SAMi SCIENCE-BASED AI/ML INSTITUTE

SmartSearch, scalable data search and aggregation in the Cloud for CCS and beyond

Vic Baker, National Energy Technology Laboratory
Kelly Rose, National Energy Technology Laboratory
Paige Morkner, National Energy Technology Laboratory

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

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

Multi-Environment

"Cloud" is compute resources where you're *not*...

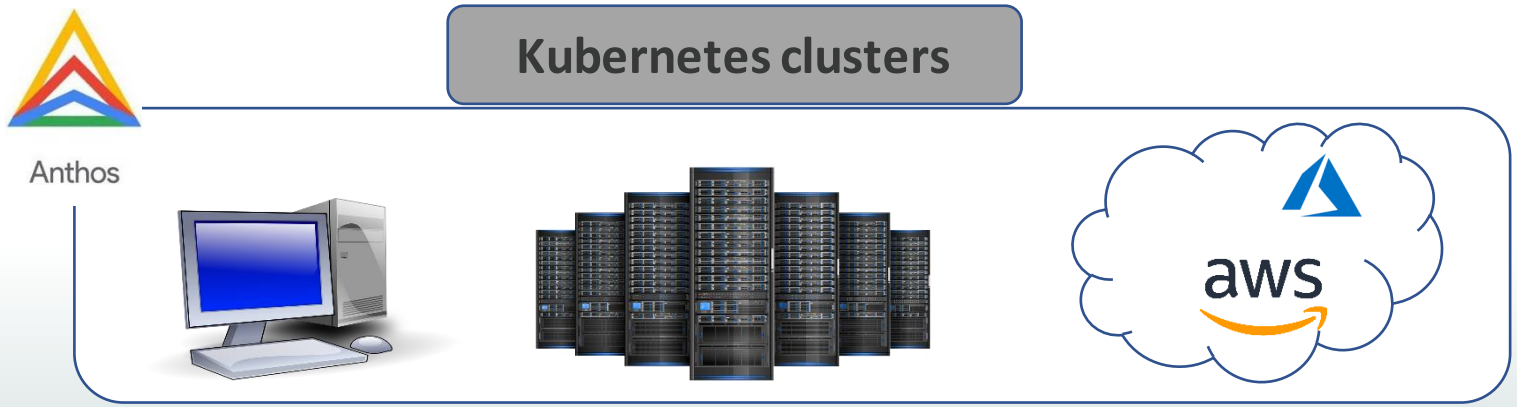
Navigating on-prem to cloud (and local dev)

- 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

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

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
- Kubeflow



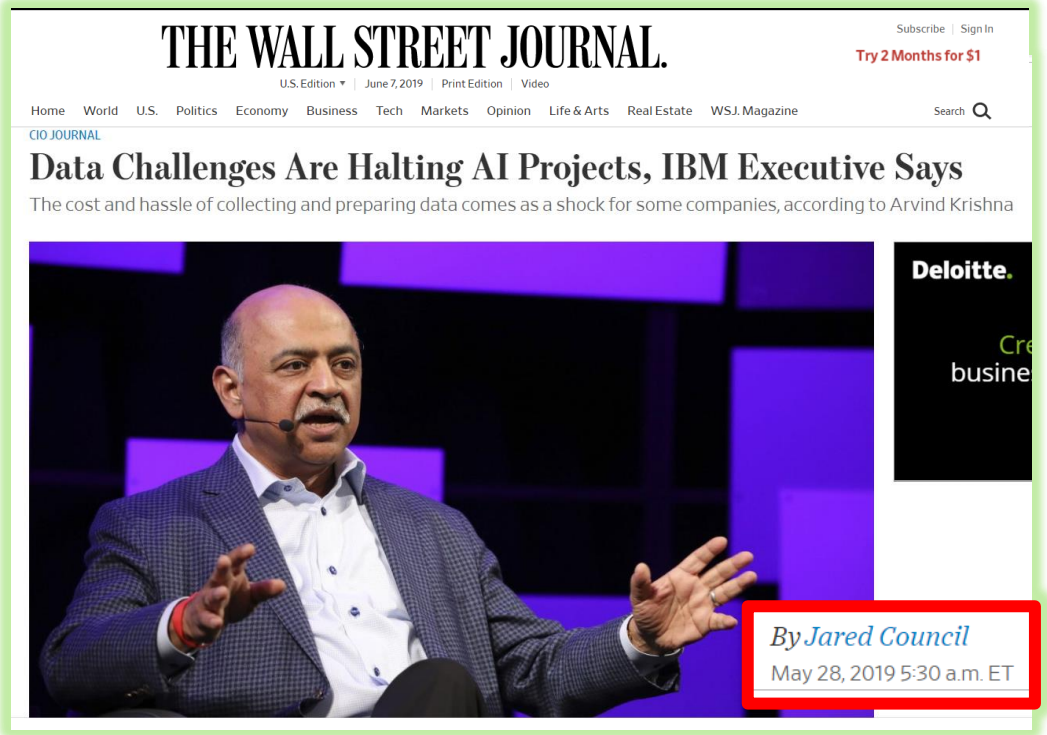
https://edx.netl.doe.gov/sami/

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.

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

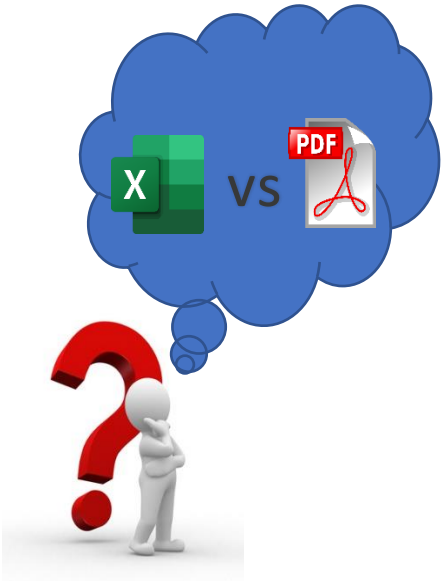
International weekly journal of science

Data are the energy for AI and analysis

SmartSearch©: Conquering the Data Avalanche

NATIONAL ENERGY TECHNOLOGY LABORATORY | SAMi SCIENCE-BASED AI/ML INSTITUTE

- **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©?

Problem:

You like these files.

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

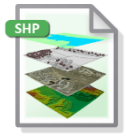
Solution:

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



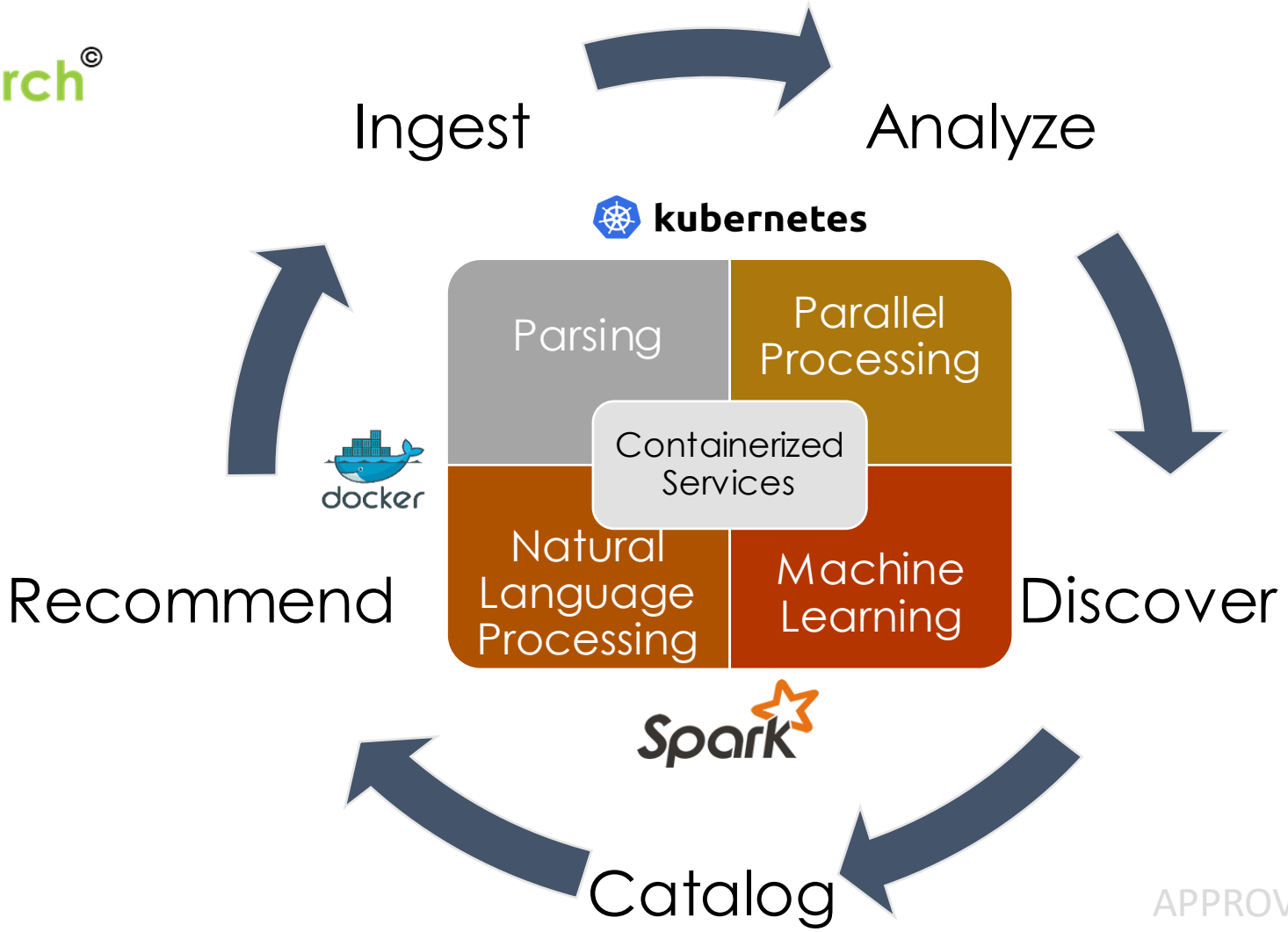
Input

72%

Discovered

APPROVED FOR PUBLIC RELEASE

How Does SmartSearch© Work?



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©

NATIONAL ENERGY TECHNOLOGY LABORATORY | SAMi SCIENCE-BASED AI/ML INSTITUTE

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 Pipelines
  - Distributed ML processing via the Spark framework
  - SmartSearch Recommendation Engine
  - LDA Topic Modeling
  - Named Entity Recognition (NER)
  - Question Answering
  - Summarization

SmartSearch© In-Use

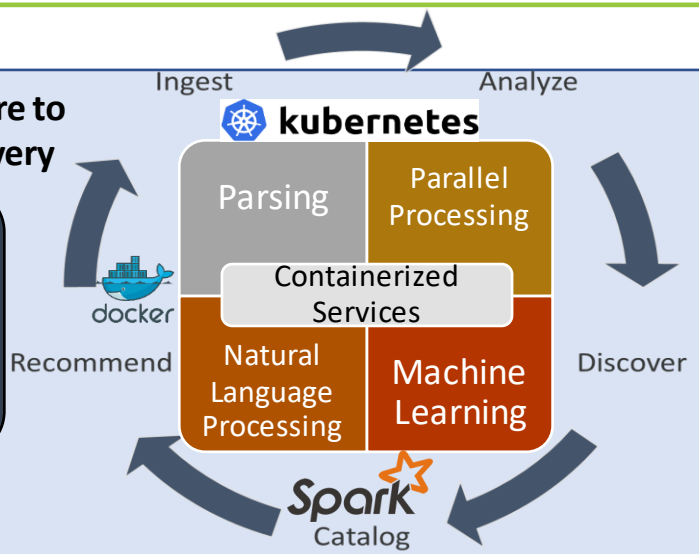
NATIONAL ENERGY TECHNOLOGY LABORATORY

AI informed approach

Challenge: data infrastructure to AI/ML enhanced data discovery

Employing AI/ML tools to find open resources

NETL SmartSearch



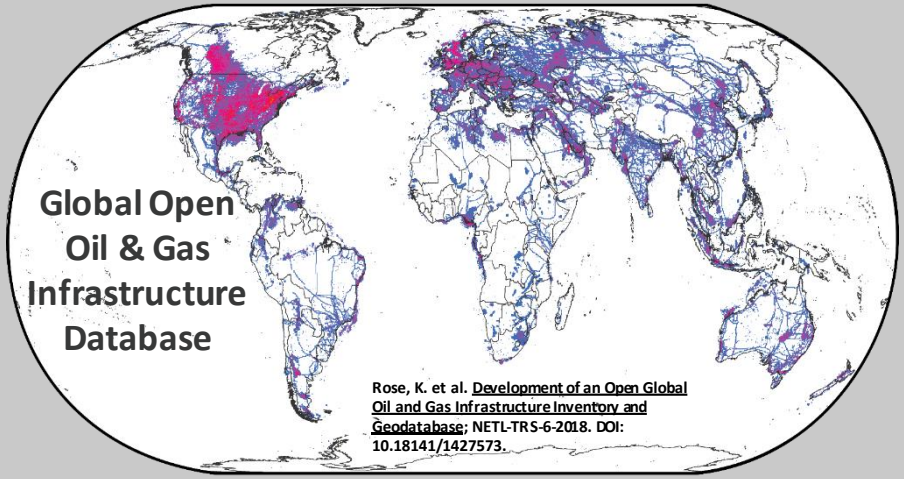
SmartSearch leverages ML+NLP to:

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

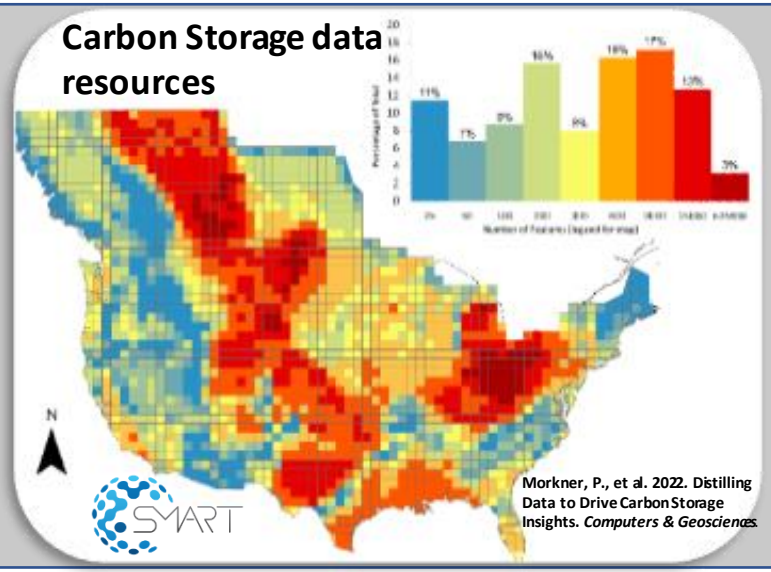
Opportunity:

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

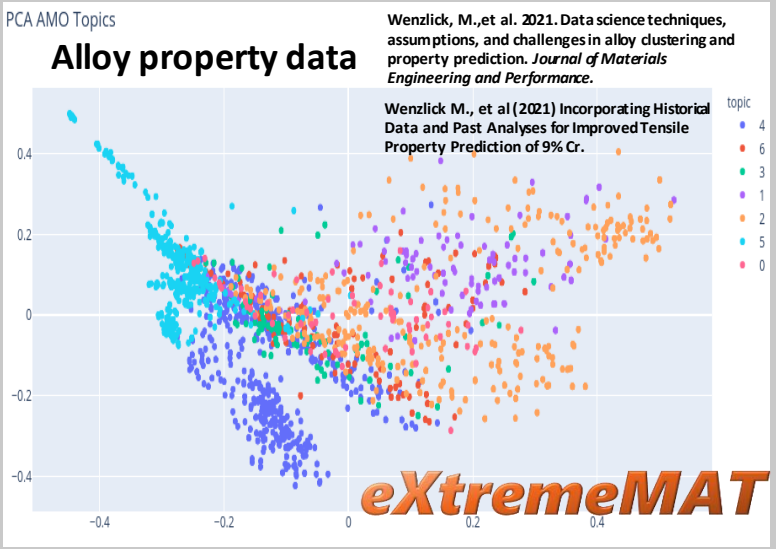
Example applications to date



Carbon Storage data resources



Alloy property data



SmartSearch© for building the Carbon Storage Open Database

NATIONAL ENERGY TECHNOLOGY LABORATORY

Collection of geospatial data from disparate websites (NETL SmartSearch)

↓

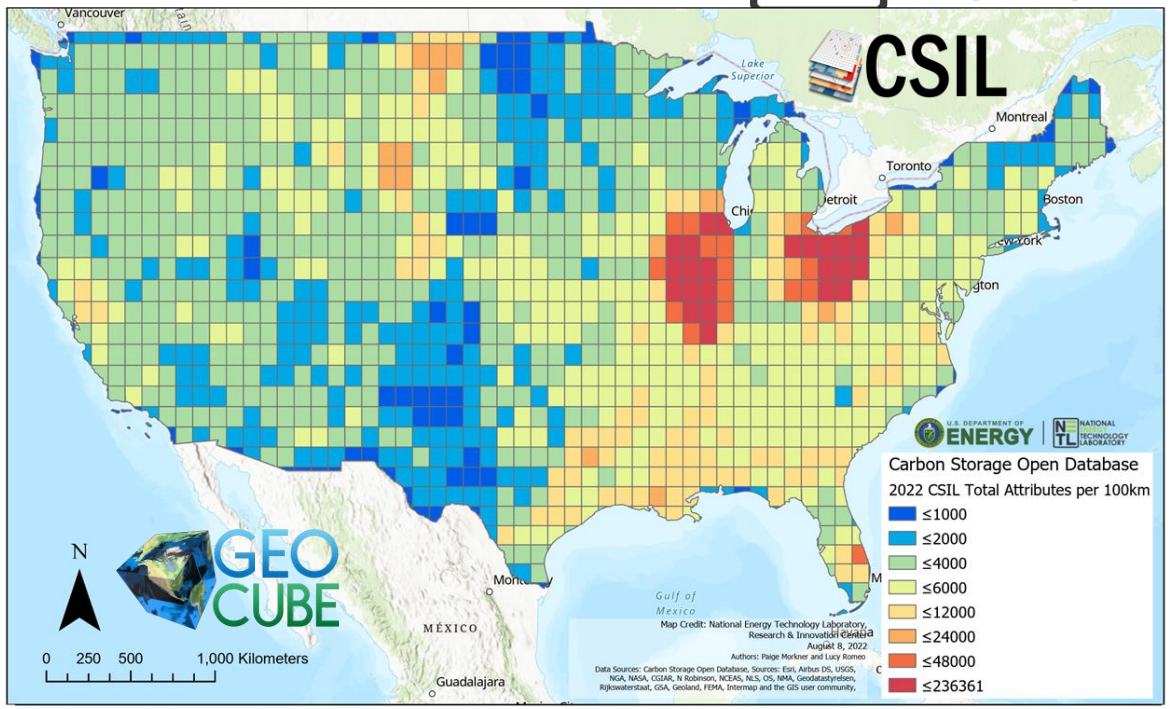
Aggregation of resources into central database

↓

Cataloging and collection of metadata

↓

Integration of metadata and data into EDX++ framework and online mapping platform

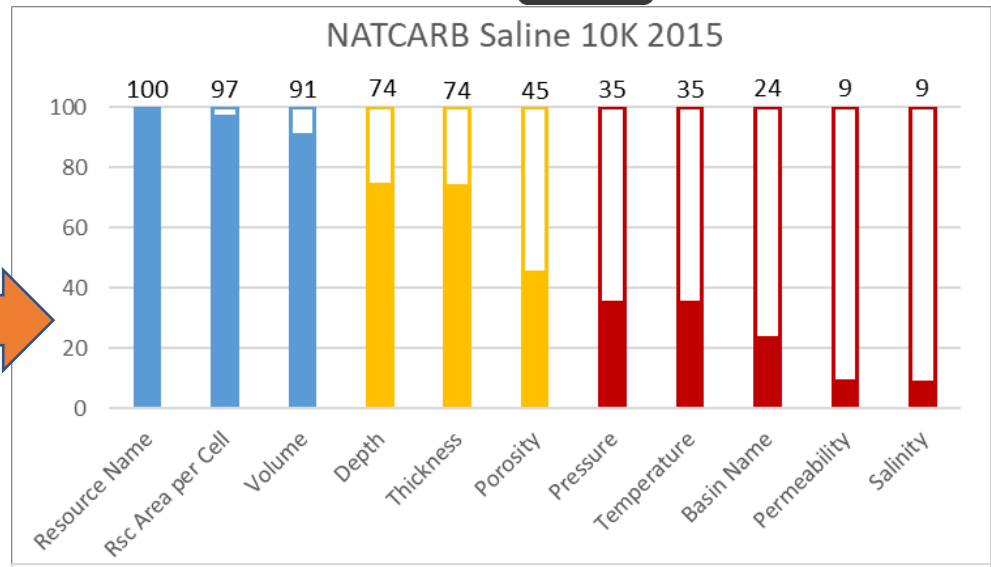


- Use of SmartSearch in finding geospatial data resources to build out collection of 892 geospatial data resources to build the Carbon Storage Open Database on GeoCube
- Morkner, P., et al. 2022. Distilling Data to Drive Carbon Storage Insights. *Computers & Geosciences*.

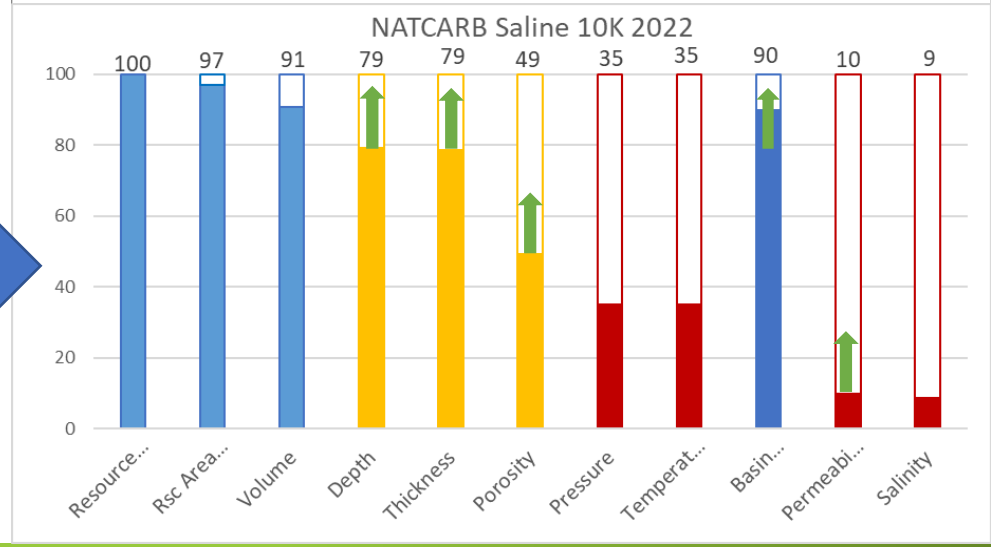
SmartSearch© for addressing data gaps in CS data

- SmartSearch was leveraged to find data to meet data gaps in the NATCARB database during EY21
- NATCARB's 2015 Saline 10km layer was used as seed data
- 1000s of results returned above 10% similarity
- Result text bodies were integrated into a postgreSQL database where keywords were used to query relevant results
- Researchers integrated thousands of new data results improving data available for:
  - Depth, thickness, porosity, basin name, and permeability

Original NATCARB Saline 10km

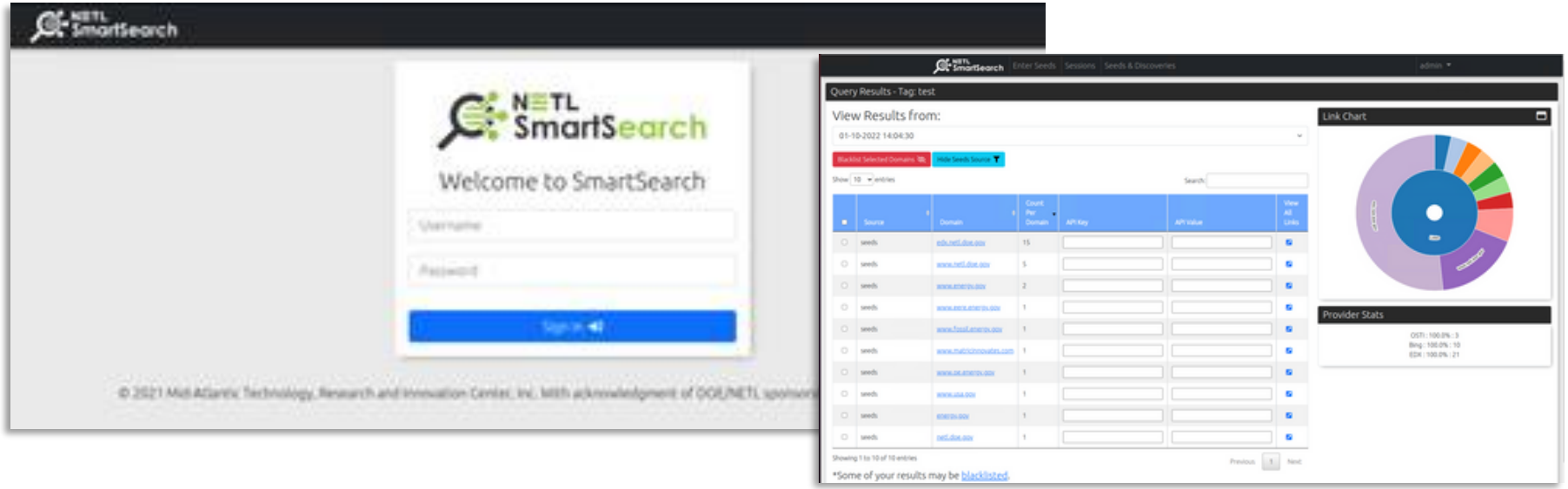


With additional data added



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)



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

Thank you!

References:

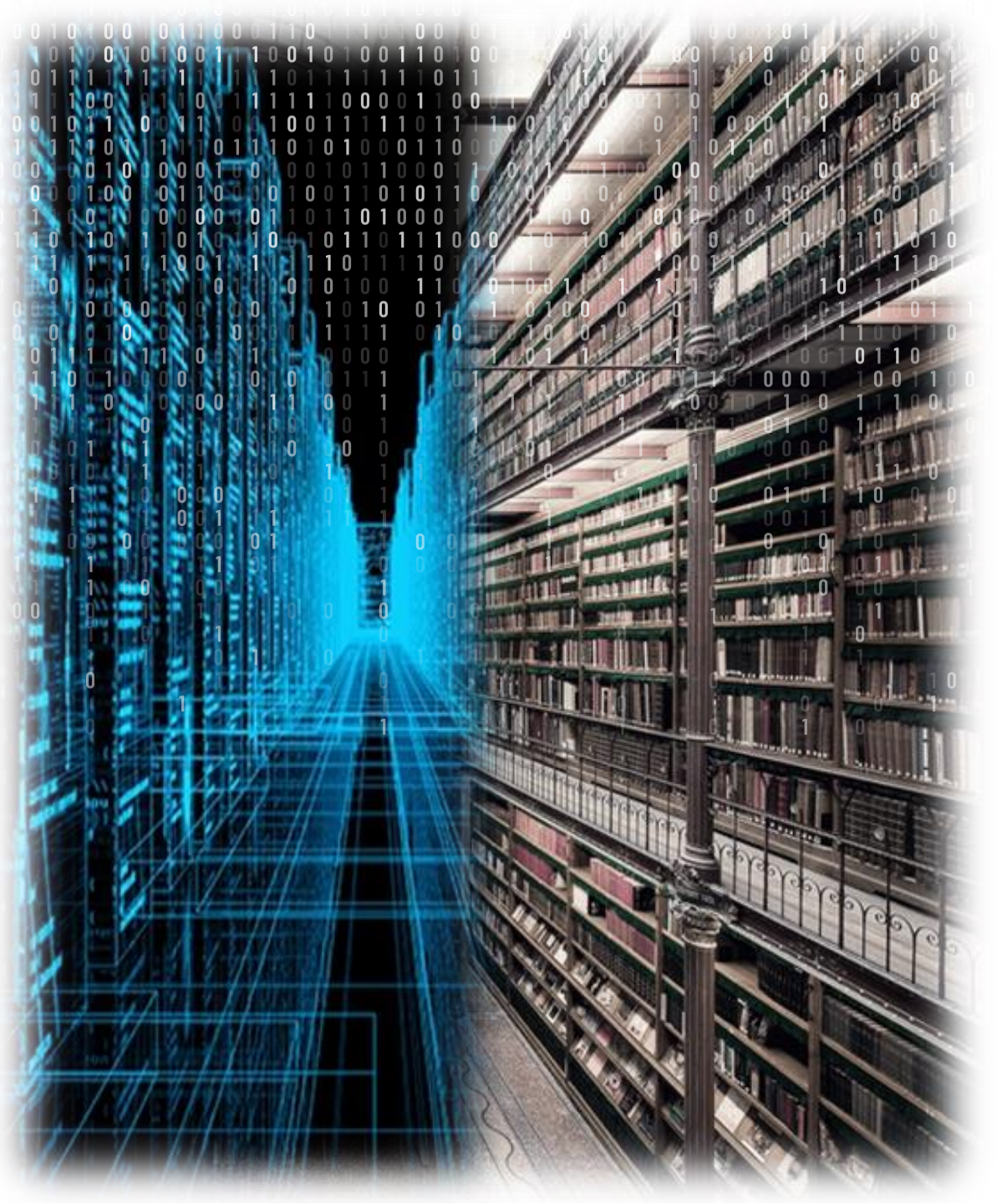
- Morkner, P., et al. 2022. Distilling Data to Drive Carbon Storage Insights. *Computers & Geosciences*.
- Rose, K. et al. Development of an Open Global Oil and Gas Infrastructure Inventory and Geodatabase; NETL-TRS-6-2018. DOI: 10.18141/1427573
- Wenzlick, M.,et al. 2021. Data science techniques, assumptions, 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

NATIONAL ENERGY TECHNOLOGY LABORATORY

Mid-Atlantic Technology, Research & Innovation Center
MATRIC