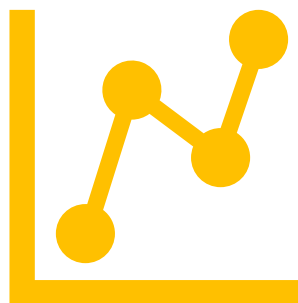
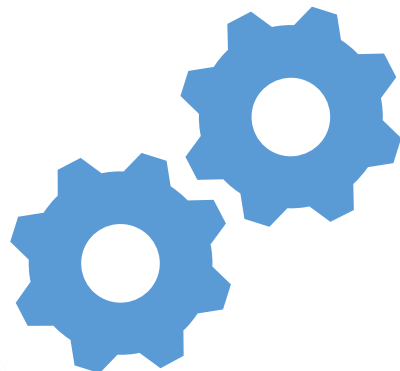




Carbon Storage's FAIR Data Repository

Catalyzing FECM's data-driven future,
a decade of contributions from the
DOE Carbon Storage Program



Inform

Analyze
& Optimize

Integrate
& Label

Explore & Transform

Move & Store

Discover & Collect

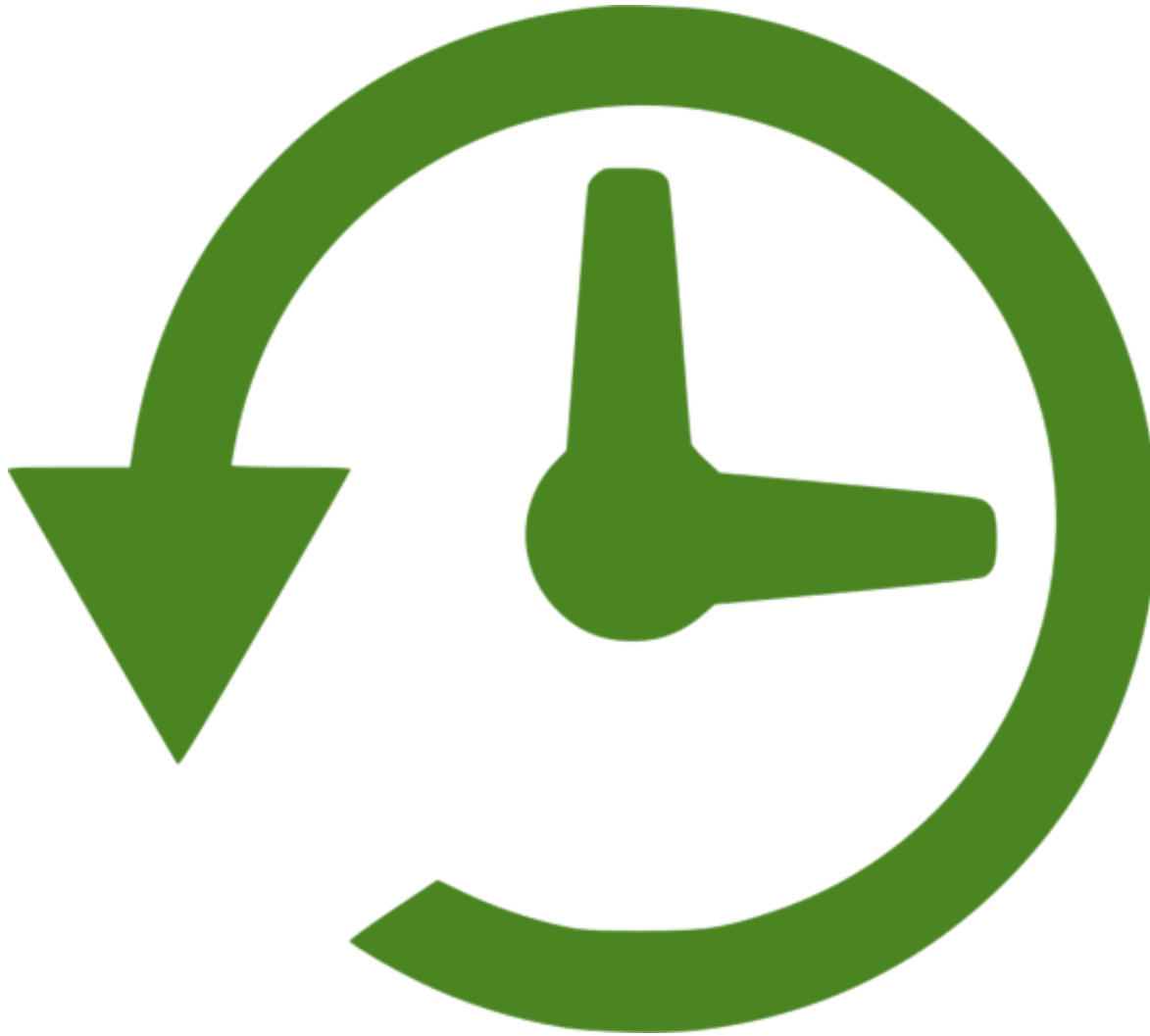


Kelly Rose, Technical Director



U.S. DEPARTMENT OF
ENERGY

Time warp... back in 2011



A few 2011 News Headlines...

- 9.0 earthquake hit east of Japan causing tsunami leading to Fukushima nuclear plant disaster
- Novak Djokovic won his first Wimbledon title
- Steve Jobs passed away...



Some Top Technological Advances of 2011

- **Mobile Internet** launched...instant messaging via Wi-Fi etc
- **Tablet Computers** take off, 1st Kindle is released...
- **Cloud Computing** goes commercial...
- **App-driven era** arrives... e.g. Snapchat launches, Uber app also first launches in SFO

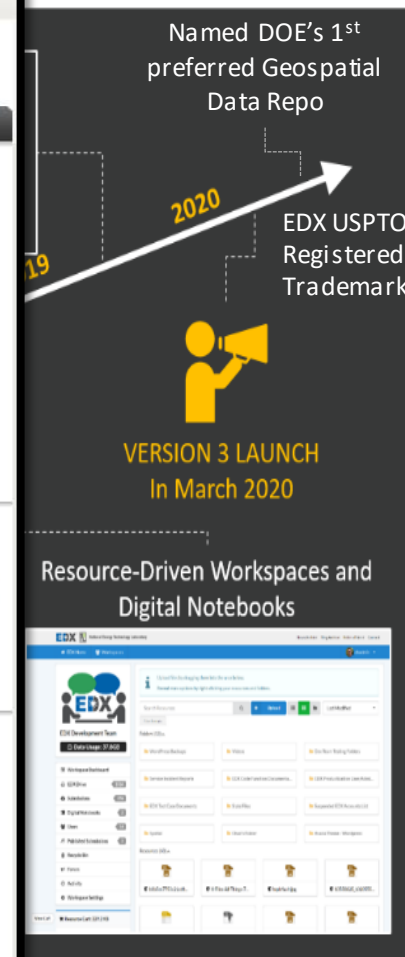
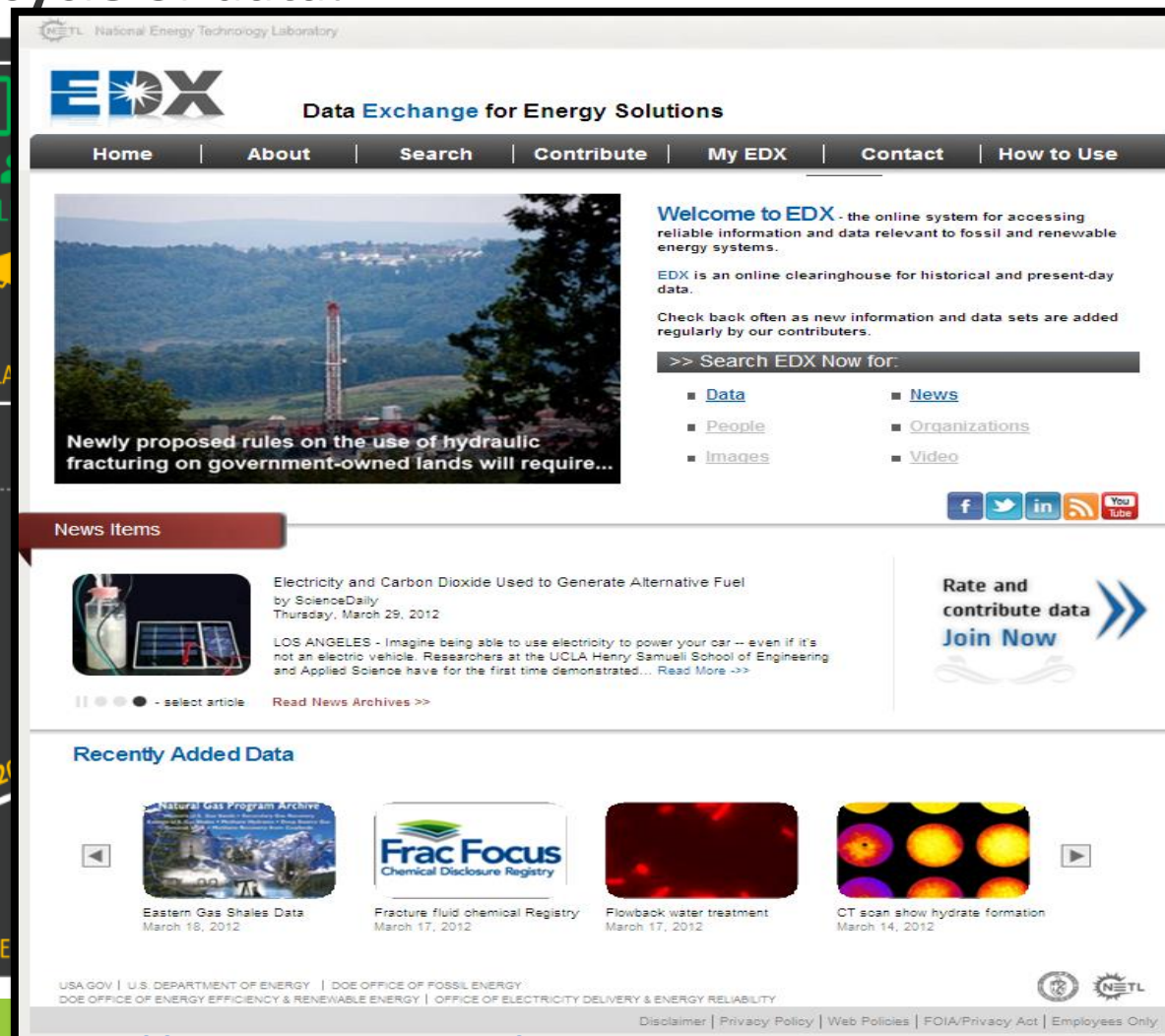
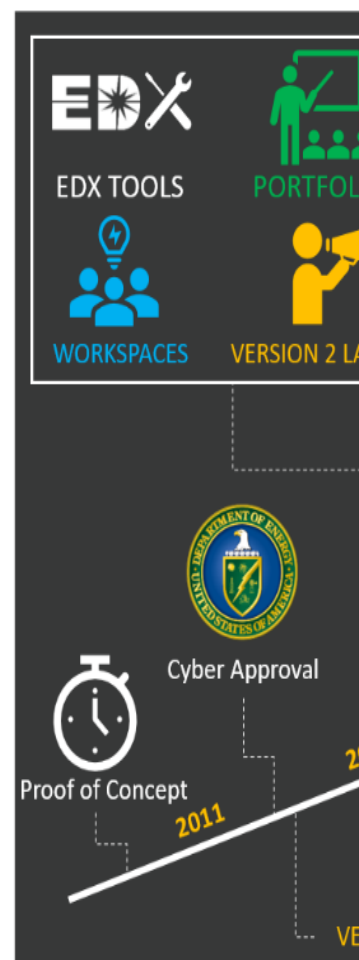
<https://www.ukfast.co.uk/blog/2011/12/14/technological-advances-of-2011/>

2011 Denotes the Launch of DOE FECM's 1st instance of



The Energy Data Exchange (EDX) identified key features needed for a system to support the entire life-cycle of data:

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



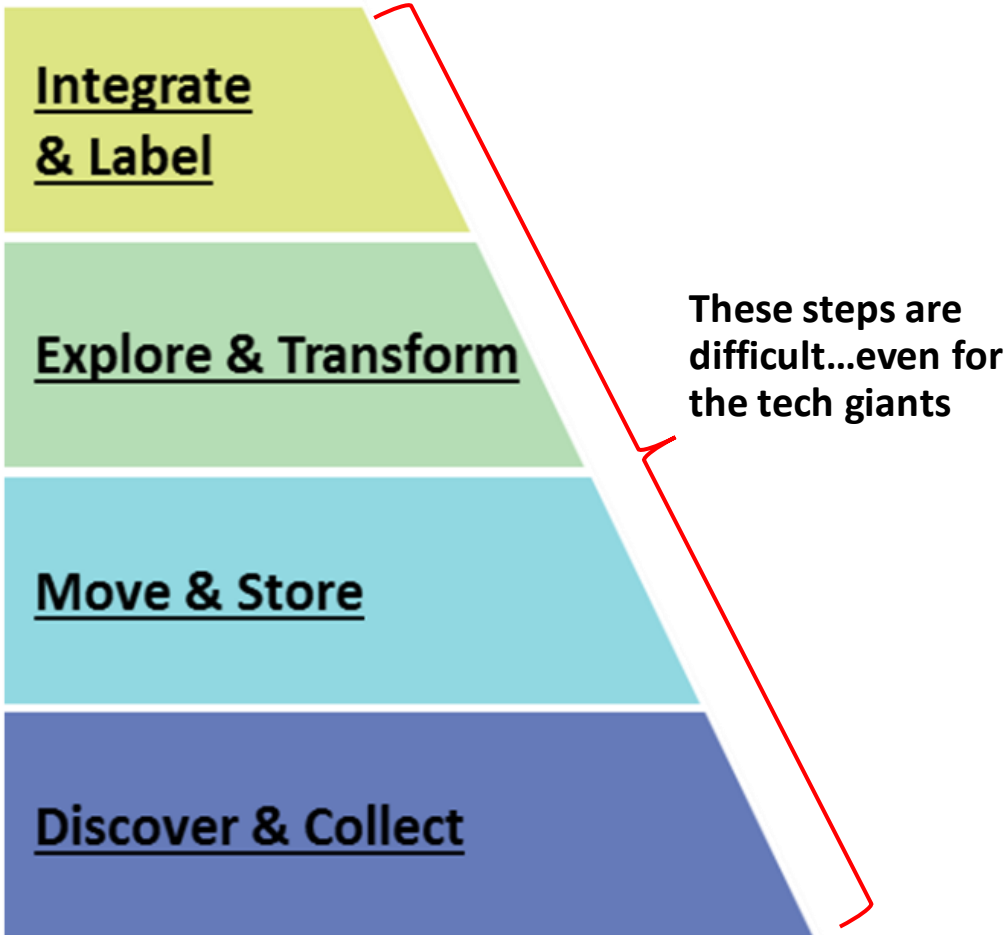
The era of FAIR Data and data science



- Data is suddenly valued
- Old data is important
- Technology improvements
- 2013/2014 Federal and DOE Orders mandate curation of federal R&D products

The “Data” Challenge

Tackling the 20:80 data access issue



THE WALL STREET JOURNAL.

U.S. Edition | June 7, 2019 | Print Edition | Video

Home World U.S. Politics Economy Business Tech Markets Opinion Life & Arts Real Estate WSJ. Magazine Search

Data Challenges Are Halting AI Projects, IBM Executive Says

The cost and hassle of collecting and preparing data comes as a shock for some companies, according to Arvind Krishna

“The world’s most valuable resource is no longer oil, but data” -The Economist

By *Jared Council*
May 28, 2019 5:30 a.m. ET

Data is the energy for AI

High-level gap:

“There is no system in place to **capture, analyze, and share AI/ML data, models, algorithms, results, and lessons learned across the DOE enterprise.** Without such a system, **every AI project has to start anew.**”

AI/ML Final Report,
Sec. of Energy Advisory Board

~99% of AI research ?

20%



ACTION

Train a model

~1% of AI research ?

80%



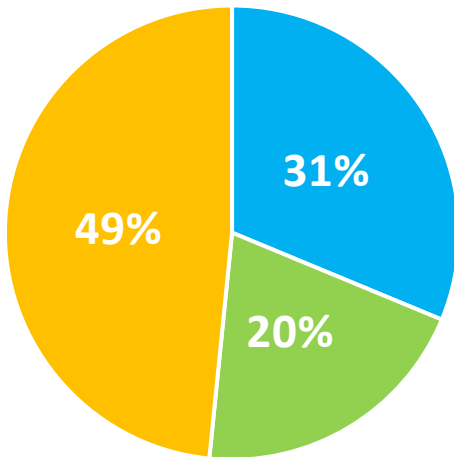
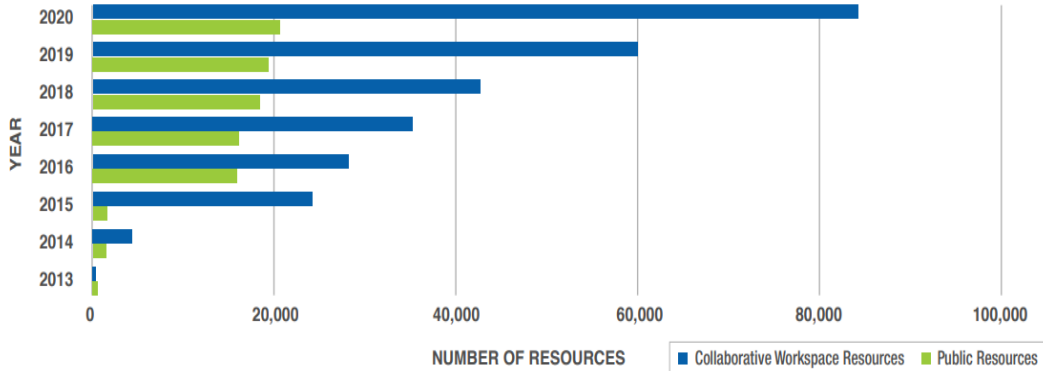
PREP

Collect and prepare high quality data

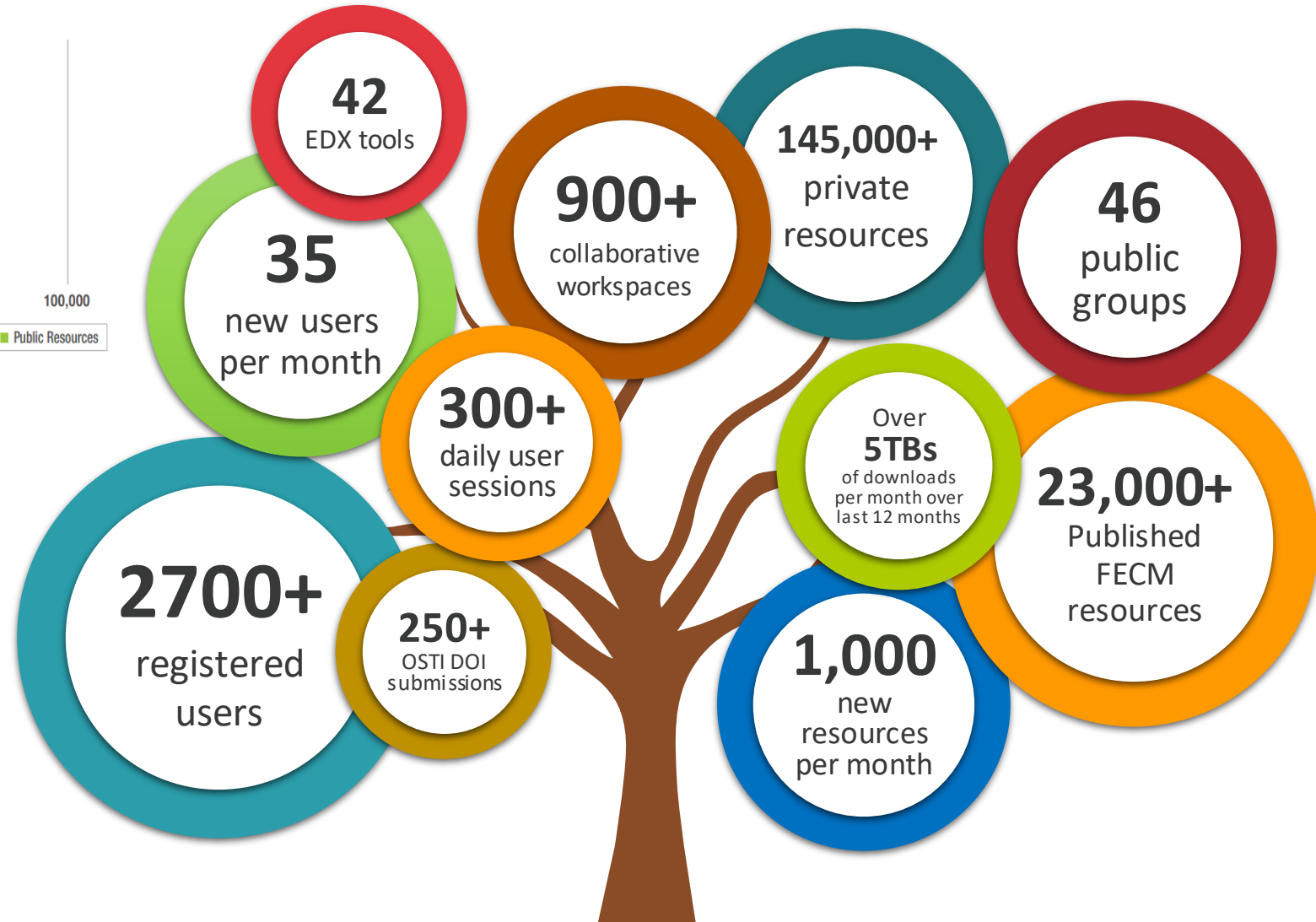
Andrew Ng, 2021

TOTAL PUBLIC RESOURCES

Over 20k resources and 800 collaborative workspaces



■ NETL ■ DOE National Lab ■ External



EDX...more than a data repository

Search

Submissions within the public search on EDX provide access to many forms of information including but not limited to **presentations**, **publications**, **tools**, and **data**.

Sort

Submissions within the public search on EDX can be sorted **spatially**, by **keyword**, and **file format** connecting users to the appropriate data and information quickly and efficiently.

Groups

Submissions within the public search on EDX can be clustered into Groups of related data. Some popular EDX Groups include the **Kimberlina Data Group**, **Appalachian Basin Data Group**, and various **RCSPs**.



Tools

EDX Tools provide access to, management of, and interaction with data through a collection of tools including **CO2 Screen**, **Natcarb Viewer 2.0**, **Offshore Risk Modeling Suite** and **NRAP Tools**.

Analyze

EDX hosted Tools like **Geocube**, **Natcarb Viewer**, and **Common Operating Platforms** allow users to find, sort, visualize, and use EDX-hosted models and tools with data via the EDX platform.

Visualize

EDX Tools provide visualization of data through various tools including **ParaView**, **Papaya**, and **RokData**.

Tiered Access Using Role-Based Security

Compliant with 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

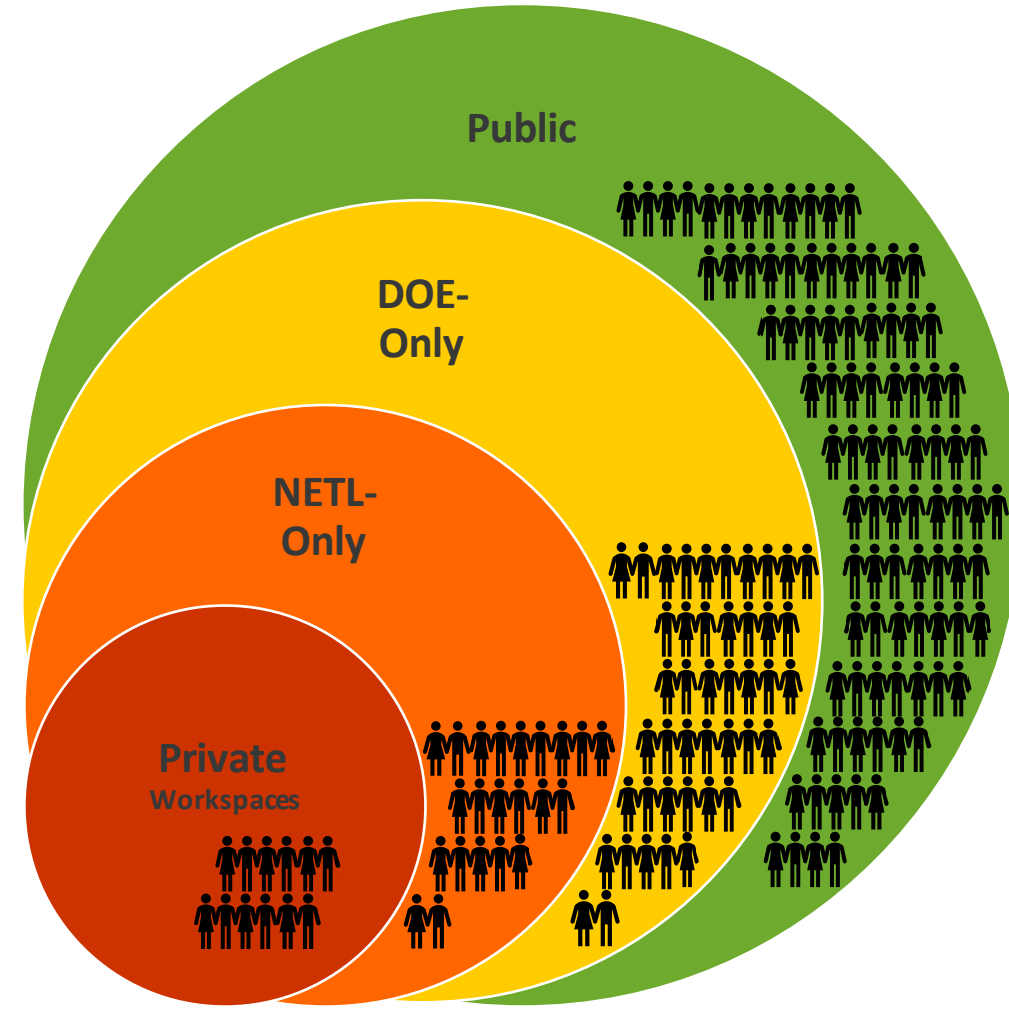


- Semi-private data
- All registered users from NETL have access

Private Workspaces



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

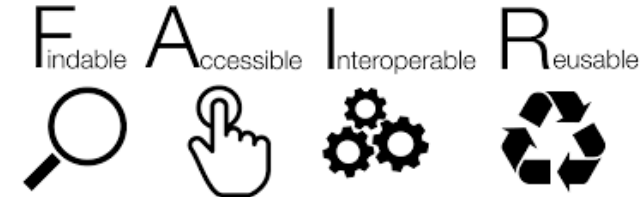


“Smart Portal” for FECM Curated Data Products

Connecting to resources beyond FECM



Sync, Connect, and Promote



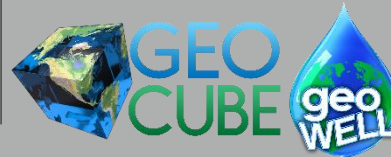
Data Federation Services

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.



Data Connection Services

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



Data Promotion Services

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



In 2021:

DOE OCIO Names EDX as 1st Priority Geospatial Data Repository

- **Compliance** with Geospatial Data Act, 2018 requirements
- Provides **Tiered Access** to resources using role-based security
- **Federated System**, connected with other DOE data repositories, including OSTI, EERE's OpenEI, & data.gov, support federation to other platforms in future
- Offers **standardized metadata & data quality metrics**, aligns with GDA responsibilities
- **Aligns to FAIR data practices**, provides an enduring link, citation, and DOI number for published resources
- Adheres to **DOE Cyber Security Policies**
- **Vetted and approved for use by DOE Chief Counsel**, all DOE National Laboratories (including NNSA labs), and DOE Headquarters
- Aligns with DOE specific rules and regulations & Federal requirements

DOE GEOSPATIAL BASEMAP NEWSLETTER

FY21 Q3 EDITION



"Impossible to map the world—we select and make graphics so that we can understand it"
— Roger Tomlinson, [note on an agenda](#), 1981

GEOSPATIAL NEWS

DOE NAMES FIRST PRIORITY GEOSPATIAL DATA RESPOSITORY



<https://edx.netl.doe.gov>



The Geospatial Executive Core members participated in training sessions in January to get background and context on the National Energy Technology Laboratory (NETL) Energy Data eXchange (EDX). From there a quorum vote took place in February, to designate (EDX) as a priority geospatial data repository for DOE. Hosted out of NETL, EDX was designed to serve as a data curation application primarily for Fossil Energy but since has grown into a collaborative space for multi-research teams to publish and release all data and datasets.

Naming priority geospatial data repository was motivated to quickly direct DOE members to DOE owned and maintained data repositories with geospatial data, that ensure the offered data align with current federal data policies and governance, transition forms, and ensure available data aligns with GDA requirements around data standards and sharing policies. Designated geospatial data repositories, and their use across DOE, will begin to standardize the application of geospatial data management practices across the agency, as well as ensure that wherever people are, their geospatial data align with GDA requirements.

EDX Spatial & GeoCube



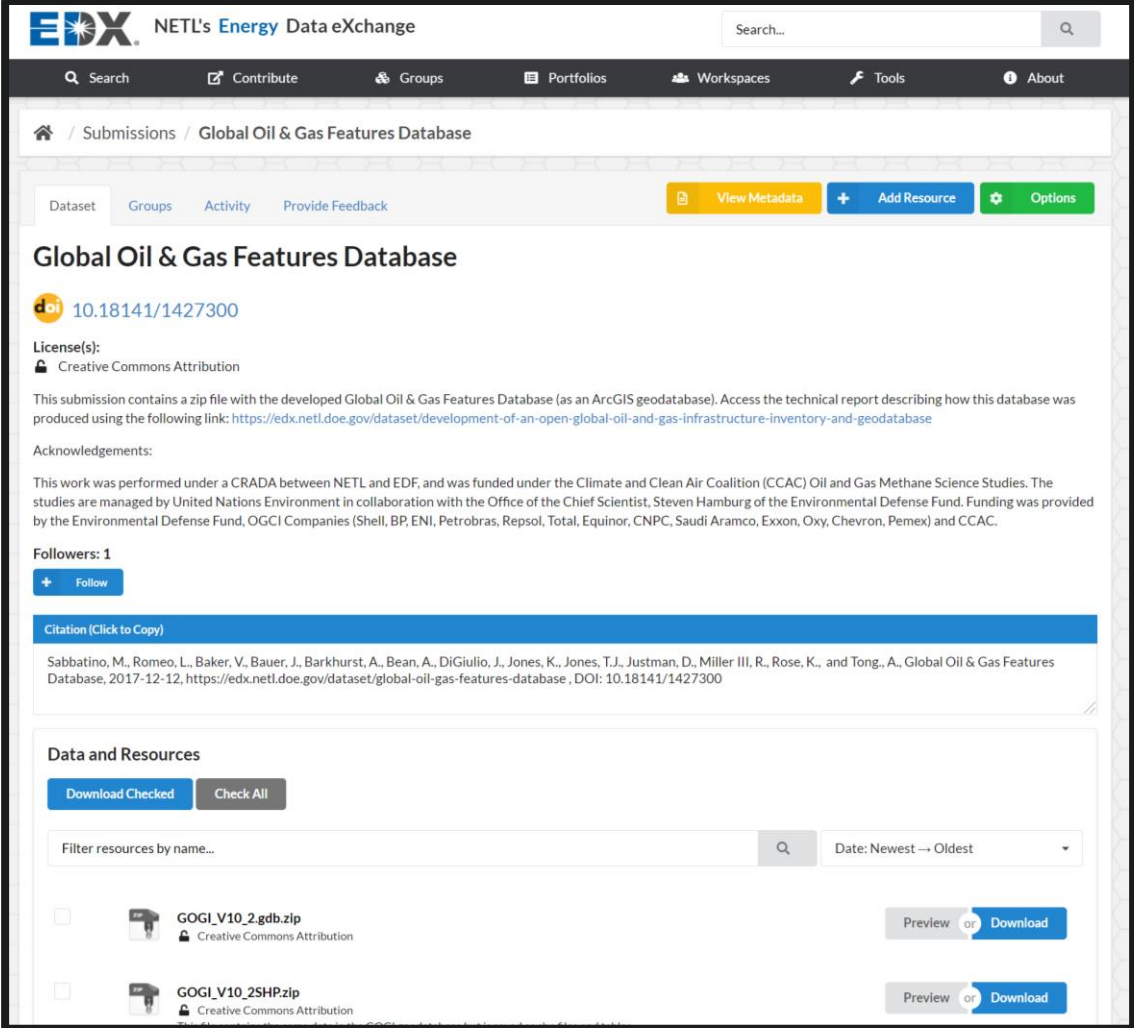
- Support discovery, access and use of geospatial data & analytical tools through EDX and EDX's web-mapping application, **GeoCube**
- Growing catalog of geospatial resources available through EDX
 - Crosscutting data for surface & subsurface FECM systems
- Serves as a Priority DOE Geospatial Data Repository
 - Aligns with geospatial management practices outlined in 2021-2025 DOE Geospatial Data Management Strategy, FGDC guidelines, and 2018 GDA covered agency requirements



“Publishing” FECM Products

Following Datacite.org citation format

- Are **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
- Published data can be organized into collections known as **EDX Groups**



EDX NETL's Energy Data eXchange


Search... [Search Icon]

Search Contribute Groups Portfolios Workspaces Tools About

Submissions / Global Oil & Gas Features Database

Dataset Groups Activity Provide Feedback View Metadata Add Resource Options

Global Oil & Gas Features Database

 10.18141/1427300

License(s):
Creative Commons Attribution

This submission contains a zip file with the developed Global Oil & Gas Features Database (as an ArcGIS geodatabase). Access the technical report describing how this database was produced using the following link: <https://edx.netl.doe.gov/dataset/development-of-an-open-global-oil-and-gas-infrastructure-inventory-and-geodatabase>



Acknowledgements:
This work was performed under a CRADA between NETL and EDF, and was funded under the Climate and Clean Air Coalition (CCAC) Oil and Gas Methane Science Studies. The studies are managed by United Nations Environment in collaboration with the Office of the Chief Scientist, Steven Hamburg of the Environmental Defense Fund. Funding was provided by the Environmental Defense Fund, OGCI Companies (Shell, BP, ENI, Petrobras, Repsol, Total, Equinor, CNPC, Saudi Aramco, Exxon, Oxy, Chevron, Pemex) and CCAC.

Followers: 1
Follow

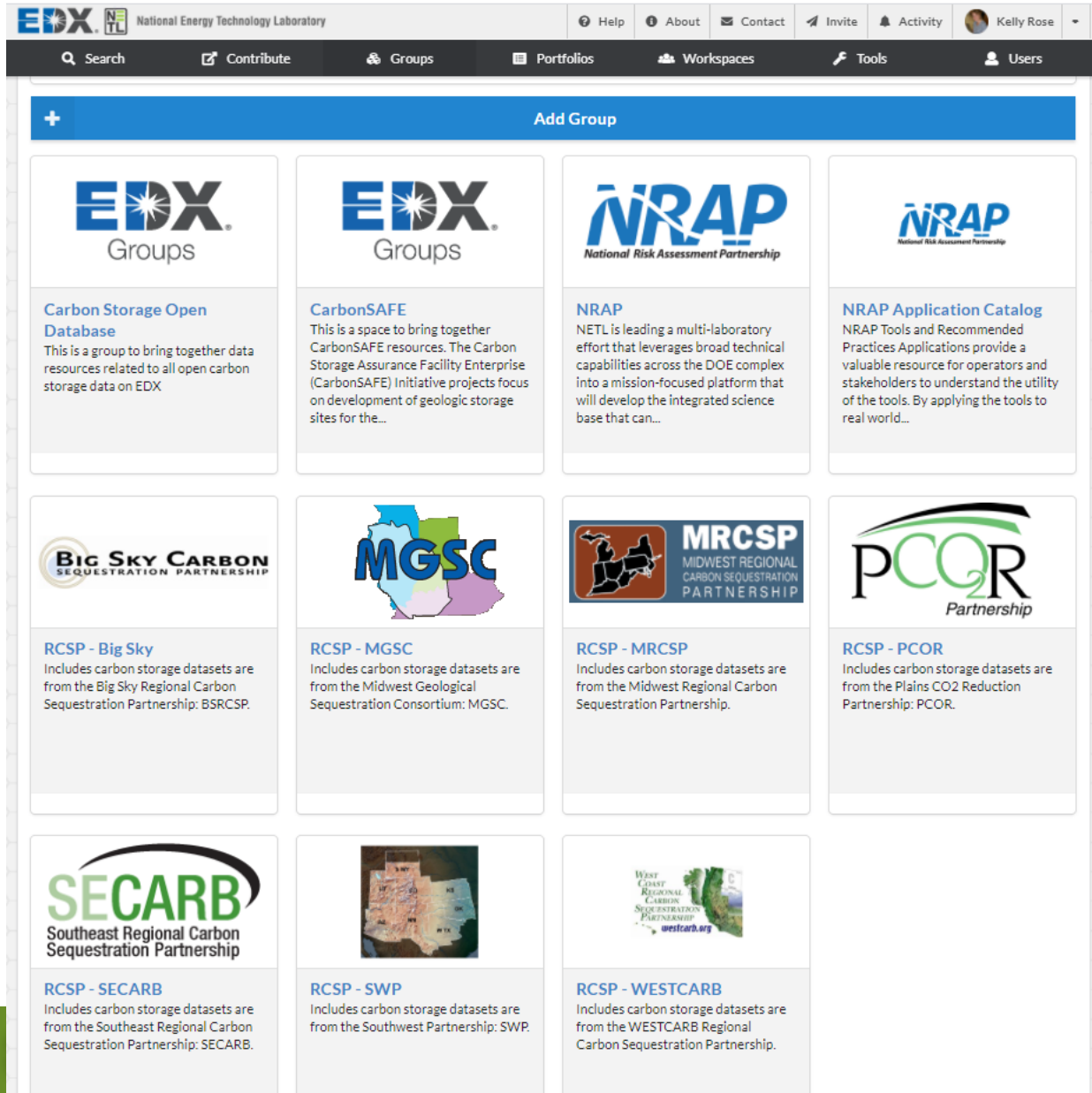
Citation (Click to Copy)
Sabbatino, M., Romeo, L., Baker, V., Bauer, J., Barkhurst, A., Bean, A., DiGiulio, J., Jones, K., Jones, T.J., Justman, D., Miller III, R., Rose, K., and Tong, A., Global Oil & Gas Features Database, 2017-12-12, <https://edx.netl.doe.gov/dataset/global-oil-gas-features-database>, DOI: 10.18141/1427300

Data and Resources
Download Checked Check All

Filter resources by name... [Search Icon] Date: Newest → Oldest

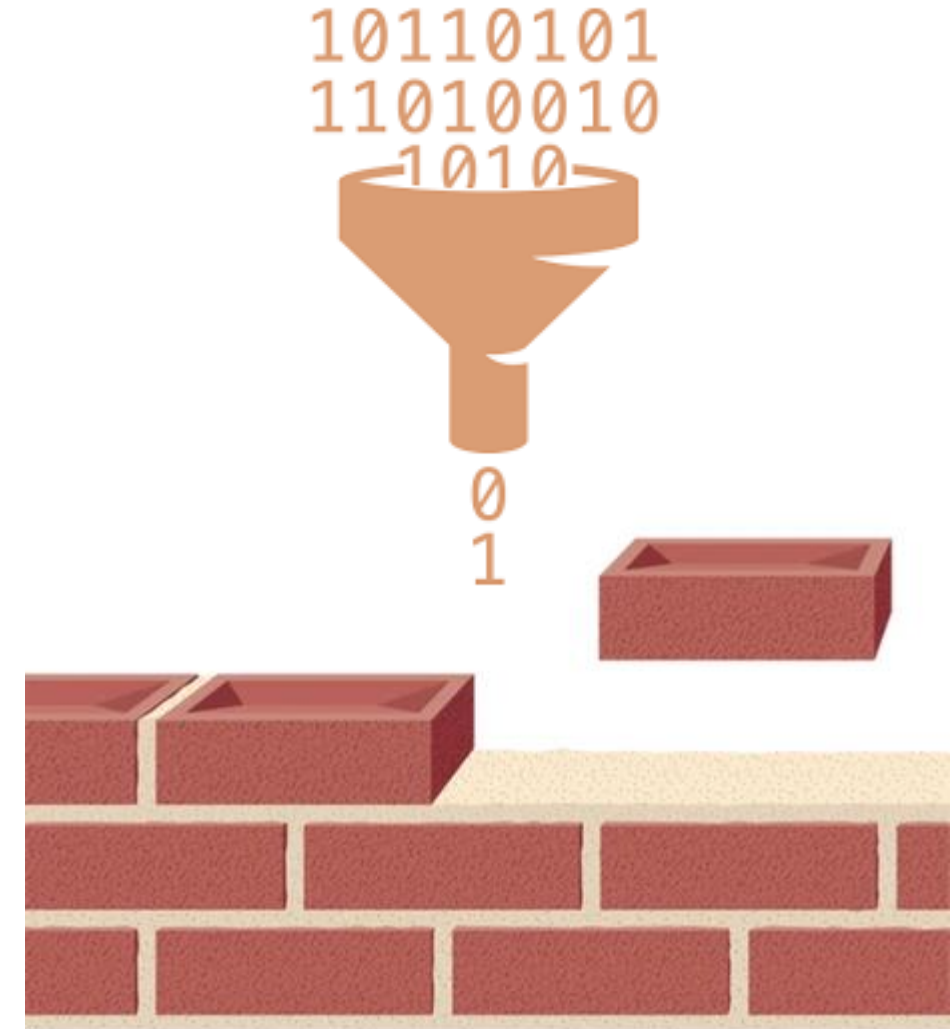
<input type="checkbox"/>	 GOGI_V10_2.gdb.zip Creative Commons Attribution	Preview or Download
<input type="checkbox"/>	 GOGI_V10_2SHP.zip Creative Commons Attribution	Preview or Download

Building a strong CS data foundation takes a community



The screenshot shows the EDX Groups page with a navigation bar at the top containing links for Search, Contribute, Groups, Portfolios, Workspaces, Tools, and Users. The main content area is titled "Add Group" and displays a grid of 12 groups, each with a logo and a brief description:

- EDX Groups**: Carbon Storage Open Database. This is a group to bring together data resources related to all open carbon storage data on EDX.
- EDX Groups**: CarbonSAFE. This is a space to bring together CarbonSAFE resources. The Carbon Storage Assurance Facility Enterprise (CarbonSAFE) Initiative projects focus on development of geologic storage sites for the...
- NRAP**: National Risk Assessment Partnership. NETL is leading a multi-laboratory effort that leverages broad technical capabilities across the DOE complex into a mission-focused platform that will develop the integrated science base that can...
- NRAP**: NRAP Application Catalog. NRAP Tools and Recommended Practices Applications provide a valuable resource for operators and stakeholders to understand the utility of the tools. By applying the tools to real world...
- BIG SKY CARBON SEQUESTRATION PARTNERSHIP**: RCSP - Big Sky. Includes carbon storage datasets are from the Big Sky Regional Carbon Sequestration Partnership: BSRCSP.
- MGSC**: RCSP - MGSC. Includes carbon storage datasets are from the Midwest Geological Sequestration Consortium: MGSC.
- MRCSP**: MIDWEST REGIONAL CARBON SEQUESTRATION PARTNERSHIP. RCSP - MRCSP. Includes carbon storage datasets are from the Midwest Regional Carbon Sequestration Partnership.
- PCQR**: Partnership. RCSP - PCOR. Includes carbon storage datasets are from the Plains CO2 Reduction Partnership: PCOR.
- SECARB**: Southeast Regional Carbon Sequestration Partnership. RCSP - SECARB. Includes carbon storage datasets are from the Southeast Regional Carbon Sequestration Partnership: SECARB.
- SWP**: Southwest Partnership: SWP. RCSP - SWP. Includes carbon storage datasets are from the Southwest Partnership: SWP.
- WESTCARB**: WESTCARB Regional Carbon Sequestration Partnership. RCSP - WESTCARB. Includes carbon storage datasets are from the WESTCARB Regional Carbon Sequestration Partnership.



The EDX Carbon Storage FAIR Timeline



2016

Decision to curate
public carbon storage
data

2016: CS program increased investment in:

- Curating FECM CS Program data products
- AI/ML tools to explore and transform data
- Integrate data into databases with FAIR standards

Data collection and assimilation:

- Manual data searches
- Automated data searches using machine learning methods

2016-Present

RCSPs contribute data to EDX
and push public

2018- 2019: Data preservation
for WESTCARB

Database development

- Data catalog development
- Metadata preservation

Natural language processing:

- Topic modeling
- Keyword identification
- Named entity recognition geotagging

Implement machine learning and natural language processing to:

2017: Virtual Sub Surface first envisioned and proposed

2017: SmartSearch, find and integrate data at scale

2019: Natural Language Processing labeling and topic modeling

2019: SmartParse, in development, created

2019: Living Database development

2020: Geotagging development

Final preparation of data for publishing:

- Integration of NLP results with metadata
- Data cleaning and preparation for upload
- Quality assessment

Ongoing efforts to curate and catalog data:

- NRAP data catalog
- Development of Carbon Storage Open Data Catalog and Database
- Development of Groups for data curation on EDX

Data publishing
on EDX and GeoCube

Results in a world-class foundation of CS resources:

- Curation of thousands of CS data products, models, and tools
- Integration of data into more reuseable, integrated collections for reuse
- Continually updated collections

2021

Creation of
model outputs

Data used for
CCS risk modeling

Data used for
CCS site screening

It started with the RCSP Phase 2



Regional Carbon Sequestration Partnerships (RCSPs)

- RCSP **public and private** resources have a combined total of 3,065 resources and 1.72TB of data
- **Publicly** released RCSP assets total almost 1 TB and over 1165 resources

RCSP EDX Group	Submissions	Resources	Data Usage (in GBs)
RCSP-Big Sky	42	136	601.07
RCSP-MGSC	18	61	99.22
RCSP-MRCSP	86	245	108
RCSP-PCOR	137	547	8.3
RCSP-SECARB	25	63	31.7
RCSP-SWP	11	33	3.7
RCSP-WESTCARB	8	70	135
RCSP TOTALS	327	1165	986.99

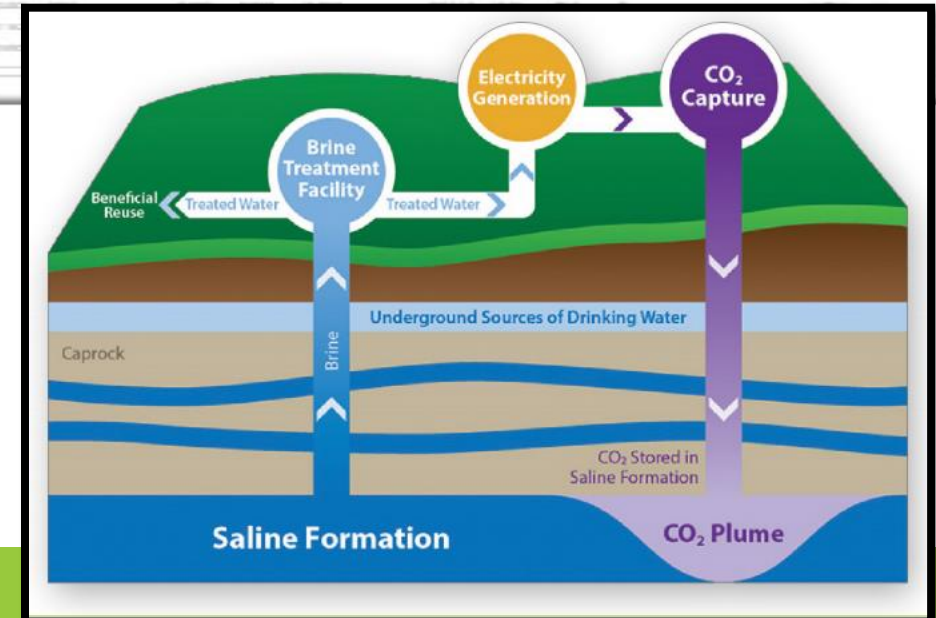
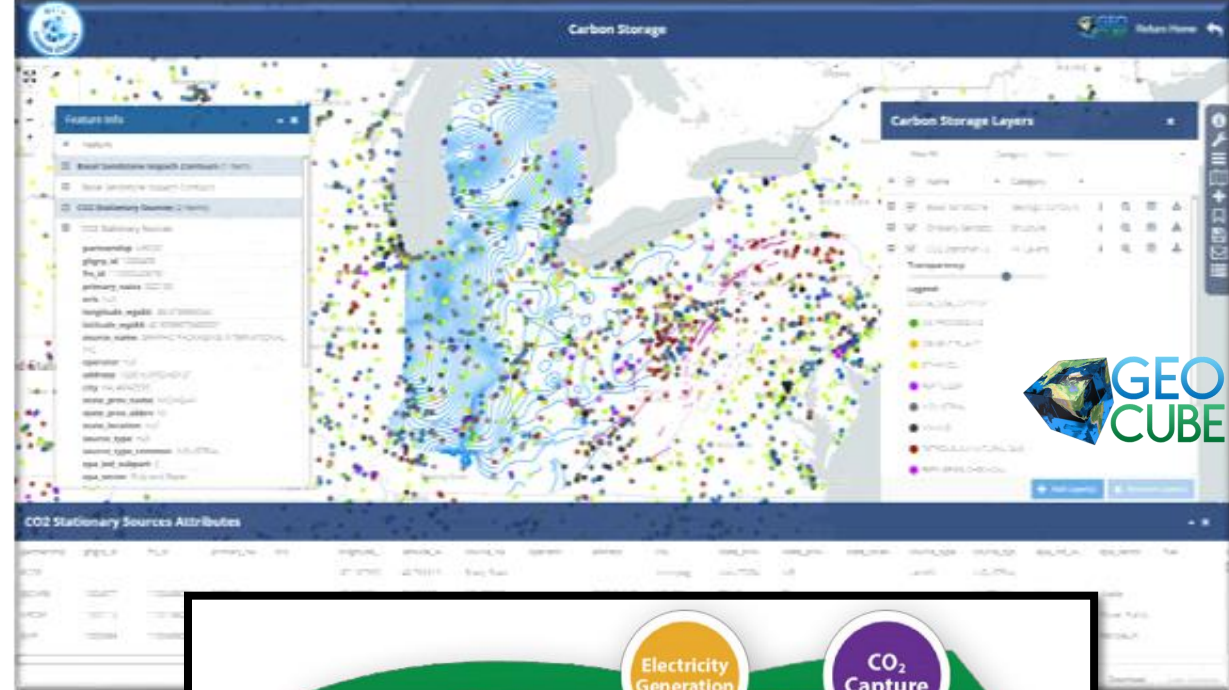
The screenshot shows the EDX interface for the SECARB group. The top navigation bar includes links for Search, Contribute, Groups, Portfolios, Workspaces, Tools, and Users. The main content area displays a list of 25 submissions. Each submission entry includes a title, a brief description, a dataset size, the number of resources, and a download button. The submissions are categorized by type (Dataset, Publication) and format (ZIP, PDF, XLSX, XML, ZIP, SHAPEFILE).



<https://edx.netl.doe.gov/group>

Carbon Storage Open Data Collection

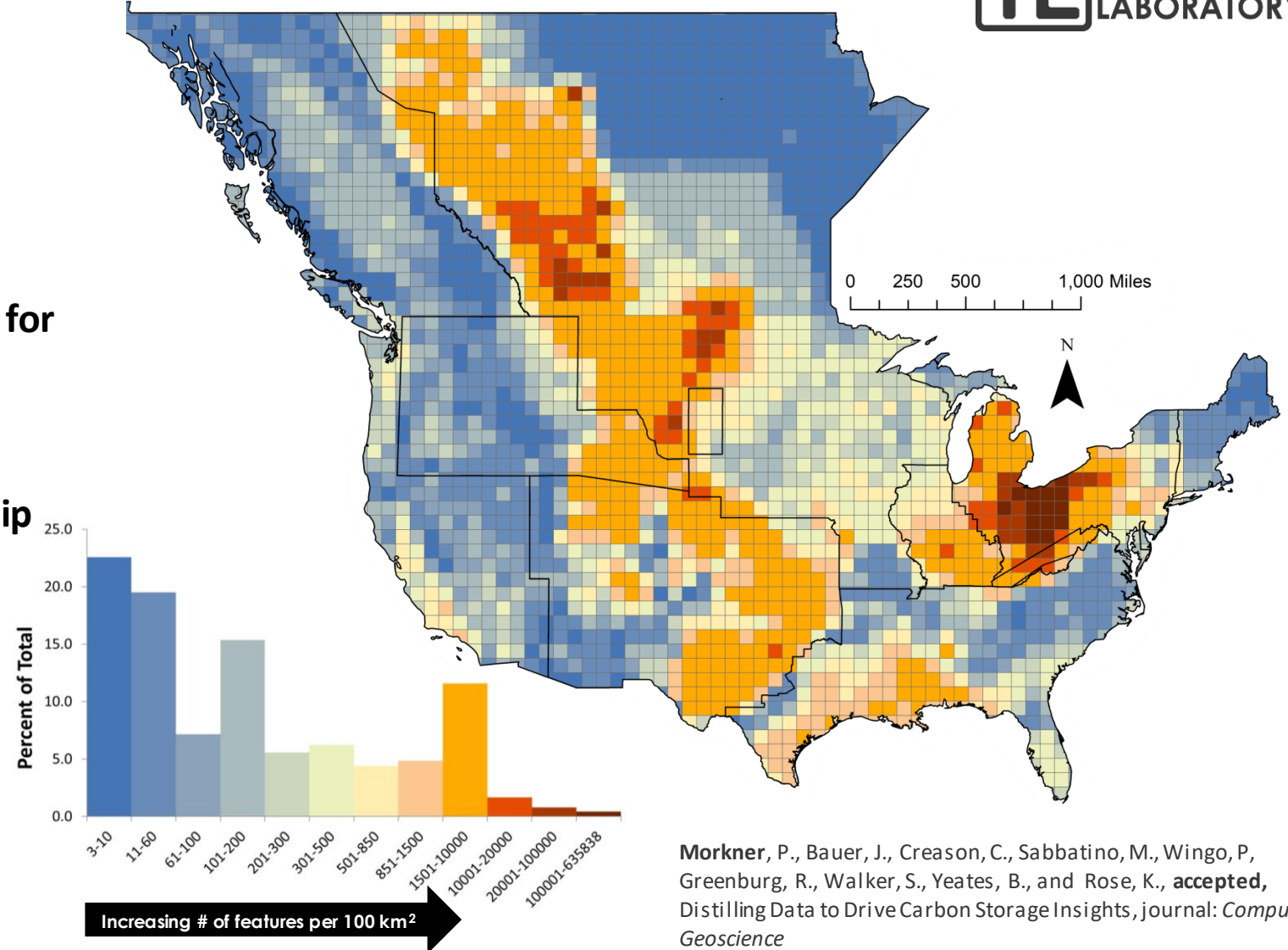
- **Current collection:**
 - 315+ spatial data layers on GeoCube
 - 1800+ text-based documents on EDX
 - These are in addition to available RCSP, CarbonSAFE, FutureGen and other data products also on EDX
- **Ongoing effort this year to update with more resources:**
 - 215 spatial data layers from EDX submissions
 - 12 outside datasets to target for **API connectivity**
 - EIA Emissions Data
 - EPA CO2 point sources
 - USGS National wells database and groundwater aquifer database
 - USGS Produced Waters database
 - USDW wells and outlines
 - EPRI subsurface models
- **Use of SmartSearch to find additional geospatial data resources**
 - Class VI permit data, surface data that would limit CCS, etc.



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
 - **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., **accepted**, Distilling Data to Drive Carbon Storage Insights, journal: *Computers & Geoscience*

CS Data Decade Summary

- 10+ years of development of EDX
- 5+ years of Carbon Storage program investments into data management
- Development of AI/ML tools to support CS data community needs
- Preserved thousands of products, from millions of dollars of research

This has led to:

- 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



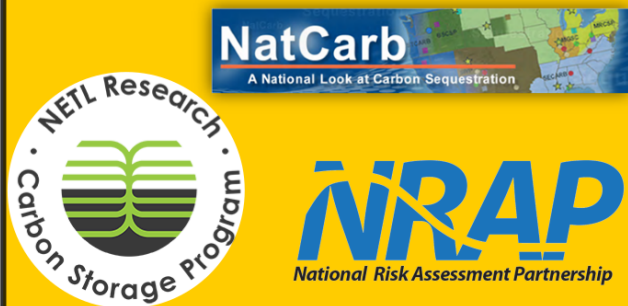
FECM's CS FAIR Data Ecosystem

Data, tools, models, computing

Carbon storage data computing resources

Data discovery, acquisition, movement, storage

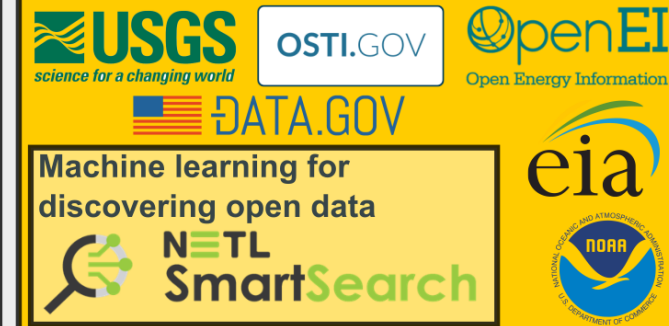
DOE funded project data



Data curation, publishing,
preservation



Collection of outside-EDX resources



Data integration, exploration, curation, analysis, optimization

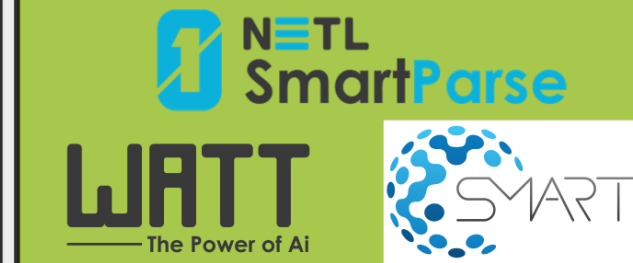
Data exploration and curation

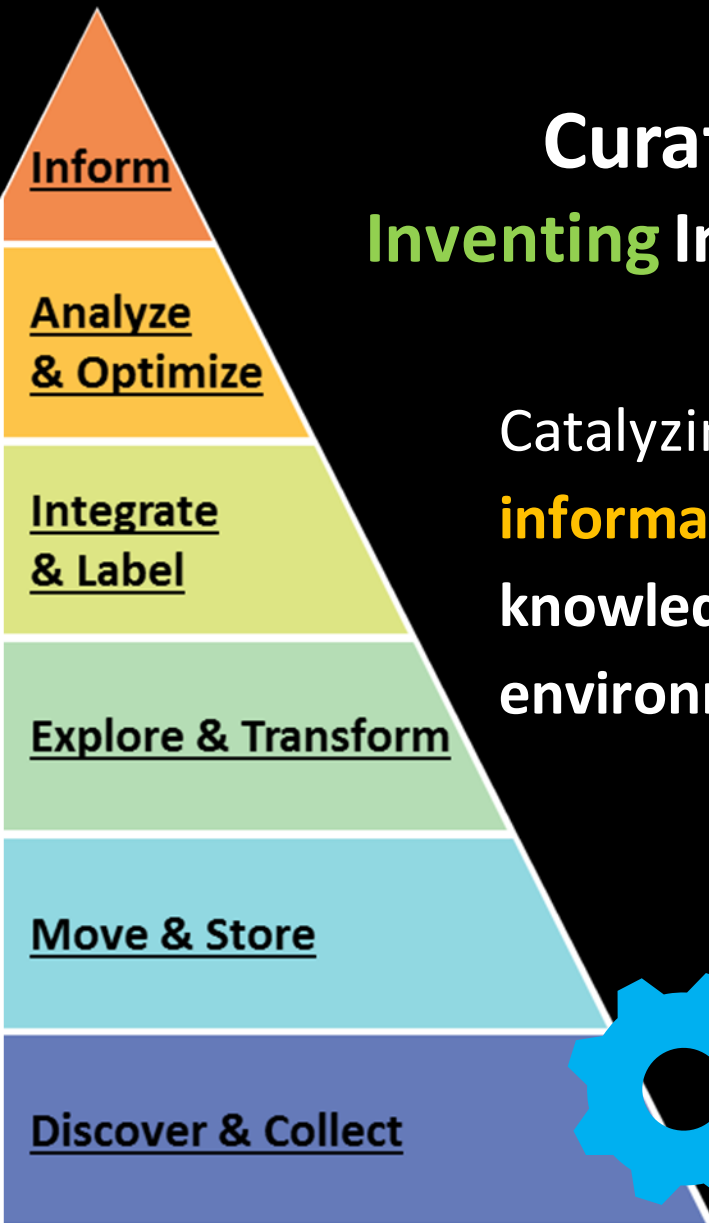


Tools



Machine learning &
natural language processing

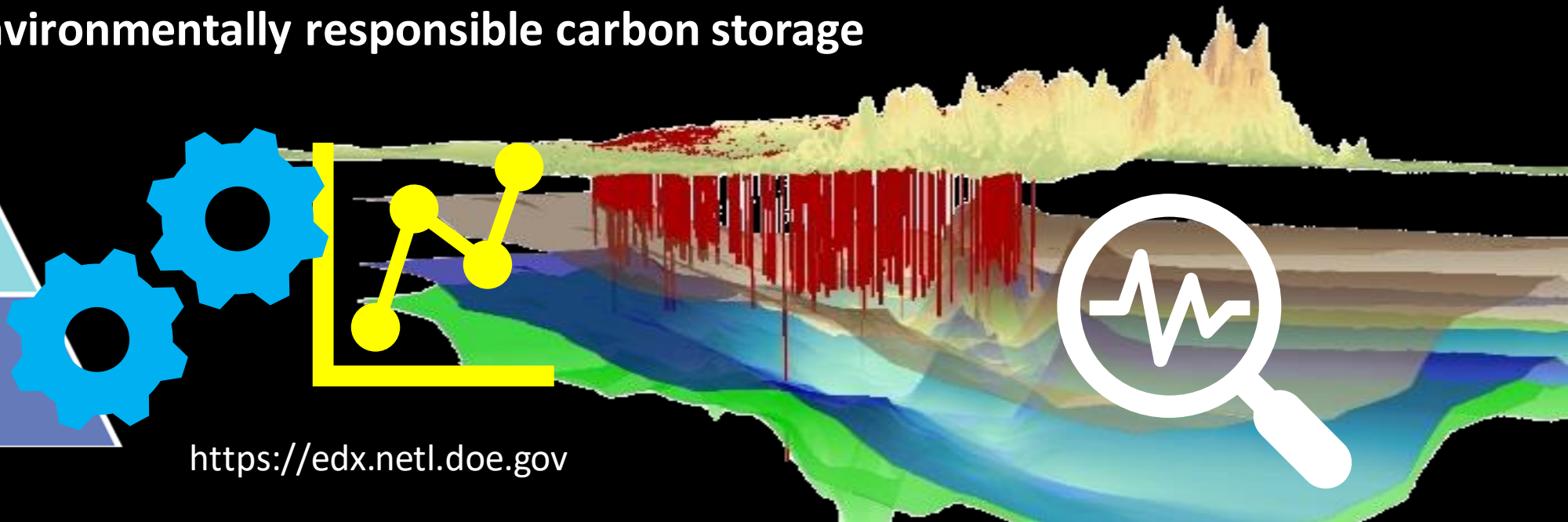




Curating products & transforming for reuse

Inventing Intelligent Solutions to Carbon Storage R&D Needs

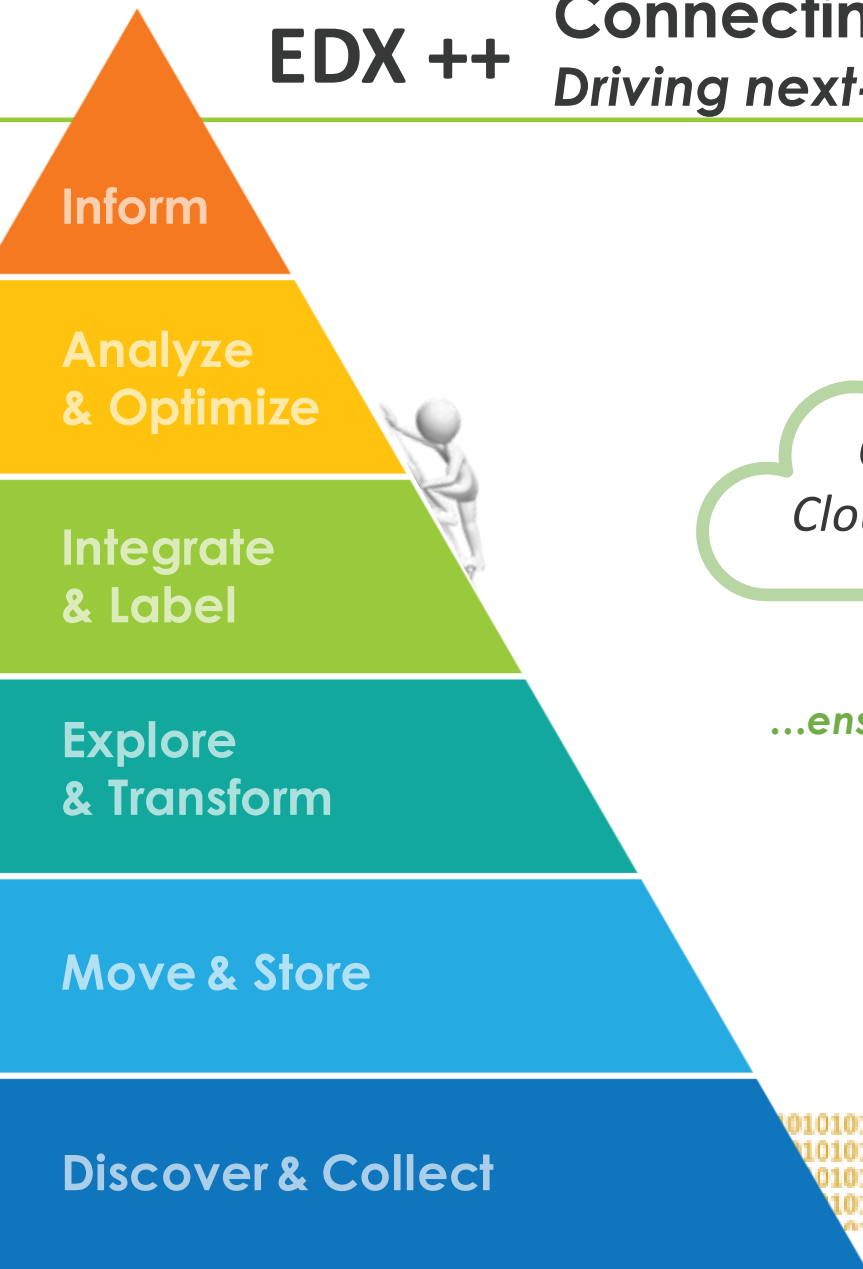
Catalyzing next-generation R&D to rapidly produce the most **up-to-date information** and produce **cutting-edge tools and models** to extract knowledge and offer insights to enable safe, affordable, and environmentally responsible carbon storage



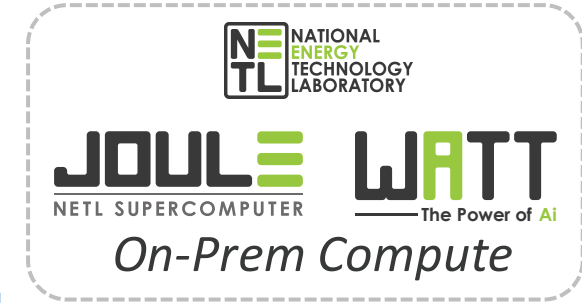
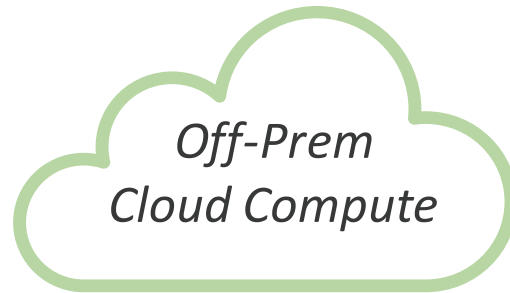
<https://edx.netl.doe.gov>

EDX ++ Connecting data with scientific computing

Driving next-gen R&D...



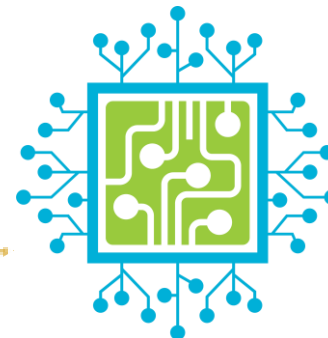
EDX++ FRAMEWORK



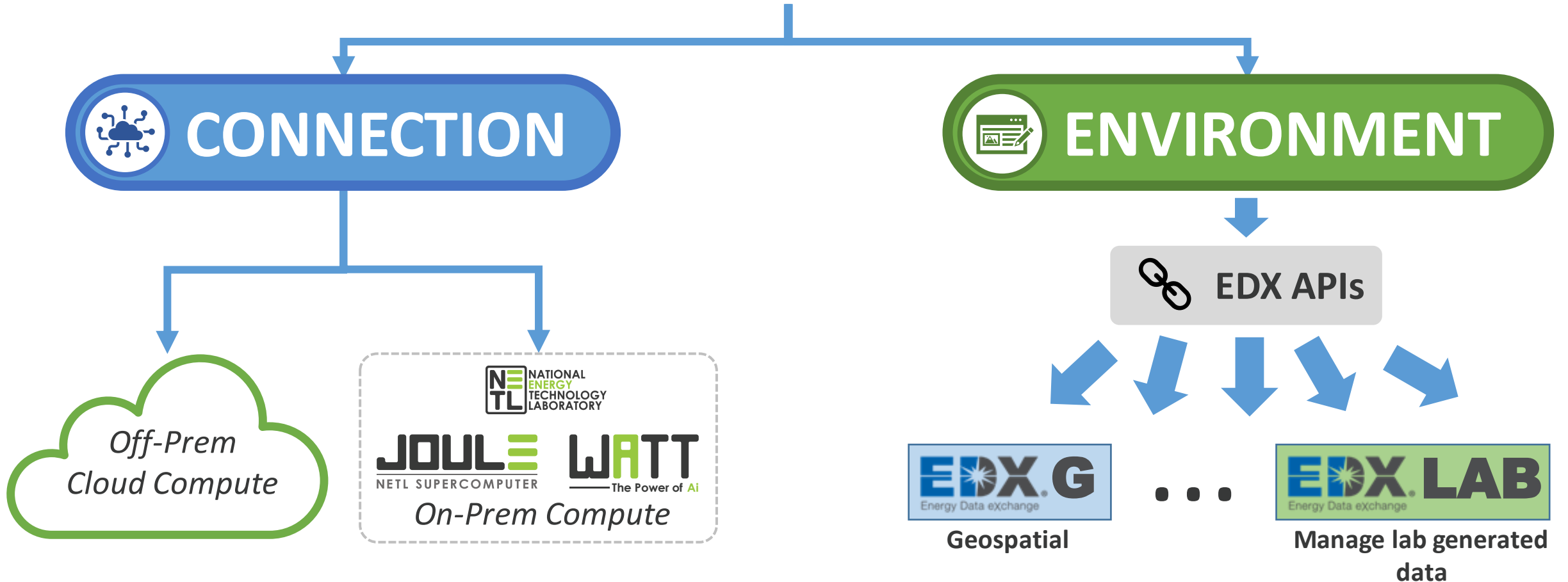
...ensuring compliance
with Federal/DOE
regulations



...ensuring preservation and
access to DOE FE knowledge
and data resources



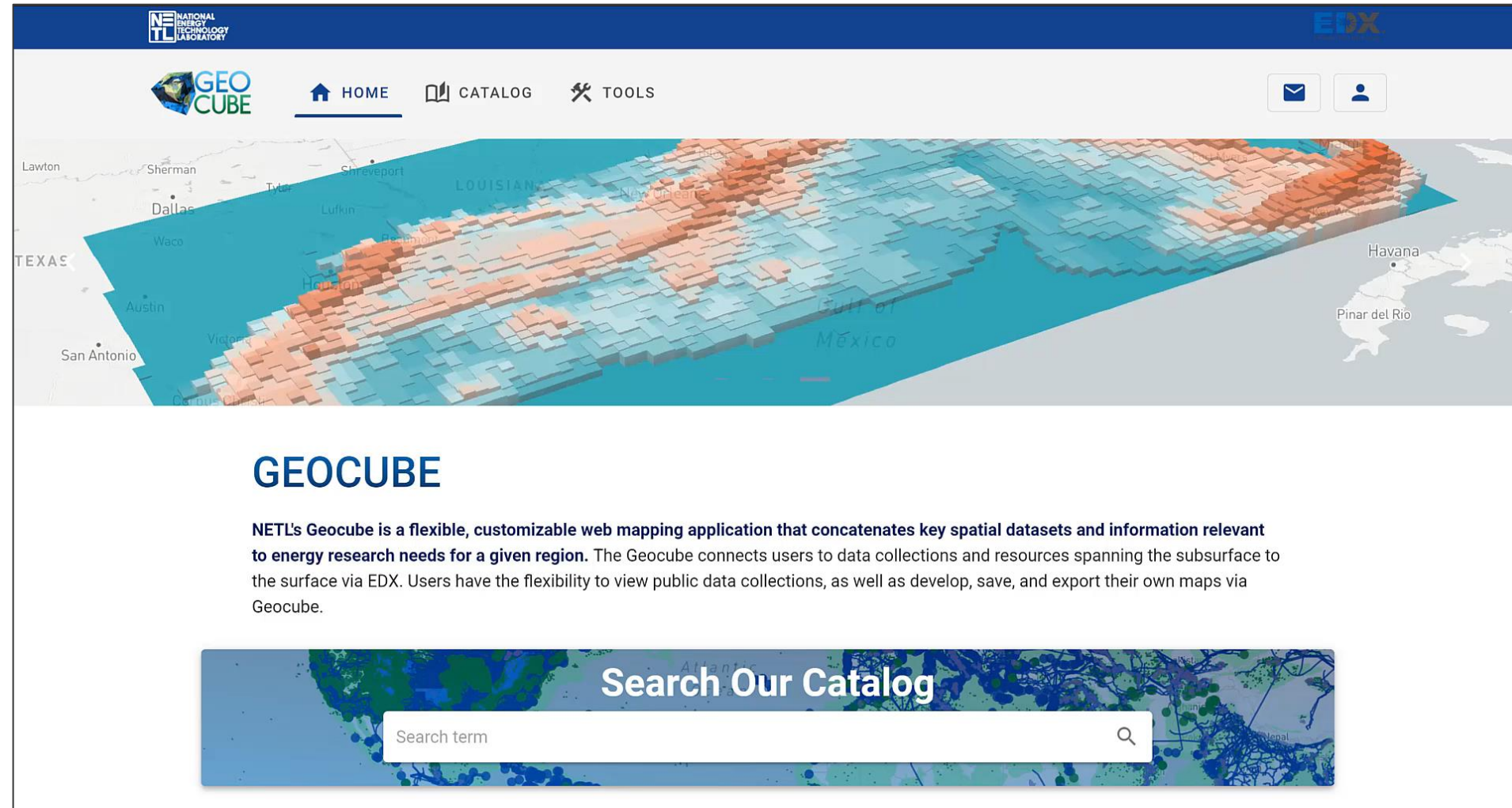
Big Data



WHAT IS YOUR DATA NEED?

EDX-G in Alpha Testing, *Coming Soon to EDX*

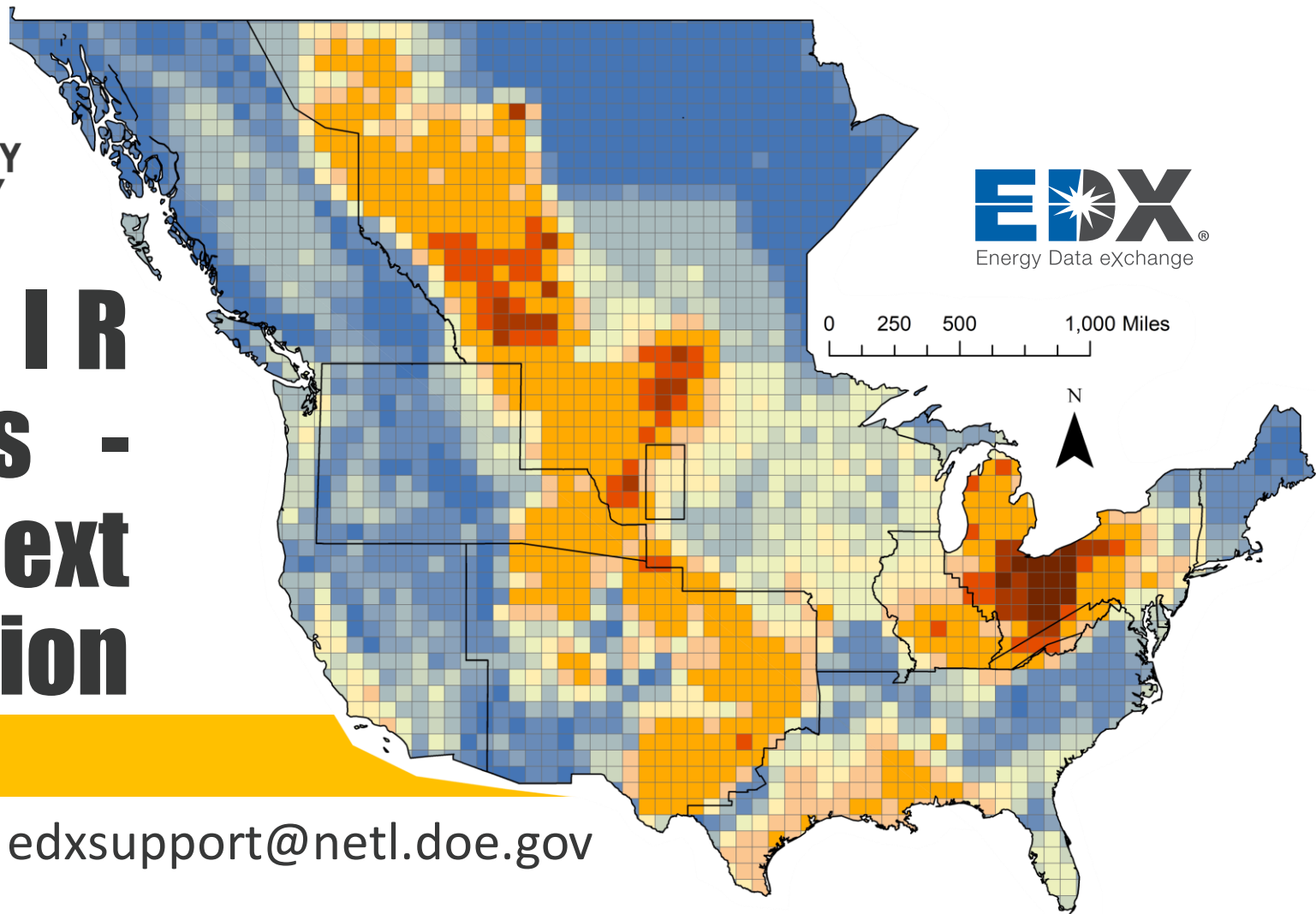
- Support discovery, access **and use** of geospatial data & analytical tools
- New user interface in **beta testing, offers:**
 - Enhanced spatial search capabilities
 - **Virtual access to CS tools, such as,**
 - CO2-SCREEN
 - Offshore CO2 Saline Calculator
 - New tool page featuring links to CCS relevant tools –
 - NRAP
 - SimCCS
 - etc



Video showing upcoming release with enhanced user interface for GeoCube, as well as expanded functionality with new data and tools integrated from EDX



FECM CS FAIR data resources - Catalyzing the next decade of innovation



Contact: edxsupport@netl.doe.gov

Contribute & Find Carbon Storage R&D Products At:

<https://edx.netl.doe.gov/>

Thanks to the FECM CS Program, EDX Dev Team, GAIA Research Group at NETL, and the numerous Carbon Storage Program contributors who were key and critical to the development of the CS FAIR Data Collection.