digital traces of routine activities such as utility consumption, web browsing or social media usage’ (Halford and Savage, 2017: 1133). Due to these data sources providing insight into the everyday behaviour of millions of people, the term has also emerged as an empirical phenomenon and emergent field of practice, especially in the social sciences.

Big data have enabled researchers to achieve what was once impossible, ask questions that could not be asked before, and reach research populations that were once unreachable (Salagnik, 2018). This paper focuses on how different types of big data can be used in social research by examining how their different ontological characteristics and associated pros and cons may impact the research process. Whilst some commentators champion big data, others remain wary of its unexpected consequences, going as far as to prophesise a 'crisis of empirical sociology' where theory is disregarded, and big data replaces all forms of knowledge (Savage and Burrows, 2007; Anderson, 2008). Several other limitations associated with using these data – such as ethical concerns and reliability issues- further obfuscate arguments for its use (Halford and Salvage, 2017). While we recognise the potential of using big data in social research, claims that these data sources would change Sociology as “we know it” are mostly unfounded (Halford and Savage, 2017; Sturgis and Luff, 2021). We conclude that traditional forms of data – such as social survey data - remain (and will continue to remain) prominent in the social sciences. While recognising big data can be helpful, especially if used in tandem with traditional survey methods, researchers should err on the side of caution when contemplating using them.

Differentiating data types

Data are 'pieces of information about phenomena and the input for (and output from) computational processes' that drive quantitative empirical enquiry (Kitchin, 2022: 4). Traditionally, data have been collected using social survey questionnaires and only in the last decade have other data sources – such as transactional and social media data – been used for social research. These data have unique characteristics that dictate how they can be handled and what they can be used to do. After defining each data source below, the following section will discuss the pros and cons of each type to provide context on how big data are different from traditional methods, explain how it has grown in prominence, and provide insight into its usefulness for researchers.

Data can be either 'small' or 'big' (Kitchin, 2022). The former includes traditional survey sources such as questionnaire data. In contrast, the latter encompasses data collected and stored digitally, such as internet search logs or data collected by administrations like hospital records. Before the emergence of big data, datasets referred to here as 'small' were simply known as data. These data are commonly recognised traditional datasets that predate big data, usually collected for the purpose of research and to investigate certain phenomena, for example through questionnaires, census data, and interview transcripts - both quantitative and qualitative (Kitchin, 2022; Connelly et al., 2016).

In contrast, companies and administrations create, collect, and store big data digitally for non-research purposes. Some examples of big data formats include - but are not limited - to transactional information, text messages, emails, audio clips, social media posts, and images. In addition, big data sources comprise administrative data collected by institutions such as tax authorities, health services, social services, and the criminal justice system (Connelly et al., 2016; Playford et al., 2016). Administrative data offer rich insight into social behaviour and provide researchers with an excellent opportunity to make policy-relevant analyses.

It should be noted that this categorisation is not fixed. Kitchin (2022) argues administrative data should be considered separately as it does not share all the essential criteria of other big data sources. However, administrative data are hugely influential for social researchers, and as such are discussed in detail below as falling under the big data umbrella.

Small data and the "withering" survey

For years, quantitative enquiry has been built on the back of small datasets and traditional survey methods. Small data sources are conventionally highly systematic and designed for statistical purposes. The population of interest is often known, and techniques are used to recruit a representative sample rather than observing a whole population. Using representative samples ensures no bias and equal representation of groups in society, allowing researchers to use statistical techniques to confidently infer findings from the broader population from which it is drawn.

Despite being referred to as 'small', these data sets can be complex and large. Understanding Society (https://www.understandingsociety.ac.uk), for example, is an extensive longitudinal survey of members of approximately 40,000 households in the United Kingdom, with various modules on different aspects of life such as employment, education, and health (University of Essex, 2022). With so many households and approx. 60,000 participants, sources like Understanding Society can still be downloaded and used by researchers with minimal training in data analysis software such as SPSS or Stata. Moreover, these datasets are usually well-structured as they have been designed for statistical purposes and with research aims in mind. Therefore, they are compatible with classical statistical methods usually taught in quantitative social science courses.

Although these data are highly structured, can be representative, and are mostly reliable, they are also expensive and burdensome to produce. Not only does collecting survey data take a lot of time and effort, but surveys also only provide data from a fixed point in time. The analysis is thus specific to the time the data was collected, and real-time observations are unavailable. Moreover, the robustness of survey methods has been questioned due to reducing response rates as people became less interested in participating (Groves, 2011; Savage and Burrows, 2007; Miller, 2017). Overall, 'sharply rising costs and declining response rates were perceived to offer population inference of uncertain accuracy at a snail's pace and eye-watering prices' (Sturgis and Luff, 2021: 692). For these reasons, researchers have turned to alternative methods to research the social world, taking advantage of the new opportunities made available during the era of big data and the digital age.

Before discussing these new innovative opportunities however, it is worth iterating that traditional survey methods associated with collecting small data remain a prominent tool in quantitative social enquiry,

even in the era of big data. Despite some commentators forecasting that the days are numbered for "withering" surveys (Sturgis and Luff, 2021), traditional methods have evolved and, in some instances, become more powerful whilst adapting to the digital age. For instance, modes of data collection have evolved to accommodate the increasing digitalisation of everyday life. Doorstep and telephone interviews and mailed questionnaires have been replaced by online web surveys that can even be conducted on a person's smartphone (Miller, 2017). Advances in computational technology have also improved access to secondary data sets collected by others. Open-access data archives such as the UK Data Service (https://ukdataservice.ac.uk) make it reasonably easy to obtain datasets for anyone, regardless of their level.

So, what is big data, and how can it be used?

It is helpful to look at how other researchers have used it to understand the different types of big data. For example, examples of data used in academic work include using Tweets to measure how mood varies across different cultures during different seasons and times of day (Golder & Macy, 2011); supermarket checkout data to examine how co-workers' productivity influenced performance (Mas & Moretti, 2009); and mobile phone call records to determine the economic circumstances of thousands of Rwandan citizens through machine learning techniques (Blumenstock et al., 2011).

These data sources differ from small data, and have three main characteristics. Regularly referred to as the '3V's', big data are (1) massive in *volume*, sometimes consisting of terabytes or petabytes of data (2) high in *velocity*, always being produced and being made instantaneously; and (3) coming in a *variety* of forms, including structured and unstructured data (Laney 2001; Zikopoulous et al., 2012). Enhancements in computer storage and processing have meant that masses of data can be collected. Before the data revolution, data was only collected and stored at one particular time. For example, small data such as surveys provide a snapshot of a particular population at the time the survey was taken. On the other hand, big data sources, such as social media data, are continuously being made and 'always on' (Salagnik, 2018: 21). Big data is not limited to structured datasets that characterise traditional survey methods. Still, this may mean that using traditional statistical techniques to analyse big data is less straightforward than it would be with smaller data sets that are designed for research purposes.

Alongside the three V's, other qualities differentiate big data. First, big data are exhaustive, meaning they can capture information from all cases of interest rather than just a population sample. For example, researchers can access all posts across an entire social media platform or all transactions made by customers of an online marketplace. Thus, there is no need to construct sampling frames or employ techniques to infer observations made during analysis, as the numbers can speak for themselves. Whilst some may consider the exhaustive nature of big data as positive, it can also be considered a limitation. Scholars of big data need to be careful they are not generalising beyond the population they observe. Access to all cases in specific sub-populations, such as Twitter users, should not be mistaken as having access to an entire *broader* population. For example, it is unlikely that all Twitter users in a specific country are representative of the *broader* population of that country (Salagnik, 2018) due to issues of access and variations in socioeconomic characteristics (such as age and class) that may influence online participation. Big data, such as social media data, can be very detailed and offer great coverage into specific sub-populations. However, it will usually be biased as it is restricted to users of certain services. In contrast, small data sets are purposefully designed to avoid bias and accurately represent the population of interest.

Second, big data are often characterised by relationality, meaning they can be linked to other data sources using unique identifiers, and although more examples of linking big data sets together are emerging, trying to link big data can often be time-consuming and messy due to their less structured nature than small data studies. However, as big data are 'always on', they are flexible, and data structures can thus be extended to new fields with minimal effort. Overall, big data can offer new opportunities to observe a specific population as it continues to expand and grow (Kitchin, 2022).

Third, big data are also non-reactive, meaning data sources are much less likely to impact the behaviour or attitude of participants. One key concern with traditional survey methods is the effect that knowledge of their participation has on participants' responses. Big data are free from this concern as data are usually collected as a by-product of another process, meaning answers may be a more accurate reflection of a person's behaviour. However, just because data are collected this way does not necessarily mean they reflect 'natural processes', especially when considering an online activity that may not always be genuine (Newman et al., 2011; Salagnik, 2018).

Finally, compared to survey data, big data tends to be incomplete, 'dirty', and inaccessible, and thus require a monumental effort and arduous data management to get ready for analysis and may not include all the data required for robust research. As big data are collected for non-research purposes, core

information such as demographics (like gender or ethnicity of participants) may be missing, restricting what Social Scientists can do with the data. Big data are also 'dirty', including junk and spam that could be misleading and non-reflective of social behaviour (Salagnik, 2018). Moreover, big data are often inaccessible as companies hold and collect it, and access may be limited by legal, commercial, or ethical concerns. Overall, obtaining and sorting these data sources for social research is a mammoth task that may require unconventional skills for Social Scientists (such as computer science skills), patience, and time.

Administration data, the overlooked big data source?

Discourse over big data is narrowly centred on online sources, as discussed above, with many neglecting administrative data over historical concerns of accessibility and breaching confidentiality (Salagnik, 2018; Playford et al., 2016). Governments have been collecting data on citizens for a long time, but Social Scientists have been restricted in accessing these sources in the past. Nonetheless, increasing digitalisation during the data revolution has accelerated the rate of data collection, storage, and analysis. In addition, and in response to increasing digitalisation, new initiatives such as the Administrative Data Research Network (UK) (ADRN) have emerged, aiming to provide 'open access to a plethora of data that have been recorded' by governments and encouraging ethical and feasible transmission of data from government to researcher (Playford et al. 2016: 3). Considering these developments, administrative data has become one of the most powerful instruments in a social scientist's tool kit.

Again, administrative data sources are not collected for research but are commonly reused by Social Scientists to explore social phenomena. For example, Farber (2015) reused data from New York City's government digital taxi meters to explore the relationship between hourly wages and hours worked. Other examples include using Scottish Population Census data in addition to administrative education data to explore curriculum provision in secondary schools (https://curriculumproject.stir.ac.uk). Administrative data fall under the big data umbrella as they meet most of the criteria discussed in the previous section. For one, administrative datasets are usually quite large in scale and vary in structure, as administrations can collect vast amounts of data on their users, such as birth, death and marriage registrations, educational qualifications, and health records for whole cities' populations, countries, or even continents (https://ec.europa.eu/eurostat/web/main/home). Yet, these data are usually more manageable than other big data sources, with standard statistical software capable of handling them for analysis.

Moreover, administrative data usually capture observations from a whole sub-population and do not rely on sampling procedures as with social surveys; they have access to enough people for the measurements to be considered as 'true population parameters', rather than estimates for making inferences from samples to the populations from which these samples were drawn. Therefore, administrative data can offer valuable insight into entire sub-populations of interest. For example, educational records will include all those who through went through the educational system in a particular county over a particular period; health records will include all those who were using health services at one time – both overall providing opportunities for detailed research on specific areas of society, rather than broader society, to be conducted.

As Kitchin (2022) argues, what distinguishes administrative data from other sources is that it is not created as rapidly as other big data sources. In some cases, data such as social media posts or online marketplace transactions are generated and recorded in real-time, with minimal delay between recording and publishing. Administrative data, on the other hand, is published more sporadically. Take, for example, data collected by governments on house prices in the UK. Although this data is updated in real-time when someone inputs into the system, results are only published weekly or monthly (Kitchin and McArdle, 2016). Admittedly, administrative data are timelier than small data sets – such as questionnaires conducted annually – but are not available to the same extent as other continuously updated online data sources. Of course, defining a data source as 'big' depends on the characteristics of the source itself. Not every dataset will meet this paper's 3Vs and other criteria. For instance, Population Census data may be considered 'big' by some – it is usually large in size (volume) and attempts to cover an entire population (exhaustive). Still, it is only collected once every ten years and therefore not high in velocity, and therefore does not meet the criteria to be considered 'big' (Kitchin, 2022). The criteria mentioned thus far should only be a rough guide for identifying different data sources. These terms are not fixed; instead, it is more about how the data are understood and handled, determining its nature.

Undisputedly, and especially within a UK context, administrative data has become an essential resource for quantitative researchers, with initiatives such as the ADRN and the Administrative Data Research UK (ADRUK) – which aim to improve the accessibility of administrative data for researchers – growing in prominence. Similarly to ADRN, ADRUK (https://www.adruk.org) links extant governmental datasets together to permit meaningful research, prepares administrative data for analysis, and trains users

on how to use it effectively. ARDUK also provides a service for researchers to safely access administrative data without concerns over breaching confidentiality or ethical standards. Through the help of these initiatives, academics can use administrative data to conduct impactful and policy-relevant research that may not have been possible when using traditional time-consuming and costly research methods such as analysis of secondary datasets, especially in instances where quick policy responses are needed. In addition, administrative data offer the benefits of large case numbers and fine-tuned data in areas highly relevant to specific policy areas, such as educational attainment or health service use.

However, regardless of the positives associated with administrative data, they are not without drawbacks. Types of administrative data of interest to Social Scientists will most likely contain sensitive information about individuals (e.g. histories of health, conditions, criminal records, or adoption records) and contain information that makes individuals easily identifiable. For this reason, the data needs to be prepared in the form that removes any possibility of identification, and the users (e.g. academic users) need to be trained how to use the data securely and ethically. Furthermore, administrative data will almost always include multiple records across cases and time points. For example, it may include tax records for one person across several years. All these factors make administration data messy; they require determination and time to properly prepare for use, by linking together different information for the same individuals and ensuring it is fully anonymised.

In fact, due to issues over anonymity, administrative data can sometimes be hard to access, even considering the initiatives established (ADRUK, ADRN) to make obtaining these datasets easier. Negotiations over data access can be tiresome and even last up to a year (Playford et al., 2016). Researchers are usually interested in using multiple administrative data sets – for example, linking individuals' health records with information about social security, e.g., whether individuals receive social benefits, but may be put off by time constraints and how long this process would take to achieve.

The debate over whether administrative data should be considered 'big' or 'small' illustrates how these concepts are not fixed and how different scholars view data within social research. Although contributing to the debate is not the purpose of this paper, we feel administrative data should be considered a type of big data as they share so many similarities. Nevertheless, researchers should not get too bogged down in abstract disputes and instead focus on the data itself. However, considering the ontological characteristics of data sources is important as the factors that differentiate types of data sources from one another will influence how the data should be handled and impact the observations being made. Therefore, when planning to conduct research using administrative or big data, these considerations should be at the forefront of decision-making.

The 'crisis' of sociology?

In response to the data revolution and the eagerness of scholars to harness the potential of big data in their work, multiple prophecies on how these changes would influence disciplines were made. Notably, Savage and Burrows (2007) feared empirical sociology would change forever and 'the dominance of the survey as the pre-eminent form of data in the social sciences [would be] usurped by new forms of digital transactional ('big') data' (Sturgis and Luff, 2021: 694). Anderson (2008) summed up arguments made at the time to suggest that empirical theorising would become redundant:

> 'In a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behaviour, from linguistics to sociology [...] Who knows why people do what they do? The point is that they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves.' (np).

Despite having prominence and gaining much attention, the 'apocalyptic vision' that big data would replace all other forms of knowledge has not endured (Sturgis and Luff, 2021: 694). Halford and Salvage (2017), for instance, highlight that certain limitations of big data cannot be overlooked and prevent it from replacing traditional methods. Big data captures only some activities of particular people, using certain devices and applications intended to record specific information, meaning these data are bound to be biased, incomplete and lacking core demographic valuable information for analysis (Halford and Salvage, 2017: 1133).

Recent empirical literature has further challenged the claims summarised by Anderson (2008). Sturgis and Luff's (2021) content analysis of 1451 of the 'highest-ranking' articles published between 1939 and 2015 across a range of social science journals (inc. economics, sociology, political science and social psychology and public opinion), found there has been little decline in the use of traditional survey data in

empirical articles. In fact, it has increased. Empirical articles including big data only constituted a small proportion of the total output.

Furthermore, in most articles that did feature big data, survey data was used in addition to or to enhance what the big data could tell us (Sturgis and Luff, 2021). Scholars (Groves, 2011; Couper, 2013; Japec et al., 2015; Salagnik, 2018) champion combining big and small data sources, as they can help mitigate their respective limitations and provide more insightful analysis that cannot be produced with either data source individually. For example, big data can be used to identify the extent of a social issue - e.g., using voting records to measure voting turnout – and combined with survey data to analyse factors that may explore the problem in more depth - e.g., creating a survey exploring how likelihood to vote varies by demographic and socioeconomic markers (Ansolabehere and Hersh, 2012).

The need for continued attention to ethical concerns

There is a common misconception that there are few ethical concerns when using big data, as participants have already provided their consent for the data to be collected. However, some important ethical considerations remain before conducting research or selecting a data source. Participating in harmful processes and unanticipated secondary use of data are two often overlooked issues. For example, by using big data, we are potentially participating in 'surveillance capitalism' – the monetisation of data collected through monitoring of people online - and thus reinforcing the asymmetric relationship between consumers and their data; exploiting individuals for information and using the same information to encourage their consumption of goods and services (Sadowski, 2019; Zuboff, 2019).

As such, by using big data for our gain, we are arguably contributing to social processes that may impact individuals' emotional and physical well-being, making them vulnerable to malicious activities of businesses and organisations and unequal power relations (Cohen, 2019; Kitchin, 2022). Although big data advocates like to highlight the potential that these data can bring to the world, we must also acknowledge the dark side of the data revolution. We *also* live in an era of surveillance capitalism in which data have become a commodity extracted from people for profit and value, without compensation (Sadowski, 2019). These data are then used to target people with individualised publicity informed by their interests – pressuring them into consuming products. Essentially, data are taken from individuals, usually without their knowledge, for free and used against them for profit. Or, in Kitchin's (2022:227) words:

> 'The users of technologically mediated services [such as apps and other online platforms] provide labour (e.g., clicking, swiping, typing, uploading) and offer the product of those labours (data related to aspects of their lives) for free to those who control the means of production.'

This creates an asymmetrical relationship between those in charge of digital technologies and platforms and the service users who must submit a certain amount of personal information to sign up. As a result, individuals are led to believe they are signing up to certain apps and other platforms for free when in fact they are paying the price dictated by the terms and service agreements made by organisations that offer these services (Kitchin, 2022). Moreover, these processes affect large numbers of people, since signing up to many services is required to participate in a consumer society. Notably, the use of technology such as smart phones is often unavoidable.

Social Scientists should be wary of participating in these processes. By using big data, we are arguably contributing to the exploitation of individuals; using their data for purposes unbeknownst to them and without compensation. We may also be actively reinforcing an exploitative system with many risks for individuals and society. For example, many argue that surveillance capitalism undermines democracy, rendering individuals as complicit entities, removing their autonomy and free will. Research about individuals without their awareness or consent has long been a cause of concern in the social sciences. These concerns have been amplified in the digital age due to mass surveillance and new opportunities for data to be collected about people and used for various reasons.

Referred to as 'unanticipated secondary use', Salagnik (2018: 290) offers a dramatic example of how Germany's Nazi regime used institutional data to facilitate genocide against Jews, Roma and other minorities across Europe during the Second World War. Using innovative technology at the time such as Hollerith's tabulating machine (developed to assist summarising information rapidly), Nazis were able to record data on individuals, and victims were identified based on criteria such as ethnicity and religion (Jasanoff, 2017). Hollerith's tabulating machine was originally built-in response to the expansion of bureaucracies in the late nineteenth century, where governments were required to create 'census-like counts' of populations and other social phenomena (such as poverty, violence, health, employment etc) to

keep track of people and goods, tax citizens, manage population, and enlist people for military service (ibid: 2). Hollerith's machine, and other forms of enumerating the population, developed to help governments manage the population but succumbed to secondary unanticipated use for notorious goals. Such historical examples illustrate why the safeguarding and ethical use of administrative data, through initiatives like ADRUK and ADRN is paramount (Verwulgen, 2017).

We are not suggesting that anyone who wants to use big data does so to mislead anyone or to participate in harming processes. Yet, these considerations remain essential for us as social researchers. If these data are used, there is a strong likelihood that concerns over unanticipated secondary use will be raised by ethic committees, and consideration of potential consequences will need to be demonstrated to them, especially where data collection is funded by private institutions. Fortunately, not only is ethical practice being promoted by ADRN and ADRUK but there is also an extensive history of ethical frameworks developed for quantitative research that scholars of big data can draw from and adapt to suit their work - read Salagnik (2018: 294-301) for a detailed overview of how to conduct good ethical research in the digital age. Furthermore, as Mills (1956) recommended, social scientists must retain a critical awareness of the potential uses of such data by political, commercial and military elites.

Overall, various datasets are available and can be used in myriad ways, but for it to be used appropriately, a systematic approach and collaboration between the data producers (such as government agencies and corporations) who have a responsibility to safeguard that data, the research councils who set up initiatives to help make the data transmittable, and researchers who must use the data responsibly, is necessary. Moreover, a suitable/appropriate level of social trust is also required from the populations from which these data sources are extracting information; people need to consent to the collection of their data, be aware of the consequences of their participation and be confident that this data will not be used in inappropriate ways. This may be more difficult in some societies than others. In democracies like the UK, for example, levels of social trust are higher than societies with used to have authoritarian regimes (Letki, 2005; Herreros, & Henar, 2008).

Conclusion and recommendations

Although the data revolution has created enormous amounts of data that can potentially be used, we must thoroughly consider whether they are appropriate or not for our research aims. Our prime concern as social researchers should be to ensure we are conducting reliable and valid analysis. We must not fall victim to the hubris and excitement around big data and use it just because we can. Not only do we need to consider whether the data source is appropriate for our research questions, but also to ensure we have the skill and tools to use it effectively. Some of the best empirical research that includes big data is conducted by computer and data scientists, who are trained differently from social researchers who may have to learn a whole new set of skills to get the full potential of these data sources. We advocate that such data should be used in combination with traditional survey data. For example, in addition to other methods (surveys or interviews), big data can be used to help to get a more nuanced picture of certain phenomenon unreached by traditional methods or used to get a broader picture of phenomenon and link between micro and macro levels.

To conclude, using big data is not without its challenges, and currently the flaws outweigh the positives. Despite these data offering many possibilities for observing large amounts of cases in real-time, the process will most likely be time-consuming, messy, and sometimes unethical. Moreover, these data are likely to be biased and partial, potentially full of spam and may generate misleading results. Not only is big data tainted by its flaws, but we must also consider how using it requires collusion in processes that we are likely to critique and dismantle as social scientists. Nevertheless, big data can still contribute to the discipline of social science and enhance traditional methods if used in meaningful and ethical ways

We do not suggest big data should be avoided in any sense and recognise their potential, but we recommend sticking to the research goals of the study and think about the most appropriate source of data for the research questions. If social surveys are enough, we recommend sticking with them and maybe incorporating big or administrative data, only if it offers some additional insights that could not be made with primary or secondary data alone. We believe that, for the most part, secondary data will be more than suitable to answer most research questions in the social sciences; but if big data are the only way to answer the research inquiry, then it should be used ethically and critically. Overall, despite the excitement surrounding big data, we refute the claim that the days of the survey are over and will remain the "gold standard" of quantitative enquiry in the future.

Acknowledgement: This paper follows the *State of data: Reflections on social research during the digital age* presentation delivered to students and staff at the faculty of Sociology at Yerevan State University on 19 May 2022. This presentation was a part of series of a knowledge exchange events conducted between the University of Stirling and Yerevan State University organised through the *Should I stay, or should I go? Sense of belonging and intentions to stay among young newcomers to Armenia* (https://armenianewcomers.stir.ac.uk) project, funded by the British Academy. We would like to thank students and staff who attended the session and contributed to the discussion, which has gone on to inform the arguments presented below. We would also like to thank Dr Sarah Wilson for contributing her ideas and suggestions for this paper during the editing process.

REFERENCES

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete, *Wire*, 23 June. Available at: www.wired.com/2008/06/pb-theory/

Ansolabehere, S., Hersh, E. (2012), Validation: What Big Data Reveal About Survey Misreporting and the Real Electorate, *Political Analysis*, 20(4), pp.437–59. DOI:10.1093/pan/mps023.

Ariba, G. (2021) *Statistics, New Empiricism and Society in the Era of Big Data*. Springer.

Blumenstock, J, E. et al. (2011). 'Risk and Reciprocity Over the Mobile Phone Network: Evidence from Rwanda' *SSRN Elibrary, http://papers.ssrn.com/papers.cfm?abstract_id=1958042*

Cohen, J, E. (2017). 'Surveillance vs. privacy inside out', *Theoretical Inquiries in law*, 20(1) pp. 1-32.

Connelly, R et al., (2016). The role of administrative data in the big data revolution in social science research, *Social science research*, 59, pp.1-12.

Couper, M. P. (2013). Is the sky falling? New technology, changing media, and the future of surveys. *Survey Research Methods*, 7(3), pp. 145–156. http://DOI.org/10.18148/srm/2013.v7i3.5751

Fraber, H, S. (2015). Why you can't find a taxi in the rain and other labor supply lessons from cab drivers, Quarterly *Journal of Economics*, 130(4), pp.1975-2026. DOI: 10.1093/gje/gjv026.

Golder, S, A. & Macy, M, W. (2011). Diurnal and Seasonal mood Vary with Work and Sleep, and Daylength across Diverse Cultures, *Science*, 333(6051), pp.1878-1881. DOI:10.1126/science.1202775.

Groves, R. M. (2011). 'Three areas of survey research'. *Public Opinion Quarterly*, 75(5), 861–871. http://DOI.org/10.1093/poq/nfr057

Halford, S. Savage, M. (2017). Speaking Sociologically with Big data: Symphonic Social Science and the Future for Big data Research, *Sociology*, 51(6), pp. 1132–1148. DOI: 10.1177/ 038038517698639.

Herreros, F, & Henar C. (2008). The state and the development of social trust. *International Political Science Review*, 29(1), pp. 53-71.

Japec, L., et al. (2015). Big data in survey research: AAPOR task force report. *Public Opinion Quarterly*, 79(4), pp. 839–880. http://DOI.org/10.1093/poq/nfv039

Jasanoff, S. (2017). Virtual, visible, and actionable: Data assemblages and the sightlines of justice', *Big Data & Society*, pp. 1-15. DOI: 10.1177/2053951717724477

Kitchin, R. (2022). *The Data Revolution.* 2nd ed. London: Sage.

Kitchin, R., McArdle, G., (2016). What makes big data, big data? Exploring the ontological characteristics of 26 datasets, *Big Data and Society*, 3(1), pp.1-10.

Laney, D. (2001). 3D management: Controlling data volume, velocity, and variety, Meta Group, https://studylib.net/doc/8647594/3d-data-management--controlling-data-volume--velocity--an... (accessed 31 July 2022).

Letki, N, Evans, G. (2005). Endogenizing social trust: democratization in East-Central Europe, *British Journal of Political Science*, 35(3), pp.515-529.

Mas, A., & Moretti, E. (2009). Peers at Work, *American Economic Review*, 99(1), pp. 112-145. DOI:10.1257/aer.99.1.112.

Miller, P. V. (2017). Is there a future for surveys?' *Public Opinion Quarterly*, 81(S1), pp. 205–212. https://DOI.org/10.1093/poq/nfx008

Mills, C., W., (1956). *The Power Elite*, Oxford University Press.

Newman, M, W et al., (2011). It's Not That I Don't Have Problems, I'm Just Not Putting Them on Facebook: Challenges and Opportunities in Using Online Social Networks for Health. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, pp. 341–50. CSCW '11. New York: ACM.

Playford, C, J et al., (2016). Administrative social science data: The challenge of reproducible research *Big data & Society*, 3(2). DOI: 2053951716684143.

Sadowski, J. (2019). When data are capital: Datafication, accumulation, and extraction', *Big data & society*, 5(1): pp. 1-12.

Salganik, M J. (2018). *Bit by Bit: Social Research in the Digital Age.* Woodstock: Princeton University Press.

Savage, M., Burrows, R. (2007). The coming crisis of empirical sociology. *Sociology*, 41(5), pp. 885–899. DOI: 0.1177/0038038507080443

Sturgis, P., Luff, R. (2021). The demise of the survey? A research note on trends in the use of survey data in

the social sciences, 1939 to 2015, *International Journal of Social Research Methodology*, 24(6), pp. 691-696, DOI:10.1080/13645579.2020.1844896

The University of Essex, Institute for Social and Economic; Research. (2021). Understanding Society: Waves 1-11, 2009-2020 and Harmonised BHPS: Waves 1-18, 1991-2009. [data collection]. 15th Edition. UK Data Service. SN: 9614, http://DOI.org/10.5255/UKDA-SN-6614-16.

Verwulgen, I. (2017). The ADRN and the public's voice: making administrative data available for research while gaining public trust: IJPDS (2017) Proceedings of the IPDLN Conference (August 2016), *International Journal of Population Data Science*, 1(1) pp. 1-155. DOI: 10.23889/ijpds.v1i1.174.

Zikopoulous, P.C., et al. (2012). *Understanding Big Data*. New York: McGraw-Hill.

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for the Future at the New Frontier of Power*. New York: Profile books.