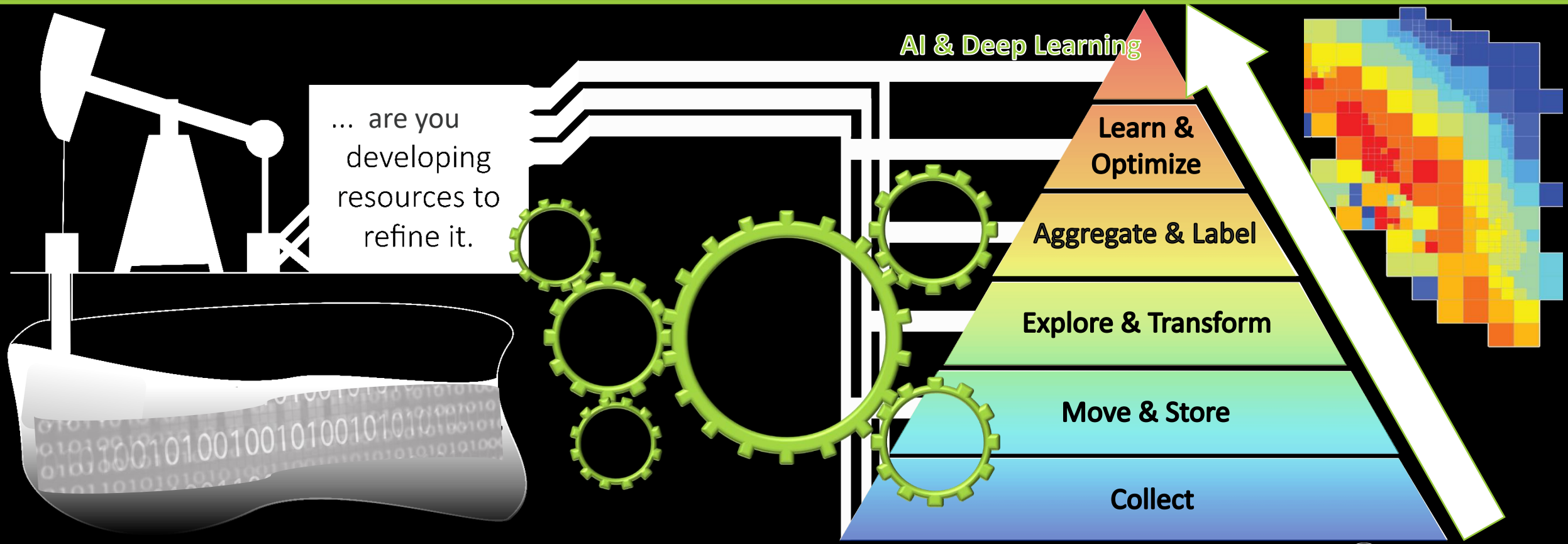
# Data is the new oil…

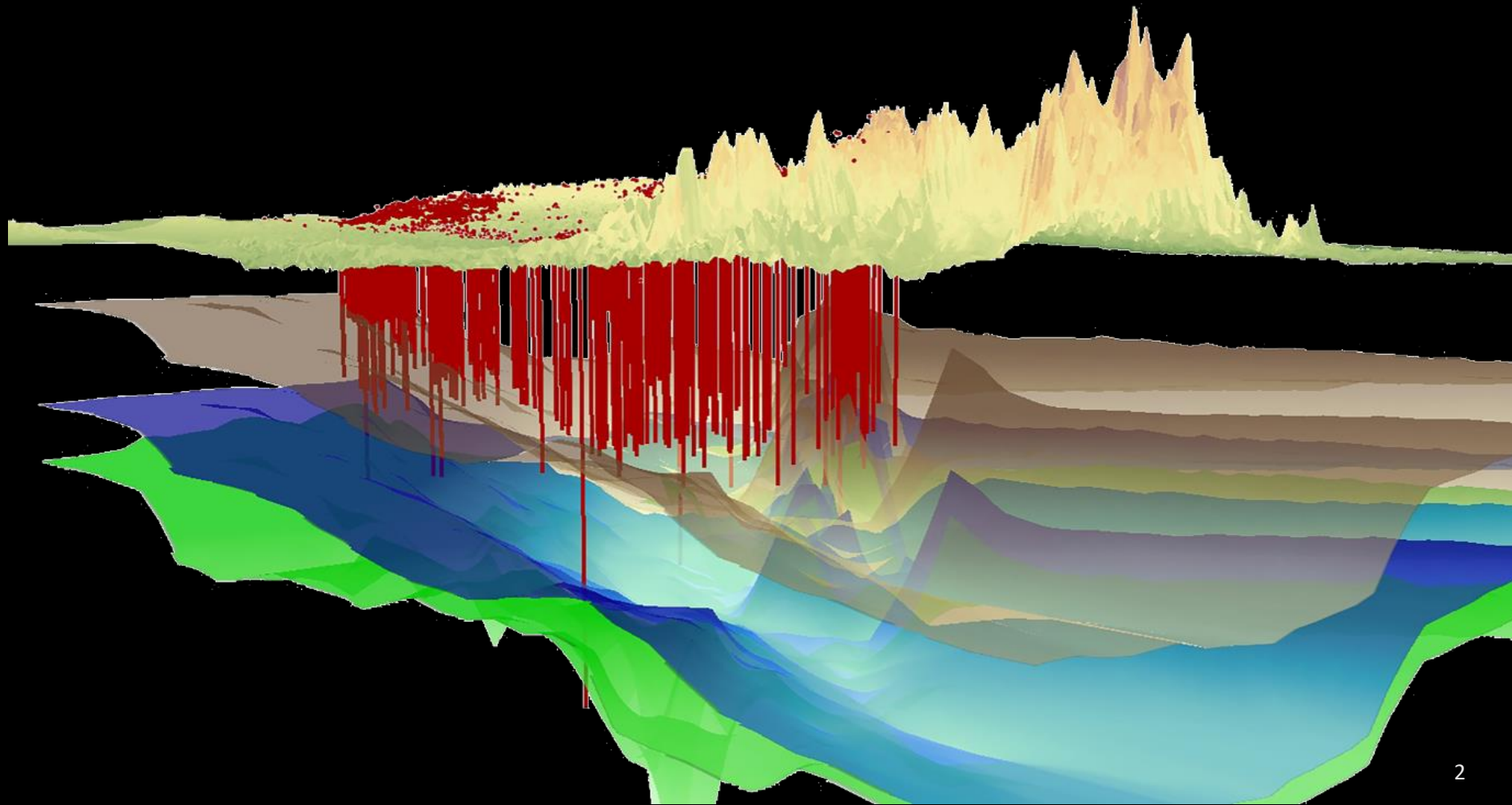
Kelly Rose, HEI Workshop
September 13, 2018



# Building a Data Framework for R&D

ML/BD solutions, tools and capabilities can be devised or implemented to streamline and automate data collection, movement, and transformation

- Collect disparate datasets and contextual information
- Incorporate data across various scales and formats
- Span surface-subsurface
- Not all data is equal, not all data are easy…but there is more out there…put it to work!



# NETL's Geo-Data Science:
# Inventing Intelligent Solutions to DOE FE Data & R&D Needs

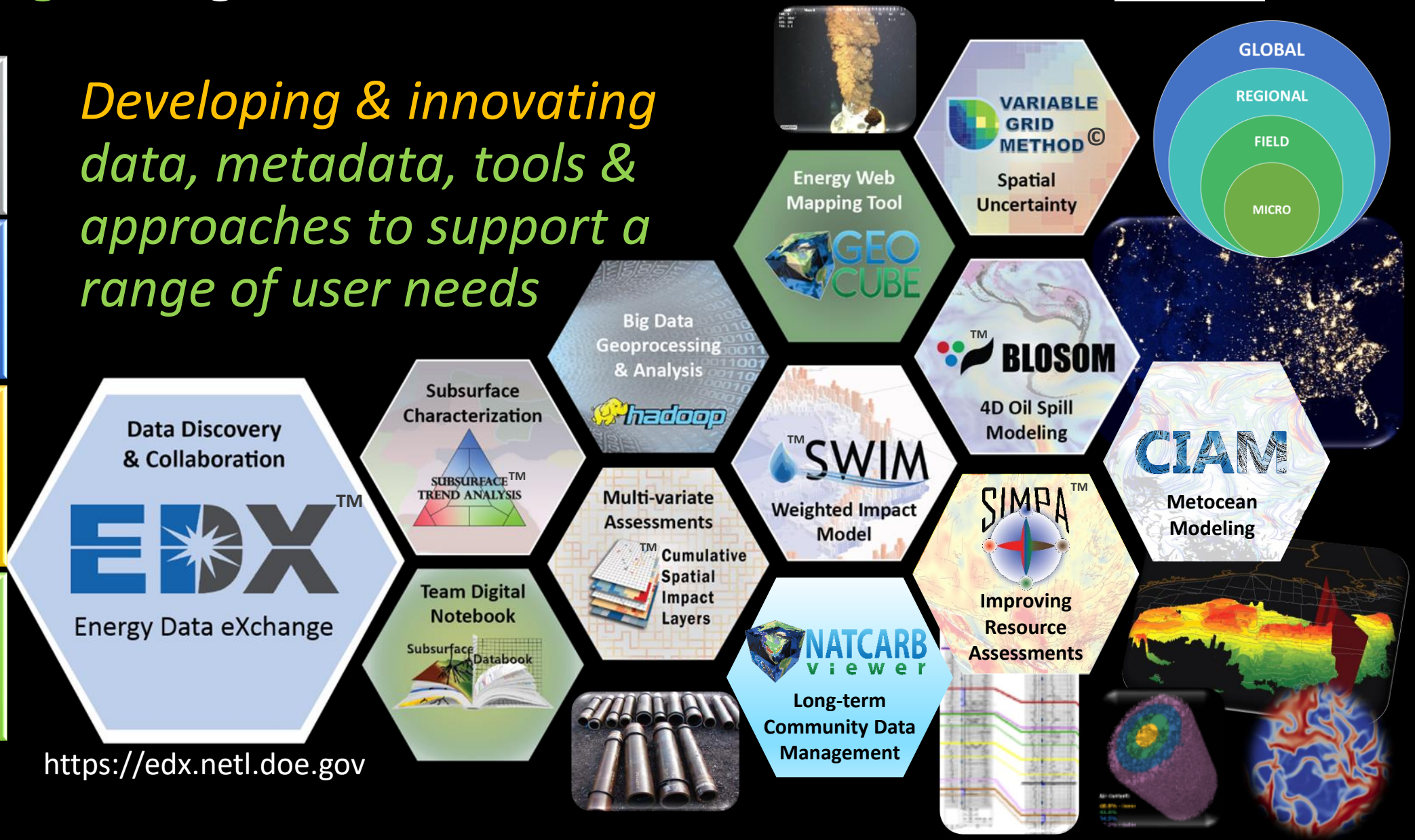
*Developing & innovating data, metadata, tools & approaches to support a range of user needs*

**Data Discovery**

**Data Curation**

**Data Interoperability**

**Data Analytics & Visualization**




https://edx.netl.doe.gov

# Scaling the Data Pyramid -

## *Building Solution for Common Data Challenges*

NATIONAL ENERGY TECHNOLOGY LABORATORY

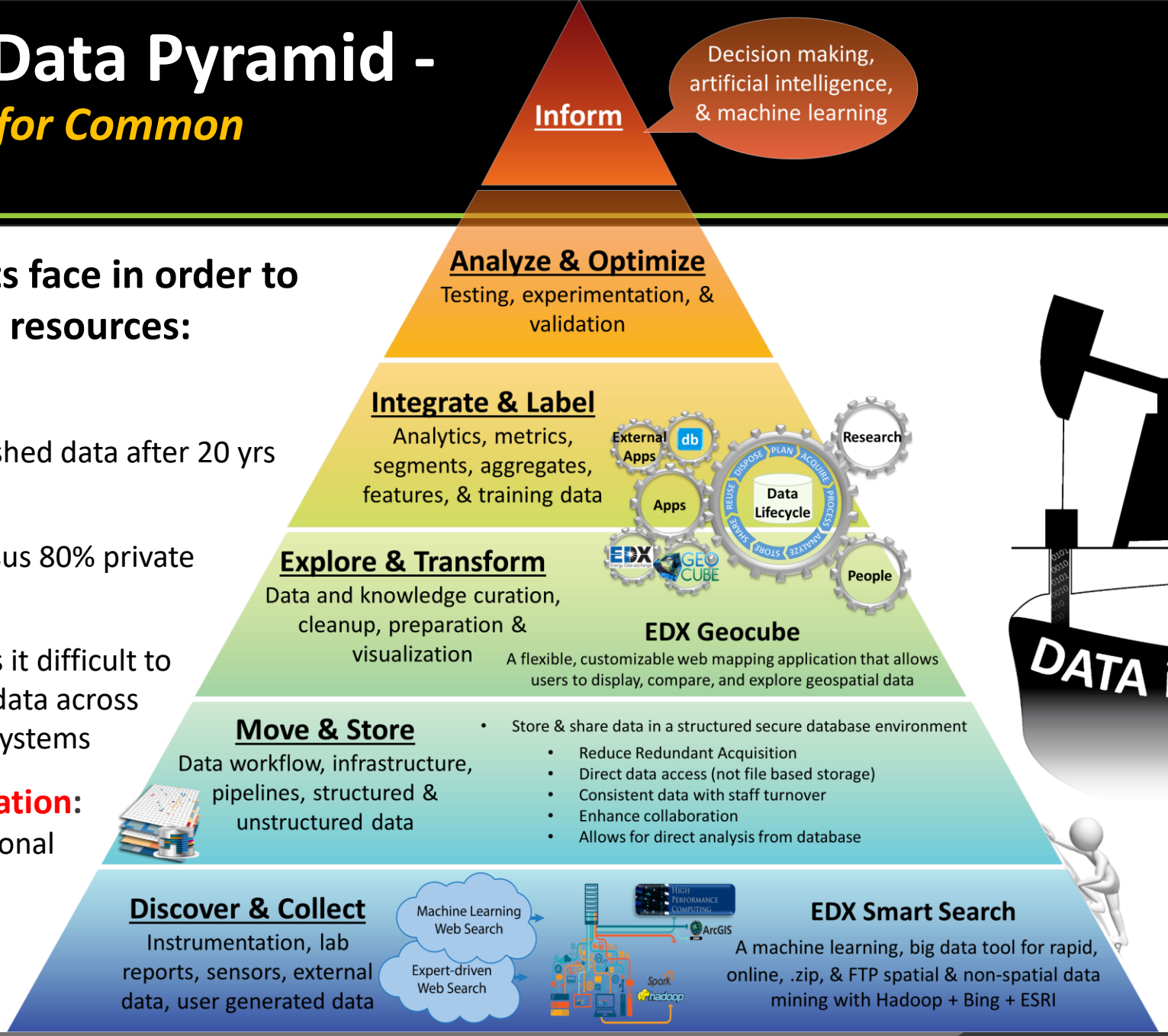
**Challenges scientists face in order to effectively use data resources:**

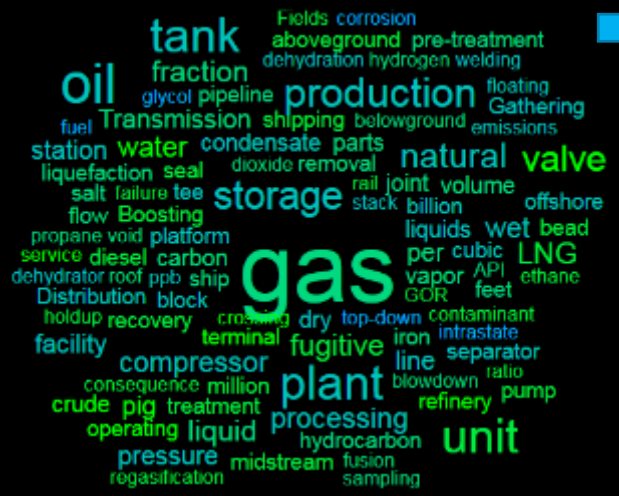
**Data Access:**
~80% loss of published data after 20 yrs

**Data Discovery:**
20% public data versus 80% private

**Data Interoperability:**
Large variety of data makes it difficult to create, exchange, and use data across different applications and systems

**Data Analytics & Visualization:**
Require advanced computational capabilities, algorithms, and large data stores to analyses these data




EDX™ Energy Data eXchange

NATIONAL ENERGY

**Converted Search Terms & Phrases into an Open O&G Spatial Database**

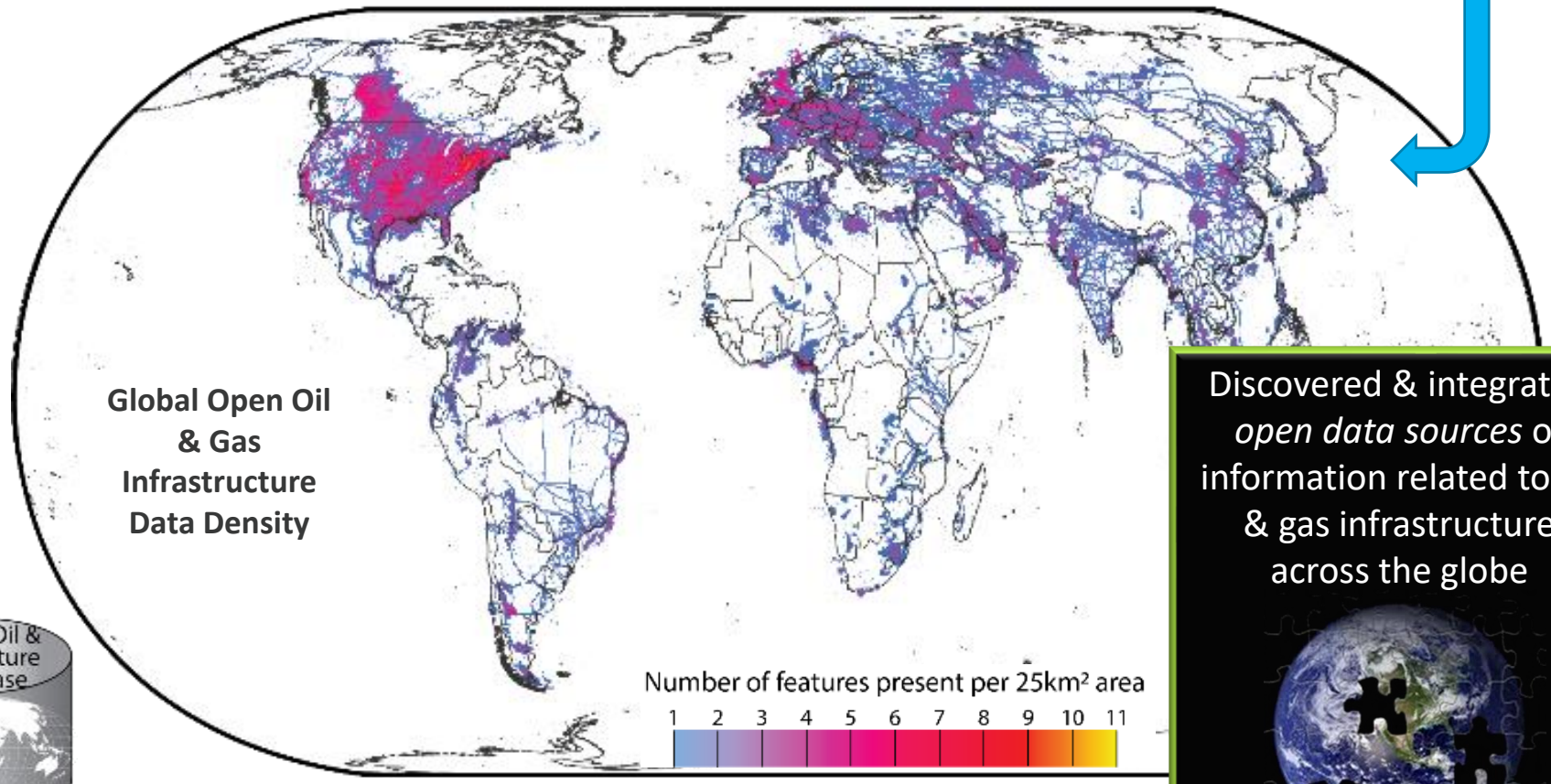
***Machine Learning Semi-Automated Approach-***

A tool that scans "seed" resources and identifies relevant keywords, then crawls the web and parses the data for integration


**4 months period of performance**

**Acquisition of disparate data:**

- **>800 datasets**
- **>4 Millions of features**




Discovered & integrated *open data sources* of information related to oil & gas infrastructure across the globe

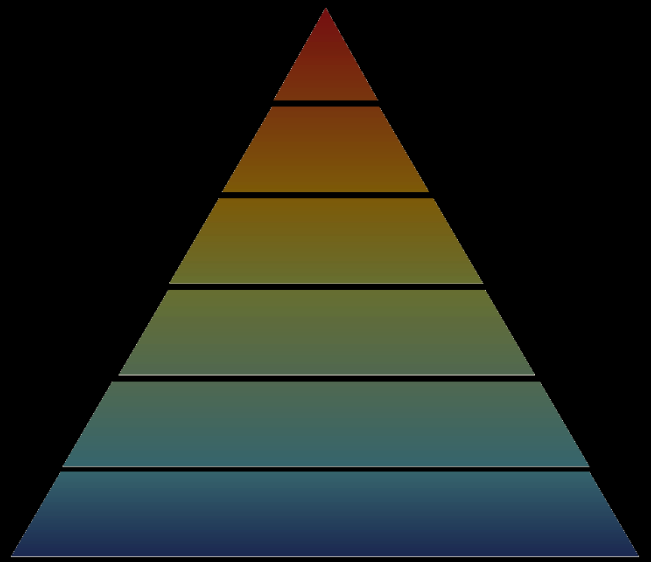


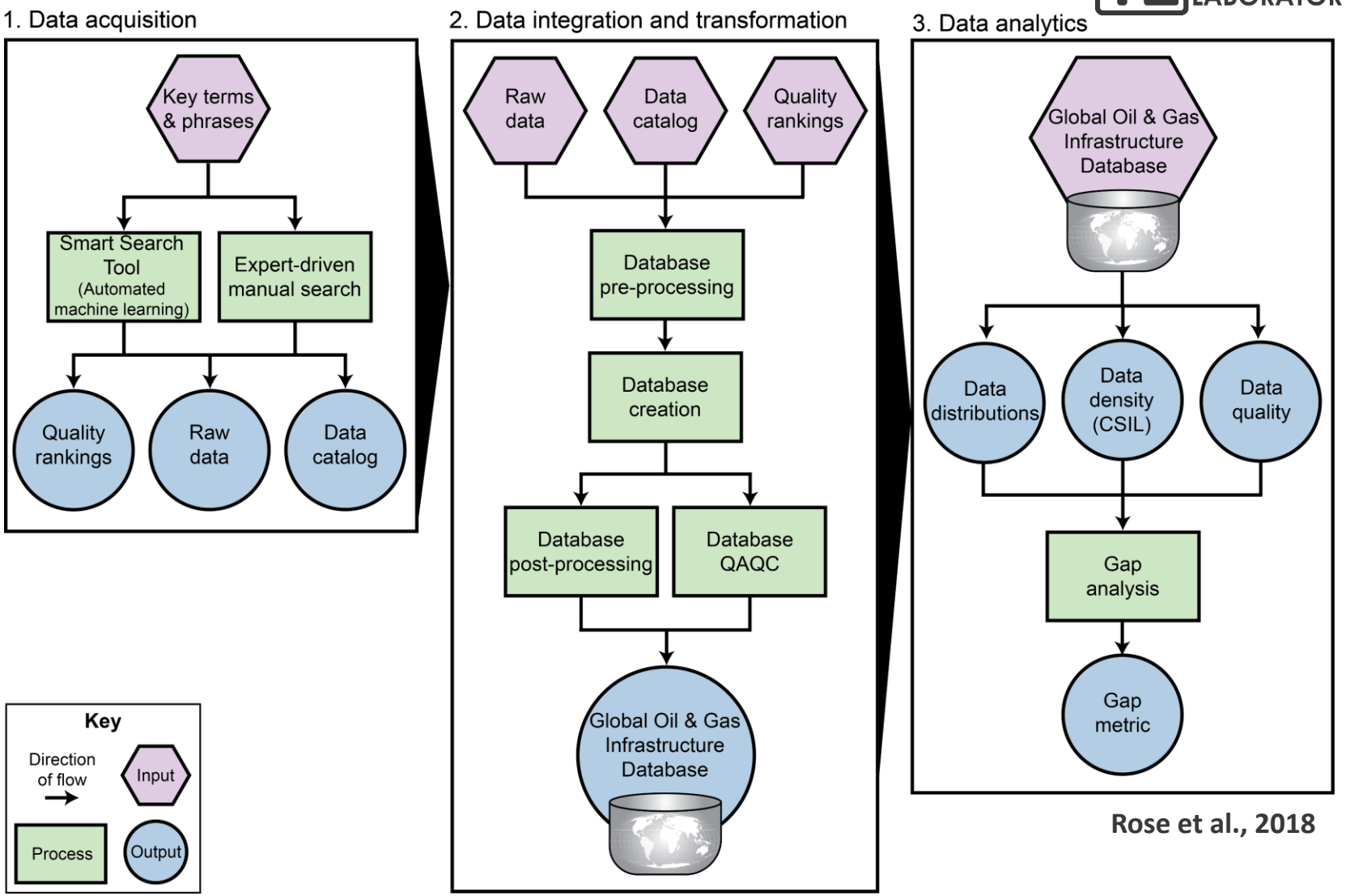

*https://edx.netl.doe.gov/dataset/global-oil-gas-features-database*

# Defining a strategy up front

Steps for data:

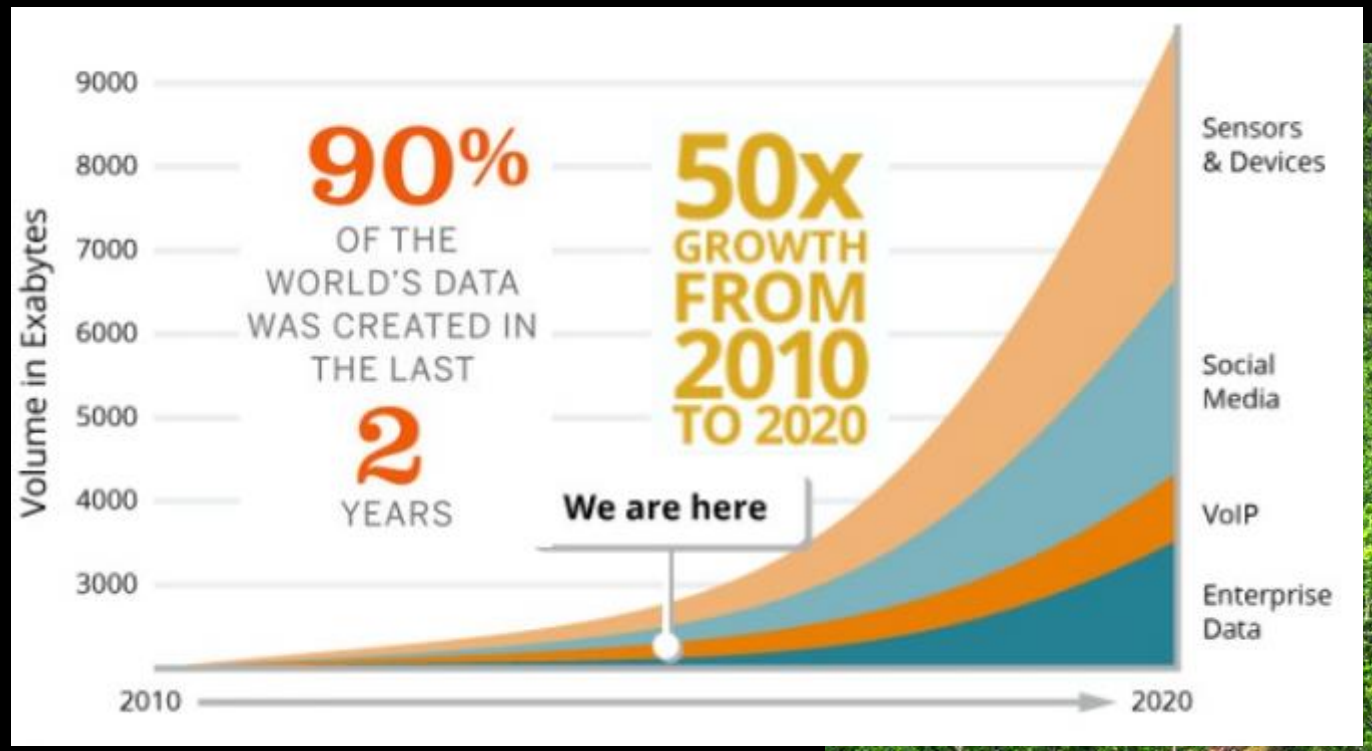
1. Acquisition,
2. Integration & transformation, and
3. Analytics

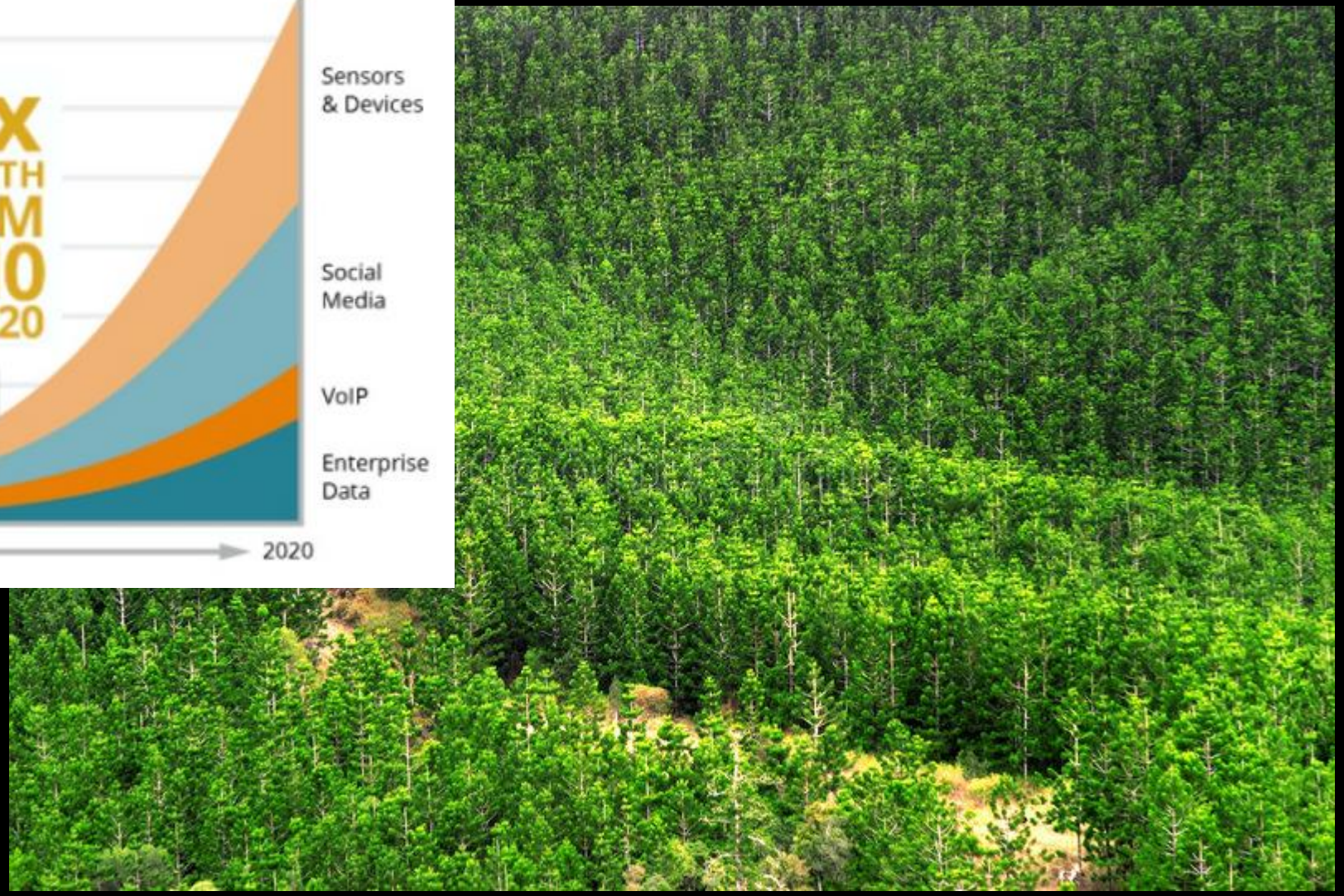





Rose et al., 2018

# As access to open, authoritative data increases science driven analyses face challenges to **efficiently find**, **integrate** and **use** these resources

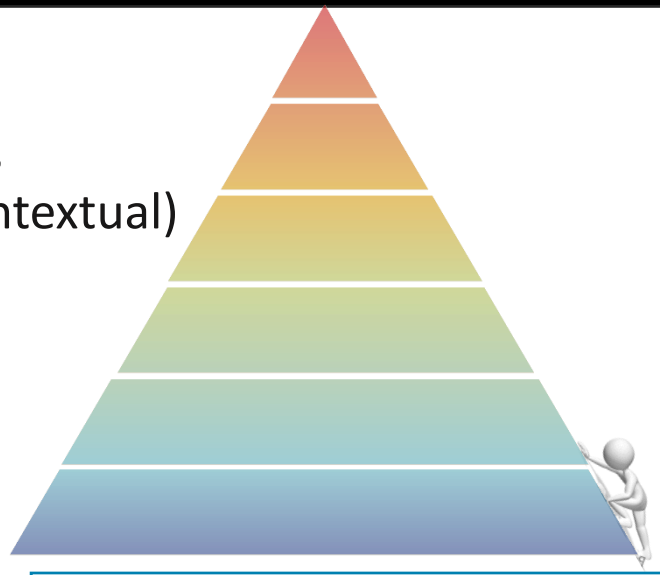


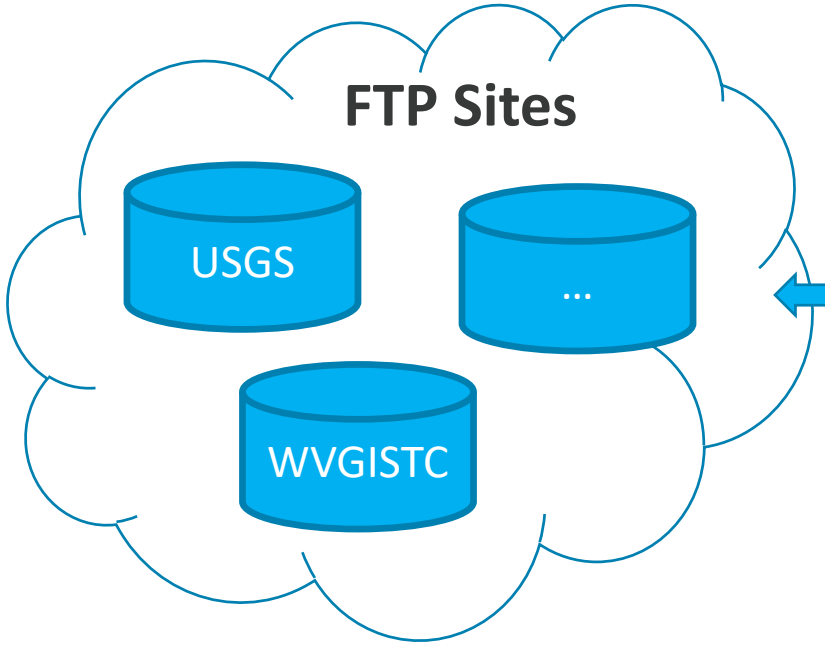



- Volume, variety, and velocity of data online is growing… exponentially
- How will you parse the tree from the forest?

# Use Case: FTP Data Mining: Hadoop + EDX

- **Problem:**
  - Need to search data in FTP silos (millions of files, spatial and contextual)
- **Solution:**
  - Index FTP silos using Hadoop






**Middleware**

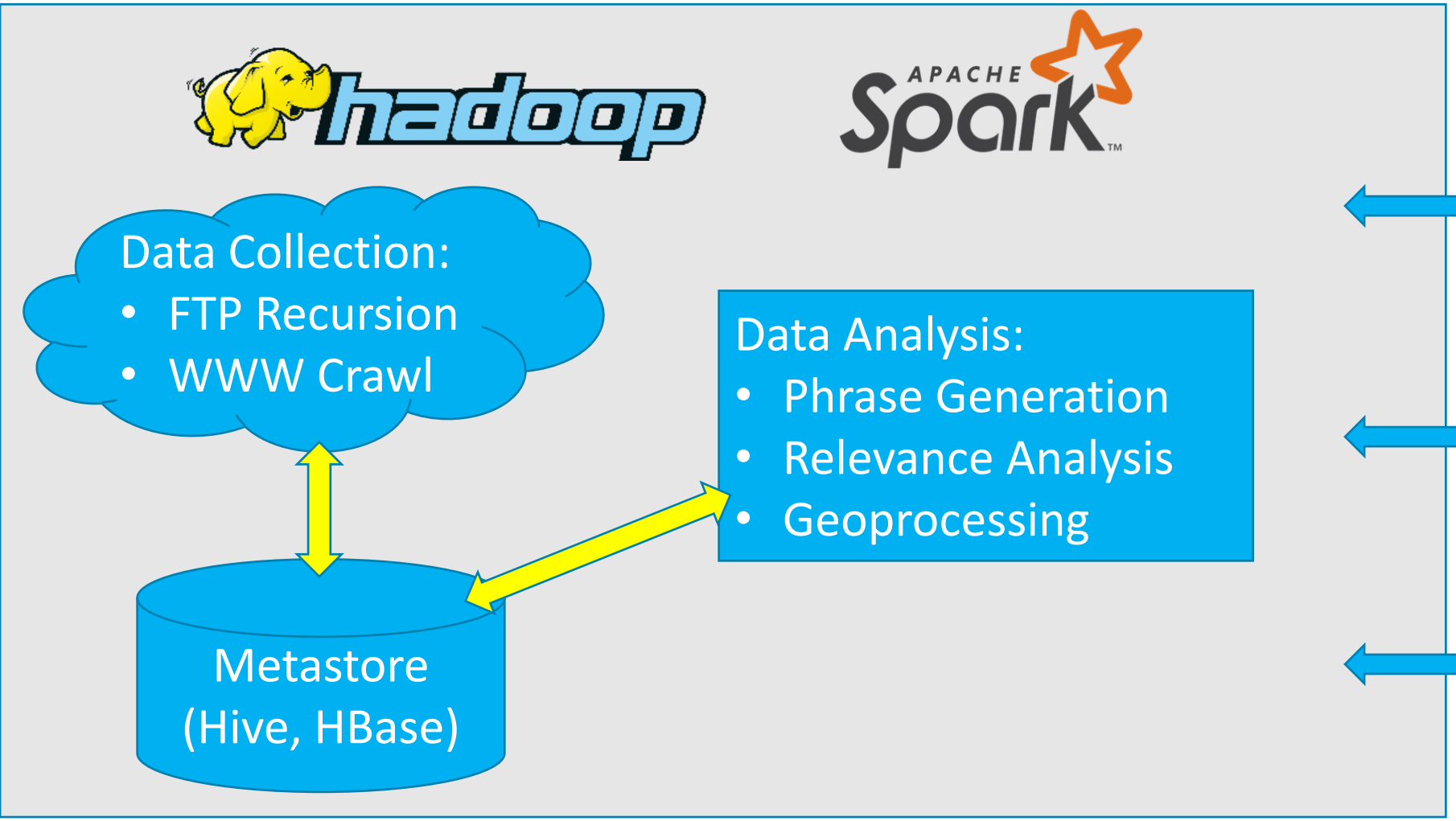
hadoop

APACHE Spark

**Client**

ESRI ArcGIS

# NETL's Big Data Discovery Ecosystem (To Date)



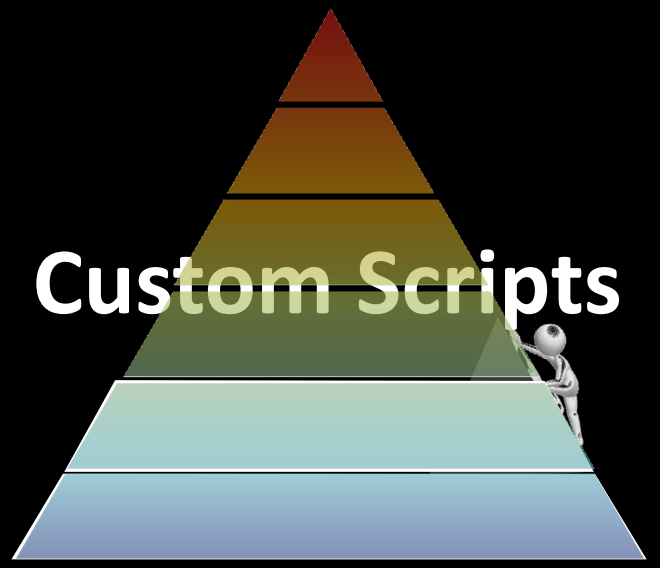

## Data Mining Clients

EDX™
A Product of NETL

ArcGIS®
ESRI

HUE

# Custom Scripts



Python source scripts used to create, translate, and integrate the GOGI geodatabase.

- Used to check and remove for duplicates
- Conduct language translation to English
- Project spatial data
- Generate updated versions of the geodatabase

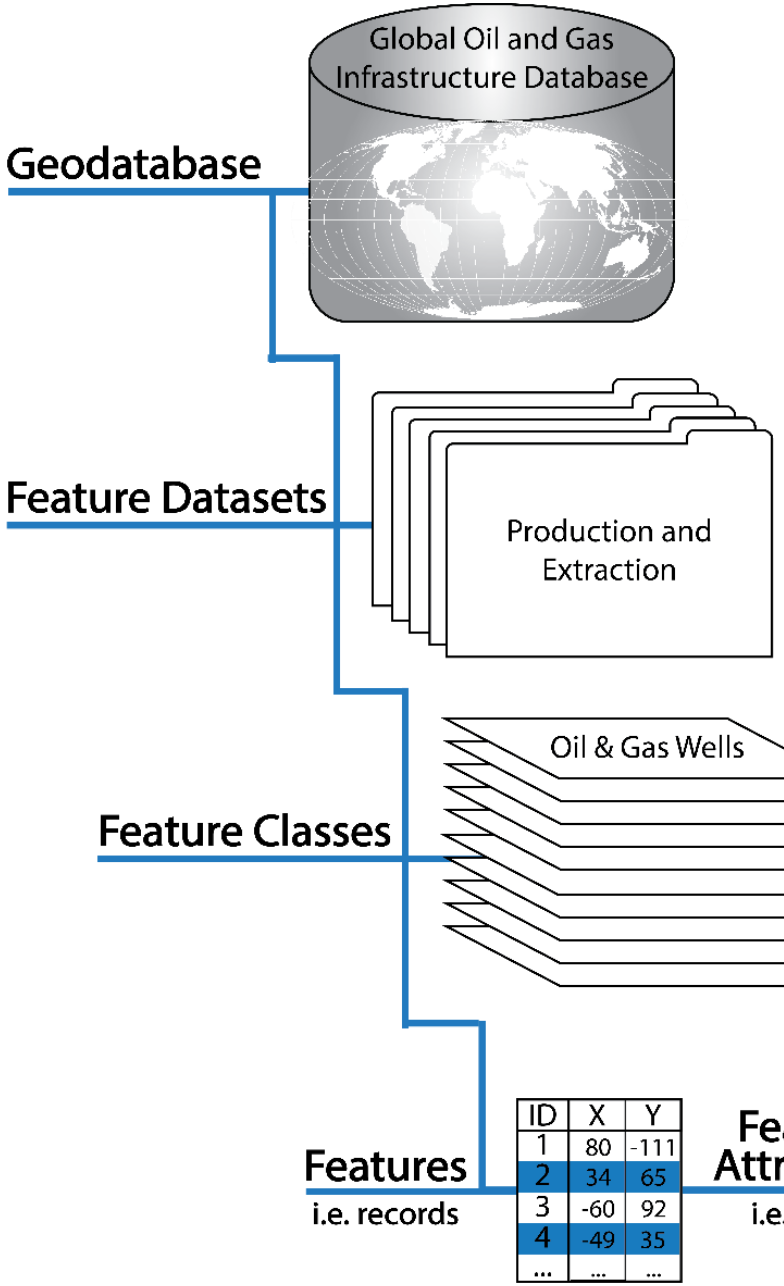
```python
#This script will scan through a geodatabase and remove points with identical locations but keep the point with thei higher Quality Rating
print("Loading arcpy please be patent...")
import arcpy
import numpy
from arcpy import env
arcpy.env.overwriteOutput = True
arcpy.env.workspace = "C:\\temp\\New_Folder\\Final\\V6_Global_Oil_and_Gas_DB_fixpipe.gdb"
datasets = arcpy.ListDatasets('*','Feature')

fclass = []
allfc = []
selAtt = "TempFC_FEAT_SEQ"

for dataset in datasets:
    fclass.append(arcpy.ListFeatureClasses(feature_dataset=dataset))
for featureclass in fclass:
    for feature in featureclass:
        allfc.append(feature)
for fc in allfc:
    print "Processing... ",fc
    arcpy.MakeFeatureLayer_management(fc,"temp_layer")
    rows = arcpy.UpdateCursor("temp_layer", '', '', '', selAtt+" D") #use this code to sort decending
    arcpy.AddField_management("temp_layer", "Remover", "SHORT", "", "", "")
    aveHigh = 0.0
    rowPrev = long(0)
    for row in
        if row.
            if
```

```python
####################Universal attribute Translator
#This script lists the attributes of a selected data table and then translates the selected attributes from non-English to english using the Google Translate API
#to use theis script you must create an account with google and generate an API Key
import arcpy
import os
import unicodedata

#generate a google api key and place it here
googleapikey = 'Place google api key here'

#os and file setup
gdbFiles = []
for path, dirs, files in os.walk(current_file_dir):
    for f in dirs:
        if f.endswith(".gdb"):
            gdbFiles.append(f)
for i in range(len(gdbFiles)):
    print i, ")    "+gdbFiles[i]
#ask user which GDB to process
numselect = int(raw_input("Ty
catalog = current_file_dir+"\
print catalog+" selected"

arcpy.env.workspace = catalog
datasets = arcpy.ListDatasets
#List available datasets
for i in range(len(datasets))
    print i, ")    "+datasets

#ask user for input to datase
numselect2 = int(raw_input("T
catalog2 = catalog+"\\"+datas
print catalog2+" selected."
arcpy.env.workspace = catalog
```

```python
##########################################
#This script Reads the data attributes and data location from a CSV file and Populates a Geodatabase with the spatial files found.
#It also includes the metadata from the CSV and populates the collected data with these attributes.
#Also a log file will be created that lists any errors that the script encounters while running.
####################REQUIREMENTS
#User must have ARCPY installed
#Makse sure the following files are in the same
#folder as this script:
#feature_classes.csv
#WGS1984.prj
#a CSV file based off the edf catalog
##########################################


import csv
import os

#os and file setup
current_file_dir = os.path.dirname(__file__)
print "Opening log.txt file..."
features = current_file_dir+"\\feature_classes.csv"
sr = current_file_dir + "\\WGS1984.prj" #This is a projection file that will help move all disparate data sets to a single spatial reference

#Loop to list csv files in current directory

csvFiles = []
for path, dirs, files in os.walk(current_file_dir):
    for f in files:
        if f.endswith(".csv"):
            csvFiles.append(f)
```

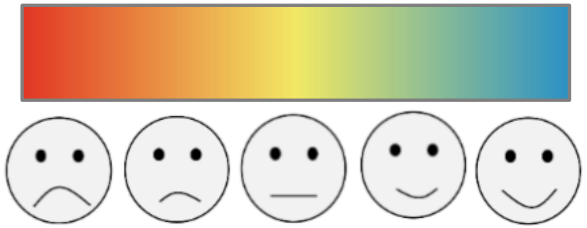
# Not all data is equal

- Quality & quantity varies
- But understanding uncertainty and gaps in data is important for data driven analytics, stakeholder decision making, and other needs




| Feature Datasets | Example Feature Classes |
|---|---|
| Transport | Ports, Railways, Pipelines |
| Facilities/ Installations | LNG, Power Plants, Processing Plants, Refineries, Stations, Storage, Terminals |
| Production/ Extraction | Oil and Gas Fields, Platforms, Well Pads, Underground Storage, Mines, and Wells |
| Geology | Sedimentary Basins |

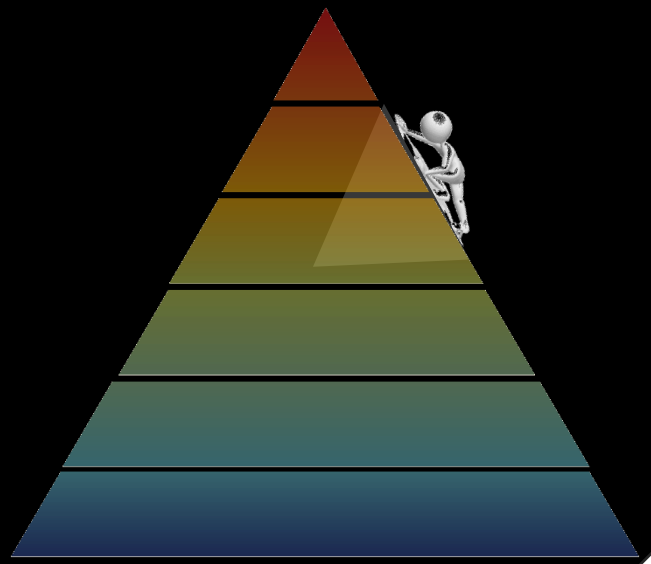
Overall quality score



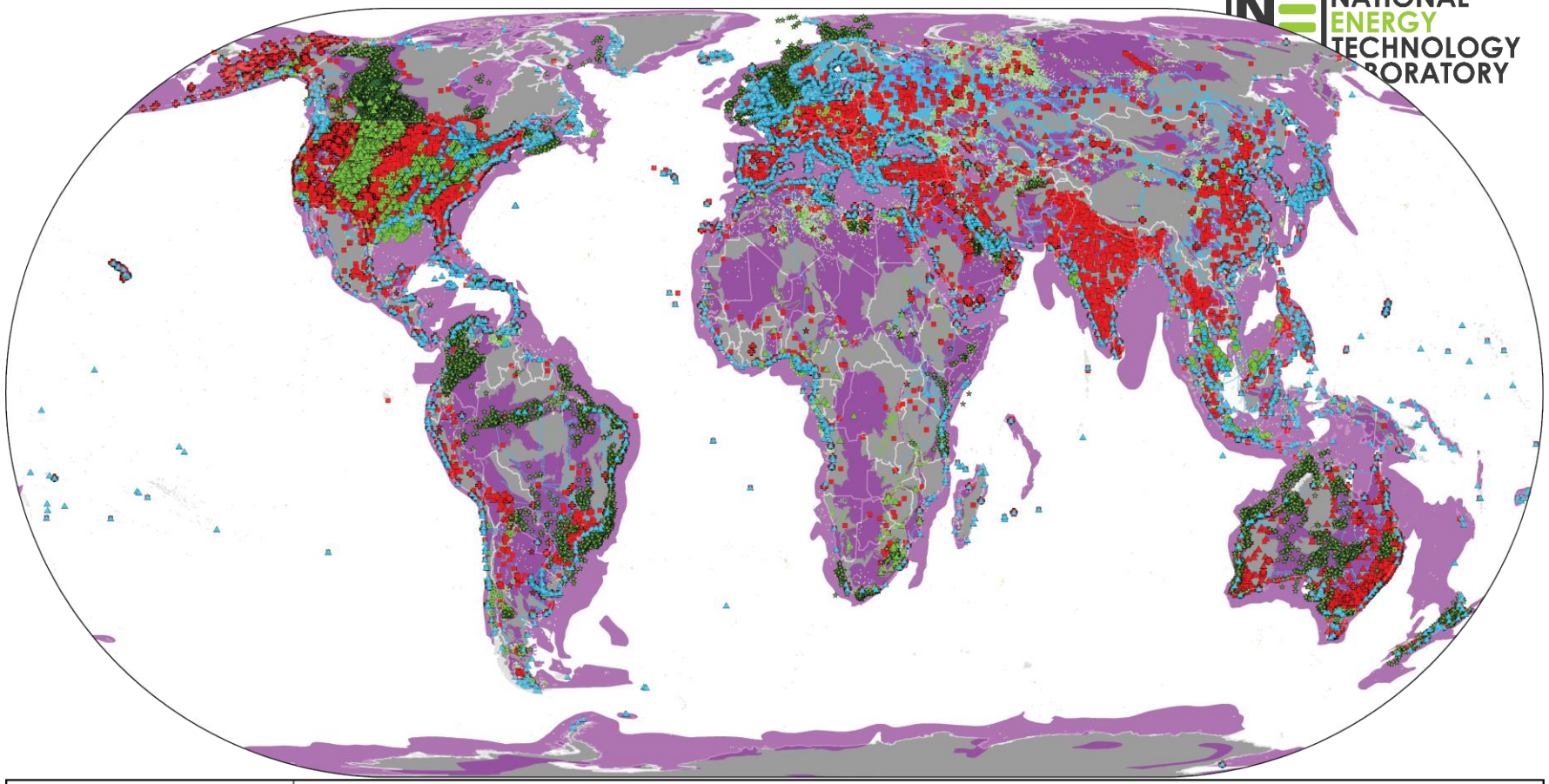
**Ranking data quality by source, spatial and temporal features helps with analytics**

# Visualization

Tip of the pyramid



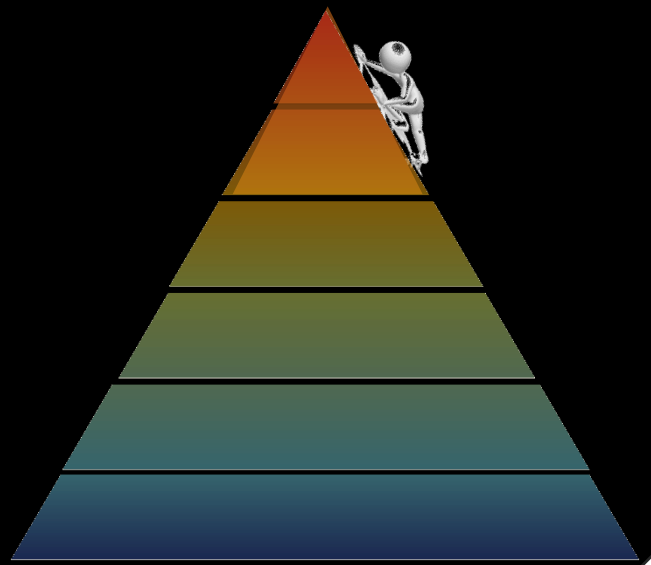
Rose et al., 2018

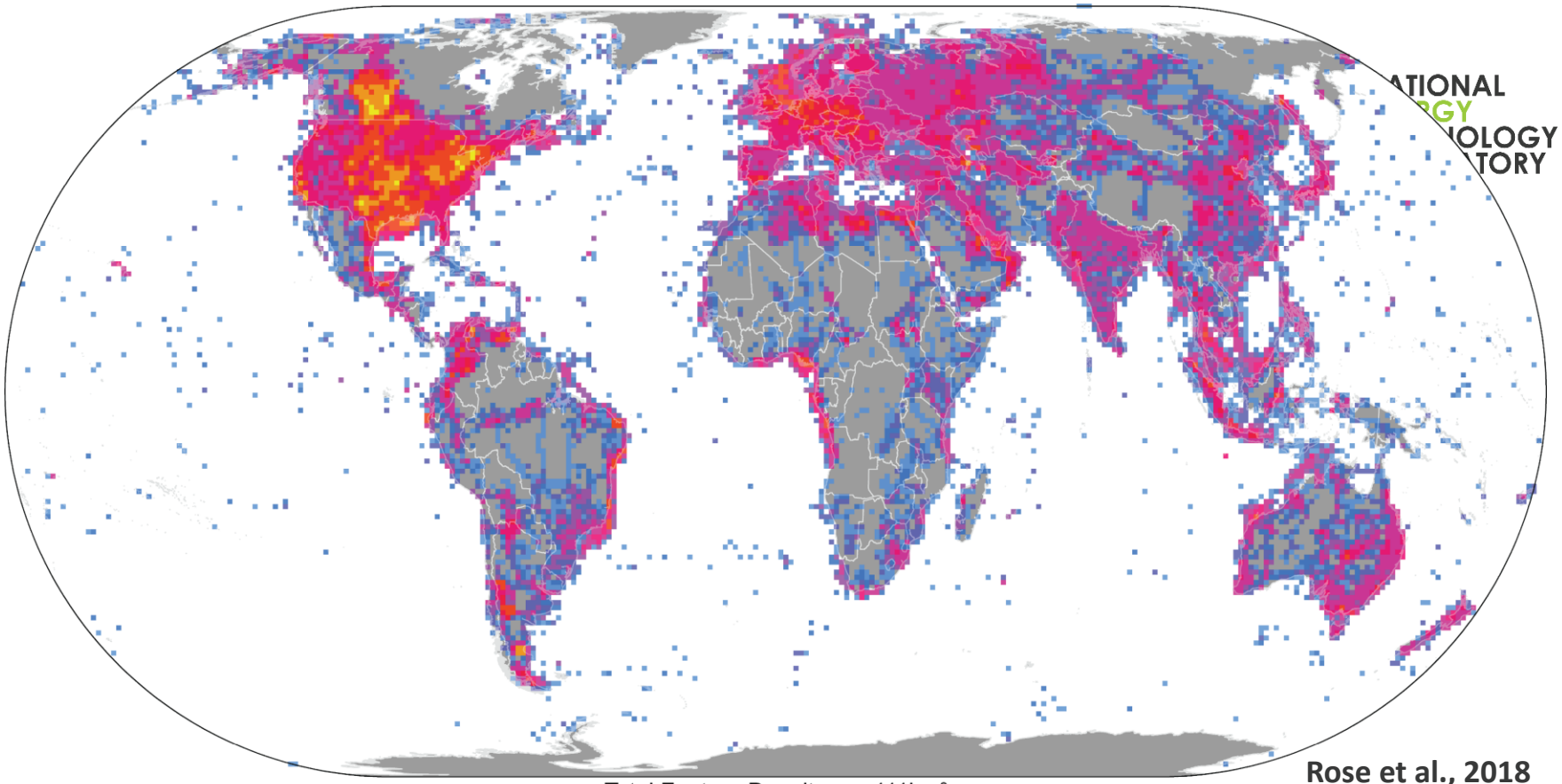


| Feature Datasets | Feature Class Symbology and Total Number of Features | | |
|---|---|---|---|
| Transport | Ports n = 3,702 | Railways n = 280,734 | Pipelines n = 94,448 |
| Facilities and Installations | Liquid Natural Gas (LNG) n = 329<br>Power Plants n = 14,097 | Processing Plants n = 1,922<br>Refineries n = 2,272 | Stations n = 13,876<br>Storage n = 26,103 |
| Production and Extraction | Fields n = 25,236<br>Platforms and Well Pads n = 9,845 | Underground Storage n = 3,731<br>Mines n = 51,602 | Wells n = 736,476<br>Well Density 25km² Grid (represents a total of n = 3,544,809 wells) 1, 2, 3–10, 11–20, 21–50, 51–1,000, 1,001–13,565 |
| Geology | Sedimentary Basins n = 1,046 | | |

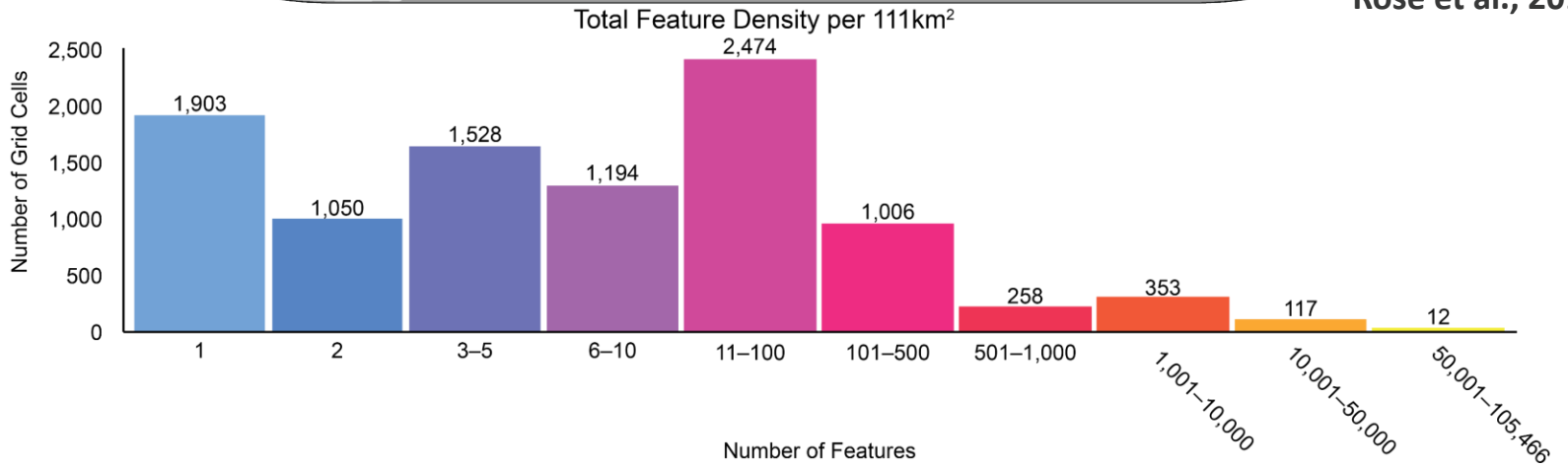
# Analytics

Tip of the pyramid



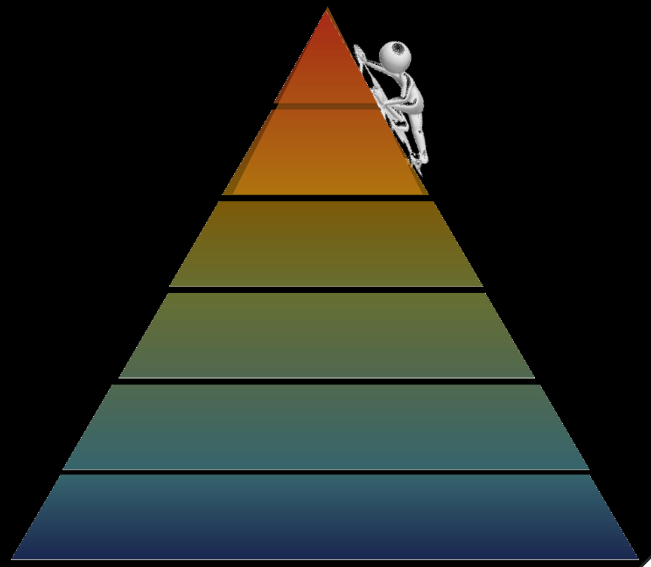


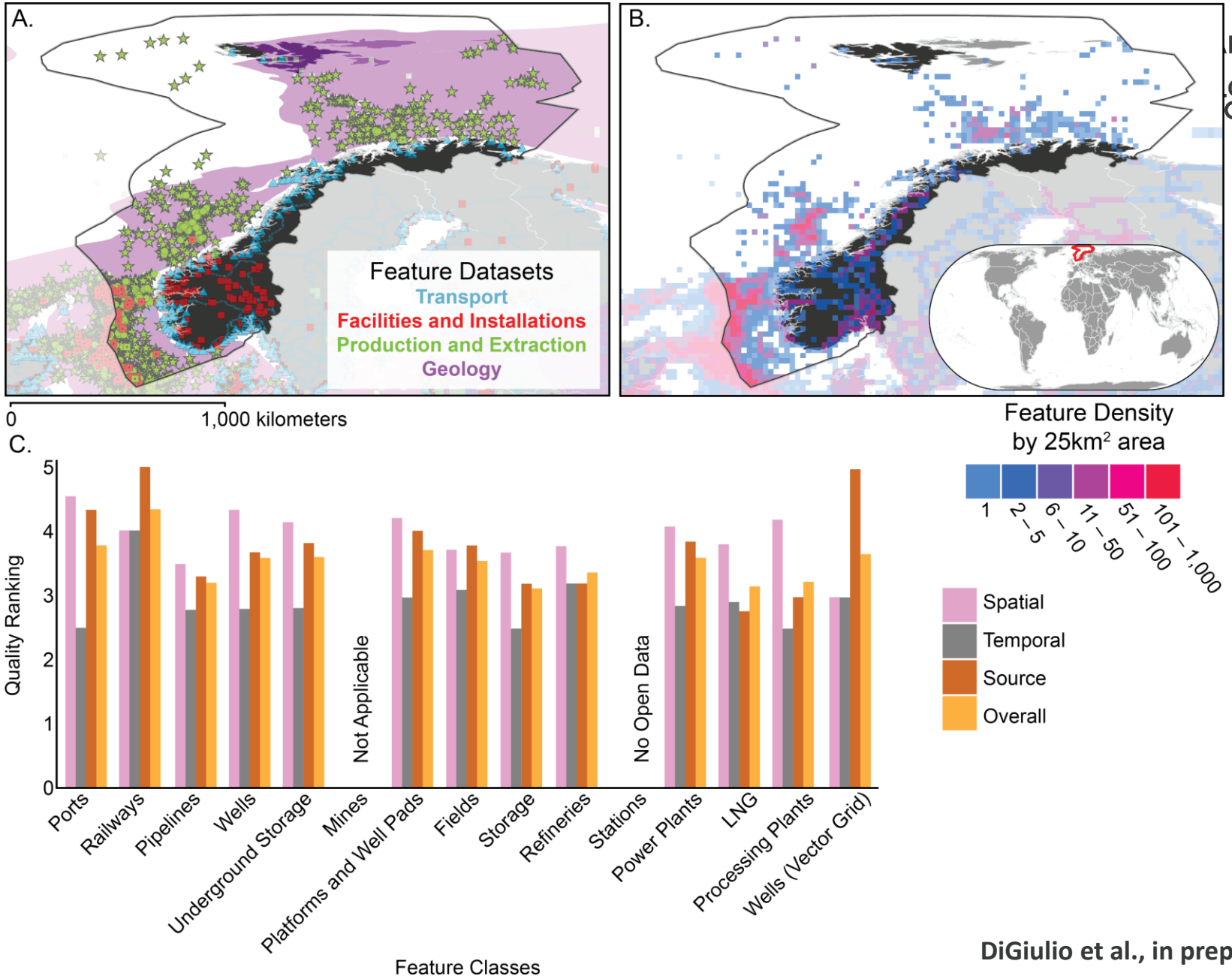
Rose et al., 2018



# Analytics!

Tip of the pyramid






DiGiulio et al., in prep

# Embrace the uncertainty and error in data

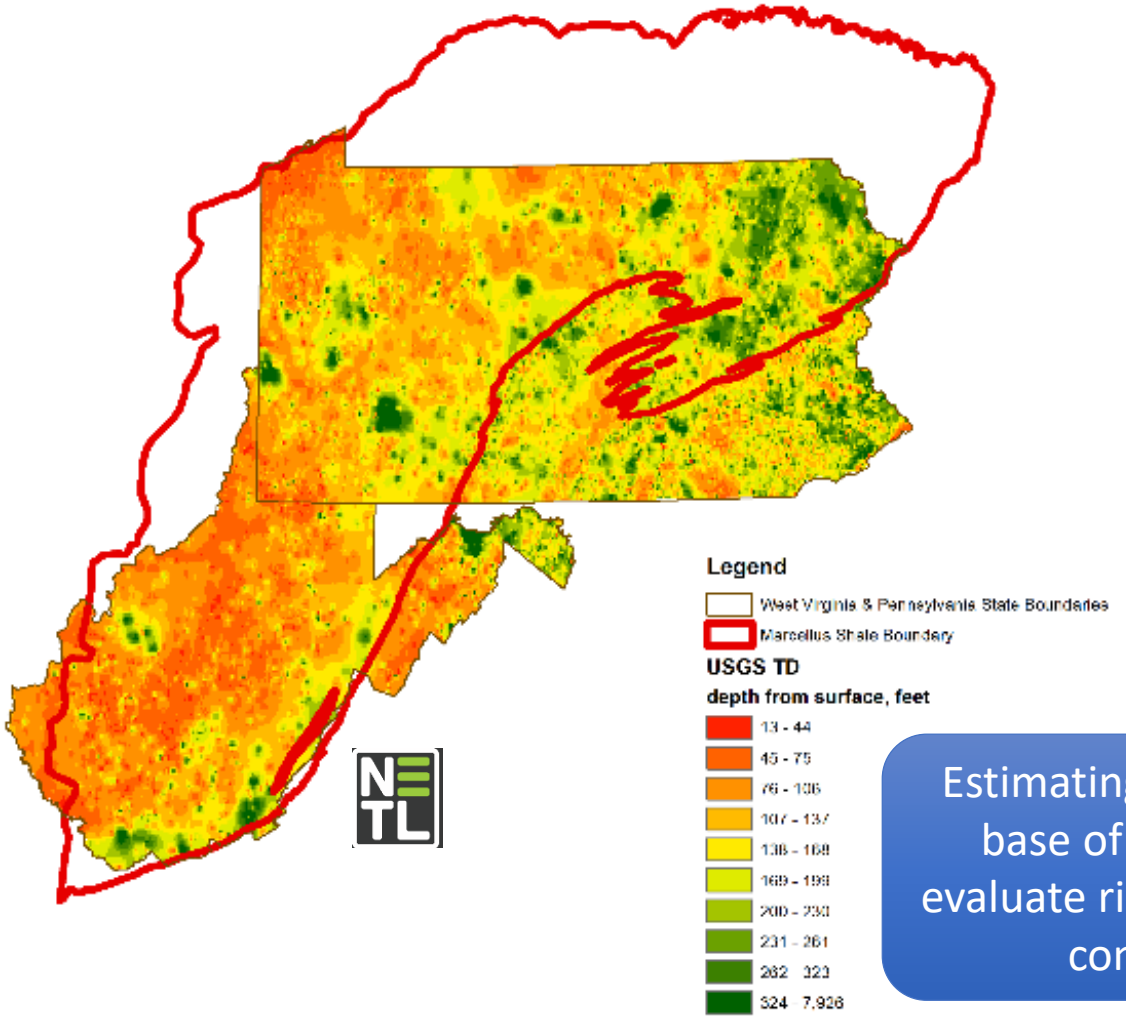
Spatio-temporal data uncertainty information is often **lacking due to difficulties** encountered:

- from the variety of potential sources and definitions,
- visualizing uncertainty, and
- communicating results

Failing to effectively communicate underlying uncertainty can lead to ***false conclusions*** and ***poor decisions*** as well as ***affect the quality*** of current and future research and products

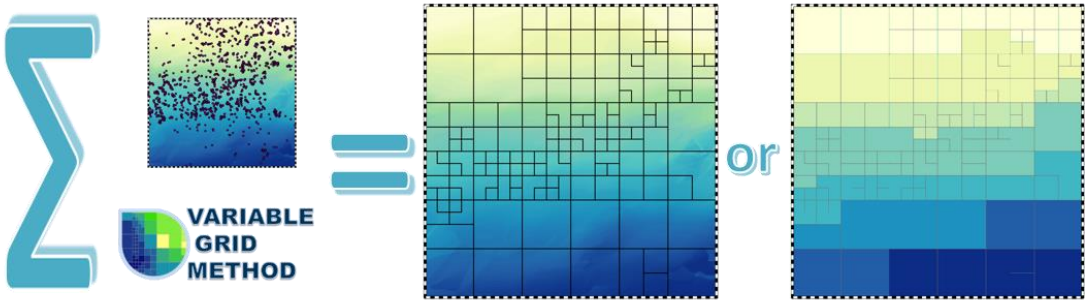
| | | | | |
|---|---|---|---|---|
| 8 PM Sun, Jul 19 | 75° | 75° | Cloudy | 0% |
| 9 PM Sun, Jul 19 | 74° | 74° | Cloudy | 0% |
| 10 PM Sun, Jul 19 | 74° | 74° | Cloudy | 20% |
| 11 PM Sun, Jul 19 | 74° | 76° | Mostly Cloudy | 10% |
| 12 AM Mon, Jul 20 | 74° | 75° | Mostly Cloudy | 5% |
| 1 AM Mon, Jul 20 | 73° | 75° | Partly Cloudy | 5% |
| 2 AM Mon, Jul 20 | 73° | 75° | Cloudy | 5% |
| 3 AM Mon, Jul 20 | 73° | 74° | Cloudy | 5% |
| 4 AM Mon, Jul 20 | 73° | 74° | Cloudy | 5% |
| 5 AM Mon, Jul 20 | 73° | 74° | Cloudy | 5% |
| 6 AM Mon, Jul 20 | 72° | 74° | Cloudy | 5% |

# Example, why uncertainty matters




Estimating the depth to the base of groundwater to evaluate risks of groundwater contamination

Allow for simultaneous visualization & quantification of spatial data and uncertainty



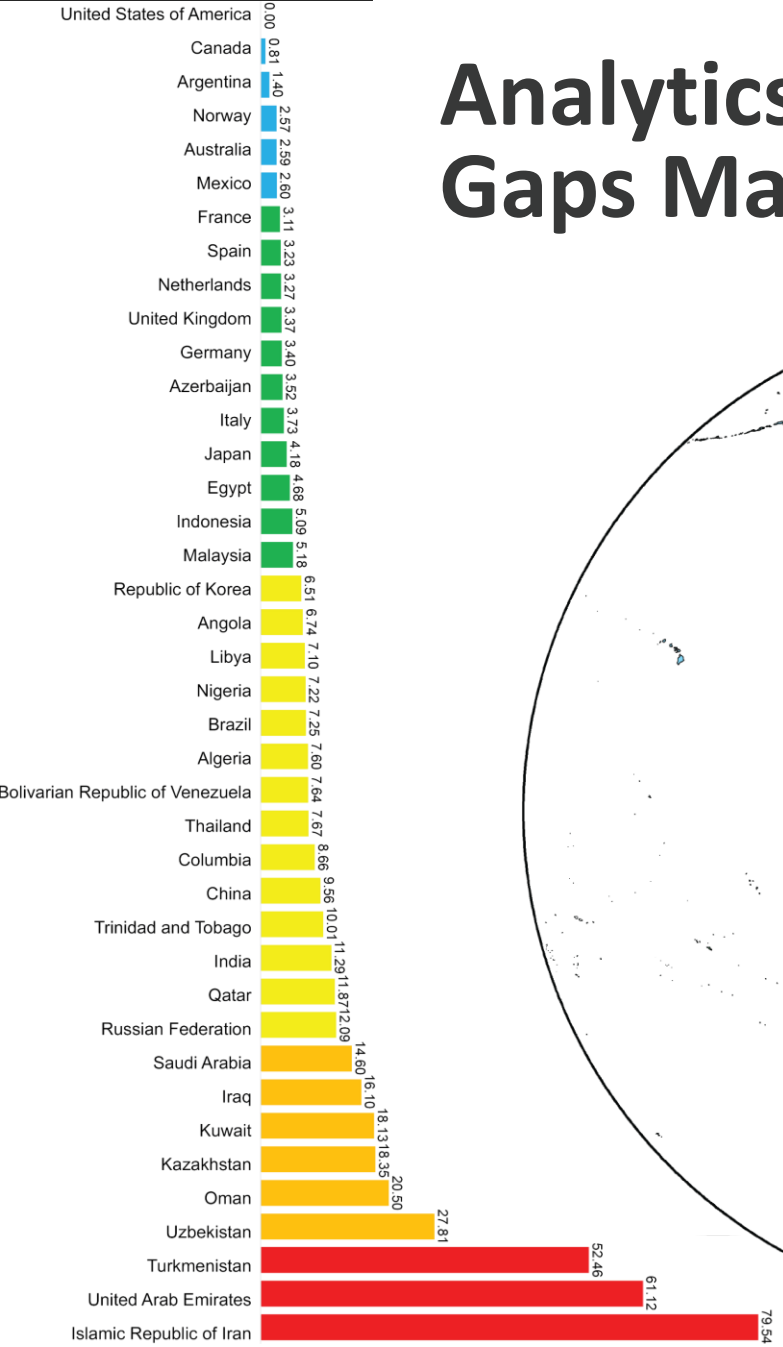

Bauer & Rose, 2015

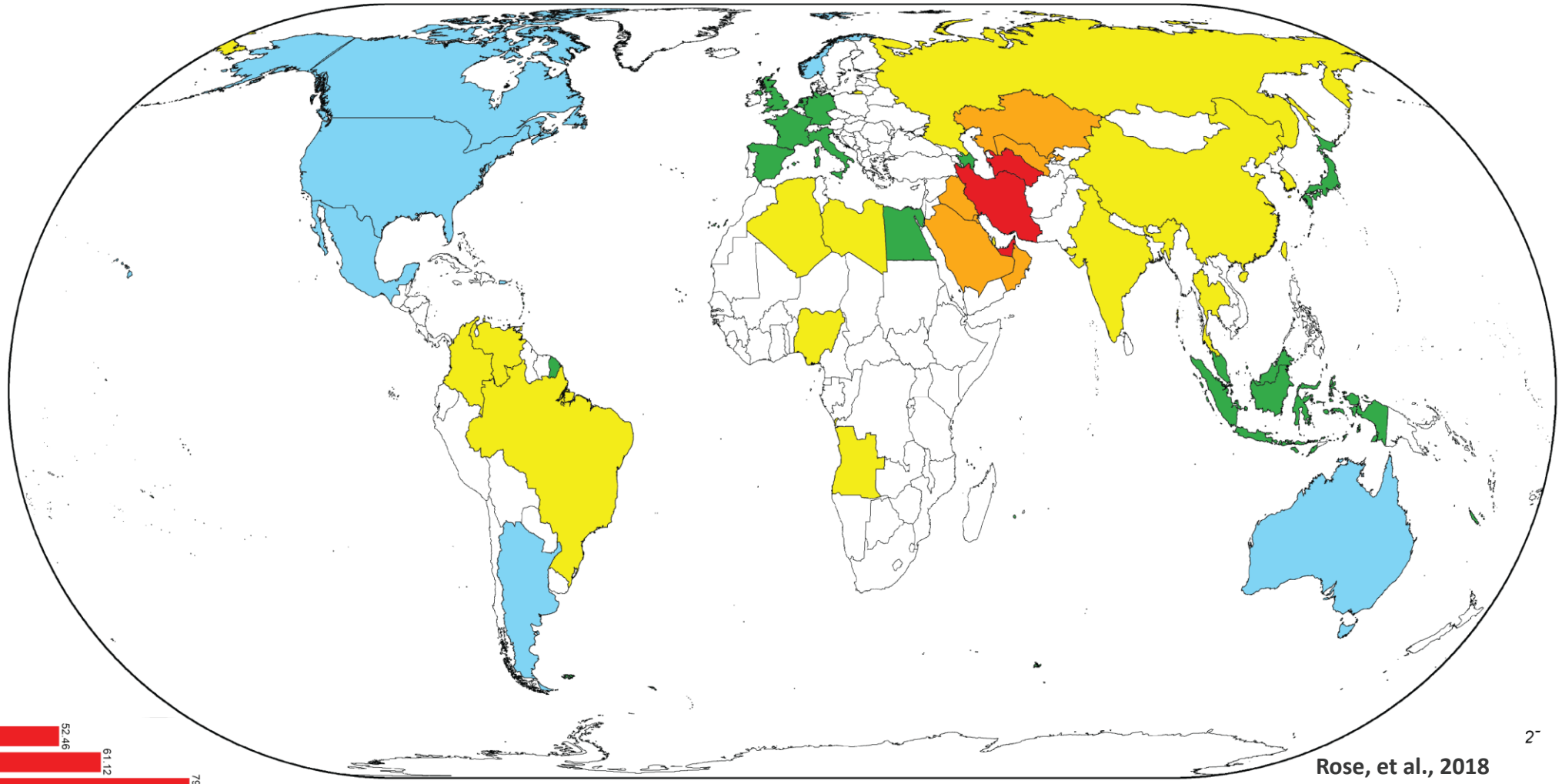
*Communicate data (via colors) and uncertainty (via grid cell size)*

**Uncertainty Viz/Quant for Spatio-Temporal Analyses Can Improve:**

- Resources evaluations
- Impact assessments
- Understanding trends in the data
- Calculating Project Feasibility
- Identifying Knowledge Gaps

# Analytics – Gaps Matter



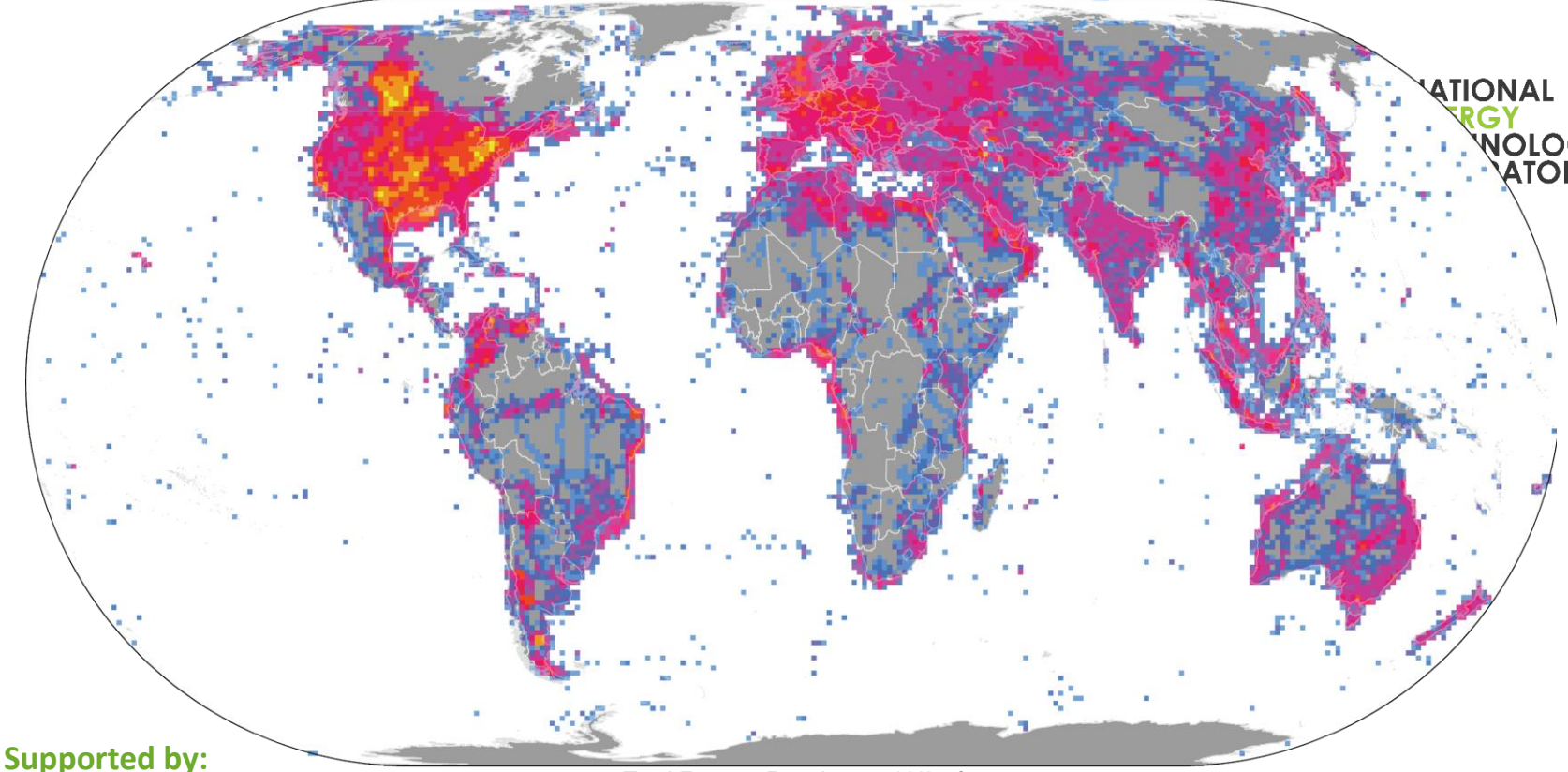



Rose, et al., 2018

# Global Oil and Gas Infrastructure (GOGI)

- 4 month performance period
- Acquisition of disparate data by country, region, & continent totaling:
  - ~800 datasets
  - 4 million+ of features
  - Attributes some regions/features

EDF used compiled database to inform decision-making about methane emissions



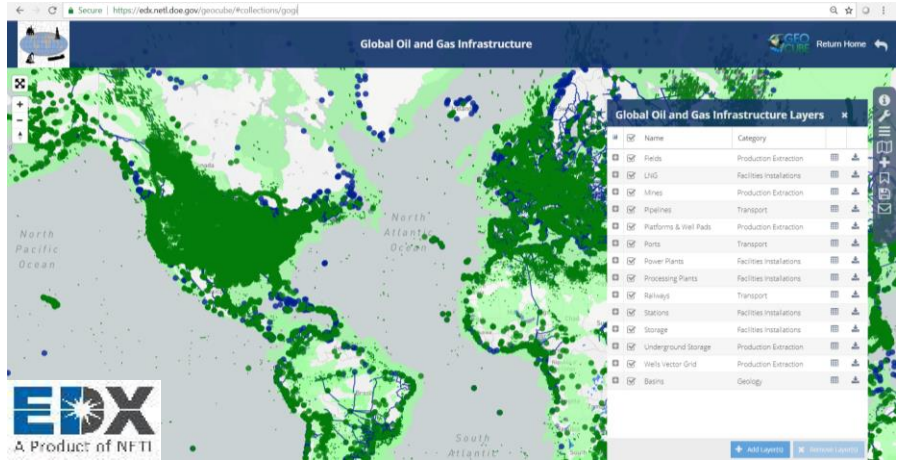
Total Feature Density per 111km$^2$

Supported by:

EDF ENVIRONMENTAL DEFENSE FUND
Finding the ways that work

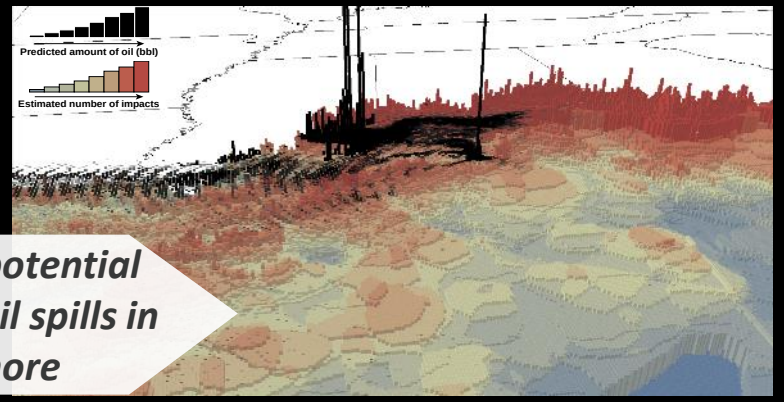
**Public Products:**

- Technical report
- GOGI Database
- Web mapping application
- Journal pub in prep: Digiulio et al., in prep, *Elementa*

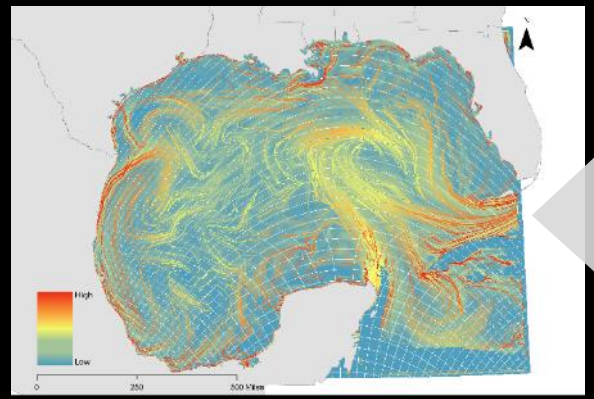


# Array of Applications

- Adapted to work with other approaches, tools, and models
- Many data formats
- Multi-scale

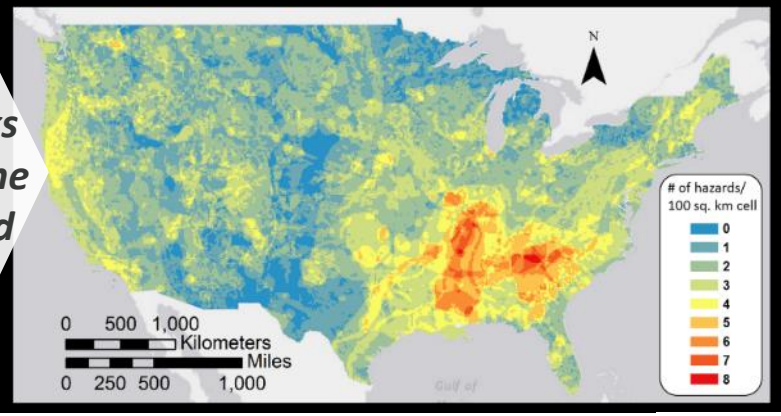



*Evaluating potential impacts of oil spills in the US Offshore*

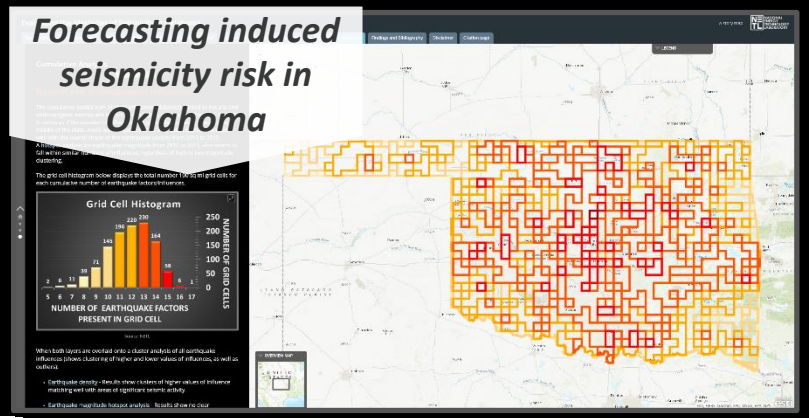



*Predicting where oil is likely to go based on oceanographic data*

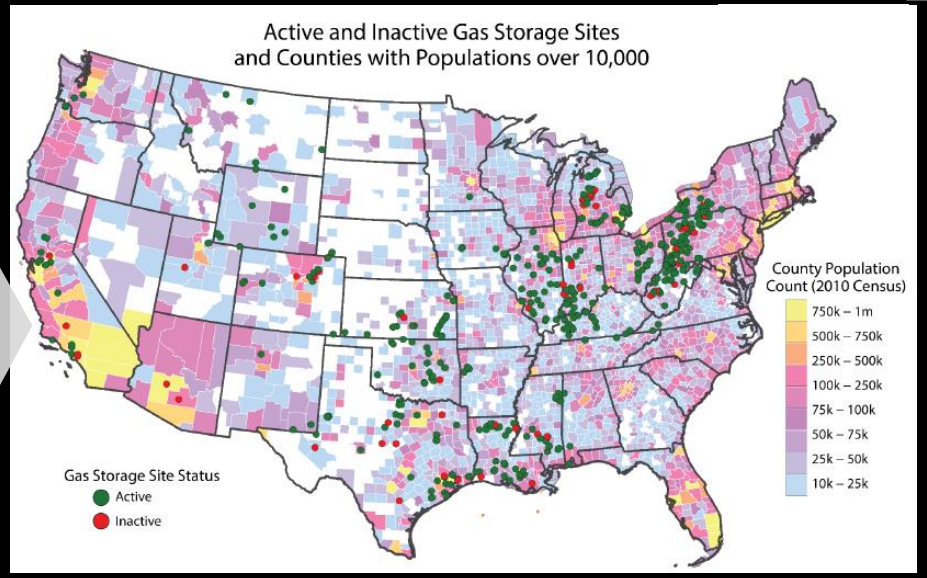
*Quantifying pipeline risks from extreme weather and geohazards*

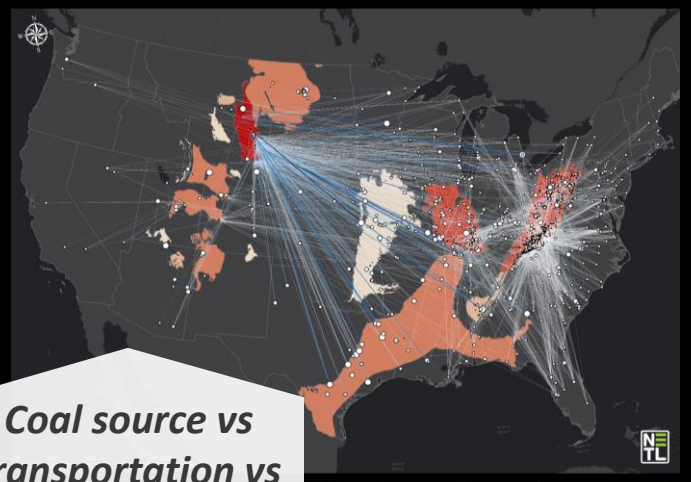



*Forecasting induced seismicity risk in Oklahoma*




*Characterizing gas storage vs population trends*



*Coal source vs transportation vs consumption*

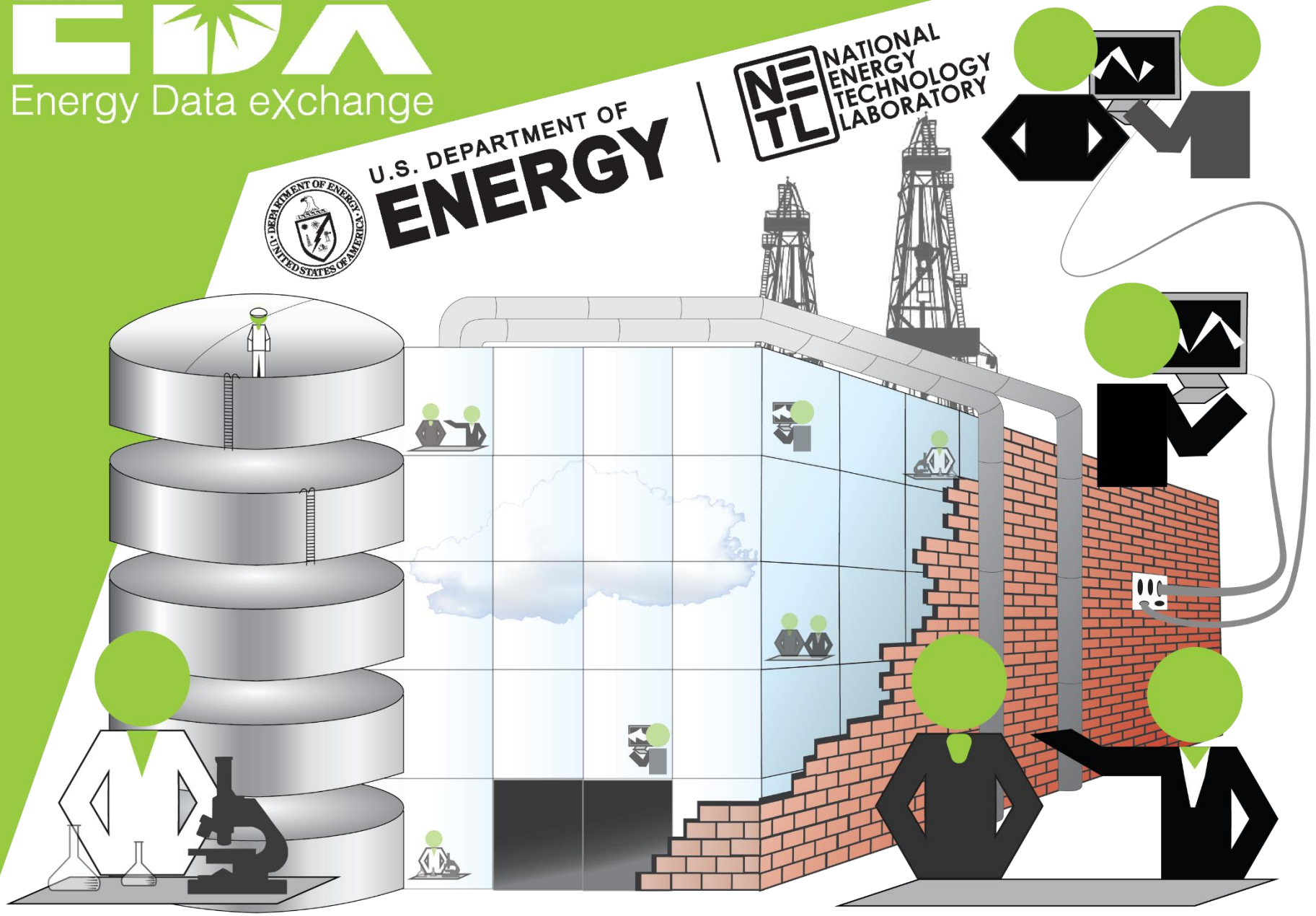
# 2018 Approach to Publishing R&D

https://edx.netl.doe.gov

EDX
Energy Data eXchange

# A Virtual Library & Laboratory for Energy Science

- Virtualizing team analytics
- Continued innovations to connect DOE FE affiliated researchers to online resources (tools, data, etc)
- Publishing data products from FE R&D for public reuse
- A virtual lab/user facility for FE R&D teams collaborate, analyze, and utilize data
- In development since 2011




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

# Numerous Data Repositories

## *Offers opportunities and challenges*

BLOG WRITE FOR DATAQUEST

https://www.dataquest.io/blog/free-datasets-for-projects/

13 SEPTEMBER 2016 / PROJECT

### 18 places to find data sets for data science projects

*This is the fifth post in a series of posts on how to build a Data Science Portfolio. You can find links to the others in this series at the bottom of the post.*

If you've ever worked on a personal data science project, you've probably spent a lot time browsing the internet looking for interesting data sets to analyze. It can be fun sift through dozens of data sets to find the perfect one, but it can also be frustrating download and import several csv files, only to realize that the data isn't that interesting after all. Luckily, there are online repositories that curate data sets and (mostly) remove the uninteresting ones.

EDX™ Energy Data eXchange

National Energy Technology Laboratory

NETL's Energy Data eXchange

Home | Search | Contribute | Groups | Portfolios | Tools | Workspaces | My EDX | About | Help | rose

Search Submissions

Spatial Search — Clear

Map data © OpenStreetMap contributors
Tiles by MapBox

Types — Clear Types
- publication 1
- dataset 1

Show More Types

Keywords — Clear Keywords
- gogi 2
- trs 1
- oil and gas ... 1

EDX | EDX (Private) | OpenEI 0 | NGDS 0 | NOAA 0 | Other

Advanced Search Builder

Filter NETL Products:

gogi — Relevance

2 submissions found for "gogi"
Submissions within the EDX system encompass many different forms of data. This includes, but is not limited to, publications, files, apps, tools, or other resources that are data driven resources.

Global Oil & Gas Features Database — NETL
This submission contains a zip file with the developed Global Oil & Gas Features Database (as an ArcGIS geodatabase). Access the technical...
ZIP

Development of an Open Global Oil and Gas Infra... — NETL
This submission contains a technical report describing the development process and visual graphics for the Global Oil and Gas Infrastructure...
PDF

Show Your Work. Share Your Work. Advance Science.
That's OPEN SCIENCE.

# ML, NLP, OCR and other tools to resurrect old data

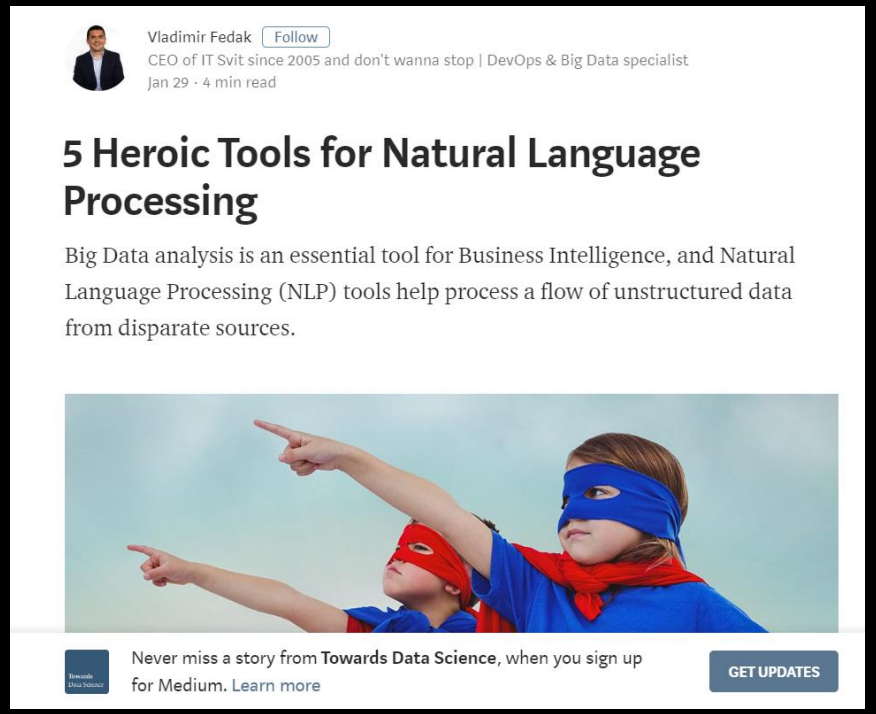
## Traditional Approach to Publishing R&D

http://www.thibaudpoirier...

Journal manuscripts, books, reports, written results

**Data tools can be used to:**

- Mine journal/patents other publications
- Convert tables and graphs back to data
- Gather images for analytics
- Scan and characterize documents



Vladimir Fedak Follow
CEO of IT Svit since 2005 and don't wanna stop | DevOps & Big Data specialist
Jan 29 · 4 min read

**5 Heroic Tools for Natural Language Processing**

Big Data analysis is an essential tool for Business Intelligence, and Natural Language Processing (NLP) tools help process a flow of unstructured data from disparate sources.

Never miss a story from **Towards Data Science**, when you sign up for Medium. Learn more

GET UPDATES

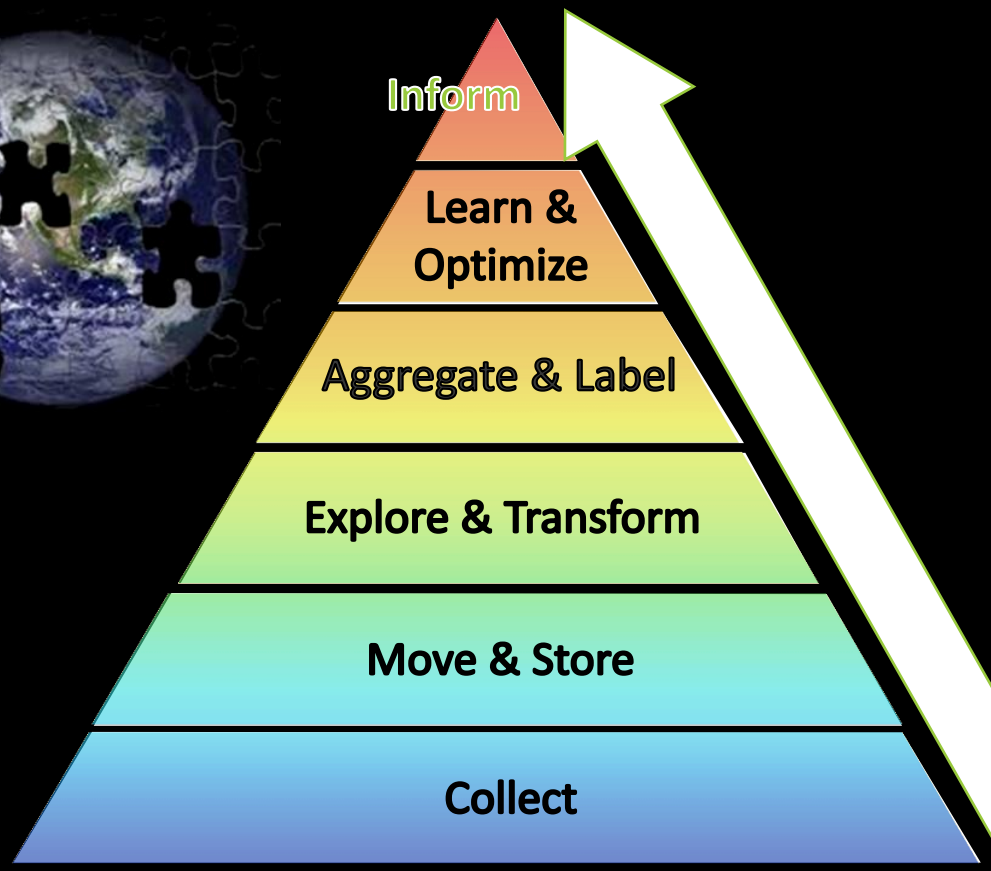
NATIONAL ENERGY TECHNOLOGY LABORATORY

# Data driven science…

## …takes a team

# Additional Data Challenges & Opportunities

- Computing science can help address subsurface systems data challenges
- How do we balance the tug of war between potential of data to innovate vs. stakeholder concerns?
- Is all data equal? What are the data priorities?
  - Fill in the data puzzle one piece at a time…
- If you don't have the data you want, are there proxy data that can fill in the gaps?
- Think about demonstrating need vs value
- Error and uncertainty are important
- Incentives to release data, data citations, journals, and scientific community standards
- Anonymization and other big data computing capabilities can help unlock sensitive data to inform




Kelly.rose@netl.doe.gov