

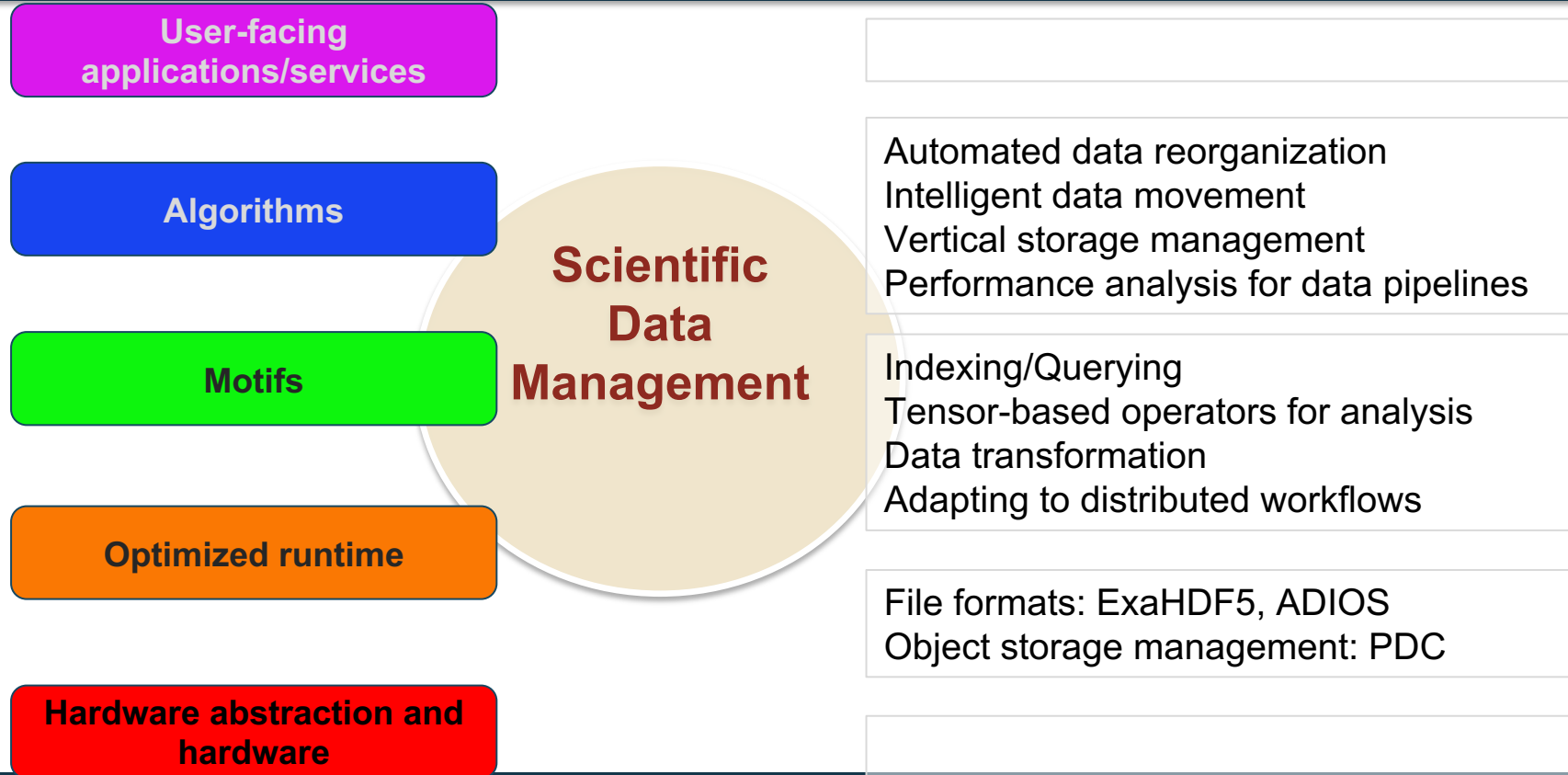
Making Scientific Data Ready for Analysis

John Wu
LBNL

Outline

- 1. Introduction to Scientific Data Management**
- 2. Data at High-Performance Computing Centers**
- 3. Data around HPC Centers: Large-scale Data Curation**

Scientific Data Management at LBNL



Outline

- 1. Introduction to Scientific Data Management**
- 2. Data at High-Performance Computing Centers**
- 3. Data around HPC Centers: Large-scale Data Curation**

Example Scientific Data Analysis at a HPC Center

-- Reducing Petabytes to Kilobytes

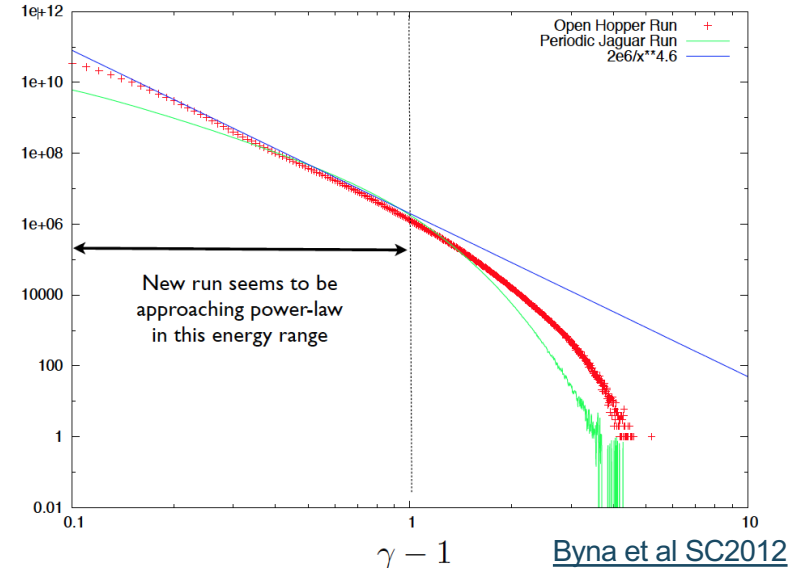
Magnetic reconnection

- ☐ Applications: magnetic confinement fusion, solar wind
- ☐ Data from simulation of trillions of ions and electrons

- **Example: space weather simulation on 120,000 hopper cores @NERSC**
 - 20,000 MPI tasks * 6 OpenMP threads
- **~35TB per timestep**
- **Total ~350TB**
- **Example science result**
 - Particle energy distribution follows the power law

Challenge

- **How to quickly and easily get from 350TB of raw data to a few kilobytes in the graph below?**



Technology 1: Efficient I/O with Proactive Data Containers

Scientific Achievement

PDC achieves efficient storage and access of data with simple object abstractions, transparently taking advantage of deep and heterogeneous HPC storage hierarchy.

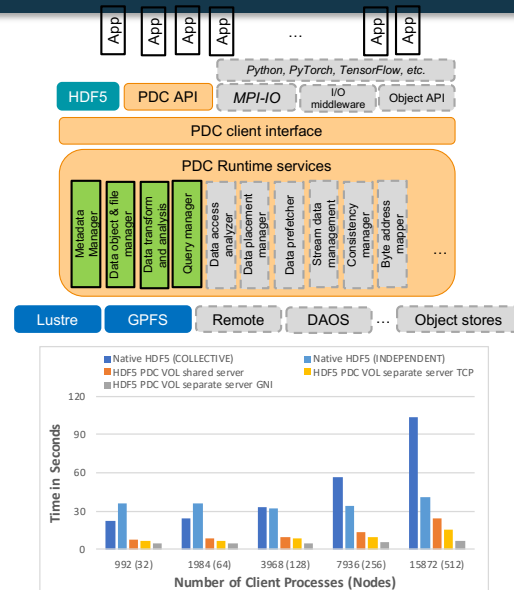
Significance and Impact

Applications store data as PDC objects, which the PDC runtime system transparently and efficiently manages in the storage hierarchy. PDC is portable over underlying HPC file systems, so users don't need special installations.

Research Details

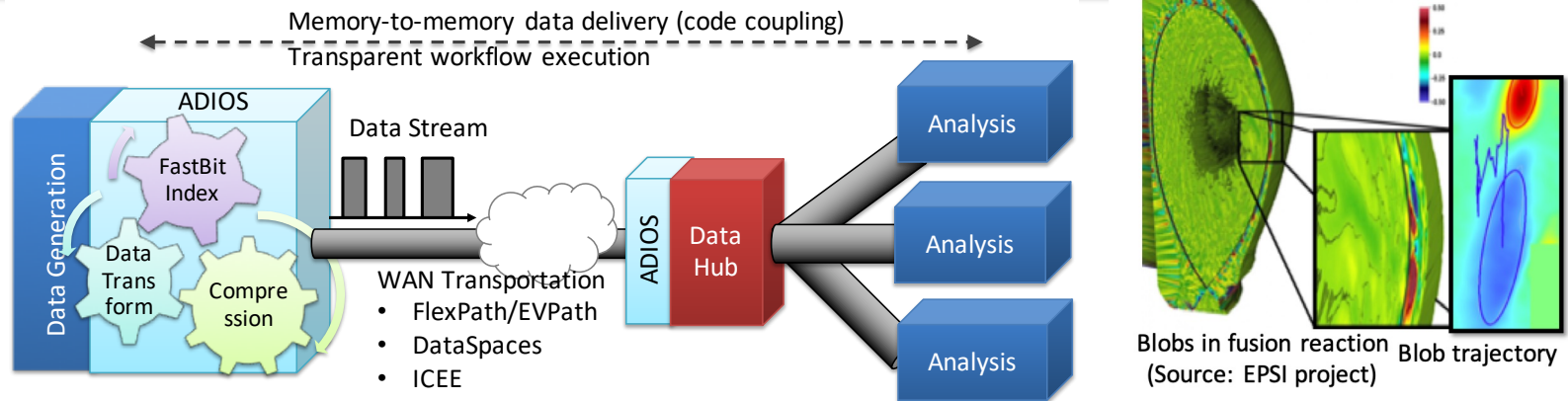
- Existing data management and I/O solutions are based on POSIX semantics and face performance challenges
- We developed a novel data management system with simple data object interfaces, efficient and transparent data movement in storage hierarchy, proactive analysis in the data path, and scalable metadata management
- PDC object management outperforms highly-tuned POSIX I/O based on HDF5 by up to **8X** for writing and by up to **22X** for reading. Searching metadata is **40X** faster than Lustre file system
- Public release available at <https://github.com/hpc-io/pdc>


H. Tang, S. Byna, et al., "Toward Scalable and Asynchronous Object-centric Data Management for HPC", 18th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid) 2018. (DOI: 10.1109/CCGRID.2018.00026)



PDC Overview and Performance: Top figure shows an overview of PDC interfaces through HDF5 and PDC's object interface and various PDC services. Gray boxes show future work. The bottom figure shows the performance of writing particle data (from 248GB to ~4TB) with HDF5 and different configurations of PDC. PDC outperforms highly-tuned POSIX I/O with HDF5 by **6.5X** on average.

Technology 2: Fast In-memory Data Movement



- **Technology: in memory data movement for stream-based distributed data process**
 - **Application: detect fusion plasma blobs:**
 - Which leak energy from tokamak plasmas
 - and damage walls of the tokamak
-  **The experimental facility may not have enough computing power for the necessary data processing**
- Distributed in transient processing**
- Makes more processing power available
 - Allows more scientists to participate in the data analysis operations and monitor the experiment remotely

Technology 3: FasTensor Simplifies Tensor Computation

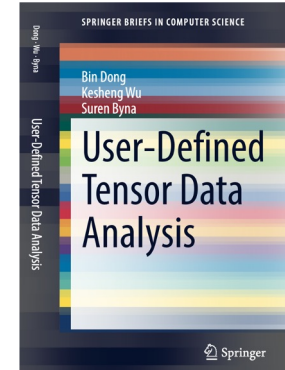
FasTensor website:

<https://sdm.lbl.gov/fastensor/>

Scientific Achievement

FasTensor, a data parallelization system for user-defined analysis, significantly reduces programming effort for various scientific analysis operations. It outperforms popular Big Data platforms such as Spark by **~50X to ~90X** in executing machine learning computations.

Book
([Springer 2021](#))



1	Introduction	1
1.1	Introduction to the Data Science	1
1.2	Data Mining	2
1.3	Machine Learning	3
1.4	Introduction to the Data Science for Science	4
2	Machine Learning Model	6
2.1	Machine Learning Model	6
2.2	Machine Learning Model	11
2.3	Machine Learning Model	11
2.4	Machine Learning Model	11
2.5	Machine Learning Model	11
2.6	Machine Learning Model	11
2.7	Machine Learning Model	11
2.8	Machine Learning Model	11
2.9	Machine Learning Model	11
2.10	Machine Learning Model	11
2.11	Machine Learning Model	11
2.12	Machine Learning Model	11
2.13	Machine Learning Model	11
2.14	Machine Learning Model	11
2.15	Machine Learning Model	11
2.16	Machine Learning Model	11
2.17	Machine Learning Model	11
2.18	Machine Learning Model	11
2.19	Machine Learning Model	11
2.20	Machine Learning Model	11
2.21	Machine Learning Model	11
2.22	Machine Learning Model	11
2.23	Machine Learning Model	11
2.24	Machine Learning Model	11
2.25	Machine Learning Model	11
2.26	Machine Learning Model	11
2.27	Machine Learning Model	11
2.28	Machine Learning Model	11
2.29	Machine Learning Model	11
2.30	Machine Learning Model	11
2.31	Machine Learning Model	11
2.32	Machine Learning Model	11
2.33	Machine Learning Model	11
2.34	Machine Learning Model	11
2.35	Machine Learning Model	11
2.36	Machine Learning Model	11
2.37	Machine Learning Model	11
2.38	Machine Learning Model	11
2.39	Machine Learning Model	11
2.40	Machine Learning Model	11
2.41	Machine Learning Model	11
2.42	Machine Learning Model	11
2.43	Machine Learning Model	11
2.44	Machine Learning Model	11
2.45	Machine Learning Model	11
2.46	Machine Learning Model	11
2.47	Machine Learning Model	11
2.48	Machine Learning Model	11
2.49	Machine Learning Model	11
2.50	Machine Learning Model	11
2.51	Machine Learning Model	11
2.52	Machine Learning Model	11
2.53	Machine Learning Model	11
2.54	Machine Learning Model	11
2.55	Machine Learning Model	11
2.56	Machine Learning Model	11
2.57	Machine Learning Model	11
2.58	Machine Learning Model	11
2.59	Machine Learning Model	11
2.60	Machine Learning Model	11
2.61	Machine Learning Model	11
2.62	Machine Learning Model	11
2.63	Machine Learning Model	11
2.64	Machine Learning Model	11
2.65	Machine Learning Model	11
2.66	Machine Learning Model	11
2.67	Machine Learning Model	11
2.68	Machine Learning Model	11
2.69	Machine Learning Model	11
2.70	Machine Learning Model	11
2.71	Machine Learning Model	11
2.72	Machine Learning Model	11
2.73	Machine Learning Model	11
2.74	Machine Learning Model	11
2.75	Machine Learning Model	11
2.76	Machine Learning Model	11
2.77	Machine Learning Model	11
2.78	Machine Learning Model	11
2.79	Machine Learning Model	11
2.80	Machine Learning Model	11
2.81	Machine Learning Model	11
2.82	Machine Learning Model	11
2.83	Machine Learning Model	11
2.84	Machine Learning Model	11
2.85	Machine Learning Model	11
2.86	Machine Learning Model	11
2.87	Machine Learning Model	11
2.88	Machine Learning Model	11
2.89	Machine Learning Model	11
2.90	Machine Learning Model	11
2.91	Machine Learning Model	11
2.92	Machine Learning Model	11
2.93	Machine Learning Model	11
2.94	Machine Learning Model	11
2.95	Machine Learning Model	11
2.96	Machine Learning Model	11
2.97	Machine Learning Model	11
2.98	Machine Learning Model	11
2.99	Machine Learning Model	11
2.100	Machine Learning Model	11
Appendix		11
A.1	Introduction to the Data Science	11
A.2	Introduction to the Data Science	11
A.3	Introduction to the Data Science	11
A.4	Introduction to the Data Science	11
A.5	Introduction to the Data Science	11
A.6	Introduction to the Data Science	11
A.7	Introduction to the Data Science	11
A.8	Introduction to the Data Science	11
A.9	Introduction to the Data Science	11
A.10	Introduction to the Data Science	11
A.11	Introduction to the Data Science	11
A.12	Introduction to the Data Science	11
A.13	Introduction to the Data Science	11
A.14	Introduction to the Data Science	11
A.15	Introduction to the Data Science	11
A.16	Introduction to the Data Science	11
A.17	Introduction to the Data Science	11
A.18	Introduction to the Data Science	11
A.19	Introduction to the Data Science	11
A.20	Introduction to the Data Science	11
A.21	Introduction to the Data Science	11
A.22	Introduction to the Data Science	11
A.23	Introduction to the Data Science	11
A.24	Introduction to the Data Science	11
A.25	Introduction to the Data Science	11
A.26	Introduction to the Data Science	11
A.27	Introduction to the Data Science	11
A.28	Introduction to the Data Science	11
A.29	Introduction to the Data Science	11
A.30	Introduction to the Data Science	11
A.31	Introduction to the Data Science	11
A.32	Introduction to the Data Science	11
A.33	Introduction to the Data Science	11
A.34	Introduction to the Data Science	11
A.35	Introduction to the Data Science	11
A.36	Introduction to the Data Science	11
A.37	Introduction to the Data Science	11
A.38	Introduction to the Data Science	11
A.39	Introduction to the Data Science	11
A.40	Introduction to the Data Science	11
A.41	Introduction to the Data Science	11
A.42	Introduction to the Data Science	11
A.43	Introduction to the Data Science	11
A.44	Introduction to the Data Science	11
A.45	Introduction to the Data Science	11
A.46	Introduction to the Data Science	11
A.47	Introduction to the Data Science	11
A.48	Introduction to the Data Science	11
A.49	Introduction to the Data Science	11
A.50	Introduction to the Data Science	11
A.51	Introduction to the Data Science	11
A.52	Introduction to the Data Science	11
A.53	Introduction to the Data Science	11
A.54	Introduction to the Data Science	11
A.55	Introduction to the Data Science	11
A.56	Introduction to the Data Science	11
A.57	Introduction to the Data Science	11
A.58	Introduction to the Data Science	11
A.59	Introduction to the Data Science	11
A.60	Introduction to the Data Science	11
A.61	Introduction to the Data Science	11
A.62	Introduction to the Data Science	11
A.63	Introduction to the Data Science	11
A.64	Introduction to the Data Science	11
A.65	Introduction to the Data Science	11
A.66	Introduction to the Data Science	11
A.67	Introduction to the Data Science	11
A.68	Introduction to the Data Science	11
A.69	Introduction to the Data Science	11
A.70	Introduction to the Data Science	11
A.71	Introduction to the Data Science	11
A.72	Introduction to the Data Science	11
A.73	Introduction to the Data Science	11
A.74	Introduction to the Data Science	11
A.75	Introduction to the Data Science	11
A.76	Introduction to the Data Science	11
A.77	Introduction to the Data Science	11
A.78	Introduction to the Data Science	11
A.79	Introduction to the Data Science	11
A.80	Introduction to the Data Science	11
A.81	Introduction to the Data Science	11
A.82	Introduction to the Data Science	11
A.83	Introduction to the Data Science	11
A.84	Introduction to the Data Science	11
A.85	Introduction to the Data Science	11
A.86	Introduction to the Data Science	11
A.87	Introduction to the Data Science	11
A.88	Introduction to the Data Science	11
A.89	Introduction to the Data Science	11
A.90	Introduction to the Data Science	11
A.91	Introduction to the Data Science	11
A.92	Introduction to the Data Science	11
A.93	Introduction to the Data Science	11
A.94	Introduction to the Data Science	11
A.95	Introduction to the Data Science	11
A.96	Introduction to the Data Science	11
A.97	Introduction to the Data Science	11
A.98	Introduction to the Data Science	11
A.99	Introduction to the Data Science	11
A.100	Introduction to the Data Science	11

Significance and Impact

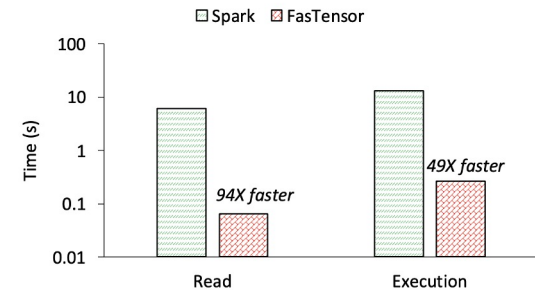
FasTensor has been evaluated using:

- Earth science for detecting earthquakes and other subsurface events
- Fusion science for tracking field evolution
- Climate data analysis with Convolutional Neural Network (CNN) to predict extreme weather events

Research Details

FasTensor programming model consists of:

- Simple data model (i.e., Stencil) abstraction well known in numerical computing
- Single operator (i.e., Transform) to execute user-defined analysis
- An execution engine for automatic parallelization



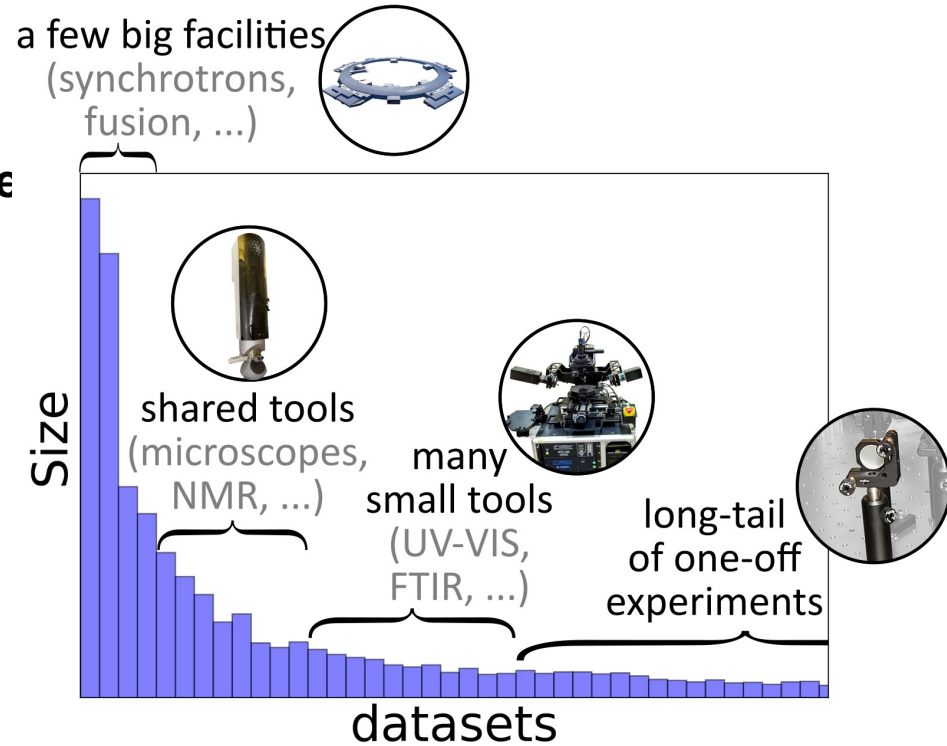
Performance comparison of FasTensor with Spark for completing CNN (CONV, Pooling and ReLU) on a 2D climate (CAM5) data

Outline

1. Introduction to Scientific Data Management
2. Data at High-Performance Computing Centers
3. Data around HPC Centers: Large-scale Data Curation

Many Data Sources Outside of HPC Centers

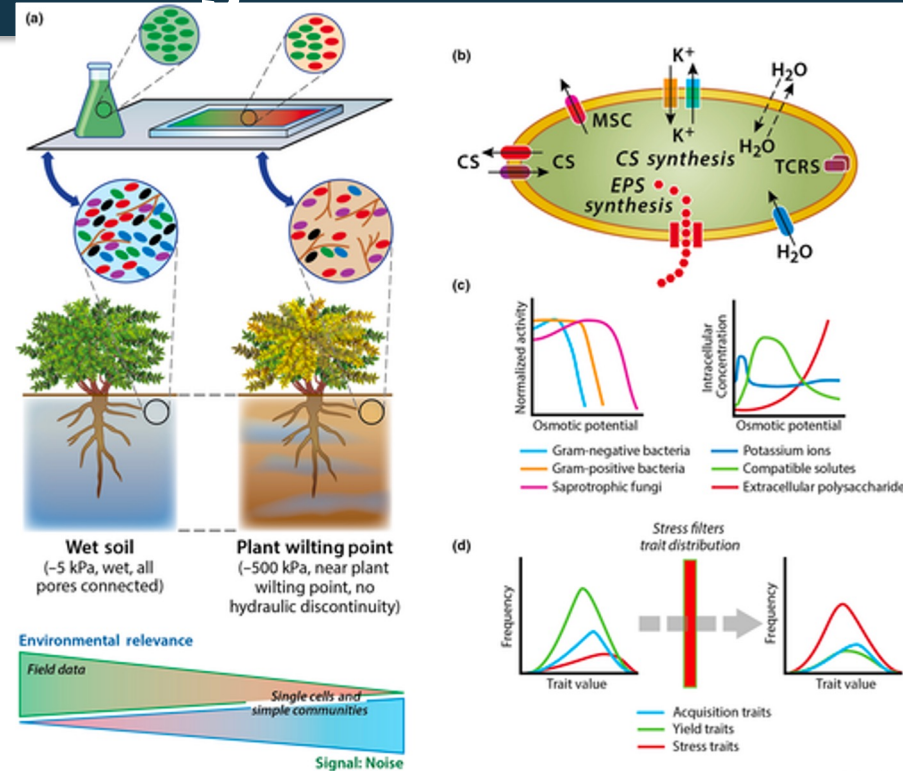
- **There are many more scientific instruments, like those used in environmental monitoring, producing relatively small amounts of data, but might be just as valuable**
 - Significant amounts of valuable data in publications
- **Because the efforts available to handle the data might be quite limited, thus need more automated solutions**
 - QA / QC
 - Metadata / provenance
 - Formatting and other data convention
 - Processing tools: automated curation
 - Integration with other data sets



Diving into Understanding Drought Processes



- Drought impacts carbon stabilization
 - Major societal impact
 - Mediated by biotic and abiotic factors
- Diverse datasets could elucidate underlying processes
 - e.g., meta-analyses, process models
 - **Data needs to be collected and integrated**



Multiscale depiction of the microbial response to drought stress.
 Malik & Bouskill. 2022. *Functional Ecology*. <https://doi.org/10.1111/1365-2435.14010>

The Need to Curate Data at Scale

Develop multi-scale capabilities to accurately sense and simulate biological-environmental feedbacks

The emergence of environmental 'big data,' available through multi-scale environmental sensing and genome sequencing technologies, can be combined with powerful computational capabilities. This combination allows the development of mechanistic models of how microbes interact within complex ecosystem networks and how these networks respond to environmental stresses to influence ecosystem functioning and productivity.

With the discovery of new processes, **Berkeley Lab researchers curate and extract information from diverse biological-environmental datasets**, taking advantage of multi-scale ecosystem sensor data and multi-omics approaches. This knowledge is used to develop the next generation of models, which can answer questions about the relationships between microbial metabolic diversity and ecosystem function. For example, our scientists explore the key ecological constraints that select for the assembly, structure, and function of microbial groups in a particular environment. In diverse settings, ranging from the root zone to permafrost to tropical forests and watersheds, we develop and use these next-generation microbe-ecosystem models to uncover how ecology, evolution, and environmental variation interact to direct the flow of energy, nutrients, and water through terrestrial ecosystems.

“...understanding multi-scale ecosystem processes often requires acquisition and integration of a variety of data”

Varadharajan et al. 2022. *Computers and Geosciences*.
<https://doi.org/10.1016/j.cageo.2021.105024>



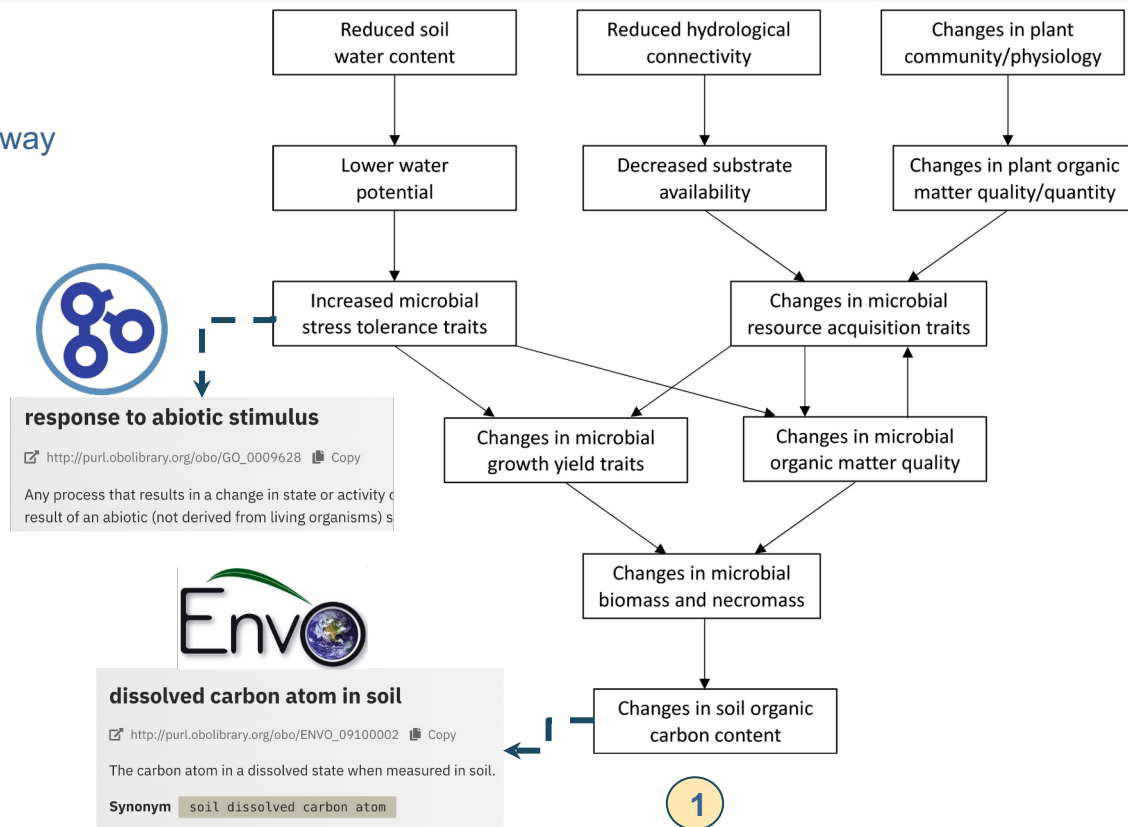
EESA Strategic Vision 2025, p13
<https://eesa.lbl.gov/about/strategic-vision>

How to curate data *at scale*?

Automating Data Curation Step 1: Expand Ontological Model

1. Curate multi-scale conceptual model

- Manually curate causally linked terms (“pathway diagram”)
- Link each box/edge to ontology concepts



Gene Ontology Causal Activity Modeling

Paul D Thomas, ..., Christopher J Mungall
Nat Genet. 2019 Oct;51(10):1429-1433

Ontology Development Kit: a toolkit for building, maintaining and standardizing biomedical ontologies

Nicolas Matentzoglou, ..., Christopher J Mungall, David Osumi-Sutherland

Database, Volume 2022, 2022, baac087



Automating Data Curation Step 2: Link Model to Datasets

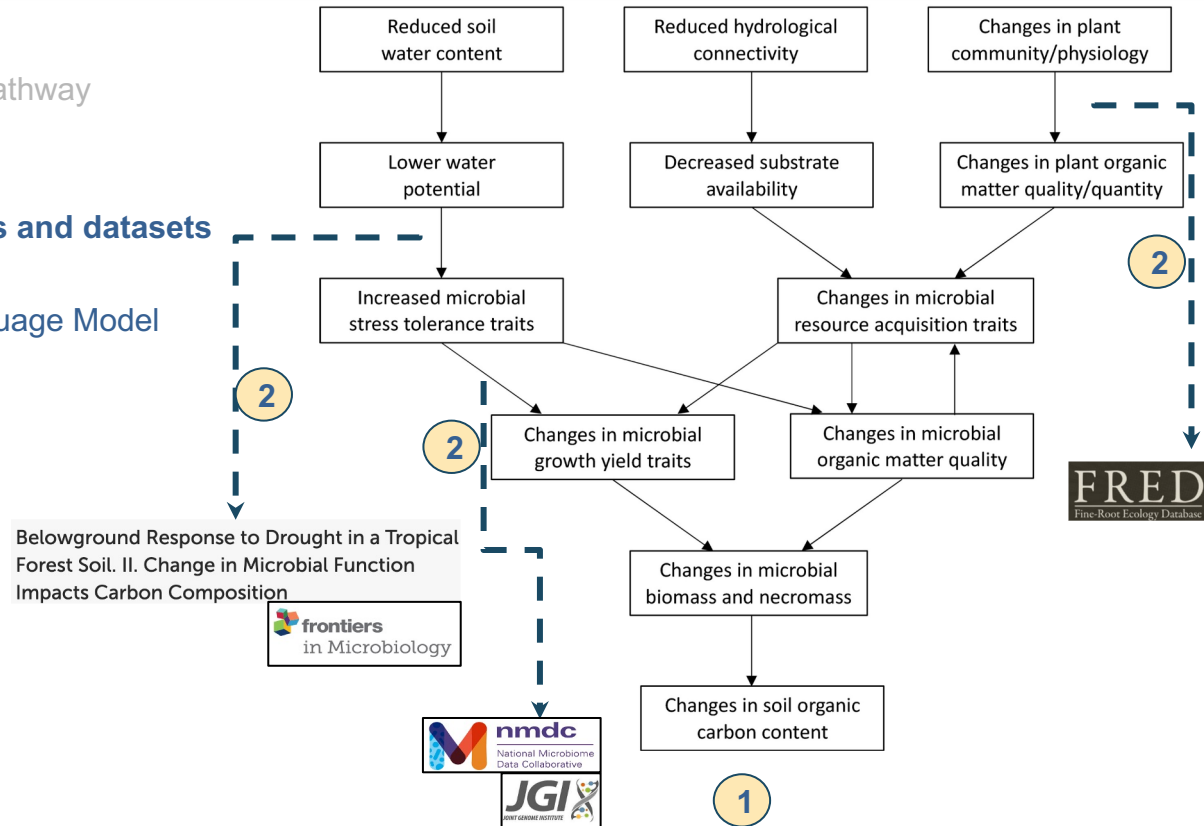
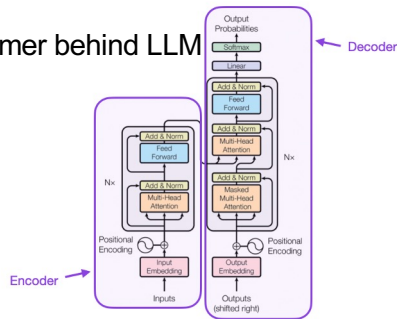
1. Curate multi-scale conceptual model

- Manually curate causally linked terms (“pathway diagram”)
- Link each box/edge to ontology concepts

2. Automatically link nodes/edges to papers and datasets

- Traditional Information-Retrieval
- Next-generation approaches: Large Language Model (LLM)

Transformer behind LLM



<https://github.com/monarch-initiative/ontogpt>
[LILLIE – doi:10.1016/j.is.2021.101938](https://doi.org/10.1016/j.is.2021.101938)

Automating Data Curation Step 3: Extract Key Variables

1. Curate multi-scale conceptual model

- Manually curate causally linked terms (“pathway diagram”)
- Link each box/edge to ontology concepts

2. Automatically link nodes/edges to papers and datasets

- Traditional Information-Retrieval
- Next-generation Large Language Model (LLM) approaches

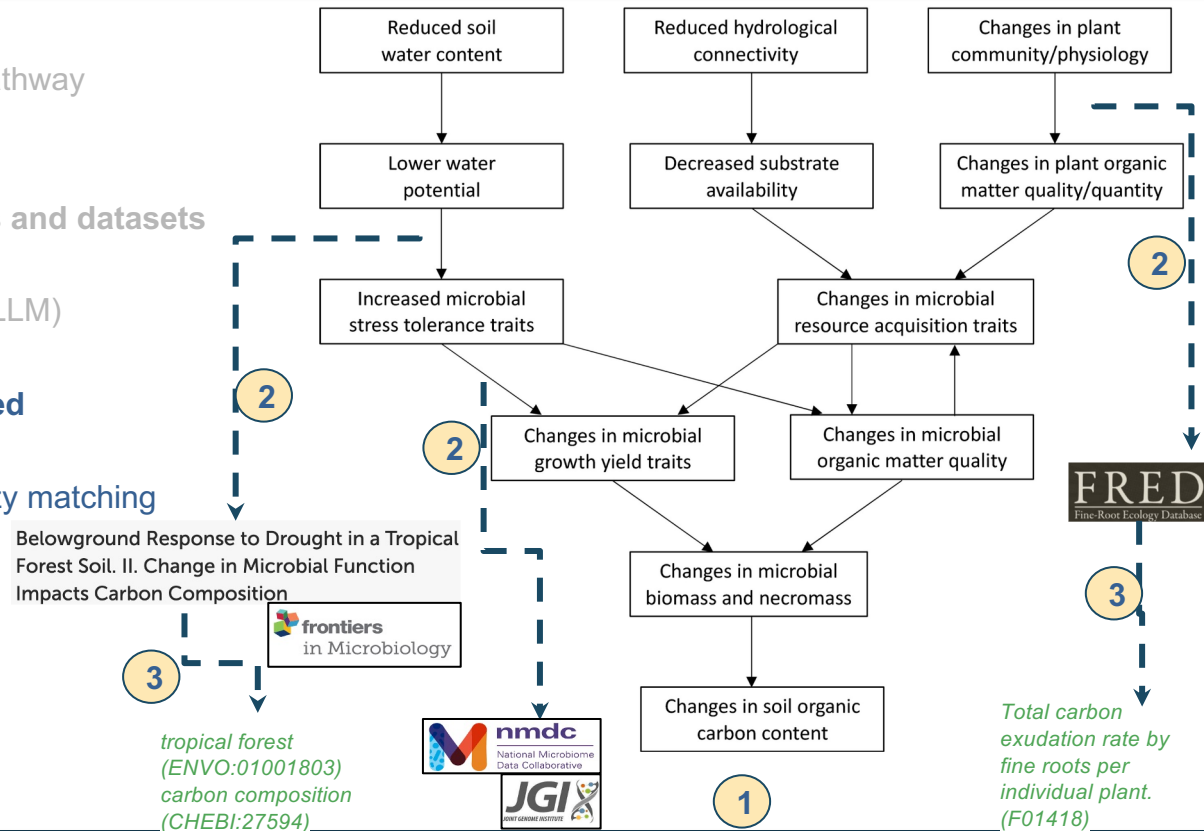
3. Automatically extract key variables studied

- E.g., moisture, drought duration
- Align to standards: entity matching → fuzzy matching

Microbiome Metadata Standards: Report of the National Microbiome Data Collaborative's Workshop and Follow-On Activities

Pajau Vangay ...Emiley A Eloë-Fadrosch

mSystems 2021 Feb 23;6(1):e01194-20



Example Information Extract Tools: OntoGPT & LILLIE

- **OntoGPT (inside LBNL, Mungall et al. 2023):** Generation of Ontologies and Knowledge Bases using GPT
 - A knowledge extraction tool that uses a large language model to extract semantic information from text.
 - This makes use of so-called *instruction prompts* in Large Language Models (LLMs) such as GPT-4.
 - SPIRES: Structured Prompt Interrogation and Recursive Extraction of Semantics
 - Zero-shot learning approach to extracting nested semantic structures from text
 - Uses text-davinci-003 from OpenAI
 - HALO: HAllucinating Latent Ontologies
 - Few-shot learning approach to generating/hallucinating a domain ontology given a few examples
 - Uses code-davinci-002 from OpenAI
- **LILLIE (outside LBNL): Information Extraction and Database Integration Using Linguistics and Learning-Based Algorithms**
 - Based on Clause-based information exchange (ClausIE), Open Information Exchange (Open IE), and Stanford CoreNLP
 - Process: Step 1 - Extract triples such as subject, predicate and object. Step 2 - Link the extracted subjects and objects to specific columns of a relational database to enrich the database.

From Keyword Search to Semantic Search: An Illustration

Google Web Images Video ^{New!} News Maps Desktop Moma mc
baker job opening Search [Advanced Search](#) [Preferences](#)

Web Results 1 - 10 of about 6,150,00

[Mt Baker School District](#)
You may also call 360-383-2075 for a applications may be downloaded from [www.mtbaker.wednet.edu/jobs/](#) - 3k - [Cached](#) - [Similar pages](#) - [Filter](#)
Mt. Baker, a school district

[CGI : Job Opening](#)
Job Seekers, Faculty & Other Researchers, Students, Journalists, Policy Makers ... Baker Institute for Animal Health, College of Veterinary Medicine ... [www.genomics.cornell.edu/jobs/view_job.cfm?id=47](#) - 15k - [Cached](#) - [Similar pages](#) - [Filter](#)

[Baker Hostetler - Staff Job Openings](#)
law business employee benefits empl regulatory litigation private wealth real [www.bakerlaw.com/Careers.aspx?Abs_WP_ID=26a8ff33-0471-4c5e-b5b7-6abdbcce0326](#) - 19k - [Cached](#) - [Similar pages](#) - [Filter](#)
Baker Hostetler, a company

[Baker & McKenzie || Careers || Current Openings ||](#)
We are always looking for talented, internationally minded people interested in building their careers with a truly global law firm. [www.bakernet.com/BakerNet/Careers/Current+Openings/](#) - 64k - [Cached](#) - [Similar pages](#) - [Filter](#)

[Current Job Opening Search](#)
Click the search button to see all job Architectural Drafting Intern, Architectural Project Leader ... [hyveenet.hy-vee.com/applynow/](#) - 75k - [Cached](#) - [Similar pages](#) - [Filter](#)
Baker, a job opening

[Law Enforcement Job Submission](#)
Advertise Your Job Openings ... -Mia Baker, Human Resources Officer, Amtrak ... You can announce your job opening to thousands of potential applicants at a ... [www.policeemployment.com/joblisting/](#) - 10k - [Cached](#) - [Similar pages](#) - [Filter](#)

Input text:

THY1 is overexpressed in human gallbladder carcinoma.

Extracted Triple:

- **Subject:**
 - Text: THY1
 - Linked Entity: *gene:thy1*
- **Predicate:** is overexpressed in
- **Object:**
 - Text: human gallbladder carcinoma
 - Linked Entity: *uberon:0002110*

Ontology-linked Triple:

gene:thy1 ; is overexpressed in ; *uberon:0002110*



LILLIE extracts useful information



Keyword search produces mixed results

Potential Collaboration Topics

- **Efficient IO for scientific data**
 - Leverage PDC (LBNL), H5Bench (LBNL), and Pegasus PMEM (Tsukuba)
 - ML workflow from cosmology?
- **Entity matching, data integration, and beyond**
 - Automating data curation requires efficient entity matching and Tsukuba scientists have worked on efficient algorithms
 - Fuzzy matching?
- **Thick-restart Lanczos method for eigenvalue computation**
- ...