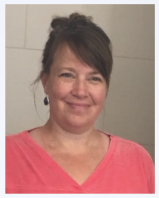
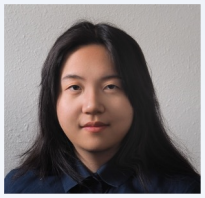
Enabling Scientific Discovery: Harnessing the Power of the National Science Data Fabric for Large-Scale Data Analysis (Session I & II)

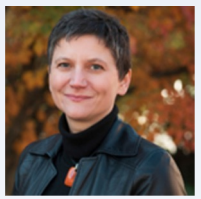
Presenters: **Valerio Pascucci, Amy Gooch, Aashish Panta, Xuan Huang, Alper Sahistan, Giorgio Scorzelli** [1]

Other contributors: Michela Taufer, Jack Marquez, Heberth Martinez , Paula Olaya, Gabriel Laboy, Jay Ashworth [2]

[1] University of Utah, [2] University of Tennessee Knoxville



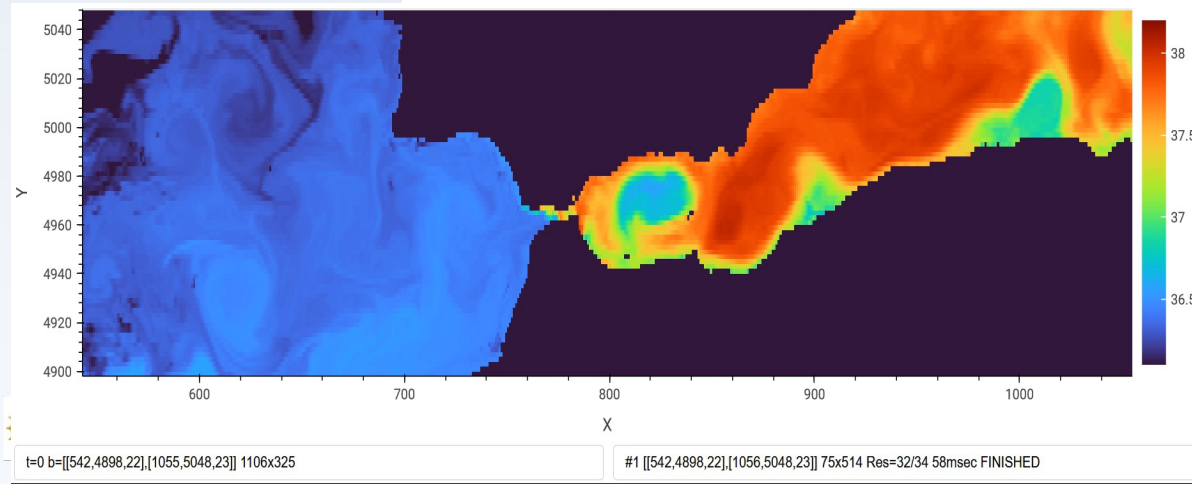
Acknowledgments

The authors of this tutorial would like to express their gratitude to:

- NSF through awards 2138811, 2103845, 2334945, 2138296, and 2331152
- Dataverse
- Seal Storage
- Rodrigo Vargas, Vargas Lab, University of Delaware
- Werner Sun, CHESS, Cornell University
- DOE SBIR Phase II award DE-SC0017152

Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

National Science Data Fabric



Schedule

The half-day tutorial is organized into four sessions:

Session I (15 mins):

This session begins with an overview of the NSDF and addresses users' challenges identified through interviews.

Session II (1 hour):

This session offers a hands-on experience with NSDF services, focusing on visualization and dashboard creation for Earth science datasets.

Session III (1 hour):

This session delves deeper into NSDF services tailored for the management and analysis of datasets exceeding 1PB.

Session IV (15 mins):

This session concludes with an interactive Q&A, allowing attendees to discuss applications of NSDF in various research fields.

Prerequisites

Tutorial

Step 0: Access to GitHub

To run this tutorial, you need to have a GitHub account.

➔ You can create one following the instructions here:

https://docs.github.com/en/get-started/start-your-journey/creating-an-account-on-github#

➔ Now you can login into GitHub

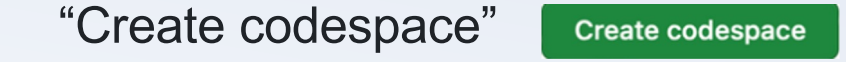
https://github.com/login

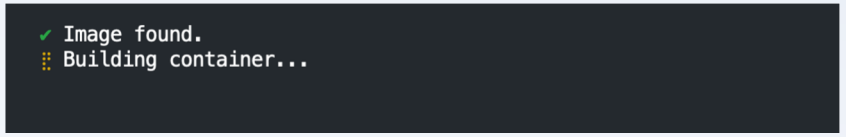
Step 1: Create Codespaces

Use your GitHub account to run this tutorial with GitHub codespaces

➔ Access this link:
NSDF Tutorial 2024

➔ Click on green button "Create codespace"





Tutorial Goals

This tutorial demonstrates end-to-end analysis of scientific data through National Science Data Fabric (NSDF) services

Tutorial Goals

- **Construct a modular workflow** that combines your application components with NSDF services
- **Upload, download, and stream data** to and from **public and private storage** solutions
- Deploy the NSDF dashboards for large-scale **data access, visualization,** and **analysis**

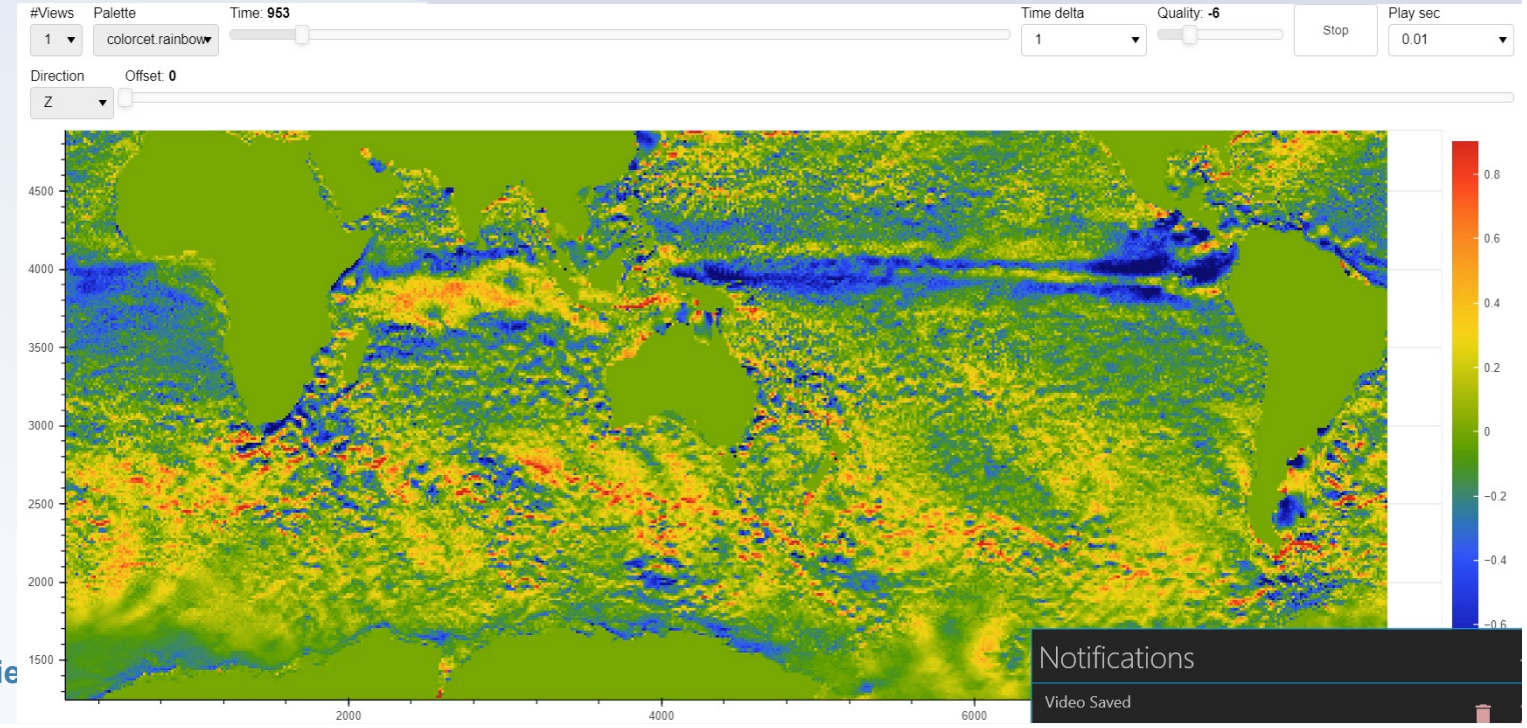
2026 IEEE SciVis CONTEST

ViSOAR

Powered by ViSUS

$1,000 Cash Prize !!!!

Visualizing the future of climate science, one dataset at a time



NSDF National Scie

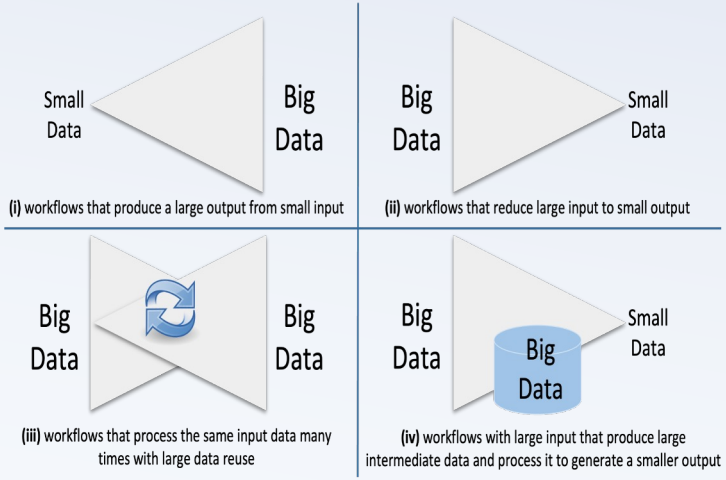
Session I: Understanding and Addressing User's Pain Points

Surveying Community Needs and Realities

NSDF Technology Addresses Key Pain Points in Data-Intensive Science

*NSDF focuses on the main classes of workflows that **CHALLENGE** data-intensive scientific investigations*

AI/ML and Data Analytics Workflows



Scalability Software stack scale from leadership computing to commodity hardware (even handheld)

Information Streams federated resources that minimize data transfers with trustworthy sharing of information

Resource Efficiency Allows teams to work effectively with limited access to human and physical resources

Data Management Standardized data and metadata management tools avoid replicated work

Accessibility Facilitate data-sharing processes for open and secure environments

Timeliness Immediate access and use of remote information without bulk data transfers

Workforce Development Trained of CI professionals

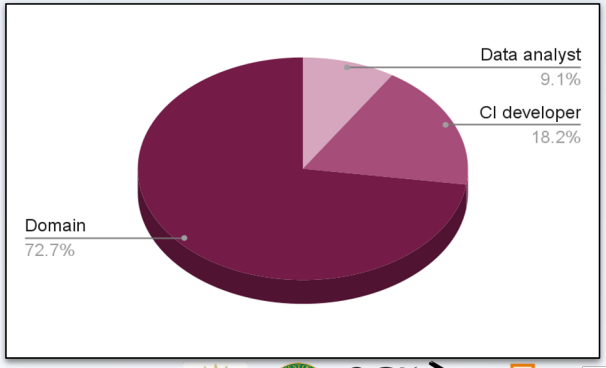
Replicability Programs/data versioning with FAIR identifiers throughout the scientific investigation

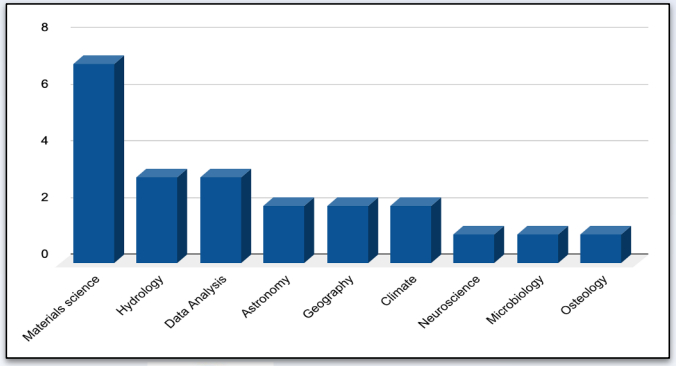
NSDF National Science Data Fabric

www.sci.utah.edu

Implementing the NSDF Vision: User Interviews

Identify Users	Target Questions	Analyze Results	Diagnose Pain Points
• **Diverse roles:** Domain scientists, CI professionals, developers • **Diverse domains:** Materials science, climate, earth sciences, astronomy, and more! • **Diverse institutions:** R1 universities, teaching colleges, MSIs, national labs, experimental facilities	• General questions about data storage, data form, metadata storage, WMS, data catalogs, programming languages • Specific questions about unique challenges related to role, domain, and institution	• Identify cross-cutting concerns • Identify concerns consistent for roles, domains, and institutions	• Distill concerns into concrete problem statements • Translate into objectives, actionable items, and milestones

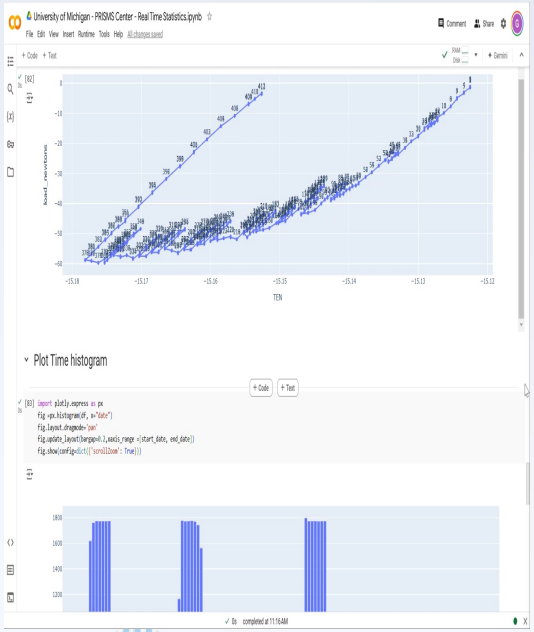




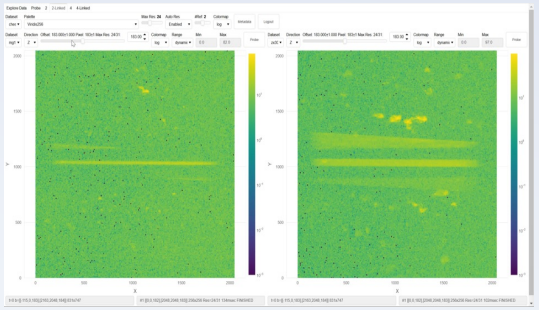
NSDF Currently Highlights 4 Main Technologies

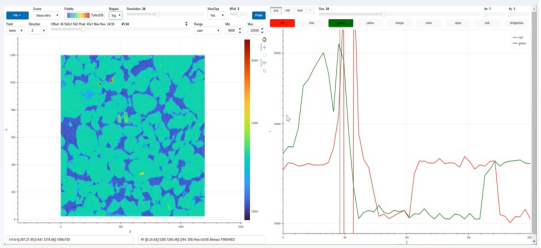
(Based on User Experiences)

Support of Experiment and Simulation Data

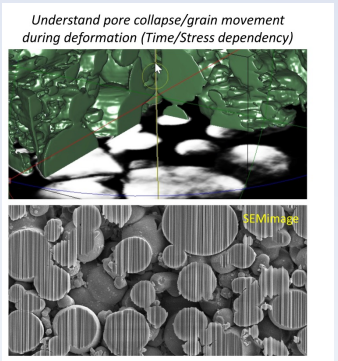


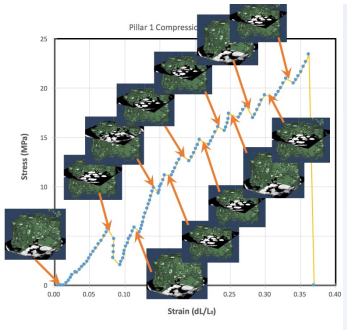
Science-Driven Data Analysis and Exploration



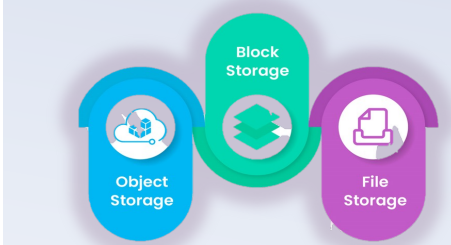


AI/ML Workflows





Data Use in Distributed Environment



NSDF National Science Data Fabric

www.sci.utah.edu

Session I: Implementing an Accessible & Tightly Integrate Data Fabric

Designing, developing, and deploying equitable services

We are building a **holistic ecosystem** to **democratize data-driven scientific discovery** by **connecting an open network of institutions**, including minority-serving institutions, with **a shared, modular, containerized data delivery environment**.

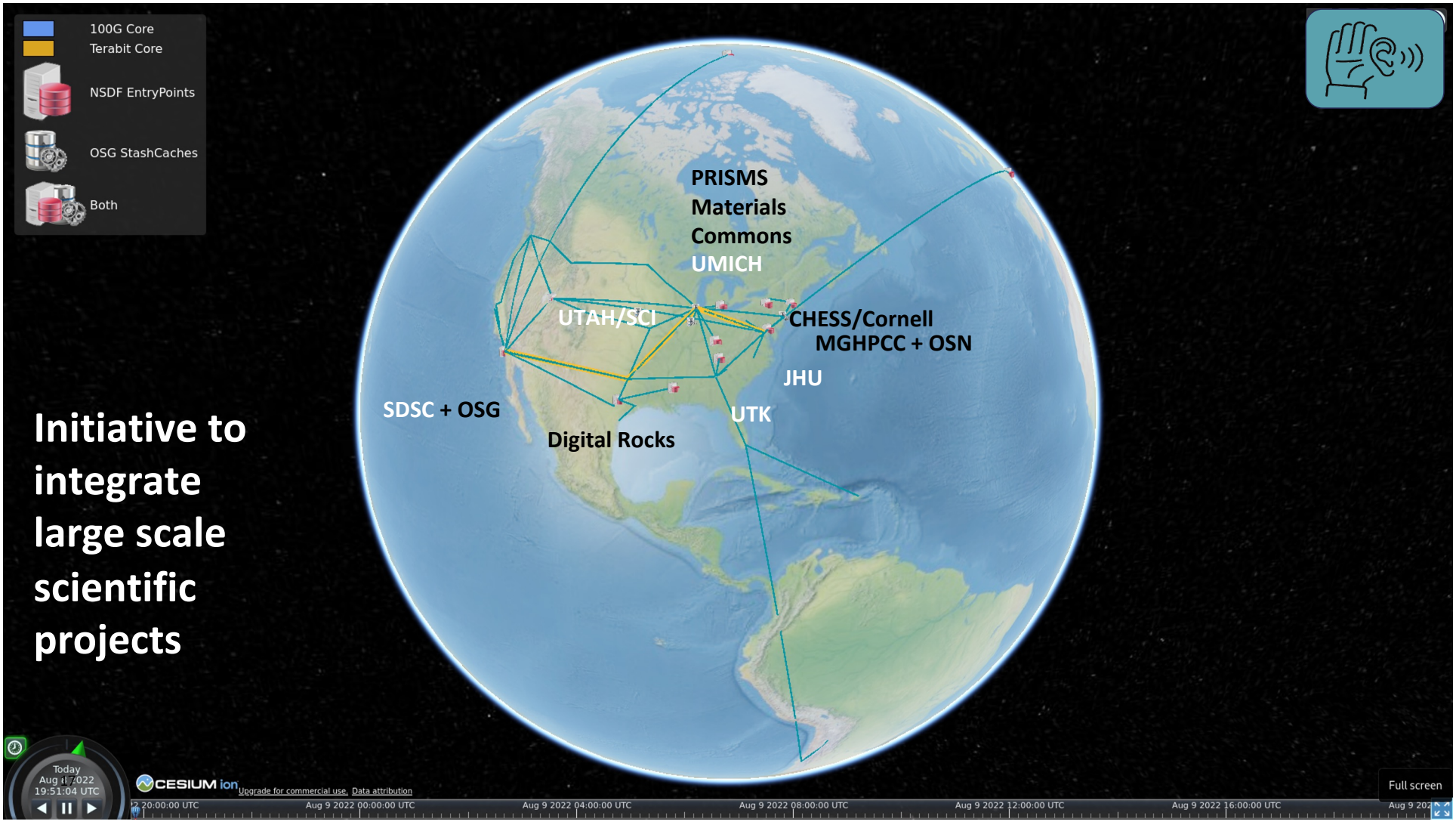
NSDF

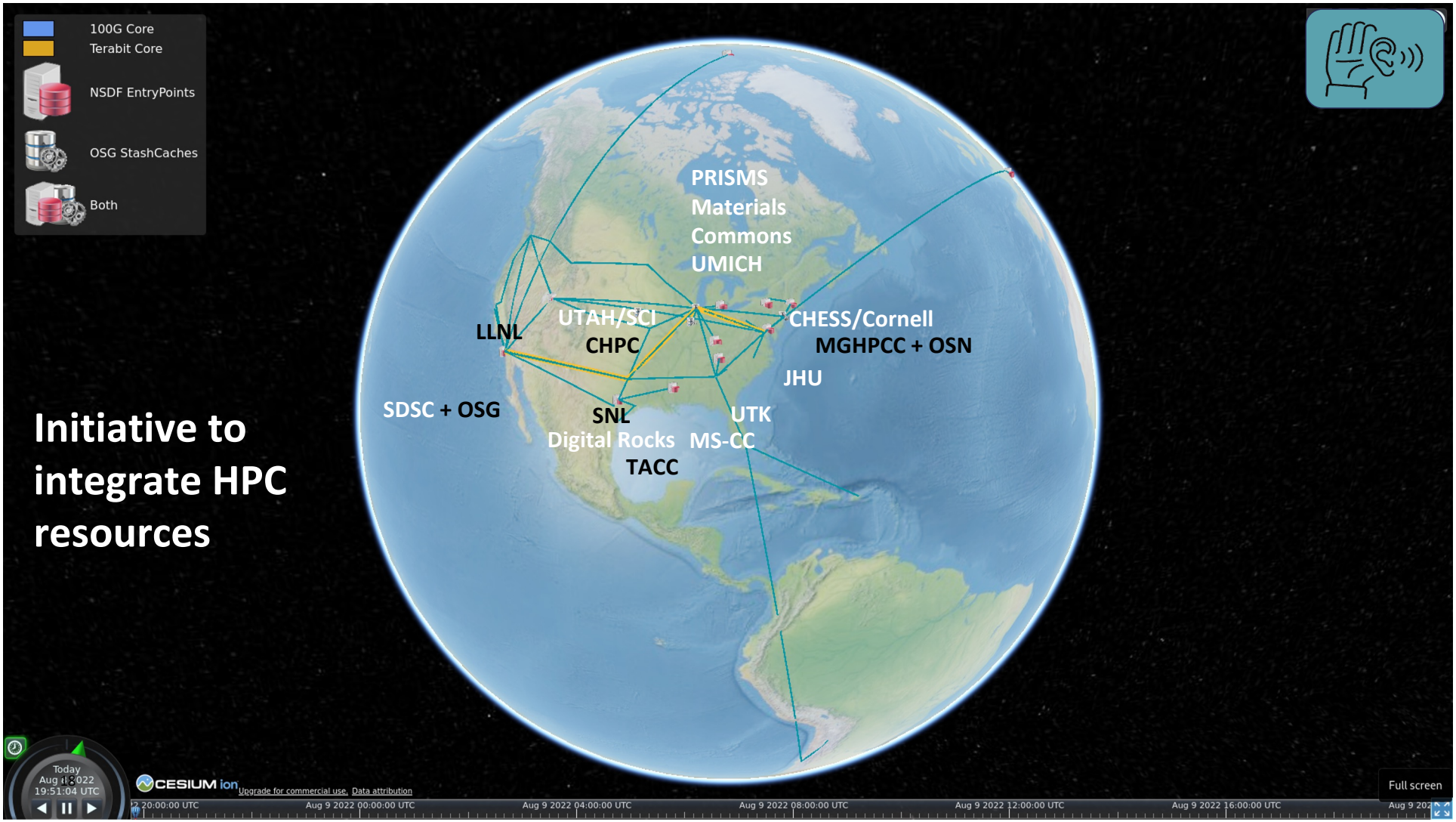
NSDF National Science Data Fabric

http://nationalsciencedatafabric.org/


100G Core
Terabit Core
NSDF EntryPoints
OSG StashCaches
Both
UMICH
UTAH
JHU
SDSC
UTK
Institutions and universities with resources to share
Today
Aug 8 2022
19:51:04 UTC
CESIUM ion
Upgrade for commercial use.
Data attribution
Aug 9 2022 00:00:00 UTC
Aug 9 2022 04:00:00 UTC
Aug 9 2022 08:00:00 UTC
Aug 9 2022 12:00:00 UTC
Aug 9 2022 16:00:00 UTC
Full screen


100G Core
Terabit Core
NSDF EntryPoints
OSG StashCaches
Both
UMICH
UTAH
UTEP
UC Denver
JHU
SDSC
UTSA
UTK
Dillan U.
Initiative to integrate minority serving institutions
Today
Aug 8 2022
19:51:04 UTC
CESIUM ion
Upgrade for commercial use.
Data attribution
Aug 9 2022 00:00:00 UTC
Aug 9 2022 04:00:00 UTC
Aug 9 2022 08:00:00 UTC
Aug 9 2022 12:00:00 UTC
Aug 9 2022 16:00:00 UTC
Full screen

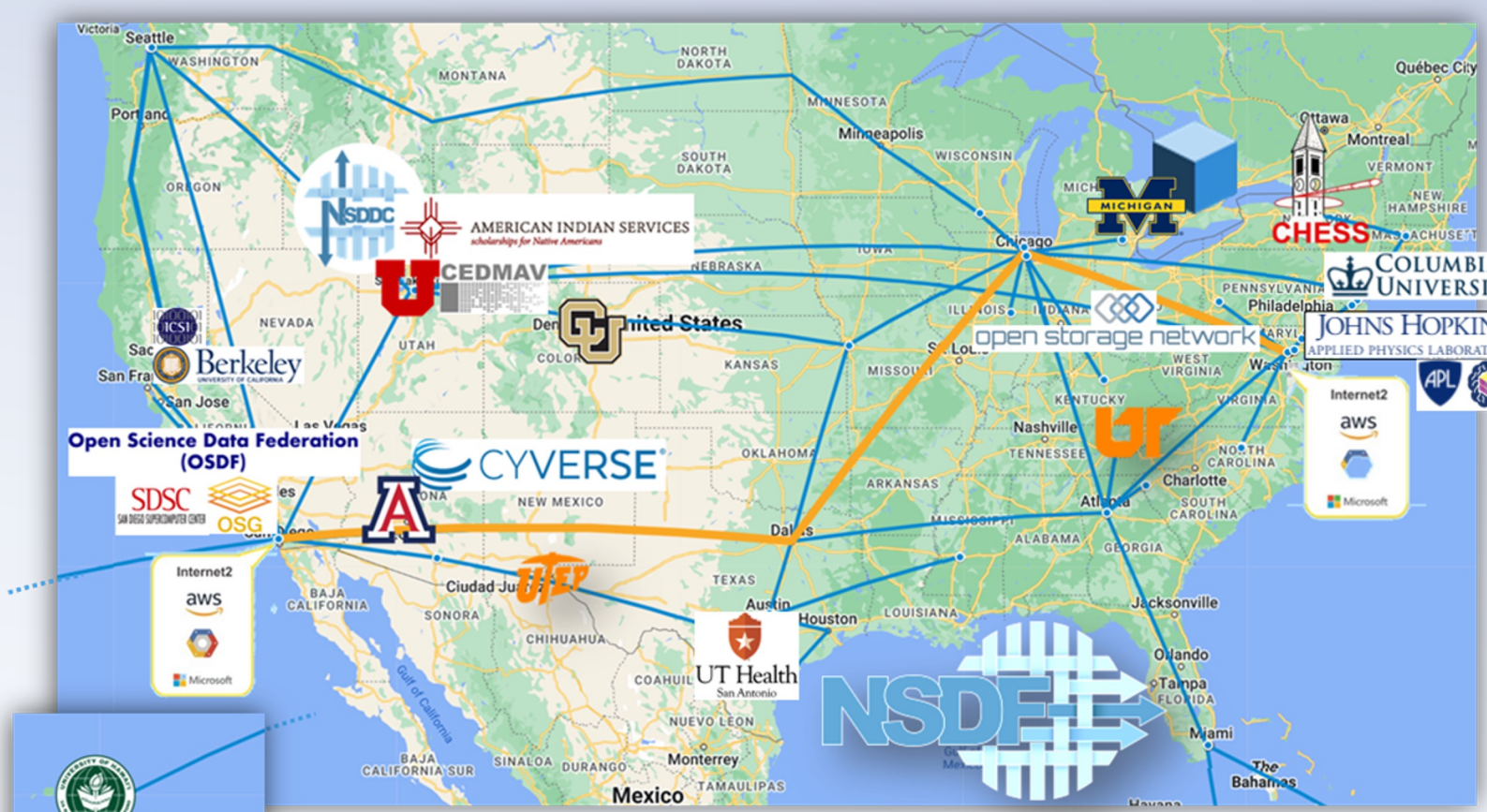

100G Core
Terabit Core
NSDF EntryPoints
OSG StashCaches
Both
PRISMS
Materials
Commons
UMICH
UTAH/SCI
CHESS/Cornell
MGHPCC + OSN
JHU
SDSC + OSG
UTK
Digital Rocks
Initiative to integrate large scale scientific projects
Today
Aug 11 2022
19:51:04 UTC
CESIUM ion
Upgrade for commercial use.
Data attribution
Aug 9 2022 00:00:00 UTC
Aug 9 2022 04:00:00 UTC
Aug 9 2022 08:00:00 UTC
Aug 9 2022 12:00:00 UTC
Aug 9 2022 16:00:00 UTC
Full screen


100G Core
Terabit Core
NSDF EntryPoints
OSG StashCaches
Both
PRISMS
Materials
Commons
UMICH
LLNL
UTAH/SCI
CHPC
CHESS/Cornell
MGHPCC + OSN
JHU
SDSC + OSG
SNL
Digital Rocks
UTK
MS-CC
TACC
Initiative to integrate HPC resources
Today
Aug 9 2022
19:51:04 UTC
CESIUM ion
Upgrade for commercial use.
Data attribution
Aug 9 2022 00:00:00 UTC
Aug 9 2022 04:00:00 UTC
Aug 9 2022 08:00:00 UTC
Aug 9 2022 12:00:00 UTC
Aug 9 2022 16:00:00 UTC
Full screen


100G Core
Terabit Core
NSDF EntryPoints
OSG StashCaches
Both
Initiative to integrate public cloud resources
PRISMS
Materials
Commons
UMICH
LLNL
UTAH/SCI
CHPC
CHESS/Cornell
MGHPCC + OSN
JHU
Internet2
SDSC + OSG
SNL
UTK
MS-CC
TACC
XSEDE/ACCESS
Jetstream2
CloudLab
Chameleon
CyVerse
Today
Aug 9 2022
19:51:04 UTC
CESIUM ion
Upgrade for commercial use.
Data attribution
Aug 9 2022 00:00:00 UTC
Aug 9 2022 04:00:00 UTC
Aug 9 2022 08:00:00 UTC
Aug 9 2022 12:00:00 UTC
Aug 9 2022 16:00:00 UTC
Full screen


100G Core
Terabit Core
NSDF EntryPoints
OSG StashCaches
Both
PRISMS
Materials
Commons
UMICH
LLNL
UTAH/SCI
CHPC
CHESS/Cornell
MGHPCC + OSN
JHU
Internet2
SDSC + OSG
SNL
UTK
MS-CC
TACC
XSEDE/ACCESS
Jetstream2
CloudLab
Chameleon
IBM Cloud
CyVerse
CloudBank
(AWS, Azure)
Storage Providers
(Seal, MinIO)
Initiative to integrate enterprise cloud and storage resources
Today
Aug 9 2022
19:51:04 UTC
CESIUM ion
Upgrade for commercial use.
Data attribution
Aug 9 2022 00:00:00 UTC
Aug 9 2022 04:00:00 UTC
Aug 9 2022 08:00:00 UTC
Aug 9 2022 12:00:00 UTC
Aug 9 2022 16:00:00 UTC
Full screen


100G Core
Terabit Core
NSDF EntryPoints
OSG StashCaches
Both
A **data fabric** must be **accessible and tightly integrated**
to coordinate data movement and use
among geographically distributed teams or organizations
IceCube
Materials
UMICH
UTAH/SCI
CHESS/Cornel
LLNL
CHPC
MGHPCC + OSN
JHU
Internet2
SDSC + OSG
SNL
UTK
MS-CC
TACC
XSEDE/ACCESS
Jetstream2
CloudLab
Chameleon
CyVerse
IBM Cloud
CloudBank
MINORITY SERVING-
CYBERINFRASTRUCTURE
CONSORTIUM
Today
Aug 9, 2022
19:51:04 UTC
CESIUM ion
Upgrade for commercial use.
Data attribution
Full screen



National Science Data Fabric

National Science Data Democratization Consortium

NSDDC

Protocol Labs

WEKA

ViSOAR
Powered by ViSUS

IBM Cloud

CLOUDFLARE

ALLUXIO

DoubleCloud

NSDF National Science Data Fabric

www.sci.utah.edu

THE UNIVERSITY OF TENNESSEE KNOXVILLE

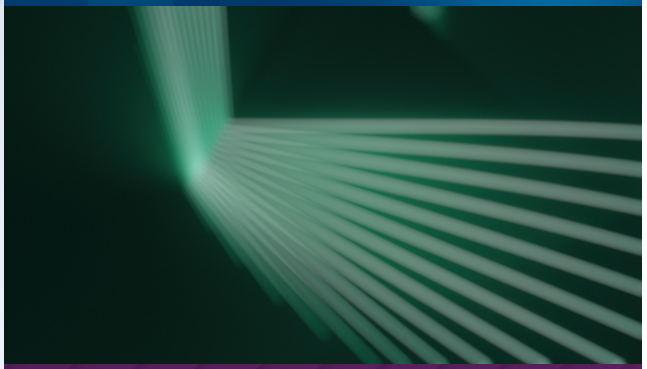
ViSOAR Powered by ViSUS

INTEL oneAPI Center of Excellence

MINIO

Click house Server for multi-petabyte multi-federation data catalog

Petabytes of cloud storage



Seal Storage Technology
Web3 Cloud Storage

WHO WE ARE
We're cloud storage and blockchain experts with over 100 years of experience in enterprise data storage from Seagate, Oracle, Cisco, and more. By seamlessly stewarding our clients into decentralized cloud storage, we're making Web3 an accessible reality for universities, research institutes, enterprises, and Web3 firms alike.

WHY SEAL
Seal provides sustainable, immutable, and affordable data storage.

- **DATA RETRIEVAL** Access data in hours vs days
- **COST EFFECTIVE** Up to 80% less than competitors
- **SECURE** Tamper proof and verifiable

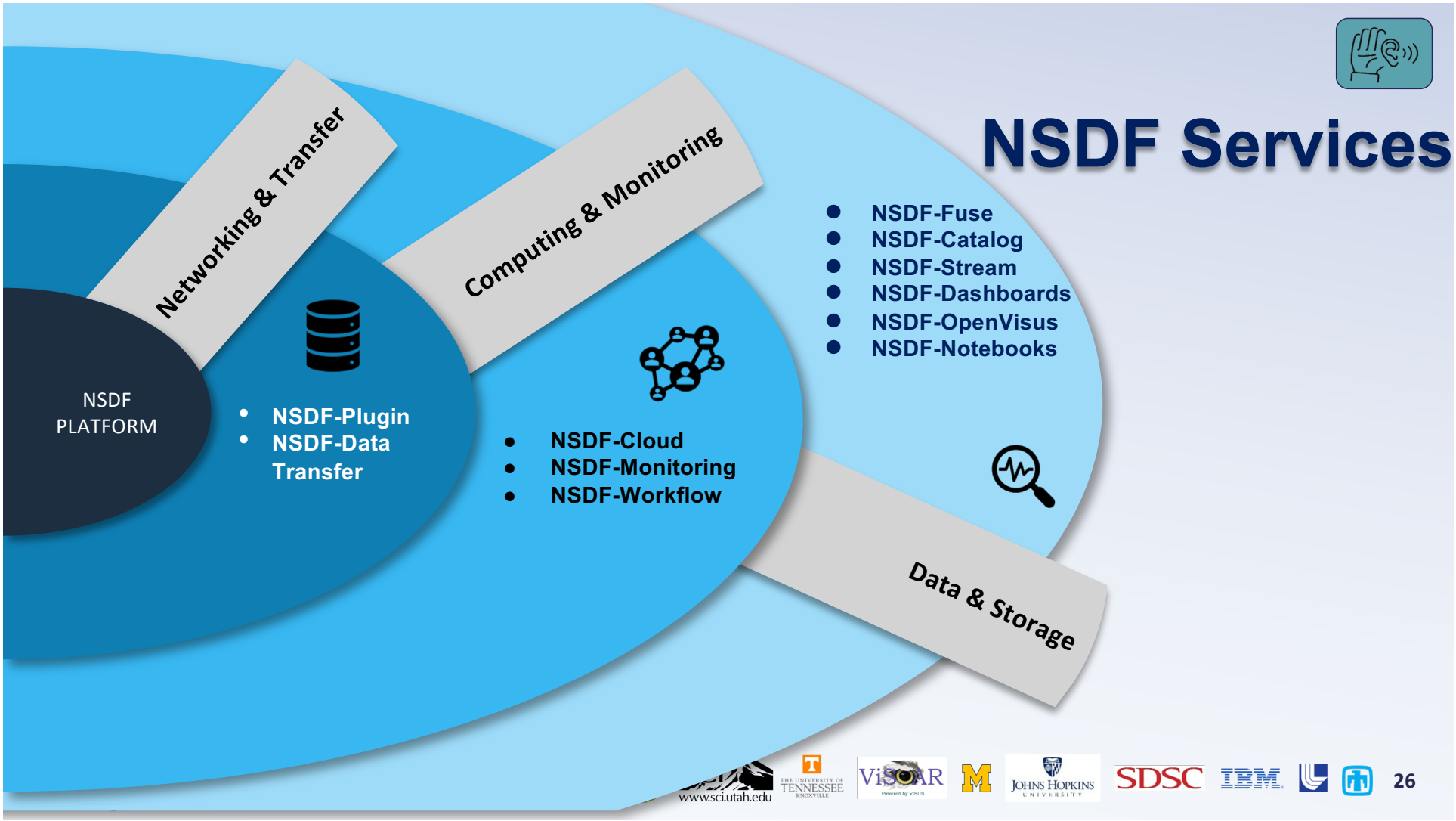
Session I: Our Services and Successful Stories

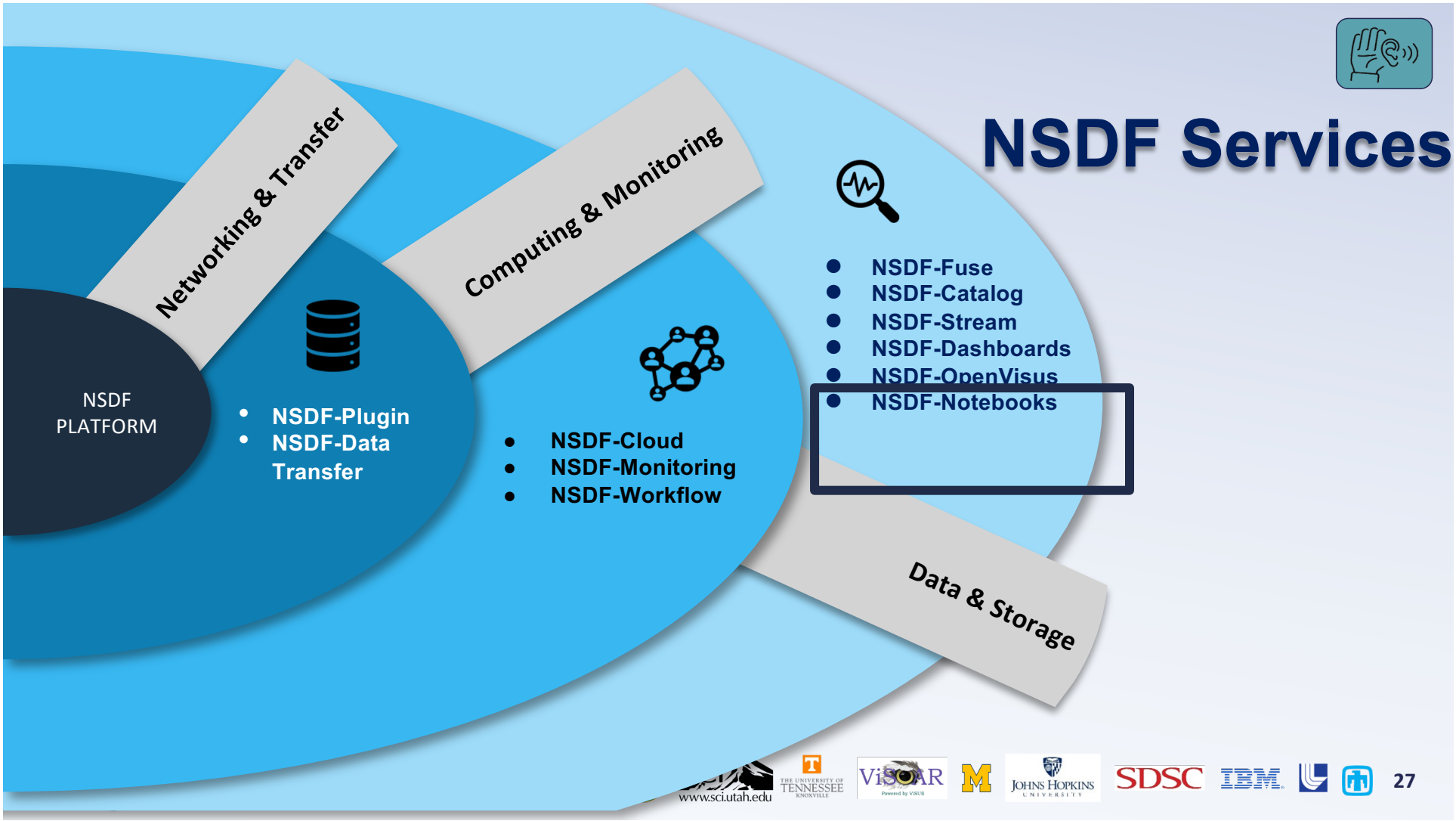
Democratizing Access and Use of Large-scale Data

A **data fabric** must be **accessible and tightly integrated** to coordinate data movement and use among geographically distributed teams or organizations

Develop a **FAIR, AI-ready, transdisciplinary software stack** that is easy to use, integrate, and scale

Develop a federated data fabric: a suite of equitable **network, computing, and storage services** interoperating across the academic and commercial cloud


NSDF Services
NSDF PLATFORM
Networking & Transfer
Computing & Monitoring
Data & Storage
NSDF-Plugin
NSDF-Data Transfer
NSDF-Cloud
NSDF-Monitoring
NSDF-Workflow
NSDF-Fuse
NSDF-Catalog
NSDF-Stream
NSDF-Dashboards
NSDF-OpenVisus
NSDF-Notebooks
www.sci.utah.edu
THE UNIVERSITY OF TENNESSEE KNOXVILLE
ViSUS
JOHNS HOPKINS UNIVERSITY
SDSC
IBM


NSDF Services
Networking & Transfer
Computing & Monitoring
Data & Storage
NSDF PLATFORM
NSDF-Plugin
NSDF-Data Transfer
NSDF-Cloud
NSDF-Monitoring
NSDF-Workflow
NSDF-Fuse
NSDF-Catalog
NSDF-Stream
NSDF-Dashboards
NSDF-OpenVisus
NSDF-Notebooks

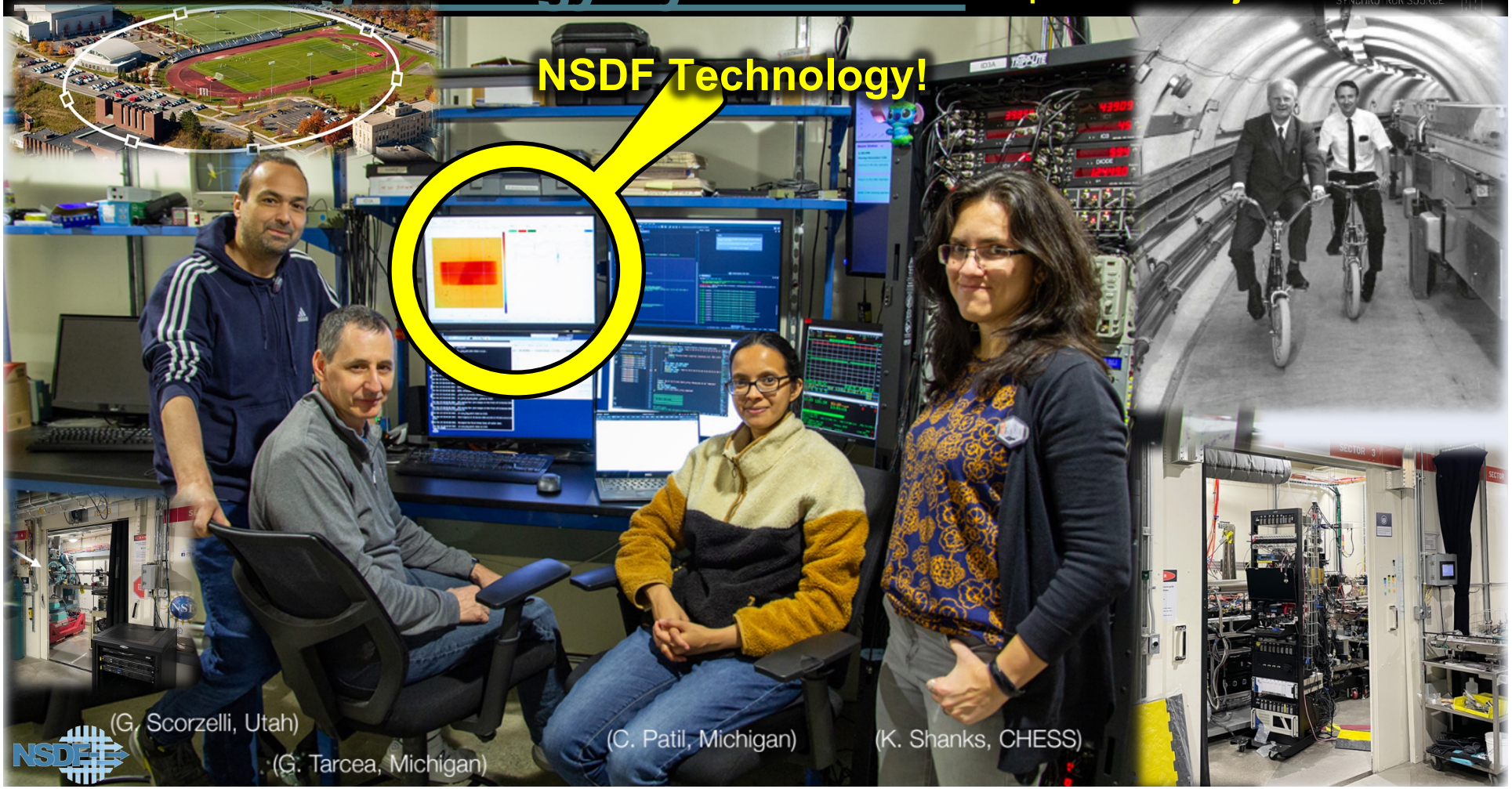
Section I: Sharing Use-Inspired Research Stories

Decentralizing Research Hubs for Transformative Scientific Discovery

Cornell High Energy Synchrotron

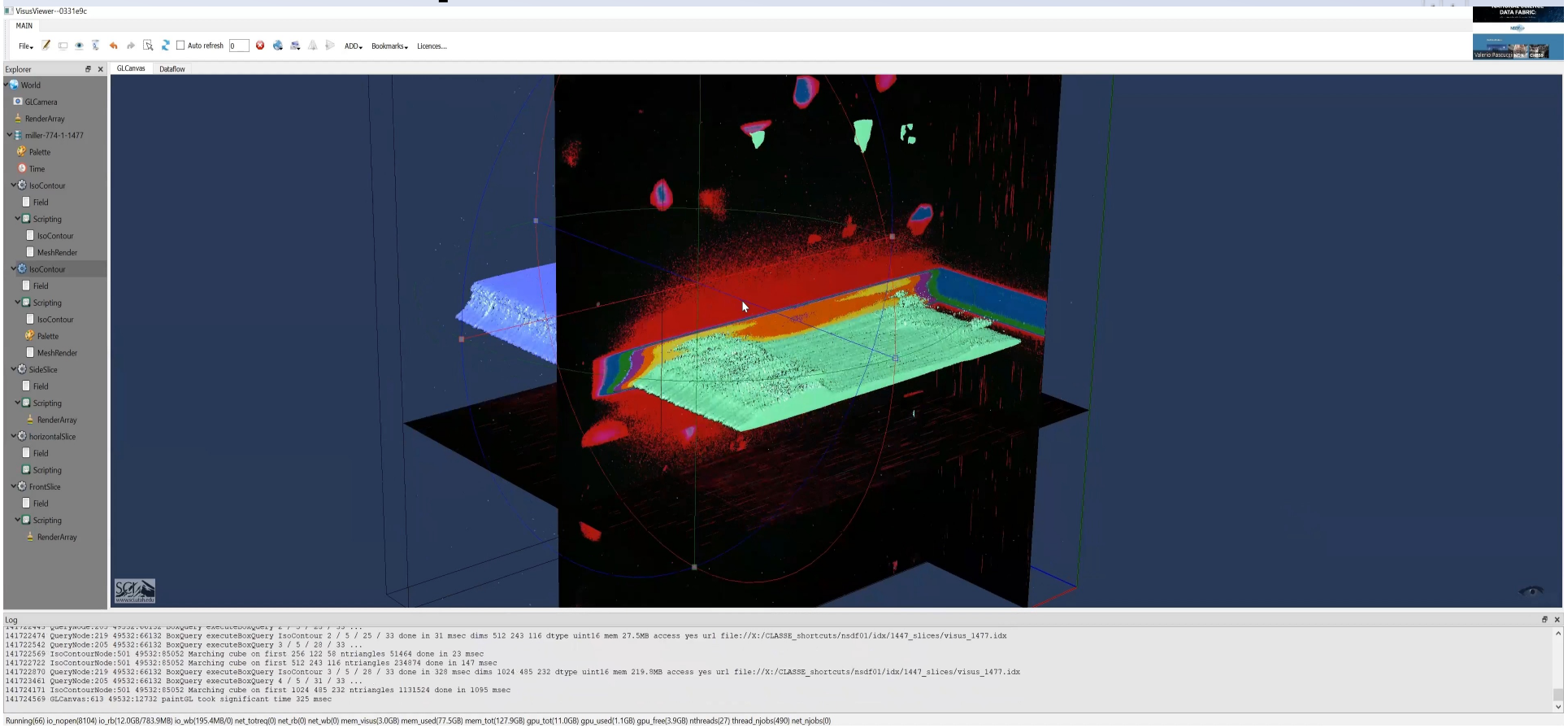
https://shorturl.at/juHL6

CHESS
CORNELL HIGH ENERGY SYNCHROTRON SOURCE



In Situ Transformation to Streamable OpenVisus Data Format

CHESS
CORNELL HIGH ENERGY SYNCHROTRON SOURCE



Kate Shanks, CHESS

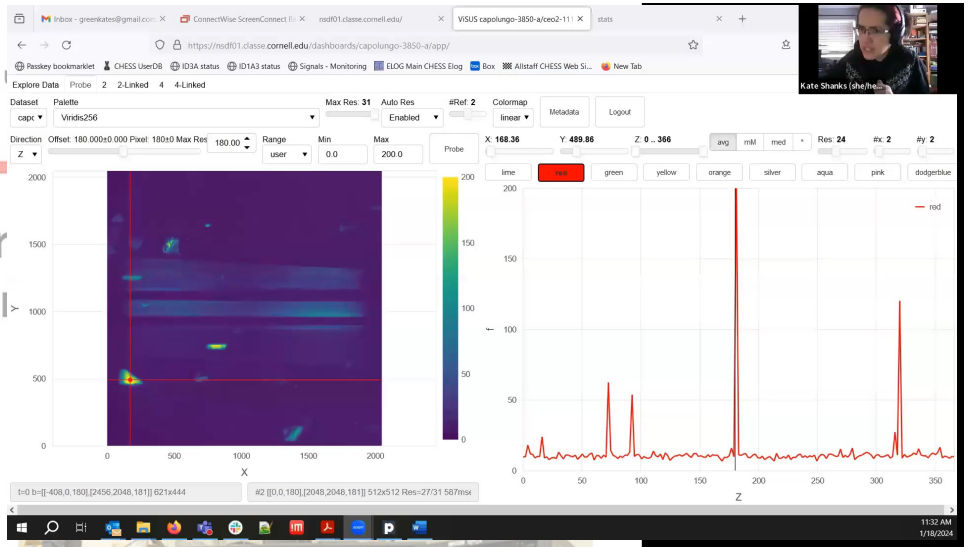
CHESS
CORNELL HIGH ENERGY SYNCHROTRON SOURCE

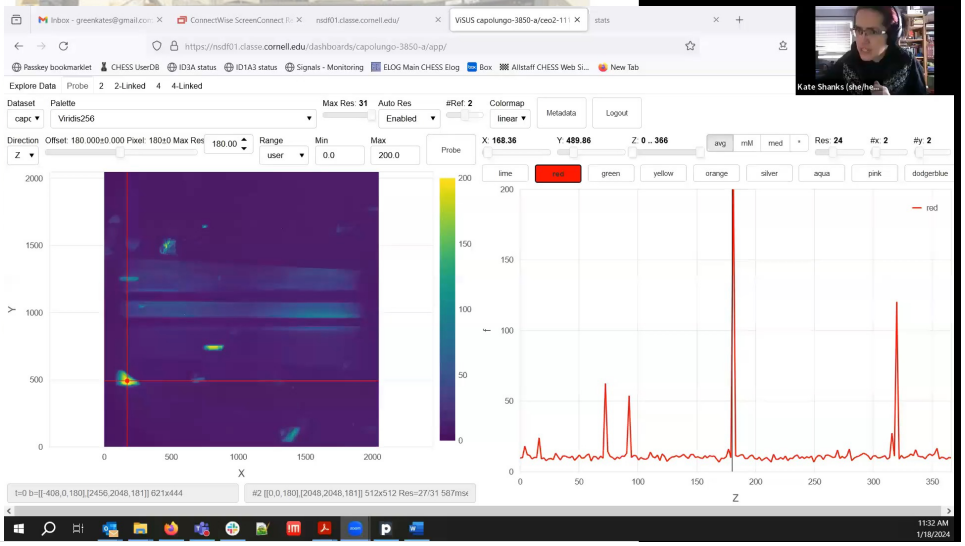
STATUS SCIENCE

National Science Data Fabr
to Democratize Data-Drive

January 11, 2024 | Savan DeSouza

med at
ntific
e Data
High
ESS)
g a
cess to
red
uting

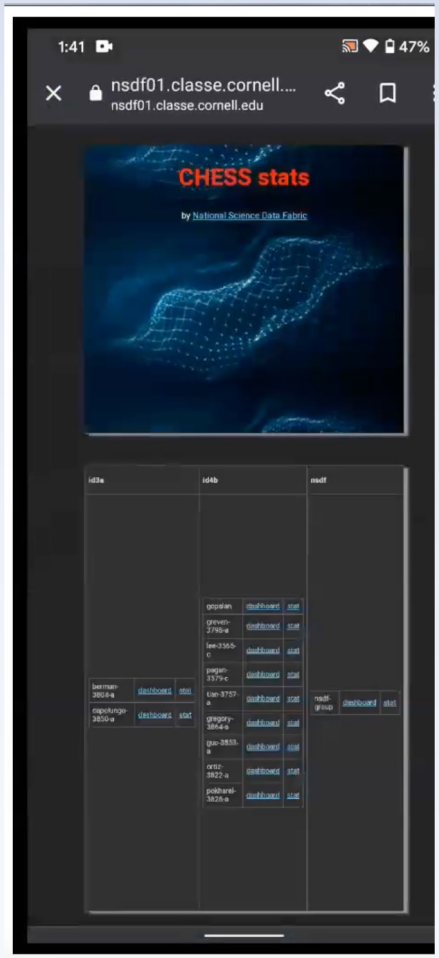


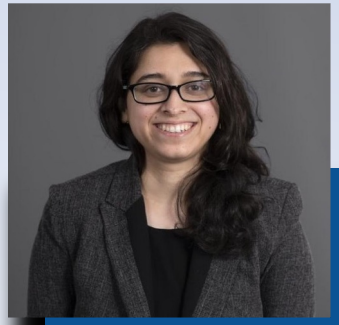


"It's much easier to pop open one of these image stacks on here, in your web browser, as opposed to trying to open it up on ImageJ. If you do that on the station computer, sometimes you crash the computer. So this decouples that and lowers the barrier to examining the datasets on the fly.."

-Kate Shanks, FAST beamline scientist

https
natio
and-
dem
scien

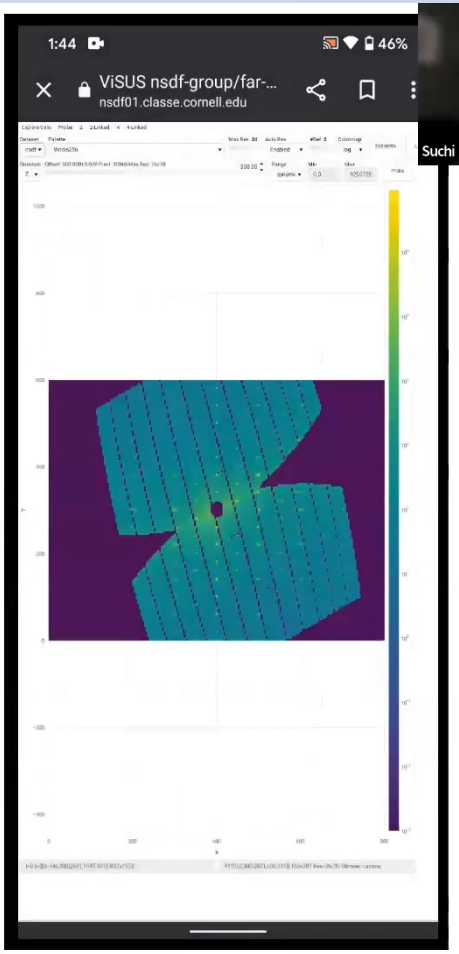


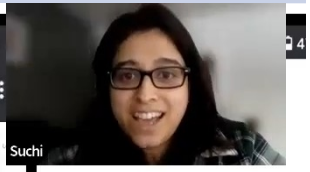


"You can access data from anywhere on any devices."

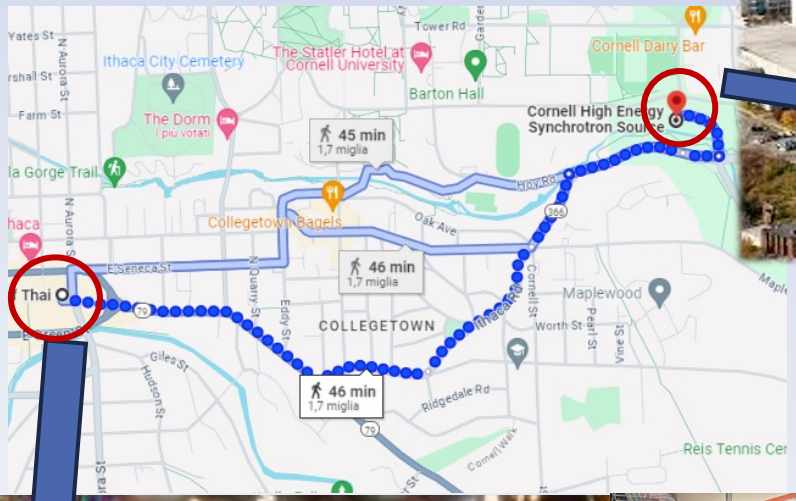
– Suchismita (Suchi) Sarkar
Staff Scientist, CHESS

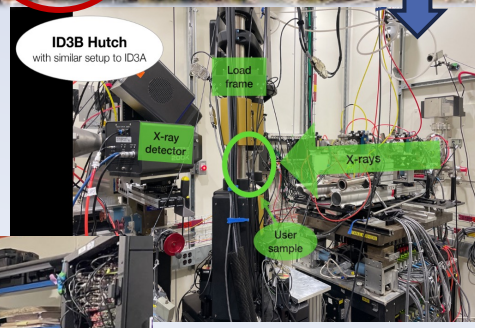
CHESS
Cornell High Energy Synchrotron Source

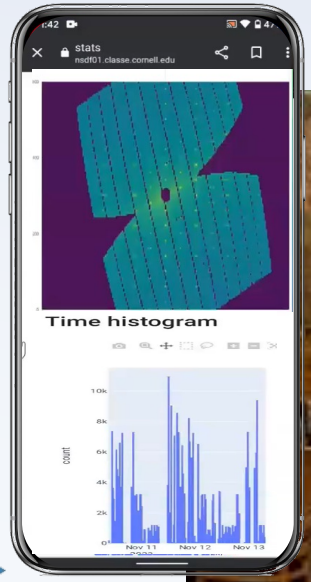


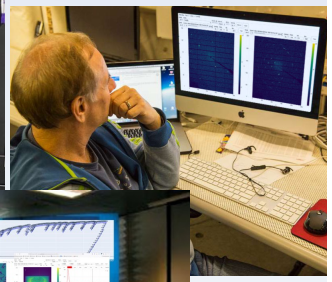


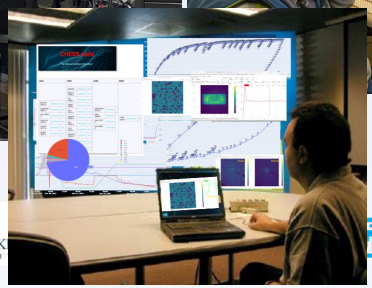
Taste of Thai
4,0 ★★★★☆ (435) · 10-20 $ · Tailandese
216 E State St







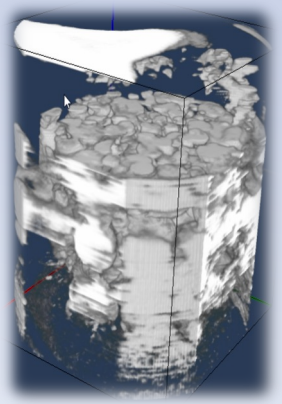


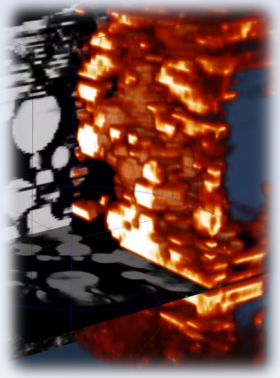


NSDF
NSF National Science

University of Utah
Brookhaven National Lab
Idaho National Lab
Micro Testing Solution

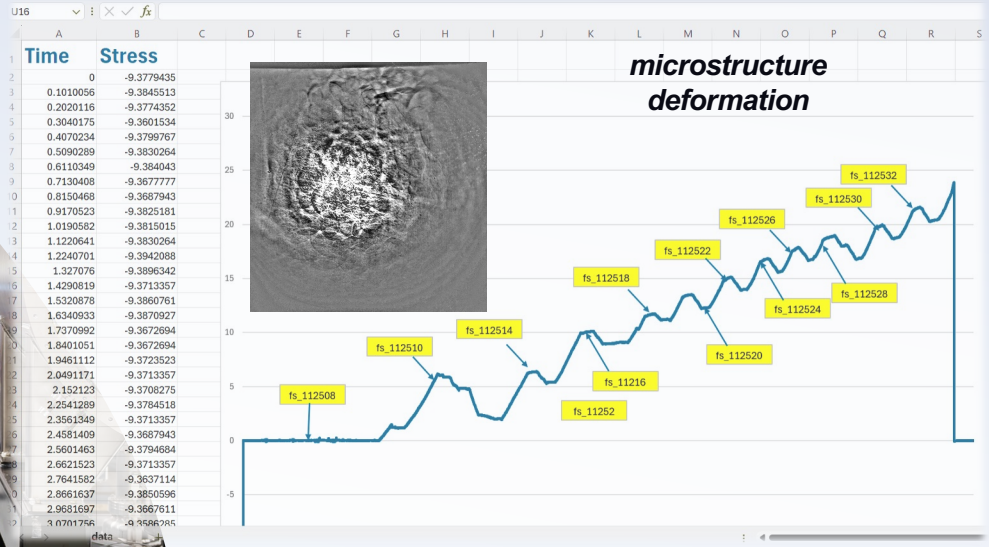
300TB cloud storage
moved **½ PB** data
>200 Cloud Instances





THE UNIVERSITY OF UTAH
MECHANICAL ENGINEERING

Pania Newell



Cloud execution of AI workflows for material characterization and segmentation

www.sci.utah.edu

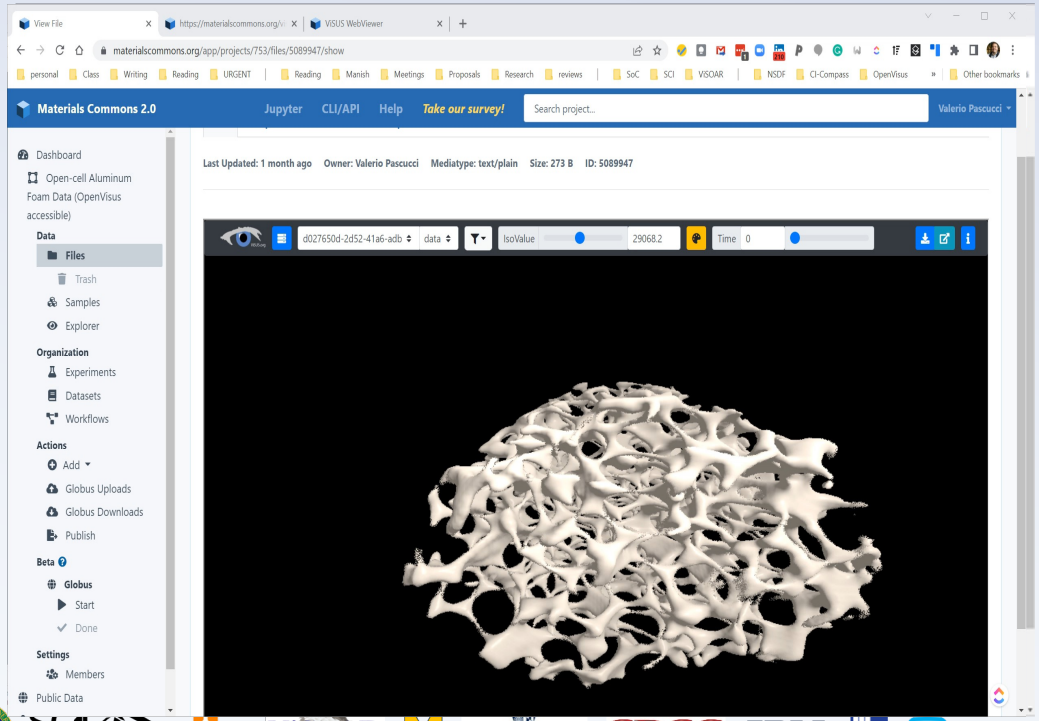
Equity in Access and Use of Community Data

- NSDF-commons
- Direct connection with CHESS light source
- Reduced resource requirements for the users
- Immediate data sharing with the broader community
- Data exploration at the portal without bulk data transfers

- Over 650 Registered Users
- Nearly **4 Million files** uploaded
- Over **26TB** of Data
- Nearly 870,000 Sample and Process Attributes
- New Features and Updates released monthly
- Over 80 Published Datasets
  - More than **14,000** views
  - Nearly **7,000** downloads

Materials Commons 2.0

PRISMS



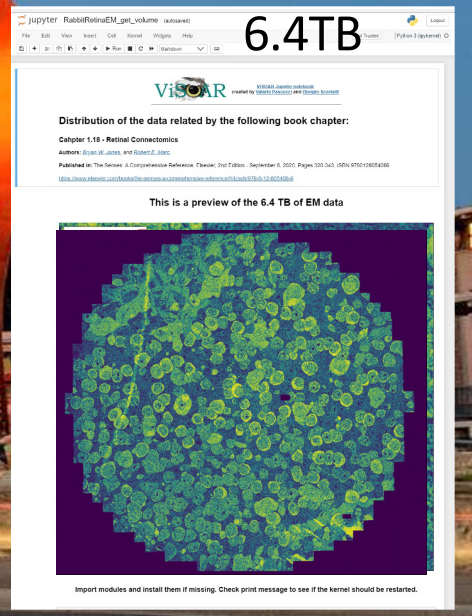
National Science Data Fabric

www.sci.utah.edu

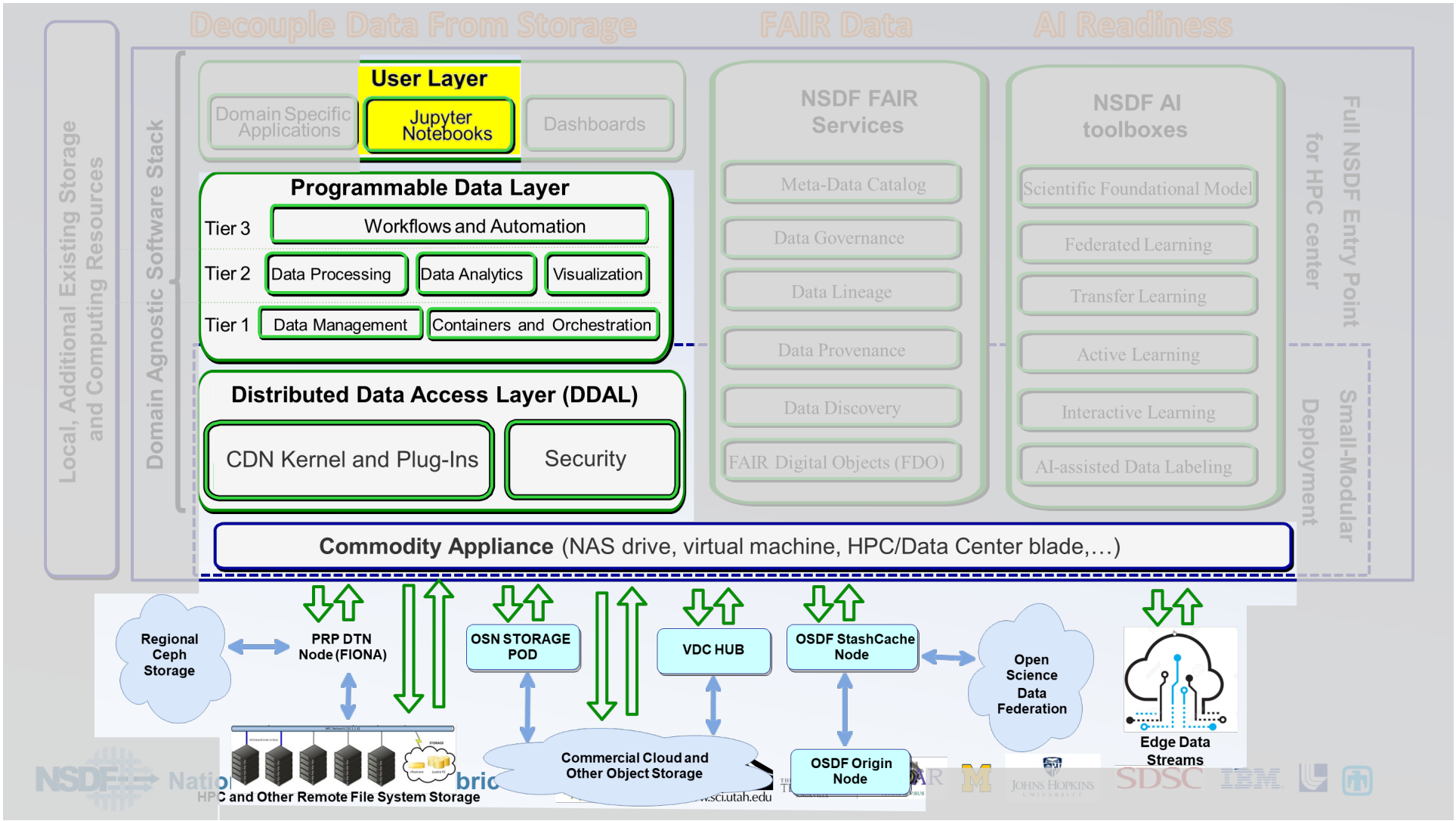
Equity in Education: UTEP

NSDF curriculum and training materials with Brian Schuster (Metallurgical, Materials, and Biomedical Engineering)

Undergraduate students at an MSI work on assignments with 6.4TB of data



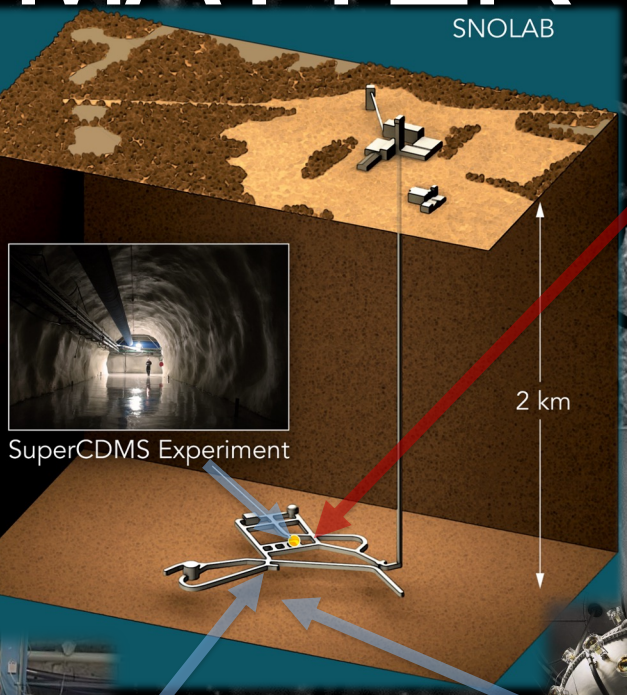
National Science Data Fabric

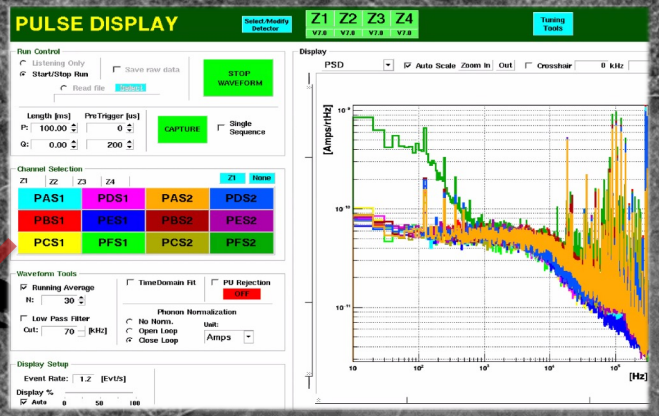

Decouple Data From Storage
FAIR Data
AI Readiness
Local, Additional Existing Storage and Computing Resources
Domain Agnostic Software Stack
User Layer
Domain Specific Applications
Jupyter Notebooks
Dashboards
Programmable Data Layer
Tier 3
Workflows and Automation
Tier 2
Data Processing
Data Analytics
Visualization
Tier 1
Data Management
Containers and Orchestration
Distributed Data Access Layer (DDAL)
CDN Kernel and Plug-Ins
Security
NSDF FAIR Services
Meta-Data Catalog
Data Governance
Data Lineage
Data Provenance
Data Discovery
FAIR Digital Objects (FDO)
NSDF AI toolboxes
Scientific Foundational Model
Federated Learning
Transfer Learning
Active Learning
Interactive Learning
AI-assisted Data Labeling
Full NSDF Entry Point for HPC center
Small-Modular Deployment
Commodity Appliance (NAS drive, virtual machine, HPC/Data Center blade,…)
Regional Ceph Storage
PRP DTN Node (FIONA)
OSN STORAGE POD
VDC HUB
OSDF StashCache Node
Open Science Data Federation
Edge Data Streams
Commercial Cloud and Other Object Storage
OSDF Origin Node
HPC and Other Remote File System Storage
www.sci.utah.edu

SEARCH FOR

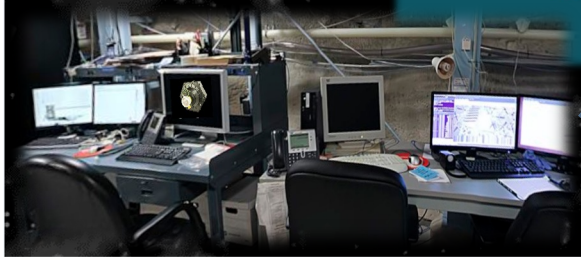
DARK MATTER

SNOLAB

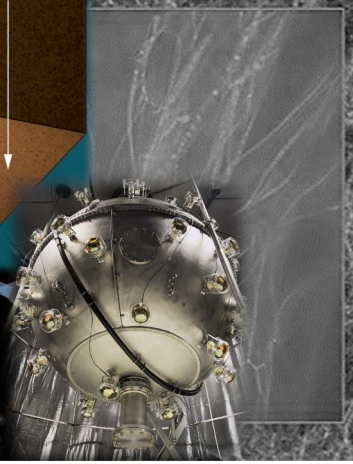




PETABYTES of multiple-channel time series



Control center



Vessel

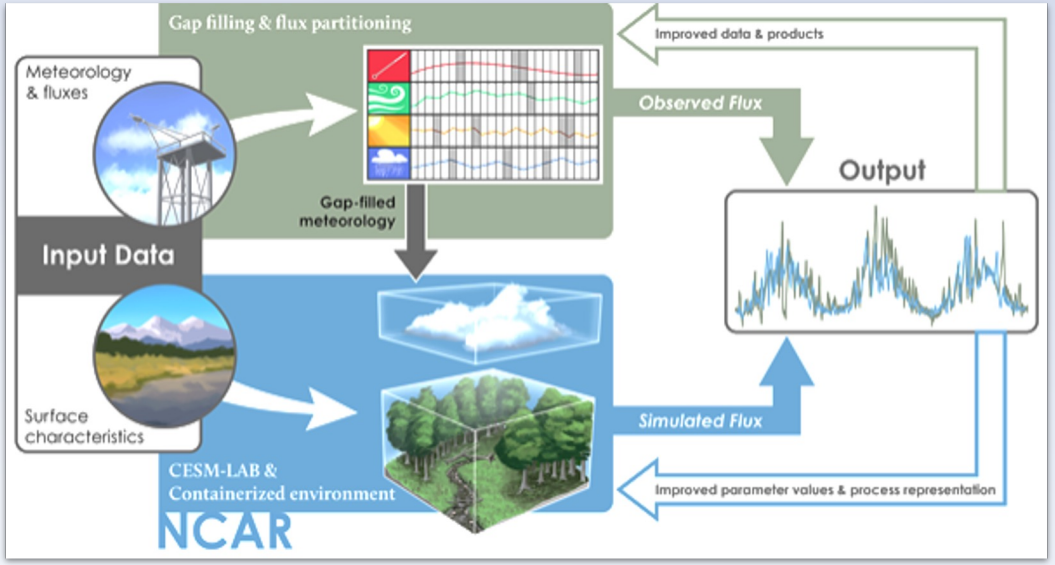
AMY ROBERTS

University of Colorado **Denver**

Access and Training at National Resources

NSDF community resources for demonstration and training in future classes:

https://ncar.nationalsciencedatafabric.org/neon-demo/v1



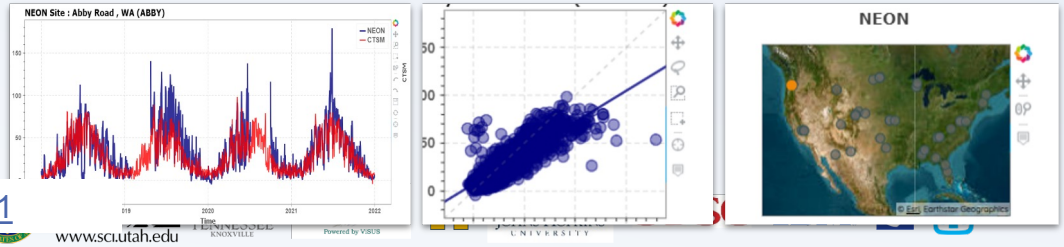
Geosci. Model Dev., 16, 5979–6000, 2023
https://doi.org/10.5194/gmd-16-5979-2023

Geoscientific Model Development
EGU

Overcoming barriers to enable convergence research by integrating ecological and climate sciences: the NCAR–NEON system Version 1

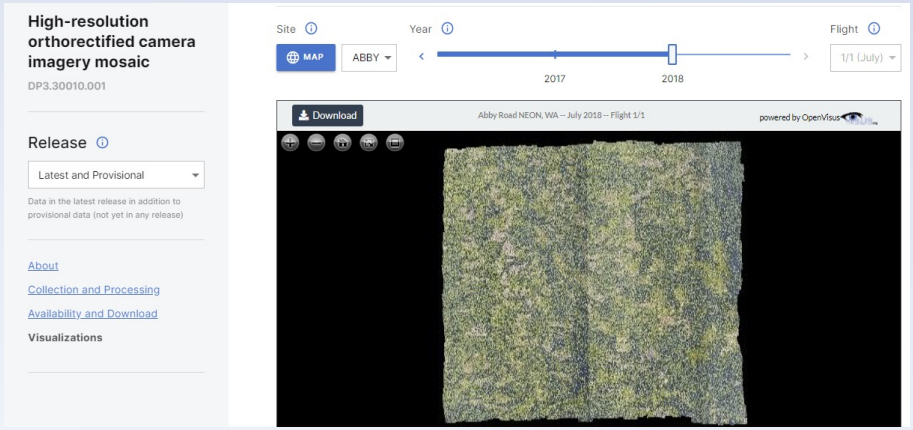
Danica L. Lombardozzi[1,★], William R. Wieder[1,2,★], Negin Sobhani[1], Gordon B. Bonan[1], David Durden[3], Dawn Lenz[3], Michael SanClements[3], Samantha Weintraub-Leff[3], Edward Ayres[3], Christopher R. Florian[3], Kyla Dahlin[4], Sanjiv Kumar[5], Abigail L. S. Swann[6], Claire M. Zarakas[6], Charles Vardeman[7], and Valerio Pascucci[8]

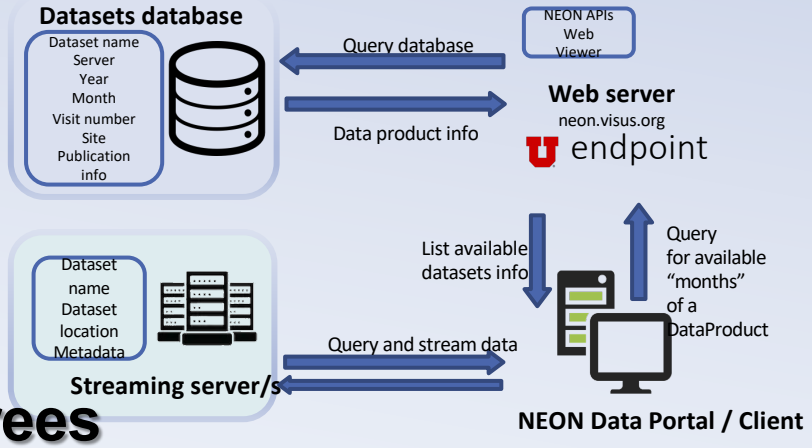
[1]Climate and Global Dynamics Laboratory, National Center for Atmospheric Research, Boulder, CO, USA
[2]Institute of Arctic and Alpine Research, University of Colorado Boulder, Boulder, CO, USA
[3]National Ecological Observatory Network, Battelle, Boulder, CO, USA
[4]Department of Geography, Environment, and Spatial Sciences, Michigan State University, East Lansing
[5]College of Forestry, Wildlife and Environment, Auburn University, Auburn, AL, USA
[6]Department of Atmospheric Sciences, University of Washington, Seattle, WA, USA
[7]Center for Research Computing, University of Notre Dame, Notre Dame, IN, USA
[8]Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, UT, USA
★These authors contributed equally to this work.



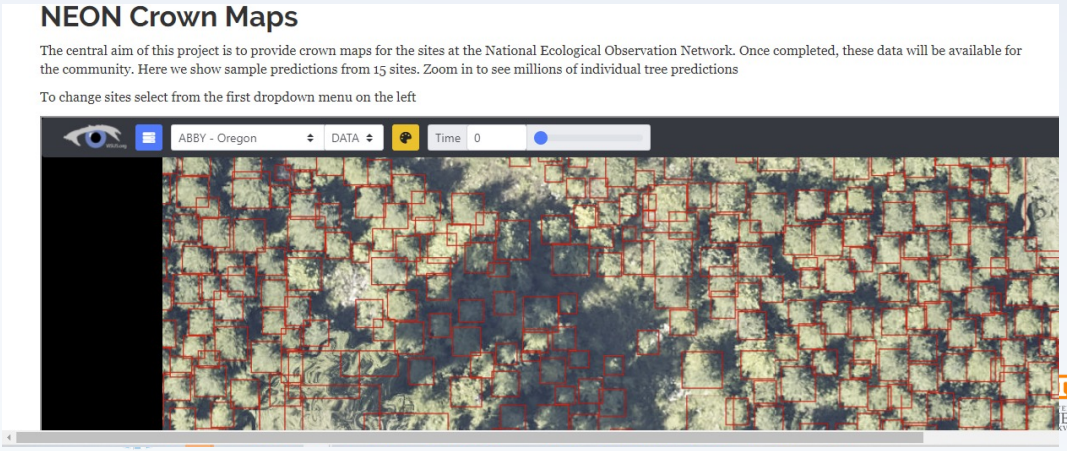
https://data.neonscience.org/data-products/DP3.30010.001

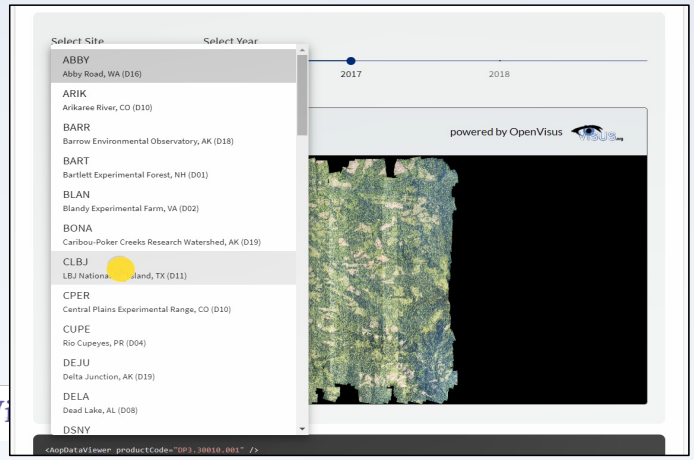
Neon Data Portal (Battelle)



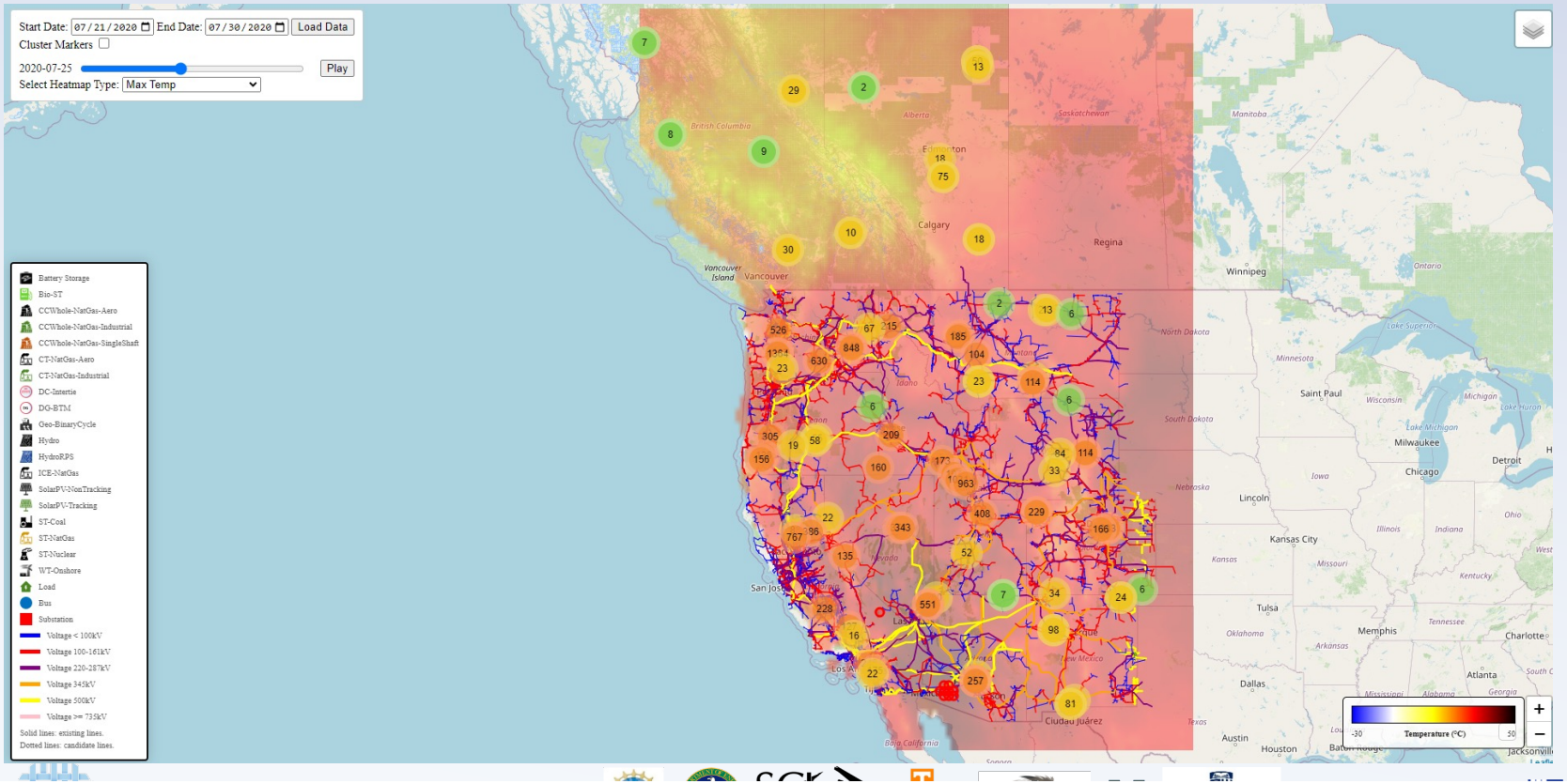


AI-based image analysis to count trees



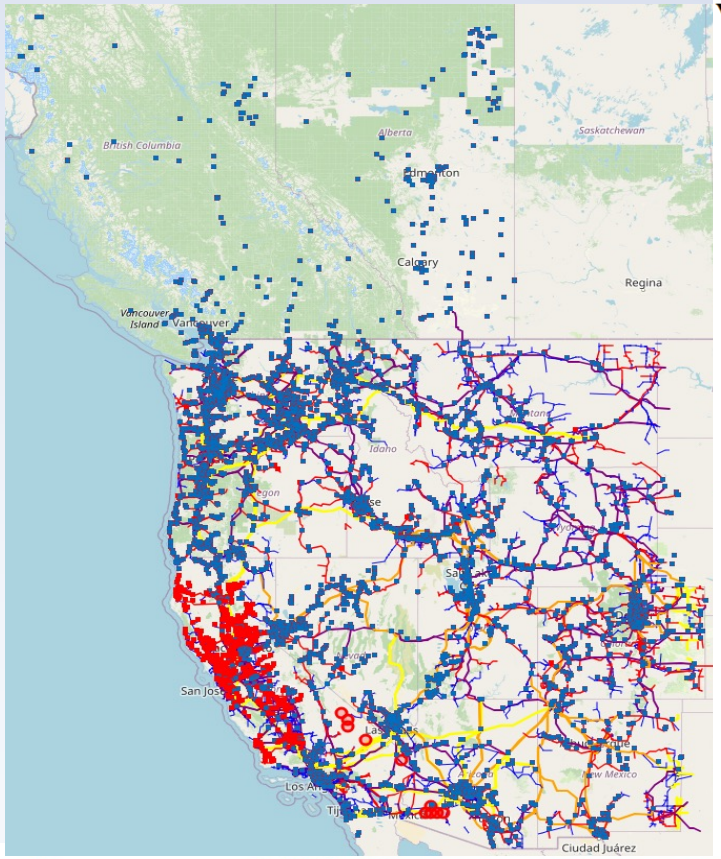


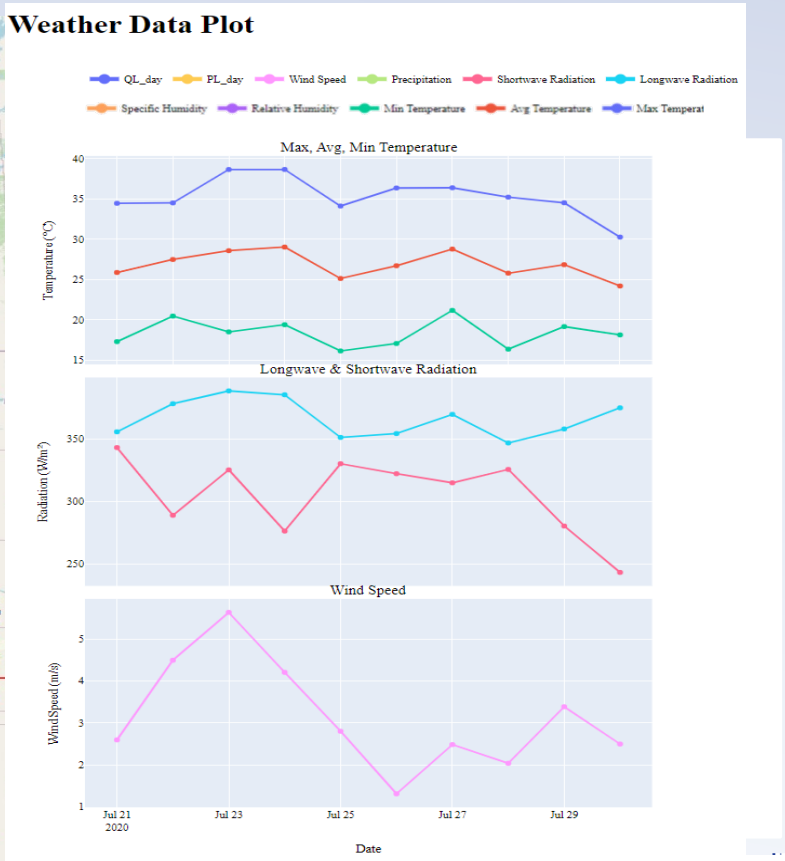
Use of the Streaming CMPI6 Data for AI-Based Weather Resiliency Model



THE UNIVERSITY OF UTAH® · UNIVERSITY OF CALGARY · DRI Desert Research Institute · THOMPSON RIVERS UNIVERSITY · UNIVERSITY OF ALBERTA · UBC · UBC · UC San Diego · THE UNIVERSITY OF NEW MEXICO · University of Regina · UNIVERSITY OF SASKATCHEWAN

Use of the Streaming CMPI6 Data for AI-Based Weather Resiliency Model

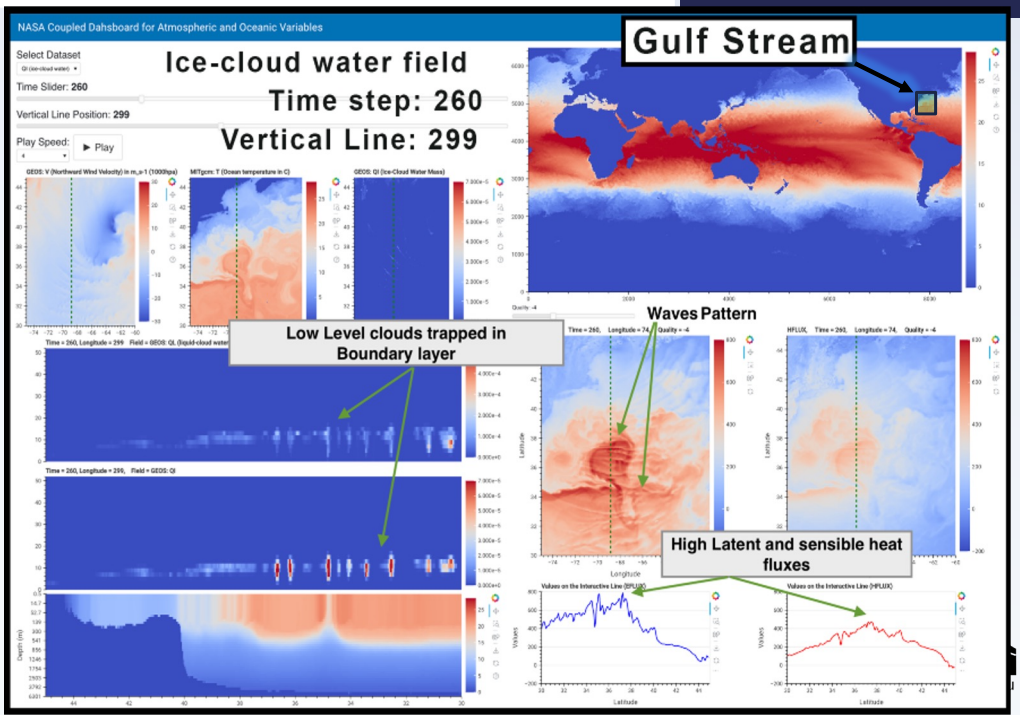




Nina McCurdy · 1st
NASA Ames Research Center
Data Analysis and Visualization Scientist at NASA Ames Research Center

Equity in Accessing Full Data in Real-Time from HCP and Cloud

Enabling streaming data access and processing in a distributed cyberinfrastructure for NASA climate modeling data (from Pleiades Supercomputer & Cloud)



LLC4320 2.8 PB Ocean dataset

Model of ocean circulation and global ocean data to study the circulation role in climate changes

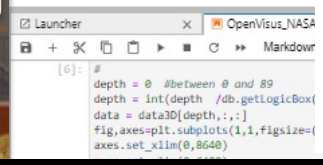
Requirements:

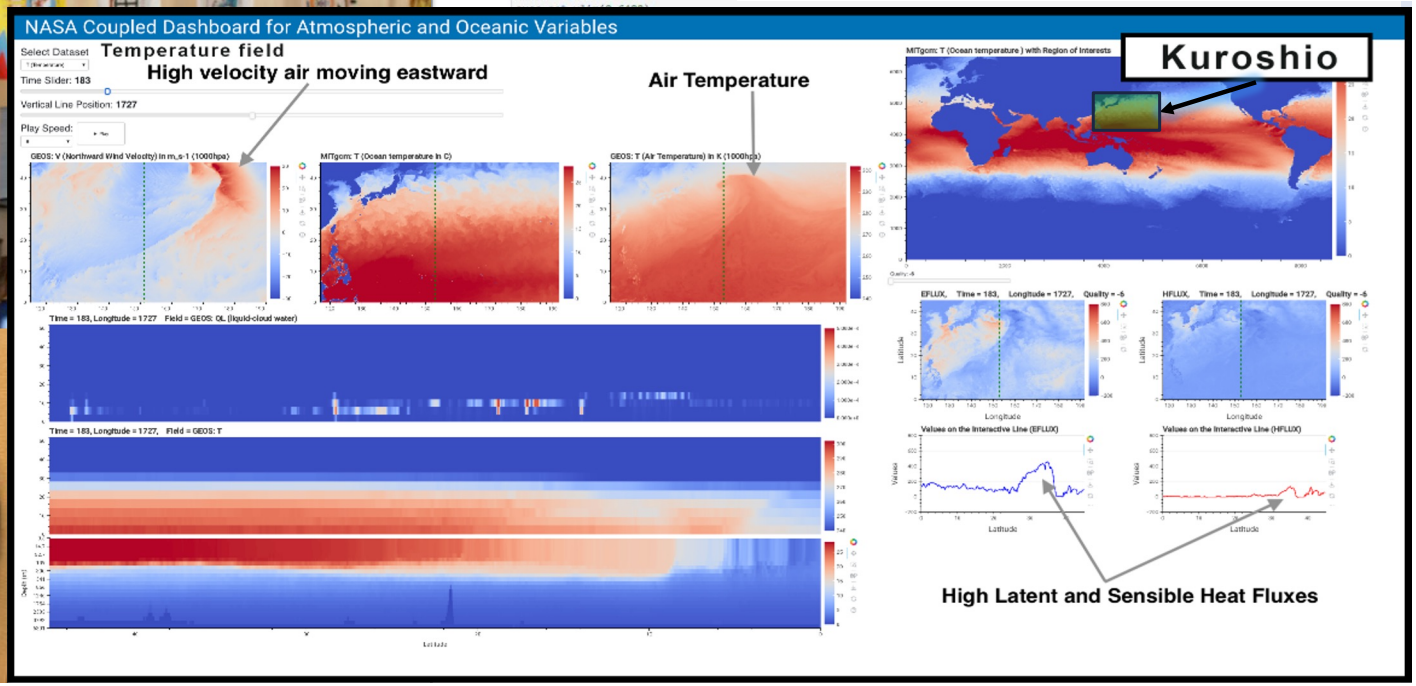
- Access the full data
- Overcome limitations in computational power
- Provide real-time processing capabilities.

Delivering Data and Training Equitably
across Earth Science Disciplines and Institutions

Gustavo Ovando-Montejo, Environment and Society Department at Utah State University, **2024 Quinney College of Natural Resources Teacher of the Year.**

USU Blanding Campus



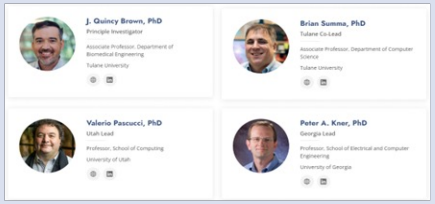


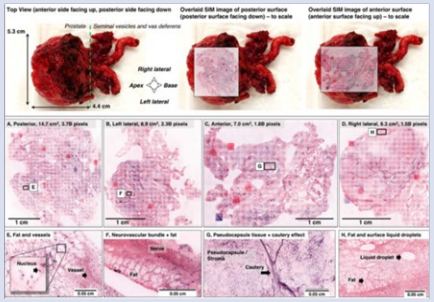
NSDF Cyberinfrastructure is core to ARPA-H **Magic-Scan**

ARPA-H
PSI

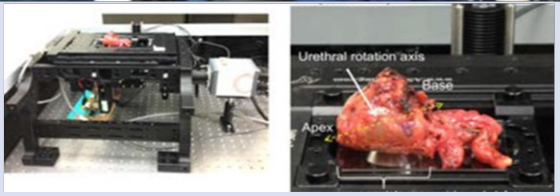
Machine-learning Assisted Gigantic-Image Cancer Margin Scanner

A project of the Advanced Research Projects Agency for Health Precision Surgical Interventions

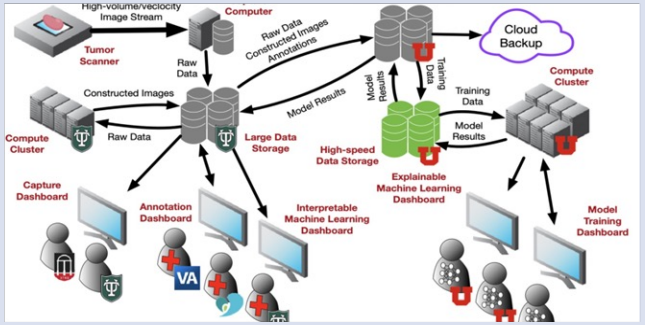




*Biden Cancer Moonshot project: up to **$23M** for an imaging system that scans a **tumor during surgery** & determines within minutes whether cancer remains.*



- key innovations in microscopy, sample automation, **cyber-infrastructure**,
- **ML model co-design** & training on Peta scale data, practical & rapid ML model deployment, and cancer detection and visualization.
- **human-centered approach** to innovation, design and development, involving end-users and stakeholders
- cost-conscious product design that optimizes benefits to **physicians, payers, and patients**.



Session II: Training on using NSDF Services for End-to-End Analysis of Scientific Data

Democratizing Access and Use of Large-scale Data

Four-Step Workflow Tutorial

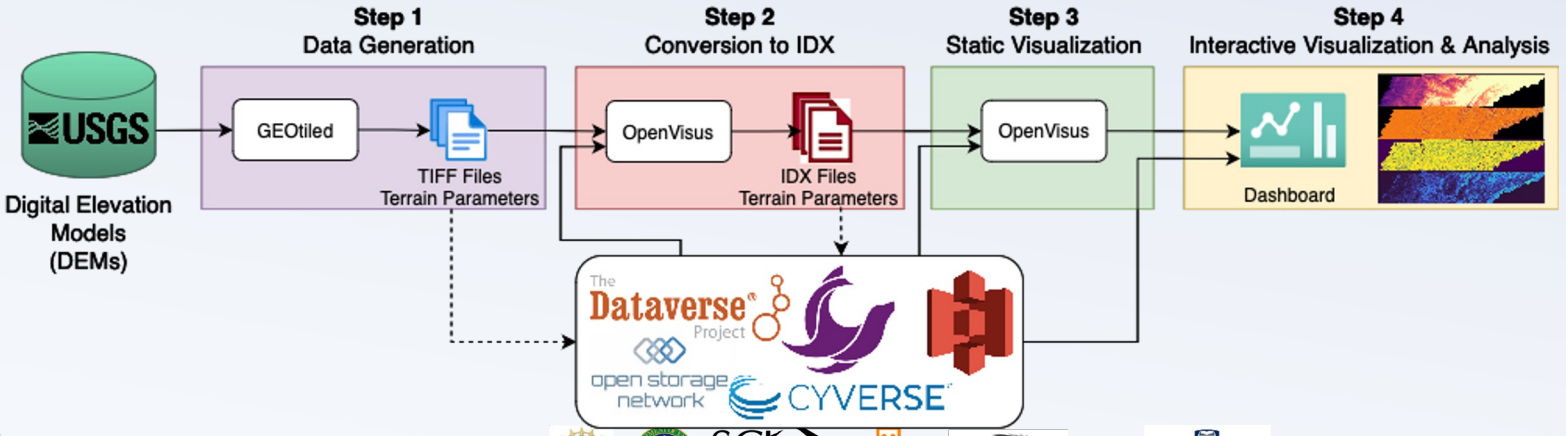
This tutorial showcases the capabilities of NSDF, guiding you through a **four-step modular workflow** that leverages OpenVisus services to analyze a geospatial dataset generated with GEOtiled.

Step 1: Data Generation
Collect DEMs from the United States Geological Survey (USGS). **Process** them with GEOtiled or upload the data from public or private storage.

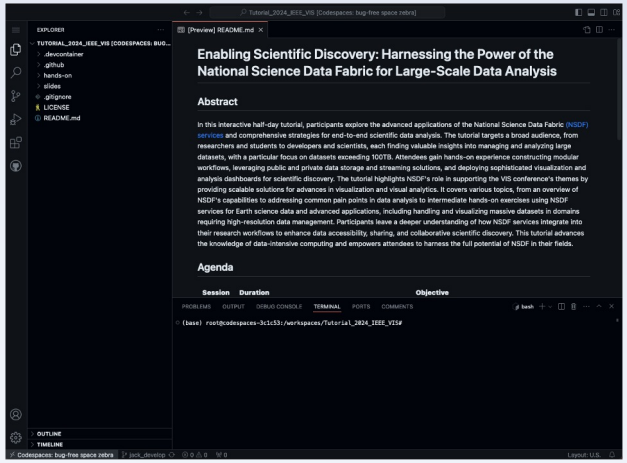
Step 2: Conversion to IDX Format
Convert files from TIFF to IDX (the format used by OpenVisus), preserving accuracy but reducing size. **Store** IDX files in public or private storage.

Step 3: Static Visualization
Statically visualize the terrain parameters in OpenVisus. **Validate** accuracy of IDX-based images with the TIFF-based images.

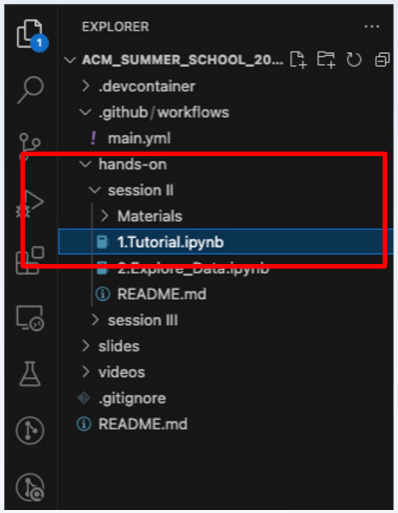
Step 4: Interactive Visualization & Analysis
Launch dashboard for interacting with large-scale data to access subregions of the original dataset for ad hoc analysis



(1) Your codespace is fully loaded. Now you should see this screen

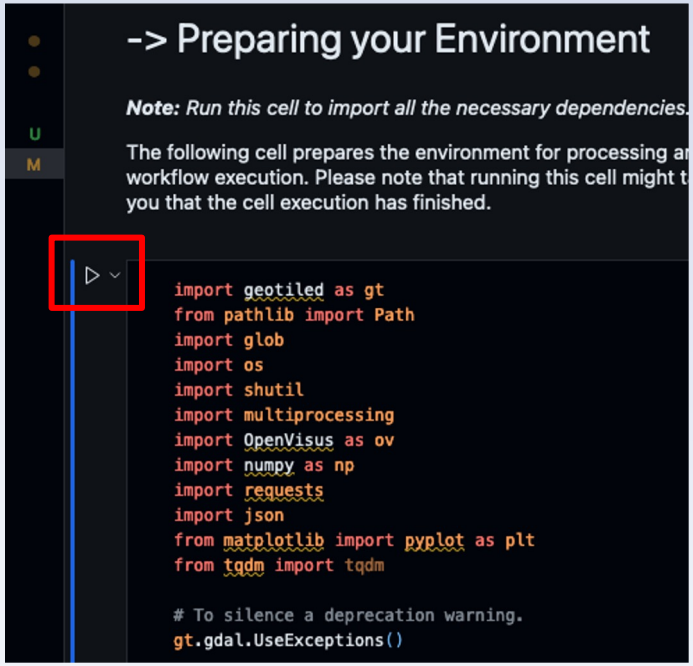


(2) Select the file *1.Tutorial.ipynb* using the side bar



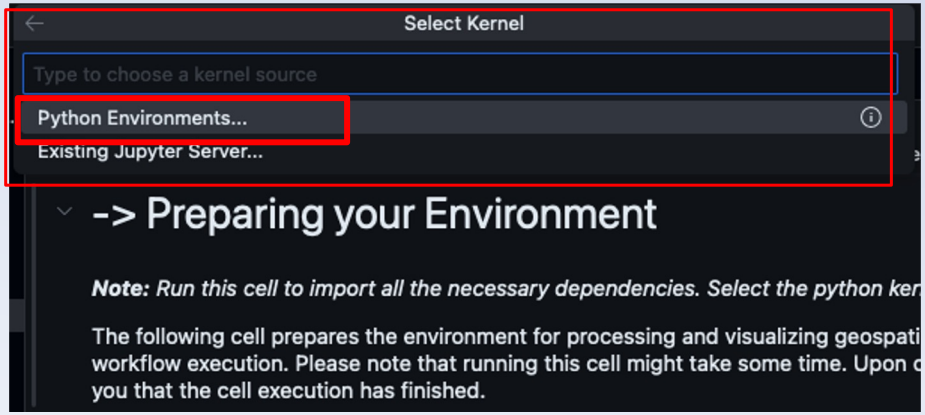
(3) Look for the *Preparing your Environment* section and press the play button

Tutorial

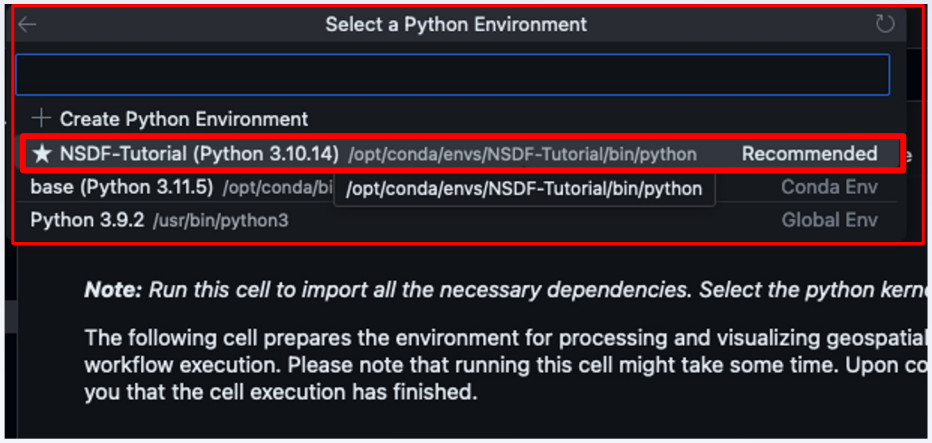


Instructions continue

(4) A list to Select Kernel should pop up. Select *Python Environments* …

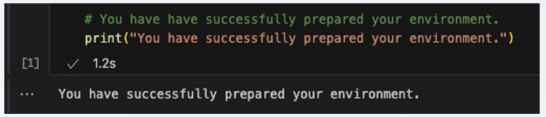


(5) Select the ★*NSDF-Tutorial* option



Tutorial

(6) After running the *Preparing your Environment* cell, you should see a message saying it was successfully prepared



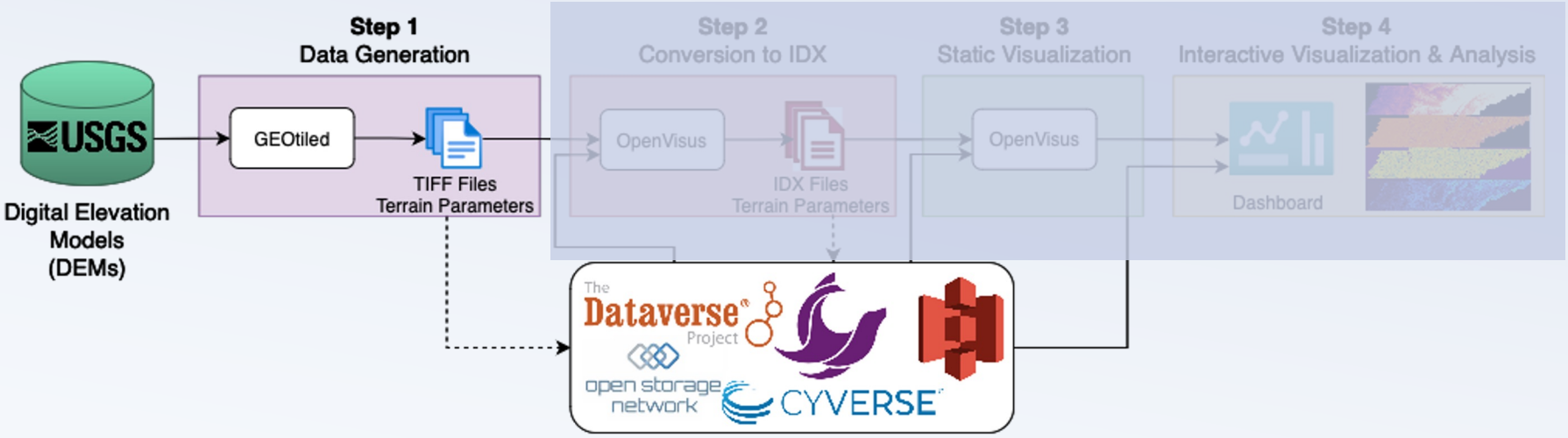
Step 1: Data Generation with GEOtiled

Step 1: Data Generation
Collect DEMs from the United States Geological Survey (USGS) and process them with GEOtiled or upload the data from public or private storage.

Step 1 provides **two options** to obtain data and generate the TIFF files before proceeding with Step 2

Option A
Generating Data Using the SOMOSPIE Application Module

Option B
Accessing data from Dataverse public commons

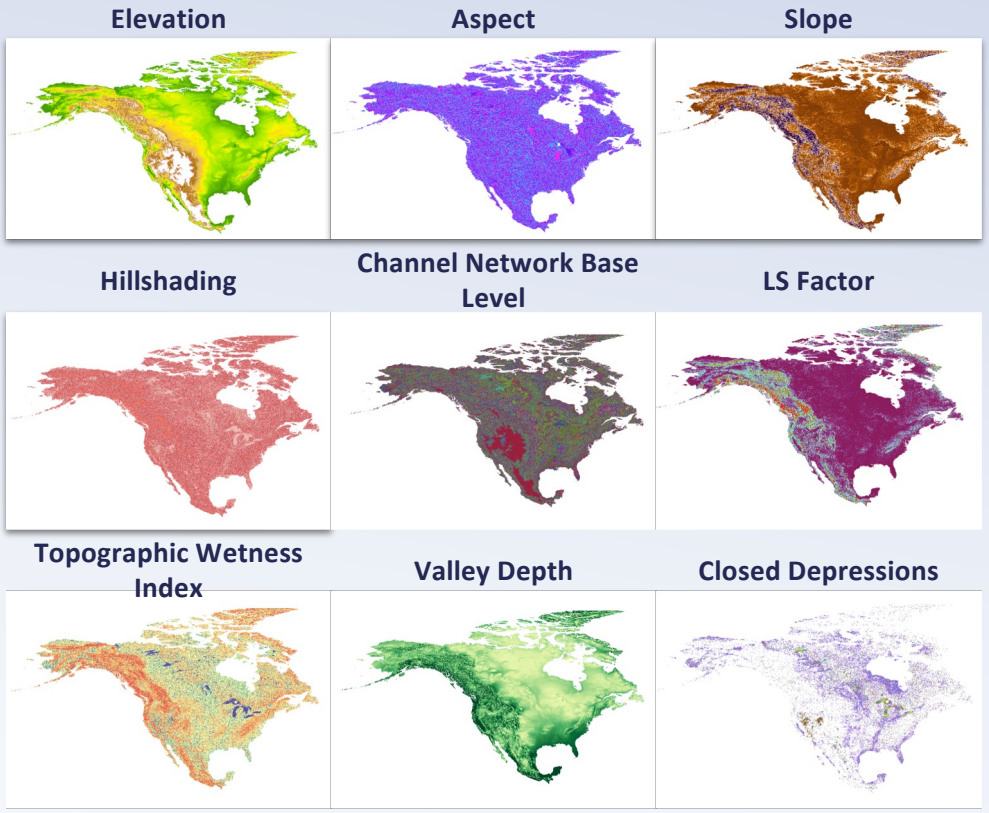


Step 1: What are Terrain Parameters?

Terrain parameters (e.g., slope, aspect, hillshading, etc.) are **descriptions of surface form derived from Digital Elevation Models (DEM).**

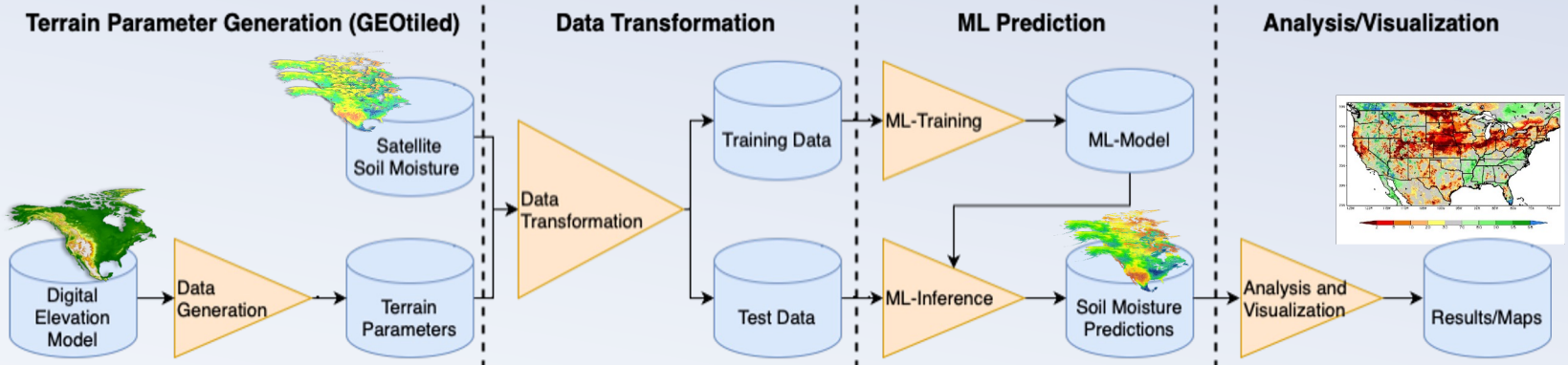
They play a **fundamental role** in applications such as **precision forestry and agriculture, and hydrology for landscape ecology.**

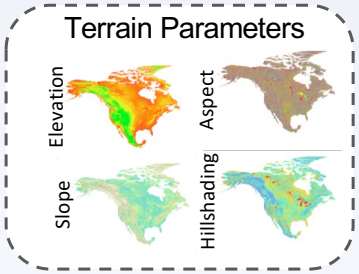
Generating **terrain parameters** at **high-resolution is computationally expensive,** hindering their accessibility by the scientific community

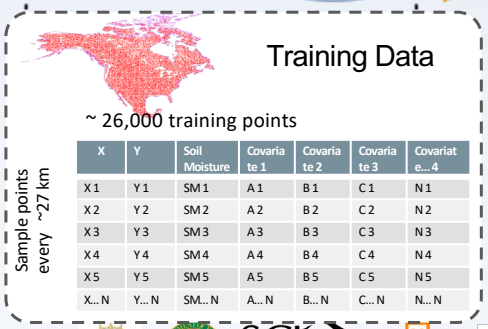


Step 1: SOMOSπ Components

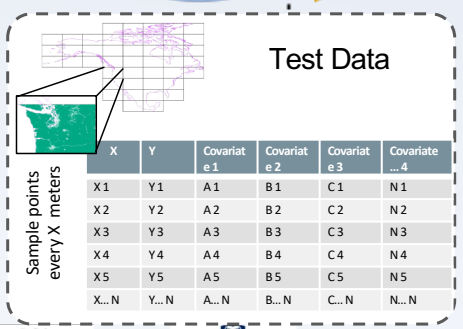
SOMOSPIE (SOil MOisture SPatial Inference Engine) has **four components** that empower scientists to generate, predict, and analyze **high-resolution topographic data**







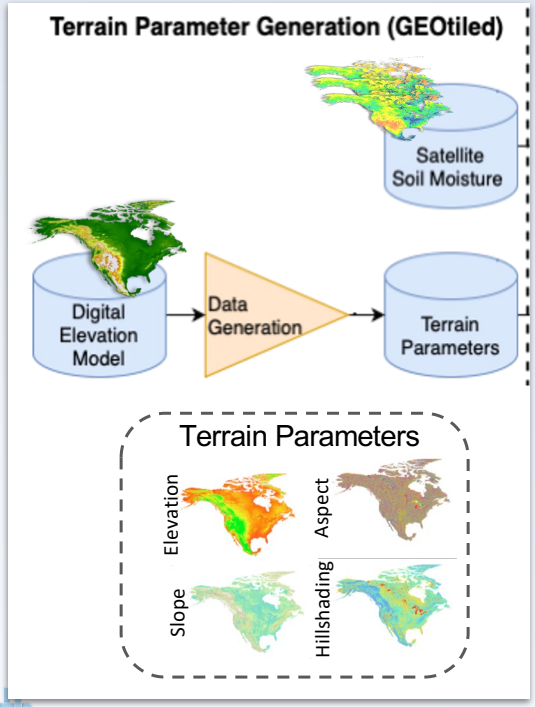
X	Y	Soil Moisture	Covariate 1	Covariate 2	Covariate 3	Covariate ...4
X1	Y1	SM1	A1	B1	C1	N1
X2	Y2	SM2	A2	B2	C2	N2
X3	Y3	SM3	A3	B3	C3	N3
X4	Y4	SM4	A4	B4	C4	N4
X5	Y5	SM5	A5	B5	C5	N5
X...N	Y...N	SM...N	A...N	B...N	C...N	N...N



X	Y	Covariate 1	Covariate 2	Covariate 3	Covariate ...4
X1	Y1	A1	B1	C1	N1
X2	Y2	A2	B2	C2	N2
X3	Y3	A3	B3	C3	N3
X4	Y4	A4	B4	C4	N4
X5	Y5	A5	B5	C5	N5
X...N	Y...N	A...N	B...N	C...N	N...N

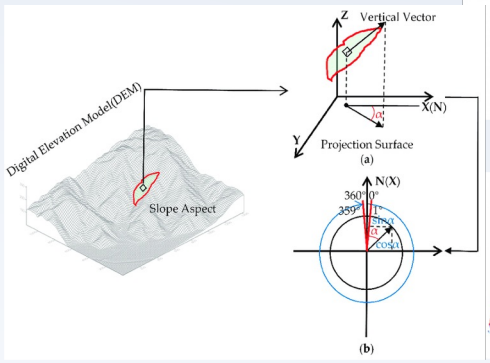
Step1: GEOtiled Terrain Generation

We expand on the first component, **GEOtiled,** that **computes high-resolution terrain parameters** using Digital Elevation Models (DEMs)

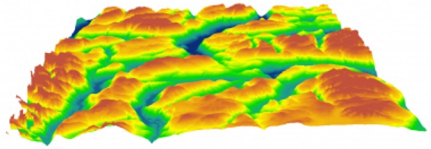


GEOtiled leverages data partitioning to accelerate the computation of terrain parameters from DEMs while preserving accuracy

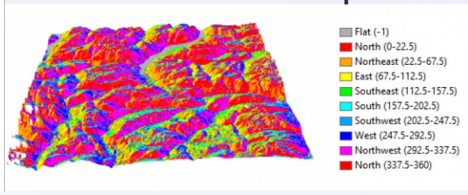
Computing High-resolution Terrain Parameters



Digital Elevation Model (DEM)

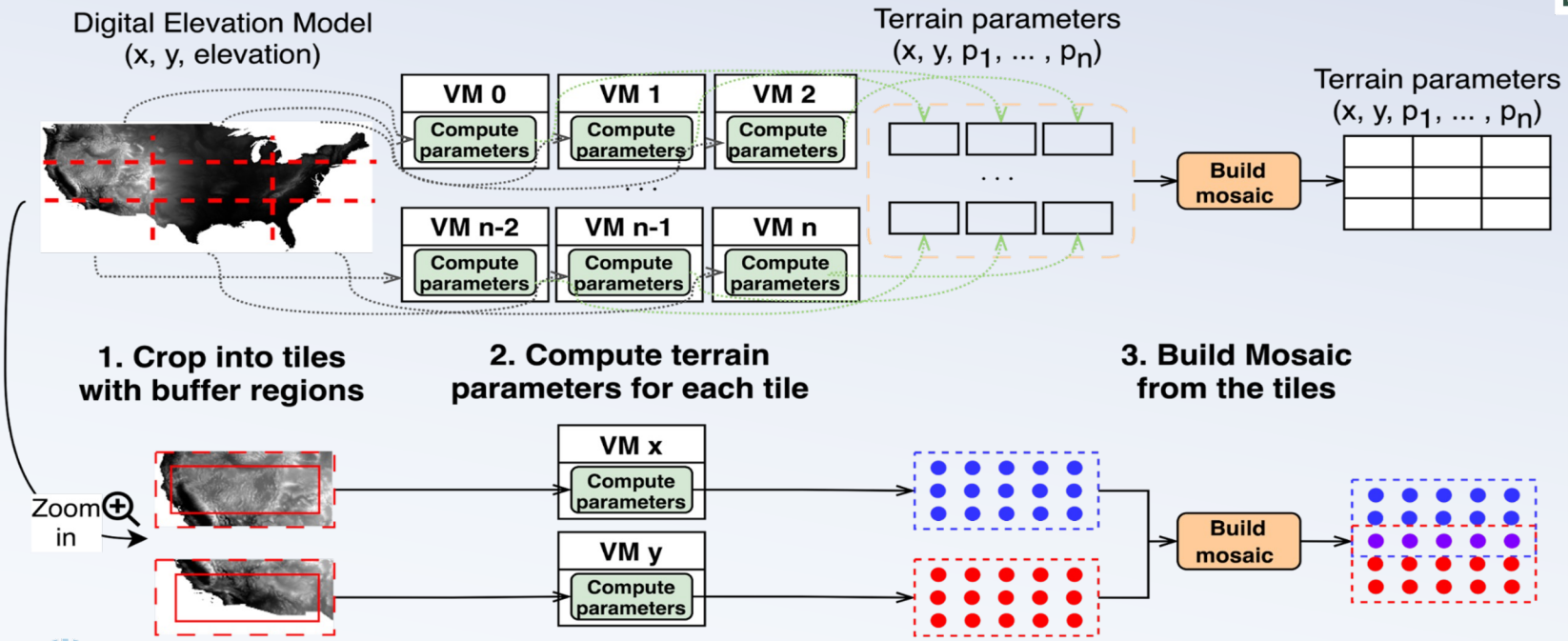


Terrain Parameter 1: Aspect



Step1: GEOtiled Terrain Generation

GEOtiled generates high-resolution terrain parameters at a large scale from a DEM. It has three stages: **(1)** we **crop DEM into tiles**, with buffer regions; **(2)** we **compute the terrain parameters** for each tile; **(3)** we **build a mosaic** from the tiles.



Step 1: Data Generation with GEOtiled

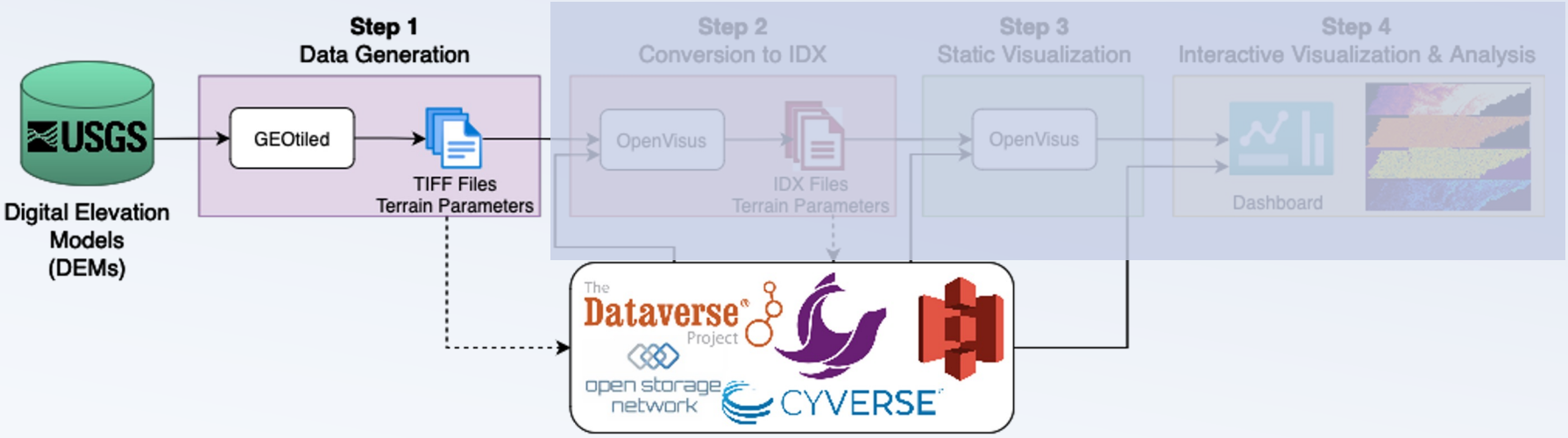
Tutorial

Step 1: Data Generation
Collect DEMs from the United States Geological Survey (USGS) and process them with GEOtiled or upload the data from public or private storage.

Step 1 provides **two options** to obtain data and generate the TIFF files before proceeding with Step 2

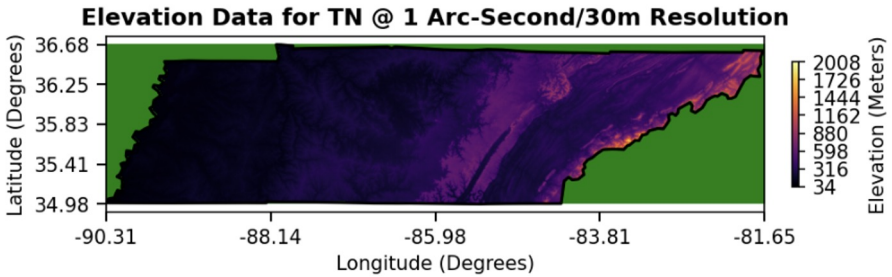
Option A (~10 mins)
Generating Data Using the SOMOSPIE Application Module

Option B (~3 mins)
Accessing data from Dataverse public commons

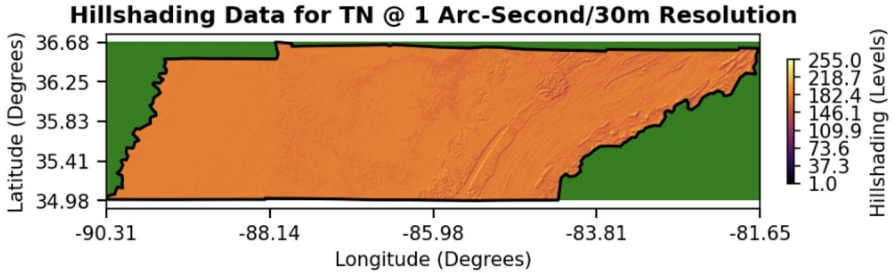


Step 1: Data Generation with GEOtiled

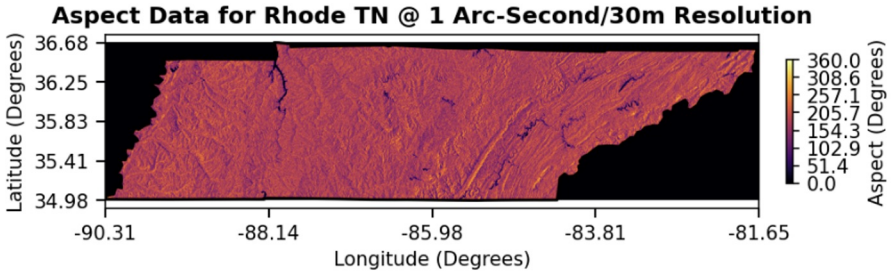
Tutorial



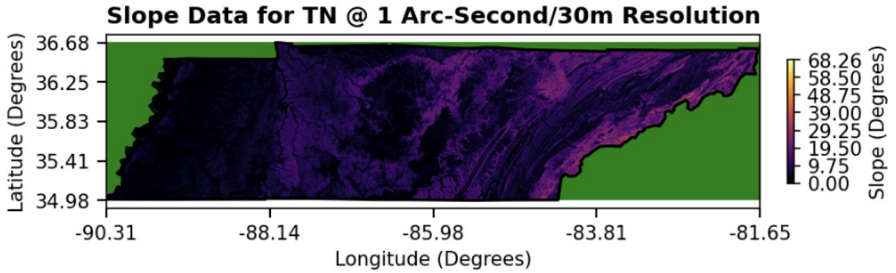
(a) Elevation - Terrain Parameter



(b) Hillshading - Terrain Parameter



(c) Aspect - Terrain Parameter



(d) Slope - Terrain Parameter

Step 2: Conversion to IDX

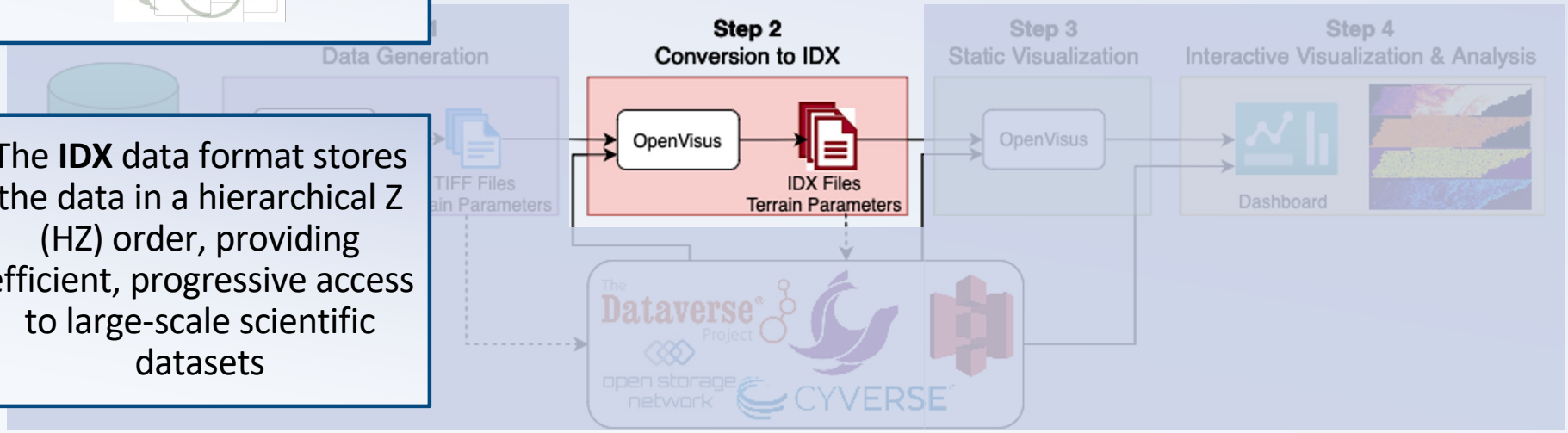
OpenVisus is a progressive cache-oblivious framework for large-scale data visualization

ViSUS

Step 2: Conversion to IDX Format
Convert files from TIFF to IDX (the format used by **OpenVisus**), preserving accuracy but reducing size. Store **IDX** files in public or private storage.

Converting **to IDX** from TIFF format **reduces file size by 20%** while preserving accuracy

The **IDX** data format stores the data in a hierarchical Z (HZ) order, providing efficient, progressive access to large-scale scientific datasets

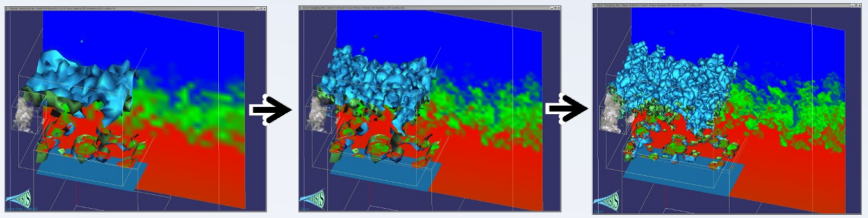


Step 2: ViSUS IDX Data Format

Why IDX?

- The IDX data format provides **efficient, cache-oblivious, and progressive access** to large-scale scientific datasets.
- Data stored in IDX format can be visualized in an **interactive environment** allowing for meaningful explorations with **minimal resources.**
- IDX provides **scalability** across a wide range of running conditions like personal computers to distributed systems.

- Conversion to IDX is **not limited** to TIFF; it will work on other data formats like **NetCDF, HDF5, RGB, raw/binary,** and so on.
- IDX supports industry-standard lossless and lossy compression algorithms such as zlib, zfp, lz4.



Fine Resolution ⟶

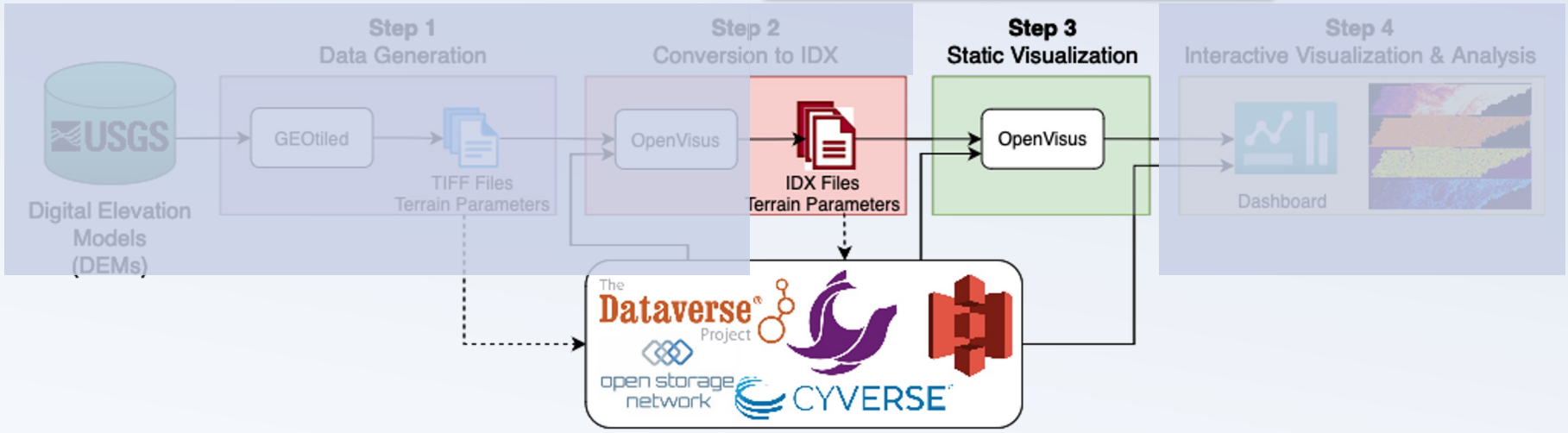
Step 3: Static Visualization

Step 3 provides **two options** to obtain data and collect the IDX files

Option A
From local storage

Option B
From Seal Storage

Step 3: Static Visualization
Statically visualize the terrain parameters in OpenVisus. Validate the accuracy of IDX-based images with TIFF-based images.



Step 3: Static Visualization

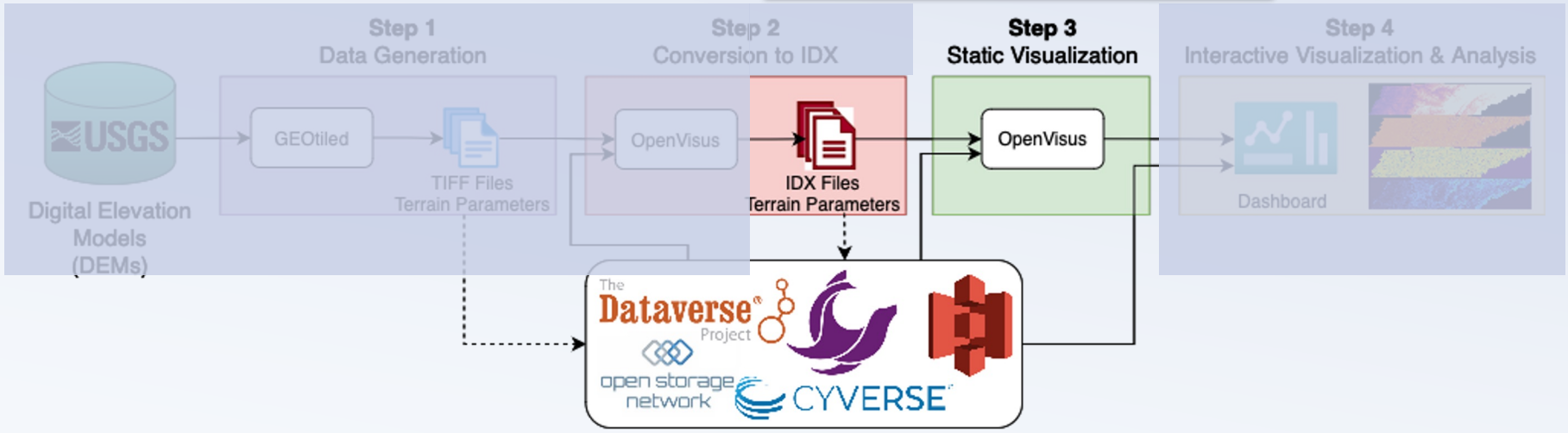
Tutorial

Step 3 provides **two options** to obtain data and collect the IDX files

Option A
From local storage

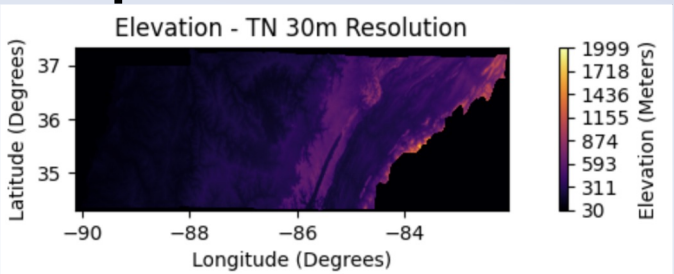
Option B
From Seal Storage

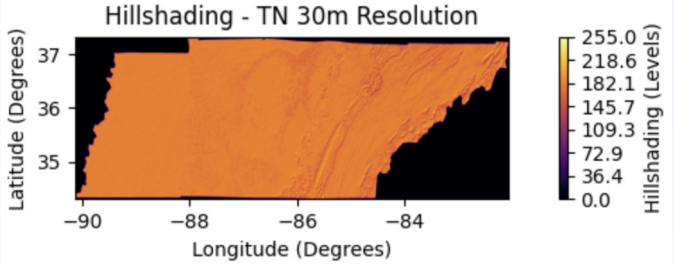
Step 3: Static Visualization
Statically visualize the terrain parameters in OpenVisus. Validate the accuracy of IDX-based images with TIFF-based images.

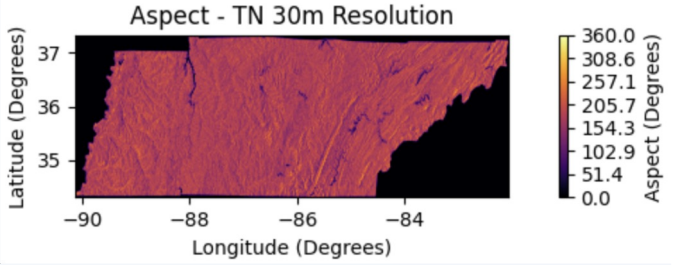


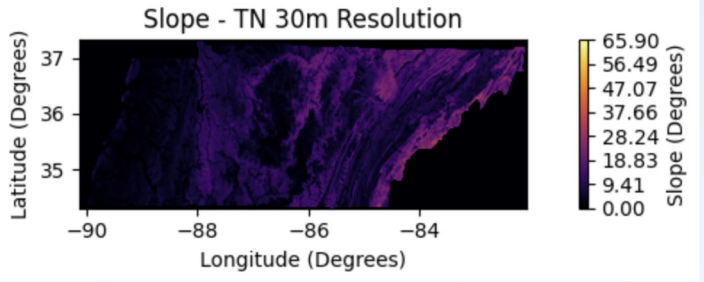
Step 3: Static Visualization

Tutorial









Option A
From local storage

Option B
From Seal Storage

Step 4: Interactive Visualization & Analysis

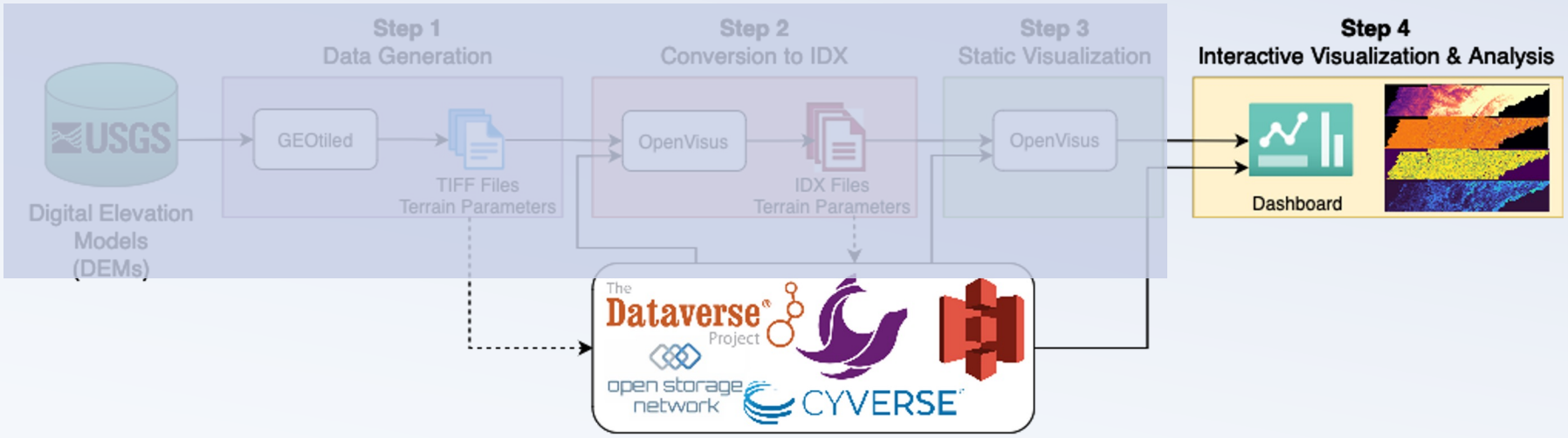
Remotely **access** large datasets, **zoom** into specific areas, **select** and **crop** subregions of interest, **save** data locally in a Python-compatible format, and **analyze** the data for scientific discovery.

Step 4 provides **two options** to obtain data and collect the IDX files

Option A
From local storage

Option B
From Seal Storage

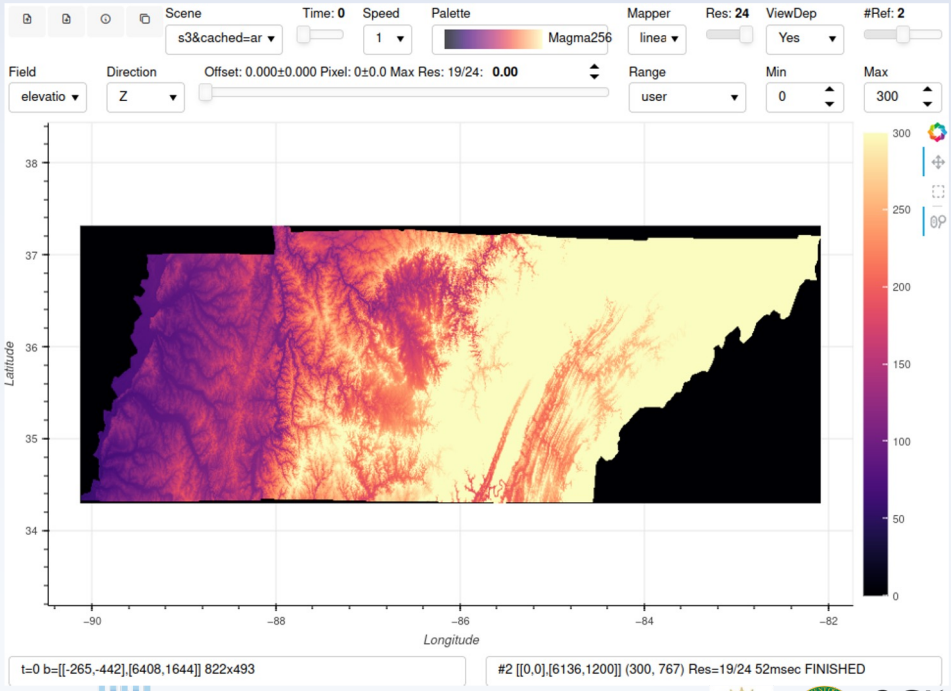
Step 4: Interactive Visualization & Analysis
Launch dashboard for interacting with large-scale data to access subregions of the original dataset for ad hoc analysis.



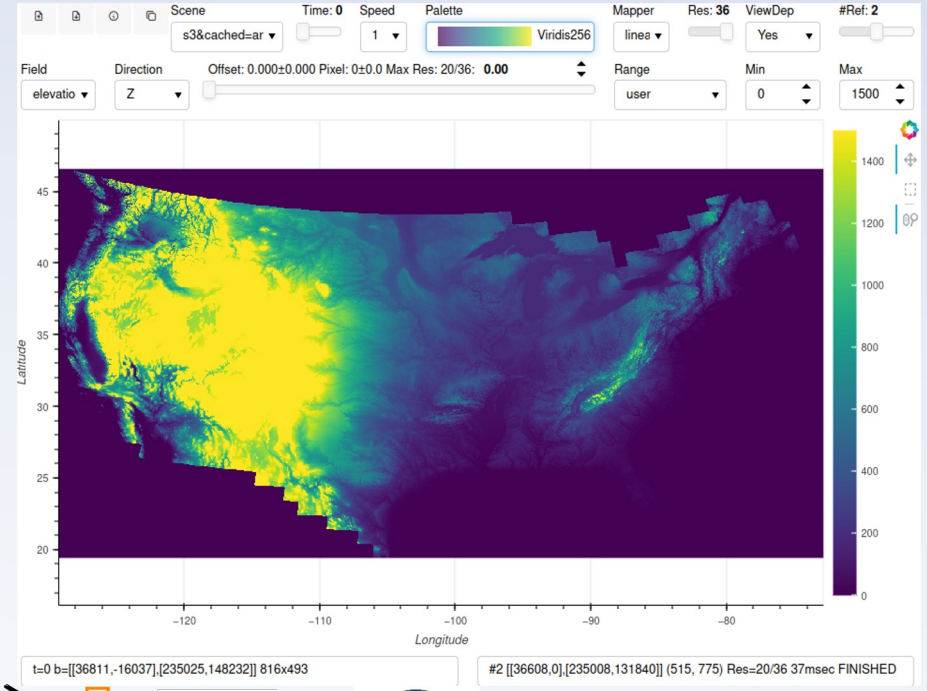
Step 4: Geographical Regions

Visualize and analyze two geographical regions at 30 m resolution

State of Tennessee - 200 MB



Contiguous United States (CONUS) - 200 GB



Step 4: Interactive Visualization & Analysis

Tutorial

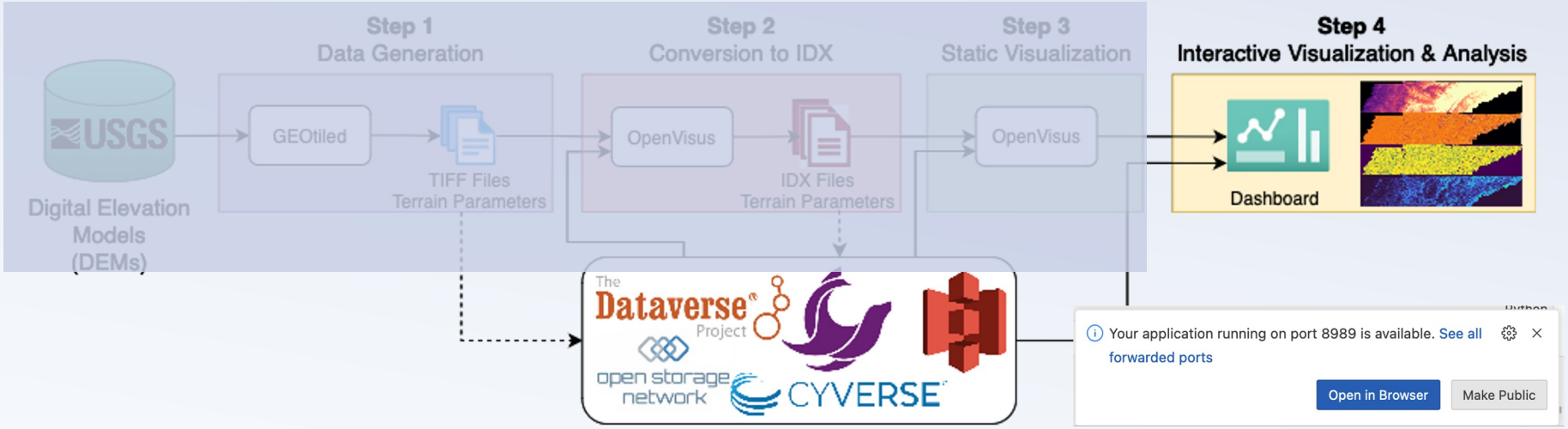
Remotely **access** large datasets, **zoom** into specific areas, **select** and **crop** subregions of interest, **save** data locally in a Python-compatible format, and **analyze** the data for scientific discovery.

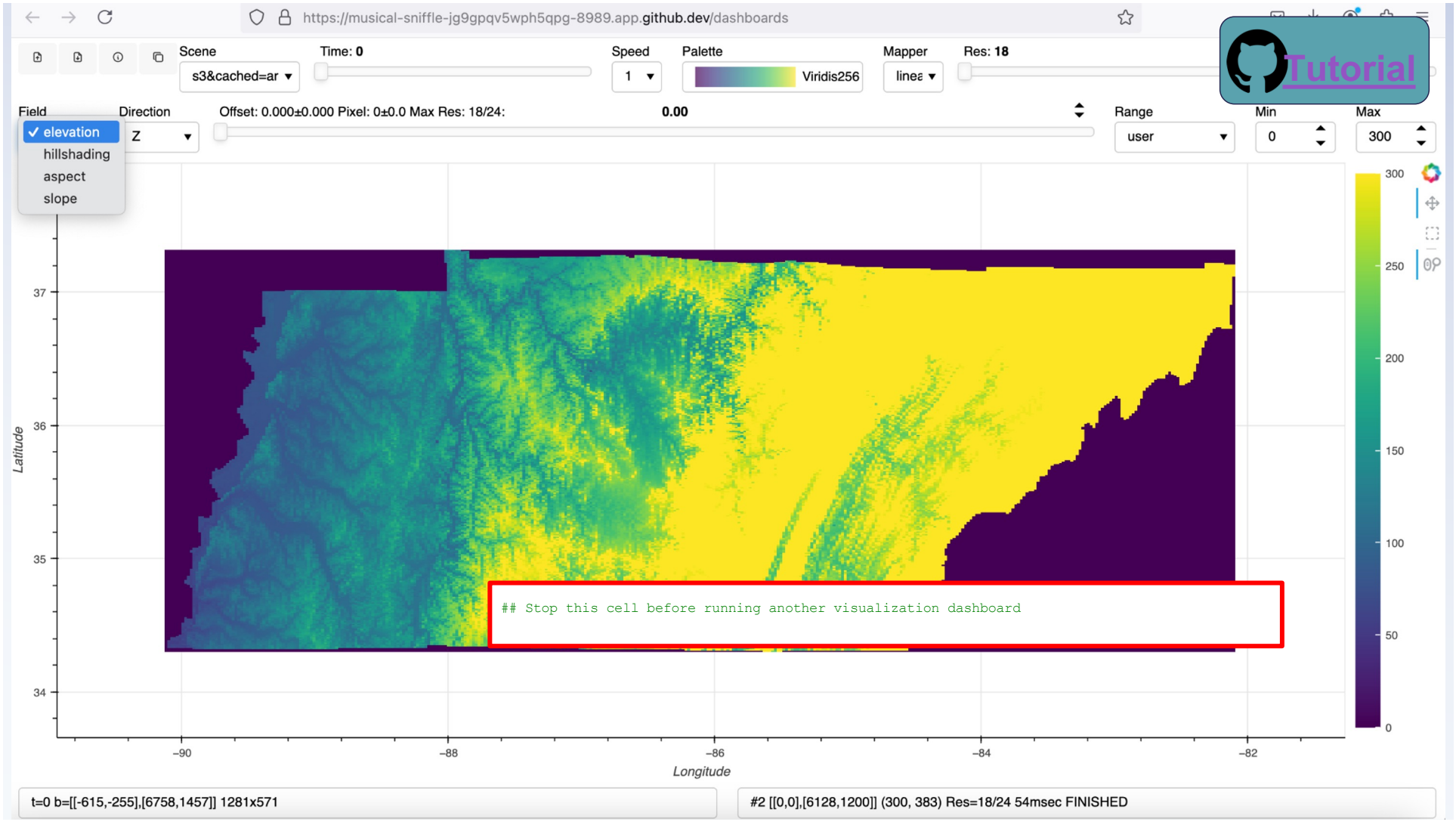
Step 4 provides **two options** to obtain data and collect the IDX files

Option A
From local storage

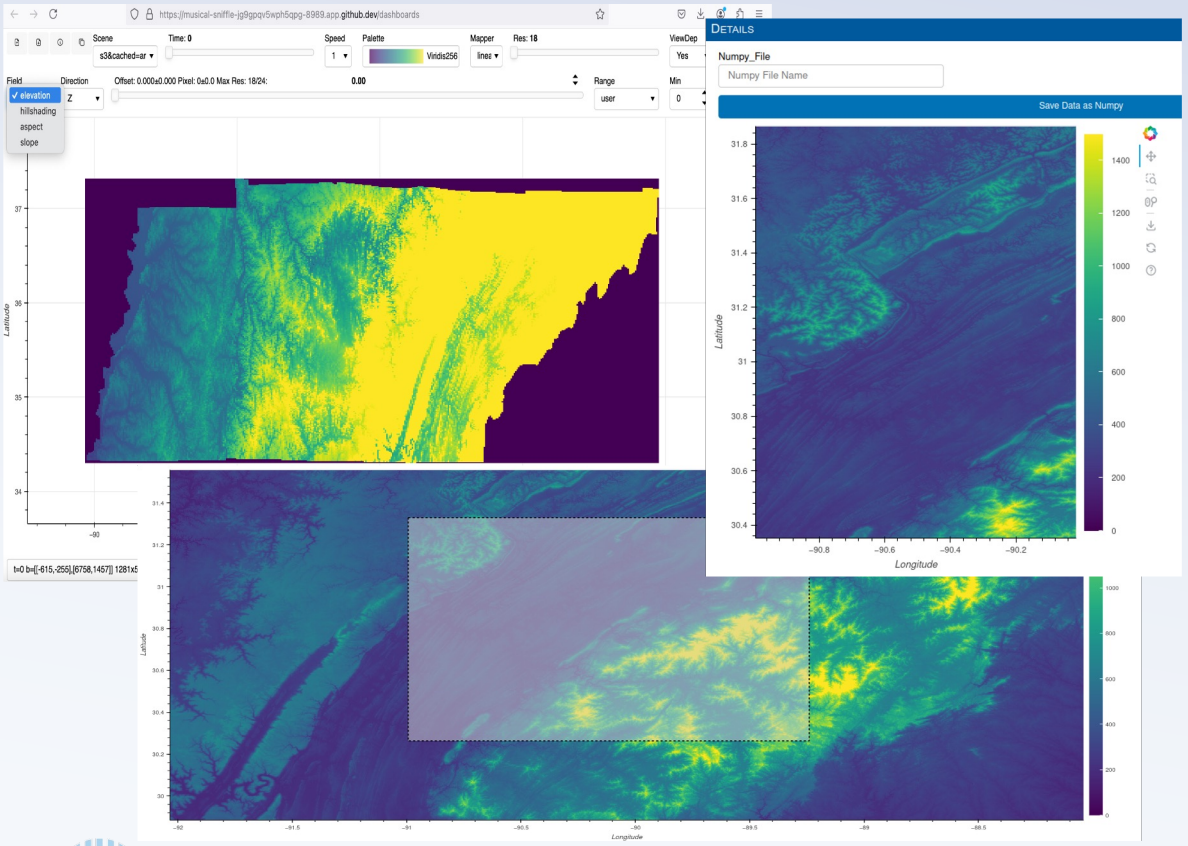
Option B
From Seal Storage

Step 4: Interactive Visualization & Analysis
Launch dashboard for interacting with large-scale data to access subregions of the original dataset for ad hoc analysis.

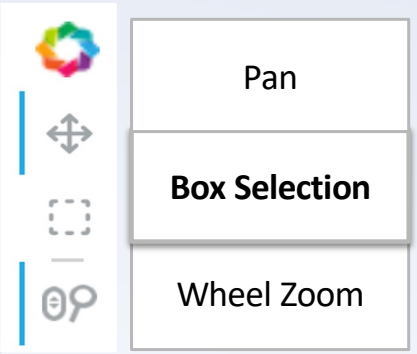



https://musical-sniffle-jg9gpqv5wph5qpg-8989.app.github.dev/dashboards
Tutorial
Scene
s3&cached=ar
Time: 0
Speed
1
Palette
Viridis256
Mapper
linea
Res: 18
Field
elevation
hillshading
aspect
slope
Direction
Z
Offset: 0.000±0.000 Pixel: 0±0.0 Max Res: 18/24:
0.00
Range
user
Min
0
Max
300
Stop this cell before running another visualization dashboard
Latitude
Longitude
37
36
35
34
−90
−88
−86
−84
−82
300
250
200
150
100
50
0
t=0 b=[[-615,-255],[6758,1457]] 1281x571
#2 [[0,0],[6128,1200]] (300, 383) Res=18/24 54msec FINISHED

Step 4: Analysis of Large-Scale Subregions



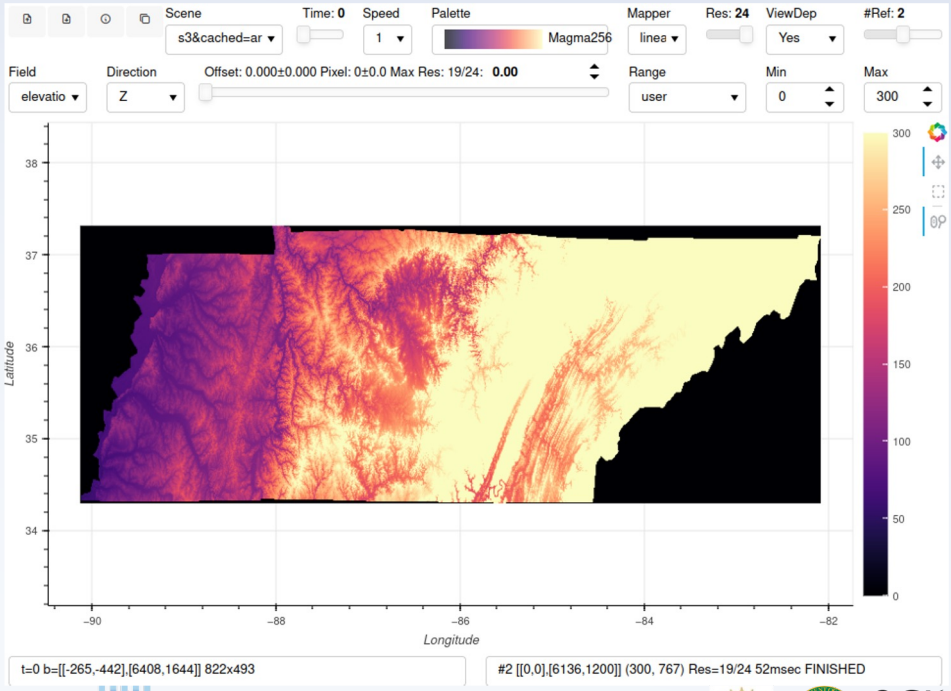
- Visualize large-scale data **remotely**
- Select and explore subregions
- Save the subregions of interest **locally** in file zoom-conus-r01



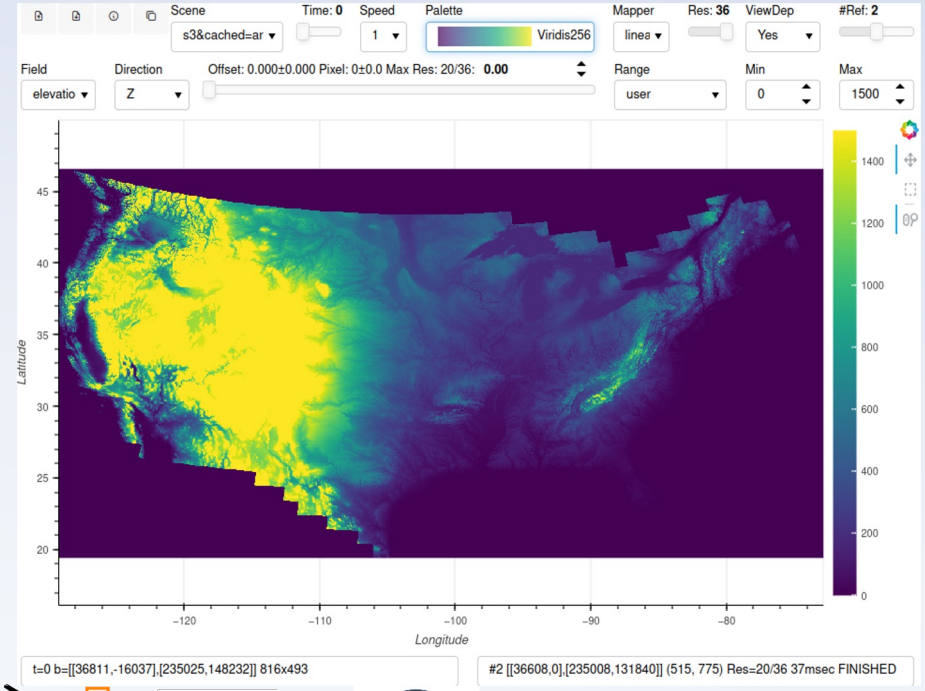
Step 4: Geographical Regions

Visualize and analyze two geographical regions at 30 m resolution

State of Tennessee - 200 MB

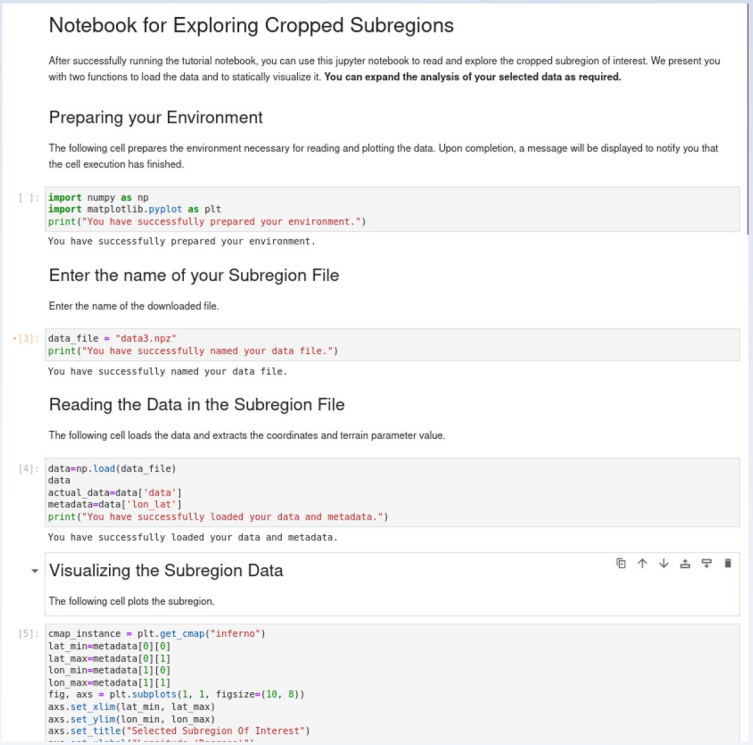


Contiguous United States (CONUS) - 200 GB



Hands on Exercise: Exploring your Subregion Data

(1) Load the downloaded subregion of interest in your local machine
(2) Compute min, max, average and std elevation
(3) Set the color bar to reflect different ranges of displayed data to reflect different needs



Notebook for Exploring Cropped Subregions

After successfully running the tutorial notebook, you can use this jupyter notebook to read and explore the cropped subregion of interest. We present you with two functions to load the data and to statically visualize it. **You can expand the analysis of your selected data as required.**

Preparing your Environment

The following cell prepares the environment necessary for reading and plotting the data. Upon completion, a message will be displayed to notify you that the cell execution has finished.

```
[ ]: import numpy as np
     import matplotlib.pyplot as plt
     print("You have successfully prepared your environment.")
```

```
You have successfully prepared your environment.
```

Enter the name of your Subregion File

Enter the name of the downloaded file.

```
[3]: data_file = "data3.npz"
     print("You have successfully named your data file.")
```

```
You have successfully named your data file.
```

Reading the Data in the Subregion File

The following cell loads the data and extracts the coordinates and terrain parameter value.

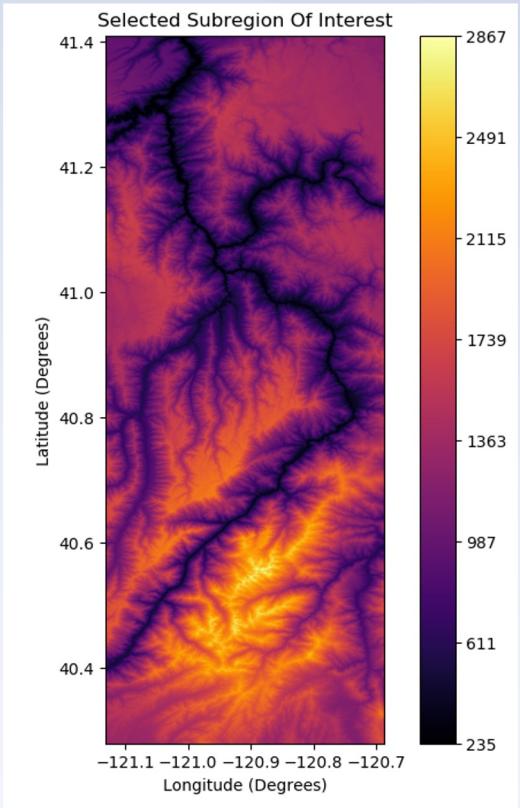
```
[4]: data=np.load(data_file)
     data
     actual_data=data['data']
     metadata=data['lon_lat']
     print("You have successfully loaded your data and metadata.")
```

```
You have successfully loaded your data and metadata.
```

Visualizing the Subregion Data

The following cell plots the subregion.

```
[5]: cmap_instance = plt.get_cmap("inferno")
     lat_min=metadata[0][0]
     lat_max=metadata[0][1]
     lon_min=metadata[1][0]
     lon_max=metadata[1][1]
     fig, axs = plt.subplots(1, 1, figsize=(10, 8))
     axs.set_xlim(lat_min, lat_max)
     axs.set_ylim(lon_min, lon_max)
     axs.set_title("Selected Subregion Of Interest")
```



Hands on Exercise: Exploring your Subregion Data

Challenge I

(1) Load the downloaded subregion of interest in your local machine

(2) Compute min, max, average and std elevation

Challenge II

(1) Load the downloaded subregion of interest in your local machine

(2) Compute min, max, average and std elevation

(3) Set the color bar to reflect the range of displayed data, from the minimum to the maximum value, providing a more accurate visual representation of the data

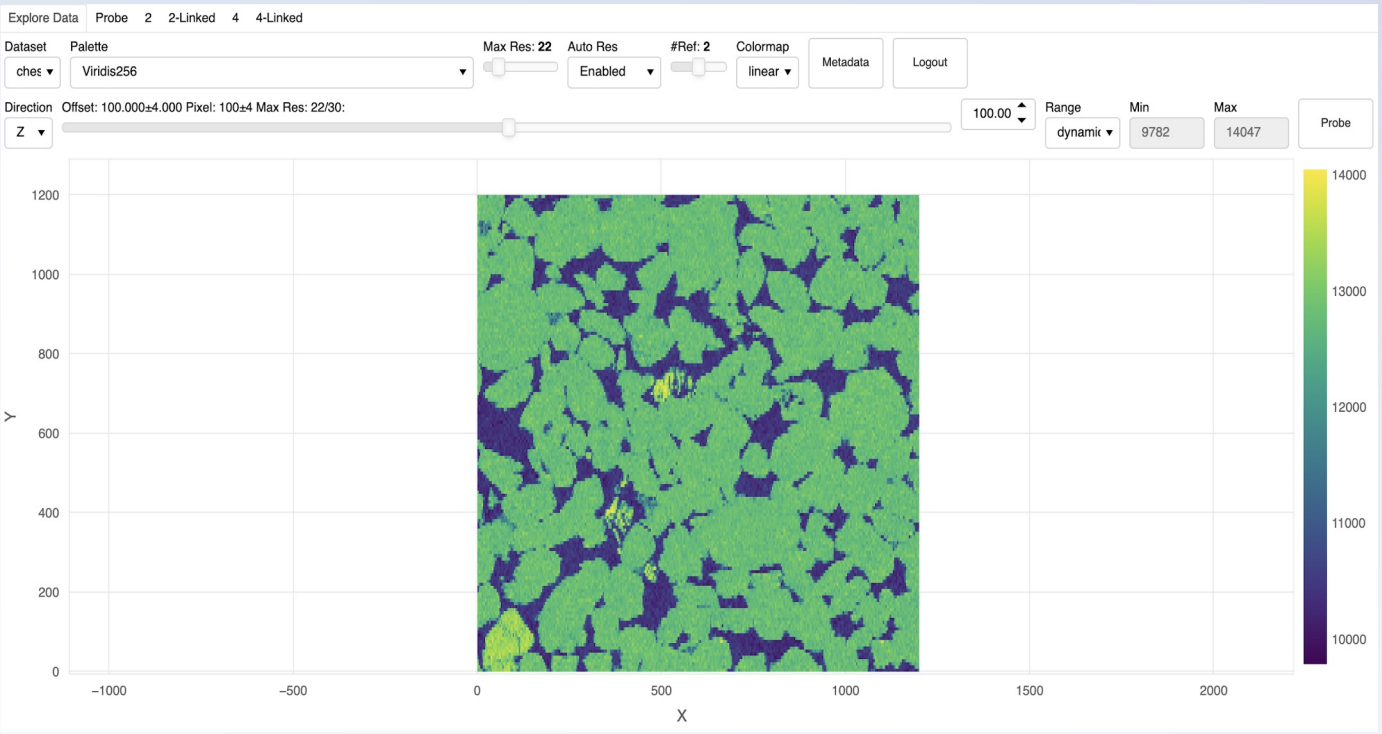
Challenge III

(1) Load the downloaded subregion of interest in your local machine

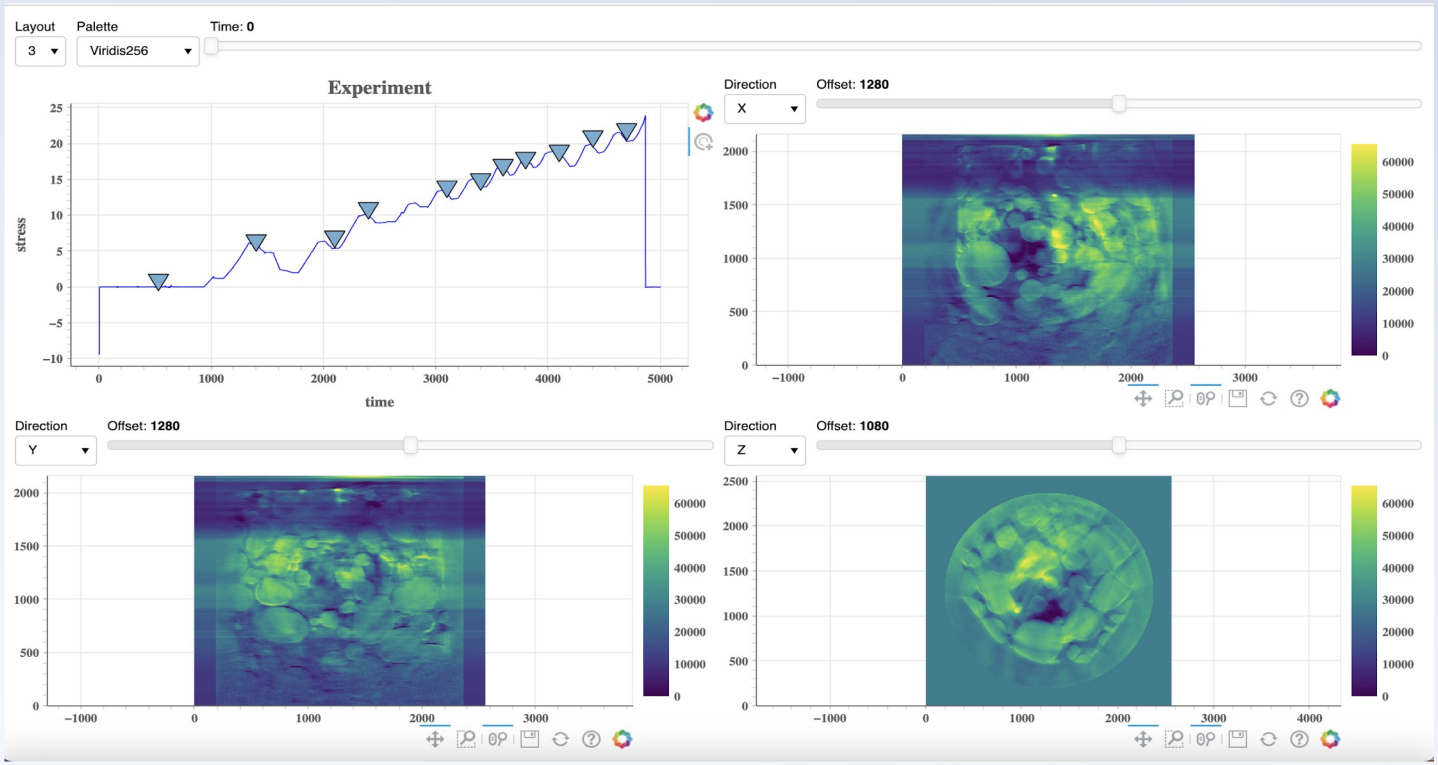
(2) Compute min, max, average and std elevation

(3) Set the color bar to reflect the range of displayed data, from the 0 to the a prefered value (e.g., 600), providing a more accurate visual representation of the data

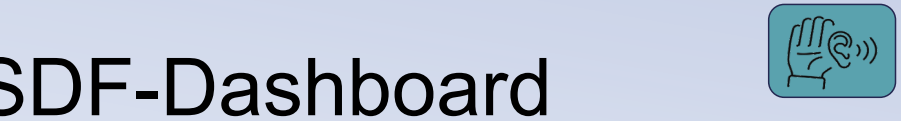
Check Out the CHESS NSDF-Dashboard

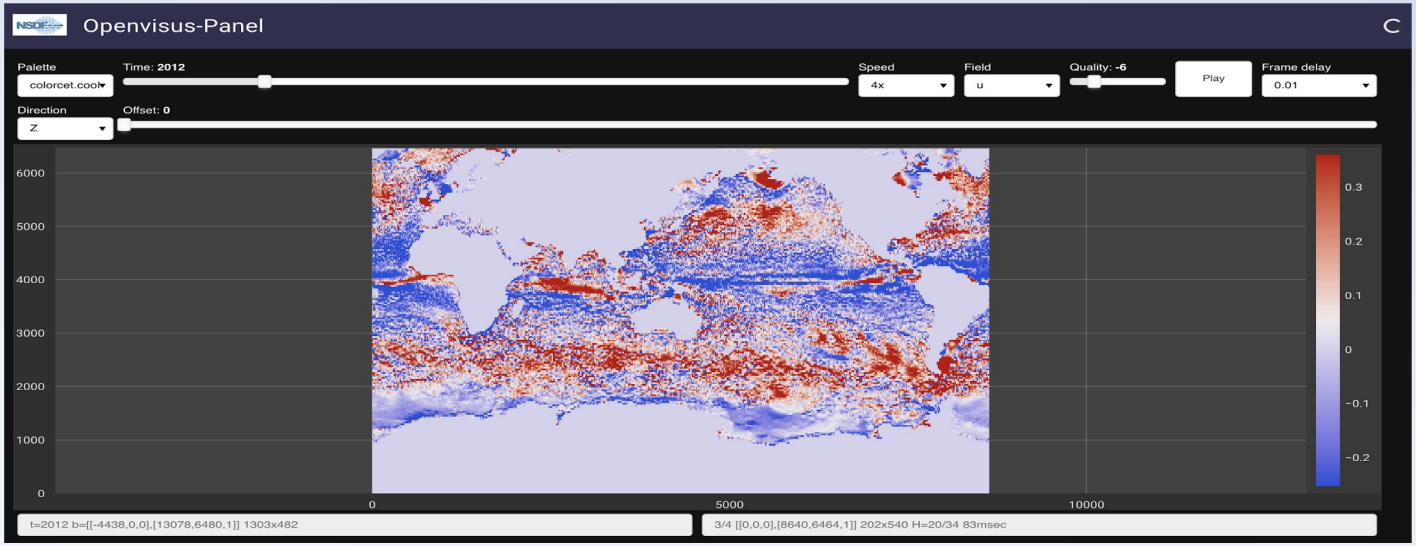


Check Out the Material Science NSDF-Dashboard



Check Out NASA Ocean NSDF-Dashboard





Discussion and Open Questions

Construct a modular workflow that combines your application components with NSDF services

Can you think of an application that is modular? Can you leverage its APIs?
Can the application take advantage of the NSDF services?

Upload, download, and stream data to and from **public and private storage** solutions

How large is the application's data? How do you access, share, and store the data?
Can the data take advantage of private and public storage?

Deploy the NSDF dashboard for large-scale **data access, visualization,** and **analysis**

What type of analysis do you perform on the data?
Can your research take advantage of an interactive dashboard?

Discussion and Open Questions

https://forms.gle/MgYDmiWr8YXo8AYT8

Tutorial Links

NSDF

GEOtiled

SOMOSPIE

OpenVisus