A data facility to generate actionable knowledge for DOE’s Climate and Environmental Sciences

Charuleka Varadharajan (cvaradharajan@lbl.gov)
Earth and Environmental Sciences Area;
Deb Agarwal (daagarwal@lbl.gov)
Computational Research Division
Lawrence Berkeley National Laboratory
Berkeley, CA, USA

There has been a paradigm shift in how we observe our natural environment. Unprecedented amounts of diverse data are being generated through the use of sensor networks, remote sensing and other imaging capabilities (e.g. drones), genome sequencing, and model simulations. However our ability to analyze and use this data in predictive models is still limited. Scientists need to be able to extract information from data that are highly diverse in spatial/temporal scales, scientific disciplines, uncertainty, and in their structure and formats. There is an opportunity to broadly enable knowledge generation for the Earth and Environmental sciences through advanced, easy-to-use computational infrastructure, for example, deep learning and data analytics; sophisticated time series and spatial analyses of diverse datasets; visualization of complex multi-spatial and temporal scale data; and model uncertainty, sensitivity, parameter estimation (Fig 1).

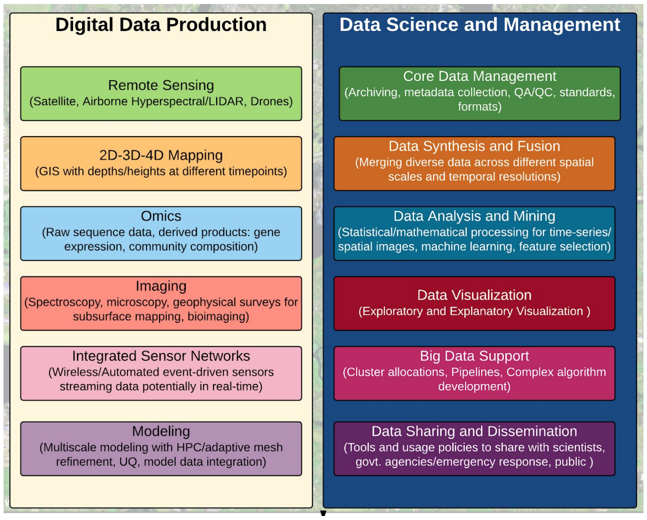
Digital Data Production
Remote Sensing
(Satellite, Airborne Hyperspectral/LIDAR, Drones)
2D-3D-4D Mapping
(GIS with depths/heights at different timepoints)
Omics
(Raw sequence data, derived products: gene expression, community composition)
Imaging
(Spectroscopy, microscopy, geophysical surveys for subsurface mapping, bioimaging)
Integrated Sensor Networks
(Wireless/Automated event-driven sensors streaming data potentially in real-time)
Modeling
(Multiscale modeling with HPC/adaptive mesh refinement, UQ, model data integration)
Data Science and Management
Core Data Management
(Archiving, metadata collection, QA/QC, standards, formats)
Data Synthesis and Fusion
(Merging diverse data across different spatial scales and temporal resolutions)
Data Analysis and Mining
(Statistical/mathematical processing for time-series/ spatial images, machine learning, feature selection)
Data Visualization
(Exploratory and Explanatory Visualization)
Big Data Support
(Cluster allocations, Pipelines, Complex algorithm development)
Data Sharing and Dissemination
(Tools and usage policies to share with scientists, govt. agencies/emergency response, public)

Figure 1: **Data generation capabilities (left panel) in the Earth and Environmental sciences have exploded over the past decade. There is an urgent need to develop computational capabilities that enable scientific users to easily utilize the data for predictive analytics (right panel).**

Centralized or federated infrastructure that supports capabilities for storing, analyzing, and integrating data with models can provide an economy of scale, and accelerate the generation of scientific insights and predictive capacity. Such a facility should support multi-scale, multi-disciplinary data and contain tools to enable scientific knowledge generation including:

- A data repository to store and distribute increasingly large and heterogeneous data with fast access mechanisms and open data licenses.
- Data integration tools that connect and synthesize distributed datasets across data systems (e.g., ESS-DIVE, ESGF, ARM, Ameriflux, NASA, USGS) and enables users to easily discover, access, and integrate big, diverse datasets.
- Multiscale data assimilation tools to enable real-time integration of observation data with simulation codes.
- Data analytics and computational capabilities for data mining and deep learning; advanced statistical and information theory algorithms for time series and spatial analyses. This includes core libraries for data preprocessing such as QA/QC, subsetting, gridding.
- Different workflow tools and science gateways.
- A computational framework that enables community development of scripts and app-based tools with analytics engines to enable users to discover, query, subset, process, analyze and store data (similar to or built on existing cloud infrastructure such as Google Cloud Platform, Amazon Web Services Cloud, Microsoft Azure Cloud).
- Interactive visualizations and narrative interfaces built using recent advances in web-based tools to enable data exploration and knowledge discovery.
- A software repository for sharing programs developed by the community (e.g. QA/QC and data processing scripts) for reproducible science offering a limited number of compatible open source licenses.

These requirements will grow in the next 10-30 years with future technologies, such as autonomous sensors and robots connected with 5G networks, new experimental user facilities (BioEPIC), and the development of physics-informed ML models.