PERSPECTIVES ON CURRICULAR PRIORITIES FOR BIG DATA ANALYTICS: A QUALITATIVE STUDY OF EDUCATORS AND PRACTITIONERS

Rossi A. Hassad
Mercy College, New York, USA
Rhassad@mercy.edu

Big Data is a universal concept that has become a metaphor for innovation and discovery; however, deriving this benefit requires going beyond traditional tools and expertise. The aim of this qualitative study (N=38) was to explore perspectives on curricular priorities for Big Data analytics from educators and practitioners. Nine themes were derived: computational and programming literacy, integrated and multidisciplinary curriculum, data quality and methodological standards, collaboration and domain-specific expertise, statistical literacy, basic statistics (probability and inference), ethico-legal considerations, critical analysis and effective communication skills, and meaning and definition of Big Data. Additionally, reflective learning was identified as an overarching theme, which can serve as a unifying pedagogical approach. These results provide a framework that appears to have content validity for informing curricular reform and development.

INTRODUCTION

This study explored the perspectives of educators and practitioners regarding curricular priorities for Big Data analytics. Big Data is ubiquitous and fast becoming a metaphor for innovation and discovery (Capurro et al., 2021); however, the promise of a "Big Data revolution" is not without major controversy (Wong & Hinnant, 2022), especially in healthcare (Bamidis et al., 2019; Househ et al., 2017; Rodriguez et al., 2018). Despite variability in its conceptualization, Big Data is generally defined by three characteristics (Cappa et al., 2021): volume (exists in massive amounts), velocity (produced at super high speeds), and variety (available in various formats). Consistent with concerns about the quality and value of Big Data, particularly its high volatility and noisy nature, more recently, importance has been given to an additional dimension, data veracity or validity, which is fundamental for producing meaningful results.

Indeed, deriving actionable and novel insights from Big Data requires a multifaceted framework involving descriptive, diagnostic, predictive, and prescriptive analytics (de Mast et al., 2022) and necessitates expertise beyond traditional technological and statistical tools and skills. Specifically, this encompasses "computational, algorithmic, statistical and mathematical techniques" (Leonelli, 2020, para. 1), together with a high level of creative thinking (Dahlstedt, 2019). In this regard, Luan et al. (2020) note that "academics, educators, policy-makers, and professionals need to engage in productive collaboration" to identify and develop "necessary competencies and essential skills important for the 21st century" (p. 3) to deal with Big Data.

CURRICULAR AND PEDAGOGICAL ISSUES

The key assumption underlying the use of Big Data is that its high dimensionality and complexity provide a fertile source of rich intelligence, albeit extracting such value is fraught with challenges and controversies. Specifically, there is an epistemological dilemma and a "divided scientific community" (Egger & Yu, 2022, p. 17) regarding what constitutes Big Data, what meaning and knowledge can be derived from it, and what tools and expertise are required to leverage and harness the benefits from it. Contrary to the traditional statistics curriculum (Horton & Hardin, 2015), dispositional competencies such as an interdisciplinary mindset (Chang et al., 2020) and holistic thinking (de Oliveira & Nisbett, 2017), both of which give importance to the data context, are considered necessary for effective Big Data analytics. Reinforcing this, Jenkins (2013) notes that a flexible and creative mindset is required to optimize the benefits of Big Data and that this "requires man, not machine" (p. 4), suggesting an emphasis on cognitive expertise rather than an overreliance on technology. Technological tools such as artificial intelligence and machine learning lack contextual reasoning ability (Singh et al., 2013) and are generally not adaptive to temporal changes. This can lead to spurious or specious outcomes.

Another controversial issue is the predominant hypothetico-deductive reasoning approach to research and data analysis, which is viewed as rigid, limiting, and not amenable to generating novel and meaningful insights and hypotheses "born from the data" rather than from theory (Kelling et al., 2009,

In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the Gap: Empowering & Educating Today's Learners in Statistics. Proceedings of the 11th International Conference on Teaching Statistics (ICOTS11 2022), Rosario, Argentina.* International Association for Statistical Education. iase-web.org

p. 613). According to Kitchin (2014), meaningful Big Data analytics is more akin to inductive reasoning. As well, concerns have been leveled at the typical "correlational nature of Big Data" and the potential for "misinterpretation that can cause serious harm" (Pentland et al., 2013, p. 5). Consequently, coverage of methodological considerations and causal reasoning toward maximizing the benefits of Big Data has been identified as a primary curricular need (Raita et al., 2021). Also, the very nature of Big Data, particularly its enormous size (together with data linkage), variety, and volatility, creates major ethical and legal challenges that should be addressed in the curriculum.

OBJECTIVE, RATIONALE, AND THEORETICAL FRAMEWORK

The objective of this qualitative study was to ascertain and explore the perceptions of educators and practitioners regarding curricular priorities for Big Data analytics. The results are intended to serve as a needs assessment for the training of statisticians and data scientists. This interpretative study was informed by ontological and epistemological assumptions about Big Data and theoretical perspectives on teaching and learning, particularly constructivism (Hassad, 2011) and gestalt principles (Hassad, 2020), which reflect the philosophical orientation that the researcher brings to this study. The constructivist paradigm (Bodner, 1986) posits that learning (and knowledge production) is a meaning-making experience rather than the pursuit of an objective reality. This model guided the interpretation of the qualitative responses and conceptualization of the curriculum. Gestalt theory (Rock & Palmer, 1990; Wagemans et al., 2012) provides insight into perceptual experience, including how the visual system organizes information, and recognizes patterns, which is central to understanding data visualization, a core aspect of Big Data analytics.

METHODOLOGY

A qualitative cross-sectional study was conducted in Fall 2020 using an anonymous online questionnaire created with Google Forms and utilizing a targeted maximum variation sample of 38 statistics and data science educators and practitioners (including recognized experts). The study received ethics approval from Canterbury Christ Church University (UK). There were three open-ended questions focused on curricular priorities and approaches for Big Data analytics; however, only results from responses to the following question are reported in this paper: *Regarding curricular and instructional strategies for training and preparing statisticians and data analysts for Big Data analytics, what do you consider to be the immediate and long-term priorities?*

The survey invitation was circulated via direct email, and through the following listservs and electronic discussion groups.

1. ASA Connect: The American Statistical Association (ASA) discussion forum for professional statisticians (PStat).
2. ALLSTAT@JISCMAIL.AC.UK: A UK-based worldwide email broadcast system for the statistical community.
3. TEACHING-STATISTICS@JISCMAIL.AC.UK: A UK-based worldwide email broadcast system concerned with the initial learning and teaching of statistics.
4. EDSTAT-L@LISTS.PSU.EDU: An email forum devoted to discussion of topics related to the teaching and learning of statistics at the college level.

Manual thematic analysis was performed. Specifically, the "open, exploratory, flexible and iterative" (Braun, & Clarke, 2019, p.5) process of reflexive thematic analysis was used, which places emphasis on an active and reflective process and on the context of the data (Damayanthi, 2019). The standard six-step process for reflexive thematic analysis (Braun, & Clarke, 2019) was followed: familiarization with the data, generating initial codes, searching for sub-themes and themes, reviewing the themes, defining and naming the themes, and producing the report. Both inductive (data driven) and deductive (theoretical) techniques were used to produce semantic (descriptive) and latent (analytical) themes, respectively. Coding and interpretation of the responses, and naming of the themes were done by the author (a statistics educator) in consultation with a statistics education colleague. Codes are derived from the responses and are the building blocks of themes. Codes can be either descriptive (the literal or surface meaning of a response) or latent (inferred or underlying meaning), and reflect the researcher's interpretation, which is driven primarily by the theoretical underpinnings of the study. A theme represents a recurring pattern or concept across the dataset, whereas a sub-theme refers to a major element or dimension of a theme. The number of sub-themes for each theme is reported as the frequency of occurrence of that theme. Note that some responses reflected multiple themes, therefore, the total frequency of occurrence of themes exceeds the number of participants.

RESULTS AND DISCUSSION

The following nine themes (Table 1) were identified as curricular priorities for the teaching of Big Data analytics.

1. *Computational and programming literacy*: Engaging students in critical thinking and problem-solving using computers and coding.
2. *Integrated and multidisciplinary curriculum*: Facilitating students to develop the ability to make meaningful connections with attention to theory, application, and context by incorporating perspectives from various related disciplines.
3. *Data quality and methodological standards*: Providing students with guidelines relating to reliability, validity, and research design for assessing the value of data for evidence-based practice.
4. *Collaboration and domain-specific expertise*: Developing students' social skills to work collaboratively in teams and focusing on the critical importance of disciplinary or content knowledge for meaningful and effective data analytics.
5. *Statistical literacy*: Developing students' abilities to interpret, critically evaluate and communicate statistical information, and make data-based arguments.
6. *Basic statistics (probability and inference)*: Developing students' knowledge and skills of the fundamentals of descriptive and inferential statistics.
7. *Ethico-legal considerations*: Engaging students in considering privacy, confidentiality, and data security concerns.
8. *The meaning and definition of Big Data*: Providing students with a clear conceptualization of Big Data that will influence how it is analyzed, interpreted, and used.
9. *Critical analysis and effective communication skills*: Ensuring that students carefully consider and evaluate all aspects of an activity or report, including possible alternative explanations for outcomes, and clearly and accurately share information with a particular target group.

Computational and programming literacy was the most prevalent theme, with 18 occurrences. This reflects the status quo because Big Data analytics is generally characterized with reference to programming and machine learning, which are sub-fields of computer science (El-Alfy & Mohammed, 2020). In contrast, the *meaning and definition of Big Data* and *critical analysis and effective communication skills*, both had the lowest frequency (2), albeit these generally relate to the data context, which is fundamental to producing and presenting meaningful and actionable results. Almost similarly ranked are *ethico-legal considerations* with three occurrences. These considerations should be accorded greater importance, especially given the necessity for data linkage and hence the omnipresence of privacy, confidentiality, and data security concerns in Big Data analytics (Alharthi et al., 2017).

Notably, *integrated and multidisciplinary curriculum* and *data quality and methodological standards* are among the stronger themes, which seems commensurate with the long-standing debate about the benefits of Big Data and how to derive and optimize its value (Günther et al., 2017). In this regard, the importance given to *collaboration and domain-specific expertise* is reassuring, especially given the dynamic and uncertain nature of Big Data. Two related themes, *statistical literacy* and *basic statistics (probability and inference)* had a combined frequency of 11 but slightly favored statistical literacy, which is an established priority for the statistics education community. Of note is that both are among the lower-ranked themes. Particularly noteworthy is the exclusive reference to basic statistics (mostly probability and inference), which is disconcerting because the core underpinning of machine learning (for example) is a statistical framework of complex algorithms that require understanding both foundational and advanced statistical methods. Otherwise, Big Data analytics will be relegated to becoming a hollow and mechanical exercise, which defeats its intended purpose. Finally, *reflective learning,* "the process of internally examining and exploring an issue of concern, triggered by an experience, which creates and clarifies meaning in terms of self and which results in a changed conceptual perspective" (Boyd & Fales,1983, p.100), was identified as an overarching theme that encompasses the concepts of critical analysis, integrated curriculum, interdisciplinary understanding, meaning, literacy, and collaboration.

Table 1. Themes, frequency of occurrence, and supporting quotes ($N = 38$)

Themes	Frequency	Representative Supporting Quotes from the Verbatim Responses
Computational and programming literacy	18	"Include a heavy dose of computer science concepts to appreciate the complexities of handling big data (speed, parallel processing, efficient algorithms) ..."
Integrated and multidisciplinary curriculum	17	"Learning how to implement theory into practice through code is vitally important..." "I see many young statisticians that can prove asymptotics but can't process raw data or code in a way that could be maintained."
Data quality and methodological standards	11	"Building a better appreciation of why particular methods might be appropriate for particular circumstances and what role uncertainty, assumptions, and data quality considerations play." "Understanding the concepts of selection bias, confounding, causal inference, etc., is essential to drawing inferences from any observational data set."
Collaboration and domain-specific expertise	9	"[M]ake sure people understand the need to know enough about the subject matter the problem is about or at least have in depth conversation with a subject matter expert to plan the study and help analyze the results …"
Statistical literacy	6	"A careful restructuring of how we teach statistics in general. We need to move further from *p*-values and rethink what data analysis is. The data need to be more important than the model or beauty of the math. We also need better collaborations."
Basic statistics (probability and inference)	5	"A foundation in basic statistics." "Keep a solid background on probability and inference, include some sampling concepts …"
Ethico-legal considerations	3	"Students understand and behave in accordance with ethical practice standards (ASA and/or ACM) …" "The implications of red tape surrounding data collection in the healthcare industry, such as the HIPAA regulations."
The meaning and definition of Big Data	2	"First, clearly defining what is meant by Big Data." "For those with a strong programming background, more emphasis on understanding the nature and context of the data."
Critical analysis and effective communication skills	2	"Students understand and can communicate clearly the limitations of both their abilities and the techniques for answering specific questions in rigorous, reproducible, and actionable ways …"

Note that some responses reflected multiple themes, therefore, the total frequency of occurrence of themes exceeds the number of participants.

CONCLUSION AND IMPLICATIONS

These results provide insight into how educators and practitioners perceive and understand the challenges of Big Data. Together, these themes constitute a framework that appears to have content validity for informing curricular reform and development for Big Data analytics. Further research and discourse are needed to address effective pedagogy and assessment toward facilitating active and authentic learning. Relevant instructional strategies include the use of case studies, multidisciplinary collaborative projects, and student peer-review of programming code and other material toward fostering real-world problem-solving skills. In particular, the importance of the data context to meaningful analysis needs greater curricular emphasis. The sequencing of topics as well as the breadth and depth of coverage of each also needs to be addressed. Of note is that the overarching theme of reflective learning can serve as a unifying pedagogical approach and help to foster mindfulness and curiosity about Big Data as well as inductive reasoning, a mindset for innovation and discovery. Finally, the methodological limitations and potential researcher biases associated with qualitative research and thematic analysis must be considered.

ACKNOWLEDGEMENTS
Special thanks to the team from Canterbury Christ Church University (UK) and Dr. Gerald Iacullo.

REFERENCES
Alharthi, A., Krotov, V., & Bowman, M. (2017). Addressing barriers to big data. *Business Horizons*, *60*(3), 285–292. https://doi.org/10.1016/j.bushor.2017.01.002

Bamidis, P. D., Konstantinidis, S. T., Rodrigues, P. P., Antani, S., & Giordano, D. (2019). Guest editorial small things and big data: Controversies and challenges in digital healthcare. *IEEE Journal of Biomedical and Health Informatics*, *23*(6), 2208–2210. https://ieeexplore.ieee.org/document/8894165

Bodner, G. M. (1986). Constructivism: A theory of knowledge. *Journal of Chemical Education*, *63*(10), 873–878. https://doi.org/10.1021/ed063p873

Boyd, E. M., & Fales, A. W. (1983). Reflective learning: Key to learning from experience. *Journal of Humanistic Psychology, 23*(2), 99–117. https://doi.org/10.1177/0022167883232011

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, *11*(4), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

Cappa, F., Oriani, R., Peruffo, E., & McCarthy, I. (2021). Big data for creating and capturing value in the digitalized environment: Unpacking the effects of volume, variety, and veracity on firm performance. *Journal of Product Innovation Management*, *38*(1), 49–67. https://doi.org/10.1111/jpim.12545

Capurro, R., Fiorentino, R., Garzella, S. & Giudici, A. (2021). Big data analytics in innovation processes: Which forms of dynamic capabilities should be developed and how to embrace digitization? *European Journal of Innovation Management*, *25*(6), 273–294. https://doi.org/10.1108/EJIM-05-2021-0256

Chang, C. N., Patterson, C. A., Harmon, W. C., Fowler, D. A., & Arroyave, R. (2020). Intellectual community as a bridge of interdisciplinary graduate education in materials data science. *MRS Advances*, *5*(7), 355–362. https://doi.org/10.1557/adv.2020.140

Dahlstedt, P. (2019). Big data and creativity. *European Review*, *27*(3), 411–439. https://doi.org/10.1017/S1062798719000073

Damayanthi, S. (2019). *Thematic analysis of interview data in the context of management controls research*. Sage Publications. https://doi.org/10.4135/9781526474858

de Mast, J., Steiner, S. H., Nuijten, W. P., & Kapitan, D. (2022). Analytical problem solving based on causal, correlational and deductive models. *The American Statistician*. Advance online publication. https://doi.org/10.1080/00031305.2021.2023633

de Oliveira, S., & Nisbett, R. E. (2017). Culture changes how we think about thinking: From "human inference" to "geography of thought." *Perspectives on Psychological Science*, *12*(5), 782–790. https://doi.org/10.1177/1745691617702718

Egger, R., & Yu, J. (2022). Epistemological challenges. In R. Egger (Ed.), *Applied data science in tourism: Interdisciplinary approaches, methodologies, and applications* (pp. 17–34). Springer. https://doi.org/10.1007/978-3-030-88389-8_2

El-Alfy, E. S. M., & Mohammed, S. A. (2020). A review of machine learning for big data analytics: Bibliometric approach. *Technology Analysis & Strategic Management*, *32*(8), 984–1005. https://doi.org/10.1080/09537325.2020.1732912

Günther, W. A., Mehrizi, M. H. R., Huysman, M., & Feldberg, F. (2017). Debating big data: A literature review on realizing value from big data. *The Journal of Strategic Information Systems*, *26*(3), 191–209. https://doi.org/10.1016/j.jsis.2017.07.003

Hassad, R. A. (2011). Constructivist and behaviorist approaches: Development and initial evaluation of a teaching practice scale for introductory statistics at the college level. *Numeracy: Advancing Education in Quantitative Literacy*, *4*(2), 1–33. https://doi.org/10.5038/1936-4660.4.2.7

Hassad, R. A. (2020). A foundation for inductive reasoning in harnessing the potential of big data. *Statistics Education Research Journal*, *19*(1), 238–258. https://doi.org/10.52041/serj.v19i1.133

Horton, N. J., & Hardin, J. S. (2015). Teaching the next generation of statistics students to "think with data": Special issue on statistics and the undergraduate curriculum. *The American Statistician*, *69*(4), 259–265. https://doi.org/10.1080/00031305.2015.1094283

Househ, M. S., Aldosari, B., Alanazi, A., Kushniruk, A. W., & Borycki, E. M. (2017). Big data, big problems: A healthcare perspective. *ICIMTH*, 36–39. https://pubmed.ncbi.nlm.nih.gov/28679881/

Jenkins, T. (2013, February 12). Don't count on big data for answers. *The Scotsman*. https://www.scotsman.com/news/opinion/columnists/tiffany-jenkins-dont-count-big-data-answers-1590688

Kelling, S., Hochachka, W. M., Fink, D., Riedewald, M., Caruana, R., Ballard, G., & Hooker, G. (2009). Data-intensive science: A new paradigm for biodiversity studies. *BioScience*, *59*(7), 613–620. https://doi.org/10.1525/bio.2009.59.7.12

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 1–12. https://doi.org/10.1177/2053951714528481

Leonelli, S. (2020, Summer). Scientific research and big data. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy.* Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/entries/science-big-data/

Luan, H., Geczy, P., Lai, H., Gobert, J., Yang, S. J., Ogata, H., Baltes, J., Guerra, R., Li, P., & Tsai, C. C. (2020). Challenges and future directions of big data and artificial intelligence in education. *Frontiers in Psychology, 11*, Article 580820. https://doi.org/10.3389/fpsyg.2020.580820

Pentland, A., Reid, T. G., & Heidbeck, T. (2013). *Big data and health: Revolutionizing medicine and public health.* World Innovation Summit for Health. https://kit.mit.edu/sites/default/files/documents/WISH_BigData_Report.pdf

Raita, Y., Camargo Jr., C. A., Liang, L., & Hasegawa, K. (2021). Big data, data science, and causal inference: A primer for clinicians. *Frontiers in Medicine*, *8*, Article 678047. https://doi.org/10.3389/fmed.2021.678047

Rock, I., & Palmer, S. (1990). The legacy of gestalt psychology. *Scientific American*, *263*(6), 84–91. https://doi.org/10.1038/scientificamerican1290-84

Rodriguez, F., Scheinker, D., & Harrington, R. A. (2018). Promise and perils of big data and artificial intelligence in clinical medicine and biomedical research. *Circulation Research*, *123*(12), 1282–1284. https://doi.org/10.1161/CIRCRESAHA.118.314119

Singh, R., Yang, H., Dalziel, B., Asarnow, D., Murad, W., Foote, D., Gormley, M., Stillman, J., & Fisher, S. (2013). Towards human-computer synergetic analysis of large-scale biological data. *BMC bioinformatics*, *14 Suppl 14*(Suppl 14), Article S10. https://doi.org/10.1186/1471-2105-14-S14-S10

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & von der Heydt, R. (2012). A century of gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization. *Psychological Bulletin*, *138*(6), 1172–1217. https://psycnet.apa.org/doiLanding?doi=10.1037%2Fa0029333

Wong, W., & Hinnant, C. (2022). Competing perspectives on the big data revolution: A typology of applications in public policy. *Journal of Economic Policy Reform*. Advance online publication. https://doi.org/10.1080/17487870.2022.2103701