The Ubiquity of Data in Science and Society

E. Wes Bethel, Lawrence Berkeley National Laboratory

March 6, 2020

Whereas the past 4 decades seen the rise of the parallel computers along with the tools and technologies to use them for modeling scientific phenomena, evolution in the next 4 decades will be driven by data: in how it is used, the tools for using it, the proliferation of new sources of data, and in the critical infrastructure for managing and enabling its lifecycle.

New uses of data. Rather than being a product of computation, data instead becomes the necessary input for new uses. The setup and design of a complex experiment can done more quickly and accurately using information about previously run similar experiments in conjunction with advanced AI-based methods. The real-time control and tuning of science experiments will require use of high throughput AI-based methods that understand the science objective and can operate quickly enough to "stay one step ahead of" the experiment in progress. Both these types of new use, which entail AI-based methods, require high-quality curated scientific data. Beyond experiment setup and optimization, data plays a key role in AI itself: without data, there are no supervised learning methods, and the quality and accuracy of such methods depends to a large degree on the quality and quantity of training data. In the future, data increasingly becomes an *asset* for science, for education, for economics, for health care, and for security.

New tools for working with data. While it is understood that the years ahead will see a rise in AI-based methods for generating models from data that are in turn used in a predictive capacity, there are a host of other types of tools that are equally important. From a practical perspective, the tools for simply being able to find data will undergo a dramatic shift, away from text-based searches to those that include more dataset introspection and advanced summarization methods that lend themselves to search. Access control mechanisms will evolve to better serve both producers and consumers of data. Methods that help to automate and improve throughput of generating metadata and curating data are required to accommodate growing data sizes.

New sources of data. The DOE SC science user facilities, already on track to generate exabytes/year of data in the near future, represent a significant challenge. Additional sources of data include distributed sensors, "Internet of Things", which will span all segments of life: economics, transportation, epidemiological/health, environmental controls, and more.

Data infrastructure and the data lifecycle. The data lifecycle is complex and profound, going from data creation to destruction. The topic of data lifecycle will become a primary focal point for the Office of Science over the next 4 decades, including issues like what infrastructure is needed, and how it, and the data that inhabits it, should be managed FAIRly [1].

[1] M. Wilkinson, et al., The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci Data* 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18