Team Data Science to Bring Data, Analysis and Domain Expertise together

John Wu, Suren Byna, Junmin Gu, Alex Sim, Houjun Tang, Vincent Dumont and Mariam Kiran

DOE research community is known for bringing in large interdisciplinary teams together to solve mission-critical problems. This approach is usually called team science. In the following decades, we foresee a similar approach is needed to organize the compute and data technologies together to better support missions. This would seamlessly bring together DOE's expertise in high-performance computing, networking, artificial intelligence, data management, and collaboration technologies, to satisfy the critical needs in physicists, biologists, chemists, geologists, climate scientists, and so on.

Some of the DOE science communities have built large collaborations in the past few decades to address their own challenges. For example, the high-energy physics community has established a large data analysis collaboration with multi-tier data storage and analysis platforms. Fusion community has built large collaborations around its experimental devices with advanced remote operations and analysis capabilities. These large-scale collaborations are critical to create and curate massive data collections for these science areas, and also provide a common analysis environment for individual scientists who might not otherwise have access to sufficient computing resources to work with the large volumes of data. Significant new scientific discoveries are generated from these collaborations. With the advances in sensing technologies, smart devices, and wider availability of FAIR (findable, accessible, interoperable, and reusable) data, we anticipate it necessary to replicate these successful large collaborations for other scientific research areas.

There are a number of factors driving the change in data from mission-critical DOE projects. For example, sensors are generating an unprecedented amount of data, these data producers are being distributed to different corners of the society, ever-more decisions are being automated with intelligent control systems, and smart devices are connecting to each other forming autonomous systems of systems. Some of such scenarios are of commercial interest, while other scenarios in science and society would require concerted research support from the government and other funding agencies. For example, the automotive industry might be interested in self-driving cars, but might not be interested in how the cars could coordinate with road systems, traffic signals, or charging stations to minimize travel time and reduce the load on the electric power grid. DOE is involved in a large set of fundamental research efforts and is in charge of monitoring the nation's critical infrastructures in energy, water, and environment, much of these efforts would benefit from a collaborative team science approach we call *team data science*.

We envision the team data science as an extension of the team science approach to bring expertise and resources from diverse fields and to build systems to enable individual scientist to conduct their research as if he/she is supported by a large team of experts. It would require work in developing collaboration software, policy for sharing resources across many organizational boundaries, and tools to seamlessly integrate a variety of cyber and physical systems. This team approach to data science could revolutionize the data management and analysis, and delivering self-driving capability to DOE scientific user facilities.