The data revolution in social science needs qualitative research

Although large-scale data are increasingly used to study human behaviour, researchers now recognize their limits for producing sound social science. Qualitative research can prevent some of these problems. Such methods can help to understand data quality, inform design and analysis decisions and guide interpretation of results.

Nikolitsa Grigoropoulou and Mario L. Small

The proliferation of large-scale data on human behaviour has helped to usher in a data revolution in the social sciences, stimulating researchers to study topics such as network evolution, political polarization and racial inequality in ways that previously were difficult or impossible[1]. But researchers now recognize that large-scale datasets from companies or agencies may have characteristics that unwittingly lead researchers to misrepresent social reality, and produce bad science[2,3]. We believe that methods often thought not to be essential to formal science — qualitative methods, such as in-depth interviewing and field observation — will help to prevent these issues. Using these methods will be necessary not only for background or context, but also to understand the quality of the data, to inform design and analysis decisions, and to guide the interpretation of results. We see at least seven reasons why qualitative research will be essential to 'big data' social science (Fig. 1).

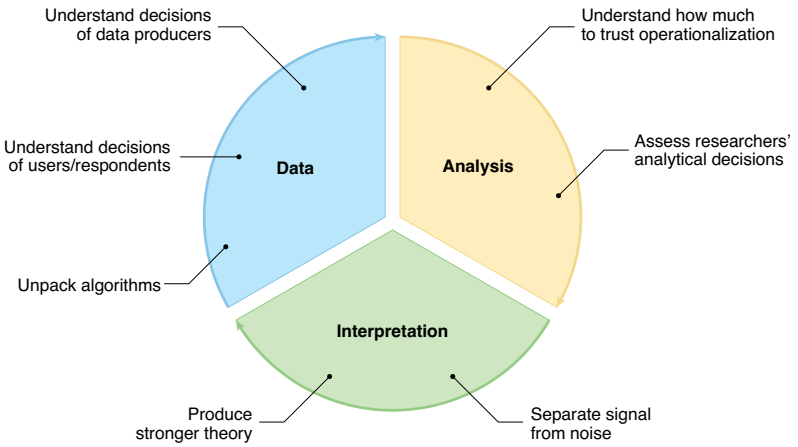

Fig. 1 | Qualitative research and big data. Seven roles for qualitative research in social science with large-scale data.

Decisions that affect data production

We need qualitative research to understand the decisions of those who produced the datasets. Conventional social science data are usually produced by researchers to generate scientific knowledge; the large-scale administrative data of today are typically not. Such data tend to be produced by companies and government agencies that, with their own interests in mind, decide what information to collect and how much to make available to researchers. Social scientists may know little of the behind-the-scenes dynamics involved, with potentially serious implications for research.

For example, Facebook created the platform Social Science One to provide researchers with access to its data. But in September 2021, news reports revealed that Facebook intentionally or unintentionally withheld information for about half of US users — those whose political affinities could not be clearly identified from their Facebook activities[4]. Thus, researchers studying topics such as political misinformation had been unknowingly working with biased samples, compromising the interpretation of their results.

The first step in analysing a large-scale dataset should be understanding exactly how the data came about — that is, understanding what the managers, programmers and others in the organization involved did to produce the data and why. Qualitative interviews with those decision-makers can be indispensable to understanding the data. Researchers evaluating data quality in health-care contexts have shown that such interviews can reveal how judgment calls, discrepancies in operational standards and other issues can compromise the quality of the resulting data[5].

Decisions made by respondents

We need qualitative research to understand the decisions of those who generated the data points. Organizations create the systems into which data are entered, but their clients, patrons and users — who ultimately generate the data points — decide what information to offer: what to post on social media, to enter on hospital forms, to report to the tax collector and so on. The psychological and social factors that influence those decisions — honesty, prejudices, fear, sense of privacy, peer pressure and so on — shape the resulting data in ways that may matter for research use.

For example, social media users differ in how much they worry about maintaining their reputations among those in their social network[6]. As a result, people who hold similar beliefs may differ in what they post or where, such that the posts available to a researcher may not reflect actual beliefs. Such issues are likely to bedevil research into political misinformation that uses social media data. In such contexts, an effective survey will be preceded by exhaustive, open-ended interviews, which — based on the slow build-up of trust between researcher and interviewee — will uncover answers to questions that researchers previously might not have known to ask.

The nature of algorithms

We need qualitative research to understand what lies behind algorithms. Algorithms are sequences of instructions that allow a computational system to perform an operation or solve a problem. Companies use algorithms to determine what users see, to flag suspicious emails, to determine credit scores and more. Thus, algorithms are essential to both user behaviour and the nature of the data.

Because algorithms are often proprietary, researchers may have little knowledge of how they affect the resulting data. Thus, it may be unclear how much a pattern observed in the data was due primarily to the algorithm itself[3,6]. For example, Twitter's algorithms have been called 'racist' because of how they have handled images of faces of Black individuals. Among other problems, if a user uploaded a photograph that was too large and contained faces of both a white and a Black individual, the algorithms were more likely to crop out the Black individual, leaving only the face of the white individual to appear on the user's timeline. Consider how this decision would affect a study. If a researcher then scraped Twitter data, they would collect the uploaded photograph, not the cropped version that users see, such that the role of the cropping algorithm would remain hidden in the dataset. The researcher could then wrongly infer that in pictures with multiple people, users ignore the faces of Black individuals — whereas in fact users were merely less likely to see such faces because of the algorithm's decisions.

Understanding algorithms is important. But uncovering the proprietary algorithmic code of a company is not always needed to assess how algorithms operate and introduce bias. Algorithms are the product of "thousands of decisions by the company's programmers"[3] and by the people who train and calibrate the models. To understand those human roles, interviews and field observations are among our most powerful tools: interviews are designed to unravel how people make decisions, and ethnographers have long embedded themselves in job sites for months at a time to understand production from the inside[7].

Linking variables to concepts

We need qualitative research to understand how much trust we can place in operationalizations. When faced with large-scale data, researchers often take a variable created with aims other than research and reinterpret it for scientific purposes. But the new meaning that they attribute to the variable may be inaccurate or inappropriate. For example, consider a recent study of how informal mentorship affects scientists' careers[8]. The researchers obtained citations to papers published in ten mostly STEM fields and interpreted the co-authorship of a paper by junior and senior scholars as an indicator of informal mentorship. They found, among other things, that co-authorship by female junior and senior scholars — that is, in their minds, female-to-female informal mentorship — was associated with fewer citations for both mentor and mentee, which prompted the authors to question policies promoting the mentoring of women by other women.

The now-retracted article was criticized for reducing informal mentoring to co-authoring, and neglecting many other dimensions. To their credit, the authors had used survey data to assess whether co-authorship was a good indicator. But the validation survey was unsuccessful, because the questions were too narrow in scope and were not pretested. A stronger approach would have first interviewed potential respondents on how they understand informal mentoring, what might be appropriate indicators and how they interpret potential survey questions. Such 'cognitive interviewing', as it is known, would have helped to produce a stronger survey instrument, probably leading the authors to interpret their results more cautiously. Operationalization lies at the heart of science, and — for variables reinterpreted for new purposes — confidence in the interpretation will often benefit from effective interviewing.

Analytical decisions

We need qualitative research to understand researchers' analytical choices. Many large-scale datasets can be analysed in nearly infinite ways. Researchers must therefore make many subjective choices. This flexibility, which has recently been termed "researcher's degrees of freedom"[9], can result in questionable findings. Solutions that critics have proposed (such as requiring hypotheses to be preregistered) have value, but they rarely help to account for why researchers make different decisions in the first place. That accounting requires studying the researchers' analytical decisions.

A recent study collected the analytical codes of 73 teams testing the same hypothesis with the same data, to examine how their decisions affected their conclusions[10]. The teams followed many different analytical pathways, which, troublingly, resulted in different and even contradictory substantive conclusions. Notably, the study could only account statistically for part of the variation based on teams' expertise, prior beliefs and expectations. They needed more insight into the idiosyncrasies behind the teams' decisions — a topic perfectly suited for in-depth interview and focus group research.

From data to theory to data

We need qualitative research to generate strong theories. The value of large-scale data has been enhanced by the growth of computational social science, which among other things has improved our ability to predict behaviour. But even state-of-the art methods applied to high-quality data may fail at prediction in the absence of good theory. Consider the results of a recent challenge[11]: 160 research teams submitted models to predict life outcomes using longitudinal data with nearly 13,000 variables on more than 4,000 US families. The teams were given background data from the first five waves of the survey, about half of the data for the outcomes in the sixth wave and wide latitude to generate their models. The results were disappointing; the best models accounted for not more than 20% of the variation in the best-predicted outcome[11]. The quantitative researchers who organized the challenge concluded that in-depth interviews, particularly with idiosyncratic cases, would be needed to provide better theories from which to generate predictions[12]. Effective theory will be necessary to predict many social outcomes, and a large dataset alone cannot supply such theory; qualitative research can generate insight about what to look for in the data and how to theorize what is being observed.

Recognizing noise

We need qualitative research to avoid wrongly inferring meaning in ambiguous patterns. Scholars have noted that large-scale data can contribute to the cognitive tendency to observe ostensibly meaningful connections when faced with an ambiguous pattern[2]. Qualitative research can help to make sense of such ambiguities. For example, a study recently used smartphone tracking data from participants at a party to examine social contact on the basis of physical proximity between phones[13]. But the researchers also realized that not all physical proximity is meaningful — people are sometimes standing next to each other but not talking — and that analysts might easily impute meaning to every contact, thus misconstruing random encounters as part of a pattern. They therefore collected in-person ethnographic data to help to discern relevant from irrelevant information, reconstruct the atmosphere of the event and formulate credible interpretations of the patterns observed.

Conclusion

Our list does not exhaust the ways qualitative research can be instrumental to capitalizing on the availability of large-scale data to produce reliable social science. Nor is every item needed in every study based on such data. But the list makes it clear that the increasing availability of large-scale datasets has made qualitative research not less, but in fact more, important. Social science requires not only the ability to detect patterns in data but also the knowledge that the data are adequate and well understood, the patterns meaningful and appropriately interpreted, and the scientists themselves aware of their potential biases and limitations. If so, then the data revolution in social science will require, at its centre, the work of interviewers and ethnographers. ❐

Nikolitsa Grigoropoulou[1] and Mario L. Small[2] ✉

[1]*SOCIUM Research Center on Inequality and Social Policy, University of Bremen, Bremen, Germany.* [2]*Department of Sociology, Columbia University, New York, NY, USA.*

✉*e-mail: mario.small@columbia.edu*

Published online: 28 March 2022
https://doi.org/10.1038/s41562-022-01333-7

References

1. Small, M. L., Akhavan, A., Torres, M. & Wang, Q. *Nat. Hum. Behav.* **5**, 1622–1628 (2021).
2. Boyd, D. & Crawford, K. *Inf. Commun. Soc.* **15**, 662–679 (2012).
3. Lazer, D., Kennedy, R., King, G. & Vespignani, A. *Science* **343**, 1203–1205 (2014).
4. Alba, D. Tracking viral misinformation. *The New York Times*, https://go.nature.com/3Kszjyw (15 September 2021).
5. Strong, D. M., Lee, Y. W. & Wang, R. Y. *Commun. ACM* **40**, 103–110 (1997).
6. Ruths, D. & Pfeffer, J. *Science* **346**, 1063–1064 (2014).
7. Christin, A. *Am. J. Sociol.* **123**, 1382–1415 (2018).
8. AlShebli, B., Makovi, K. & Rahwan, T. *Nat. Commun.* **11**, 5855 (2020).
9. Simmons, J. P., Nelson, L. D. & Simonsohn, U. *Psychol. Sci.* **22**, 1359–1366 (2011).
10. Breznau, N. et al. Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. Preprint at MetaArXiv, https://doi.org/10.31222/osf.io/cd5j9 (2021).
11. Salganik, M. J. et al. *Proc. Natl Acad. Sci. USA* **117**, 8398–8403 (2020).
12. Salganik, M. J., Maffeo, L. & Rudin, C. *Harv. Data Sci. Rev.* **2.3**, https://doi.org/10.1162/99608f92.eecdfa4e (2020).
13. Blok, A. et al. *Big Data Soc.* **4**, https://doi.org/10.1177/2053951717736337 (2017).

Acknowledgements

The authors thank the University of Bremen Excellence Chairs programme for support.

Competing interests

The authors declare no competing interests.