FACETS
a multidisciplinary open access science journal

EDITORIAL

Data science

Paul D. McNicholas[a]*

[a]Department of Mathematics & Statistics, McMaster University, Hamilton, ON L8S 4K1, Canada

*paulmc@mcmaster.ca

FACETS is introducing a Data Science section as a platform to facilitate the dissemination of high-quality research focused on data. Given the explosion of data, and new data types, across virtually all areas of research endeavour in recent years, this is the right time for such a platform to emerge. Furthermore, considering the inherently interdisciplinary nature of data science, *FACETS* is the right venue. The assembled team of Subject Editors for the Data Science section gives excellent coverage of both theoretical and applied aspects of data science.

The advent of a *FACETS* Data Science section is an opportune time to consider data science as a field as well as the various types of submissions we wish to see. A little historical context will be provided before moving to the very broad interpretation of data science that *FACETS* will take. Although the activities associated with data science have been around for many years (see, e.g., Donoho 2017), it is useful to start by considering the origins of the term data science and its evolution to what we know today as data science.

The use of the term data science goes at least as far back as the 1996 meeting of the International Federation of Classification Societies (IFCS). As Hayashi (1998, p. 40) explained:

> The roundtable discussion "Perspectives in classification and the Future of IFCS" was held at the last Conference under the chairmanship of Professor H.-H. Bock. In this panel discussion, I used the phrase 'Data Science'. There was a question, "What is 'Data Science'?" I briefly answered it. This is the starting point of the present paper.

The "present paper" referred to above is the contribution of Hayashi (1998, p. 40), wherein data science is described as follows:

> Data science is not only a synthetic concept to unify statistics, data analysis and their related methods but also comprises its results.

Hayashi (1998, p. 40) went on to specifically contrast data analysis and mathematical statistics:

> ... mathematical statistics have been prone to be removed from reality. On the other hand, the method of data analysis has developed in the fields disregarded by mathematical statistics and has given useful results to solve complicated problems based on mathematico-statistical methods (which are not always based on statistical inference but rather are descriptive).

Although the foregoing excerpts from Hayashi (1998) are important for historic context, they do not serve as an accurate contextualization of modern data science. However, before moving entirely beyond this important early work, it is useful to draw parallels—both historic and modern—with the sentiments captured therein.

Arguments for viewing data analyses through a more practical lens go back to well before the entry of the term data science into the mainstream parlance. One of the more famous examples centres around

Citation: McNicholas PD. 2019. Data science. FACETS 4: 131–135. doi:10.1139/facets-2019-0005

Received: January 21, 2019

Accepted: April 8, 2019

Published: May 13, 2019

Published by: Canadian Science Publishing

the disagreement between Fisher and Gossett (the latter being "Student" of *t* test fame) over the notion of significance. Many fascinating accounts of this disagreement have been given, including interesting work by Ziliak (2008). In brief, one might say that the practical and data-context-specific approach favoured by Gossett stood in contrast to the notion of a result being either statistically significant or not. The notion of statistical significance has, to date at least, stood the test of time and is eloquently summarized by Ziliak (2008, p. 200):

> Yet against "Student's" wishes and periodic warnings, it was this same extraordinary Fisher, "Student's" younger friend and colleague, who invented and campaigned for the 5 percent rule of statistical significance. Today, Fisher's preferred interpretation of "Student's" test is customary if not enforced in most sciences, journals, and even courts of law.

Interestingly, the notion of a 5% level of statistical significance is central to the current discourse around *p*-values (see, e.g., Wasserstein and Lazar 2016; Goodman 2019; Kmetz 2019; Startz 2019; Valentine et al. 2019). It is notable, and perhaps inevitable, that the subject of the famous disagreement between Fisher and Gossett has gained increased attention at a time when data science—and related attitudes that emphasize practical and data-context-specific considerations—has burgeoned into a sufficiently important field to penetrate into the public discourse. It may yet be that, around a century later, Gossett wins the argument.

The relationship between data science and statistics was taken up by others shortly after the work of Hayashi (1998). For instance, Cleveland (2001, p. 25) wrote:

> A very limited view of data science is that it is practiced by statisticians. The wide view is that data science is practiced by statisticians and subject matter analysts alike, blurring exactly who is and who is not a statistician.

Although the relationship between data science and statistics is important in understanding the origins of data science, it is important to note that the field of data science is now very broad and reaches far beyond these origins. In a recent monograph, McNicholas and Tait (2019, p. 1–2) discussed the relationship between statistics and data science:

> On the one extreme, some might view data science—and data analysis, in particular—as a retrogression of statistics; yet, on the other extreme, some may argue that data science is a manifestation of what statistics was always meant to be. In reality, it is probably an error to try to compare statistics and data science as if they were alternatives.

McNicholas and Tait (2019, p. 2) went on to take the view that

> … statistics plays a crucial role in data analysis, or data analytics, which in turn is a crucial part of the data science mosaic.

Beyond statistics, computer science plays a very important role in data science today; e.g., machine learning has become very important for data analytics, and data preprocessing can be very computationally demanding. The same is true of optimization, which is a field practiced by people with a variety of backgrounds including management science and mathematics. Crucially, however, there is much more to data science than data analytics. For example, data security and privacy are issues of tremendous practical importance (e.g., Abouelmehdi et al. 2018). Data ethics, and the responsible use of data in general, are fundamental for data science (see Floridi and Taddeo 2016, for example). Communication, including effective data visualization, is another key part of the data science mosaic (e.g., Perkel 2018). The foregoing is far from an exhaustive list of data science topics but it should serve to illustrate how broad the field has become. The breadth of data science is also apparent in

the extremely diverse range of its applications including work in cyber security (e.g., Buczak and Guven 2016), health care (e.g., Spruit and Lytras 2018), finance (e.g., Giudici 2018), and public policy (e.g., Matheus et al. 2018).

The breadth of the field of data science will be reflected in the papers sought by the *FACETS* Data Science section. However, before further discussion on the new section, it will be helpful to consider the relationship between big data and data science. Similar to data science, the term big data has no one universally accepted meaning. Some definitions are based on three, or more, words beginning with the letter V. Puts et al. (2015) give an interesting discussion of the three-V definition as well as addressing the difference between big data and administrative data. Very roughly, the three-V definition defines big data in terms of size (volume), diversity (variety), and streaming (velocity). Whether one of the three Vs will suffice, according to this paradigm, for data to be big data is open for debate. However, this debate will not be enriched herein because to do so might distract from the key point, i.e., there is more to data science than big data. In fact, experience suggests that some of the most difficult data problems do not have any of the Vs. Examples include data that are too few, rather than too many, as well as datasets containing missing data. Both of these situations are difficult, and important, in data science.

However one may wish to define data science, the key must always be data. For a piece of work to be considered data science, we require only that data are at its heart. At the time of writing, the *FACETS* Data Science Section has three subject areas: Data Science Theory and Methods, Data Science Applications, and Research Data Management. Because data must be at the heart of the work, Data Science Theory and Methods and Data Science Applications are very much related. Consequently, it might be difficult for authors to determine which of these subject areas is most suitable for a particular manuscript. As a guideline, if the novelty principally concerns methodology, then the Data Science Theory and Methods area should be chosen whereas, if the novelty lies in the application, then Data Science Applications is more suitable. The historically important description of data science given by Hayashi (1998), and related discussion herein, might give one the impression that methodological statistics work is unwelcome; however, this is not the case. All theory and methodology will be given full consideration provided that data are at the core of the work. As one would expect, work concerning data storage, security, privacy, etc. should be submitted to the Research Data Management subject area. Papers on other topics in data science, such as data ethics and the responsible use of data, are welcome and should also be submitted to the Research Data Management area.

Acknowledgements

The author is grateful to the Subject Editors for the Data Science Section as well as the Editor-in-Chief and several colleagues for their thoughtful comments on this editorial.

Author contributions

PDM drafted and revised the manuscript.

Competing interests

PDM is the Section Editor of the Data Science Section of *FACETS*.

References

Abouelmehdi K, Beni-Hessane A, and Khaloufi H. 2018. Big healthcare data: preserving security and privacy. Journal of Big Data, 5(1). DOI: 10.1186/s40537-017-0110-7

Buczak AL, and Guven E. 2016. A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials, 18(2): 1153–1176. DOI: 10.1109/COMST.2015.2494502

Cleveland WS. 2001. Data science: an action plan for expanding the technical areas of the field of statistics. International Statistical Review/Revue Internationale De Statistique, 69(1): 21–26. DOI: 10.1111/j.1751-5823.2001.tb00477.x

Donoho D. 2017. 50 years of data science. Journal of Computational and Graphical Statistics, 26(4): 745–766. DOI: 10.1080/10618600.2017.1384734

Floridi L, and Taddeo M. 2016. What is data ethics? Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2083): 20160360.

Giudici P. 2018. Financial data science. Statistics and Probability Letters, 136: 160–164. DOI: 10.1016/j.spl.2018.02.024

Goodman SN. 2019. Why is getting rid of *p*-values so hard? Musings on science and statistics. The American Statistician, 73(Suppl. 1): 26–30. DOI: 10.1080/00031305.2018.1558111

Hayashi C. 1998. What is data science? Fundamental concepts and a heuristic example. *In* Data science, classification, and related methods. *Edited by* C Hayashi, K Yajima, HH Bock, N Ohsumi, Y Tanaka, and Y Baba. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Tokyo, Japan. pp. 40–51.

Kmetz JL. 2019. Correcting corrupt research: recommendations for the profession to stop misuse of *p*-values. The American Statistician, 73(Suppl. 1): 36–45. DOI: 10.1080/00031305.2018.1518271

Matheus R, Janssen M, and Maheshwari D. 2018. Data science empowering the public: data-driven dashboards for transparent and accountable decision-making in smart cities. Government Information Quarterly (in press). DOI: 10.1016/j.giq.2018.01.006

McNicholas PD, and Tait PA. 2019. Data science with Julia. Chapman & Hall/CRC Press, Boca Raton, Florida.

Perkel JM. 2018. Data visualization tools drive interactivity and reproducibility in online publishing. Nature, 554(7690): 133–134. PMID: 29388968 DOI: 10.1038/d41586-018-01322-9

Puts M, Daas P, and de Waal T. 2015. Finding errors in big data. Significance, 12(3): 26–29. DOI: 10.1111/j.1740-9713.2015.00826.x

Spruit M, and Lytras M. 2018. Applied data science in patient-centric healthcare: adaptive analytic systems for empowering physicians and patients. Telematics and Informatics, 35(4): 643–653. DOI: 10.1016/j.tele.2018.04.002

Startz R. 2019. Not *p*-values, said a little bit differently. Econometrics, 7(1): 11. DOI: 10.3390/econometrics7010011

Valentine KD, Buchanan EM, Scofield JE, and Beauchamp MT. 2019. Beyond *p* values: utilizing multiple methods to evaluate evidence. Behaviormetrika, 46: 121–144. DOI: 10.1007/s41237-019-00078-4

Wasserstein RL, and Lazar NA. 2016. The ASA's statement on *p*-values: context, process, and purpose. The American Statistician, 70: 129–133. DOI: 10.1080/00031305.2016.1154108

Ziliak ST. 2008. Retrospectives: Guinnessometrics: the economic foundation of "Student's" *t*. Journal of Economic Perspectives, 22(4): 199–216. DOI: 10.1257/jep.22.4.199