SAMi SCIENCE-BASED AI/ML INSTITUTE

Cloud, Containers, and Discovery

GRID June 1, 2022

Vic Baker, Kelly Rose, Joe Obradovich, Dan McFarland

```
    plt.
Run Cell | Run Abo
    #%%# Split-out va
    array = dataset.value
    X = array[:,0:4]
    Y = array[:,4]
    validation_size = 0.20
    seed = 7
    X_train, X_validation, Y_train, Y_validati
```

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

Multi-Cloud (...really means Multi-Environment)

- ‘Cloud’ is where you’re ***not*** at
  - Cloud Service Providers
  - On-Prem Compute
  - Edge computing (sensors, cell phone, etc)
- Leverage APIs for service-to-service communications
- Networking:
  - peered networks
  - commodity networks
  - dedicated / shared interconnects
  - consider egress fees (leverage compute where data lives)
- Utilize cloud-native concepts

YOU ARE HERE
aws
Cloud Service Providers
On-Prem
EDX
Energy Data eXchange
Remote Services

NATIONAL ENERGY TECHNOLOGY LABORATORY

MATRIC

SCIENCE-BASED AI/ML INSTITUTE

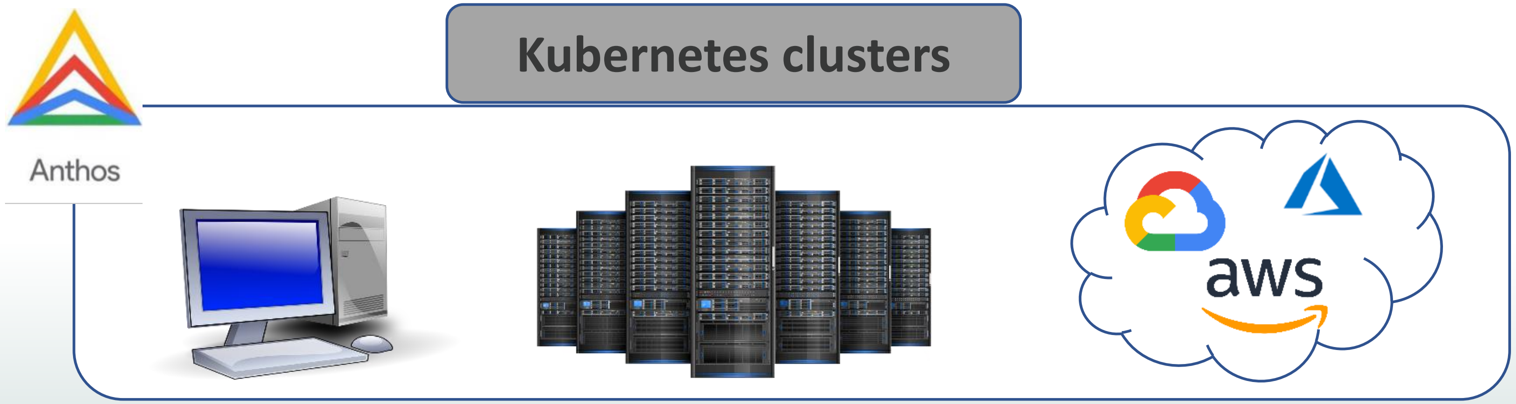
https://edx.netl.doe.gov/sami/

Navigating on-prem to cloud (and local dev)

- One of the biggest challenges: migrating from an “on-prem” development / deployment environment to cloud – HOW???
- Utilize “cloud native” concepts
  - Containerization
  - Microservices
  - Scalability
- Leverage enabling tools:
  - Deployment:
    - Kubernetes
    - Helm
    - Terraform
  - Development:
    - VS Code
    - Containerized environments

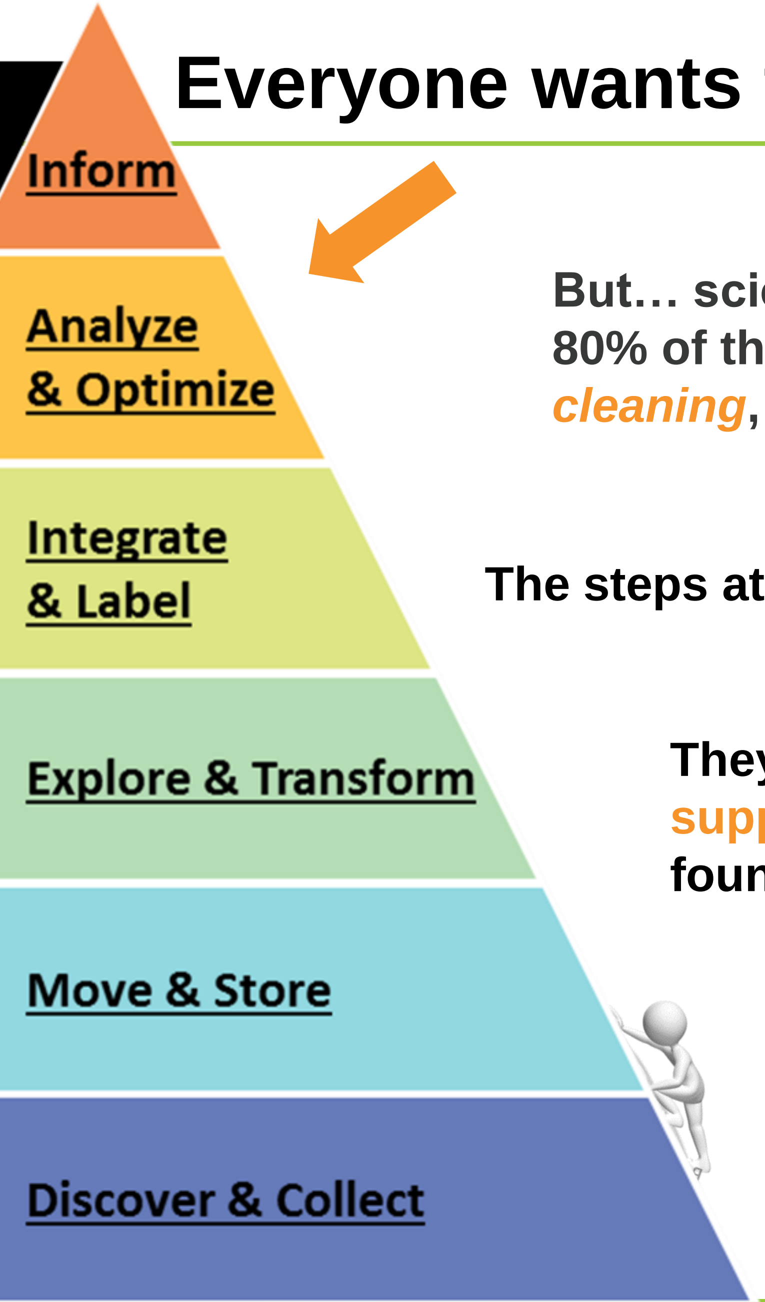
Kubernetes

- Supported by all major cloud providers
- Configurable autoscaling based on load thresholds (CPU, RAM usage)
    - Services
    - Nodes
- Managed service deployments via Helm
- Seamless local, on-prem, cloud deployments via kubectl
- Manage multi-environment clusters via GCP Anthos



https://edx.netl.doe.gov/sami/

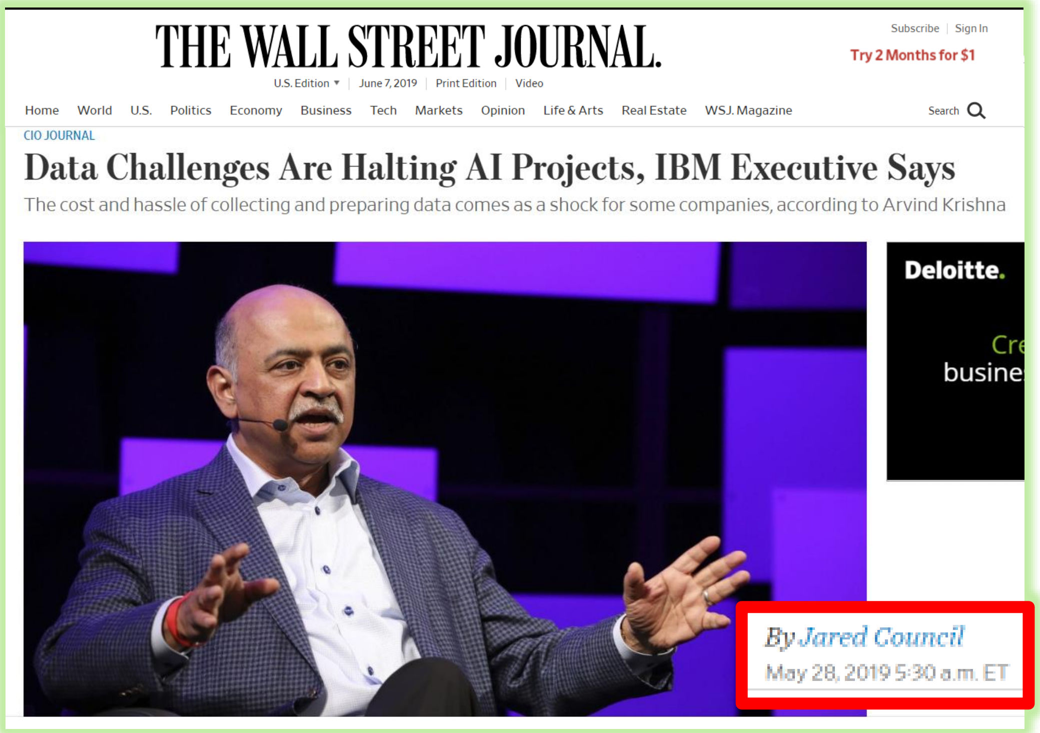
Everyone wants to jump to the top of the pyramid…



But… scientists still spend nearly 80% of their time *acquiring*, *cleaning*, and *organizing* data

The steps at the *bottom* must come first

They require leadership & support to ensure a solid data foundation for future R&D



THE WALL STREET JOURNAL.

Subscribe | Sign In

Try 2 Months for $1

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

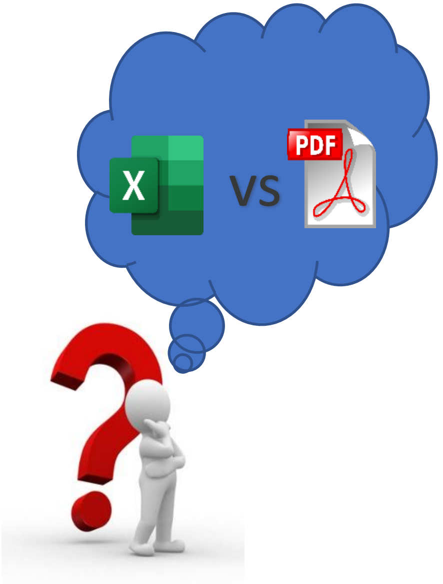
"Invest 5% of research funds in ensuring data are reusable. Funders hold the stick: they should disburse no further funding without a data stewardship plan."
- *Nature*, February 2020

nature
International weekly journal of science

Data are the energy for AI and analysis

SmartSearch©: Conquering the Data Avalanche

- **How do you currently search?**
  - Type in a few keywords
  - Skim the top few results
  - Type in more keywords and try again



- **How do you find and connect to something relevant?**
  - Open a file / web page
  - Read it (skim it)
  - Decide if it’s relevant

What is SmartSearch©?

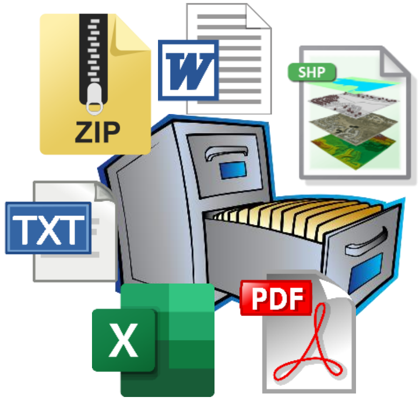
NATIONAL ENERGY TECHNOLOGY LABORATORY

SAMi SCIENCE-BASED AI/ML INSTITUTE

Problem:

You like these files.

You want to find more data relevant to the content of these files



Solution:

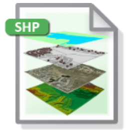
NETL SmartSearch©

SmartSearch automates data discovery by …

1) **Analyzing** content you like

2) **Finding** new content via www, local, enterprise data stores

3) Telling you **how relevant** the discovered data is to what you like



Input

Discovered

Benefits of SmartSearch©

- **Infinitely Scalable (Automated) Data Discovery**
  - Analyze millions+ of files and generate comparison metrics
  - Generate topic models, categorization, recommendations
  - Desktop, cluster, cloud

- **Treat geospatial data like a document**
  - Automatically extract text from geospatial data (shapefiles, geodatabases)
  - Compare textual vs geospatial data to identify relevancy
- **Search for meta tags within HTML body of discovered web sites**
  - i.e., find map tags
- **Analyze archive files – even archives within archives (zips within zips, etc)**
  - Process every file – docs, spatial, etc

AI/ML in SmartSearch©

NETL SmartSearch

- SmartSearch built via cloud native design principles
- Combines 'cluster of clusters', Spark, Kubernetes, and data lakes for massively scalable compute infrastructure
- Natural Language Processing via SparkNLP
    - Distributed NLP processing via the Spark framework
    - Implemented via SparkML Pipelines
    - Provides thousands of pretrained models and pipelines (Glove, Bert, Onto, etc)
    - Custom models can be implemented and trained within same distributed framework
- Machine Learning via SparkML
    - Distributed ML processing via the Spark framework
    - SmartSearch Recommendation Engine
    - LDA Topic Modeling
    - Named Entity Recognition (NER)
    - Question Answering
    - Summarization

SmartSearch© In-Use

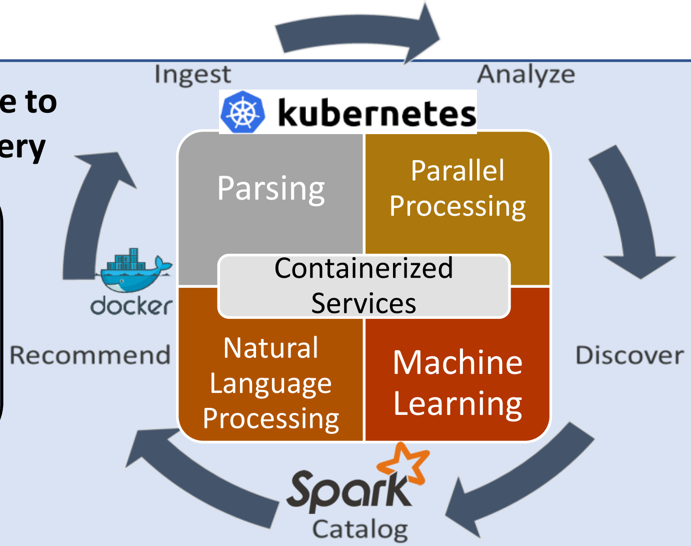
NATIONAL ENERGY TECHNOLOGY LABORATORY

AI informed approach

Challenge: data infrastructure to AI/ML enhanced data discovery

Employing AI/ML tools to find open resources

NETL SmartSearch



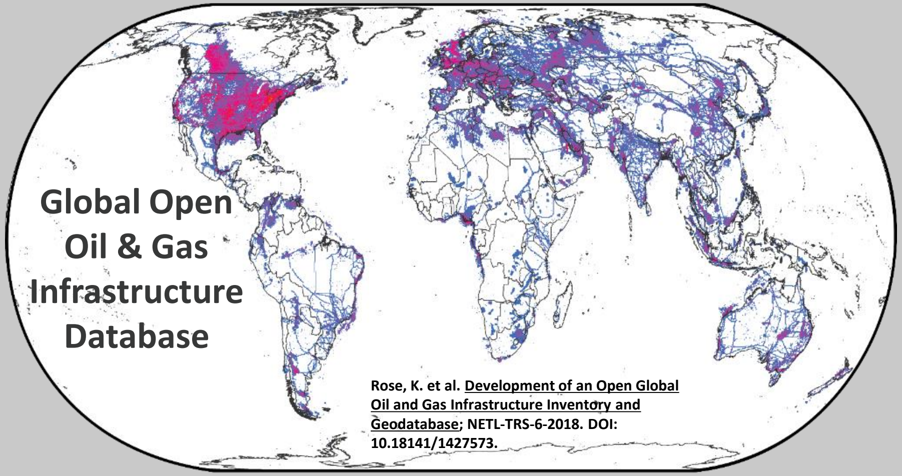
SmartSearch leverages ML+NLP to:

1) **Analyzing** content you like
2) **Finding** new content via www, local, enterprise data stores
3) Telling you **how relevant** the new data is to what you like

Opportunity:

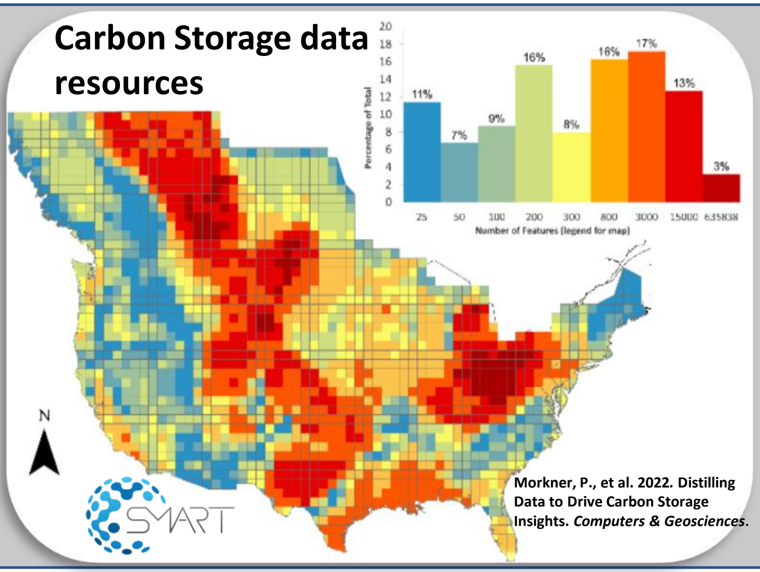
Infinitely scalable to return text, graphical, tabular, image, html, spatial, etc result

Example applications to date



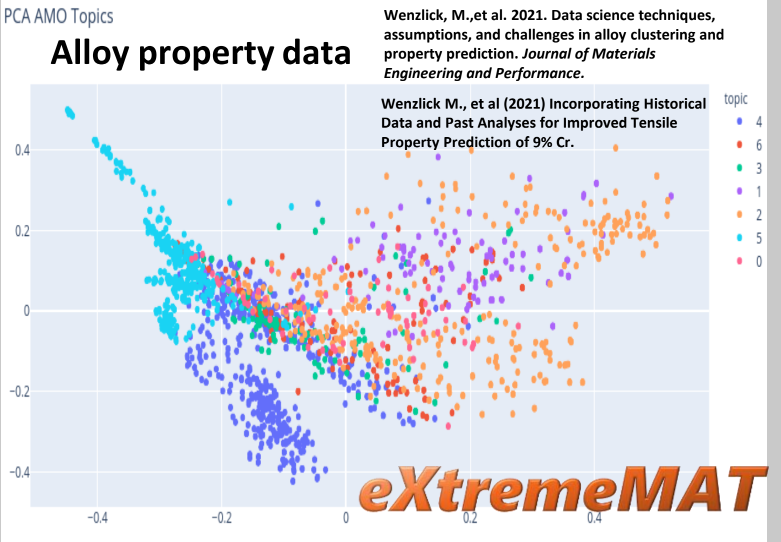
Global Open Oil & Gas Infrastructure Database

Rose, K. et al. Development of an Open Global Oil and Gas Infrastructure Inventory and Geodatabase; NETL-TRS-6-2018. DOI: 10.18141/1427573.



Carbon Storage data resources

Morkner, P., et al. 2022. Distilling Data to Drive Carbon Storage Insights. *Computers & Geosciences*.



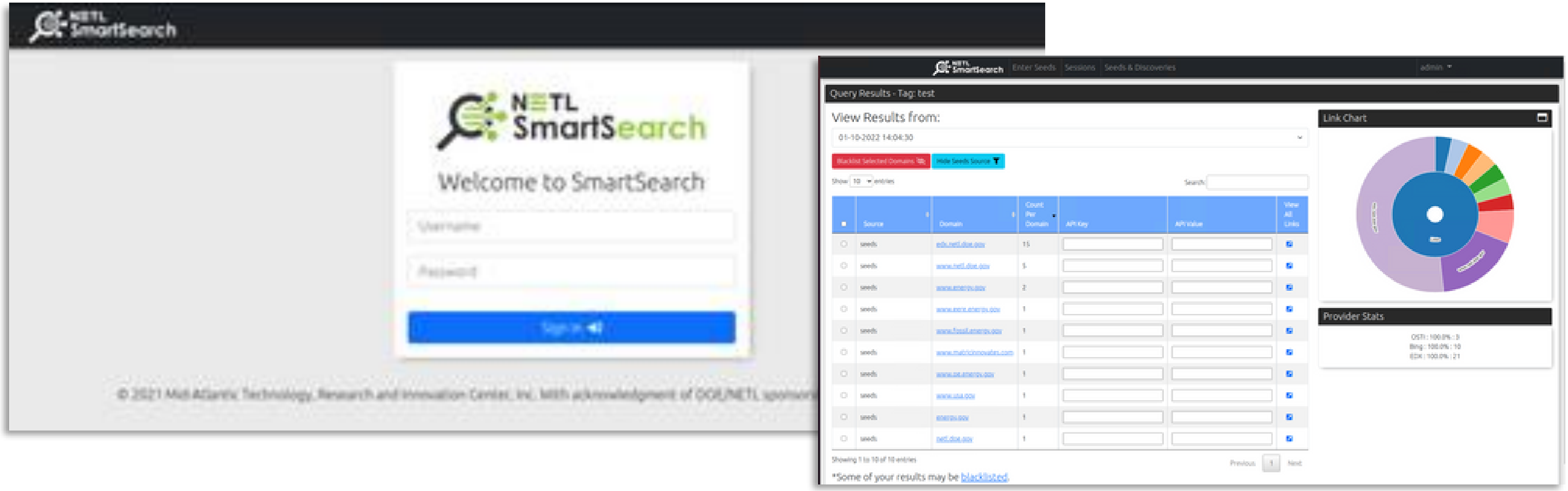
Alloy property data

Wenzlick, M.,et al. 2021. Data science techniques, assumptions, and challenges in alloy clustering and property prediction. *Journal of Materials Engineering and Performance.*

Wenzlick M., et al (2021) Incorporating Historical Data and Past Analyses for Improved Tensile Property Prediction of 9% Cr.

Summary

- SmartSearch© automates the data discovery process by using a scalable compute environment coupled with NLP and ML
- Will be integrated within EDX
- SmartSearch supports ongoing research projects
- SmartSearch used to evaluate cloud providers (GCP, AWS, Azure)



© 2021 Mid-Atlantic Technology, Research and Innovation Center, Inc. With acknowledgment of DOE/NETL sponsor

Query Results - Tag: test

View Results from: 01-10-2022 14:04:30

Show 10 entries   Search:

	Source	Domain	Count Per Domain	API Key	API Value	View All Links
○	seeds	edx.netl.doe.gov	15			
○	seeds	www.netl.doe.gov	5			
○	seeds	www.energy.gov	2			
○	seeds	www.eere.energy.gov	1			
○	seeds	www.fossil.energy.gov	1			
○	seeds	www.matricinnovates.com	1			
○	seeds	www.ne.energy.gov	1			
○	seeds	www.usa.gov	1			
○	seeds	energy.gov	1			
○	seeds	netl.doe.gov	1			

Showing 1 to 10 of 10 entries

*Some of your results may be blacklisted.

OSTI : 100.0% : 3
Bing : 100.0% : 10
EDX : 100.0% : 21

https://edx.netl.doe.gov/sami/

Thank you!

References:

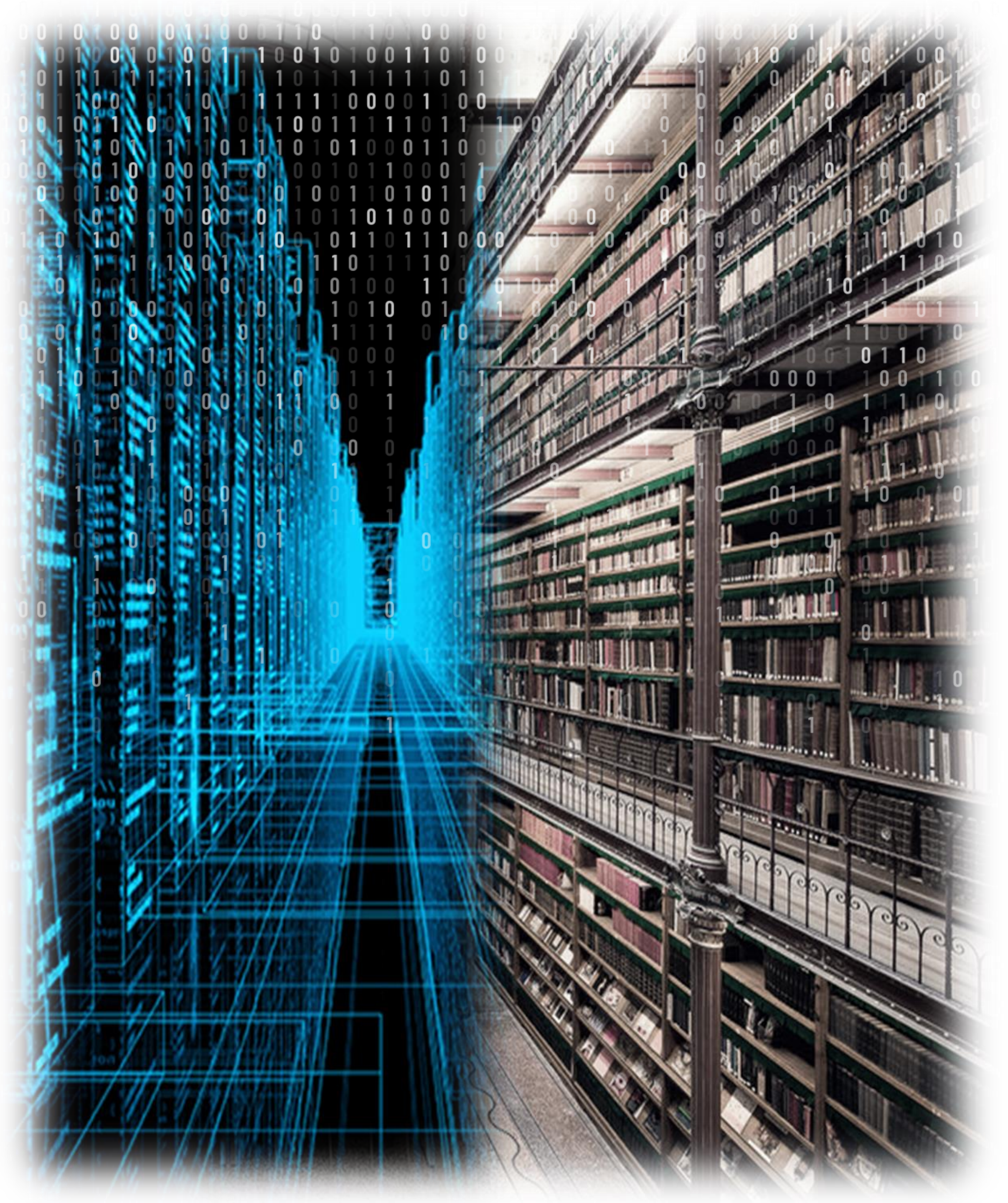
- Morkner, P., et al. 2022. Distilling Data to Drive Carbon Storage Insights. *Computers & Geosciences*.
- Rose, K. et al. Development of an Open Global Oil and Gas Infrastructure Inventory and Geodatabase; NETL-TRS-6-2018. DOI: 10.18141/1427573
- Wenzlick, M.,et al. 2021. Data science techniques, assumptions, and challenges in alloy clustering and property prediction. *Journal of Materials Engineering and Performance.*
- Wenzlick M., et al (2021) Incorporating Historical Data and Past Analyses for Improved Tensile Property Prediction of 9% Cr.

CONTACT:

Vic Baker

vic.baker@netl.doe.gov | vic.baker@matricinnovates.com

SAMI – SAMI (doe.gov)



SAMi
SCIENCE-BASED
AI/ML INSTITUTE

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 authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.*

NETL
NATIONAL
ENERGY
TECHNOLOGY
LABORATORY

Mid-Atlantic Technology, Research & Innovation Center
MATRIC ®

"Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Networking and Information Technology Research and Development Program."

The Networking and Information Technology Research and Development (NITRD) Program

Mailing Address: NCO/NITRD, 2415 Eisenhower Avenue, Alexandria, VA 22314

Physical Address: 490 L'Enfant Plaza SW, Suite 8001, Washington, DC 20024, USA Tel: 202-459-9674, Fax: 202-459-9673, Email: nco@nitrd.gov, Website: https://www.nitrd.gov