BERKELEY LAB
Bringing Science Solutions to the World

U.S. DEPARTMENT OF ENERGY
Office of Science

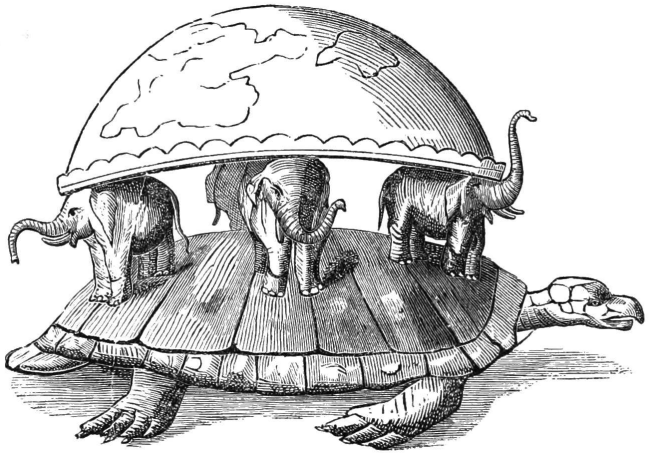
Scientific Data Management and AI/ML

Dan Gunter, LBNL

AI Energy Storage Workshop, 16 April 2024

Overview

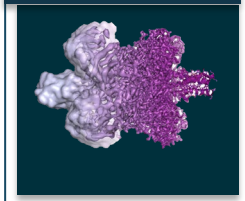
- AI approaches are completely dependent on data
  - Trust in the data determines trust in the result
    - This is the same challenge we have for scientific data analysis in general
- **Data management approaches for scientific data should form the foundation for our data management approaches for AI/ML as well**
- LBNL has had success with scientific data management
  - Considering the entire scientific data lifecycle
  - Understanding and engaging the user community



The Scientific Data Lifecycle™*

*our division / Deb Agarwal *did* do a lot to promote this view within DOE research

Acquire

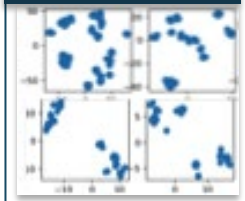


Collect from sensors, experiments, simulations

Transfer

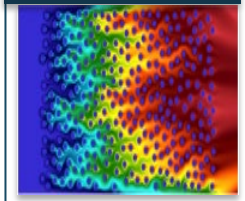
Move from instrument to center

Clean



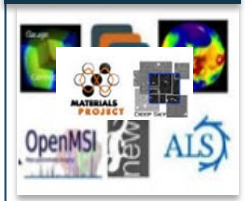
Organize, annotate, filter, encrypt, compress

Use/Reuse



Analyze, mine, model, learn, infer, derive, predict

Publish



Disseminate & aggregate, using portals, databases

Preserve

Index, curate, age, track provenance, purge

These do not necessarily occur in this exact order

For example, planned NCEM* workflow:

*National Center for Electron Microscopy

Acquire → **Clean** → **Transfer** → NERSC: **Clean** → **Use**

Data cleaned/filtered by an edge device

Data cleaned before use at HPC facility

ESS-DIVE*: community archive of environmental data

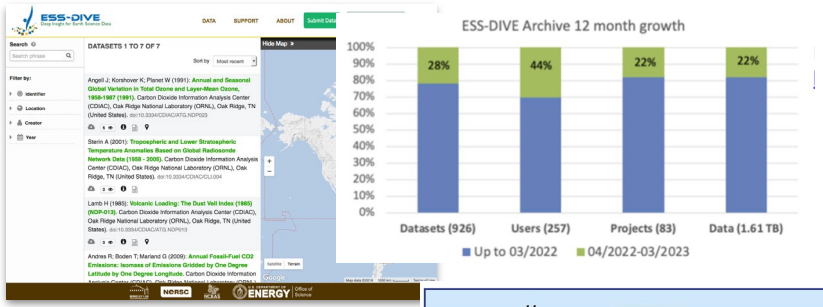
*Environmental System Science Data Infrastructure for a Virtual Ecosystem

► Provides long-term stewardship of the DOE's Environmental System Science data: https://data.ess-dive.lbl.gov through a Findable, Accessible, Interoperable, and Reusable (**FAIR**) web service portal

► Adopt and develop **data standards** through community engagement (eg. Sample IDs)

► **Store** data for long-term archival with **DOIs** for citation along with **searchable metadata** and data **downloads**.

► **Data at Scale** with Globus support, scalable containers, automated APIs, replication on DataONE federation



ESS-DIVE User Portal

Work with users to develop data standards

C. Varadharajan, S. Cholia, C. Snavely, V. Hendrix, C. Procopiou, D. Swantek, W. J. Riley, and D. A. Agarwal (2019), Launching an accessible archive of environmentaldata,*Eos, 100*

AmeriFlux: Building a High-Quality Carbon Flux Dataset for the Americas

Scientific Achievement

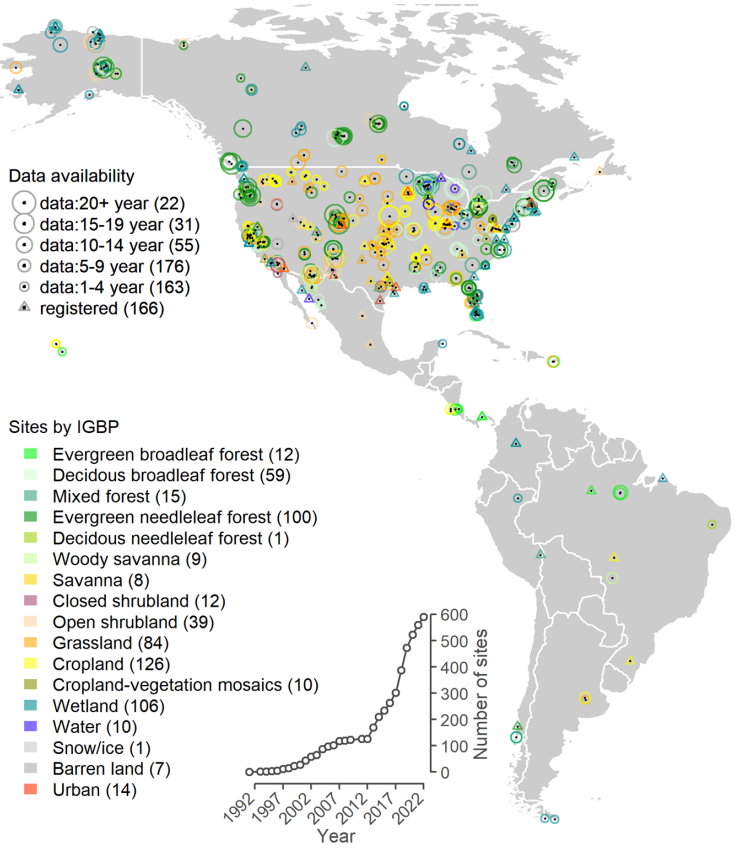
AmeriFlux is a network of PI-managed measurement sites in the Americas measuring ecosystem CO2, water, and energy fluxes to address earth science research. The diversity of each site's characteristics and operations presents multiple challenges, such as data standardization, quality assurance, and sharing. LBNL researchers are developing advanced quality assessment and flux partitioning processing that will scale to a large number of sites in the network. The network has grown from 60 to more than 650 sites (about 15% of sites outside the US).

Significance and Impact

The network has over 18,000 registered users and features two main data products. AmeriFlux BASE is a standardized half-hourly/hourly flux-met data product with more than 30,000 unique downloads. The data product features over 3390 site years and more than 80 sites with decade-long records. AmeriFlux FLUXNET is a gap-filled, partitioned data product that is fully compatible with the FLUXNET2015 data product. FLUXNET2015 has been uniquely downloaded more than 17,000 times, and the AmeriFlux FLUXNET more than 2,200 times.

Technical Approach

- Automated quality and format assessment of data and issue ticket tracking for communication with users.
- ONEFlux* processing, including turbulence filtering, gap-filling, partitioning of CO2 fluxes into ecosystem respiration and gross primary production, and uncertainty estimates. ONEFlux generates the AmeriFlux FLUXNET product.



DOE BER

https://ameriflux.lbl.gov
https://fluxnet.org

AMERIFLUX MANAGEMENT PROJECT

Takeaway: Data Management Challenges are Socio-Technical

AmeriFlux and ESS-DIVE are successful because they strongly engage both data providers and data users in their communities

- "UX" processes are very useful for this: structured understanding of users in their actual context (not some idealized workflow)
  - Helps identify shared "pain points" to prioritize always-limited resources
- Create/use data standards -- but with proper incentives on both sides
- Preservation is.. hard
  - Ignoring it will make it go away, but not in a good way

AI/ML Data Challenges Are (also) Socio-Technical

- Semi-opaque algorithms mining huge, semi-opaque datasets to generate results (that may be fed back into those same datasets)
  - What could possibly go wrong?
- Some Lessons Learned from ESS-DIVE and AmeriFlux*
  - Prioritize community engagement
  - Pay attention to known incentives and pain points
  - Take a user experience approach to building understanding
  - Plan for this to be an *ongoing long-term process*

✗ Build it and they will come

✓ Build and maintain, and adapt, and shepherd it and they will ***stay***

Not atypical script/data workflow

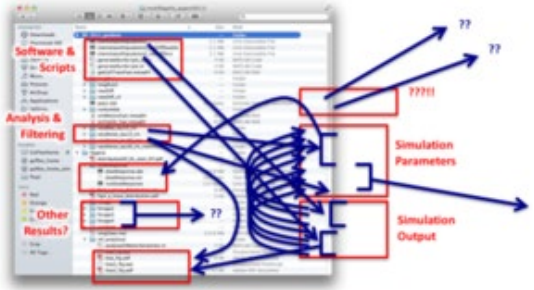
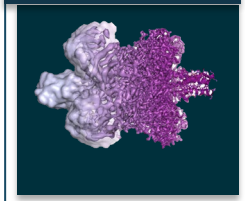
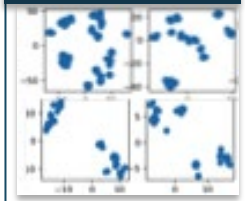
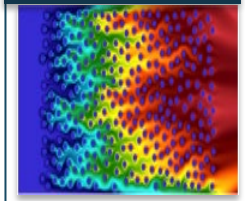
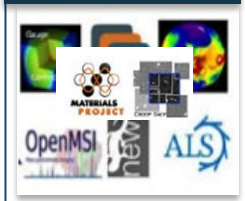


Image courtesy Paramvir Dehal, LBNL

*and other successful projects

Consider the Entire Data Lifecycle

Acquire	Transfer	Clean	Use/Reuse	Publish	Preserve
					
Collect from sensors, experiments, simulations	Move from instrument to center	Organize, annotate, filter, encrypt, compress	Analyze, mine, model, learn, infer, derive, predict	Disseminate & aggregate, using portals, databases	Index, curate, age, track provenance, purge

Users need a lot of help here!

Tendency to focus here

Community benefits

- LBNL has deep experience on technical and socio-technical challenges of scientific data management
- This applies equally to AI/ML analyses and innovations, which are *completely dependent* on the quality of the data on which they are based

Thank You

Questions?