Received: 21 March 2023 | Accepted: 22 March 2023

DOI: 10.1002/qre.3341

REVIEW ARTICLE

WILEY

Review: Is design data collection still relevant in the big data era? With extensions to machine learning

Laura Freeman

Virginia, Tech Research Center, Arlington, Texas, USA

Correspondence
Laura Freeman, Virginia Tech Research Center, Arlington, Virginia, USA.
Email: laura.freeman@vt.edu

KEYWORDS
big data, design of experiments, machine learning

Dr. Anderson-Cook and Lu's paper on the synergies between big data (BD) and designed data collection (DDC) is an outstanding synthesis of how multiple data collection strategies can benefit from one another throughout the data lifecycle.[1] The paper's 11 opportunities show how all phases of data collection (planning, during collection, and after collection) can benefit from thinking about the design of the data and what questions it will be used to answer. A fundamental truth that appears throughout the paper, but deserves calling out, is that the reason we collect data fundamentally is to inform decision-making, and that decision starts with a problem that needs to be solved.

I especially appreciate how the paper is structured to strategies that can be applied before BD, during actual BD, and after the collection of BD. This construct highlights the importance of a data lifecycle. It also makes the point that there is never a wrong time in the data lifecycle to re-evaluate if you have the right data to address the problem that needs to be solved.

Another point that resonates with me in this article, is that data science in the era of BD is really a team problem. Anderson-Cook and Lu emphasize expert knowledge of the process under study (what problem we are solving), statistical expertise in DDC, and engineering problem-solving skills are all needed. Based on my experience, the team is even bigger, expertise is needed in computer science, databases, computer and network architectures, data pipelines, data security, networks security, and modelers of all types (statistical, machine learning, deep learning, computational, etc.), and data visualization are all important members of the team. The exact makeup of the team ties back to the problem that needs to be solved and who it needs to be communicated to.

Finally, I would like to propose that the case for DDC is even stronger in the era of BD, and it goes beyond just synergies between DDC and BD. BD has significant cost, security, and other negative aspects associated with it to include the challenge of finding the signal in the noise of messy data, the cost of storing and disseminating BD, and answering the question of when can data be deleted. I am sure many other statisticians and data scientists will resonate with a frequent problem that I encounter in my work, where data has been collected or recovered, often at a significant cost, and the conjecture is made, "we have all of this data, there must be some key insights in there," or, "we have all of this data collected to answer problem Y, surely you can repurpose it to solve problem Z." I fondly refer to this scenario as the BD rabbit hole, and I have been down the BD rabbit hole numerous times in my career. So far without fail, the end result is typically an exhausted analyst and a disappointed decision-maker, because the data was never designed to answer the problem they are trying to address.

I concur with Anderson-Cook and Lu that neither BD nor DDC can replace the other, but I propose two additional lenses to view the DDC and BD discussion through that I believe focus the conversation to not the synergies between DDC and BD, but rather, how DDC should always be used in support of any BD collection effort. These two additional concepts that bear discussion include information quality and the cost of BD.

1 | INFORMATION QUALITY

Two separate, but related lines of work have emphasized the concept that all data is not equal in its ability to answer questions and that more attention should be placed on the quality of data collected and retained. Wilkinson et al. defined the four guiding principles for scientific data: findable, accessible, interoperable, reusable (FAIR) within the context of enabling reuse of scientific data in scholarly studies.[2] Others have expanded on the principles to include concepts such as security, trustworthiness, and data pedigree. The Department of Defense, where I focus many of my research efforts, has adopted a Data Strategy that provides a set of data quality goals, requiring all DoD data to be visible, accessible, understandable, linked, trustworthy, interoperable, and secure (VAULTIS).[3] All of these guiding principles are useful in thinking about the synergies between BD and DDC, in that DDC can be structured to increase the likelihood that the principles are achievable.

Another line of research has termed the concept of information quality. This line of research arguably dates back to around 1990, which a heavy emphasis emerging in the late 1990s.[4] Over time, with the increased collection of BD, a shift has occurred from the concept of data quality to information quality.[5]

Kenett and Shmueli define information quality as "the potential of a data set to achieve a specific goal by using a given empirical analysis method.[6]" Two different views on information quality that have a high correlation to FAIR and VAULTIS are shown in Table 1.

There are also numerous frameworks for assessing information quality. Lee et al. developed a validated scale approach including 65 information quality assessment items.[8] Kenett and Shmueli take a different approach and expand on the framework for information quality by mathematically linking the analysis goal to the to the data via a utility function.[6] A key takeaway from all of these dimensions of information quality is that the most important quality dimensions come back to the problem that the data is being collected to solve.

The concept of information quality is important to consider when thinking about the benefits of DDC in the BD era in that DDC efforts have the ability to purposely improve information quality and focus on quality aspects that are most important to the analysis goal. For example, in the network traffic streaming problem presented by Anderson-Cook and Lu, if the data is being collected to identify potential malicious traffic in the network, then prioritizing information extracted from aggregated packet data such as NetFlow will provide more timely data over full packet capture (e.g., pcap) due to the decreased size of the data. However, timeliness might come at the cost of losing important information about the network traffic contained in full packets that might be useful in identifying malicious behavior. In a DDC these tradeoffs can be assessed and purposefully integrated into the process of data collection.

Moreover, DDC provides the opportunity to leverage statistical approaches for increasing information quality. Kenett and Shmueli highlight randomization, blocking, replication, blinding, placebo treatments, and linking of the collection protocol to the dataset as strategies for increasing the information quality of DDCs.[6]

2 | COST OF BIG DATA

In addition to increasing the quality of data collected, designed data strategies provide an opportunity to ensure adequate data in terms of quantity and quality data are maintained overtime. Data collection strategies should always tie back to the problem, and we should seek to collect and store the minimally sufficient data to answer the problems at hand. Notably, BD has a significant cost associate with the collection, curation/processing, dissemination, use/analysis of, storage, and deposition of the data both in hardware and non-hardware domains. Local to me, one only needs to drive out the Dulles Toll Road in VA to see the number of data storage facilities and the power, space, and cost associated with storing data. A local article captured the data center footprint at 26 million square feet in May of 2021, with potential to more than double in the future.[9]

Tallon addresses the trade space between value, risk, and cost of BD from a corporate perspective and emphasizes that data governance practices are key to striking the right balance.[10] Figure 1 below shows the conceptual value of data

TABLE 1 Differing perspectives on the dimensions of information quality.

Information quality dimensions (Kenett[6])	Information quality dimensions (Miller[7])
1. Data resolution	1. Relevance
2. Data structure	2. Accuracy
3. Data integration	3. Timeliness
4. Temporal relevance	4. Completeness
5. Chronology of data and goal	5. Coherence
6. Generalizability	6. Format
7. Operationalization	7. Accessibility
8. Communication	8. Compatibility
	9. Security
	10. Validity

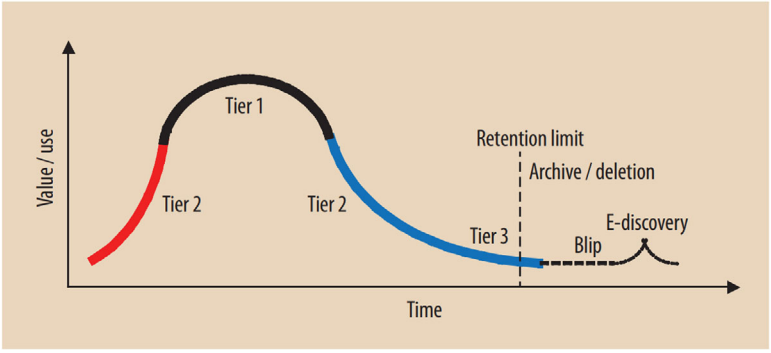

FIGURE 1 Information life-cycle curve from Tallon.[10]

over time as it pertains to informational decision-making. As more government organizations promote data strategies and leveraging data in government decision-making, we need strategies to reduce data holdings when their value has dwindled. This is especially important in the government space, as there are competing policies and guidance that require data and record retention, for example, federal records management policies. Policies need to be developed on culling data to the minimal set required to answer questions over time.

Importantly, statistical methods such as efficient sampling and statistical process control that Anderson-Cook and Lu highlight provide useful tools in reducing data stored over time as its value decreases. Additionally, techniques such as change point analysis, dimensionality reduction, and even the concept of a sufficient statistic provide context to what data is important to retain near-term and long-term.

3 | EXTENSIONS TO MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

Finally, I will conclude with a few exciting intersections of BD and DDC in the context of machine learning (ML) and artificial intelligence (AI). In my work, I focus on the application spaces of complex networked systems with an emphasis in cyber network security and the adoption of ML and AI. The current wave of ML/AI has been termed statistical learning enabled AI.[11] I like this terminology because it emphasizes that the learning process is statistical and based on observed data. Initial phases of the current wave of ML/AI were fueled by the availability of BD coupled with large-scale computing.[12] However, more recent publications have begun to focus on the need for DDC in improving current ML/AI methodologies. A few examples that illustrate this shift include:

- Systematic selection of training and testing datasets for ML/AI applications. Cody et al. shows that leveraging combinatorial coverage, a concept derived from experimental design and software testing, can be used to systematically split existing datasets into training and testing sets that improve coverage diversity and/or test the generalizability of ML/AI algorithms.[13]

- Adjusting training and testing data set selection to account for class imbalance. Lian et al. shows how designed experiments can improve the robustness of AI through the use of mixture designs and an exploratory framework to systematically investigate how the data quality issues such as imbalance among class labels, and distribution shift between training data and test data can be used to improve classification accuracy for convolutional neural network (CNN) and XGboost algorithms.[14]
- Improving transfer learning outcomes with small retraining datasets. Ahamed et al. shows that ML/AI outcomes (i.e., classification accuracy and F1) can be improved by selecting targeted data sets over randomly selected data sets of the same size. They leverage design coverage based in experimental design to achieve the realized increases in performance.[15]
- Balanced data reduces overfitting and enhanced generalizability. Yao et al. provides a material science example that is subject to imbalanced data to show that for artificial neural network (ANN) common ML challenges like overfitting and the generalizability of the algorithm are improved when compared to the original data sets.[16]

I expect these examples are just a small sample of an emerging body of work that will highlight the importance of design data collection in the ML/AI space. I will conclude by thanking and congratulating Anderson-Cook and Lu for providing an insightful article that will continue to advance the discussion on the value of data, BD, and the importance of DDC.

DATA AVAILABILITY STATEMENT

None.

ORCID

Laura Freeman https://orcid.org/0000-0001-7108-3921

REFERENCES

1. Insert reference for the Is Designed Data Collection Still Relevant in the Big Data Era?
2. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*. 2016;3(1):1-9.
3. Deputy Secretary of Defense. *DoD Data Strategy*. 2020. Accessed September 2, 2023. https://media.defense.gov/2020/Oct/08/2002514180/-1/-1/0/DOD-DATA-STRATEGY.PDF
4. Eppler MJ, Wittig D. Conceptualizing information quality: a review of information quality frameworks from the last ten years. *IQ*. 2000;20:83-96.
5. Strong DM, Lee YW, Wang RY. 10 potholes in the road to information quality. *Computer*. 1997;30(8):38-46.
6. Kenett RS, Shmueli G. On information quality. *J R Stat Soc*. 2014;177:3-38.
7. Miller H. The multiple dimensions of information quality. *Inf Syst Manag*. 1996;13(2):79-82.
8. Lee YW, Strong DM, Kahn BK, Wang RY. AIMQ: a methodology for information quality assessment. *Inf Manag*. 2002;40(2):133-146.
9. Data Center Dynamics News. Accessed September 2, 2023. https://www.datacenterdynamics.com/en/news/loudoun-county-is-home-to-26-million-sq-ft-of-data-centers-dulles-cloud-south-could-add-another-56-million-sq-ft-if-it-passes/
10. Tallon PP. Corporate governance of big data: perspectives on value, risk, and cost. *Computer*. 2013;46(6):32-38.
11. Launchbury J. A DARPA Perspective on Artificial Intelligence. Accessed September 2, 2023. https://www.darpa.mil/attachments/AIFull.pdf
12. Freeman L, Rahman A, Batarseh FA. Enabling artificial intelligence adoption through assurance. *Soc Sci*. 2021;10(9):322.
13. Cody T, Lanus E, Doyle DD, Freeman L, Systematic training and testing for machine learning using combinatorial interaction testing. 2022 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW), pp. 102-109, 2022; IEEE.
14. Lian J, Freeman L, Hong Y, Deng X. Robustness with respect to class imbalance in artificial intelligence classification algorithms. *J Qual Technol*. 2021;53(5):505-525.
15. Ahamed SF, Aggarwal P, Shetty S, Lanus E, Freeman LJ, ATTL: an automated targeted transfer learning with deep neural networks. 2021 IEEE Global Communications Conference (GLOBECOM), pp. 1-7, 2021 IEEE.
16. Yao Y, Sullivan IV T, Yan F, Gong J, Li L. Balancing data for generalizable machine learning to predict glass-forming ability of ternary alloys. *Scr Mater*. 2022;209:114366.

How to cite this article: Freeman L. Review: Is design data collection still relevant in the big data era? With extensions to machine learning. *Qual Reliab Eng Int*. 2023;39:1102–1106. https://doi.org/10.1002/qre.3341

AUTHOR BIOGRAPHY

Dr. Laura Freeman, is a Research Associate Professor of Statistics and dual hatted as the Deputy Director of the Virginia Tech National Security Institute and Assistant Dean for Research for the College of Science. Her research leverages experimental methods for conducting research that brings together cyber-physical systems, data science, artificial intelligence (AI), and machine learning (ML) to address critical challenges in national security. She develops new methods for test and evaluation focusing on emerging system technology. Dr. Freeman has a BS in Aerospace Engineering, a MS in Statistics and a PhD in Statistics, all from Virginia Tech. Her PhD research was on design and analysis of experiments for reliability data.