Big Science and Little Science in Open and Distance Digital Education

Heather Kanuka

Contents

Abstract

This chapter provides a discussion of big science and little science. An overview of the definitions and uses of each is provided, as well as data collection and analysis practices, inclusive of a range of digital data analysis tools for research projects in open, distance, and digital education. A discussion is also provided on the promises, opportunities, controversies, and complications of big data and little data, as well as the possibilities of working with both forms of data collection. Insights based on the literature are highlighted, providing suggestions for practice when working with big data and/or little data. The chapter concludes with questions and suggestions for further research and implications for open, distance, and digital education that arise from the literature.

H. Kanuka (✉)
University of Alberta, Edmonton, AB, Canada
e-mail: hakanuka@ualberta.ca

O. Zawacki-Richter, I. Jung (eds.), *Handbook of Open, Distance and Digital Education*,
https://doi.org/10.1007/978-981-19-0351-9_13-1

Keywords

Big data · Little data · Digital tools · Open and distance digital education

Introduction

In God we trust. All others must bring data. (W. Edwards Deming, Hanson, 2019)

Student's digital activities generate an enormous amount of data which has led to the pursuit of how to analyze these data to determine if and/or how the information can be used to enhance learning environments (Sin & Muthu, 2015). Often referred to as big science, the central aim is to draw meaningful information from a large volume of data by eliminating the noisy data which can then be used to make better, faster, and smarter decisions (Dahdouh, Dakkak, Oughdir, & Messaoudi, 2018). This information, in turn, can be used to enhance ODDE systems. For example, big data can provide information ranging from enrolment and attrition to course materials and student activities. Indeed, according to Atasoy, Bozna, Sönmez, Akkurt, Büyükköse, and Fırat (2020), big science "can solve everyday problems . . . [in] education, enable personalized learning for each learner, offer a new type of evaluation and assessment and allow continuous feedback and feedforwards" (p. 145). What big science cannot do, however, is determine if the data have any kind of impact on learner outcomes (O'Brian, 2017). As Prinsloo, Archer, Barnes, Chetty, and van Zyl (2015) note, ". . . it is clear that in order for big(ger) data to be better data, a number of issues need to be addressed" (p. 284). The issues Prinsloo et al. note revolve around the problem that big data analysis can provide patterns *about* what students do online, but the data cannot *interpret* the patterns and/or determine how the data links to learning theory (see also Maldonado-Mahauad, Pérez-Sanagustín, Kizilcec, Morale, & Munos-Gama, 2018). To address these issues, data triangulation has been suggested, which could include qualitative research methodologies – or little science, which can provide explanatory power using thick, rich data. But like big data, little data also have limitations (e.g., inability to generalize, researcher privileging, sample bias, etc.)

ODDE research that includes the breadth and depth that both big and little science offers provides a more complete set of findings than either can provide singly. ODDE researchers, for example, can use the analysis of big data patterns to gain information on what is occurring, which can then be effectively used with little data (qualitative) methods to provide insights on why. Or, alternatively, ODDE research can use the insights arising from qualitative methods to determine if the data are generalizable to a wider population.

Big Science, Big Data

Apparently, one of the hottest things anyone can become these days is a data scientist (Fruhlinger, 2019). A data scientist collects and analyses big datasets of structured and unstructured data. Most often, a data scientist will have knowledge of computer science, statistics, and mathematics. They use their knowledge and skills to find patterns, identify trends, and manage data – or, quite simply, make sense of an extremely large amount of messy data that do not easily fit into existing database software. Mills (2018) notes that big data has "captured the imagination of researchers worldwide, with a proliferation of digital media rendering extremely large datasets more rapidly searchable, analysable and shareable" (p. 591). Josh Wills (a senior director of data science at Cloudera) describes himself as ". . . a data janitor. That's the sexiest job of the twenty-first century" (Harnham Blog and News, n.d.). Dan Ariely also notes the allure of big data in a tweet (Ariely, 2013): "Big Data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it." An overview of the research literature on big data indicates that, as Ariely aptly tweeted, not only does no one really know how to do it; there is little consensus on what it is, as well as how to define it.

The following section provides a synopsis of the literature on how big data are described and/or defined by researchers and practitioners, as well as their uses. As this section illustrates, one cannot assume there is a shared understanding of definitions for big data, uses, and/or how the data are collected and analyzed. When the ODDE researcher is choosing and using big data, also referred to as big science, it is essential at the onset to decide on a definition and intended use, with a clear and lucid description of how the data will be collected and analyzed.

Big Data Defined

Put simply, big data are human artifacts generated and shared through technological environments where (almost) anything can be captured digitally and collected as data, which Mayer-Schönberger and Cukier (2013) referred to as "datafication." The phrases "big data" and "big science" have been around since the 1990s, though it would appear no one is certain exactly when it emerged and/or who coined the phrases (Big Data Fundamentals, 2019). While still a relatively new construct in the research area, big data has already become a somewhat prosaic, all-encompassing phrase used to describe a variety of different purposes about enormously gigantic data sources. Big data has been used to describe everything from the collection and aggregation of large amounts of data to vast amounts of digital analysis aimed to identify patterns in human behavior for researchers and industries alike (Favaretto, Clercq, & Elger, 2020). This is in addition to uses aimed to improve science and research; optimize performance; improve health care; enhance machine and device performance, security, and law enforcement; and, most recently, provide essential information on the pandemic which has guided public health decisions. The use of big data has shown to have the potential to identify key information for improved decision-making processes, and, as such, it is easy to understand why it has also attracted the attention of ODDE researchers.

The most frequently cited definition of big data is by the National Science Foundation (NSF, 2012). The National Science Foundation states:

> The phrase "big data" in this solicitation does not refer just to the volume of data, but also to its variety and velocity. Big data includes large, diverse, complex, longitudinal, and/or distributed data sets generated from instruments, sensors, Internet transactions, email, video, click streams, and/or all other digital sources.

The problem with this definition is that, as Favaretto et al. (2020) note, it is "loaded with conceptual vagueness" (para. 1). Essentially, big data is comprised of ". . . any collection of data or datasets so complex or large that traditional data management approaches become unsuitable" (IPSOS Encyclopedia, 2016). A study conducted by Favaretto et al. investigated researchers' understanding of the big data phenomenon over the last decade. The findings of this study revealed that many of their participants were uncertain how to define big data, though there was some agreement on using the traditional "Vs" definition – though, again, there was no agreement on the number of Vs. Depending on who one reads, the Vs definition of big data includes two to seven of the following: *volume (big,* extremely big, data – or very large datasets consisting of terabytes, petabytes, zettabytes of data – or larger), variety (multiple datasets that include structured and unstructured data – such as pictures, voice recordings, tweets, etc.), veracity (trustworthiness which includes the increasingly complex data structure, anonymities, imprecision, or inconsistency in large datasets), velocity (high volume of incoming data with nonhomogeneous structure), value (extracting data that lead to the discovery of a critical causal effect that results in an important new discovery), variability (the meaning of the data are constantly changing), and visualization (presentation of data that is readable) (Sivarajah, Kamal, Irani, & Weerakkody, 2017).

In the study by Favaretto et al. (2020), most of the participants (who were big data researchers) preferred a practical definition that linked to practice such as the processes of data collection and processing. Also noted in the findings is that the participants, on the whole, had an uneasiness with respect to the use of the term big data, recognizing that this field is a "shifting and evolving cultural phenomenon. Moreover, the currently enacted use of the term as a hyped-up buzzword might further aggravate the conceptual vagueness of big data" (para. 4). As Favaretto et al. also emphasize, "big data is a term that has invaded our daily world. From commercial applications to research in multiple fields, big data holds the promise of solving some of the world's most challenging problems" (Introduction, Para. 2). Given the shifting ways big data are collected, analyzed, and used, it would seem a definition of big data needs to be linked to its use. At this point in time, researchers are using big data to "analyse and group data, create correlations, look for clusters and essentially gain insights into data, that we cannot get from standard reporting of the tools and systems creating and storing the data" (Sivarajah et al., 2017, p. 266). Given how big data are currently being used in ODDE research, it can be defined as the analysis of an extremely large group of data that generates correlations and/or clusters, providing insights otherwise unobtainable from standard collection and analysis. While there is no agreed-upon threshold for big data (upper and lower limits depend on the kind of data collected and time span of the data collected), “standard collection” of data can be understood as data created and collected that can be analyzed through traditional data management approaches (e.g., existing database management software such as Oracle, FoxPro, FileMaker Pro, Microsoft Access, etc.). Big data, then, are large amounts of data (e.g., terabytes, petabytes exabytes, and larger) created and collected overtime, and the data are analyzed using big data analytic software (e.g., Domo, Grow, Toucan Toco Data, Python, R, etc.).

How Big Is Big Data?

Quite simply, big data is very, very big. Just how big” very, very big” is is difficult to determine because the ways the data are collected, analyzed, and used are constantly changing. A report provided by Dobre and Xhafa (2014) determined that the world produces about 2.5 quintillion bytes of data. How big is quintillion byte of data? One exabyte equals one quintillion bytes, so one exabyte equals one billion. A Google search on the Internet indicates (depending on the site visited) that in 2020, we (people who use digital technologies) created about 1.7 megabytes of data every second and by the end of 2020, approximately 44 zettabytes comprised the entire digital universe. The eighth edition of Domo’s “Data Never Sleeps” report estimated that we created 2.5 quintillion data bytes, daily, in 2020 (fyi: there are 18 zeros in a quintillion). Raconteur (2021) estimates there will be 483 exabytes of data generated each day by 2025. According to the World Economic Forum (2019), there are 40 times more bytes in 483 exabytes than there are observable stars in the universe.

There appears to be no end in sight on the ways big data continues to challenge our imagination with respect to limits and by association the ways in which an ODDE researcher can use big data to gain relational information about ODDE. For example, Wen, Zhang, and Shu (2019) assert that through the use of a chaos optimization and cognitive learning model they developed, it is possible to gather information about student attributes (e.g., motivation, task demands, efficacy, interaction, time on tasks, learning styles, etc.) to potentially improve the ODDE learning experience. As Wen et al. illustrate (see also Huda, Maseleno, Atmotiyoso, Siregar, Ahmad et al., 2018), it is possible to optimize the chaos, of large, incomplete, noisy, fuzzy, big data to uncover potentially useful information which can be used to not only enhance the learning experience but also assist in market strategies, risk reduction, administrative tasks (e.g., registrations), resource and infrastructure management, and policy decisions. Another example is the perennial issue of student attrition in ODDE which, as O’Brian (2017) aptly notes, is often only identified after an exam is missed or a student is no longer logging into the learning system. It is possible that big data analysis can enable early identification of students who are at risk (dropout, flunkout, time-out), providing opportunities for interventions. A study by Zhang, Gao, and Zhang (2021), for example, used clickstream data to investigate student attrition in ODDE. Their findings revealed that introductory learning resources, scaffolding, and embedded assessment can mitigate attrition. Kyritsi, Zorkadis, Stavropoulos, and Verykios (2019) also found that the use of discussion fora is correlated to higher achievements on course assignments, quizzes, and exams; this, in turn, could reduce attrition due to failure.

As these examples illustrate, the greatest contribution of big data is the ability to gather predictive data which can assist in strategic decision-making, as well as guide students through their programs and course selections. In turn, this could also improve student success and satisfaction, increase the quality of teaching resources, and lower costs (Dahdouh et al., 2018; Rienties, Cross, Marsh & Ullmann, 2017).

To assist in understanding how to use the enormous amounts of data to enhance ODDE, data scientists use data visualization tools. Data visualization tools provide a representation of data in a graph, chart, or other visual formats that illustrates relationships of the data through the use of images. The visual relationships, then, allow us to identify and interpret trends and patterns, which can provide predictive analysis. For example, as Atasoy et al. (2020) note, it is obvious that a better understanding of the student (e.g., demographics, grades, attendance, log data, interaction, time spent in online, and responses to interventions and learning designs) would benefit students and “thus the educational institution’s retention and success rate” (p. 147). According to Atasoy et al., it is possible to use this information to:

> . . . predict learners’ performance, identify undesirable learning behaviors and emotional states, ascertain and monitor learners at risk and provide appropriate help for learners. It can also stipulate learners with learning features that will make their learning experience more personal and engaging, encourage reflection and development and stronger descriptions of patterns . . . there will be personalized theories and philosophies that fit each learner and application of a student-centric, inquiry-based model of analytics will put the tools and premises of analytics into the hands of learners and empower them as metacognitive agents of their own learning . . . Also, the collection of large amounts of data, big data, can help educators and system makers to identify patterns which will enable tailored education for each individual. By this way, pedagogy and andragogy can break their chains; become free from “one-size-fits-all” principles. (pp. 159–160)

Data Deluge

The data deluge phenomenon refers to the tsunami of complex, unstructured, and structured data available alongside a perception that we can simply, and easily, mine whatever data we are interested in, analyze it, and voila: we have novel insights and significant findings from an unprecedented scale of large data available**.** This is a misguided assumption.

Mayer-Schönberger and Cukier (2013) describe a transition that is occurring in research practices from causal inference approaches to analyzing data to data analysis practices based on the advantages of conducting correlational analysis with extremely large datasets. There is no question, as Gejingting, Ruiqiong, Wei, Libao, and Zhenjun (2019) observe, big data are capable of providing powerful functions for correlation analysis. The strength in correlational analysis of big data is the probability meanings; hence, if the correlation coefficient is large, it can establish probability with a high degree of accuracy. However, there are well-known limitations with correlational research, including the well-known limitation that not all correlations are meaningful (e.g., just because two variables are correlated does not mean a causation relationship exists between them). It is true that big data will return results on (almost) anything the researcher asks. Unfortunately, if researchers ask the wrong question or are just "going fishing" for significant relationships, big data will return significant results – regardless of whether causation exists or not. It is also well-known that big data are prone to data breaches, there are a lot of data behind firewalls that are not available for data analysis resulting in skewed and/or an incomplete analysis of the data, and the tools used to collect big data are inexact. As Fan, Han, and Liu highlight (2014): "…the massive sample size and high dimensionality of big data introduce unique computational and statistical challenges, including scalability and storage bottleneck, noise accumulation, spurious correlation, incidental endogeneity, and measurement errors" (para. 1), leading to mistaken statistical inferences and incorrect scientific conclusions.

For reasons noted above, several ODDE researchers have cautioned about the possible perils of working with big data. Unsupported assertions with unbridled enthusiasm about big data have been challenged and continue to experience increased criticism. The following is such a quote by an enthusiastic researcher who declared that big data will end the need for theory and make scientific methods obsolete:

> … massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behaviour, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves. (Anderson, 2008)

In response to such assertions, Crawford (n.d.) published myths about big data arguing, among other things, it is a mistaken assumption that when the numbers are large enough, the data speak for themselves. Other criticisms include discrimination, asynchronous power between social groups, and invasion of privacy (e.g., Leurs & Shepherd, 2017; Mills, 2018). Boyd and Crawford (2012) have asked critical questions about the analysis and use of big data, including the following: Are big data changing our definition of knowledge? Are big data misleading us with respect to objectivity and accuracy? Are big data better data? Are big data meaningful without context? While the data are available for collection, is it ethical? Is the use of big data creating a new digital divide? On a darker side, Leurs and Shepherd (2017) question who, exactly, benefits from the correlative analysis of big data? And who suffers? They describe the issues with "runaway data that asymmetrically order our social … institutions through hidden algorithmic practices that tend to further entrench inequality by seeking to predict risk" (p. 211).

To be clear, big data are remarkable at ubiquitously collecting a vast array of human behaviors available in a digital format. However, meaningful research is more than just a matter of getting a ticket dump and using data analytic and visualization tools. It is essential to know and understand the context, who is contributing to the data, who is not, how it is being used, and what processes it is supporting. Big data, in and of itself, is meaningless.

Little Science, Little Data

As illustrated above by Anderson (2008), there are practitioners and researchers who believe that big data will render the data arising from small-scale research (most data collected in qualitative studies would be considered small-scale research) inadequate and perhaps even become an obsolescent form of data collection. Of course, these assertions have been challenged, most often countering with the argument that complex research questions about human behavior and society require identification of patterns within contextualized data and these data are, typically, located in the minds, artifacts, and/or documents of individuals and organizations – not always available in a digital format (Mills, 2018). Access to these data relies on the willingness of individuals and organizations to share this information (e.g., opinions, perspectives, documents, etc.). Furthermore, researchers are (typically) awarded funding, and published in competitive journals, when new insights from original data are produced, providing solutions for current issues and problems (Borgman, 2015). As Mills notes:

> … big data has potential for optimizing and advancing the efficiency of research and scholarship, more than ever before there is the need for reason, theorization, problem-solving, originality, and social justice in determining what questions can be served by the data, and whose interests they serve. A ready supply of statistics and the vast scale of data in the digital world is not particularly useful for answering the kinds of research questions that people in the social sciences are asking. (p. 595)

By way of an example, a problem in Canada is the provision of access to ODDE opportunities in rural and remote communities who continue to have limited and/or unreliable Internet access. Big data cannot provide insights to issues where these kinds of digital black spots exist. Hence, there are contexts and environments that big data cannot capture; the need for small science will always exist.

Little Data Defined

Unlike big data, there is little controversy with respect to understanding little data. Little data, or qualitative research, is (mostly) an agreed-upon construct. While all research involves collecting, analyzing, interpreting, and writing the results of a study, qualitative research involves an inquiry process of understanding a social or human problem, based on building a complex, holistic picture, formed with words, reporting detailed views of informants and conducted in a natural setting (Creswell, 1994). Denzin and Lincoln (1994) elaborate further, describing qualitative research as a multi-method, interpretive, and naturalistic research approach. In its simplest sense, then, qualitative research seeks to understand individuals' social reality.

How Small Is Small Data?

Because qualitative researchers collect words, documents, artifacts, and/or information as their data, quantifying the data and determining statistical significance are not a concern (Onwuegbuzie & Leach, 2007). Rather, qualitative researchers are concerned about gathering enough data to achieve "conceptual power" (Constantinou, Georgiou, & Perdikogianni, 2017) which, in turn, provides detailed descriptions to ensure the findings are transferable, rather than generalizable. As such, the quality of data is more important than the quantity of data collected. Where the waters get muddy in qualitative research is just how big should the data be? And how small is too small? There is certainly no shortage of scholarly literature on this front or pedestrian opinions available on the Internet.

Qualitative data can include historical documents, observations, visual data, books, and texts – to list a few. However, by far, the most frequent data collected by qualitative researchers are the words provided by purposively selected individuals and/or group(s) of people. The issue revolving around how many individuals or groups of individuals are needed to achieve rigor, credibility, and trustworthiness is where there is less consensus. Depending on whom one reads, sample sizes involving individuals and/or groups can be as small as one person (Baker & Edwards, 2012) in, for example, biographical research. Alternatively, recommendations by Becker et al. (2002) argue that "In the case of 2-4-h interviews . . . [the] rule of thumb is that fewer than 60 interviews cannot support convincing conclusions and more than 150 produce too much material to analyse effectively and expeditiously" (p. 23), while others conclude there are no rules. Baker and Edwards conducted interviews with experts in the field, asking them "how many qualitative interviews is enough?" With few exceptions, the answers by the experts selected for this study involved explaining that it depends, concluding as one participant mused:

> But in general the old rule seems to hold that you keep asking as long as you are getting different answers, and that is a reminder that with our little samples we can't establish frequencies but we should be able to find the range of responses . . . the best answer is to report fully how it was resolved. (Bakers & Edwards, pp. 3–4)

Data Saturation

Data saturation is a term which is used for what the above participant refers to as "the old rule." According to Glaser and Strauss (1967), data are saturated when the topics or themes drawn from the researcher(s)' dataset are repeated and the data ceases to provide new information or themes relating to the research problem. There is general agreement in the research community on how data saturation is defined, as well as consensus that data saturation contributes to ensuring the data collection and analysis are robust and valid. What is rarely in the literature on data saturation, as well as in published research studies, is a description of how data saturation is achieved. It should also be noted that what is actually saturated is not the data, per se, but the categories/topics and themes. As Constantinou et al. (2017) note, words cannot be saturated because the words used will be different across participants; what researchers actually analyze are the commonalities of the words and their meanings among participants. Technically, then, it is themes saturation, not data saturation. This noted, it has been argued that thematic saturation can be attained the same way: when there is cessation of new themes and categories. Yet, as Morse (2015) observes, how to achieve themes saturation is not always well understood by researchers, noting it is typically comprised of an abstract description, vacant of a detailed process.

Constantinou et al. (2017) reviewed the literature on different approaches for reaching saturation; they found a limited number of papers on how to conduct saturation. Depending on the processes described, this literature indicates that saturation can be achieved after 8–17 interviews (e.g., Bowen, 2008; Francis et al., 2010; Guest, Bunce, and Johnson, 2006). The processes presented for saturation were deemed in several ways as inadequate, with the biggest issue revolving around the question of interview order. Specifically, if the interviews were conducted in a different sequence (what they refer to as order-induced error), the researcher cannot be certain whether saturation would have been achieved within 8–17 interviews. Constantinou et al. offered an alternative method to achieve saturation which involves reordering the interviews multiple times. While Constantinou et al. provide a solution for the order-induced error, what continues to be unclear is as follows: How many participants are enough? It is reasonable to assume that the larger the sample size, the greater the number of topics and themes that will emerge. Hence, the issue about whether the sample size and selection are an accurate representation is not resolved with saturation, irrespective of the methods proposed. Based on the proposed methods for saturation and the literature critiquing these processes, it would appear saturation does not, de facto, contribute to the credibility or trustworthiness of qualitative research. In agreement with Wray, Markovic, and Manderson (2007), in reality, no data are ever truly saturated.

An alternative to thematic (or data) saturation is a statistical calculation for sample size proposed by Fugard and Potts (2015). While debates have been ongoing about the use of a statistical calculation for sample size in qualitative research, this may be a useful way for ODDE small data researchers to consider sample size within the context of the study before the data have been collected (a priori) rather than after the data have been collected (a posteriori). As noted previously, saturation is determined based on data analysis redundancy or cessation of new theoretical insights. As such, sample size is determined a posteriori. Fugard and Potts have proposed sample size can be determined a priori based on the contexts, similar to determining sample size in midsized quantitative research, such as survey methodology. Fugard and Potts proposed that sample sizes are comparable to those found in the literature, for example, ". . . to have 80% power to detect two instances of a theme with 10% prevalence, 29 participants are required. Increasing power, increasing the number of instances or decreasing prevalence increases the sample size needed" (p. 669).

Fugard and Potts (2015) acknowledge that the statistical calculation they have developed (and is open access; see Appendix) is not sufficient, in and of itself, for qualitative research. Rather, it is to be used in combination with other contextual considerations. As such, the statistical calculation proposed and developed by Fugard and Potts can be used as a practical tool for ODDE small data researchers to plan sample size involving thematic analysis, a priori. The tool is easy to use; the calculations are provided so qualitative researchers who are unfamiliar with statistical calculations should not have problems determining a sample size.

To be clear, Fugard and Potts (2015) do not propose that their tool will provide thematic saturation; rather, it is to be used as a useful estimate when planning for a qualitative research project (e.g., funding and ethics). Given the issues with determining saturation, using Fugard and Pott's statistical calculation is a viable tool worth considering in ODDE research. As Fugard and Potts note, it should be used with consideration of the context, and while not stated by Fugard and Potts, it could also be used alongside a saturation method, whereby sample size is estimated a priori and saturation is conducted a posteriori.

In Consideration of Big and Little Science for ODDE

Up to this point, big data and little data have been presented as separate forms of research. However, as discussed, both have possibilities and problems with respect to the kinds of insights obtained. Given the vast range of topics and practices in ODDE, ODDE researchers are well-positioned to generate meaningful research questions that can be effectively answered using both big and little datasets. ODDE researchers can use the analysis of big data patterns to gain information on what is occurring, which can be used in tandem with qualitative methods to gain better insights on why. Big data analytics, for example, can provide essential information about what ODDE students do online, where their activities are located, and what courses they are enrolling in, but it cannot explain why ODDE students leave their programs of study or why they select certain educational institutions, nor understand ODDE students' opinions and thoughts about their educational experiences. Qualitative research can gather data that provide insights into ODDE that shape how researchers can gain further understandings of ODDE. For example, if the ODDE researcher is interested in back channel text-based communication in asynchronous MOOC courses, discourse analysis (a method for studying written language in relation to its social context) would likely be the research method chosen. The analysis of discourse in a MOOC course would be difficult and time-consuming to conduct and would require substantive resources and a large research team. However, data visualization tools could be used to establish patterns and relationships, which could then be followed up with ethnographic observations and interviews to make the links with big data patterns and in-depth data from individual students or cases. Another example could be collecting big data from social network analysis (SNA) to build on distance learning theories. In particular, SNA could determine the relationships between the actors that facilitate the flow of information. Based on the relationships generated by SNA, ODDE researchers could follow up with ethnographic observations of the textual communication in distance education courses for richer understandings of relationships. As Mills (2018) notes, small datasets that use qualitative methods are useful for refining (and/or generating) theories that are used by researchers to explain the data. This is important in that what data are collected will always have "an element of arbitrariness, and data are not truth in themselves. They are simply sources of evidence that can be used to assert a certain view of reality" (Mills, 2018, p. 599). Mills also notes that the data researchers collect belongs to the subjects and are constructed in situ and must be collected accordingly.

Conclusion

This chapter provides a discussion on the possibilities and problems of little science and big science. An often-overlooked aspect by new and experienced ODDE researchers is to acknowledge we do not have shared understandings of what big data and little data are. An important aspect presented in this chapter is that when conducting research in ODDE with big and/or little data, ODDE researchers need to provide working definitions. This chapter also highlighted some of the limitations of the use of big and little datasets; however, it is certainly not an exhaustive description of all the problems and limitations. ODDE researchers who enter into research projects who are aware of the limitations are best prepared to provide either alternatives or additional research practices to compensate for the limitations, as well as to clearly and fully explain the limitations providing readers with a full understanding of the trustworthiness of the findings. All research is flawed.

Finally, the possibilities of gaining insights about ODDE through the building on and/or blending of big and little datasets are limited only by our imagination. Through the use of big and little datasets, we can gain further information and meaningful insights about persistent problems in ODDE, such as the following: Why is attrition so high in self-directed/self-regulated distance education? What distance education theories provide the greatest explanatory power for at-risk ODDE students? Are there specific characteristics of students at risk? And if so, are there strategies that can assist at-risk students? What are the characteristics of successful distance education students? What kinds of communication platforms provide the best support for ease of group communication for ODDE students? What kinds of online learning activities are effective at supporting critical, creative, and complex skills? Is a blended asynchronous and synchronous communication format more effective than a non-blended format for ODDE? Do student characteristics impact the kinds of communication effectiveness? In what ways do discipline impact communication effectiveness?

When big and little datasets are used to investigate ODDE, we have the ability to gain information about what our open and distance education students are doing and why they are doing what they are doing.

Cross-References

▶ Classic Theories of Distance Education
▶ Learning Analytics in Open, Distance, and Digital Education (ODDE)
▶ Managing Innovation in Teaching in ODDE
▶ Research Trends in Open, Distance, and Digital Education

Appendix

Big data open access tools

There are few options for ODDE researchers who wish to use open access software for big data collection and analysis. However, there are several tools that offer free use and/or free trial options. The three most commonly used tools providing these options are:

Domo (domo.com)
Grow (grow.com)
Toucan Toco Data (toucantoco.com)

Little data open source tools

Computing the sample size proposed by Fugard and Potts (2015) is provided in the appendix of their paper (pp. 483–484). The following is the example provided by Fugard and Potts:

To compute the sample size required for a power of 80% to find a theme prevalence of 0.1, and 2 instances, run:
sampSizeForQual(0.8, 0.1, 2)
This gives the answer 29.

Fugard and Potts also note that this code may be run even if R is not installed. Two sites that are open access for qualitative researchers wishing to determine sample size a priori are:

R-Fiddle (r-fiddle.org)
Ideone (ideone.com/oT4BRE)

Both are open access.

References

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

Ariely, D. [@danariely]. (2013, January 6). *Big data is like teenage sex: Everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it* [Tweet; thumbnail link to article]. Twitter. https://twitter.com/danariely/status/287952257926971392?lang=en

Atasoy, E., Bozna, H., Sönmez, A., Akkurt, A. A., Büyükköse, G. T., & Fırat, M. (2020). Active learning analytics in mobile:Visions from PhD students. *Asian Association of Open Universities Journal, 15*(2), 145–166. Available: https://www.emerald.com/insight/content/doi/10.1108/AAOUJ-11-2019-0055/full/html.

Baker, S. & Edwards, R. (2012). How many qualitative interviews is enough? Expert voices and early career reflections on sampling and cases in qualitative research. National Centre for Research Methods Review Paper. Available: http://eprints.ncrm.ac.uk/2273/4/how_many_interviews.pdf

Becker, H., Berger, P., Luckmann, T., Burawoy, M., Gans, H., Gerson, K. Gerson, K. Gerson, Glaser, B. Strauss, A., Horowitz, R., Horowitz, R., Inciardi, J., Horowitz, R. Pottieger, A., Lewis, O. Liebow, E., Mead, G.H., & Mills, C.W. (2002). Observation and interviewing: Options and choices in qualitative research in: Qualitative research in action. Sage Research Methods. https://doi.org/10.4135/9781849209656.n9

Big Data Fundamentals. (2019). Big data framework. Available: https://www.bigdataframework.org/short-history-of-big-data/

Borgman, C. (2015). *Big data, little data, no data. Scholarship in the networked world*. Cambridge, MA: MIT Press.

Bowen, G. A. (2008). Naturalistic inquiry and the saturation concept: A research note. *Qualitative Research, 8*(1), 137–152. https://doi.org/10.1177/1468794107085301.

boyd, d., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society, 15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878.

Constantinou, C. S., Georgiou, M., & Perdikogianni, M. (2017). A comparative method for themes saturation (CoMeTS)in qualitative interviews. *Qualitative Research, 17*(5), 571–588. https://doi.org/10.1177/1468794116686650.

Crawford, K. (n.d.). *Think again big data*. Available: https://foreignpolicy.com/2013/05/10/think-again-big-data/

Creswell, J. W. (1994). *Research design: Qualitative and quantitative approaches*. Thousand Oaks, CA: Sage.

Dahdouh, K., Dakkak, A., Oughdir, L., & Messaoudi, F. (2018). Big data for online learning systems. *Education and Information Technologies, 23*, 2783–2800. https://doi.org/10.1007/s10639-018-9741-3.

Denzin, N. K., & Lincoln, Y. S. (1994). *Handbook of qualitative research*. Thousand Oaks, CA: SAGE.

Dobre, C., & Xhafa, F. (2014). Intelligent services for big data science. *Future Generation Computer Systems, 37*, 267–281. https://doi.org/10.1016/j.future.2013.07.014.

Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National Science Review, 1*(2), 293–314. https://doi.org/10.1093/nsr/nwt032.

Favaretto, M., De Clercq, E., Schneble, C. O., & Elger, B. S. (2020). What is your definition of big data? Researchers' understanding of the phenomenon of the decade. *PLoS One, 15*(2), e0228987. https://doi.org/10.1371/journal.pone.0228987.

Francis, J. J., Johnston, M., Robertson, C., Glidewell, L., Entwistle, V., Eccles, M. P., & Grimshaw, J. M. (2010). What is an adequate sample size? Operationalising data saturation for theory-based interview studies. *Psychology and Health, 25*(10), 1229–1245. https://doi.org/10.1080/08870440903194015.

Fruhlinger, J. (2019). "Data Scientist" is the hottest profession of 2019 according to job-listing data. *The Business of Business*. Available: https://www.businessofbusiness.com/articles/massive-increase-in-demand-for-data-science-jobs-in-2019/

Fugard, A. J., & Potts, H. W. (2015). Supporting thinking on sample sizes for thematic analyses: A quantitative tool. *International Journal of Social Research Methodology, 18*(6), 669–684. https://doi.org/10.1080/13645579.2015.1005453.

Gejingting, X., Ruiqiong, J., Wei, W., Libao, J., & Zhenjun, Y. (2019). Correlation analysis and causal analysis in the era of big data. *Materials Science and Engineering, 563*. https://doi.org/10.1088/1757-899X/563/4/042032.

Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory: Strategies for qualitative research*. New York, NY: Aldine Publishing Company. Available: http://www.sxf.uevora.pt/wp-content/uploads/2013/03/Glaser_1967.pdf.

Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. *Field Methods, 18*(1), 59–82. https://doi.org/10.1177/1525822X05279903.

Hanson, H. L. (2019). Big data. *IBM Nordic Blog.* [online]. Available: https://www.ibm.com/blogs/nordic-msp/in-god-we-trust-all-others-must-bring-data/

Harnham Blog and News. (n.d.). Available: https://www.harnham.com/us/a-data-janitor-the-sexiest-job-of-the-21st-century

Huda, M., Maseleno, A ., Atmotiyoso, P., Siregar, M., Ahmad, R., Jasmi, K. A., . . . & Basiron, B. (2018). Big data emerging technology: Insights into innovative environment for online learning resources. *International Journal of Emerging Technologies in Learning, 13*(1), 23–36. https://doi.org/10.3991/ijet.v13i01.6990.

IPSO Encylopedia. (2016). Big Data. Available: https://www.ipsos.com/en/ipsos-encyclopedia-big-data

Kyritsi, K. H., Zorkadis, V., Stavropoulos, E. C., & Verykios, V. S. (2019). The pursuit of patterns in educational data mining as a threat to student privacy. *Journal of Interactive Media in Education, 1*(2), 1–10. https://doi.org/10.5334/jime.502.

Leurs, K., & Shepherd, T. (2017). Datafication and discrimination. In M. T. Schäfer & K. van Es (Eds.), *The datified society* (pp. 211–232). Amsterdam, Netherlands: Amsterdam University Press. https://doi.org/10.1515/9789048531011-018.

Maldonado-Mahauad, J., Pérez-Sanagustín, M., Kizilcec, R. F., Morales, N., & Munos-Gama, J. (2018). Mining theory-based patterns from Big data: Identifying self-regulated learning strategies in Massive Open Online Courses. *Journal of Computers in Human Behavior, 80*, 179–196. https://doi.org/10.1016/j.chb.2017.11.011.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. New York, NY: Houghton Mifflin Harcourt.

Mills, K. A. (2018). What are the threats and potentials of big data for qualitative research? *Qualitative Research, 18*(6), 591–603. https://doi.org/10.1177/1468794117743465.

Morse, J. (2015). Data were saturated. . .. *Qualitative Health Research, 25*(5), 587–588. https://doi.org/10.1177/1049732315576699.

National Science Foundation (NSF). (2012). Core techniques and technologies for advancing big data science & engineering (BIGDATA)**.** Available: https://www.nsf.gov/events/event_summ.jsp?cntn_id=124058&org=NSF

O'Brian, M. M. (2017). What should be done with collected online professional learning information? *The Quarterly Review of Distance Education, 17*(4), 39–48.

Onwuegbuzie, A., & Collins, K. M. T. (2007). A typology of mixed methods sampling designs in social science research. *The Qualitative Report, 12*(2), 281–316. Available: https://nsuworks.nova.edu/tqr/vol12/iss2/9/.

Prinsloo, P., Archer, E., Barnes, G., Chetty, Y., & van Zyl, D. (2015). Big(ger) data as better data in open distance learning. *The International Review of Research in Open and Distance Learning, 16*(1), 284–306. https://doi.org/10.19173/irrodl.v16i1.1948.

Ranconteur. (2021). A day in data. *Data Analytics*. Available: https://www.raconteur.net/topic/technology/data-analytics/

Rienties, B., Cross, S., Marsh, V., & Ullmann, T. (2017). Making sense of learner and learning big data: Reviewing five years of data wrangling at the Open University UK, open learning. *The Journal of Open, Distance and e-Learning, 32*(3), 279–293. https://doi.org/10.1080/02680513.2017.1348291.

Sin, K., & Muthu, L. (2015). Application of big data in education data mining and learning analytics – A literature review. *ICTACT Journal on Soft Computing, 05*(04), 1035–1049. https://doi.org/10.21917/IJSC.2015.0145.

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of big data challenges and analytical methods. *Journal of Business research, 70,* 263–286. [online]. Available: https://www.sciencedirect.com/science/article/pii/S014829631630488X

Wen, J., Zhang, W., & Shu, W. (2019). A cognitive learning model in distance education of higher education institutions based on chaos optimization in big data environment. *Journal of Supercomputing, 75*, 719–731. https://doi.org/10.1007/s11227-018-2256-2.

World Economic Forum. (2019). *How Much Data is Generated Each Day?* see https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/

Wray, N., Markovic, M., & Manderson, L. (2007). Researcher saturation: The impact of data triangulation and intensive-research practices on the researcher and qualitative research process. *Qualitative Health Research, 17*, 1392–1402. https://doi.org/10.1177/1049732307308308.

Zhang, J., Gao, M., & Zhang, J. (2021). The learning behaviours of dropouts in MOOCs: A collective attention network perspective. *Computers & Education, 167*. https://doi.org/10.1016/j.compedu.2021.104189.