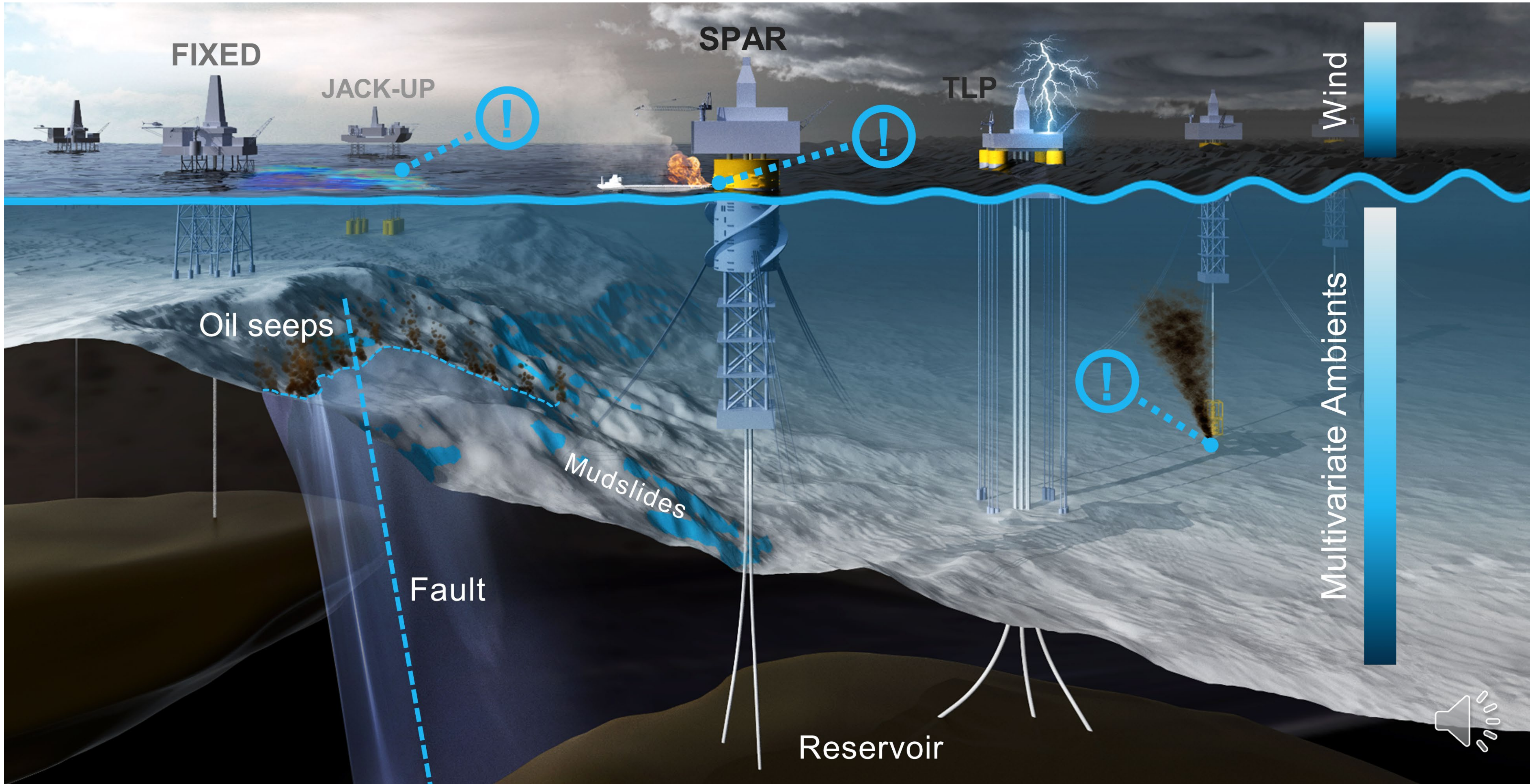
Making the Most of Data: Feature Engineering for Applied Supervised Machine Learning

Thomas Martin
Research Scientist
Research & Innovation Center

AIIM

Advanced
Infrastructure
Integrity
Modeling



Presented to DOE Data Days
June 2, 2022 - LLNL

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.

Authors and Key Personnel

PIs

Lucy Romeo[1,2] and Jennifer Bauer[1]

Key Personnel

Rodrigo Duran[3], Alec Dyer[1,2], Isabelle Pfander[1,2], Thomas Martin[1,2], Chukwuemeka Okoli[1,2], Kelly Rose[1], Michael Sabbatino[1,2], Madison Wenzlick[1,2], Patrick Wingo[1,2], Dakota Zaengle[1,2]

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) Theiss Research, 7411 Eads Avenue, La Jolla, CA 92037, USA

Thomas.Martine@netl.doe.gov
Lucy.Romeo@netl.doe.gov
Jennifer.Bauer@netl.doe.gov

Setting the Stage

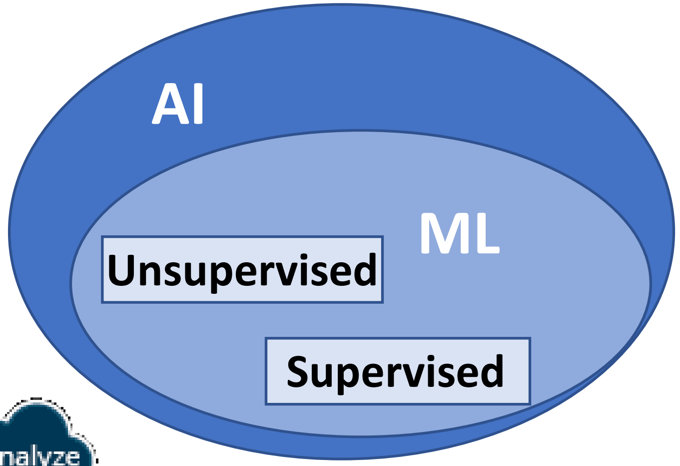
NATIONAL ENERGY TECHNOLOGY LABORATORY

Machine Learning (ML)

- **Supervised ML** – machine is trained, taught with labeled examples
- **Unsupervised ML**– machine creates its own labels (i.e. clustering)
- **Big Data & Big Data Computing** – Large volumes, variety, variability, velocity of data and the computing engineering & systems to handle them

Features – Variables or attributes (ex. continuous or categorical)

Feature Engineering – Select, transform, process, and visualize input features of a given dataset



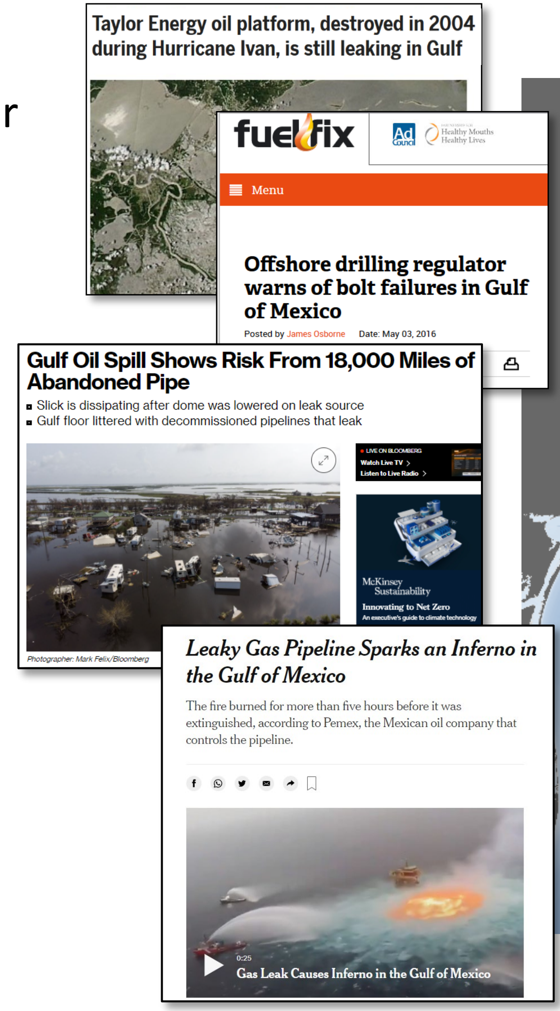
JOULE
NETL SUPERCOMPUTER

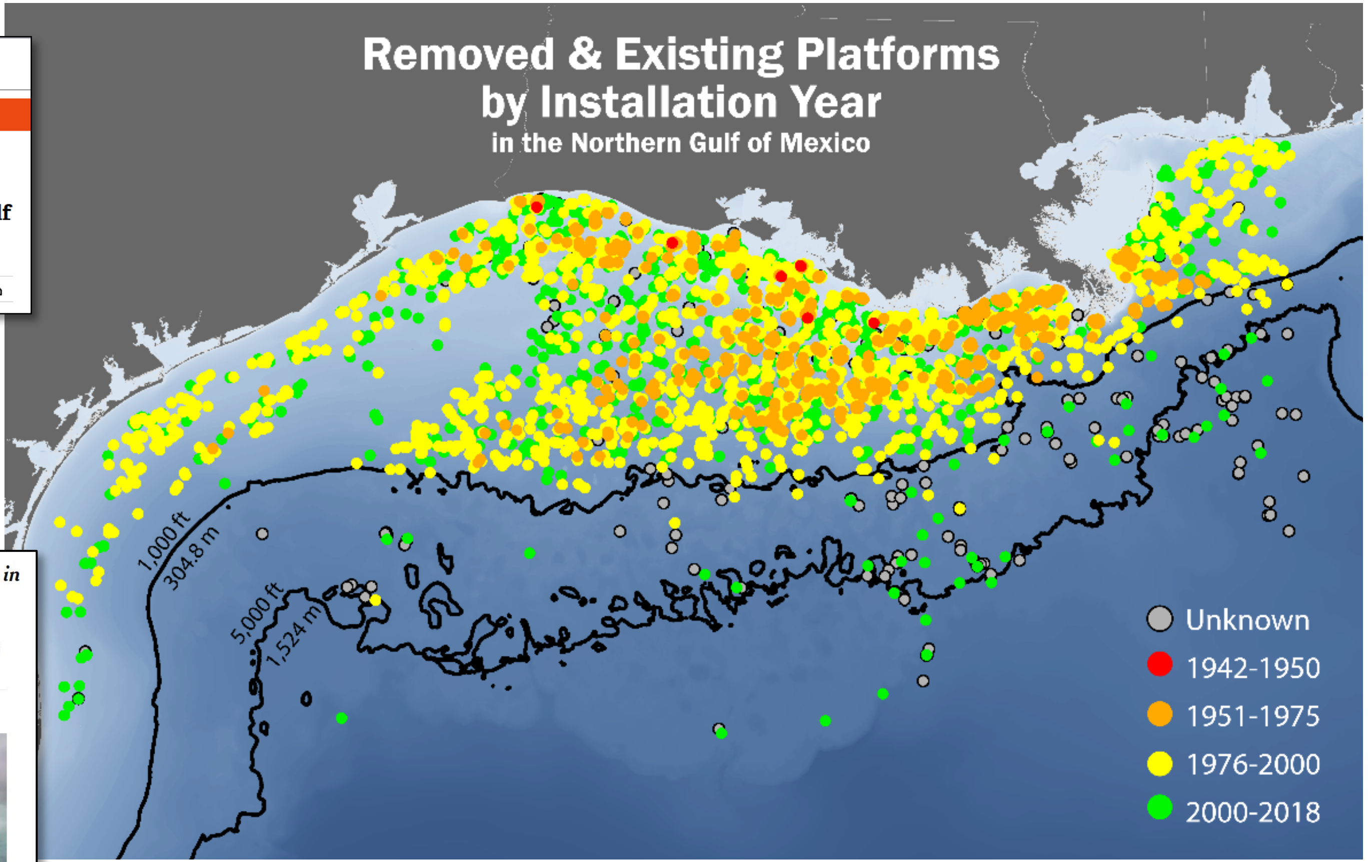
WATT
The Power of Ai

https://edx.netl.doe.gov/offshore/portfolio-items/assessing-current-and-future-infrastructure-hazards/

Offshore Infrastructure Hazards

- Aging infrastructure
- Operational wear-and-tear
- Offshore environment:
  - Extreme weather
  - Climate change
  - Corrosion hazards
  - Geohazards
- **Need:**
  - Identify & prevent hazards
  - Inform safe lifespan extension strategies
  - Environmentally prudent planning in low-carbon economy





Typical platform design life, *20-30 years*

>60% of platforms >30 years old

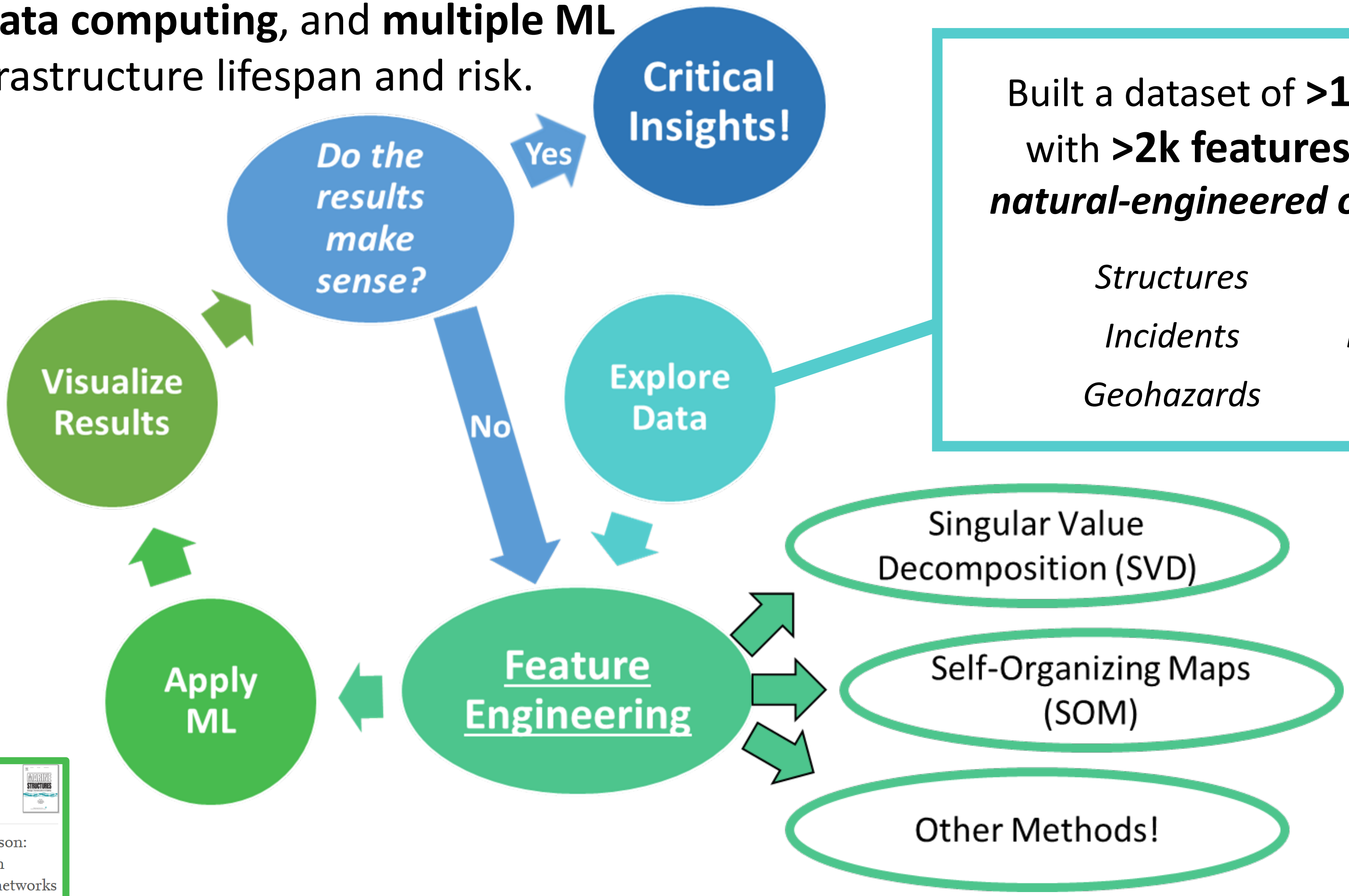
https://edx.netl.doe.gov/offshore/portfolio-items/assessing-current-and-future-infrastructure-hazards/

How AIIM Operates

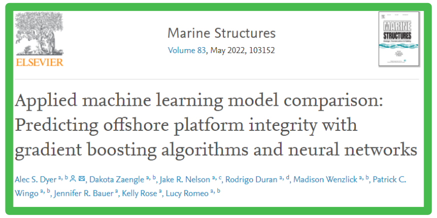
Utilizes **big data**, **big data computing**, and **multiple ML models** to forecast infrastructure lifespan and risk.

Key points:

- Data **analysis** and **visualization** at every step
- Subject Matter Expert **QAQC**
- A focus on the final product being **explainable**, **logical**, and **defendable**



Nelson et al., 2021



Marine Structures
Volume 83, May 2022, 103152

Applied machine learning model comparison: Predicting offshore platform integrity with gradient boosting algorithms and neural networks

Alec S. Dyer, Dakota Zaengle, Jake R. Nelson, Rodrigo Duran, Madison Wenzlick, Patrick C. Wingo, Jennifer R. Bauer, Kelly Rose, Lucy Romeo

Dyer et al, 2022

https://edx.netl.doe.gov/offshore/portfolio-items/assessing-current-and-future-infrastructure-hazards/

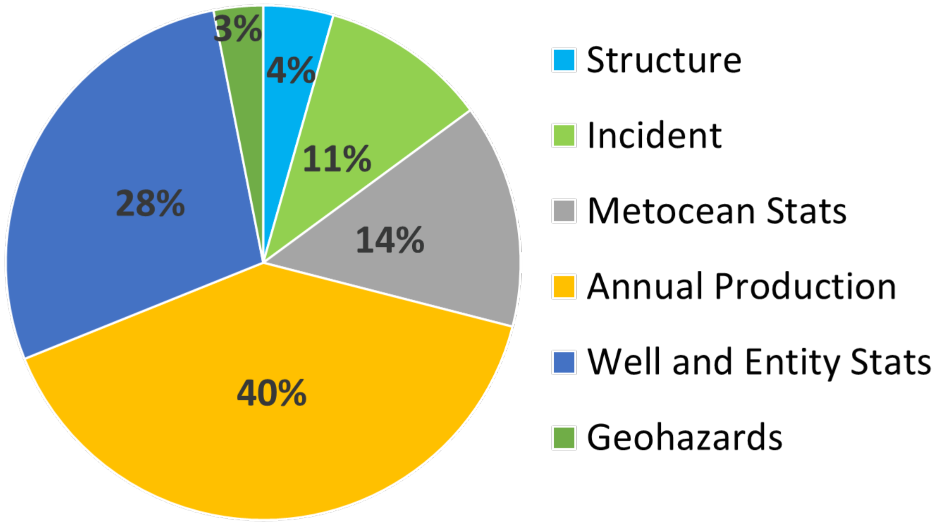
Getting to Know the *Integrated* Data

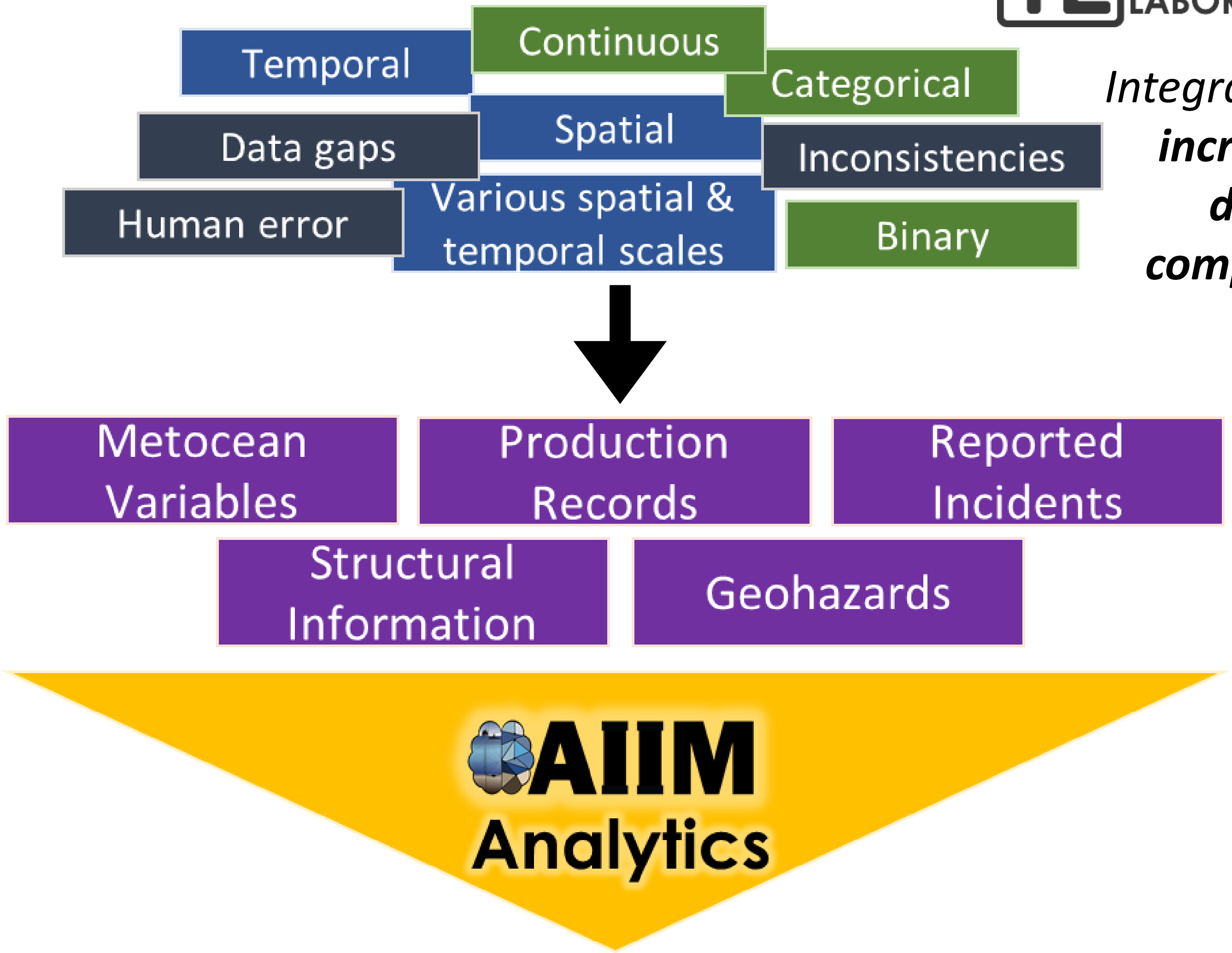
Challenges & Opportunities

> 11,000 platform records * > 2,000 features
= **>22,000,000 data values**

~50% of the dataset has ~90% coverage on a per feature basis

Feature Breakdown





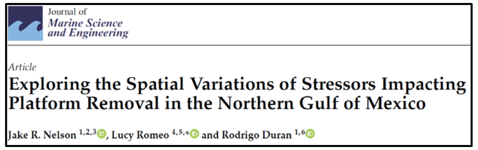
*Integration has **increased data complexity***

https://edx.netl.doe.gov/offshore/portfolio-items/assessing-current-and-future-infrastructure-hazards/

Singular Value Decomposition (SVD)

SVD efficiently **identifies** and **summarizes important information** in a **correlation** or **covariance** matrix

[Nelson et al, 2021](#)



Journal of Marine Science and Engineering

Article

Exploring the Spatial Variations of Stressors Impacting Platform Removal in the Northern Gulf of Mexico

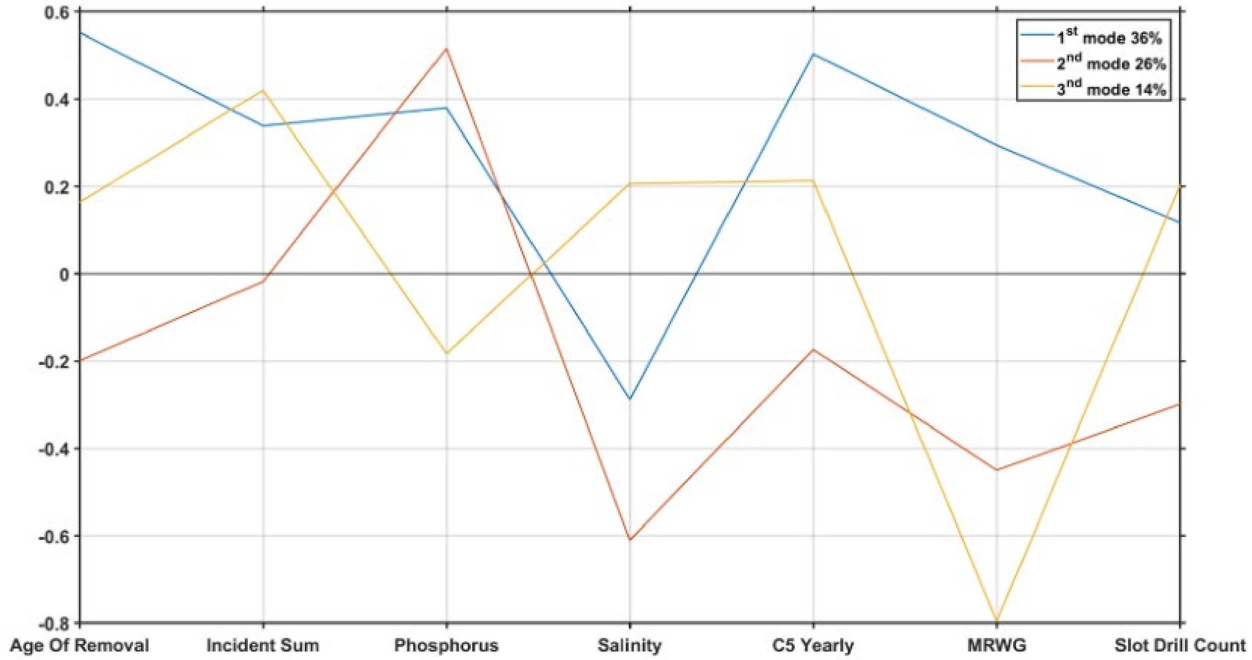
Jake R. Nelson [1,2,3], Lucy Romeo [4,5,*] and Rodrigo Duran [1,6]

Pros

- Interpretable
- Appropriate for time series and continuous spatial data
- Most efficient way to summarize data in a matrix (Eckart-Young Theorem)

Cons

- Does not work with categorical features
- Incomplete data requires pre-processing
- Expert opinion needed to select features



First three right singular vectors of a data correlation matrix, showing relations between **input variables** and the **target variable** “Age of Removal”. 76% percent of features explained by 6 features.

https://edx.netl.doe.gov/offshore/portfolio-items/assessing-current-and-future-infrastructure-hazards/

Self Organizing Maps (SOM)

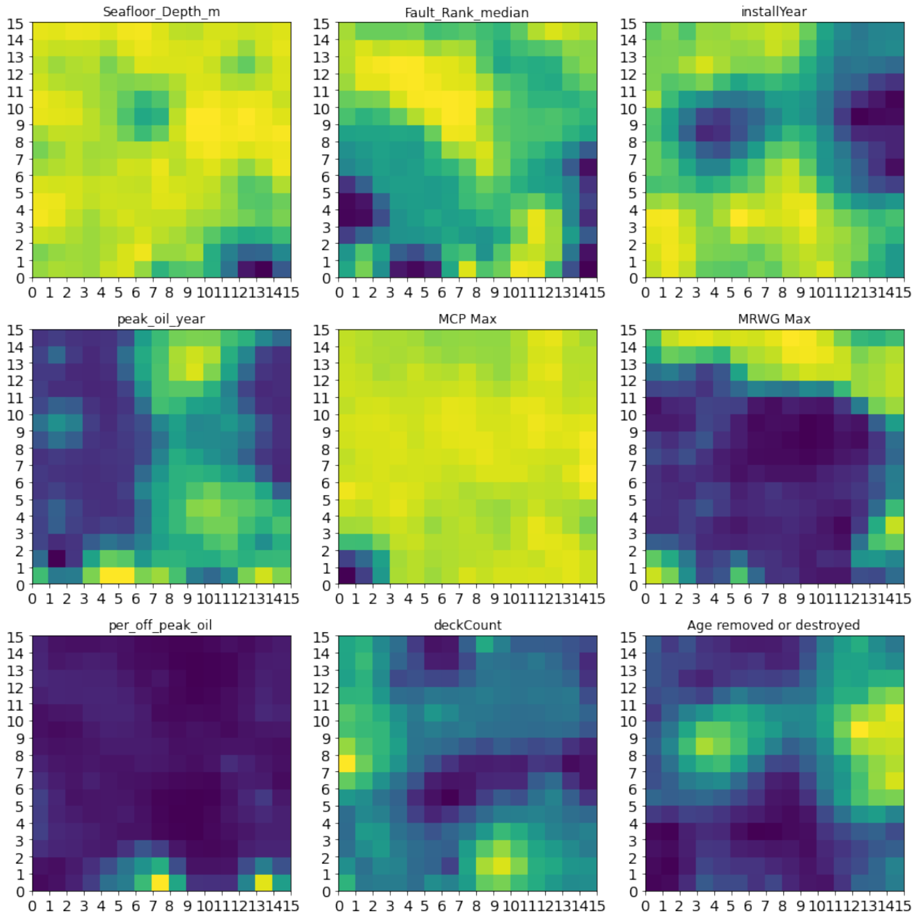
SOM is an *unsupervised* ML technique that is a specific type of neural network. SOMs identify non-linear feature relationships.

Pros

- Can be used with nonlinear features
- Relatively fast
- Threshold for different features is user-selected
- Can be used to create composite features

Cons

- Can’t be used with categorical data
- Expert opinion needed to select features
- Like all neural networks, complete and pre-processed data helps with convergence and speed



Self Organizing Map Weights compared to target variable (lower right)

https://edx.netl.doe.gov/offshore/portfolio-items/assessing-current-and-future-infrastructure-hazards/

Letting the Model Decide – Feature Importance

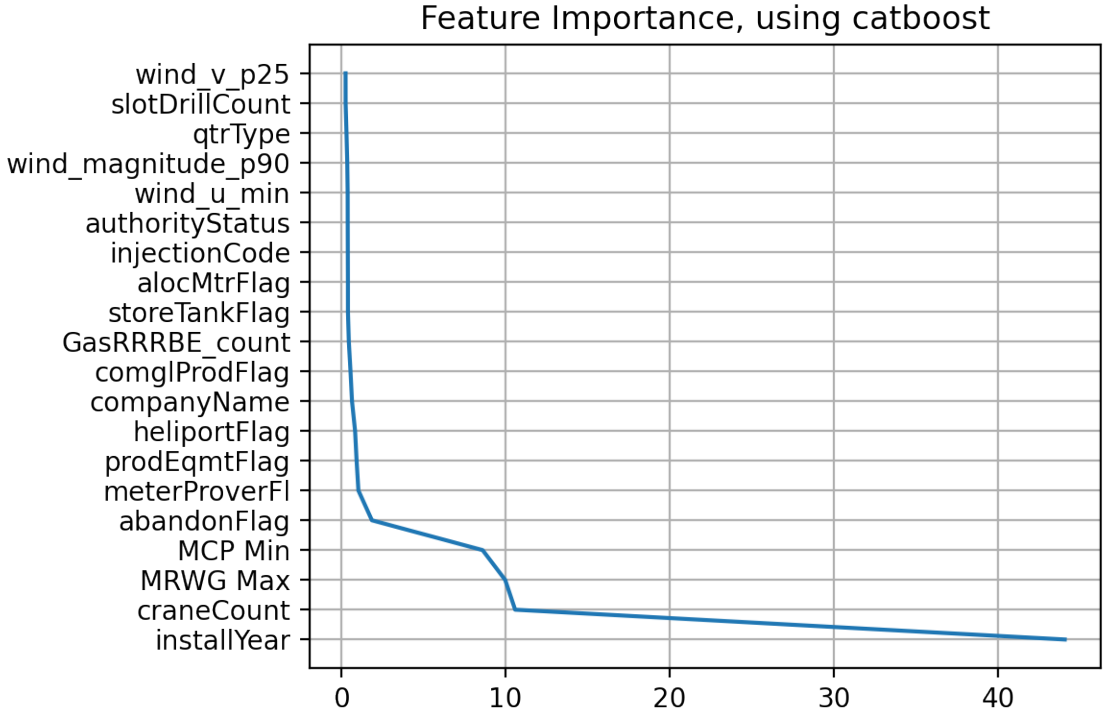
Gradient Boosted Decision Trees (GBDTs) are a common and well-used ML algorithm. This is one method to assess every features importance.

Pros

- Handles all data types
- Easily interpretable
- Great ML model to be used for prediction as well

Cons

- Shared feature importance (potentially collinearity w/other input variables).
- Scores are presented quantitively, easy to overinterpret.
- Model accuracy has limited impact on feature importance.



Feature Importance from a GBDT (using CatBoost). This specific model removed many features while still retaining similar accuracy. Low importance features could be further removed.

https://edx.netl.doe.gov/offshore/portfolio-items/assessing-current-and-future-infrastructure-hazards/

Comparison of Methods

NATIONAL ENERGY TECHNOLOGY LABORATORY

Overcoming incomplete, complex, multivariate data

SVD

- **Best** for **numeric data** (time series, spatial)

AIIM

SOM

- **Best** for deciding between **closely related non-linear features**

GBDT

- Best for **categorical, incomplete data**

Using any method alone will give an incomplete picture

Identified variables containing duplicate information.

Highlighted storm-related features as important.

Confirmed age variables are key.

Confirmed findings from SVD testing.

Using top 5–10 variables does not degrade model performance.

Continued interpretation of environmental loadings and age variables is key.

https://edx.netl.doe.gov/offshore/portfolio-items/assessing-current-and-future-infrastructure-hazards/

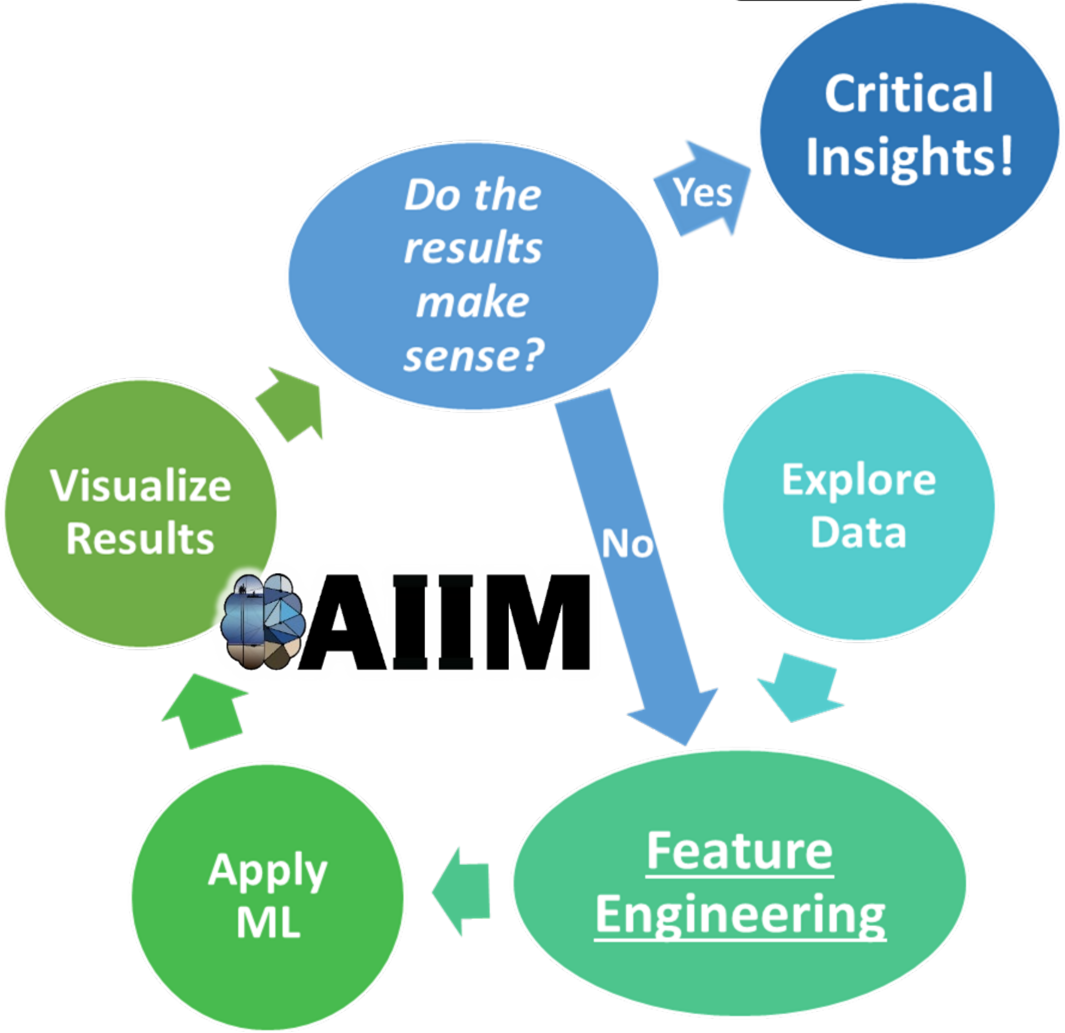
Key Findings

Thomas Martin
thomas.martin@netl.doe.gov

NATIONAL ENERGY TECHNOLOGY LABORATORY

Feature Engineering Matters!

- Reduced input features from >2,000 to >20
- ***Minimizing*** complexity & error, ***maintaining*** accuracy
- Insights to inform **safe infrastructure reuse & removal**
- Identify hazards to support **environmental** and **operational risk prevention**



Next Steps:

- Finalize feature engineering
- Expand to evaluate pipelines & wellbores
- Develop & compare additional models
- Build an interactive AIIM modeling and visualization tool

EDX® Energy Data eXchange

https://edx.netl.doe.gov/offshore/portfolio-items/assessing-current-and-future-infrastructure-hazards/

NETL RESOURCES

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

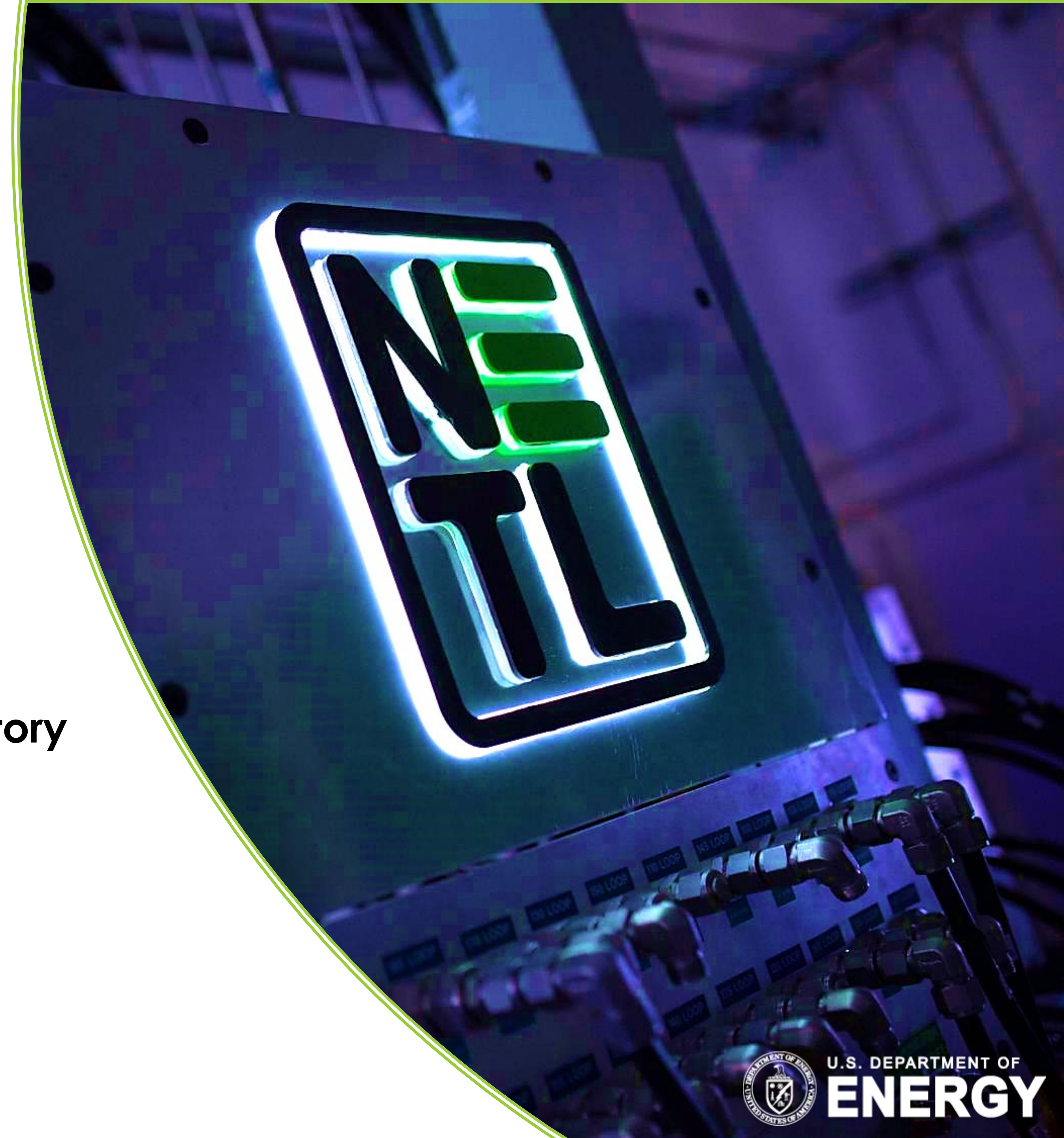
@NETL_DOE

@NETL_DOE

@NationalEnergyTechnologyLaboratory

Thomas Martin
thomas.martin@netl.doe.gov

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



U.S. DEPARTMENT OF ENERGY

References

- Dyer, A.S., Zaengle, D., Nelson, J.R., Duran, R., Wenzlick, M., Wingo, P.C., Bauer, J.R., Rose, K., and Romeo, L. (2022). Applied machine learning model comparison: Predicting offshore platform integrity with gradient boosting algorithms and neural networks, Marine Structures, Volume 83, 103152. https://doi.org/10.1016/j.marstruc.2021.103152.

- Nelson, J. R., Romeo, L., & Duran, R. (2021). Exploring the Spatial Variations of Stressors Impacting Platform Removal in the Northern Gulf of Mexico. *Journal of Marine Science and Engineering*, *9*(11), 1223.

- Nelson, J., Dyer, A., Romeo, L., Wenzlick, M., Zaengle, D., Duran, R., Sabbatino, M., Wingo, P., Barkhurst, A., Rose, K., Bauer, J. Evaluating Offshore Infrastructure Integrity; DOE/NETL-2021/2643; NETL Technical Report Series; U.S. Department of Energy, National Energy Technology Laboratory: Albany, OR, 2020; p 70. doi.org/10.2172/1780656

https://edx.netl.doe.gov/offshore/portfolio-items/assessing-current-and-future-infrastructure-hazards/