Big Data

Francisco Rowe[1]

Defining Big Data

The term *Big Data* can be traced back to the mid-1990s, used to describe the manipulation and analysis of very large data sets. Big Data are often formally defined in reference to distinctive traits. Seven key attributes are used to characterise Big Data (Kitchin, 2021). In the late 2000s, three traits ('the 3Vs') were typically used, describing Big Data as: large in *volume* encompassing terabytes or petabytes of data; high in *velocity* being generated in real or near-real time; and, diverse in *variety* in types. More recently, four attributes have been added to this representation, with Big Data being characterised as: *exhaustive* in scope, seeking to capture complete populations or systems; fine-grained in *resolution*; *relational* containing common fields that enable fusion with other data sets; and, *flexible*, enabling data sets to be expanded. Traditional "small data" are typically constrained across these attributes. They are produced in highly controlled settings employing statistical sampling techniques in ways that constrain their size, resolution, scope, velocity, variety and flexibility.

Most Big Data sources contain locational information. Not surprisingly Big Data have thus become instrumental in human geography over the last decade, revolutionising many areas of the discipline (Kitchin, 2014). Studies unleashing the attributes of Big Data have contributed to expanding existing theories, developing new explanations, adopting new analytical tools and infrastructures, and advancing new areas of research, such as urban science, online harm and misinformation. These studies have also greatly improved our understanding of key societal issues, ranging from socio-economic inequality to the spread of infectious diseases. More fundamentally they have influenced policy practice and governance to tackle real-world challenges, notably during the unfolding COVID-19 pandemic.

Sources and forms of Big Data

Technological advances in computational power, storage and network platforms have enabled the emergence of Big Data. These technological innovations have facilitated the production, processing, analysis and storage of large volumes of digital data. Information that previously could not be stored, or used to be captured using analog devices can now be recorded digitally. We can now digitally generate, store, manage and analyse data that were previously very challenging to access, such as books, newspapers, photographs and art work.

---

[1] *Corresponding author*: F.Rowe-Gonzalez@liverpool.ac.uk

An unprecedented amount of social data is also now available and generated continuously through interactions on digital devices and platforms.

Kitchin (2013) identified three broad systems as key sources of Big Data: directed, automated and volunteered systems. Directed systems involve digital administrative platforms operated directly by a human recording data on places or people e.g. closed-circuit television, digital school registers, immigration control, biometric scanning and health records. Automated systems include digital technologies which automatically and autonomously record and process data with little human intervention e.g. mobile phone, electronic smartcard ticketing, energy smart meter sensors, satellite, email, banking and retail networks. Volunteered systems involve platforms in which humans contribute data through interactions on social media networks (e.g. Twitter and Facebook) or crowdsourcing projects (e.g. OpenStreetMap and Wikipedia).

Opportunities of Big Data

Big Data offer an extraordinary potential for research in human geography. Compared to traditional data sources, Big Data can provide much more sophisticated, wider scale, finer grained understandings about our social and economic systems. Data sets of larger size now means that we have a greater ability to zoom into smaller populations and infrequent data points. Yet, the most notable promises of Big Data are probably the greater breadth, depth, scale and timeliness that they offered.

Big Data offer higher geographic and temporal granularity. Much of the Big Data are spatially and temporally referenced offering great opportunities for enhancing our geographic understanding of social processes and their changes in short time intervals. Mobile phone data represent a key example. Mobile phones are continuously collecting and storing locational information with timestamps. Such information has been key to expand our understanding of human mobility and migration flows over space and time. We have now the ability to observe exact locations and times. For example, theoretical ideas formulated in the context of time-space geography can now be explored and tested, based on daily individual-level data for entire population systems and cities.

Big Data also provide extensive coverage. Traditionally, random sampling has been used to gather data because collecting data on entire populations is normally unfeasible due to logistical and financial challenges. National censuses are a rare exception. Big Data can potentially provide information to study entire systems of populations or geographical areas. For instance, Twitter posts offer complete population coverage at a global scale to enhance our understanding of differences across populations and geographic areas in sentiment expressions on Twitter (Rowe et al., 2021a).

Real or near-real time availability is also a key feature of Big Data. Traditionally data are often collected through analogue devices, such as surveys and questionnaire, processed and be made available weeks, month or even years after collection. Digital trace data can now be streamed uninterruptedly in real or near-real time. Google and Apple mobility data are notable examples of this feature. Google and Apple have continuously released data on

mobility. This information is frequently updated as new data are made available, and has been key to evaluating public health interventions during the COVID-19 pandemic, such as lockdowns and social distancing measures.

Challenges of Big Data

Big data also, however, pose major epistemological, methodological and ethical challenges. Understanding these challenges is important as they are likely to shape the future of human geography. In this section, I therefore elaborate on the key challenges around the use of Big Data.

Epistemological challenges

Big Data has raised fundamental questions about the positioning of human geography in science. Big Data have pushed us to rethink the discipline in the context the rise of new forms of positivist and empiricist reasoning outside social sciences (Kitchin, 2014). Efforts have been made to rethink, integrate and position human geography as a key actor in the development of principles and practices that can guide the use of Big Data in the study of social, political and spatial issues. Valuable critical reflections have proposed ways in which human geography could be reconfigured to leverage on data-driven thinking (Miller and Goodchild, 2015). Arguments have also been made for human geography to become a key player in integrating quantitative geography thinking with Data Science in the development of explicitly spatial analytical approaches (Singleton and Arribas-Bel, 2021).

However, human geography has been relatively slow in crystallising these propositions. Other disciplines and fields have started to undertake spatial analysis, and new areas of study have emerged integrating social sciences with computer science and engineering. Computational social science is a key emerging area which has been influential in setting the research agenda on the application of computational approaches to social science questions. Unfortunately human geography has not remained underrepresented within this emerging research area (Springer Nature Labs, 2021). Such developments are key to attract research funding, secure exclusive data access, and influence the business and policy space.

Yet, human geography can still occupy a primary role in shaping the research agenda. Naivety has existed outside human geography believing that Big Data can provide new insights for themselves and do not require domain-specific knowledge. Social processes have been assumed to follow laws of physics, often ignoring decades of social science scholarship and resulting in reductionist analyses failing to take account for cultural, political, policy, demographic and economic factors. The challenge for human geography is to shape the research agenda. The complexities in the use of Big Data require strengthening multidisciplinary collaboration. Drawing on a long history of interdisciplinary experience attracting scholars and epistemologies from various fields, human geography provides an ideal environment to foster such collaboration.

Big Data have also sparked a need for the elaboration of existing theories and development of new ones. Human geography theories have traditionally and largely been static and

prescriptive, often neglecting the evolving and specific local nature of human behaviour and social interactions. Big Data now provide detailed spatial and temporal information to capture these dynamic processes. Theorisation is needed to generate relevant hypotheses and constructs that can appropriately be measured using Big Data. Additionally existing questions often focus on the middle of the population distribution but now Big Data offer an opportunity to expand our line of enquiry and focus on marginalised, smaller populations.

New theories are also needed to understand new interactions between human behaviour and digital technology. Digital technologies shaping human behaviour in new ways. For instance, Twitter and Facebook influences our behaviour by shaping our social network i.e. suggesting who to follow and what to read; Google recommending search terms; and, Netflix suggesting movies and TV shows. Additionally, digital traces encode our behaviour and our awareness of surveillance practices feeds into what is measured creating a human-technology interaction loop. People can now deliberately influence what data are collected by using virtual private networks to hide locational information, or avatars to avoid facial recognition. The data altered by these processes are not affected at random across the population. Only certain individuals have the knowledge and skills to intervene digital platforms in this way. Theories are needed to better understand these interactions and differences across populations.

Methodological challenges

Big Data also pose methodological challenges. Human geographers require new methods to handle, analyse and store data sets of millions or billions of observations which are continuously being generated in a variety of forms. Traditional statistical methods were designed to identify significant relationships in small sample sizes with known properties e.g. p-values to assess statistical significance have become less relevant. Big data are not collected for research purposes. They are an unintended consequence of administrative processes or social interactions and need to be reengineered for research. Handling Big Data requires a wider and new digital skills set, largely based on machine learning, artificial intelligence and coding, in addition to greater knowledge of computing technology (e.g. Jupyter notebooks, Github and Docker). Except for a few centres, current university geography programmes and infrastructures are largely unprepared to deliver the required training. A multidisciplinary approach is needed to integrate computational training into human geography.

Inference and causality represent an additional methodological challenge. A dimension of these issues is confounding which relates to the capacity to identify typical human behaviour from those resulting from rules governing a digital ecosystem. Individuals may express their own genuine opinions on social media, but digital platforms can also influence these views through algorithmic decisions. Digital platforms influence how, where, and why we move, consume, act and interact. Algorithmic decisions may operate in pervasive and opposite ways across the population, reinforcing social polarisation and political leanings. Drawing inferences about attitudes and opinions from images and human expressions is a further challenge (Rowe et al., 2021b). Computers have a hard time accurately decoding this information as expressions and images contain sarcasm, irony and hyperbole. Approaches on how to address these issues are needed.

Biases are also a major obstacle. Generalising findings from Big Data because they may represent a subset of the population of interest; or, reflect systematic differences in access to technology. Digital traces tend to be left by relatively wealthy people in high-income countries, biasing attempts to draw global conclusions. Equally research on Twitter, the most widely cited emerging data source, should be critically assessed before generalising any findings. For example, it is only used by 26% of the UK population and much less in most other countries. Undermining generalisation is also the existence of silos as different platforms elicit systematically different behaviours. Individual interactions on Twitter are likely to differ from those on Instagram or Zoom and these systematic biases are captured in the data. Generalisation is further complicated as technological sensors may introduce biases recording behaviour divided across humans (two people using the same email account) or across sensors (the same individual watching AppleTV on a smartphone and a desktop).

Lack of standards to guide data analysis remains a major issue. First, identifying the demographic characteristics of individual users is challenging. Developing robust approaches to infer these attributes from digital trace data, or extract and fusion them from other sources is a key challenge. Second, while progress has been made, distinguishing human from non-human activity particularly identifying bots remains difficult. Third, there is a need to establish standards on how to label data, train and validate models, particularly given the fact that this is becoming a prominent market (e.g. Amazon Web Services (AWS) Marketplace). Data and models are often labelled and trained, with limited validation and metadata based on specific populations or geographical areas in a single language; that is, typically male white individuals from industrialised countries using English. These practices have raised concerns around discrimination and applicability of such models to the wider society, particularly non-white female populations and other geographic settings in less developed countries due to poor predictive performance and large-scale societal implications. Additionally, validation standards are also needed to assess new ways of measuring existing concepts using Big Data. Often traditional ways of measuring concepts are seen as the gold standard. Yet, by doing this, we may simply replicate their shortcomings (Lazer et al., 2021).

Ethical challenges

Big Data have also raised a number of ethical questions. A major issue is the transparency paradox. On the one hand, we want to access the data collected by digital platforms and understand better their inner-workings; on the other, we want to ensure privacy. The challenge is on how to securely access Big Data – which is complicated by the fact that most data are generated by private corporations and have no obligation to freely share the data they collect. Access to Big Data is thus rarely available to academics, and current practices are largely unregulated. When data are accessible, raw data are often unavailable owing to privacy and intellectual property concerns, or may become unavailable in the future. This impedes reproducibility and replication of the results. When happens, accessibility is typically granted through a patchwork system with some data available through public application programming interfaces (APIs); other data only by working with and often physically in the company; and still other data through personal connections and one-off arrangements. Innovative partnerships have emerged providing secure data access e.g. the

Consumer Data Research Centre (CDRC) in the UK. Yet, no platform provides access to information on the extensive randomised control experiment that they conduct, which could enable inferences of the influences of their algorithms on human behaviour.

Box 1
Outlook
Restructuring practices and standards in human geography

- More actively seek to shape the research agenda on the application of computational approaches to social science questions
- Embrace a new era of data-driven thinking
- Infuse quantitative geography thinking into Data Science to develop spatially explicit machine learning and AI approaches
- Adapt university training programmes and computing infrastructure for social scientists
- Produce new theories of human behaviour and elaborate on existing explanations

Expanding the toolkit for human geography research

- Expand training in coding, machine learning, artificial intelligence and the basics of computing technology
- Reorganise academic structures to enable connections between researchers sharing interests in computational approaches and multidisciplinary collaboration
- Address causality and bias issues present in digital trace data before making inferences
- Establish data analysis standards to guide data integration, labelling, training and validation

Ensuring privacy and enabling data access

- Expand existing secure data centres for granting access, monitoring outputs, and enforcing compliance with privacy and ethics rules
- Develop enforceable guidelines around research ethics, transparency, researcher autonomy, and replicability
- Design anonymisation approaches that enable research on small and marginalised populations

Establish ethical guidelines on the use of data capturing publicly visible behaviours and network information

Particularly about privacy, numerous ethical considerations remain. While general agreement exists about a need to anonymise individual records so that they are not identifiable, robust approaches to achieve this are lacking. Differential privacy approaches have been developed to add noise to the data and thus ensure some level of anonymity. Yet, a trade-off exists because adding noise to enhance privacy diminishes the utility of the data. This is particularly a problem for research based on small, marginalised populations as they are disproportionally affected by this procedure. Recently the concept of open data products enable insights from highly sensitive, controlled and/or secure data which may not be accessible otherwise (Arribas-Bel et al., 2021). Additionally, there is a lack of clarity about privacy expectations for data capturing publicly visible behaviours e.g. tweets. Some universities require ethics approval to work with such data, but this approach varies greatly across institutions, and there is not clear guidance on, for example, if public information, such

as user screen names and posted messages can be displayed in publications. Even if ethical procedures exist, they do not consider the network structure of digital platforms, and that when users disclose information, they are generally sharing information about their social network contacts e.g. friends, followers and users they interact with. Guidance is needed to ensure the appropriate use of this information.

Human Geography and Big Data

Big data present unprecedented opportunities for human geography to transform our understanding of the social world and human behaviour, as well as have societal implications shaping politics, behaviours, elections, policy and the economy. During the COVID-19 pandemic alone, researchers have been able to access large streams of mobile phone, satellite imagery, credit card and social media to understand the social, health and economic implications of the SARS–CoV–2 spread in near-real time across the globe. Human geographers have made significant contributions to mapping, modelling and understanding of these ramifications using digital trace data. Nonetheless, Big Data also poses major epistemological, methodological and ethical challenges to the discipline, to effectively unleash their full potential. While some progress has been made, three broad challenges remain and involve: reorganising standards and practices in our discipline, expanding the research skill set and institutional structures, and enabling data access while ensuring privacy. Box 1 identifies the key challenges and actions to tackle going forward.

References

Arribas-Bel, D., Green, M., Rowe, F. and Singleton, A., (2021). Open data products-A framework for creating valuable analysis ready data. Journal of Geographical Systems, pp.1-18.

Green, M., Pollock, F. D., and Rowe, F. (2021). New forms of data and new forms of opportunities to monitor and tackle a pandemic. In COVID-19 and Similar Futures, pages 423–429. Springer.

Kitchin, R. (2013). Big data and human geography: Opportunities, challenges and risks. Dialogues in human geography, 3(3):262–267.

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. Big data & society, 1(1):2053951714528481.

Kitchin, R. (2021). The Data Revolution: A Critical Analysis of Big Data, Open Data and Data Infrastructures. SAGE.

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

Miller, H. J. and Goodchild, M. F. (2015). Data-driven geography. GeoJournal, 80(4):449–461.

Rowe, F., Mahony, M., Graells-Garrido, E., Rango, M., and Sievers, N. (2021a). Using twitter to track immigration sentiment during early stages of the COVID-19 pandemic.

Rowe, F., Mahony, M., Graells-Garrido, E., Rango, M., and Sievers, N. (2021b). Using twitter data to monitor immigration sentiment.

Singleton, A. and Arribas-Bel, D. (2021). Geographic data science. Geographical Analysis, 53(1):61–75.

Springer Nature Labs (2021). Data page: Explore computational social science using our experimental app. https://doi.org/10.1038/d42859-021-00054-7.

Further Reading

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science*, 355(6324):483– 485.

Batty, M. (2013). Big data, smart cities and city planning. *Dialogues in Human Geography*, 3(3):274–279.

Boyd, D., and Crawford, K., (2012). Critical questions for big data. Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5):662–679.

Blazquez, D. and Domenech, J. (2018). Big Data sources and methods for social and economic analyses. *Technological Forecasting & Social Change*, 130:99–113.

Chen, C., Ma, J., Susilo, Y., Liu, Y., and Wang, M. (2018). The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C*, 68:285–299.

Hilbert, M., (2016(. Big data for development: A review of promises and challenges. *Development Policy Review*, 34(1):135–174.

Kandt, J., and Batty, M. (2021). Smart cities, big data and urban policy: Towards urban analytics for the long run. *Cities*, 109:102992.

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