



**BERKELEY LAB**

Bringing Science Solutions to the World



Office of Science

# Applied Research for Collaborative Science

An overview of the Usable Data Systems, Integrated Data Systems, and Sustainable Software Engineering groups

Dan Gunter (UDS group lead) - 06/13/2023



# Table of contents

## 01 Overview

About

What we do

Engagement approach

## 02 Projects

**Domain Science**

- Earth Sciences
- Biology
- Physics
- Process Engineering

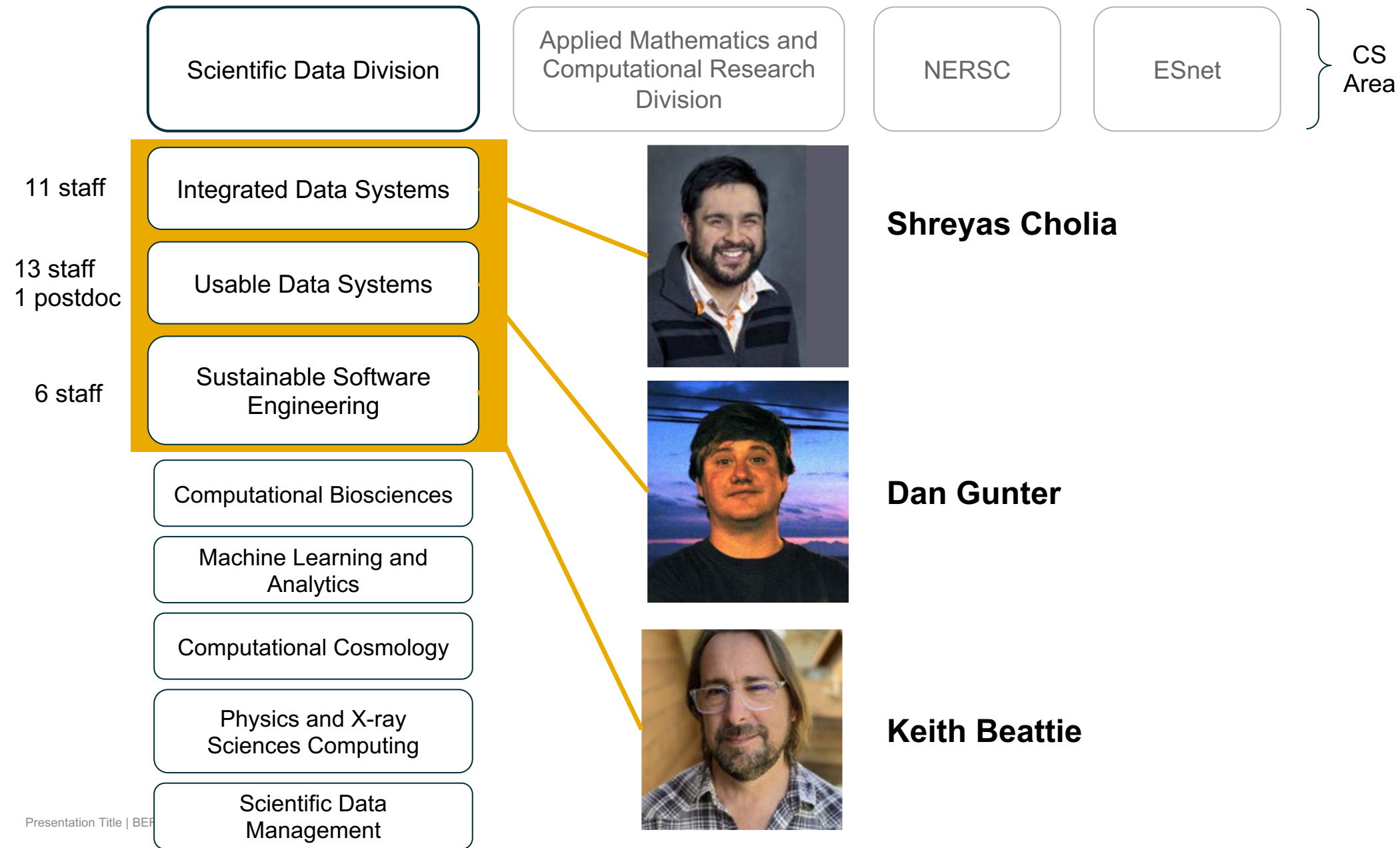
**Research**

- Data Science and AI/ML
- User experience (UX)
- Cybersecurity

## 03 Conclusion

# Overview

# About the groups



# What we do


Deliver leading-edge, innovative methods for solving data-intensive science problems

## Methods

- UI and UX
- Software Engineering
- AI/ML
- Optimization
- Data archives & pipelines

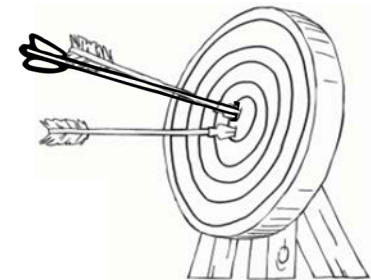


## Values

- Production code over proof-of-concept
  - Being flexible over being experts
- 
- Adapting the solution over adapting the problem

## Goal

Support end-to-end scientific mission through engagement, software engineering and research



# Engagement strategy

We work closely with partners on the domain science side



Co-develop software with stakeholders, identify owner on the science team

Leverage diversity in areas of expertise to form a cross-functional team

User centric design practices drive development

Iterate with stakeholders and rapidly build working prototypes

Implement and disseminate modern software engineering best-practices to ensure sustainability

Partner with NERSC and ESnet to address workflow and data scaling challenges

# Projects

# Domain Science

↳ **Earth Sciences**



# ESS-DIVE: Environmental Systems Science Data Infrastructure for a Virtual Ecosystem

## Scientific Achievement

ESS-DIVE provides long-term stewardship and enables broad usage of data from research in the DOE's Environmental System Science program using the FAIR principles. Key areas of innovation include:

- enhanced support for project-based data.
- scalable support for large data.
- support for new reporting formats.
- enabling linked data across external repositories.

Metrics (Significant Growth!):

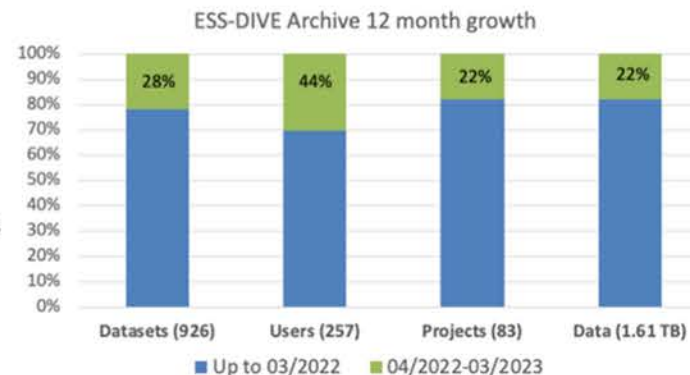
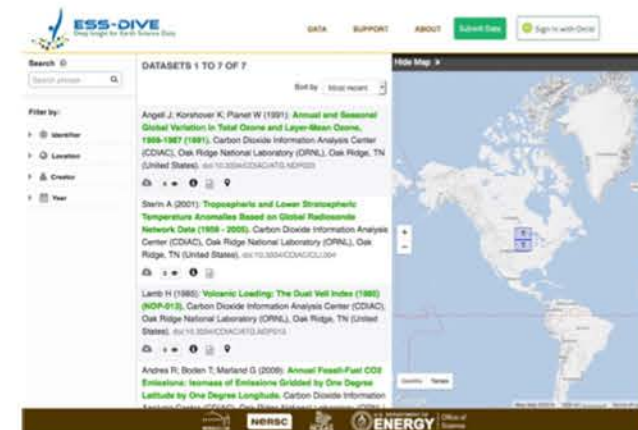
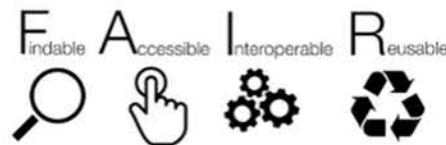
- Datasets: 926 packages / 1.6TB, Users: 257, Projects: 83

## Significance and impact

- Enabling **project-based views and collaboration**.
- Ecosystem of **linked data** across repositories.
- **Large hierarchical datasets** using a tiered storage model.
- Dedicated **storage platform** for analysis and synthesis.
- **13 Reporting Formats** including, newly added **UAV format**.
- **Active Community Engagement**: Webinars, PI Meeting Tutorials, Community Workshops.

## Technical Approach

- User research approach led to project enhancements for data search, collaboration, sharing, and user management.
- **Linked Metadata and cross-BER Sample Interoperability Working group** with ESS-DIVE, NMDC, JGI, EMSL, and KBase.
- Ceph object storage platform for synthesis, integration, and analysis of ESS data (initial 0.5 PB - easy to scale up).
- Tiered Storage model allows for hierarchical folder structures and high performance uploads & downloads via web or Globus.
- Paper in Nature Scientific Data on community reporting formats.



ESS-DIVE data portal:  
<https://data.ess-dive.lbl.gov>

**PI(s)/Facility Lead(s):** Shreyas Cholia, Charuleka Varadharajan, Deb Agarwal

**Collaborating Institutions:** -

**BER Program:** EESSD Data Management

**BER PM:** Justin Hnilo

**Publication(s) for this work:** Crystal-Ornelas et al. Enabling FAIR data in Earth and environmental science with community-centric (meta)data reporting formats. Sci Data 9, 700 (2022). <https://doi.org/10.1038/s41597-022-01606-w>

**Datasets:** <http://data.ess-dive.lbl.gov>

Clear all filters

Search

Filter by:

Project

Identifier

Region description

Creator

Year

Access

DATASETS 1 TO 25 OF 26

1 2 Next

Sort by: Most recent

Lathrop E ; Nutt M ; Wilson C ; Bolton R ; Perkins G ; Harris R (2022): **Soil moisture, physical and chemical properties coincident with airborne SAR data collections for 2017 and 2019, Seward Peninsula, Alaska.** Next-Generation Ecosystem Experiments (NGEE) Arctic, ESS-DIVE repository. Dataset. doi:10.5440/1854940

📄 📍

32

Piliouras A ; Rowland J (2019): **Arctic delta Landsat image classifications.** Watershed Function SFA, ESS-DIVE repository. Dataset. doi:10.15485/1505624

📄 📍

98

1.5K

Bennett K ; Bolton R ; Busey B ; Lathrop E ; Dann J ; Miller G ; Nutt M ; Wilson C (2021): **End-of-Winter Snow Depth, Temperature, Density, and SWE Measurements at Teller Road Site, Seward Peninsula, Alaska, 2019.** Next-Generation Ecosystem Experiments (NGEE) Arctic, ESS-DIVE repository. Dataset. doi:10.5440/1798170

📄 📍

51

Dengel S ; Chafe O ; Cook P ; Torn M (2020): **NGEE Arctic Soil Micro-warming Experiment Temperature Profiles, Council Road Mile Marker 71, Seward Peninsula, Alaska, 2017-2019.** Next-Generation Ecosystem Experiments (NGEE) Arctic, ESS-DIVE repository. Dataset. doi:10.5440/1634215

📄 📍

68

Wilson C ; Bolton R ; Busey R ; Lathrop E ; Dann J ; Bennett K (2021): **End-of-Winter Snow Depth, Temperature, Density and SWE Measurements at Kougarok Road Site, Seward Peninsula, Alaska, 2018.** Next-Generation Ecosystem Experiments (NGEE) Arctic, ESS-DIVE repository. Dataset. doi:10.5440/1593874

📄 📍

77

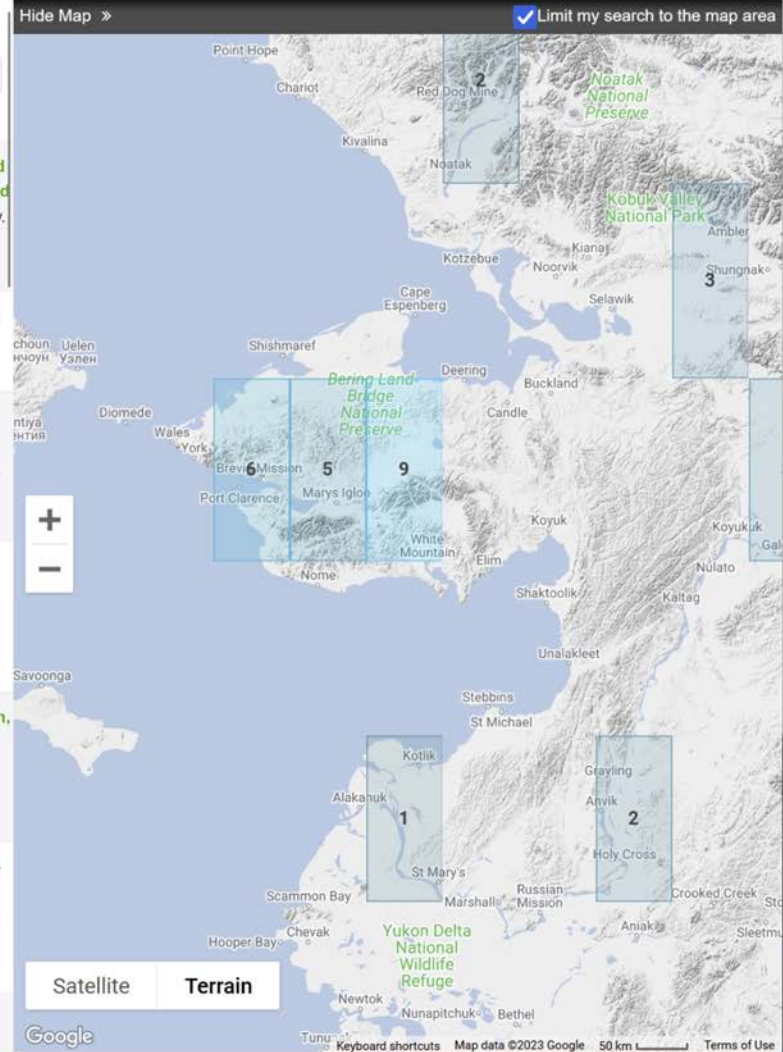
Wilson C ; Bolton R ; Busey R ; Lathrop E ; Dann J ; Charsley-Groffman L ; Bennett K (2021): **End-of-Winter Snow Depth, Temperature, Density and SWE Measurements at Teller Road Site, Seward Peninsula, Alaska, 2016-2018.** Next-Generation Ecosystem Experiments (NGEE) Arctic, ESS-DIVE repository. Dataset. doi:10.5440/1592103

📄 📍

5

330

Nutt M ; Wilson C ; Lathrop E ; Graham D ; Kholodov A ; Conroy N ; Perkins G ; DelVecchio J ; Rowland J (2022): **NGEE Arctic Soil Pit and Core Inventory for Samples Collected Between**



# 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 CO<sub>2</sub>, water, and energy fluxes to address earth science research. Berkeley Lab 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 >600 sites (107 sites outside of the US).

## Significance and Impact

- **Data products:** AmeriFlux BASE – 3157 site-years (>80 sites with decade-long records), AmeriFlux FLUXNET (gap-filled, partitioned data) – 665 site-years
- **Users:** >14,000 registered users
- **Downloads:** >30,000 unique downloads for AmeriFlux BASE data product, >17,000 unique downloads of AmeriFlux sites in FLUXNET2015, >2200 download for AmeriFlux FLUXNET product (compatible with FLUXNET2015)

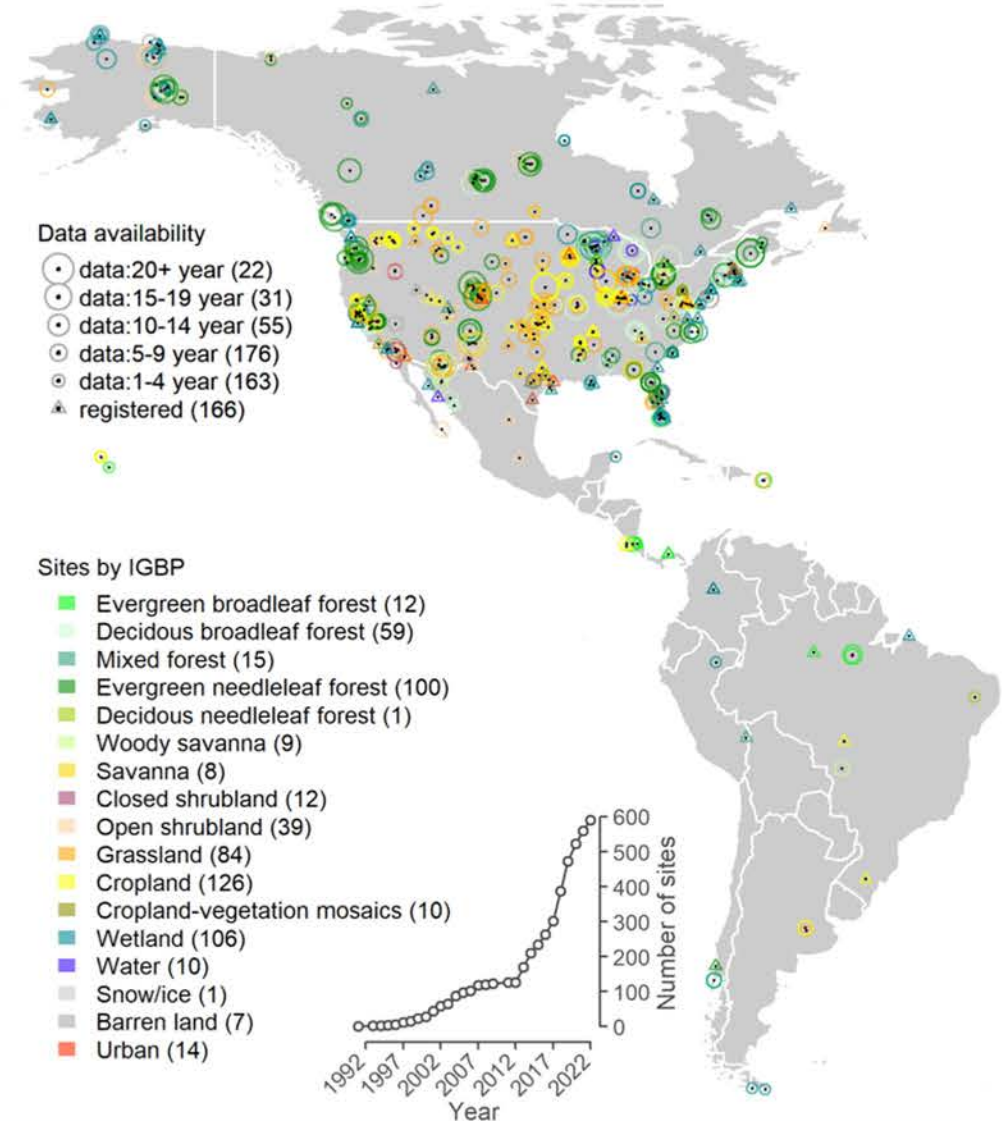
## Technical Approach

- Automated quality and format assessment of data and issue ticket tracking for communication with users.
- ONEFlux\* processing, including turbulence filtering, gap-filling, partitioning of CO<sub>2</sub> fluxes into ecosystem respiration and gross primary production, and uncertainty estimates. ONEFlux generates the AmeriFlux FLUXNET product.

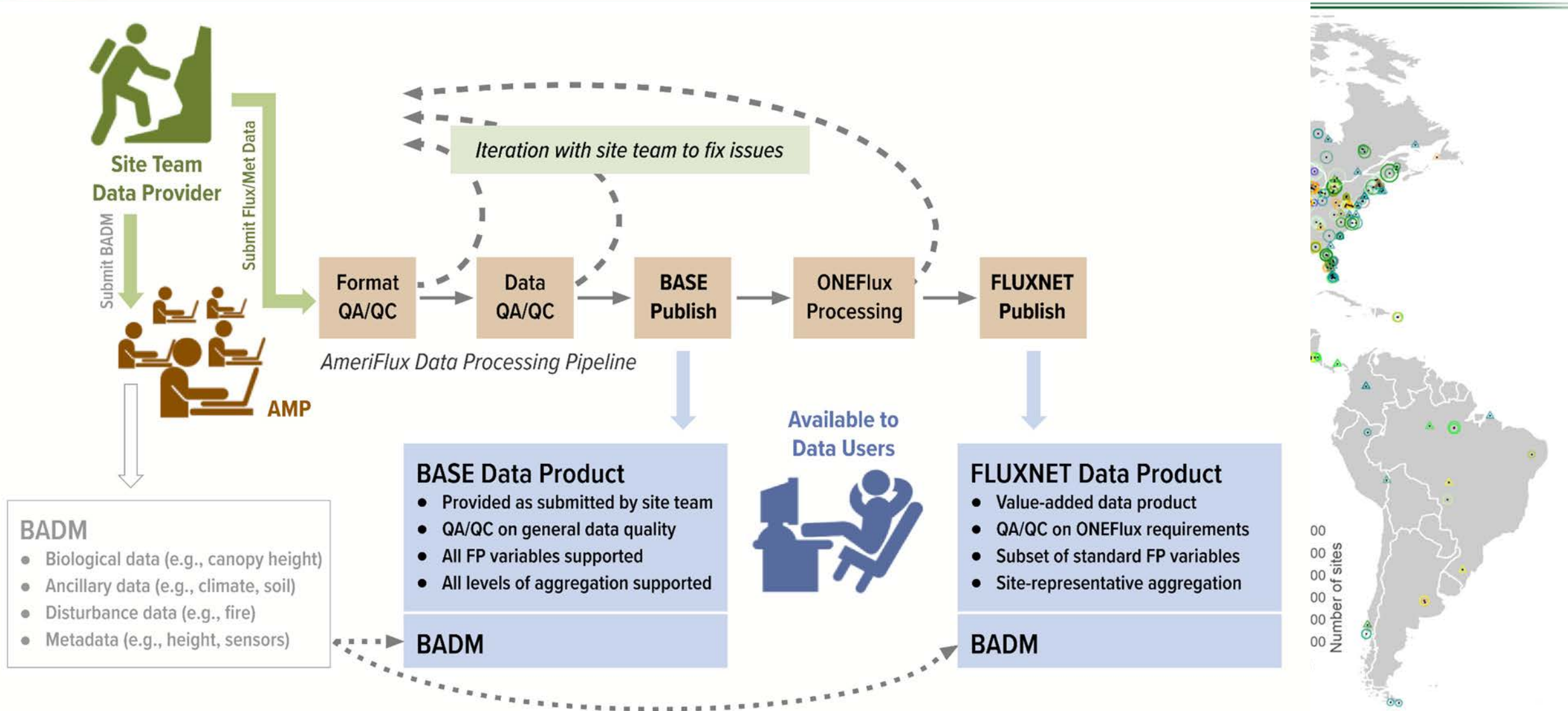
PI: Margaret Torn (Berkeley Lab) Technical POC: You-Wei Cheah (Berkeley Lab)

Collaborating Institutions: UC Berkeley

Publication(s) for this work: \*Pastorello et al. The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. Scientific Data, 7, 225 (2020). <https://www.nature.com/articles/s41597-020-0534-3>



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



# NGEE-Tropics Data Archive: Curating and Preserving Environmental Data from the Tropics

## Scientific Achievement

Developed successful data curation and archival procedures tailored to the needs of data collected in the Tropics, which have unique quality and processing challenges such as long gaps and extreme sparsity in data availability.

## Significance and Impact

- 134 data packages, including 111 public, encompassing a broad variety of data types (shown in the word cloud, with font size indicating the number of packages with data of that type).
- Portal has 498 unique users and 5700 unique data package downloads.

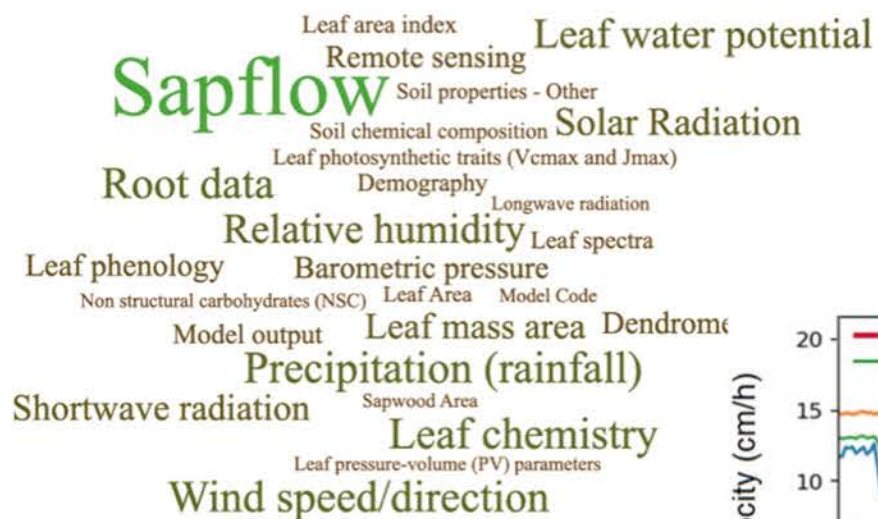
## Technical Approach

Data curation and quality assurance/quality control (QA/QC) processes tailored to data types and analysis goals. Also, part of recent features of NGEE-Tropics Data Archive: automated DOI issuing; implementation of community data review best practices; support for data versioning, draft datasets, raw data packages, partner-specific data policy, and others.

PI: Jeffrey Chambers (Berkeley Lab), Technical POC: Gilberto Pastorello (Berkeley Lab)

Partnering Institutions: Brookhaven National Laboratory, National Institute for Amazon Research, International Institute of Tropical Forestry, Los Alamos National Laboratory, Oak Ridge National Laboratory, Pacific Northwest National Laboratory, National Aeronautics and Space Administration, National Center for Atmospheric Research, Smithsonian Tropical Research Institute, USDA-Forest Service, and 30+ other collaborating institutions

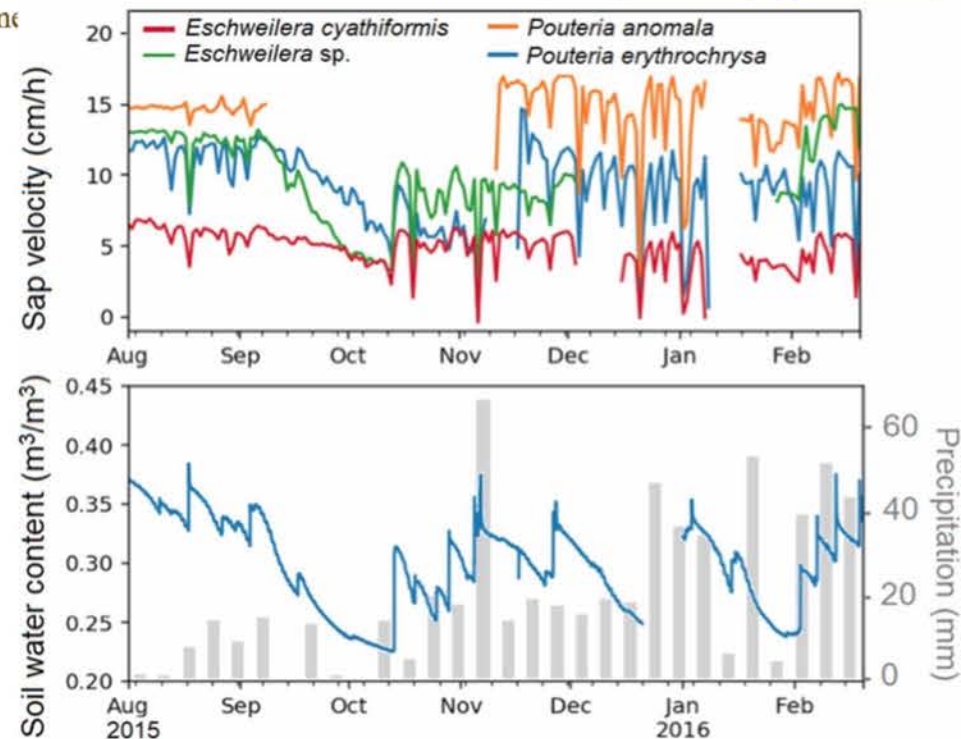
Publication(s) for this work: 120+ Data Packages for NGEE-Tropics: <https://ngt-data.lbl.gov/doi/>



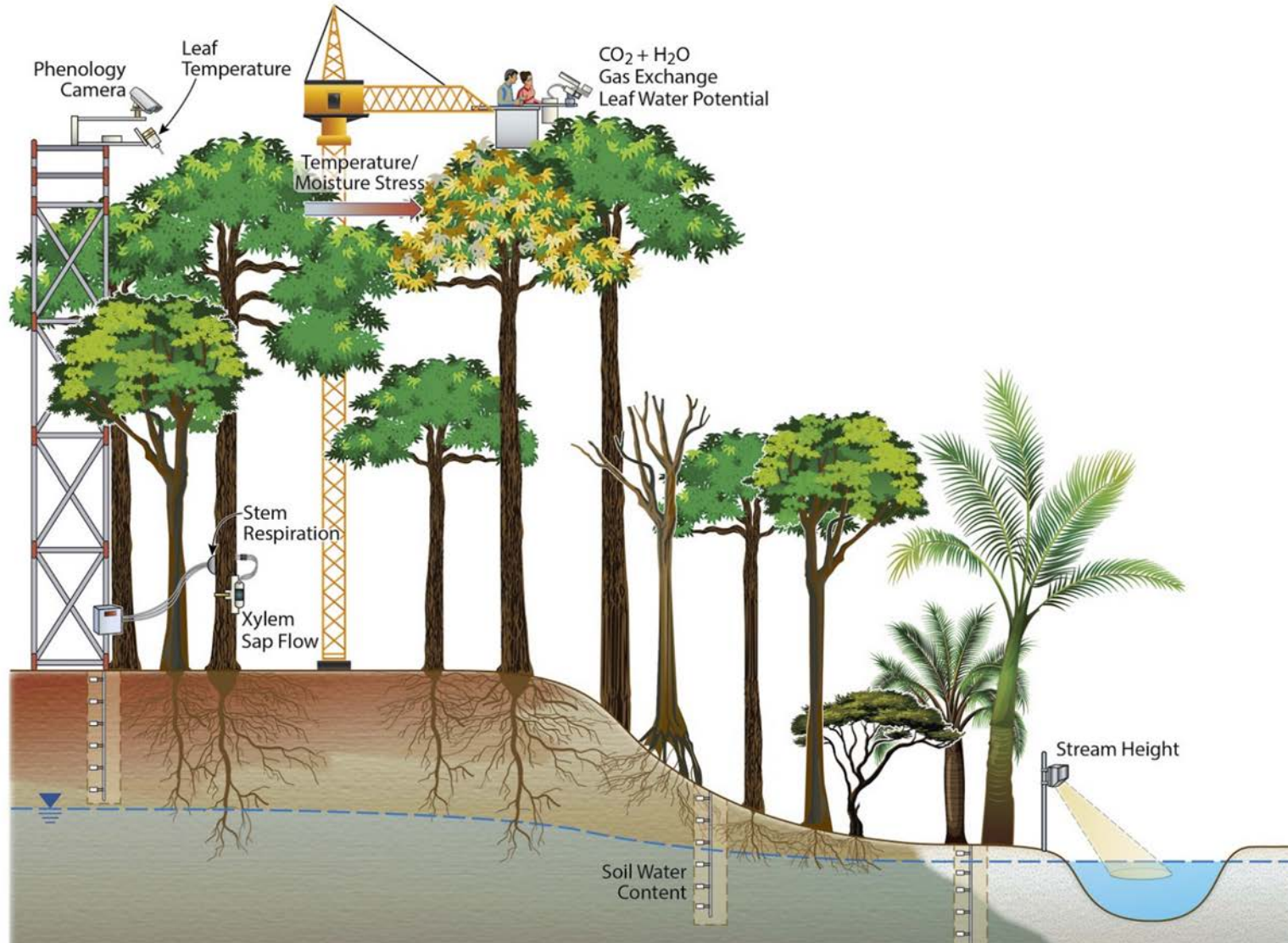
Example of data prepared and QA/QC'd by project team and stored with NGEE-Tropics Archive, exploring soil moisture, radiation, and sapflow relationships, showing a rare shift from radiation- to water-limited conditions for vegetation dynamics in the Central Amazon.

Paper: Meng et al. (2022): [10.1088/1748-9326/ac6f6d](https://doi.org/10.1088/1748-9326/ac6f6d)

Dataset: Gimenez et al. (2021) [10.15486/ngt/1570380](https://doi.org/10.15486/ngt/1570380)

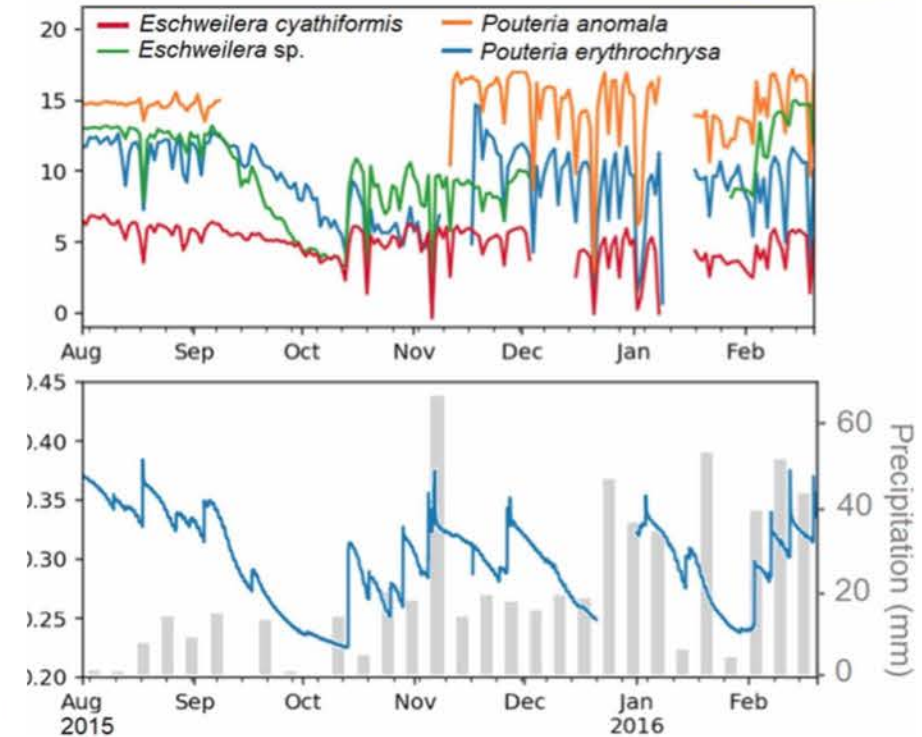


# NGEE-Tropics Data Archive: Curating and Preserving Environmental Data from the Tropics



Example of data prepared and QA/QC'd by project team and stored with NGEN-Tropics Archive, exploring soil moisture, radiation, and sapflow relationships, showing a rare shift from radiation- to water-limited conditions for vegetation dynamics in the Central Amazon.

**Paper:** Meng et al. (2022): [10.1088/1748-9326/ac6f6d](https://doi.org/10.1088/1748-9326/ac6f6d)  
**Dataset:** Gimenez et al. (2021) [10.15486/ngt/1570380](https://doi.org/10.15486/ngt/1570380)



# Domain Science

↳ **Biology**

# Data Portals and Workflows for the National Microbiome Data Collaborative

## Scientific Achievement

The team developed three core products – Data Portal, Submission Portal and NMDC Edge – along with standardized workflows and metadata standards.

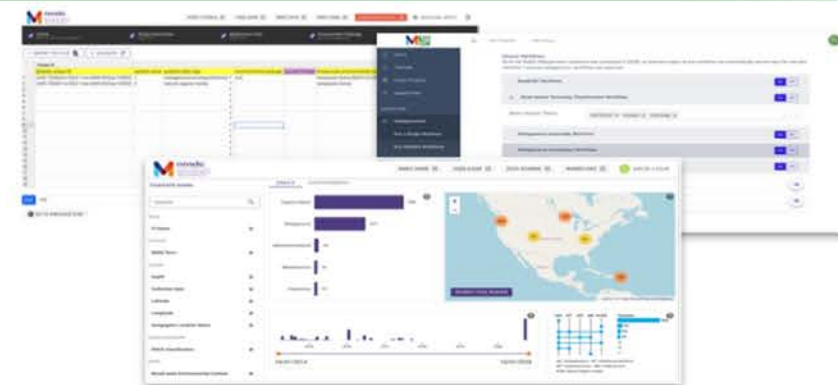
## Significance and Impact

This project supports a Findable, Accessible, Interoperable, and Reusable (FAIR) microbiome data-sharing network, using infrastructure, data standards, and community building that addresses pressing challenges in environmental sciences. It furthers National Microbiome Data Collaborative’s goal to connect data, people, and ideas to advance microbiome innovation and discovery.

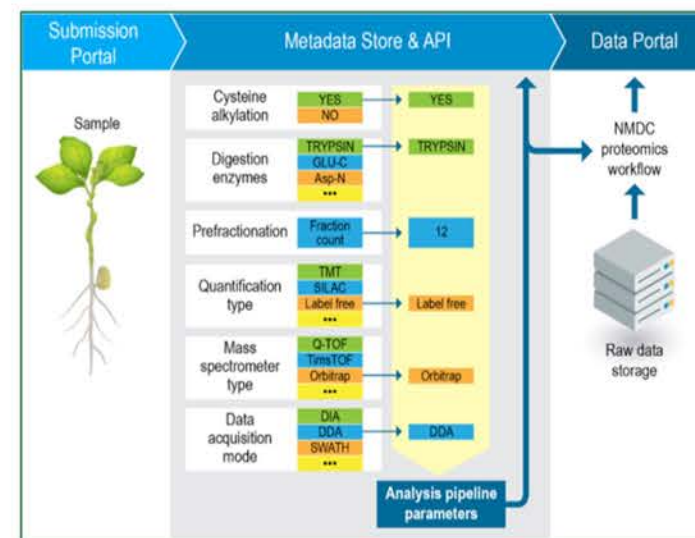
## Technical Approach and Results

- Increased data in the Data Portal (14 studies and 2,449 samples);
- New workflows available in NMDC EDGE: metaproteomics and plasmid/virus detection;
- New features in the Submission Portal & support for JGI and EMSL user projects;
- Onboarded new Ambassador cohort and supported 50 Champions;

PI: Emiley Eloie-Fadrosch, Berkeley Lab POC: Shreyas Cholia, Berkeley Lab  
Collaborating Institutions: Berkeley Lab, LANL, PNNL  
BER Program: BSSD  
BER PM: Ramana Mudupu  
Publication: Keliher et al. “Cohort-based learning for microbiome research community standards,”  
Nature Microbiology. 2023 April 17. doi:10.1038/s41564-023-01361-7.



NMDC products: Submission Portal, NMDC Edge, Data Portal.



NMDC Workflows: From samples to multi-omics data.



# Domain Science

↳ **Process Engineering**

## Scientific Achievement

IDAES represents a paradigm shift for modeling and analyzing complex energy and industrial processes as the only fully equation-oriented platform with integrated support for steady-state design, optimization, dynamic operations, data reconciliation, parameter estimation, and uncertainty quantification.

## Significance and Impact

Over the next decade, hundreds of billions of dollars will be invested in new 21st century energy systems and processes that are more dynamic and interconnected than ever before. The IDAES Integrated Platform will enable companies, technology developers, and researchers to model, design, and optimize these complex integrated energy systems and industrial processes, accelerating their development and deployment to enable more rapid decarbonization.

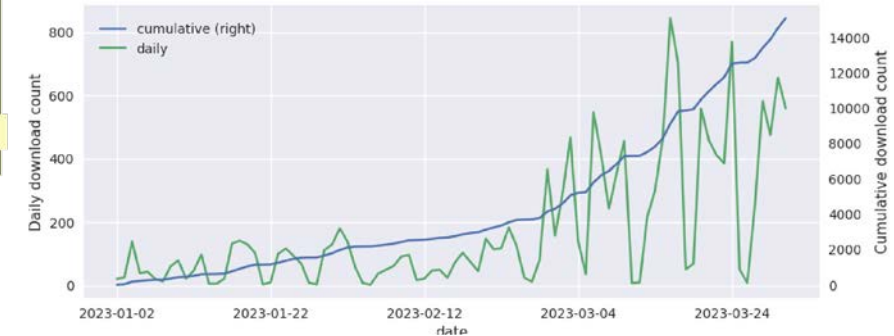
## Technical Approach

- Surrogate modeling integrated with the equation-oriented optimization (PySMO)
- Lead software development, test, and release for 40+ member team
- Develop graphical, notebook, and API user interfaces
- Training and stakeholder outreach

**PI(s)/Facility Lead(s):** David Miller (NETL); Dan Gunter, Berkeley Lab PI  
**Collaborating Institutions:** NETL, Berkeley Lab, SNL, CMU, WVU, U. ND, GA Tech  
**Publication(s) for this work:** Lee, et al., *J. of Adv. Manufacturing and Processing*, April 2021, <https://doi.org/10.1002/amp2.10095>  
**Code Developed or Datasets:** <https://github.com/IDAES/idaes-pse>



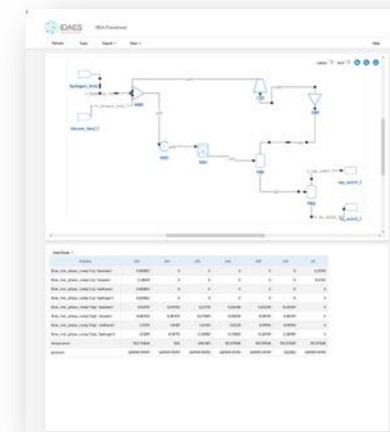
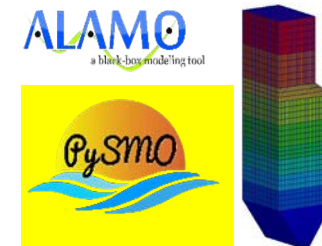
## Software release engineering



Open Source: <https://github.com/IDAES/idaes-pse>

Lee, et al., *J. of Adv. Manufacturing and Processing* (2021)

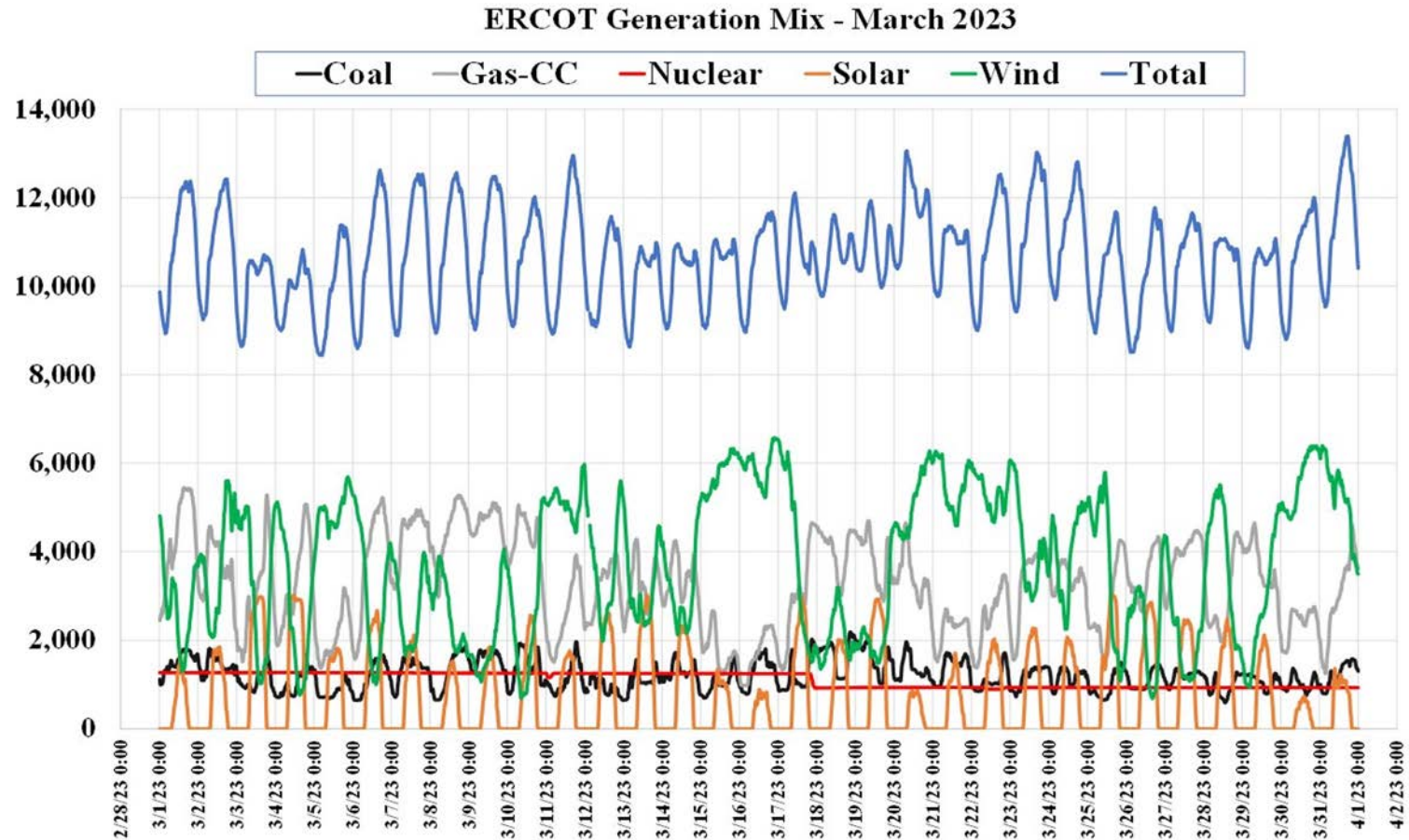
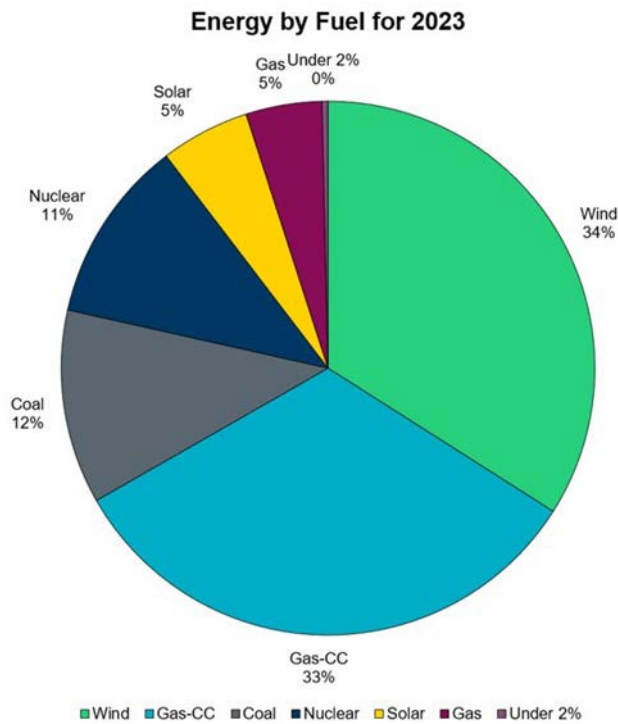
## AI/ML Surrogate Modeling



Uis and UX  
research for  
advanced  
interfaces

# Evolving energy grid increasingly requires flexibility

Data for Electric Reliability Council of Texas (ERCOT) ISO



# Software release engineering for team science

IDAES is an example of a large team with diverse backgrounds:

- Chemical engineers
- Staff
- Students
- Professors
- Computer Scientists
- Mathematical modelers



# National Alliance for Water Innovation (NAWI)

## Water treatment Technoeconomic Assessment Platform (WaterTAP)

### Scientific Achievement

WaterTAP is an open-source advanced process systems engineering tool for desalination and water treatment processes. The National Alliance for Water Innovation (NAWI) modeling project at Berkeley Lab focuses on mathematically simulating electrified water desalination processes to support the next-generation water infrastructures relying on more renewable energy.

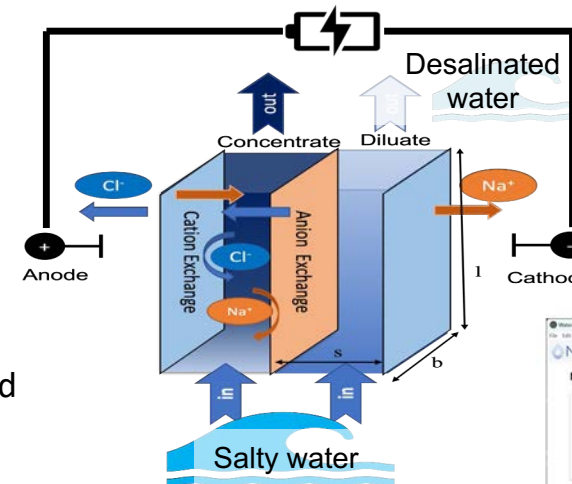
### Significance and Impact

This work is a major modeling effort of the \$100M DOE Office of Energy Efficiency and Renewable Energy's National Alliance for Water Innovation project. WaterTap also provides modeling for the IEDO and SETO programs. WaterTap is built on the IDEAS algebraic optimization framework.

### Technical Approach

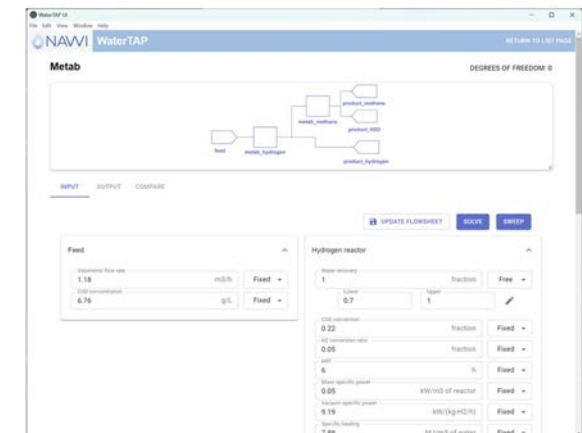
- More than 50 assumed performance models (zero-order models)
- Solve the system by numerical methods optimization and discretization methods to predict the desalination performance and optimize the system design by minimizing water production cost as an objective function.
- WaterTAP UI: Run and compare model optimizations with different inputs

PI(s)/Facility Lead(s): Dan Gunter, Berkeley Lab/ Tim Bartholomew  
Collaborating Institutions: NETL, Berkeley Lab, NREL, ORNL  
Program Manager: Steve McKnight  
Publication(s) for this work: N/A  
Code is available at: <https://github.com/watertap-org/watertap>



Electrified water desalination (electrodialysis)

Graphical UI for running model optimizations



### Funding Sources

- NAWI – desalination (TRL 2-4)
- IEDO – water resource recovery (TRL 4-7)
- SETO – solar driven desalination (TRL 4-7)

# WaterTAP user interface (1)

WaterTAP UI

File Edit View Window Help

NAWI WaterTAP RETURN TO LIST PAGE

### NF-DSPM-DE with bypass

DEGREES OF FREEDOM: 3

INPUT OUTPUT COMPARE

RESET FLOWSHEET SOLVE SWEEP

#### Feed

Volumetric flow rate	3612.22	L/h	Fixed
Ca <sub>2+</sub> concentration	257.99	mg/L	Fixed
SO <sub>4</sub> <sup>2-</sup> concentration	1010.98	mg/L	Fixed
HCO <sub>3</sub> <sup>-</sup> concentration	384.99	mg/L	Fixed
Na <sub>+</sub> concentration	738.98	mg/L	Fixed
Cl <sup>-</sup> concentration	890.93	mg/L	Fixed
K <sup>+</sup> concentration	9	mg/L	Fixed
Mg <sub>2+</sub> concentration	90	mg/L	Fixed

#### NF design

NF pump pressure	3	bar	Free
Lower	1	Upper	
NF area	50	m <sup>2</sup>	Free
Lower	0	Upper	1000
NF water recovery	0.09	fraction	Free
Lower	0	Upper	1

#### NF performance metrics

NF bypass	0.5	fraction	Free
-----------	-----	----------	------

#### System constraints

Product quality	200	mg/L	Fixed
-----------------	-----	------	-------

# WaterTAP user interface (2)

The screenshot displays the WaterTAP user interface for a process named "NF-DSPM-DE with bypass". The interface includes a process flow diagram, a navigation menu (INPUT, OUTPUT, COMPARE), and several data panels. The process flow diagram shows a "Feed" entering a "Splitter", which then feeds into a "Pump" and a "Nanofiltration (NF)" unit. The "NF" unit has a "Waste" outlet and a "Treated" outlet that goes to a "Mixer". A "Bypass" line also connects the "Splitter" directly to the "Mixer". The "DEGREES OF FREEDOM" is indicated as 3.

**OUTPUT** (selected)

- NF design**
  - NF pump pressure **6.36** bar
  - NF area **174.97** m<sup>2</sup>
  - NF water recovery **0.54** fraction
- NF performance metrics**
  - NF bypass **0.04** fraction
- System streams quality**
  - Product hardness **200** mg/L
  - Feed hardness **1016.29** mg/L
  - Disposal hardness **2035.11** mg/L
- Process cost and operating metrics**
  - System cost **0.12** \$/m<sup>3</sup>
  - System energy consumption **0.34** kWhr/m<sup>3</sup>
- NF intrinsic rejection**
  - Ca<sub>2+</sub> intrinsic rejection **96.12** %
  - SO<sub>4</sub><sub>2-</sub> intrinsic rejection **100** %
  - HCO<sub>3</sub><sub>-</sub> intrinsic rejection **82.66** %
  - Na<sub>+</sub> intrinsic rejection **80.08** %
  - Cl<sub>-</sub> intrinsic rejection **71.07** %
  - K<sub>+</sub> intrinsic rejection **78.77** %
  - Mg<sub>2+</sub> intrinsic rejection **96.76** %
- NF observed rejection**
  - Ca<sub>2+</sub> obs. rejection **88.6** %
  - SO<sub>4</sub><sub>2-</sub> obs. rejection **99.99** %
  - HCO<sub>3</sub><sub>-</sub> obs. rejection **58.86** %
  - Na<sub>+</sub> obs. rejection **63.01** %
  - Cl<sub>-</sub> obs. rejection **52.47** %
  - K<sub>+</sub> obs. rejection **67.62** %
  - Mg<sub>2+</sub> obs. rejection **83.59** %

**SAVE CONFIGURATION**

# Research

↳ **Cybersecurity**



# Trusted CI, the NSF Cybersecurity Center of Excellence

## Scientific Achievement

Scientific cyberinfrastructure brings unique challenges for cybersecurity due to its open nature, use of unique instruments, large and complex data sets, and rich ecosystems of collaboration across countries and between disciplines.

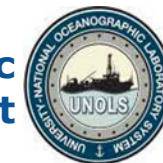
## Significance and Impact

The mission of Trusted CI is to lead in the development of an NSF Cybersecurity Ecosystem with the workforce, knowledge, processes, and cyberinfrastructure that enables trustworthy science and NSF's vision of a nation that is a global leader in research and innovation.

## Technical Approach

- Sean Peisert is Deputy Director and Co-PI of Trusted CI.
- LBNL led 2020 studies on data confidentiality and integrity in science.
- LBNL led a 2021 study on scientific software assurance.
- LBNL leading 2022 study on securing operational technology in science.
- LBNL leading effort on building security into design and procurement for NSF maritime and polar Major Facilities during construction.
- LBNL leads maintenance of the Open Science Cyber Risk Profile (OSCRP)

PI(s): Jim Basney (Director and PI, UIUC), Sean Peisert (Deputy Director and Co-PI, LBNL), Kelli Shute (Exec. Director and Co-PI, IU), Barton Miller (Co-PI, UW)  
Collaborating Institutions: UIUC, Indiana University, Pittsburgh Supercomputing Center, UW Madison  
NSF Program: Office of Advanced Cyberinfrastructure (OAC)  
NSF PM: Robert Beverly  
Publication(s) for this work: A. Adams, E. K. Adams, D. Gunter, R. Kiser, M. Krenz, S. Peisert, and J. Zage, "Roadmap for Securing Operational Technology in NSF Scientific Research," Trusted CI Report, Nov. 2022. DOI: 10.5281/zenodo.7327987



## U.S. Academic Research Fleet

*LBNL is leading the 2023 Trusted CI effort supporting acceptance testing of the NSF-funded Research Class Research Vessels (RCRVs) at Oregon State University, the U.S. Antarctic Program (USAP)'s design of the Antarctic Research Vessel (ARV), and the Scripps Institution of Oceanography's design of the California Coastal Research Vessel (CCRV).*

# Research

↳ **Data Science and AI/ML**

# Automated Learning of Metadata for Scientific Texts through Natural Language Processing

## Scientific Achievement

Using genomic proposals from the Joint Genome Institute (JGI), developed automated and semi-automated techniques for generating relevant training labels that can be used for validating the machine learning (ML) keywords generated for unlabeled texts.

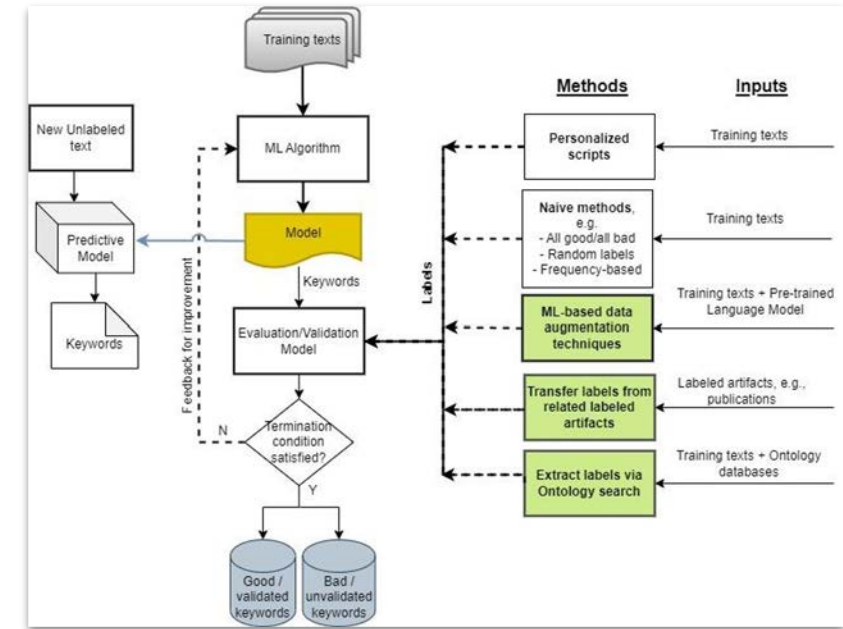
## Significance and Impact

Non-human labeling approaches for the validation of ML-generated keywords is critical to reducing dependence on data contributors for manual labeling of artifacts, and to improving the ability of users to search for unlabeled texts (such as proposals, technical reports). The new automated techniques for labeling is a significant step toward the full automation of the metadata extraction process.

## Technical Approach

- **Exploitation of artifact relationships:** transferring metadata between data artifacts known to be linked to each other directly, e.g papers and proposals.
- **Large Language Modeling (LLM):** exploiting recent developments in LLM and prompt engineering capabilities for topic classification and summarization.

PI(s)/Facility Lead(s): Lavanya Ramakrishnan, Berkeley Lab  
 ASCR Program: Machine Learning and Understanding for High Performance Computing Scientific Discovery  
 ASCR PM: Robinson Pino  
 Publication: Amusat, O.O. et al., "Automated Annotation of Scientific Texts for ML-based Keyphrase Extraction and Validation" (2023; in preparation)



Schematic representation of automated metadata generation techniques for unlabeled texts: The current methods – naive approaches/personalized scripts (shown in white) are insufficient to obtain good quality labels. The new approach (in green) provides new, more robust approaches for validating the ML keywords.

Dataset	Text Type	F-1@10
Inspec	Paper abstracts	0.295
Semeval2010	Full papers	0.219
<b>Our work</b>	<b>Proposals</b>	<b>0.252</b>

Results from exploiting artifact relationships. Performance resembles state-of-the-art results for labeled scientific texts.

# What are Jupyter, JupyterLab, JupyterHub?

Interactive open-source web applications



Allows you to create and share documents, “notebooks,” containing:

- Live code
- Equations
- Visualizations
- Narrative text
- Interactive widgets

- You can use Jupyter notebooks for:
- Data cleaning and data transformation
  - Numerical simulation
  - Statistical modeling
  - Data visualization
  - Machine learning
  - Workflows and analytics frameworks
  - etc.

A screenshot of the JupyterLab web interface. The interface is divided into several panes. On the left, there is a sidebar with a 'Files' pane showing a list of notebooks and files, with 'Lorenz.ipynb' selected. Below the sidebar are 'Running', 'Commands', 'Cell Tools', and 'Tabs' sections. The main area contains a notebook titled 'Lorenz.ipynb'. The notebook has a text cell with the title 'In this Notebook we explore the Lorenz system of differential equations:' followed by the equations  $\dot{x} = \sigma(y - x)$ ,  $\dot{y} = \rho x - y - xz$ , and  $\dot{z} = -\beta z + xy$ . Below this is another text cell explaining that the function will be called to view solutions for specific parameters, resulting in trajectories swirling around two attractors. This is followed by a code cell with the following Python code:

```
In [4]: from lorenz import solve_lorenz
t, x_t = solve_lorenz(N=10)
```

The notebook also features an 'Output View' pane with three sliders for parameters: sigma (set to 10.00), beta (set to 2.67), and rho (set to 28.00). Below the sliders is a 3D visualization of the Lorenz attractor, showing its characteristic butterfly shape. To the right of the visualization is a code cell with the following Python code:

```
9 def solve_lorenz(N=10, max_time=4.0, sigma=10.0, beta=8./3, rho=28.0):
10     """Plot a solution to the Lorenz differential equations."""
11     fig = plt.figure()
12     ax = fig.add_axes([0, 0, 1, 1], projection='3d')
13     ax.axis('off')
14
15     # prepare the axes limits
16     ax.set_xlim((-25, 25))
17     ax.set_ylim((-35, 35))
18     ax.set_zlim((5, 55))
19
20     def lorenz_deriv(x_y_z, t0, sigma=sigma, beta=beta, rho=rho):
21         """Compute the time-derivative of a Lorenz system."""
22         x, y, z = x_y_z
23         return [sigma * (y - x), x * (rho - z) - y, x * y - beta * z]
24
25     # Choose random starting points, uniformly distributed from -15 to 15
26     np.random.seed(1)
27     x0 = -15 + 30 * np.random.random((N, 3))
28
```

# Scalable Collaborative Interactive Reproducible Analytics (SCIRA)

## Scientific Achievement

Establishing the patterns and tools to encourage collaborative, interactive computing at scale, enabling reproducible science. Deployed key prototype demos with **Jupyter** to showcase **secure real-time collaboration** Jupyter notebooks and a **registry** of containerized Jupyter environments launched across HPC and cloud environments

## Significance and Impact

- Developed enhancements to Jupyter/Jupyterhub to support **secure user authorization and traceability** in collaborative notebook environments.
- Enables a secure model for HPC centers to support **Real-Time Collaboration** in Jupyter Notebooks for team science workflows
- Demonstration of notebook sharing capabilities for NERSC systems.
- Collaboration with Dockstore to enable prototype Jupyter container registry to launch predefined reproducible notebook environments

## Technical Approach

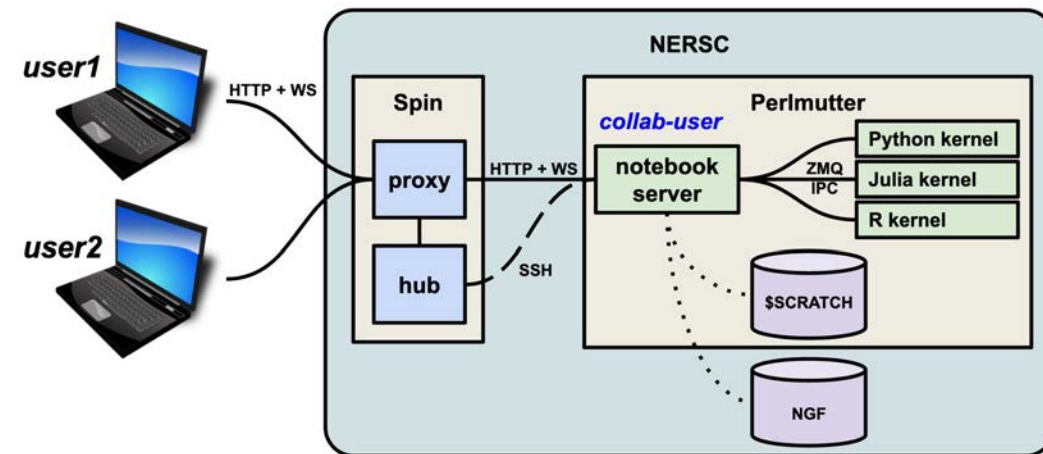
- Enabled key security features in Jupyterhub to facilitate secure, traceable collaborative notebooks with multiple editors. Allows for “Google Docs” type sharing capabilities while maintaining traceability of individual actions.
- Leveraged Dockstore (a mature searchable bioinformatics workflow registry) which supports many workflow languages and multiple launch platforms. Added support to Dockstore for registering and launching Jupyter based workflows on Binder a cloud platform and on JupyterHub HPC clusters.

PI(s)/Facility Lead(s): Lavanya Ramakrishnan

Collaborating Institutions: U.C. Berkeley, Simula, Project Jupyter

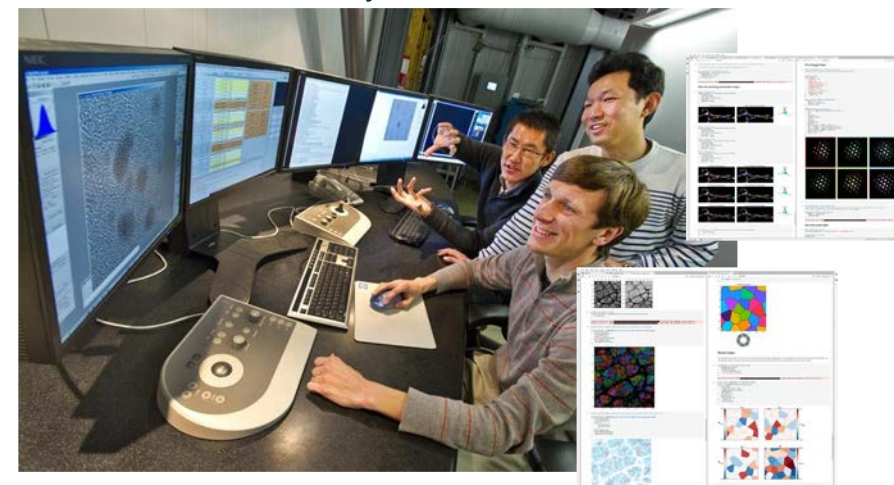
ASCR Program: Data and Visualization

ASCR PM: Margaret Lentz



*Secure model for multi-user real-time collaboration in Jupyter Notebooks.*

*Use Case: Improved ability to integrate external collaborators in the data science process, i.e. they can instantaneously see collected data and perform data collaborative analyses.*



# Science Capsule: Integrating Jupyter Analyses and Data Lifecycle Events

## Scientific Achievement

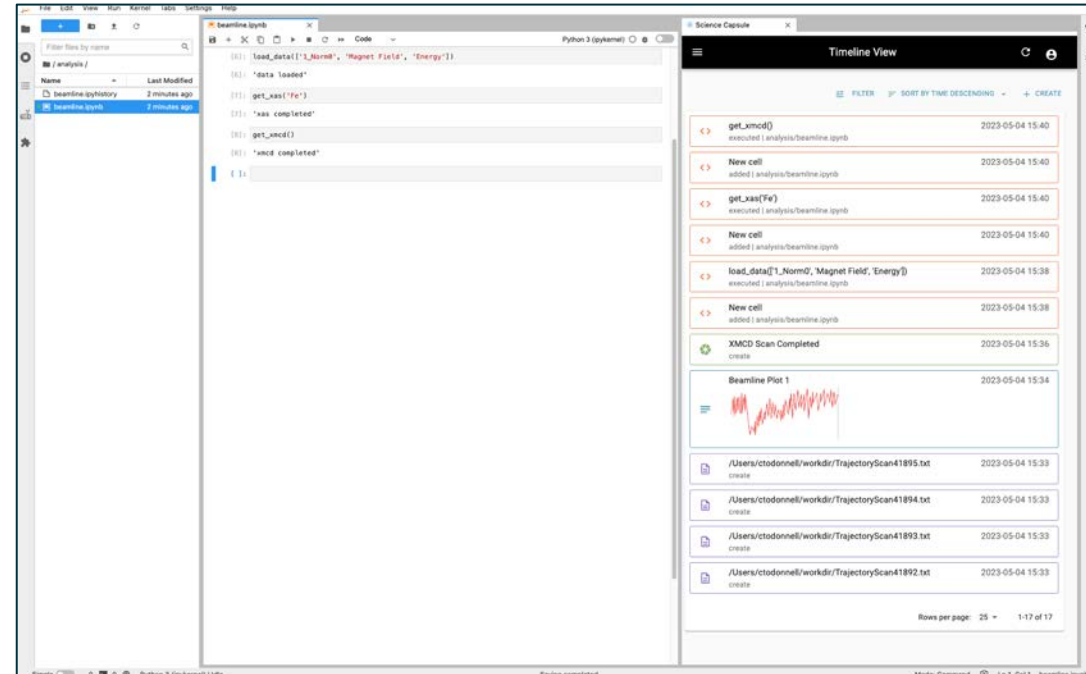
Enabling reproducibility of scientific pipelines through a framework that captures data changes, notes and artifacts, and analyses run in the scientific software ecosystem, including JupyterLab.

## Significance and Impact

Science Capsule captures the scientific lifecycle on a number of different platforms and environments that scientists work on, including analysis work in Jupyter. The integration of Science Capsule and JupyterLab provides the foundation for capturing the analysis provenance from scientific workflows.

## Technical Approach

- User research informed the need for a small-profile, unobtrusive application that could run side by side with other science applications and Jupyter.
- A React app was developed, published as a stand-alone package, and imported as a custom JupyterLab extension.
- The JupyterLab extension allows users to have a combined view of their own Jupyter analysis notebooks and Science Capsule events, including events from our modified Verdant extension that capture notebook events.



*The React app is part of the Science Capsule framework that runs both as a stand-alone application and as a JupyterLab Extension. Integration with JupyterLab allows for capturing the run events and details coming from Jupyter analysis tasks. This provides insights into the history of how functions were parameterized along with how data was changed in the process.*

PI: Lavanya Ramakrishnan  
ASCR Program: Computer Science  
ASCR PM: Margaret Lenz

<https://bitbucket.org/sciencecapsule/sciencecapsule>

# Research

↳ **User Experience (UX)**

# STRUDEL Software typology and user experience framework

## Scientific Achievement

The Scientific software research for User experience, Design, Engagement, and Learning project (STRUDEL) is developing a typology and design system for scientific software to democratize improving user experience, software quality, and software sustainability.

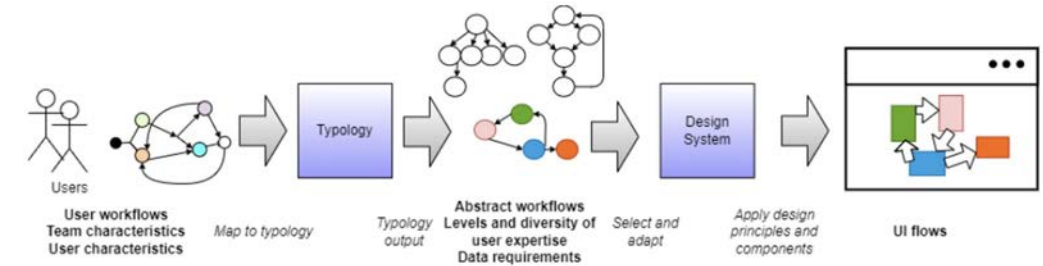
## Significance and Impact

Frequently, user experience (UX) work in the sciences is small-scale, if it exists at all. Scaling up UX and improving software sustainability of core scientific software are intertwined. Democratizing UX, software quality, and software sustainability will help improve the efficiency, accuracy, and satisfaction for science end-users and the stakeholders providing solutions.

## Technical Approach

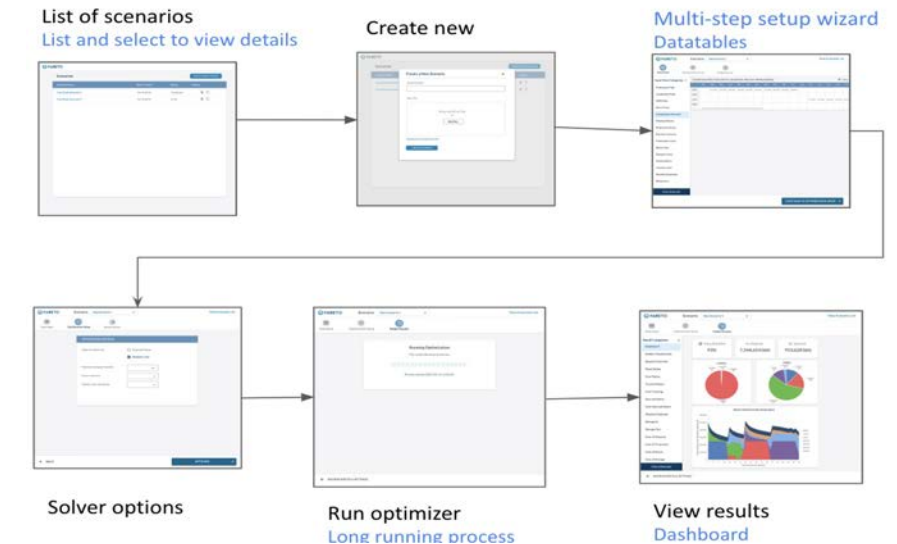
- Studied project user interfaces to classify them into various workflows and component patterns, forming the base for a Scientific Design Library for building tools that are reusable and adaptable.
- User interviews about varied projects and their software development experiences revealed dimensions influencing project investment decisions that will shape decision patterns for the typology.
- Effectiveness of the typology design library will be tested by using it in the development of a new project.

PIs: Lavanya Ramakrishnan, Dan Gunter, Sarah Poon (Berkeley Lab)  
Alfred P. Sloan Foundation  
Program Manager: Elizabeth Vu



*By mapping project characteristics to typology categories, we will identify common patterns and workflow abstractions that will be represented as UI flows (how a user progresses through an interface) in our design system.*

### PARETO - Optimization Workflow

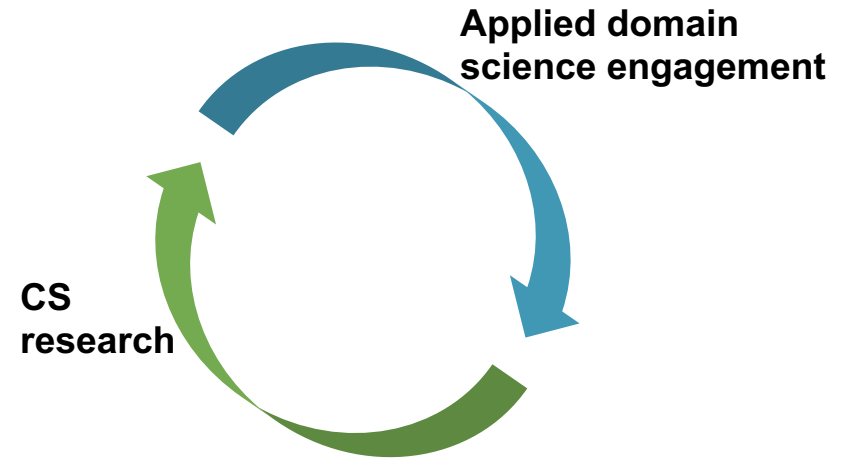


*Documenting user journeys in existing software informs the common patterns seen in scientific software. This makes it possible to generalize solutions and guidelines that scientists can take and modify to improve the speed of software development and the quality of scientific software.*



# Conclusion

# The "virtuous cycle" of engagement and research



We have found our greatest successes by working as *embedded* members of scientific teams.

This requires building of trust and respect by understanding the team and the community before proposing solutions, and by "sticking around" to see through those solutions after classical computer science work is done.

Ideas generated during these engagements often lead to very meaningful research goals.

# Find out more

Some points of contact for you to reach out and find out more about what our groups do in these projects.

ESS-DIVE - Val Hendrix, Shreyas Cholia

AmeriFlux / NGEE Tropics - Gilberto Pastorello

NMDC - Shreyas Cholia

IDAES - Mayo Amusat (surrogate modeling), Dan Gunter (UI), Keith Beattie, Ludovico Bianchi (software release)

WaterTAP - Mayo Amusat, Xiangyu Bi (modeling), Mike Pesce (UI)

TrustedCI - Sean Peisert

ScienceSearch, ScienceCapsule - Anna Giannakou, Lavanya Ramakrishnan

SciRA - Shreyas Cholia

STRUDEL - Sarah Poon, Drew Paine

# Thank You

Questions?