Next-Generation Architectures for Knowledge Centers to Advance DOE's Climate and Environmental Sciences

William D. Collins (wdcollins@lbl.gov)
Director, Climate and Ecological Sciences Division;
Lawrence Berkeley National Laboratory
Berkeley, CA, USA

Traditional data repositories have served as a tremendous resource for scientists, both by providing equal access to investigators from across the global community while safe-guarding and documenting the properties and provenance of the data for decades to come.

However, the literally exponential growth in data being collected autonomously and generated from supercomputer models poses a serious challenge to the utility of these repositories. The amount of digital data doubles every two years, with volumes expected to reach 163 zetabytes by 2025. NASA alone expects its EOSDIS archive to grow by a factor of 10, to 250 PB. Simply storing this torrent of information along with its metadata risks relegating much of it to "write once, read never" status.

To meet this growth, it would be prudent to re-architect traditional data repositories as "Knowledge Centers", with several important new, or currently underutilized, capabilities, including:

Open-source sharing of methodologies: Several existing data systems including KBase are building libraries of user-contributed analytical methods. This is very much in the spirit of the open-source contributions to R, python, etc. but devoted to the analysis of the KBase's biological data. New Knowledge Centers should build upon this highly successful model coupled to AI assistants that can help users make maximum use of the growing software knowledge base.

Integral uncertainty quantification capabilities: One of the central challenges in synthesizing disparate observations is to construct an accurate and comprehensive estimate of uncertainty to the resulting synthetic data set. Inference and extrapolation from data also require detailed knowledge of the underlying uncertainties. The Knowledge Centers should build on DOE's extensive expertise in UQ to automate robust uncertainty quantification and its utilization in hypothesis testing to the greatest extent possible.

Integration and execution of process models as a key component of uncertainty reduction: Typically process models (models based on first principles that simulate the real world at the limit of our ability to observe it) are run after, not while, observations are collected. When the models and data disagree, the attribution of the error to underlying cause is a major conundrum, i.e., is the error due to externalities (insufficient information on initial and boundary conditions), parametric error in the model, or, more seriously, structural error in the model? Having these models linked with much faster emulators of them (constructed using ML) for queries while experiments are underway would be very useful.

Containerized software, to ensure portability and extensibility: The underlying code for the next-generation Knowledge Centers should be built in a containerized framework, so that federations of sister Centers can be readily constructed on geographically distributed servers or in the cloud. Software frameworks to enable compute-hardware portability should be extended to include storage-hardware portability as, e.g., the era of rotating disk media comes to an end.

Real-time support of field observatories: These centers should be equipped with APIs that can be queries by networks of autonomous sensors, field experiment control centers, and other systems collecting and managing observing systems. This would enable DOE's observatories to determine (a. how best to deploy mobile observing systems, based on experience under similar environmental conditions with the same or analogous systems; and (b. how to maximize the progress contributed by new data toward a field experiment's scientific objectives based upon the data collected so far.

AI assistants for data interrogation: It seems increasingly likely that the exponentiating volume of metadata itself may overwhelm human analysts – keyword searches that turn up millions of hits, for example. We will need a mechanism to generate "meta-meta-data". I.e., smart speakers for work, not just for home.

Next-Generation Architectures for Knowledge Centers to Advance DOE's Climate and Environmental Sciences