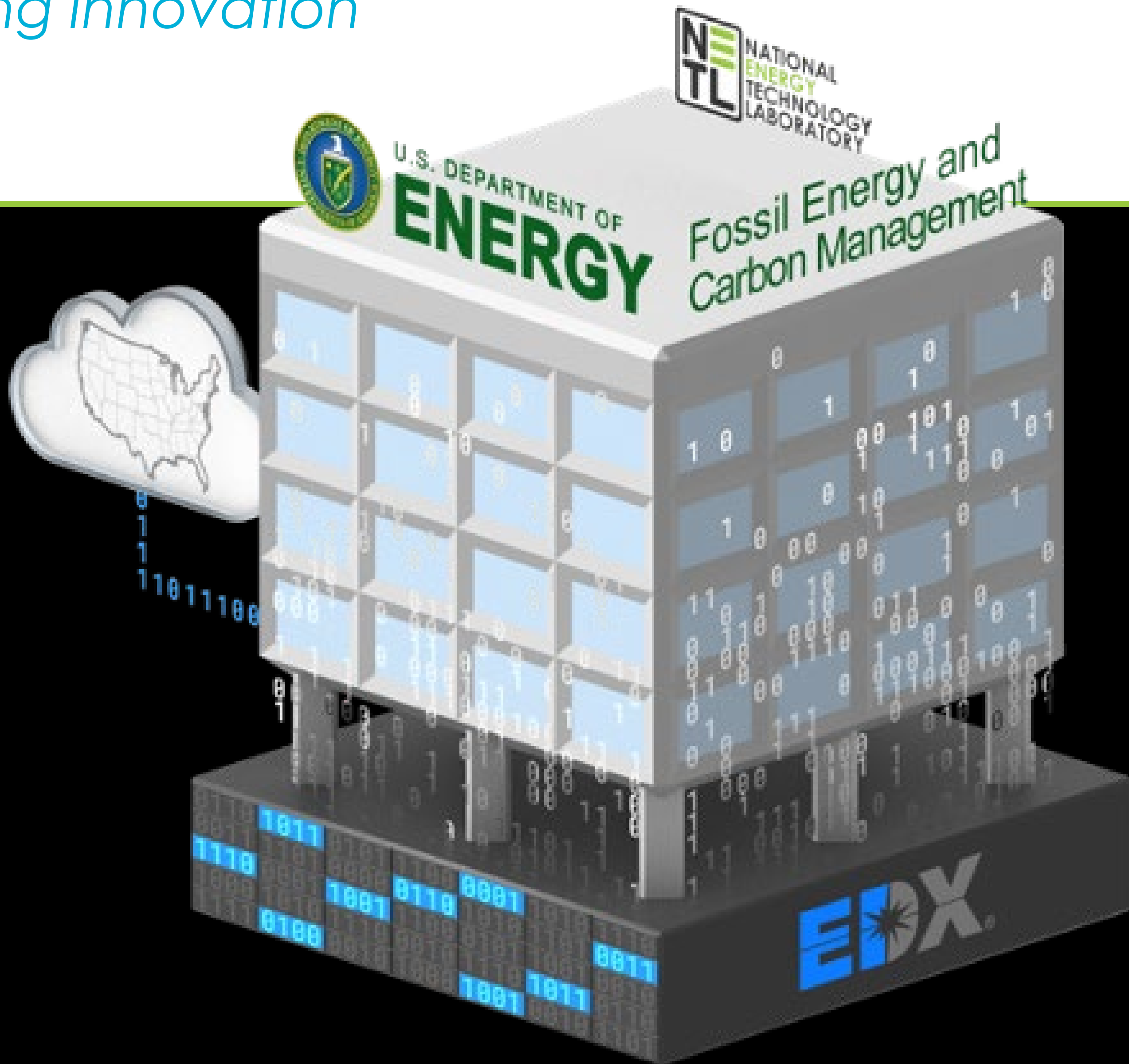
Using Cloud to Host Large Open-Source Datasets

Catalyzing Collaboration & Accelerating Innovation


NETL
NATIONAL ENERGY TECHNOLOGY LABORATORY
U.S. DEPARTMENT OF ENERGY
Fossil Energy and Carbon Management
EDX

EDX®

Energy Data eXchange

Chad Rowan

National Energy Technology Laboratory (NETL)

Computational Science & Engineering

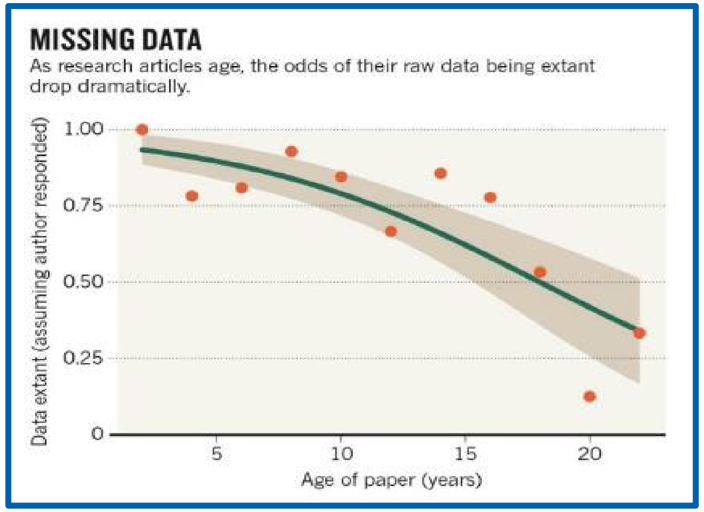
Advanced Computing & Artificial Intelligence

Disclaimer

This work 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.

FECM R&D is Impeded by Common Challenges

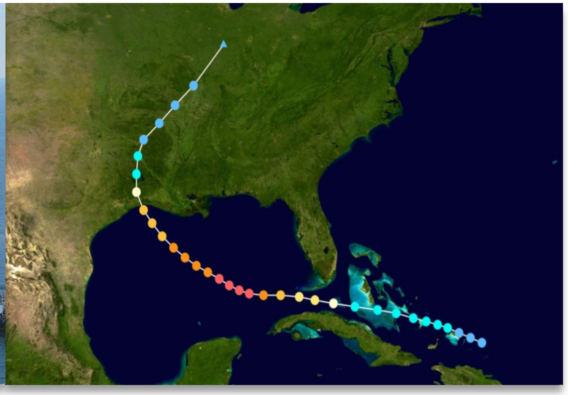
- Finding and accessing relevant datasets
- Publishing & preserving R&D data products
- Accessing previously developed R&D data
- Sharing secure R&D scale data resources among team
- Collaborating across multi-organizational teams
- Need to access prior R&D data products to accelerate next-generation innovations



MISSING DATA

As research articles age, the odds of their raw data being extant drop dramatically.

Image from : http://www.nature.com/news/scientists-losing-data-at-a-rapid-rate-1.14416



The Bigger the Data the Bigger the Challenges

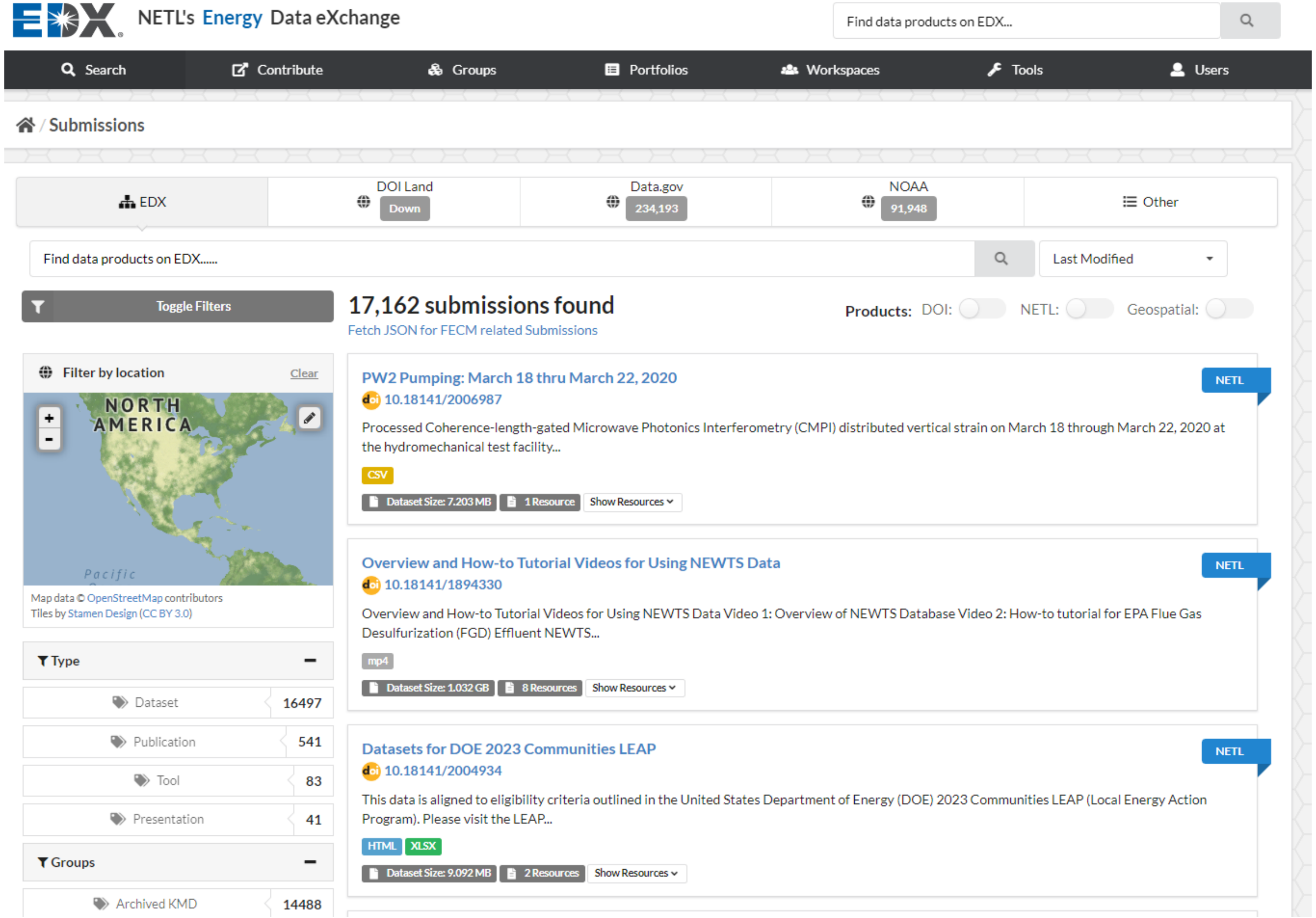
- Datasets are increasingly becoming larger
- Large datasets are more difficult to curate and make publicly available
- Large datasets are not exempt from federal data publishing guidelines

How has FECM addressed these issues?

EDX®

Energy Data eXchange

a web-hosted, virtual library and laboratory that supports the NETL/FECM community


NETL's Energy Data eXchange
Find data products on EDX...
Search
Contribute
Groups
Portfolios
Workspaces
Tools
Users
Submissions
EDX
DOI Land Down
Data.gov 234,193
NOAA 91,948
Other
Find data products on EDX......
Last Modified
Toggle Filters
Products: DOI: NETL: Geospatial:
Filter by location
Clear
NORTH AMERICA
Pacific
Map data © OpenStreetMap contributors
Tiles by Stamen Design (CC BY 3.0)

17,162 submissions found

Fetch JSON for FECM related Submissions

Type	
Dataset	16497
Publication	541
Tool	83
Presentation	41

Groups	
Archived KMD	14488

PW2 Pumping: March 18 thru March 22, 2020 NETL

10.18141/2006987

Processed Coherence-length-gated Microwave Photonics Interferometry (CMPI) distributed vertical strain on March 18 through March 22, 2020 at the hydromechanical test facility...

CSV

Dataset Size: 7.203 MB | 1 Resource | Show Resources

Overview and How-to Tutorial Videos for Using NEWTS Data NETL

10.18141/1894330

Overview and How-to Tutorial Videos for Using NEWTS Data Video 1: Overview of NEWTS Database Video 2: How-to tutorial for EPA Flue Gas Desulfurization (FGD) Effluent NEWTS...

mp4

Dataset Size: 1.032 GB | 8 Resources | Show Resources

Datasets for DOE 2023 Communities LEAP NETL

10.18141/2004934

This data is aligned to eligibility criteria outlined in the United States Department of Energy (DOE) 2023 Communities LEAP (Local Energy Action Program). Please visit the LEAP...

HTML XLSX

Dataset Size: 9.092 MB | 2 Resources | Show Resources

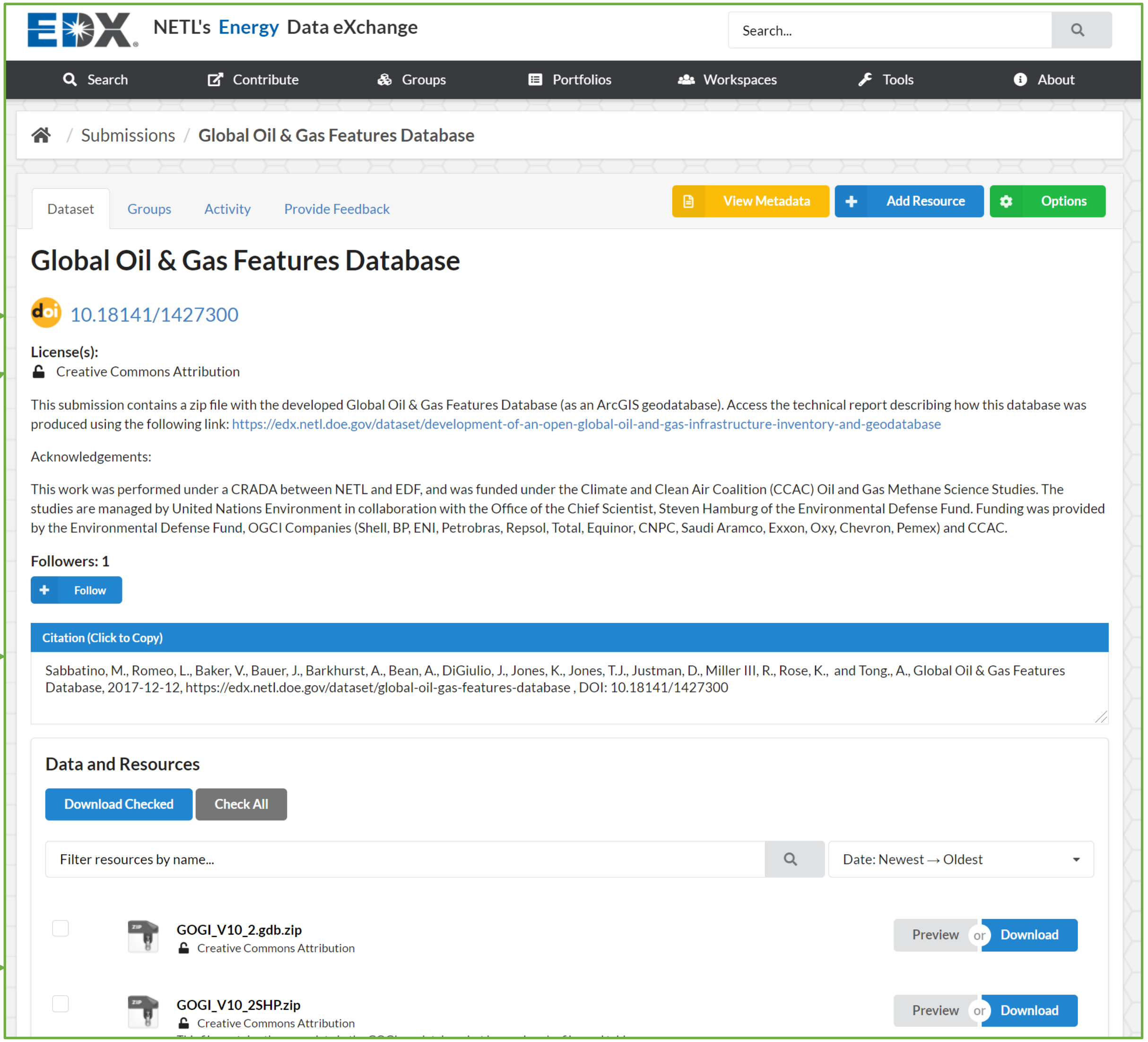
Advantages of publishing data products

OSTI DOI Number

Data License

Data Citation

Data Access



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

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

GOGI_V10_2.gdb.zip
Creative Commons Attribution

GOGI_V10_2SHP.zip
Creative Commons Attribution

Many journals require models, tools and data be publicly available prior to journal publication.

Energy Data eXchange

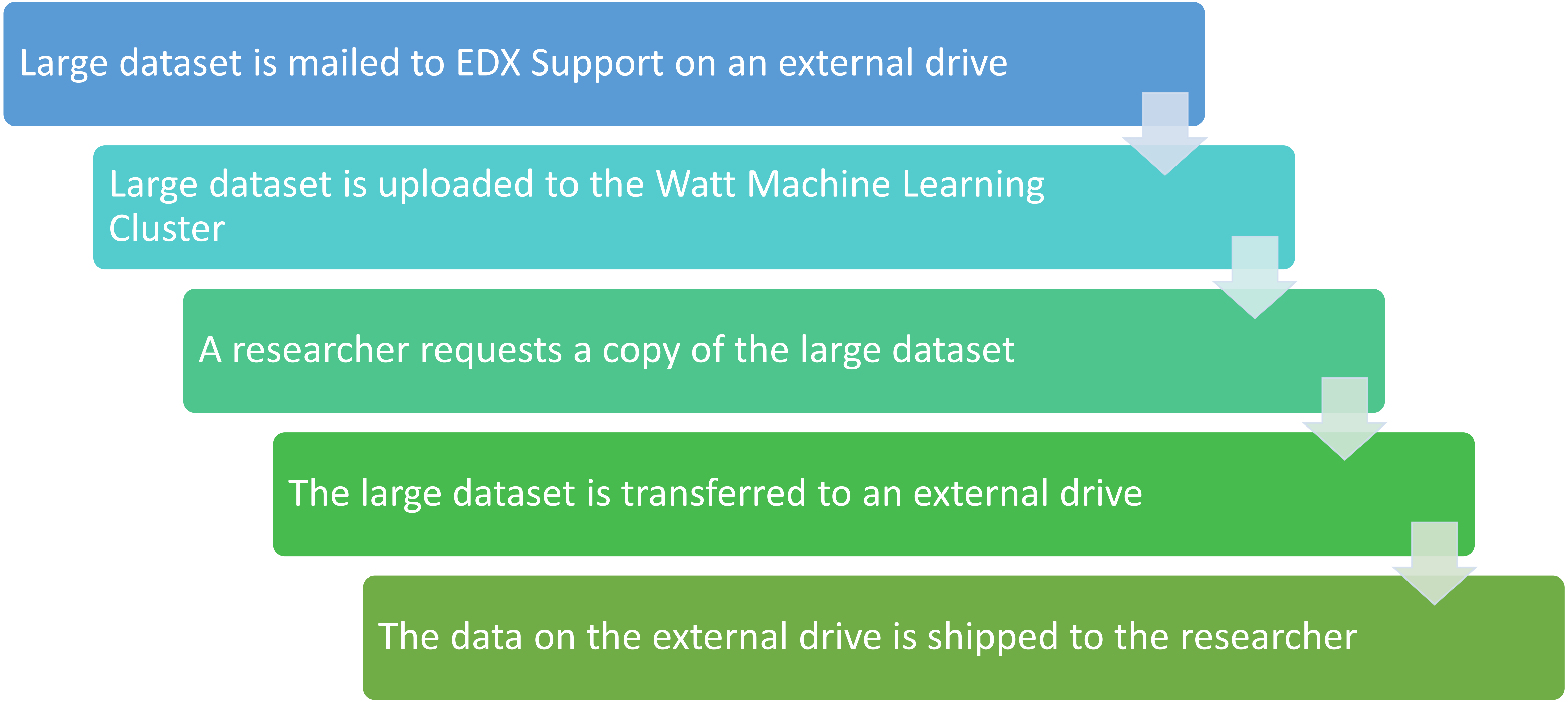
can help!

Our Big Data Journey

- In 2011, we thought big data was a few gigabytes
- Transferring gigabyte files across the Internet to EDX was slow, but doable
- In the last few years, we have datasets that have grown to over 100 terabytes
- Transferring terabyte files to our on-prem EDX server over the Internet was not feasible

This Photo by Unknown Author is licensed under CC BY-SA

The Recent Past



Large dataset is mailed to EDX Support on an external drive

Large dataset is uploaded to the Watt Machine Learning Cluster

A researcher requests a copy of the large dataset

The large dataset is transferred to an external drive

The data on the external drive is shipped to the researcher

We needed a more nimble approach to Big Data

- Rather than upgrade our on-prem hardware we started exploring cloud options
- We were quickly introduced to Cloud Open Data Programs
- All major cloud service providers have Open Data Programs

What do cloud Open Data Programs provide?

- Free hosting and egress of large, publicly accessible data
- Increased access to ODP hosted datasets
- Faster upload/download speeds
- Access to cloud tools
- Access to cloud compute
- Access to cloud data analytics

What is the benefit of hosting Carbon Storage data in an ODP?

- ODP is **FREE** for large, public datasets
- Provides **ACCESS** to large, public datasets that were historically difficult to share
- Increases **VISIBILITY** and **DISCOVERABILITY** of large, public datasets
- Facilitates **CLOUD COMPUTE** at the source of the data

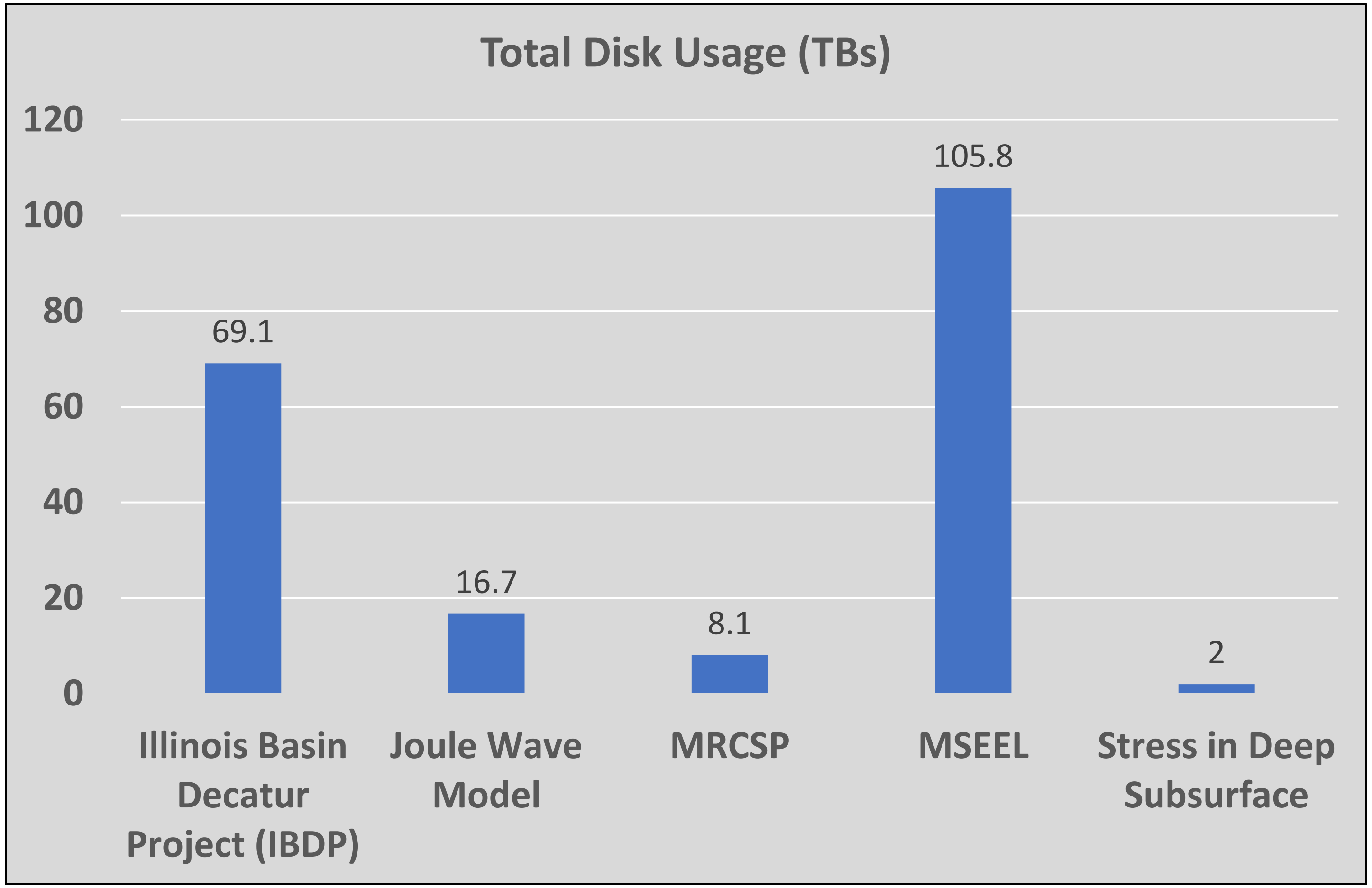
Google's Open Data Program

- After reviewing Open Data Programs from all of the cloud service providers we started our ODP journey with Google
- Google shipped a 300TB transfer appliance to transfer our first collection of large datasets
- The data was transferred to the appliance and sent back to Google where it was uploaded to a GCP bucket
- The EDX Team is currently working with Google engineers to apply metadata, making the dataset discoverable for use/re-use

What was included in the first ODP package?

NATIONAL ENERGY TECHNOLOGY LABORATORY

- ✓ 5 datasets
- ✓ Over 200TBs of data
- ✓ Over 24M data files



Why are Open Data Programs Free?

- Large datasets are desirable
- CSPs know if they host some of your data for free they have a better chance of hosting your other data at a cost
- CSPs can market cloud tools for compute and visualization that incur a cost

Open Data Programs facilitate the concept of EDX++

Freedom for users to use any cloud service providers

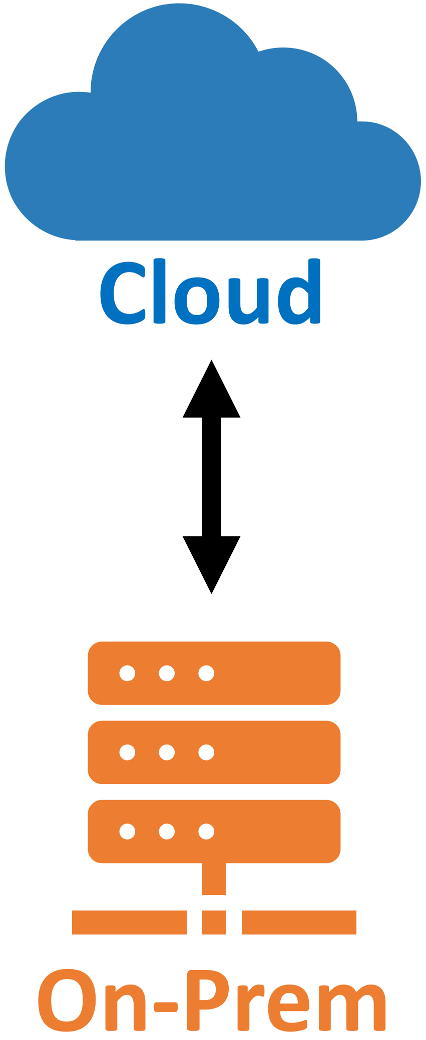
- Compute at the source of the data
- Utilize APIs to move data
- Web hosted applications

Improves flexibility and performance

- Does not limit users to one storage & compute platform
- Compute occurs at the data source

Resilience

- Redundancy across multiple regions
- Strategic alignment for data transfer and compute across multiple cloud service providers



ODPs Facilitate FECM R&D Data Use and Reuse

Carbon Storage Program

- **Free Hosting and Egress:**
  - Over 200 terabytes of data
  - Over 24 Million data files
  - Supporting use/re-use of current and future Carbon Storage research efforts

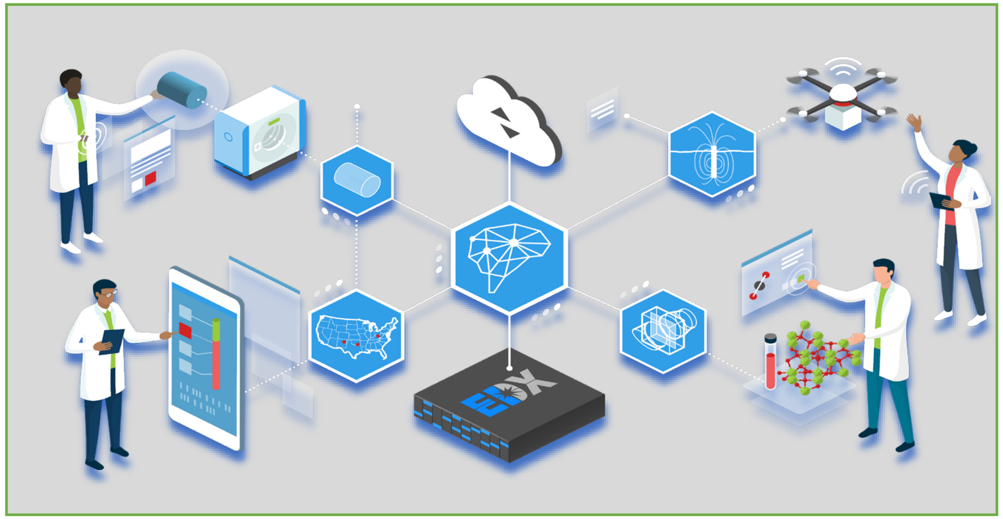
SAMi
SCIENCE-BASED AI/ML INSTITUTE
NETL Research · Carbon Storage Program ·
NRAP
National Risk Assessment Partnership
SMART
SECARB
Southeast Regional Carbon Sequestration Partnership
PCO2R
Partnership
EDX4CCS
MGSC
BIG SKY CARBON SEQUESTRATION PARTNERSHIP
CarbonSafe
BIPARTISAN INFRASTRUCTURE LAW
S WY
UT
CO
KS
OK
AZ
NM
W TX
MRCSP
MIDWEST REGIONAL CARBON SEQUESTRATION PARTNERSHIP

just a few examples

The future is now

FECM is embracing research challenges by providing state of the art solutions

- Evolving into a multi-cloud solution
- Accelerating AI/ML
- Tackling data compute in the cloud and on-prem
- Improving transfer speed, security, and pipe



Any questions?

Key Resources

- EDX Reference Shelf
- Focused training for research teams (Request Training)
- EDX Training Videos (pre-recorded)
- Robust API Documentation
- **Contact** EDXSupport@netl.doe.gov or SAMI@netl.doe.gov