Data Literacy as a meta-skill: options for Data Science curriculum implementation

Pavel Glukhov[1,2*], *Andrey* Deryabin[2], and *Aleksandr* Popov[1,2]

[1]Moscow City University, Institute of System Projects, Moscow, Russia
[2]Russian Presidential Academy of National Economy and Public Administration, Federal Institute for Education Development, Moscow, Russia

Abstract. Data science is affecting an increasingly wide area of everyday life but general education in Russia has not yet reacted to the new challenges associated with this aspect of digitalization. The changes in technologies, the economy, and society over the last two decades have formed a new agenda for teaching mathematics and information technologies, as well as media education and social sciences. Education in all these fields requires a reconsideration of the content and methods of teaching due to the increasing importance of data science and artificial intelligence in the context of fundamental changes in the economy and the labor market. As many areas of human life are changing, there is a need to formulate new types and kinds of educational results, at which modern pedagogy should be aimed. A modern way of meeting such challenges is to distinguish new literacies (media literacy, environmental literacy, functional literacy, etc.). The article deals with the concept of data literacy, examines its content and composition, and substantiates its relevance as an educational result consistent with digitalization trends that one can observe in modern society. A distinction is made between approaches to in-depth and general studies of data science. A description is given of various types of tasks aimed at developing data literacy among students in the context of their setting on different educational material. The authors consider possible ways of deploying programs aimed at mastering data science by students without the need to formalize it into a separate discipline or school subject.

Keywords: data science, new literacies, data literacy, education.

1 Introduction

Today, we live in a digital society, one of the most important characteristics of which is “datafication” – when many processes of economic and social development are based on the collection and processing of large amounts of data (big data), including personal data. Big data is data, the volume, diversity, and growth rate of which exceed the capabilities of traditional systems for storing and processing information that existed at the turn of the 2000s. In a certain sense, a modern person becomes a source of data, which is expressed in what the person buys, what links the person clicks on the Internet, how much time the

[*] Corresponding author: gluhovpav.pav@gmail.com

person spends on viewing information, what geolocations the person visits, etc. Becoming the owner of a particular information product (a smartphone, application, or software), one enters a relationship with different algorithms that collect information about their actions and interests. The question arises about information security and whether specific people are aware of this aspect. When we discuss young people who pick up a smartphone at a fairly early age, this issue becomes even more relevant. However, we should not talk about isolating young people from modern advances in information technology. The formation of specific types of literacy is seen as a more rational task, which at least allows increasing the level of awareness, critical attitude, and reflection towards the operation of modern technical means and information. The goal of the study is to determine the content of a specific type of literacy as a possible new educational outcome focused on a reflective attitude towards the phenomenon of big data in modern society.

2 Methods

The study is an overview in which we tried to analyze and summarize modern concepts of educational outcomes related to the phenomenon of big data. We selected the sources that allowed us to consider literacy in the field of big data from the perspective of general developmental and not (pre-) professional educational results.

3 Results

Although the analysis of big datasets is more often referred to highly specialized and professional fields of activity, data science tools seem to be more widely available than it might seem at first glance. Today, the concept of data literacy is acquiring clear boundaries and can be described as a general ability, applicable in the framework of social and ethical issues, allowing one to develop informed rational decisions and formulate an evidence-based position in the modern digital world [1]. Pangrazio and Sefton-Green [2] rightly emphasize that data should be treated critically and with caution as the datasets used (systematized datasets) can undergo too many processing cycles and be used by different agents in different contexts, with different goals and motives. Data literacy presupposes the use of quantitative and qualitative data analysis tools but a much more important skill here is the use of such tools in a specific context [3].

Data literacy is seen as a set of different characteristics possessed by a modern member of the information society and can include configurations of the following abilities:

- assessing the consequences of using and providing data about oneself and one’s life when working with modern mobile devices, payment systems, social networks, various information services and software, etc.;
- reconstructing the context of the presented information systematized by someone to the level of data used in it and methods of their application;
- search, discovery, extraction, and processing of data necessary for more rational decision-making in life, which may be associated with the choice of a city of residence and a place of work or with forecasting the financial situation of one’s own family.

4 Discussion

One of the key issues related to data literacy is the redundancy and feasibility of using methods and tools for teaching data science. What can we eliminate in the classical ways of studying data science to reach the level of formation of general developmental educational results that are widely used at different (including early) stages of development?

Most often, we can find a wide range of offers in the field of specialized educational programs on data science. There are popular online courses that try to respond to staffing demand. Here are examples of courses on Coursera [4] and edX [5]. We can also see a lot of professional training programs in the field of Data Science at the undergraduate and graduate levels [6-8]. Such programs strongly focus on the technical side of the data analysis process, which makes one consider them highly specialized training programs. This approach may be redundant for solving the problems of mass formation of the relevant type of literacy in as in this case we pay less attention to the overall impact of data on the life of society, economy, and culture, which is also noted by experts [9]. In turn, those examples that affect the general contour of teaching data science to adolescents have different durations and feature a different range of content [10-15].

According to some data science educators, for the successful development of this field, one must possess not just complex knowledge of specialized software but, more importantly, a combination of basic knowledge in this area with the ability to analyze [16]. The dependence of more and more areas of modern life on big data sets actualizes these personality characteristics in a general developmental context. At this point, we can discuss not just the possibility of developing skills in working with big data but rather general analytical skills that allow one to be critical of information that uses big data or arrays of specialized data. This may be due to the increase in such a phenomenon as speculation on complexity for the sake of dubious justification of any social and economic phenomena through manipulation of big datasets or corpora of statistical data. We saw a vivid demonstration of this phenomenon during the proliferation of conflicting information about the spread of COVID-19. People tend to trust charts and statistics as these means are pseudoscientific and operate with actual numbers. However, despite the presence of figures that reflect the facts, the information can be applied incorrectly (for example, the data used can be obtained by incorrect methods or could go through many cleanup cycles, etc.). In this sense, teaching data literacy is, among other things, the process of forming competent criticality, which allows one not only to doubt the reliability of the information but also to verify the reliability of the data that underlie the information provided to the public.

Cuoco et al. [17] write about the importance of forming a “mental habit” of accessing data as a way of thinking, questioning, and solving problems. Finzer [18] is also a firm believer that during school teaching one must instill specific thinking, a view of the world through the prism of data. The researcher notes that “thinking in data” is a kind of meta-skill that should be developed in school through the application of data science methods within existing subjects and not through the introduction of a special discipline. This vision is also shared by authors writing about the development of educational programs in data science and statistics at the high school level [6, 19]. At the same time, today, big data and data science practically do not stand out from the entire array of distance digital technologies in education. These subjects are included in the basic curricula of the subjects “Information technology” and “Technology” as individual components, taught in a detached, objectified mode. Rarely, computer data analysis and machine learning are shown to students as tools for solving educational problems, including design-related and creative tasks [20].

5 Conclusion

Data literacy is associated not so much with big data as with data in general, methods of data processing, and general analytical capabilities which allow one to reconstruct the basis of information that affects them and form a valid basis for making one’s own rational decisions. It should also be understood that data literacy differs from digital and information literacy but correlates with them, specifying and complementing the former two in a certain way. If digital literacy allows a person to efficiently and safely use digital technologies and resources according to the context of the assigned tasks, then data literacy allows one to see and understand what data is generated and used in the process of interacting with digital technologies. Depending on the behavior when interacting with a particular service, one may receive different results based on the types of data collected during such interaction. For example, the cost of a taxi ride that one orders using applications depends on various variables, including how often they use the app, the order history, the trip cancellation history, the willingness to pay more or less in peak situations, demand for taxis, etc. Depending on the personal sets of such data collected throughout the entire process of user interaction with the program, the cost of identical trips can vary significantly for two different users. It is important to understand the methods and logic of data processing, considering that these methods change depending on which company services use them and for what tasks.

Information literacy is formed around the ability to be critical of the surrounding information and to reflexively seek and apply information to solve certain problems. On the one hand, data literacy makes it possible to establish what lies at the basis of a complex set of information, for example, using statistics. In such cases, data literacy makes it possible to assess the correctness and redundancy of the statistics used. On the other hand, data literacy allows one to reconstruct the sets of information underlying the algorithms that recommend particular information or advertisement. In this case, there is an illustrative precedent associated with Cambridge Analytica which collected personal data from millions of users of the social network Facebook to simulate electoral behavior to further select relevant political ads that influence the “undecided” voters’ final political preferences. It can be assumed that a person who understands the methods of collecting and processing data will become suspicious if they observe an increase in the recommended content associated with a particular political party.

The application of data science in education already allows two different educational outcomes for which it can work: general developmental and (pre-) professional. Data literacy should be viewed as a meta-skill formed by appropriate approaches, without singling it out as a separate discipline or subject, unless one aims to provide professional or highly specialized training. Data literacy can be considered as a self-sufficient educational result but integrated into other types and kinds of educational activities.

References

1. M. Schield, IASSIST Quarterly, **28(2)**, 6–11 (2004). Accessed on: December 16, 2020. [Online]. Available: https://iassistquarterly.com/public/pdfs/iqvol282_3shields.pdf
2. L. Pangrazio, J. Sefton-Green, Learning, Media and Technology, **45(2)**, 208–220 (2019). https://doi.org/10.1080/17439884.2020.1707223
3. J.P. Gibson, T. Mourad, American Journal of Botany, **105(12)**, 1953–1956 (2018). https://doi.org/10.1002/ajb2.1195
4. Coursera Inc., Nauka o dannih (2020). Accessed on: December 16, 2020. [Online]. Available: https://www.coursera.org/browse/data-science
5. edX Inc., Data science courses on edX (2020). Accessed on: December 16, 2020. [Online]. Available:https://www.edx.org/course/subject/data-science
6. R. De Veaux et al., Annual Review of Statistics and Its Application, **4(1)**, 15–30 (2017). https://doi.org/10.1146/annurev-statistics-060116-053930
7. B.S. Baumer, The American Statistician, **69(4)**, 334–342 (2015). https://doi.org/10.1080/00031305.2015.1081105

8. P.E. Anderson, J.F. Bowring, R. McCauley, G.J. Pothering, Ch.W. Starr, *An undergraduate degree in data science: Curriculum and a decade of implementation experience*, in Proceedings of the 45th ACM Technical Symposium on Computer Science Education (SIGCSE '14), ACM, March 2014, Atlanta Georgia, USA, 145–150 (2014). https://doi.org/10.1145/2538862.2538936
9. B. Heinemann et al., *Drafting a data science curriculum for secondary schools,* in Proceedings of the 18th International Conference on Computing Education,. University of Helsinki, November 2018, Koli, Finland (2018). https://doi.org/10.1145/3279720.3279737
10. S. Srikant, V. Aggarwal, *Introducing data science to school kids,* in Proceedings of the The 48th ACM Technical Symposium on Computer Science Education, SIGCSE, March 2017, Seattle Washington, USA (2017). https://doi.org/10.1145/3017680.3017717
11. C. Bryant et al., *A middle-school camp emphasizing data science and computing for social good*, in Proceedings of the 50th ACM Technical Symposium on Computer Science Education, SIGCSE, February 2019, Minneapolis MN, USA (2019). https://doi.org/10.1145/3287324.3287510
12. A. Dryer, N. Walia, A. Chattopadhyay, *A middle-school module for introducing data-mining, big-data, ethics and privacy using rapidminer and a Hollywood theme,* in Proceedings of the 49th ACM Technical Symposium on Computer Science Education, ACM, February 2018, Baltimore, USA (2018). https://doi.org/10.1145/3159450.3159553
13. S. Datta, V. Nagabandi, *Integrating data science and R programming at an early stage*, in IEEE 4th International Conference on Soft Computing & Machine Intelligence (ISCMI), IEEE, 23-24th November 2017, Port Louis, Mauritius (2017). https://doi.org/10.1109/ISCMI.2017.8279587
14. R. Mariescu-Istodor, I. Jormanainen, *Machine Learning Exercises for High School Students,* in Proceedings of the 19th Koli Calling International Conference on Computing Education Research, November 2019, Koli, Finland (2019). Accessed on: December 16, 2020. [Online]. Available: http://www.cs.columbia.edu/~CS4HS/talks/ml_for_hs.pdf
15. A. Wolff, M. Wermelinger, M. Petre, International Journal of Human Computer Studies, **129**, 41-54 (2019). http://dx.doi.org/doi:10.1016/j.ijhcs.2019.03.006
16. Data science is the science of data. How to become a data scientist from ground zero. Future2day.ru (2019) Accessed on: December 16, 2020. [Online]. Available: https://future2day.ru/data-science/
17. A. Cuoco, E.P. Goldenberg, J. Mark, Journal of Mathematical Behavior, **15(4)**, 375–402 (1997). https://doi.org/10.1016/S0732-3123(96)90023-1
18. W. Finzer, Technology Innovations in Statistics Education, **7(2)** (2013). Accessed on: December 16, 2020. [Online]. Available: https://escholarship.org/uc/item/7gv0q9dc
19. J. Hardin, et al., The American Statistician, **69(4)**, 343–353 (2015). https://doi.org/10.1080/00031305.2015.1077729
20. G.A. Mamedova, L.A. Zeinalova, R.T. Melikova, Open Education, **6**, 41–48 (2017). https://doi.org/10.21686/1818-4243-2017-6-41-48