Federated Data-Science for Accelerated Fusion Research: Why do we need it and how do we achieve it?

CS Chang
Princeton Plasma Physics Laboratory
cschang@pppl.gov

In this talk, I will present a motivation for federated data-science for accelerated fusion research and a high-level view of the necessary computer science, applied mathematics, and physics tools.

Complex multi-physics phenomena govern the tokamak plasma dynamics, which interact with each other nonlinearly and nonlocally in a wide-range of space-time multiscale. Collaboration among experts who are geographically separated all over the world is necessary for accelerated development of commercial fusion reactors. In today's tokamaks, hundreds or thousands of diagnostic sensors are employed to yield localized information on the multiscale information in their own part of the space-time. All the sensor data have their own value, variety and veracity, and are to be integrated together with proper mathematical and computational tools in understanding the multi-scale plasma dynamics. Number of variables associated with the individual sensor signals is simply too large for a manual integration unless a sophisticated data science is employed, which include machine learning, neural network, and artificial intelligence.

Moreover, the volume and the velocity of the data production from these sensors are too high for an efficient near real-time international collaboration. Data-science based reduction tools, fast network, and an automated framework is needed for the collaborative integration, analyses and discoveries of the multiscale physics, which could improve next experiments and, thus, help to achieve an accelerated progress.

Execution of the data science at both the experimental and analysis sites require hierarchical computers – from leadership class HPCs for predictive simulations, to distributed resources for near real-time analysis that can lead to experimental steering. Large-scale numerical studies simulating the experiments also produce a large number of synthetic diagnostic data, which need to be orchestrated together with data from the large number of experimental diagnostic instruments in the automated framework.

The problem will grow to be a much bigger issue for ITER and future reactors. More advanced applied mathematics and computer science technologies need to be developed for a centralized federation of the ITER physics collaboration in near real-time; which include predicting and allocating the available resources dynamically, executing a federation workflow, using the state of the art machine learning and AI technologies, applying the streaming data analysis techniques, indexing and reducing the streaming ITER and simulation data, capturing the provenance, and quantifying the uncertainties.

An example for such a federated data-science activity will be introduced, which is just being launched for a collaborative KSTAR tokamak research between NFRI and my PPPL SciDAC team; via the data handling and computational capability of NERSC, the data-science technology of ORNL, and the network service provided by ESnet and KISTI.