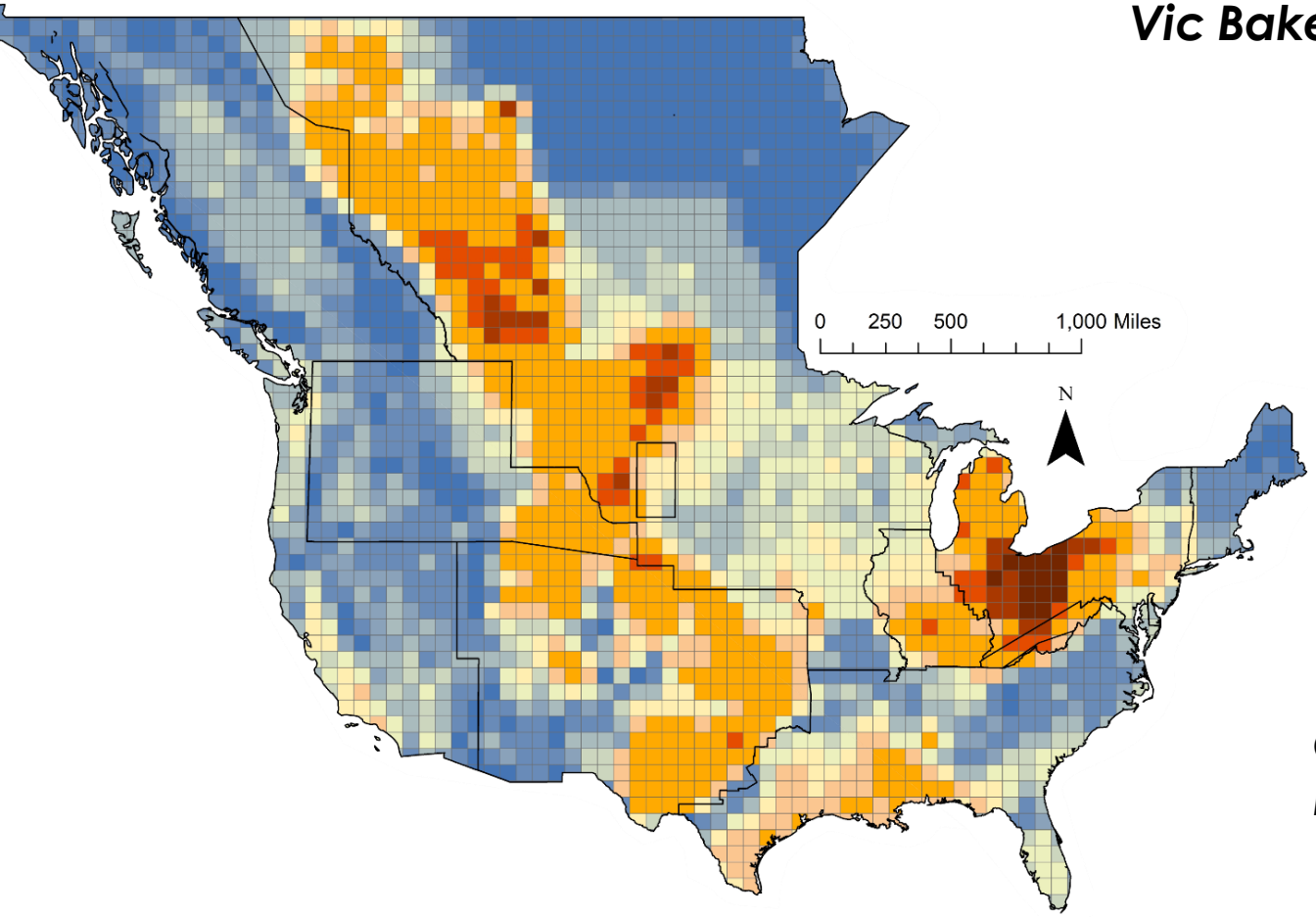
NETL's SmartSearch Deep Learning Tool

Scalable Data Search and Parse for Carbon Storage Data Discovery



Vic Baker[1,2], Jennifer Bauer[1], Paige Morkner[2,3], Kelly Rose[1]

[1] *NETL*, [2] *MATRIC*, [3] NETL Support Contractor

NETL SmartSearch

Carbon Management and Oil and Gas Research Project Review Meeting - Carbon Storage

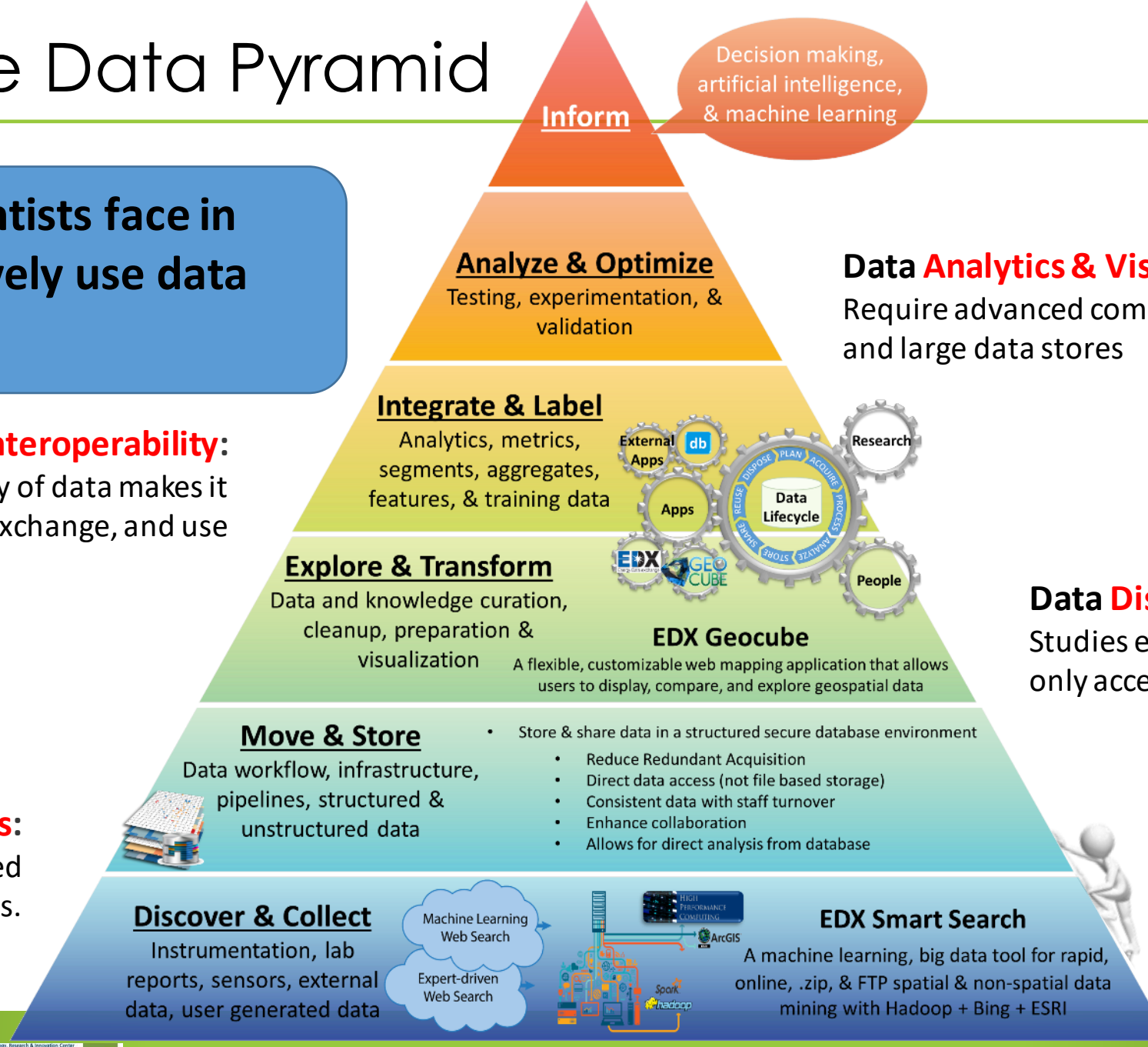
August 4 - 11, 2021

Scaling the Data Pyramid

Challenges scientists face in order to effectively use data resources:

Data Interoperability:
Large variety of data makes it difficult to create, exchange, and use

Data Access:
~80% loss of published data after 20 yrs.



Data Analytics & Visualization:
Require advanced computational capabilities and large data stores

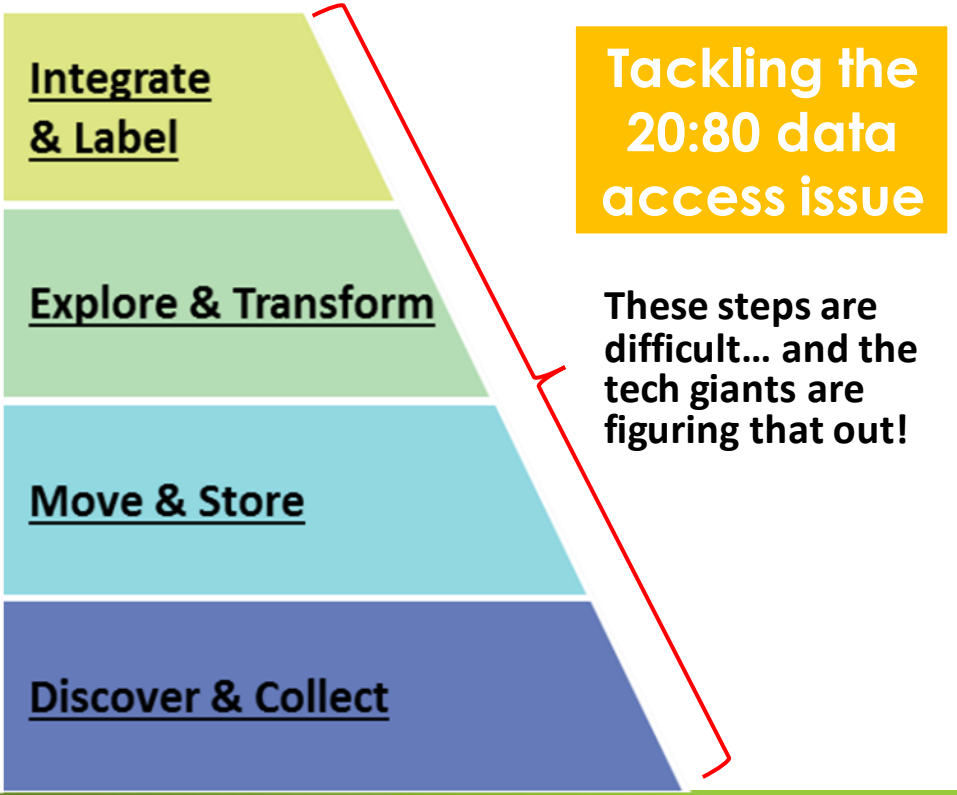
Data Discovery:
Studies estimate avg data project only accesses 20% of relevant data

Scaling the data pyramid is not simple

The "Big Data" Problem

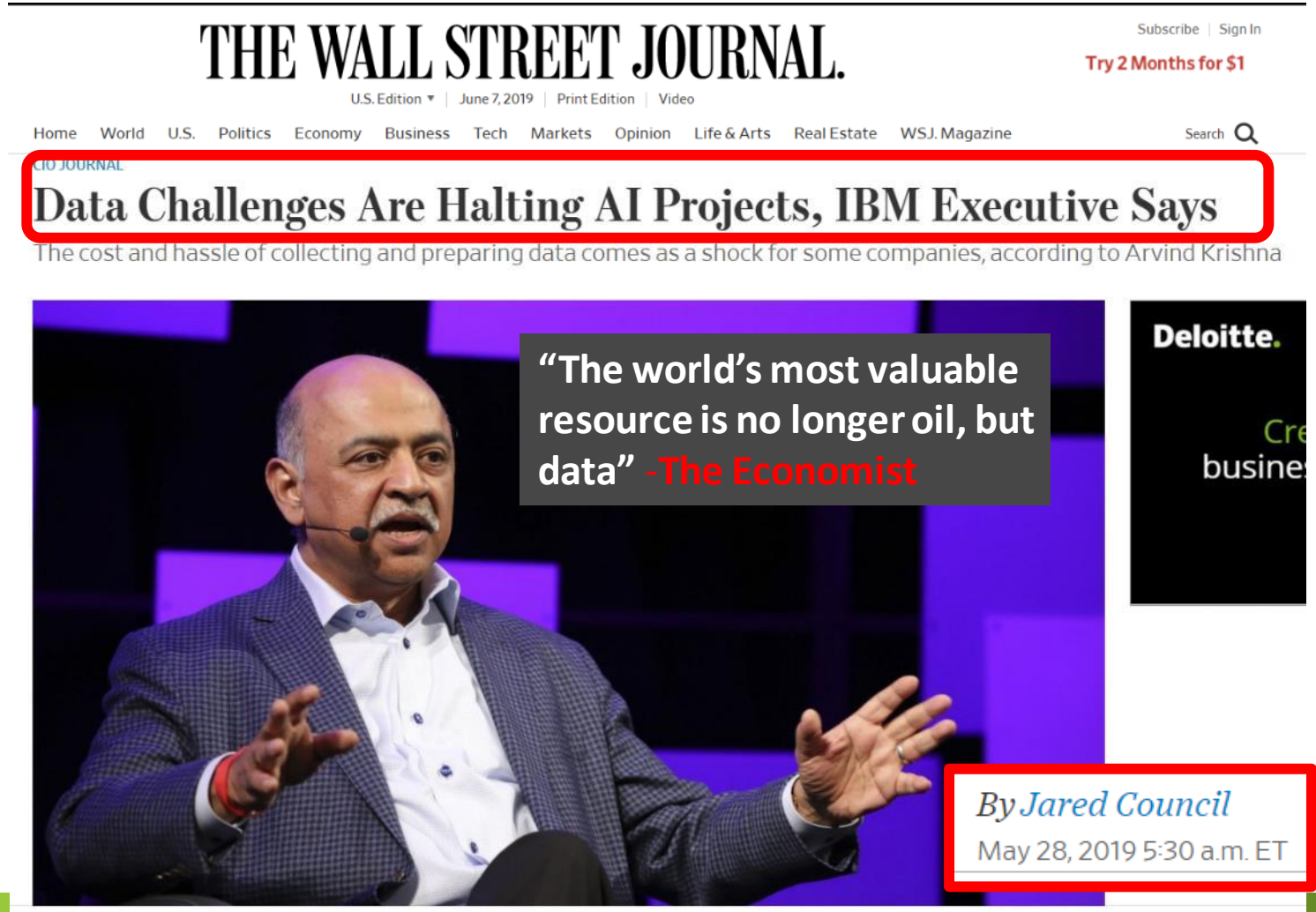
The value of data

So much data… how do we handle all of it?



Tackling the 20:80 data access issue

These steps are difficult… and the tech giants are figuring that out!



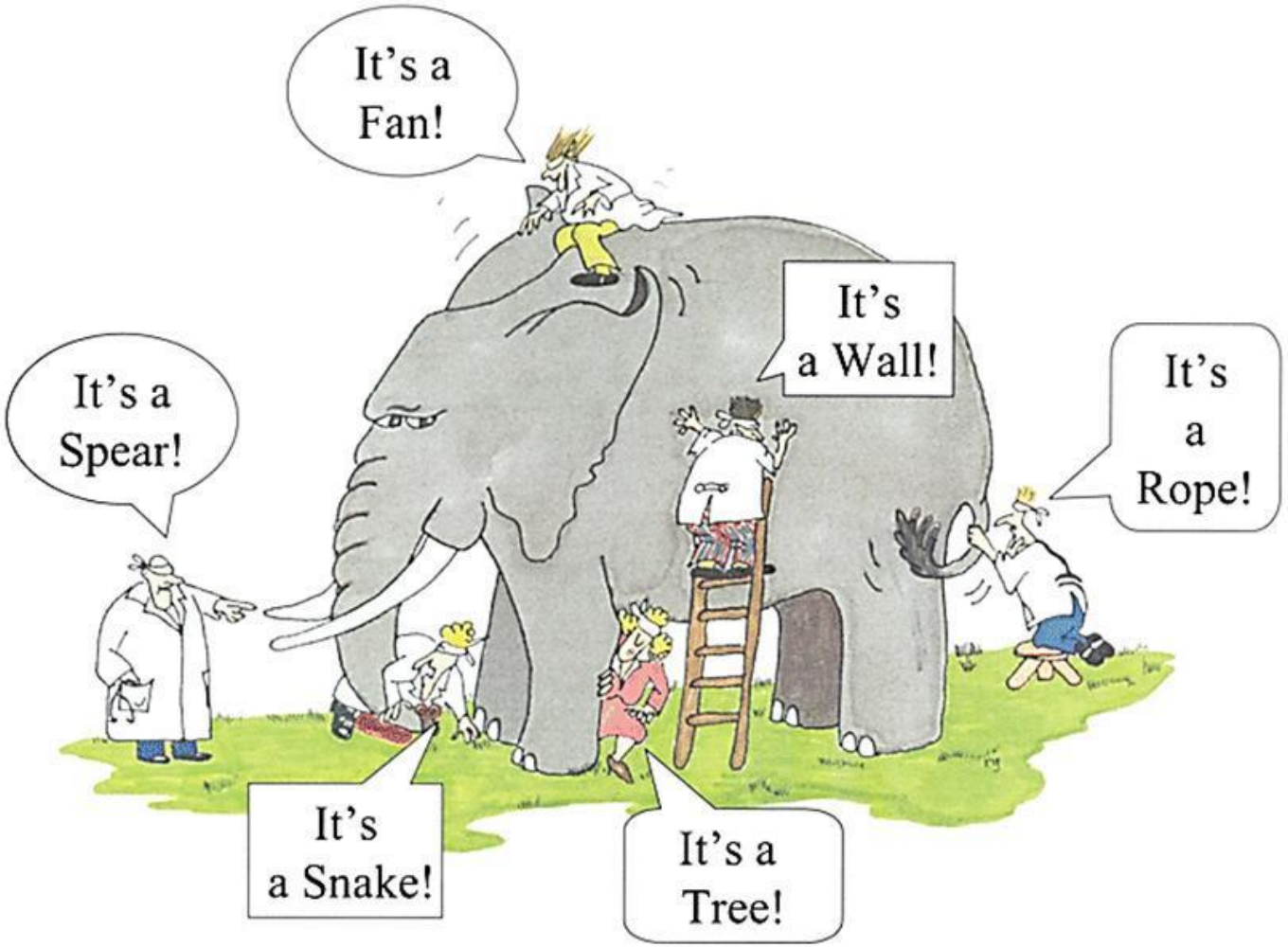
Data Discovery Challenges

Data is often unstructured, mixed:

- Spatial, contextual
- FTP, WWW, local filesystems, storage area networks, etc.

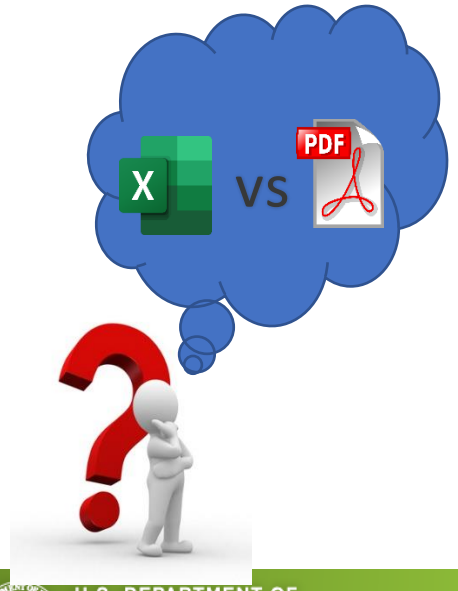
Convoluted ways to search for and identify data:

- Hard to identify all the data, i.e., see the whole “Elephant”, without falling down the “rabbit hole”



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 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

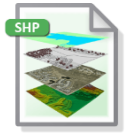
Solution:

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 new data is to what you like



Input

72%

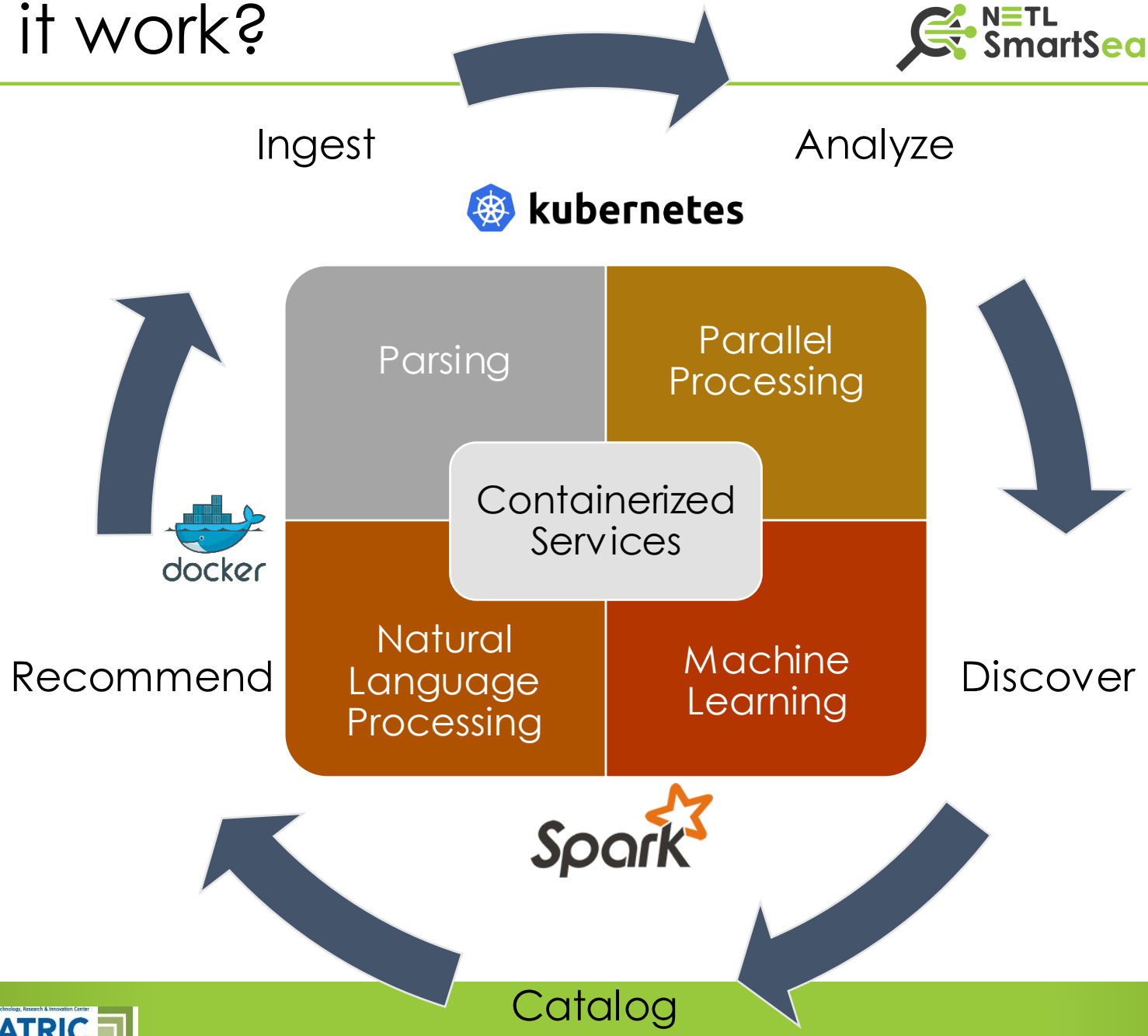
Discovered

Benefits of SmartSearch

- Infinitely Scalable (Automated) Data Discovery
  - Analyze millions+ of files and generate comparison metrics
  - Generate topic models, categorization
  - Desktop, cluster, cloud
- Treat geospatial data like a document
  - Automatically extract text from geospatial data (shapefiles, gec
  - Compare textual vs geospatial data to identify relevancy

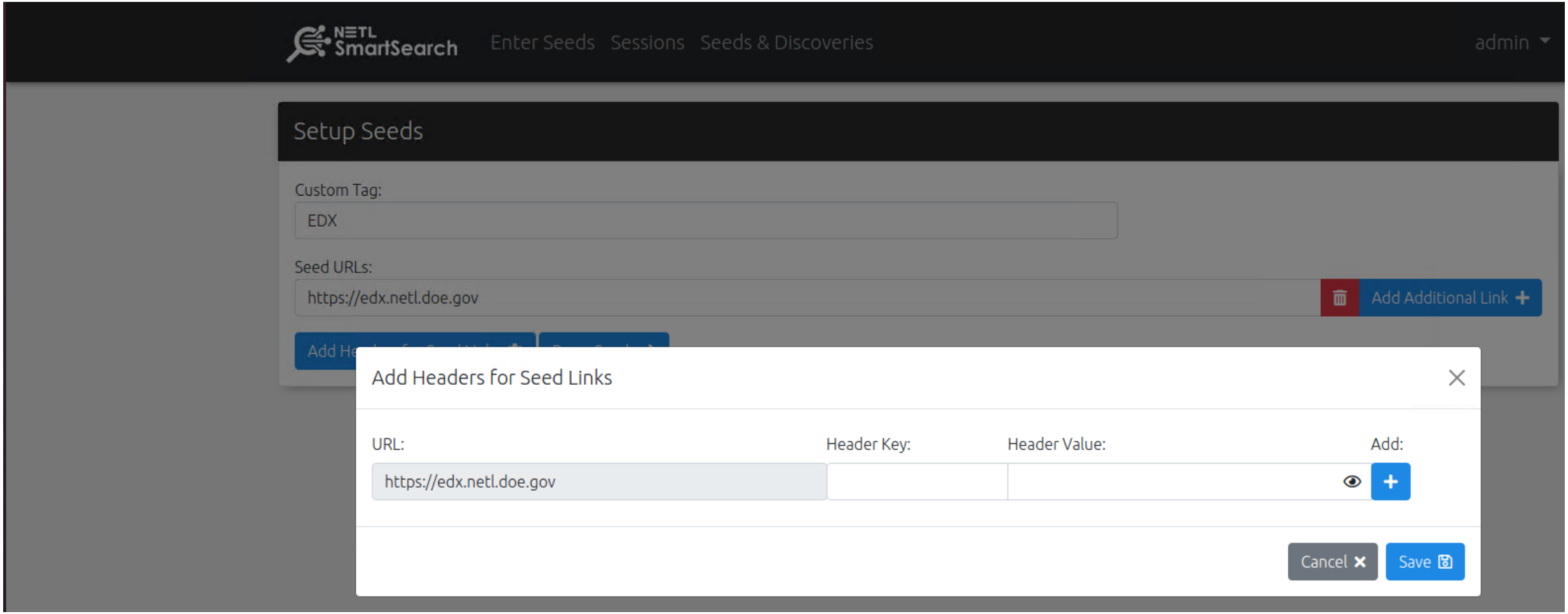
- Search for meta tags within HTML body of discovered web sites
  - i.e., find map tags
- Analyze archive files – even archives within archives (zips within zips, etc.)
  - Process every file – docs, spatial, etc.

How does it work?



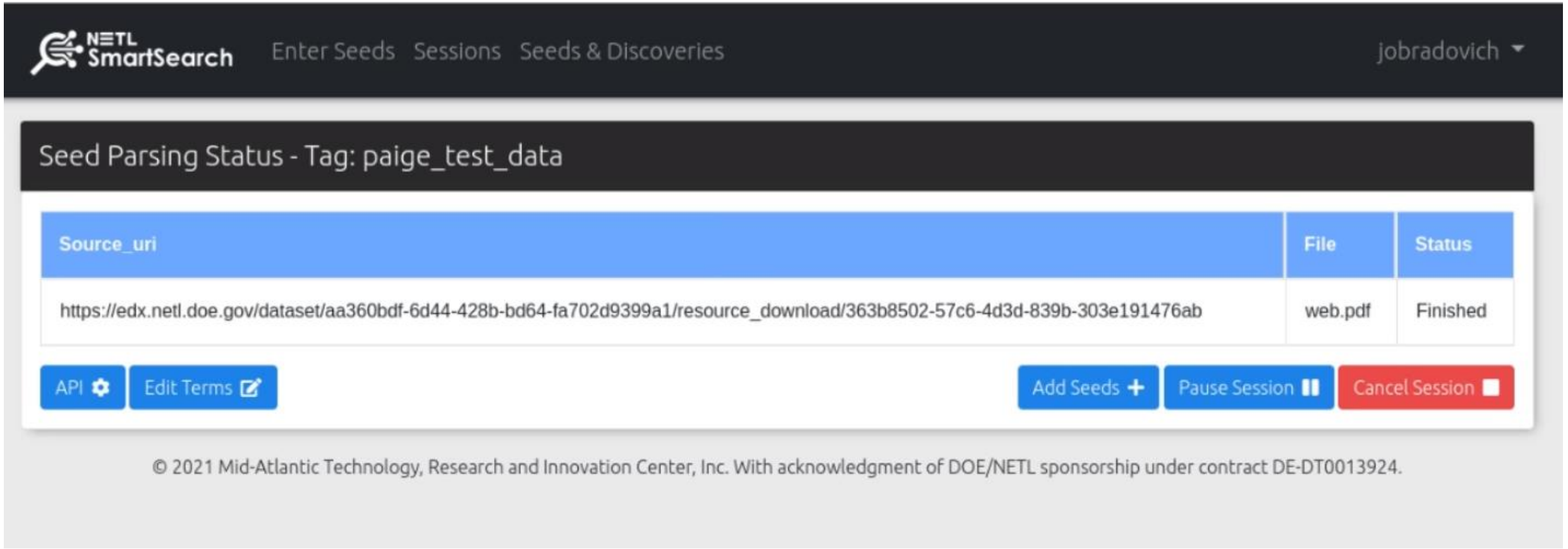
Seed Entry

Users input urls to web resources, along with any relevant API keys



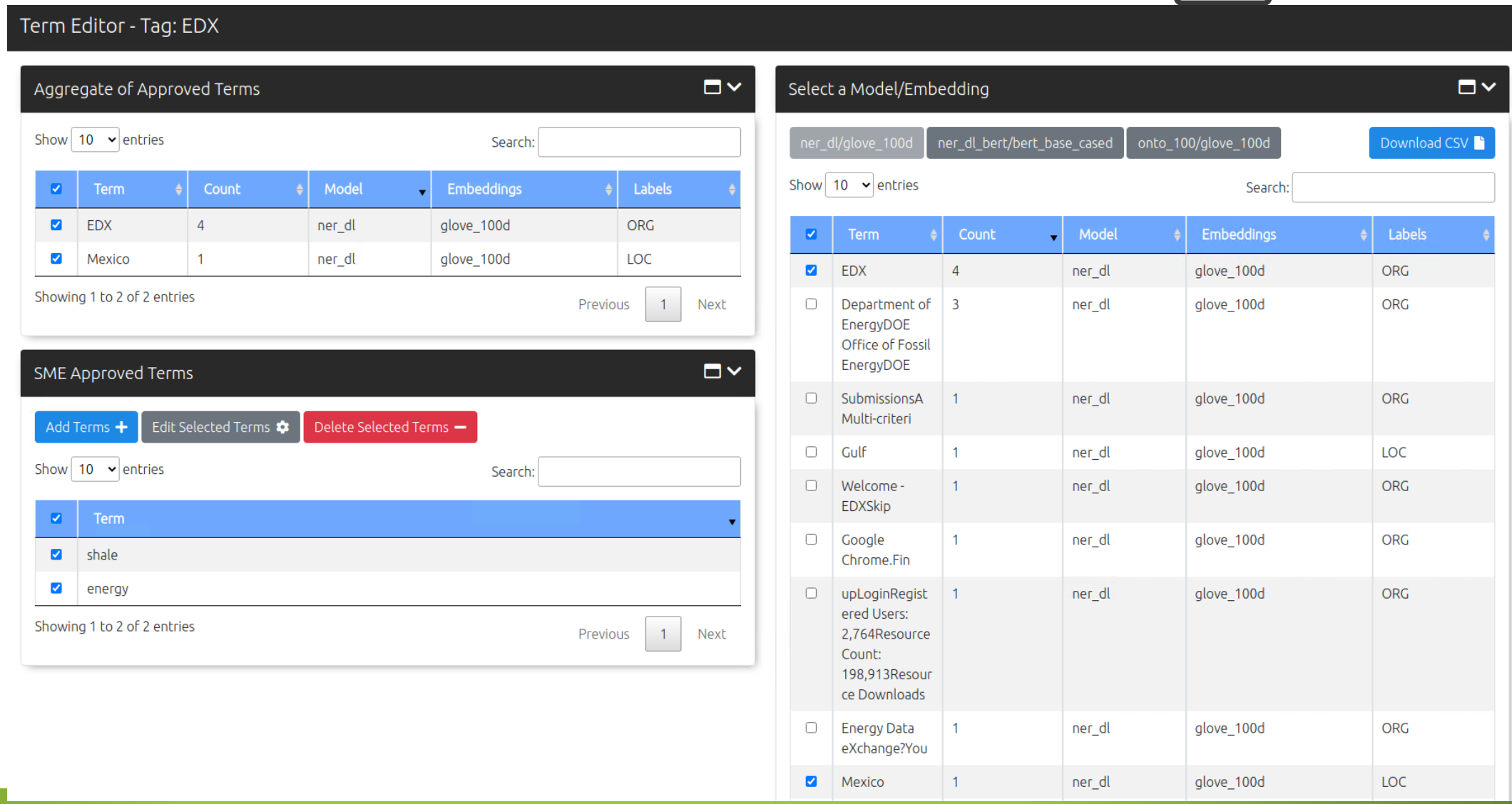
SmartSearch UI

Initial seed data provided to SmartSearch – document here is parsed and ready to be processed via NLP – key terms, locations



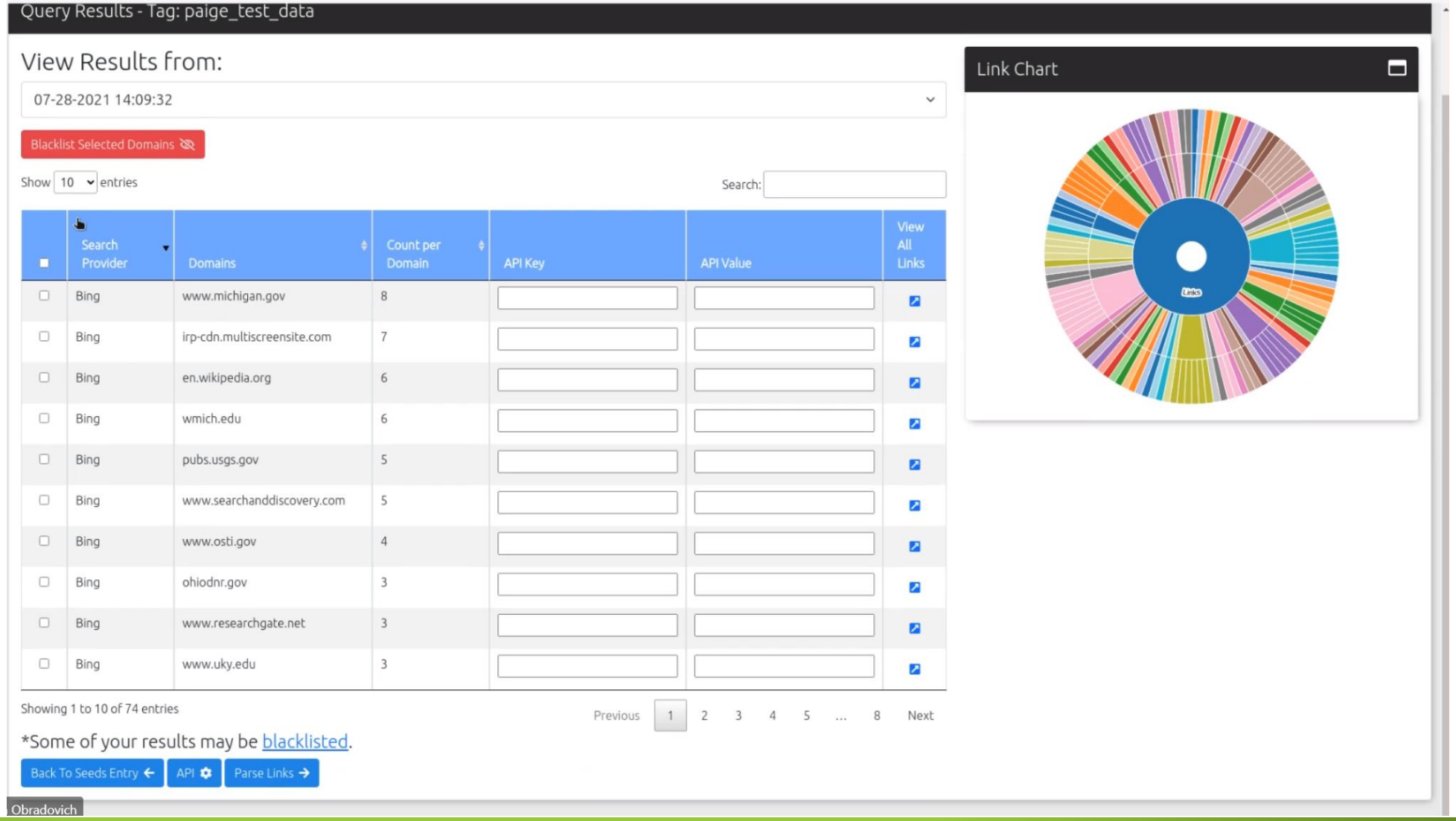
SmartSearch UI

- Seed document keywords being extracted.
- Produces a list of terms which can be approved by user
- User can also provide approved, supplemental terms to aid search



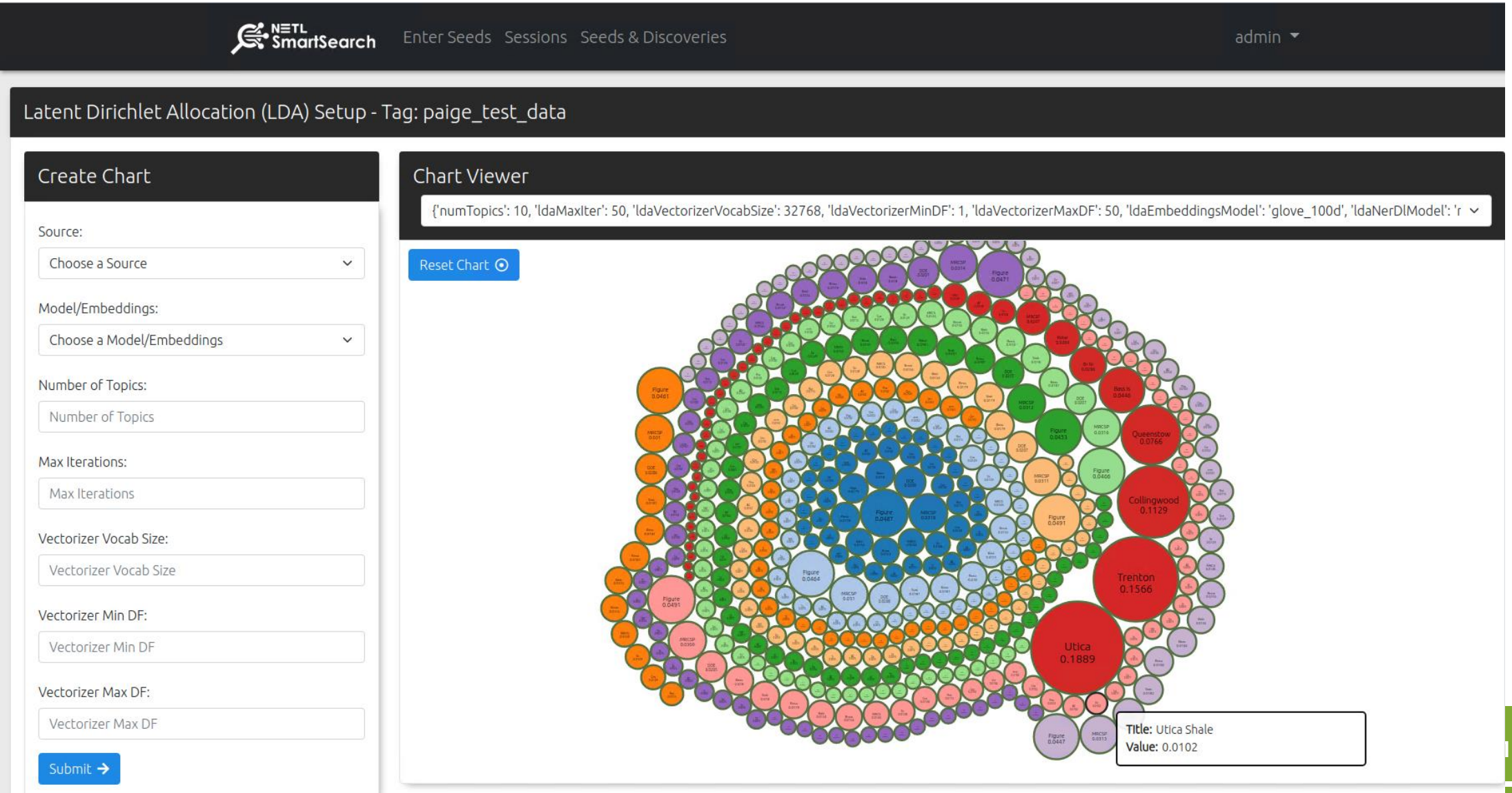
SmartSearch UI

- Seed parsed terms are then used to search the WWW for other relevant results
- Results are returned of webpages that likely contain relevant results



SmartSearch UI

Latent Dirichlet Allocation – a type of NLP topic modeling – can be calculated and visualized



SmartSearch UI

NETL SmartSearch

NATIONAL ENERGY TECHNOLOGY LABORATORY

Results of relevant resources discoveries are returned with a Similarity score (0.1 < x < 1)
(Min threshold customizable)

NETL SmartSearch   Enter Seeds   Sessions   Seeds & Discoveries   admin

Relevance Analysis

Toggle column: Seed Session ID | Seed Tag | Seed Token | Seed Source URI | Seed File | Seed Body | Discovered Session ID | Discovered Tag | Discovered Token | Discovered Source URI | Discovered File | Discovered Body | Similarity

Show 10 entries   Search:

Seed Session ID	Seed Tag	Seed Token	Seed Source URI	Seed File	Seed Body	Discovered Session ID	Discovered Tag	Discovered Token	Discovered Source URI	Discovered File	Discovered Body	Similarity
sida4c26dfd7…	paige_test_data	f4e62d4c31a…	https://edx.n…	web.pdf	Volume VI.IMi…	sida4c26dfd7…	paige_test_data	af7578627c3…	https://www.…	web.html	Leveraging R…	0.28418
sida4c26dfd7…	paige_test_data	f4e62d4c31a…	https://edx.n…	web.pdf	Volume VI.IMi…	sida4c26dfd7…	paig					
sida4c26dfd7…	paige_test_data	f4e62d4c31a…	https://edx.n…	web.pdf	Volume VI.IMi…	sida4c26dfd7…	paig					
sida4c26dfd7…	paige_test_data	f4e62d4c31a…	https://edx.n…	web.pdf	Volume VI.IMi…	sida4c26dfd7…	paige_test_data	edceba41eea…	https://www.…	web.html	State Charlto…	0.28102
sida4c26dfd7…	paige_test_data	f4e62d4c31a…	https://edx.n…	web.pdf	Volume VI.IMi…	sida4c26dfd7…	paige_test_data	864054cc560…	https://www.…	web.html	State Charlto…	0.28095
sida4c26dfd7…	paige_test_data	f4e62d4c31a…	https://edx.n…	web.pdf	Volume VI.IMi…	sida4c26dfd7…	paige_test_data	af7578627c3…	https://www.…	web.html	Leveraging R…	0.27846
sida4c26dfd7…	paige_test_data	f4e62d4c31a…	https://edx.n…	web.pdf	Volume VI.IMi…	sida4c26dfd7…	paige_test_data	af7578627c3…	https://www.…	web.html	Leveraging R…	0.27846
sida4c26dfd7…	paige_test_data	f4e62d4c31a…	https://edx.n…	web.pdf	Volume VI.IMi…	sida4c26dfd7…	paige_test_data	af7578627c3…	https://www.…	web.html	Leveraging R…	0.27846
sida4c26dfd7…	paige_test_data	f4e62d4c31a…	https://edx.n…	web.pdf	Volume VI.IMi…	sida4c26dfd7…	paige_test_data	af7578627c3…	https://www.…	web.html	Leveraging R…	0.27846
sida4c26dfd7…	paige_test_data	f4e62d4c31a…	https://edx.n…	web.pdf	Volume VI.IMi…	sida4c26dfd7…	paige_test_data	e2d800b77fc…	https://www.…	web.html	Leveraging R…	0.27843

Leveraging Regional Exploration to Develop Geologic Framework for CO2 Storage in Deep Formations in Midwestern United States (Technical Report) | OSTI.GOVskip to main contentSign InCreate AccountU.S. Department of Energy Office of Scientific and Technical Information Search terms: Advanced search options Advanced Search OptionsAdvanced Search queries use a traditional Term Search. For more info, s…

Showing 1 to 10 of 241 entries

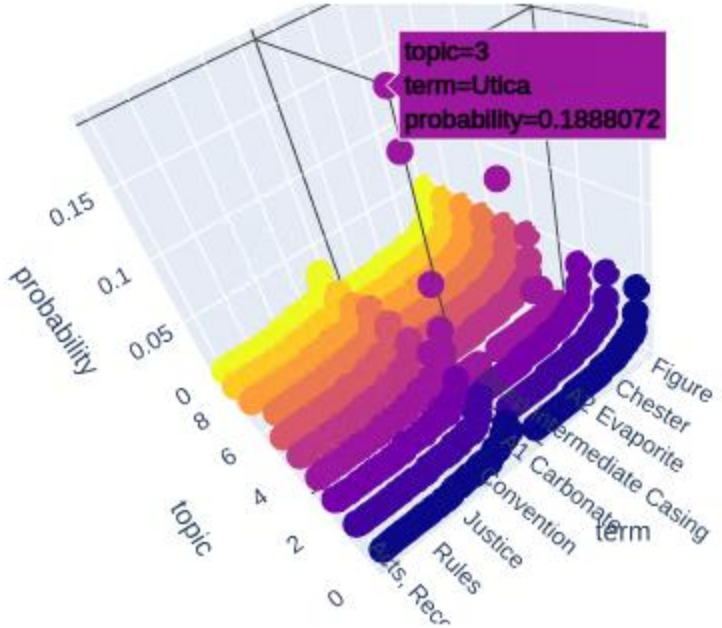
Previous 1 2 3 4 5 … 25 Next

Download

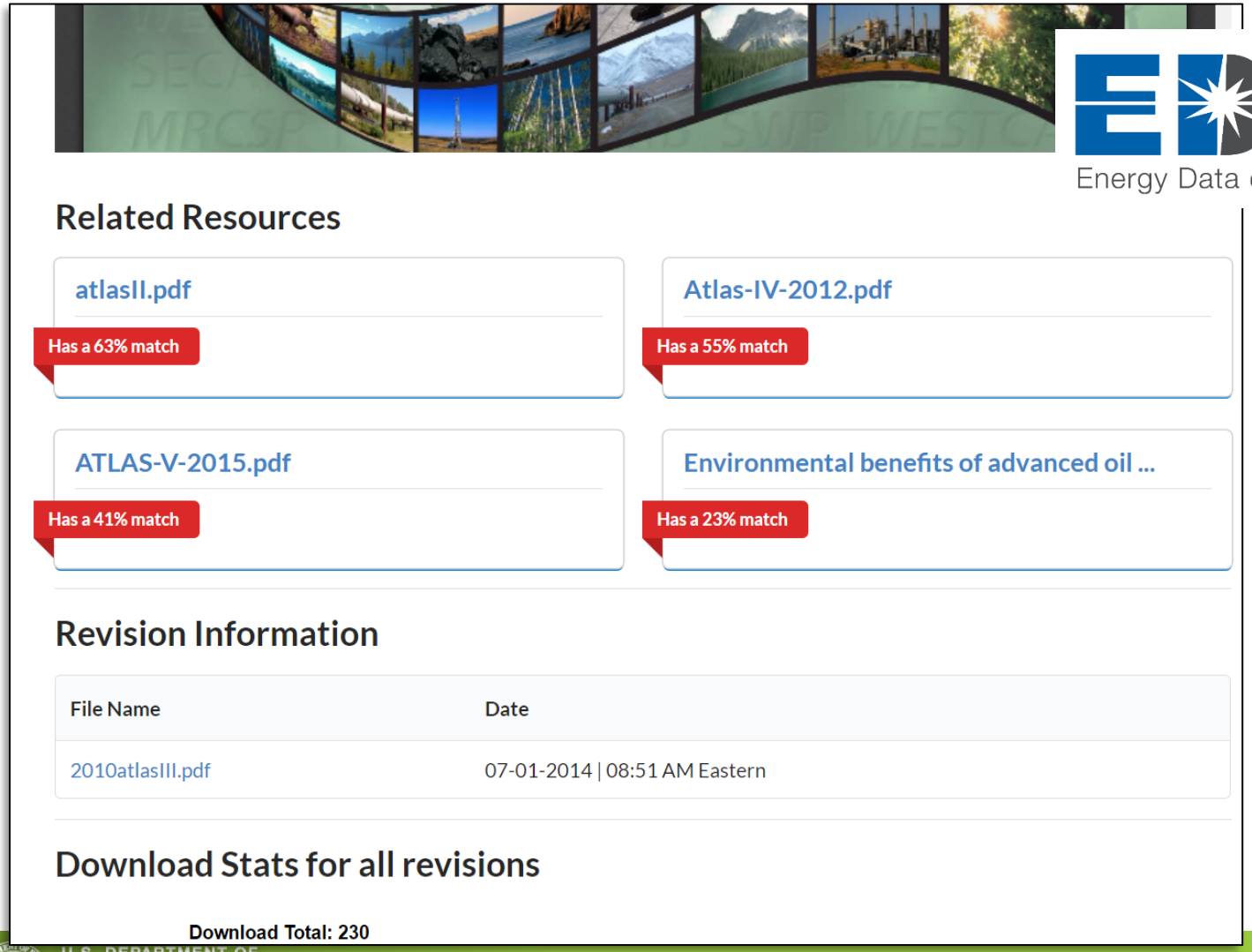
Visualizations

SmartSearch offers rich visualization capabilities

spark-recommendations-strip-graph



Ultimately, launch SmartSearch in EDX as a **Deep Analysis** Recommendation Engine



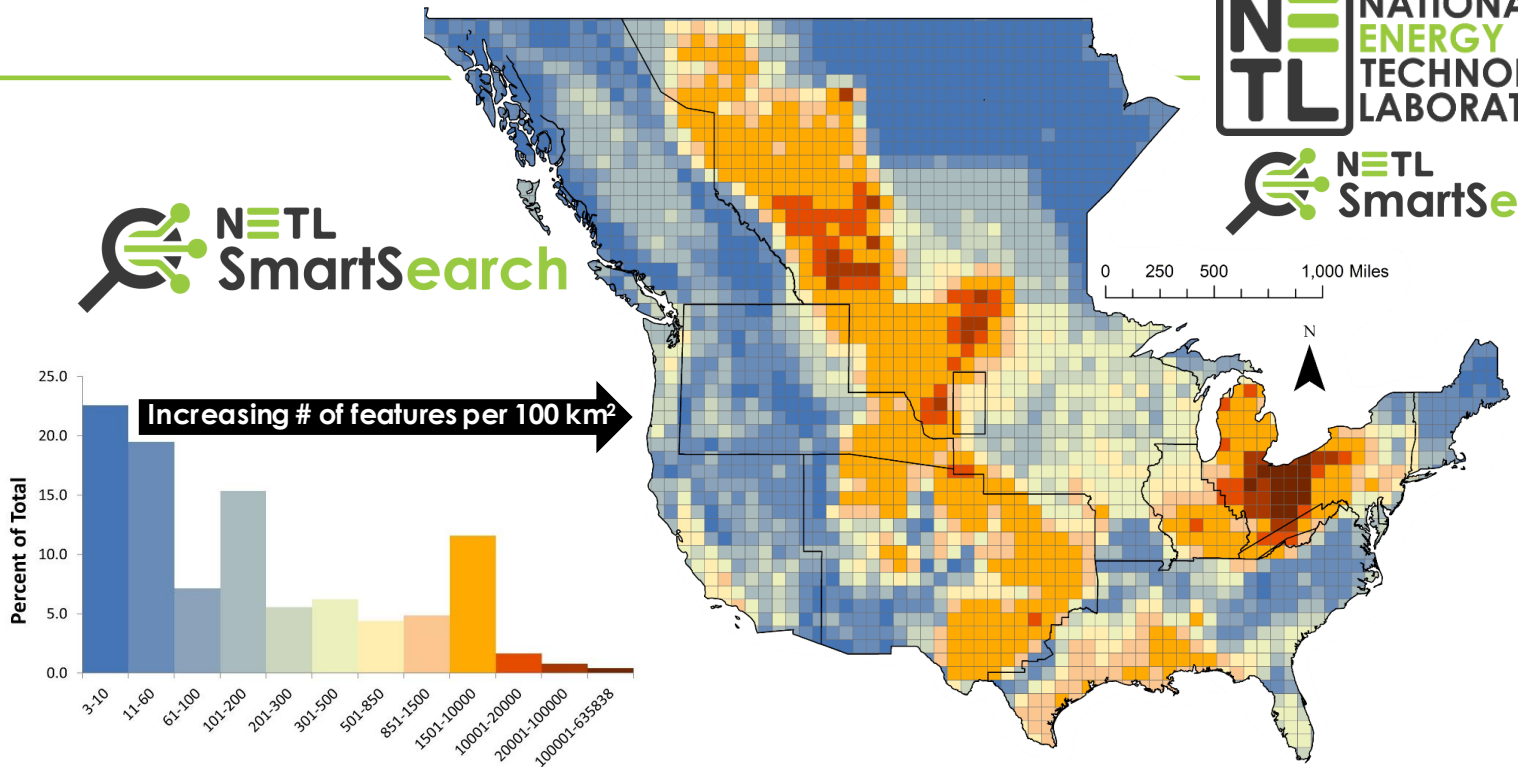
Anticipate releasing SmartSearch v1 in next few months via EDX

- Perform deep contextual analysis
- Machine learning, natural language processing
- Generates correlation matches of contextually similar content
- Expanded to include spatial and webcrawl assets
- Implemented using Spark, Scala, Python, Kubernetes
- Ideal for cluster – RAM, CPU, and bandwidth intensive

Building A **Big Data Ecosystem** for CS R&D Data Discovery

NATIONAL ENERGY TECHNOLOGY LABORATORY

NETL SmartSearch



CS data driven research requires:

- Lots of data
- Incorporating different data types & formats,
- Integrating data from multiple locations (web, local, databases)

Traditional Search methods impede our efforts:

- Search engine limits context to a few terms
- Labor intensive to conduct data searching
- Even more difficult to find relevant data

SmartSearch in Press:

- Baker, D.V., Rose, K., Bauer, J., and Rager, D., 2016, **Computational Advances and Data Analytics to Reduce Subsurface Uncertainty**, ARMA 16-493, June 26-29, 2016, 16 pgs.
- Baker, D.V., Bauer, J., and Rose, K., 2018. **Developing a Smarter Way to Search – Parsing the online "forest"** to find data for your research needs via EDX. U.S. DOE Mastering the Subsurface through technology innovation, partnership and collaboration: Carbon Storage and Oil and Natural Gas Technologies review meeting
- Rose, K.; Bauer, J.; Baker, V.; et al., 2018, **Development of an Open Global Oil and Gas Infrastructure Inventory and Geodatabase**; NETL-TRS-6-2018; NETL Technical Report Series; U.S. Department of Energy, National Energy Technology Laboratory: Albany, OR, 2018; p 594; DOI: 10.18141/1427573.

NETL RESOURCES

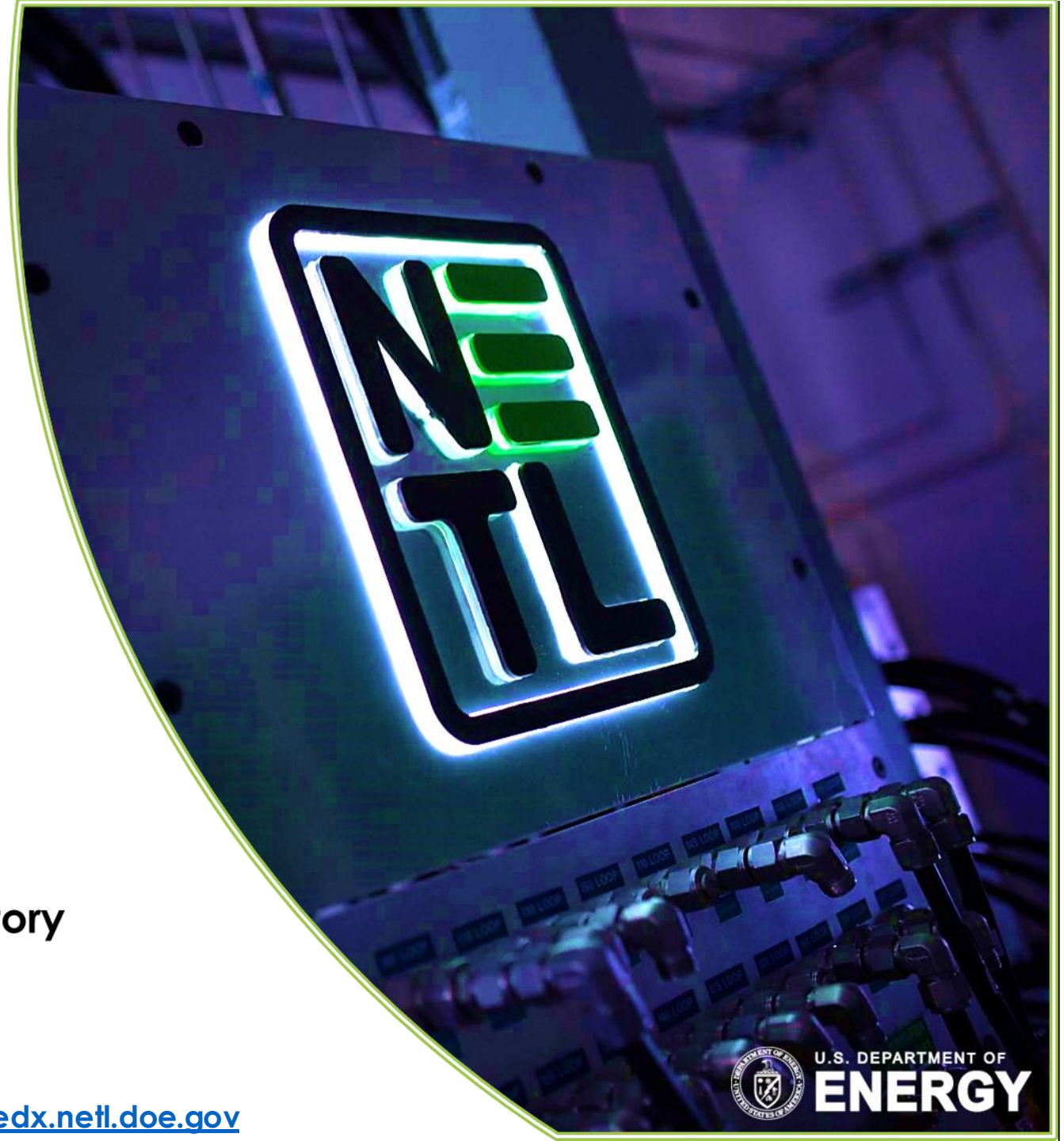
VISIT US AT: **www.NETL.DOE.gov**

@NETL_DOE

@NETL_DOE

@NationalEnergyTechnologyLaboratory

vic.baker@matricinnovates.com
jennifer.bauer@netl.doe.gov



https://edx.netl.doe.gov