The Evolving Data Lifecycle:
Implications for Cyberinfrastructure and Domain Sciences

Dr. Rahul Ramachandran

The ever increasing amounts of information generated by sensors, model simulations, analyses, and other sources have led to an abundance of scientific data. In order to process, manage, and extract useful information from this deluge of data, the development of new tools and algorithms has become necessary. This data deluge in the sciences has also shifted the focus of computing applications from simulations to a new paradigm called data intensive science.

Data intensive science presents new opportunities in both research and education. One such opportunity exists in the need for the construction of a cyberinfrastructure consisting of hardware, networks, software, and middleware to support the advanced data acquisition, storage, management, integration, analytics, and visualization needed to expand the data-centered paradigm in science. As these new technological capabilities become integrated into a larger cyberinfrastructure ecosystem, they can tremendously accelerate scientific research. A new curriculum is now needed in data science for educating a new breed of researchers who can integrate both computer science and domain science knowledge.

A scientific research process can be represented as a data lifecycle consisting of a series of stages through which a piece of data passes during its lifetime. These stages include data processing, discovery, archiving, and finally use, which by itself encompasses access, integration, visualization and analysis, and sharing.

The sheer magnitude and complexity of data now available allows access to measurements spanning multiple scales of length, time, and disciplines. As a result, the data lifecycle itself is evolving in its complexity, requiring different stages than those outlined above and creating new needs in existing stages. Each stage within this expanded lifecycle presents new challenges and opportunities that require new computational solutions.

Specific examples of ongoing research from several stages of the data lifecycle will be presented in this talk.