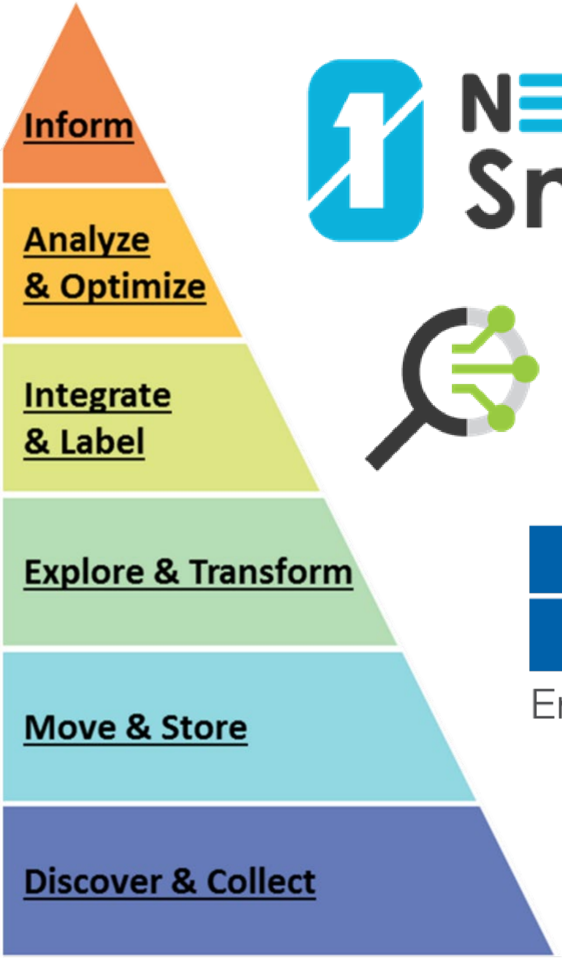
Advanced Data Extraction to Support a Living Database

NATIONAL ENERGY TECHNOLOGY LABORATORY

Michael Sabbatino

NETL Support Contractor
Research Innovation Center



Inform

Analyze & Optimize

Integrate & Label

Explore & Transform

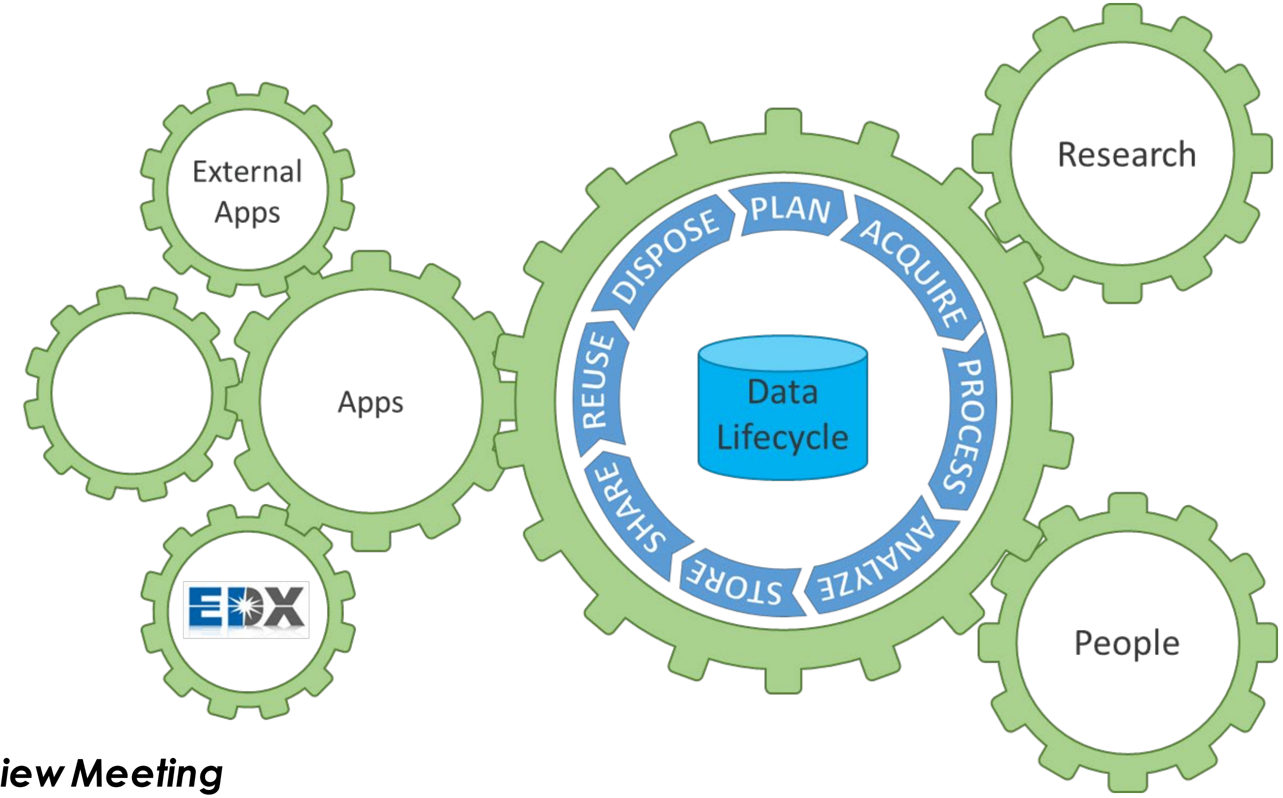
Move & Store

Discover & Collect

NETL SmartParse

NETL SmartSearch

EDX™
Energy Data eXchange



2022 Carbon Management Project Review Meeting

August 17, 2022

Disclaimer

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 the 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

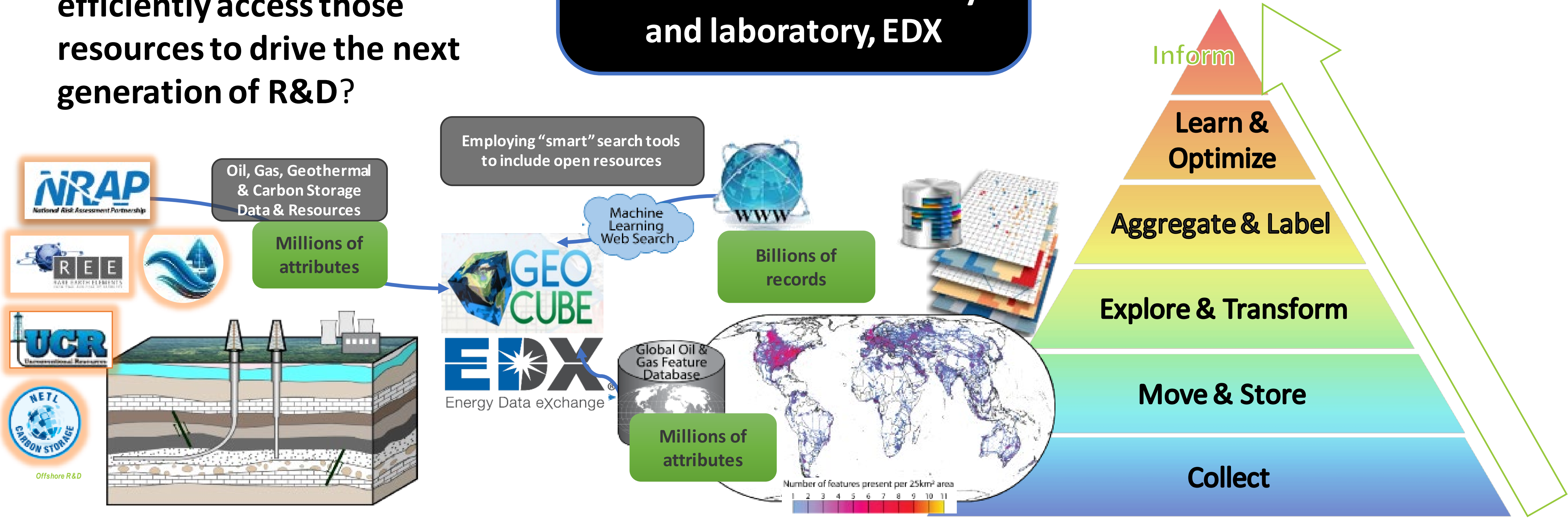
[3] NETL Support Contractor, 3610 Collins Ferry Rd., Morgantown, WV, 26505, USA

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 advanced AI/ML tools via DOE's virtual data library and laboratory, EDX

NETL SmartParse



Employing "smart" search tools to include open resources

Oil, Gas, Geothermal & Carbon Storage Data & Resources

Millions of attributes

Machine Learning Web Search

Billions of records

Global Oil & Gas Feature Database

Millions of attributes

Inform

Learn & Optimize

Aggregate & Label

Explore & Transform

Move & Store

Collect

Carbon Storage Data Lifecycle

NATIONAL ENERGY TECHNOLOGY LABORATORY

Collection

NETL SmartSearch

- SmartSearch
- Expert-driven research
- EDX submissions

Metadata development and capture

NETL SmartParse

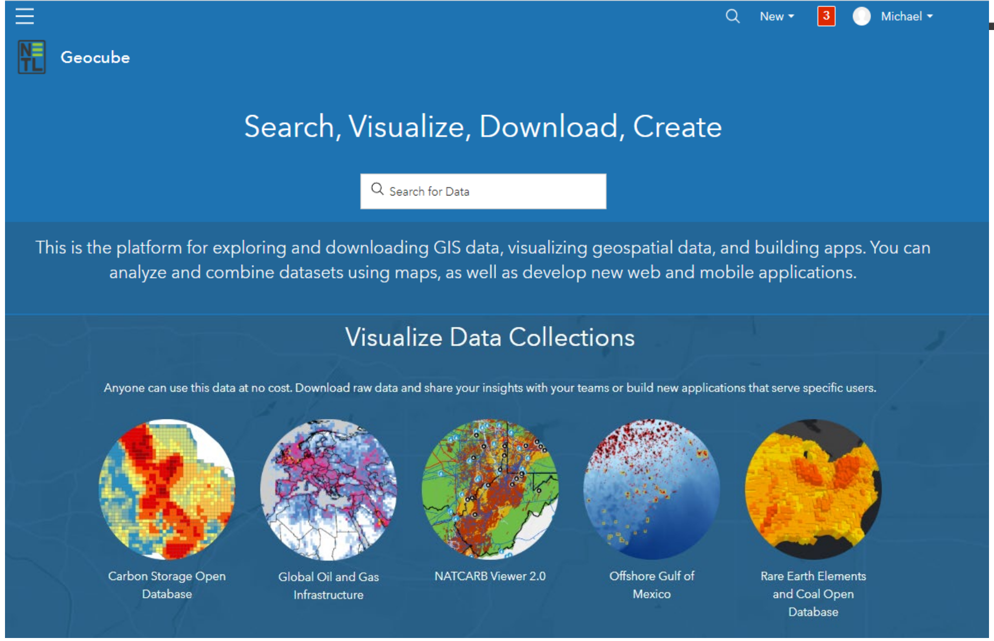
- Cataloging
- ReadMe file development
- Natural language processing for keywords, topic modeling, geographic association

Quality Assessment

- Data ranking
- Data assessment method scoring

Data Organization and publishing

- Private workspaces
- Submission packaging
- ArcGIS Enterprise with EDX
  - GeoCube data integration
  - Collaborative Map Applications



Search, Visualize, Download, Create

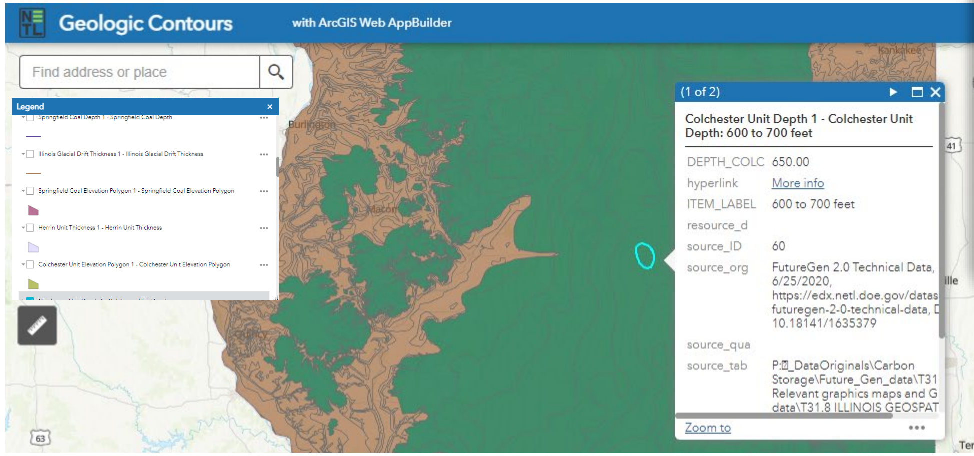
Search for Data

This is the platform for exploring and downloading GIS data, visualizing geospatial data, and building apps. You can analyze and combine datasets using maps, as well as develop new web and mobile applications.

Visualize Data Collections

Anyone can use this data at no cost. Download raw data and share your insights with your teams or build new applications that serve specific users.

Carbon Storage Open Database
Global Oil and Gas Infrastructure
NATCARB Viewer 2.0
Offshore Gulf of Mexico
Rare Earth Elements and Coal Open Database



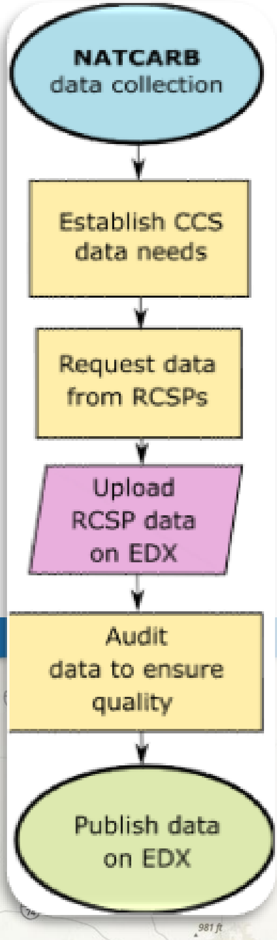
Geologic Contours
with ArcGIS Web AppBuilder
Find address or place
Legend
Springfield Coal Depth 1 - Springfield Coal Depth
Illinois Glacial Drift Thickness 1 - Illinois Glacial Drift Thickness
Springfield Coal Elevation Polygon 1 - Springfield Coal Elevation Polygon
Herrin Unit Thickness 1 - Herrin Unit Thickness
Colchester Unit Elevation Polygon 1 - Colchester Unit Elevation Polygon
Burlington
Macomb
Kankakee
41
63
Indiana
Indianapolis
Terre Haute
981 ft

(1 of 2)

Colchester Unit Depth 1 - Colchester Unit Depth: 600 to 700 feet

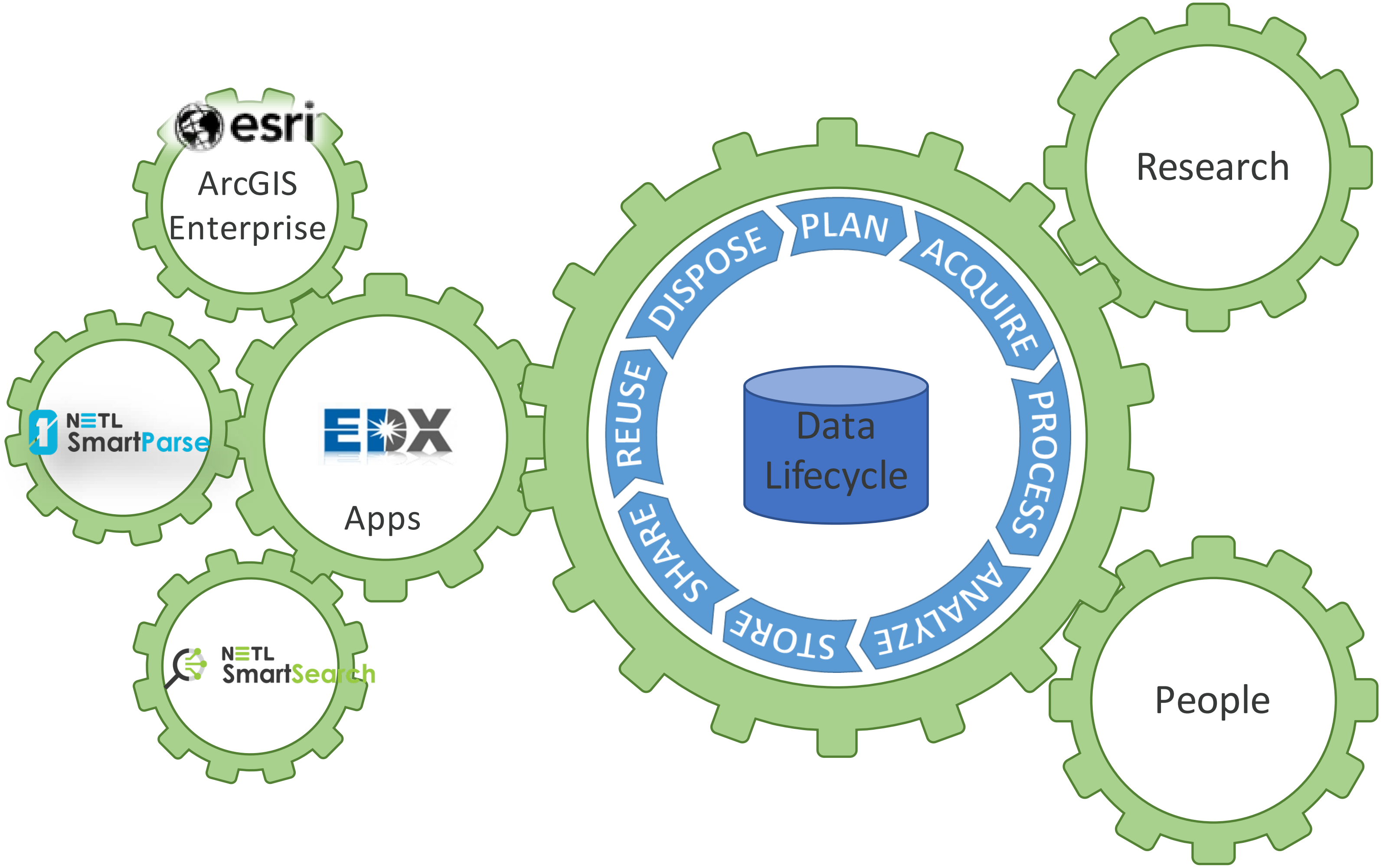
DEPTH_COLC 650.00
hyperlink More info
ITEM_LABEL 600 to 700 feet
resource_d
source_ID 60
source_org FutureGen 2.0 Technical Data, 6/25/2020, https://edx.netl.doe.gov/datas futuregen-2-0-technical-data, D 10.18141/1635379
source_qua
source_tab P:_DataOriginals\Carbon Storage\Future_Gen_data\T31 Relevant graphics maps and G data\T31.8 ILLINOIS GEOSPAT

Zoom to



Expanding Integrations for a Living Database

- **Expanded to Utilize ArcEnterprise GIS data management System hosted on AWS servers**
  - **Store/Share Geospatial Data**
  - **Host New Geocube Web Application**
  - **Custom Interactive Map Applications**
- **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**



Types of Carbon Storage Data

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

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

Image Extraction:
- Documents
- Presentations
- Maps
- Posters

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

EDX Energy Data eXchange

Feeding the data hippo!

Advanced AI/ML Tools for CS Data Lifecycle

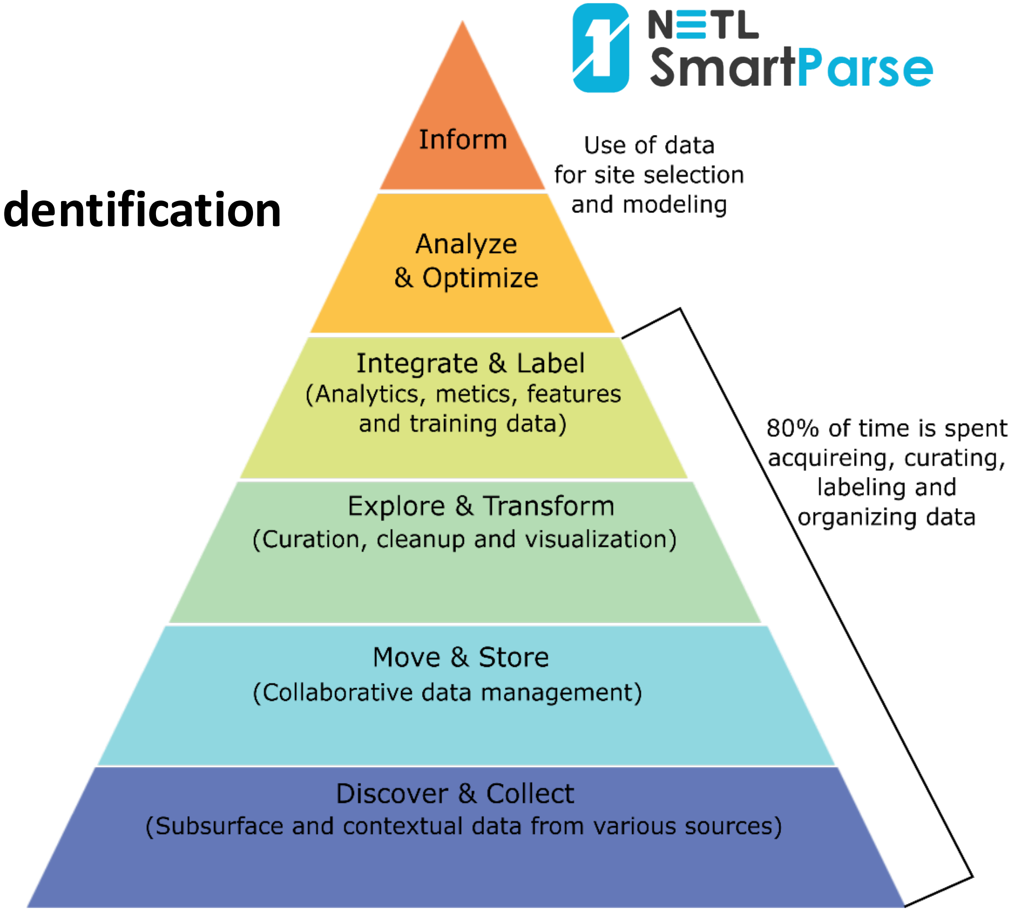
NETL SmartSearch

NETL SmartParse

Challenge: Continue increasing available data while enhancing metadata, searchability, and curation.

Solutions:

- **Natural Language Processing:**
  - **text-based resource classification, organization, keyword identification**
  - **metadata extraction and preservation**
  - **geographic association (for searchability)**
- **Image Classification and Text Extraction**
  - **Identify images from papers, posters, and documents**
  - **Classify images and extract text**
  - **Extract image metadata**
- **ArcGIS Enterprise on AWS:**
  - **Geographic database development (Geocube)**
  - **Interactive map creation and collaboration**
  - **Integration with EDX**



Inform

Use of data for site selection and modeling

Analyze & Optimize

Integrate & Label (Analytics, metics, features and training data)

Explore & Transform (Curation, cleanup and visualization)

Move & Store (Collaborative data management)

Discover & Collect (Subsurface and contextual data from various sources)

80% of time is spent acquireing, curating, labeling and organizing data

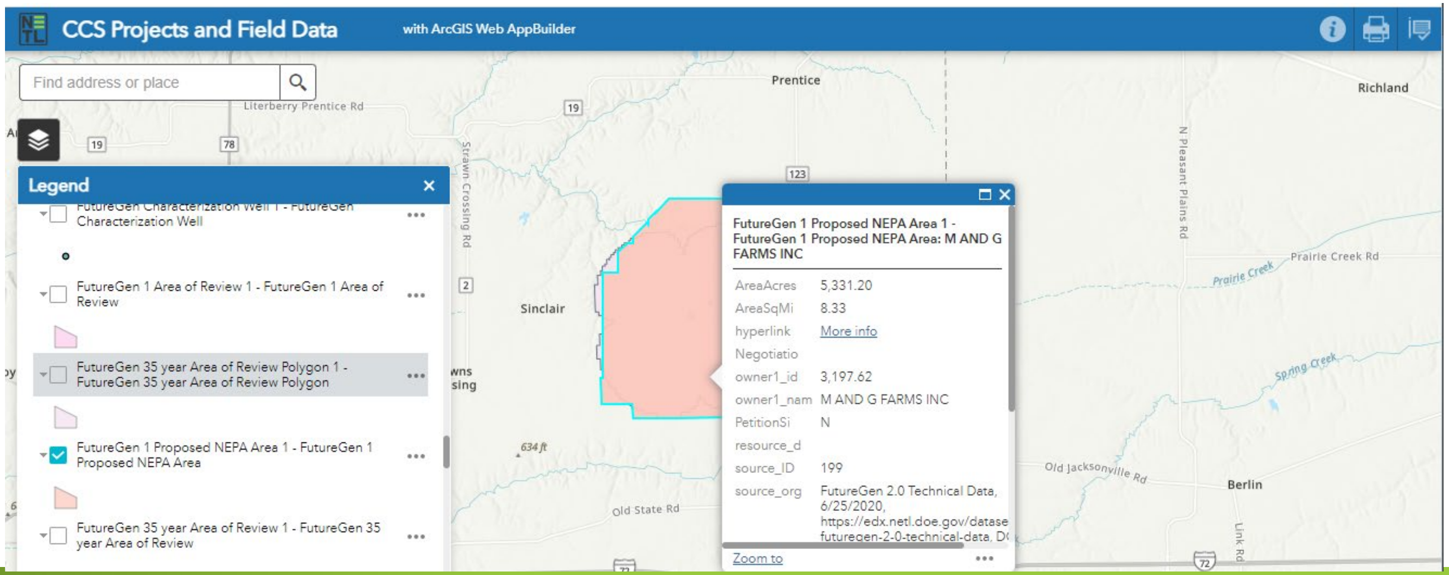
Data Cleaning for ML, AI and Spatial Analysis

Identify data to be collected

Includes:

- **Data, Papers, Catalogs of Data, Online Sources , and Metadata**

Data collected and processed using Python tools to move, quantify and label data



Legend

□ FutureGen Characterization Well 1 - FutureGen Characterization Well

□ FutureGen 1 Area of Review 1 - FutureGen 1 Area of Review

□ FutureGen 35 year Area of Review Polygon 1 - FutureGen 35 year Area of Review Polygon

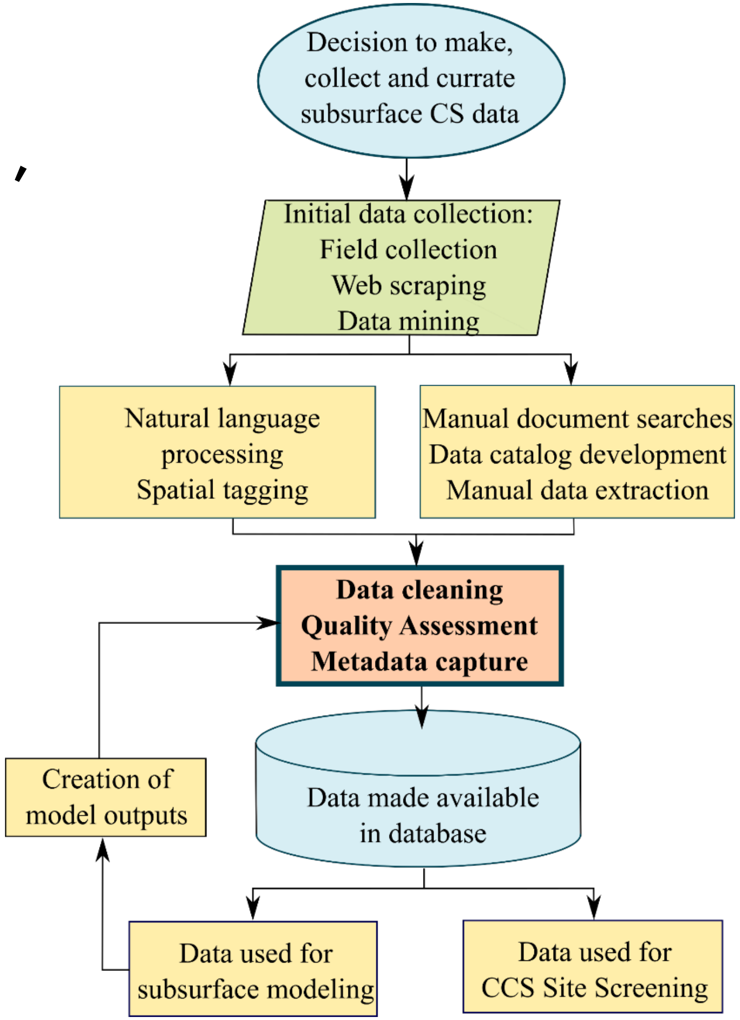
☑ FutureGen 1 Proposed NEPA Area 1 - FutureGen 1 Proposed NEPA Area

□ FutureGen 35 year Area of Review 1 - FutureGen 35 year Area of Review

FutureGen 1 Proposed NEPA Area 1 - FutureGen 1 Proposed NEPA Area: M AND G FARMS INC

AreaAcres	5,331.20
AreaSqMi	8.33
hyperlink	More info
Negotiatio	
owner1_id	3,197.62
owner1_nam	M AND G FARMS INC
PetitionSi	N
resource_d	
source_ID	199
source_org	FutureGen 2.0 Technical Data, 6/25/2020, https://edx.netl.doe.gov/datase futuregen-2-0-technical-data, D(

Zoom to



Decision to make, collect and currate subsurface CS data

Initial data collection: Field collection Web scraping Data mining

Natural language processing Spatial tagging

Manual document searches Data catalog development Manual data extraction

Data cleaning Quality Assessment Metadata capture

Data made available in database

Creation of model outputs

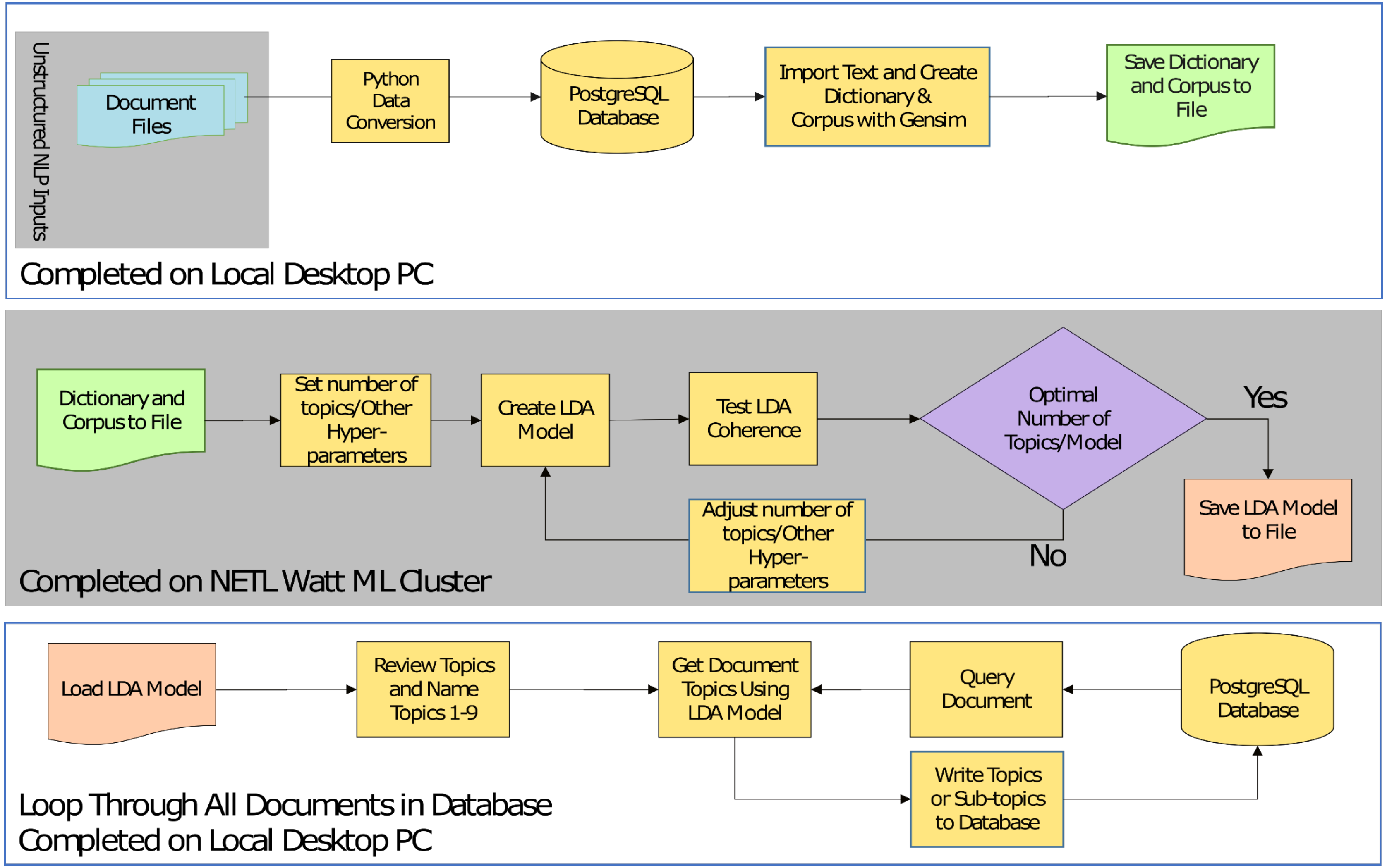
Data used for subsurface modeling

Data used for CCS Site Screening

Natural Language Processing (NLP) Unsupervised ML for Document Classification

NETL SmartParse

- Latent Dirichlet allocation **(LDA) model based on corpus of 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 with metadata on EDX**
- **Parse geographic location to associate with each document** – when possible



Unstructured NLP Inputs

Document Files

Python Data Conversion

PostgreSQL Database

Import Text and Create Dictionary & Corpus with Gensim

Save Dictionary and Corpus to File

Completed on Local Desktop PC

Dictionary and Corpus to File

Set number of topics/Other Hyper-parameters

Create LDA Model

Test LDA Coherence

Optimal Number of Topics/Model

Yes

Save LDA Model to File

No

Adjust number of topics/Other Hyper-parameters

Completed on NETL Watt ML Cluster

Load LDA Model

Review Topics and Name Topics 1-9

Get Document Topics Using LDA Model

Query Document

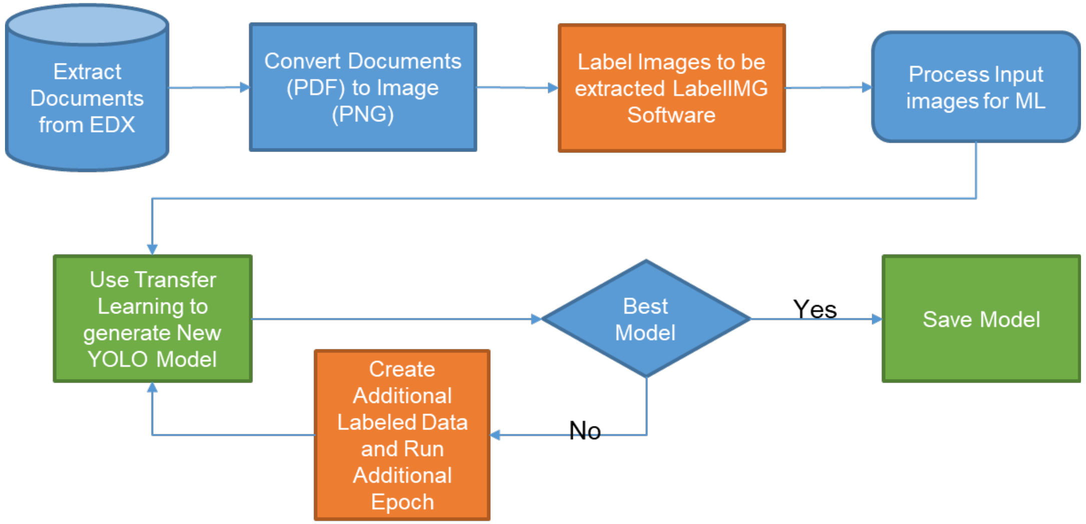
PostgreSQL Database

Write Topics or Sub-topics to Database

Loop Through All Documents in Database
Completed on Local Desktop PC

Machine Learning Image Data Extraction

- 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


Extract Documents from EDX
Convert Documents (PDF) to Image (PNG)
Label Images to be extracted LabelIMG Software
Process Input images for ML
Use Transfer Learning to generate New YOLO Model
Best Model
Yes
Save Model
No
Create Additional Labeled Data and Run Additional Epoch

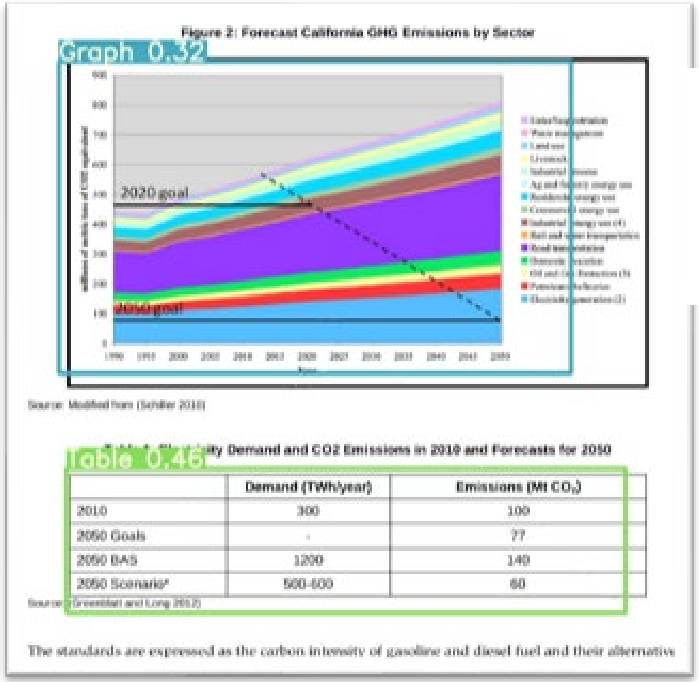


Figure 2: Forecast California GHG Emissions by Sector

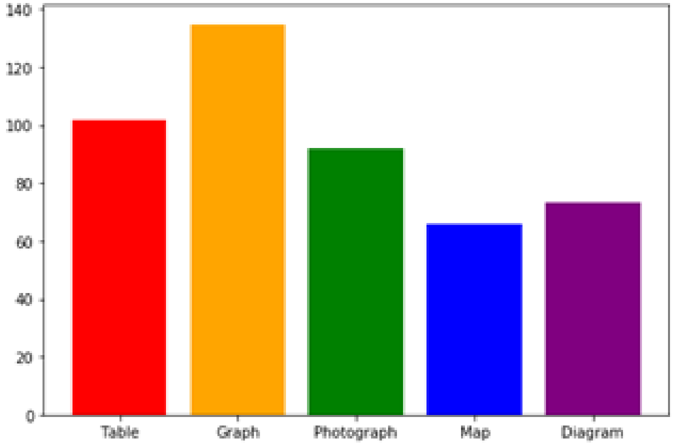
Source: Modified from (Schiller 2010)

...ity Demand and CO2 Emissions in 2010 and Forecasts for 2050

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

Source: Greenblatt and Long 2012)

The standards are expressed as the carbon intensity of gasoline and diesel fuel and their alternative



Images and Tables Targeted for Data for Extraction

Machine Learning Data Extraction

Utilize Object Identification ML Models to Extract Additional Data

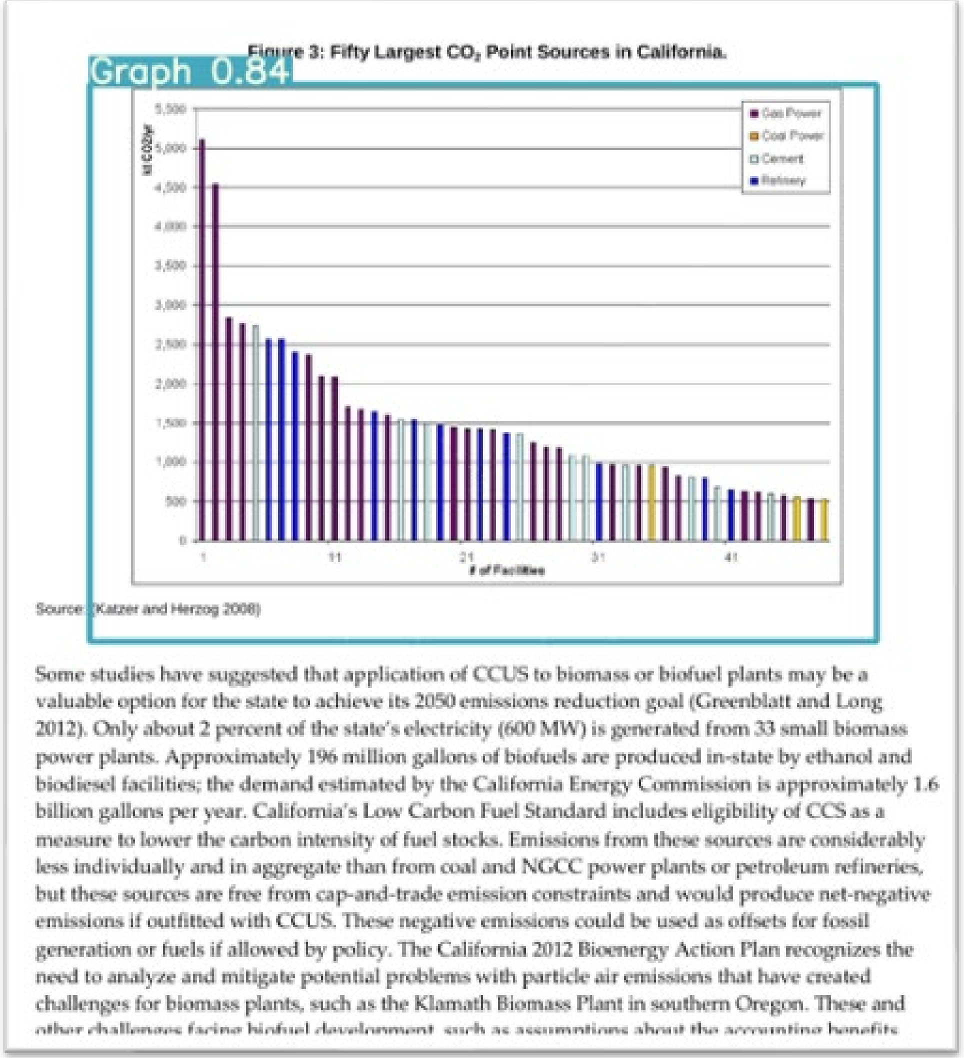
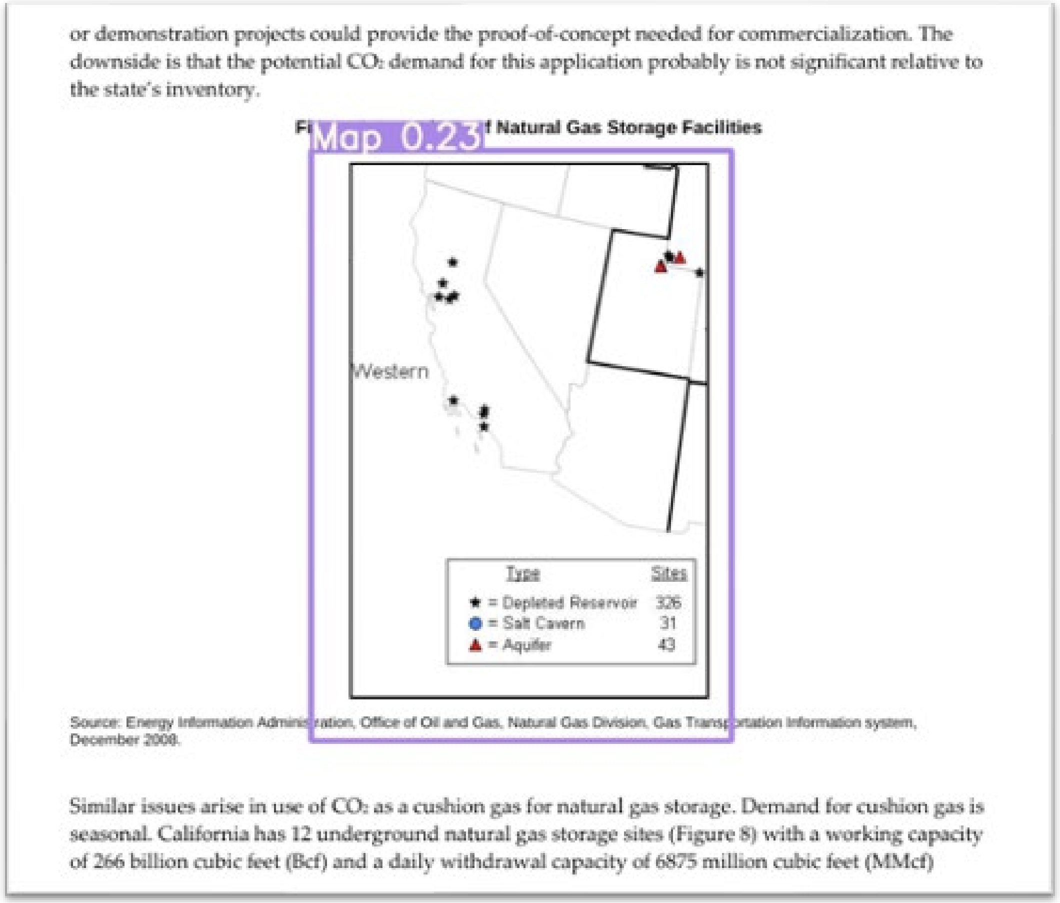


Figure 3: Fifty Largest CO_2 Point Sources in California.

Source: (Katzer and Herzog 2008)

Some studies have suggested that application of CCUS to biomass or biofuel plants may be a valuable option for the state to achieve its 2050 emissions reduction goal (Greenblatt and Long 2012). Only about 2 percent of the state's electricity (600 MW) is generated from 33 small biomass power plants. Approximately 196 million gallons of biofuels are produced in-state by ethanol and biodiesel facilities; the demand estimated by the California Energy Commission is approximately 1.6 billion gallons per year. California's Low Carbon Fuel Standard includes eligibility of CCS as a measure to lower the carbon intensity of fuel stocks. Emissions from these sources are considerably less individually and in aggregate than from coal and NGCC power plants or petroleum refineries, but these sources are free from cap-and-trade emission constraints and would produce net-negative emissions if outfitted with CCUS. These negative emissions could be used as offsets for fossil generation or fuels if allowed by policy. The California 2012 Bioenergy Action Plan recognizes the need to analyze and mitigate potential problems with particle air emissions that have created challenges for biomass plants, such as the Klamath Biomass Plant in southern Oregon. These and other challenges facing biofuel development, such as assumptions about the accounting benefits



or demonstration projects could provide the proof-of-concept needed for commercialization. The downside is that the potential CO_2 demand for this application probably is not significant relative to the state's inventory.

Fi f Natural Gas Storage Facilities

Source: Energy Information Adminis ation, Office of Oil and Gas, Natural Gas Division, Gas Trans ortation Information system, December 2008.

Similar issues arise in use of CO_2 as a cushion gas for natural gas storage. Demand for cushion gas is seasonal. California has 12 underground natural gas storage sites (Figure 8) with a working capacity of 266 billion cubic feet (Bcf) and a daily withdrawal capacity of 6875 million cubic feet (MMcf)

Machine Learning Data Extraction

Utilize Object Identification ML Models to Extract Additional Data

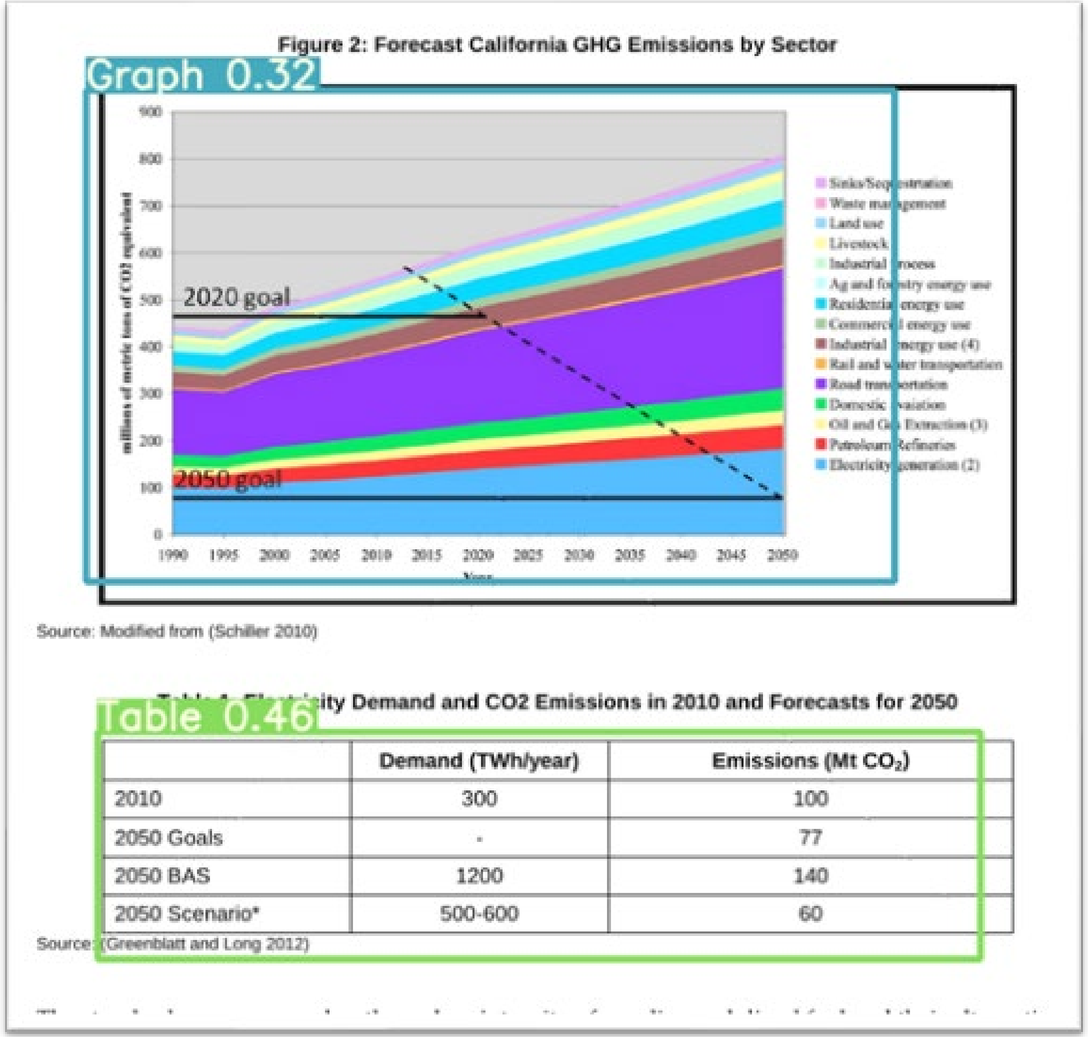


Figure 2: Forecast California GHG Emissions by Sector

Source: Modified from (Schiller 2010)

Table 4: Electricity Demand and CO2 Emissions in 2010 and Forecasts for 2050

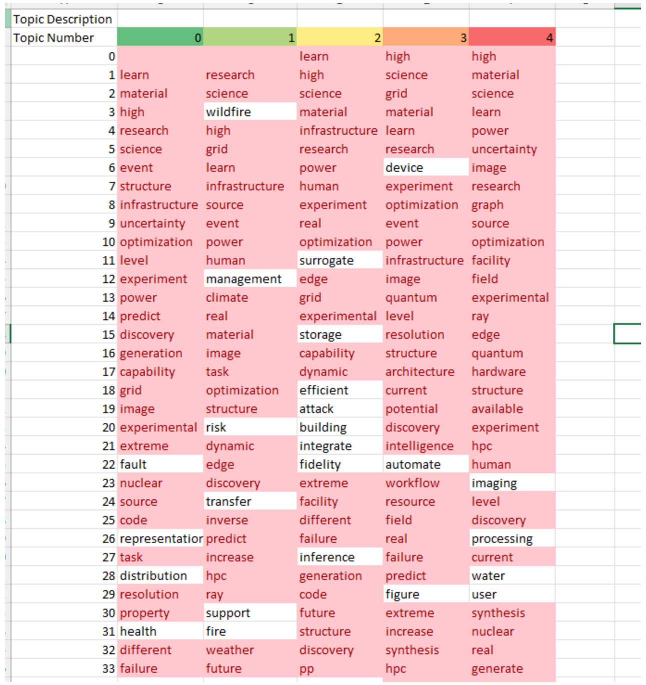
	Demand (TWh/year)	Emissions (Mt CO_2)
2010	300	100
2050 Goals	-	77
2050 BAS	1200	140
2050 Scenario*	500-600	60

Source: (Greenblatt and Long 2012)

Use Case: Abstract White Paper Data Extraction

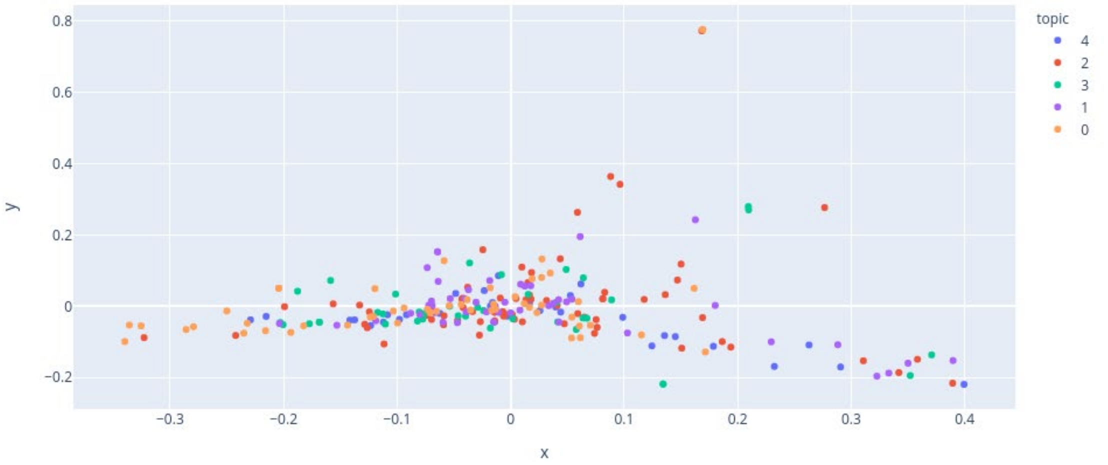
NLP and Machine learning to Classify and Text and Images

- 10 topic, 5 topic and variable
- PCA Analysis
- Keywords and Custom Stop Word List

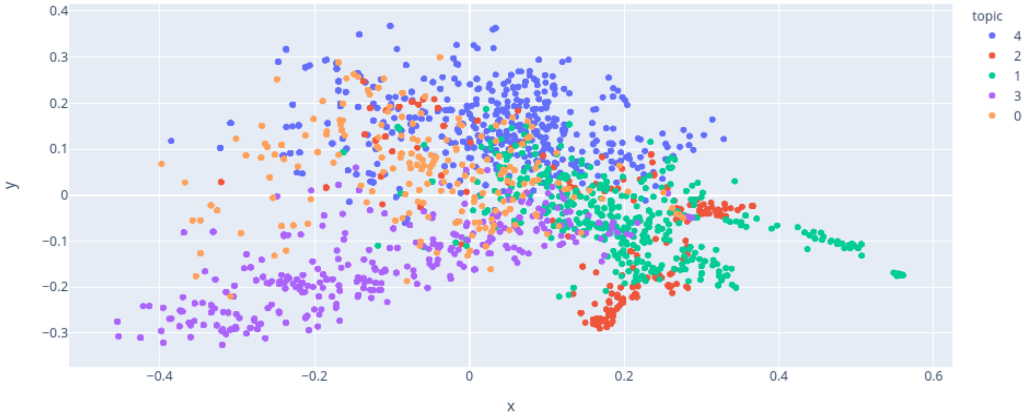


Topic Description					
Topic Number	0	1	2	3	4
0			learn	high	high
1	learn	research	high	science	material
2	material	science	science	grid	science
3	high	wildfire	material	material	learn
4	research	high	infrastructure	learn	power
5	science	grid	research	research	uncertainty
6	event	learn	power	device	image
7	structure	infrastructure	human	experiment	research
8	infrastructure	source	experiment	optimization	graph
9	uncertainty	event	real	event	source
10	optimization	power	optimization	power	optimization
11	level	human	surrogate	infrastructure	facility
12	experiment	management	edge	image	field
13	power	climate	grid	quantum	experimental
14	predict	real	experimental	level	ray
15	discovery	material	storage	resolution	edge
16	generation	image	capability	structure	quantum
17	capability	task	dynamic	architecture	hardware
18	grid	optimization	efficient	current	structure
19	image	structure	attack	potential	available
20	experimental	risk	building	discovery	experiment
21	extreme	dynamic	integrate	intelligence	hpc
22	fault	edge	fidelity	automate	human
23	nuclear	discovery	extreme	workflow	imaging
24	source	transfer	facility	resource	level
25	code	inverse	different	field	discovery
26	representatior	predict	failure	real	processing
27	task	increase	inference	failure	current
28	distribution	hpc	generation	predict	water
29	resolution	ray	code	figure	user
30	property	support	future	extreme	synthesis
31	health	fire	structure	increase	nuclear
32	different	weather	discovery	synthesis	real
33	failure	future	pp	hpc	generate

PCA Papers 5 Topics 241 Papers



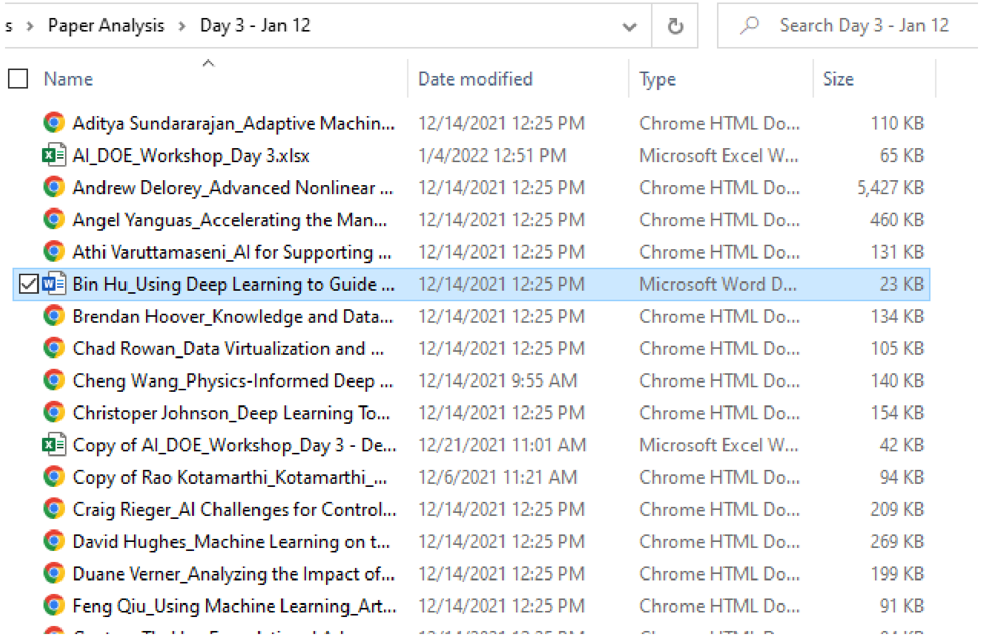
PCA AMO 5 Topics 3037 Documents



Use Case: Abstract White Paper Data Extraction

NLP and Machine learning to Classify and Text and Images

- Data Input Zipped Papers, Spreadsheets, and Images
- Process with NLP (Gensim) to Create topic model
- Convert PDF to JPG for preprocessing
- Used trained Yolo model using transfer learning
- Extracted images from papers and classify



s › Paper Analysis › Day 3 - Jan 12

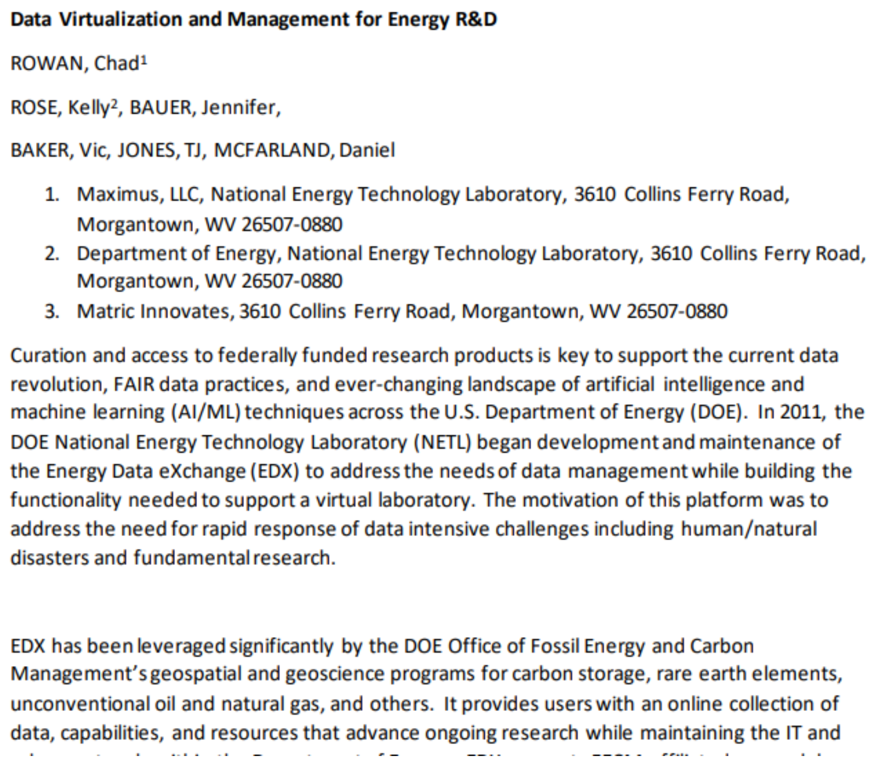
Search Day 3 - Jan 12

□ Name	Date modified	Type	Size
Aditya Sundararajan_Adaptive Machin...	12/14/2021 12:25 PM	Chrome HTML Do...	110 KB
AI_DOE_Workshop_Day 3.xlsx	1/4/2022 12:51 PM	Microsoft Excel W...	65 KB
Andrew Delorey_Advanced Nonlinear ...	12/14/2021 12:25 PM	Chrome HTML Do...	5,427 KB
Angel Yanguas_Accelerating the Man...	12/14/2021 12:25 PM	Chrome HTML Do...	460 KB
Athi Varuttamaseni_AI for Supporting ...	12/14/2021 12:25 PM	Chrome HTML Do...	131 KB
☑ Bin Hu_Using Deep Learning to Guide ...	12/14/2021 12:25 PM	Microsoft Word D...	23 KB
Brendan Hoover_Knowledge and Data...	12/14/2021 12:25 PM	Chrome HTML Do...	134 KB
Chad Rowan_Data Virtualization and ...	12/14/2021 12:25 PM	Chrome HTML Do...	105 KB
Cheng Wang_Physics-Informed Deep ...	12/14/2021 9:55 AM	Chrome HTML Do...	140 KB
Christoper Johnson_Deep Learning To...	12/14/2021 12:25 PM	Chrome HTML Do...	154 KB
Copy of AI_DOE_Workshop_Day 3 - De...	12/21/2021 11:01 AM	Microsoft Excel W...	42 KB
Copy of Rao Kotamarthi_Kotamarthi_...	12/6/2021 11:21 AM	Chrome HTML Do...	94 KB
Craig Rieger_AI Challenges for Control...	12/14/2021 12:25 PM	Chrome HTML Do...	209 KB
David Hughes_Machine Learning on t...	12/14/2021 12:25 PM	Chrome HTML Do...	269 KB
Duane Verner_Analyzing the Impact of...	12/14/2021 12:25 PM	Chrome HTML Do...	199 KB
Feng Qiu_Using Machine Learning_Art...	12/14/2021 12:25 PM	Chrome HTML Do...	91 KB

Use Case: Abstract White Paper Data Extraction

NLP and Machine learning to Classify and Text and Images

- Data Input Zipped Papers, Spreadsheets, and Images
- Process with NLP (Gensim) to Create topic model
- Convert PDF to JPG for preprocessing
- Used trained Yolo model using transfer learning
- Extracted images from papers and classify



Data Virtualization and Management for Energy R&D

ROWAN, Chad[1]

ROSE, Kelly[2], BAUER, Jennifer,

BAKER, Vic, JONES, TJ, MCFARLAND, Daniel

1. Maximus, LLC, National Energy Technology Laboratory, 3610 Collins Ferry Road, Morgantown, WV 26507-0880
2. Department of Energy, National Energy Technology Laboratory, 3610 Collins Ferry Road, Morgantown, WV 26507-0880
3. Matric Innovates, 3610 Collins Ferry Road, Morgantown, WV 26507-0880

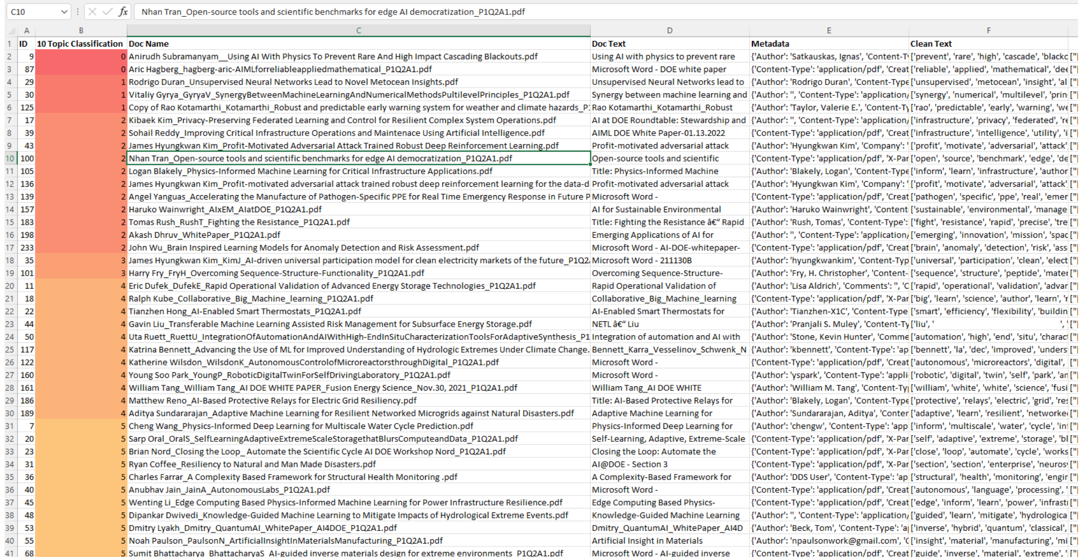
Curation and access to federally funded research products is key to support the current data revolution, FAIR data practices, and ever-changing landscape of artificial intelligence and machine learning (AI/ML) techniques across the U.S. Department of Energy (DOE). In 2011, the DOE National Energy Technology Laboratory (NETL) began development and maintenance of the Energy Data eXchange (EDX) to address the needs of data management while building the functionality needed to support a virtual laboratory. The motivation of this platform was to address the need for rapid response of data intensive challenges including human/natural disasters and fundamental research.

EDX has been leveraged significantly by the DOE Office of Fossil Energy and Carbon Management's geospatial and geoscience programs for carbon storage, rare earth elements, unconventional oil and natural gas, and others. It provides users with an online collection of data, capabilities, and resources that advance ongoing research while maintaining the IT and

Use Case: Abstract White Paper Data Extraction

NLP and Machine learning to Classify and Text and Images

- Data Input Zipped Papers, Spreadsheets, and Images
- Process with NLP (Gensim) to Create topic model
- Convert PDF to JPG for preprocessing
- Used trained Yolo model using transfer learning
- Extracted images from papers and classify

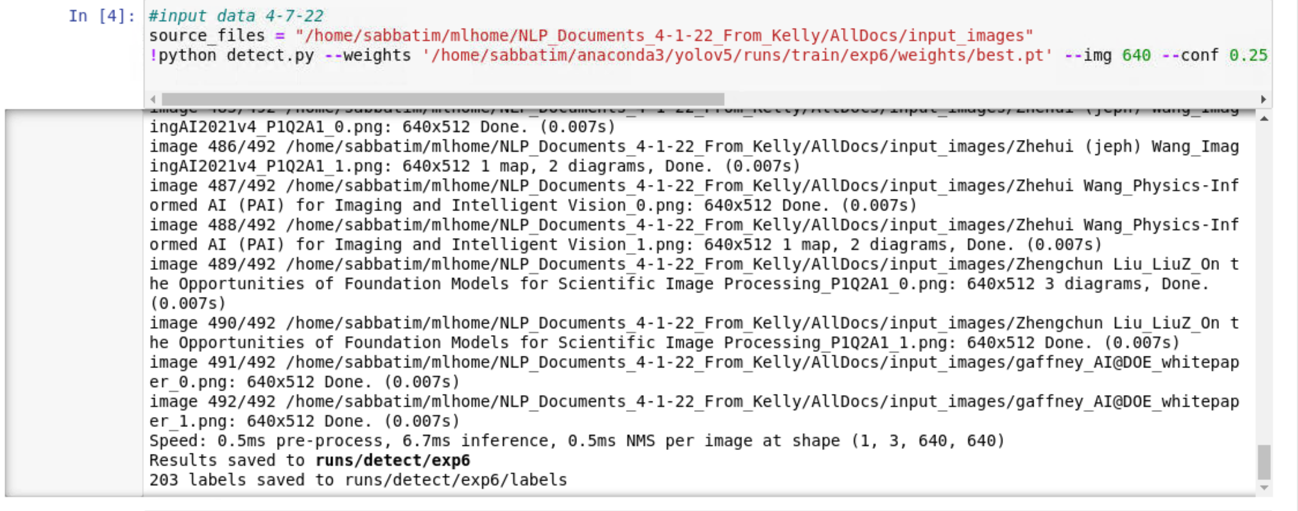


C10 Nhan Tran_Open-source tools and scientific benchmarks for edge AI democratization_P1Q2A1.pdf

ID	10 Topic Classification	Doc Name	Doc Text	Metadata	Clean Text	
9	0	Anirudh Subramanyam__Using AI With Physics To Prevent Rare And High Impact Cascading Blackouts.pdf	Using AI with physics to prevent rare	{'Author': 'Satkauskas, Ignas', 'Content-T	['prevent', 'rare', 'high', 'cascade', 'blackc	["
87	0	Aric Hagberg_hagberg-aric-AIMLforreliableappliedmathematical_P1Q2A1.pdf	Microsoft Word - DOE white paper	{'Content-Type': 'application/pdf', 'Creat	['reliable', 'applied', 'mathematical', 'dee	["
29	1	Rodrigo Duran_Unsupervised Neural Networks Lead to Novel Metocean Insights.pdf	Unsupervised Neural Networks lead to	{'Author': 'Rodrigo Duran', 'Content-Type	['unsupervised', 'metocean', 'insight', 'al	["
30	1	Vitaliy Gyrya_GyryaV_SynergyBetweenMachineLearningAndNumericalMethodsPultilevelPrinciples_P1Q2A1.pdf	Synergy between machine learning and	{'Author': '', 'Content-Type': 'application/	['synergy', 'numerical', 'multilevel', 'prin	["
125	1	Copy of Rao Kotamarthi_Kotamarthi_Robust and predictable early warning system for weather and climate hazards_P1	Rao Kotamarthi_Kotamarthi_Robust	{'Author': 'Taylor, Valerie E.', 'Content-Ty	['rao', 'predictable', 'early', 'warning', 'we	["
17	2	Kibaek Kim_Privacy-Preserving Federated Learning and Control for Resilient Complex System Operations.pdf	AI at DOE Roundtable: Stewardship and	{'Author': '', 'Content-Type': 'application/	['infrastructure', 'privacy', 'federated', 're	["
39	2	Sohail Reddy_Improving Critical Infrastructure Operations and Maintenace Using Artificial Intelligence.pdf	AIML DOE White Paper-01.13.2022	{'Content-Type': 'application/pdf', 'Creat	['infrastructure', 'intelligence', 'utility', 'i	["
43	2	James Hyungkwan Kim_Profit-Motivated Adversarial Attack Trained Robust Deep Reinforcement Learning.pdf	Profit-motivated adversarial attack	{'Author': 'Hyungkwan Kim', 'Company': '	['profit', 'motivate', 'adversarial', 'attack'	["
100	2	Nhan Tran_Open-source tools and scientific benchmarks for edge AI democratization_P1Q2A1.pdf	Open-source tools and scientific	{'Content-Type': 'application/pdf', 'X-Par	['open', 'source', 'benchmark', 'edge', 'de	["
105	2	Logan Blakely_Physics-Informed Machine Learning for Critical Infrastructure Applications.pdf	Title: Physics-Informed Machine	{'Author': 'Blakely, Logan', 'Content-Type	['inform', 'learn', 'infrastructure', 'author	["
136	2	James Hyungkwan Kim_Profit-motivated adversarial attack trained robust deep reinforcement learning for the data-d	Profit-motivated adversarial attack	{'Author': 'Hyungkwan Kim', 'Company': '	['profit', 'motivate', 'adversarial', 'attack'	["
139	2	Angel Yanguas_Accelerating the Manufacture of Pathogen-Specific PPE for Real Time Emergency Response in Future P	Microsoft Word -	{'Content-Type': 'application/pdf', 'Creat	['pathogen', 'specific', 'ppe', 'real', 'emer	["
157	2	Haruko Wainwright_AIxEM_AIatDOE_P1Q2A1.pdf	AI for Sustainable Environmental	{'Author': 'Haruko Wainwright', 'Content	['sustainable', 'environmental', 'manage	["
183	2	Tomas Rush_RushT_Fighting the Resistance_P1Q2A1.pdf	Title: Fighting the Resistance â€“ Rapid	{'Author': 'Rush, Tomas', 'Content-Type':	['fight', 'resistance', 'rapid', 'precise', 'tre	["
198	2	Akash Dhruv_WhitePaper_P1Q2A1.pdf	Emerging Applications of AI for	{'Author': '', 'Content-Type': 'application/	['emerging', 'innovation', 'mission', 'spac	["
233	2	John Wu_Brain Inspired Learning Models for Anomaly Detection and Risk Assessment.pdf	Microsoft Word - AI-DOE-whitepaper-	{'Content-Type': 'application/pdf', 'Creat	['brain', 'anomaly', 'detection', 'risk', 'ass	["
35	3	James Hyungkwan Kim_KimJ_AI-driven universal participation model for clean electricity markets of the future_P1Q2	Microsoft Word - 211130B	{'Author': 'hyungkwankim', 'Content-Typ	['universal', 'participation', 'clean', 'elect	["
101	3	Harry Fry_FryH_Overcoming Sequence-Structure-Functionality_P1Q2A1.pdf	Overcoming Sequence-Structure-	{'Author': 'Fry, H. Christopher', 'Content-	['sequence', 'structure', 'peptide', 'mater	["
11	4	Eric Dufek_DufekE_Rapid Operational Validation of Advanced Energy Storage Technologies_P1Q2A1.pdf	Rapid Operational Validation of	{'Author': 'Lisa Aldrich', 'Comments': '', 'C	['rapid', 'operational', 'validation', 'advar	["
18	4	Ralph Kube_Collaborative_Big_Machine_learning_P1Q2A1.pdf	Collaborative_Big_Machine_learning	{'Content-Type': 'application/pdf', 'X-Par	['big', 'learn', 'science', 'author', 'learn', 'r	["
22	4	Tianzhen Hong_AI-Enabled Smart Thermostats_P1Q2A1.pdf	AI-Enabled Smart Thermostats for	{'Author': 'Tianzhen-X1C', 'Content-Type	['smart', 'efficiency', 'flexibility', 'buildin	["
44	4	Gavin Liu_Transferable Machine Learning Assisted Risk Management for Subsurface Energy Storage.pdf	NETL â€“ Liu	{'Author': 'Pranjali S. Muley', 'Content-Ty	['liu', ' ', '	["
50	4	Uta Ruett_RuettU_IntegrationOfAutomationAndAIWithHigh-EndInSituCharacterizationToolsForAdaptiveSynthesis_P1	Integration of automation and AI with	{'Author': 'Stone, Kevin Hunter', 'Comme	['automation', 'high', 'end', 'situ', 'charact	["
117	4	Katrina Bennett_Advancing the Use of ML for Improved Understanding of Hydrologic Extremes Under Climate Change.	Bennett_Karra_Vesselinov_Schwenk_N	{'Author': 'kbennett', 'Content-Type': 'ap	['bennett', 'la', 'dec', 'improved', 'unders	["
122	4	Katherine Wilsdon_WilsdonK_AutonomousControlofMicroreactorsthroughDigital_P1Q2A1.pdf	Microsoft Word -	{'Content-Type': 'application/pdf', 'Creat	['autonomous', 'microreactors', 'digital',	["
160	4	Young Soo Park_YoungP_RoboticDigitalTwinForSelfDrivingLaboratory_P1Q2A1.pdf	Microsoft Word -	{'Author': 'yspark', 'Content-Type': 'appli	['robotic', 'digital', 'twin', 'self', 'park', 'an	["
161	4	William Tang_William Tang_AI DOE WHITE PAPER_Fusion Energy Science_Nov.30, 2021_P1Q2A1.pdf	William Tang_AI DOE WHITE	{'Author': 'William M. Tang', 'Content-Ty	['william', 'white', 'white', 'science', 'fusi	["
186	4	Matthew Reno_AI-Based Protective Relays for Electric Grid Resiliency.pdf	Title: AI-Based Protective Relays for	{'Author': 'Blakely, Logan', 'Content-Type	['protective', 'relays', 'electric', 'grid', 'res	["
189	4	Aditya Sundararajan_Adaptive Machine Learning for Resilient Networked Microgrids against Natural Disasters.pdf	Adaptive Machine Learning for	{'Author': 'Sundararajan, Aditya', 'Conter	['adaptive', 'learn', 'resilient', 'networke	["
7	5	Cheng Wang_Physics-Informed Deep Learning for Multiscale Water Cycle Prediction.pdf	Physics-Informed Deep Learning for	{'Author': 'chengw', 'Content-Type': 'app	['inform', 'multiscale', 'water', 'cycle', 'in	["
20	5	Sarp Oral_OralS_SelfLearningAdaptiveExtremeScaleStoragethatBlursComputeandData_P1Q2A1.pdf	Self-Learning, Adaptive, Extreme-Scale	{'Content-Type': 'application/pdf', 'X-Par	['self', 'adaptive', 'extreme', 'storage', 'bl	["
23	5	Brian Nord_Closing the Loop_ Automate the Scientific Cycle AI DOE Workshop Nord_P1Q2A1.pdf	Closing the Loop: Automate the	{'Content-Type': 'application/pdf', 'X-Par	['close', 'loop', 'automate', 'cycle', 'works	["
31	5	Ryan Coffee_Resiliency to Natural and Man Made Disasters.pdf	AI@DOE - Section 3	{'Content-Type': 'application/pdf', 'X-Par	['section', 'section', 'enterprise', 'neuros	["
36	5	Charles Farrar_A Complexity Based Framework for Structural Health Monitoring .pdf	A Complexity-Based Framework for	{'Author': 'DDS User', 'Content-Type': 'ap	['structural', 'health', 'monitoring', 'engir	["
40	5	Anubhav Jain_JainA_AutonomousLabs_P1Q2A1.pdf	Microsoft Word -	{'Content-Type': 'application/pdf', 'Creat	['autonomous', 'language', 'processing', '	["
45	5	Wenting Li_Edge Computing Based Physics-Informed Machine Learning for Power Infrastructure Resilience.pdf	Edge Computing Based Physics-	{'Content-Type': 'application/pdf', 'Creat	['edge', 'inform', 'learn', 'power', 'infrastı	["
48	5	Dipankar Dwivedi_Knowledge-Guided Machine Learning to Mitigate Impacts of Hydrological Extreme Events.pdf	Knowledge-Guided Machine Learning	{'Author': '', 'Content-Type': 'application/	['guided', 'learn', 'mitigate', 'hydrologica	["
53	5	Dmitry Lyakh_Dmitry_QuantumAI_WhitePaper_AI4DOE_P1Q2A1.pdf	Dmitry_QuantumAI_WhitePaper_AI4D	{'Author': 'Beck, Tom', 'Content-Type': 'a	['inverse', 'hybrid', 'quantum', 'classical',	["
55	5	Noah Paulson_PaulsonN_ArtificialInsightInMaterialsManufacturing_P1Q2A1.pdf	Artificial Insight in Materials	{'Author': 'npaulsonwork@gmail.com', 'C	['insight', 'material', 'manufacturing', 'mi	["
68	5	Sumit Bhattacharya_BhattacharyaS_AI-guided inverse materials design for extreme environments_P1Q2A1.pdf	Microsoft Word - AI-guided inverse	{'Content-Type': 'application/pdf', 'Creat	['guide', 'inverse', 'material', 'extreme', '	["

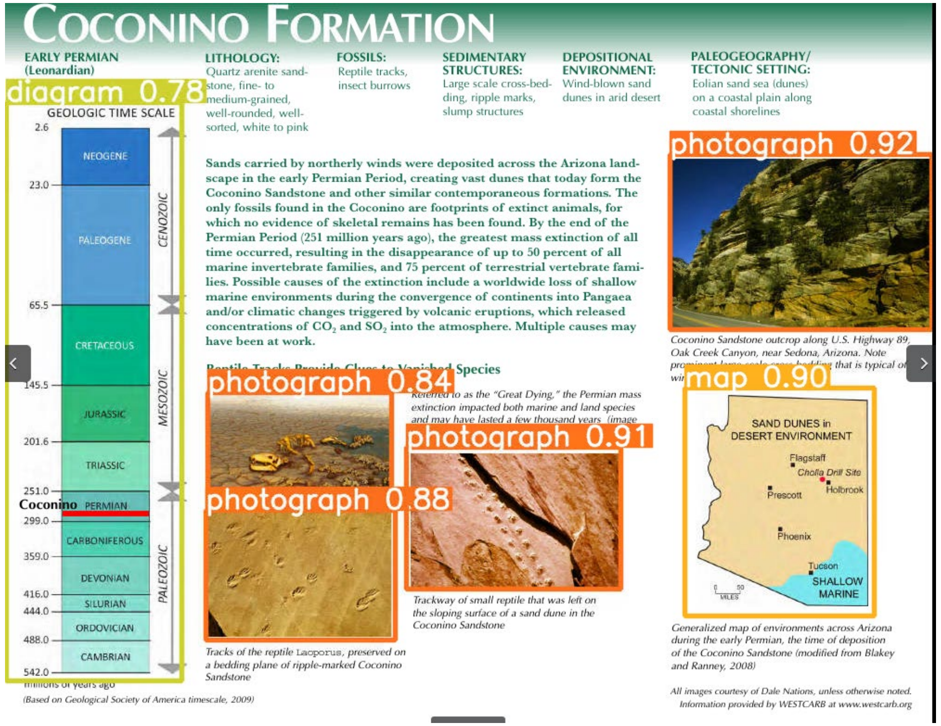
Use Case: Abstract White Paper Data Extraction

Image Classification Demo/Results



```
In [4]: #input data 4-7-22
        source_files = "/home/sabbatim/mlhome/NLP_Documents_4-1-22_From_Kelly/AllDocs/input_images"
        !python detect.py --weights '/home/sabbatim/anaconda3/yolov5/runs/train/exp6/weights/best.pt' --img 640 --conf 0.25
```

```
ingAI2021v4_P1Q2A1_0.png: 640x512 Done. (0.007s)
image 486/492 /home/sabbatim/mlhome/NLP_Documents_4-1-22_From_Kelly/AllDocs/input_images/Zhehui (jeph) Wang_Imag
ingAI2021v4_P1Q2A1_1.png: 640x512 1 map, 2 diagrams, Done. (0.007s)
image 487/492 /home/sabbatim/mlhome/NLP_Documents_4-1-22_From_Kelly/AllDocs/input_images/Zhehui Wang_Physics-Inf
ormed AI (PAI) for Imaging and Intelligent Vision_0.png: 640x512 Done. (0.007s)
image 488/492 /home/sabbatim/mlhome/NLP_Documents_4-1-22_From_Kelly/AllDocs/input_images/Zhehui Wang_Physics-Inf
ormed AI (PAI) for Imaging and Intelligent Vision_1.png: 640x512 1 map, 2 diagrams, Done. (0.007s)
image 489/492 /home/sabbatim/mlhome/NLP_Documents_4-1-22_From_Kelly/AllDocs/input_images/Zhengchun Liu_LiuZ_On t
he Opportunities of Foundation Models for Scientific Image Processing_P1Q2A1_0.png: 640x512 3 diagrams, Done.
(0.007s)
image 490/492 /home/sabbatim/mlhome/NLP_Documents_4-1-22_From_Kelly/AllDocs/input_images/Zhengchun Liu_LiuZ_On t
he Opportunities of Foundation Models for Scientific Image Processing_P1Q2A1_1.png: 640x512 Done. (0.007s)
image 491/492 /home/sabbatim/mlhome/NLP_Documents_4-1-22_From_Kelly/AllDocs/input_images/gaffney_AI@DOE_whitepap
er_0.png: 640x512 Done. (0.007s)
image 492/492 /home/sabbatim/mlhome/NLP_Documents_4-1-22_From_Kelly/AllDocs/input_images/gaffney_AI@DOE_whitepap
er_1.png: 640x512 Done. (0.007s)
Speed: 0.5ms pre-process, 6.7ms inference, 0.5ms NMS per image at shape (1, 3, 640, 640)
Results saved to runs/detect/exp6
203 labels saved to runs/detect/exp6/labels
```



COCONINO FORMATION

EARLY PERMIAN (Leonardian)

LITHOLOGY: Quartz arenite sandstone, fine- to medium-grained, well-rounded, well-sorted, white to pink

FOSSILS: Reptile tracks, insect burrows

SEDIMENTARY STRUCTURES: Large scale cross-bedding, ripple marks, slump structures

DEPOSITIONAL ENVIRONMENT: Wind-blown sand dunes in arid desert

PALEOGEOGRAPHY/ TECTONIC SETTING: Eolian sand sea (dunes) on a coastal plain along coastal shorelines

millions of years ago

(Based on Geological Society of America timescale, 2009)

Sands carried by northerly winds were deposited across the Arizona landscape in the early Permian Period, creating vast dunes that today form the Coconino Sandstone and other similar contemporaneous formations. The only fossils found in the Coconino are footprints of extinct animals, for which no evidence of skeletal remains has been found. By the end of the Permian Period (251 million years ago), the greatest mass extinction of all time occurred, resulting in the disappearance of up to 50 percent of all marine invertebrate families, and 75 percent of terrestrial vertebrate families. Possible causes of the extinction include a worldwide loss of shallow marine environments during the convergence of continents into Pangaea and/or climatic changes triggered by volcanic eruptions, which released concentrations of CO_2 and SO_2 into the atmosphere. Multiple causes may have been at work.

Coconino Sandstone outcrop along U.S. Highway 89, Oak Creek Canyon, near Sedona, Arizona. Note prominent large-scale cross-bedding that is typical of wi

Reptile Tracks Provide Clues to Vanished Species

Referred to as the "Great Dying," the Permian mass extinction impacted both marine and land species and may have lasted a few thousand years (image

Tracks of the reptile Laoporus, preserved on a bedding plane of ripple-marked Coconino Sandstone

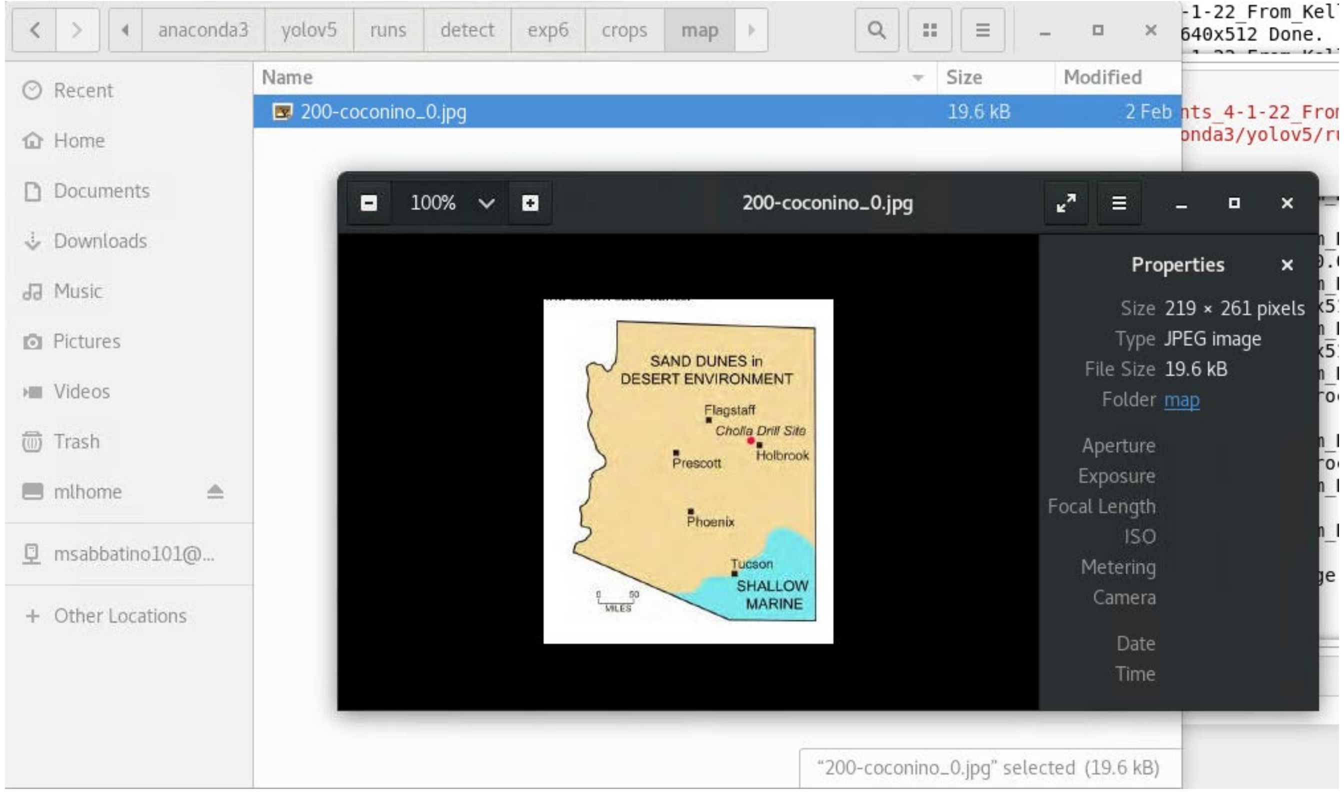
Trackway of small reptile that was left on the sloping surface of a sand dune in the Coconino Sandstone

Generalized map of environments across Arizona during the early Permian, the time of deposition of the Coconino Sandstone (modified from Blakey and Ranney, 2008)

All images courtesy of Dale Nations, unless otherwise noted. Information provided by WESTCARB at www.westcarb.org

Use Case: Abstract White Paper Data Extraction

Image Classification Demo/Results



```
-1-22_From_Kel
640x512 Done.
nts_4-1-22_Fro
onda3/yolov5/ru
```

Name	Size	Modified
200-coconino_0.jpg	19.6 kB	2 Feb

200-coconino_0.jpg

Size 219 × 261 pixels
Type JPEG image
File Size 19.6 kB
Folder map

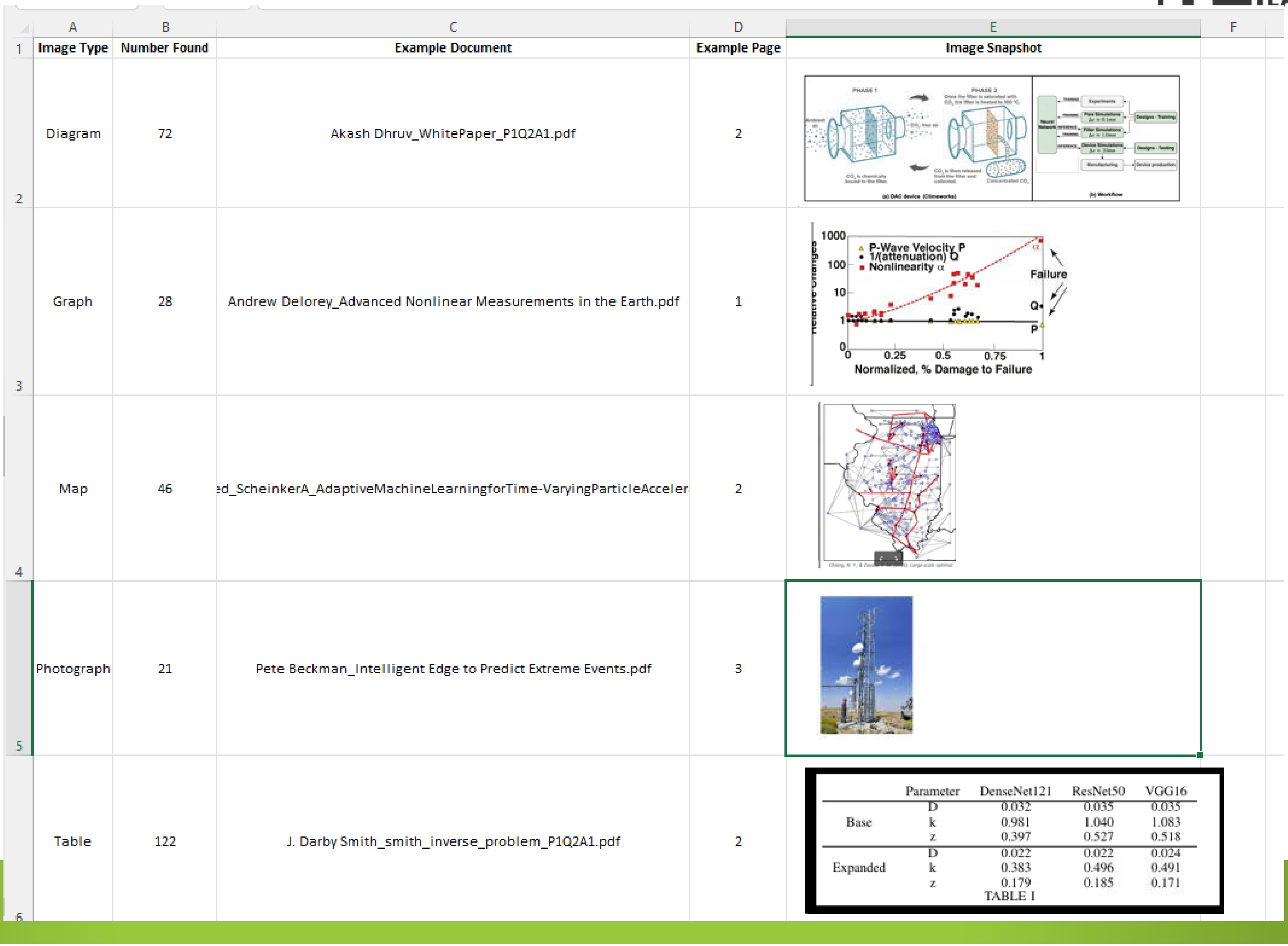
Aperture
Exposure
Focal Length
ISO
Metering
Camera

Date
Time

“200-coconino_0.jpg” selected (19.6 kB)

Use Case: Abstract White Paper Data Extraction

Image Classification Demo/Results



	A	B	C	D	E	F
1	Image Type	Number Found	Example Document	Example Page	Image Snapshot	
2	Diagram	72	Akash Dhruv_WhitePaper_P1Q2A1.pdf	2		
3	Graph	28	Andrew Delorey_Advanced Nonlinear Measurements in the Earth.pdf	1		
4	Map	46	ed_ScheinkerA_AdaptiveMachineLearningforTime-VaryingParticleAcceler	2		
5	Photograph	21	Pete Beckman_Intelligent Edge to Predict Extreme Events.pdf	3		
6	Table	122	J. Darby Smith_smith_inverse_problem_P1Q2A1.pdf	2		

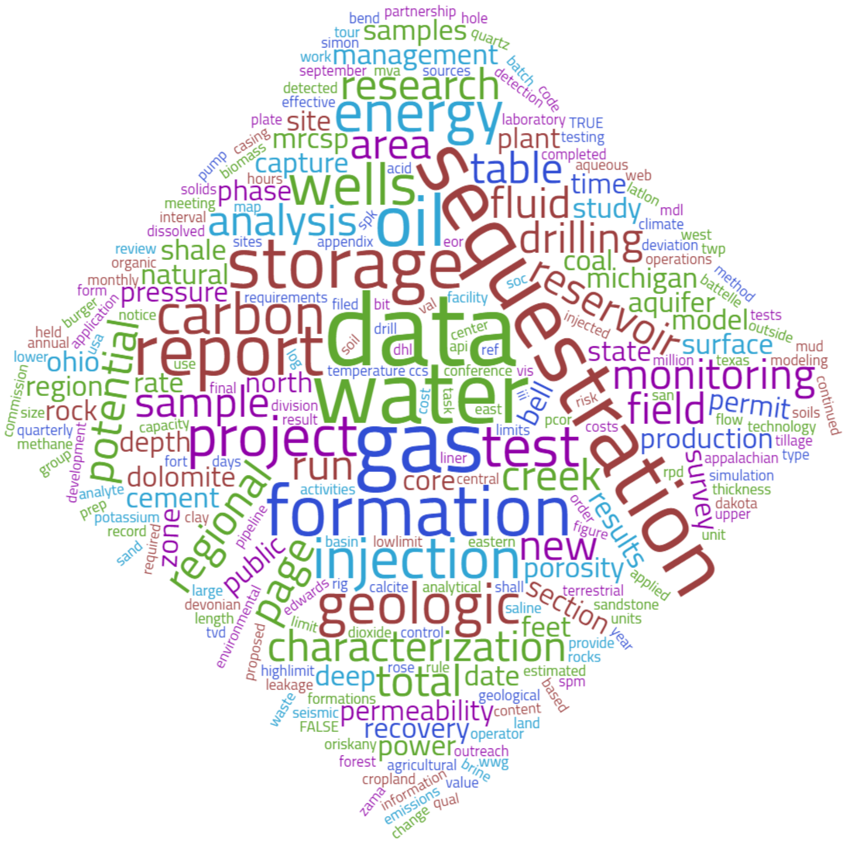
	Parameter	DenseNet121	ResNet50	VGG16
Base	D	0.032	0.035	0.035
	k	0.981	1.040	1.083
	z	0.397	0.527	0.518
Expanded	D	0.022	0.022	0.024
	k	0.383	0.496	0.491
	z	0.179	0.185	0.171

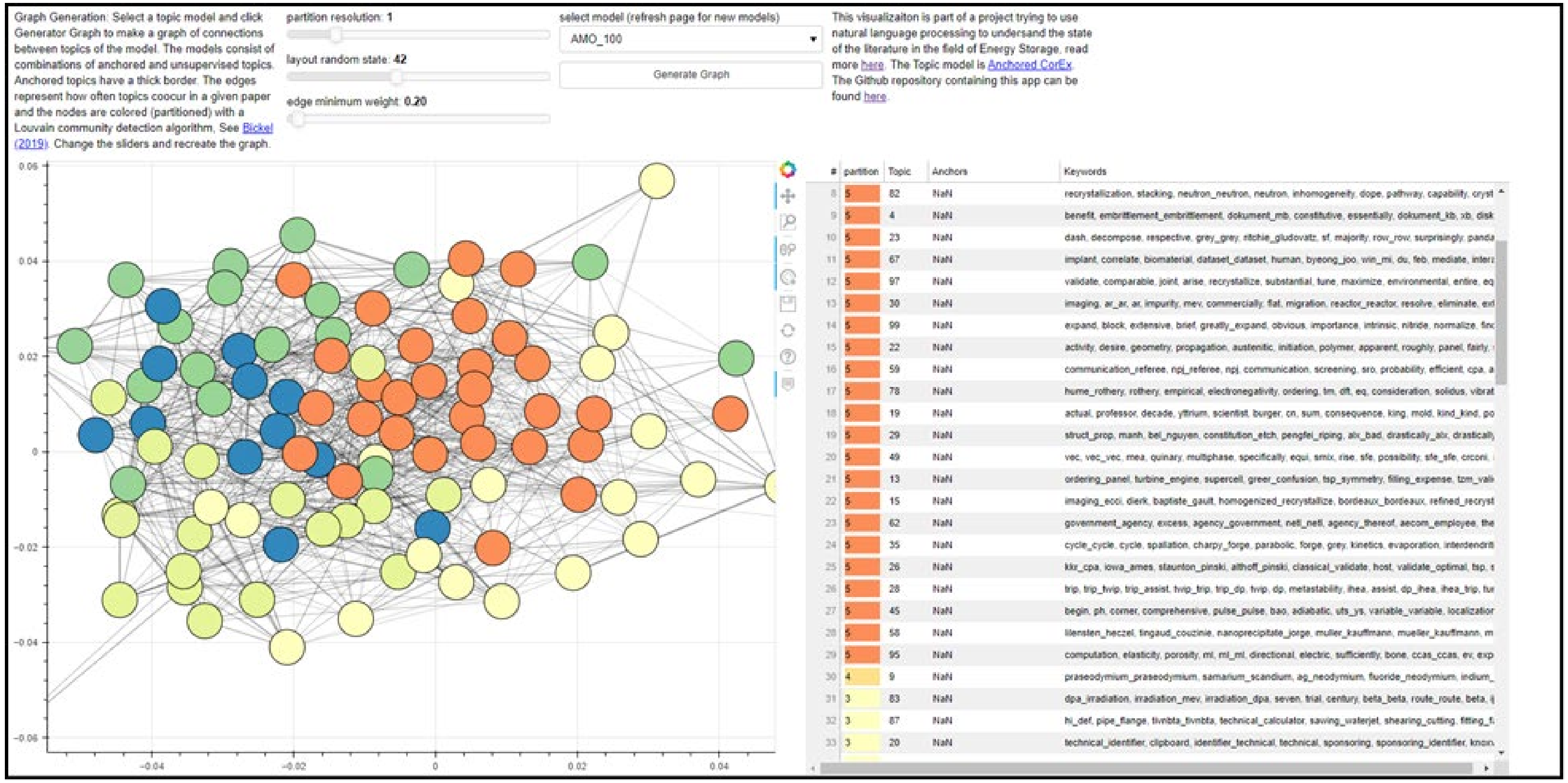
TABLE I

Synergy Opportunities

Collaborative cross-project technology

- Use material same NLP tech
- Using other NLP Models Louvian Community Detection





Graph Generation: Select a topic model and click Generator Graph to make a graph of connections between topics of the model. The models consist of combinations of anchored and unsupervised topics. Anchored topics have a thick border. The edges represent how often topics coocur in a given paper and the nodes are colored (partitioned) with a Louvain community detection algorithm, See Bickel (2019). Change the sliders and recreate the graph.

partition resolution: 1

layout random state: 42

edge minimum weight: 0.20

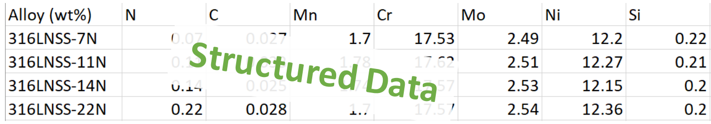
select model (refresh page for new models): AMO_100

This visualizaiton is part of a project trying to use natural language processing to undersand the state of the literature in the field of Energy Storage, read more here. The Topic model is Anchored CorEx. The Github repository containing this app can be found here.

#	partition	Topic	Anchors	Keywords
8	5	82	NaN	recrystallization, stacking, neutron_neutron, neutron, inhomogeneity, dope, pathway, capability, cryst
9	5	4	NaN	benefit, embrittlement_embrittlement, dokument_mb, constitutive, essentially, dokument_kb, xb, disk
10	5	23	NaN	dash, decompose, respective, grey_grey, ritchie_gludovatz, sf, majority, row_row, surprisingly, panda
11	5	67	NaN	implant, correlate, biomaterial, dataset_dataset, human, byeong_joo, win_mi, du, feb, mediate, inter
12	5	97	NaN	validate, comparable, joint, arise, recrystallize, substantial, tune, maximize, environmental, entire, eq
13	5	30	NaN	imaging, ar_ar, ar, impurity, mev, commercially, flat, migration, reactor_reactor, resolve, eliminate, ext
14	5	99	NaN	expand, block, extensive, brief, greatly_expand, obvious, importance, intrinsic, nitride, normalize, find
15	5	22	NaN	activity, desire, geometry, propagation, austenitic, initiation, polymer, apparent, roughly, panel, fairly,
16	5	59	NaN	communication_referee, npj_referee, npj, communication, screening, sro, probability, efficient, cpa, a
17	5	78	NaN	hume_rothery, rothery, empirical, electronegativity, ordering, tm, dft, eq, consideration, solidus, vibrat
18	5	19	NaN	actual, professor, decade, yttrium, scientist, burger, cn, sum, consequence, king, mold, kind_kind, po
19	5	29	NaN	struct_prop, manh, bei_nguyen, constitution_etch, pengfei_riping, alx_bad, drastically_alx, drastically
20	5	49	NaN	vec, vec_vec, mea, quinary, multiphase, specifically, equi, smix, rise, sfe, possibility, sfe_sfe, crconi,
21	5	13	NaN	ordering_panel, turbine_engine, supercell, greer_confusion, tsp_symmetry, filling_expense, tzm_vali
22	5	15	NaN	imaging_ecci, dierk, baptiste_gault, homogenized_recrystallize, bordeaux_bordeaux, refined_recryst
23	5	62	NaN	government_agency, excess, agency_government, netl_netl, agency_thereof, aecom_employee, the
24	5	35	NaN	cycle_cycle, cycle, spallation, charpy_forge, parabolic, forge, grey, kinetics, evaporation, interdendriti
25	5	26	NaN	kkr_cpa, iowa_ames, staunton_pinski, althoff_pinski, classical_validate, host, validate_optimal, tsp, s
26	5	28	NaN	trip, trip_twip, trip_assist, twip_trip, trip_dp, twip, dp, metastability, ihea, assist, dp_ihea, ihea_trip, tw
27	5	45	NaN	begin, ph, corner, comprehensive, pulse_pulse, bao, adiabatic, uts_ys, variable_variable, localization
28	5	58	NaN	lilenstein_heczel, tingaud_couzinie, nanoprecipitate_jorge, muller_kauffmann, mueller_kauffmann, m
29	5	95	NaN	computation, elasticity, porosity, ml, ml_ml, directional, electric, sufficiently, bone, ccas_ccas, ev, exp
30	4	9	NaN	praseodymium_praseodymium, samarium_scandium, ag_neodymium, fluoride_neodymium, indium_
31	3	83	NaN	dpa_irradiation, irradiation_mev, irradiation_dpa, seven, trial, century, beta_beta, route_route, beta, i
32	3	87	NaN	hi_def, pipe_flange, tivnbta_tivnbta, technical_calculator, saving_waterjet, shearing_cutting, fitting_f
33	3	20	NaN	technical_identifier, clipboard, identifier_technical, technical, sponsoring, sponsoring_identifier, know

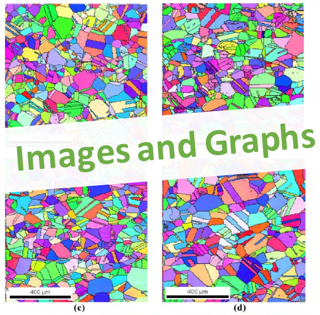
Supporting Data Collection, Curation & Analysis in Other Areas

Data mining, including…



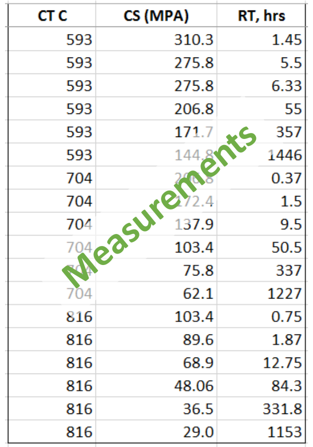
Alloy (wt%)	N	C	Mn	Cr	Mo	Ni	Si
316LNSS-7N	0.07	0.027	1.7	17.53	2.49	12.2	0.22
316LNSS-11N	0.1	[illegible]	1.78	17.62	2.51	12.27	0.21
316LNSS-14N	0.14	0.025	1.74	17.57	2.53	12.15	0.2
316LNSS-22N	0.22	0.028	1.7	17.57	2.54	12.36	0.2

Structured Data



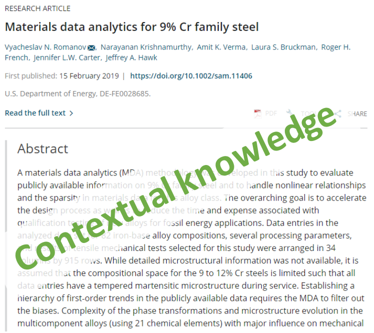
Images and Graphs

Fig. 1. Orientation imaging micrographs of solution annealed 316LN SS containing nitrogen (wt.%) of (a) 0.07, (b) 0.11, (c) 0.14 and (d) 0.22 N. Nearly equiaxed grains and annealing twins have been observed.



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	206.8	0.37
704	172.4	1.5
704	137.9	9.5
704	103.4	50.5
704	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

Measurements



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.

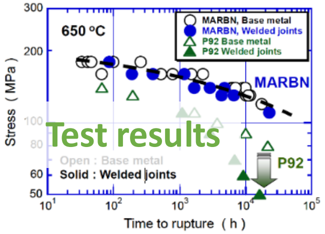
Read the full text >

Abstract

A materials data analytics (MDA) methodology was developed in this study to evaluate publicly available information on 9% Cr family steel and to handle nonlinear relationships and the sparsity in materials data for this alloy class. The overarching goal is to accelerate the design process as well as to reduce the time and expense associated with qualification testing of new alloys for fossil energy applications. Data entries in the analyzed data set for 82 iron-base alloy compositions, several processing parameters, and results of tensile mechanical tests selected for this study were arranged in 34 columns by 915 rows. While detailed microstructural information was not available, it is assumed that the compositional space for the 9 to 12% Cr steels is limited such that all data 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

MARBN : 9Cr-3W-3Co-VNb, 120 - 150 ppm B & 60 - 90 ppm N
P92 : 9Cr-0.5Mo-1.8W-VNb, 20 ppm B & 500 ppm N

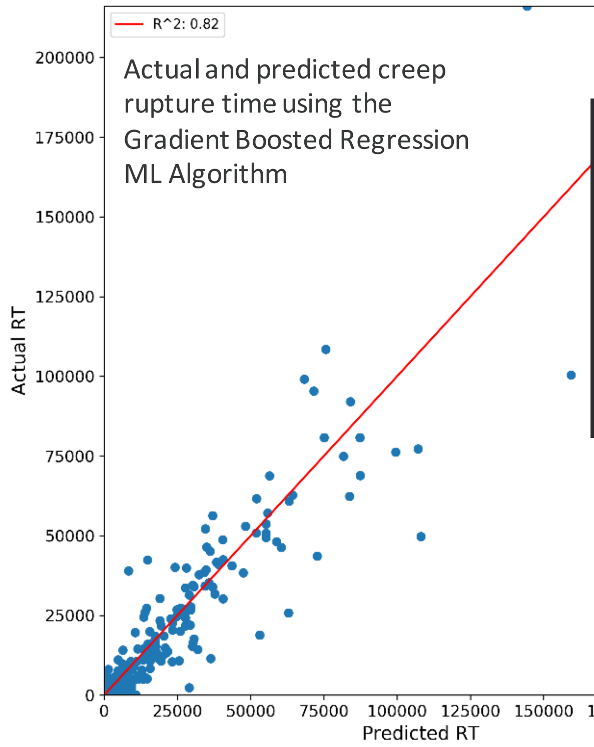


Test results

Move & Convert…

EDX™
Energy Data eXchange

…& use in predictive analytics for alloy behavior



Actual and predicted creep rupture time using the Gradient Boosted Regression ML Algorithm

Evaluating machine learning models to:

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

Lessons Learned

Machine Learning, Artificial Intelligence, and Natural Language Processing are Difficult

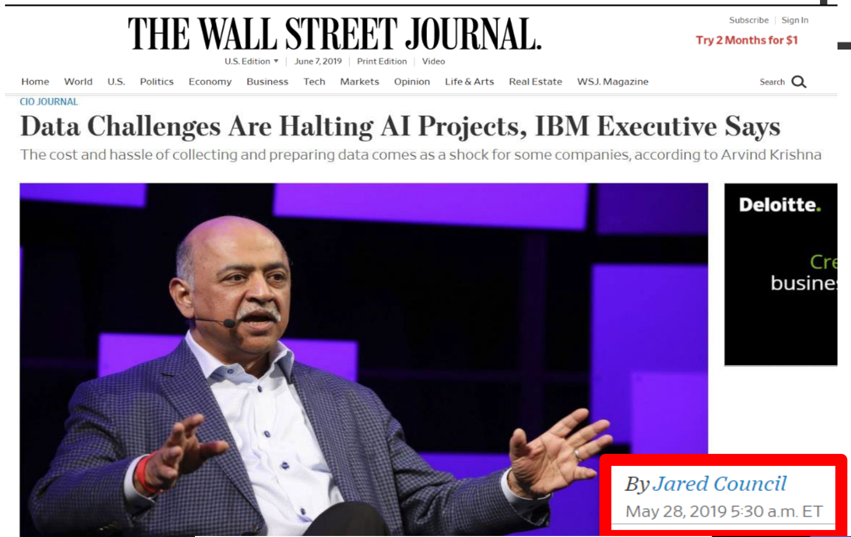
- Whatever happened to Watson?

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 the Research Network
- The database would improve if deployed on a cloud service or other shared environment



Subscribe | Sign In

Try 2 Months for $1

THE WALL STREET JOURNAL.

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

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

CIO JOURNAL

Data Challenges Are Halting AI Projects, IBM Executive Says

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

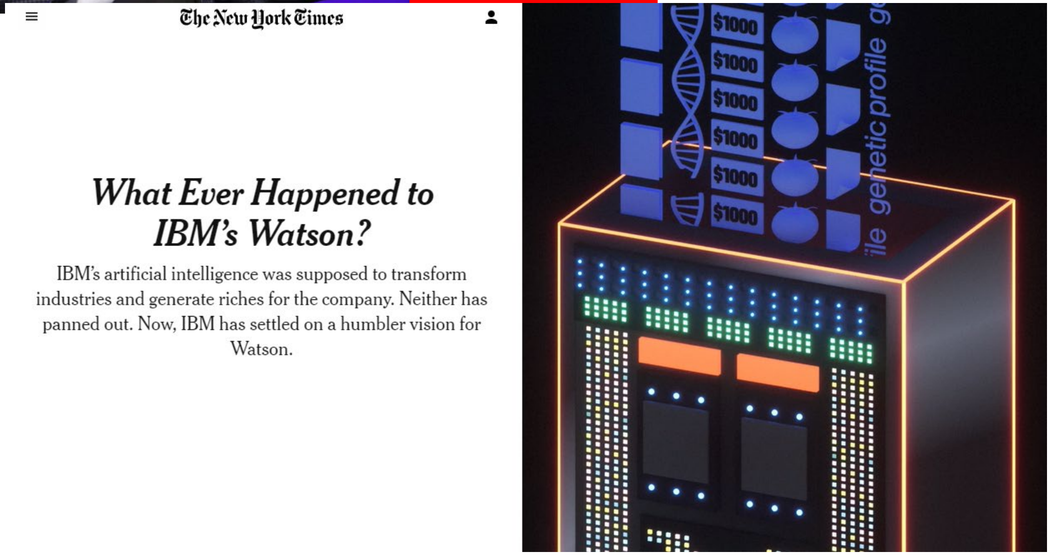
Deloitte.

Cre

busine

By Jared Council

May 28, 2019 5:30 a.m. ET

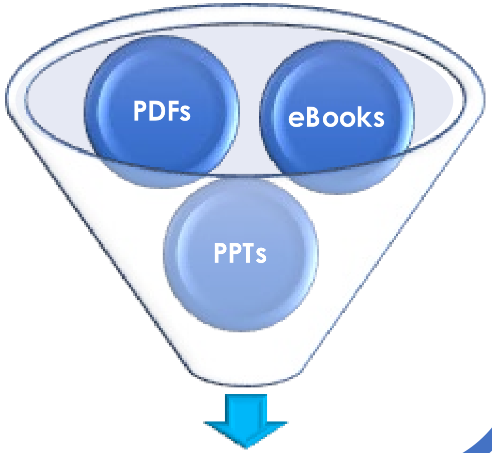


The New York Times

What Ever Happened to IBM's Watson?

IBM's artificial intelligence was supposed to transform industries and generate riches for the company. Neither has panned out. Now, IBM has settled on a humbler vision for Watson.

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



Structured data

Summary

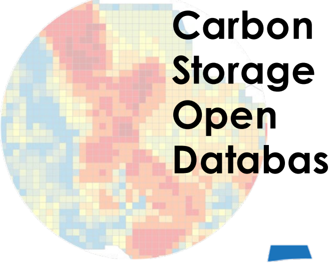
FE and Carbon Storage program investments into data curation and management have led to the development of AI/ML tools and the preservation of millions of dollars of research products which benefits ongoing and future research. This has led to:

- 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

What's next: EDX4CCS

NETL Research · Carbon Storage Program ·

NatCarb
A National Look at Carbon Sequestration



Carbon Storage Open Database

SMART

NRAP
National Risk Assessment Partnership

NETL SmartSearch

NETL SmartParse

GEO CUBE

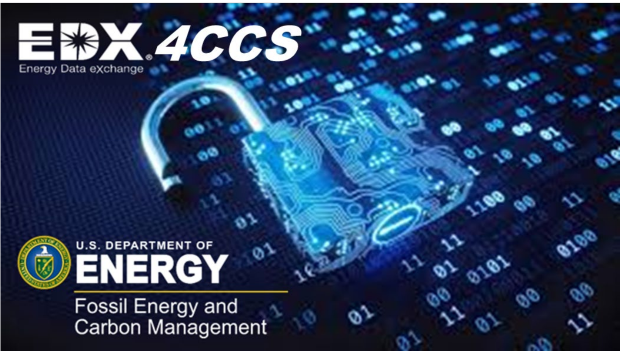
EDX®

EDX4CCS

Data, Integration, generation, and deployment to feed SMART, NRAP, and regulatory models

Tools, Develop and/or integrate the deployment of tools for data interaction and visualization, decision-support such as for pipelines, regulatory permitting, resource characterization, data visualization, and more

Core CCS EDX DisCO2ver platform, Broader community virtualized data computing platform, and central EDX CCS data and tool hub


EDX 4CCS
Energy Data eXchange
U.S. DEPARTMENT OF ENERGY
Fossil Energy and Carbon Management

Thank you!

NETL Resources

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

@NETL_DOE

@NETL_DOE

@NationalEnergyTechnologyLaboratory

Appendix

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

Benefit to the Program

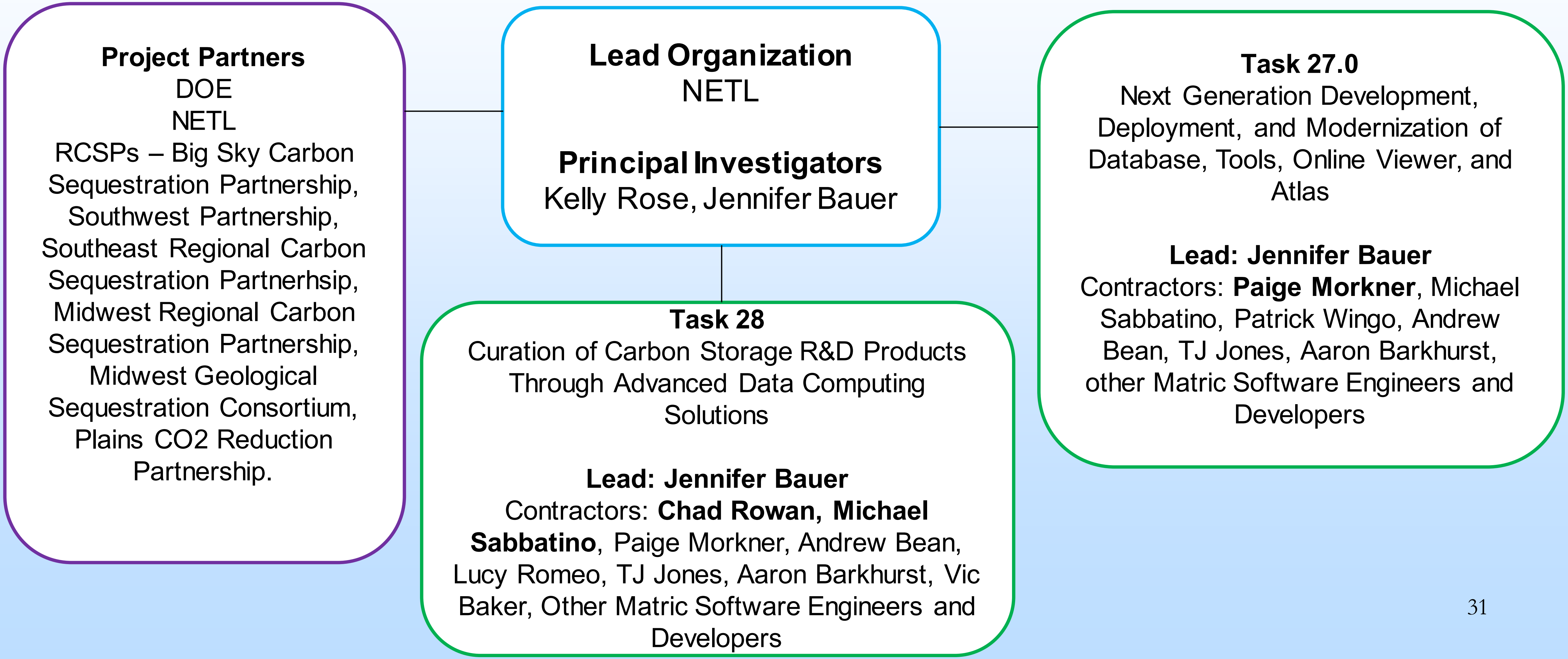
- Task 27 supports the development of data, materials, maps, analyses, and figures for the Carbon Storage Atlas, Natcarb Viewer, and Natcarb database. This includes the release of new data insights to the GCS community, through the sixth edition of the Carbon Storage Atlas, and through bi-annual updates 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 of 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

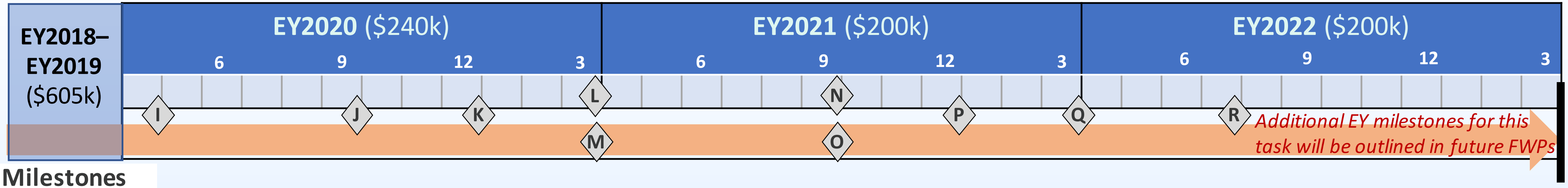
Lead: Jennifer Bauer
Contractors: **Paige Morkner**, Michael Sabbatino, Patrick Wingo, Andrew Bean, TJ Jones, Aaron Barkhurst, other Matric Software Engineers and Developers

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.