Driving Data Management for Science Using the 20 Questions Approach

Deb Agarwal, Lawrence Berkeley National Laboratory

In science, data analysis capabilities are rapidly emerging as the fourth paradigm along with experiment, theory, and simulation (http://research.microsoft.com/en-us/collaboration/fourthparadigm/). Scientific data has become ubiquitous and 'big' with cheap sensors deployed everywhere, large numbers of satellites measuring the earth, genome sequencing in bulk, ultra-high resolution instruments and cameras, high fidelity simulations, and cheap storage. These data provide an unprecedented opportunity to improve models and connect and study phenomena across disciplines and areas. To achieve this vision, we need to move beyond collecting the data, to enabling broad usage of data within and across disciplines. Jim Gray's 20 questions provided a valuable methodology for interacting with astrophysics users to understand how they want to use data. In this approach, the user provides the top 20 simple questions they want to be able to ask once the data is available. These questions then help to drive the data organization and interfaces. We have been working with a wide array of science disciplines and have been evolving a modified version of the 20 questions approach to engage with these communities. This talk will discuss our experiences in enabling collaborative science on data from instruments, sensors, and models and some of the remaining challenges still to be solved.