



Online Research Methods and Social Theory

In: The SAGE Handbook of Online Research Methods

By: Grant Blank

Edited by: Nigel G. Fielding, Raymond M. Lee & Grant Blank

Pub. Date: 2017

Access Date: May 21, 2020

Publishing Company: SAGE Publications Ltd

City: 55 City Road

Print ISBN: 9781473918788

Online ISBN: 9781473957992

DOI: <https://dx.doi.org/10.4135/9781473957992>

Print pages: 628-641

© 2017 SAGE Publications Ltd All Rights Reserved.

This PDF has been generated from SAGE Research Methods. Please note that the pagination of the online version will vary from the pagination of the print book.

Online Research Methods and Social Theory

Grant Blank

The enormous growth in online activities has created new opportunities for research. These opportunities are theoretical as well as methodological. The theoretical opportunities have been present in prior chapters but never emphasized; this chapter brings theory into focus without losing sight of methods. Specifically, the chapter discusses the explanatory power of theory based on online methodologies to address important social issues. Using this goal, it describes two themes common in the preceding chapters: big data and the qualitative data revolution. Each theme presents problems as well as opportunities and the goal of this chapter is to explore how methods and theory work together to define and mitigate the problems as well as exploit the opportunities.

The link between methods and theory has a history almost as long as modern science itself. It begins at the dawn of empirical science, over 350 years ago. One of the earliest scientific communities was formed around Robert Boyle, leading a group of experimentalists who were exploring the relationships between pressure, temperature, and volume. Their primary technology was a vacuum pump that they called an 'air pump'. Their findings were codified into what we now know as 'Boyle's Laws'. (Much of the following discussion is drawn from Shapin and Shaffer, 1985 and Zaret, 1989).

These early scientists are interesting, not just because they developed some of the earliest experimental research methods using the advanced technology of the day and not just for their exploration of the relations between pressure, temperature, and volume, but also because their work is intimately linked to social theory. The mid-1600s in England was a period of political turmoil: there was the English Civil War, the Regicide (1649), the creation of the Republic (1649–53), Oliver Cromwell's Protectorate (1653–58), and the Restoration of the Stuart monarchy in 1660. As they attempted to understand this period of bitter political, social, and religious conflict, many Englishmen came to the conclusion that the fundamental source of social conflict was differing views of religious truth. The implication of this assumption was that if everyone believed in the same religion then these extraordinary conflicts would end.

The Restoration of the Monarchy in 1660 had restored a central political authority, but it did not dampen religious strife. As a result, the 1660s were marked by a new crisis of civil authority. The issue was the role of religious belief, particularly the Protestant and Puritan emphasis on an individual's personal religious beliefs. This created a problem that became a major source of tension and conflict.

The problem was that it is very difficult to settle disputes when everyone relies on their own personal vision of truth. Under such circumstances, how can anyone determine whose personal vision is fairer? More just? Or, in any sense, better? In fact, when people believe that their highly individual versions of truth are the only correct version then political compromise and accommodation becomes very difficult. Thoughtful Englishmen saw society and politics splitting into a large number of semi-hostile groups, each suspiciously defending its

personal vision of the truth. This was not attractive, for it looked like the jealous incompatibility of these visions might make a cohesive society with normal politics impossible.

In this social environment, the experimental scientists offered an alternative vision of community. This community claimed to have created an understanding of conflict and social unity that stood in stark contrast to the disorder plaguing English society. Their signal achievement was that they were able to settle disputes and achieve consensus without resorting to violence and without powerful individuals imposing their beliefs on others.

Facts were uncovered by experiments, attested to by competent observers. When there was a disagreement it could be settled by appeal to facts made experimentally manifest and confirmed by competent witnesses from within the community. Stable agreement was won because experimentalists organized themselves into a defined and bounded society that excluded those who did not accept the fundamentals of good order. Consensus agreement on facts was an accomplishment of that community. It was not imposed by an external authority. In this sense, facts were social; they were made when the community freely assented.

This did not imply consensus was always easily reached. Indeed, Hooke's vehement disagreements with Newton and others anticipated a long line of hostile quarrels among scientists. This is another respect in which the early experimentalists formed something that looks like science.

Despite their internal disagreements and despite the inevitable tensions of ego and competition for status, in the context of strife-ridden post-Civil War society the model of a community committed to joint discovery of facts was an attractive alternative. It contributed to the political support required to set up the early institutions that supported and fostered the development of science: the Royal Society and its journal.

The point is that a fundamental link between social theory and research methods was embedded in the culture of science from the very beginning. Neither methods nor theory existed independently of each other. This chapter investigates how theory and methods influence each other in online work. This chapter does not develop new theory; instead, it suggests two possibilities. First, online developments are creating new opportunities for substantive theory just as they are creating new data and new methodology. The new theoretical opportunities come in part from the new social forms and new communities being created by online technologies. They also come from the fact that online research can offer a novel perspective that casts new light on older, pre-existing social forms. Second, certain online methods have an affinity for certain kinds of theoretical explanations. This is an interesting limitation because it means that some theories cannot be developed with certain online methods.

A volume on methods leads naturally to a particular *kind* of theory. This is not the 'grand theory' of Marx, Weber, and Durkheim; instead, it is the middle-range theory or substantive theory that is commonly used in conjunction with standard methodological tools like statistical hypothesis testing. This sort of theory has several relationships to methods. Theoretical concepts are operationalized in scales or indices; in survey questions; by counting or describing attributes of people, organizations, or websites; or by coding qualitative

data into appropriate categories. Relations between concepts are described by hypotheses, which may form the basis for inferential tests. Related hypotheses can be collected into theories that may be modeled with statistical, computational, or mathematical methods. The theory and all its components remain fairly concretely tied to empirical data and to measurement. The payoff from the use of this kind of theory is often a clearer understanding of contemporary social problems or issues. Thus this chapter discusses the explanatory power of theory based on online methodologies to address important social issues.

In the course of this task I draw together many common methodological themes from prior chapters. This is a personal reading of these chapters and no one should infer that my opinions are shared by the authors themselves or by other editors. I found that the papers in this volume each attempt to deal with new opportunities offered by gathering data online while suggesting ways to cope with special problems posed by online research. I generally draw on online examples with some comparison to offline work. I found two common themes. Each theme reflects attempts to deal with new problems or opportunities in online work. They are:

1. Big data
2. The qualitative data revolution

Big Data

Since the first edition of this Handbook, the term 'big data' has become a popular way to describe one of the distinctive characteristics of online research. Online data come not only in large quantities, but also very quickly and in immense variety, and so big data are often seen as consisting of the 'three Vs': volume, velocity and variety (Laney, 2001). The sources of big data are many. Any electronic transaction leaves a digital trace. In cashless financial transactions, communication via email, text messages, mobile telephone call records, wikis, games, photo or video sharing sites, or online interactions with official government agencies, many formerly ephemeral aspects of people's lives are digitally captured and stored. For anyone familiar with the painful cost of collecting data offline, the extent and easy availability of ready-made digital data is breathtaking. Adding to the available data are a number of 'open data' or 'open government' initiatives that seek to make digital data collected by government available to anyone who wants to download it (Longo, 2011; Sieber and Johnson, 2015). The declining cost of storage and ease of accessibility are driving rapid increases in digital record gathering and storage, reinforcing these trends. These developments are likely to continue and they will encourage much more use of online data in future research.

Most of this is unobtrusive, as Dietmar Janetzko's chapter on nonreactive data collection says, in the sense that people are recorded as they go about their ordinary lives. They create data that can be used for research, although it is not generated as research data. Any digitally stored data can be used for research.

Digital records accumulate, leaving minutely detailed records of social interaction. Welser et al. (2007b: 117) comment that such records 'present social scientists with an opportunity to study in unprecedented scale

and scope the dynamics, structure and results of social interaction'. There are many research possibilities here and they are widely recognized, including in five chapters in this volume: Baram-Tsabari *et al.*'s chapter on data mining *and* big data; Bright's chapter on big data; Innes *et al.*'s chapter on social media; Hookway and Snee's chapter on the blogosphere; and Thelwall's chapter on sentiment analysis (see also Lazer *et al.*, 2009).

Some have argued big data has little relation to theory and that it signals the end of theory (Anderson, 2008; Cukier and Mayer-Schönberger, 2013). Anderson argues the case: 'Petabytes allow us to say: "Correlation is enough". We can stop looking for models. We can analyze the data without hypotheses about what it might show'. We can see why this is implausible by looking at the best book ever written about big data, *Moneyball* (Lewis, 2003). Michael Lewis tells the story of how undervalued athletes at one of the poorest teams in professional baseball, the Oakland Athletics, won so many games. Big data are at the center of this story. Big data have a track record in baseball. Detailed baseball statistics go back over a century. You can reasonably wonder, with decades' worth of detailed data available to people who spend their entire professional lives in the game, can anything possibly remain novel or undiscovered? Under general manager Billy Beane, the Athletics discovered that baseball professionals had been looking at the wrong data for decades. For example, batting averages, thought to measure ability to get on base, were not the same as the on-base percent, a more accurate measure of potential runs. Time-honored tactics like bunting and base stealing were ineffective and did not contribute to runs scored. The core of the problem was bad theory. In baseball, as in social science, theory is key because theory tells you what data to look at. If you don't have a good theory, you will look at the wrong data and you will be misled, as baseball professionals were for decades.

It is notable that the people who argue against theory are not themselves experienced data analysts. Perhaps they do not give enough weight to the fact that the initial form of the data is rarely the form that yields the most useful information. Data have to be merged, pooled, scaled, reorganized and transformed in order to make them useful. The on-base percent, for example, is a combination of several statistics. Administrative data is organized and stored using data structures that meet administrative reporting needs, not those that facilitate social science data analysis. Good theory tells you how to convert your data into meaningful numbers. Data critically depend on theory in the sense that theory tells you how to (re-)construct your data so that they become meaningful. Correlations are no better than the data they are based on. There is no such thing as data without theory. Big data without theory is nonsense.

Although big data does not diminish the value of theory, it has a number of subtle, often serious problems. Some are worth describing here because their solutions have theoretical implications. To begin with, many of these data are not available because they are proprietary. Corporations collect them for their economic purposes. Private companies are usually unwilling to supply datasets to researchers because of privacy and competitive fears. Once they give data to a researcher, it is out of their control. The data could be mined for important competitive information if it fell into their competitors' hands; therefore, giving proprietary data to a researcher requires a major leap of faith and trust, with no likely business benefit. It isn't likely to happen easily or often. An example of proprietary data used for research is Marc Sanford's (2007) retail scanner

data. After 14 months of persuasion, requiring what Sanford describes as 'countless hours on the phone' and signing several legal agreements designed to limit the use of the data, and ensure security and confidentiality, Sanford was given over 750 million records. An example of public email is the Enron data (see Klimpt and Yang, 2004; Culotta *et al.*, 2004), discussed in chapters in this Handbook by Janetzko and also Eynon *et al.* It consists of about 200,000 emails exchanged between 151 top executives. The emails were released as part of the court cases that followed the Enron accounting fraud. These are exceptions that prove the rule.

In addition to having access to proprietary data, corporations have another advantage when it comes to use of big data. They can treat anyone as a potential customer, and this vastly simplifies how they use big data. Take Amazon as an example. Whenever I visit Amazon.com, it can treat me as a buyer. It doesn't care that my real motive for looking at a book is because I want the ISBN or because I need to find the correct spelling of the third author's name for a list of references. Amazon's theory that explains my visit is simple and straightforward: I intend to buy. Outside of business schools and marketing, social science theories are usually more complex. For social scientists, both meaning and motives matter. Many actions can have multiple meanings and so it becomes much harder to account for people's actions. The theory required to explain action becomes much more complex. Online big data is often transactional data, and so all researchers know are the things people did. They have no access to attitudes, emotions, motives or meaning. This means big data are much less useful for any theory where meaning plays an important role (but see later for suggestions for extracting meaning from textual data). Furthermore, Amazon loses nothing if its theory is wrong; it can treat me as a potential buyer with no consequences. This isn't true in social sciences. A wrong theory weakens scientists' professional reputations and it diminishes their ability to predict or understand.

Even when something like an Open Government initiative (Smith *et al.*, this volume) solves the proprietary data problem and makes big data available for research, there are three serious problems. First, during the 1980s and 1990s many social sciences went through a methodological debate about the relative value of quantitative and qualitative data. There is no space here to describe the ontological, epistemological, and political issues revealed in the debate (see Sale *et al.*, 2002); my point is that the debate led to a much clearer sense of the relative strengths and weaknesses of each. Quantitative data are best able to generalize to populations using reliable measures. Qualitative data often have greater nuance, a clearer understanding of the importance of context and are better at establishing the mechanisms that generate behavior. Big data are another form of quantitative data and they suffer from exactly the same weaknesses as other quantitative data. They lack nuance, context and are generally unable to capture the subtle context of people's lives (for more detailed discussion of this debate, see Bryman, 1984; Fine and Elsbach, 2000; Lieberman, 2010).

Second, other forms of quantitative data have notable strengths not shared by big data. Big data researchers tend to emphasize correlations between factors because that is something that big data can do really, really well. However, establishing causality requires that data be collected according to research designs that isolate the effects of individual variables. This is critical in certain contexts; many theories make causal claims, and public policy needs specific research designs to establish causal links between policy interventions and relevant outcomes. Big data may be more useful to monitor existing programs as part of an ongoing

established policy. Except for cases where natural or designed experiments are possible, establishing causality is not a strength.

One solution to the first and second problems is the use of mixed methods. Combining qualitative and big data methods can be done in several ways. One, called *complementarity*, uses the strengths of one to offset the weaknesses of the other. A second, called *confirmation or triangulation*, attempts to confirm the findings from one methodology by the other method. Both advantages are very strong and they have contributed to widespread use of mixed methods throughout the social sciences. (Methodologists will recognize that I am oversimplifying, see Small, 2011, for a more complex summary of the use and value of mixed methods).

The third problem is that most digital traces are collected and stored for administrative purposes and their content reflects the needs and convenience of bureaucrats or accountants, not the requirements of good research design or social theory. They typically do not contain fundamental variables that social scientists incorporate into their theories. Education and occupation, for example, are often not included in financial records. Marital status, religion, and ethnicity are rarely available. Everywhere attitudes are missing. When data are voluntarily deposited on social media, websites, blogs, listservs, Usenet, and online games, they only contain the information that users think important, which is inconsistent from person to person. It is usually impossible to get income, education, marital status, gender, or age. While there are interesting attempts to infer race and gender based on names, these have large errors and they remain in the proof of concept stage (Mislove *et al.*, 2011; Sloan *et al.*, 2013). Big data is not a substitute for reliable measures of variables that make a big difference in people's lives. With many theoretically important variables unavailable, people are, at best, thinly described. How much can researchers know about a social setting when they know so little about the people? Big data may be a mile wide, but it is an inch deep. Big data is thin data.

Geography is one solution to this third problem. Some kinds of big data have a geographic basis – they are, for example, coded with postcode or zip code information – and they can be merged with other geographically coded data like census data or police crime statistics. The unit of analysis becomes a spatial unit, like American census tracts, British census Output Areas, London Boroughs, or Chicago neighborhoods. The variables are something like the proportion of each gender or the proportion of each marital status in each borough. This can add race, ethnicity, education, income, crime rates, house prices, or whatever substantive variables are missing. This is how Sanford (2007) added theoretically important variables to his retail scanner data. The availability of key substantive variables for geographical units is leading to a surge in spatially based social analysis and theory (see chapters by Zook *et al.*, Martin *et al.* and Smith *et al.*, this volume). We may be entering the golden age of geography.

In almost all cases, big data require time-consuming, highly skilled work to put datasets into a condition where interesting substantive problems can be addressed. They are usually stored in a format designed for efficient administrative storage and reporting. This is almost never the form needed for statistical analysis or network analysis. Appropriate data have to be extracted from the existing tables and merged, disaggregated or aggregated to theoretically meaningful levels and into a format that software used for statistical analysis or network analysis will accept. Such tasks are time-consuming even for skilled people and they require serious

data management skills, which are usually not taught as part of graduate training and are, in fact, rare among social scientists. There will be social science uses for some of this new data, but they depend on creative, imaginative thought to make them workable. A notable example is Marc Smith's NetScan work collecting Usenet mail headers (see Welser *et al.*, 2007a; Welser *et al.*, 2007b). This example is in many respects typical of other attempts to create useful social data and theory from electronic traces. The dataset is a collection of every possible Usenet message header, 1.2 billion of them. This is the social science version of collecting the entire haystack of data.

An alternative strategy complements the 'collect everything' approach of NetScan. It involves the use of sampling using a mobile phone application. Mihaly Csikszentmihalyi's experience sampling method (ESM) stands in sharp contrast to both qualitative and quantitative 'collect everything' research strategies. The ESM is a data gathering technique that exploits mobile technology; subjects are given a mobile phone app (originally a beeper). They are sent a short set of questions at random times during the day. The questions ask what they are doing and how they feel. This gives a continuous stream of data about individuals' moods, opinions, activities, and interactions with other people. Hektner *et al.* (2007) has details and summarizes the entire research stream. This method too has weaknesses, but they are different weaknesses. They are mostly the well-known problems of questionnaire research: response rates, sampling bias, reliability, validity, and others.

The use of the ESM has a variety of advantages. The questionnaires give more direct access to internal states; to attitudes, emotions and meanings. Questionnaires can be designed to ask about theoretically grounded empirical categories. Finally, the data are based on a random sample of times of the day. The point is, instead of spending the time and money to collect everything, and having to spend more time deciding what you really want, and then throwing away all the data you collected that you decided that you don't need, you can simply collect what you wanted to begin with. Samples are really valuable; they are much faster and easier to collect and to analyze. You lose very little by employing a sample. I think it is a reasonable methodological question: under what circumstances is there value in collecting more than a random sample? Why not collect only the data you need in the first place?

I was born in Missouri, which is called the 'show-me' state. This nickname supposedly comes from Missourians habit of asking people to 'show me' the evidence. There is something to be said for this; in the context of online methodology, show me the theoretical payoff. It is hard to find a concept as striking or as influential as Csikszentmihalyi's (1991) idea of flow, developed out of his studies using the ESM.

In addition to administrative or government-collected data, big data has a second form: online researchers who can collect their own data have it easy. Internet and online surveys are cheap and fast compared to the offline alternatives. Responses can be automatically checked for consistency and stored directly in a dataset, ready for analysis; indeed, simple analyses, such as descriptive statistics, frequencies and standard cross-tabulations, can be automatically produced. This can largely eliminate the time-consuming, difficult, and costly steps of data cleaning and data input. Five chapters in this Handbook discuss this: Vehovar and Manfreda's overview of online surveys; Fricker's sampling methods for online surveys; Toepoel's Internet

survey instrument design; Kaczmirek's Internet survey software tools; and Dillman *et al.*'s mixed mode surveys. The special problem of online survey research is that there is no way to construct a sampling frame. There is no online equivalent to random digit dialing or postal address files; therefore, it is not generally possible to select online respondents according to some randomized process. Even if Internet usage reached saturation levels, for most populations the sampling frame problem would remain intractable. There are exceptions, as Fricker points out in his chapter, for example a survey of an organization where the organization has a complete list of its members and everyone has an email address. But this is unusual. A more general solution to this problem, discussed by both Fricker and Dillman *et al.*, is to use mixed mode research. Sometimes online and offline data collection can be combined to overcome the lack of a random sample in online research while still retaining most of the advantages of low cost and easy administration.

Without a sampling frame, almost all Internet surveys rely on self-selected respondents collected into large, pre-constituted panels. This yields cheap, speedy results, but the problems are serious. Respondents are often recruited from advertisements on web pages and they are paid for participation. This produces a pool of respondents guaranteed to be biased. Surveys usually attempt to correct their selection bias by using quota sampling of the panel and using post-stratification weights. These attempts are only partially successful. The core problem is that this process is almost guaranteed to produce biased data. The Pew Research Center (Keeter *et al.*, 2015) compared an unusually high quality online sample with a randomly selected offline sample. It found that there were significant differences, particularly in responses to Internet-related variables. Some of these problems became embarrassingly public in the polling errors predicting the 2015 British general election outcome (Curtice, 2016; Sturgis *et al.*, 2016). The accuracy of Internet polls may be sufficient for marketing purposes, but scholarly use remains problematic.

For other online data, like clicks, Tweets, and email messages, sample sizes can be, at least potentially, extremely large. Website click-through data can have sample sizes of over 100 million cases (remember, that is the sample!). As the Hindmarsh chapter on real-time video analysis points out, cheap cameras and inexpensive disk storage increase the feasibility of video recordings. The resulting data can contain records of individuals, almost unprecedented in their details.

The low cost of online data collection and the possibility for using video have been widely noticed. Less widely recognized is the fact that the low cost and easy access to subjects also applies to ethnographic research. Six chapters describe the implications of online research for various forms of qualitative data collection: Hine's chapter on virtual ethnography; O'Connor and Madge's chapter on online interviewing; Hookway and Snee's chapter on blogs; Abrams and Gaiser's chapter on virtual focus groups; the Hindmarsh chapter on real-time video; and the Silver and Bulloch chapter on new affordances of CAQDAS. Enormous amounts of qualitative data can be collected very quickly. For blogs, listservs, social media, or email the digital form is the only form in which the data exist. These data do not need to be converted to digital form by transcription and this eliminates major costs, time delays, and sources of error. The new wealth of data opens a real opportunity for all kinds of innovative research.

There is no free lunch in online data collection. One price of simple, low-cost access to subjects is a

set of complicated, difficult ethical questions. Ethics are discussed in several chapters, but they are comprehensively addressed in the Eynon *et al.* chapter on ethics and ethical governance. There are two primary protections for social science research subjects, anonymity and informed consent, and under online conditions both are more difficult to achieve. The same easy access to online data and the ease of matching individual respondents to other datasets that makes online data collection so much simpler also makes it much easier for someone to break anonymity and discover the identity of individual respondents. Files in all their versions are often preserved on backups, which are often automatically created and not always under control of the researcher. Someone could obtain an early version of a file with identifiers still in place, possibly without the researcher even knowing. An obvious situation where this could occur is when a government agency is interested in a research project that collected data on illegal actions, such as criminal behavior, drug use, illegal immigration, or terrorism.

What will we do with all these Data?

The signal characteristic that distinguishes online from offline data collection is the enormous amount of data available online. As we think about our newfound wealth of data, I want to raise two skeptical questions, one for qualitative data and the second for quantitative data.

The sources of massive amounts of qualitative data that are being and could be collected include automated security cameras, social media content, as well as purposefully collected data like video and audio tapes. These data promise a remarkably fine-grained, detailed picture of people in all kinds of social situations. Given this fact, here is the place to raise the key question: so what? What is the payoff? Does the availability of this enormous volume of data promise soon-to-come advances in our understanding of society, politics, and culture, along with much better theory? My best-guess answer to this question is 'probably not'.

To see why, it is necessary to realize that detailed qualitative data is not new. Ecological psychologists (e.g. Barker and Wright, 1951, 1954) collected it over 60 years ago. Barker and his students created and published minute-by-minute records of the activities of children from morning to night. We can reasonably ask, if there is going to be a major payoff from data-intensive studies of social life, why haven't we heard more about the ecological psychologists? Why aren't they more important? Why didn't people pay attention?

One answer is that the theory Barker developed was exceptionally closely tied to concrete social settings and situations. Barker studied under Kurt Lewin and he was influenced by Lewin's theories of the importance of the environment in predicting behavior. Barker himself argued that behavior was radically situated, meaning that accurate predictions about behavior require detailed knowledge of the situation or environment in which people find themselves. His work often consisted of recording how expected behavior is situational: people act differently in different behavioral settings, for example in their roles as students or teachers in school or as customers in a store. In his theory of behavior settings, Barker is fairly explicit that he believes broad, 'grand' theories cannot usefully predict behavior. Since Barker focuses so tightly on behavior in a very local setting, his research doesn't generalize very well to other settings. This is an implication built into the

idea of behavioral settings and it is intentional. Ecological psychology has sustained its intellectual ground as a school, but its results turn out to be fairly limited in their applications. This is a disadvantage. Other researchers, looking for theories that help them in their research, will not find rich sources of insightful ideas that they can use. Since few other researchers found the ecological psychologists' work useful, it was not widely adopted.

The ecological psychologists actually observed people, and observers can gain insight into emotions because they are often visually inescapable. Transaction data may be voluminous but transactions provide no realistic way to infer internal emotional states. This is a serious limitation because much human action depends on meaning. The same actions can have multiple meanings and, for different people, they can have completely different meanings. Without the ability to gain access to meanings it is very hard to develop or test any theory based on meaning, emotions, or any mental states.

Of course, no researcher has complete access to internal states like emotions or meaning. To a greater or lesser extent, meanings, motives, or other mental states always have to be inferred. It is a matter of the degree to which meaning can be inferred from particular kinds of data. The point here is that many forms of big data supply much less direct access to internal states than some other kinds of data.

The problems of fine-grained data raise a key question: under what circumstances are detailed, fine-grained data about social life useful? Most answers to this question describe the use of case studies in data-intensive research. Burawoy (1991) developed the 'extended case method' as a way to use detailed case studies to identify weaknesses in existing theory, and to extend and refine theory, for example by describing subtypes of a phenomenon. Ragin (1987) links case studies to the study of commonalities where comparable cases are studied to construct a single composite portrait of the phenomenon. Case studies are holistic and they emphasize causal complexity and conjunctural causation. This use of cases is similar to that used by anthropological ethnography, where holistic understanding is a central goal. What can we learn from these two examples? Both emphasize the importance of case selection.

If you can only study a few cases, then how you select those cases is key. Flyvbjerg (2006) summarizes four case selection strategies to maximize the researcher's ability to understand the phenomenon (see also Ragin, 1992). Average or ordinary cases are not rich sources of information; instead, extreme or deviant cases may reveal more about the relevant actors and mechanisms. These cases are chosen to emphasize a central aspect of a phenomenon. Often a theoretical sampling strategy is used: choose cases as different as possible. If this 'maximum difference' strategy is followed, then any commonalities discovered are much more likely to be fundamental to the phenomenon rather than artifacts of a biased selection of cases. Third, critical cases have special characteristics or properties that make them unusually relevant to the problem. For example, a case can be chosen because it seems most likely to *disconfirm* the hypothesis of interest. If the hypothesis is *confirmed* for this case, then the researcher can argue that it is likely to be true in all less critical cases. The researcher argues 'This case had the best chance of falsifying my argument and it failed; hence, my argument must be true'. Finally, cases can be selected because they form an exemplar. These cases form the basis for exemplary research that shows how a particular paradigm (Kuhn, 1970) can be applied in

a concrete research setting. An example is Geertz's (1973) study of the 'deep play' of the Balinese cockfight. The importance of case selection underlines that there is no substitute for good research design. For fine-grained data to be useful, it must be carefully chosen to illuminate issues of broader interest.

Research design is usually driven by a theoretical understanding of what to investigate. Theory is important because it gives direction and focus to research. It identifies important issues and categories. It suggests the kinds of research settings and data that could speak to those issues. It suggests relevant related concepts to investigate. This can be overstated and many theories are also vague and incomplete. In practice there are limits to the ability of theory to guide empirical research. But within these limits, theories play a major role in research design.

There are some research settings where individual cases are of great interest: the French Revolution is one example. But under most circumstances, cases are more interesting as examples of a more general phenomenon. Here theory plays a key role: theory connects cases. It is broader than individual cases and so it tells us which cases are examples of the same event or situation and which cases are different. Powerful theories can link disparate settings that seem to have little in common, and show how they are actually examples of the same phenomenon. For researchers, the categories and links between categories that make up a theory supply conceptual tools to help them think about their research, their research site(s), and their data. Theories are good to think with. As noted earlier in the discussion of ecological psychology, one of the most valuable payoffs from theory is that theories developed in one setting may serve as sources of creative ideas for researchers working in other settings.

The Qualitative Data Revolution

Although our ability to record social transactions has increased dramatically, our ability to analyze the recorded data has not expanded nearly so fast. On one hand, certain types of analysis are much easier today. All types of quantitative analysis, for example, have benefited. One effect of the additional computational power is that many more models can be examined and model diagnostics are easier. Statistics have always been a way to summarize data. In general, the ideas of central tendency, spread, and other statistical concepts can summarize a large dataset about as effectively as a small one. As datasets become larger, the nature of the statistical summaries does not change. Networks and data archives give researchers convenient access to statistics and data that they could never use before (see Corti and Wathan's chapter on archives and secondary analysis, and Carmichael's chapter on secondary analysis of qualitative data). The use of games for research purposes has blossomed (see Verhagen *et al.*'s chapter) and simulations of social settings have become much easier (see Elsenbroich's chapter).

In general, computer power makes possible much more thorough explorations of data using statistical, computational, and graphical techniques. Graphical analysis and visualization has blossomed remarkably with the increase in computing power. Computers draw all kinds of diagrams and plots so much faster than they can be drawn by hand. The advances in online research methods have been almost wholly positive for

quantitative researchers.

Qualitative research is different. Collection of qualitative data has always been extremely slow and difficult. No longer. Vast quantities of text are now available in digital form. Silver and Bulloch's chapter on new affordances of CAQDAS highlights the collection of new forms of data: web pages, online role playing games, emails, blogs, video, still images, etc. Carmichael's chapter on secondary qualitative analysis describes developments in secondary analysis that further increase the availability of data. There have also been advances in qualitative, non-statistical analysis. Brent's chapter on artificial intelligence and Hindmarsh's chapter on video analysis point to some of those advances. The process has been improved by the use of qualitative analysis software like NVivo, Atlas.ti, Qualrus, or QDA Miner. The software adds reliability and speed, as well as new capabilities like Boolean retrievals and semi-automated coding of text. There is a lot of potential here. Qualitative analysis of text has always been a particularly strong method to gain access to attitudes, motives, and meaning.

Qualitative analysts have mostly reacted to their newfound wealth of data by ignoring it. They have used their new computerized analysis possibilities to do more detailed analysis of the same (relatively small) amount of data. In spite of the deluge of textual data and the advances in artificial intelligence and in software, qualitative analysis has not changed much. Most current forms of qualitative analysis are too slow and require too much hand work. They don't scale well and so they cannot be used to analyze the volumes of readily available text. Qualitative analysis was once a theory-rich, data-poor field. Now the biggest bottleneck for qualitative work has shifted. It used to be data collection but that is no longer true. Qualitative analysis is no longer limited by the difficulty of getting text; it is now data-rich but methodology has not kept up. In a nutshell, the need is for a way to summarize large amounts of text in a theoretically meaningful way. The enormous amount of text requires an automated summary process.

Considerable work has been done on the automated processing of text. I am aware of projects working with web pages, email, blogs, and online versions of newspapers. The most sophisticated work is proprietary, owned by corporations or governments. Major statistical software companies have produced text mining software, for example SAS Institute's Text Miner, SPSS's LexiQuest products, and Stata's link to QDA Miner. These are designed to analyze unstructured text with machine learning, natural language processing, and visualization tools. Popping's chapter (this volume) on analysis of text covers these approaches and more. The academic work is based on qualitative analysis software, such as NVivo, MAXQDA, or Qualrus. There is also software specifically developed to support mixed mode research, like QDA Miner. Looking at these products, it is clear that any sort of truly automated processing of text is in the future, except in some highly restricted domains where controlled vocabularies can be used.

We need a procedure that is (1) automated; (2) scalable to large amounts of text; and (3) sensitive to meaning. Meaning can only be derived from the context. Determining context and meaning is not something that computers do well. This has been a crippling limit on the use of automated categorizing of text or other qualitative data (see Thelwall's chapter). One technique promises a scalable way to summarize text that also handles context. The contextual possibility is crude by the standards of qualitative research, but it is an

approach to meaning. By this, I do not mean that computers can determine meaning – that is still a problem for the analyst – but there are techniques that are context-sensitive and that is a crucial first step. Topic models are a promising class of text analysis methods. They provide a (relatively) automated procedure to code the content of a body of text into a (relatively) small set of substantively meaningful categories called 'topics'.

A 'topic' is understood as a group of words that appear together frequently. Topics can be thought of as a single theme or a coding category. The model of authorship for topic models is that authors have a bag of words that they use to write about a topic. As they write they reach into their bag and choose the most appropriate words. Different topics generally have different bags of words, although some words may be used in multiple topics. Topic models attempt to reconstruct the bags of words that authors use for each topic. The primary output is the list of words associated with each topic. The lists of words are believed to reflect the hidden structure of topics in the text.

The task of the analyst is to work backwards from the words to identify the original topic. This is an inductive process. Topics are not labeled by the program. There is no sense in which any topic modeling algorithm 'understands' the words. The analysis must determine the meaning of the text. The goal of the analyst is to find topics that are substantively meaningful and theoretically interpretable. For examples of social science use of topic models, see Mohr and Bogdanov (2013) and their references.

Topic models are a class of procedures that try to group things together. They are classification algorithms. Many possible algorithms are available, from well-known procedures like principal components and cluster analysis to recently developed procedures like latent Dirichlet allocation. Relatively little research has been done to investigate how or whether different algorithms produce different results on real data. Topic model algorithms are beginning to be integrated into qualitative analysis software. SAS Institute's Text Miner and Provalis' QDA Miner both have topic modeling available. There is a serious need for qualitative methods that scale to large amounts of text and are sensitive to context and meaning. These are tools that can complement traditional qualitative methods. I look forward to a productive convergence of computer science and social science to advance social theory.

The Irony of Big Data

Big data holds great promise, but we are just beginning to understand what it is useful for. There is great irony in the current situation. The tools to handle numerical forms of data are well-developed and widely available, but they have little ability to help us research meaning. On the other hand, textual big data gives us much better access to meaning, but the methodologies for handling text are weak and underdeveloped. Their potential is mostly unrealized. For big data to really fulfill its promise, we really need to do both. Lewis (2003) has shown how baseball was transformed when big data was combined with imaginative new theories. Can the same be done in the social sciences?

References

- Anderson, C.** (2008) 'The end of theory: the data deluge makes the scientific method obsolete'. *Wired*, June 23. <http://www.wired.com/2008/06/pb-theory/> (accessed on December 15, 2015)
- Barker, R. G. and Wright, H. F.** (1951) *One boy's day: a specimen record of behavior*. New York, NY: Harper.
- Barker, R. G. and Wright, H. F.** (1954) *The Midwest and its children: a psychological ecology of an American town*. Evanston IL: Row, Peterson.
- Bryman, A.** (1984) 'The debate about quantitative and qualitative research: a question of method or epistemology?', *British Journal of Sociology*, 35 (1): 75–92.
- Burawoy, M.** (1991) 'The extended case method'. In **M. Burawoy** (ed.), *Ethnography unbound: power and resistance in the modern metropolis*. Berkeley, CA: University of California Press, pp. 271–89.
- Csikszentmihalyi, M.** (1991) *Flow: the psychology of optimal experience*. New York, NY: HarperCollins.
- Cukier, K. and Mayer-Schönberger, V.** (2013) *Big data: a revolution that will transform how we live, work and think*. New York, NY: Houghton-Mifflin.
- Culotta, A., Bekkerman, R. and McCallum, A.** (2004) 'Extracting social networks and contact information from email and the Web'. *Proceedings of the First Conference on Email and AntiSpam* (CEAS) July 30–31, Mountain View, CA. Available at http://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1035&context=cs_faculty_pubs (accessed July 28, 2016).
- Curtice, J.** (2016) *British Social Attitudes. The benefits of random sampling: lessons from the 2015 UK general election*. London: NatCen. Available at <http://www.bsa.natcen.ac.uk/media/39018/random-sampling.pdf> (accessed February 25, 2016).
- Fine, G. A. and Elsbach, K. D.** (2000) 'Ethnography and experiment in social psychological theory building: tactics for integrating qualitative field data with quantitative lab data', *Journal of Experimental Social Psychology*, 36: 51–76.
- Flyvbjerg, B.** (2006) 'Five misunderstandings about case-study research', *Qualitative Inquiry*, 12 (2): 219–45.
- Geertz, C.** (1973) 'Deep play: notes on the Balinese cockfight'. In **C. Geertz** (ed.), *The interpretation of cultures*. New York, NY: Basic Books, pp. 412–54.
- Hektner, J. M., Schmidt, J. A. and Csikszentmihalyi, M.** (2007) *Experience sampling method: measuring the quality of everyday life*. Thousand Oaks, CA: Sage Publications.
- Keeter, S., McGeeney, K. and Weisel, R.** (2015) *Coverage error in Internet surveys: who web-only surveys*

miss and how that affects the results. Washington DC: Pew Research Center. Available at http://www.pewresearch.org/files/2015/09/22_coverage-error-in-internet-surveys.pdf (accessed July 28, 2016).

Klimt, B. and Yang, Y. (2004) 'Introducing the Enron corpus'. *Proceedings of the First Conference on Email and Anti-Spam* (CEAS) July 30–31, Mountain View, CA. Available at <http://nl.ijs.si/janes/wp-content/uploads/2014/09/klimtyang04a.pdf> (accessed July 28, 2016).

Kuhn, T. S. (1970 [1962]) *The structure of scientific revolutions*. 2nd edn. Chicago, IL: University of Chicago Press.

Laney, D. (2001) *3D data management: controlling data volume, velocity, and variety*. META Group, February 6. Available at <https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf> (accessed July 29, 2016).

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J. and Gutmann, M. (2009). 'Life in the network: the coming age of computational social science', *Science*, 323 (5915): 721.

Lewis, M. (2003) *Moneyball: the art of wining an unfair game*. New York, NY: W.W. Norton.

Lieberman, E. S. (2010) 'Bridging qualitative-quantitative divide: best practices in the development of historically oriented replication databases', *Annual Review of Political Science*, 13: 17–59.

Longo, J. (2011) '#Opendata: digital-era governance thoroughbred or new public management Trojan Horse?', *Public Policy & Governance Review*, 2 (2): 38.

Mislove A., Lehmann S., Ahn Y., Onnela J. and Rosenquist J. (2011). 'Understanding the demographics of Twitter users'. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, July 17–21, Barcelona, pp. 554–57.

Mohr, J. and Bogdanov, P. (2013) 'Introduction – topic models: what they are and why they matter', *Poetics*. 41: 545–69. doi:10.4135/9780857020055.n7.

Ragin, C. C. (1987) *The comparative method: moving beyond qualitative and quantitative strategies*. Berkeley, CA: University of California.

Ragin, C. C. (1992) "'Casing" and the process of social inquiry'. In **C. C. Ragin and H. S. Becker** (eds.), *What is a case? Exploring the foundations of social inquiry*. Cambridge: Cambridge University Press, pp. 217–26.

Sale, J. E. M., Hohfeld, L. H. and Brazil, K. (2002) 'Revisiting the quantitative-qualitative debate: implications for mixed-methods research', *Quality & Quantity*, 36: 43–53.

Sanford, M. (2007) Consumption and the urban milieux: using consumption as a measure of similarity for

defining urban neighborhoods. PhD dissertation, unpublished, University of Chicago.

Shapin, S. and Schaffer, S. (1985) *Leviathan and the air-pump: Hobbes, Boyle, and the experimental life*. Princeton, NJ: Princeton University Press.

Sieber, R. E. and Johnson, P. A. (2015) 'Civic open data at a crossroads: dominant models and current challenges', *Government Information Quarterly*. doi:10.1016/j.giq.2015.05.003.

Sloan, L., Morgan, J., Housley, W., Williams, M., Edwards, A., Burnap, P. and Rana, O. (2013) 'Knowing the Tweeters: deriving sociologically relevant demographics from Twitter', *Sociological Research Online*, 18 (3): 7. Available at <http://www.socresonline.org.uk/18/3/7.html> (accessed July 29, 2016). doi:10.5153/sro.3001.

Small, M. L. (2011) 'How to conduct a mixed methods study: recent trends in a rapidly growing literature', *Annual Review of Sociology*, 37: 57–86. doi:10.1146/annurev.soc.012809.102657.

Sturgis, P., Baker, N., Callegaro, M., Fisher, S., Green, J., Jennings, W., Kuha, J., Lauderdale, B. and Smith, P. (2016) *Report of the Inquiry into the 2015 British general election opinion polls*. London: Market Research Society and British Polling Council. Available from http://eprints.ncrm.ac.uk/3789/1/Report_final_revised.pdf (accessed July 29, 2016).

Welser, T., Gleave, E., Fisher, D. and Smith, M. (2007a) 'Visualizing the signatures of social roles in online discussion groups', *Journal of Social Structure*, 8 (2). Available at <https://www.cmu.edu/joss/content/articles/volume8/Welser/> (accessed December 15, 2015).

Welser, T., Smith, M., Fischer, D. and Gleave, E. (2007b) 'Distilling digital traces: computational social science approaches to studying the Internet'. In **N. Fielding, R. Lee and G. Blank** (eds.), *Handbook of online research methods*. London: Sage Publications, pp. 116–41. doi:10.4135/9780857020055.n7.

Zaret, D. (1989) 'Religion and the rise of liberal democratic ideology in seventeenth-century England', *American Sociological Review*, 54 (April): 163–79.

<http://dx.doi.org/10.4135/9781473957992.n36>