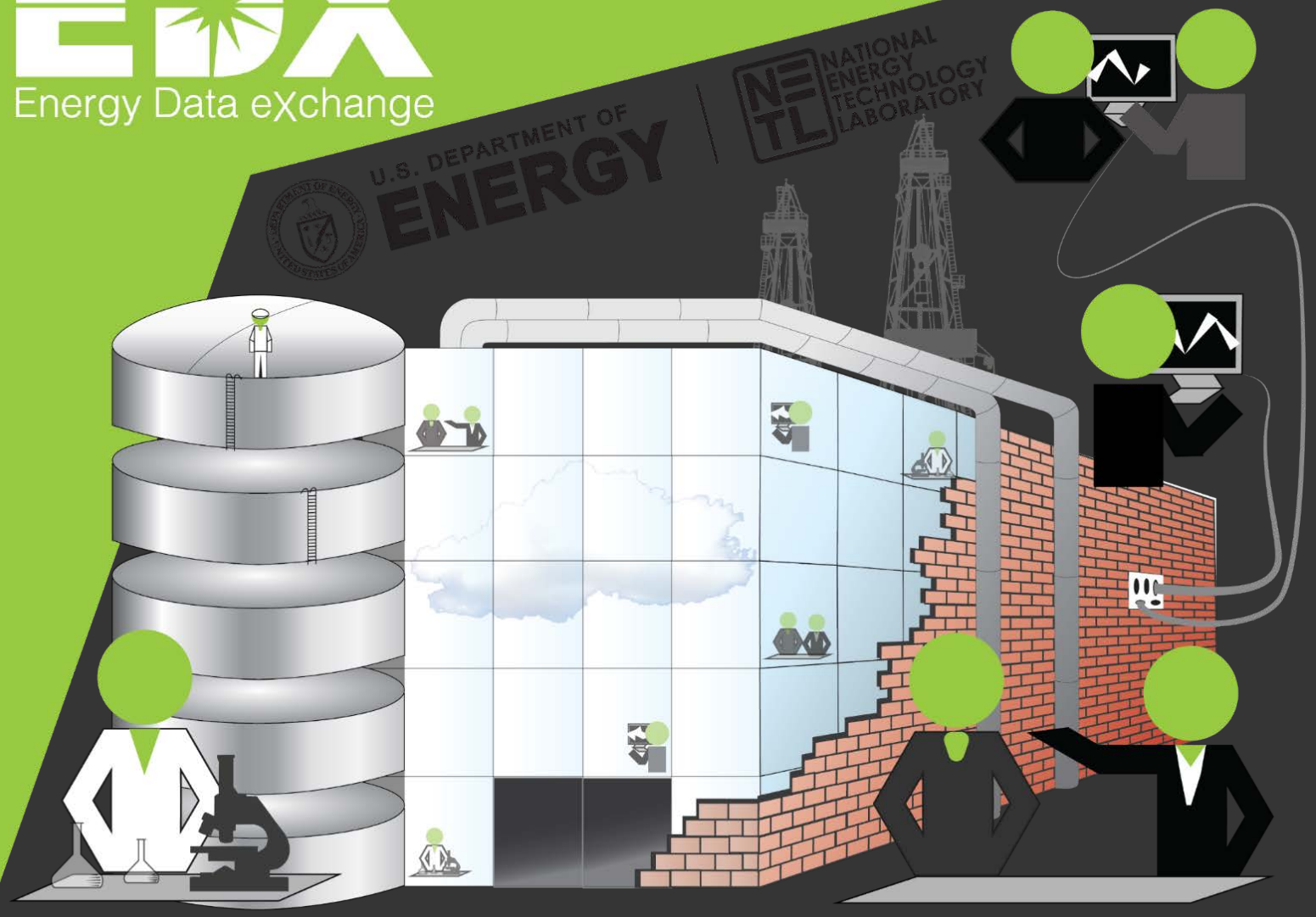
EDX
Energy Data eXchange



NATIONAL ENERGY TECHNOLOGY LABORATORY

# EDX-Natcarb

## A *Virtual* Data Library & Laboratory for Carbon Storage Science


**Kelly Rose[1], Vic Baker[2], Jenny Digiulio[3,1], TJ Jones[2], Michael Sabbatino[3,1], Alex Tong[1,4], Patrick Wingo[3,1]**

[1]National Energy Technology Laboratory, [2]MATRIC, [3]AECOM, [4]ORISE


**August 2017**

**Solutions for Today | Options for Tomorrow**

U.S. DEPARTMENT OF ENERGY

# Current project objectives

- **Support development and update of two geologic data systems for CS/SubTER R&D:**
  - National Carbon Sequestration Database (NATCARB) and EDX, are being used to integrate public data as an internal research tool for $CO_2$ storage site characterizations and resource assessments
- **Support EDX and NATCARB growth to include results from the Regional Partnerships and Core R&D Programs and support development of future editions of the Carbon Storage Atlas.**
  - These both focus on development and maintenance of these systems as a curation and access resource for resources used by NETL Carbon Storage and DOE FE R&D affiliated researchers as a whole.
  - Support ingestion and curation of RCSP knowledge and data products
  - Support and streamline Natcarb Atlas VI production
  - Modernize and update Natcarb Atlas tool, pair with other open data and tools to meet user needs and experience

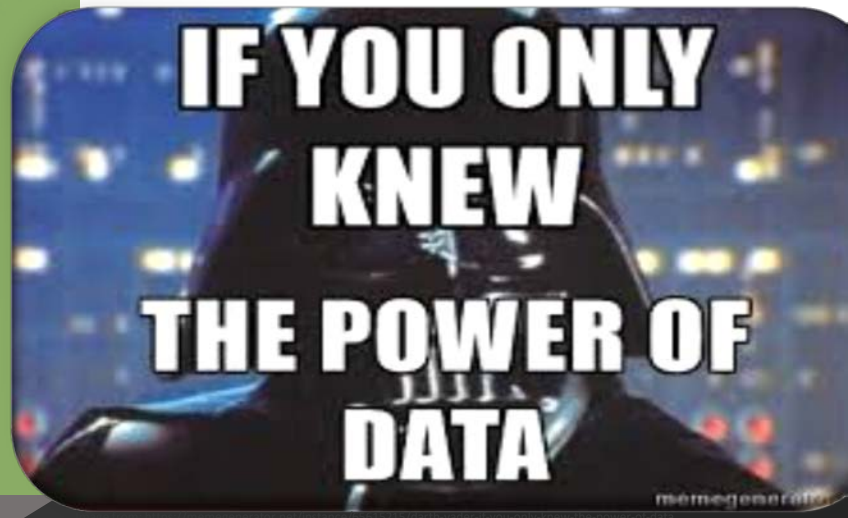
# Data are key to R&D, but access is challenging

**"The world's most valuable resource is no longer oil, but data"** -**The Economist**

**"I want you to think about data as the next natural resource"**-**Ginni Rometty**, IBM CEO


http://successflow.co.uk/blog/2015/11/27/data-is-the-new-oil-but-do-you-have-the-resources-to-refine-it/

- **Volume of data is growing:** Scientific data is projected to exceed more than 40,000 exabytes by 2020.
- **Scientists losing data at a rapid rate:** Decline can mean 80% of data are unavailable after 20 years.
- **Finding older R&D data is hard:** As published research ages, access to the underlying datasets decreases.
- **20%** of world's data are stored online while **80%** are being privately held.

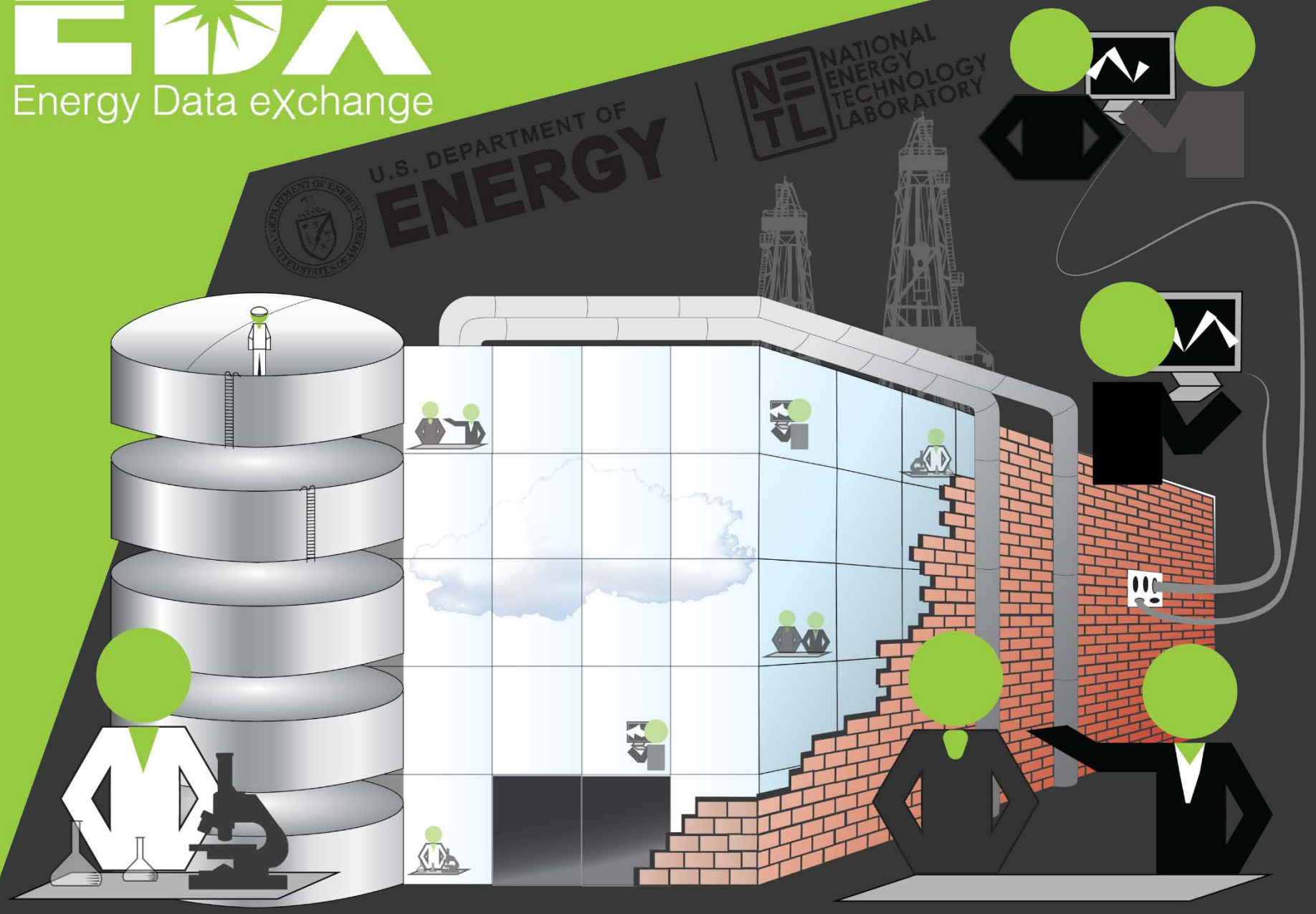



https://www.facebook.com/groups/ifyouonlyknewthepowerofdata/

# A Virtual Library & Laboratory for Energy Science

- Virtualizing team analytics
- Continued innovations to connect NETL researchers to online resources
- Increasing # of tools and apps for use in team workspaces
- **In development since 2011**

EDX
Energy Data eXchange




U.S. DEPARTMENT OF ENERGY | NATIONAL ENERGY TECHNOLOGY LABORATORY

# EDX Highlights

## Members (Internal and External to NETL)

- **Over 1,100 Registered Members** (40% NETL, 60% External Collaborators), (56% Gov, 22% Academia, 22% Private)
- An average of over **500GBs of downloads per month** since July 2016

## Published Data, Tools, Publications, and Presentations

- **Over 16,265** published data files
- **Over 327,528** resources, EDX + federated (OpenEI, NGDS, Data.gov, NOAA)
- **18 EDX Tools** in Support of Science-Based Decision Making
- **15 EDX Groups**
- **7 Research Portfolios**

## Secure, Private Collaboration

- **Over 372 Research Projects** with EDX Collaborative Workspaces
- **Over 32,000** secure, private data files

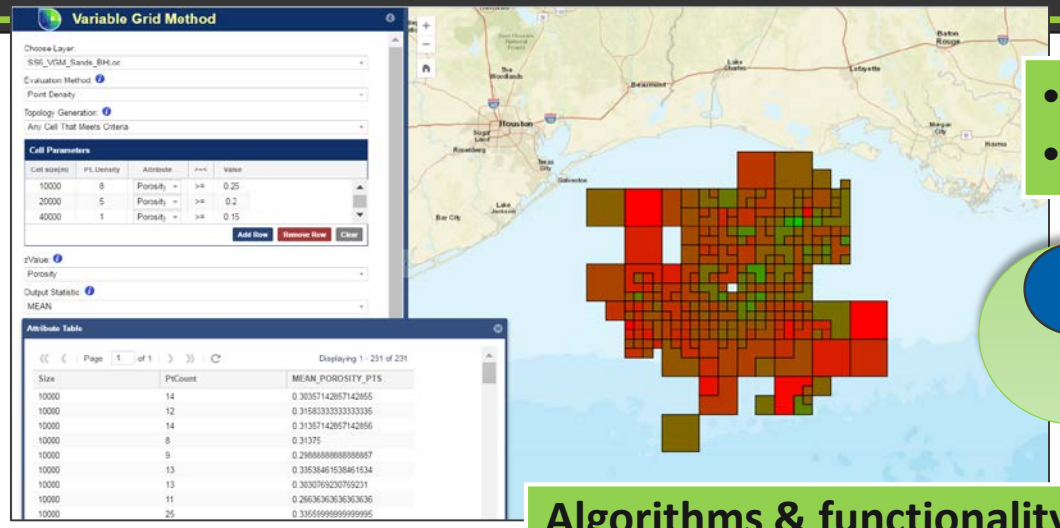
# EDX – Inventing Solutions to DOE FE Data R&D Needs

*Data Analytics*

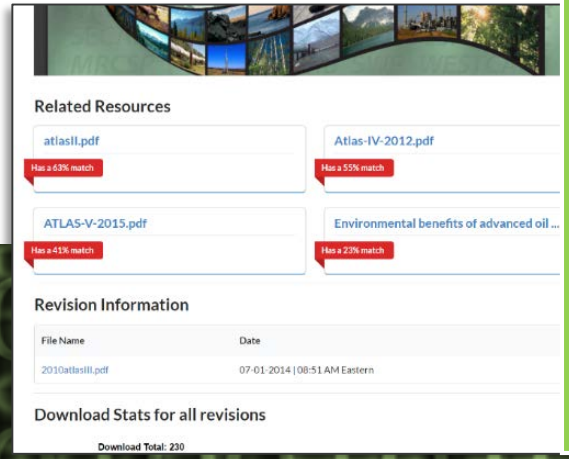
*Data Discovery*

*Describing Data*

*Curating Data*

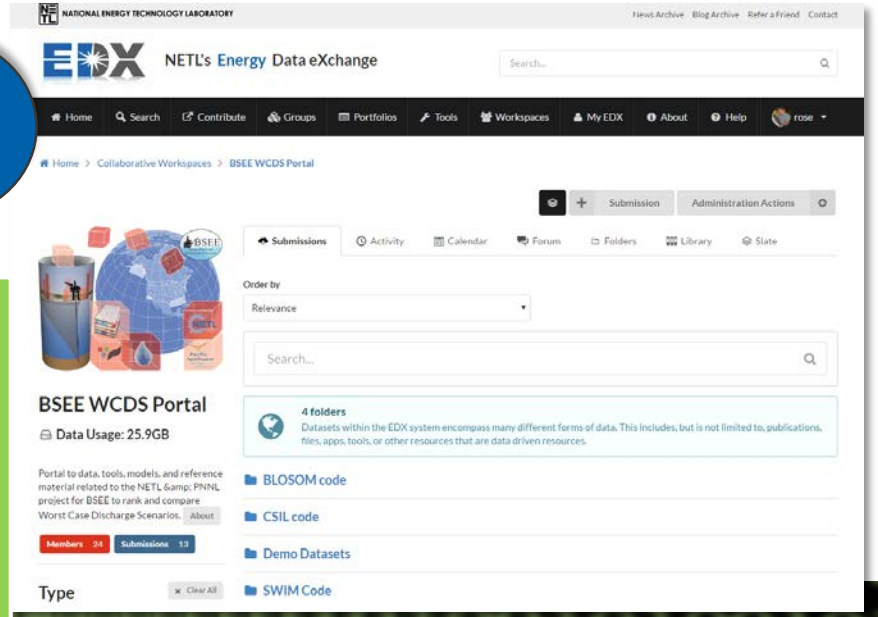



- **Secure team sharing**
- **Integrating data, tools & resources for R&D**

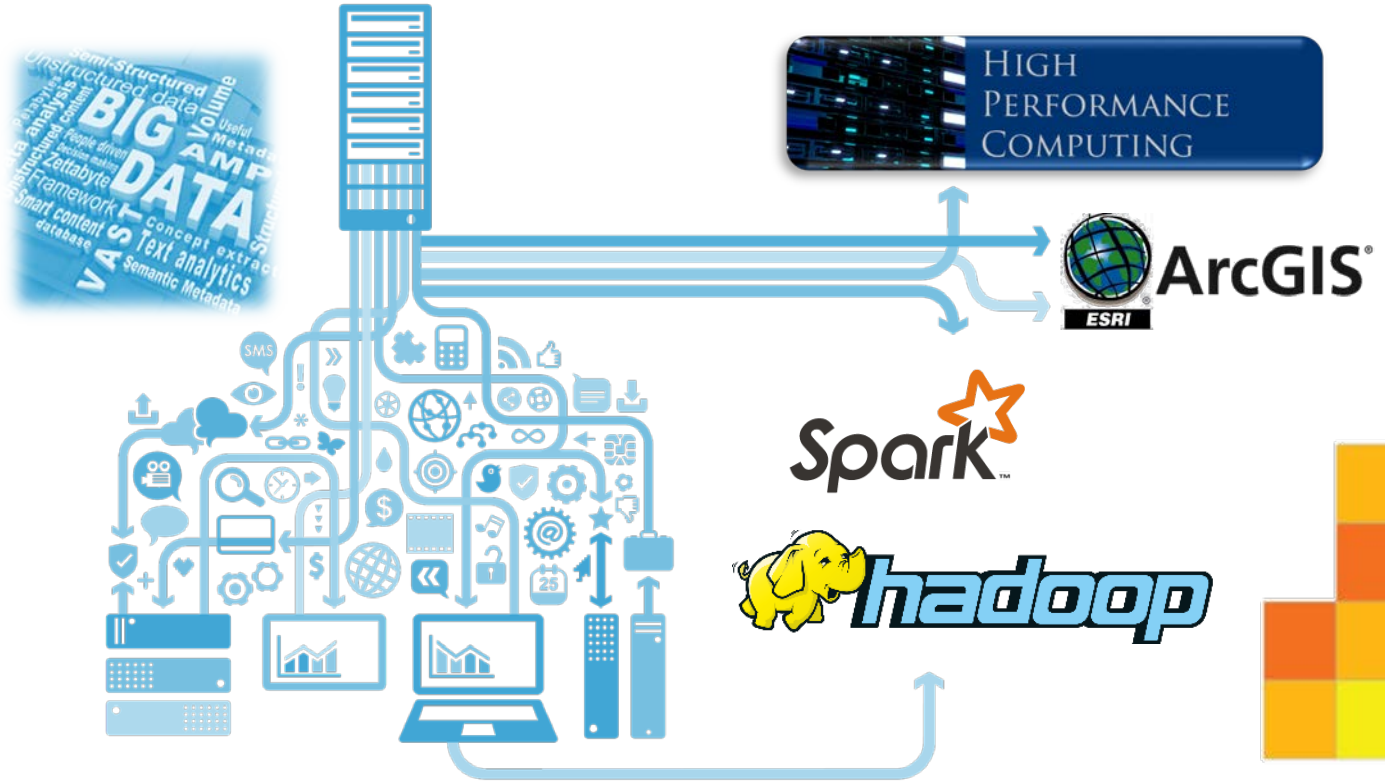



**Algorithms & functionality:**

- Custom "smart search" tool in development
- Digital spatial team "notebook"
- Auto-indexing algorithm, provides analysis of your search and helps recommend other items



# Example machine learning, big data tool for advanced FTP Data Mining: Hadoop + ESRI

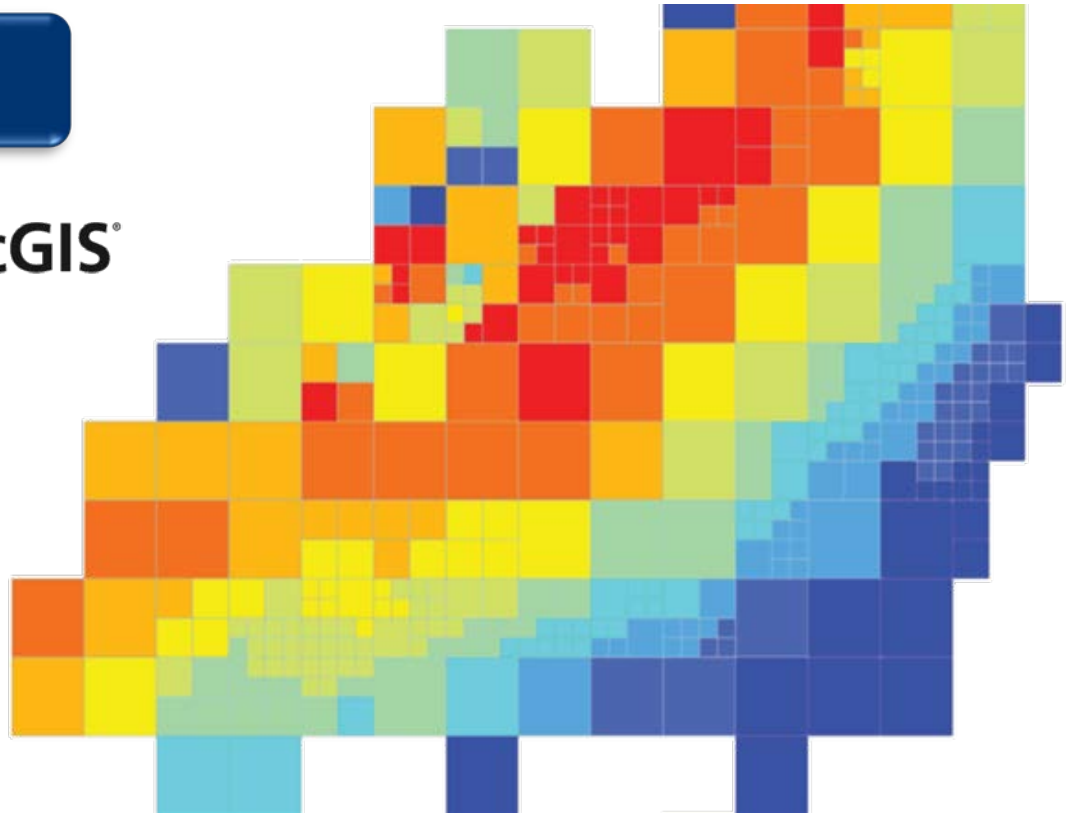



HIGH PERFORMANCE COMPUTING

ArcGIS
ESRI

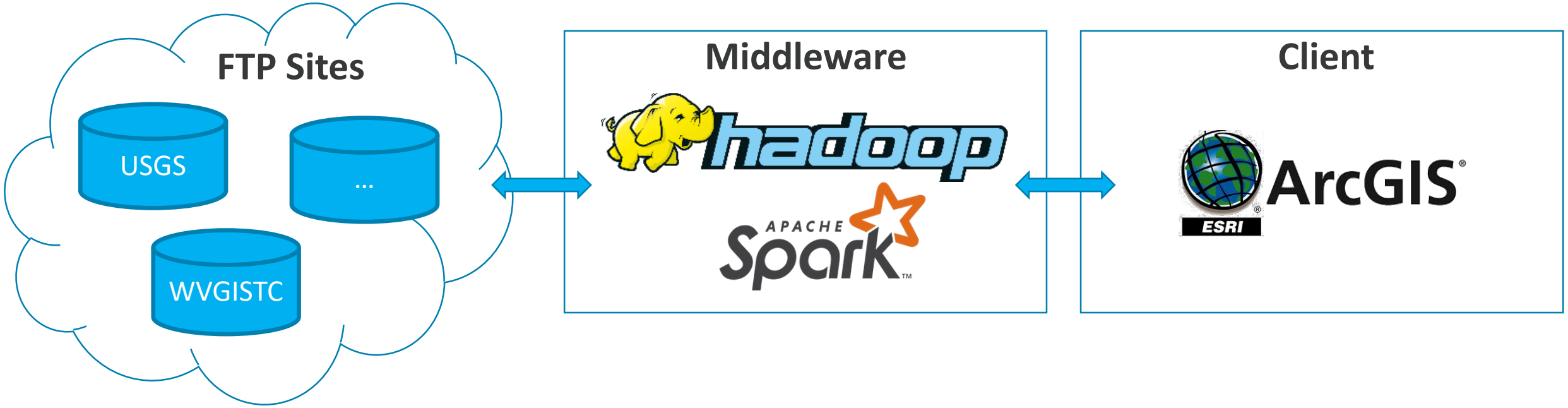
Spark

hadoop

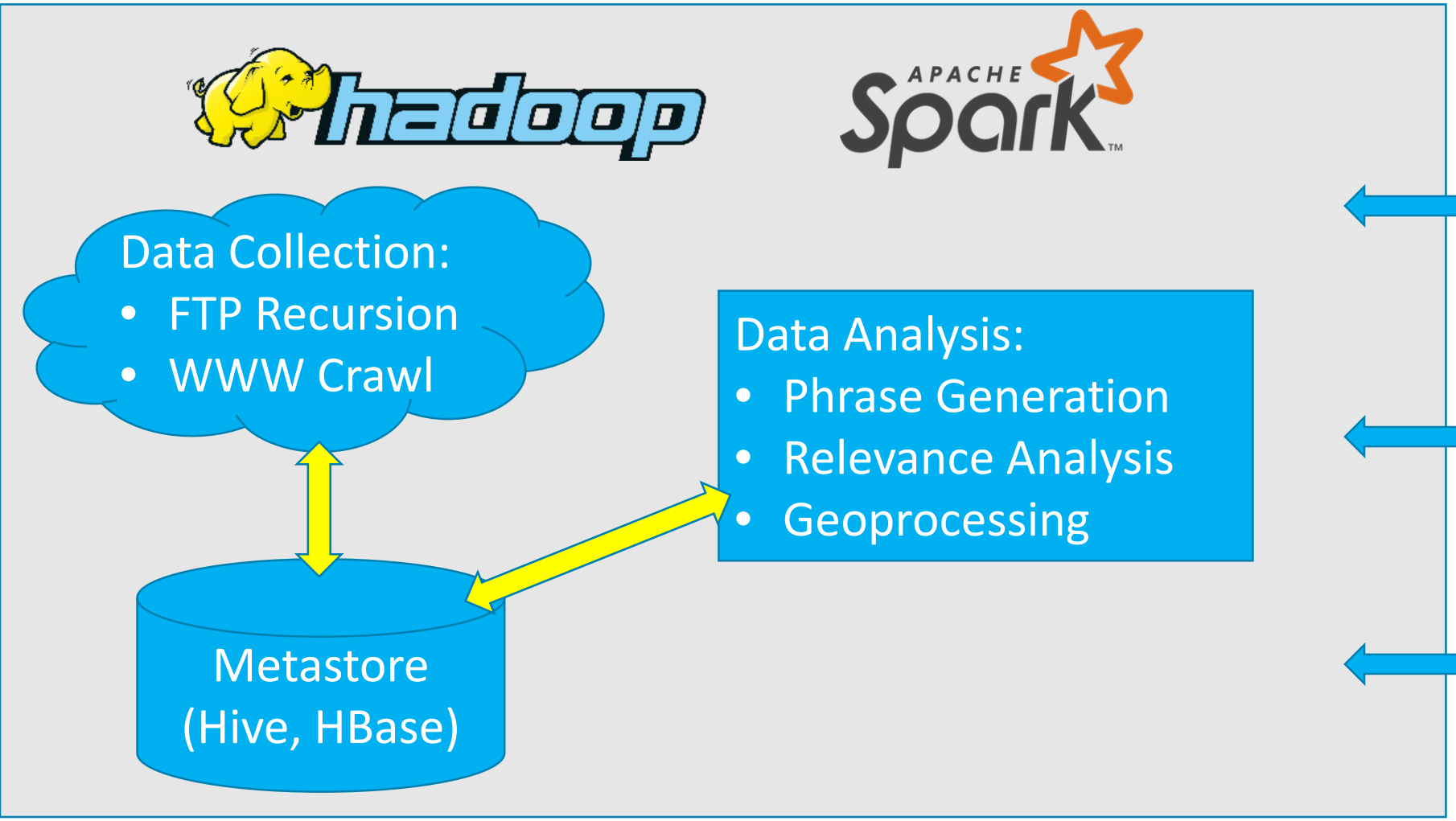


# Use Case: FTP Data Mining: Hadoop + ESRI

- **Problem:**
  - Need to search data in FTP silos (millions of files, spatial and contextual)
- **Solution:**
  - Index FTP silos using Hadoop and query using ESRI ArcMap



# NETL's Big Data Discovery Ecosystem (To Date)




## Data Mining Clients

EDX
A Product of NETL

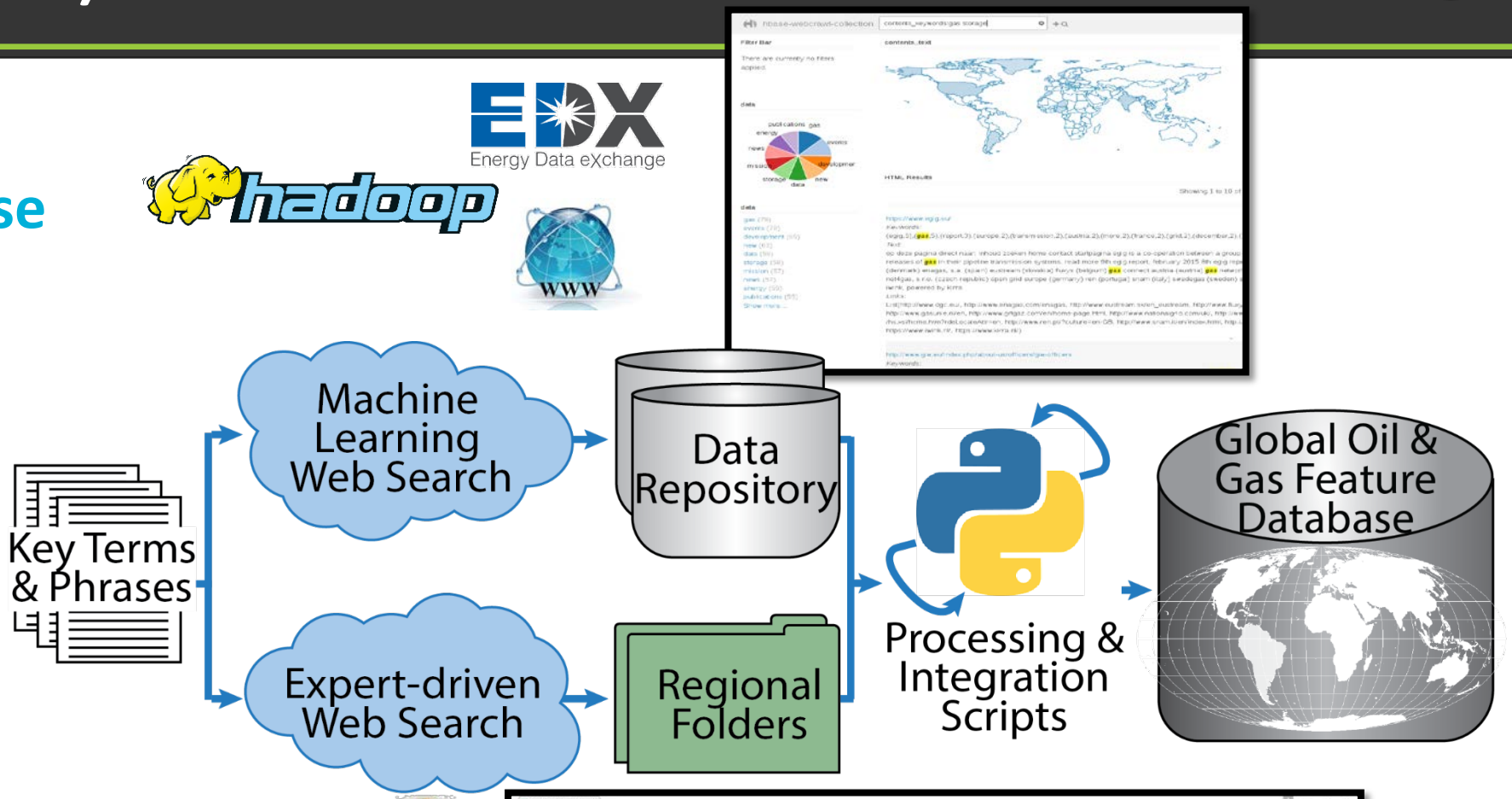
ArcGIS
ESRI

HUE

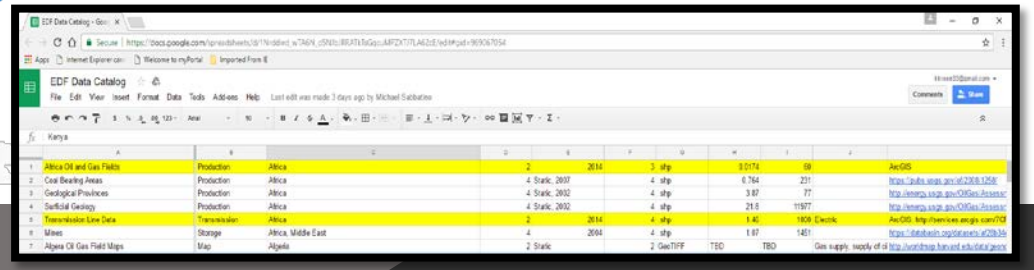
# Beyond Well Data - Building an Open Global Oil & Gas Infrastructure (GOGI) Database

## 2 methods used to produce the database over 4 months

- **Machine learning web search leveraging NETL's custom built, big data computing tool**
- **Expert drive web search to manually identify datasets**





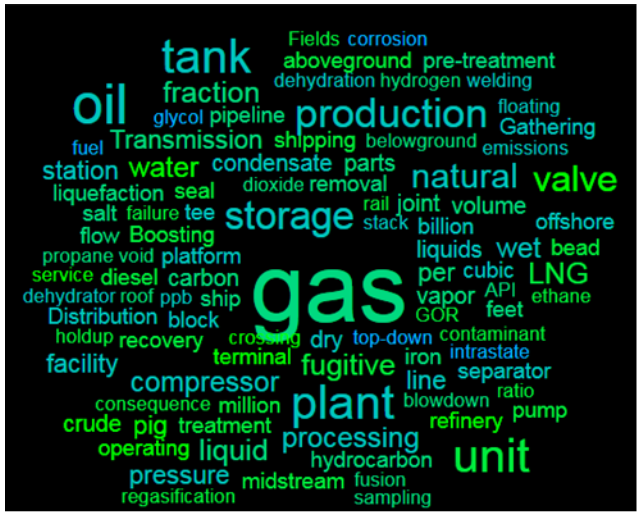
CRADA with: EDF ENVIRONMENTAL DEFENSE FUND® Finding the ways that work

# Combined these approaches resulted in:

**Acquisition of disparate data by country, region, & continent totaling:**

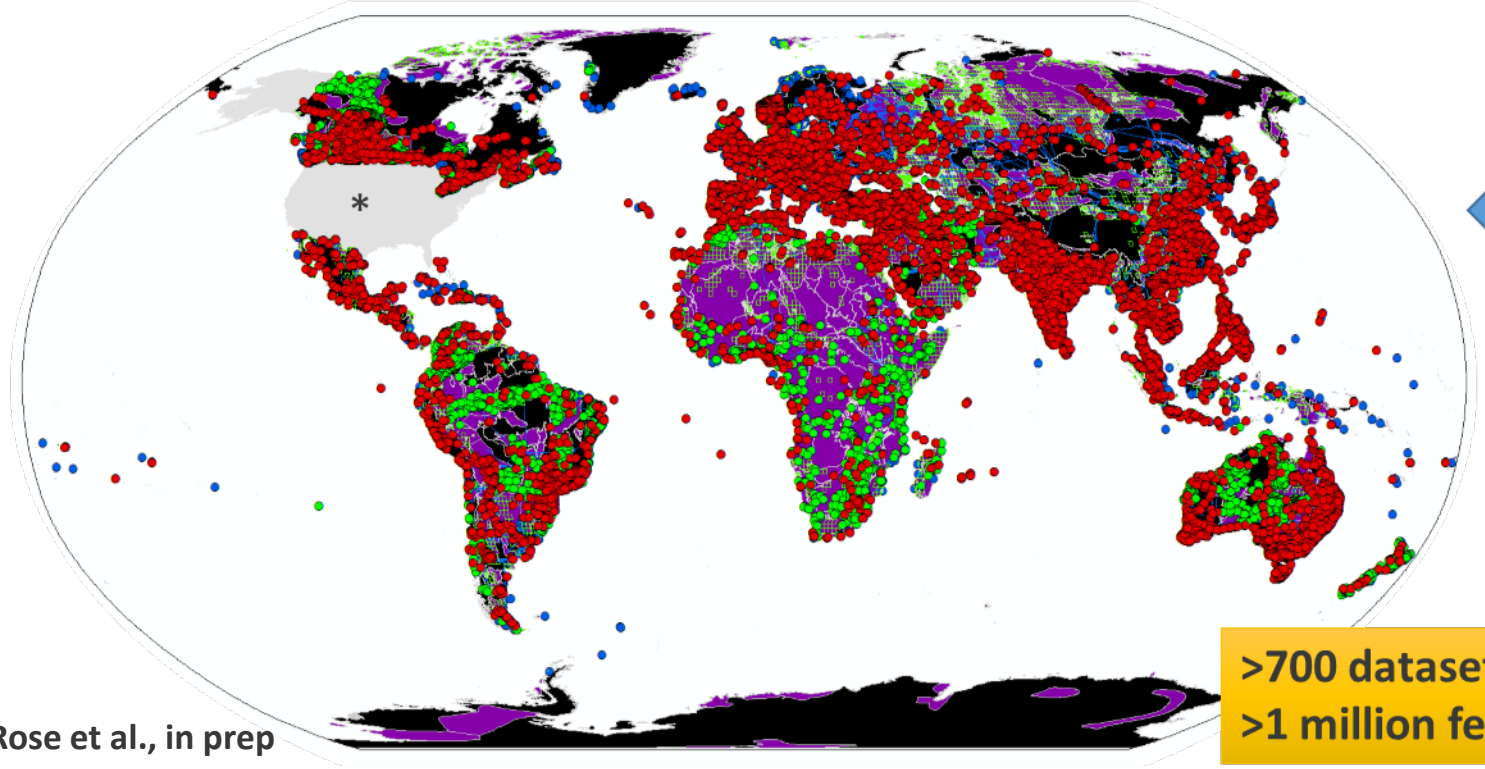
- >700 datasets
- >1 million features
- Attributes for some regions/features

- Dataset = Collection of data from a single source that represents real world objects
- Feature Type = A collection of one kind of feature (e.g. wells)
- Feature = a record for a single resource (i.e. – a well, a pipeline, a port, etc)




Convert Search Terms & Phrases into an Open O&G Spatial Database

Categorical Terms = 500
Geographical Terms = 5151
Spatial and Non-spatial = 380
**Total # of Search Terms = 6049**



Rose et al., in prep

Global Oil & Gas Feature Database

**>700 datasets**
**>1 million features**

# Base CKAN Features

- Content searching and indexing
- Raw data and metadata storage
- Public contribution workflow
- Public group functionality
- Geospatial searching
- API features to federate communication with other CKAN nodes (data.gov, openEI, NGDS, etc.)
- Data history and activity traceability info for each submission
- Data visualization for text and image data.
- User login

A Product of NETL

# EDX Custom Solutions Added to CKAN (1 of 2)

**What makes EDX different from other CKAN systems? 6 Years of data innovations**

- Collaborative Workspaces
- Slate, team digital notebook
- EDX suggested submissions and related resources
- Review process (Submissions, Users, Tools, Groups)
- Mobile support
- News
- Latest submissions
- Sign-up approval and activation process
- Portfolios
- Tools
- Libraries
- Calendars
- Private forums
- Draft process modification
- System administration blogs
- Geocube (connected to EDX datasets)
- Rate datasets modifications
- Custom statistics
- Auto generated citations
- Multi file upload/download
- Document previewing
- Zip file previewer and individual file extractor
- Drag and drop for uploading
- Two-factor authentication
- Heavily customized system admin capabilities
- Account workflow modifications to Password Reset
- Help customization and searchability
- External agency search feature (NOAA, USGS, EIA, BOEM, PHMSA, etc.)
- Advanced search builder
- Resource filter search
- EDXWiki

# EDX Ongoing & Future Development Focus Areas

- Automated metadata identification
- Enhanced search capabilities
- Analytics tools, plug & play for research
- Full OSTI integration
- Data review process automation
- 3D spatial viewing
- GIS persistent sessions
- Customizable collaborative workspaces
- Plug and play app/tools in CWs
- Testing & integrating cloud computing capabilities for EDX
- Continued integration of big data & HPC computing capabilities

**Data Analytics**

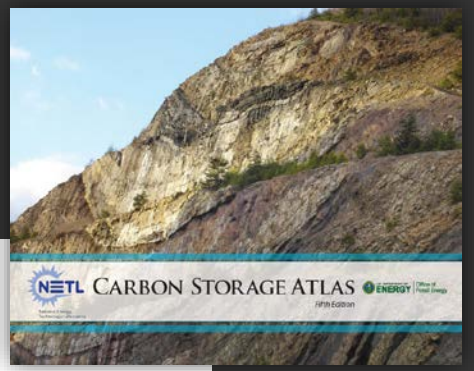
**Data Discovery**

**Describing Data**

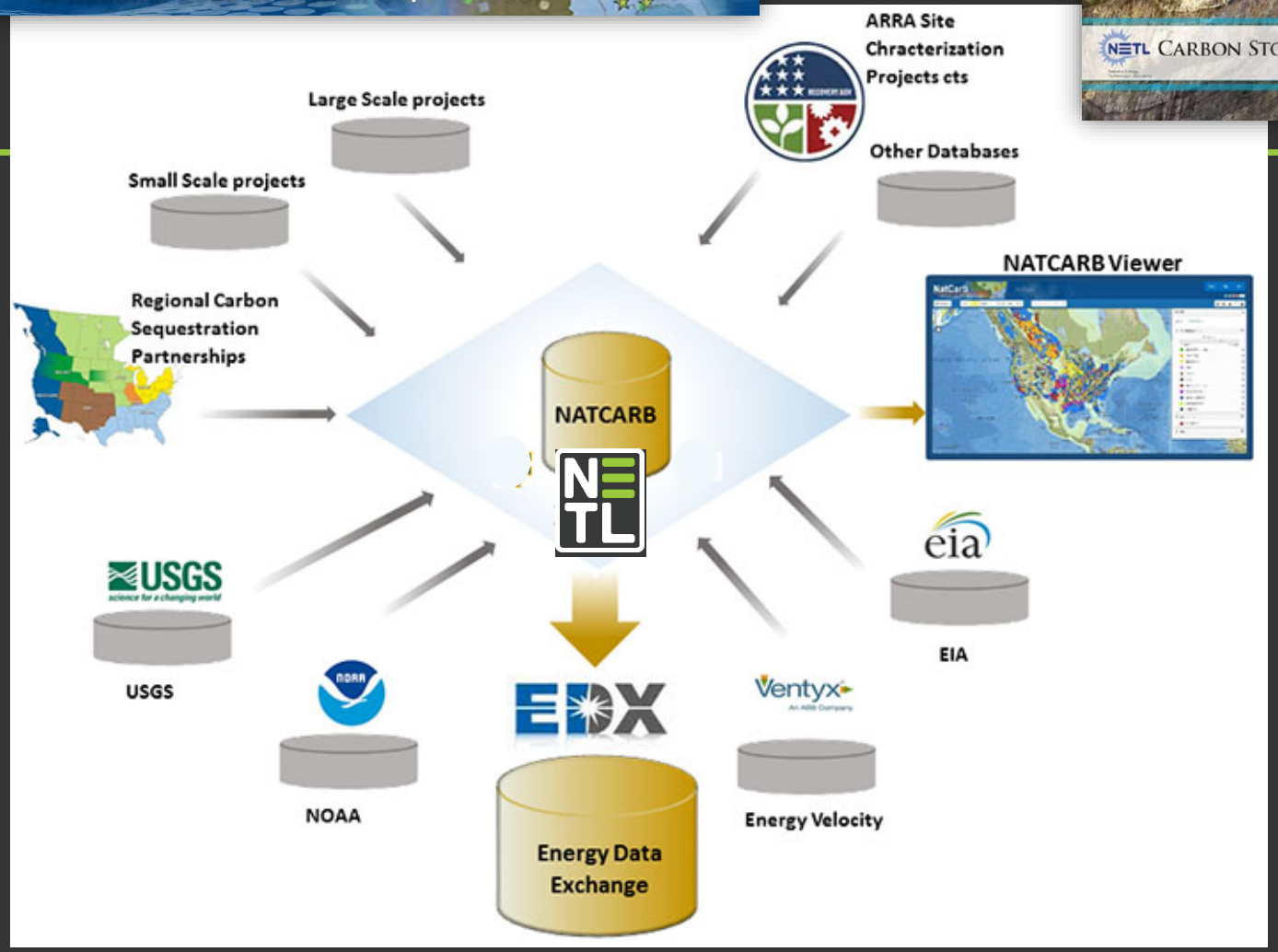
**Curating Data**

A Product of NETL

NatCarb
A National Look at Carbon Sequestration




NATIONAL ENERGY TECHNOLOGY LABORATORY




# Building a subsurface data framework for DOE R&D

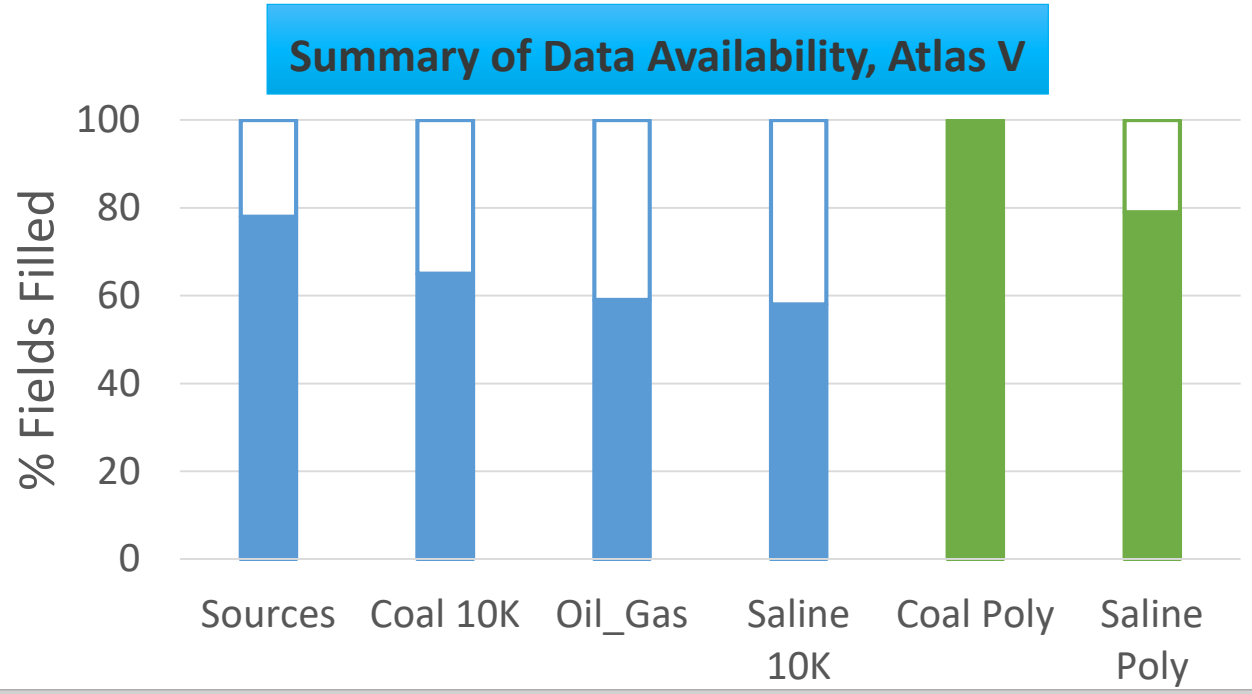
## RCSP Knowledge & Data for Natcarb Next Generation

Solutions for Today | Options for Tomorrow

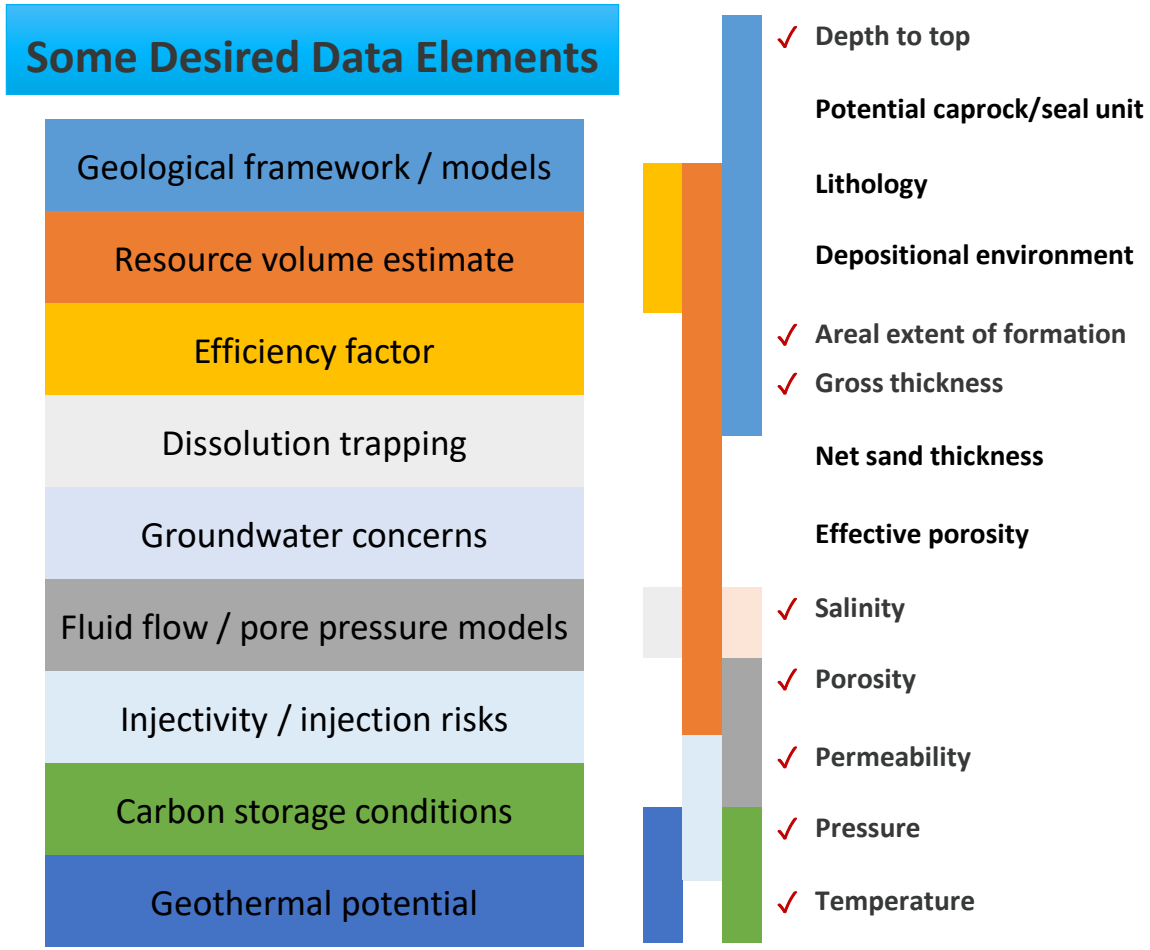
U.S. DEPARTMENT OF ENERGY

# Audited & Reviewed Natcarb Past

- **Audited content received vs desired**
- **Audited workflows for data processing**
- **Audited Natcarb tool**



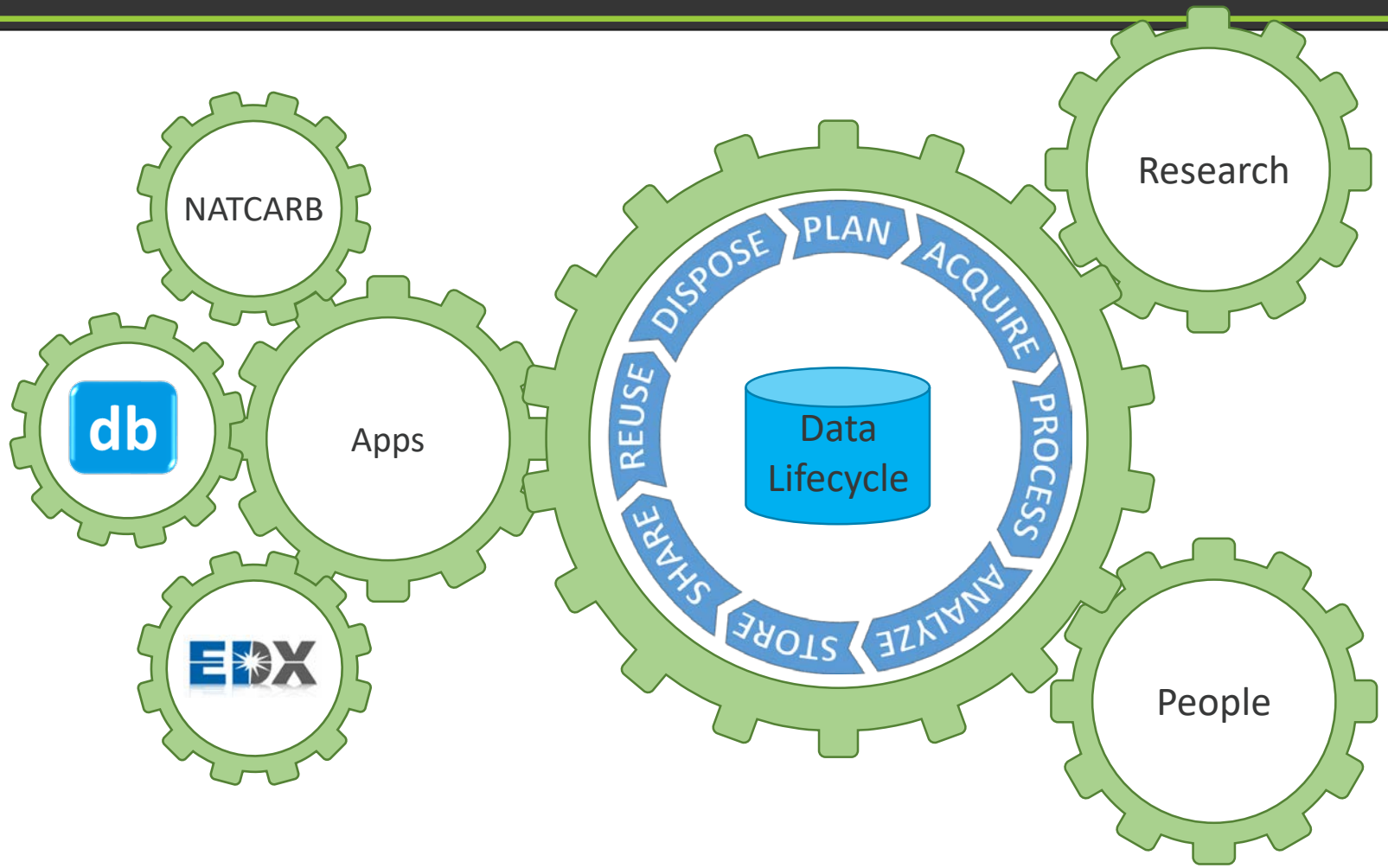

Except for the Coal Polygon layer, **only ~60-80 %** of the attribute cells contain information




✓ = Already requested from RCSPs

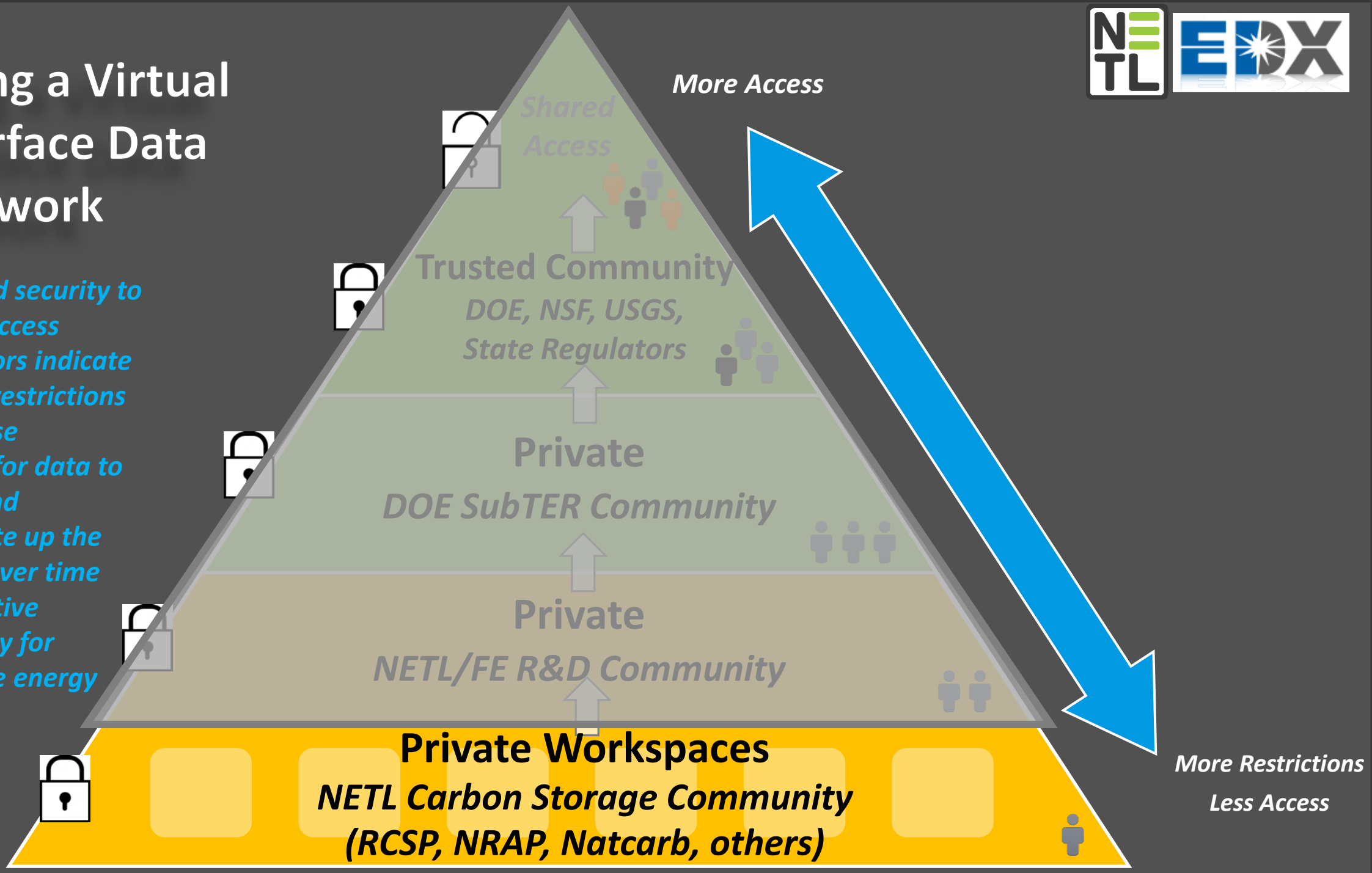
# Why Data Curation Matters - Research Data Lifecycle

- **Data Ecosystem**
- **Store and Share Data in a Structured Secure Environment**
  - Reduce Redundant Acquisition
  - Reduce Reuse Recycle
  - Consistent Data with Staff Turnover
  - Enhanced Collaboration
- **Curation of data and knowledge**



# Building a Virtual Subsurface Data Framework

- *Role based security to manage access*
- *Contributors indicate "license" restrictions on data use*
- *Potential for data to mature and matriculate up the pyramid over time*
- *Collaborative community for subsurface energy R&D*



# Why Data Curation Matters

## Spurs innovation

### City of Los Angeles – GeoHub

***Open Data sharing for economic development***

**Free-Range Data**

- By connecting datasets across departments
- Fewer Stovepipes, More Networks
- Search for data…mash up *[or]* combine maps, get insights, make better decisions

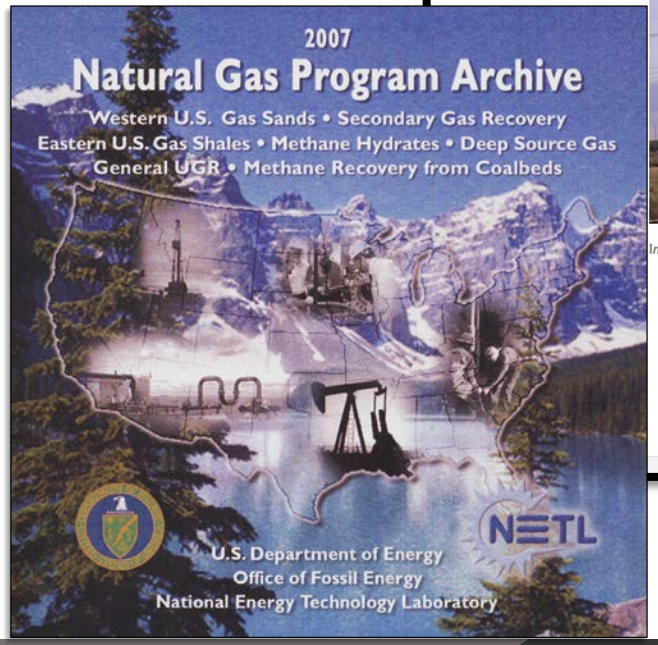
**Economic Benefits**

- **Startups** represent not only potential economic development but also collaboration opportunities for solving some of the city's biggest problems
- **Developers** can access the city's data, along with open APIs, to build apps that they can bring to market.

# Why Data Curation Matters

## Spurs innovation

- **Not just about Amazon, Google, shopping histories etc.**
- **Data is valuable to research**
- **Provides a <u>foundation</u> for new innovation, fill in knowledge gaps, etc.**
  - E.G., DOE's own shale gas R&D from the 1970's -90's helped spur the natural gas revolution in 2007 – present worldwide


https://www.wired.com/insights/2014/07/data-new-oil-digital-economy/

SHARE

SHARE 192

TWEET

COMMENT 0

EMAIL

PARTNER CONTENT JORIS TOONDERS, YONEGO

**DATA IS THE NEW OIL OF THE DIGITAL ECONOMY**

Image: verifex/Flickr

DATA IN THE 21st Century is like Oil in the 18th Century: an immensely, untapped valuable asset. Like oil, for those who see Data's fundamental value and learn to extract and use it there will be huge rewards.

We're in a digital economy where data is more valuable than

# RCSP's Knowledge & Data Has Opportunity to Transform DOE R&D Landscape

Data drives innovation and supports advanced R&D tools, technologies, models and analyses

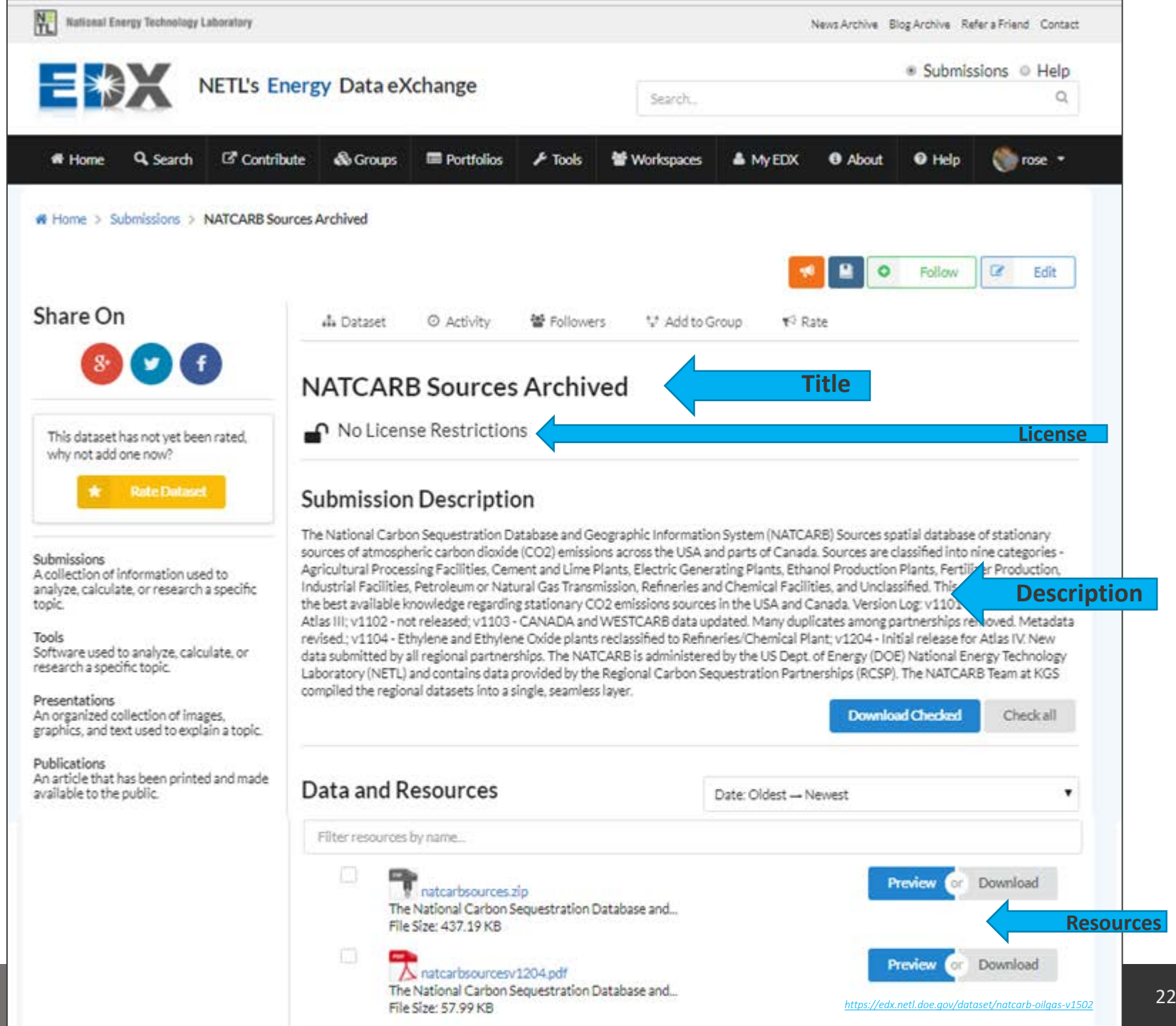
By building a virtual subsurface data framework for DOE R&D…

- Stop recreating data "wheels"
- Understand what is known and where there are gaps
- Leverage EDX's public and <u>private</u> capabilities to enable data sharing for DOE R&D community benefit

# Now two EDX options for curating RCSP/Natcarb data

## Conventional data resource submission

- **Data resource = dataset, tool, model, app, pub, presentation**



# DataBook?

## DataBook is a virtual, team digital notebook

Formatting of individual components is handled visually (clicking and dragging)


Provides a platform for team members to collaborate and present data

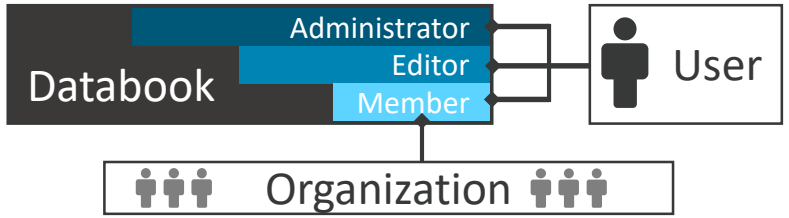
Multimedia support: text, image, audio, video, map data, and more

No fixed organization of data on page

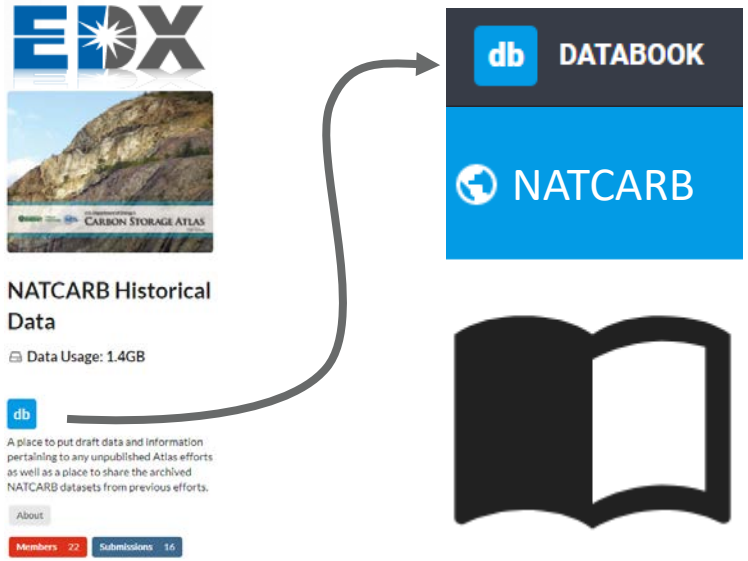
Hosted within EDX collaborative workspaces

# How DataBook works

- Three different tiers of access.
  - **User:** Single account, with different tiers:
    - **Admin:** Full read/write access; can modify entire databook and user roster.
    - **Editor:** Full read/write access to content; can modify databook.
    - **Member:** Read only access to databook content.
  - **Organization:** Read-only access to all users within an organization (determined by email address). Equivalent access as **Member** user role.

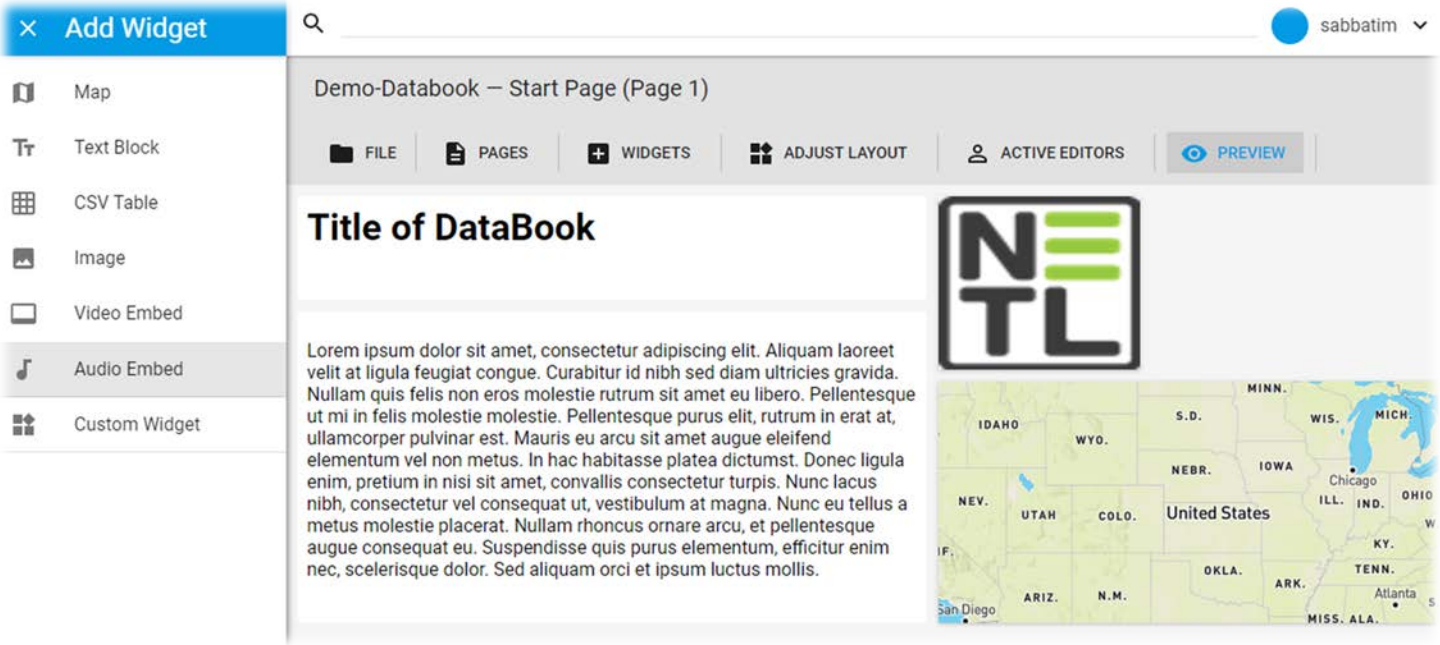



- Hosted within EDX
  - Ability to create databook(s?) within EDX workspaces
    - All users and associated permissions are imported into Databook on first click.
  - Future enhancements will link additional data between databook and Collaborative Workspaces.



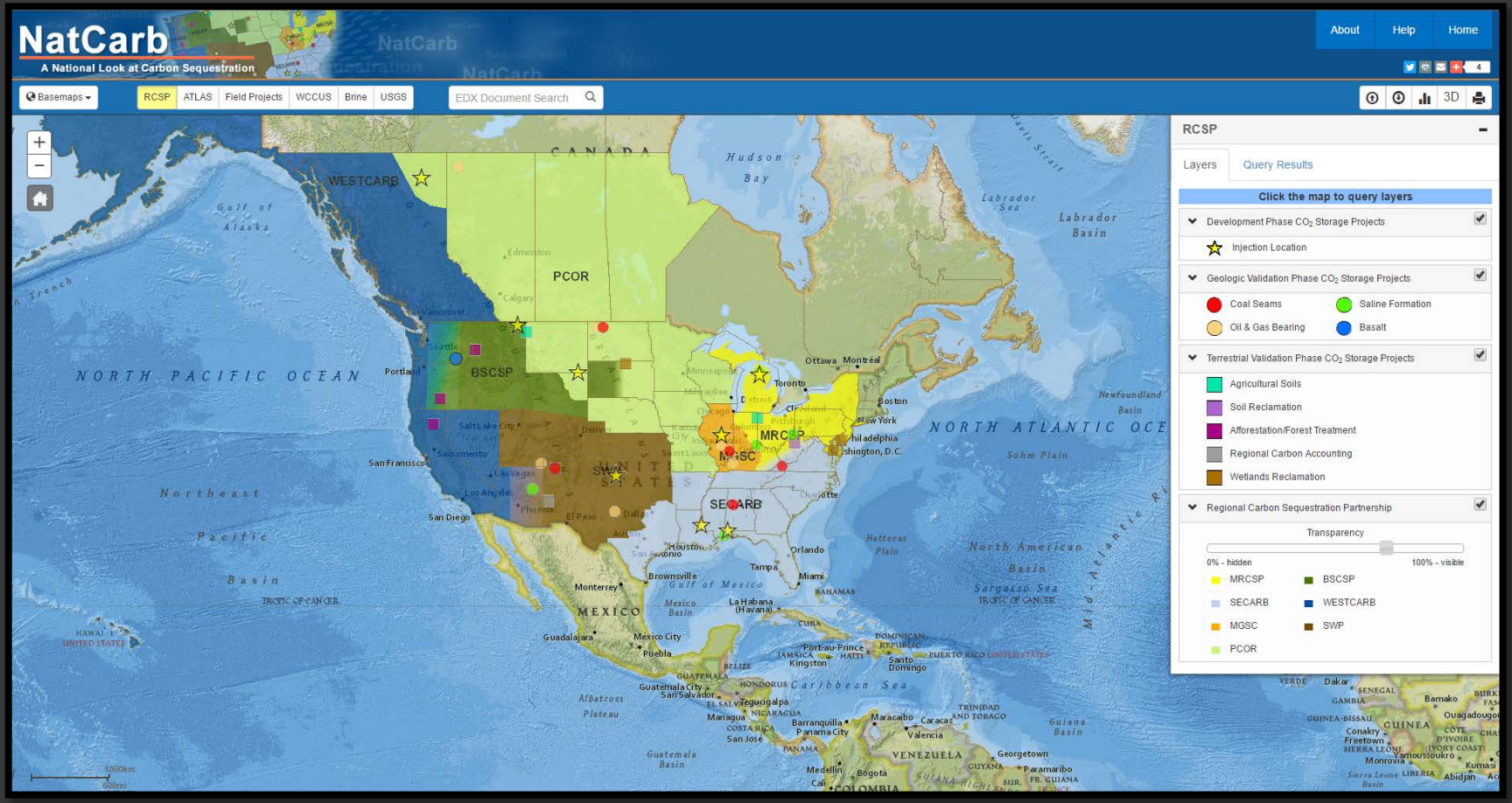
# How DataBook works

- **Widget driven. Widgets allow content of different types to be added to DataBook**
  - Text –Titles and text notes
  - Data Tables –Tabular or .csv data with basic spreadsheet functions
  - Image -.png, .jpg or other image file loaded onto DataBook
  - Map –Widget to view spatial with basic GIS functionality
  - Audio –External link to audio source
  - Video –External link to video source
- **DataBook for R&D and Natcarb**
  - New DataBooks can be initiated in any collaborative workspace
  - For Natcarb Atlas update, DataBooks with prescribed templates will be provided requesting specific data inputs



# NatCarb Tool now

- Dependent on manual processes for production of Atlas content both online & paper
- Limited to Atlas specific products and data

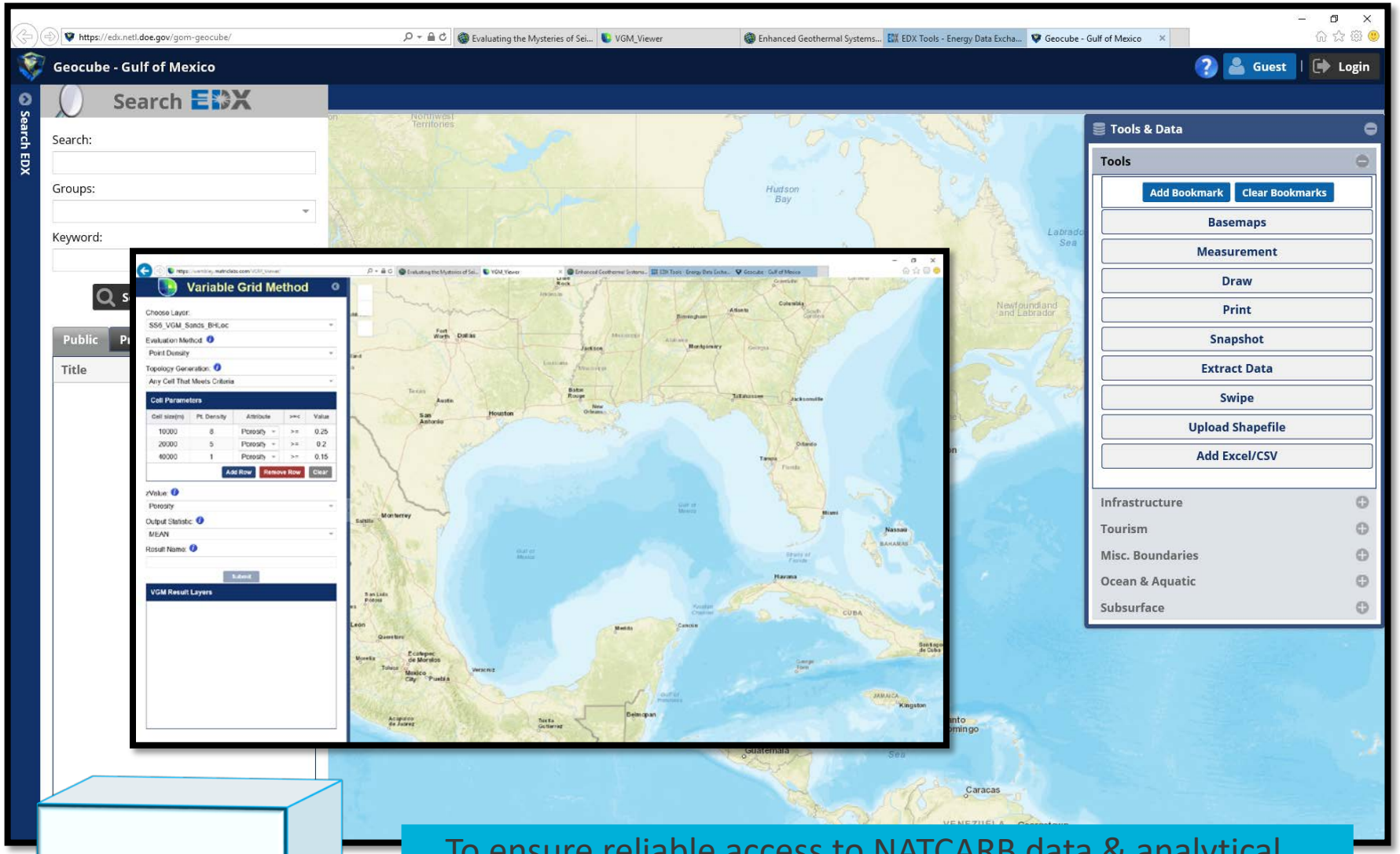


# NATCARB

## THE NEXT GENERATION

# NatCarb Tool – *Next Generation*

- Integrating with EDX's Geocube web mapping tool
- Maintains the current Natcarb URL
- But **freshens the look, feel and functionality**
- Integrates with other EDX and Geocube **resources** for improved discovery & analytics




To ensure reliable access to NATCARB data & analytical resources, we plan in next 6 months to integrate Natcarb into its own instance via Geocube

## New Team Data Tool via EDX

# db DataBook

**What is DataBook?** DataBook is a web-based collaborative environment for teams to create and publish interactive data "notebooks." DataBook curates team knowledge to develop a living, evolving data and information foundation for R&D.

EDX Energy Data eXchange

# User Session

## Drop In Event, Anyone Welcome!

**Date** – Wednesday, August 2, 2017
**Time** – 1:20 to 5:40 pm
**Location – Sheraton Station Square Hotel, 2nd Floor Executive Board Room**

**Bring your laptop & questions**
Talk to EDX Experts
Learn how to **customize EDX** for your needs

## Thank you!!!

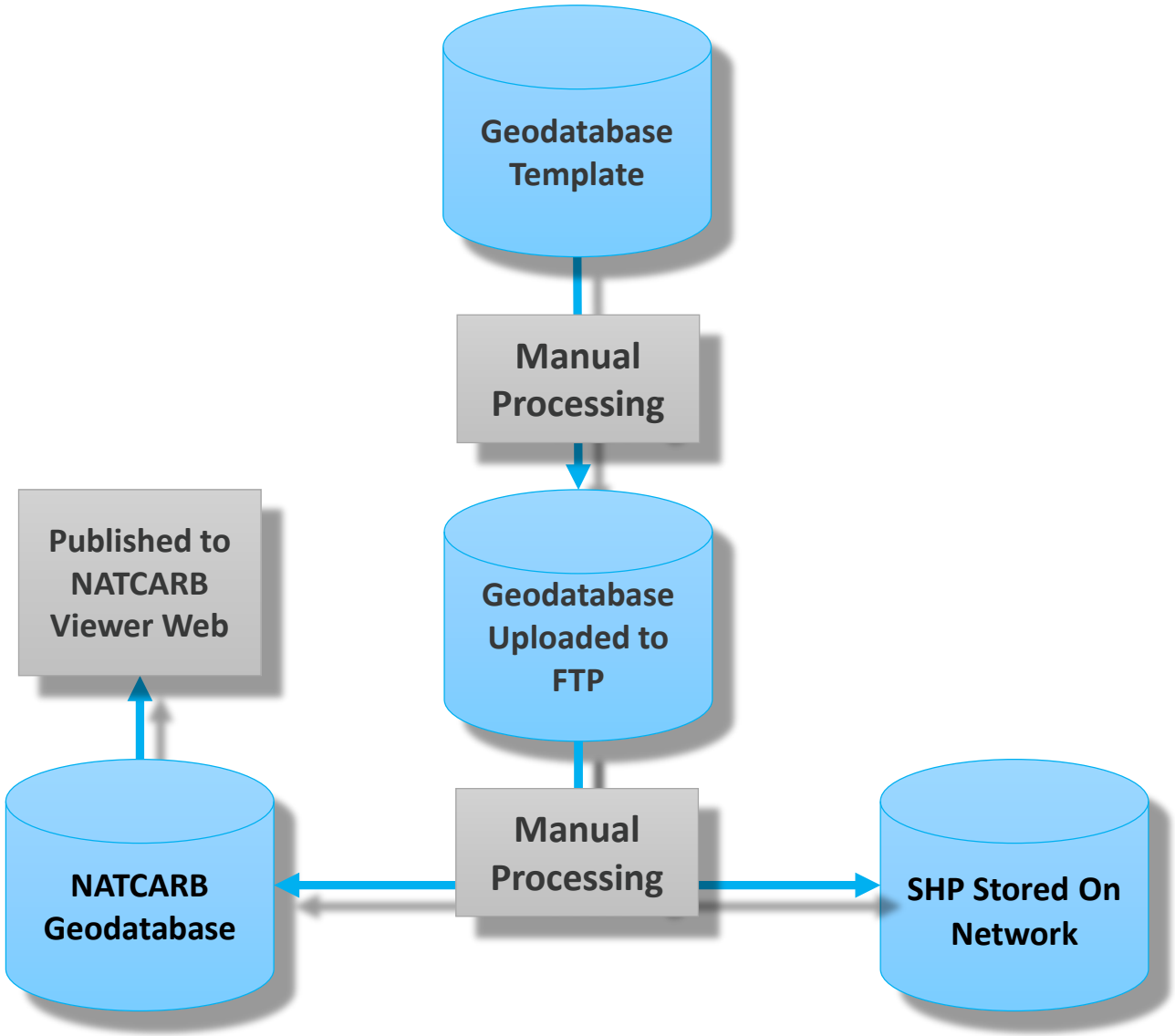
Kelly.rose@netl.doe.gov

EDXsupport@netl.doe.gov

**Pubs & Presentations:**


- Baker, V., et al., *in prep*, Big data computing and machine learning for efficient data discovery, *Big Data Research*
- Baker, D.V., Rose, K., Bauer, J.R., and Justman, D. Big Data Computing for GIS Data Discovery. Esri User Conference, San Diego, CA, July 10-14, 2017. http://www.esri.com/events/user-conference.
- Bauer, J.R., Rose, K., Baker, D.V., and Barkhurst, A. Big Data, Big Uncertainty – Taming Uncertainty in Big Data Spatio-Temporal Analytics with the Variable Grid Method. American Association of Geographers (AAG) Annual Meeting, Boston, MA, April 5-9, 2017. http://www.aag.org/cs/annualmeeting.
- Rose, K., Bauer, J.R., Baker, D.V., Justman, D., Romeo, L., Mark-Moser, M., and Miller, M. Data driven spatial methods for subsurface & infrastructure resources. Esri User Conference, San Diego, CA, July 10-14, 2017. http://www.esri.com/events/user-conference.
- Rose, K., et al., 2017, Working Smarter Not Harder – Developing a Virtual Subsurface Data Framework for US Energy R&D, invited talk, American Geophysical Union Annual Meeting, **IN035. Increasing the bandwidth of imaging-data-to-research pipelines**
- Rose, K., et al., 2017, A smarter way to search, share and utilize open-spatial online data for energy R&D – Custom machine learning and GIS tools in U.S. DOE's virtual data library & laboratory, EDX, invited talk, American Geophysical Union Annual Meeting, **IN055. Spatial Data Infrastructure for Earth and Space Sciences: Analyzing, Visualizing, and Sharing Spatio-temporal Earth Science Data Small and Big**

# Previous NATCARB Data Flow



# NATCARB & RCSP Data

- Beyond supporting Atlas products and curation of data….

- We propose in FY18/19 to also evaluate options for how to <u>use</u> data to support Carbon Storage R&D

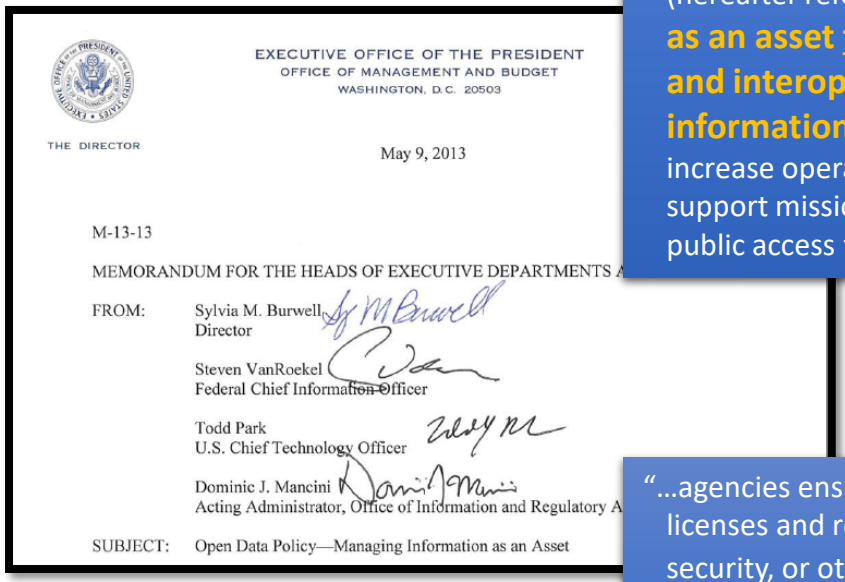
- Questions about NATCARB
  - What is covered by whom going forward? Heard Westcarb is going away
  - What data is coming in from NATCARB or RCSPs? Size, volume, format, restrictions etc?
  - What map/data products does program envision requiring to support next update of Atlas?

# 2013 Executive Order

## *Open Data Policy – Managing Information as an Asset*

***Memorandum for the Heads of Executive Departments and Agencies: Open Data Policy —Managing Information as an Asset*, May 9, 2013, accessed June 25, 2013.**

- Federal government must manage information throughout its lifecycle
- Must properly safeguard systems & information
- This will increase efficiencies, reduce costs, improve services, support mission needs, & increase public access to government information products
- Effective information management throughout it's lifecycle promotes interoperability and openness
- Ensure information stewardship
- Modernize information systems to maximize interoperability and information access
- Maintain internal and external data inventories
- Clarify information management responsibilities



EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF MANAGEMENT AND BUDGET
WASHINGTON, D.C. 20503

THE DIRECTOR

May 9, 2013

M-13-13

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS A

FROM: Sylvia M. Burwell
Director

Steven VanRoekel
Federal Chief Information Officer

Todd Park
U.S. Chief Technology Officer

Dominic J. Mancini
Acting Administrator, Office of Information and Regulatory A

SUBJECT: Open Data Policy—Managing Information as an Asset

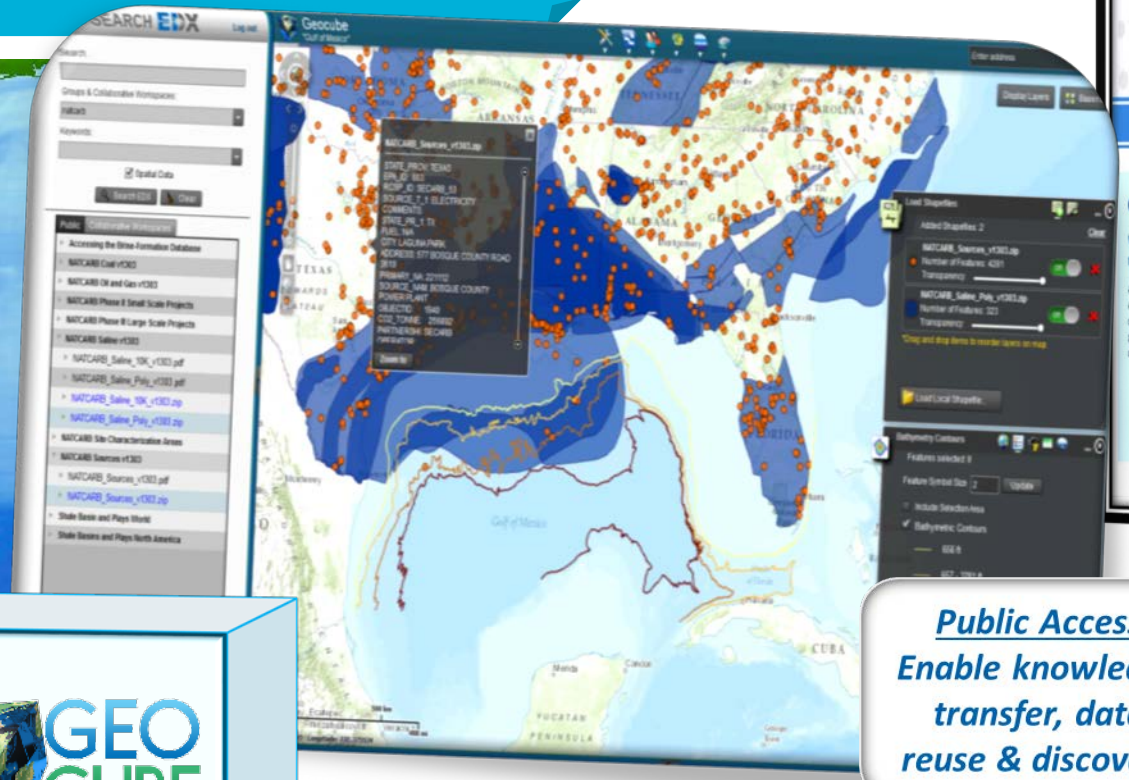
https://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf

"Information is a valuable national resource and a strategic asset to the Federal Government, its partners, and the public. In order to ensure that the Federal Government is taking full advantage of its information resources, executive departments and agencies (hereafter referred to as "agencies") **must manage information as an asset throughout its life cycle to promote openness and interoperability, and properly safeguard systems and information.** Managing government information as an asset will increase operational efficiencies, reduce costs, improve services, support mission needs, safeguard personal information, and increase public access to valuable government information."

"…agencies ensuring information stewardship through the use of open licenses and review of information for privacy, confidentiality, security, or other restrictions to release. Additionally, it **involves agencies building or modernizing information systems in a way that maximizes interoperability and information accessibility, maintains internal and external data asset inventories, enhances information safeguards, and clarifies information management responsibilities.** "

# Data Access & Analytics thru EDX

To ensure reliable access to these datasets, we leverage NETL's Energy Data eXchange (EDX) and online tools, like Geocube, to **access & serve key datasets**



GEO CUBE

***Public Access***
***Enable knowledge transfer, data reuse & discovery***

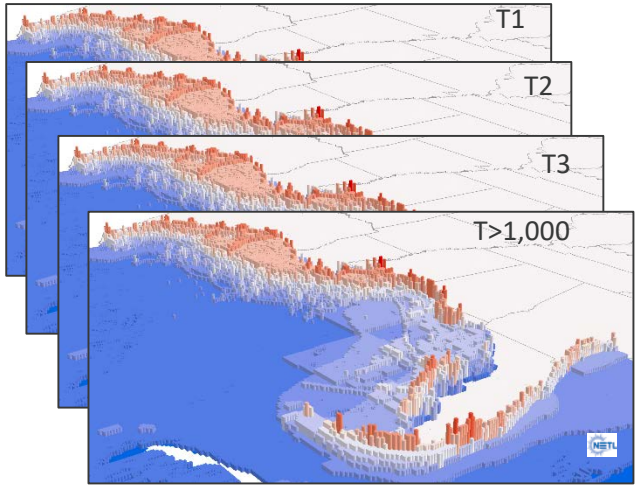
***Secure/Private Access***
***Support research development, collaboration, & online analytics***

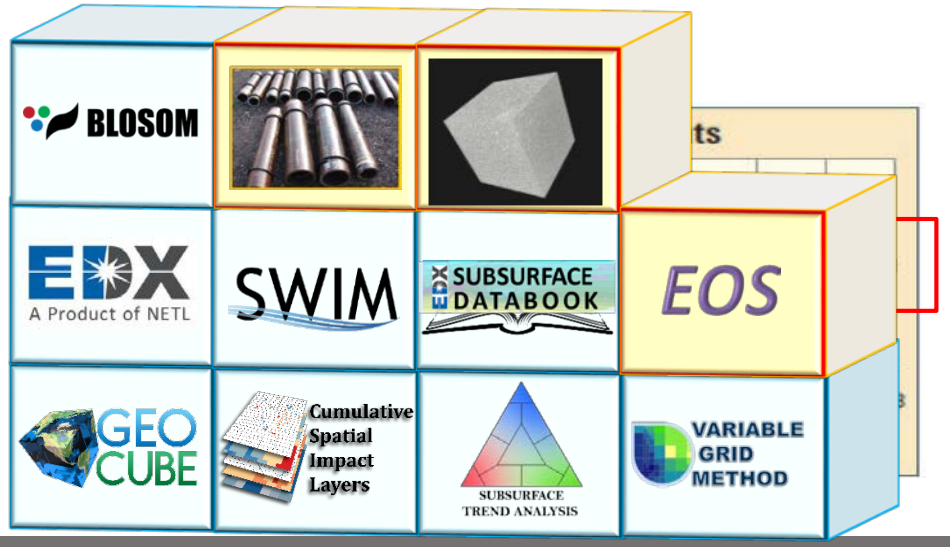
# Online Analytics – Using EDX Hosted Data for R&D

- Integrating data, tools and models to support informed decision making & analyses
- Prepare, predict, prevent


EDX
A Product of NETL

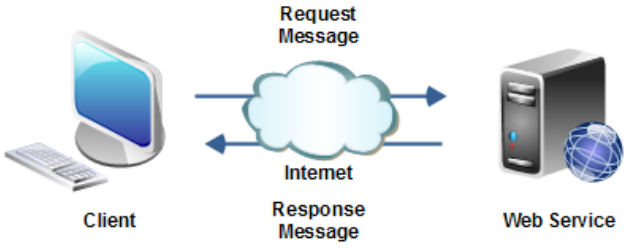
Developing **an online, common operating platform**, serving web-based tools, and big data geoprocessing for analytics




- Utilizing risk suite for monte carlo-style **assessments of GOM spatio-temporal risks**
- Inform decision making

# Web services, the power of mining & sharing




**Web services connect online tools & systems with data**

- **Connecting** data from it's **primary home** for community use
- Ensures most up to date information is always available