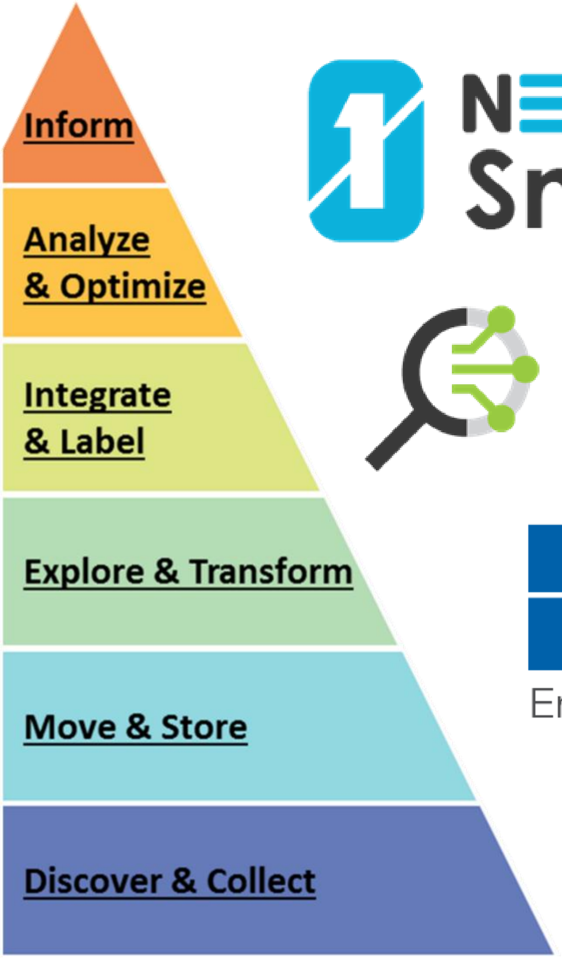
Using SmartParse NLP Tools to Develop a Living Database for Carbon Storage Data

NATIONAL ENERGY TECHNOLOGY LABORATORY

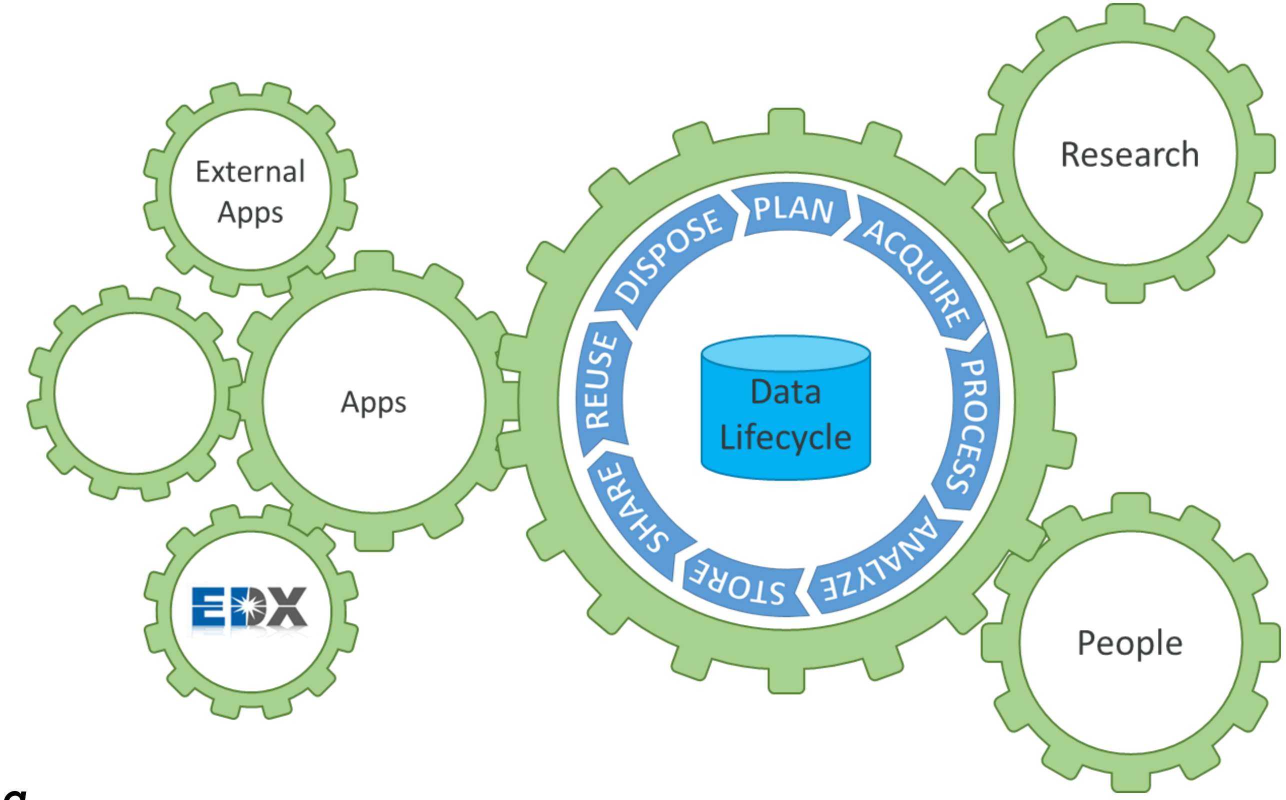
Michael Sabbatino and Paige Morkner
NETL Support Contractor
Research Innovation Center



NETL SmartParse

NETL SmartSearch

EDX™ Energy Data eXchange



NETL Carbon Storage Virtual Meeting
August 5, 2021

Disclaimer

NATIONAL ENERGY TECHNOLOGY LABORATORY

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

Author Information

Michael Sabbatino[1,2], Paige Morkner[1,2], Jennifer Bauer[1], Kelly Rose[1]

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

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

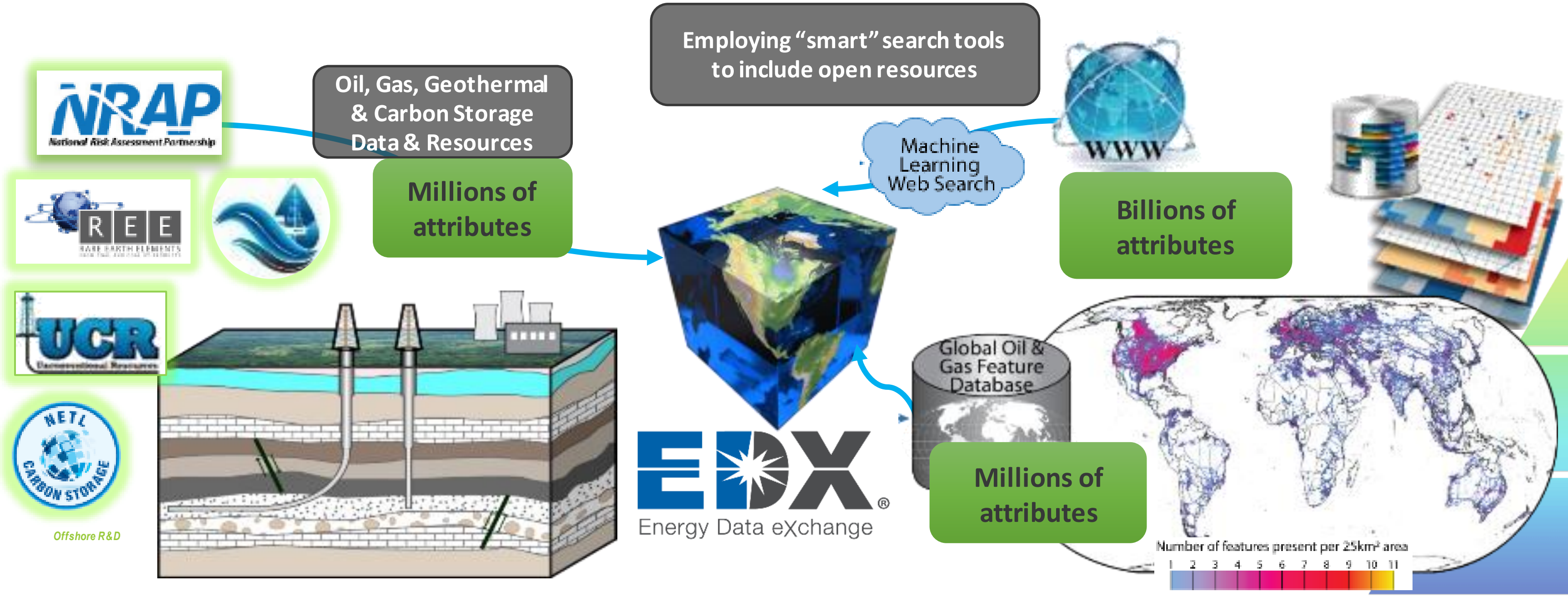
[3] NETL Support Contractor, 3610 Collins Ferry Rd., Morgantown, WV, 26505, USA

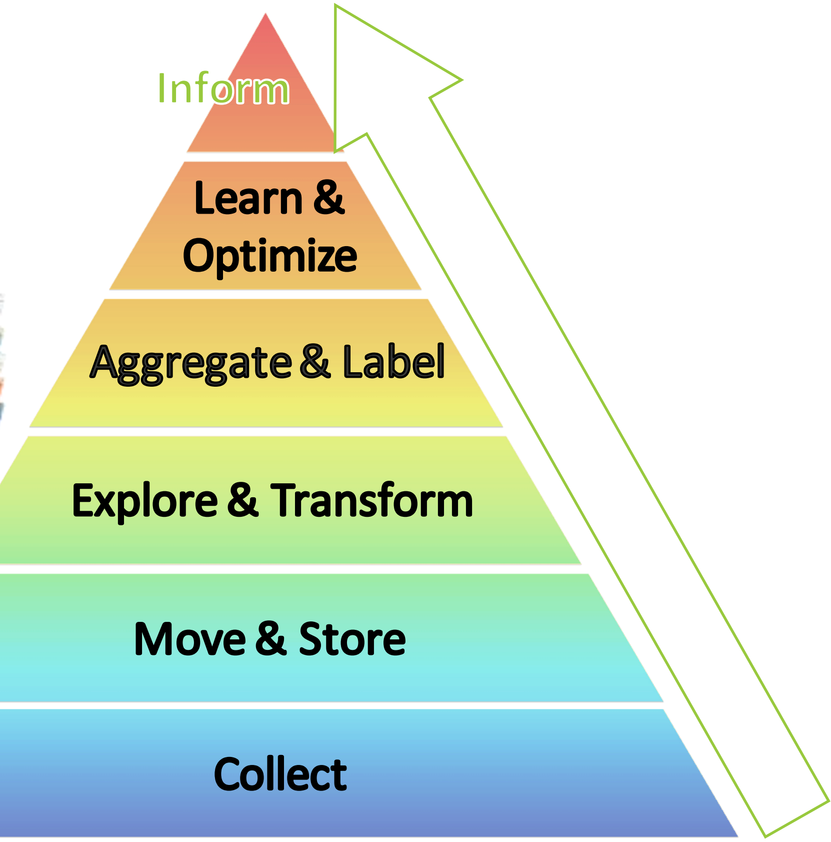
Research is data-driven

- Millions of dollars of research and data are available from carbon storage efforts
- How can we **preserve** and **efficiently access** those resources **to drive the next generation** of R&D?

Address the needs of the community through AI/ML enhanced methods via DOE's virtual data library and laboratory, EDX

NETL SmartParse





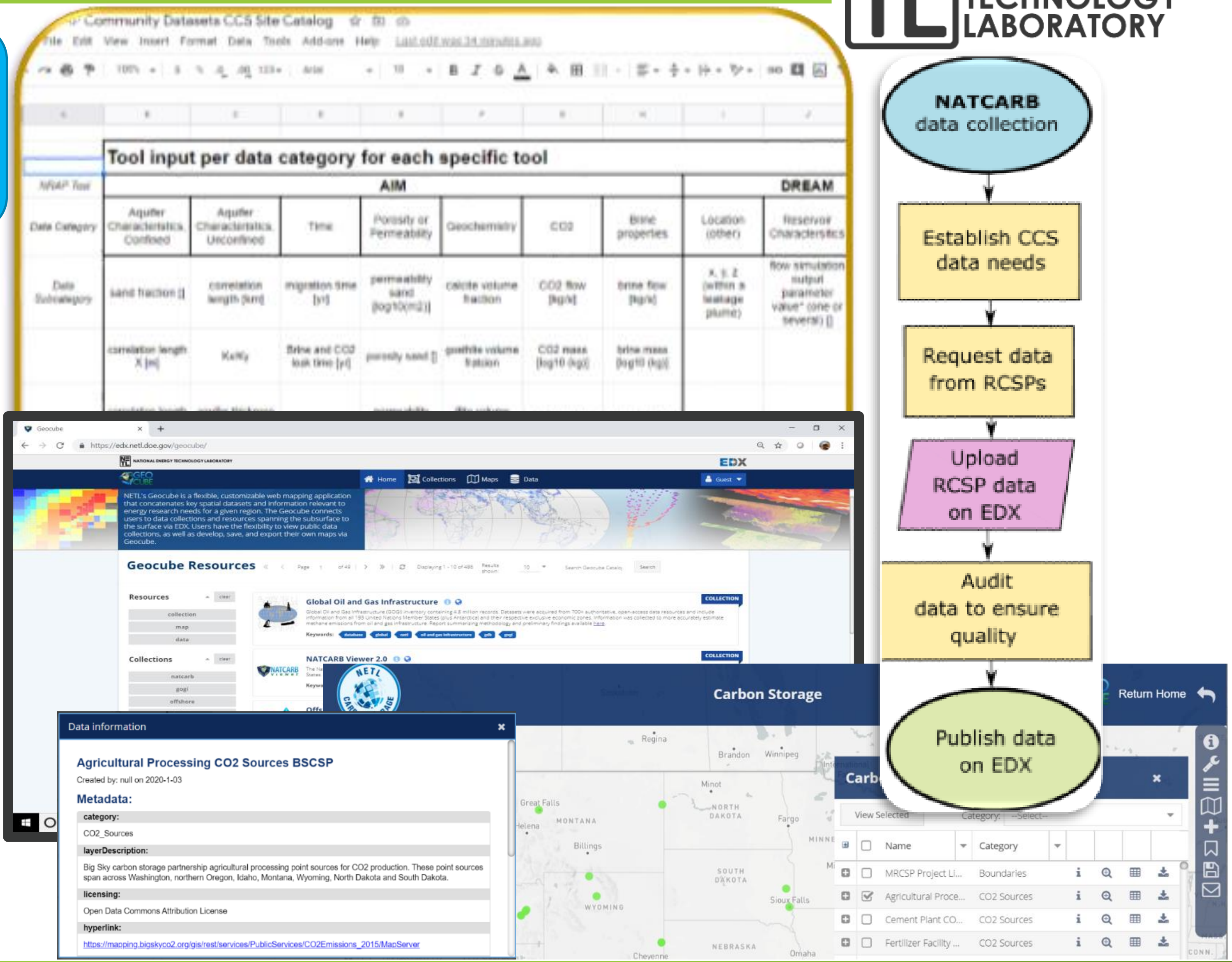
Supporting the whole life cycle of carbon storage data

Collection
- SmartSearch
- Expert-driven research
- EDX submissions

Metadata development and capture
- Cataloging
- ReadMe file development
- Natural language processing for keywords, topic modeling, geographic association

Quality Assessment
- Data ranking
- Data assessment method scoring

Data Organization and publishing
- Private workspaces
- Groups
- Submission packaging
- GeoCube data integration



Using AI/ML Tools for CS Data Curation

NATIONAL ENERGY TECHNOLOGY LABORATORY

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

Solutions:

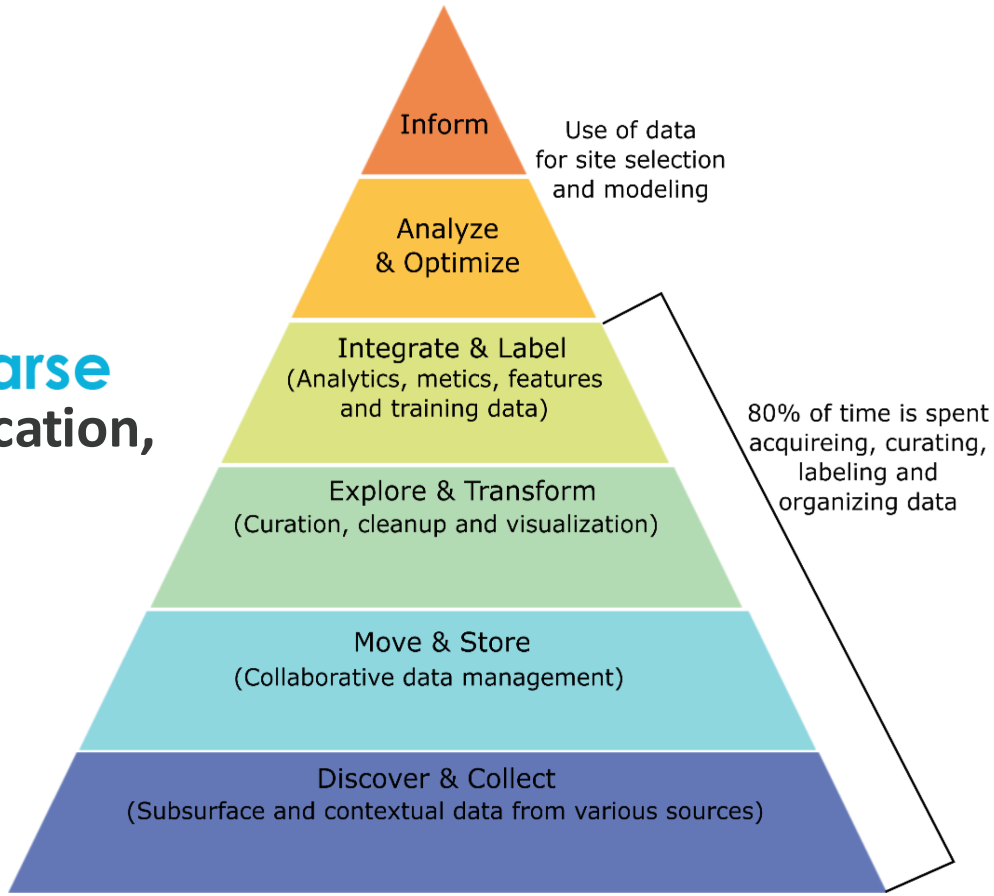
- Open-source data scraping efforts

NETL SmartSearch

- Cataloging for metadata extraction and preservation
- Geographic database development to make searches easier (GeoCube)

NETL SmartParse

- Natural language processing for text-based resource classification, organization, keyword identification (metdata building) and geographic association (for searchability)
- SmartParse NLP has integrations with New API for EDX
- ML image object recognition and data extraction



Types of Carbon Storage Data

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

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

Image Extraction:
- Documents
- Presentations
- Maps
- Posters

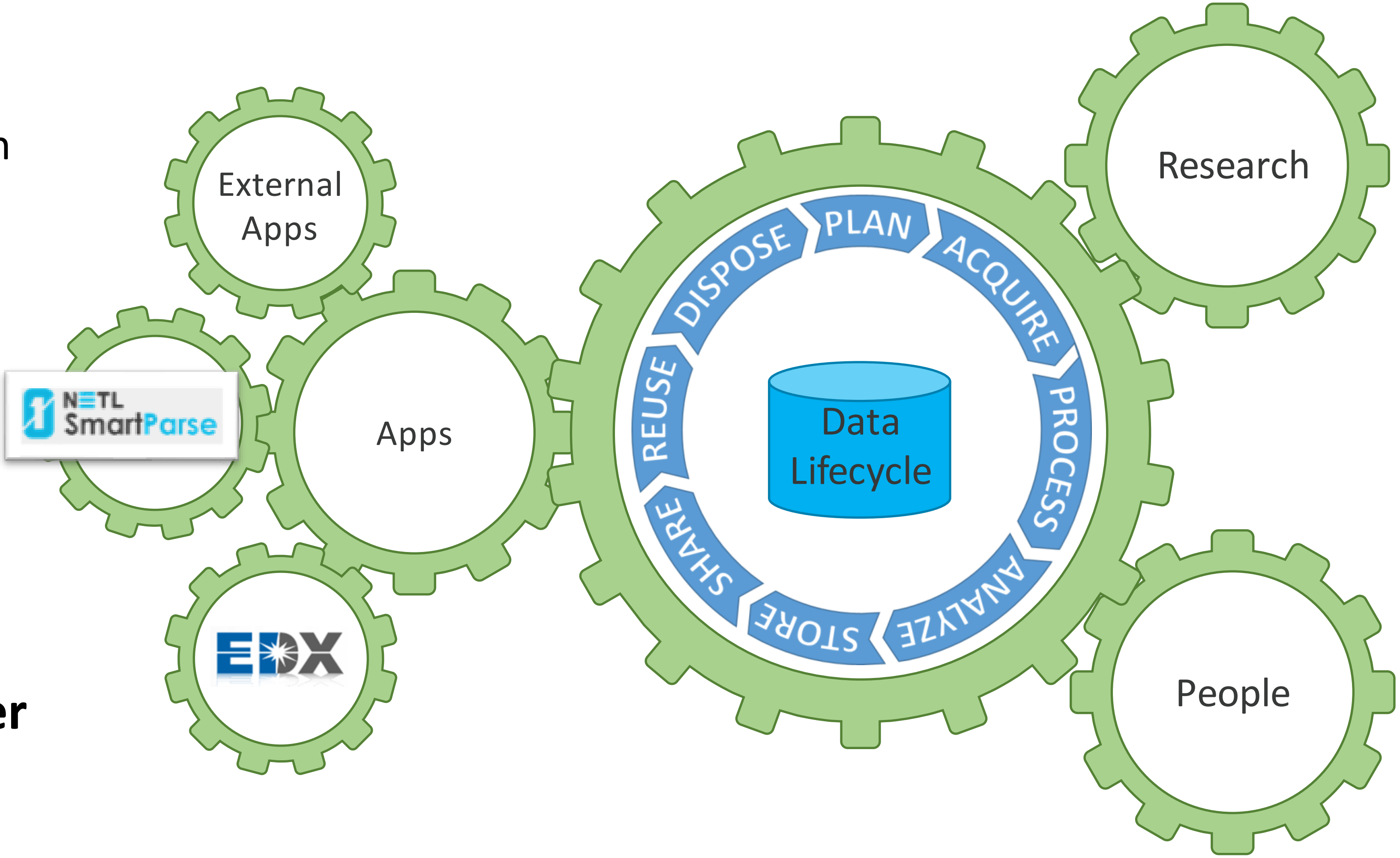
Other types of data:
- Tools
- Applications
- APIs
- LivingDatabase

Energy Data eXchange

Feeding the data hippo!

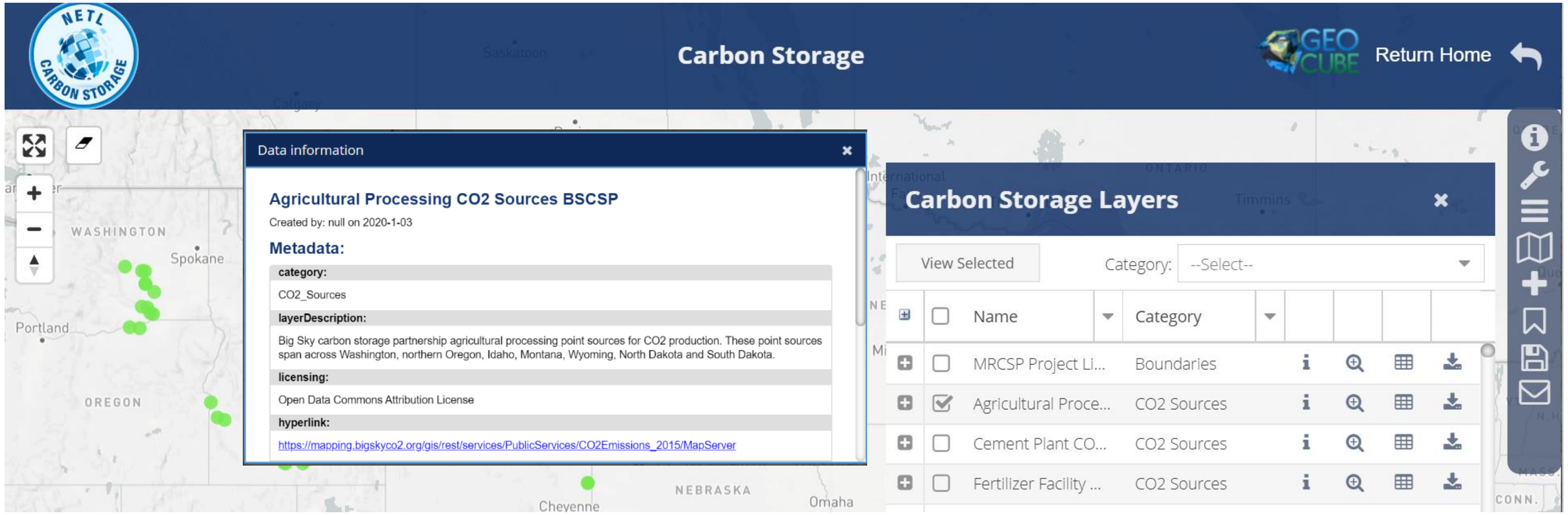
Living Database

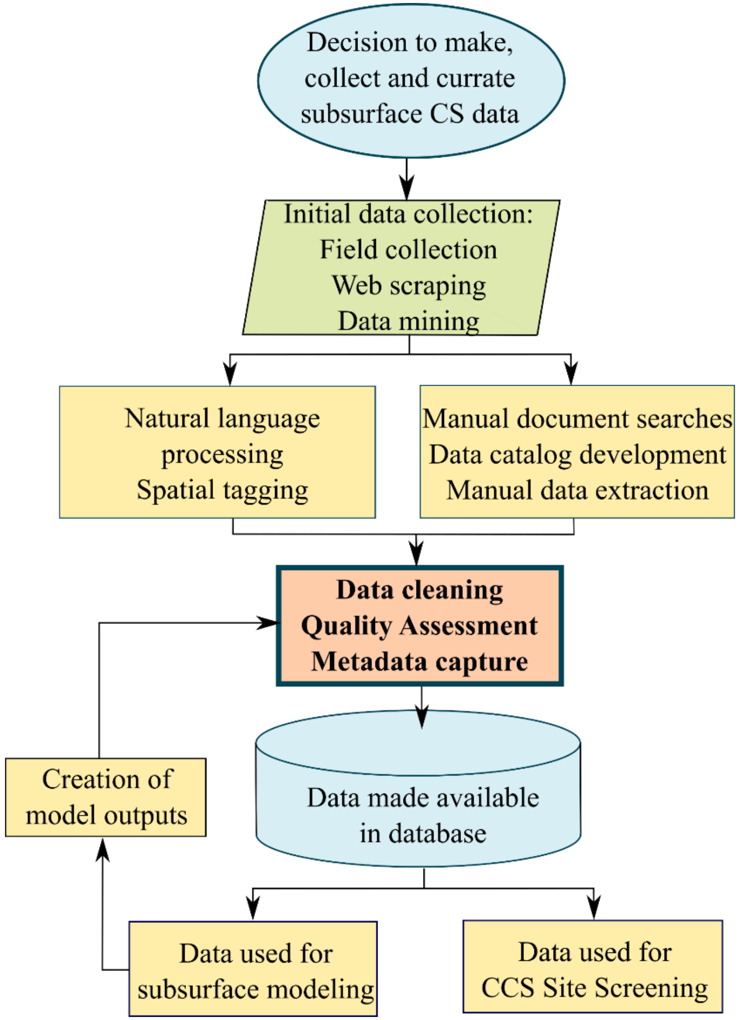
- **Store and Share Data in a Structured Secure Database Environment**
  - Reduce Redundant Acquisition
  - Direct Data Access (not file based storage)
  - Consistent Data with Staff Turnover
  - Enhanced Collaboration
- **Curation of data and knowledge**
- **Allows Direct Analysis from Database**
- **Available On Research network and Watt ML Cluster**



Data Cleaning for ML, AI and Spatial Analysis

- **Use Python scripts to automate data cleaning and help rapidly add structure, labels and metadata for datasets**
- **Metadata development for open carbon storage database**
  - Use of ArcREST data, geographic location, and attribute table to **develop metadata for layers in Geocube**
- **Link to EDX to capture additional metadate for datasets**



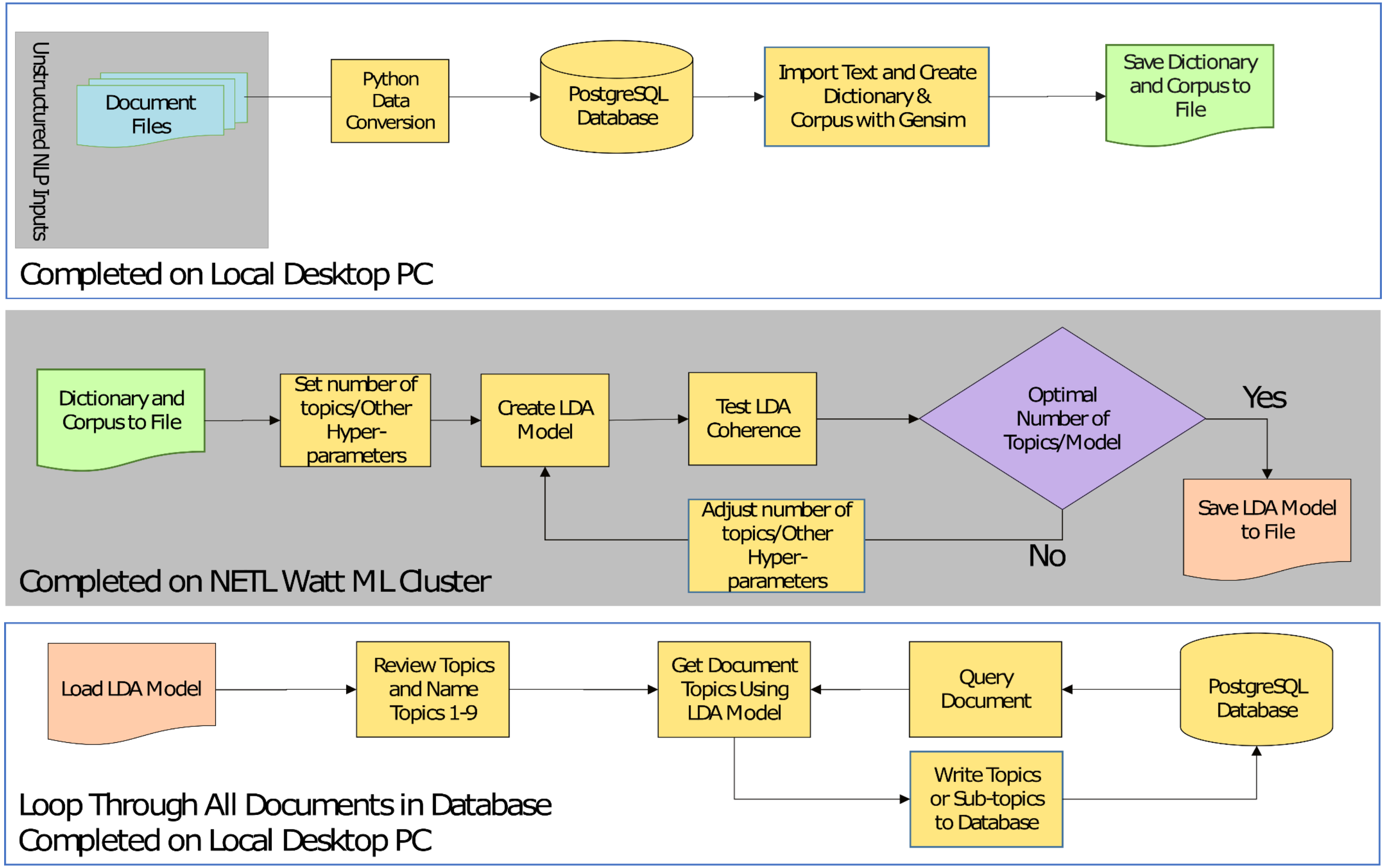


Natural Language Processing (NLP) Unsupervised ML for Document Classification

NETL SmartParse

NATIONAL ENERGY TECHNOLOGY LABORATORY

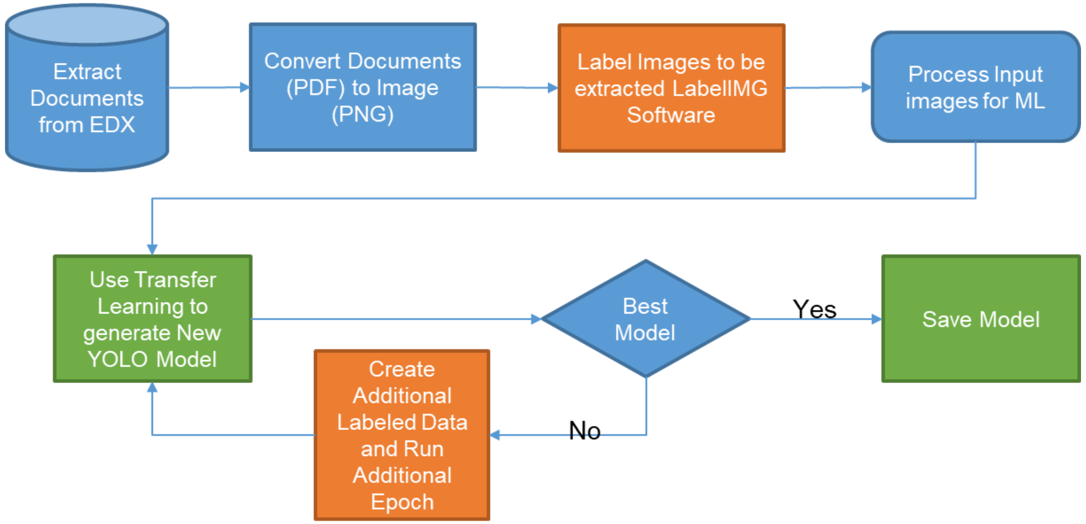
- Latent Dirichlet allocation **(LDA) model based on corpus of 2071 text-based documents**
- Topic names assigned by subject-matter experts
- **Each document is classified** by % of each topic it's associated with
- **Each document has 50+ keywords identified** and can be **associated metadata on EDX**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

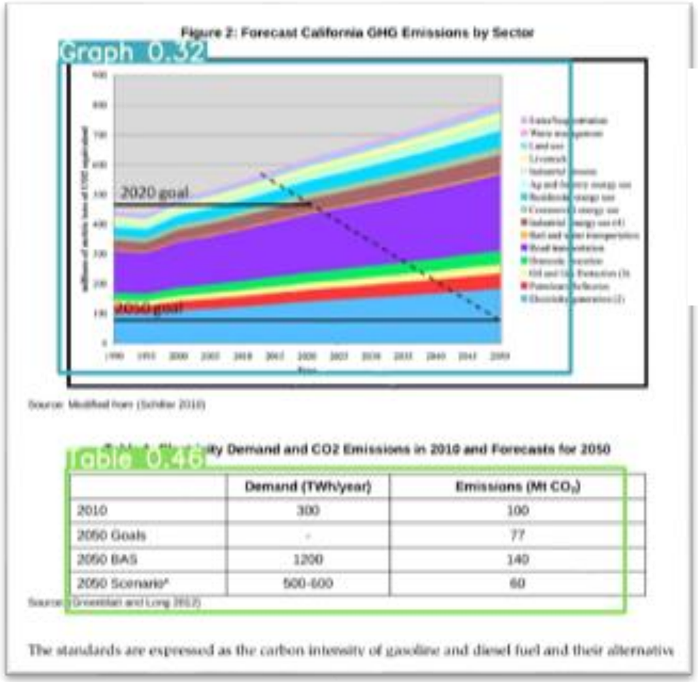


Machine Learning Image Data Extraction

NATIONAL ENERGY TECHNOLOGY LABORATORY

- Object Detection Model Development Process
  - Use transfer learning to train object detection model for specific image and data types
  - Detect Graphs, Diagrams, Photos, Maps, and Tables
  - Image Labeling and process Developed with help from Mickey Leland Energy Fellowship

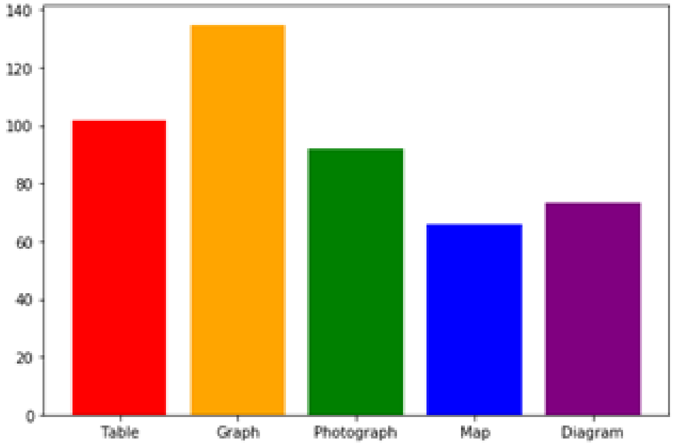




NETL SmartParse

Mickey Leland Energy Fellowship 1995 25 2020

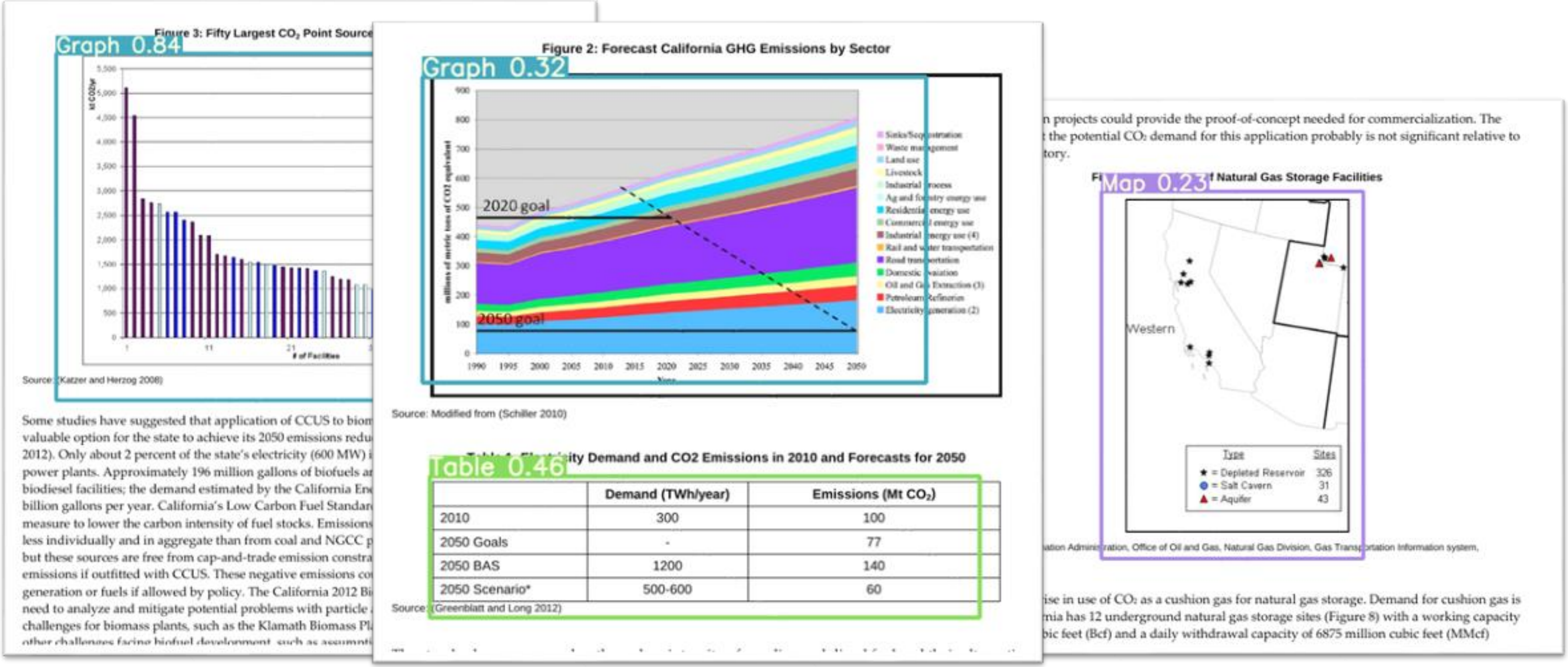
WATT The Power of Ai



Images and Tables Targeted for Data for Extraction

Machine Learning Data Extraction

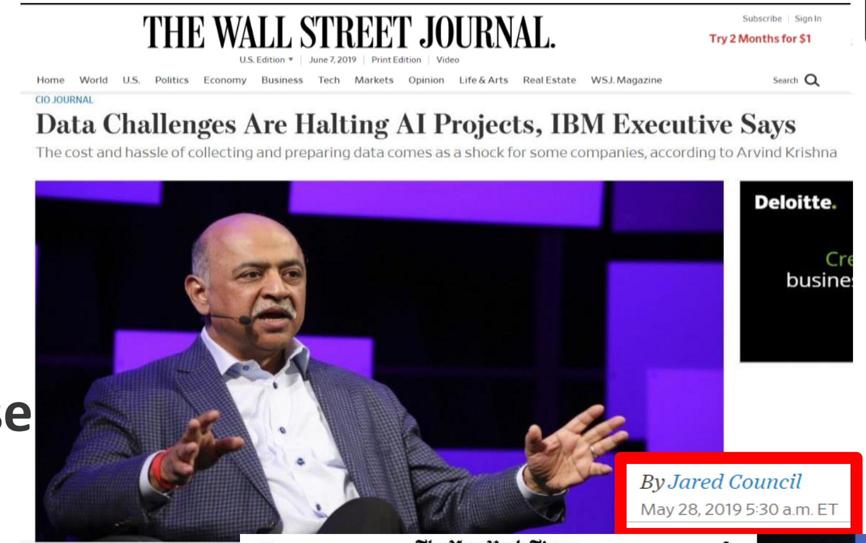
Utilize Object Identification ML Models to Extract Additional Data

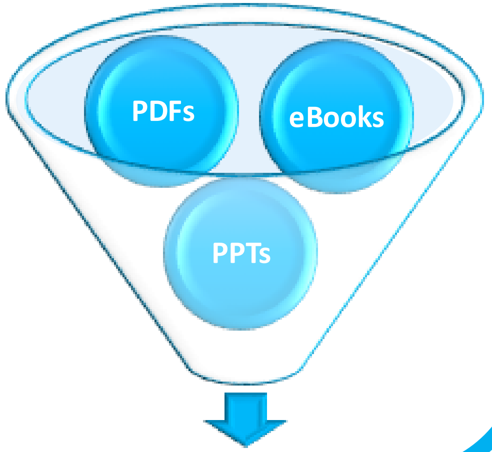


	Demand (TWh/year)	Emissions (Mt CO_2)
2010	300	100
2050 Goals	-	77
2050 BAS	1200	140
2050 Scenario*	500-600	60

Lessons Learned

- **Machine Learning, Artificial Intelligence and Natural Language processing are Difficult**
  - Whatever happened to Watson?
- **Lack of Labeled Training Data**
  - Training data is time consuming to develop and can be costly
- **Data availability is limited with Living Database**
  - Currently deployed on Research network
  - The database would improve if deployed on a cloud service or other shared environment



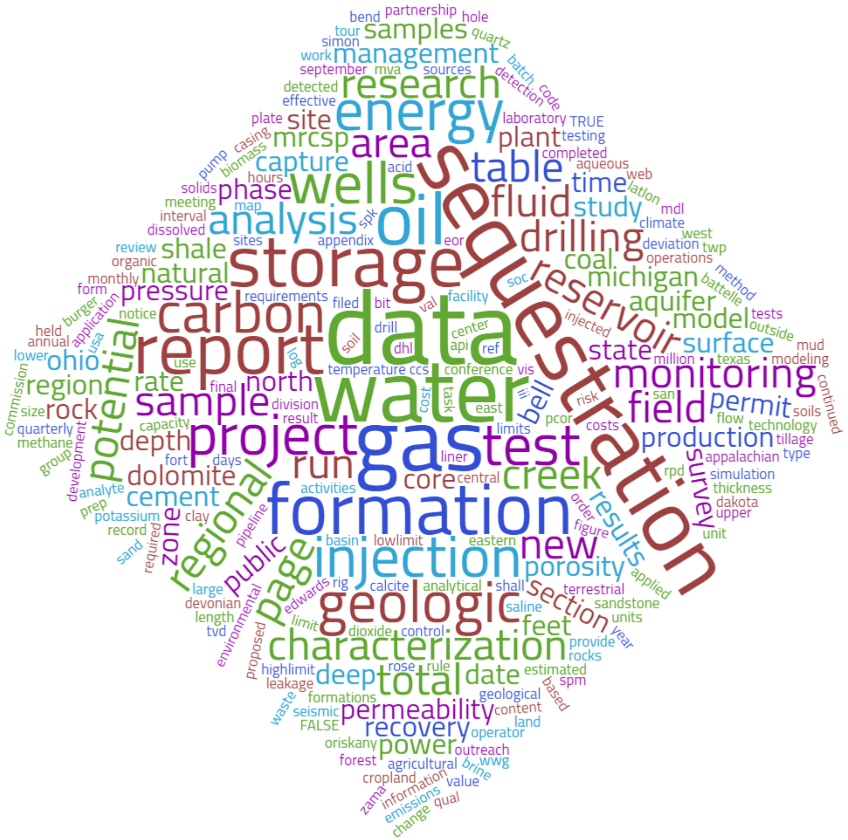


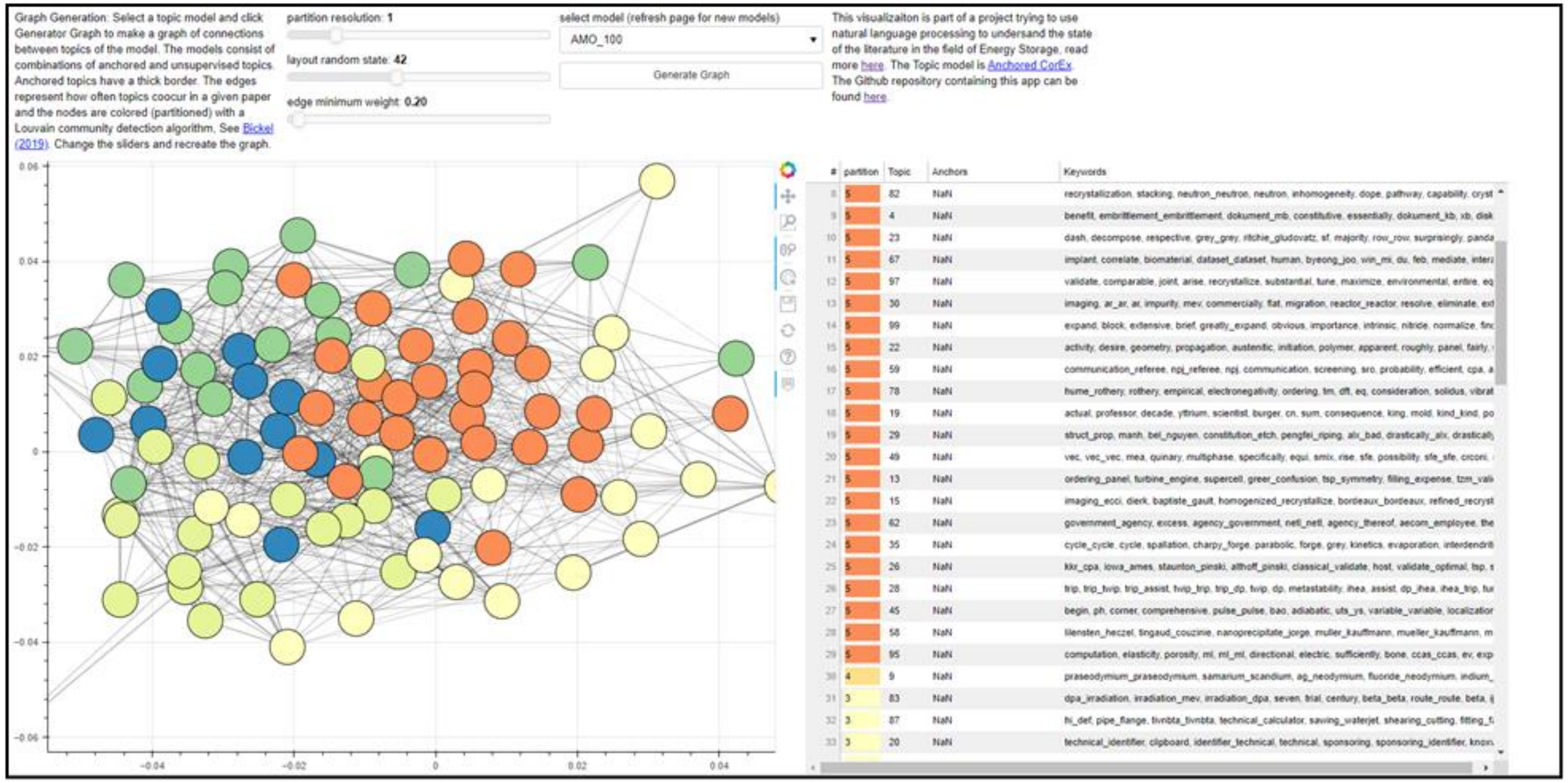
Structured data

https://www.nytimes.com/2021/07/16/technology/what-happened-ibm-watson.html

Synergy Opportunities

- **Collaborative cross project technology**
  - Use material same NLP tech
  - Using other NLP Models Louvian Community Detection





Supporting Data Collection, Curation & Analysis in Other Areas

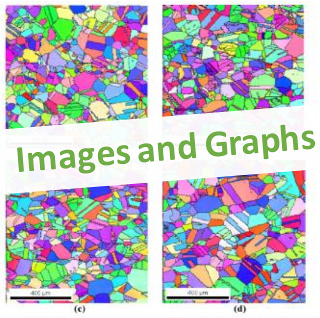
NATIONAL ENERGY TECHNOLOGY LABORATORY

Data mining, including…

Structured Data

Alloy (wt%)	N	C	Mn	Cr	Mo	Ni	Si
316LNSS-7N	0.07	0.027	1.7	17.53	2.49	12.2	0.22
316LNSS-11N	[illegible]	[illegible]	1.78	17.62	2.51	12.27	0.21
316LNSS-14N	0.14	0.025	[illegible]	[illegible]	2.53	12.15	0.2
316LNSS-22N	0.22	0.028	[illegible]	17.57	2.54	12.36	0.2

Images and Graphs



RESEARCH ARTICLE

Materials data analytics for 9% Cr family steel

Vyacheslav N. Romanov, Narayanan Krishnamurthy, Amit K. Verma, Laura S. Bruckman, Roger H. French, Jennifer L.W. Carter, Jeffrey A. Hawk

First published: 15 February 2019 | https://doi.org/10.1002/sam.11406

U.S. Department of Energy, DE-FE0028685.

Read the full text >

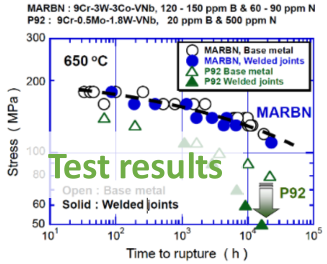
Abstract

A materials data analytics (MDA) [illegible] study to evaluate publicly available infor[illegible] andle nonlinear relationships and the sparsi[illegible] The overarching goal is to accelerate the desi[illegible] e and expense associated with [illegible] sil energy applications. Data entries in the [illegible] e alloy compositions, several processing parameters, [illegible] anical tests selected for this study were arranged in 34 [illegible] ws. While detailed microstructural information was not available, it is [illegible] the compositional space for the 9 to 12% Cr steels is limited such that all [illegible] entries have a tempered martensitic microstructure during service. Establishing a hierarchy of first-order trends in the publicly available data requires the MDA to filter out the biases. Complexity of the phase transformations and microstructure evolution in the multicomponent alloys (using 21 chemical elements) with major influence on mechanical

Contextual knowledge

Measurements

CT C	CS (MPA)	RT, hrs
593	310.3	1.45
593	275.8	5.5
593	275.8	6.33
593	206.8	55
593	171.7	357
593	144.8	1446
704	[illegible]	0.37
704	172.4	1.5
704	137.9	9.5
704	103.4	50.5
[illegible]	75.8	337
704	62.1	1227
816	103.4	0.75
816	89.6	1.87
816	68.9	12.75
816	48.06	84.3
816	36.5	331.8
816	29.0	1153

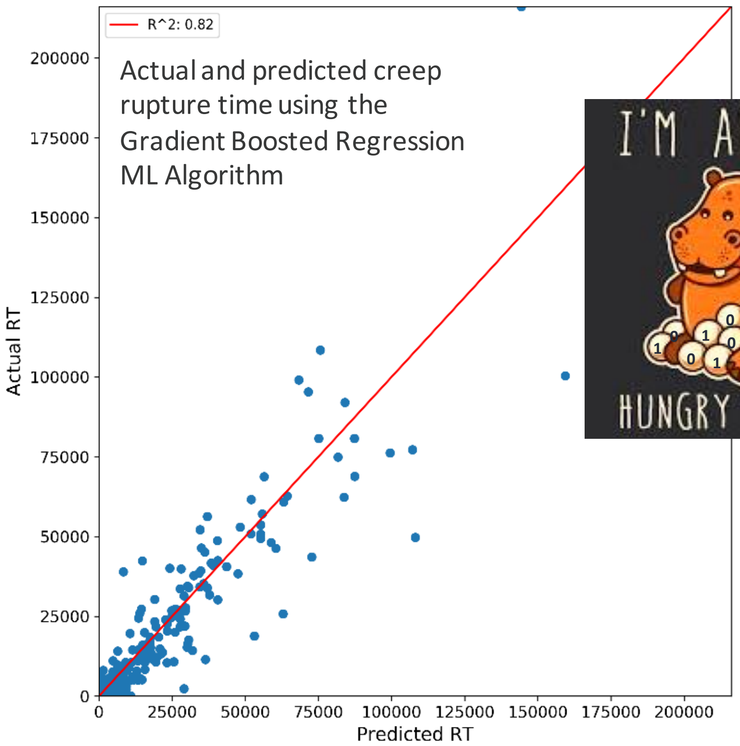


Test results

Move & Convert…

EDX™ Energy Data eXchange

…& use in predictive analytics for alloy behavior

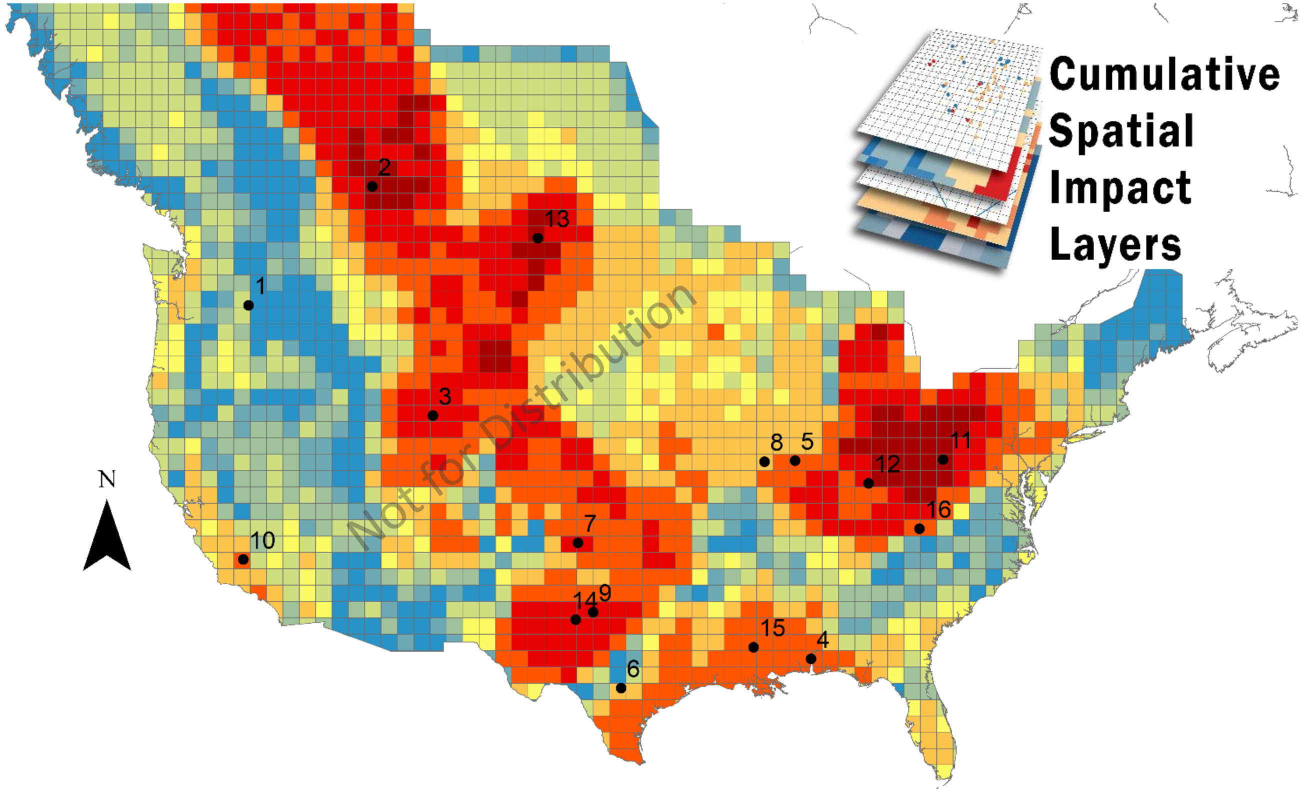


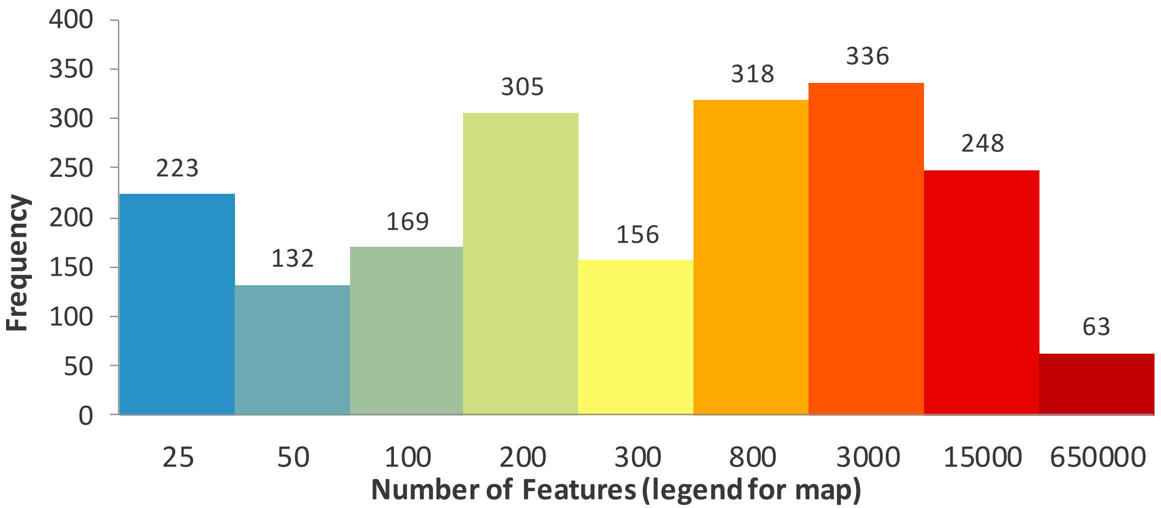
Evaluating machine learning models to:

- address data gaps
- identify key features in lifetime behavior of the alloy

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., **in review,** Distilling Data to Drive Carbon Storage Insights, journal: *Computers & Geoscience*

CCS Projects Datasets Cataloged

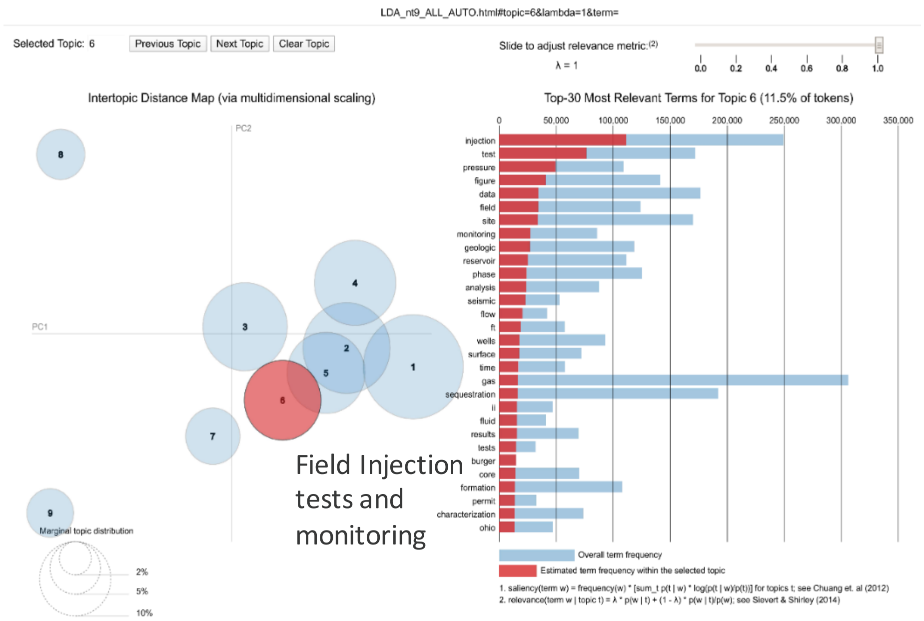
1. **Big Sky Validation Phase - Wallula Basalt Pilot Project**
2. **CAMi - Field Research Station**
3. **CarbonSAFE – Wyoming**
4. **Citronelle (SECARB)**
5. **Decatur**
6. **Edwards Aquifer**
7. **Farnsworth - Anadarko Basin**
8. **FutureGen**
9. **High Plains Aquifer**
10. **Kimberlina (WESTCARB)**
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)**

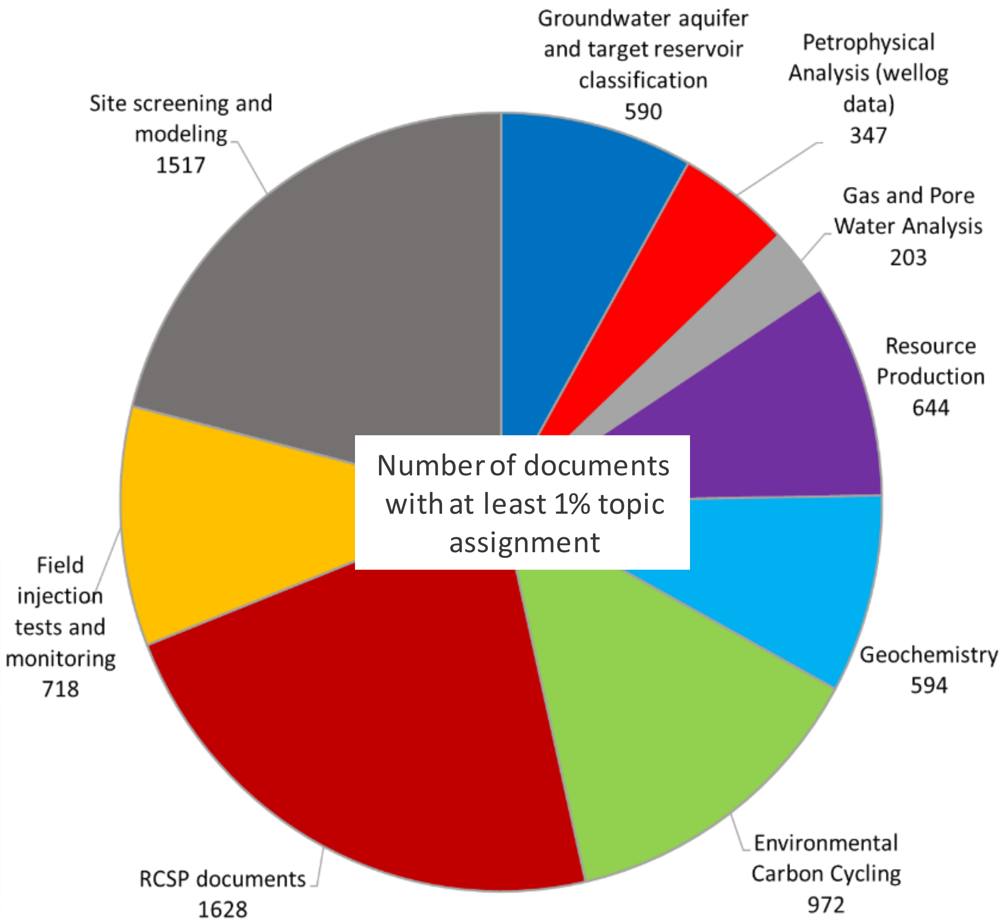
Results: Natural Language processing

Keywords and geographic associations

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

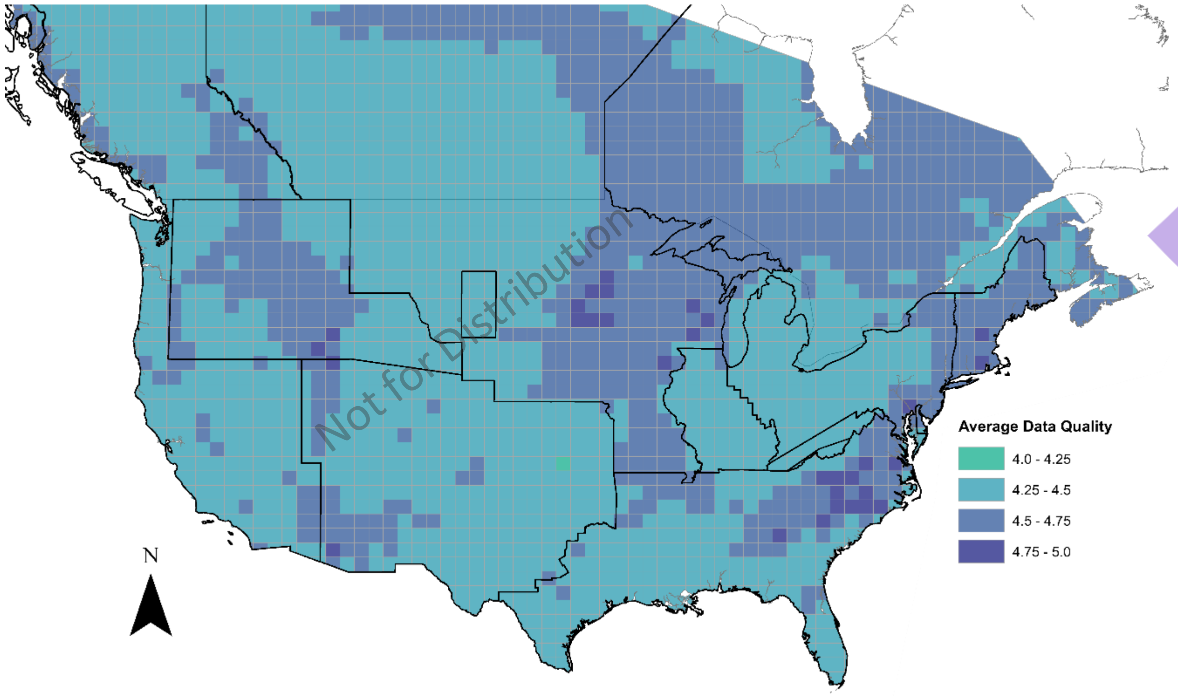
spaCy

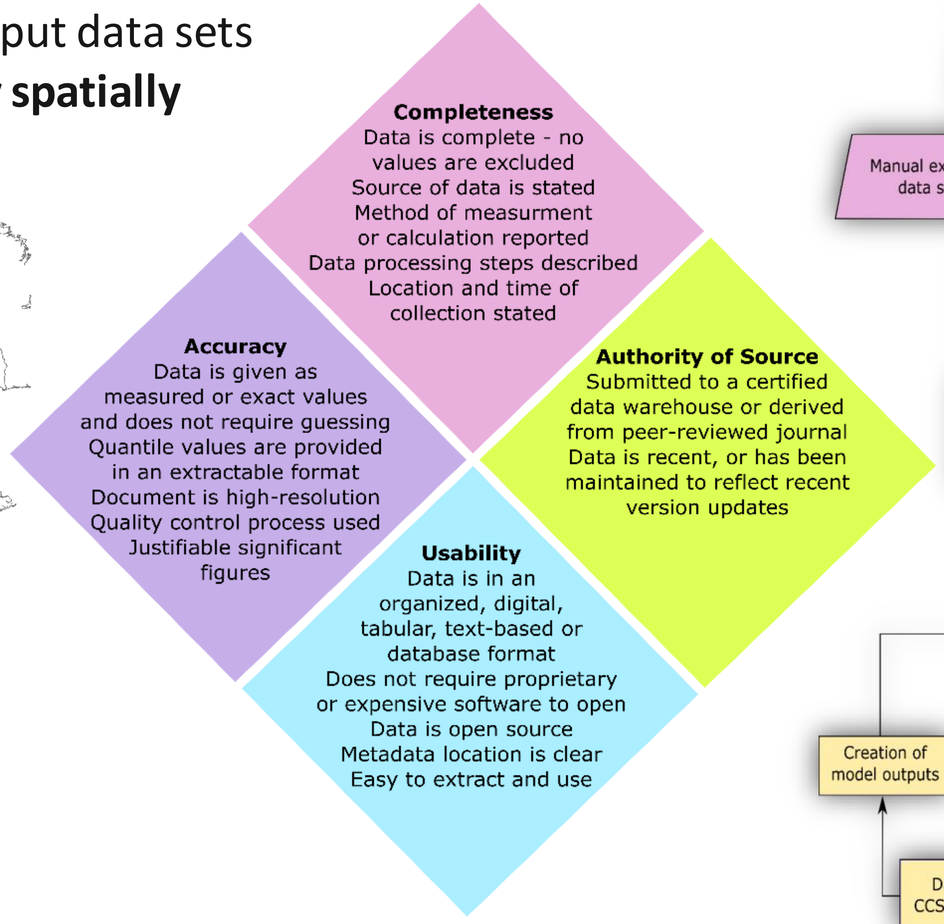


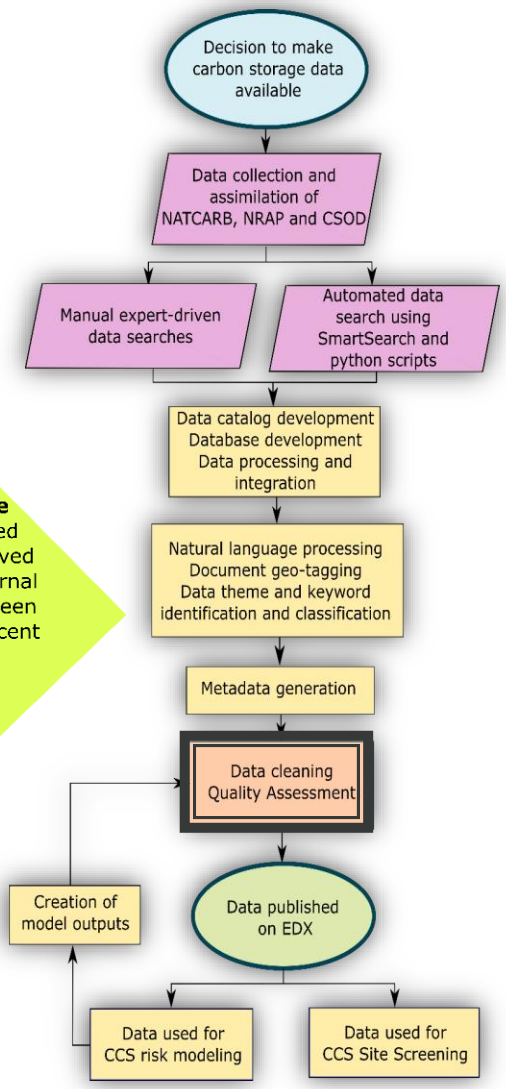


Results: Data Quality assessment method development and spatial trends in CS data quality

- **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**
- **Manuscript outlining method in prep**







18

Summary

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

Next Steps

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:

- **Continue collecting and adding data to EDX, Geocube, and LivingDatabase**
- **Develop additional integrations between SmartSearch, SmartParse, and EDX**
- **Improve ML models and NLP analysis utilizeing additional libraries, developing more training data, and applications**
- **Share and expand technology and data resources across NETL projects to improve and expand data curation**

EDX™ Energy Data eXchange

NETL SmartSearch

NETL SmartParse

GEO CUBE

NETL CARBON STORAGE

NRAP National Risk Assessment Partnership

NatCarb A National Look at Carbon Sequestration

Thank you!

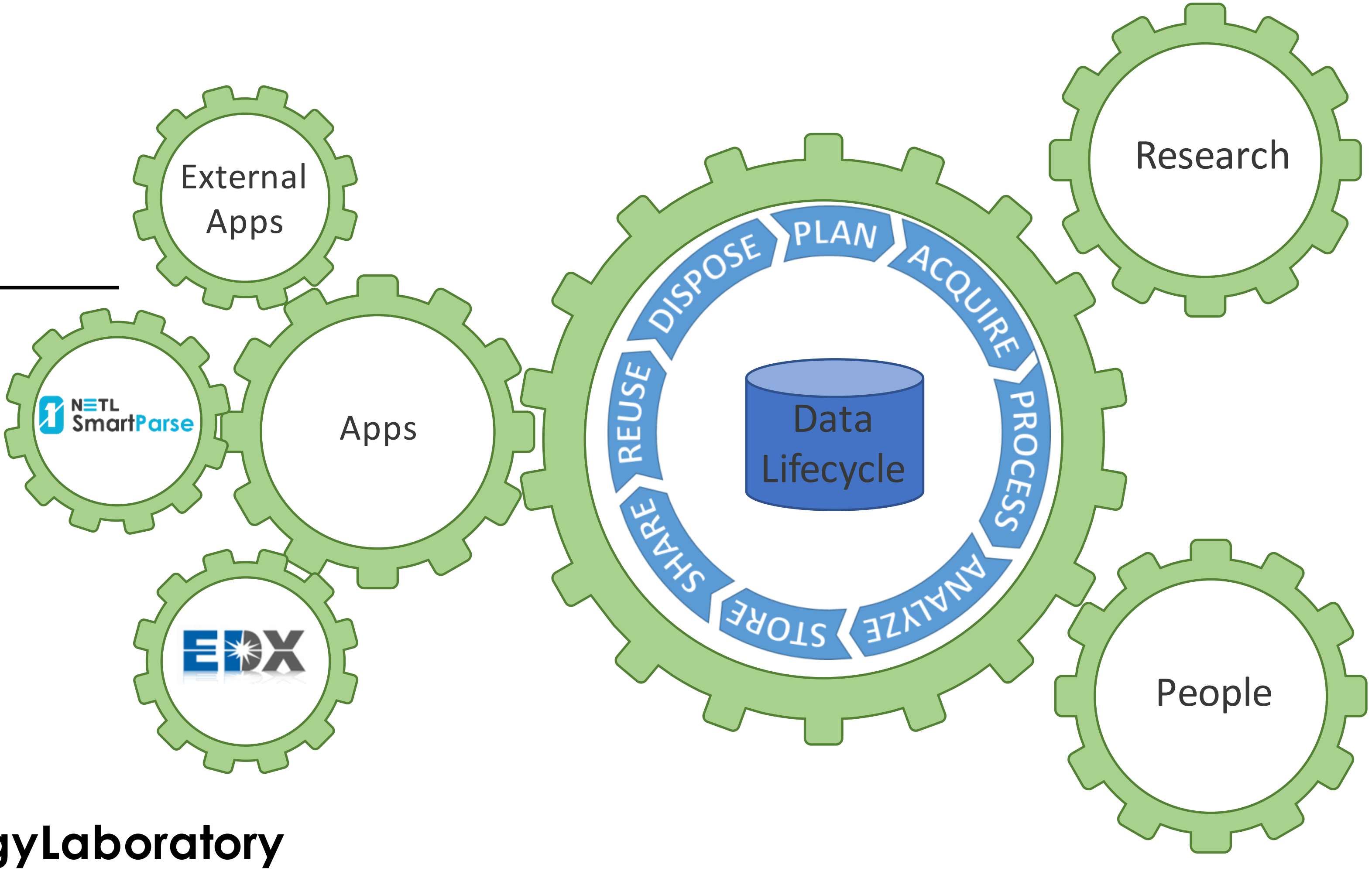
NETL RESOURCES

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

@NETL_DOE

@NETL_DOE

@NationalEnergyTechnologyLaboratory



CONTACT:
Michael Sabbatino and Paige Morkner
Michael.Sabbatino@netl.doe.gov, Paige.Morkner@netl.doe.gov

U.S. DEPARTMENT OF ENERGY

Appendix

– These slides will not be discussed during the presentation, but are mandatory.

Benefit to the Program

- 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.

Project Overview

Goals and Objectives

- 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

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

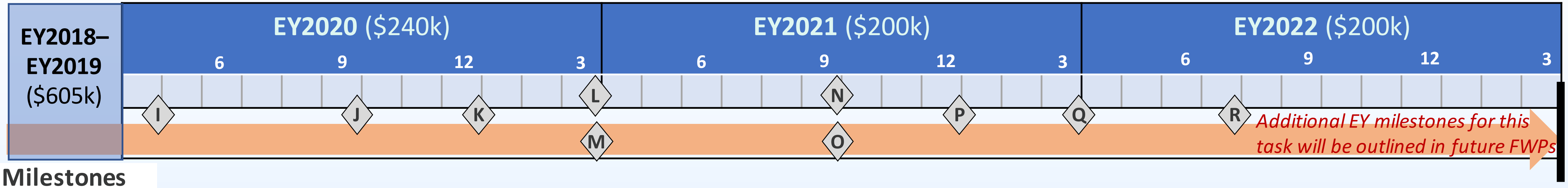
Lead: Jennifer Bauer
Contractors: **Paige Morkner**, Michael Sabbatino, Patrick Wingo, Andrew Bean, TJ Jones, Aaron Barkhurst, other Matric Software Engineers and Developers

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

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

Task 28.0: Project Timeline Overview

Curation of Carbon Storage R&D Products Through Advanced Data Computing Solutions (PIs: Michael Sabbatino, Jennifer Bauer)



Milestones

Number	Expected Completion Date	Milestone Description
EY20.28.I	04/30/2020	Push to public on EDX appropriate **MGSC** Partnership data products.
EY20.28.J	09/30/2020	Deploy LivingDatabase beta version capability in EDX, private side, for CS teams (e.g., RCSPs) use and testing.
EY20.28.K	12/31/2020	Integration of CSP data products that are spatially related through enhanced EDX spatial search and discovery tool on GeoCube.
EY20.28.L	03/31/2021	Deploy NETL SmartSearch version 2 algorithm in EDX to support automated gathering of open, CS relevant data.
EY20.28.M	03/31/2021	Deploy LivingDatabase version 1 capability in EDX, private side, for CS teams (e.g., RCSPs) use and testing.
EY21.28.N	09/30/2021	Develop and test SmartSearch and SmartParse beta integration.
EY21.28.O	09/30/2021	Complete testing of Living Database dashboard tools.
EY21.28.P	12/31/2021	Create additional training data for SmartParse image, graph, and table extraction model improvement.
EY21.28.Q	03/31/2022	Develop beta Living Database user interface and dashboard.
EY22.28.R	07/29/2022	Ingestion and push to public on EDX appropriate **SW Regional Partnership** data products.

Chart Key
- Milestone
- Project Completion
- Go/No-Go Timeframe

Key Accomplishments/Deliverables

- 2018–Present, Addition of **Big Sky**, PCOR, Midwest CS Partnership, SECARB, and MGSC data and resources on EDX, for a combined total of 3,037 and 1.64 TB of data
- 2018–2020, Big data computing cluster, Watt, set up and work to directly link EDX with these computing capabilities
- 2019–2021, Test and validate SmartSearch for use with commercial cloud & EDX to evaluate capabilities to assimilate relevant CS data; including work as part of an NDA with Google and collaboration with DOE-HQ OCIO
- 2020–2021, Develop Living Database logic to host and storge large volumes of CS data
- 2021–2022, Deploy beta instance of Living Database front end and dashboard tools
- 2022, Addition of any final RCSP and other CS resources to EDX

Value Delivered

- **Collecting, curating, and cataloging** data from all regional CS partnerships and 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 CS 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 28.0 is integrating data into an existing tool with no development of a technology. Therefore, no TRL is assigned.

Bibliography

– List peer reviewed publications generated from the project per the format of the examples below.

- Morkner, P., Bauer, J., Creason, C., Bean, A., and Rose, K., "A Data Quality Assessment Method to Support Carbon Storage," in preparation . Target journal: *Nature Scientific Data*. (Tasks 27.0, 28.0)
- Morkner, P., Creason, C., Sabbatino, M., Wingo, P., DiGiulio, J., Jones, K., Greenburg, R., Bauer, J., and Rose, K., "Distilling Data to Drive Carbon Storage Insights," accepted pending final revisions, *Computers and Geosciences*. (Tasks 27.0, 28.0)
- Barkhurst, A., Morkner, P., Bauer, J., Rose, K. GeoCube, TRS report, in prep, target completion Fall 2021.