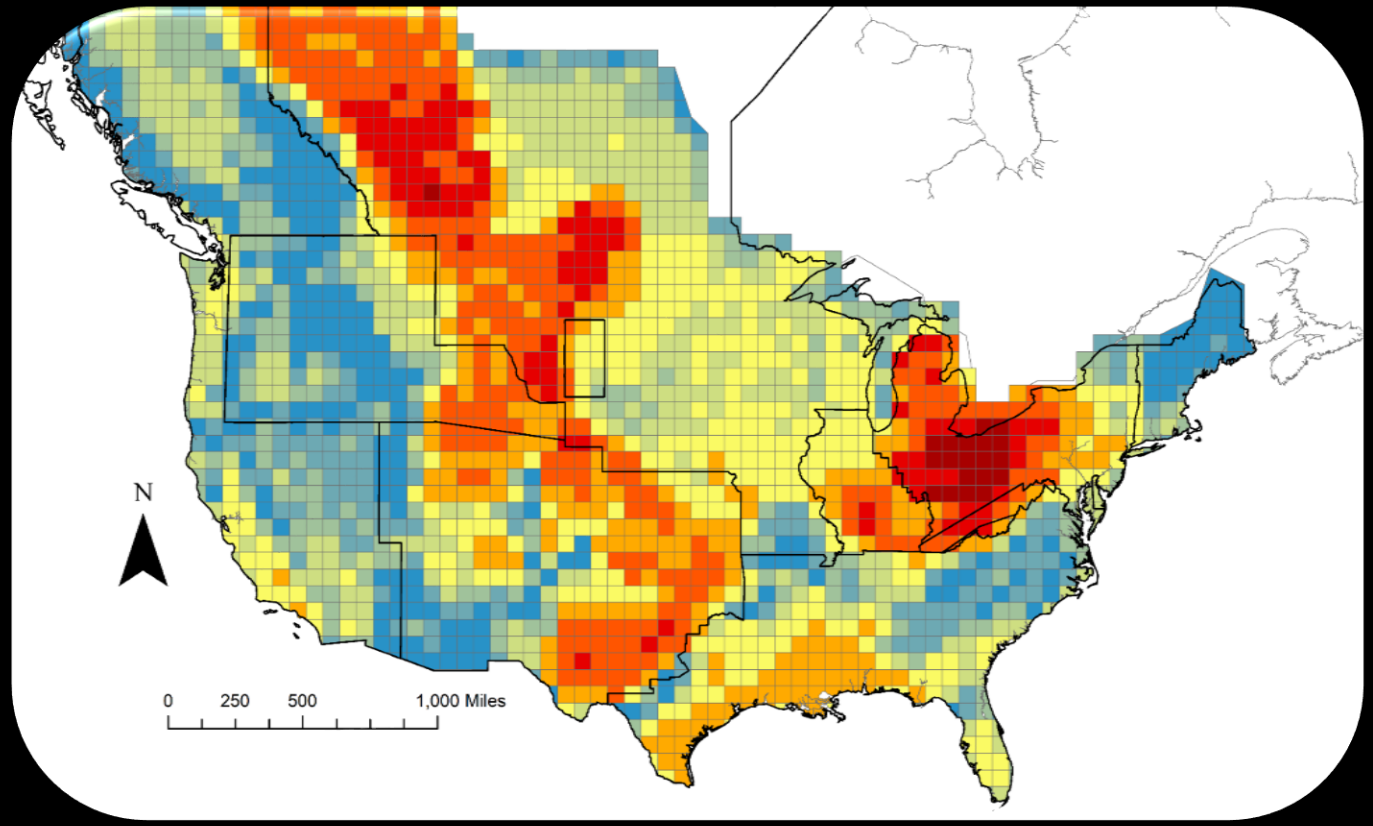
Using AI/ML to Curate Thousands of Carbon Storage Data Assets via EDX

Paige Morkner[1,2], Chad Rowan[1,3], Kelly Rose[1], Jennifer Bauer [1], Michael Sabbatino [1,2], Patrick Wingo[1,2], Andrew Bean[1,2]

NATIONAL ENERGY TECHNOLOGY LABORATORY

EDX Energy Data eXchange

GEO CUBE

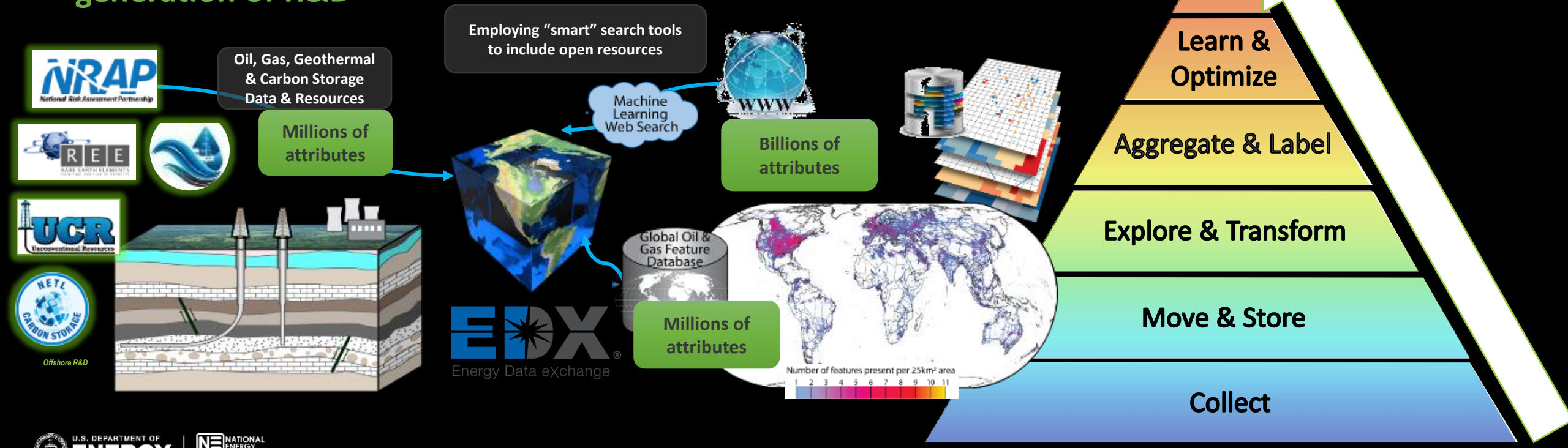


[1] National Energy Technology Laboratory, 1450 SW Queen Ave, Albany, OR, 97321, USA
[2] Leidos Research Support Team, 1450 SW Queen Ave, Albany, OR, 97321, USA
[3] Attain, 3610 Collins Ferry Rd. Morgantown, NETL 26505

Research is data-driven

- Millions of dollars in R&D products are now publicly available from carbon storage efforts
- There is a need to preserve and efficiently access those resources to drive the next generation of R&D

The virtual spatial and subsurface (VSS) data framework seeks to address the needs of the community through AI/ML enhanced methods via DOE's virtual data library and laboratory, EDX

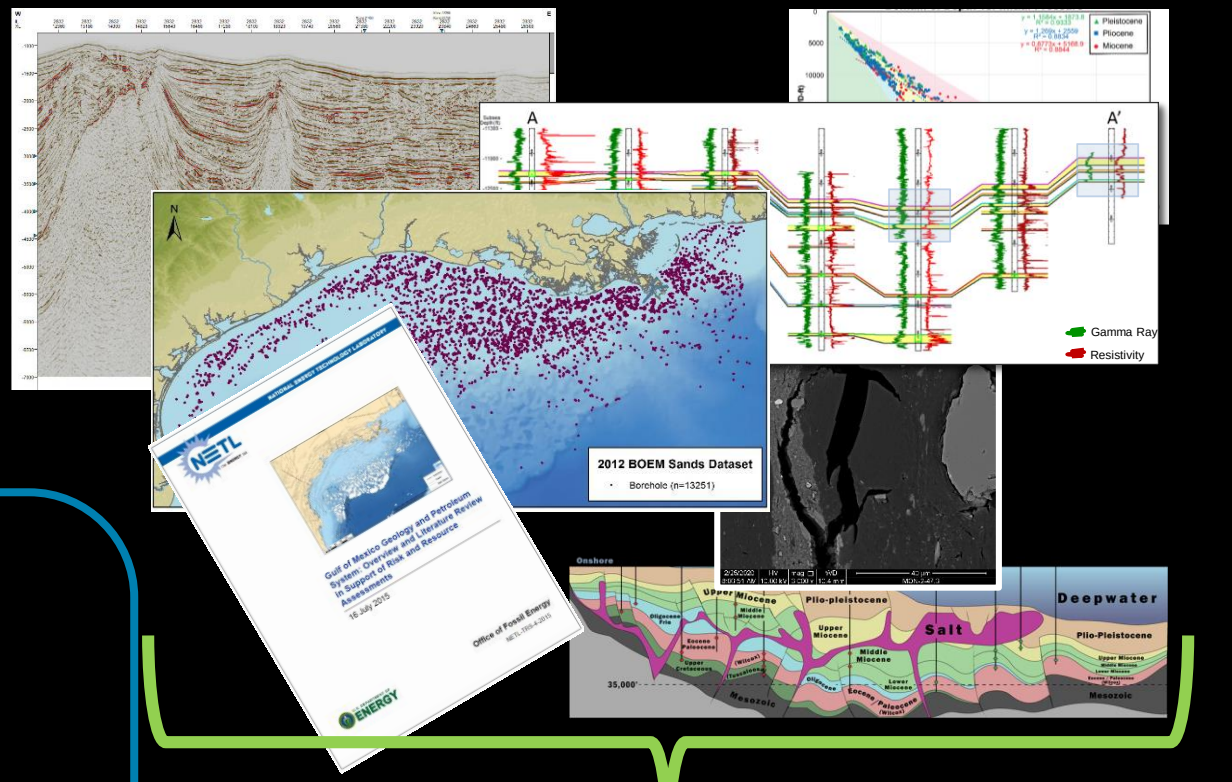


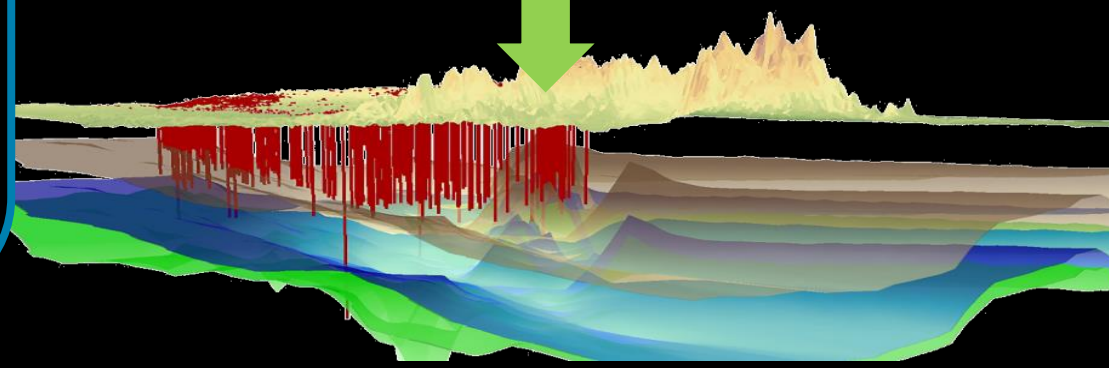
Using AI/ML, millions of data features and attributes have been integrated and preserved across the USA in support of advanced carbon storage projects

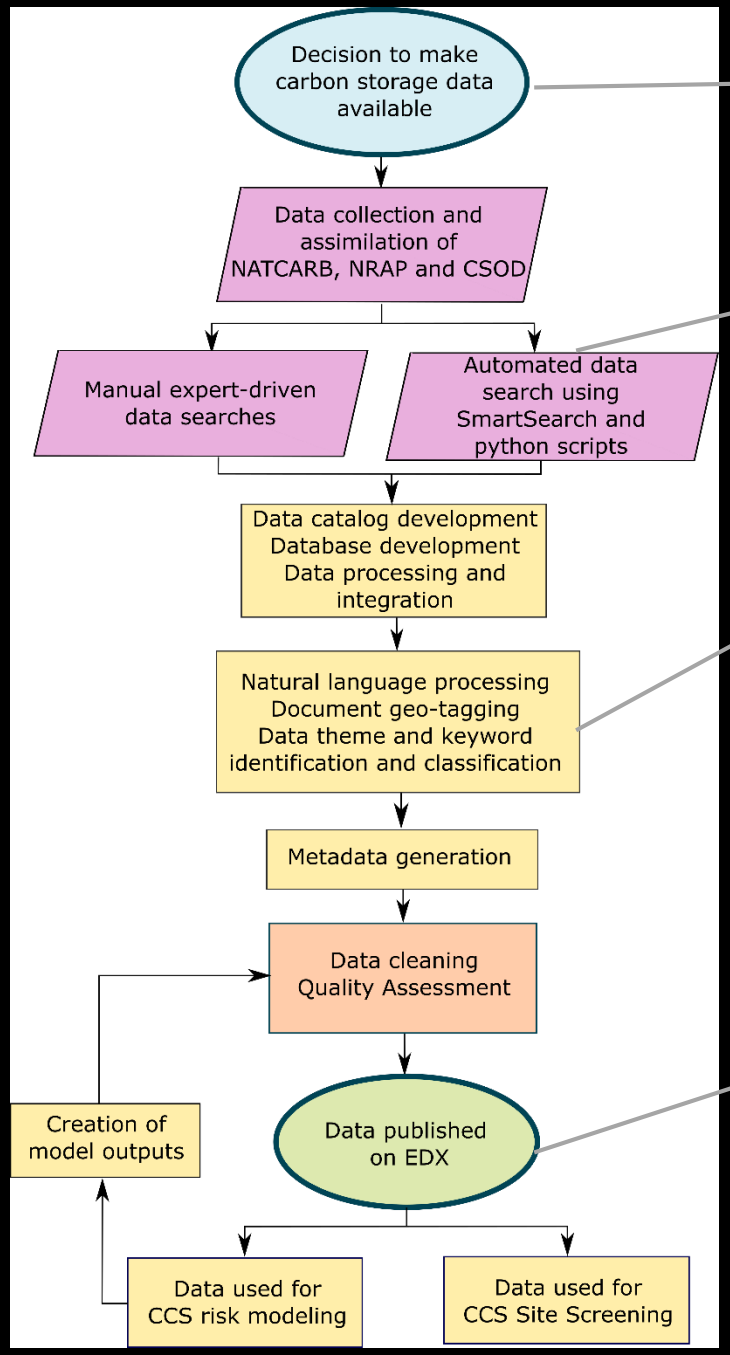
This effort has already aided SMART-CS, NRAP and outside entities (e.g. major industry operator) to drive subsurface modeling, machine learning, and insights for a range of end user needs

EDX supports:

- RCSP, CarbonSafe, NRAP data ingestion
- Data mining to aggregate authoritative, open source resources relevant to CS researchers
- Integration of other FE resources
- Access, visualization, and interaction with CS data collections via NETL EDX mapping platforms Natcarb Viewer and Geocube
- Reuse of data by new FE projects via EDX Collaborative Workspaces and more…







2016: CS program invested in helping their funded projects:

- Curate their data products
- Explore and transform data into data products
- Integrate data into databases with FAIR standards

2016-Present
RCSPs contribute data to EDX and push public

2018- 2019: Data rescue efforts for WESTCARB

Implement machine learning and natural language processing to:

2017: Virtual Sub Surface first envisioned and proposed

2017: SmartSearch, in beta testing, gathers and extracts relevant resources

2019: Natural Language Processing labeling and topic modeling

2019: SmartParse, in development, created

2019: Living Database development

2020: Geotagging development

Ongoing efforts to curate and catalog data:

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

All these results in:

- **Publishing of large amounts of data publicly on EDX to support missions across the FE portfolio**
- **Integration of data into GeoCube and Keyword search on EDX for enhanced searchability**
- **Use of Living Database to continually update data**

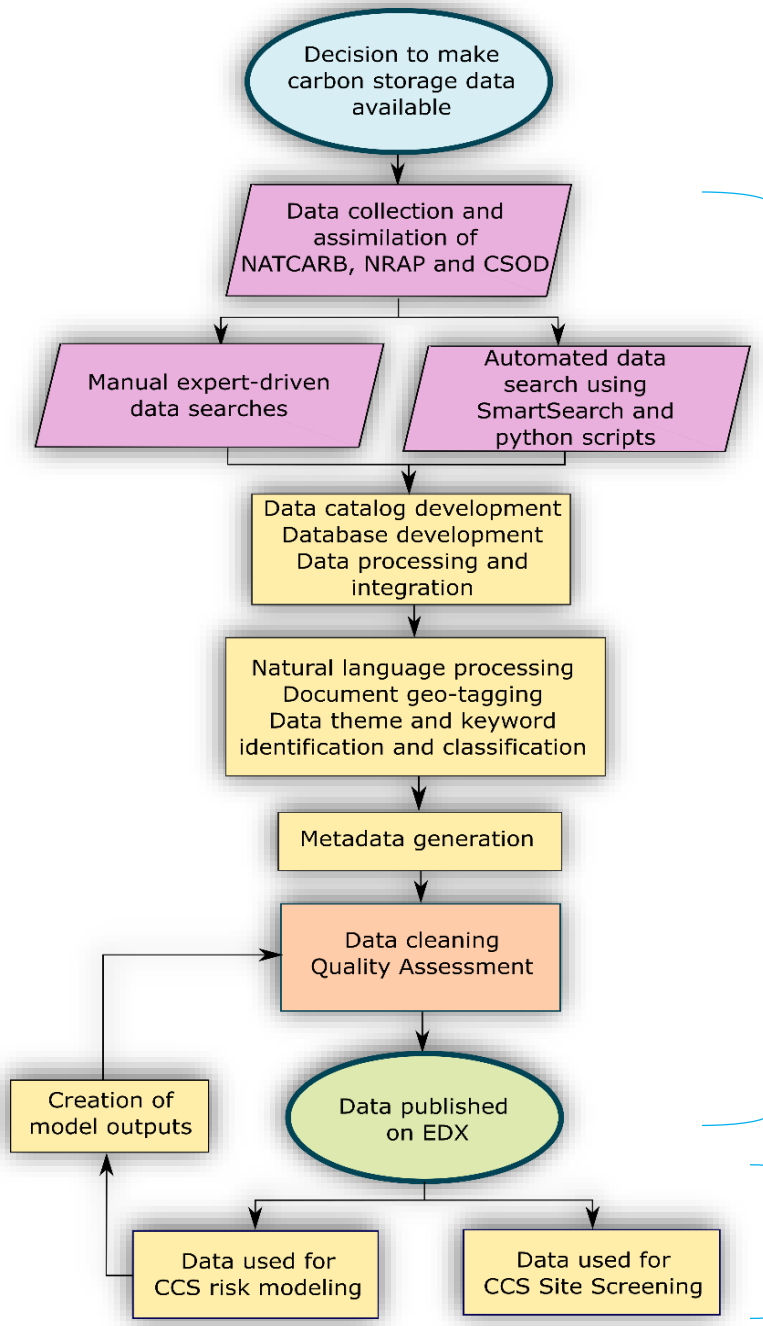
2016

2020

Morkner et al., submitted

The life cycle of data from collection to release to utilization

- Use of Energy Data eXchange (EDX) for data curation and collaboration
- The amount of data being published creates a need for intuitive data curation to increase discoverability and usability



Data for CS applications undergoes a process:

- **collection**
- **cataloging**
- **metadata development**
- **quality assessment**
- **publishing**

Then can data be reused for CS applications such as **modeling, simulations and site screening for projects like SMART-CS and NRAP**

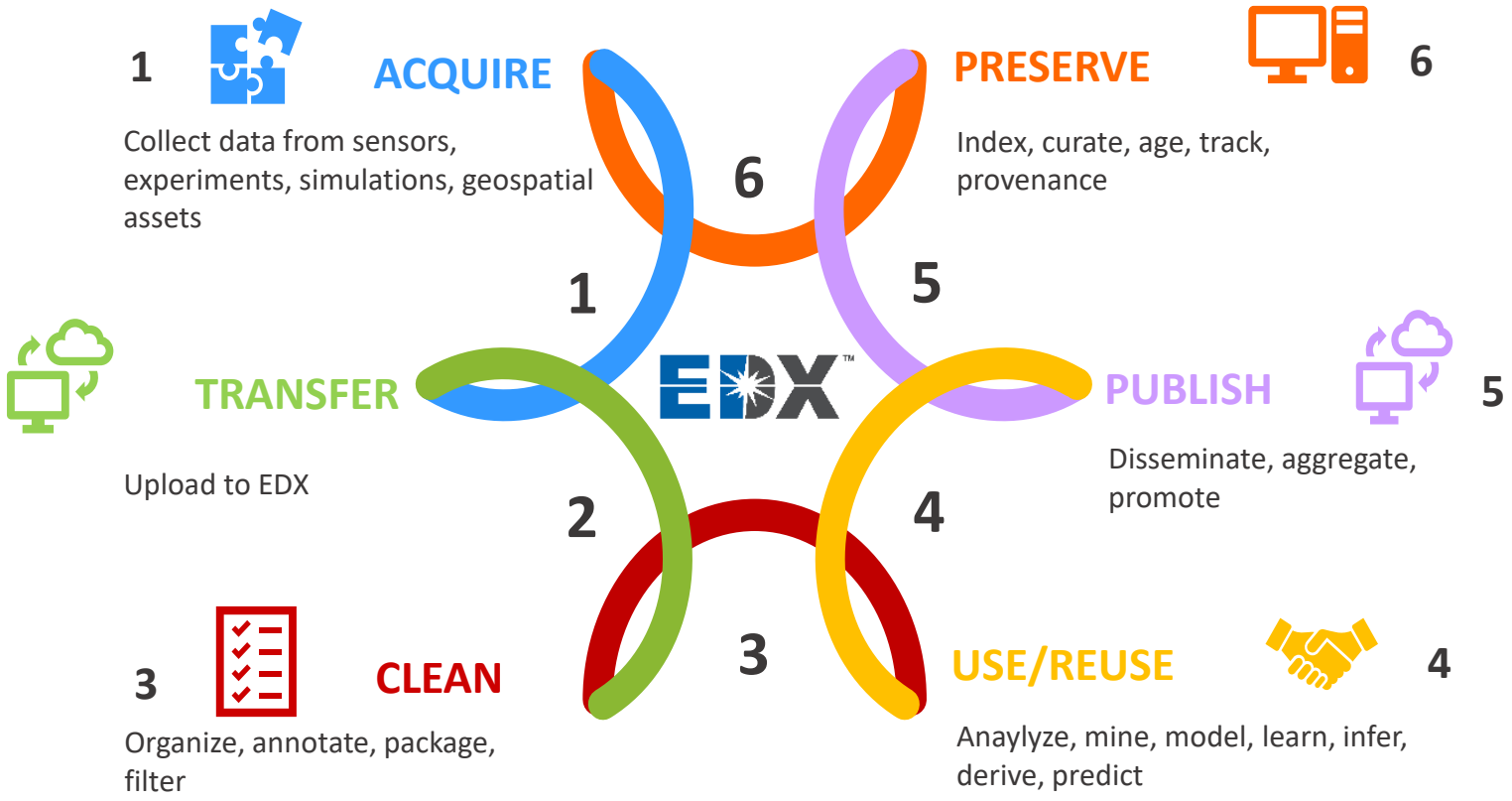
Morkner, P., Bauer, J., Creason, C., Sabbatino, M., Wingo, P, Greenburg, R., Walker, S., Yeates, B., and Rose, K., **in review,** Distilling Data to Drive Carbon Storage Insights, journal: *Computers & Geoscience*

EDX Supports the Entire Life-Cycle of Data

- EDX supports the entire life-cycle of data, presentations, publications, and tools.
- EDX has evolved to meet the needs of the DOE FE user community.
- EDX ensures users and resources within the platform align to Federal and DOE regulations and policies
- EDX is utilizing technologies such as machine learning, natural language processing and its very own Smart Search to enhance user data discoverability, integration, labeling and transformation.

edx.netl.doe.gov

Private Collaboration/Public Dissemination



www.presentationgo.com

Tiered Access Using Role-Based Security

Public

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

DOE-Only Workspaces

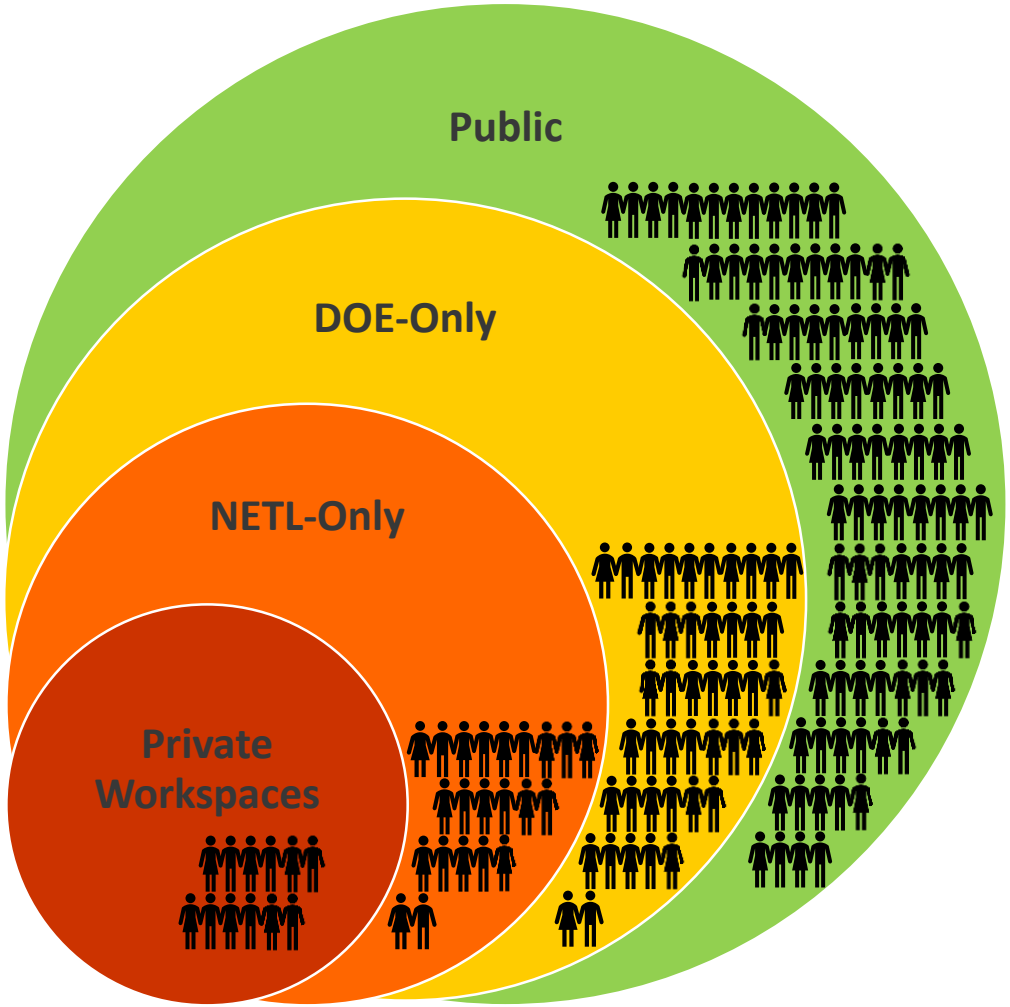
- Semi-private data
- All registered users from DOE Labs and DOE HQ have access

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



Find, Sort, Visualize, and Interact with Data

EDX
A Product of NETL

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** , **CSIL** and **NRAP Tools**.

Spatial

EDX Tools like **Geocube**, **Natcarb Viewer**, and **Blosom** allow users to find, sort, visualize, and interact with geospatial data.

Visualize

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

www.presentationgo.com

Types of Carbon Storage Data

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 Curated "To Date…"

these numbers keep growing

RCSP Data by the #'s:

- **1.7TB source data**
- **3065+ resources**
  - *879 Published EDX resources*
  - *632.8 GB of Published EDX resources*
  - *3,185 open data resources*
- **>4 million federated spatial data records**
- **BSCSP** and **PCOR** EDX **on track**
- **SECARB** EDX data ingestion at **Cranfield** CCS site **complete** and **pushed to public**
- **MRCSP & MGSC in progress**

NRAP Community Datasets CCS Site Catalog Curation to date:

- **19 sites cataloged**
- **7552 records including 6241 spatial**
- **Cataloging includes open source publications, data, and EDX submissions**
- Catalog will be integrated into EDX by end of September
- **Continual updates to catalog as new resources are published on EDX**
- **Releases of FutureGen and Kimberlina datasets**

Carbon Storage Open Database:

- **Scraped from public websites and ArcREST servers**
- **315 Spatial layers in EDX's GeoCube**
- **1846 text-based documents**

NATIONAL ENERGY TECHNOLOGY LABORATORY

NRAP National Risk Assessment Partnership

EDX Energy Data eXchange

NETL CARBON STORAGE

NatCarb A National Look at Carbon Sequestration

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

RCSP Collaborative Workspaces

RCSP public and private resources have a combined total of 3,065 resources and 1.72TB of data.

PRIVATE				PUBLIC		
CW Name	# of Submissions Waiting to Go Public	# of Resources	Data Usage	Submissions	Resources	Data Usage
Big Sky	0	87	1000.1GB	31	101	600.3GB
GOM-Carb	15	33	277.2MB	-	-	-
MRCSP Phase 1	0	4	32.7MB	3	4	32.7MB
MRCSP Phase 2	27	71	2.0GB	17	17	334.5MB
MRCSP Phase 3	15	56	70.39GB	-	-	-
PCOR Master Workspace	20	1377	4.4GB	120	711	1.9GB
SECARB – Anthropogenic Test	153	550	8.38GB	-	-	-
SECARB – Early Test	0	0	0.0GB	17	37	30.2GB
Southwest	0	6	33.1MB	3	9	53.7MB
WESTCARB	0	2	1.2GB	-	-	-
TOTAL	**230**	**2186**	**1086.8GB**	**191**	**879**	**632.8GB**

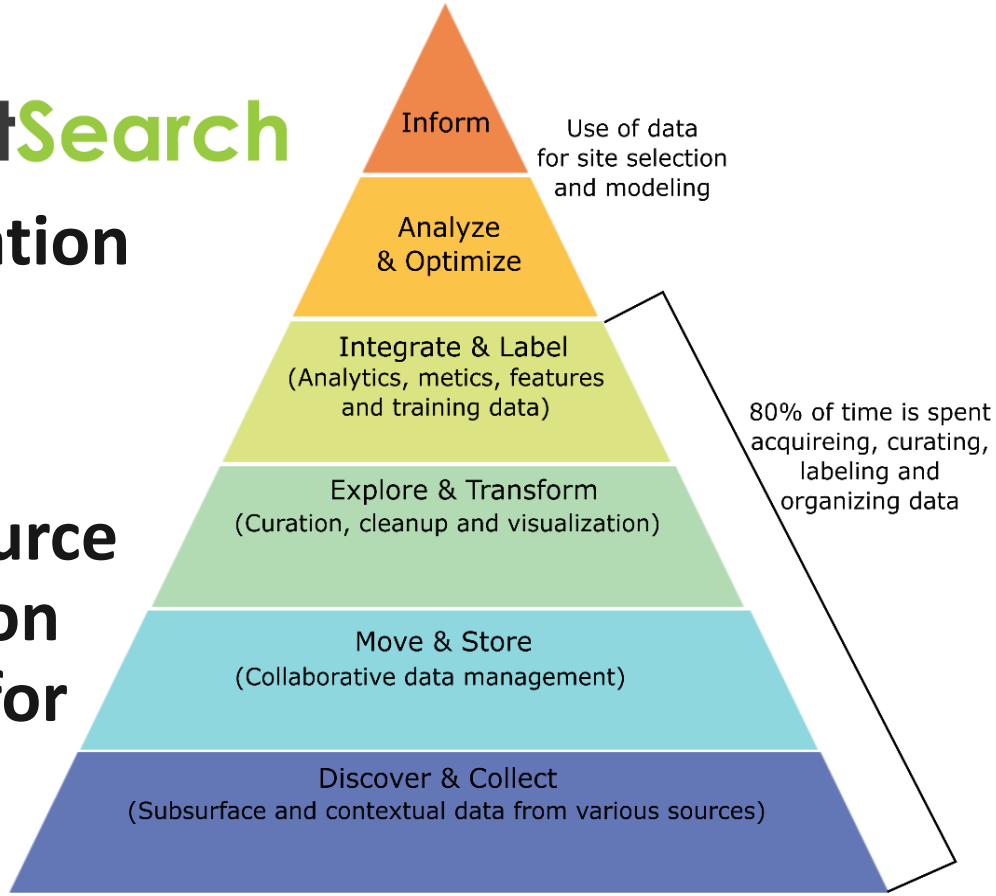
*updated 8/25/20

Use of AI/ML Tools for CS Data Curation

Challenge: Making available data discoverable, searchable, and easy to reuse

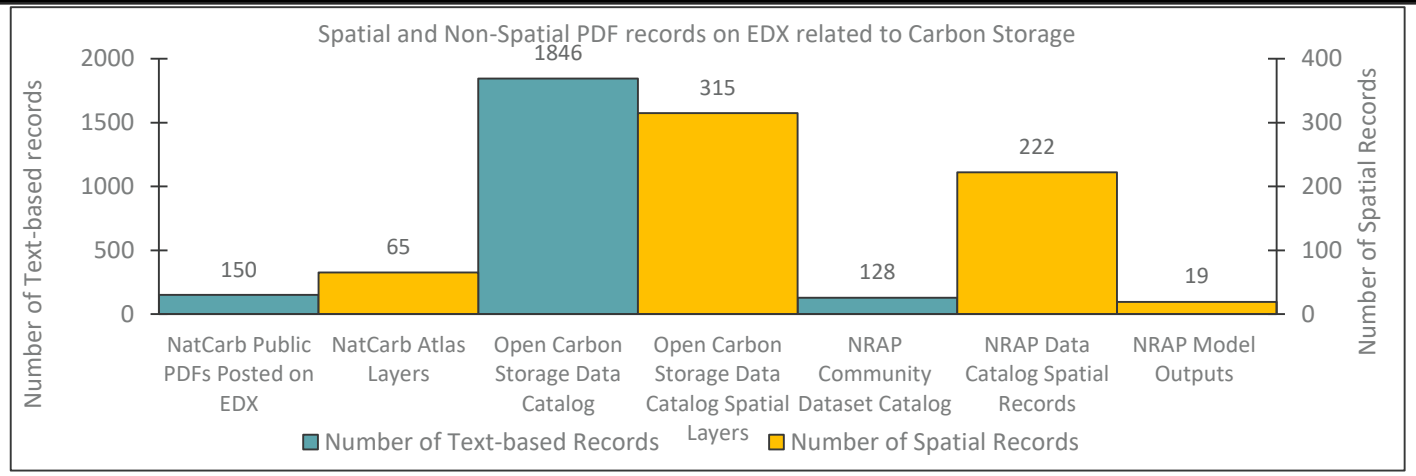
Solutions:

- **Open-source data scraping efforts**
- **Cataloging for metadata extraction and preservation**
- **Geographic database development to make searches easier (GeoCube)**
- **Natural language processing for text-based resource classification, organization, keyword identification (metdata building) and geographic association (for searchability)**



Carbon Storage Data Collection

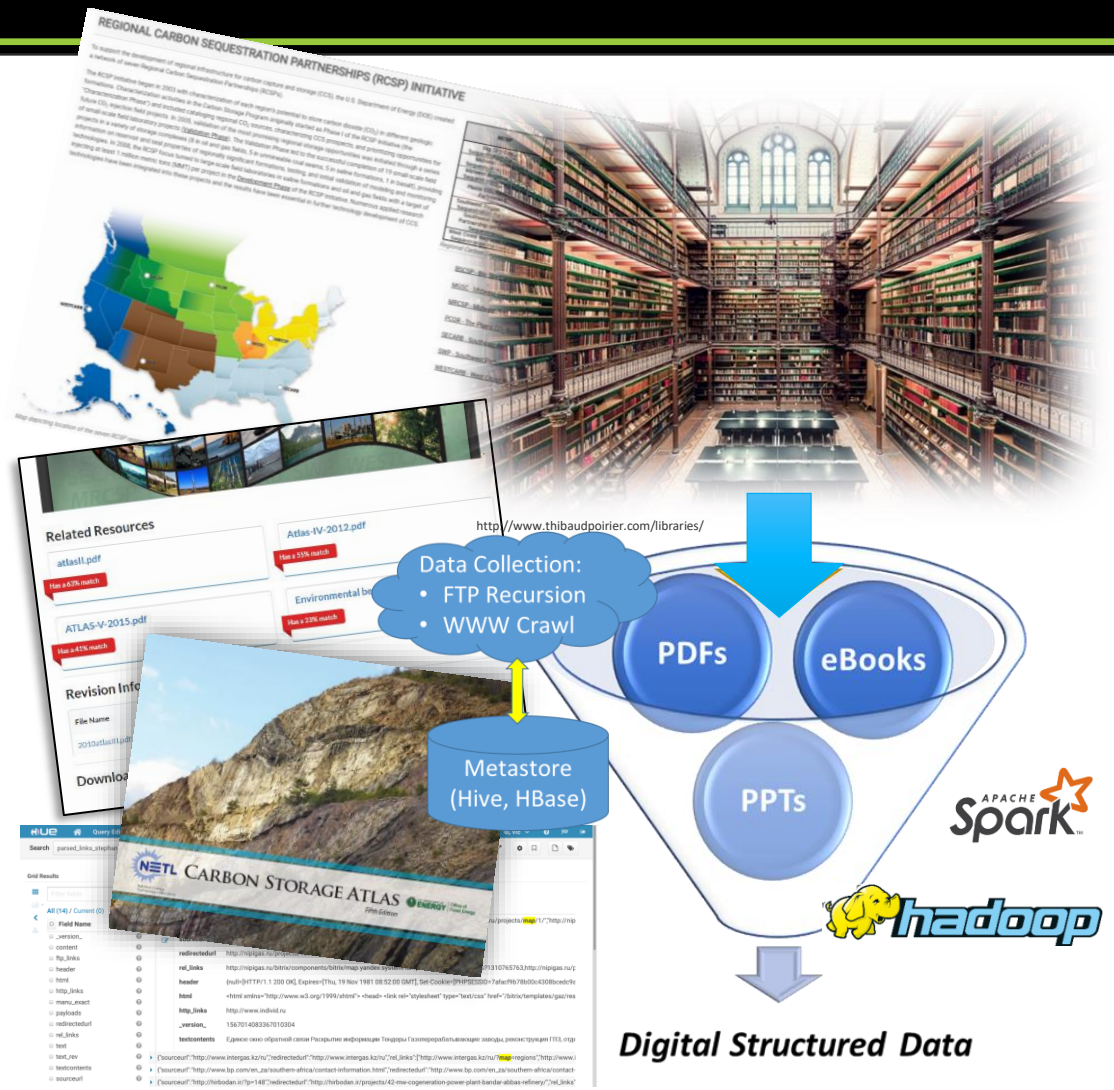
- **Scraping of data from public websites**
  - Development of **Carbon Storage Open Database** on EDX using python scripts
  - Scraping of data using **SmartSearch**
- **Addition of data directly to EDX by RCSPs and NRAP and others**
  - EDX Groups used for data collection organization
- **Data publishing from projects such as FutureGen and Kimberlina**
- **Living Database** – ML/AI tool for updating data in real time, in coordination with SmartSearch



GEO CUBE NATCARB viewer EDX Energy Data eXchange

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*

NETL's SmartSearch, a big data, algorithm



SmartSearch core features:

- Massively Parallel Operations:
  - Parsing many file formats (at scale) -- uses Tika, can be extended to additional formats beyond Tika
  - Parsing of zips (parses content of nested zips)
  - Cluster based large scale data management
  - ML processing (both Spark ML and Spark NLP) for 1) parallel NLP processing, 2) ML (recommendation engine) using Sparse and Dense Vectors and easily implement myriad of ML pipelines.
- Web crawling / indexing
- FTP crawling / indexing
- Data discovery -- combine above with web APIs (search engine, USPTO, etc) to automate data discovery and identify relevant data from seed(s)
- Developing Spark ML with GPU on NETL ML cluster.
- Developing integrations with Databricks GPU Spark ML processing.

Developing Data Catalogs

Identifying Data Relevant to & Produced By the CS R&D Community

- **Builds a data foundation for CS community**
  - Inventories **what is available and where**
- **Documentation of resource metadata and inputs**
  - Metadata preservation about data such as extractable data attributes, file size, spatial extent, etc
- **ML tools can be used** to set up foundation for cataloging and **to discover data for catalog integration**

NRAP Community Datasets CCS Site Catalog

File Edit View Insert Format Data Tools Add-ons Help Last edit was 34 minutes ago

Tool input per data category for each specific tool

NRAP Tool	AIM							DREAM	
Data Category	Aquifer Characteristics, Confined	Aquifer Characteristics, Unconfined	Time	Porosity or Permeability	Geochemistry	CO2	Brine properties	Location (other)	Reservoir Charactersitics
Data Subcategory	sand fraction []	correlation length [km]	migration time [yr]	permeability sand [log10(m2)]	calcite volume fraction	CO2 flow [kg/s]	brine flow [kg/s]	x, y, z (within a leakage	flow simulation output parameter value* (one or
	correlation length X [m]	Kx/Ky							
	correlation length Y [m]	aquifer thickness [m]							
	density sand [kg/m3]	horizontal hydraulic gradient []							

NATCARB - RCSP Open Data Catalog

File Edit View Insert Format Data Tools Add-ons Help Last edit was made 10 days ago by Michael Sabbatino

RCSP	Data Availability	Status	Data Type	Data Products	Host (spatial data only)
Big Sky Carbon Sequestration Partnership (**BSCSP**)	Y	Acquired	Spatial	feature	ArcGIS REST Services (BSCSP)
	Y	Acquired	Non-Spatial	newsletters, graphics, press releases, posters, journal articles, videos	-
Midwest Geological Sequestration Consortium (**MGSC**)	N	N/A	Spatial	N/A	N/A
	Y	Acquired	Non-Spatial	maps, posters	-
Midwest Regional Carbon Sequestration Partnership (**MRCSP**)	Y	Acquired	Spatial	feature, table, raster	ArcGIS REST Services (ODNR)
	Y	Acquired	Non-Spatial	reports, fact sheets, presentations	-
Plains CO2 Reduction Partnership (**PCOR**)	Y	Acquired	Spatial	feature, table, raster	ArcGIS REST Services (PCOR)
	Y	Pending	Non-Spatial		-

Example: NRAP Community Datasets CCS Site Catalog

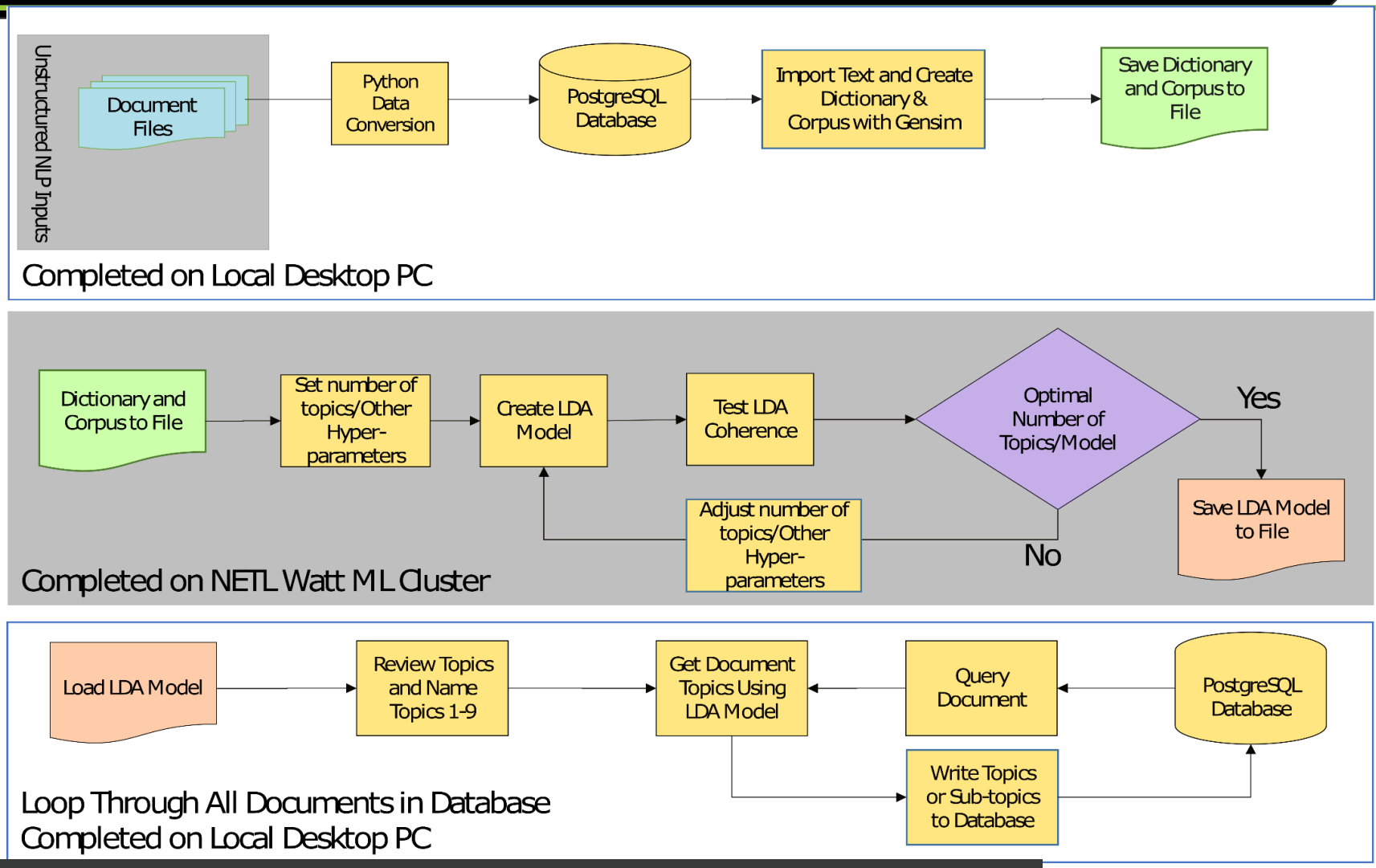
- To date: **19 sites have been cataloged, incorporating open source publications, data, and EDX submissions**

SITE	RECORDS (ALL)	SPATIAL
1. Big Sky - Basalt Injection	37	3
2. CaMI Field Research Site	15	0
3. CarbonSAFE - Wyoming (Rock Springs Uplift)	14	0
4. Citronelle	38	31
5. Decatur	34	17
6. Edwards Aquifer	40	11
7. Farnsworth - Anadarko Basin	31	0
8. Future_Gen	7087	6109
9. High Plains Aquifer	6	2
10. Kimberlina	65	51
11. MRCSP - Appalachian Basin Test	19	0
12. MRCSP - Cincinnati Arch Test	36	0
13. PCOR - Williston Basin Oil Field Test	24	0
14. SACROC Oil Field Site	29	4
15. SECARB - Cranfield Site	68	13
16. SECARB - Central Appalachian Basin Test	9	0
17. Kevin Dome (BSCSP)	41	40
18. Bell Creek (PCOR)	66	14
19. Cristian County CarbonSAFE (MRCSP)	3	0
TOTAL RESOURCES CATALOGED	**7552**	**6241**

Natural Language Processing (NLP)

A case study used for CS text-based resources

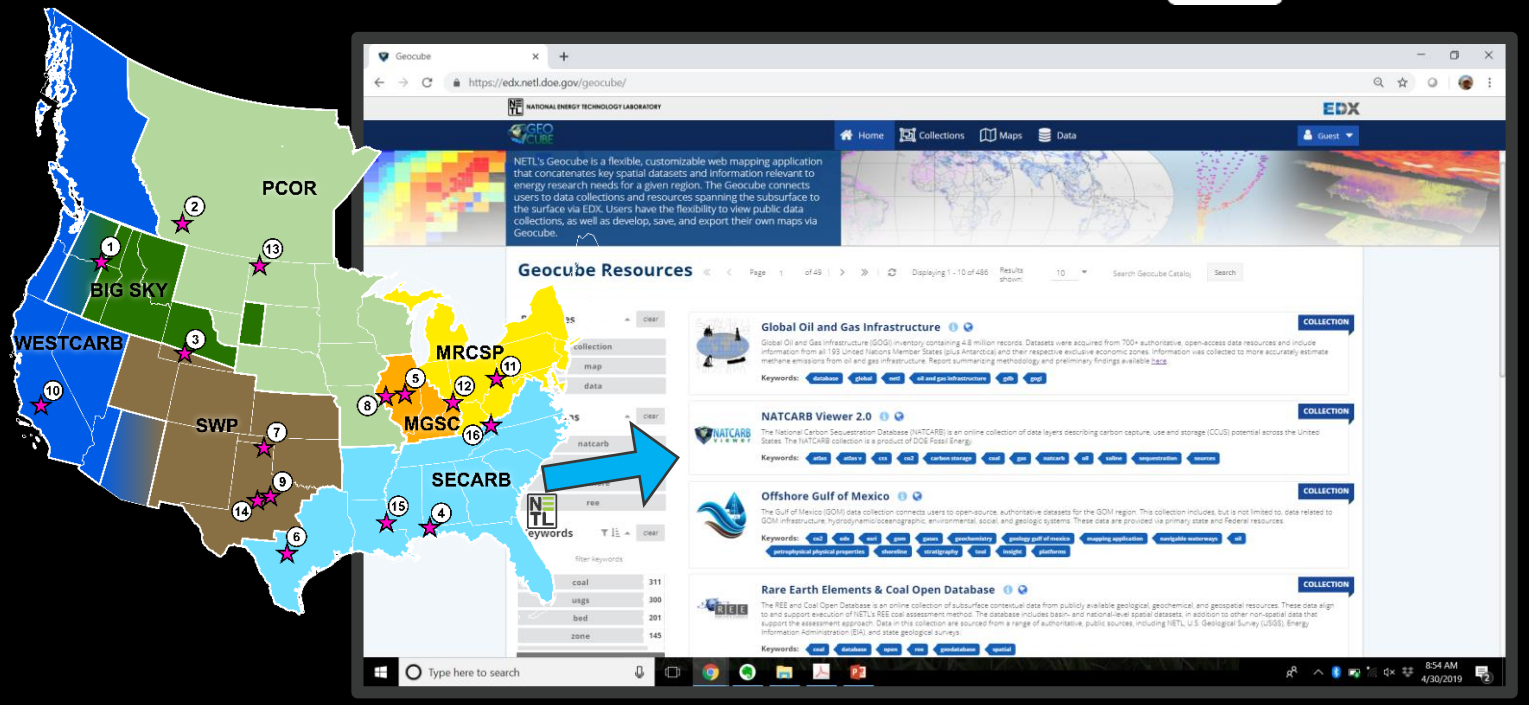
- Latent Dirichlet allocation **(LDA) model based on corpus of 2071 text-based documents**
- Topic names assigned by subject-matter experts
- **Each document is classified** by % of each topic it's associated with
- **Each document has 50+ keywords identified** and can be **associated metadata on EDX**
- **Parse geographic location to associate with each document** – when possible

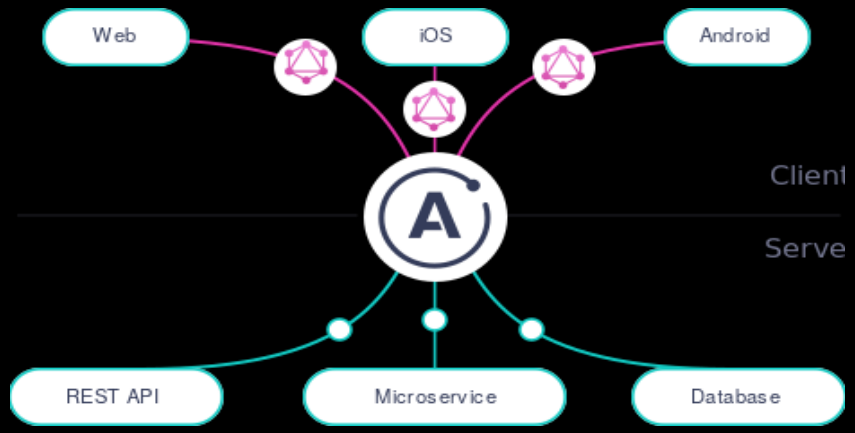


Morkner, P., Bauer, J., Creason, C., Sabbatino, M., Wingo, P, Greenburg, R., Walker, S., Yeates, B., and Rose, K., **in review,** Distilling Data to Drive Carbon Storage Insights, journal: *Computers & Geoscience*

GeoCube 3.0 – CS spatial data search & visualization

- **Upcoming GeoCube enables spatial search capabilities – users can rapidly access and visualize data on an interactive map for the world**
  - **Natcarb Viewer 2.0**
  - **Make accessible relevant supplemental resources from across the FE portfolio - Drives citation and reuse of data**
- **Data inputs** needed for tools and models can be derived from datasets
- **Data outputs** from modeling and field studies can be searched for by site or region
- **Outside resources and modeling tools** can be incorporated



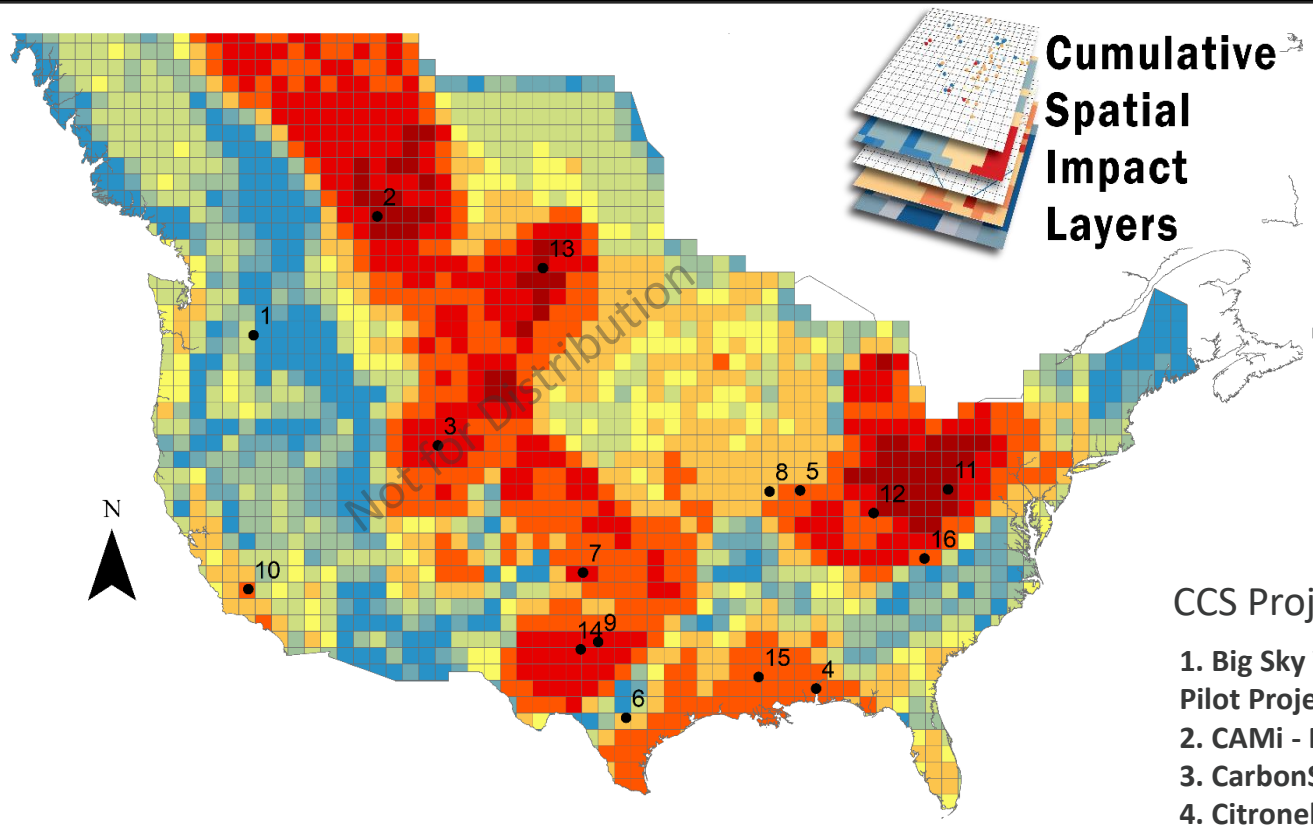


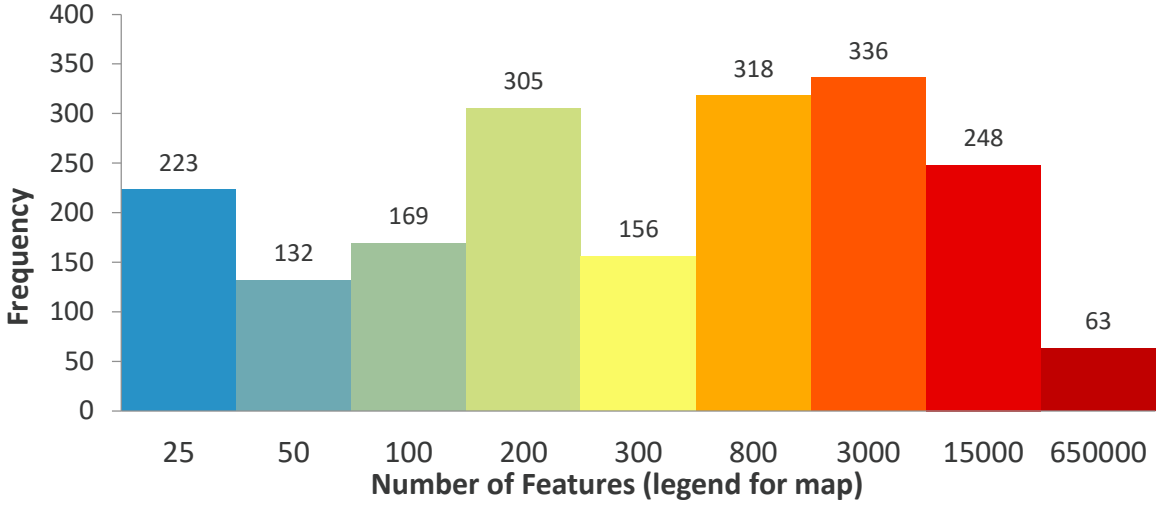
GraphQL API

New GraphQL/Aollo Server API provides a unified interface for querying numerous backing data sources

Results: Spatio-temporal trends in CS data

Bringing together NatCarb, NRAP catalog and the Carbon Storage Open Database





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

CCS Projects Cataloged

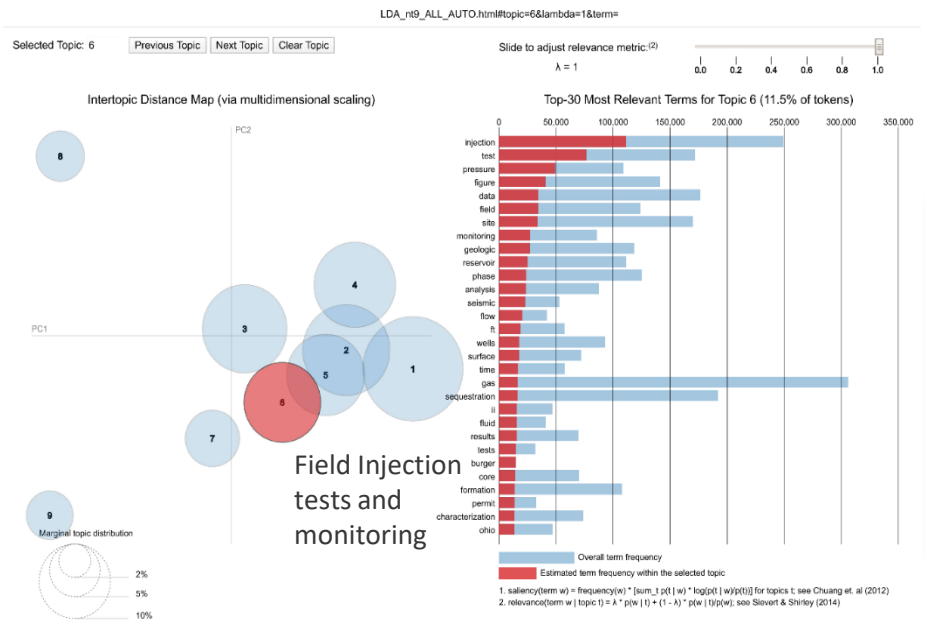
1. **Big Sky Validation Phase - Wallula Basalt Pilot Project**
2. **CAMi - Field Research Station**
3. **CarbonSAFE – Wyoming**
4. **Citronelle (SECARB)**
5. **Decatur**
6. **Edwards Aquifer**
7. **Farnsworth - Anadarko Basin**
8. **FutureGen**
9. **High Plains Aquifer**
10. **Kimberlina (WESTCARB)**
11. **Appalachian Basin Test (MRCSP)**
12. **Cincinnati Arch Test (MRCSP)**
13. **Williston Basin Oil Field Test (PCOR)**
14. **Scurry Area Canyon Reef Operations**
15. **Cranfield Site (SECARB)**
16. **Central Appalachian Basin Test (SECARB)**

Paige Morkner, Paige.Morkner@netl.doe.gov
PI: Kelly Rose, Kelly.Rose@NETL.doe.gov

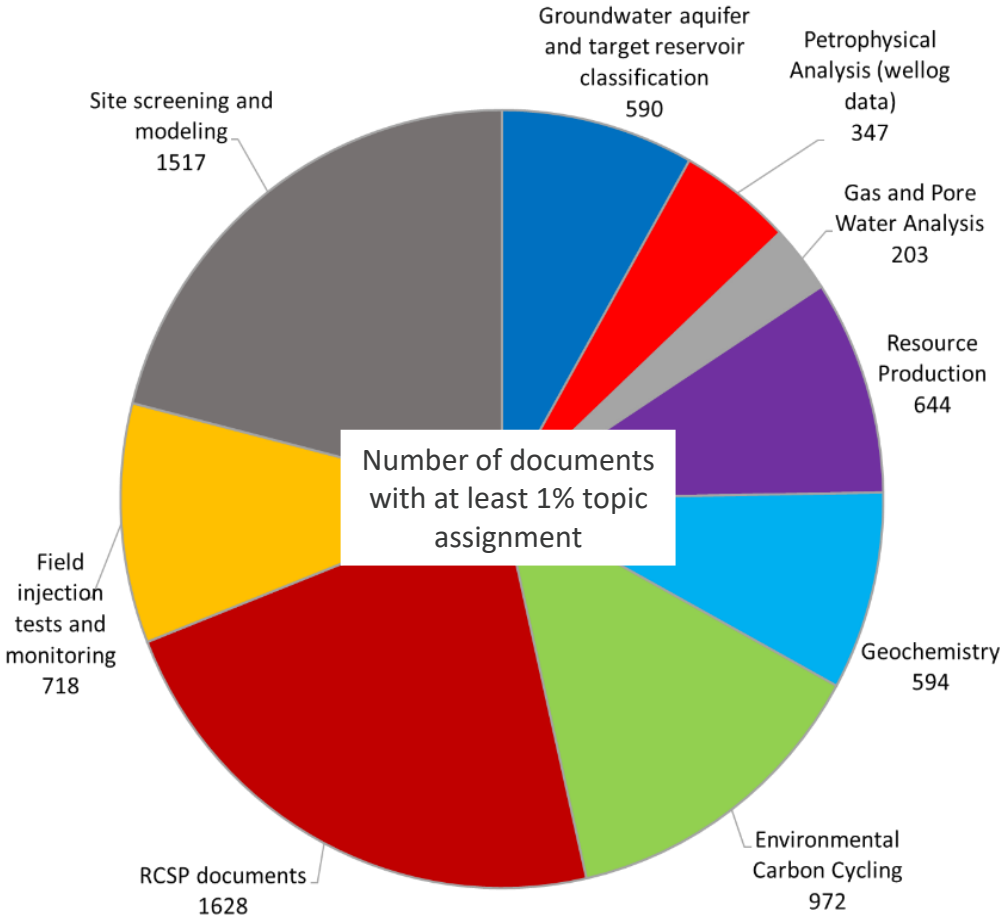
Results: Natural Language processing

Keywords and geographic associations

- Produced a **9 topic LDA model** – grouping similar papers
- Produced **keywords** associated with resources
- Geographic location recognition (in progress)
- Integration into EDX through NETL SmartParse



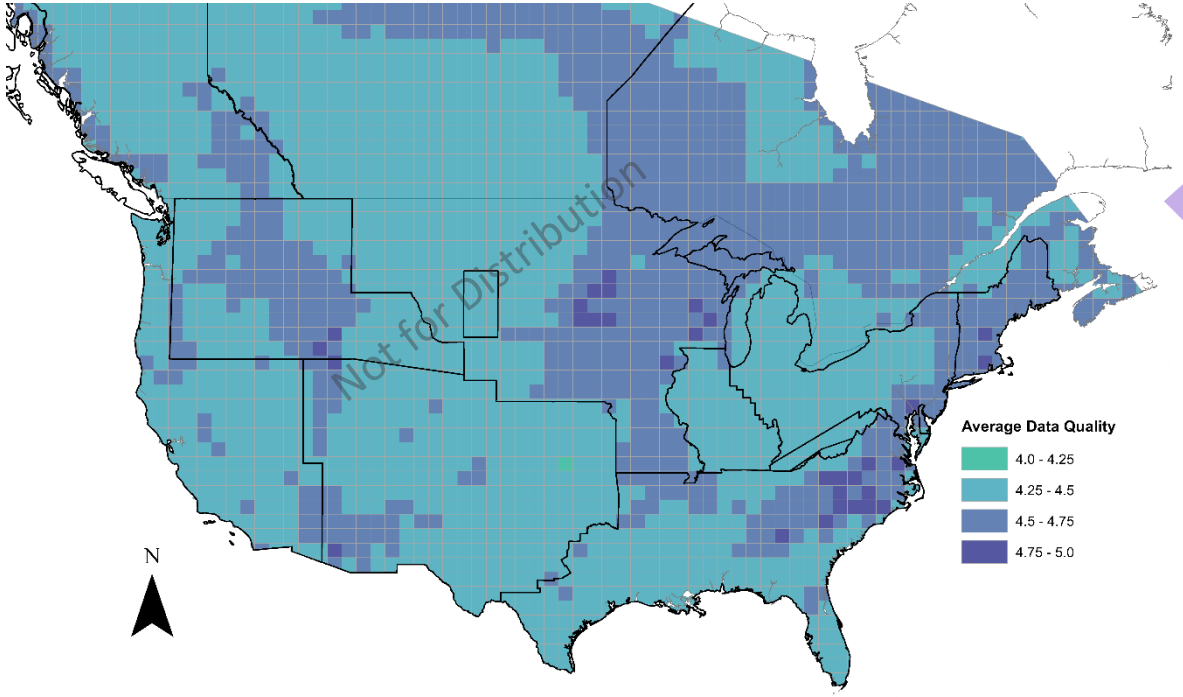
spaCy

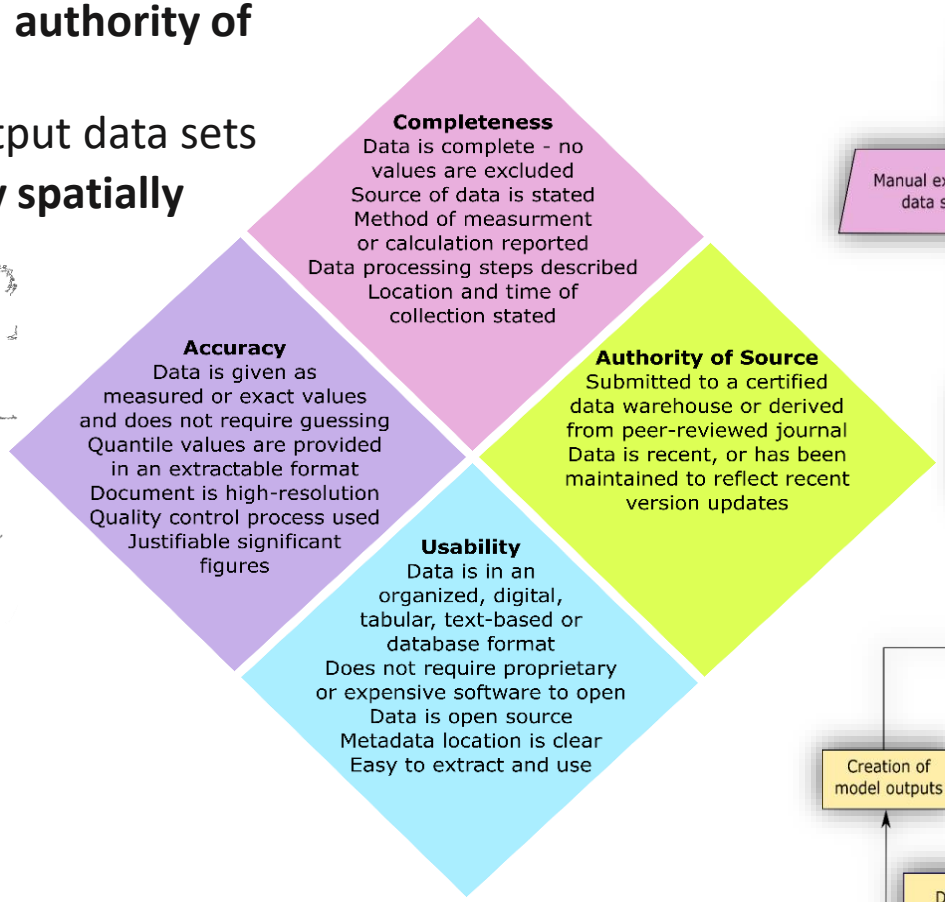


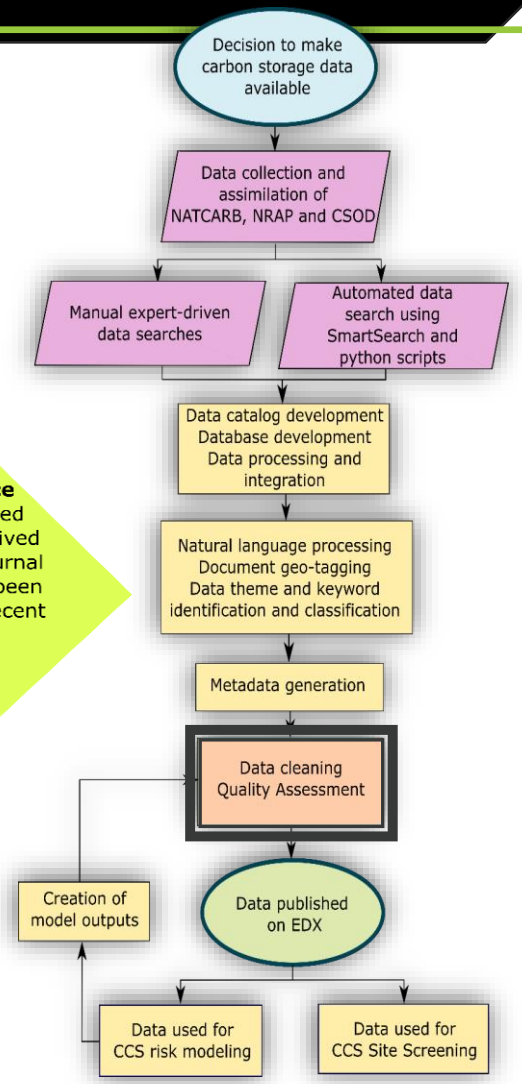
Morkner, P., Bauer, J., Creason, C., Sabbatino, M., Wingo, P, Greenburg, R., Walker, S., Yeates, B., and Rose, K., **in review,** Distilling Data to Drive Carbon Storage Insights, journal: *Computers & Geoscience*

Results: Data Quality assessment method development and spatial trends in CS data quality

- **5-point data quality assessment method** developed
- Quality based on **completeness, accuracy, usability,** and **authority of source**
- **Applicable to many subsurface data sets** and model output data sets
- Combined with CSIL can be **used to analyze data quality spatially**





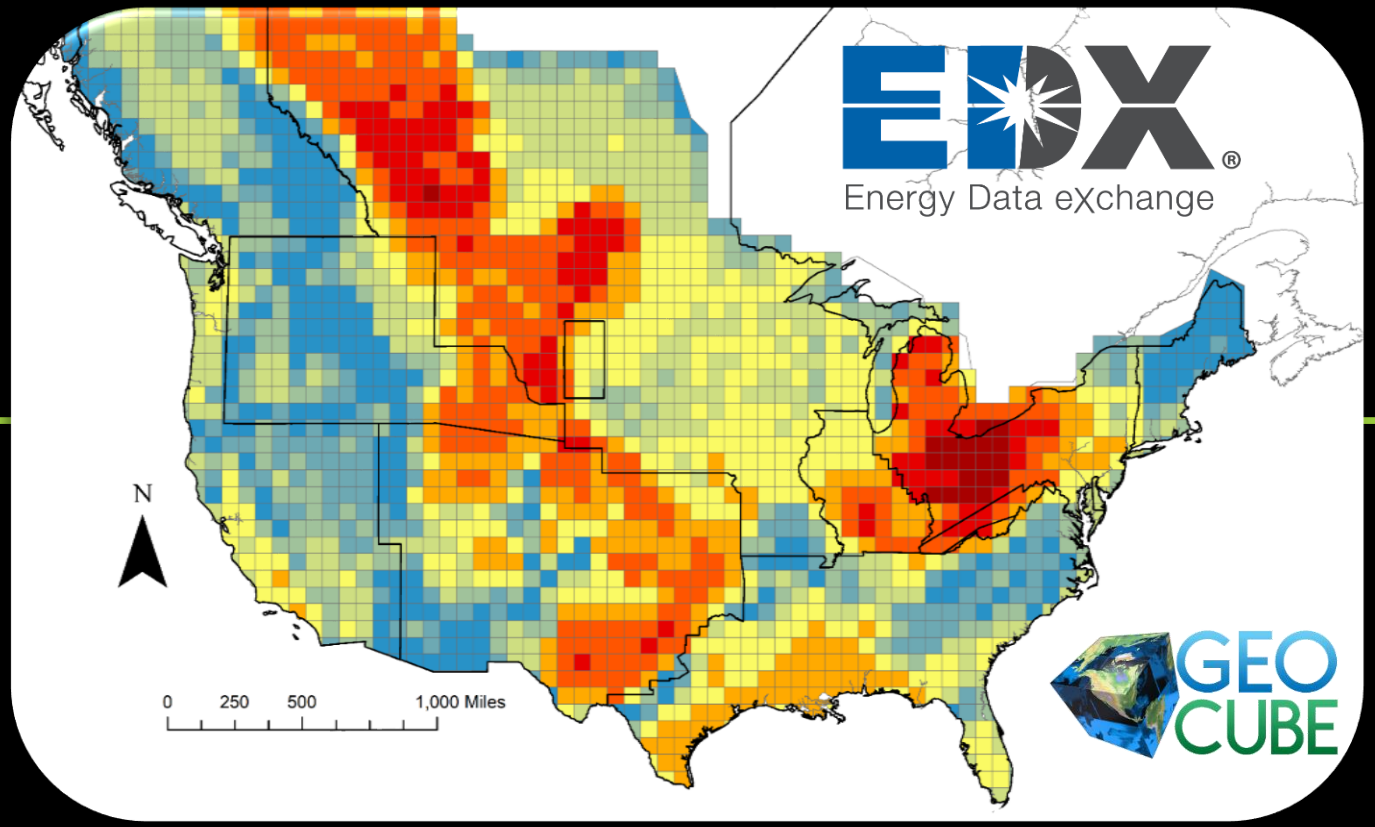


Morkner, P., Bauer, J., Creason, C., Sabbatino, M., Wingo, P, Greenburg, R., Walker, S., Yeates, B., and Rose, K., **in review,** Distilling Data to Drive Carbon Storage Insights, journal: *Computers & Geoscience*

Summary and Next Steps

FE and Carbon Storage program investments into data curation and management has led to 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:

- **A better understanding of CS relevant open- data density and data quality throughout US and Canada**
- **Improved access through the integration of CS data resources on EDX into GeoCube, SmartSearch and SmartParse (EDX version of NLP tools presented here) for further searchability with spatial searches and keyword searches**
  - Updates to GeoCube for enhanced spatial searchability and integration of modeling tools to come
- **EDX AI/ML data discovery, labeling, integration tool developments trained to support Carbon Storage, SMART-CS, and NRAP**
  - Deployment of AI/ML algorithms to allow on-demand data discovery and integration, ready-made for each end-user needs



Thank you!

NATIONAL ENERGY TECHNOLOGY LABORATORY

Contacts:

Paige Morkner, paige.Morkner@netl.doe.gov

Chad Rowan, chad.rowan@netl.doe.gov

Kelly Rose, kelly.rose@netl.doe.gov

https://edx.netl.doe.gov/group/?q=rcsp&sort=title+asc

https://edx.netl.doe.gov/geocube/#collections/carbonstorage

Acknowledgement: Parts of this technical effort were performed in support of the National Energy Technology Laboratory's ongoing research under the Carbon Storage Field Work Proposal DE-FE-1022465 by NETL's Research and Innovation Center, including work performed by Leidos Research Support Team staff under the RSS contract 89243318CFE000003 and the ITSS contract DE-DT0013924 Information Technology Support Services.

Using AI/ML to Curate Thousands of Carbon Storage Data Assets via EDX

Carbon Storage DE FE-1022465

Paige Morkner[1,2] and Chad Rowan[1,3], Kelly Rose, Jennifer Bauer, Michael Sabbatino, Andrew Bean

1. National Energy Technology Laboratory
2. Leidos Research Support Team, NETL
3. Attain, NETL

U.S. Department of Energy

National Energy Technology Laboratory

Carbon Capture Front End Engineering Design Studies and CarbonSafe

2020 Integrated Review Webinar

August-17-19 2020

Program Overview (1-2 Slides)

- Funded by DOE as part of Carbon Storage DE FE-1022465, Tasks 27 and 28
- RSS Contract and ITSS contract researchers
- Ongoing performance dates 2018-2022
- Project Participants
  - PI: Kelly Rose
  - LRST: Paige Morkner, Michael Sabbatino, Andrew Bean, Lucy Romeo, Patrick Wingo
  - ITSS: Chad Rowan, TJ Jones, Aaron Barkhurst, Vic Baker

Technology Section

- Task 27 supports the development of data, materials, maps, analyses, and figures for the Carbon Storage Atlas, Natcarb Viewer, and Natcarb database. This includes release of new data insights to the GCS community, through the sixth edition of the Carbon Storage Atlas, and through bi-annual updates to the Natcarb Viewer and Natcarb database.
- Task 28 focuses on addressing CS R&D data curation challenges associated with ingesting, describing, and curating data products from DOE FE to ensure enduring access and more efficient utilization of those resources using AI/ML enhanced approaches to support future CS R&D. Ultimately, this effort will result in tools, data resources, and virtual capabilities for the CSP and community to facilitate efficient CS data discovery, integration, and curation using NETL's EDX
- Use of EDX and development of tools to support the collection, curation, organization, labeling, and publishing large quantities of data for carbon storage. Whether laboratory, field, or computational, CS R&D is both a producer and consumer of data resources (datasets, tools, models, etc.). However, while the volume of open, online data is increasing exponentially, scientists struggle to find, access, and make operable data products from previous R&D projects due to insufficient and/or burdensome online data curation tools and outdated techniques.

FY2020 Accomplishments

a. **Energy Data eXchange hits major milestones – June 2020 surpassed 2 million resources downloaded** mark, and continues to support the preservation of over 2 billion dollars of federally funded research products
b. **EDX brand granted registered trademark by USPTO**
c. Completed and submitted manuscript outlining spatio-temporal trends of open CS data on EDX. "Morkner, P., Bauer, J., Creason, C., Sabbatino, M., Wingo, P., Greenburg, R., Walker, S., Yeates, D., Rose, K. *Submitted.* **Distilling Data to Drive Carbon Storage Insights**. *Computers & Geosciences***.**
d. Made advances in connecting SmartSearch and Living Database to make real-time updates to databases
e. Release of Five Kimberlina Oil Field model simulations on EDX - https://edx.netl.doe.gov/group/kimberlina-data
f. Release of 64GB of FutureGen 2.0 Subsurface Technical Data to EDX - https://edx.netl.doe.gov/group/futuregen-data

Progress and Current Status of Project

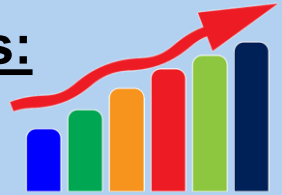
Outline:

- Carbon Storage Data Curation to date
- EDX Data statistics for carbon storage
- Use of AI/ML Tools for Data Curation
- NETLs SmartSearch algorithem
- Natural Language Processing
- GeoCube
- Results of carbon storage data spatio-temporal analysis
- Results of natural language processing
- Results of data quality analysis method development and spatial analysis

Carbon Storage Data Curated "To Date…"

these numbers keep growing

RCSP Data by the #'s:



- **1.7TB source data**
- **3065+ resources**
  - *879 Published EDX resources*
  - *632.8 GB of Published EDX resources*
  - *3,185 open data resources*
- **>4 million federated spatial data records**
- **BSCSP** and **PCOR** EDX **on track**
- **SECARB** EDX data ingestion at **Cranfield** CCS site **complete** and **pushed to public**
- **MRCSP & MGSC in progress**

NRAP Community Datasets CCS Site Catalog Curation to date:

- **19 sites cataloged**
- **7552 records including 6241 spatial**
- **Cataloging includes open source publications, data, and EDX submissions**
- Catalog will be integrated into EDX by end of September
- **Continual updates to catalog as new resources are published on EDX**
- **Releases of FutureGen and Kimberlina datasets**

Carbon Storage Open Database:

- **Scraped from public websites and ArcREST servers**
- **315 Spatial layers in EDX's GeoCube**
- **1846 text-based documents**

NatCarb
A National Look at Carbon Sequestration

RCSP Collaborative Workspaces

RCSP public and private resources have a combined total of 3,065 resources and 1.72TB of data.

PRIVATE				PUBLIC		
CW Name	# of Submissions Waiting to Go Public	# of Resources	Data Usage	Submissions	Resources	Data Usage
Big Sky	0	87	1000.1GB	31	101	600.3GB
GOM-Carb	15	33	277.2MB	-	-	-
MRCSP Phase 1	0	4	32.7MB	3	4	32.7MB
MRCSP Phase 2	27	71	2.0GB	17	17	334.5MB
MRCSP Phase 3	15	56	70.39GB	-	-	-
PCOR Master Workspace	20	1377	4.4GB	120	711	1.9GB
SECARB – Anthropogenic Test	153	550	8.38GB	-	-	-
SECARB – Early Test	0	0	0.0GB	17	37	30.2GB
Southwest	0	6	33.1MB	3	9	53.7MB
WESTCARB	0	2	1.2GB	-	-	-
TOTAL	230	2186	1086.8GB	191	879	632.8GB

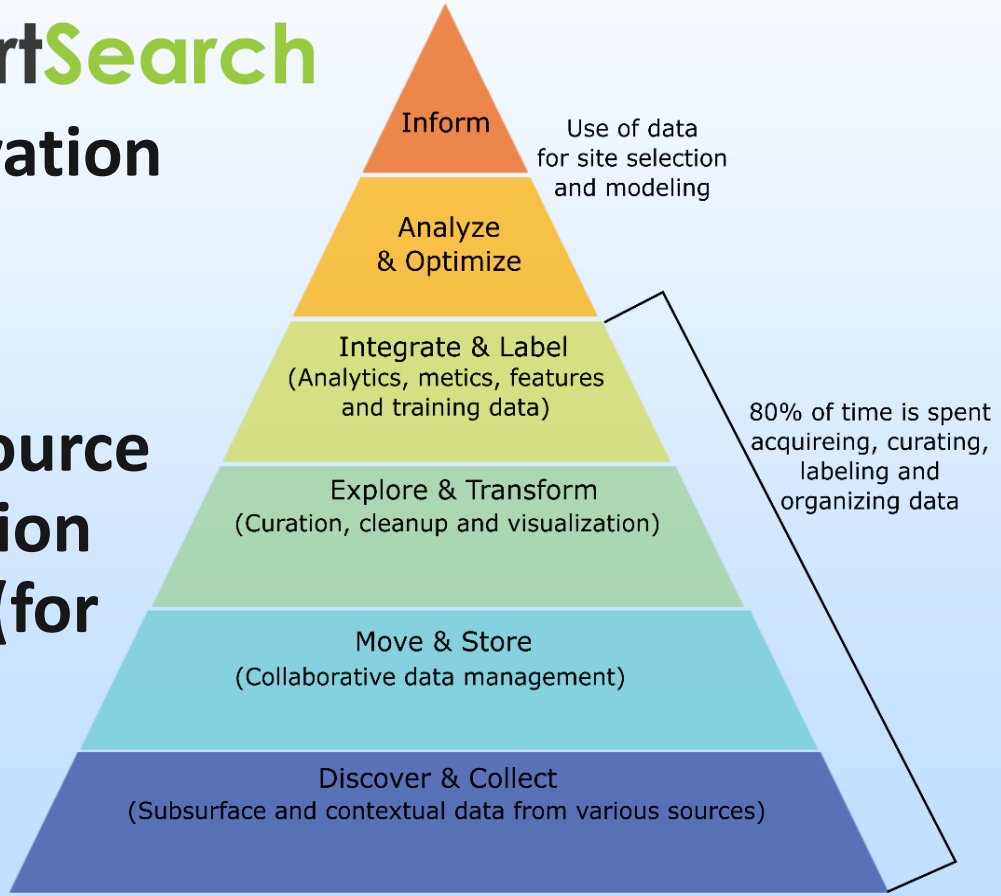
*updated 8/25/20

Use of AI/ML Tools for CS Data Curation

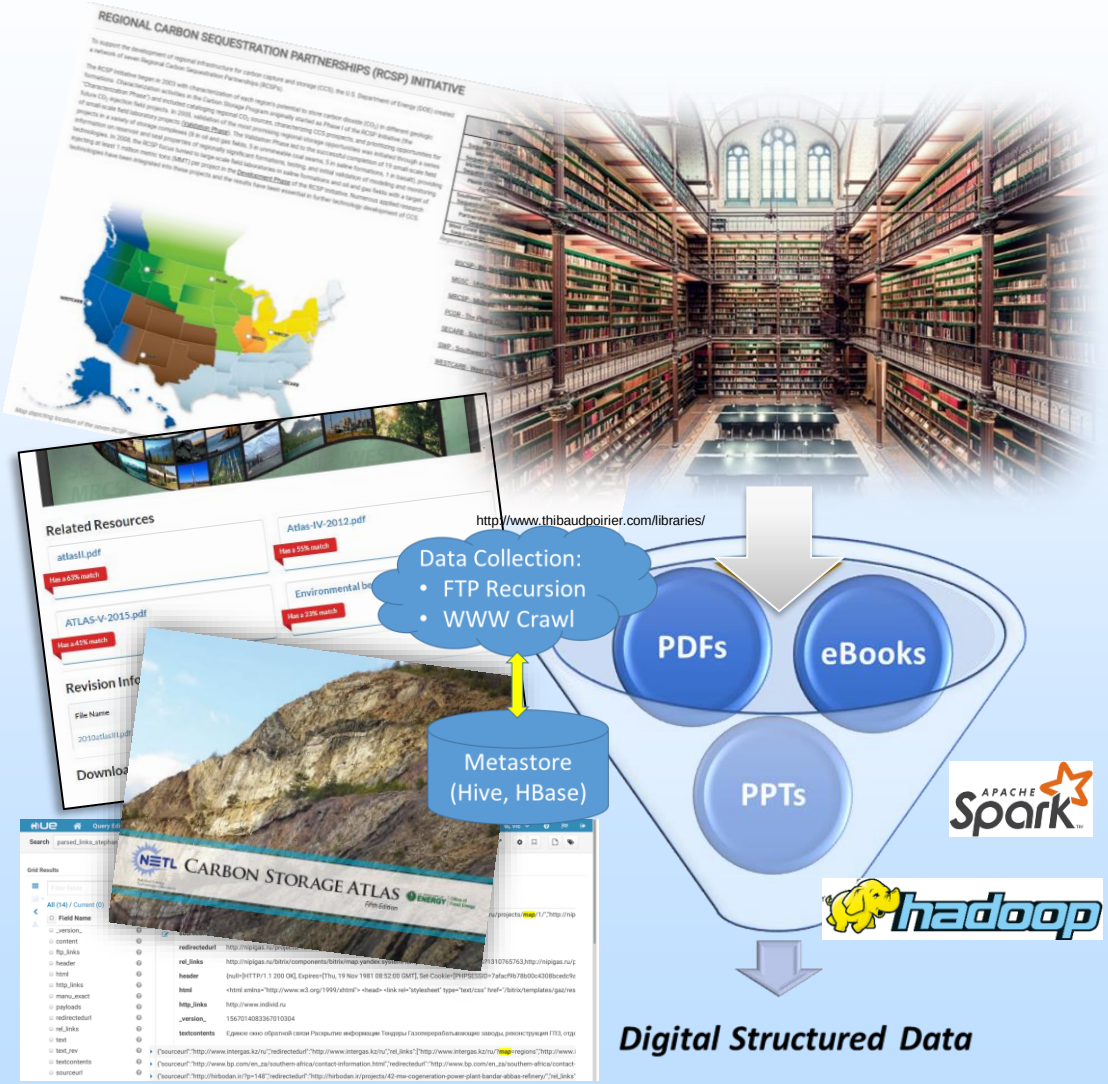
Challenge: Making available data discoverable, searchable, and easy to reuse

Solutions:

- **Open-source data scraping efforts** NETL SmartSearch
- **Cataloging for metadata extraction and preservation**
- **Geographic database development to make searches easier (GeoCube)**
- **Natural language processing for text-based resource classification, organization, keyword identification (metdata building) and geographic association (for searchability)** NETL SmartParse



NETL's SmartSearch, a big data, algorithm

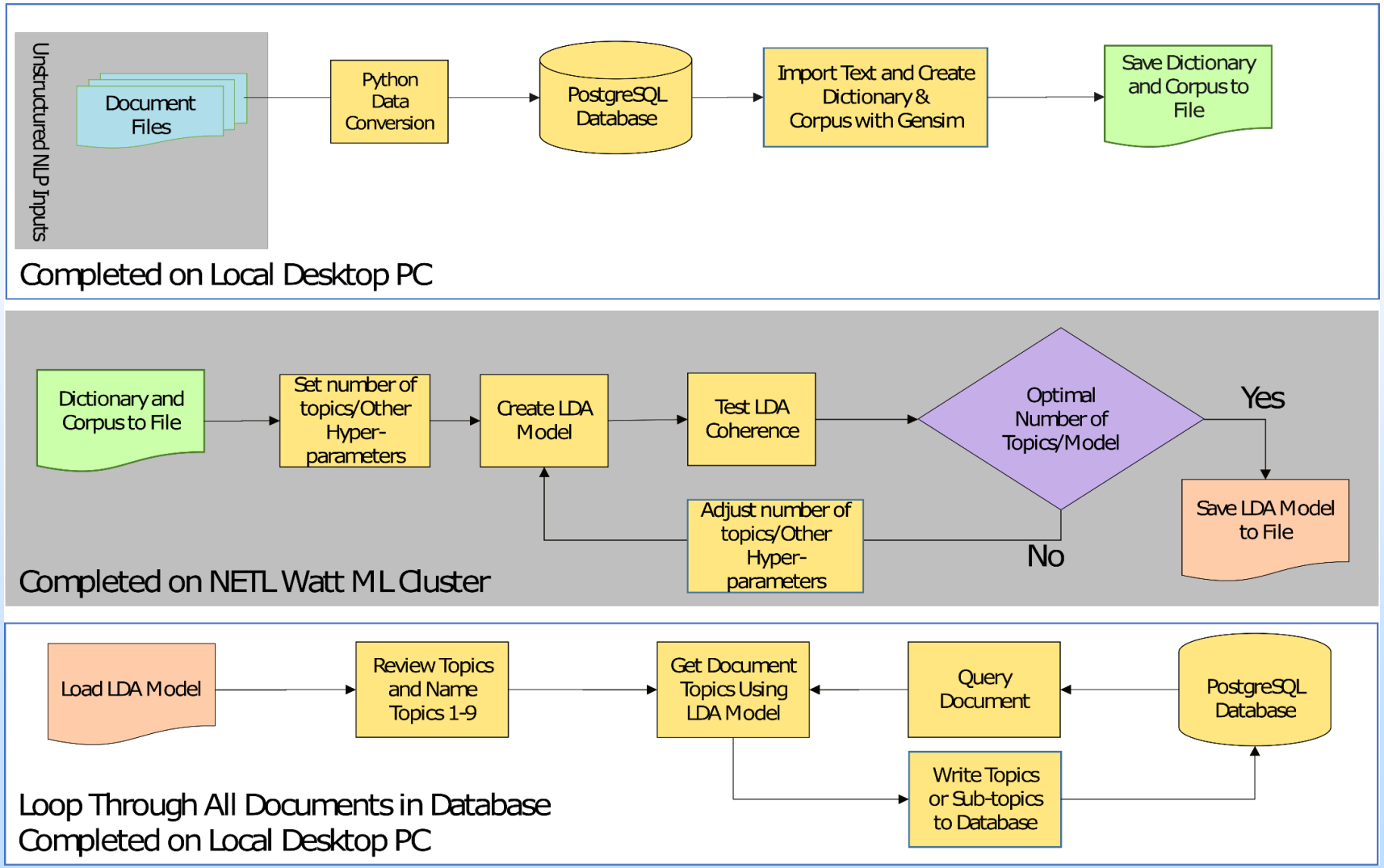


SmartSearch core features:

- Massively Parallel Operations:
  - Parsing many file formats (at scale) -- uses Tika, can be extended to additional formats beyond Tika
  - Parsing of zips (parses content of nested zips)
  - Cluster based large scale data management
  - ML processing (both Spark ML and Spark NLP) for 1) parallel NLP processing, 2) ML (recommendation engine) using Sparse and Dense Vectors and easily implement myriad of ML pipelines.
- Web crawling / indexing
- FTP crawling / indexing
- Data discovery -- combine above with web APIs (search engine, USPTO, etc) to automate data discovery and identify relevant data from seed(s)
- Developing Spark ML with GPU on NETL ML cluster.
- Developing integrations with Databricks GPU Spark ML processing.

Natural Language Processing (NLP)
A case study used for CS text-based resources

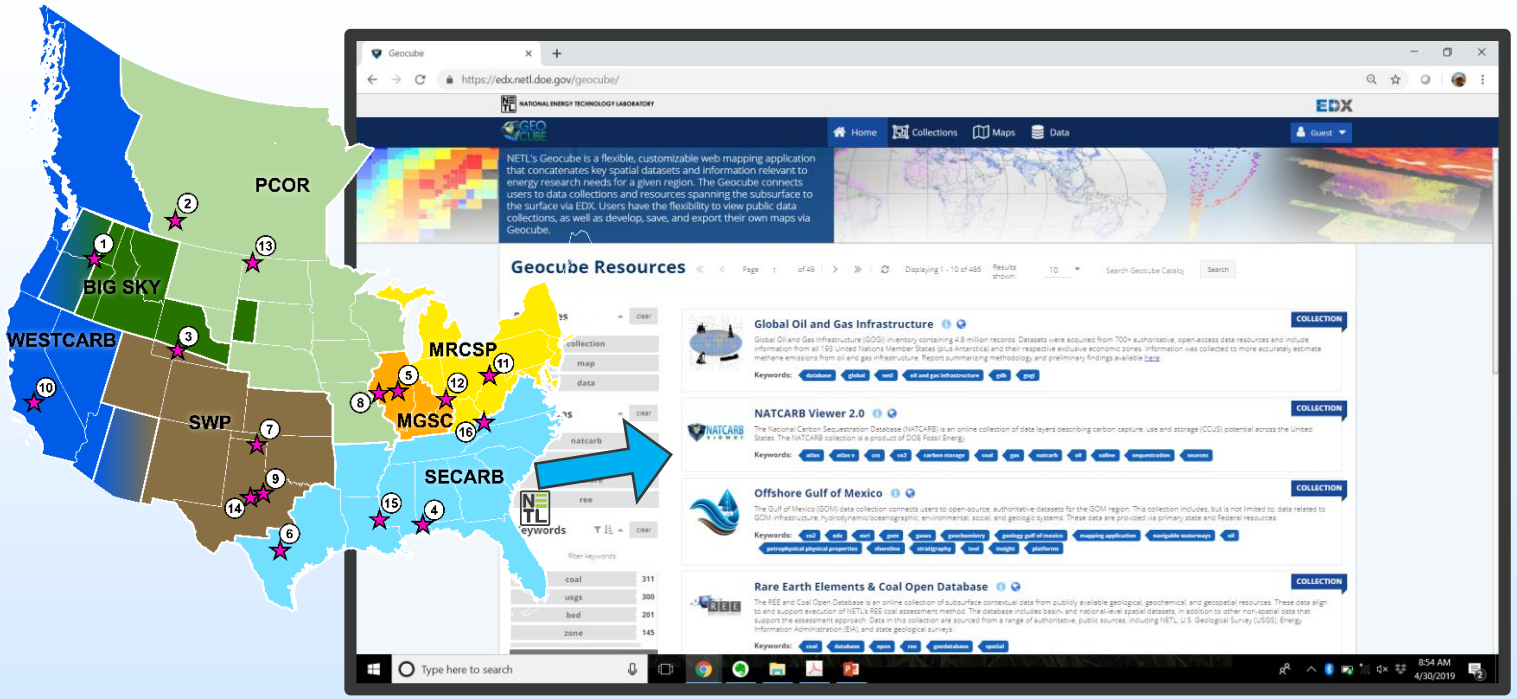
- Latent Dirichlet allocation **(LDA) model based on corpus of 2071 text-based documents**
- Topic names assigned by subject-matter experts
- **Each document is classified** by % of each topic it's associated with
- **Each document has 50+ keywords identified** and can be **associated metadata on EDX**
- **Parse geographic location to associate with each document** – when possible

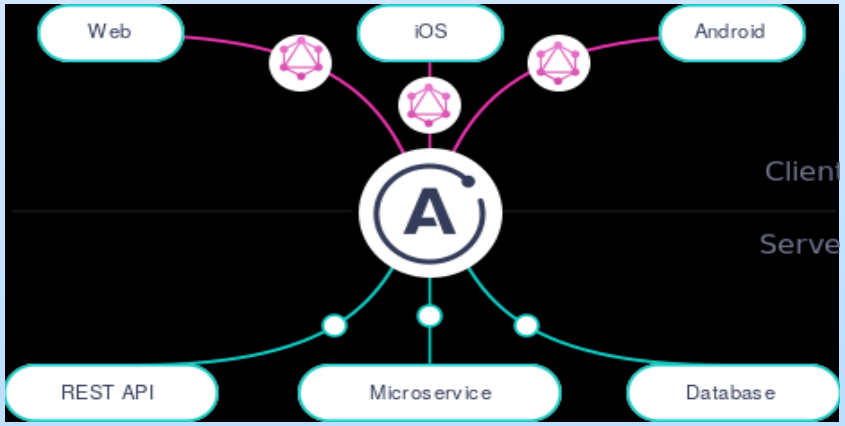


Morkner, P., Bauer, J., Creason, C., Sabbatino, M., Wingo, P, Greenburg, R., Walker, S., Yeates, B., and Rose, K., **in review,** Distilling Data to Drive Carbon Storage Insights, journal: *Computers & Geoscience*

GeoCube 3.0 – CS spatial data search & visualization

- **Upcoming GeoCube enables spatial search capabilities –** users can rapidly access and visualize data on an interactive map for the world
  - Natcarb Viewer 2.0
  - **Make accessible relevant supplemental resources from across the FE portfolio** - Drives citation and reuse of data
- **Data inputs** needed for tools and models can be derived from datasets
- **Data outputs** from modeling and field studies can be searched for by site or region
- **Outside resources and modeling tools** can be incorporated



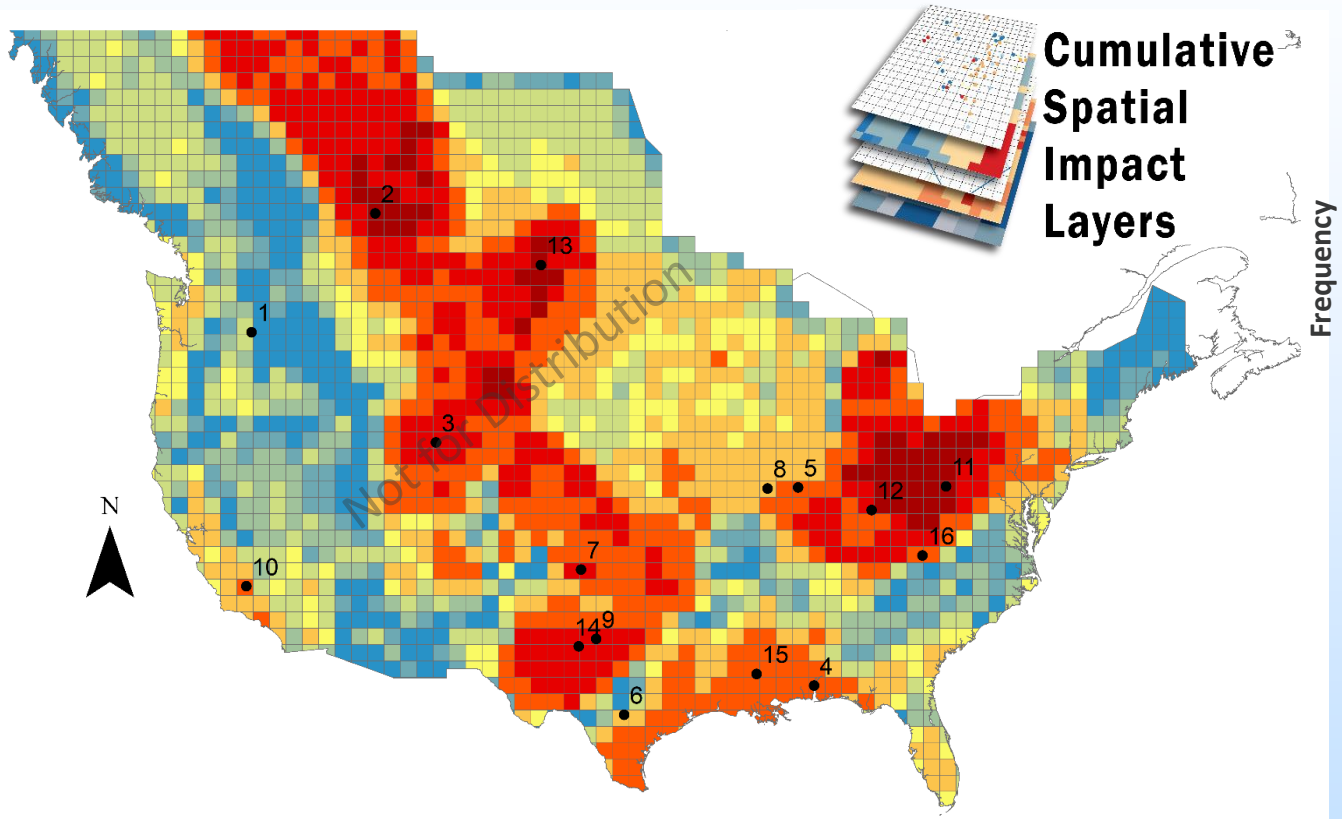


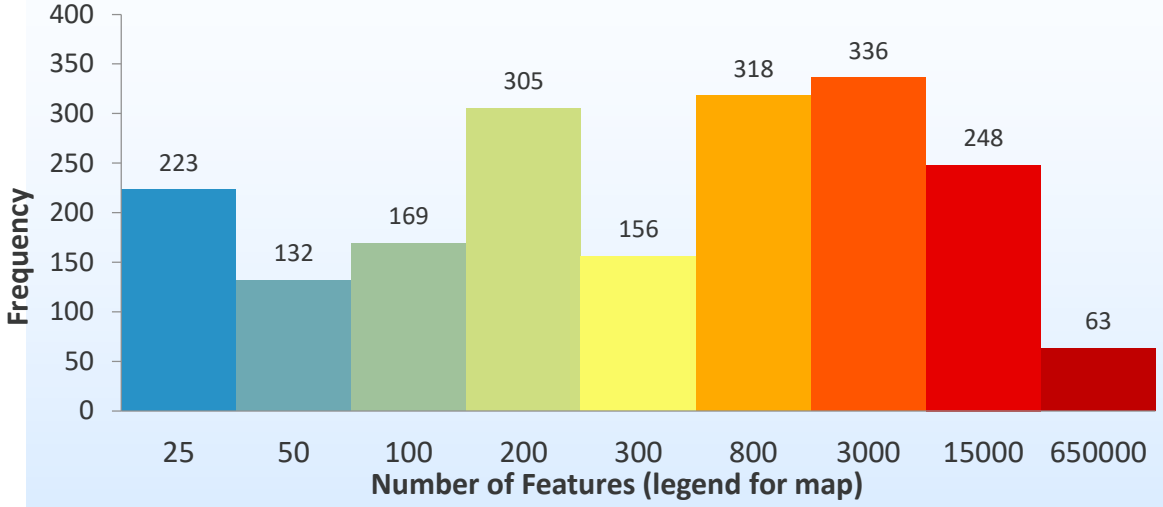
GraphQL API

New GraphQL/Aollo Server API provides a unified interface for querying numerous backing data sources

Results: Spatio-temporal trends in CS data

Bringing together NatCarb, NRAP catalog and the Carbon Storage Open Database





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

CCS Projects Cataloged

1. Big Sky Validation Phase - Wallula Basalt Pilot Project
2. CAMi - Field Research Station
3. CarbonSAFE – Wyoming
4. Citronelle (SECARB)
5. Decatur
6. Edwards Aquifer
7. Farnsworth - Anadarko Basin
8. FutureGen
9. High Plains Aquifer
10. Kimberlina (WESTCARB)
11. Appalachian Basin Test (MRCSP)
12. Cincinnati Arch Test (MRCSP)
13. Williston Basin Oil Field Test (PCOR)
14. Scurry Area Canyon Reef Operations
15. Cranfield Site (SECARB)
16. Central Appalachian Basin Test (SECARB)

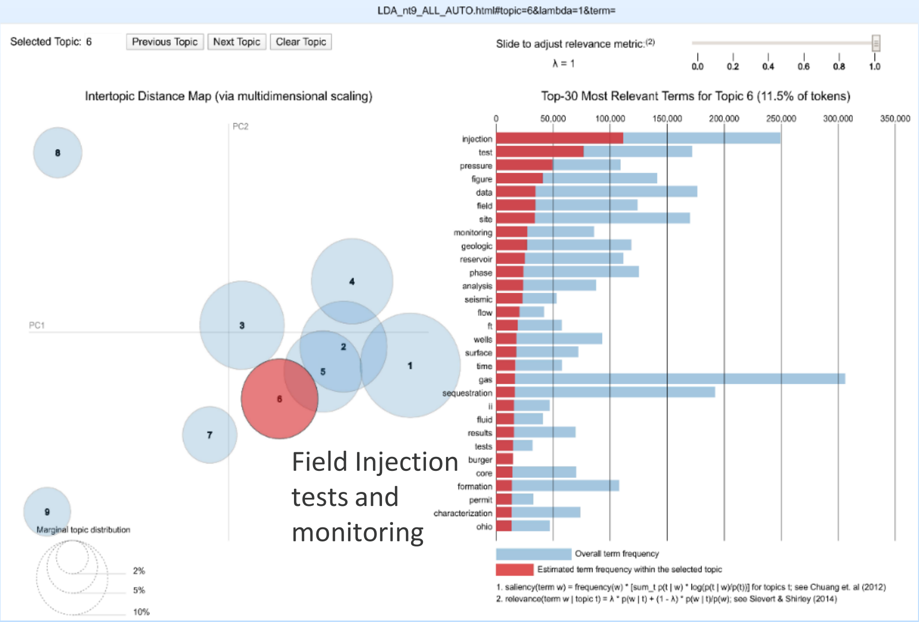
Paige Morkner, Paige.Morkner@netl.doe.gov
PI: Kelly Rose, Kelly.Rose@NETL.doe.gov

Results: Natural Language processing

Keywords and geographic associations

NETL SmartParse

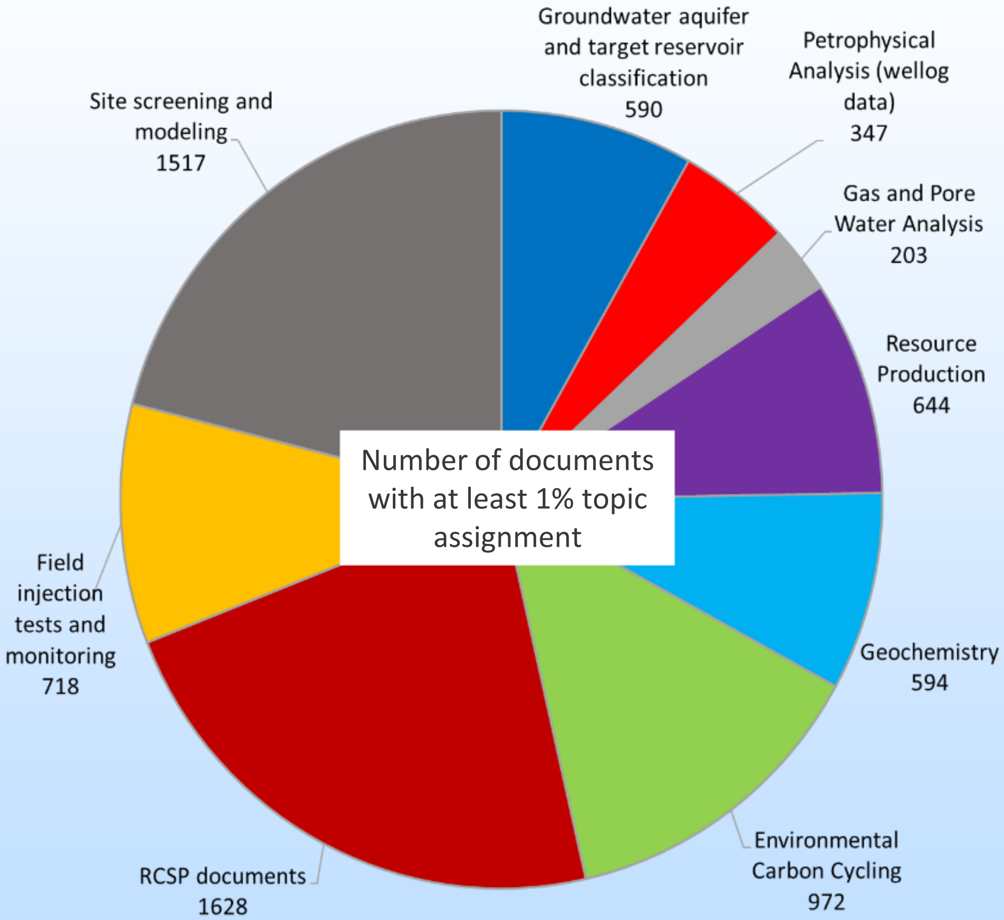
- Produced a **9 topic LDA model** – grouping similar papers
- Produced **keywords** associated with resources
- Geographic location recognition (in progress)
- Integration into EDX through NETL SmartParse



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Field Injection tests and monitoring

spaCy

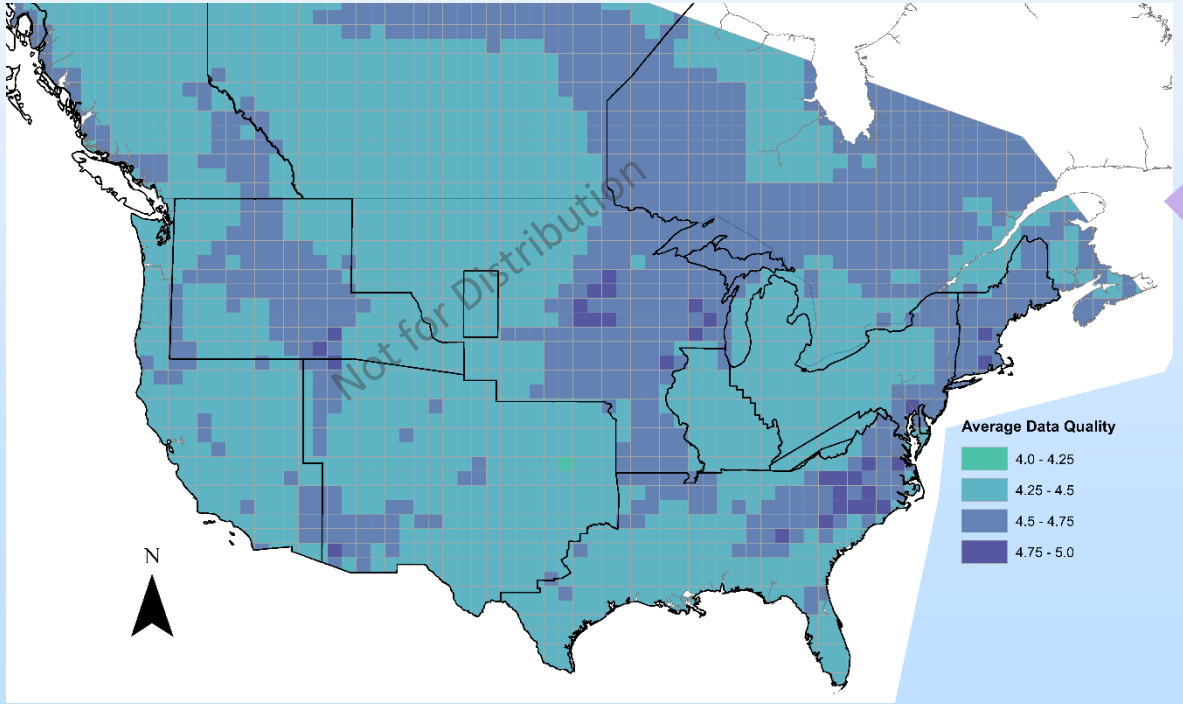


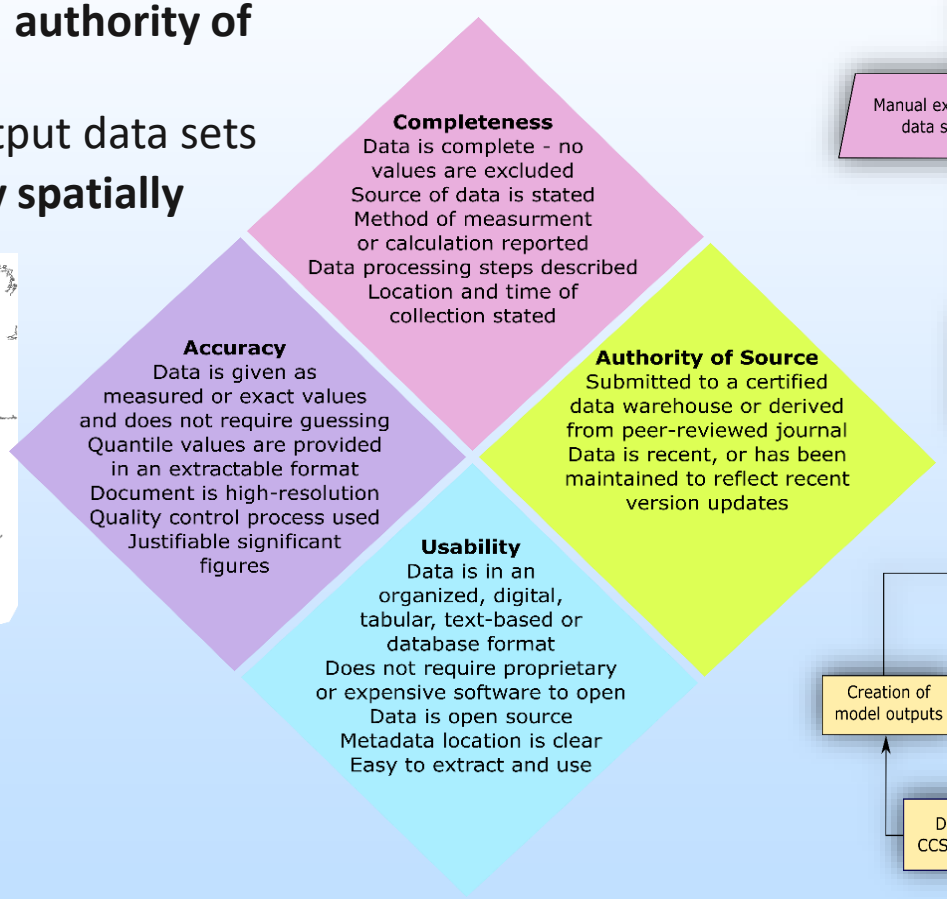
Number of documents with at least 1% topic assignment

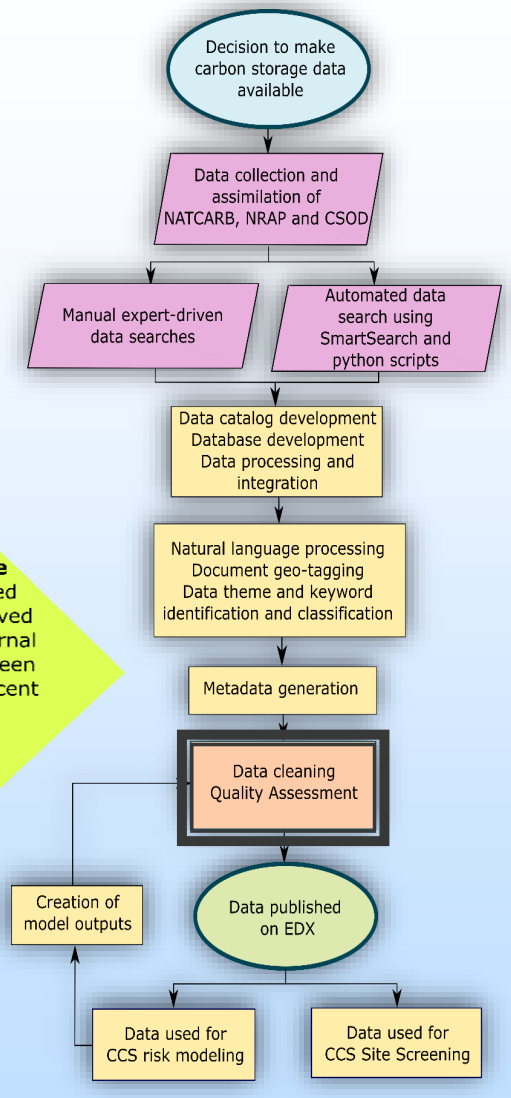
Morkner, P., Bauer, J., Creason, C., Sabbatino, M., Wingo, P, Greenburg, R., Walker, S., Yeates, B., and Rose, K., **in review,** Distilling Data to Drive Carbon Storage Insights, journal: *Computers & Geoscience*

Results: Data Quality assessment method development and spatial trends in CS data quality

- **5-point data quality assessment method** developed
- Quality based on c**ompleteness, accuracy, usability,** and **authority of source**
- **Applicable to many subsurface data sets** and model output data sets
- Combined with CSIL can be **used to analyze data quality spatially**







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 and Next Steps

FE and Carbon Storage program investments into data curation and management has led to 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:

- **A better understanding of CS relevant open- data density and data quality throughout US and Canada**
- **Improved access through the integration of CS data resources on EDX into GeoCube, SmartSearch and SmartParse (EDX version of NLP tools presented here) for further searchability with spatial searches and keyword searches**
  - Updates to GeoCube for enhanced spatial searchability and integration of modeling tools to come
- **EDX AI/ML data discovery, labeling, integration tool developments trained to support Carbon Storage, SMART-CS, and NRAP**
  - Deployment of AI/ML algorithms to allow on-demand data discovery and integration, ready-made for each end-user needs

Organization Chart
Carbon Storage Data

Project Partners
DOE
NETL
RCSPs – Big Sky Carbon Sequestration Partnership, Southwest Partnership, Southeast Regional Carbon Sequestration Partnerhsip, Midwest Regional Carbon Sequestration Partnership, Midwest Geological Sequestration Consortium, Plains CO2 Reduction Partnership.

Lead Organization
NETL

Principal Investigators
Kelly Rose, Jennifer Bauer

Task 28
Curation of Carbon Storage R&D Products Through Advanced Data Computing Solutions

Lead: Kelly Rose
Contractors: Chad Rowan, Michael Sabbatino, Paige Morkner, Andrew Bean, Lucy Romeo, TJ Jones, Aaron Barkhurst, Vic Baker, Other Matric Software Engineers and Developers

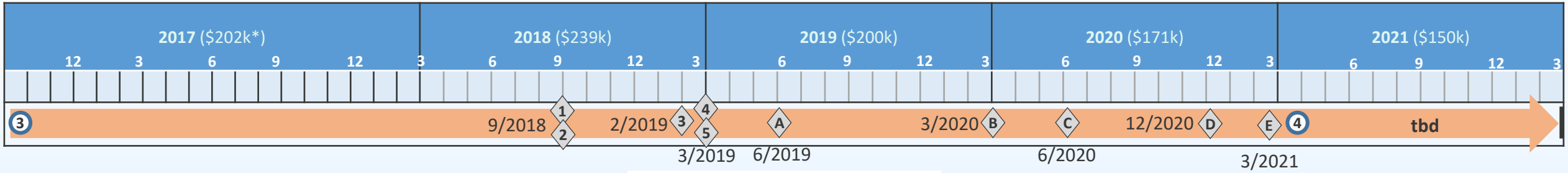
Task 27.0
Next Generation Development, Deployment, and Modernization of Database, Tools, Online Viewer, and Atlas

Lead: Jennifer Bauer
Contractors: Paige Morkner, Michael Sabbatino, Patrick Wingo, Andrew Bean, TJ Jones, Aaron Barkhurst, other Matric Software Engineers and Developers

Task 27: Project Timeline Overview

Key Team Members: PI – Jen Bauer, Kelly Rose CO PI – Paige Morkner

Natcarb - Next Generation Development, Deployment, and Modernization of Database, Tools, Online Viewer, and Atlas



Milestones

EY18:

1. Outline design and content for online version of Carbon Storage Atlas (9/2018)
2. Assess open CS datasets and resources. Generate catalog with priorities for integration into EDX to support CS data analytics and R&D (9/2018)
3. List of enhancements made to Natcarb tool, including links to scripts through EDX (2/2019)
4. Deploy CS virtual surface/subsurface data resources in EDX Geocube v2 tool on EDX and in EDX supported virtual geodatabase (3/2019)
5. Public release of interactive, online version of Carbon Storage Atlas (3/2019)

EY19:

A. Outline framework for integrating Spatial Analytical capabilities within Natcarb tool (6/2019)
B. Summarize methods used and key spatio-temporal analytical findings from expanded CS data available on EDX (3/2020)

EY20:

C. Identify tools and models that will be targeted for integration and inclusion within the Natcarb Viewer (6/2020)
D. Outline report/manuscript on updated technical capabilities of Natcarb Viewer (12/2020)
E. Release update of Natcarb Viewer and Natcarb Database to EDX (3/2021)

Impact

Key Accomplishments/Deliverables	Value Delivered
Barkhurst, A., Bauer, J., Rose, K., Chittum, J., Rowan, C., Romeo, L. Geocube, 2018-09-23, https://edx.netl.doe.gov/dataset/geocube, DOI: 10.18141/1471973 Bauer, J., Rowan, C., Barkhurst A., Digiulio J., Jones K., Sabbatino M., Rose K., Wingo P. Natcarb, 2018-09-27, https://edx.netl.doe.gov/datasets/natcarb, DOI: 10.18141/1474110	• Produce a robust subsurface data framework that provides improved data access, data discoverability, and ease of use within the carbon storage community. • Integrate online, advanced analytics and models to help facilitate research across the carbon storage community. • Develops process and tools for production of future interactive, online Carbon Storage Atlases

* Task 27 is updating content into an existing tool with no development of a technology. Therefore, no TRL is assigned.

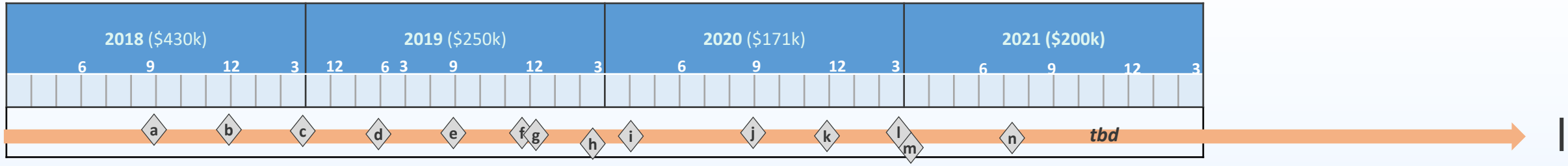
Chart Key

TRL Score | Go / No-Go Timeframe | Project Completion | Milestone

Task 28: Project Timeline Overview

Key Team Members: PI – Kelly Rose - CO PI – Chad Rowan, Mike Sabbatino

Curation of Carbon Storage R&D Products Through Advanced Data Computing Solutions



Milestones

Number	Expected Completion Date	Description
A	09/28/2018	Ingestion and stage on EDX for public release appropriate **Big Sky and PCOR** data products for CSP Managers to QAQC.
B	12/31/2018	Push to public on EDX appropriate **Big Sky and PCOR** data products.
C	03/29/2019	Ingestion and stage on EDX for public release appropriate **SECARB** Partnership data products for CSP managers to QAQC.
D	06/30/2019	Push to public on EDX appropriate **SECARB** Partnership data products.
E	09/30/2019	Ingestion and stage for public release appropriate **MRCSP** data products for CSP managers to QAQC.
F	12/31/2019	Push to public on EDX appropriate **MRCSP** data products.
G	12/31/2019	Ingestion and stage on EDX for public release appropriate **MGSC** Partnership data products for CSP Managers to QAQC.
H	03/31/2020	**Stretch:** Deploy NETL SmartSearch version 1 algorithm in EDX to support automated gathering of open, CS relevant data.
I	04/30/2020	Push to public on EDX appropriate **MGSC** Partnership data products.
J	9/30/2020	Deploy LivingDatabase beta version capability in EDX, private side, for CS teams (e.g., RCSPs) use and testing.
K	12/31/2020	Integration of CSP data products that are spatially related through enhanced EDX spatial search and discovery tool on Geocube
L	03/31/2021	Deploy NETL SmartSearch version 2 algorithm in EDX to support automated gathering of open, CS relevant data.
M	03/31/2021	Deploy LivingDatabase version 1 capability in EDX, private side, for CS teams (e.g., RCSPs) use and testing.
N	07/29/2022	Ingestion and push to public on EDX appropriate **SW Regional Partnership** data products.

Key Accomplishments/Deliverables

- 2017, EDX data ingestion tools & training to support curation of RCSP resources
- 2017, Developed web-based team digital notebook called "DataBook" & deployed on EDX
- 2017, Audited and gathered, using web scraping tools, open source datasets for RCSP websites 2018, Deployment of SmartSearch v1 to support automated gathering of open, CS relevant data
- 2018, Addition of Big Sky & PCOR data to EDX
- 2018, Constructed geodatabase hierarchy using inputs from cataloged RCSPs spatial datasets, and open data resources
- 2018, Big data computing cluster via EDX upgrades deployed
- 2019, Addition of Midwest CS Partnership & SECARB data resources to EDX & Natcarb Tool
- 2019, Addition of MGSC data resources to EDX & Natcarb Tool
- 2019, NDA signed with Google in relation to SmartSearch
- 2020, RCSP public and private resources hosted on EDX have a combined total of 3,037 and 1.64 TB of data.
- 2022, Addition of any final resources to EDX & Natcarb Tool

-- Impacts --

Value Delivered

- **Collecting, curating, and cataloging** data from all regional carbon storage partnerships & open-sources
- **Developing capabilities** to query curated data
- **Delivering** EDX's public-private capabilities, including growing access to its **big data computing** cluster and Amazon Web Services (AWS) **cloud services**, seek to facilitate more effective research **for DOE FE subsurface scientists.**
- **Pairing EDX hosted carbon storage data resources and products with other online capabilities**, data, custom ML algorithms and capabilities to enhance user experience and provide research teams with the resources needed to make subsurface energy research more efficient, reduce redundancy, and drive innovation.

* Task 28 is integrating data into an existing tool with no development of a technology. Therefore, no TRL is assigned.

Chart Key

- # TRL Score
- Go / No-Go Timeframe
- Project Completion
- Milestone