WHAT MAKES DATA SCIENCE EDUCATION UNIQUE?: A LITERATURE REVIEW

Hiroto Fukuda
Okayama University of Science, Japan
hfukuda@xmath.ous.ac.jp

Although data science education research has been accumulating, the meaning of data science itself in such research is varied, as there is no common definition. Thus, the purpose of this paper is to clarify the several characteristics that make data science education unique. Through a review of previous studies on data science education, it was clarified that the definition of 'data science in data science education' is multifaceted, moreover, there is no mention of the meaning of 'data science education'. This paper attempts to explore this challenge and discusses the statistical inquiry cycle, PPDAC cycle (Problem – Plan – Data – Analysis - Conclusion) (Wild & Pfannkuch, 1999) as a theoretical framework for this purpose. As a result, the several characteristics are as follows: trans-scientific problems (Problem), interdisciplinary approach (Plan), big data (Data), technology use & teamwork (Analysis), and fruitful disagreement (Conclusion).

INTRODUCTION

'Data' symbolises our society, including its future, and thus, data literacy is an indispensable skill to judge and use essential data correctly from the large amount of data that they will encounter throughout their lives (cf. Fukuda, 2020; Ben-Zvi, Makar, & Garfield, 2018). With the increasing popularity of data science, data scientists are attracting attention and being a data scientist is said to be the sexiest job of 21st century (Davenport & Patil, 2012). Although data science education research has been accumulating, the meaning of data science itself in such research is varied, as there is no common definition. Thus, the purpose of this paper is to clarify the several characteristics that make data science education unique.

For this literature review, the author first reviewed previous studies on data science education, which mention the meaning of data science and confirmed whether the meaning of data science is unclear. The author used the statistical inquiry cycle proposed by Wild and Pfannkuch (1999) as a framework to achieve the research purpose, as this framework is similar to data science in terms of the inquiry procedure.

LITERATURE REVIEW

The term 'data science' was first used in a lecture by Wu (1997). Subsequently, Cleveland (2001) proposed data science as a new discipline, since then, data science has rapidly advanced in the society. Thereafter, Yan and Davis (2019) laid the background for data science, while stating that, 'despite the fact that data science has become so popular, and we are using products enabled by data science on almost a daily basis, there is currently no consensus on the definition of data science' (Yan & Davis, 2019, p. 99). Indeed, Yan and Davis (2019) define data science as 'the science of learning from data' (p. 99). However, Engel (2017) defines it as 'a set of skills and techniques that include statistics, data mining, computer science, domain expertise, and communication' (p. 47), which are represented as a Venn diagram (Figure 1).

Furthermore, Taylor (2016) introduced the following definition of data science, although the source of the quote is unknown: 'work that takes more programming skills than most statisticians have, and more statistical skills than a programmer has' (n.d.). Something similar mentioned in Yan and Davis's (2019) and Kauermann's (2018) works, where data science is defined as consisting of statistics and computer science, and is structured in a way that the two have a complementary relationship, as shown in Figure 2.

As mentioned above, although the meaning of 'data science in data science education' is polysemous, each researcher approaches their definition either linguistically or graphically. Moreover, the meaning of 'data science education' has not been mentioned in any of these studies. Therefore, in the next section, the author considers data science from an educational perspective and clarifies the characteristics that make data science education unique.

In: R Helenius, E Falck (Eds.), *Statistics Education in the Era of Data Science*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
Aug-Sept 2021, Online conference. iase-web.org/Conference_Proceedings.php

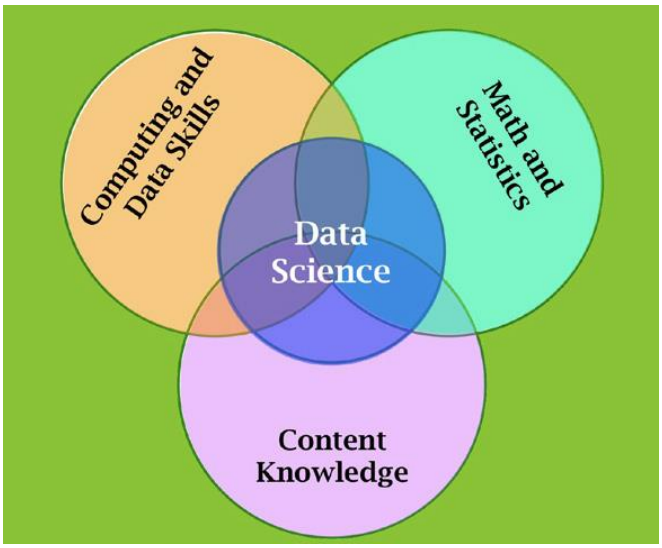

Figure 1. Data science as interdisciplinary field (Engel, 2017, p. 47)

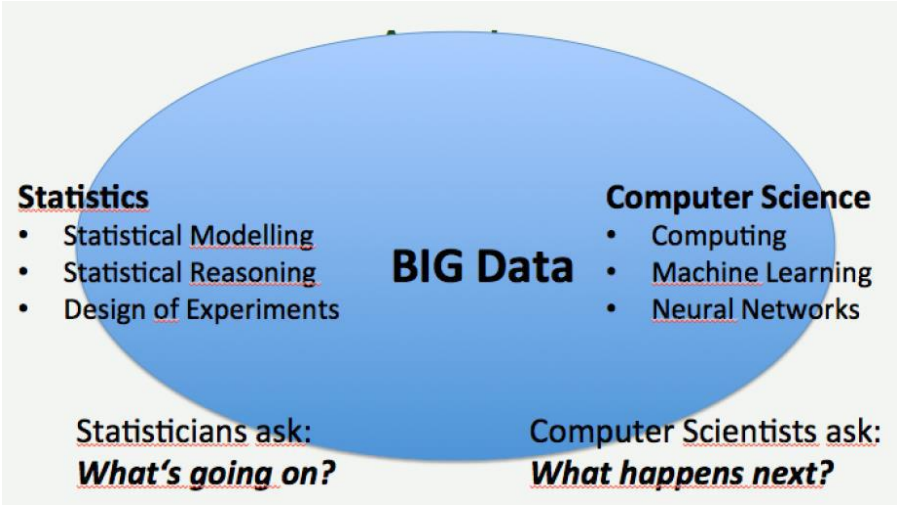

Figure 2. Statistics and computer science in data science (Kauermann, 2018, p. 18)

RESULT

In this study, the author uses the statistical inquiry cycle proposed by Wild and Pfannkuch (1999) as a framework. This statistical inquiry cycle consists of problem, plan, data, analysis, and conclusion (PPDAC cycle), which was proposed based on the inquiry of many statisticians (Figure 3). Although the PPDAC cycle is an inquiry method used by statisticians, the procedure of statistical inquiry performed by programmers is similar, thus, it is a significant investigative tool in data science education as well. Therefore, the author decided to adopt it as the theoretical framework for this study.

In the following paragraphs, each element of the PPDAC cycle is discussed. The first element is the problem. The author describes the characteristics of the problems addressed in data science education. For instance, the problem associated with the new coronavirus, the mutant strains of which are now spreading around the world, is said to be a trans-scientific problem. Trans-scientific problems are described as 'questions which can be asked of science and yet which cannot be answered by science' (Weinberg, 1972, p. 209). In today's society of highly complex information, not only problems related to the new coronavirus, but several problems can be considered as trans-scientific problems with complex structures. In addition, dealing with such problems from a trans-scientific context can provide the foundations to conduct a data scientific inquiry.

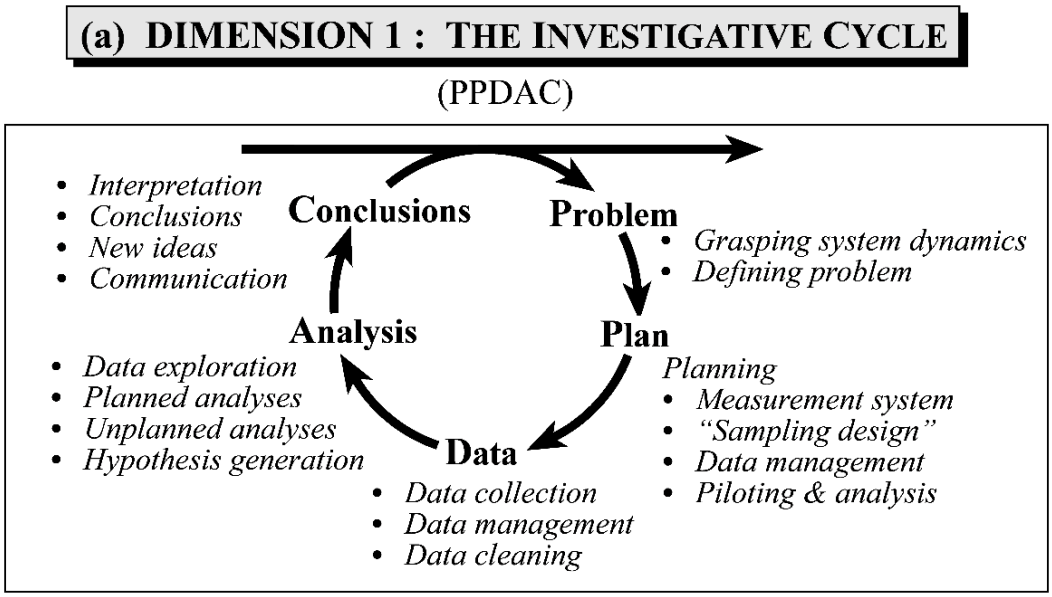

Figure 3. The investigative cycle (Wild & Pfannkuch, 1999, p. 226)

The second element is the plan. A data scientific inquiry cycle within a trans-scientific context requires not only statistical and mathematical knowledge but also scientific, social, and similar realms of knowledge. Therefore, an interdisciplinary approach must be adopted to complete an authentic data scientific inquiry cycle (cf. Zieffler, Garfield, & Fry, 2018; Gal & Garfield, 1997).

As for the third element, dataset, the problem of the new coronavirus is a clear example, wherein, in order to carry out the analysis, data from all over the world are collected. In today's technologically advanced and information-oriented society, data is not just a few numbers that can be counted manually, and thus, students need to acquire the skills for processing big data (Ridgway, 2016). This is evident from the central position of big data in Figure 2.

The fourth element is analysis. The processing of big data cannot be done with pens and paper alone, and requires the use of technology for investigation (cf. Ben-Zvi, Gravemeijer, & Ainley, 2018; Saldanha & McAllister, 2016). In addition, as evident in the case of the new coronavirus, the solution to trans-scientific problems does not depend only on experts in a particular field but requires the collective wisdom of experts from various fields, and interpretation and decision-making in light of the various trade-offs by politicians and the public. Therefore, data analysis involves team effort and not individual effort.

Finally, the fifth element is conclusion. In trans-scientific problems, each solution is not an absolute or unique solution. In other words, the solution of one researcher may be in direct conflict with the solution provided by another. When this happens, they need to work as a team to find a new solution or 'fruitful disagreement', by developing perspective for each other's solutions and finding areas of compromise and sympathy for them, rather than 'sterile confrontation' where they simply reject each other's solutions (Noe, 2012; Rorty, 1981).

CONCLUSION AND IMPLICATIONS

The purpose of this paper is to clarify the several characteristics that make data science education unique. Through a review of previous studies on data science education, it was clarified that the definition of 'data science in data science education' is multifaceted, moreover, there is no mention of the meaning of 'data science education'. This paper attempts to explore this challenge and discusses the PPDAC cycle as a theoretical framework for this purpose. Table 1. shows the characteristics that make data science education unique.

The PPDAC cycle, a normative theory of statistical inquiry, continues to be relevant for more than 20 years after it was first proposed, because it is based on the inquiry methods of statisticians. However, the properties of the elements of the PPDAC cycle are time-dependent, and therefore, these properties in today's data science education were clarified, and it can be said that new

Table 1. The characteristics that make data science unique

Viewpoint	Characteristics
Problem	Trans-scientific problems
Plan	Interdisciplinary approach
Data	Big data
Analysis	Technology use & Teamwork
Conclusion	Fruitful disagreement

implications have been given to the normative theory of the PPDAC cycle. While there are various views on the definition of data science, this paper does not define it, but clarifies the characteristics of data science education, which is original. Furthermore, it is novel that these characteristics are organised from the viewpoint of the PPDAC cycle.

ACKNOWLEDGEMENTS

This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number 19K14219.

REFERENCES

Ben-Zvi, D., Gravemeijer, K., & Ainley, J. (2018). Design of statistics learning environments. In D. Ben-Zvi, K. Makar, & J. Garfield (Eds.), *International handbook of research in statistics education* (pp. 473-502). Cham, Switzerland: Springer International Publishing.

Ben-Zvi, D., Makar, K., & Garfield, J. eds. (2018). *International handbook of research in statistics education*. Cham, Switzerland: Springer International Publishing.

Cleveland, W. S. (2001). Data science: An action plan for expanding the technical areas of the field of statistics. *International Statistical Review*, *69*(1), 21-26.

Davenport, T. H., & Patil, D. J. (2012). Data scientist: The sexiest job of the 21st century, *Harvard Business Review*, *90*, 70-76.

Engel, J. (2017). Statistical literacy for active citizenship: A call for data science education. *Statistics Education Research Journal*, *16*(1), 44-49.

Fukuda, H. (2020). *Research towards a principle for the statistics curriculum in Japan from the perspective of context*, Doctoral dissertation. Hiroshima University. *https://ir.lib.hiroshima-u.ac.jp/files/public/4/49358/20200721115046156763/k8061_3.pdf*

Gal, I., & Garfirld, J. (1997). Curricular goals and assessment challenges in statistics education. In I. Gal & J. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 1-14). Amsterdam, The Netherlands: IOS Press.

Kauermann, G. (2018). Data science master's degree. In R. Biehler, L. Budde, D. Frischemeier, B. Heinemann, S. Podworny, C. Schulte, & T. Wassong (Eds.), *Paderborn symposium on data science education at school level 2017: The collected extended abstracts* (pp. 18-20). Paderborn: Universitätsbibliothek Paderborn. http://doi.org/10.17619/UNIPB/1-374

Noe, K. (2012). For the sake of "fruitful disagreement". *Trends in the Sciences*, *17*(5), 46-50. (in Japanese)

Ridgway, J. (2016). Implications of the data revolution for statistics education. *International Statistical Review*, *84*(3), 528-549.

Rorty, R. (1981). *Philosophy and the mirror of nature*. New Jersey, US: Princeton University Press.

Saldanha, L., & McAllister, M. (2016). Building up the box plot as a tool for representing and structuring data distributions: An instructional effort using Tinkerplots and evidence of students' reasoning. In D. Ben-Zvi & K. Makar (Eds.), *The teaching and learning of statistics: International perspectives* (pp. 235-245). Switzerland: Springer International Publishing.

Taylor, D. (2016). Battle of the data science Venn diagrams. KDNuggets News. https://www.kdnuggets.com/2016/10/battle-data-science-venn-diagrams.html

Weinberg, A. M. (1972). Science and trans-science. *Minerva*, *10*(2), 209-222.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*(3), 223-265.

Wu, C.-F. J. (1997). Statistics = data science?, in *H. C. Carver Professorship Lecture, Ann Arbor*, MI: The University of Michigan. file:///C:/Users/Owner/Downloads/datascience.pdf

Yan, D., & Davis, G. E. (2019). A first course in data science. *Journal of Statistics Education*, *27*(2), 99-109.

Zieffler, A., Garfield, J., & Fry, E. (2018). What is statistics education?. In D. Ben-Zvi, K. Makar, & J. Garfield (Eds.), *International handbook of research in statistics education* (pp. 37-70). Cham, Switzerland: Springer International Publishing.