Harvard Data Science Review • 1.1

The Data Life Cycle

Jeannette M. Wing

Published on: Jun 23, 2019

Updated on: Oct 04, 2019

DOI: [10.1162/99608f92.e26845b4](https://doi.org/10.1162/99608f92.e26845b4)

ABSTRACT

To put data science in context, we present phases of the data life cycle, from data generation to data interpretation. These phases transform raw bits into value for the end user. Data science is thus much more than data analysis, e.g., using techniques from machine learning and statistics; extracting this value takes a lot of work, before and after data analysis. Moreover, data privacy and data ethics need to be considered at each phase of the life cycle.

Keywords

analysis, collection, data life cycle, ethics, generation, interpretation, management, privacy, storage, story-telling, visualization

Data science is the study of extracting value from data. “Value” is subject to the interpretation by the end user and “extracting” represents the work done in all phases of the *data life cycle* (see Figure 1).¹

© Jeannette M. Wing 2017

Figure 1. The Data Life Cycle

The cycle starts with the *generation* of data. People generate data: every search query we perform, link we click, movie we watch, book we read, picture we take, message we send, and place we go contribute to the massive digital footprint we each generate. Walmart collects 2.5 petabytes of unstructured data from 1 million customers every hour ([DeZyre, 2015](#)). Sensors generate data: more and more sensors monitor the health of our physical infrastructure, e.g., bridges, tunnels, and buildings; provide ways to be energy efficient, e.g., automatic lighting and temperature control in our

rooms at work and at home; and ensure safety on our roads and in public spaces, e.g., video cameras used for traffic control and for security protection. As the promise of the Internet of Things plays out, we will have more and more sensors generating more and more data. At the other extreme from small, cheap sensors, we also have large, expensive, one-of-a-kind scientific instruments, which also generate unfathomable amounts of data. The latest round of the Intergovernmental Panel on Climate Change (IPCC) will produce up to 80 petabytes of data ([Balaji et al., 2018](#)). The Large Synoptic Survey Telescope is expected to build over a period of 10 years a 500 petabyte database of images and a 15 petabyte catalog of text data ([LSST Project Office, 2018](#)). The total amount of Large Hadron Collider data already collected is close to one exabyte ([Albrecht et al., 2019](#)).

After generation comes *collection*. Not all data generated is collected, perhaps out of choice because we do not need or want to, or for practical reasons because the data streams in faster than we can process. Consider how data are sent from expensive scientific instruments, such as the IceCube Neutrino Detector at the South Pole. Since there are only five polar-orbiting satellites, there are only certain windows of opportunities to transmit restricted amounts of data from the ground to the air ([IceCube South Pole Neutrino Observatory, 2019](#)). Suppose we drop data between the generation and collection stages: could we possibly miss the very event we are trying to detect? Deciding what to collect defines a filter on the data we generate.

After collection comes *processing*. Here we mean everything from data cleaning, data wrangling, and data formatting to data compression, for efficient storage, and data encryption, for secure storage.

After processing comes *storage*. Here the bits are laid down in memory. Today we think of storage in terms of magnetic tape and hard disk drives, but in the future, especially for long-term, infrequently accessed storage, we will see novel uses of optical technology ([Anderson et al., 2018](#)) and even DNA storage devices ([Bornholt et al., 2016](#)).

After storage comes *management*. We are careful to store our data in ways both to optimize expected access patterns and to provide as much generality as possible. Decades of work in database systems have led us to optimal systems for managing relational databases, but the kinds of data we generate are not always a good fit for such systems. We now have structured and unstructured data, data of many types (e.g., text, audio, image, video), and data that arrive at different velocities. We need to create and use different kinds of metadata for these dimensions of heterogeneity to maximize our ability to access and modify the data for subsequent analysis.

Now comes *analysis*. When most people think of what data science is, what they mean is data analysis. Here, we include all the computational and statistical techniques for analyzing data for some purpose: the algorithms and methods that underlie artificial intelligence (AI), data mining, machine learning,² and statistical inference, be they to gain knowledge or insights, build classifiers and predictors, or

infer causality. For sure, data analysis is at the heart of data science. Large amounts of data power today's machine learning algorithms. The recent successes of the application of deep learning to different domains, from image and language understanding to programming ([Devlin et al., 2017](#)) to astronomy ([Gupta, Manuel, Matilla, Hsu, & Haiman, 2018](#)) are astonishing.

Beyond analysis, data *visualization* helps present results in a clear and simple way that a human can readily understand and visualize. Here a picture is worth not a thousand words (that comes later) but a thousand petabytes! It is at this stage in the data life cycle when we need to consider, along with functionality, aesthetics, and human visual perception to convey the results of data analysis.

Also, it is not enough just to show a pie chart or bar graph. By *interpretation*, we provide the human reader an explanation of what the picture means. We tell a story explaining the picture's context, point, implications, and possible ramifications.

Finally, in the end, we have the human. The human could be a scientist, who, through data, makes a new discovery. The human could be a policymaker who needs to make a decision about a local community's future. The human could be in medicine, treating a patient; in finance, investing client money; in law, regulating processes and organizations; or in business, making processes more efficient and more reliable to serve customers better.

The diagram omits the arrows that show the many feedback loops in the data life cycle. Inevitably, after we present some observations to the user based on data we generated, the user asks new questions and these questions require collecting more data or doing more analysis.

Underlining this diagram is the importance of using data responsibly at each phase in the cycle. We must remember to consider privacy and ethical concerns throughout, from privacy-preserving collection of data about individuals to ethical decisions that humans or machines will need to make based on automated data analysis. The importance of these concerns cannot be overstated. Indeed, it is an opportunity for ethicists, humanists, social scientists, and philosophers to join forces with the technologists and together define the field of data science. Just as business, law, journalism, and medicine provide ethical training for their students, so must we in data science.

References

Albrecht, J., Alves, A., Amadio, G., Andronico, G., Anh-Ky, N., Aphecetche, L., ..., Yazgan, E. (2019). A Roadmap for HEP Software and Computing R&D for the 2020s, The HEP Software Foundation, *Computer Software for Big Science*, 3(7) Retrieved from <https://arxiv.org/abs/1712.06982>

- Anderson, P., Black, R., Cerkauskaitė, A., [Chatzieleftheriou, A.](#), [Clegg, J.](#), Daint, C., ..., Wang L. (2018). Glass: A New Media for a New Era?, *Proceedings of the 10th USENIX Workshop on Hot Topics in Storage and File Systems (HotStorage 18)*. Retrieved from <https://www.microsoft.com/en-us/research/uploads/prod/2018/07/hotstorage18-paper-anderson.pdf>
- Balaji, V., Taylor, K.E., Jukes, M. Lawrence, B.N., Durack, P.J., Lautenschlager, M., ..., Williams, D. (2018). Requirements for a global data infrastructure in support of CMIP6, *Geoscientific Model Development*, 11(9) 11, 3659-3680. Retrieved from <https://www.geosci-model-dev.net/11/3659/2018/>
- Bornholt J., Lopez, R., Carmea, D.M., Ceze, L., Seelig, G., & Strauss K. (2016). A DNA-Based Archival Storage System, *Proceedings of the International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*. Retrieved from <https://homes.cs.washington.edu/~luisceze/publications/dnastorage-asplos16.pdf>
- Devlin, J., Uesato J., Bhupatiraju, S., Singh, R., Mohamed, A., & Kohli, P. (2017). RobustFill: Neural Program Learning under Noisy I/O. *Proceedings of the 34th International Conference on Machine Learning*. <https://arxiv.org/abs/1703.07469>
- DeZyre (2015). How Big Data Analysis helped increase Walmarts Sales turnover?. Retrieved from <https://www.dezyre.com/article/how-big-data-analysis-helped-increase-walmarts-sales-turnover/109>
- Gupta, A., Manuel, J., Matilla, Z., Hsu, D., & Haiman, Z. (2018). Non-Gaussian information from weak lensing data via deep learning. *Physical Review D*. Retrieved from <https://arxiv.org/abs/1802.01212>
- IceCube South Pole Neutrino Observatory (2019). Data Movement. Retrieved from <https://icecube.wisc.edu/science/data/datamovement>
- Jordan, M. (2019). Artificial intelligence—the revolution hasn’t happened yet, *Harvard Data Science Review*, volume 0. Retrieved from <https://hdsr.mitpress.mit.edu/pub/wot7mkc1>
- LSST Project Office (2018). LSST and Big Data, Fact Sheets. Retrieved from <https://docushare.lsst.org/docushare/dsweb/Get/Document-14554>
- Wing, J.M. (2018). The Data Life Cycle, Data Science Institute, Columbia University. Retrieved from <https://datascience.columbia.edu/data-life-cycle>
- Wing, J.M., Janeia, V.P., Kloefkorn, T., & Erickson, L.C. (2018). Data Science Leadership Summit, Workshop Report, National Science Foundation. Retrieved from <https://dl.acm.org/citation.cfm?id=3293458>

This article is © 2019 by Jeannette M. Wing. The article is licensed under a Creative Commons Attribution (CC BY 4.0) International license (<https://creativecommons.org/licenses/by/4.0/legalcode>), except where otherwise indicated with respect to particular material included in the article. The article should be attributed to the author identified above.

Footnotes

1. The picture and prose are extracted from the blog post “[The Data Life Cycle](#)” (Wing 2018), a variation of which appears in (Wing, Janeia, Kloefkorn, & Erickson 2018). [↵](#)
2. AI, machine learning, and data science are often erroneously confused as synonyms. This article shows that there is more to data science than AI and machine learning; similarly, there is more to AI and machine learning than data science (Jordan, 2019). [↵](#)