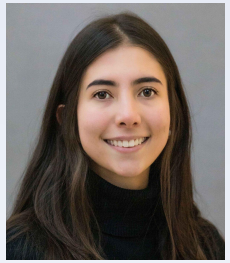
Enabling Scientific Discovery: Harnessing the Power of the National Science Data Fabric for Large-Scale Data Analysis (Session I & II)

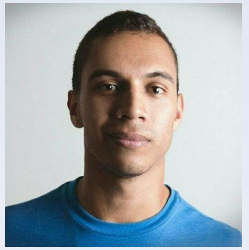
Presenter: Michela Taufer[1]
Other contributors: Heberth Martinez[1], Paula Olaya[1],Asshish Panta[2], Amy Gooch[2], Jack Marquez[1], Gabriel Laboy[1], Jay Ashworth[1], Giorgio Scorzelli[2] and Valerio Pascucci[2]

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










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.

Schedule

The half-day tutorial is organized into four sessions:

Session I (30 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 100TB.

Session IV (30 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

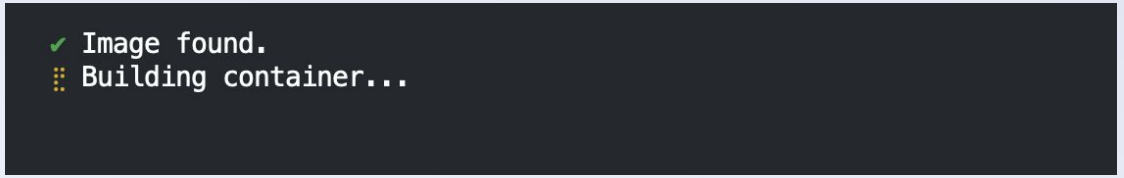
Step 1: Create Codespaces

Use your GitHub account to run this tutorial with GitHub codespaces

- ➔ Access this link:

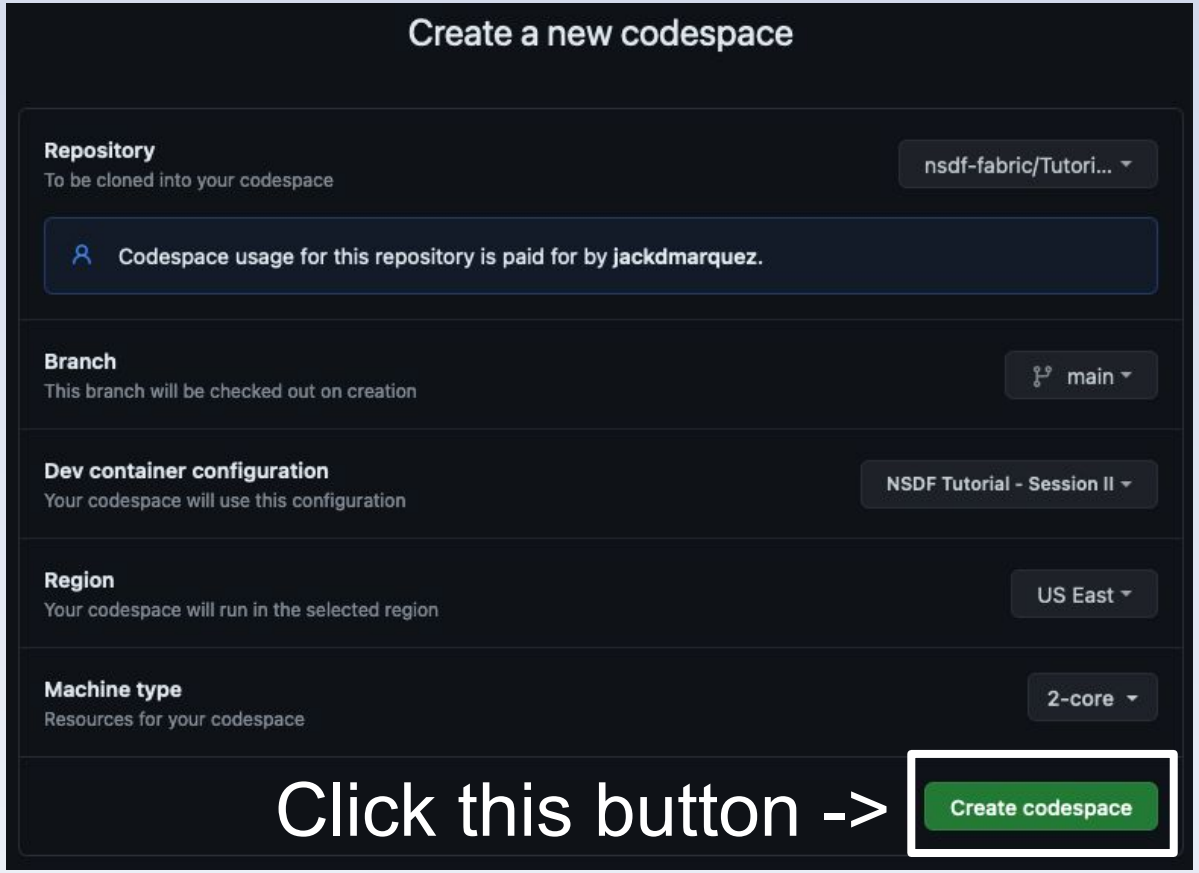
  NSDF Tutorial 2024

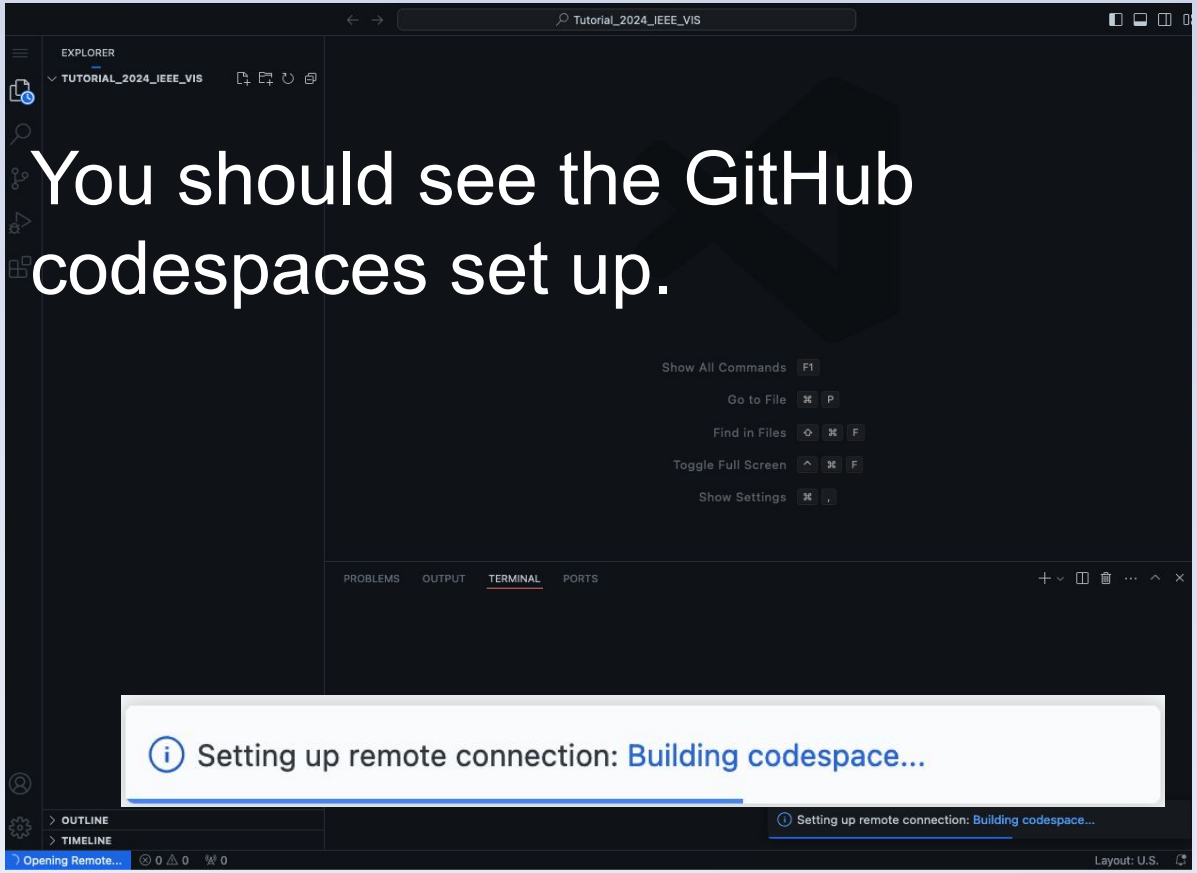
- ➔ Click on green button "Create codespace"



Creating the GitHub Codespace

Tutorial





➔ Let's start:
[NSDF Tutorial 2024]

Loading GitHub Codespace can take from 1 to 5 minutes

Tutorial Goals

This tutorial demonstrates end-to-end analysis of scientific data through 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**

Session I: Understanding and Addressing User's Pain Points

Surveying Community Needs and Realities

Identifying Pain Points: User Interviews

Identify Users

- 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

Target Questions

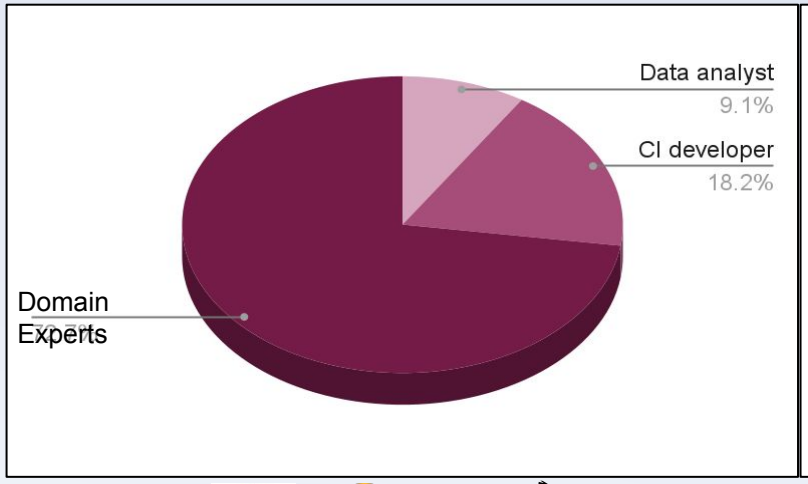
- General questions about data storage, data form, metadata storage, WMS, data catalogs, programming languages
- Specific questions about unique challenges related to role, domain, and institution

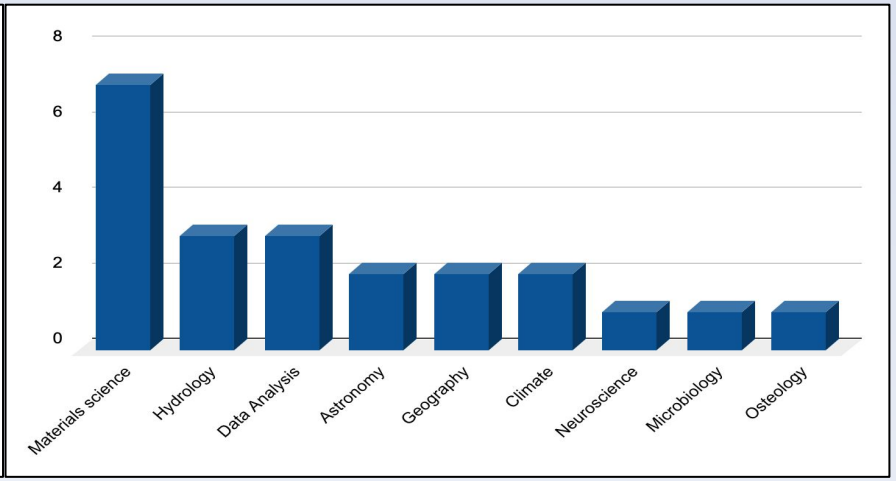
Analyze Results

- Identify cross-cutting concerns
- Identify concerns consistent for roles, domains, and institutions

Diagnose Pain Points

- Distill concerns into concrete problem statements
- Translate into objectives, actionable items, and milestones





Identifying Pain Points: Testimonials

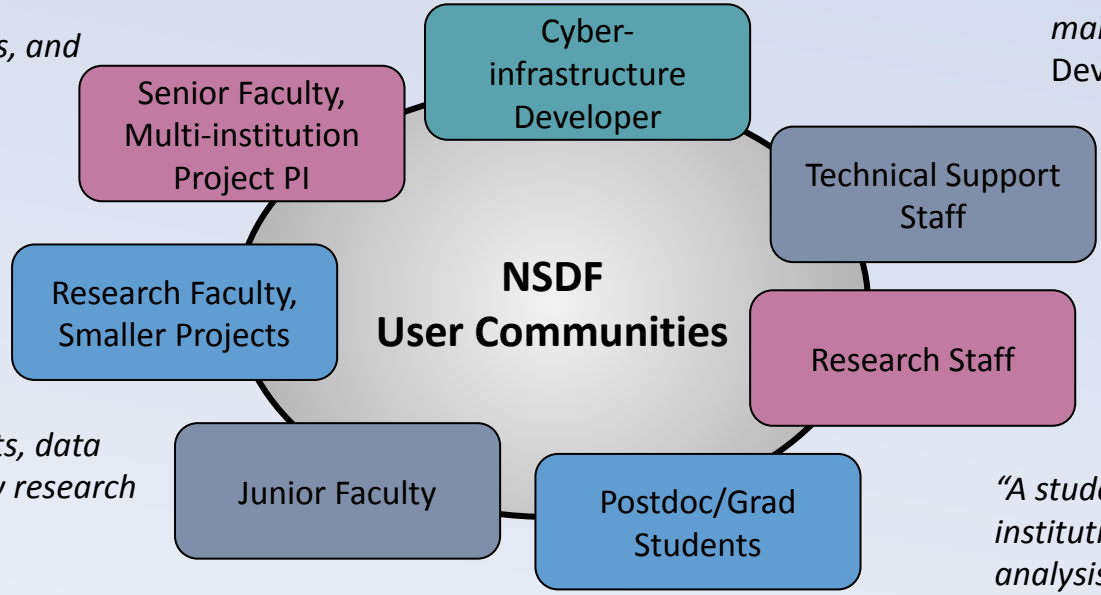
"If we make it easy, people will share their data. We need to extend the scalability of our infrastructure through community supplied storage." - Project PI, Materials Science

"We don't have a plan to scale storage if our cyberinfrastructure takes off." - Cyberinfrastructure developer, Hydrology

"We can't scale to PB data without massive funding or infrastructure. We need centralized point of access to federated data" - Cyberinfrastructure Developer, Materials Science

"We lack personnel to do basic development and maintenance of our systems" - Cyberinfrastructure Developer, Astronomy

"The time and effort for using public repositories, and limited realized gains limits our data sharing" - Faculty, Materials Science



"We can't hire enough system maintainers and *have research funding"* - Data Analyst Group Lead

"Remote quality control during acquisition would let us better use beamline time" - Faculty, Materials Science

"Jupyter notebook access to data (TBs) would reduce barrier to entry " - Research Staff, Climate Science

"Before I had funding to run my own experiments, data shared by a friend at a national lab launched my research career" - Faculty, Materials Science

"A student copies GBs of data from the scientist to my institution. I download to my laptop to prototype analysis. It is cumbersome and limits testing." - Research Staff, Data Analysis

"I'm perplexed by the lack of urgency around reproducible and replicable processes for data management" - Faculty, Geography

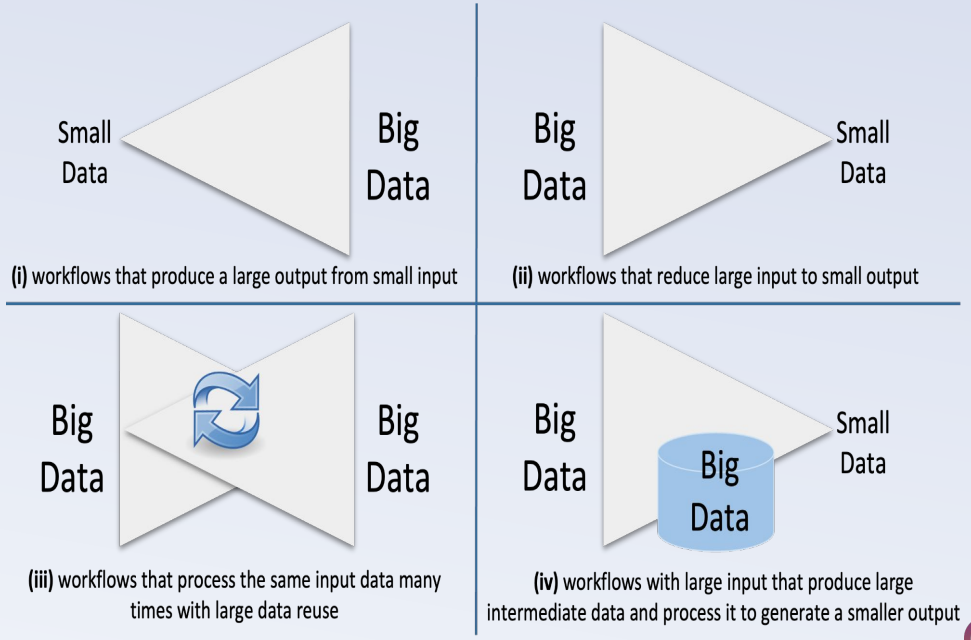
"Our old data is kept on (external) drives. It's hard to keep things organized" - Graduate Student, Materials Science

"Our long-term storage is a shelf of external hard drives" - Research Staff, Neuroscience

"We move data (from light source) by flying back with TB hard drives" - Graduate Student, Materials science

Pain Points Inform Data Science Strategies

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



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

What do Scientists Need?

Information streaming: knowledge exploration at run time

Findable, Accessible, Interoperable, Reusable (FAIR) data

Equity in access and use

Access and use of data from remote facilities (experimental, computing, and repository)

Remote queries with minimal data movement to mitigate bandwidth and storage issues

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**.

Institutions and universities with resources to share


100G Core
Terabit Core
NSDF EntryPoints
OSG StashCaches
Both
UTAH
UMICH
JHU
SDSC
UTEP
UC Denver
UTK
UTSA
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
UTK
SDSC + OSG
Digital Rocks
Initiative to integrate large scale scientific projects
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
UTAH/SCI
LLNL
CHPC
CHESS/Cornell
MGHPCC + OSN
JHU
SDSC + OSG
SNL
UTK
Digital Rocks
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
PRISMS
Materials
Commons
UMICH
UTAH/SCI
LLNL
CHPC
CHESS/Cornell
MGHPCC + OSN
JHU
Internet2
SDSC + OSG
SNL
UTK
MS-CC
TACC
XSEDE/ACCESS
Jetstream2
CloudLab
Chameleon
CyVerse
Initiative to integrate public cloud 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
PRISMS
Materials
Commons
UMICH
UTAH/SCI
LLNL
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

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

National Science Data Fabric

National Science Data Democratization Consortium: Engaging Industry Partners

MINIO

SEAL

WEKA

CLOUDFLARE

NSDDC

IBM Cloud

DoubleCloud

intel

ALLUXIO

Partnering with Existing NSF Initiatives

Chameleon

open storage network

NATIONAL DATA PLATFORM

FABRIC

CYVERSE®

WIRED GLOBAL CENTER
U.S.-Canada Center on Climate-Resilient Western Interconnected Grid

CICompass

ACCESS

CloudLab

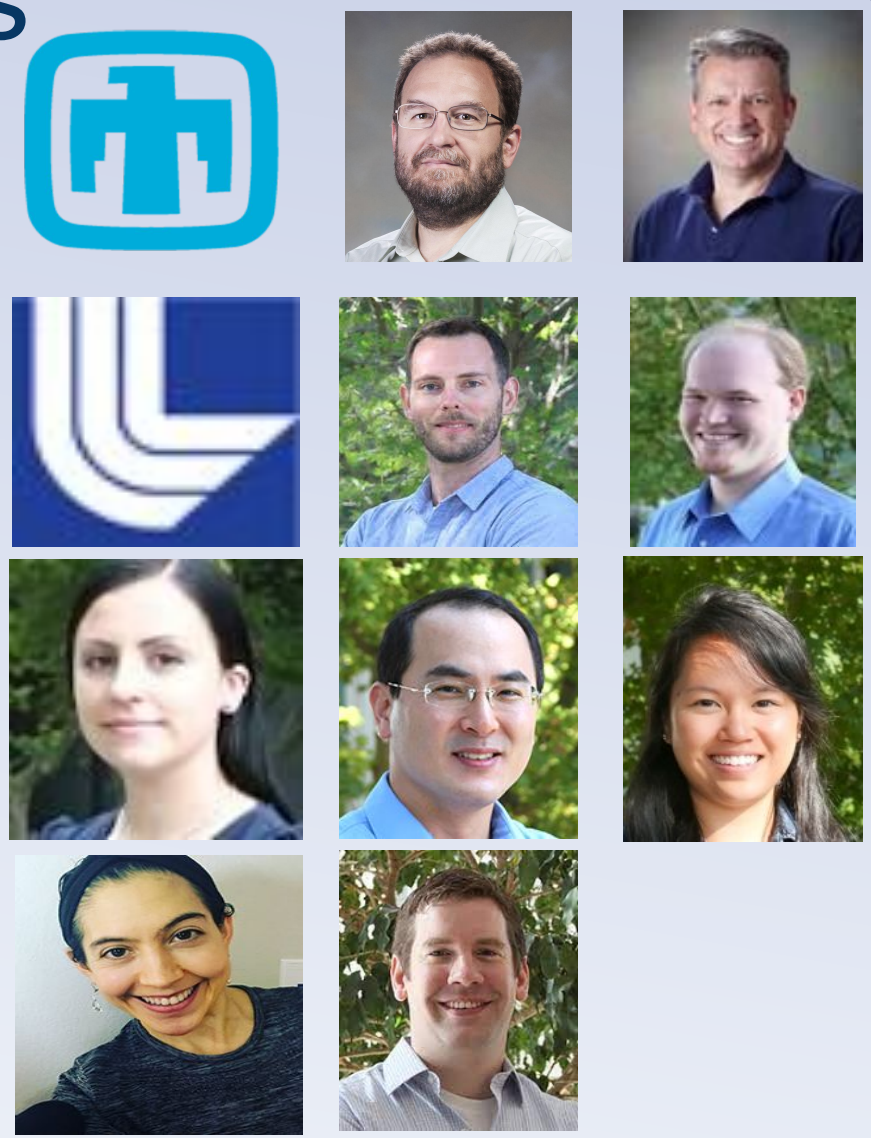
CloudBank

Open Science Grid

Partnerships with DoE Labs

- **Sandia National Laboratories**
  - Workflow containerization (Trustworthy Computing; Data Democratization)

- **Lawrence Livermore National Laboratory**
  - Thicket project (Large Scale Computing and Performance)
  - Flux project (Scheduling and Resource Management)
  - Fractale (Convergence of HPC, Cloud, and Edge)



Session I: Our Services and Successful Stories

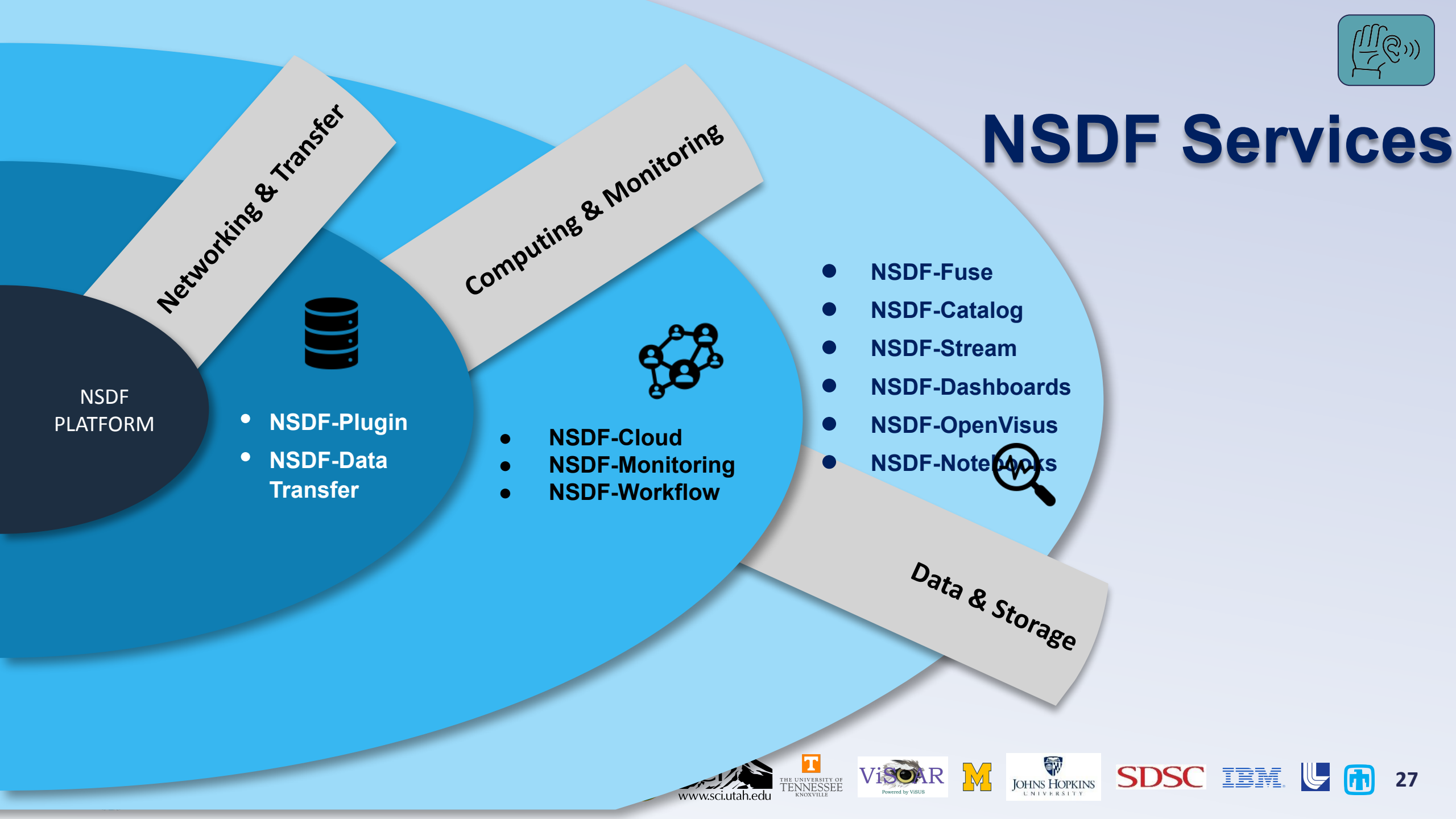
Democratizing Access and Use of Large-scale Data

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

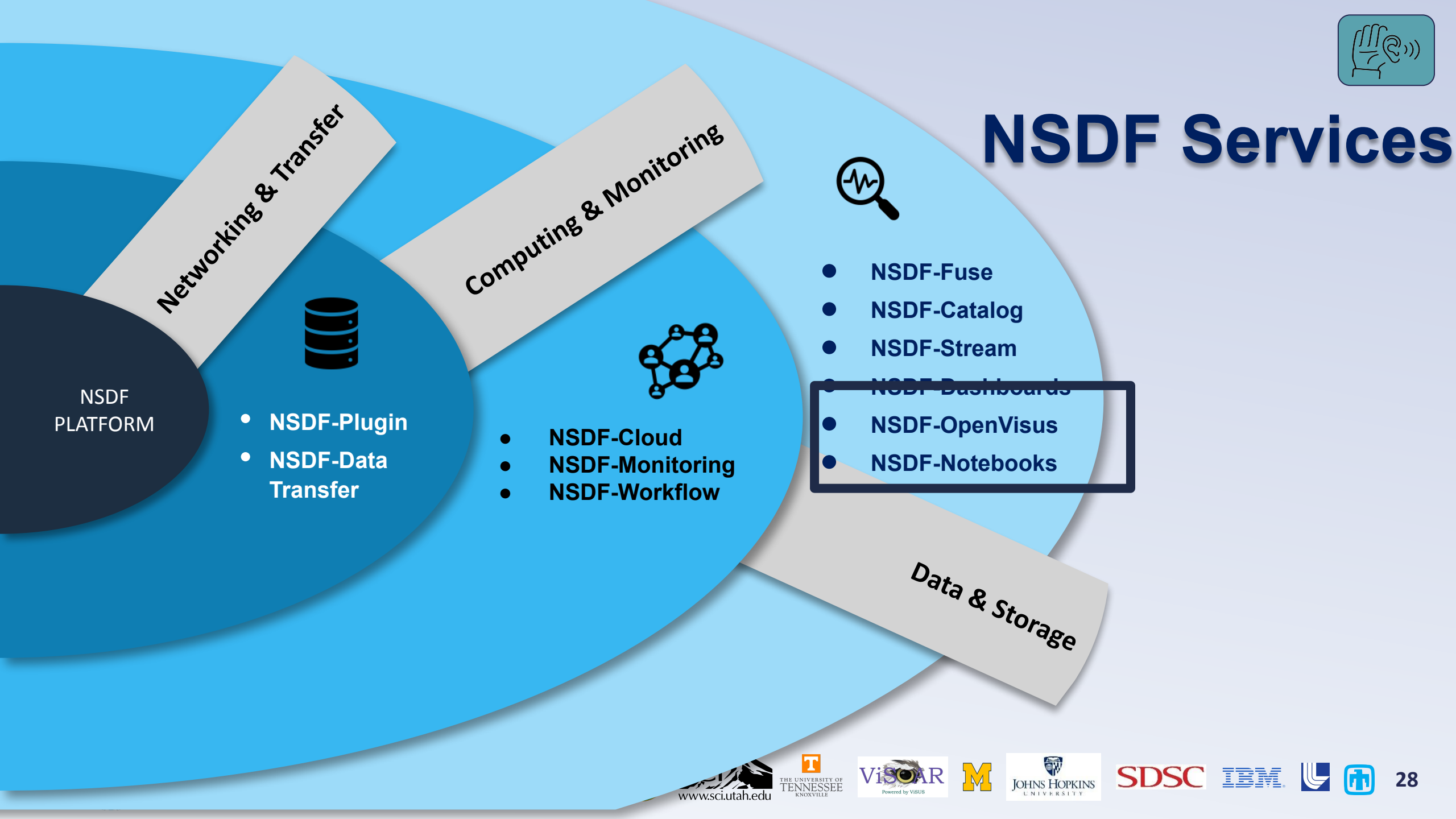
Develop a **FAIR, AI-ready, transdisciplinary software stack** that is easy to use, integrate, and scale

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

NSDF Services



NSDF Services



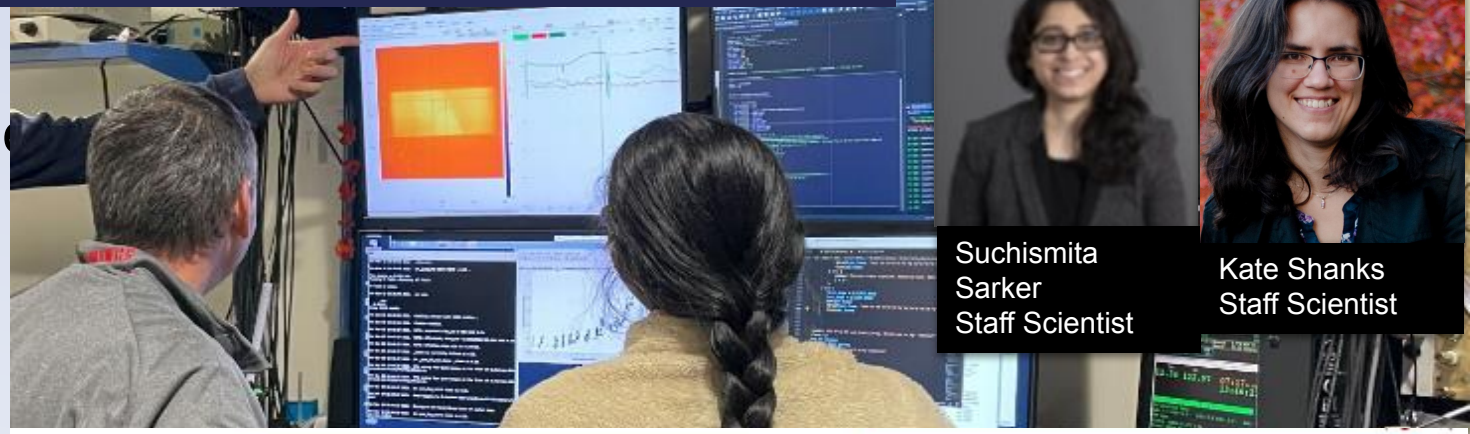
Section I: Sharing Use-Inspired Research Stories

Decentralizing Research Hubs for Transformative Scientific Discovery

Equity in the Use of Experimental Facilities

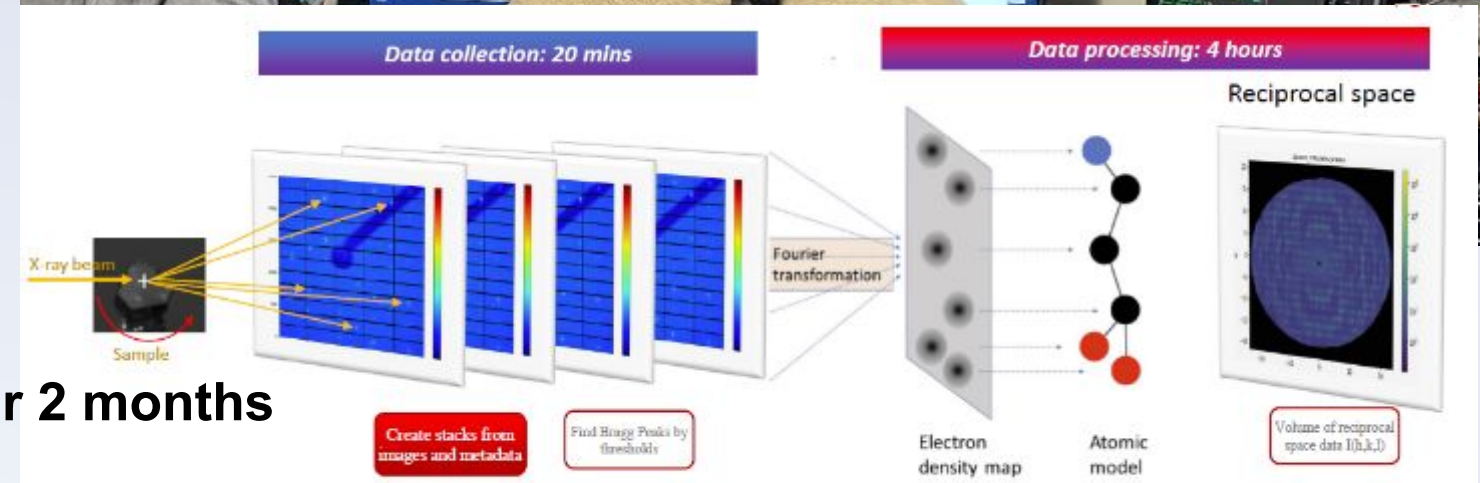
Establishing a comprehensive workflow from experimental facilities to the end-user data analysis.

- Cornell High Energy Synchrotron Source: Quantum Materials Beamline (6 of the 8 lines)
- Real-time data access and sharing
- Steering experiments
- AI-driven workflows
- Remote team collaboration
- Optimize effective use of scientists' time and national resource
- Publish data with no delay (e.g., Materials Commons) □ in real-time



Suchismita Sarker
Staff Scientist

Kate Shanks
Staff Scientist

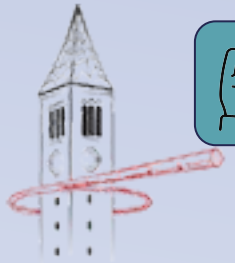


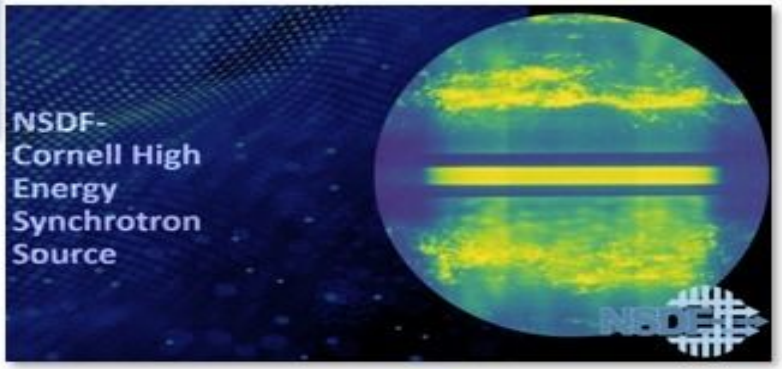
200TB data, 10M files in 2 beamlines over 2 months

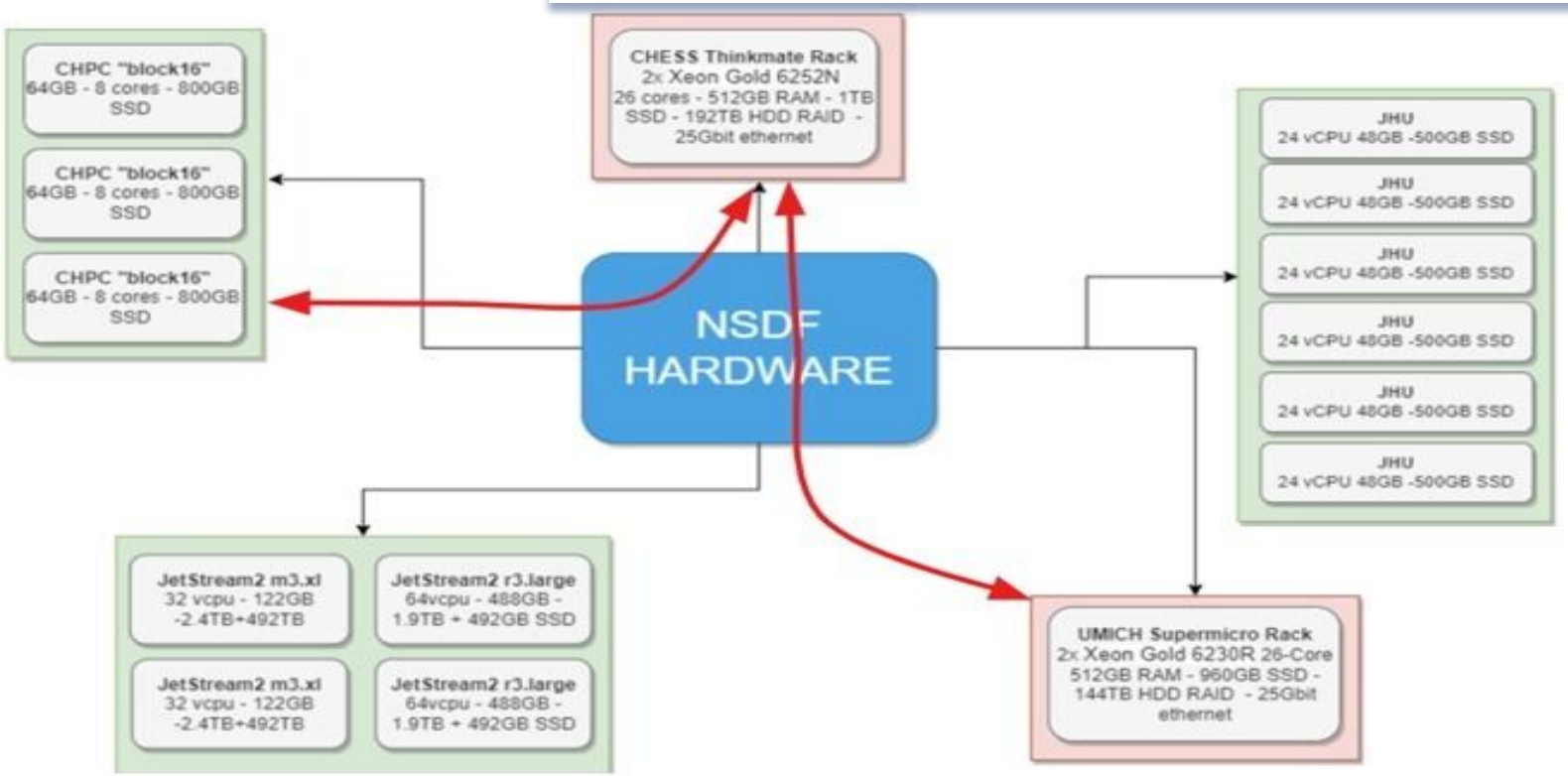
Direct Data Access & Steering from Cornell High Energy Synchrotron Source (CHESS)

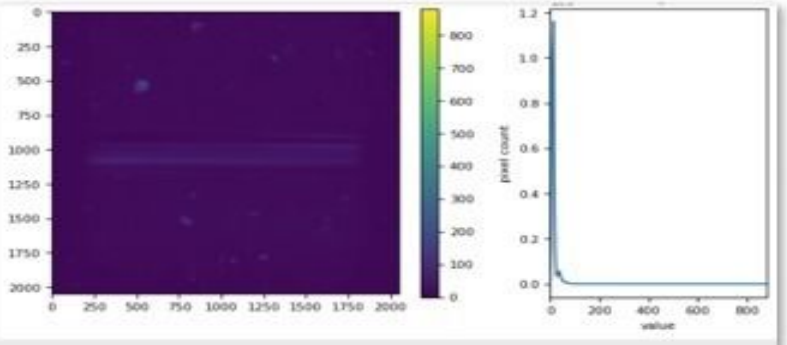
Werner Sun · 2nd
IT Director, Cornell Laboratory for Accelerator-Based Sciences a
Education at Cornell University
Ithaca, New York, United States · Contact info

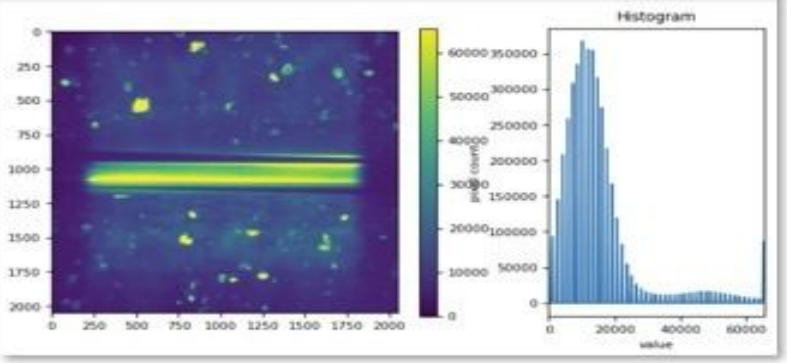
CHESS
CORNELL HIGH ENERGY SYNCHROTRON SOURCE





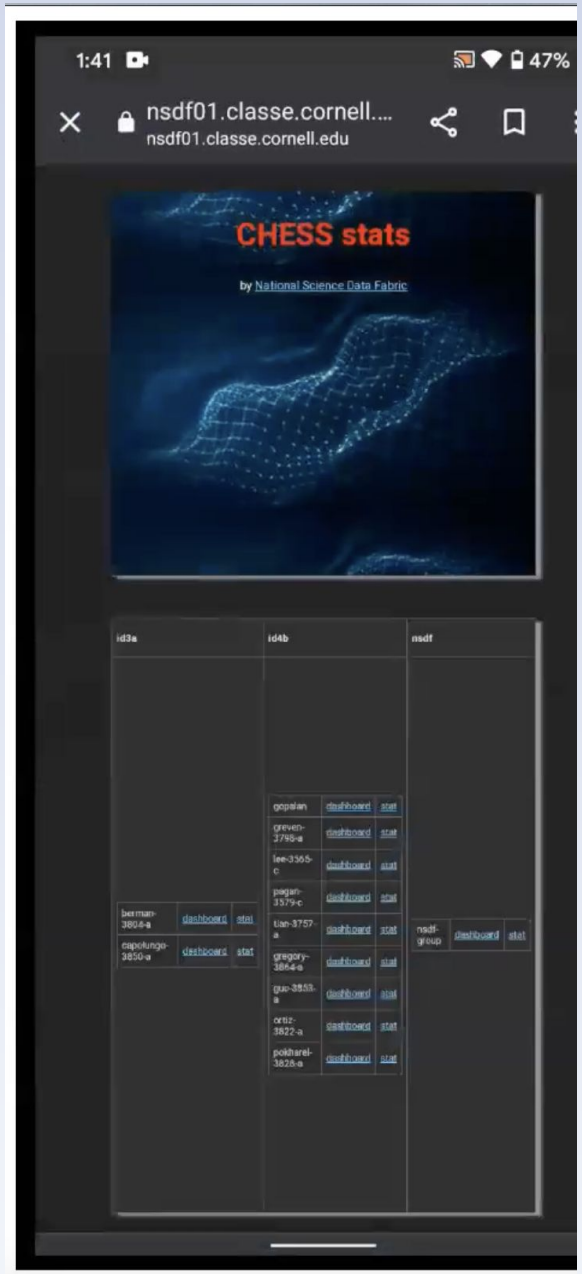






~1500 datasets
Uncompressed size 7.7TiB
Compressed size 2.4TiB

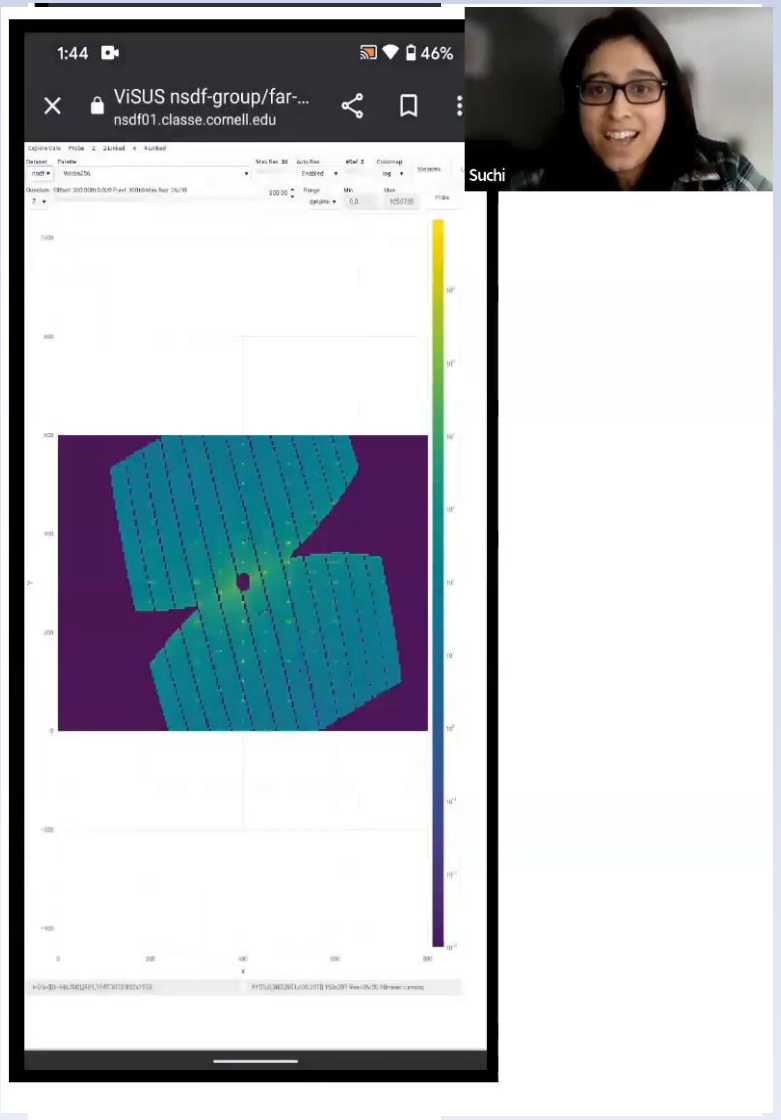
NSDF National Science Data Fabric



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

– Suchismita (Suchi) Sarkar
Staff Scientist, CHESS

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

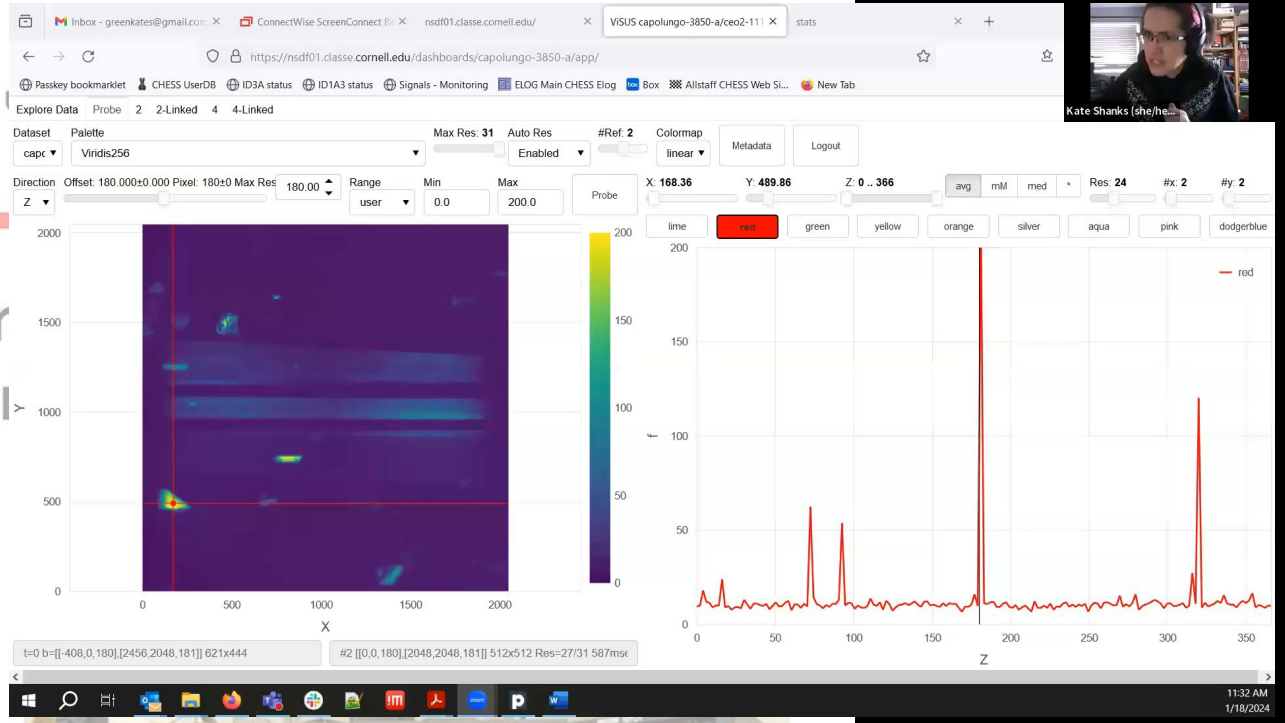
January 11, 2024 | Savan DeSouza

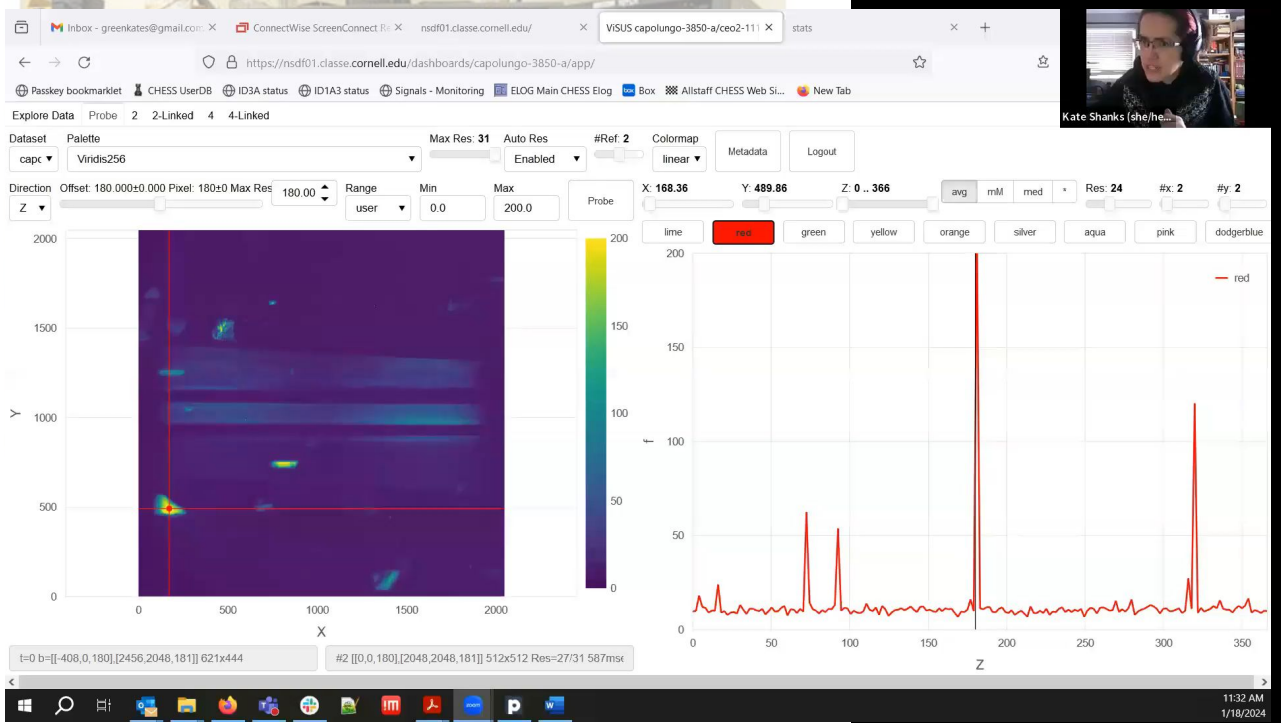
med at
ntific
e Data
High
ESS)
g a
cess to
red
uting

“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..”

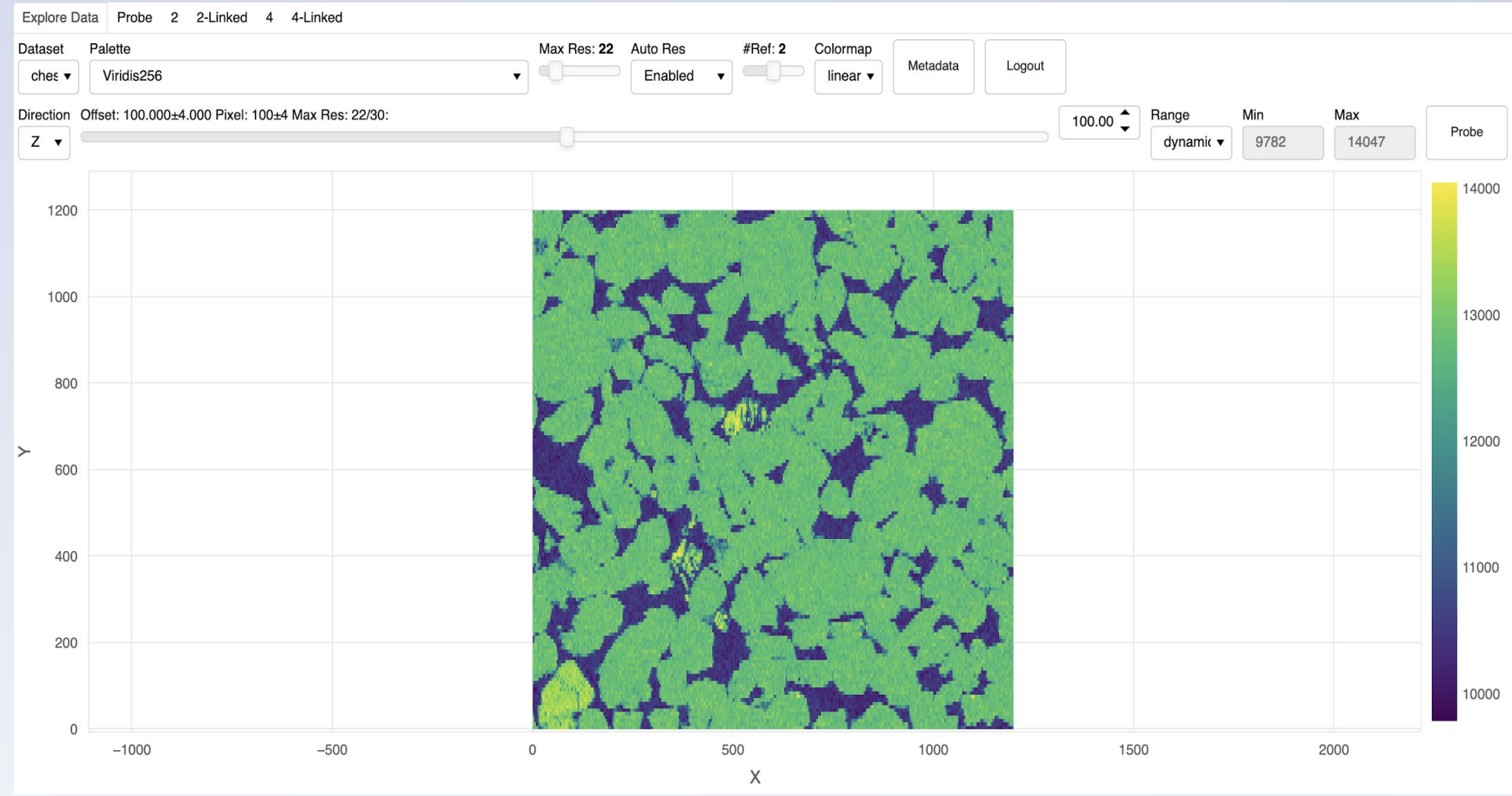
-Kate Shanks, FAST beamline scientist

https
natio
d-ch
ze-da
ery



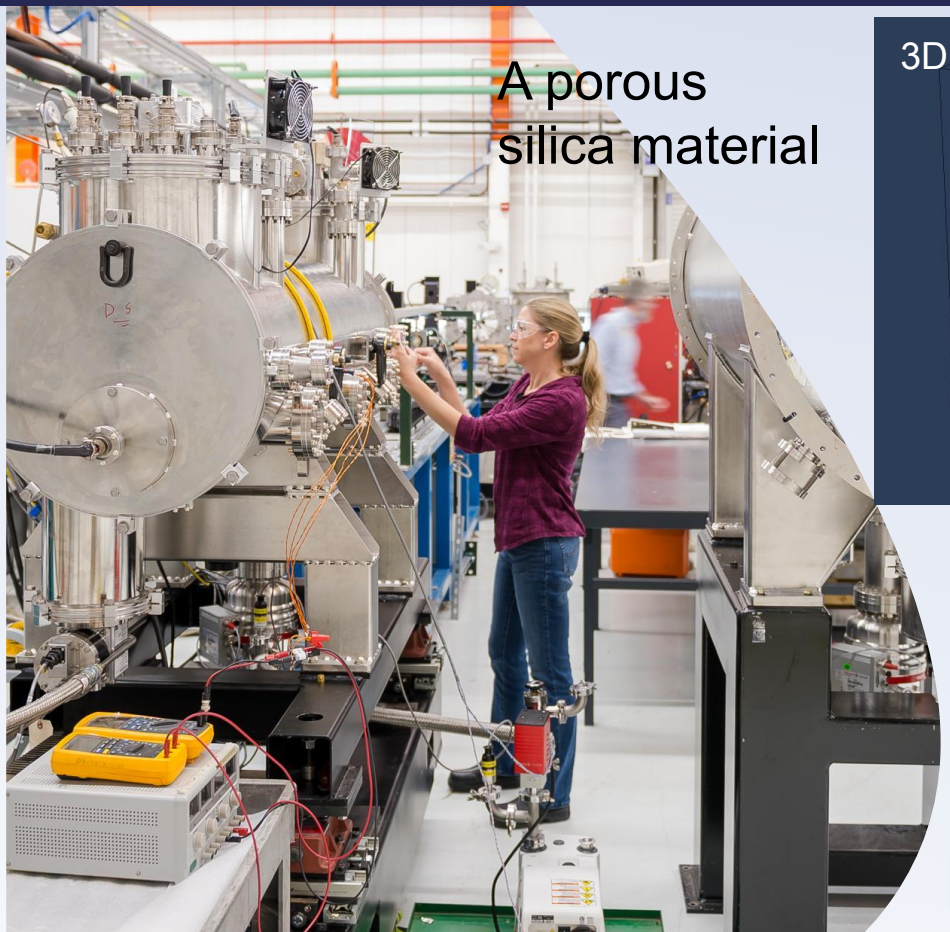


Check Out the CHESS NSDF-Dashboard

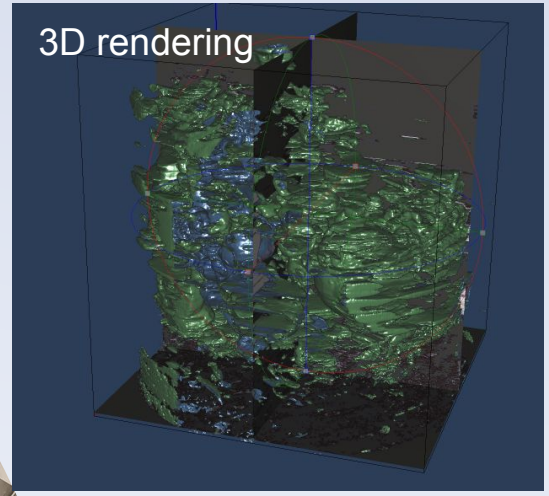


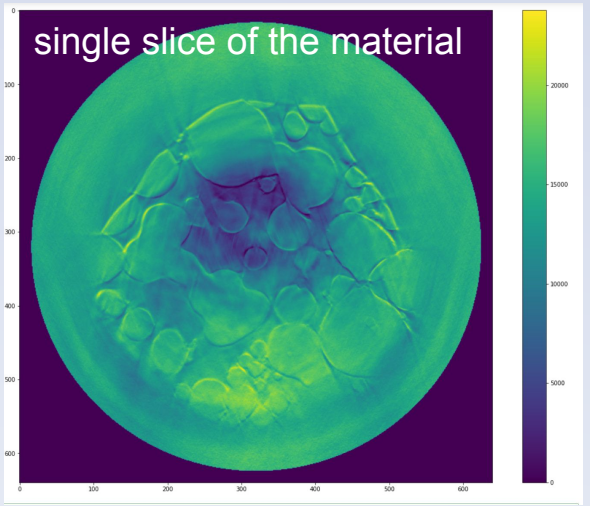
Equity in Research with National Labs

Facilitating rapid processing of 100+ terabytes of data by compact virtual laboratories, achieving in days what would take months for moving data between national labs and universities.



A porous silica material





Pania Newell
Professor in Mechanical Engineering at UoU

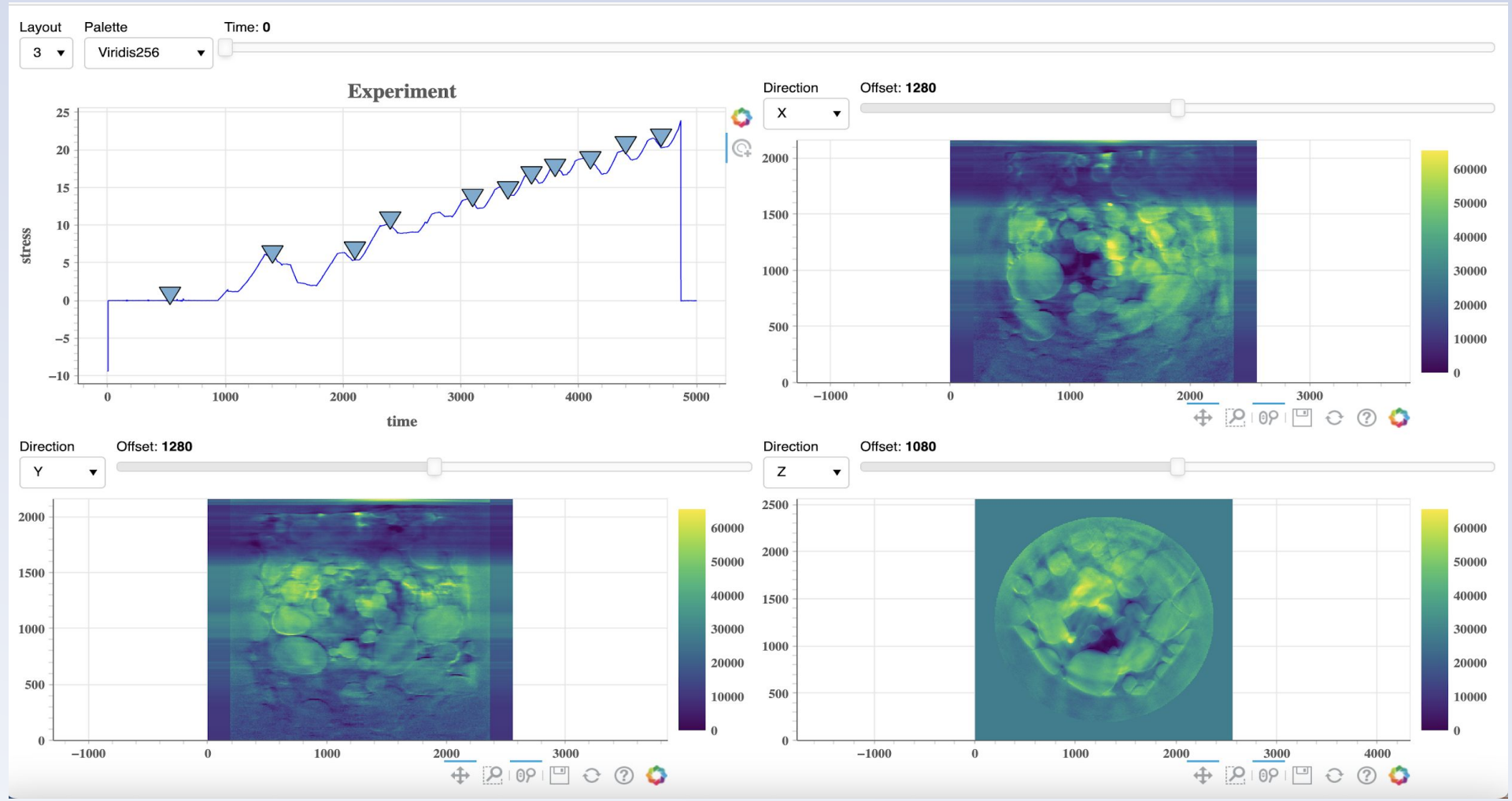
THE UNIVERSITY OF UTAH
MECHANICAL ENGINEERING

Analysis of **the internal structure of porous silica materials** at different stress levels

Over **400TB of generate data** and more than 200 machines used on CloudLab, Chameleon, & AWS

http://services.nationalsciencedatafabric.org/materialscience

Check Out the Material Science NSDF-Dashboard



National Science Data Fabric

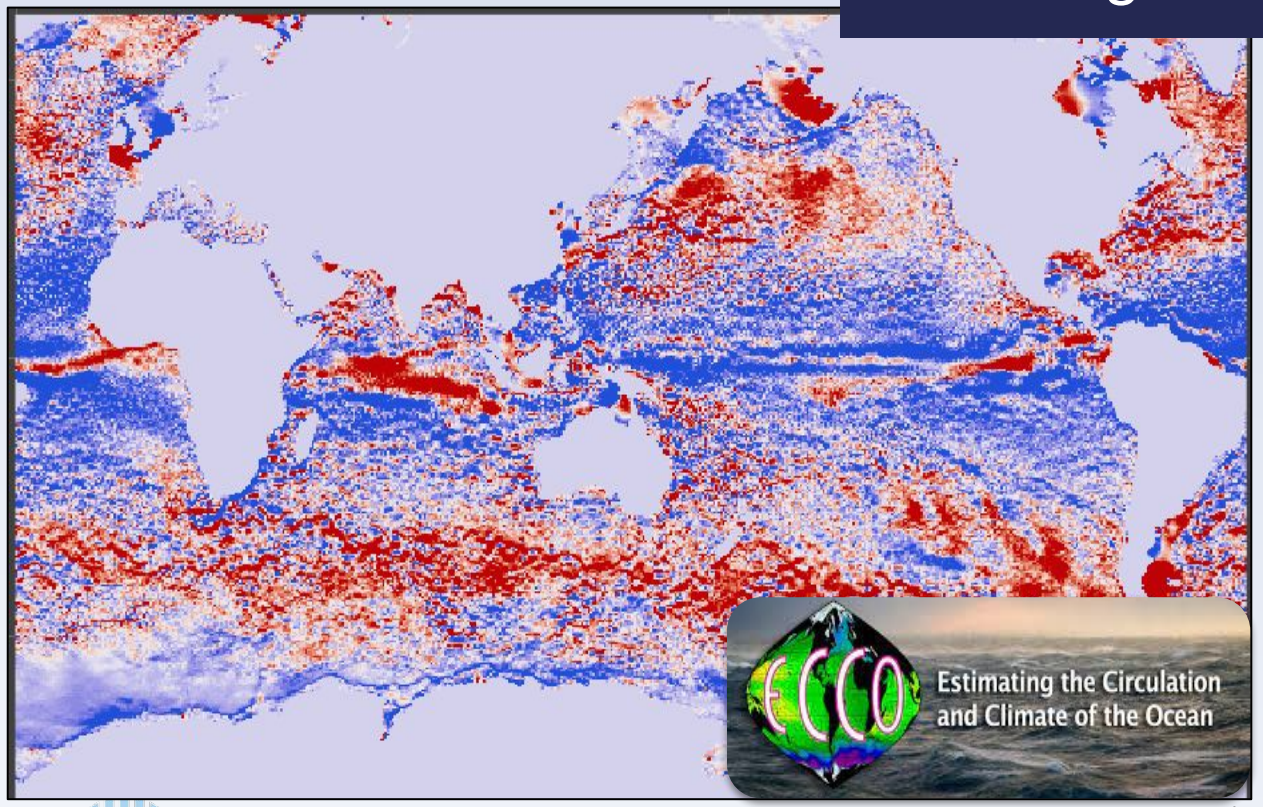
Nina McCurdy · 1st

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)



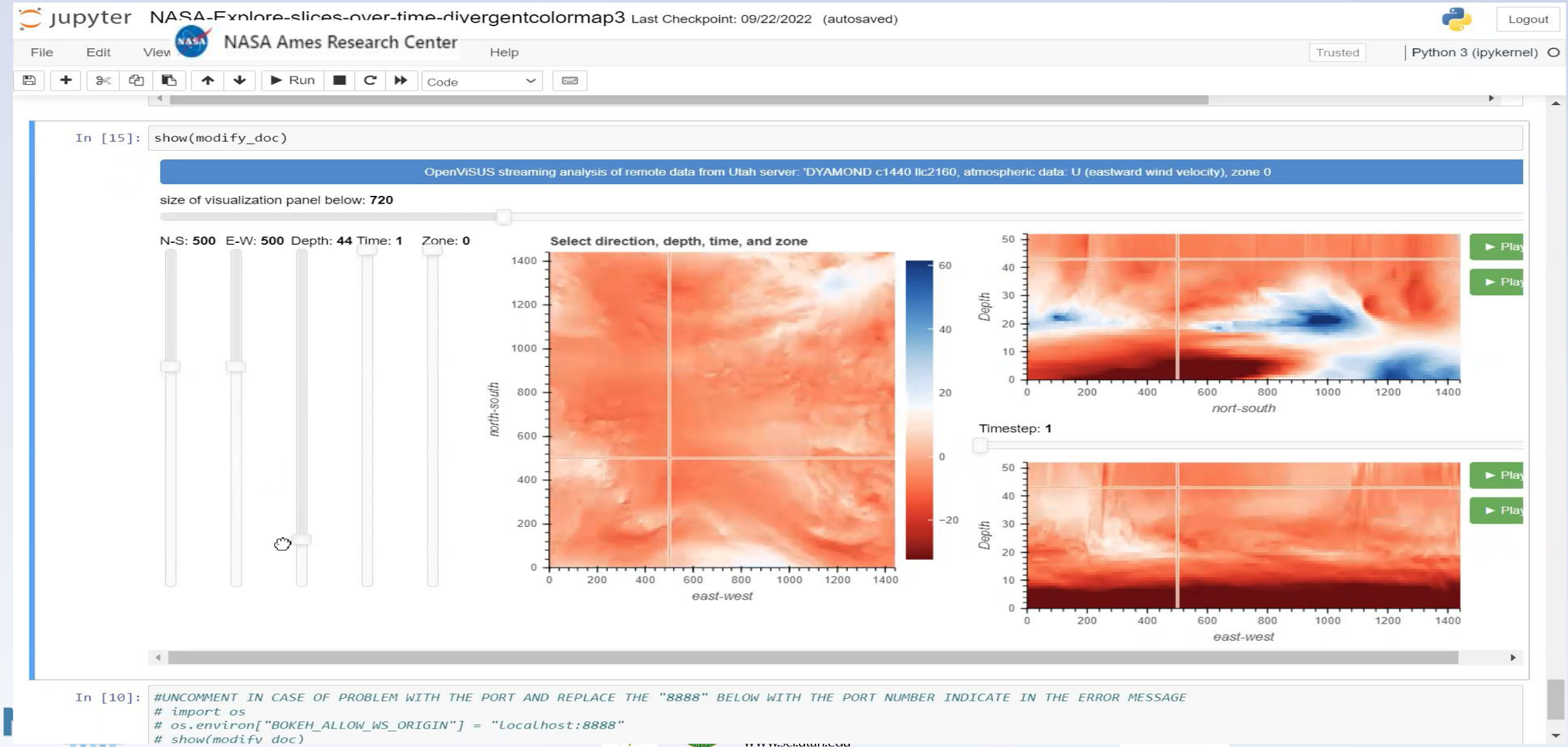
LLC4320 2.8 PB Ocean dataset

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

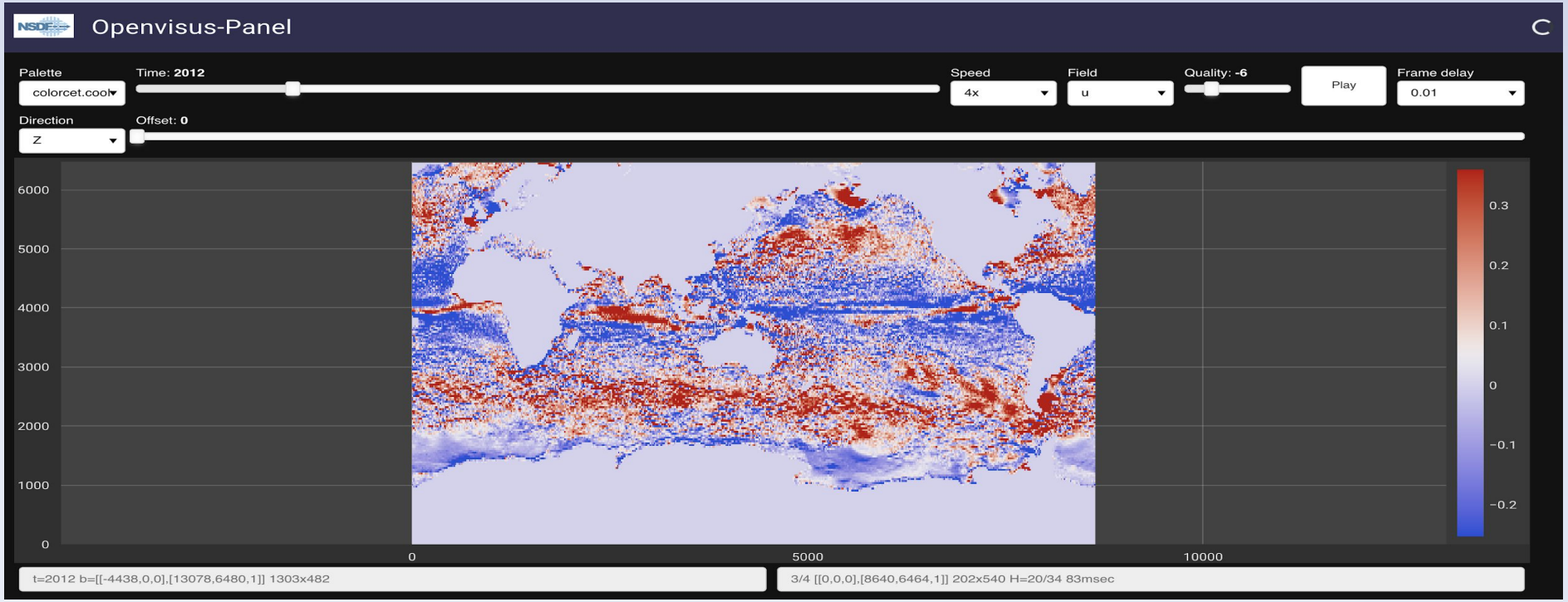
Requirements:

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

Streaming Data Access for Run-Time Analysis



Check Out NASA Ocean NSDF-Dashboard



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

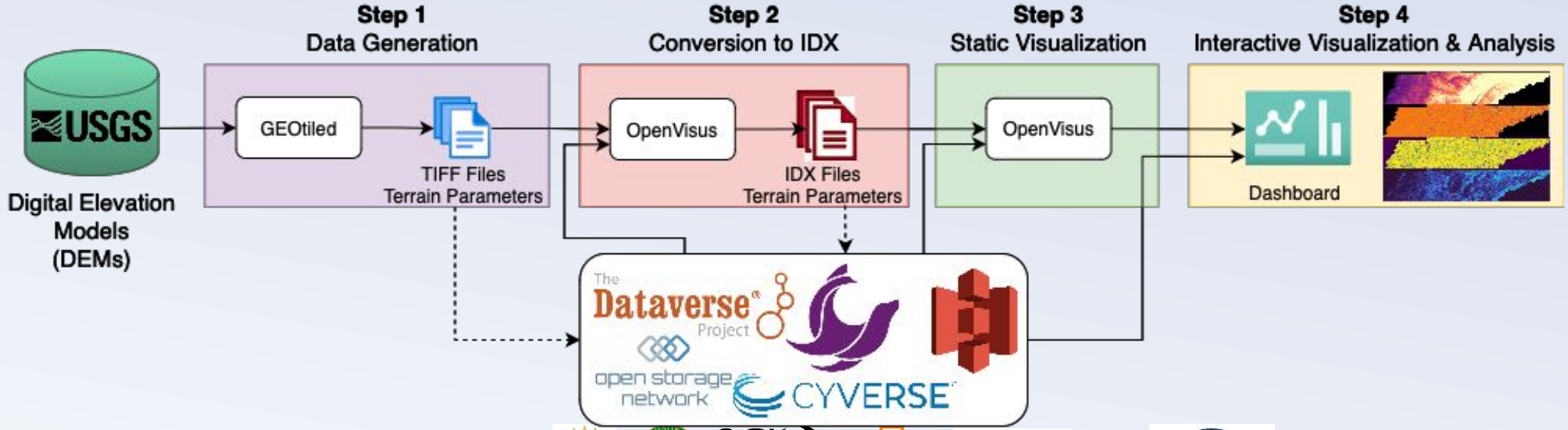
This tutorial showcases the capabilities of NSDF, guiding you through a **four-step modular workflow** that leverages OpenVisus services to analyze a geospatial dataset generated with GEOtiled.

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.

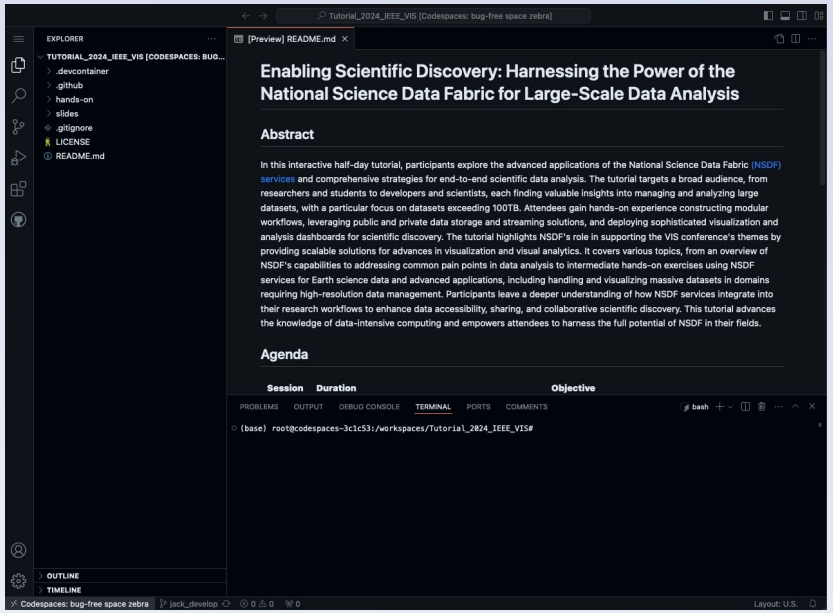
Step 2: Conversion to IDX Format
Convert files from TIFF to IDX (the format used by OpenVisus), preserving accuracy but reducing size. **Store** IDX files in public or private storage.

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

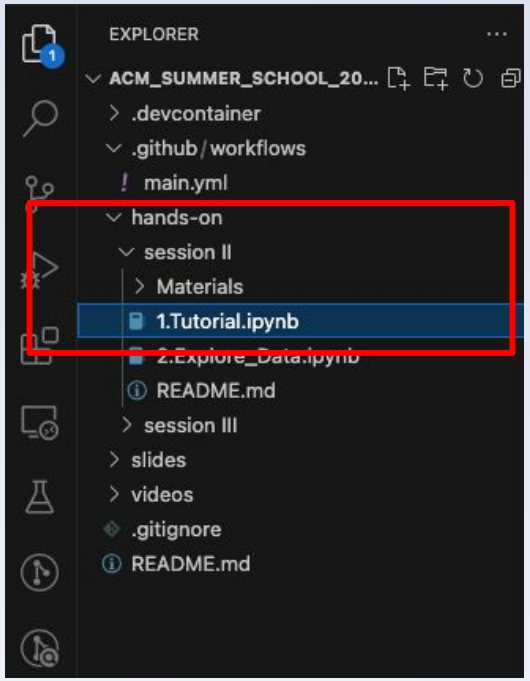
Step 4: Interactive Visualization & Analysis
Launch dashboard for interacting with large-scale data to access subregions of the original dataset for ad hoc analysis



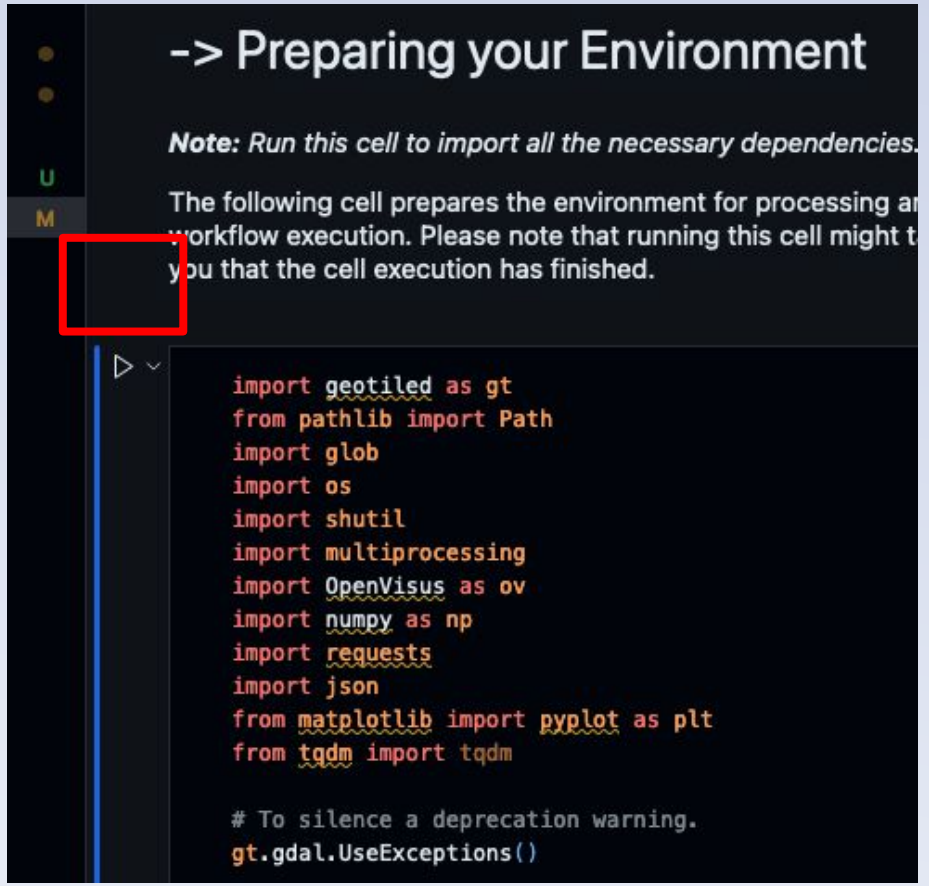
(1) Your codespace is fully loaded. Now you should see this screen



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

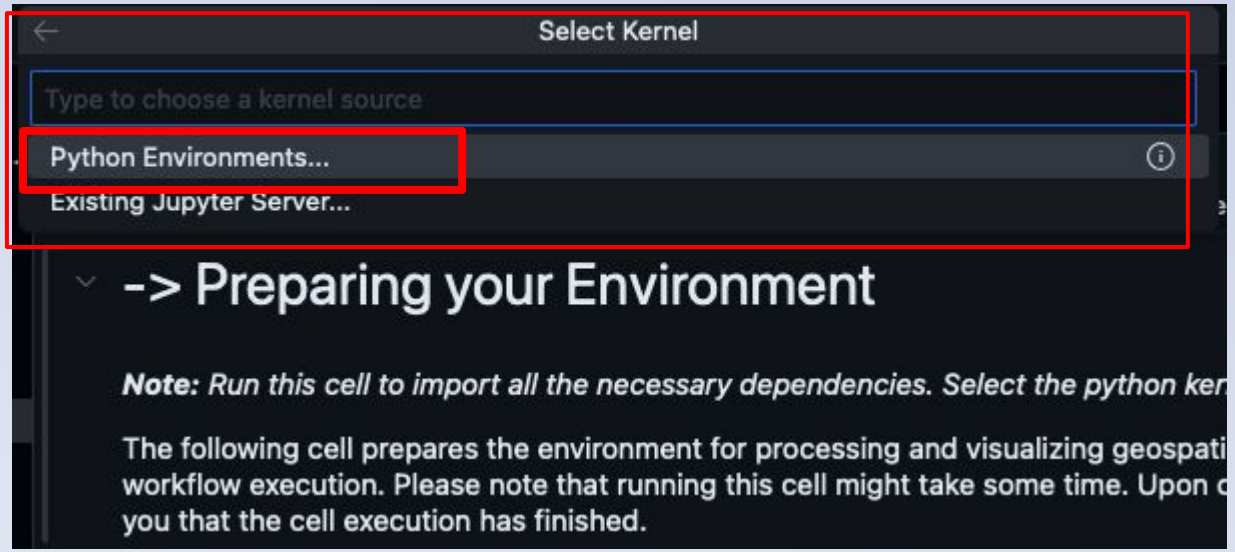


(3) Look for the *Preparing your Environment* section and press the play button

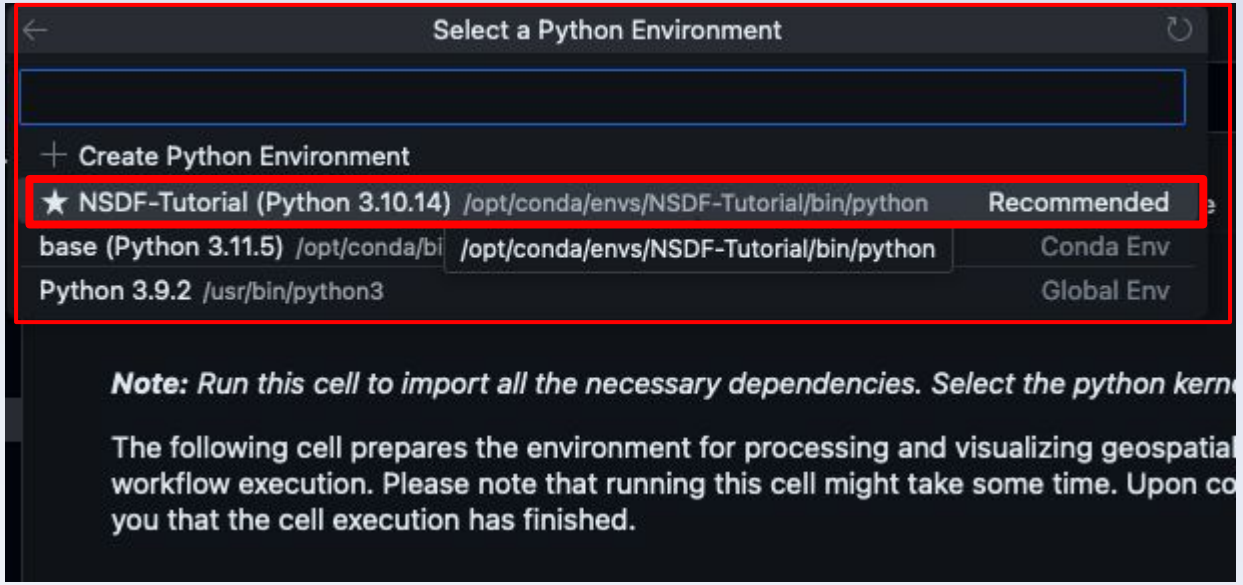


Instructions continue

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

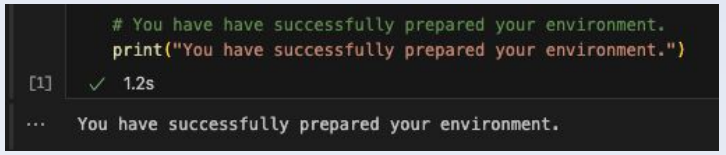


(5) Select the ★*NSDF-Tutorial* option



Tutorial

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



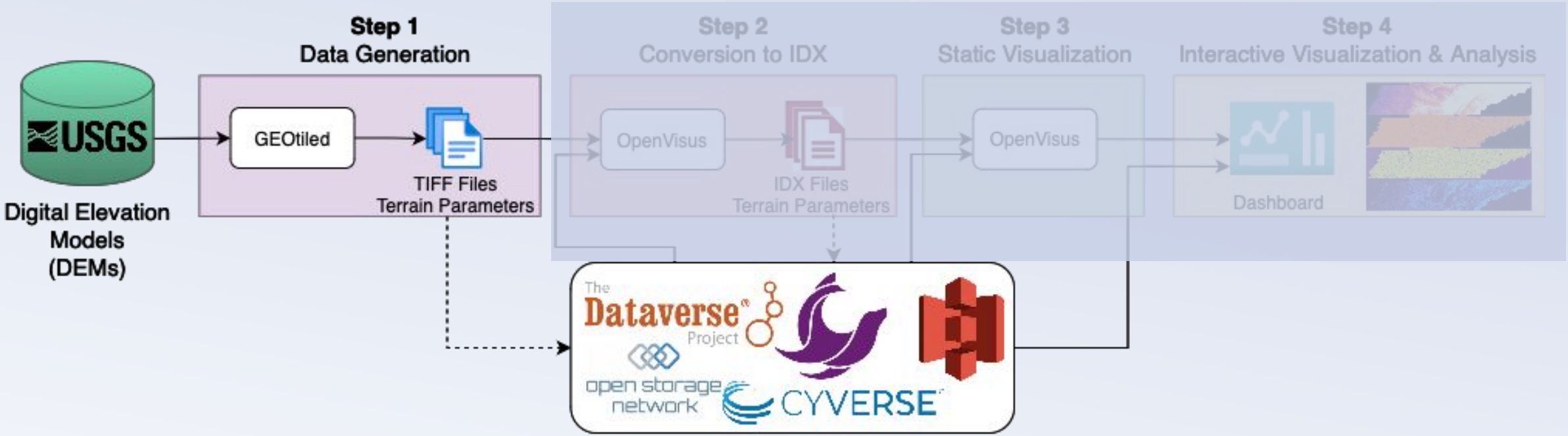
Step 1: Data Generation with GEOtiled

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

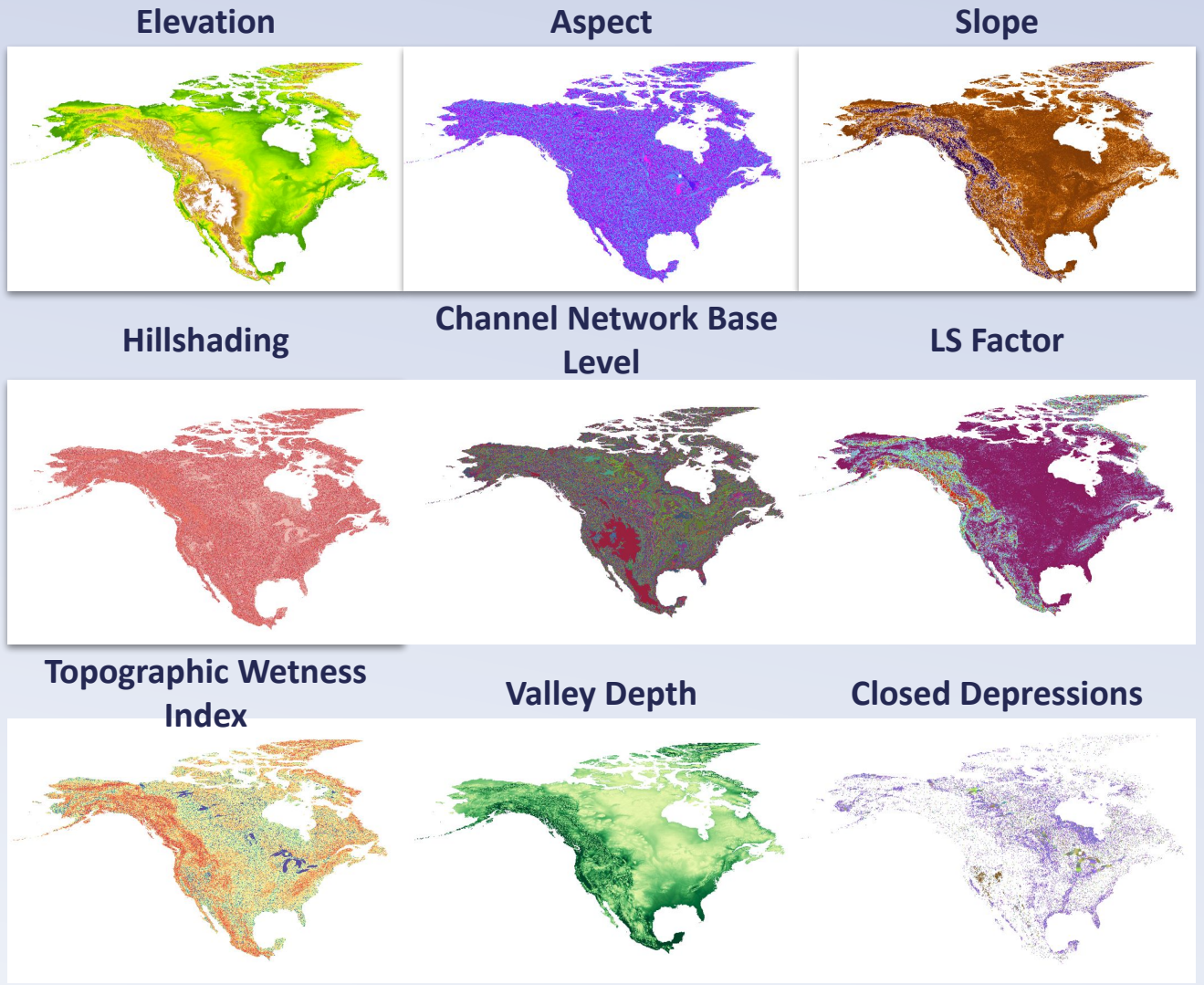


Step 1: What are Terrain Parameters?

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

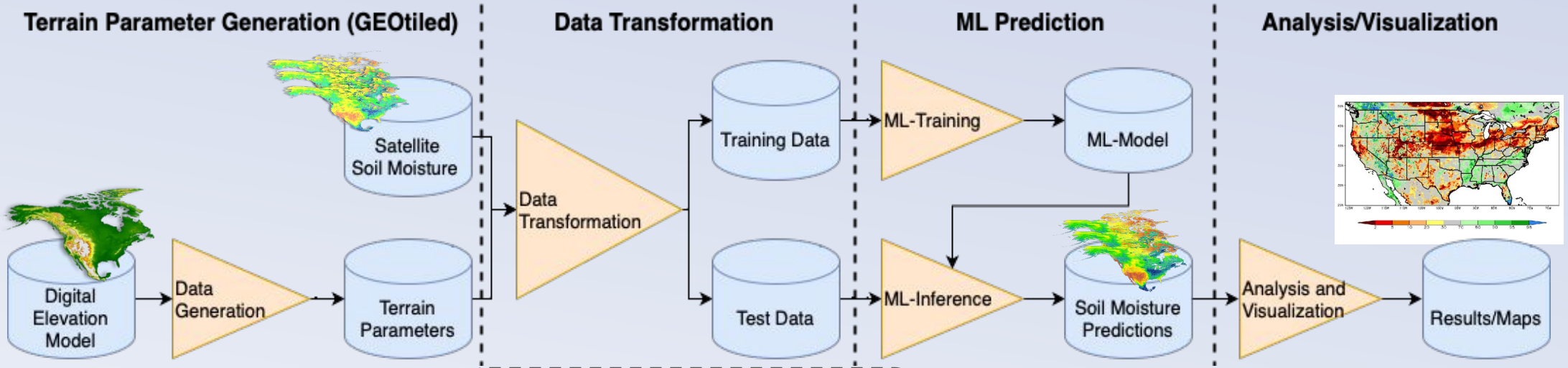
They play a **fundamental role** in applications such as **precision forestry and agriculture, and hydrology for landscape ecology.**

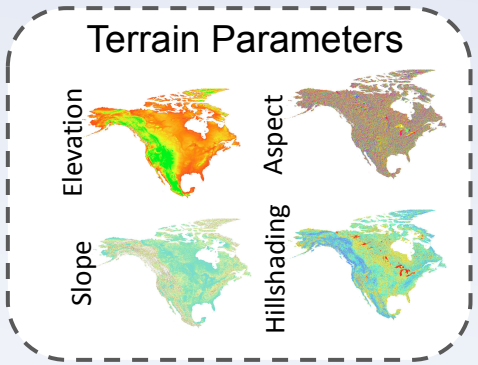
Generating **terrain parameters** at **high-resolution is computationally expensive,** hindering their accessibility by the scientific community

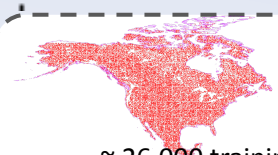


Step 1: SOMOSπ Components

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





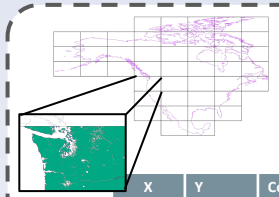


Training Data

~ 26,000 training points

Sample points every ~27 km

X	Y	Soil Moisture	Covariate 1	Covariate 2	Covariate 3	Covariate ... 4
X 1	Y 1	SM 1	A 1	B 1	C 1	N 1
X 2	Y 2	SM 2	A 2	B 2	C 2	N 2
X 3	Y 3	SM 3	A 3	B 3	C 3	N 3
X 4	Y 4	SM 4	A 4	B 4	C 4	N 4
X 5	Y 5	SM 5	A 5	B 5	C 5	N 5
X... N	Y... N	SM... N	A... N	B... N	C... N	N... N



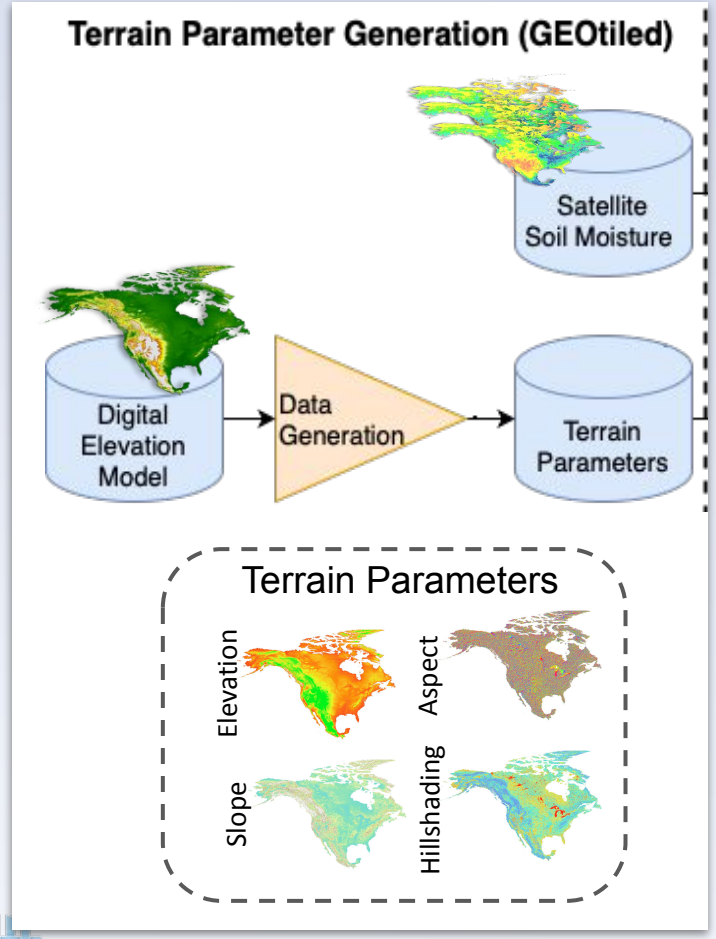
Test Data

Sample points every X meters

X	Y	Covariate 1	Covariate 2	Covariate 3	Covariate ... 4
X 1	Y 1	A 1	B 1	C 1	N 1
X 2	Y 2	A 2	B 2	C 2	N 2
X 3	Y 3	A 3	B 3	C 3	N 3
X 4	Y 4	A 4	B 4	C 4	N 4
X 5	Y 5	A 5	B 5	C 5	N 5
X... N	Y... N	A... N	B... N	C... N	N... N

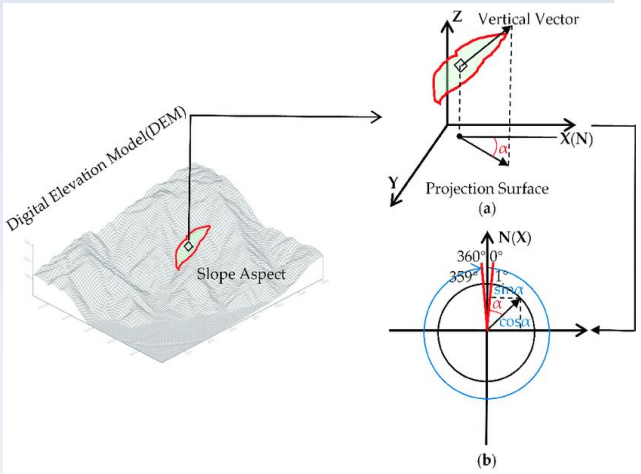
Step1: GEOtiled Terrain Generation

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

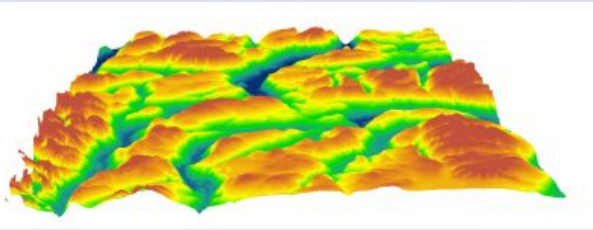


GEOtiled leverages data partitioning to accelerate the computation of terrain parameters from DEMs while preserving accuracy

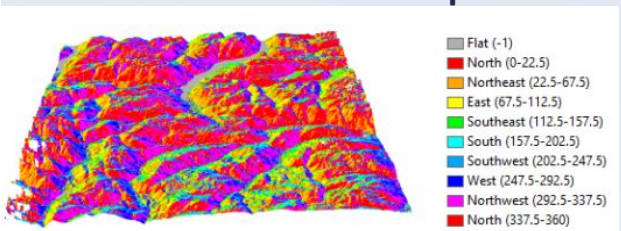
Computing High-resolution Terrain Parameters



Digital Elevation Model (DEM)

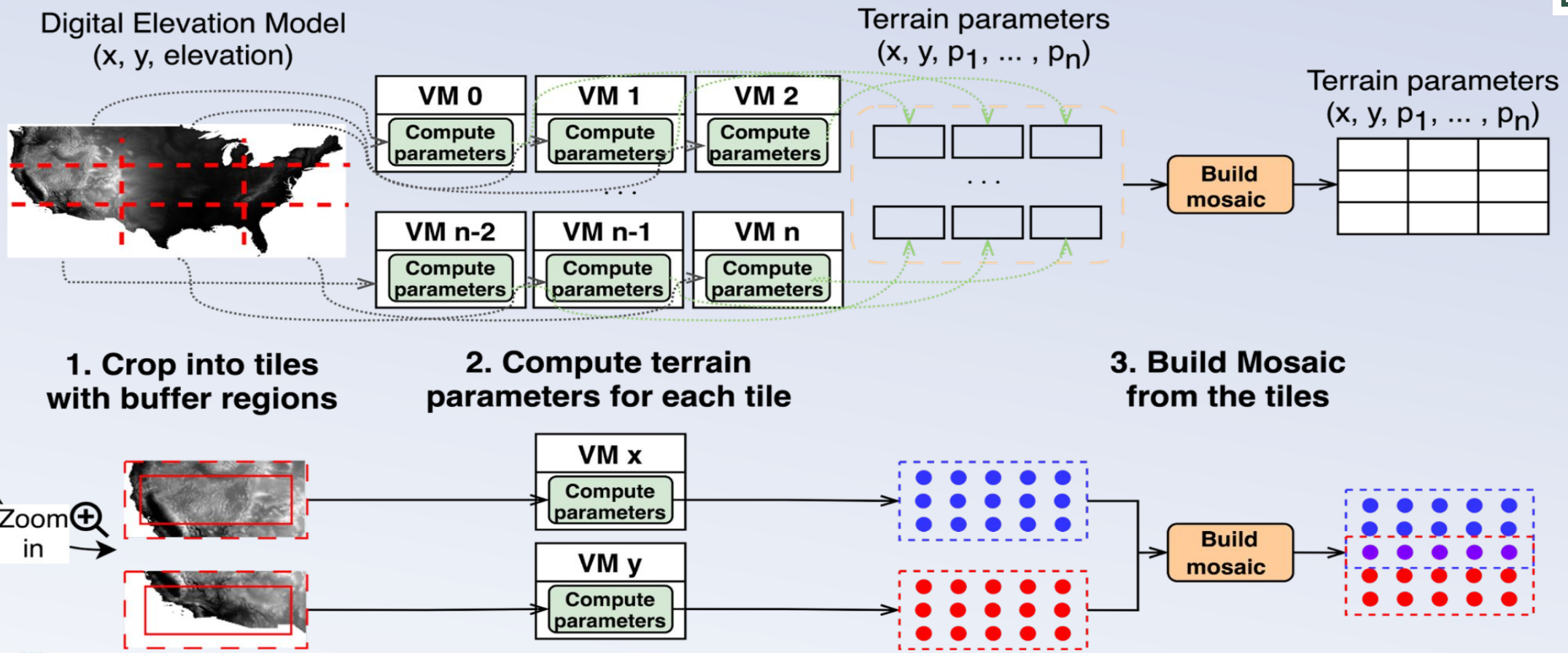


Terrain Parameter 1: Aspect



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

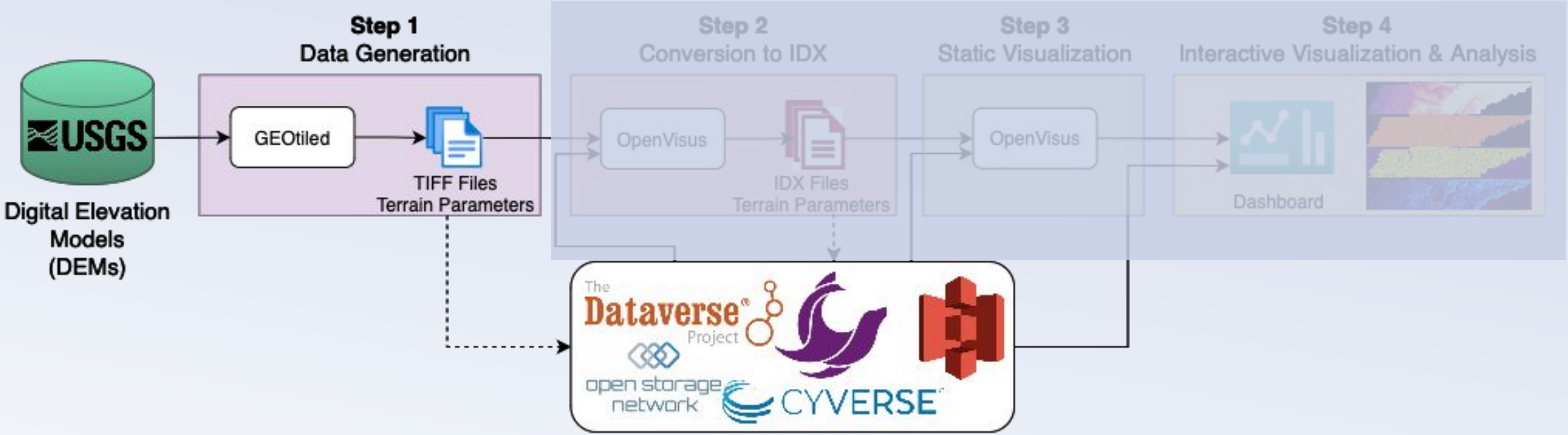
Tutorial

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

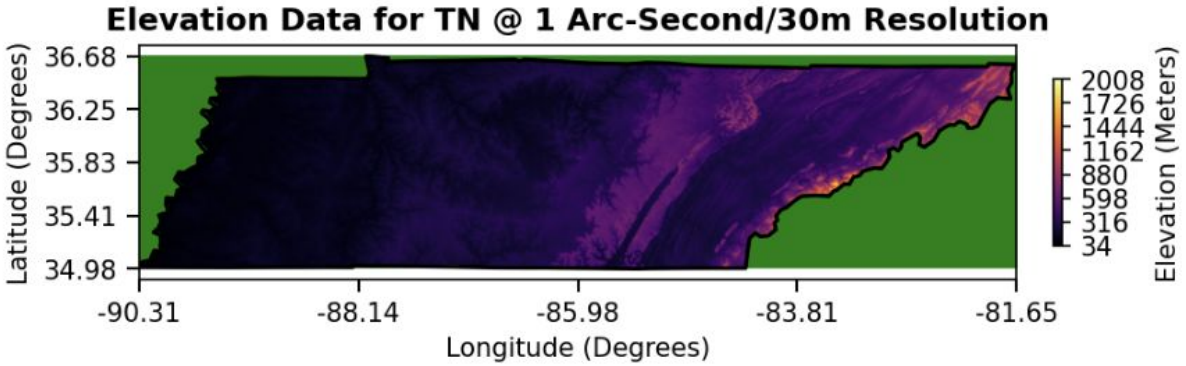
Option A (~10 mins)
Generating Data Using the SOMOSPIE Application Module

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

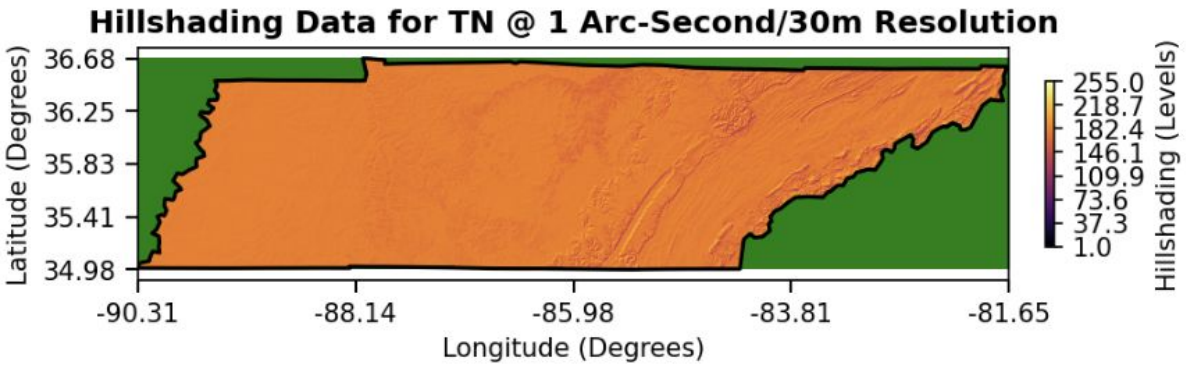


Step 1: Data Generation with GEOtiled

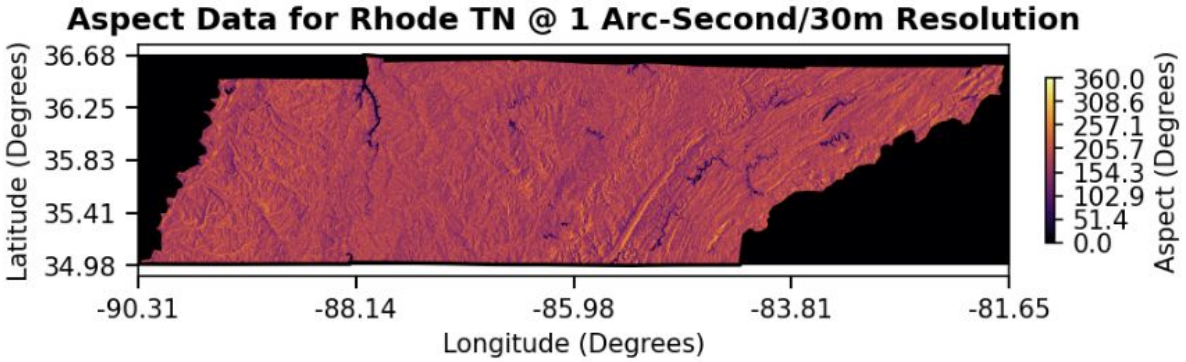
Tutorial



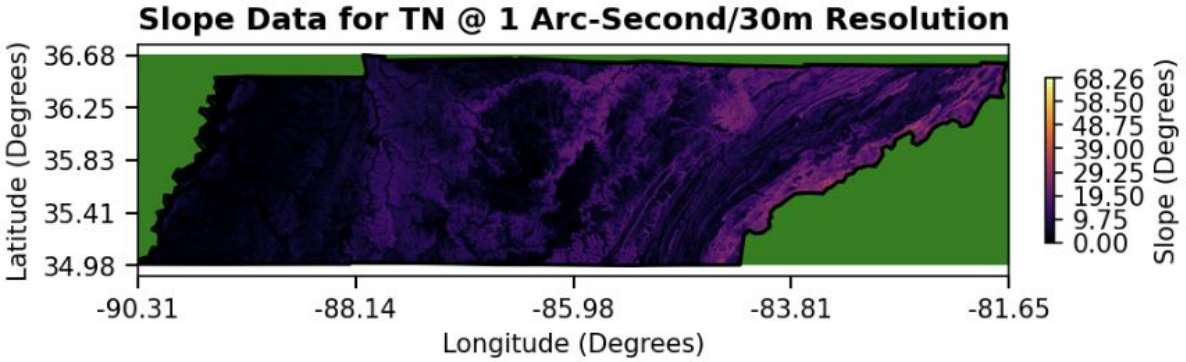
(a) Elevation - Terrain Parameter



(b) Hillshading - Terrain Parameter



(c) Aspect - Terrain Parameter



(d) Slope - Terrain Parameter

Step 2: Conversion to IDX

Tutorial

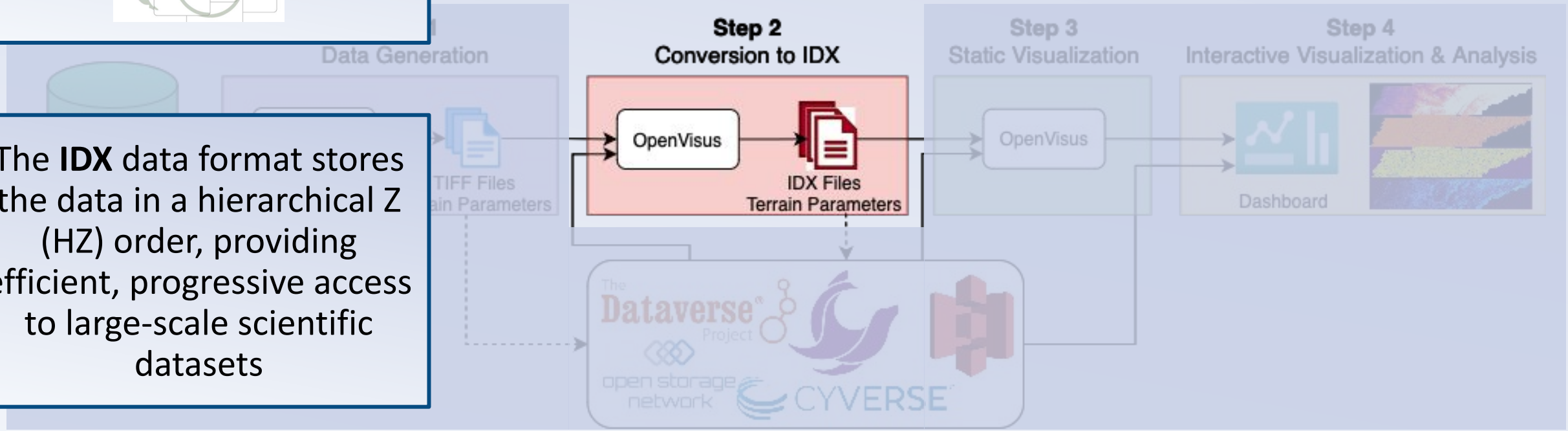
OpenVisus is a progressive cache-oblivious framework for large-scale data visualization

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

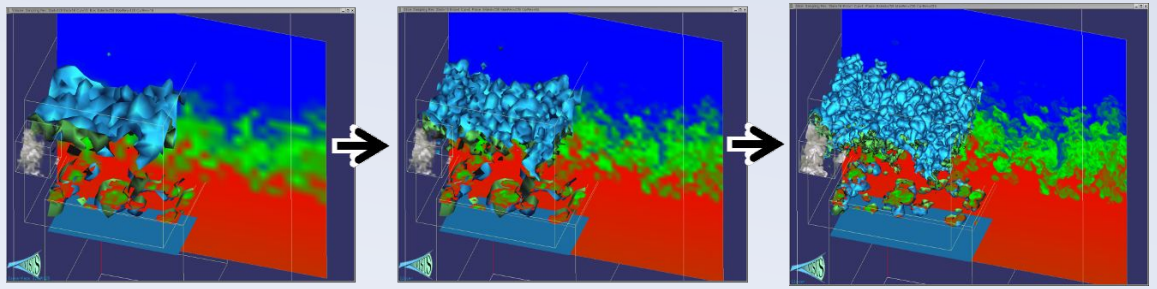


Step 2: ViSUS IDX Data Format

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

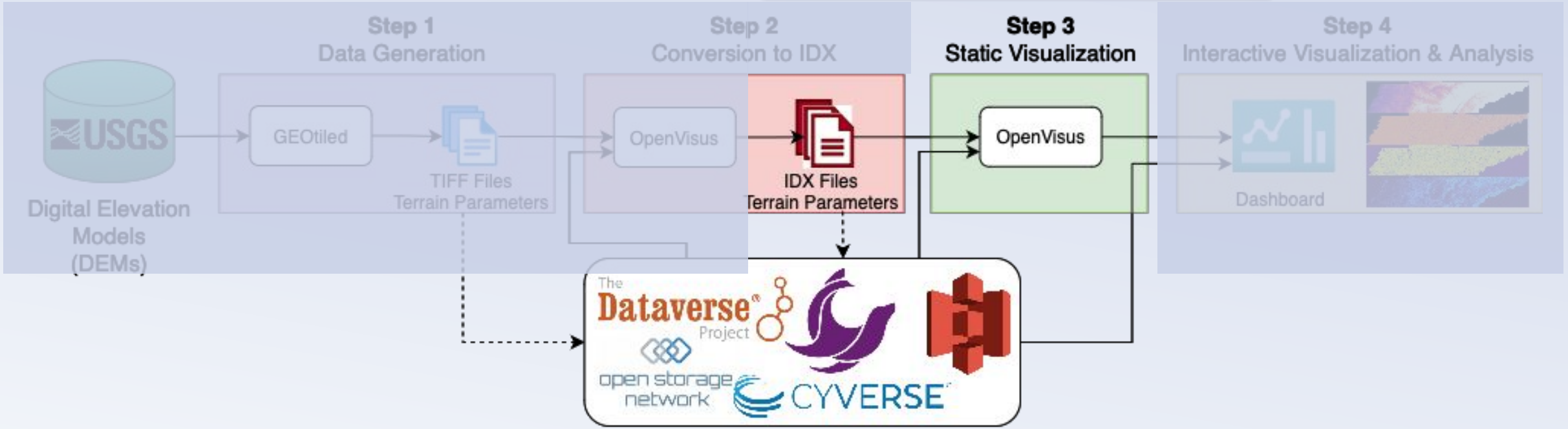
Tutorial

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

Option A
From local storage

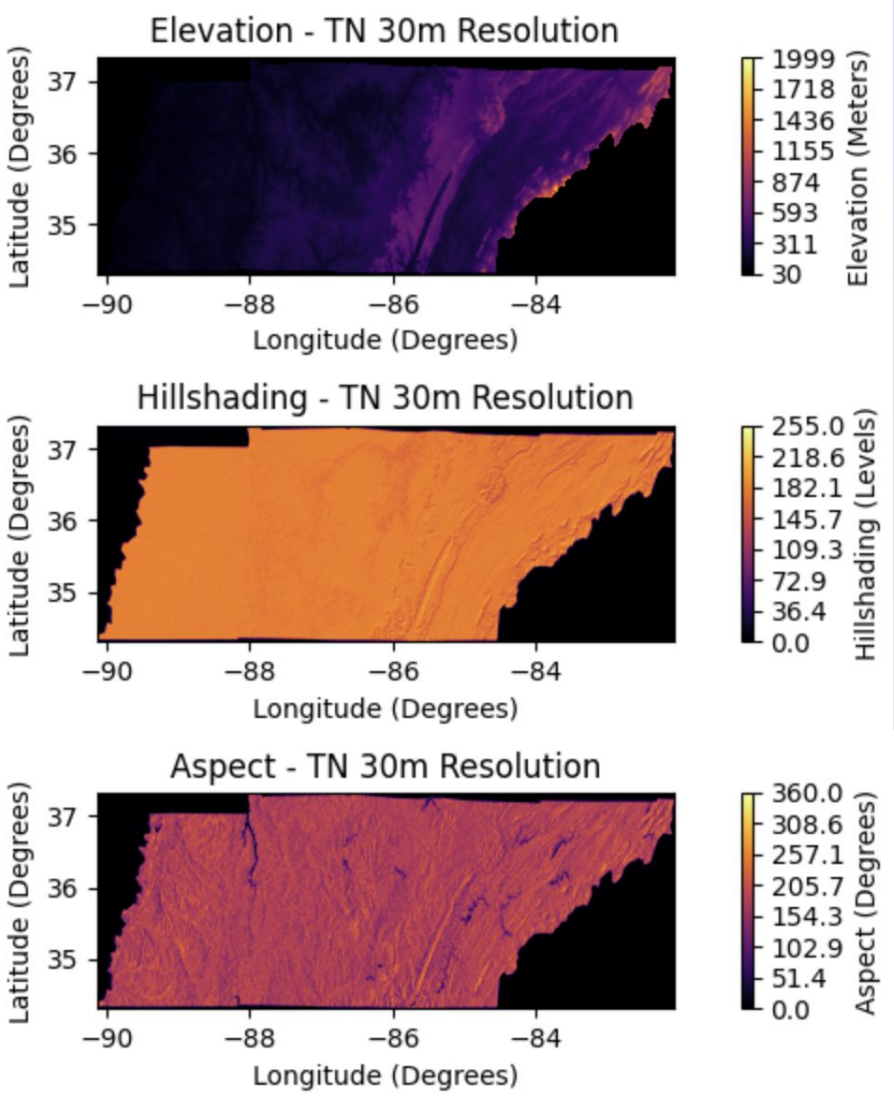
Option B
From Seal Storage

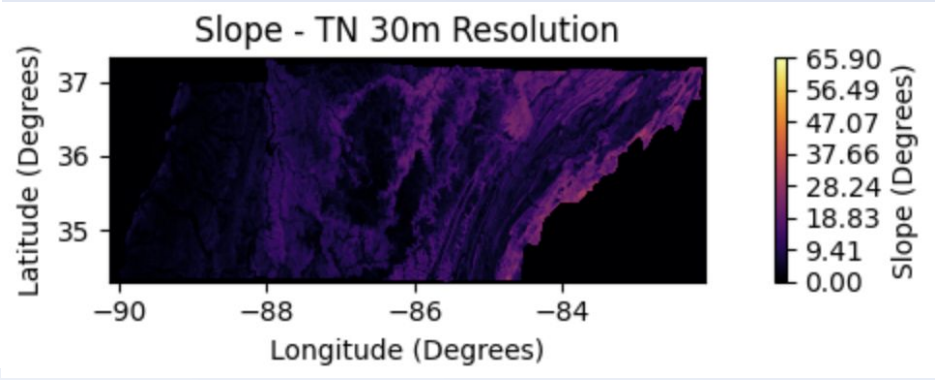
Step 3: Static Visualization
Statically visualize the terrain parameters in OpenVisus. Validate the accuracy of IDX-based images with TIFF-based images.



Step 3: Static Visualization

Tutorial



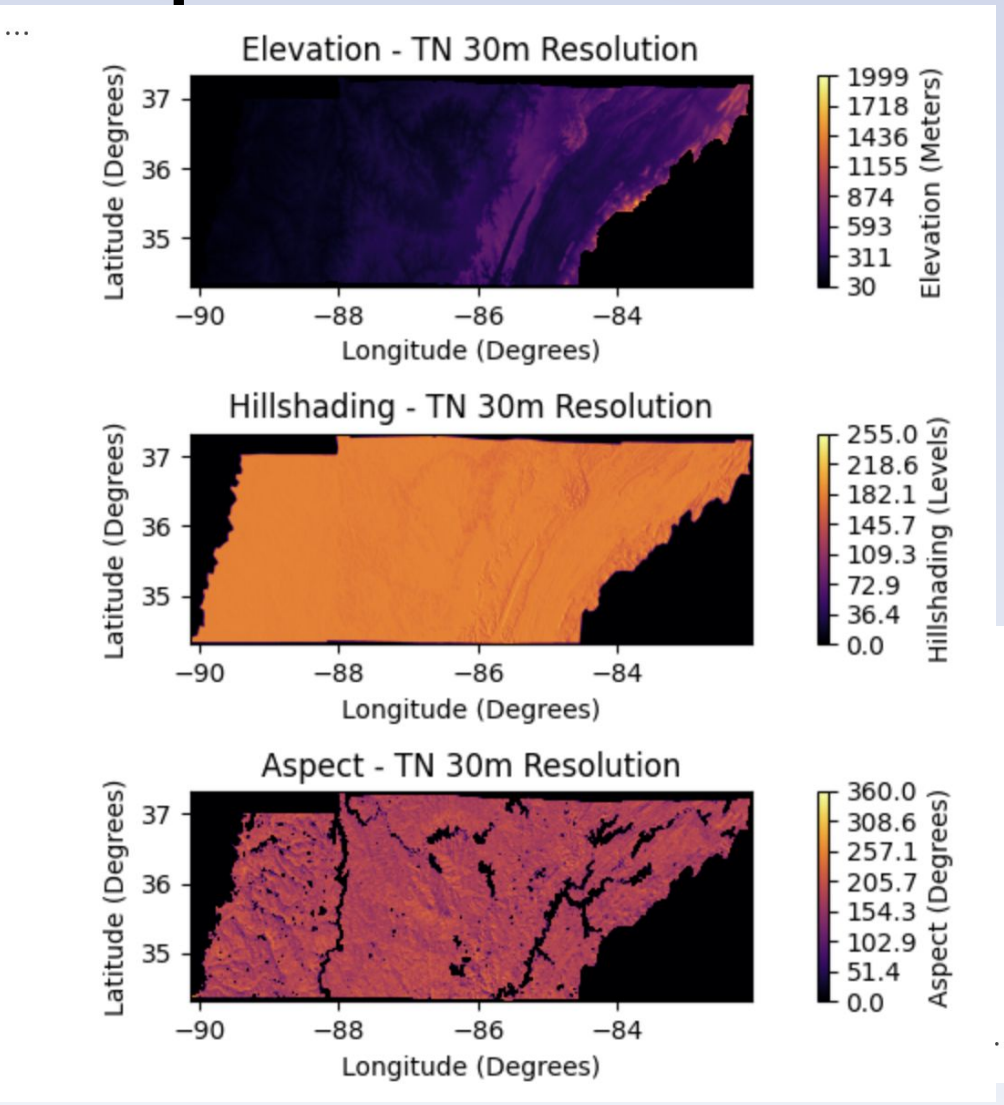


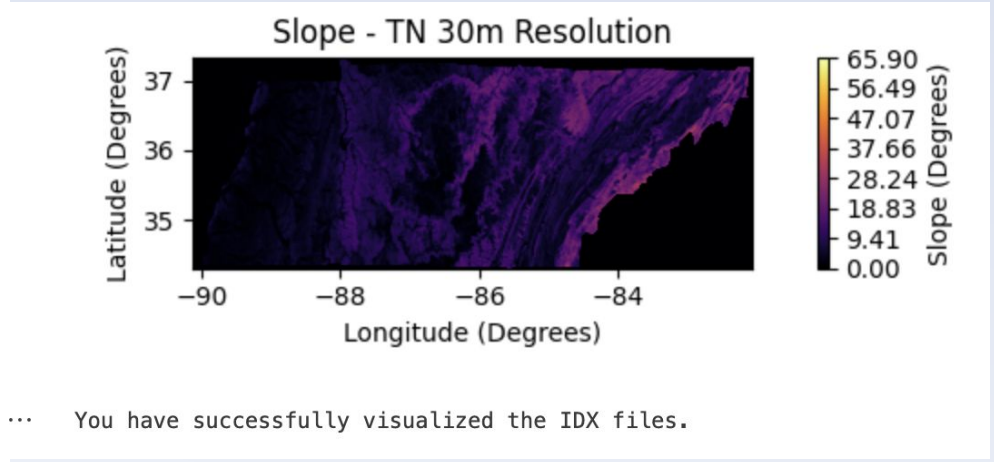
Option A
From local storage

Step 3: Static Visualization

Tutorial

Option B
From Seal Storage






```
...  You have successfully visualized the IDX files.
```

Step 4: Interactive Visualization & Analysis

Tutorial

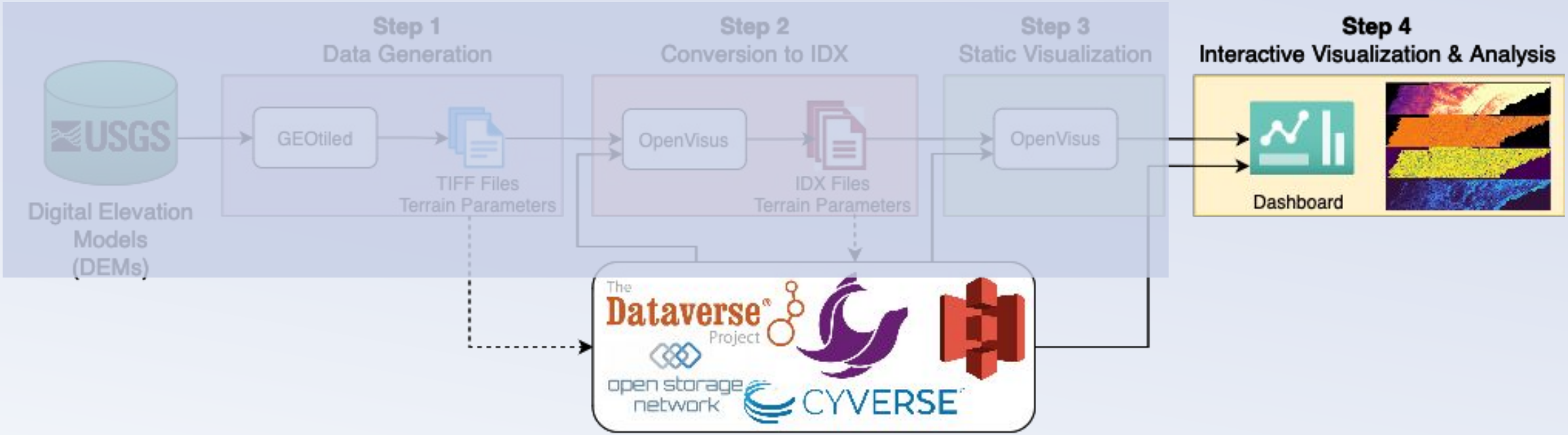
Remotely **access** large datasets, **zoom** into specific areas, **select** and **crop** subregions of interest, **save** data locally in a Python-compatible format, and **analyze** the data for scientific discovery.

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

Option A
From local storage

Option B
From Seal Storage

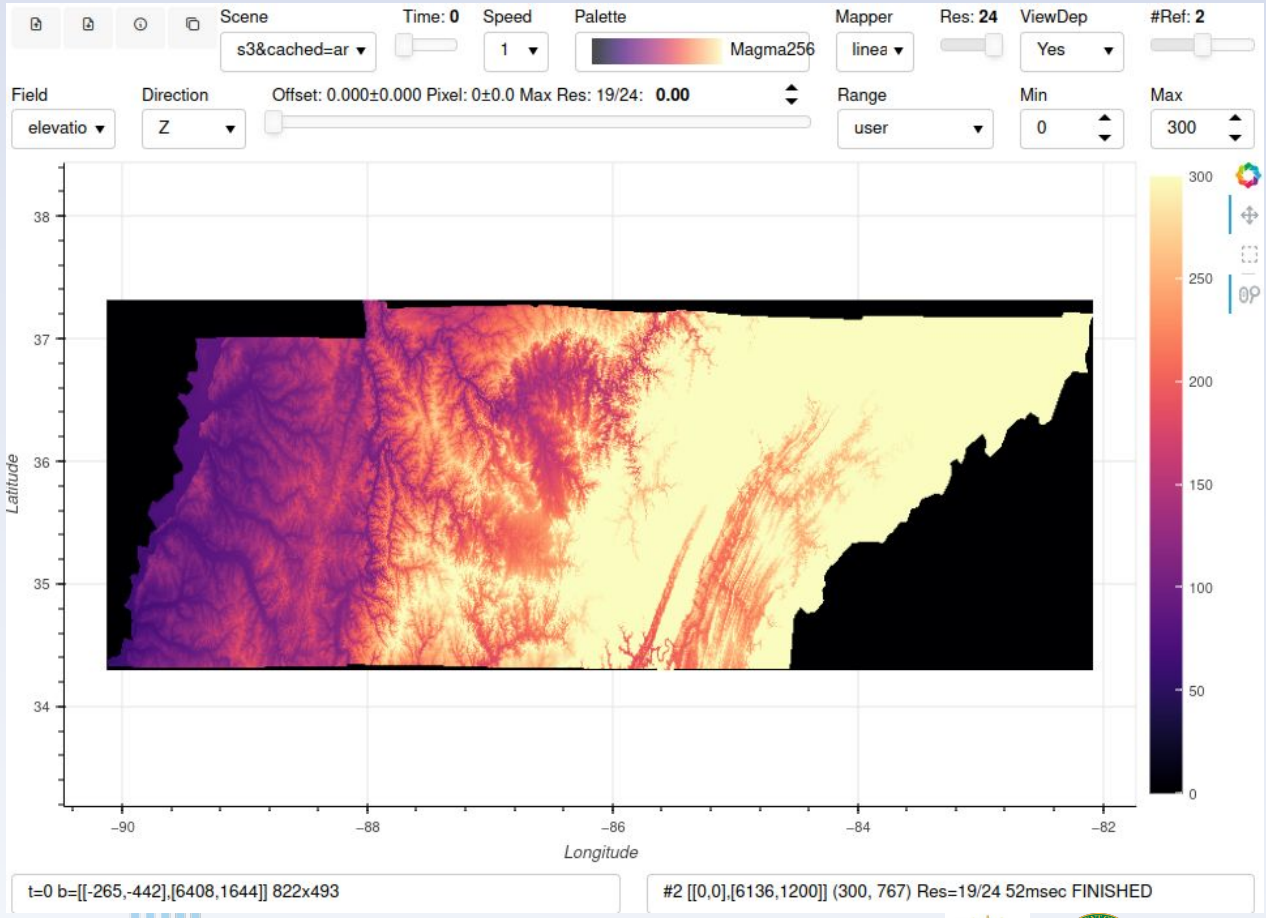
Step 4: Interactive Visualization & Analysis
Launch dashboard for interacting with large-scale data to access subregions of the original dataset for ad hoc analysis.



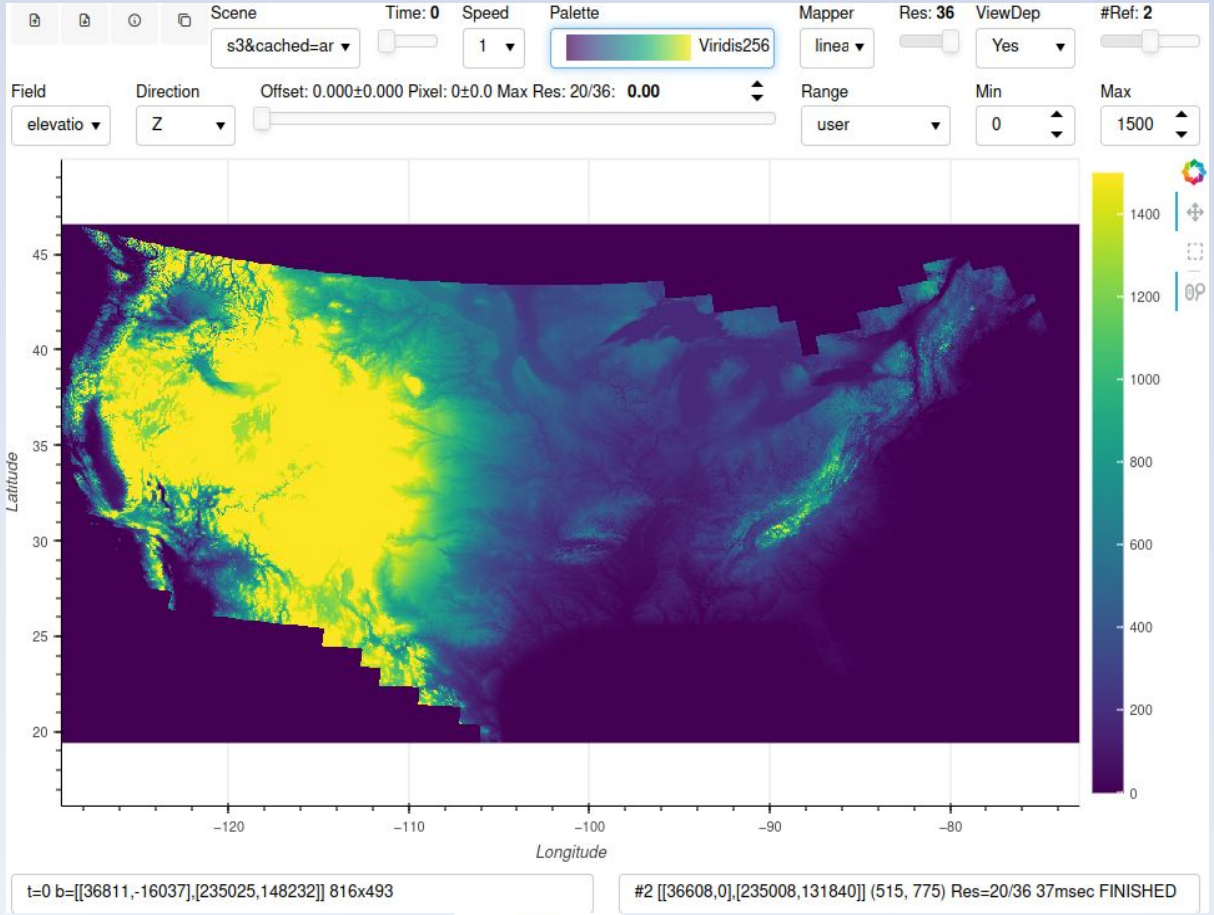
Step 4: Geographical Regions

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

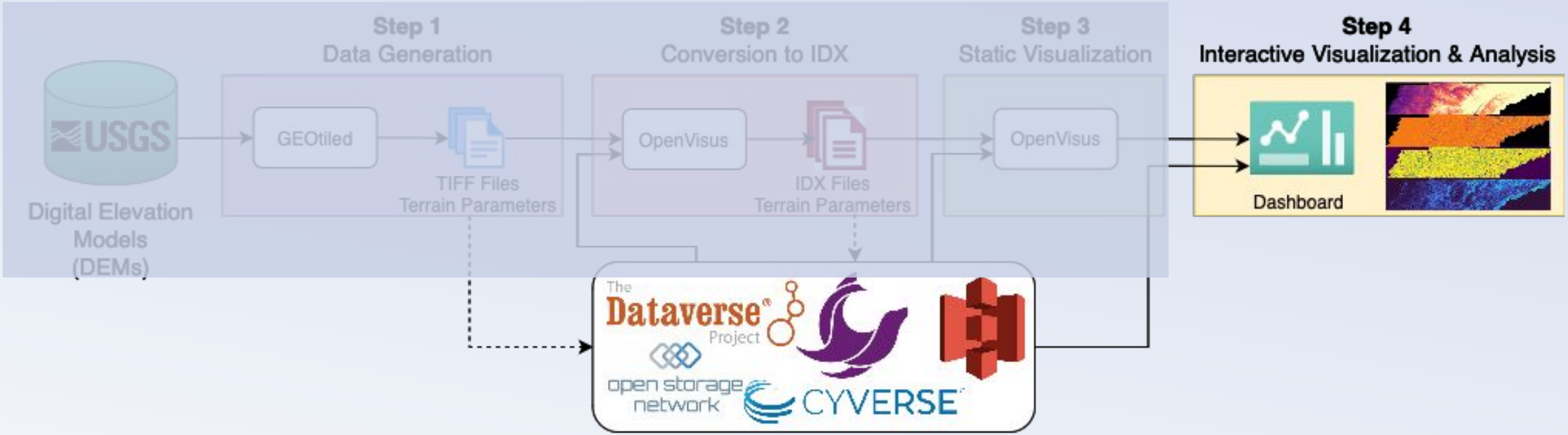
Remotely **access** large datasets, **zoom** into specific areas, **select** and **crop** subregions of interest, **save** data locally in a Python-compatible format, and **analyze** the data for scientific discovery.

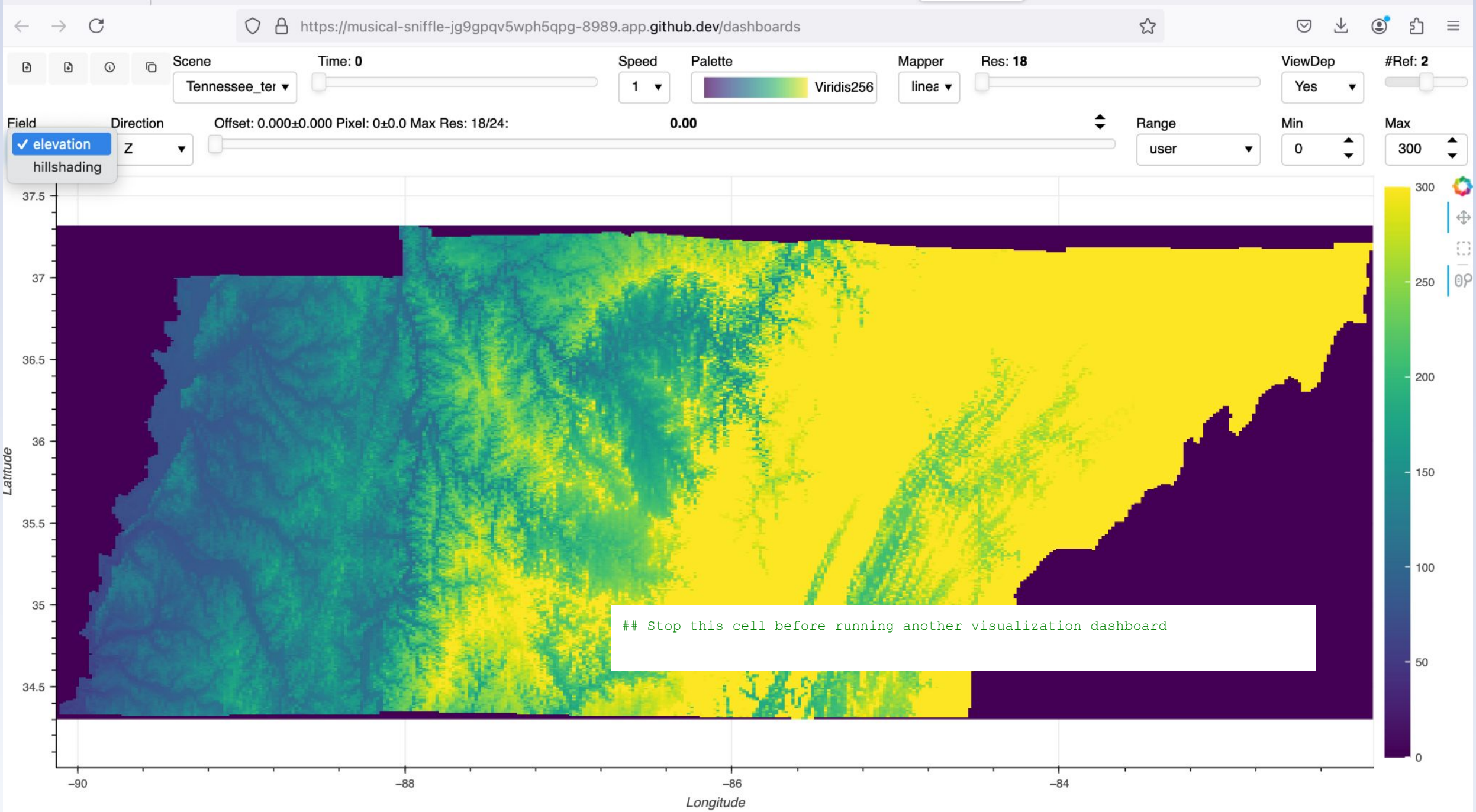
Step 4 provides **two options** to obtain data and collect the IDX files

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
Scene
Tennessee_ter
Time: 0
Speed
1
Palette
Viridis256
Mapper
linea
Res: 18
ViewDep
Yes
#Ref: 2
Field
elevation
hillshading
Direction
Z
Offset: 0.000±0.000 Pixel: 0±0.0 Max Res: 18/24:
0.00
Range
user
Min
0
Max
300
37.5
37
36.5
36
35.5
35
34.5
Latitude
−90
−88
−86
−84
Longitude
300
250
200
150
100
50
0
Stop this cell before running another visualization dashboard
t=0 b=[[0,-123],[6454,1385]] 1272x571
#2 [[0,0],[6448,1260]] (315, 403) Res=18/24 38msec FINISHED

Step 4: Interactive Visualization & Analysis

Tutorial

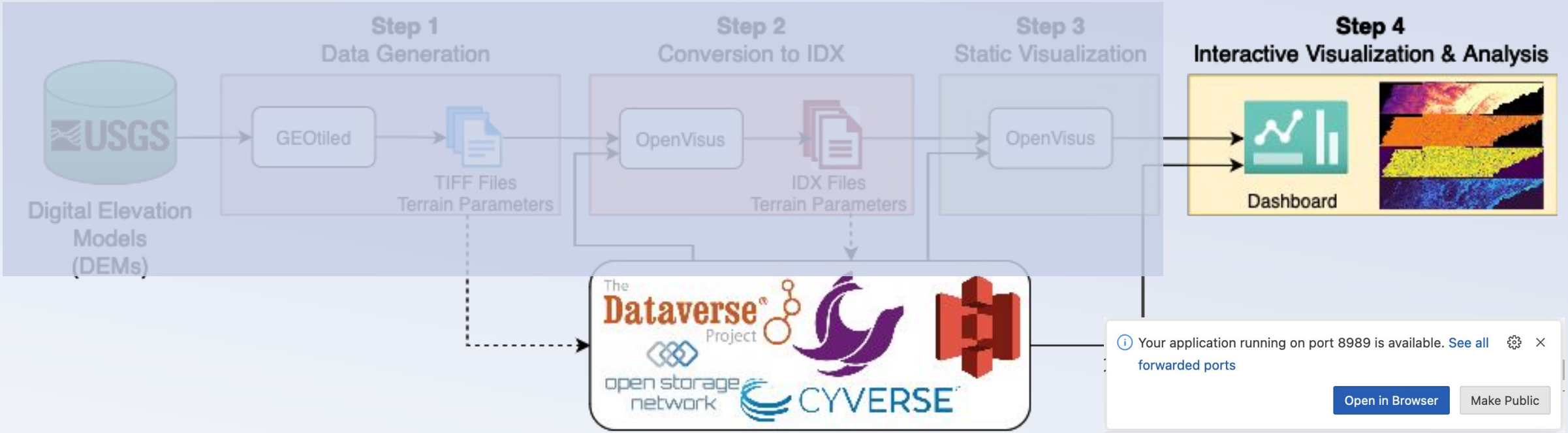
Remotely **access** large datasets, **zoom** into specific areas, **select** and **crop** subregions of interest, **save** data locally in a Python-compatible format, and **analyze** the data for scientific discovery.

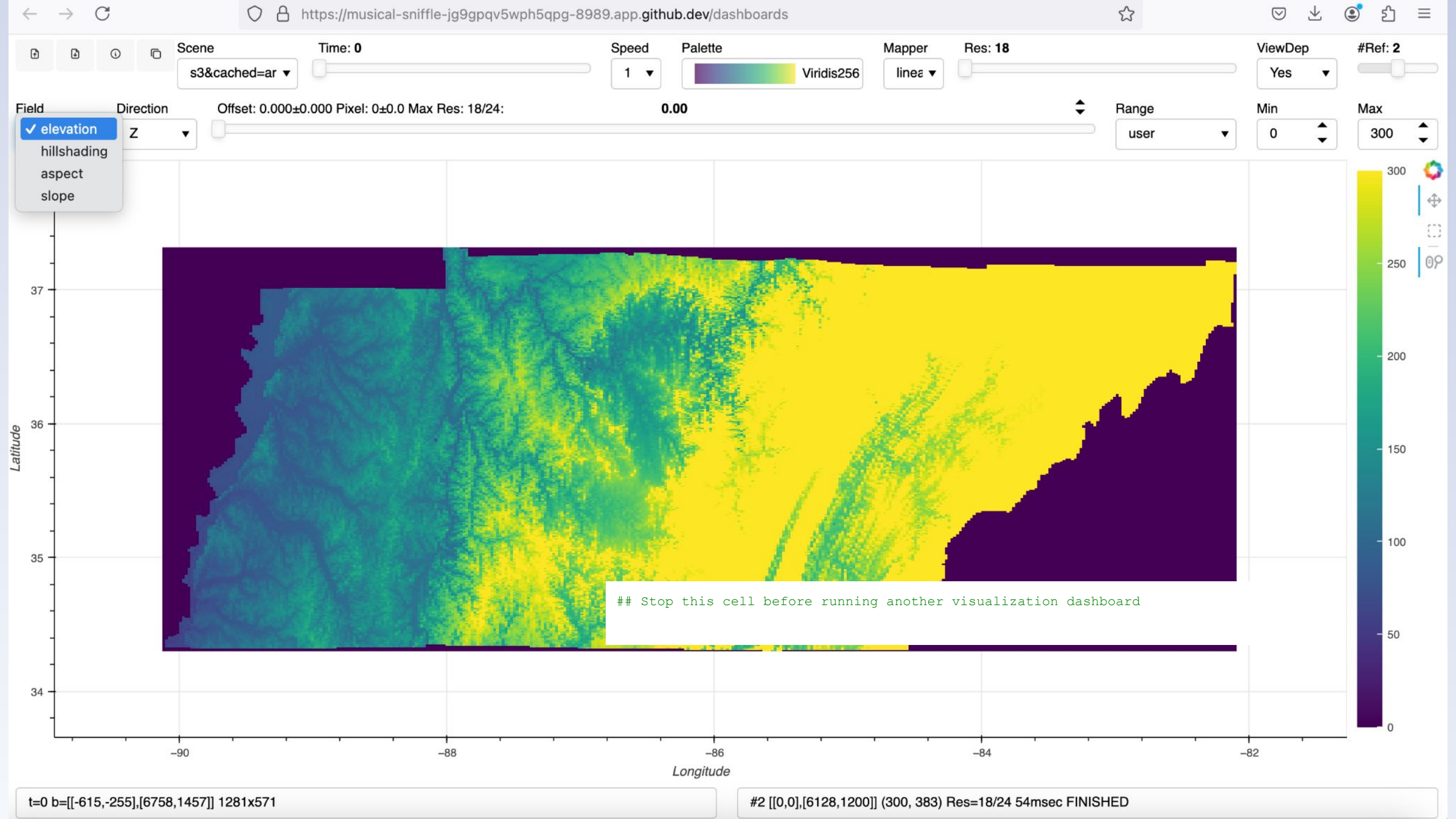
Step 4 provides **two options** to obtain data and collect the IDX files

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.

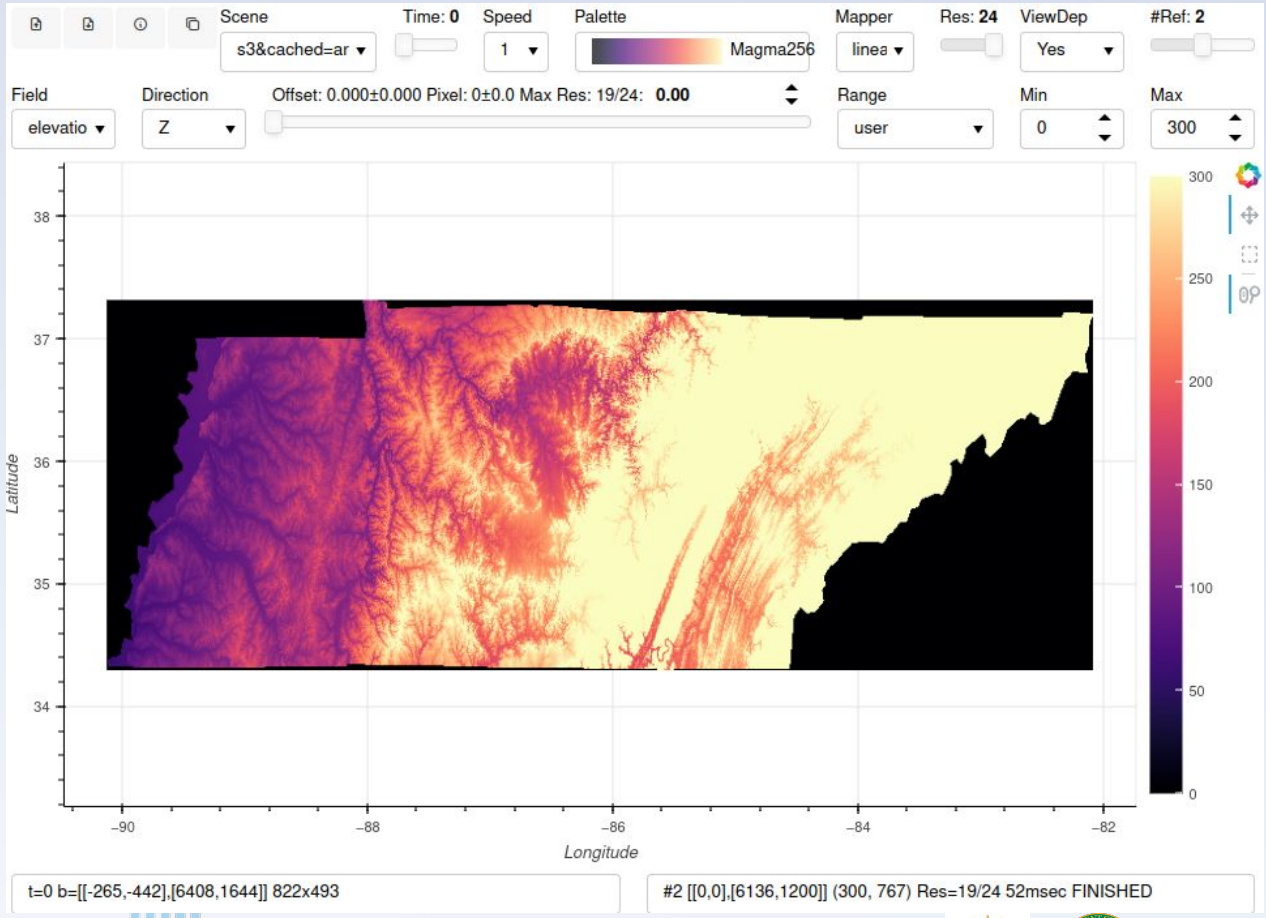



https://musical-sniffle-jg9gpqv5wph5qpg-8989.app.github.dev/dashboards
Scene
s3&cached=ar
Time: 0
Speed
1
Palette
Viridis256
Mapper
linea
Res: 18
ViewDep
Yes
#Ref: 2
Field
elevation
hillshading
aspect
slope
Direction
Z
Offset: 0.000±0.000 Pixel: 0±0.0 Max Res: 18/24:
0.00
Range
user
Min
0
Max
300
300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

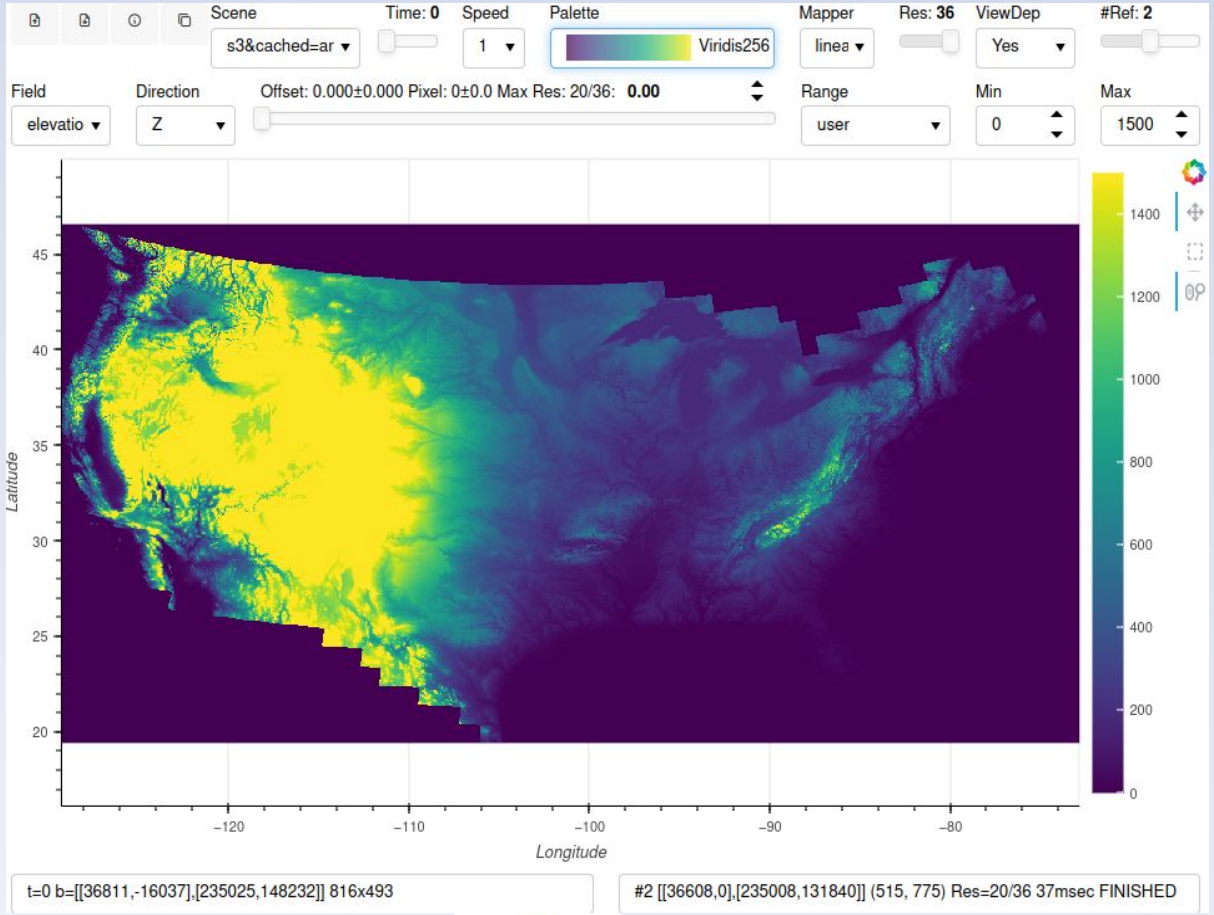
Step 4: Geographical Regions

Visualize and analyze two geographical regions at 30 m resolution

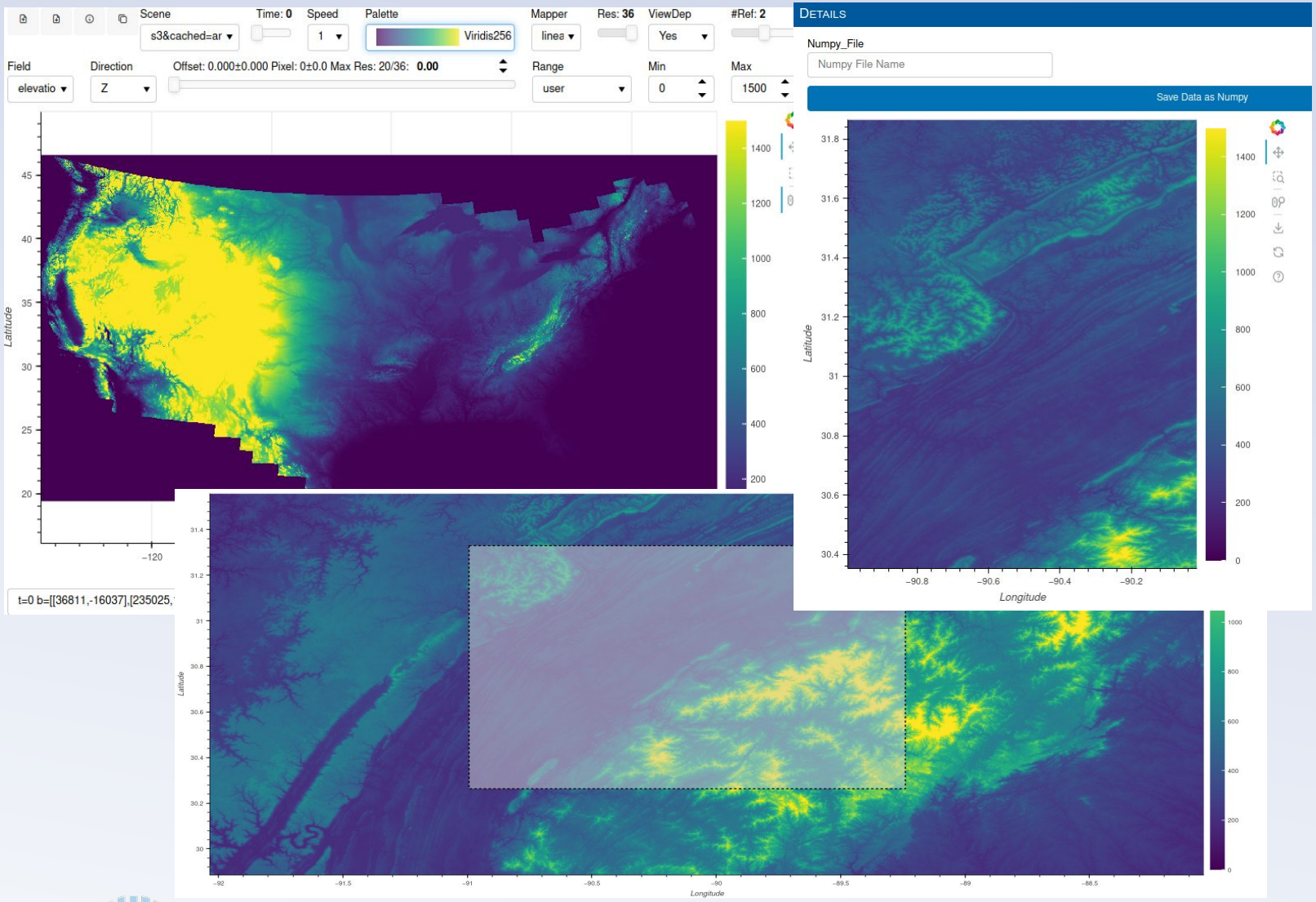
State of Tennessee - 200 MB



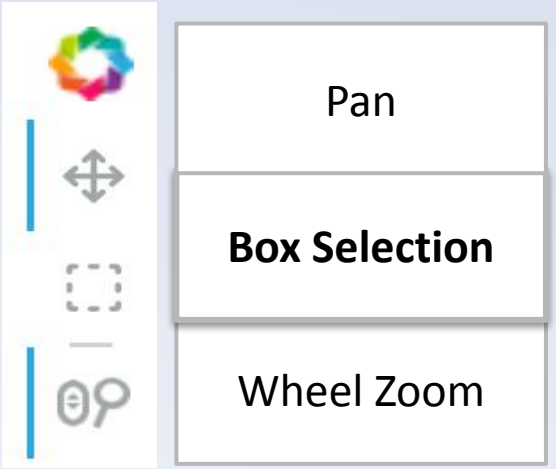
Contiguous United States (CONUS) - 200 GB



Step 4: 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



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 the range of displayed data, from the minimum to the maximum value, providing a more accurate visual representation of the data

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.

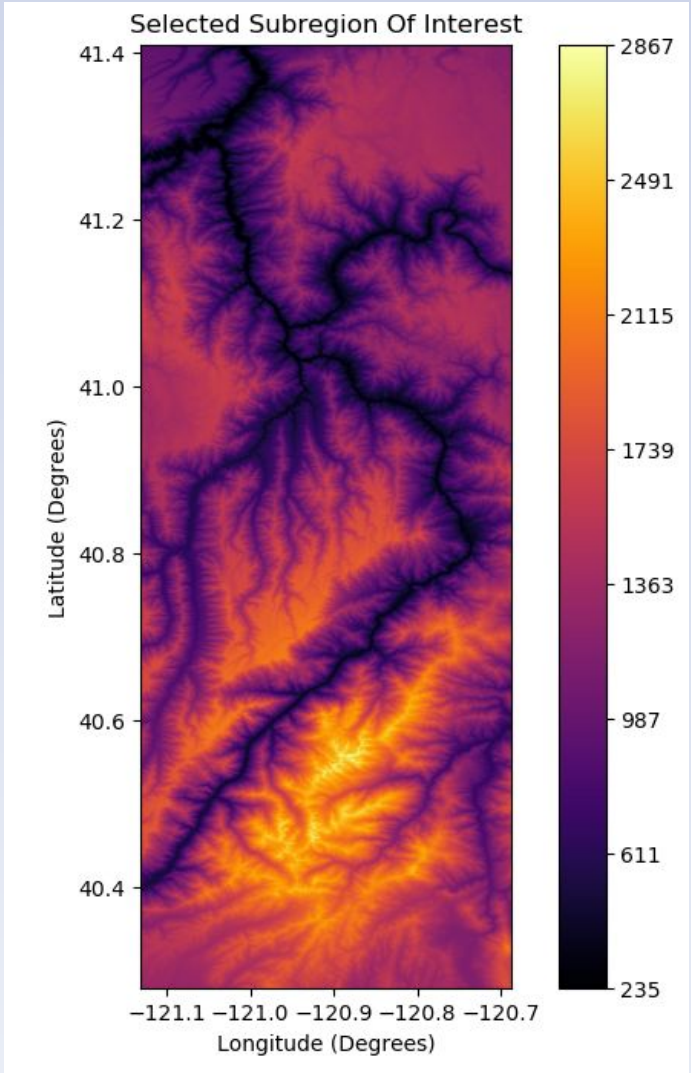
```
[4]: data=np.load(data_file)
     data
     actual_data=data['data']
     metadata=data['lon_lat']
     print("You have successfully loaded your data and metadata.")
```

```
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

(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

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?

Tutorial Links

NSDF

GEOtiled

SOMOSPIE

OpenVisus