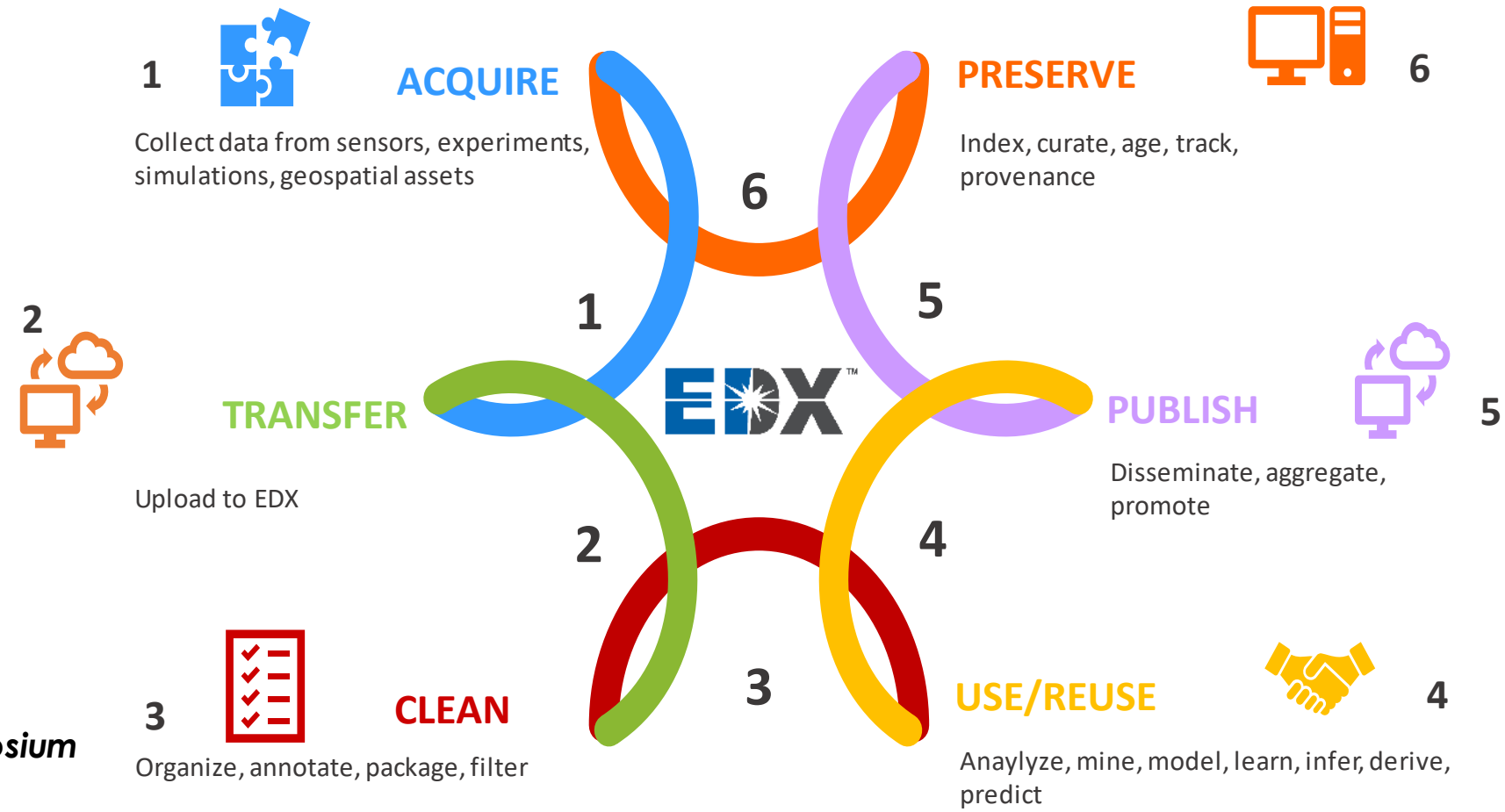
Fostering Data Curation Throughout the Entire Carbon Storage Data Life Cycle via the Energy Data eXchange and GeoCube

Paige Morkner
NETL Support Contractor
Research Innovation Center



CCUS Database Virtual Symposium
July 27, 2021

Disclaimer

This project was funded by the United States Department of Energy, National Energy Technology Laboratory, in part, through a site support contract. *Neither the United States Government nor any agency thereof, nor any of their employees, nor the support contractor, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.*

Author Information

Paige Morkner[1,2], Jennifer Bauer[1], Kelly Rose[1], Chad Rowan[1,3], Aaron Barkhurst[1,3]

[1] National Energy Technology Laboratory, 1450 Queen Avenue SW, Albany, OR, 97321 USA

[2] NETL Support Contractor, 1450 Queen Avenue SW, Albany, OR, 97321, USA

[3] NETL Support Contractor, 3610 Collins Ferry Rd., Morgantown, WV, 26505, USA

Supporting the Life Cycle of Data Curation is Key to Scientific Discovery

NATIONAL ENERGY TECHNOLOGY LABORATORY

Access

Researchers need access to pertinent, relevant data and the ability use/reuse data beyond its original purpose.

Sharing

Researchers need to share data resources and collaborate across multi-organizational private teams and publish the results to the public.

Most focus on the visible iceberg!

What is hidden below the surface?

Lost Data

Data is lost at an alarming rate. EDX provides a mechanism to capture data and support it throughout its entire life cycle.

Inadequate Resources

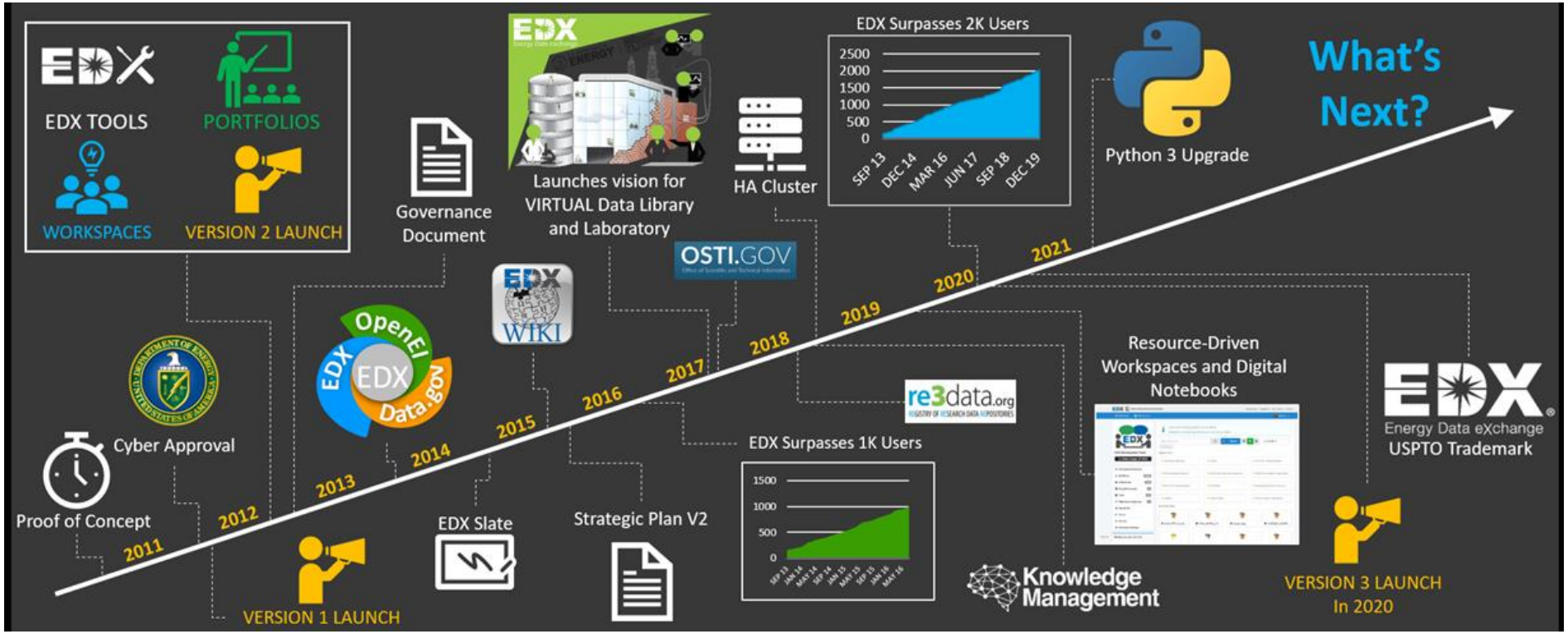
Data is often limited by hardware, software, technology, and other resources. EDX equips its users with tools and functionality to combat these issues.

Scientific discovery is restricted by lost data

What is Needed to Support the Entire Life Cycle of Data?

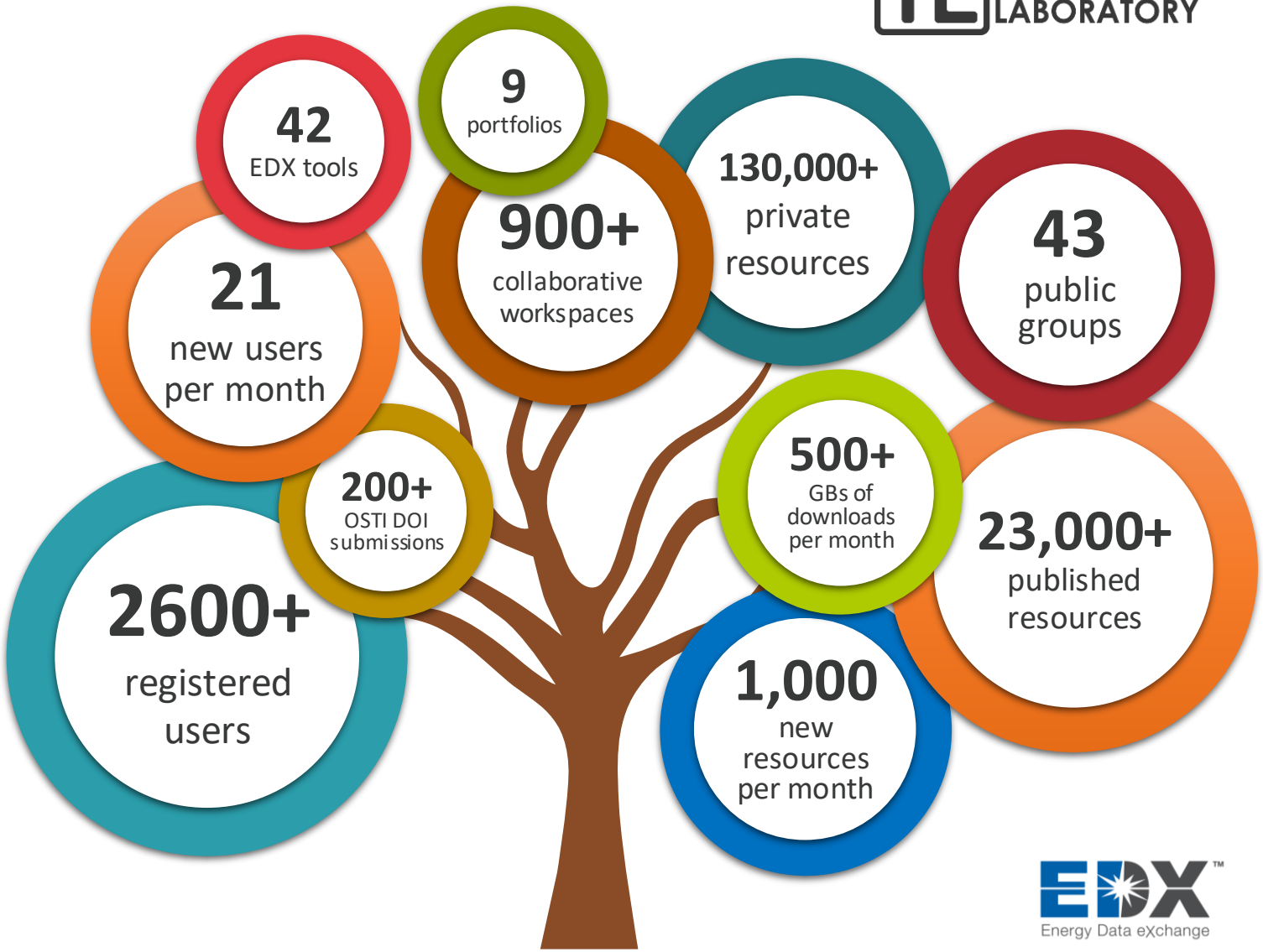
- **Private and multi-organizational collaboration**
- **Public dissemination** of data, publications, presentations, and tools
- **Secure and accessible** platform to internal and external users
- DOE and Federal **regulatory compliance**
- **Scalable architecture**
- **Agile development process** to meet the needs of users

In 2011, the Energy Data Exchange (EDX) Team identified the following key features needed for a system to support the entire life cycle of data



What is EDX?

- EDX is a data repository and digital laboratory that supports the entire life cycle of data, presentations, publications and tools
- EDX evolves to meet the needs of the DOE FECM community
- EDX is a solution for efficient and easy access to authoritative and relevant data resources
- EDX conforms to DataCite metadata citation standards
- EDX supports discovery and aggregation of authoritative, open-source resources
- Geocube supports access, visualization, and interaction with geospatial data



Tiered Access Using Role-Based Security

Utilizing NETL and DOE Cyber Security Protocols

Public

- Published data with a citation
- Registered and non-registered users have access

DOE-Only Workspaces

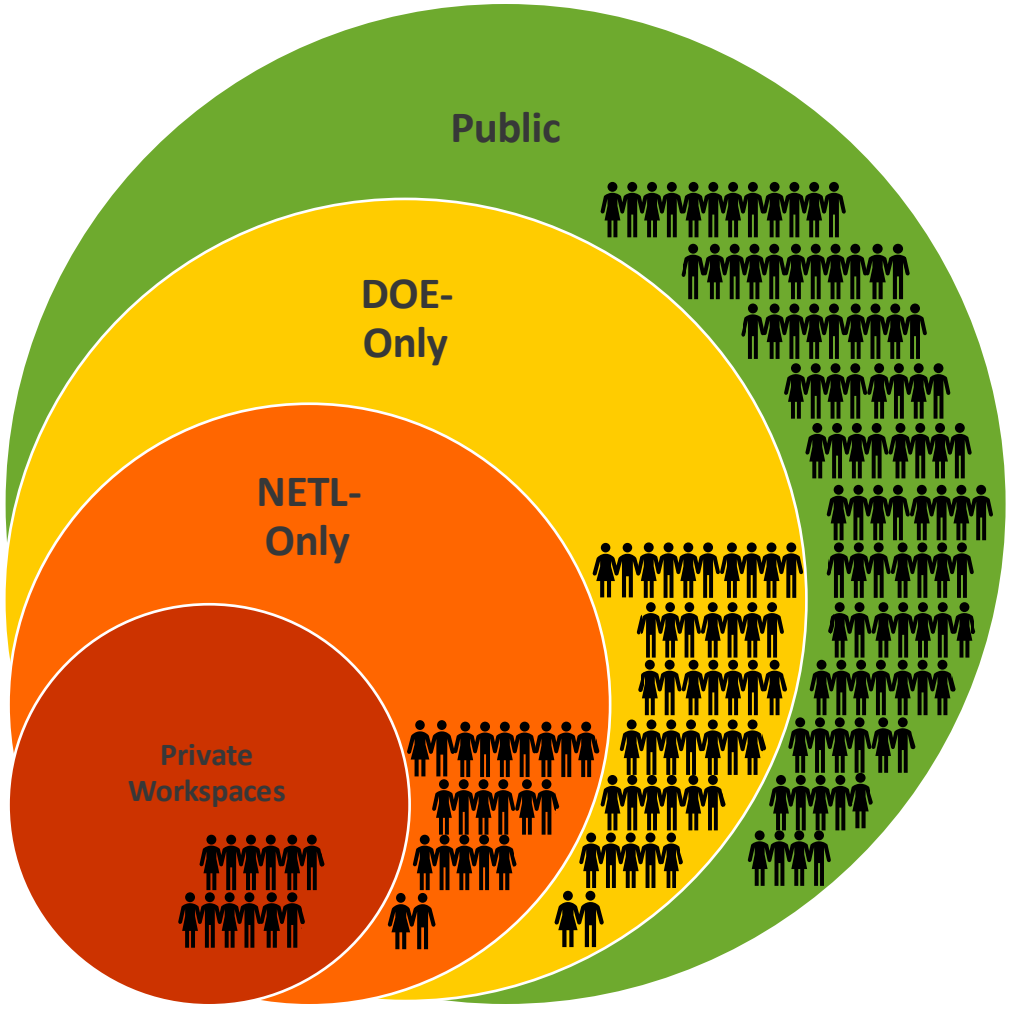
- Semi-private data
- All registered users from DOE Labs and DOE HQ have access

NETL-Only Workspaces

- Semi-private data
- All registered users from NETL have access

Private Workspaces

- Private data
- Admins add/remove registered users and assign roles



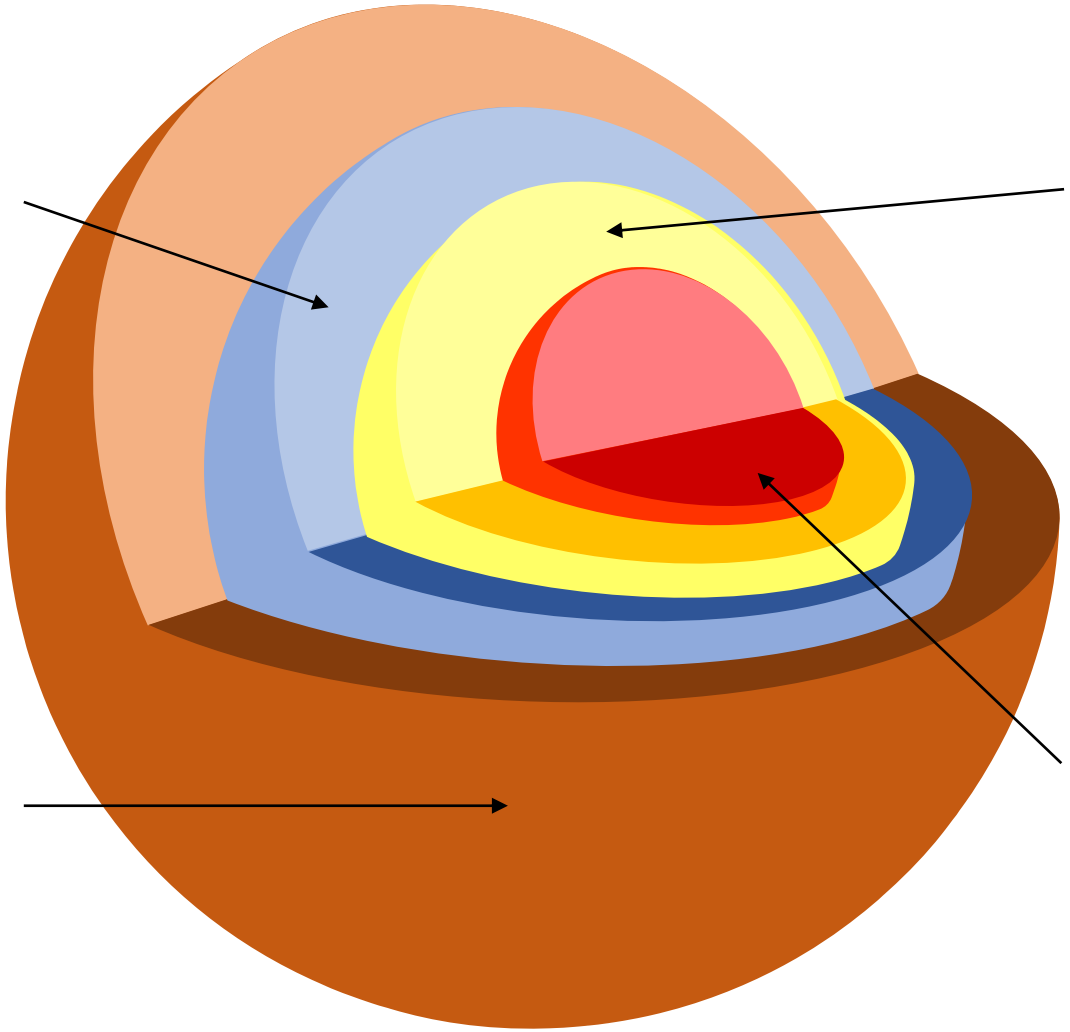
EDX Multi-Layer Data Security

Groups

Groups allow users to group together public submissions under a community or topic.

Public

EDX is DOE FE's virtual data library and laboratory for public curation of FE R&D data, publications, presentations, and tools.



Moratorium

Moratorium allows users to publish submissions and undergo the typical data review process. However, submissions will not be released to the public until the moratorium date has matured.

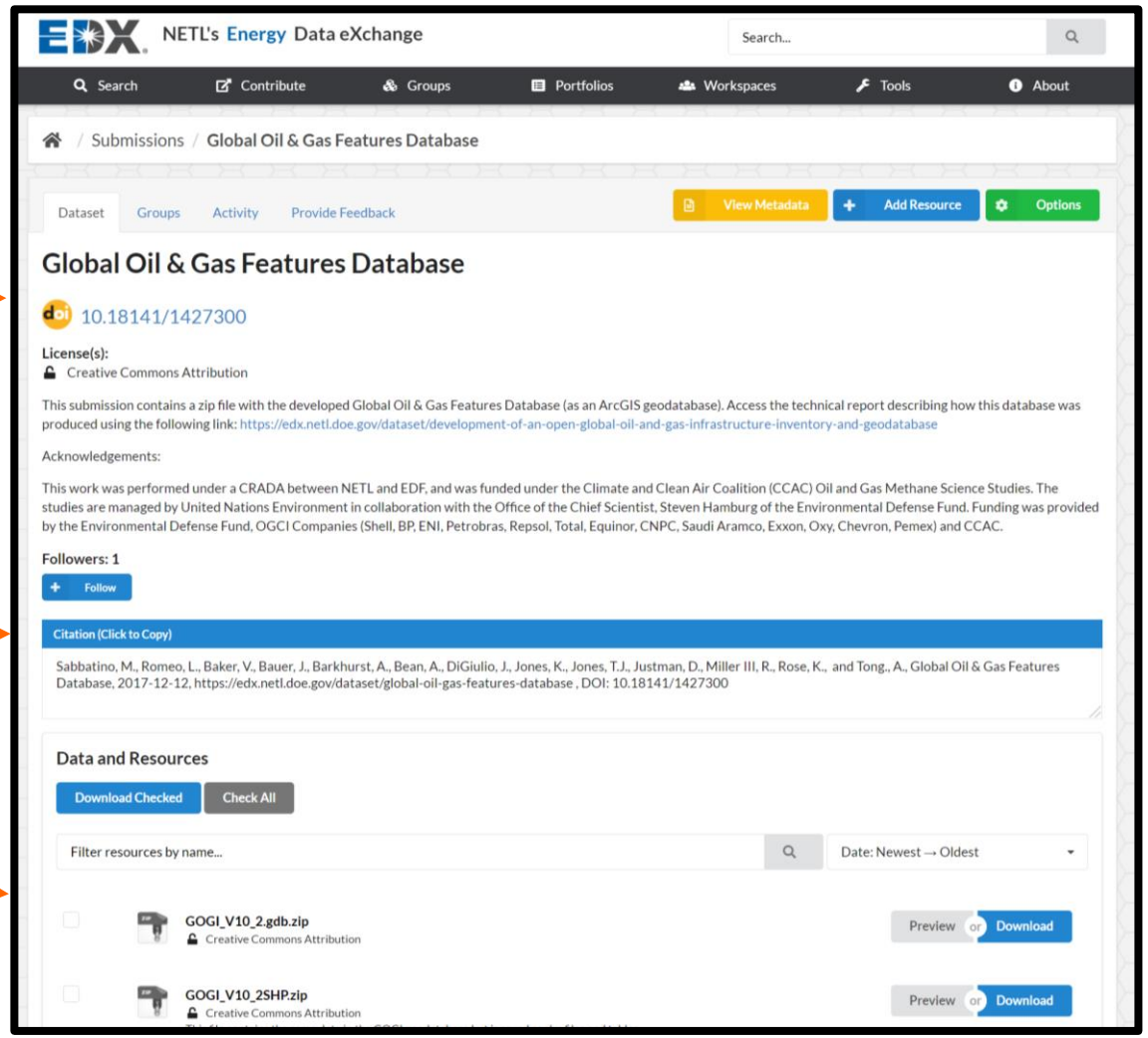
Collaborative Workspaces

Workspaces are a secure environment that offer researchers a platform to access and share private data and data-driven products in support of science-based decision making.

Published Data

Accessible to Both Registered and Non-Registered Users

- Published data is accessible to registered and non-registered users of EDX
- Published data can obtain an OSTI DOI number making it more discoverable in data repositories such as OSTI.gov, data.gov, and Google scholar
- Published data is assigned a data citation
- Each published resource includes a license restriction defined by the contributor



“Smart Portal” for FECM Curated Data Products

NATIONAL ENERGY TECHNOLOGY LABORATORY

Connecting to Resources Beyond EDX

EDX™ Energy Data eXchange

Sync, Connect, and Promote

Findable Accessible Interoperable Reusable

Data Federation Services

Federated services are built upon the same **open-source platform (CKAN)** as EDX and provide data synchronization and advanced searching capabilities making data products more discoverable in other systems.

NGDS National Geothermal Data System

NOAA National Oceanic and Atmospheric Administration U.S. Department of Commerce

OpenEI Open Energy Information

Data Connection Services

Data connection services **connect FE users to the primary source of data** like USGS, EPA, state data, GeoWELL and GeoCube.

EPA

USGS science for a changing world

GEO CUBE geo WELL

Data Promotion Services

Data promotion services help **disseminate FE data products to external data repositories** like OSTI, DOE CODE, and Google Scholar.

OSTI

DOE CODE

Google Scholar

EDX Spatial & GeoCube

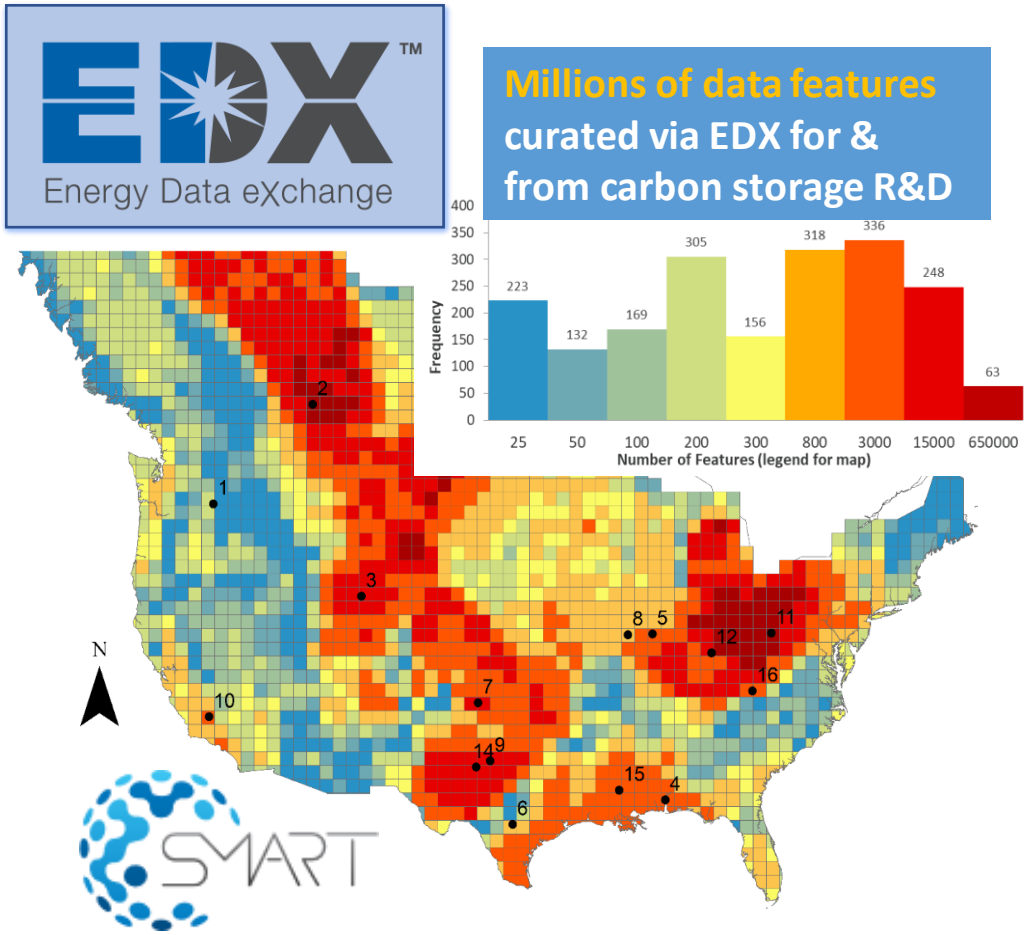
EDX Energy Data eXchange

GEO CUBE

- **Support discovery, access and use of geospatial data & analytical tools**
- **Growing catalog of geospatial resources available**
  - From traditional formats as well as EDX processing to unlock additional place-based insights for EDX resources
- **Serves as a Priority DOE Geospatial Data Repository**
  - Aligns with geospatial management practices outlined in 2021-2025 DOE Geospatial Data Management Strategy, FGDC guidelines, and 2018 GDA covered agency requirements

FECM Data Success Stories (Examples)

- Carbon Storage Program Products
  - EDX is preserving DOE funded R&D products totaling >3,065 resources
  - In reuse by SMART-CS initiative as well as industry and other stakeholders
  - Undergoing transformation for rehosting with open-data assets, EDX spatial carbon storage database contains millions of features (see figure to right)
- Knowledge Management Database
  - Preserved over 14,000 archived oil and gas documents as well as topical reports
- Global Oil and Gas Infrastructure (GOGI) Database
  - Developed using *SmartSearch* tool in under 4 months
  - Accessed and integrated over 6 millions features from open-source hosts, and integrated oil and gas infrastructure data into single database hosted on EDX
  - Downloaded more than 1,000 times, used by Harvard, EDF, UNEP and others in follow on R&D
- FutureGen 2.0 Technical Data Preservation
  - Multi-year effort to publish previously inaccessible unlimited rights data
  - Over 200GBs of archived data of which 115 GBs has been published for unrestricted use/reuse



Morkner, P. …and **Rose, K.**, **In review,** Distilling Data to Drive Carbon Storage Insights, *Computers & Geoscience*

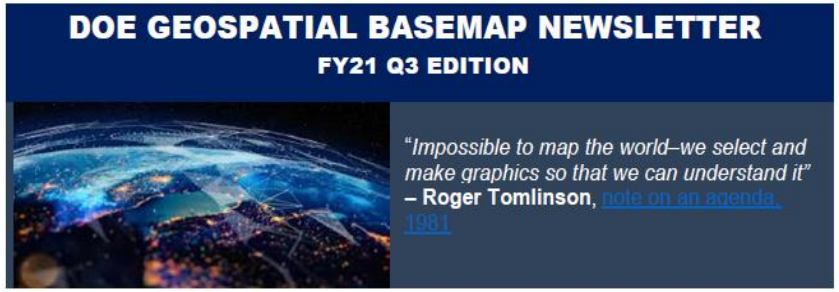
EDX Inside DOE, but Beyond FECM (Examples)

NATIONAL ENERGY TECHNOLOGY LABORATORY

EDX Unanimously Approved as OCIO's First Priority DOE Geospatial Data Repository February 2021

EDX has appeal and capabilities relevant beyond FECM:

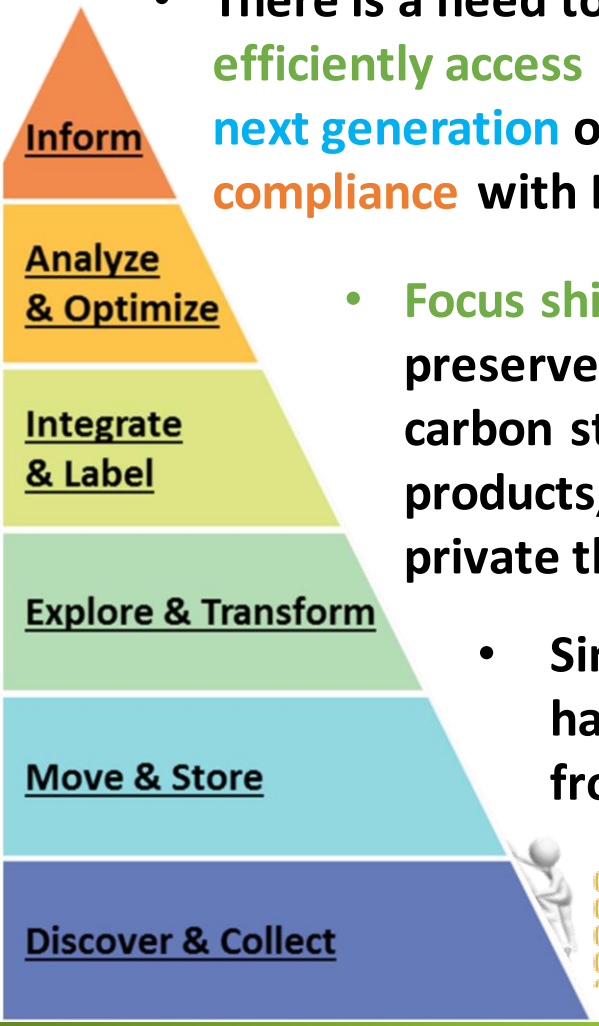
- **DOE OS/OSTI**, EDX connects to OSTI's DOI generator, and OSTI pulls metadata about EDX resources into its cataloging services
- **DOE AITO,** Briefed AITO Director on EDX; Pending coordination with AITO's AIX database manager on opportunities with EDX and AIX
- **DOE Chief Counsel**, has provided reviews of EDX workflows, public and private, to ensure IP protocols are robust
- **DOE OCIO**, EDX *SmartSearch* AI/ML data search tool partnering with OCIO to help test Google Cloud Platform at scale (*in progress*)
- **DOE OCIO**, EDX named as priority geospatial data curation platform recommendation #1 for all of DOE
- **DOE Environmental Management & Legacy Management,** through DOE's Subsurface Interagency Coordination Team (IACT), EM/LM managers have encouraged curation of R&D Products on EDX
- **DOE EERE,** Federation between EDX and EERE's OpenEI platform for enhanced DOE R&D public product discoverability; Ongoing EERE AMO materials project that is leveraging EDX and *SmartSearch* to drive materials data breakthroughs
- **DOE OTT**, OTT funded use EDX *SmartSearch* for DOE product cataloging and COVID-19 outreach efforts; Briefed OTT manager on EDX DOE tiered community capabilities for potential knowledge community need of OTT

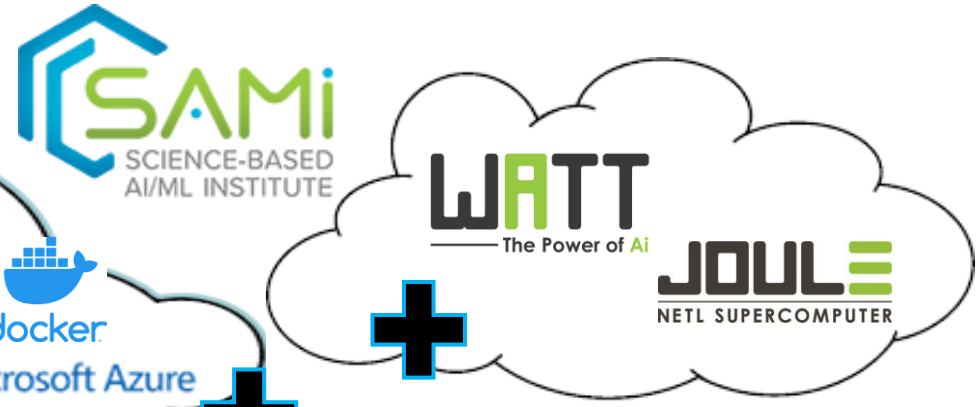


Leveraging EDX Next-Gen DOE Data Computing for Carbon Storage Data

NATIONAL ENERGY TECHNOLOGY LABORATORY

- There is a need to preserve and efficiently access resources to drive the next generation of R&D while ensuring compliance with DOE regulations
  - Focus shifted in 2016 to preserve and curate carbon storage data products, both public and private through the EDX
    - Since then, millions of dollars in R&D products have been preserved and made publicly available from carbon storage efforts





EDX
Energy Data eXchange

Big Data

Open Carbon Storage Data Collection

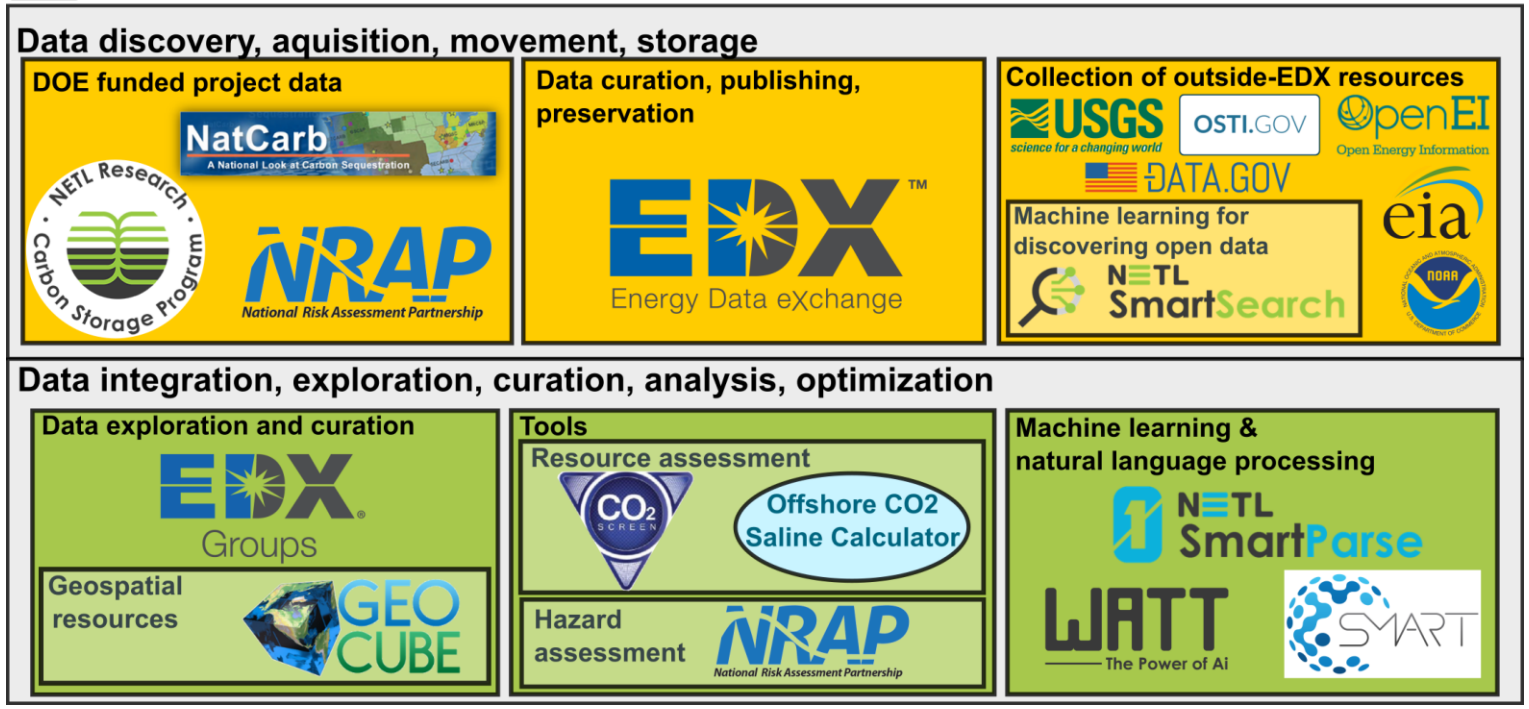
Data sources include:

- National carbon sequestration database - **NATCARB**
- **Regional Carbon Sequestration Partnership data**
- Field projects like **FutureGen 2.0 and CarbonSAFE**
- **National Risk Assessment Partnership** data, models, tools

Data Types:

- **Geospatial (shapefiles and rasters)**
- **Well logs**
- **Seismic data**
- **Text-based resources**
- **Tools, models**
- **Data catalogs**
- **Both surface and subsurface geology**

Carbon storage data computing resources



- **85+ TB source data and growing**
- **3455+ resources**
  - *872 Published EDX resources*
  - *632.7 GB of Published EDX resources*
- **1800+ text-based resources on EDX**
- **Data curated into Groups on EDX**

Challenges for Open Carbon Storage Data

NATIONAL ENERGY TECHNOLOGY LABORATORY

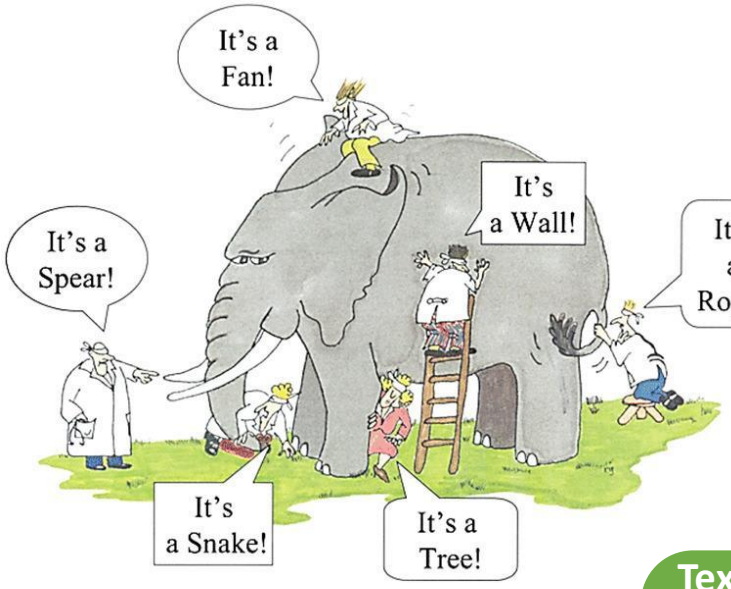
Data is disparate in nature

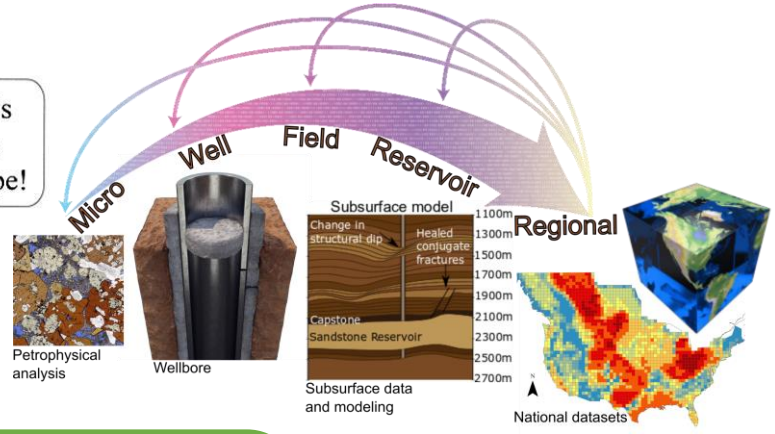
Carbon storage data consists of **structured** and **unstructured** data types consisting of **subsurface** and **surface** geologic data, models, tools

Data exists at **many different scales**

The way we search for an identify data is convoluted

Hard to identify all the data, i.e., see the whole "elephant", without falling down the "rabbit hole"





Spatial data:
- Shapefiles (field, basin, regional scale)
- Datasets
- Models

Text-based Data
- Documents
- Publications
- Power points
- Memos
- Posters

Other types of data:
- Tools
- Applications
- APIs

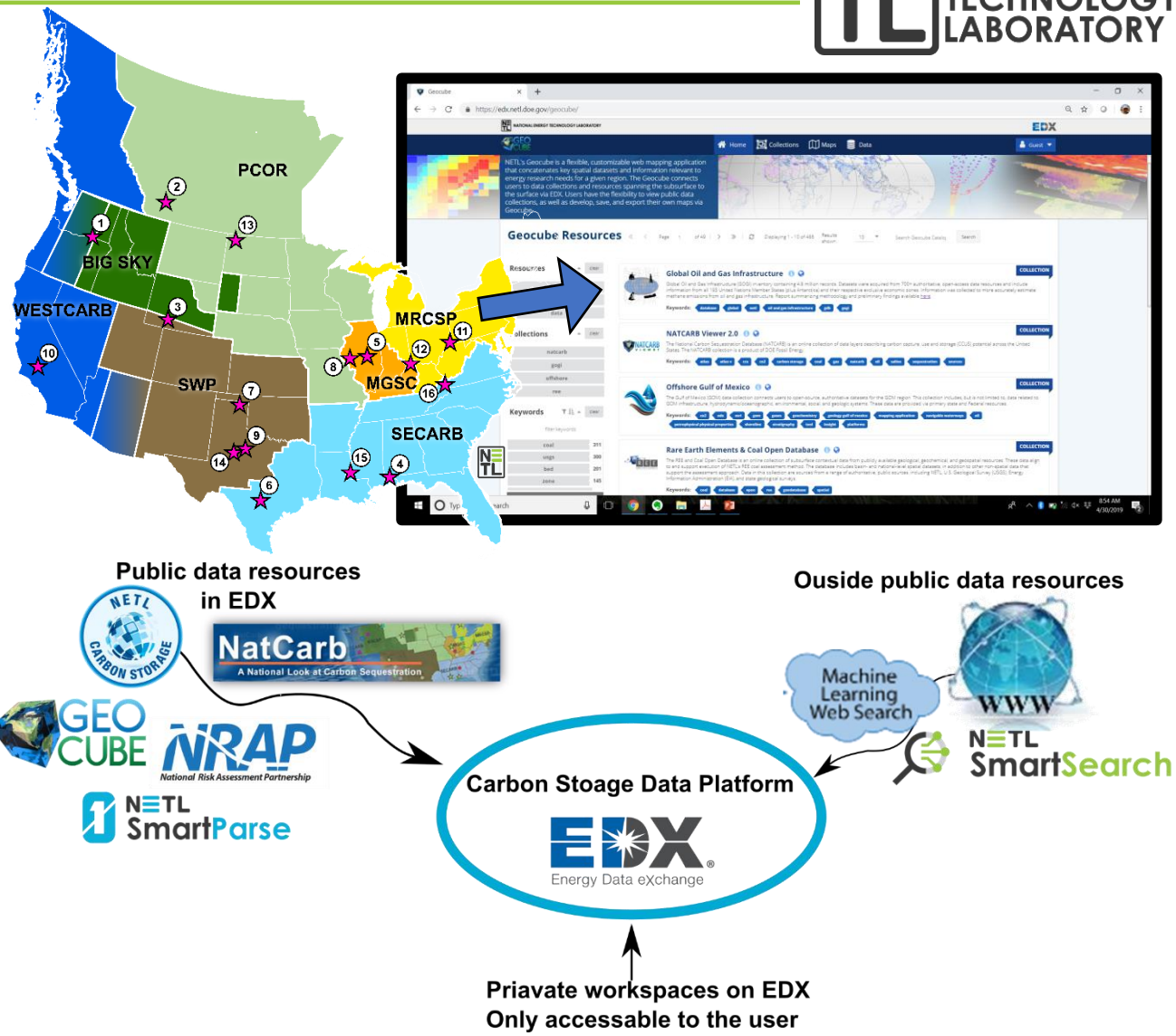
Feeding the data hippo!

Carbon storage data curation goal: To find data curation solutions that use a holistic view and bring together diverse data types

Carbon Storage Data Search & Visualization on EDX

GeoCube enables easy **spatial search capabilities** – users can **rapidly access and visualize data** on an interactive map of the world

- Hosts NATCARB database and 315+ other carbon storage relevant data layers
- Search for relevant data resources at any spatial scale
- Bring together carbon storage data resources with tools to perform analysis
- Integrates tools:
  - CO2-SCREEN, Offshore CO2 Calculator
- Links to NRAP tools and others
- Limit results with keyword search criteria
- Make accessible relevant supplemental resources from across the FE portfolio - drives citation and reuse of data



Tools on EDX for Data Discovery, Labeling and Curation

NATIONAL ENERGY TECHNOLOGY LABORATORY

NETL SmartSearch

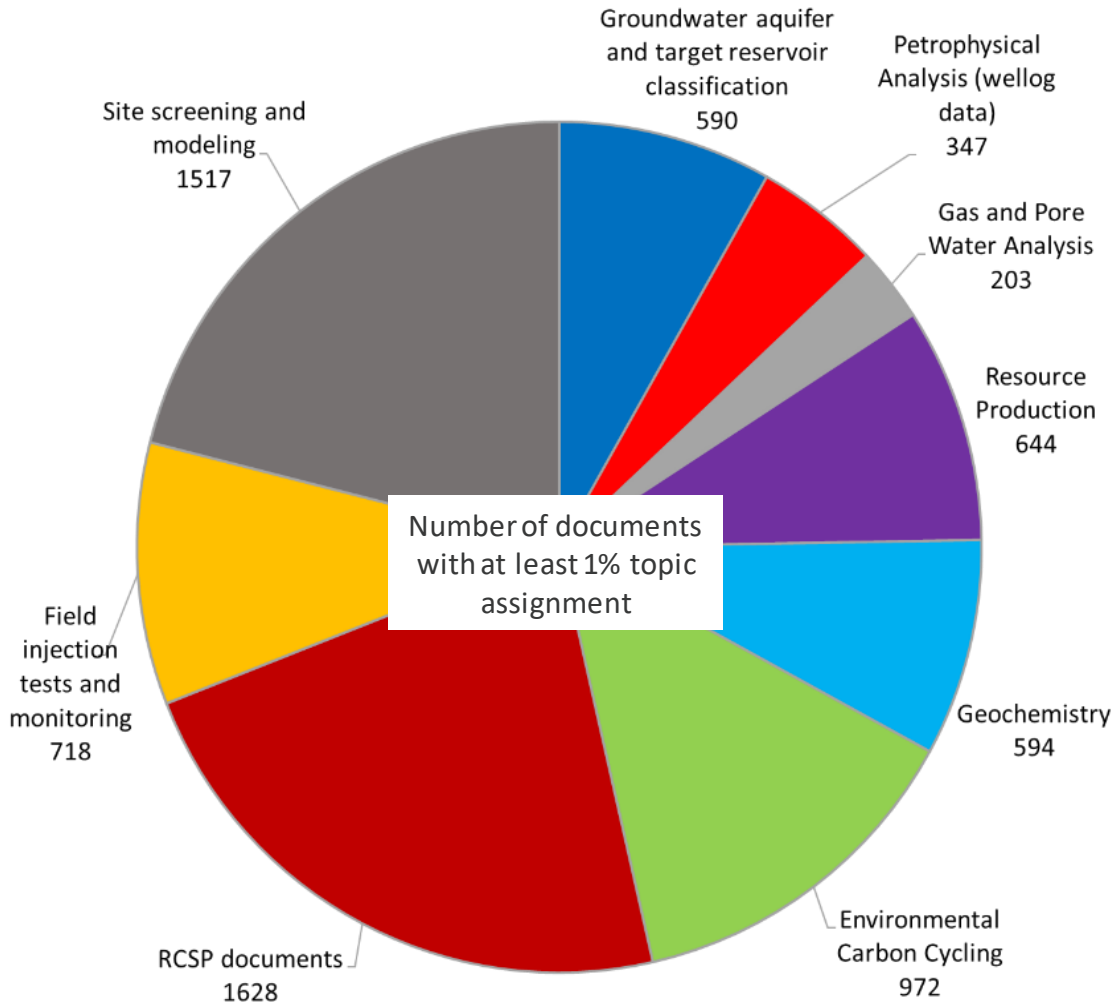
Web crawler too – in beta testing on EDX

- **Searches for resources on WWW based on seed (data, document, etc.)**
- **Searches for resources similar to what you feed it**
- **Can target specific language in documents**

NETL SmartParse

Natural Language Processing tool for cloud computing on EDX

- **Parses through language in unstructured data (such as documents, ppts, etc.) to achieve:**
  - **Topic modeling**
  - **Keyword Identification**
  - **Geographic named entity recognition**



Morkner, P., Bauer, J., Creason, C., Sabbatino, M., Wingo, P, Greenburg, R., Walker, S., Yeates, B., and Rose, K., in review, **Distilling Data to Drive Carbon Storage Insights,** *in review at Computers & Geoscience*

FutureGen Subsurface Technical Data Release

NATIONAL ENERGY TECHNOLOGY LABORATORY

Data rescue effort – over 217 GB (zipped) of data preserved on EDX private workspace in 2016

A subset of data was vetted for sensitivity and publicly released in 2020

6,800+ files (84GB zipped) containing data:

Subsurface Characterization

- 2D seismic, borehole geophysical logs, gravity, plug analyses
- Regional well and core data
- Interpretations of above data

Field tests

- Hydrogeological, geomechanical, well integrity tests

Monitoring Data

- Groundwater, soil/gas, ecological, atmospheric monitoring

Models

- USDW, reservoir models

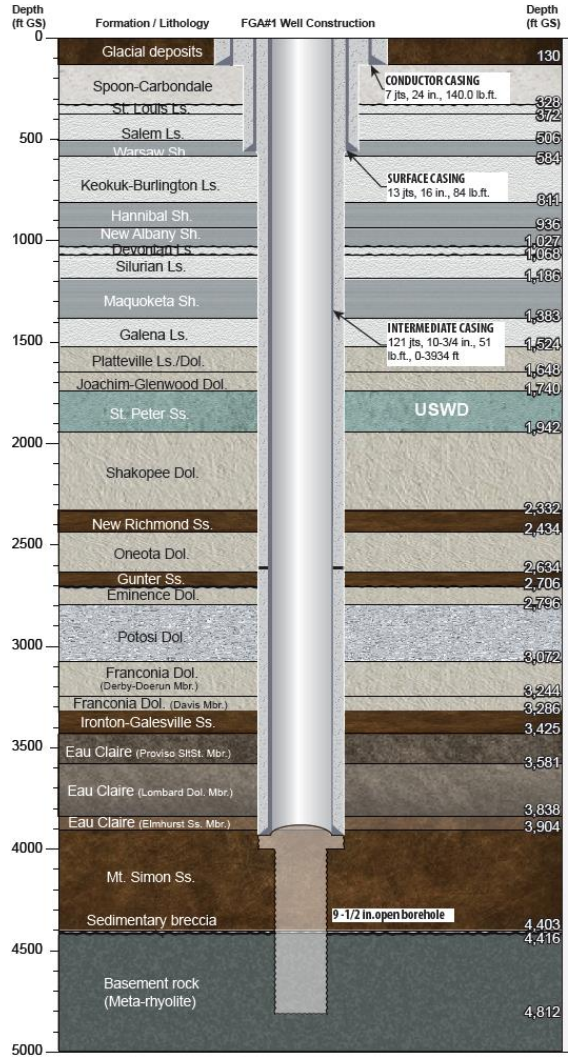
Geospatial data

- Seismicity, structural, thickness, boundary shapefiles

FutureGen ALLIANCE

- 52 2D seismic lines
- 94 well logs
- 212 spreadsheets
- 218 shapefiles
- 300+ reports
- 800+ figures, core photos

https://edx.netl.doe.gov/group/futuregen-data

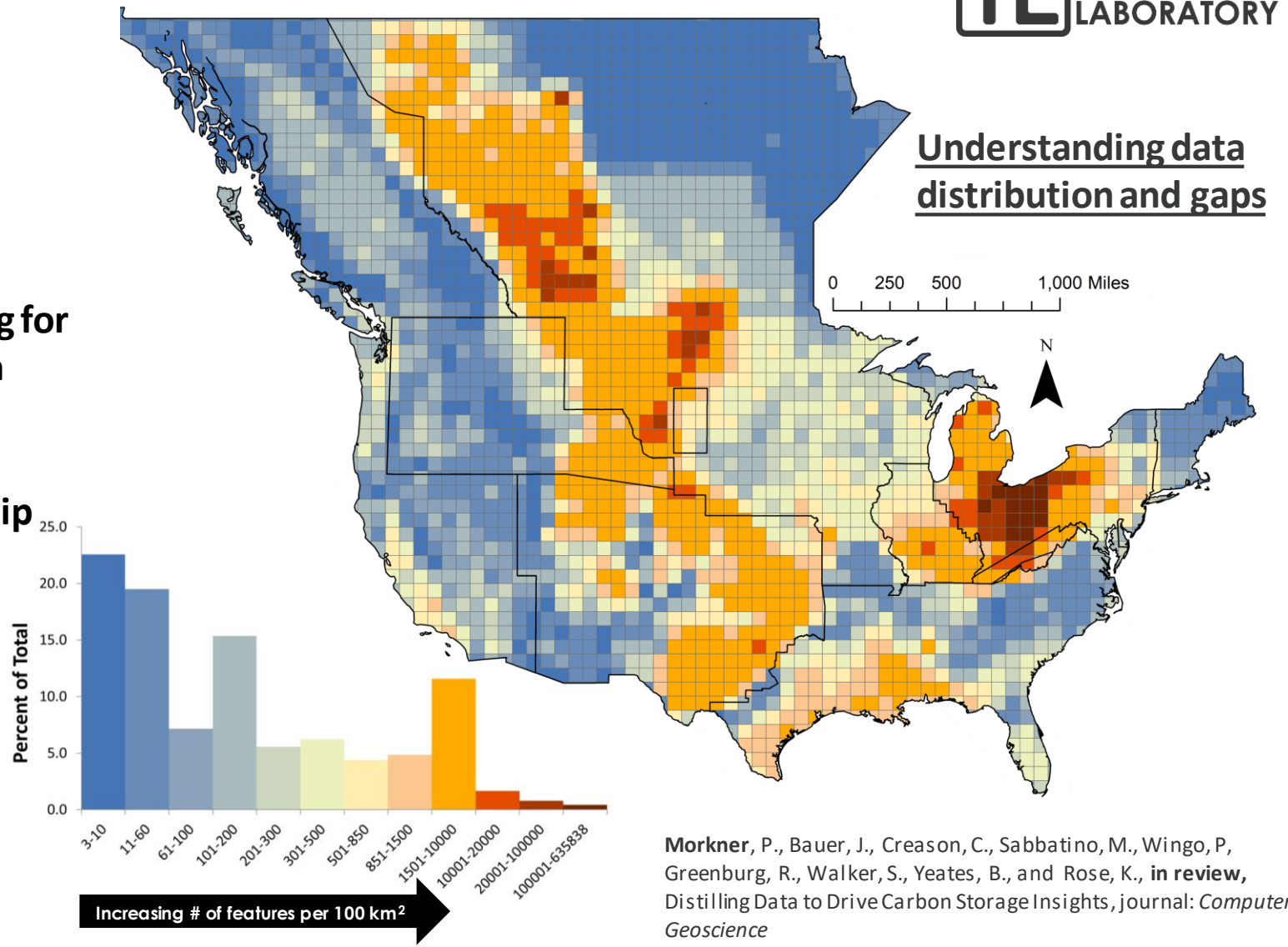


FGA#1 Stratigraphy and Well Construction

Use of CS Data to Drive R&D Products and Insights

Available CS data has been used for:

- **Site screening, reservoir modeling, and potential storage estimation by**
  - State geologic surveys
  - **CarbonSAFE projects**
  - Industry groups like EPRI
  - **Science-informed Machine Learning for Accelerating Real-Time Decisions in Subsurface Applications (SMART) initiative**
  - **National risk assessment partnership**
- Spatial data density analysis
- **Natural language processing topic model** development



Morkner, P., Bauer, J., Creason, C., Sabbatino, M., Wingo, P, Greenburg, R., Walker, S., Yeates, B., and Rose, K., **in review,** Distilling Data to Drive Carbon Storage Insights, journal: *Computers & Geoscience*

Summary

10+ years of development of EDX and 5+ years of FECM and Carbon Storage program investments into data curation and management has resulted in the development of AI/ML tools and the preservation of millions of dollars of research products which benefits ongoing and future research. This has led to:

- Preservation and publishing of thousands of DOE funded resources
- A better understanding of CS relevant open-data resources throughout the U.S. and Canada
- Improved access through the integration of CS data resources on EDX
  - Including within specialty apps like GeoCube for spatial data visualization and curation
- Development, testing and use of EDX-driven AI/ML data discovery, labeling, integration capabilities trained to support Carbon Storage, SMART-CS, and NRAP
  - e.g., SmartSearch and SmartParse (EDX version of NLP tools presented here) for further searchability with spatial searches and keyword searches

Thank you!

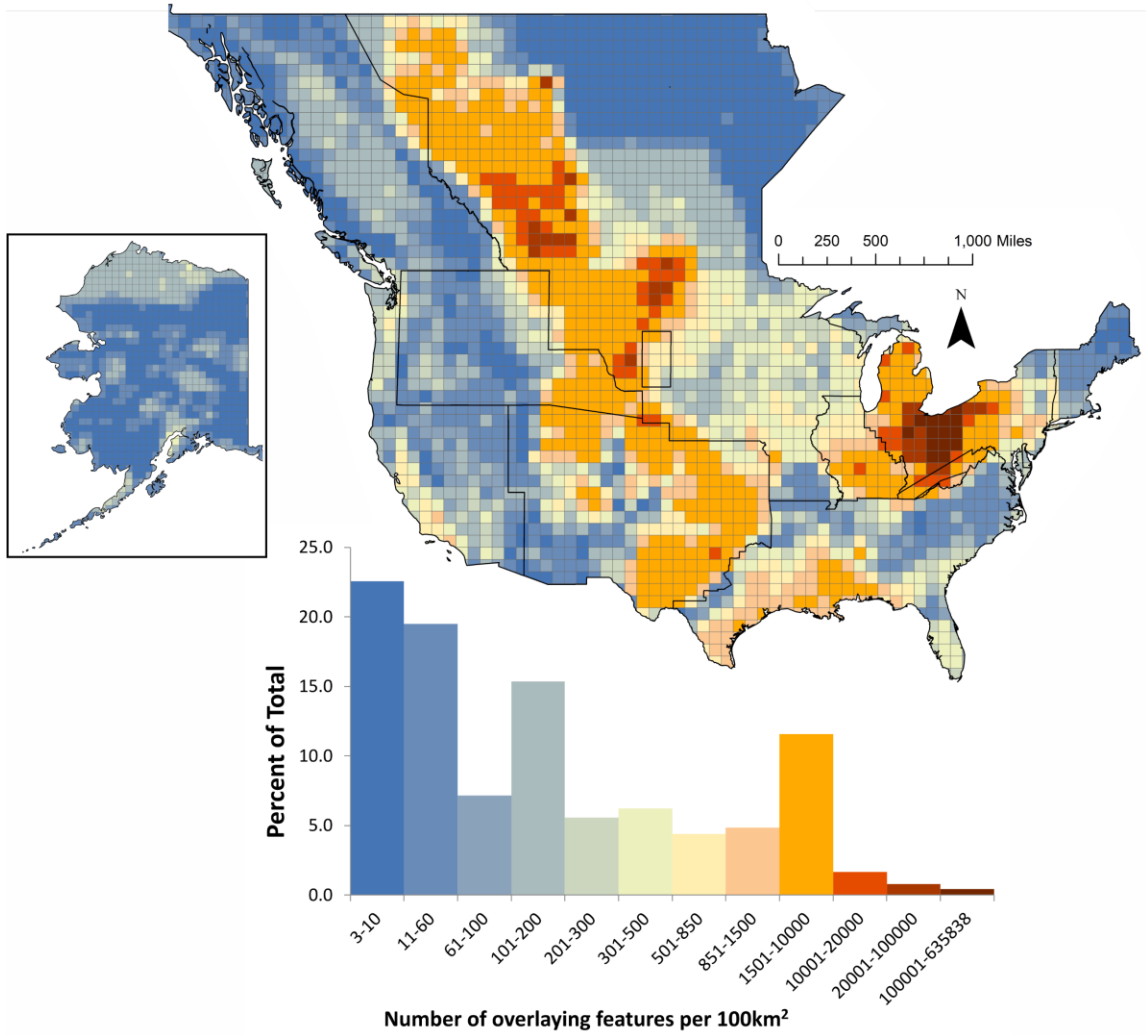
NETL RESOURCES

VISIT US AT: **www.NETL.DOE.gov**

@NETL_DOE

@NETL_DOE

@NationalEnergyTechnologyLaboratory



CONTACT:
Paige Morkner
Paige.Morkner@netl.doe.gov

U.S. DEPARTMENT OF ENERGY