From Federated Data to Digital Twins: Computational Infrastructure for Fusion Data Science

Fusion data science is increasingly bottlenecked not by models but by data access and context: experimental data is scattered across machines, formats, and archives, and researchers can spend weeks retrieving and preparing it before any analysis begins. This talk describes the computational and machine-learning infrastructure our team is building to remove that bottleneck. The Fusion Data Platform (FDP) provides federated, secure access to experimental data from multiple tokamaks, retrievable from a laptop or an HPC center alike, together with scalable parallel retrieval and mapping to the IMAS schema for cross-machine analysis. Building on FDP, we are developing an agentic layer that makes fusion data directly consumable by AI agents, backed by the Fusion Energy Data Ecosystem and Repository (FEDER) knowledge graph: a semantic layer that links diagnostics, physics quantities, signals, and provenance to give agents the context they need to discover and retrieve the right data. I will show how these capabilities connect to the DIII-D digital twin effort and outline where a broader, federated fusion data ecosystem is heading.