BMC Research Notes

EDITORIAL Open Access

Advancing methods in big data capture, integration, classification and liberation

Eftim Zdravevski[1*] and Ivan Miguel Pires[2,3]

Abstract
This special issue focuses on the importance of advancing research techniques for managing and analyzing data in today's data-rich landscape. In this editorial, we set the context and invite contributions for a *BMC* Collection of articles titled 'Advancing methods in data capture, integration, classification and liberation'. The collection emphasizes the need for efficient ways to standardize, cleanse, integrate, enrich, and liberate data, highlighting recent advancements in research methods and industrial technologies that facilitate this. We invite researchers to submit their best work to the collection and to showcase the latest advancements and additions to research techniques.

Main text

The exponential growth of data presents both opportunities and challenges [1]. Data can provide valuable insights and lead to new discoveries. Still, the sheer volume of data generated daily can be overwhelming and challenging to manage, process, and use [2].

Over the span of its use, any data set must undergo a series of crucial processes for it to be utilized successfully. Firstly, data must be captured from a given source and stored in a usable format. In turn, data sets from varying sources may need to be integrated and consolidated, increasing the value, accuracy, or reliability of the answers we seek to gain [2]. Next, data classification can help improve data usability and discoverability. This process refers to categorizing or labeling based on some pre-defined criteria or based on some automatically identified patterns or relationships by machine learning processes.

Furthermore, the process of making data more accessible, usable, and shareable is often encompassed by data liberation. This term can implicate removing data access or licensing restrictions, converting data into open or standardized formats, or making data available through APIs or other interfaces that facilitate integration and analysis. Data liberation aims to enable more people to use data to gain insights and make informed decisions. These challenges are relevant for open and reproducible research. Likewise, this has significant implications for organizations with siloed data and business users who require instant access to data and insights from it [3].

The recent growth of sensor technologies has made it possible to capture real-time data in various new contexts, including wearable devices, manufacturing, smart homes, smart cities, autonomous vehicles, smart agriculture [4] and medicine [5], to name a few. Moreover, these advancements coincide with the exponential growth in machine learning applications [6], causing completely novel use cases to emerge and raising new challenges. This collection seeks to publish research note articles that

*Correspondence:
Eftim Zdravevski
eftim.zdravevski@finki.ukim.mk
[1]Faculty of Computer Science and Engineering, University Ss Cyril and Methodius, Skopje 1000, Macedonia
[2]Instituto de Telecomunicações, Universidade da Beira Interior, Covilhã 6200-001, Portugal
[3]Polytechnic Institute of Santarém, Santarém, Portugal

address novel or improved data capture methods, better processes for data classification and contextualization, methods for integrating data from a variety of sources, improving automatic tagging of data, identifying sensitive information, recognizing relationships between data, and more efficient ways for users to view and export their data. With the vast choices of technology stacks, significant challenges exist in minimizing the total cost of ownership of cloud infrastructures and choosing optimal, cost-effective solutions [7].

We are also seeking articles that report improvements to existing methods and techniques or introduce new methodologies to the field. For example, novel data capture techniques could include the use of machine learning algorithms for automated data acquisition or the use of wearable devices to collect data in real-time. Improvements in data classification and contextualization could involve more sophisticated natural language processing techniques, while the integration of data from multiple sources may require new approaches to data modeling and data warehousing. Finally, better methods for data liberation could include more user-friendly interfaces or improved tools for data visualization.

We also welcome commentaries discussing the history and future of significant techniques or methodologies in this field as well as the controversies they may present and how they can be addressed. These commentaries should provide a thoughtful and well-informed perspective on the relevant issues.

We invite researchers to submit their work to this collection, which promises to showcase the latest advancements and additions to research techniques in the data science field. With the rapidly changing landscape of data science, we believe that this collection will contribute to a deeper understanding of the field and facilitate further research in this critical area. We hope that the papers published in this issue will lead to a better understanding of the challenges and opportunities that researchers face when working with data and inspire new and innovative techniques for data management and analysis.

Acknowledgements
E.Z. acknowledges the support of the Ss. Cyril and Methodius University in Skopje, Faculty of Computer Science and Engineering, Macedonia. I.M.P. acknowledges the support of Instituto de Telecomunicações, and the Polytechnic Institute of Santarém, Portugal.

Author Contribution
E.Z. wrote the draft of the Editorial text and I.M.P. provided feedback and improved it. All authors reviewed the manuscript.

Funding
I.M.P. acknowledges the funding for this work by FCT/MEC through national funds and co-funded by FEDER - PT2020 partnership agreement under the project UIDB/50008/2020.

Data Availability
Considering that this article is a Guest Editor Editorial, and there is no data and material used in its preparation, there is nothing to declare in this section.

Declarations

Ethics approval and consent to participate
Considering that this article is a Guest Editor Editorial, there is no need for ethical approval and consent to participation.

Consent for publication
Considering that this article is a Guest Editor Editorial, there is no need for consent to publication.

Competing interests
The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; nor in the decision to publish the results.

Received: 13 April 2023 / Accepted: 20 April 2023
Published online: 27 April 2023

References

1. Aykroyd RG, Leiva V, Ruggeri F. Recent developments of control charts, identification of big data sources and future trends of current research. Volume 144. Technological Forecasting and Social Change; 2019. pp. 221–32.
2. Zdravevski E, Lameski P, Apanowicz C, Ślęzak D. 2020. From Big Data to business analytics: The case study of churn prediction. Applied Soft Computing, 90, p.106164.
3. Hashem IAT, Yaqoob I, Anuar NB, Mokhtar S, Gani A, Khan SU. The rise of "big data" on cloud computing: review and open research issues. Inform Syst. 2015;47:98–115.
4. Su Y, Wang X. Innovation of agricultural economic management in the process of constructing smart agriculture by big data. Sustainable Computing: Informatics and Systems. 2021 Sep;1:31:100579.
5. Wang M, Li S, Zheng T, Li N, Shi Q, Zhuo X, Ding R, Huang Y. Big data health care platform with multisource heterogeneous data integration and massive high-dimensional data governance for large hospitals: Design, development, and application. JMIR Medical Informatics. 2022 Apr 13;10(4):e36481.
6. Roh Y, Heo G, Whang SE. A survey on data collection for machine learning: a big data-ai integration perspective. IEEE Transactions on Knowledge and Data Engineering. 2019 Oct 8;33(4):1328-47.
7. Grzegorowski M, Zdravevski E, Janusz A, Lameski P, Apanowicz C, Ślęzak D. Cost optimization for big data workloads based on dynamic scheduling and cluster-size tuning. Big Data Research. 2021;25:100203.

Publisher's Note