BLOG@CACM

Why Is It Hard to Define Data Science?

By Koby Mike and Orit Hazzan
December 2, 2022

Comments (1)

VIEW AS: SHARE:

If you ask a group of data scientists what data science is, you would probably hear different definitions. Indeed, although many attempts have been made to define data science, such a definition has not yet been reached. One reason for the difficulty to reach a single, consensus definition for data science is its multifaceted nature: it can be described as a science, as a research paradigm, as a research method, as a discipline, as a workflow, and as a profession. One single definition just cannot capture the diverse essence of data science. In this blog we attempt to present the essence of each of these perspectives.

Data science as a science

Empirical science has been always about data. Kepler used data about the movement of the planets collected by Tycho Brahe to prove Copernicus' theory of the solar system. Was Kepler the first data scientist when he looked for patterns and models in raw data? While Kepler used data to achieve insights, data science today is more than an empirical science. That is, *data science views the data itself as a natural resource and deals with methods for extracting value out of this data* (Simberloff et al., 2005). While science focuses both on understanding the world and on developing tools and methods to perform research, data science focuses on understanding *data* and developing tools and methods to perform research on *data* (Skiena, 2017).

Data science as a research paradigm

Data science also introduces a new scientific paradigm. The first scientific paradigm, established thousands of years ago, is *empirical science*, in which scientists describe natural phenomena. The second scientific paradigm, applied hundreds of years ago, is the *theoretical paradigm*, in which scientists build models of nature. The third scientific paradigm was introduced only several decades ago and is the *computational paradigm*, in which scientists simulate complex phenomena using algorithms and computers. The fourth scientific paradigm, according to Gray (2007), is *data exploration*, in which data is captured or simulated, and then analyzed by scientists to infer new scientific knowledge. Following Gray, the National Institute of Standards and Technology (NIST) claimed that data science is the current evolution of the fourth paradigm and described data science as *"the conduct of data analysis as an empirical science, learning directly from data itself. This can take the form of collecting data followed by open-ended analysis without preconceived hypothesis (sometimes referred to as discovery or data exploration)"* (Chang et al., 2015, p. 7). In fact, this perspective views data science as the application of the grounded theory paradigm (Glaser & Strauss, 1967) for quantitative research.

Data science as a research method

Data science integrates research tools and methods taken from statistics and computer science that can be used to conduct research in various application domains, such as social science and digital humanities.

Drug discovery is one area that illustrates how machine learning is applied as a research method that

SIGN IN for Full Access

User Name

Password

» Forgot Password?
» Create an ACM Web Account

SIGN IN

MORE NEWS & OPINIONS

States' Push to Protect Kids Online Could Remake the Internet
The New York Times

Where Is the Research on Cryptographic Transition and Agility?
David Ott, Kenny Paterson, Dennis Moreau

The Urgency of the Technological, Geopolitical, and Sustainability Races
Marc Duranton

includes "*target validation, identification of prognostic biomarkers and analysis of digital pathology data in clinical trials*." (Vamathevan et al., 2019, p. 1).

Another example is the application of machine learning methods in social science research. In such research, complex human-generated data, such as posts on social networks, are used to map social phenomena. Grimmer et al. (2021) reviewed current use of machine learning in social science research and stated that *"inclusion of machine learning in the social sciences requires us to rethink not only applications of machine learning methods but also best practices in the social sciences… [machine learning] is used to discover new concepts, measure the prevalence of those concepts, assess causal effects, and make predictions. The abundance of data and resources facilitates the move away from a deductive social science to a more sequential, interactive, and ultimately inductive approach to inference."* (p. 1). As can be seen, data science is viewed in this case as a research method that transforms the research process from deductive to inductive, in line with the perspective on data science as a research paradigm presented above.

Data science as a discipline

Data science integrates knowledge and skills from several disciplines, namely computer science, mathematics, statistics, and an application domain. One way to present such a relationship is using a Venn diagram. Conway (2010) was the first to propose a Venn diagram for data science as a discipline;, many other Venn diagrams were proposed for the discipline of data science following Conway (Taylor, 2016). Figure 1 shows our Venn diagram for data science.

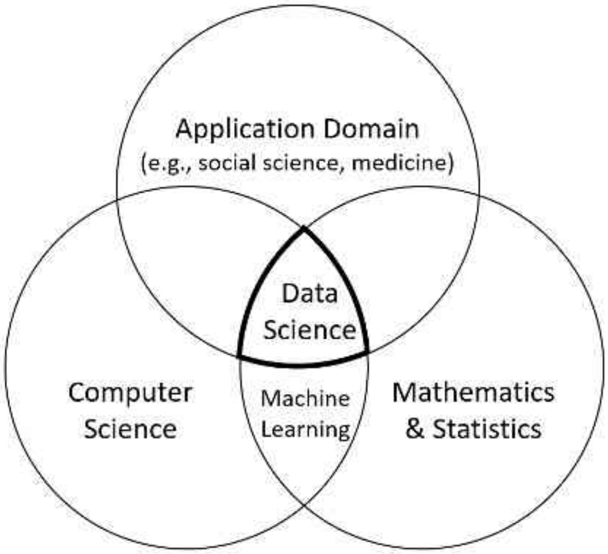

Figure 1: The data science Venn diagram (the authors' version)

Researchers recognize three levels of integration between two or more distinct disciplines: *multidisciplinarity, interdisciplinarity, and transdisciplinarity* (Alvargonzález, 2011). *Multidisciplinarity* is the lowest level of integration. In multidisciplinary education, learners are expected to gain knowledge and understanding in each discipline separately. *Interdisciplinarity* represents a higher level of integration than multidisciplinarity. In interdisciplinary education, after learners gain basic knowledge and understanding in each discipline separately, they are expected to understand the interconnections between the disciplines and to be able to solve problems that require applying different knowledge and methods from each discipline. In *transdisciplinarity,* boundaries among several disciplines transcend to create a holistic approach.

The gradual transition of data science from multidisciplinarity to interdisciplinarity introduces challenges concerning the definition of data science as a discipline that encompasses the different bodies of knowledge, traditions, and cultures that originate in the different fields. Data science may eventually become a transdisciplinary domain.

Data science as a workflow

The 2015 National Science Foundation (NSF) report, summarizing the NSF-sponsored workshop on data science education, introduced a definition of data science that reflects the perspective of data science as a workflow: *"Data science is a process, including all aspects of gathering, cleaning, organizing, analyzing, interpreting, and visualizing the facts represented by the raw data"* (Cassel & Topi, 2015, p. iii). Data science is indeed commonly presented as an iterative workflow for generating value and data-driven actions from data (see Figure 2).

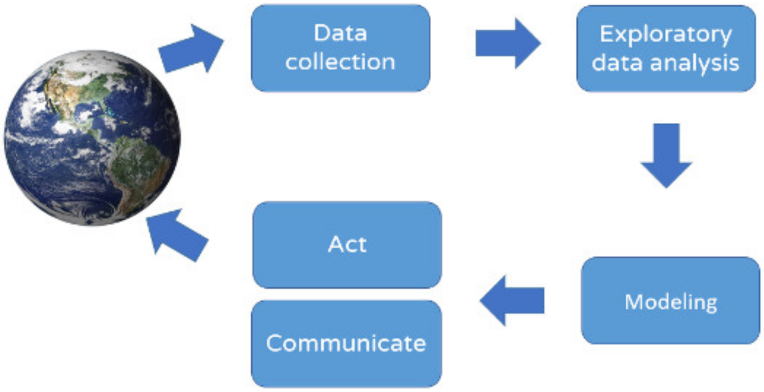

Figure 2. Data science workflow (the authors' version)[1]

Data science as a profession

Irizarry (2020) proposed that the term *data science* was coined in order to improve communication between human resource recruiters in the industry and work applicants. According to Irizarry, "*As the demand in* [sic] *employees capable of completing data-driven projects increased, the term data scientist quickly became particularly prominent because it helped recruiters specify the type of employee they wanted."* (Irizarry, 2020, Section 1, para 4).

Accordingly, the definition of data science can be derived from the description of the profession of data scientist. And thus, for example, in their paper "Data scientist: The sexiest job of the 21st century", Davenport and Patil (2012) describe the profession of data scientist as *"a high-ranking professional with the training and curiosity to make discoveries in the world of big data… More than anything, what data scientists do is make discoveries while swimming in data. It's their preferred method of navigating the world around them"* (ibid, A New Breed, para 1).

Conclusion

Data science is emerging fast and a single definition of it has not yet been reached. In this blog, we presented six facets of data science, each highlighting a different perspective of the field. This multifaceted nature of data science may partially explain the difficulties associated with the efforts to define it.

References

Alvargonzález, D. (2011). Multidisciplinarity, interdisciplinarity, transdisciplinarity, and the sciences. *International Studies in the Philosophy of Science*, *25*(4), 387–403. https://doi.org/10.1080/02698595.2011.623366

Cassel, B., & Topi, H. (2015). *Strengthening data science education through collaboration: Workshop report 7-27-2016*. Arlington, VA.

Chang, W. L., Grady, N., & others. (2015). *Nist big data interoperability framework: Volume 1, big data definitions*.

Conway, D. (2010). The data science venn diagram. *Datist*. http://www.dataists.com/2010/09/the-data-science-venn-diagram/

Davenport, T. H., & Patil, D. (2012). Data scientist: The sexiest job of the 21st century. *Harvard Business Review*, *90*(5), 70–76.

Glaser, B.G., & Strauss, A. L. (1967). The Discovery of Grounded Theory: Strategies for Qualitative Research. New York: Aldine de Gruyter.

Gray, J. (2007). *EScience – A transformed scientific method*. http://research.microsoft.com/en-us/um/people/gray/talks/NRC-CSTB_eScience.ppt

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021). Machine learning for social science: An agnostic approach. *Annual Review of Political Science*, *24*, 395–419.

Irizarry, R. A. (2020). *The role of academia in data science education*.

Simberloff, D., Barish, B., Droegemeier, K., Etter, D., Fedoroff, N., Ford, K., Lanzerotti, L., Leshner, A., Lubchenco, J., Rossmann, M., & others. (2005). Long-lived digital data collections: Enabling research and education in the 21st century. *National Science Foundation*.

Skiena, S. S. (2017). *The data science design manual*. Springer.

Taylor, D. (2016). Battle of the Data Science Venn Diagrams. *KDnuggets*. https://www.kdnuggets.com/battle-of-the-data-science-venn-diagrams.html/

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., & others. (2019). Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, *18*(6), 463–477.

[1] The image of Earth was originally posted to Flickr by DonkeyHotey at https://flickr.com/photos/47422005@N04/5679642883. It was reviewed on 4 December 2020 by FlickreviewR 2 and was confirmed to be licensed under the terms of the cc-by-2.0.

Koby Mike *is a Ph.D. graduate from the Technion's Department of Education in Science and Technology under the supervision of Professor Orit Hazzan. He is currently a post-doc at Bar-Ilan University.* ***Orit Hazzan*** *is a professor at the Technion's Department of Education in Science and Technology. Her research focuses on computer science, software engineering, and data science education. For additional details, see https://orithazzan.net.technion.ac.il/.*

Comments

Ian Dinwoodie

February 06, 2023 03:58

In colleges, Data Science is often designed to update obsolete Statistics curricula, which tend to be old calculus methods on random samples, and so a new name is required.

Displaying **1** comment