Review of Accounting Studies (2022) 27:1117–1124
https://doi.org/10.1007/s11142-022-09708-x

Practical issues to consider when working with big data

Lorien Stice-Lawrence[1]

Accepted: 20 July 2022 / Published online: 5 August 2022

Abstract
Increasing access to alternative or "big data" sources has given rise to an explosion in the use of these data in economics-based research. However, in our enthusiasm to use the newest and greatest data, we as researchers may jump to use big data sources before thoroughly considering the costs and benefits of a particular dataset. This article highlights four practical issues that researchers should consider before working with a given source of big data. First, big data may not be conceptually different from traditional data. Second, big data may only be available for a limited sample of individuals, especially when aggregated to the unit of interest. Third, the sheer volume of data coupled with high levels of noise can make big data costly to process while still producing measures with low construct validity. Last, papers using big data may focus on the novelty of the data at the expense of the research question. I urge researchers, in particular PhD students, to carefully consider these issues before investing time and resources into acquiring and using big data.

Keywords Big data · Alternative data · Emerging technologies · Research design

JEL classifications: A1 · B4 · C55 · G00 · M00 · M4

1 Introduction

Researchers have increasing access to alternative or "big data" sources, giving rise to an explosion in the use of this type of data in economics-based research. However, researchers embarking on new projects using big data sometimes fail to recognize that big data, while novel, is not a panacea for the problems faced when working

✉ Lorien Stice-Lawrence
sticelaw@usc.edu

[1] University of Southern California, Los Angeles, CA, USA

with more traditional data sources. This article highlights key issues that researchers should consider *before* working with a particular source of big data in order to ensure that any prospective project has the potential to convincingly test an economically interesting and impactful research question. Many of the issues I discuss are common to *any* source of new data, not just big data, and are therefore familiar to experienced data veterans. As a result, my remarks are targeted primarily at PhD students. Blankespoor et al. (2022, also in this issue) is presented as one example of a study that has largely overcome the issues with big data presented here.

"Big data" refers to data that is larger and more complex than traditional machine-readable data sources; so much so that traditional approaches are often inadequate and new techniques are needed to process and analyze the data. When defining big data, Oracle, the world's largest database management company, refers to terabytes or even petabytes of data (Oracle 2022). In actuality, much of the data frequently referred to as "big data" in business research is really not that big, especially because researchers often only use small slivers of what is available. As a result, the term "alternative data" is perhaps a more accurate way to describe the broad set of new data sources increasingly being used. For parsimony, I refer to either alternative data or big data collectively as "big data".

In this paragraph, I briefly describe several types of big data sources that have been used by researchers in accounting and financial economics. However, the high-level issues I discuss in Sect. 2 apply to the use of big data in other disciplines as well. One of the earliest sources of big data used in accounting and finance research was *textual* content, for instance company disclosures, such as annual reports, press releases, and conference call transcripts, but also other sources of text, including news articles, website content, product reviews, and analyst reports (e.g., Loughran and McDonald 2016). Textual data can be large and is inherently unstructured, necessitating additional processing before it can be used in statistical analyses. Similarly, *images*, such as those of executives' faces or their signatures (e.g., Ham et al. 2017), as well as *audio-visual* content, such as conference calls or roadshow presentations (e.g., Hobson et al. 2012), require specialized processing techniques. In addition, research is increasingly making use of *social media* data: company and customer tweets, friend networks, and content likes, to name a few (e.g., Lee et al. 2015). *Location and movement* data captured by cell phone signals, satellite images, and taxi data can approximate physical movements of customers, shipments, or investors (e.g., Kang et al. 2021) while *transaction and point-of-sale* data can approximate what is being purchased, how much, and when (e.g., Blankespoor et al. 2022). Similarly, *digital footprint* data from web search and browsing, app usage, and online transactions can track users' online movement and behavior (e.g., Froot et al. 2017). More broadly, *IoT (Internet of Things)* devices such as smart and connected devices in home, healthcare, manufacturing, retail, and transportation settings generate a vast amount of data. While most IoT data in accounting and finance up to this point is restricted to cell phone data, this data will likely play a much larger role in our research going forward. Although this summary of big data sources is incomplete, and many of these categories are non-mutually exclusive, an overall takeaway is that "big data" is an umbrella term that encompasses many different types of data.

This article does not intend to give technical software tips or in-depth analysis of each of the many types of big data, and I refer readers interested in more extensive descriptions of alternative data sources and their potential applications to Cong et al. (2021) and Teoh (2018). Instead, I focus on the common issues that can arise when dealing with *any* of these big data sources. Because the issues that I discuss, if ignored, can limit the econometric rigor and contribution of research, successfully published papers using big data are those that have *overcome* these issues through careful research designs and well-thought-out research questions. This sample selection of papers in the public record may give the false impression to researchers considering working in this area, in particular PhD students, that these sorts of issues may simply "fall into place" after the researcher has found a sufficiently interesting source of big data. However, what are not as visible are the many projects that were eventually abandoned because the authors were *not* able to overcome the hurdles discussed below. As a result, this article is not a review of published research using big data but instead highlights the high-level issues researchers should consider at the initial project planning and selection stage. In Sect. 2, I outline the four key issues to consider when working with big data and apply this approach to Blankespoor et al. (2022).

2 Key issues to consider when conducting research using big data

2.1 Is it conceptually different?

The first question to ask when considering whether to use a particular source of big data is whether the data is conceptually different from more traditional data. This is especially important if prior research has already examined a similar research question. In other words, prospective researchers need to decide whether they believe the use of big data allows them to answer *new* questions, or if instead it allows them to answer *old* questions with more precision, for example because of increased statistical power. In most cases, the potential contribution and impact of the former is greater than the latter. Although a new source of big data may be unique in many ways, it may capture a very similar construct as data used in prior research. In that case, a researcher may inadvertently end up testing an old question without realizing it. The similarities with prior data may not just be at the construct level; big data can still suffer from the same empirical pitfalls as traditional data. For example, Teoh (2018) points out that studies using alternative data are still subject to endogeneity concerns. As a result, researchers should carefully review the prior literature and clearly articulate to themselves how their big data source differs from previously used data at the construct level, or how it overcomes an empirical challenge.

The largest contribution of using big data comes when researchers can address questions that were previously unanswerable with more traditional data sources. As a case in point, Blankespoor et al. (2022) in this issue uses real-time transaction data to estimate the performance information available to managers within a given fiscal quarter. They aggregate individual credit and debit card transaction to the firm-week level so that they can estimate a firm's performance during the quarter as of a given

week. They then examine how managers' disclosure choices dynamically respond to performance throughout the quarter, which managers observe in real time as the quarter progresses. This research question would be impossible to address using traditional performance data, which provides only a summary measure of aggregate firm performance for the entire quarter. Big data can provide us with the opportunity to study behaviors that may have occurred before its existence but were previously unobservable.

In addition, big data may allow us to answer new questions if individuals and firms behave differently in the presence of big data. For example, Zhu (2019) examines how the availability of big data about a given firm disciplines managers to make better decisions on behalf of shareholders. However, in some cases, examining whether individuals behave differently in the presence of big data is conceptually identical to examining whether individuals behave differently in the presence of *data.* Big data is just one type of data that is made up of many signals rather than just a few. As a result, the new availability of alternative data sources in the last two decades is conceptually similar to the introduction of the internet and widely accessible electronic databases (such as EDGAR) in the 1990s or even the original introduction of CRSP and Compustat in the 1960s. That being said, big data is more timely and more granular than traditional databases and has the ability to track behavior at the individual level in ways that have never been seen before. Consequently, its presence can have different effects on behavior than coarser data. Zhu (2019) exploits this aspect of big data by showing that firm outsiders can monitor firms more effectively when they have access to real-time big data than when they had access only to traditional financial reports.

2.2 Lots of data, few subjects

In spite of the name, big data can still suffer from small sample problems. Researchers often end up with a surprisingly small number of subjects when the data is aggregated to the unit of interest (e.g., to the firm level). For example, Blankespoor et al. (2022) begins with 1.6 billion individual transactions but aggregates these to measure the performance of only 243 firms, all in the retail sector. The result is that studies using big data can suffer from a lack of statistical power. Sometimes such sample limitations are inherent to the data: only retail firms have point-of-sale data, and only healthcare companies have patient medical records. These sorts of sample limitations aren't problematic if the sample *is* the population of interest (i.e., retail or healthcare, respectively). If the goal is to predict hospital readmissions, it makes sense to only examine healthcare facilities. The problem comes when trying to make inferences outside of the original sample. Studies that wish to answer research questions about broad economic phenomena need to ensure that the economic forces of interest are not affected by unique attributes of the sample in which these research questions are tested.

Blankespoor et al. (2022) uses a sample of retail firms to document that managers respond to real-time information by suppressing negative information at the beginning, but not the end, of the fiscal quarter. However, if disclosure responses to real-time information vary across industries, then the results of this study may not

generalize. This could be the case if the disclosure response to real-time information varies based on operating cycle length. Firms with relatively short operating cycles (such as retail firms) may initially choose to suppress negative real-time performance information if they expect that they may be able to take corrective actions and change performance during the quarter. On the other hand, firms in industries with relatively long operating cycles (for example, manufacturers of large machinery like Boeing) may not be able to respond quickly enough to materially change performance within a single quarter. As a result, those firms would have less incentive to delay bad news. Retail firms might also have more incentive to suppress bad news if their greater brand visibility leads to disproportionately negative investor reactions to bad news.

Generalizability is not a new problem, but it often gets little attention by authors in the face of novel data. After all, 1.6 billion transactions is a lot! Additionally, while authors should acknowledge them, generalizability problems in a single paper may not be critical if *readers* are careful when making inferences, especially if follow-on work can validate the results in other settings. However, generalizability problems can compound if research within a given area repeatedly uses big data sources with similar sample selection issues, giving a misleading impression of generalizability. Many sources of big data currently used by researchers in accounting and finance are somewhat or completely focused on retail firms, namely credit card transactions, brand awareness, parking lot images, and foot traffic data. Hopefully this disproportionate representation of retail firms will eventually decrease as access to new sources of big data expands. In an ideal world, researchers could avoid generalizability problems by using big data sources without sample selection issues. Practically speaking, researchers should identify ways in which their sample selection might limit generalizability and design tests to mitigate these concerns.

2.3 Noise and cleaning

Unfortunately, quantity does not always equal quality when it comes to data, and big data is no exception. Many sources of alternative data suffer from high levels of noise or require a large amount of processing before any useable information can be extracted. For example, textual, image, and audiovisual data are all unstructured and require careful analysis that often relies heavily on subjective research design or processing choices (Loughran and McDonald 2016). Unlike when they use CRSP or Compustat, researchers using big data sources are often the first to use the data and entirely responsible to process and vet it. The resources required to do so can be significant. Although the authors of Blankespoor et al. (2022) are modest in describing the difficulty of their data cleaning process, analyzing 1.6 billion transactions was undoubtedly a headache. Unlike corporate users of big data, research teams at academic institutions often consist of only three or four individuals, usually with access to either high-powered personal computers (often inadequate for large amounts of data) or university-hosted computing servers subject to a variety of restrictions and bottlenecks. In other words, the costs to clean and process big data are far from trivial. Further, processing and cleaning choices can materially impact the results of statistical tests (Denny and Spirling 2018) and yet are often undisclosed, study-specific, and difficult for others to replicate.

A considerable risk that researchers bear when they decide to incur the costs of processing big data is that the resulting data may *still* suffer from noise and measurement error. Unfortunately, it is frequently unclear what the quality of the data will be before these costs are incurred. A bigger problem is that the data, even if perfectly measured, might still only be loosely tied to the construct of interest. For example, consider a hypothetical scenario in which a researcher uses Google Trends data to gauge sentiment about a topic that will be voted on at a shareholder meeting in order to predict how management might preemptively respond. Unfortunately for this hypothetical researcher, Google Trends is based on searches by all users, not just those who own stock in a given company, and a large proportion of votes are actually cast by proxy voters. As a result, Google Trends data might not be a good proxy for the attitudes of those who will ultimately cast votes. Although Google Trends can be an excellent barometer for attitudes and interest in many settings, it is not necessarily suited to every research question. Ultimately, no matter how carefully researchers clean and vet their data, big data is not a cure-all for construct validity problems.

2.4 Interesting data, boring questions

Unfortunately, it is difficult to reverse engineer an interesting (or impactful) research question. Frequently researchers are so enthusiastic about a new source of big data that they jump into cleaning and validating the data before they have clearly articulated a research question. Sometimes researchers may even have an idea for a specific construct of interest that they think they can measure with the available data. However, if it was difficult to generate a research question *before* processing the data, it is often just as difficult generating a research question *after* processing the data. As a result, some researchers end up writing papers that focus entirely on validating and describing a given data source in the absence of economic intuition. A related tendency is to write papers that develop methodological approaches without discussing applications or testing new research questions. If the researcher cannot generate any ideas on how a particular source of big data or a particular innovative methodology could be used to test interesting economic questions, then it is unlikely that follow-on researchers will be able to do so either.

The temptation to jump immediately to data collection is especially high for PhD students who hope to develop their research question as they go and are intent on obtaining data with high barriers to entry. However, this is the population in which the costs of wasted effort are potentially the highest. Researchers should carefully consider their research question *before* collecting data. If possible, they should generate the research question first, and then search for the best data. It is much easier to scale back a grandiose research question to meet the limits of the data than to scale up an uninteresting research question to match an interesting dataset.

2.5 Other issues

The analysis of big data relies more heavily on black box methodologies such as machine learning than does traditional data. These models can provide statistical power at the expense of economic interpretability (Loughran and McDonald 2016;

Rudin 2019). Further, the use of some alternative and big data sources can lead to ethical concerns not applicable to traditional data. For example, the use of highly detailed data at the individual level, such as geolocation, web browsing, app usage, and social media data, raises concerns about individual privacy. This is compounded by the fact that such data is most available for those that are the least literate in data privacy. One could argue that tracking such behavior at an aggregate level protects individual identities. But is there ever a point at which certain behaviors are invasive to study, even in aggregate? And what stops other researchers from targeting and revealing individuals? Another ethical consideration centers around the use of leaked or hacked data. In some cases, we may feel an ethical duty to study corporate behavior that firms have purposefully tried to conceal. On the other hand, using this data may serve to further the unclear agenda of those who leaked or hacked the data in the first place.

3 Conclusions

In many ways, big data is not inherently different from other types of data. However, researchers, especially PhD students, can forget this in the excitement of learning about a new dataset. This article highlights four practical issues to consider when conducting economics-based research using big data. First, a particular source of big data may not be conceptually different from traditional data. As a result, studies that simply replicate prior results using big data may lack contribution, especially if the new data suffers from the same issues as prior data (e.g., endogeneity). Second, big data sources may only be available for a limited number of entities, especially when aggregated to the unit of interest, leading to limited statistical power and generalizability. Third, high levels of noise and the large volume of data can make big data costly to process; however, an arduous data cleaning process itself does not ensure that empirical proxies are tied to the constructs of interest. Last, interesting research questions are difficult to reverse engineer after the fact, and researchers who invest heavily in big data before generating a research question may end up with a paper that focuses on validating data in the absence of economic intuition. Researchers who keep these four issues in mind can potentially save themselves considerable resources (not to mention heartache!) by avoiding low-impact, high-cost projects and by instead focusing on research questions with the greatest potential for contribution.

Acknowledgements This article is an expansion of discussant remarks delivered at the 2021 Review of Accounting Studies Conference. Many thanks to Patricia Dechow, Earl K. Stice, Forester Wong, and conference participants.

References

Blankespoor, E., B. E. Hendricks, J. D. Piotroski, and C. Synn. 2022. Real-time revenue and firm disclosure. Review of Accounting Studies 27 (3).

Denny, M. J., and A. Spirling. 2018. Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis* 26 (2): 168–189.

Cong, L. W., B. Li, and Q. T. Zhang. 2021. Alternative data in fintech and business intelligence. In *The Palgrave Handbook of FinTech and Blockchain*, eds. M. Pompella, and R. Matousek, 217–242. Cham: Palgrave Macmillan.

Froot, K., N. Kang, G. Ozik, and R. Sadka. 2017. What do measures of real-time corporate sales say about earnings surprises and post-announcement returns? *Journal of Financial Economics* 125 (1): 143–162.

Ham, C., M. Lang, N. Seybert, and S. Wang. 2017. CFO narcissism and financial reporting quality. *Journal of Accounting Research* 55 (5): 1089–1135.

Hobson, J. L., W. J. Mayew, and M. Venkatachalam. 2012. Analyzing speech to detect financial misreporting. *Journal of Accounting Research* 50 (2): 349–392.

Kang, J. K., L. Stice-Lawrence, and Y. T. F. Wong. 2021. The firm next door: Using satellite images to study local information advantage. *Journal of Accounting Research* 59 (2): 713–750.

Lee, L. F., A. P. Hutton, and S. Shu. 2015. The role of social media in the capital market: Evidence from consumer product recalls. *Journal of Accounting Research* 53 (2): 367–404.

Loughran, T., and B. McDonald. 2016. Textual analysis in accounting and finance: A survey. *Journal of Accounting Research* 54 (4): 1187–1230.

Oracle Corporation. 2022. *What is Big Data?* Oracle.com. https://www.oracle.com/big-data/what-is-big-data/. Accessed June 23, 2022.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1 (5): 206–215.

Teoh, S. H. 2018. The promise and challenges of new datasets for accounting research. *Accounting Organizations and Society* 68: 109–117.

Zhu, C. 2019. Big data as a governance mechanism. *The Review of Financial Studies* 32 (5): 2021–2061.