Cyberinfrastructure for Scientific Data Preservation and Image Similarity Search

Yichen Guo[1,2], Yifan Zhang[3], Julian Goddy[2], Kio Polson[4], Kaushik Jagini[3], Joshua Brown[5], Marina Potapova[6], Chad Peiper[4], Jane Greenberg[4], Joshua Agar[2], Jeff Heflin[3]

[1]Lehigh University Department of Materials Science and Engineering; [2]Drexel University Department of Mechanical Engineering and Mechanics; [3]Lehigh University Department of Computer Science and Engineering; [5]Drexel University College of Computing and Informatics; [5]Oak Ridge National Laboratory Data Lifecycles and Scalable Workflows Group; Department of Biodiversity, [4]Earth and Environmental Science Academy of Natural Sciences, Drexel University

Cyberinfrastructure Challenges for Experimental Sciences

Most Data is Underanalyzed

- Data analysis takes much longer than acquisition → Analysis takes weeks-months
- Data is generally only accessible by originator

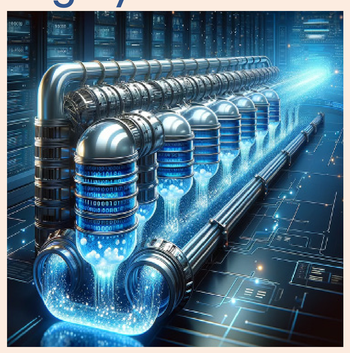
Data is not FAIR

- Science is distributed; it is rare that data is collated → Most data is saved in folders in local file systems
- Sharing between institutions is challenging

Cyberinfrastructure is not Science

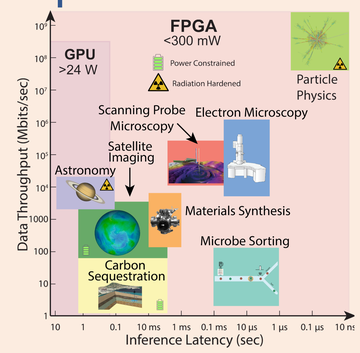
- Experimental scientists have training for functional computational literacy → Minimal support for software development
- Software contributions are undervalued

Computation is Rarely Highly-Available



- Compute infrastructure is designed for simulations not experiments → Experiments cannot wait in a queue
- Need for high-availability infrastructure

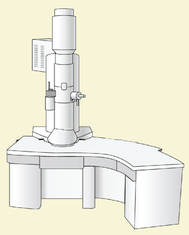
Non-Deterministic Computational Latency



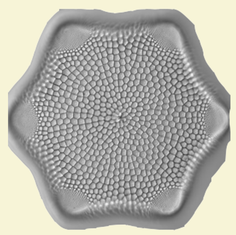
- Experimentalists rarely deploy deterministic low-latency computation → excluding dynamic process control
- Software, algorithm, hardware codesign

Scientific Data Ingestion

Atomic Force Microscopy



Electron Microscopy



Diatom Herbarium

Data → *Metadata*

Composable Containerized Backend

DataFed Client

- Data and metadata preservation
- Access control
- Disaster recovery

File Transfer

globus

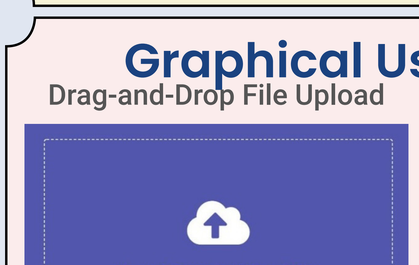
- High-performance file transfer
- Management and logging
- Fine-grained access control

docker Compose

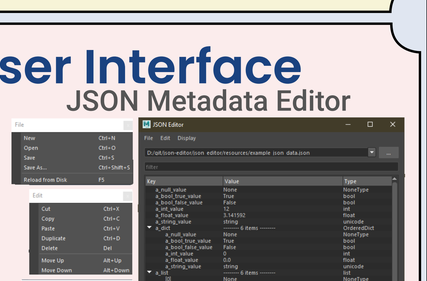
Testing

Graphical User Interface

Drag-and-Drop File Upload



JSON Metadata Editor

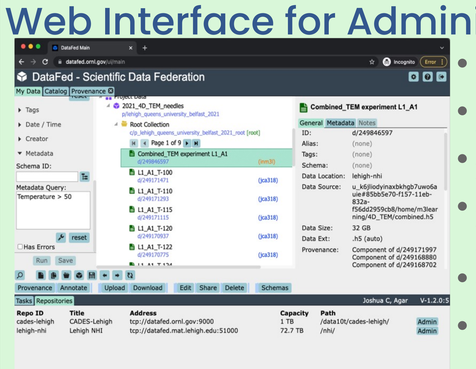


Metadata Extraction Engine

Development

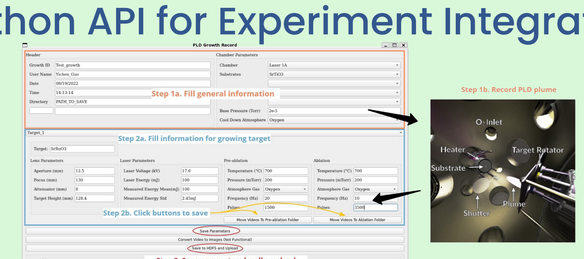
DataFed

Web Interface for Administration and Search



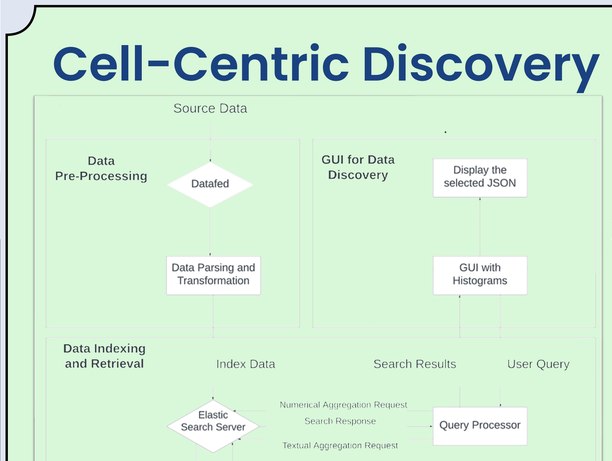
- Federated scientific data management system
- Read, write, and admin control at the user and group level
- Automated file collation and transfer via Globus
- Secure access controlled file transfer between institutional firewalls
- Standard schemes as complex graph relational queries
- Fully functional command line interface and Python API

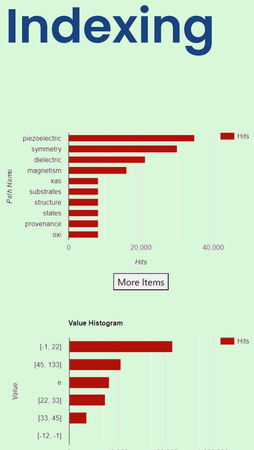
Python API for Experiment Integration



Production

Cell-Centric Discovery Indexing



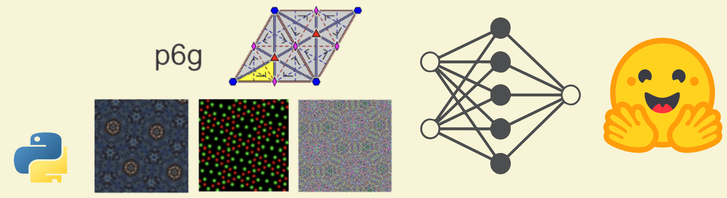


- Novel approach to exploring collections of semi-structured data by extending the cell-centric indexing approach
- Easy data access and retrieval without knowledge of schema or data organization
- User-friendly interface for data exploration and retrieval

Production

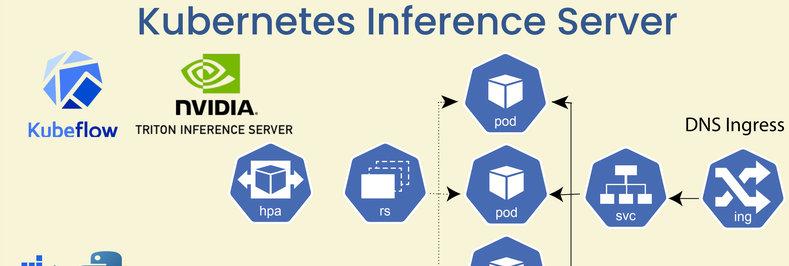
AI Similarity Engine

Domain Specific Models



- Model libraries trained on collections of scientific images
- Models trains on domain specific tasks (e.g., classifying wallpaper symmetry

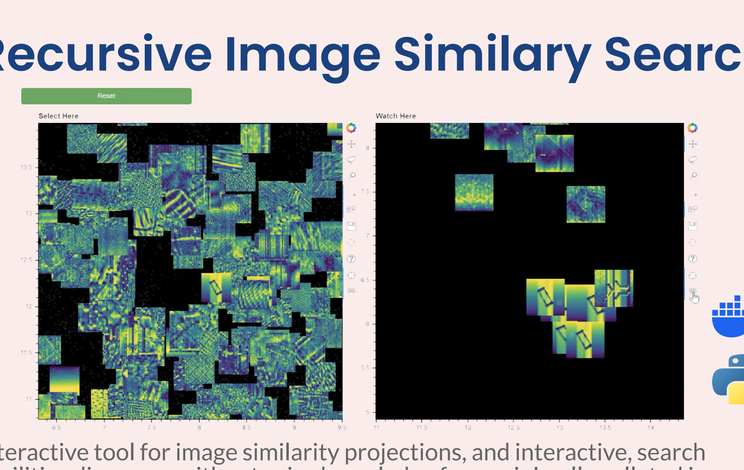
Kubernetes Inference Server



Testing

- Horizontally scalable low-latency inference servers for dynamic loads
- Allows for rolling model deployment and updates with 0 down time

Recursive Image Similary Search



- Interactive tool for image similarity projections, and interactive, search
- Facilities discovery without prior knowledge from minimally collated image data

Development

Funding

Primary

Elements: CRISPS: Cell-Centric Recursive Image Similarity Projection Searching
PIs: Joshua Agar, Jeff Heflin; 2209135

Secondary

NSF: MRI: Development of Heterogeneous Edge Computing Platform for Real-Time Scientific Machine Learning (2215789)
NSF: MRI: Development of a Platform for Accessible Data-Intensive Science and Engineering (2320600)
DOE: Real-time Data Reduction Codesign at the Extreme Edge for Science
ARL: Collaborative for Hierarchical Agile and Responsive Materials (CHARM) (W911NF-19-2-0119)