NS&T ML-AI

Big Data *Machine Learning* *Artificial Intelligence*

This symposium involves audio/video recording of all presentations, discussions, and comments. The recording for this symposium will be made available to the public. By receiving this notification, your participation consents to recording your interactions with the symposium and public release.

Your audio and video is muted by default

Use the "Chat" feature to ask questions. All questions will be addressed at the end of each presentation (time permitting)

Use the "Chat" feature to let us know if you have technical difficulties

For low-quality connections, switch off video and do not use VPN, if possible

A separate audio PIN will be provided when you sign in for the phone-in audio option

Webinar will begin at 11:00 am MST

Welcome to the

Artificial Intelligence and Machine Learning Symposium 7.0

February 10, 2022

IDAHO NATIONAL LABORATORY

Big Data, Machine Learning, Artificial Intelligence

"'Data – What is it good for!"

Agenda – ML/AI Symposium 7.0

February 10, 2022 – 11:00 AM to 1:00 PM MDT

Time	Presentation Subject	Speaker(s)
11:00-11:10	"Data" What is it good for!	**Curtis Smith**
11:10-11:30	Addressing Data Issues and Data Collection to Support AI Development	**Jeremy Renshaw**
11:30-11:45	Operating Nuclear Power Plant Data for AI/ML Applications	**Zhegang Ma**
11:45-12:00	Large Language Models in The Nuclear Domain	**Bradley Fox & Jerrold Vincent**
12:00-12:10	Considerations of Data Integration in the Nuclear Power Industry	**Ahmad Al Rashdan**
12:10-12:20	INL Strategic Plan: Data goal	**Eric Whiting**
12:20-12:30	Physics-informed Machine Learning for Engineering Applications with Sparse Data: BWR Moisture-Carryover Prediction	**Haoyu Wang**
12:30-12:40	Non-Invertible Deceptive Infusion of Data (DIOD) Methodology for Critical Data Communication	**Hany Abdel-Khalik**
12:40-12:50	Analysis and handling of big data in cosmology: AI/ML to the rescue	**Katrin Heitman**
12:50-1:00	Improving the quality of Imbalanced datasets using Generative Machine Learning Models	**Jared Wadsworth**

NS&T ML-AI

Big Data   Machine Learning   Artificial Intelligence

Thank you

February 10, 2022

Dr. Curtis Smith, Director
Nuclear Safety and regulatory
Research Division

Welcome to the AI/ML Symposium 7.0 – "Data" What is it good for!

"Data" What is it good for!

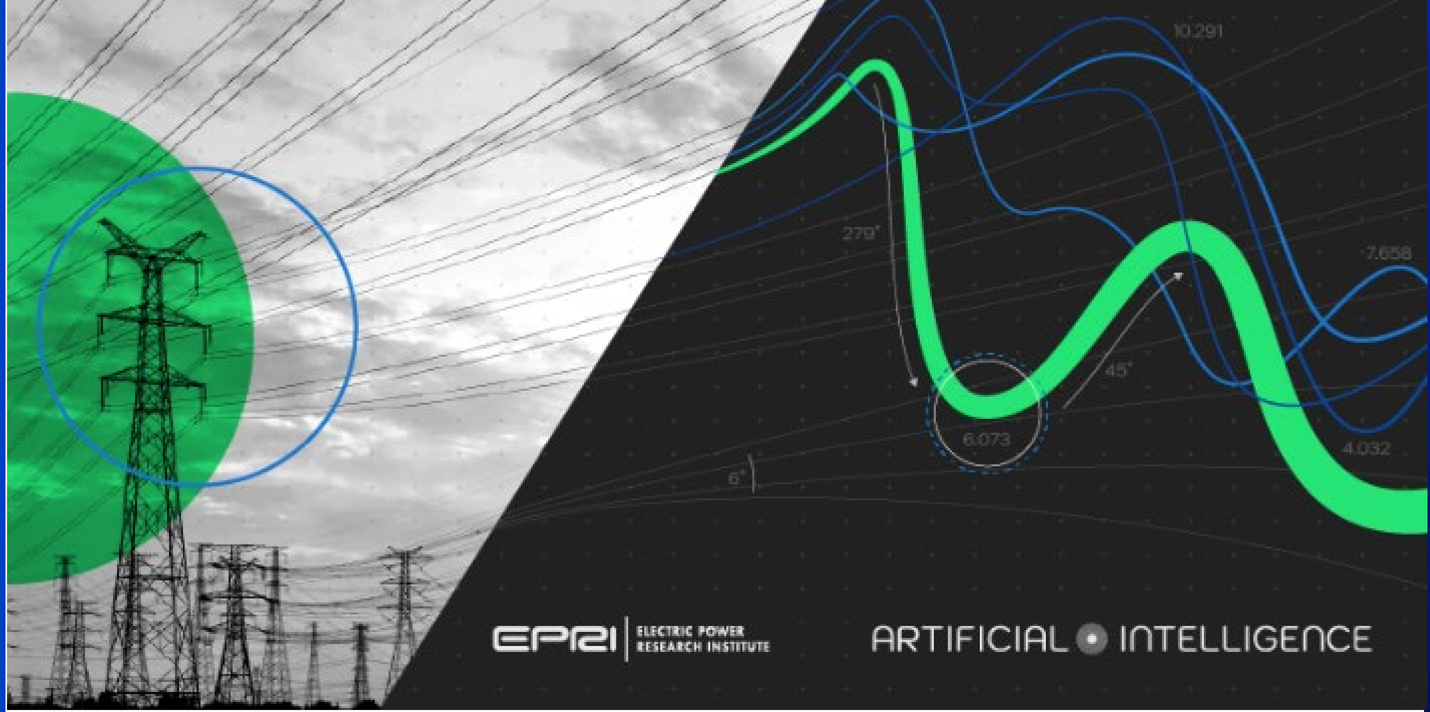
- About two years ago, we started the x.0 symposiums on Artificial Intelligence (AI) and Machine Learning (ML), with a focus on science and engineering
- In that time, AI/ML has continued to evolve and be applied to complex tasks
- What has not really changed is the need for **data** in AI/ML
  - Hence the focus of the 7.0 symposium
- Unlike the “War” Motown song sung by Edwin Starr, data is absolutely worth something
- Today, we will hear from a variety of speakers on the need and use of data for various applications and domains

“And I told him, AI and ML aren’t the thing. They’re the thing that gets us to the thing.”

(See Halt and Catch Fire)

Curtis.Smith@inl.gov

Thank you and enjoy the symposium!



Addressing Data Issues and Data Collection to Support AI Development

Jeremy Renshaw
Sr. Program Manager, Artificial Intelligence
jrenshaw@epri.com

ai.epri.com | ai@epri.com

www.epri.com

Key Point on Importance of Data

Improving the data used in your AI/ML model will (typically) improve the model performance more than improving the AI model itself

Importance of Data for Artificial Intelligence

- Data plays a critical role in AI/ML model results
- Data must be of sufficient **quality, quantity, and cover the anticipated range of conditions**
- Data is the "fuel" for the AI engine, but we are much more likely to feed our AI bad data than put bad gas in our car.
- Having a "great" algorithm on "fair" data is worse than having a "fair" algorithm on "great" data

Many algorithms using AI that solved major challenges were available years or decades before they were implemented. The limiting factor was data availability.

Notable Examples of AI Failures Due to Data Issues

An AI system was used to flag suspected fraudulent transactions in financial data

The AI algorithm was trained with vast tomes of high-quality data

The data had high quality, large quantity, and covered a wide range of conditions

However, when it went active, the AI algorithm immediately flagged **every single transaction** on a particular island as fraudulent.

What went wrong?

Notable Examples of AI Failures Due to Data Issues

= Area Used by Humans to Classify the Image

= Area Used by Computers to Classify the Image

Image Classified as: Dog

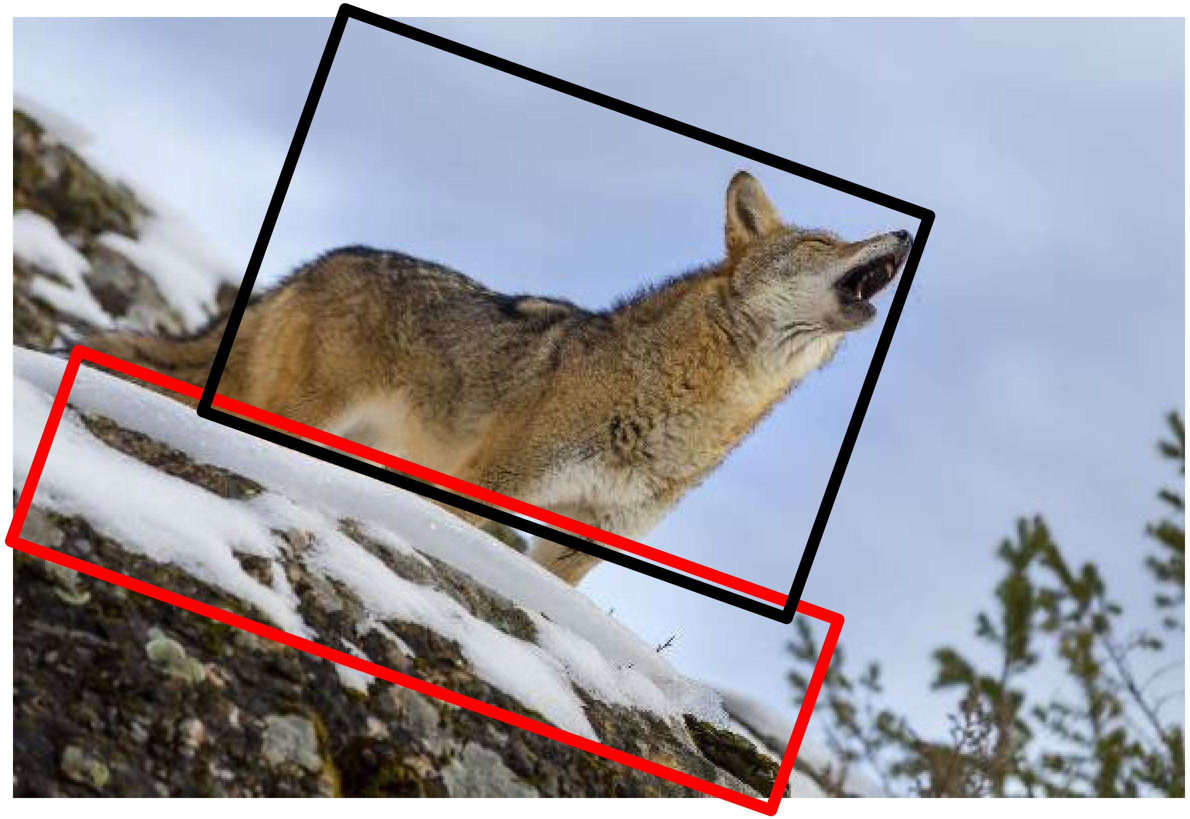


Image Classified as: Wolf

Notable Examples of AI Failures Due to Data Issues

- Image Classified as: Dog

- Image Classified as: Wolf

What is needed for AI to be successful in Power Industry?

Training Data Sets:

- Statistically Significant
- Wide Range of Conditions
- Secured and Governed
- Anonymized

Power Industry Experts:

- Understand AI Basics
- Know where AI is Applicable
- Aware of, and engaged with, AI Community and sharing data

EPRI

AI Community:

- Aware of Power Industry Issues
- Understand the Physics
- Have access to Data Sets

Understand AI Performance:

- Criteria for AI Applicability
- Unbiased Technically Sound Evaluation of AI Solutions

What is needed for AI to be successful in Power Industry?

Training Data Sets:

- Statistically Significant
- Wide Range of Conditions
- Secured and Governed
- Anonymized

Power Industry Experts:

- Understand AI Basics
- Know where AI is Applicable
- Aware of, and engaged with, AI Community and sharing data

EPRI

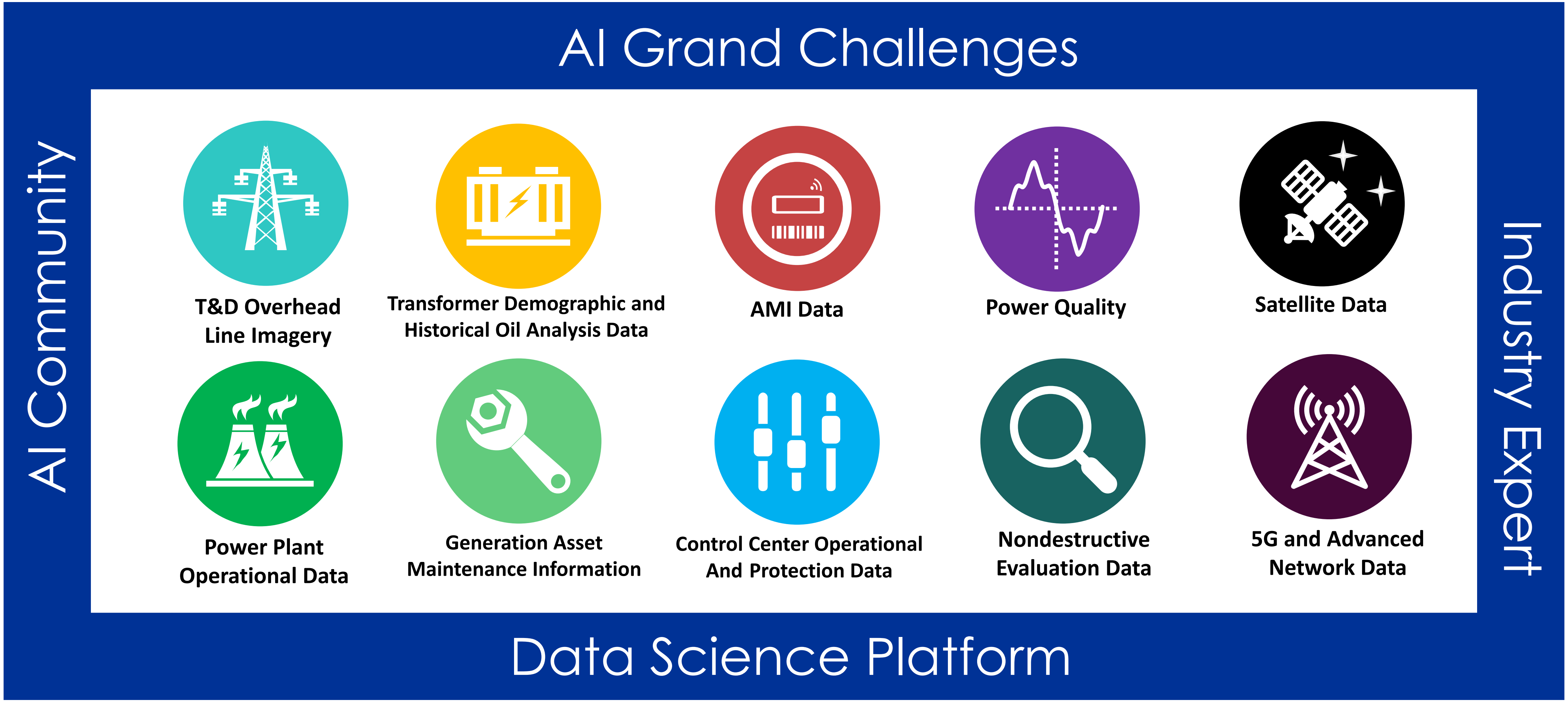
AI Community:

- Aware of Power Industry Issues
- Understand the Physics
- Have access to Data Sets

Understand AI Performance:

- Criteria for AI Applicability
- Unbiased Technically Sound Evaluation of AI Solutions

AI.EPRI: Be the AI catalyst for tomorrow's energy network



AI.EPRI.com

ai@epri.com

Building an AI-Electric Power Community

Collecting, Curating and Sharing Data, and Developing Solutions

Deepening AI Expertise in the Electric Power Industry

EPRI ARTIFICIAL • INTELLIGENCE

Our

AI GRAND CHALLENGES

- Grid-Integrated Smart Cities
- Energy System Resilience
- Environmental Impacts
- Intelligent and Autonomous Plants
- AI-Enhanced Cybersecurity

Conclusions

- Data is a critical aspect of AI/ML models that cannot be overlooked
- Data must be of sufficient quantity, quality, and cover the range of conditions
- WATCH OUT for biases in the data and expect the unexpected
- Understand and try to mitigate potential pitfalls and unintended consequences caused by your training data

- AND Don't forget!!!
- Improving the data used in your AI/ML model will (typically) improve the model performance more than improving the AI model itself

Contact:

ai.epri.com | ai@epri.com


Together…Shaping the Future of Energy™

February 10, 2022

Zhegang Ma

Zhegang.Ma@inl.gov

Operating Nuclear Power Plant Data for AI/ML Applications

INL – AI & ML Symposium 7.0

Zhegang Ma, Sai Zhang, Han Bao, Andrea Mack
Idaho National Laboratory

Min Xian
University of Idaho

Introduction

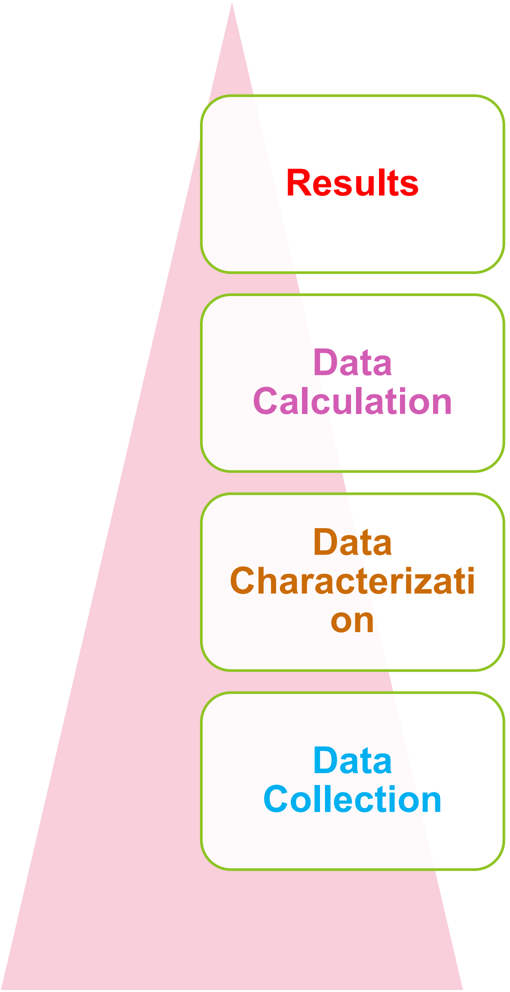
- Since 1990s, Idaho National Laboratory (INL) has been providing technical assistance to the Nuclear Regulatory Commission (NRC) on data collection and computation activities associated with nuclear power plant operating experience (OpE) information
- **Two "Classical" NRC OpE Projects (2000 – Current)**
  - **Reactor Operating Experience Data (RxOpED) for Risk Applications**
    - Integrated Data Collection and Coding System (IDCCS)
    - Capture, update, and maintain data needed to support data computation activities
    - Web display methods ➡ NRC Reactor Operating Experience Data (NROD), **nrod.inl.gov**

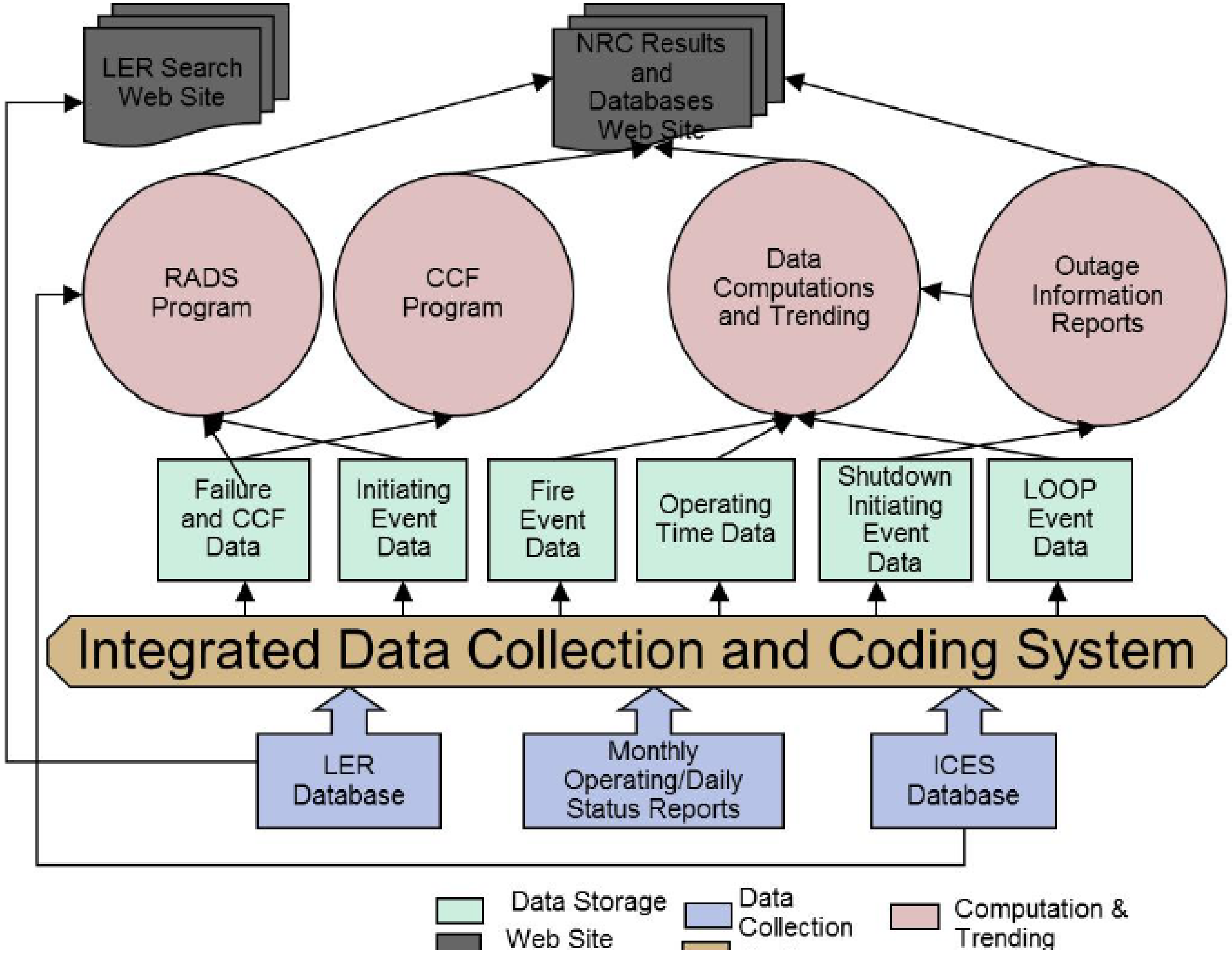
Introduction (cont.)

- Two “Classical” NRC OpE Projects
  - **Computational Support for Risk Applications (CSRA)**
    - Maintain and update industry and plant-specific system and component reliabilities, initiating events frequencies, system/train unavailability, and common-cause failure (CCF) parameter estimates
    - Update component performance and system reliability studies
    - Probabilistic risk assessment (PRA) data calculations web site → Reliability and Availability Data System (RADS), **rads.inl.gov**
    - Web pages that display updated calculation results → NRC Reactor Operational Experience Results and Databases, **nrcoe.inl.gov**

nrcoe.inl.gov is available to the public.
nrod.inl.gov and rads.inl.gov include proprietary data from Institute of Nuclear Power Operations (INPO) and are available to INPO members only.

Introduction (cont.)





Introduction (cont.)

- **One “Exploratory” NRC OpE Project (2020-2021)**
  - **Feasibility Study of Advanced Computational Predictive Capabilities Using Artificial Intelligence (AI), Machine Learning (ML) and Analytics in OpE**
    - Explore advanced computational tools and techniques for operating nuclear plants
    - Assess the using of AI/ML in commercial nuclear industry
    - Explore potential applications of AI/ML in nuclear power plants
    - An overview of nuclear data & sources was conducted to support the above tasks

Introduction (cont.)

- **Data vs Information** (Merriam-Webster; D. Kelly and C. Smith, Bayesian inference for probabilistic risk assessment: A practitioner's guidebook, 2011)
- **Data**
  - Basic, unrefined, and generally observable information
  - Factual information used as a basis for reasoning, discussion, or calculation
- **Information**
  - Processed, more refined, and often inferred data
  - Knowledge obtained from investigation, study, or instruction
- We used the term “data” here in a general sense that it could include “information”

Nuclear Data & Sources – “Classical”

- **U.S. Nuclear Industry**
  - Licensee Event Reports (LERs) – primary source of initiating events (IEs)
    - Reactor trip
    - Turbine trip
    - Loss of offsite power (LOOP)
    - Steam generator tube rupture
  - INPO Data – equipment failure data
    - Pump failed to run
    - Valve failed to open
    - Total demands
    - Total run time

Nuclear Data & Sources – “Classical” (cont.)

- **U.S. Nuclear Industry (cont.)**
  - Monthly Operating Reports – Reactor critical years, shutdown years
  - Event Notification Reports …
- **U.S. NRC**
  - OpE Studies/Trends
    - Component reliabilities: failure probability, failure rate
    - Initiating event frequency
    - CCF parameters
    - Component/system reliability and trend analysis
  - Inspection Reports
  - Preliminary Notifications
- **International Nuclear Industry**

Nuclear Data & Sources – “Exploratory”

- Broader data
  - Observed data
  - Synthetic data
  - Processed data
- OpE data could be plant-specific, generic (national), and generic (international)
- OpE data could be operational data, maintenance data, regulatory data, and other data

Nuclear Data & Sources – "Exploratory" (cont.)

- **Plant-Specific Operational Data**
  - Process instrumentation and control (I&C) data
  - Plant logs
  - Plant condition reports/corrective action programs/internal plant failure reports
- **Plant-Specific Maintenance Data**
  - Maintenance and replacement records
  - Inspection, calibration, and surveillance test records

Nuclear Data & Sources – “Exploratory” (cont.)

- **Plant-Specific Regulatory Data**
  - LERs
  - Daily/monthly/quarterly/annual reports
  - Regular or special inspection reports
  - Preliminary notification reports
  - Significant enforcement actions
- **Plant-Specific Miscellaneous**
  - Plant design and license-related documents
  - Plant operating guidance documents
  - Technical specifications
  - Plant procedures and guidelines
  - Plant business data

Nuclear Data & Sources – “Exploratory” (cont.)

- **Generic (National) Data - anonymized raw data or processed data**
  - INPO IRIS database
  - NRC IDCCS database – NROD web app
  - NRC LERSearch
  - NRC reliability/IE/LOOP/unavailability/CCF database- RADS web app
  - NUREG/CR-6928 and updates for generic component reliability and IE frequency
  - NRC LOOP reports, IE reports, component and system reliability reports
  - Department of Energy (DOE) generic component failure database for sodium reactor PRAs
  - EPRI reports on pipe rupture frequencies, components, shutdown IE frequencies
  - Human performance data

Nuclear Data & Sources – “Exploratory” (cont.)

- **Generic (International) Data**
  - **International Atomic Energy Agency (IAEA)** OpE feedback, component performance data, reliability data for research reactor PRA
  - **World Association of Nuclear Operators (WANO)** plant performance data, performance analysis program
  - **Organisation for Economic Co-operation and Development/Nuclear Energy Agency (OECD NEA)**
    - OpE feedback
    - Fire incidents records exchange project
    - Component performance data
    - **International common-cause data exchange project**
    - Component operational experience, degradation & aging program
    - Cable aging data and knowledge project

Nuclear Data & Sources – How to Better Utilize

- For “classical” data with conventional statistical methods, how can AI/ML be utilized to provide new insights?
- For “exploratory” data including existing but less utilized data, and new data brought by advanced technologies such as advanced sensors, how can AI/ML be utilized to develop new methodologies and provide new directions?

Large Language Models in The Nuclear Domain

Bradley Fox & Jerrold Vincent

https://nuclearn.ai

Jerrold Vincent

Co-Founder & CFO Nuclearn.ai

Jerrold holds a B.S. in Business Economics and an M.S. in Computer Science. Prior to Co-Founding Nuclearn, Jerrold spent ten years in Utility Data Science and Business Intelligence at Palo Verde Generating Station.

Bradley Fox

Co-Founder & CEO Nuclearn.ai

Brad holds a B.S. Materials Science & Engineering. Prior to Nuclearn Brad spent six years in Nuclear Engineering and six years in Data Science & Software at Palo Verde Generating Station.

NuclearN

Inventors of US Patent 11080127 for *Methods and apparatus for detection of process parameter anomalies*

Recipients of 2020 Nuclear Energy Institute's Top Innovative Practice Award for *Process Automation using Machine Learning*

Current Work

Assessment Readiness (INPO, WANO, etc)

CAP Automation

Multi-Task Large Language Models

CAP Program Human-AI Interface Enhancements

What are large language models?

Form of NLP, using specialized neural networks trained on HUGE amounts of data for modeling natural language

Broad (English), domain specific (Nuclear) or task specific (Q&A)

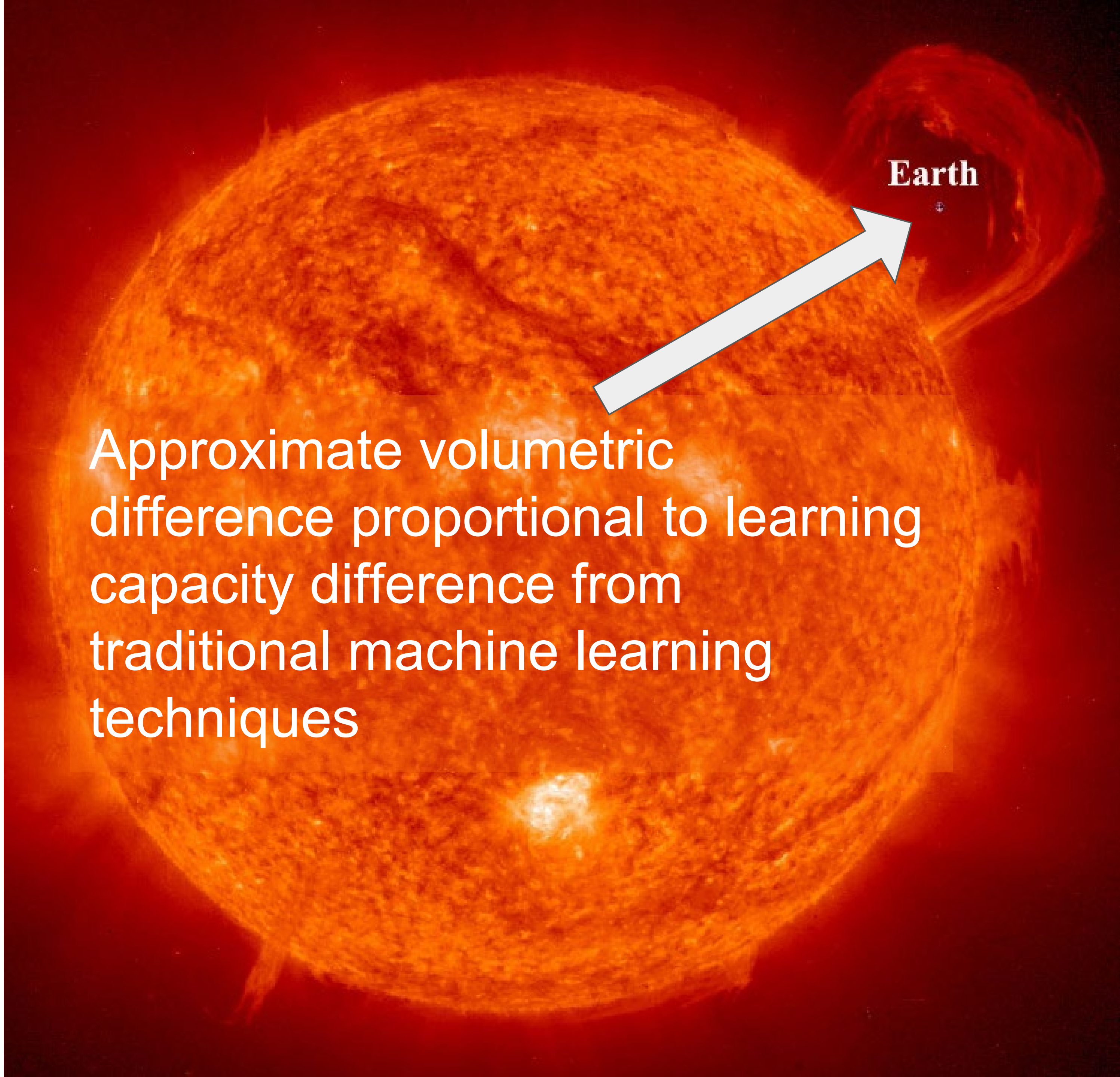
Single model can answer questions, generate novel passages, classify text, perform translations, summarize content

Generative:

$$p(\overbrace{t_1, t_2, \ldots, t_N}^{\text{Token sequence}}) = \prod_{k=1}^{N} p\left(t_k \mid t_1, t_2, \ldots, t_{k-1}\right)$$

Sequence probability

Conditional probabilities



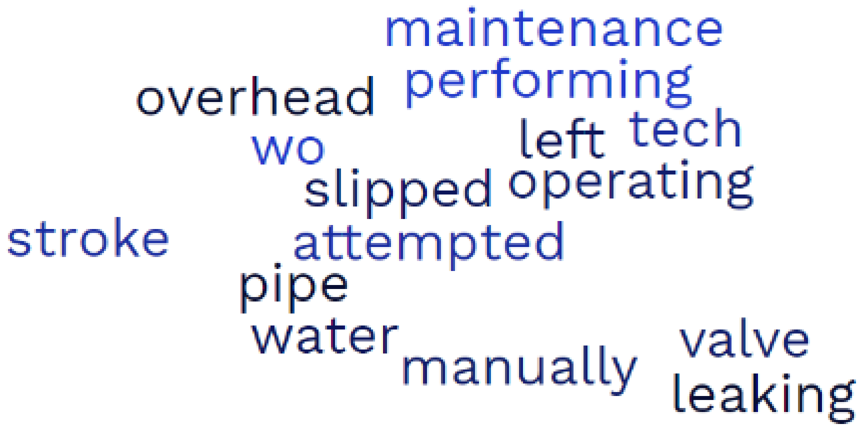
Approximate volumetric difference proportional to learning capacity difference from traditional machine learning techniques

Revolution in Natural Language Approaches

Move data pipeline complexity and feature engineering into the language model

<u>Traditional Approach</u>

- Manually clean text to reduce number of extraneous words and identify “phrases” and “keywords” that matter
- Train Naive Bayes/Boosted Tree/Simple Neural Network on features
- Accuracy is typically lower than humans



<u>Large Language Model Era</u>

- Pre-trained models can perform many tasks without any additional training
- Models can be “fine-tuned” to specific problems to achieve superior performance
- Increased context windows help understanding: 1k - 4k token window
- SuperGLUE NLP Benchmark increase from 44.5 using BOW models to 91.0 [1]

After performing WO 1234567, maintenance tech attempted to stroke the valve. While manually operating the valve, the tech slipped on water left from a leaking overhead pipe.

1.https://super.gluebenchmark.com/leaderboard

Differences in Natural Language Approaches

Everything becomes language

- Reframe problems as text.
- i.e. “*A large metal component with a bonnet, stem and actuator is a {blank}*”

Few or Zero Shot (No task specific training)

- Tasks are designed as ‘prompts’
- *I.e. “The main feed pump is in the turbine building. <sep> The atmospheric dump valve is in the main steam supply building <sep> The reactor is in the {blank}”*

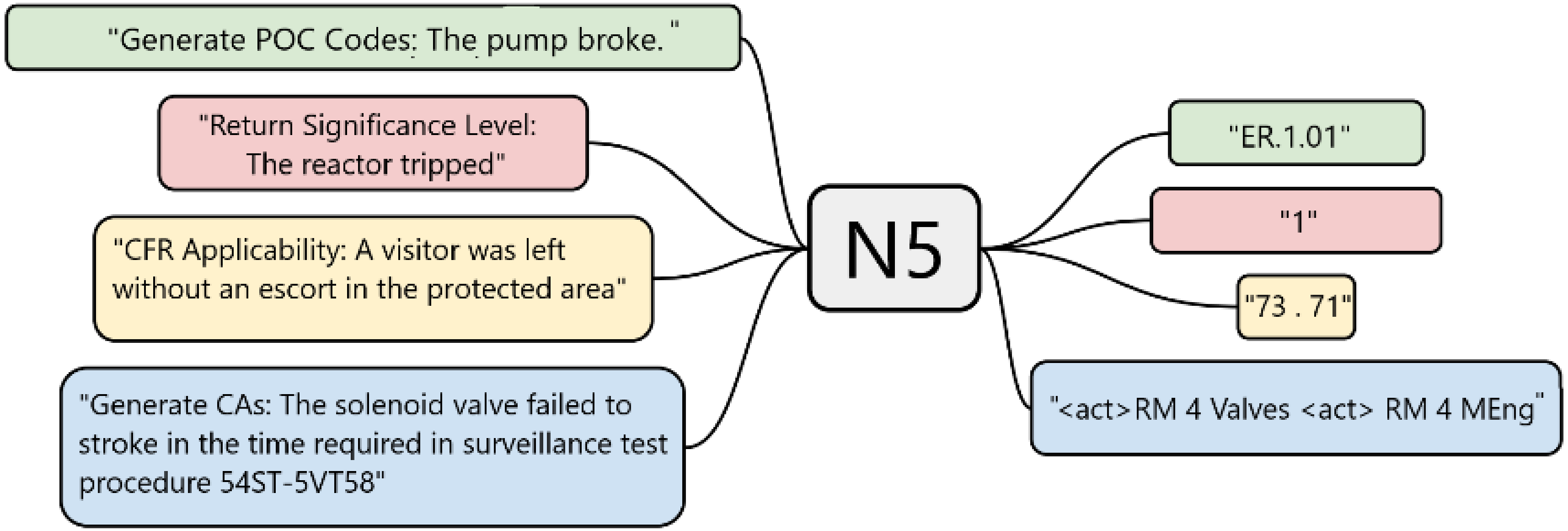
Fine Tuning

- Familiarize model with specific domain language
- Unsupervised or with engineered prompts
- Model updates weights and is specialized in fine-tuned task

What can we do with these models?

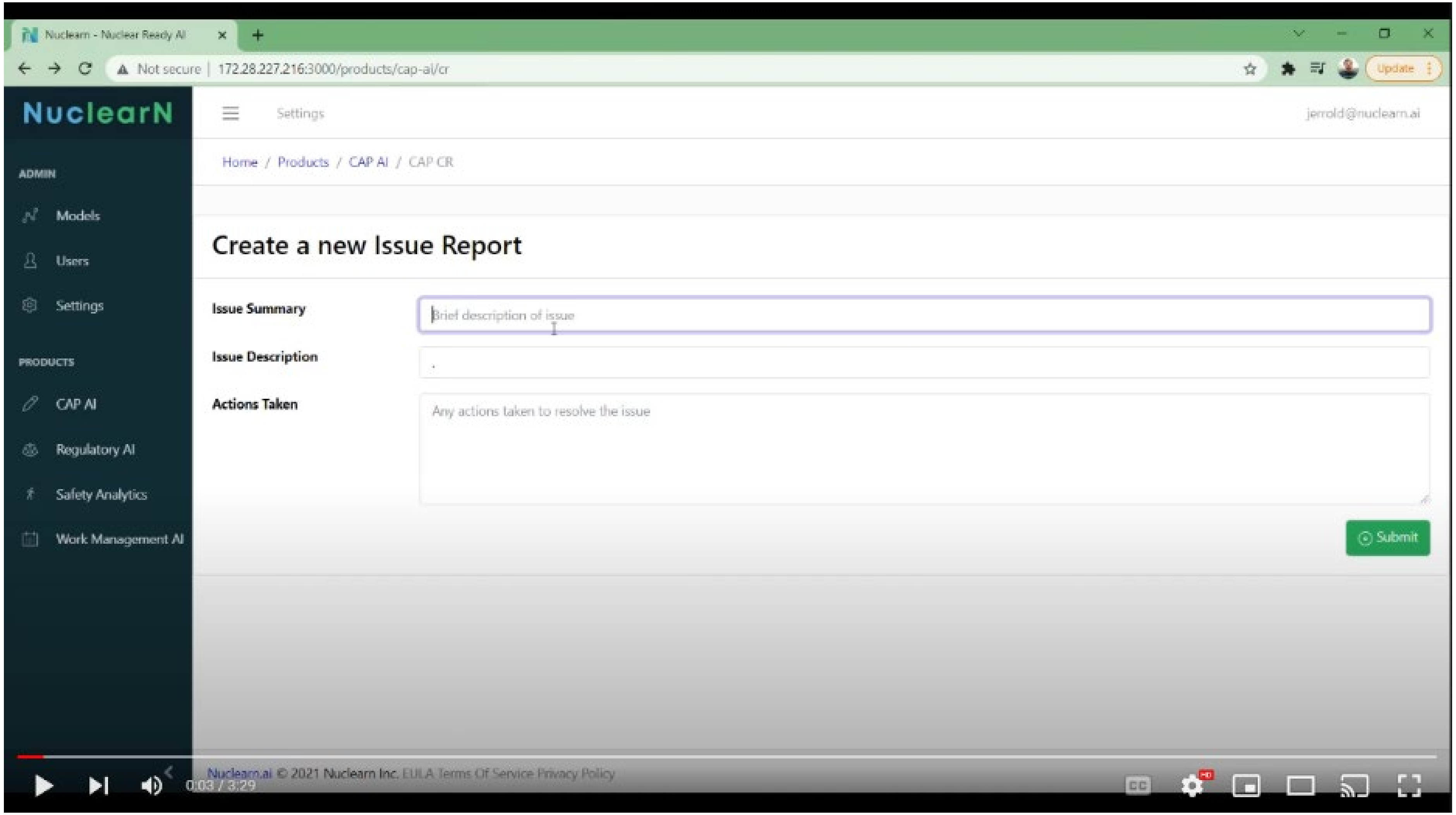
- More accurately auto-screen a higher proportion of issues utilizing improved classification abilities
- Improve the quality of reports using intelligent autocomplete with Nuclear-specific terms and phrases
- Evaluate whether an issue report contains sufficient information as it is being written

Multi Task Framework



Example Application

Writing and evaluating a condition report for quality using an LLM



Nuclearn - Nuclear Ready AI

Not secure | 172.28.227.216:3000/products/cap-ai/cr

NuclearN

Settings

jerrold@nuclearn.ai

ADMIN

- Models
- Users
- Settings

PRODUCTS

- CAP AI
- Regulatory AI
- Safety Analytics
- Work Management AI

Home / Products / CAP AI / CAP CR

Create a new Issue Report

Issue Summary

Foreign material found in condenser

Issue Description

While performing WO #1234567-01, *it was discovered that a piece*

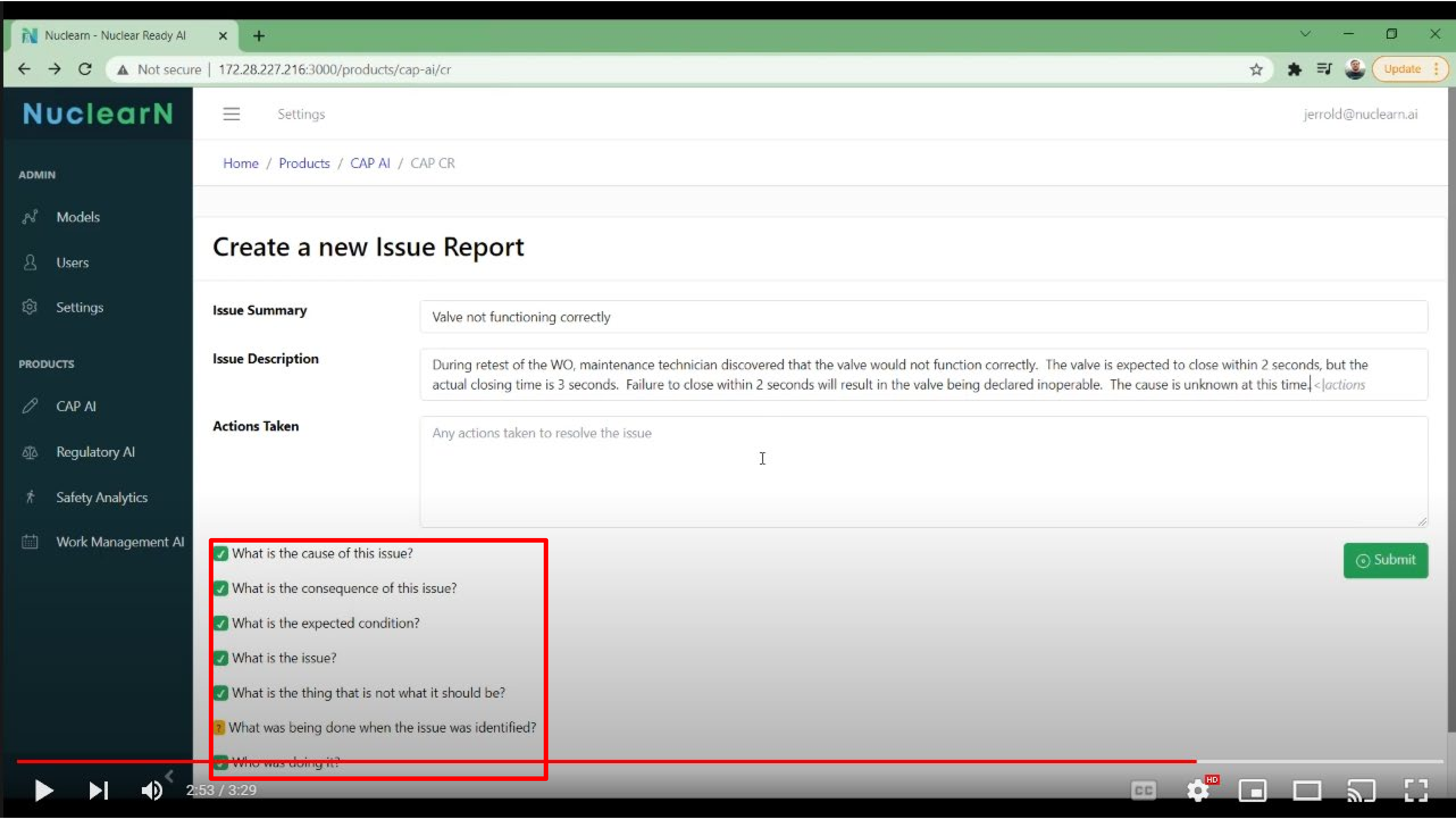
Actions Taken

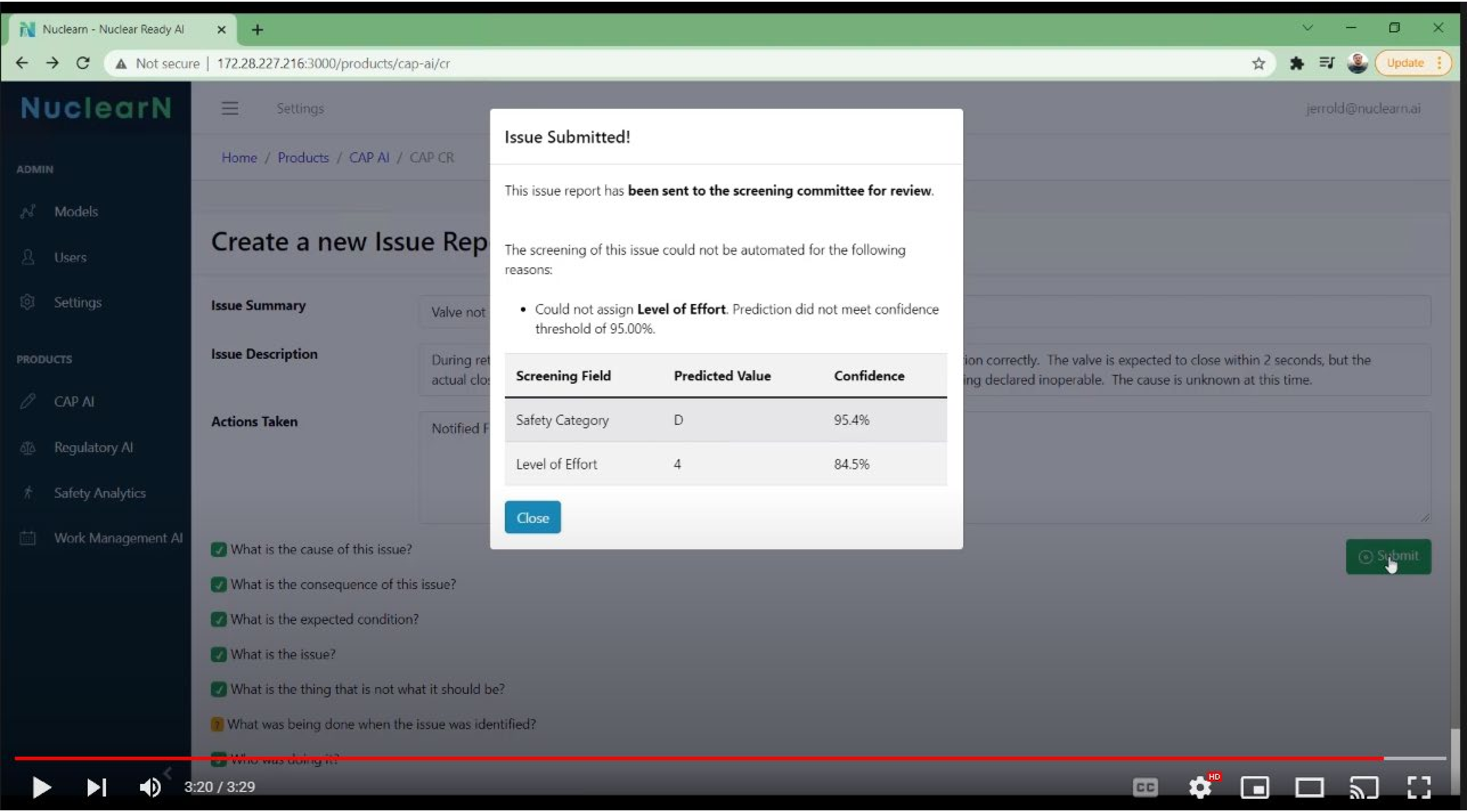
Any actions taken to resolve the issue

- ✓ What is the cause of this issue?
- ✗ What is the consequence of this issue?
- ✗ What is the expected condition?
- ✓ What is the issue?
- ✓ What is the thing that is not what it should be?
- ✗ What was being done when the issue was identified?
- ✗ Who was doing it?

Submit

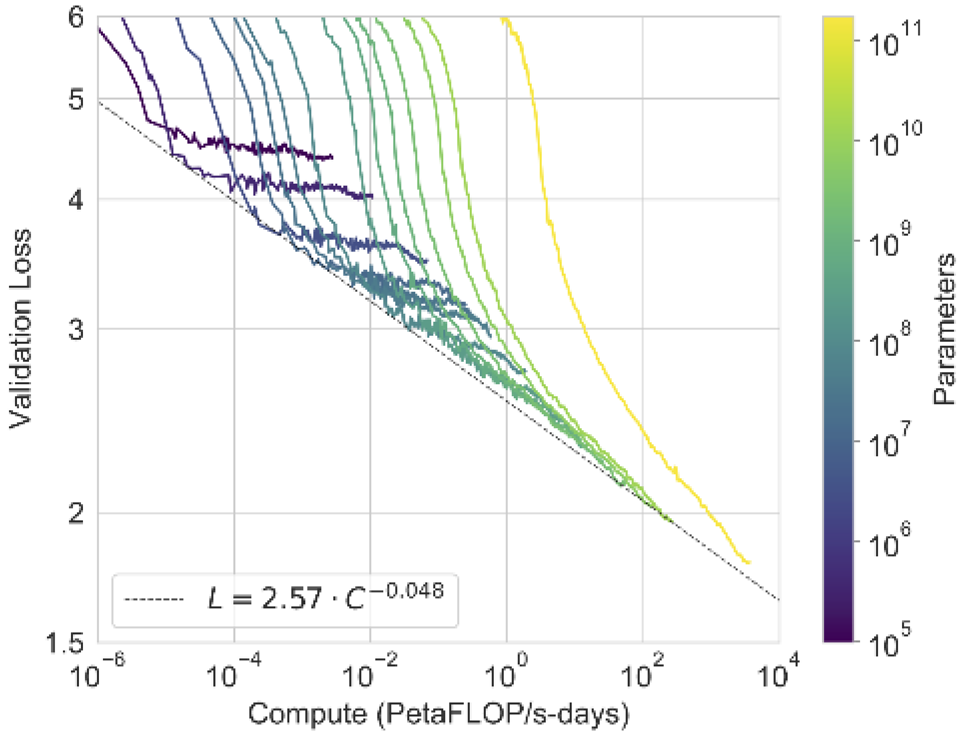
0:34 / 3:29


Nuclearn - Nuclear Ready AI
Not secure | 172.28.227.216:3000/products/cap-ai/cr
Update
NuclearN
Settings
jerrold@nuclearn.ai
ADMIN
Models
Users
Settings
PRODUCTS
CAP AI
Regulatory AI
Safety Analytics
Work Management AI
Home / Products / CAP AI / CAP CR
Create a new Issue Report
Issue Summary
Valve not functioning correctly
Issue Description
During retest of the WO, maintenance technician discovered that the valve would not function correctly. The valve is expected to close within 2 seconds, but the actual closing time is 3 seconds. Failure to close within 2 seconds will result in the valve being declared inoperable. The cause is unknown at this time. <|actions
Actions Taken
Any actions taken to resolve the issue
What is the cause of this issue?
What is the consequence of this issue?
What is the expected condition?
What is the issue?
What is the thing that is not what it should be?
What was being done when the issue was identified?
Who was doing it?
Submit
2:53 / 3:29


Nuclearn - Nuclear Ready AI
Not secure | 172.28.227.216:3000/products/cap-ai/cr
Update
NuclearN
Settings
jerrold@nuclearn.ai
ADMIN
Models
Users
Settings
PRODUCTS
CAP AI
Regulatory AI
Safety Analytics
Work Management AI
Home / Products / CAP AI / CAP CR
Create a new Issue Rep
Issue Summary
Valve not
Issue Description
During ret
actual clo
ion correctly. The valve is expected to close within 2 seconds, but the
ing declared inoperable. The cause is unknown at this time.
Actions Taken
Notified F
What is the cause of this issue?
What is the consequence of this issue?
What is the expected condition?
What is the issue?
What is the thing that is not what it should be?
What was being done when the issue was identified?
Who was doing it?
Submit
Issue Submitted!
This issue report has been sent to the screening committee for review.
The screening of this issue could not be automated for the following reasons:
Could not assign Level of Effort. Prediction did not meet confidence threshold of 95.00%.
Screening Field | Predicted Value | Confidence
Safety Category | D | 95.4%
Level of Effort | 4 | 84.5%
Close
3:20 / 3:29

Large Language Models are still improving.

- Next generation predicted to be 200x size of current generation
- Models will achieve superhuman performance on a broad range of natural language and general AI tasks
- Services such as Github Copilot already leverage advanced auto-complete functionality for millions of users
- Gartner predicts that by 2025 generative AI will account for 10% of all data produced worldwide



For the first time in the history of Machine Learning, there is no evidence of decreasing returns from increasing model size. The only limiting factor is compute resources.

Future Work and Research

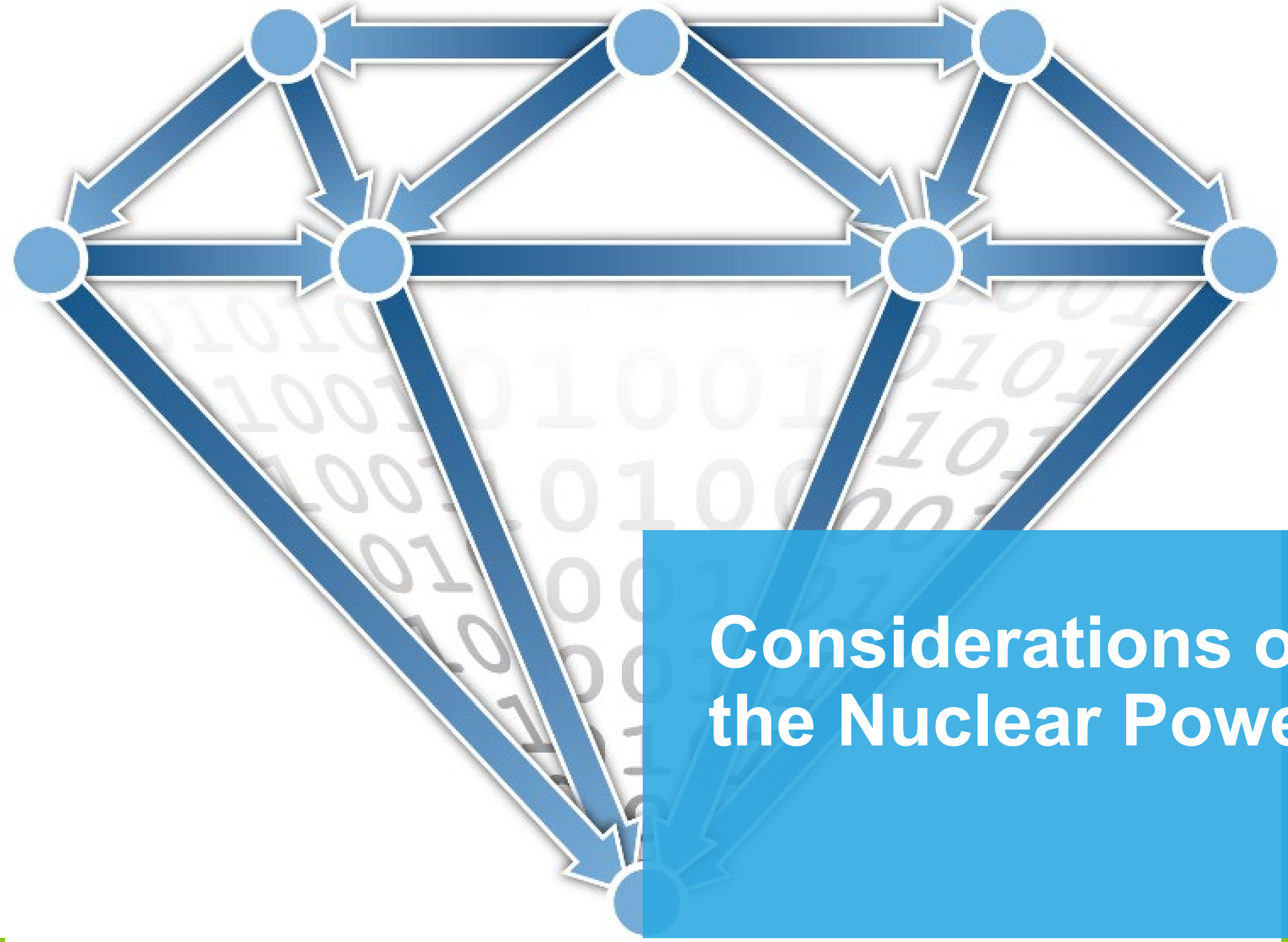
- Train even larger LLMs
- Auto-completion and sequencing of procedures and work instructions, including generation of entire work steps from unstructured text
- Open Domain Q&A- “Query” large Nuclear texts for answers (e.g. FSAR, design documents, etc.)
- Conversational AI user interface
- Automatic summarization of site schedules and daily issues

Questions
?

jerrold@nuclearn.ai
brad@nuclearn.ai

NuclearN

https://nuclearn.ai



Feb 10, 2022

Ahmad Al Rashdan
Senior R&D Scientist

Considerations of Data Integration in the Nuclear Power Industry

Idaho National Laboratory

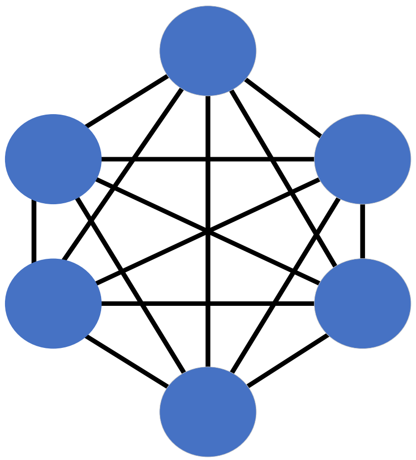
Acknowledgements

- Idaho National Laboratory:
  - *Chris Ritter*
  - *Jeren Browning*
  - *John Darrington*

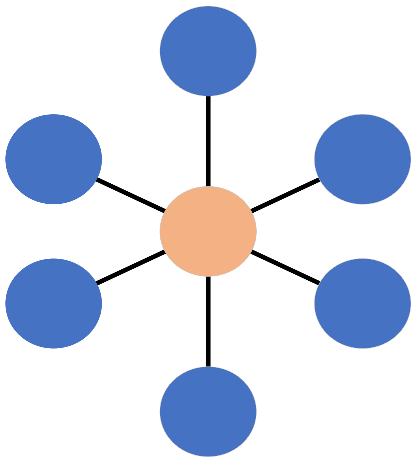
Current Practice

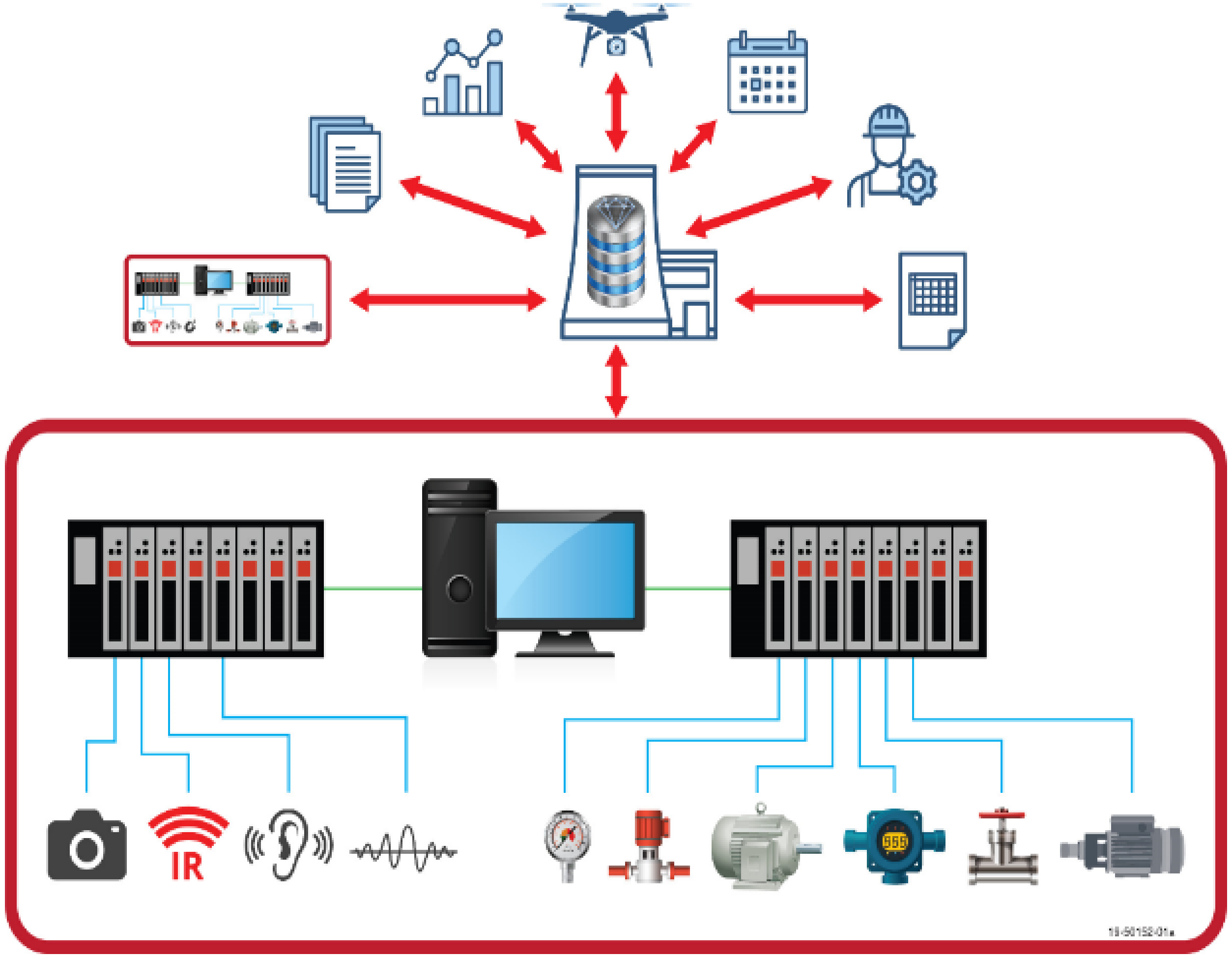
- Nuclear power plant data are stored in isolated forms in different systems with many structures and tools that are used independently.
- No significant data and methods exchange across the industry for research

Mesh

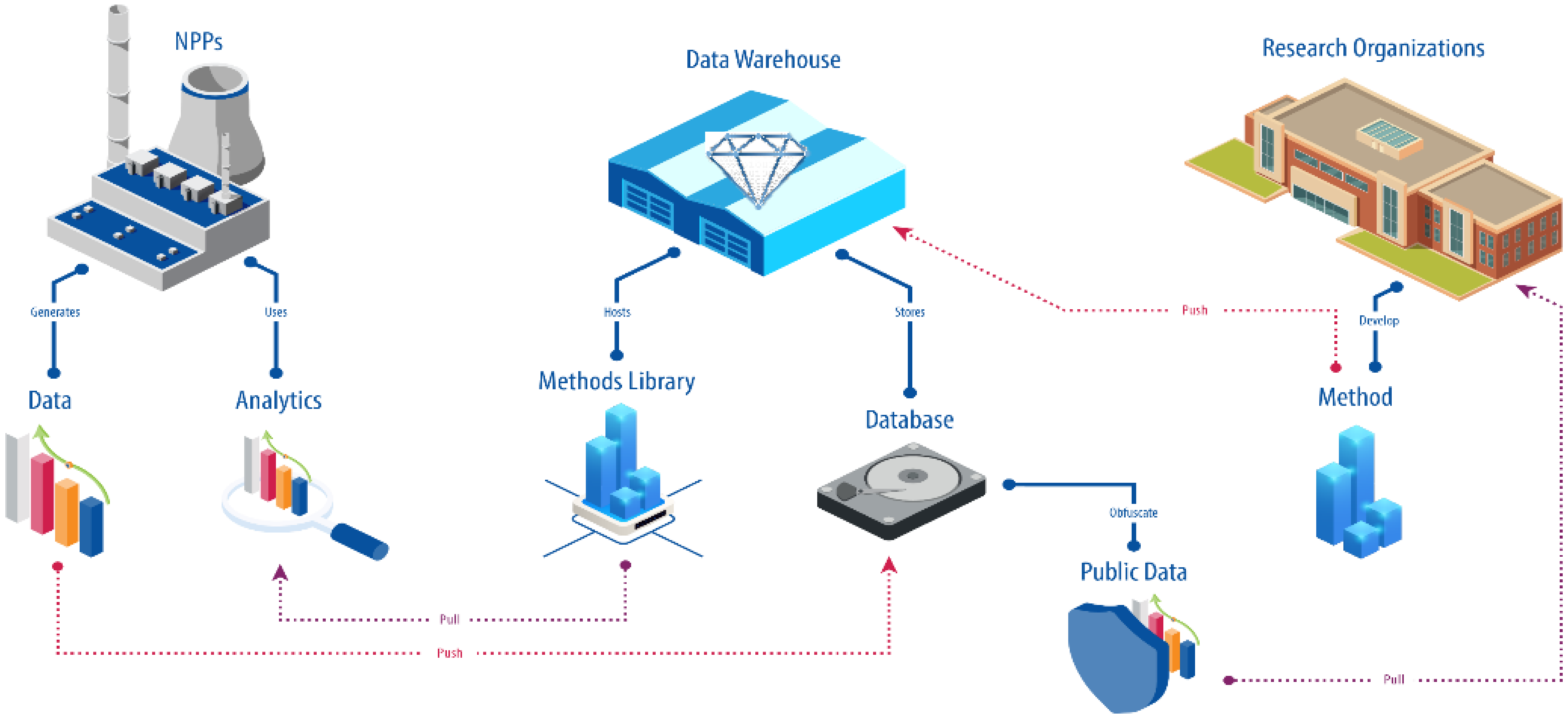


Star





End-State Vision



Closing the Gap

- DIAMOND is a data model that was developed to enable data sharing across
  - various nuclear power plant data and tools into one data warehouse.
  - the nuclear power industry and other stakeholders including research community.

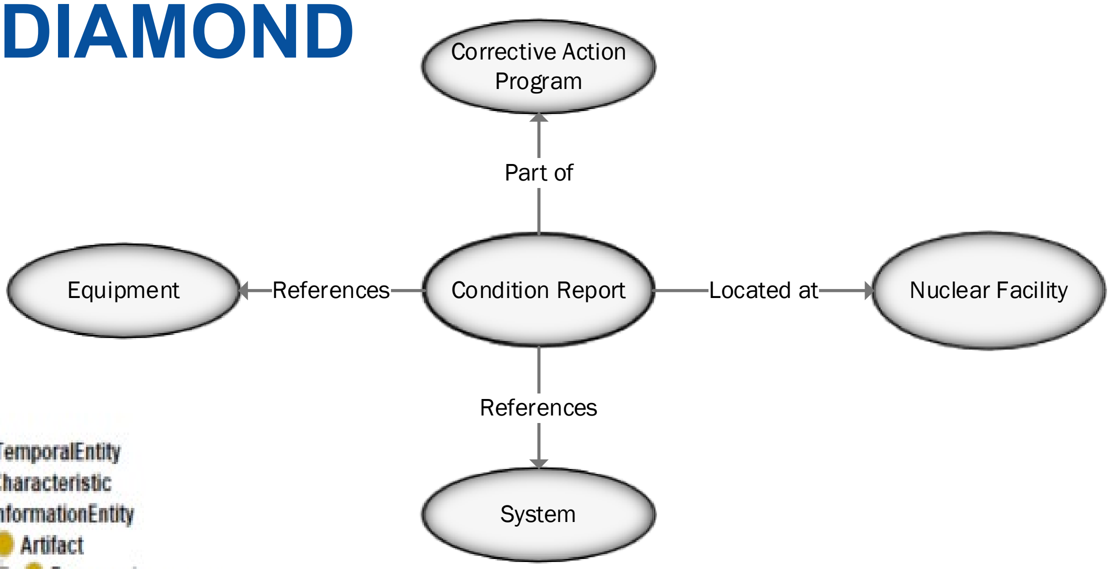
  https://github.com/idaholab/DIAMOND
- Deep-Lynx is an intelligent data warehouse tool that manages data in a centralized schema.
  - It provides users the ability to holistically query and understand data via the defined schema.

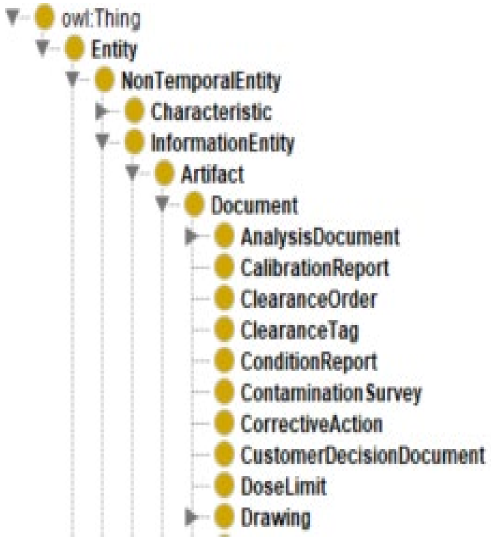
  https://github.com/idaholab/Deep-Lynx

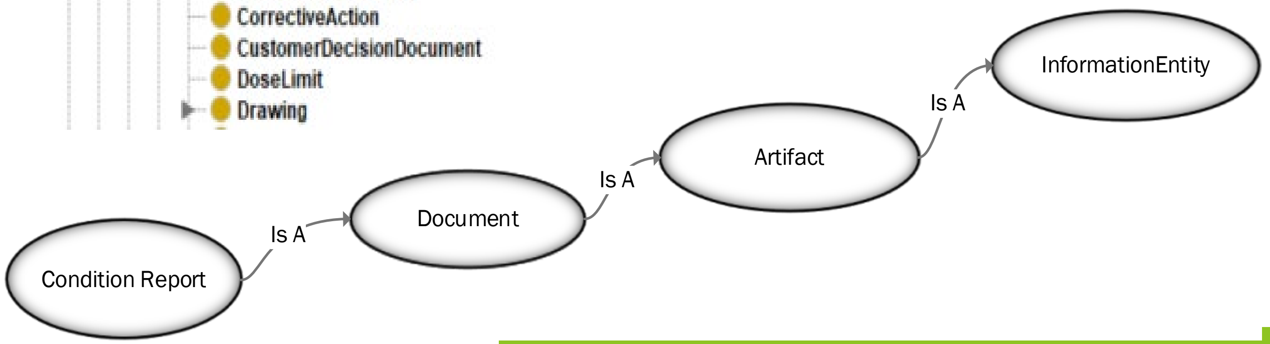
Data Integration Aggregated Model and Ontology for Nuclear Deployment- DIAMOND

- Consists of classes, object properties (relationships), and data attributes that are incorporated into a hierarchical tree structure.
- Adopted commonly used models such as Basic Formal Ontology (BFO) and Lifecycle Modeling Language (LML).
- Based on an evolving-development approach, meaning it was established with a core set of data objects and is populated with a preliminary level of detail.

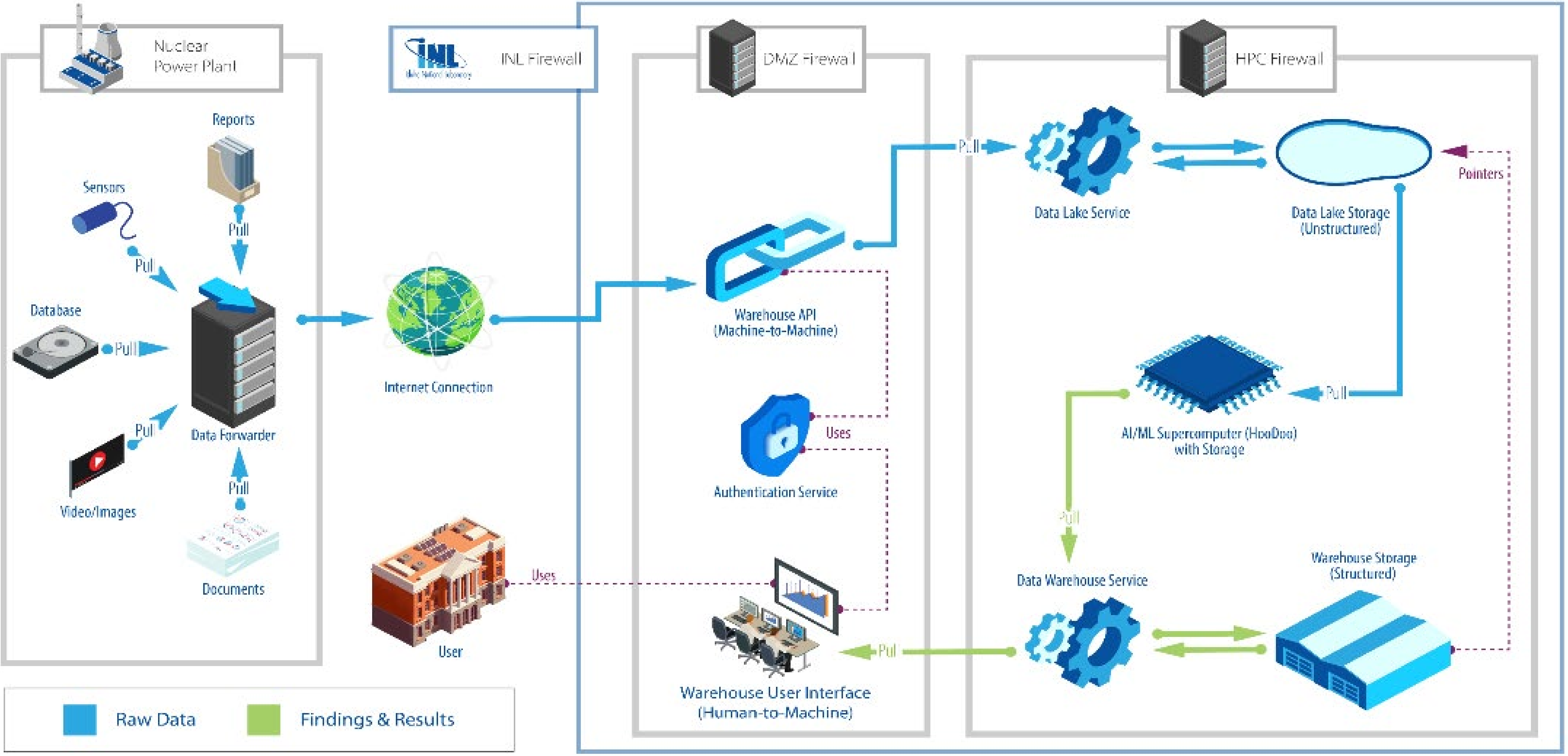
https://github.com/idaholab/DIAMOND







What's next?



Nuclear Power Plant
INL Firewall
DMZ Firewall
HPC Firewall
Reports
Sensors
Pull
Pull
Database
Pull
Data Forwarder
Pull
Video/Images
Pull
Documents
Internet Connection
User
Uses
Warehouse API (Machine-to-Machine)
Authentication Service
Uses
Warehouse User Interface (Human-to-Machine)
Pull
Pull
Data Lake Service
Data Lake Storage (Unstructured)
Pointers
AI/ML Supercomputer (HooDoo) with Storage
Pull
Pull
Data Warehouse Service
Warehouse Storage (Structured)
Raw Data
Findings & Results

Idaho National Laboratory

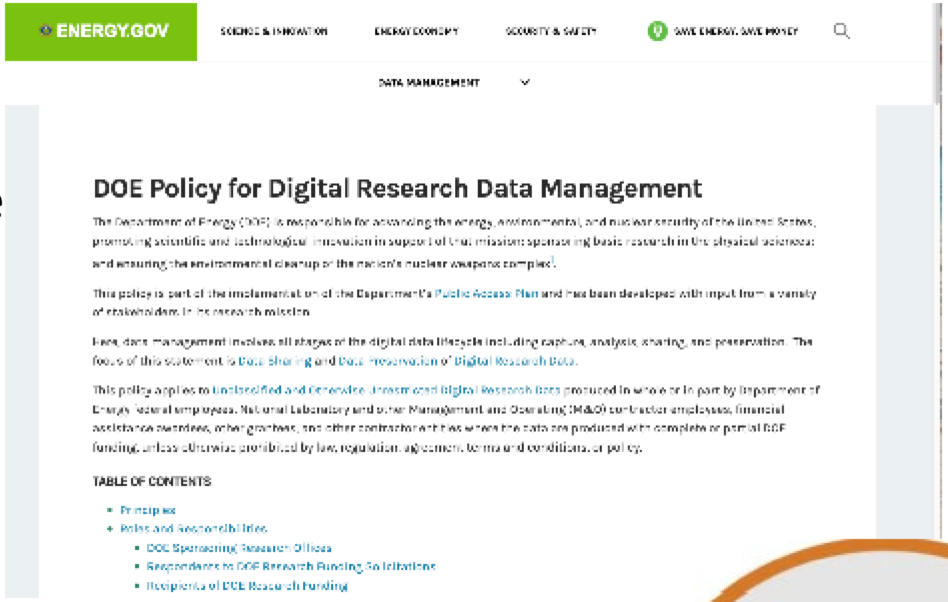
Research Data Management

27 January 2022

Eric Whiting
Division Director Advanced Scientific Computing
Nuclear Science & Technology

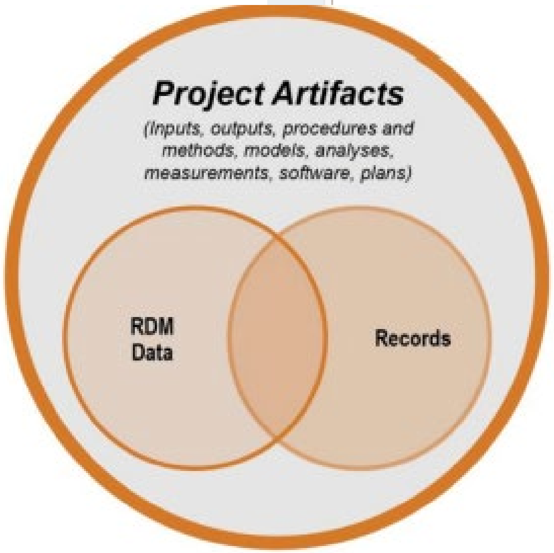
Idaho National Laboratory

Research Data Management: Why?



Requirement

- This policy applies to Unclassified and Otherwise Unrestricted Digital Research Data produced in whole or in part by Department of Energy federal employees, National Laboratory and other Management and Operating (M&O) contractor employees, financial assistance awardees, other grantees, and other contractor entities where the data are produced with complete or partial DOE funding, unless otherwise prohibited by law, regulation, agreement terms and conditions, or policy.

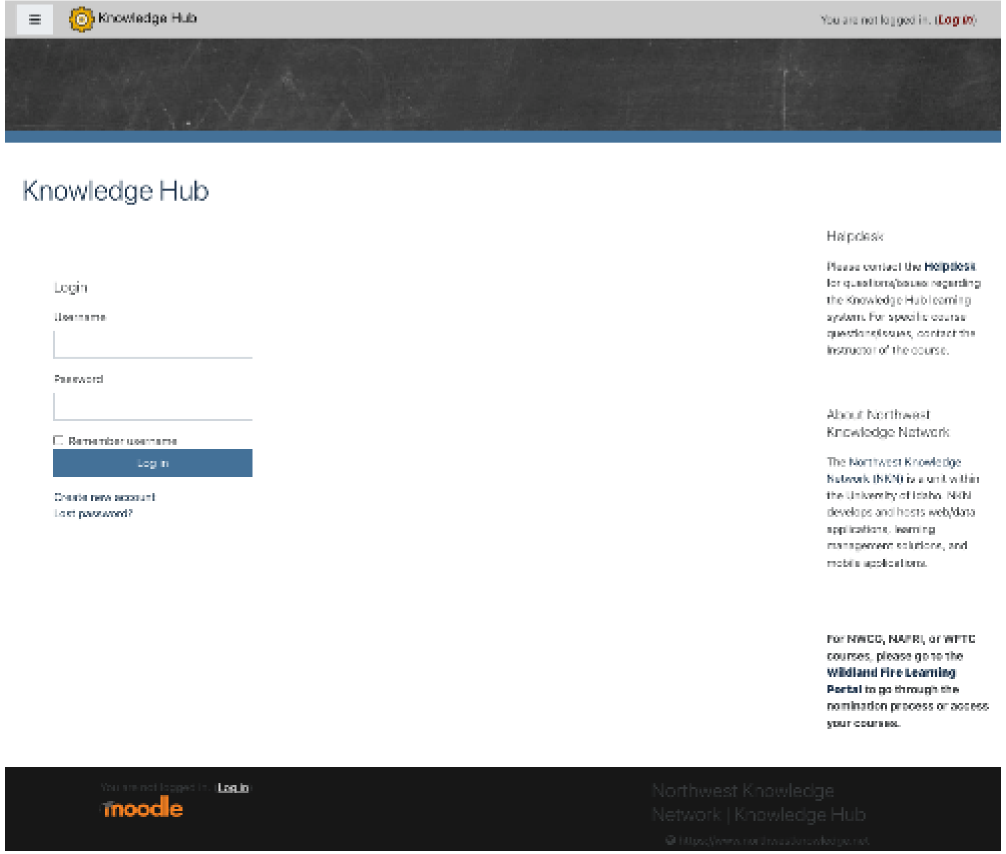


Findable, Accessible, Interoperable, Reusable

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.

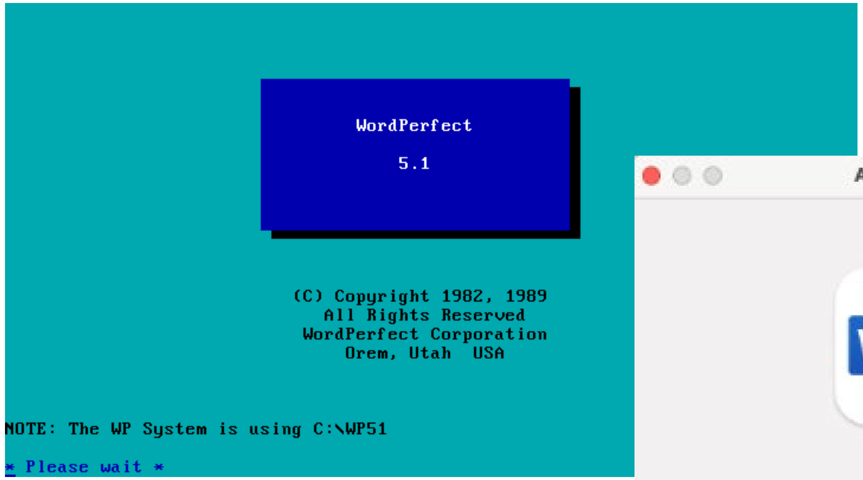
Findable, **Accessible,** Interoperable, Reusable

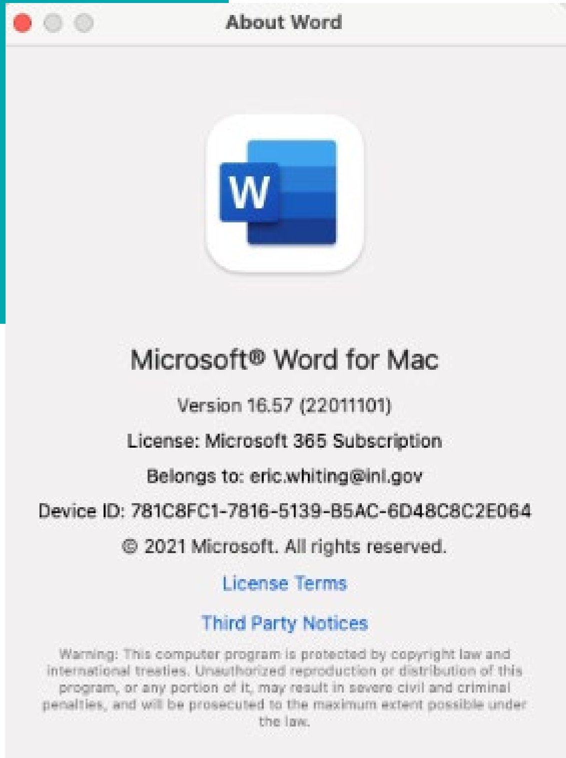
Once the user finds the required data, she/he/they need to know how they can be accessed, possibly including authentication and authorization.



Findable, Accessible, Interoperable, Reusable

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.





Findable, Accessible, Interoperable, Reusable

The ultimate goal of FAIR is to optimize the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

Research Data Management: How?

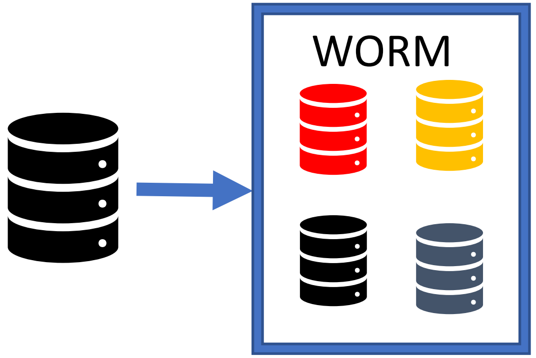
RDM often leverages and assortment of tools. Some researchers have program-specific repositories and policies, other RDM efforts are more ad-hoc. Storage locations include enterprise document management systems, cloud hosted platforms, scientific computing storage, tape drives, portable disk drives, thumb drives, and local instrument storage. It is unlikely that these solutions as implemented are fully compliant with DOE requirements and FAIR principles.

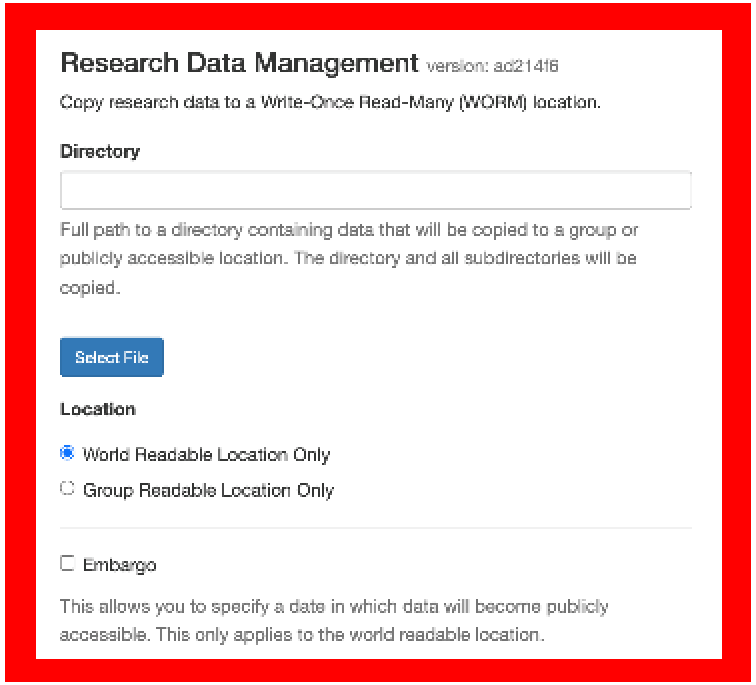
Efforts have been undertaken to manage INL enterprise data on a global scale with a ‘data lake’ architecture. This effort will include some aspect of RDM, but research data typically is of a size and format to make it incompatible with these systems.

INL Advanced Scientific Computing has deployed an initial prototype for perpetual storage of research data in order to meet some immediate needs for RDM. This puts data close to compute for analysis and potential future reuse.

High Performance Computing offers Perpetual Research Data Management and Storage

- HPC has created a **Write-Once Read-Many** (WORM) data storage system available through https://ondemand.hpc.inl.gov for storing and curating scientific data.
- Scientific data submitted will be maintained in perpetuity.
- A cryptographic hash is created with each submission to easily verify all data remain unaltered.
- An embargo access date on data can be provided at submission.
- To request permission to submit data to the WORM, send a message to hpcsupport@inl.gov.





HPC Research Data Management System for Scientific Data curation now available.

For more information:
Matthew.Anderson2@inl.gov

Research Data Management: Next Steps

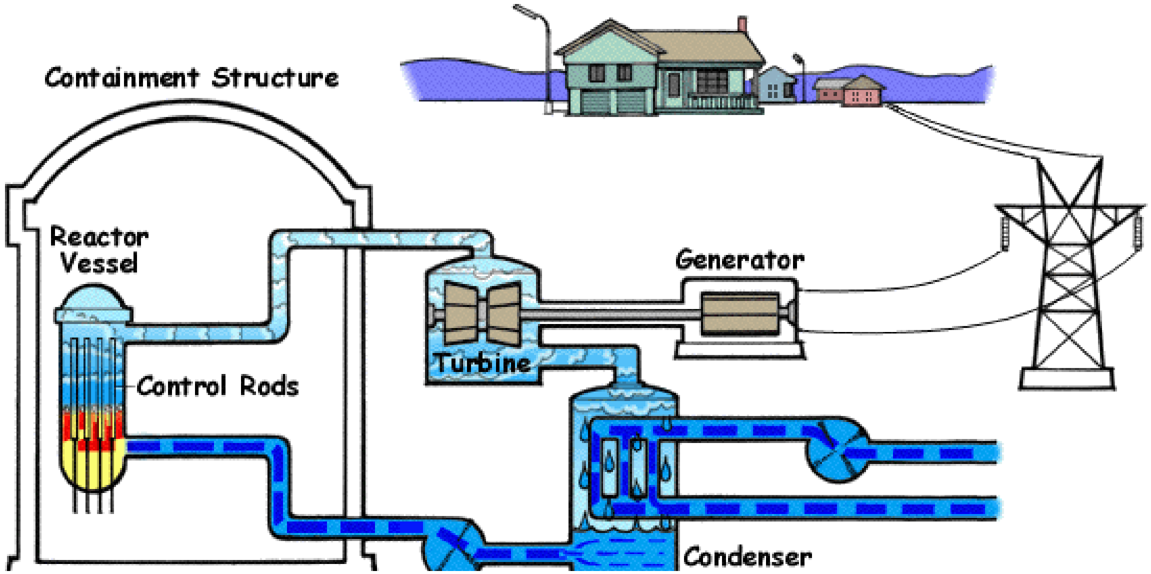
- Evaluate best practices from other DOE labs and programs
    - Northwest Knowledge Network
    - EDX
    - DataOne

- Capture INL needs and requirements

- Develop an INL RDM strategy

- Deploy simple RDM tools with minimal impact to workflows

PHYSICS-INFORMED MACHINE LEARNING FOR ENGINEERING APPLICATIONS WITH SPARSE DATA: BWR MOISTURE-CARRYOVER PREDICTION



HAOYU WANG
Nuclear Engineer
Argonne National Laboratory

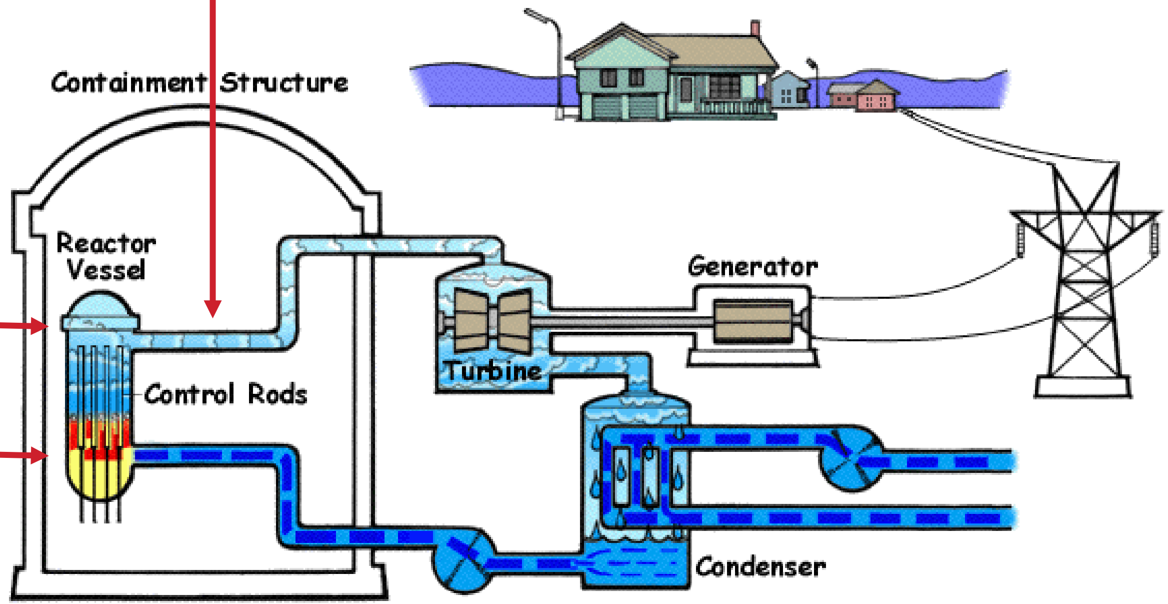
AI & ML Symposium 7.0
February 10, 2022

MOISTURE CARRYOVER AND EFFECTS

Un-separated liquid droplets:
Moisture Carryover (MCO)

Steam Drying

Coolant Boiling



Excessive MCO will cause:

- Higher impact and corrosion on turbine components;
- Elevated radiation dose to on-site personnel.

Goal:

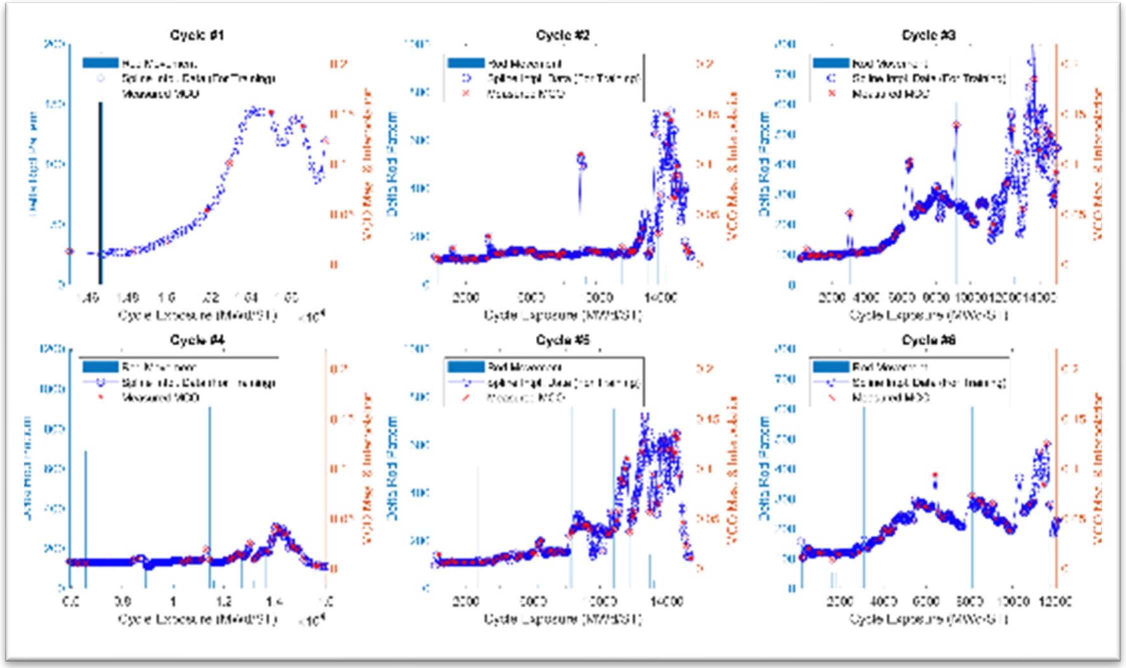
- Model MCO using plant measured data;
- Predict MCO level for un-started cycle.

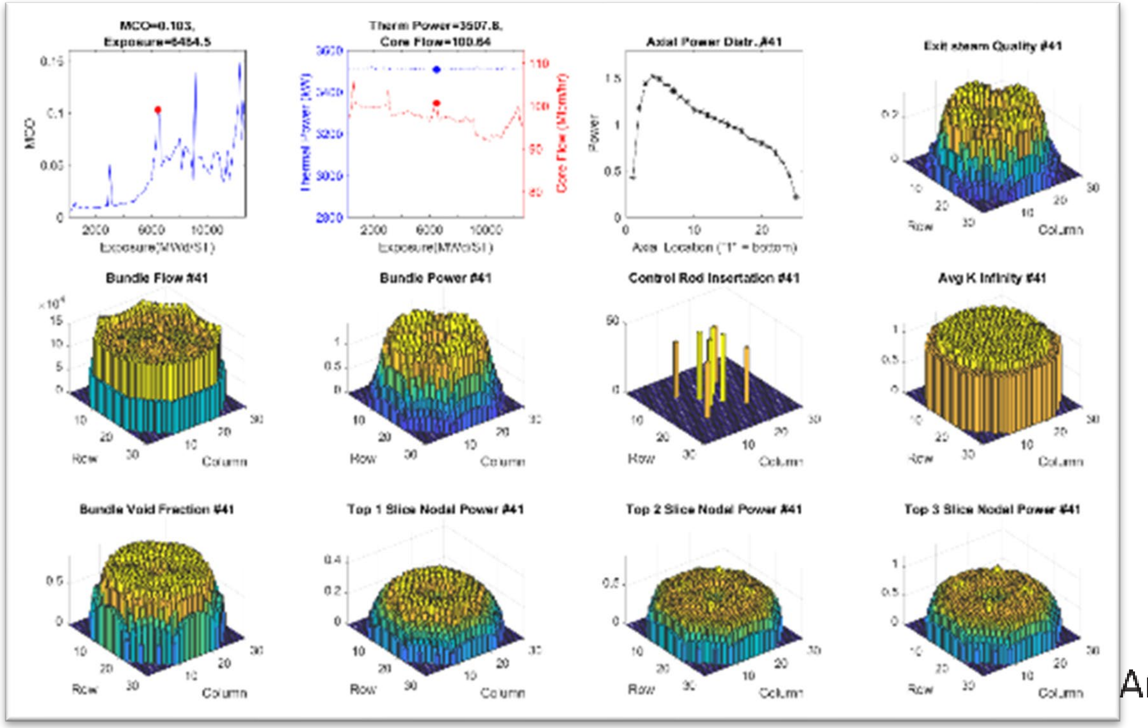
CHALLENGES

1. Limited and sparse entries of expensive data points
    - *6 completed cycles;*
    - *601 experimental measurements;*
    - *~$2,000 USD / measurement.*

2. Excessive number of candidate features
    - *7,000+ process variables;*
    - *Covers the power, steam quality, rod, flow distribution over entire core*

3. Need for accurate predictions for future un-started cycle



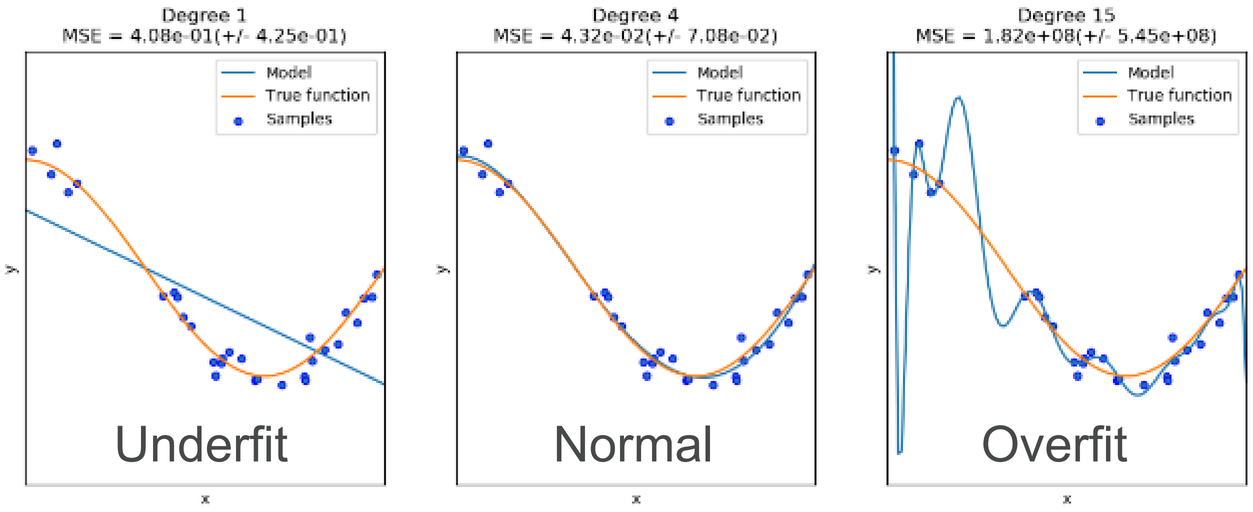


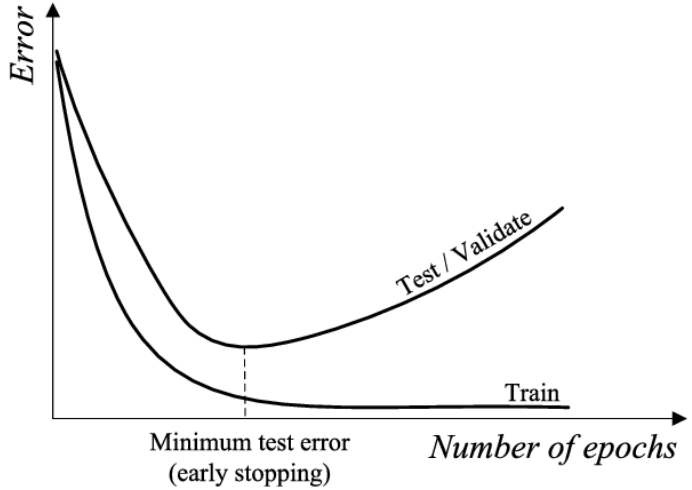
Argonne NATIONAL LABORATORY

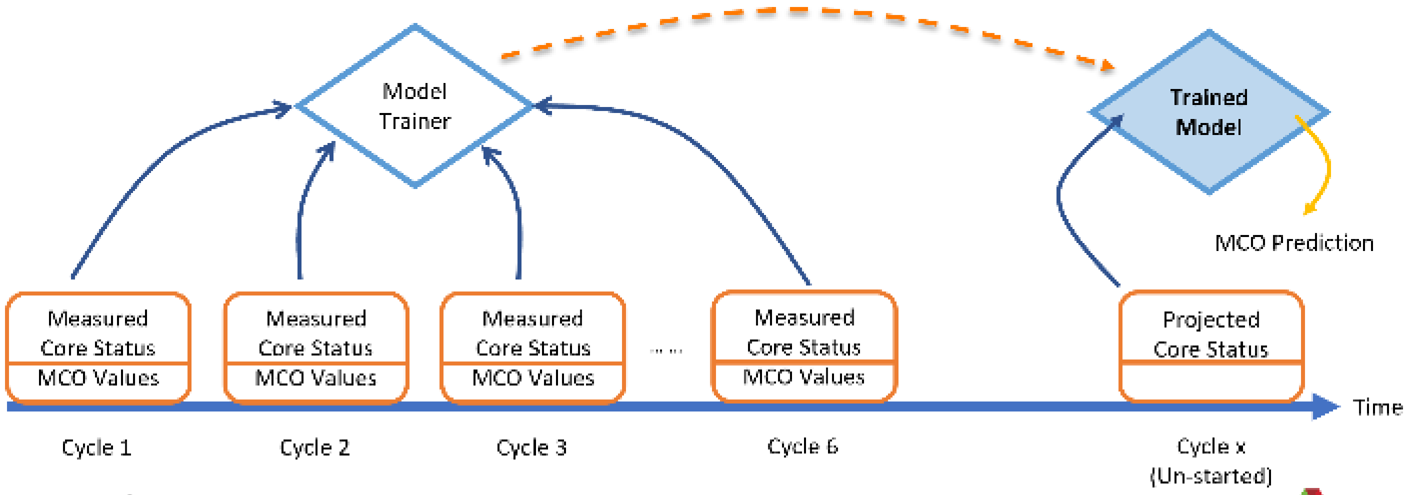
DANGER OF OVERFITTING

With limited amount of data entries, number of features and model complexity needs to be constrained.

In addition, the error balance between training and prediction needs to be considered.







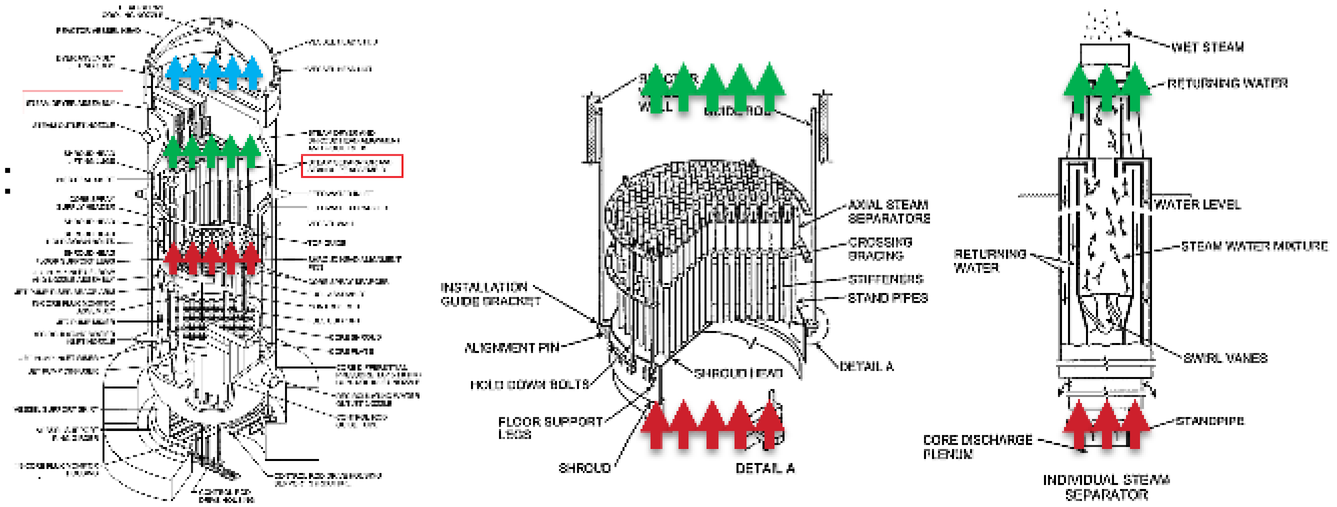
METHODOLOGY

1. Physics-informed feature and model selection:

Lower initial steam quality (***Q***),Higher ***MCO*** : $MCO \sim \frac{1}{Q^m}\ (m > 0)$

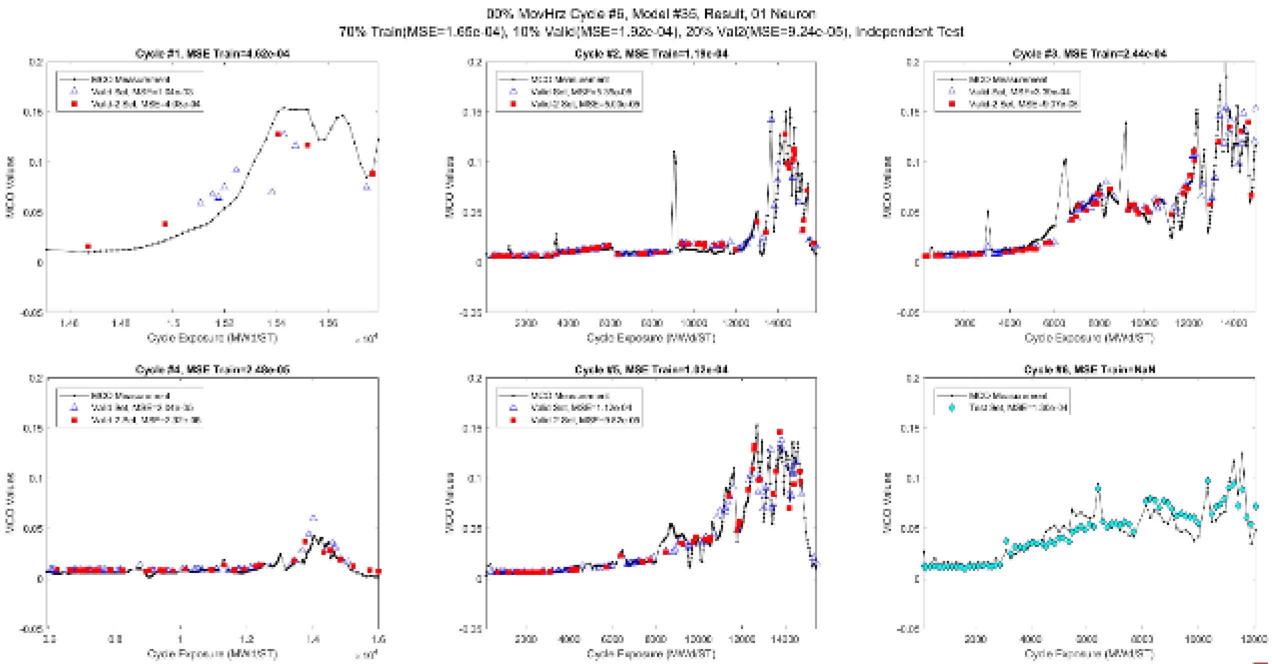
Too low or too high flow rate (***V_L***), Higher ***MCO*** : $MCO \sim \frac{1}{{V_L}^{n1}} + {V_L}^{n2} (n1, n2 > 0)$

Non-linear summation nature of MCO: Neural network



2. Hyper-parameter optimization (Genetic Algorithm), balancing training and prediction error:

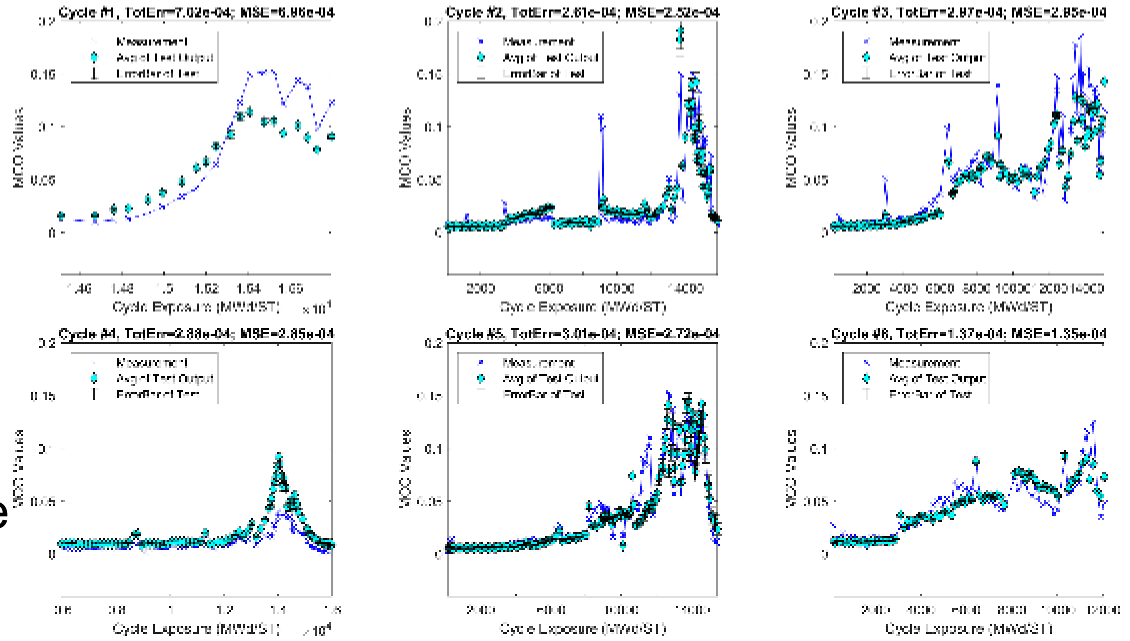
- Leave-one-out and Cross-test: train on 5 cycles, test on 1, then rotate;
- Optimize for overall minimum cross-test error.



RESULTS

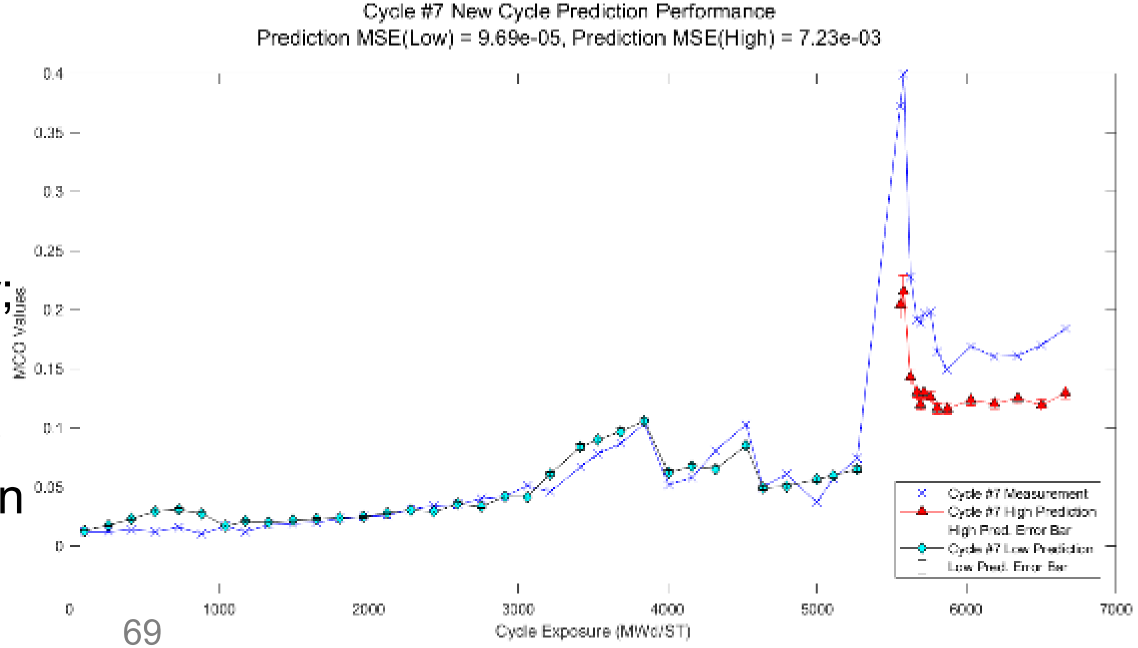
1. Leave-one-out and Cross-test result:

- Hyper parameters were optimized towards minimized overall cross-test error;
- Physics-informed feature with optimized hyper parameters can capture the baseline trends and spikes in each MCO trajectory



2. New cycle prediction:

- The trained model can predict the MCO in the training range (< 0.15%) with low error;
- The sudden spike is poorly predicted, which is caused by a severely asymmetric in-core flow and rod distribution never seen in the training data.



Argonne NATIONAL LABORATORY

CONCLUSION

- Plant experimental data + AI is solving real problems in nuclear energy:
  - Physics-informed feature selection on sophisticated systems;
  - AI modeling and hyper parameter optimization on sparse reactor data;
  - Target oriented cross-test scheme;
  - Even, prediction of the future.

- Challenges:
  - Data diversity and Model reliability;
  - Time and Cost during data collection, and cost-effectiveness;
  - New-physics supported by Data;

REFERENCES

- H. Wang, J. T. Gruenwald, J. Tusar & R. Vilim "Moisture-carryover performance optimization using physics-constrained machine learning." Progress in Nuclear Energy 135 (2021): 103691.

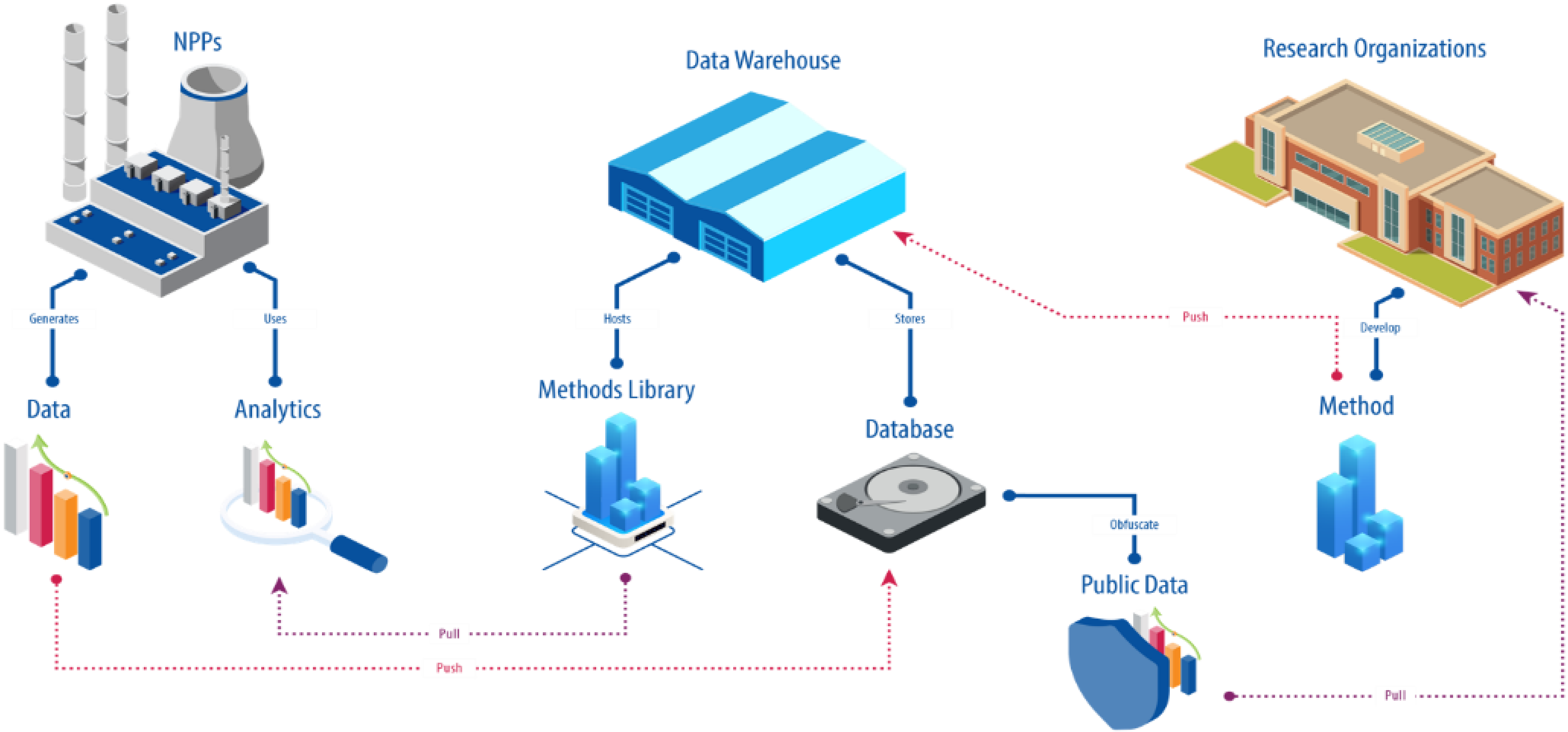
THANK YOU

Deceptive Infusion of Data (DIOD):
Novel Data Masking Paradigm for High-Value Systems

Arvind Sundaram & Hany Abdel-Khalik Purdue University
In collaboration with INL's Ahmad Al Rashdan & Mohammad Abdo

INL AI/ML Symposium 7.0, Feb 10, 2022

Data in Nuclear



Credits: Ahmad Al Rashdan, Idaho National Laboratory

DIOD Paradigm: Key Objectives

› How to successfully mask industrial data while promoting collaboration?

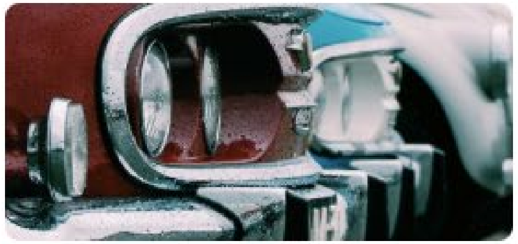
– Protect privacy of owner

– Preserve data utility with respect to AI task

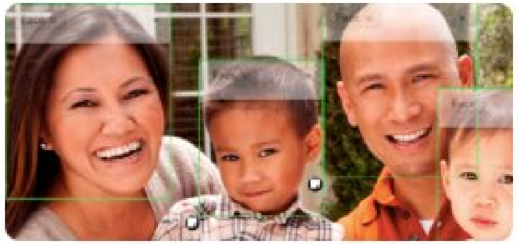
– Prevent reverse-engineering efforts, i.e., control what you want them to see

https://www.tolpagorni.com/blog/professional-networking-and-collaboration

Open-Source Data



Financial Tweets
David Wallach



Face Detection in Images
DataTurks

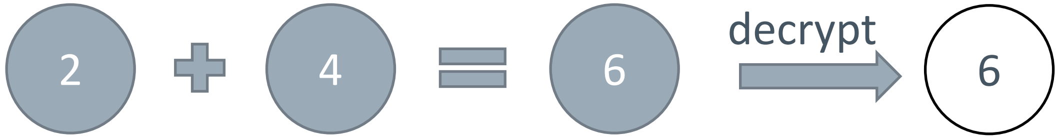
Star Trek Scripts
Gary Broughton

Avocado Prices
Justin Kiggins

www.kaggle.com, https://openml.org

Current R&D efforts

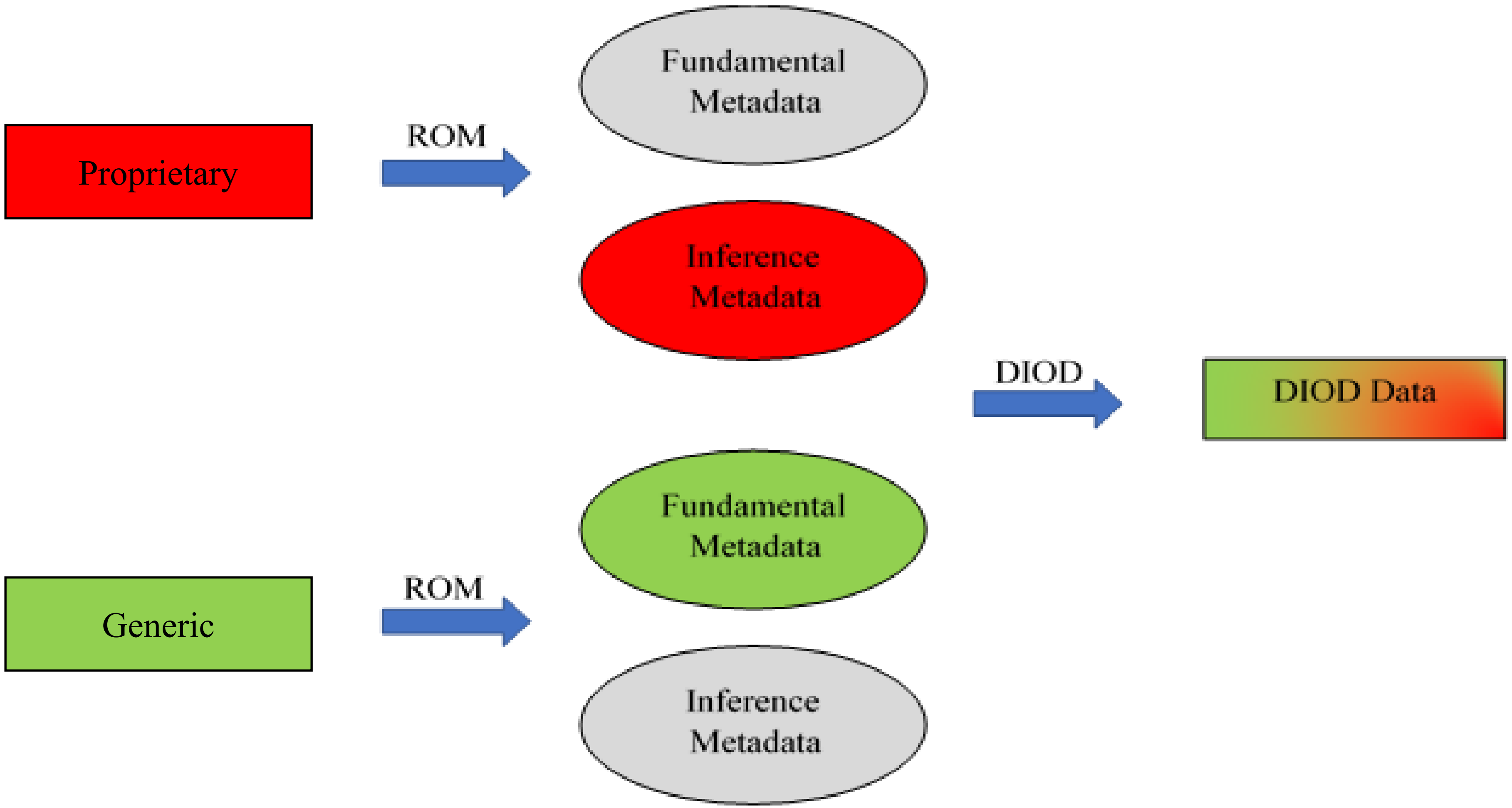
- Industrial Data
  - Differential privacy
    - Insert noise to cause uncertainty in data
    - Affects the statistical properties of the data
  - Privacy-preserving computation
    - Allows users to perform computations on data in encrypted form
    - Encryption is extremely expensive, not scalable for vast amounts of data



DIOD Data Masking Paradigm

- Splits dataset into fundamental metadata and inference metadata
  - Fundamental metadata denotes information pertaining to system identity
  - Inference metadata denotes information relevant for target AI/ML task
- Obfuscates proprietary system identity by mounting inference metadata onto fundamental metadata of a different generic system; generate DIOD version of data
- Cannot reverse-engineer DIOD data to decipher system identity as transformation is one-way
- Efficient and scalable after an initial one-time investment into constructing ROMs
- Can be applied to obscure sensitive data while maintaining inference – classification, regression, clustering etc.

Masking: Fundamental Metadata



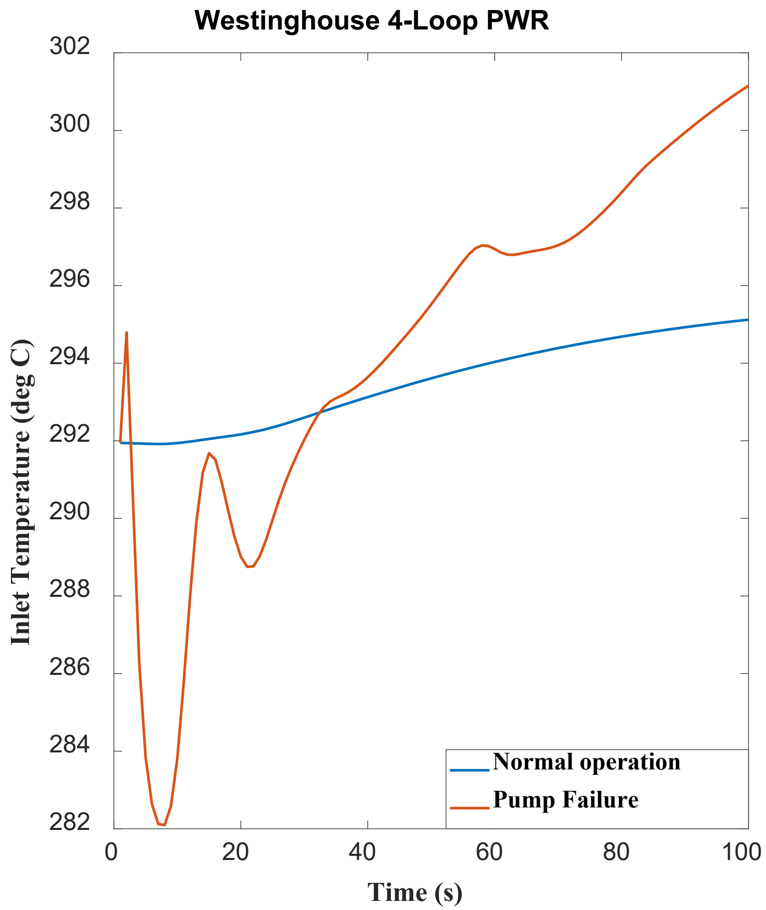
Mutual Information in DIOD

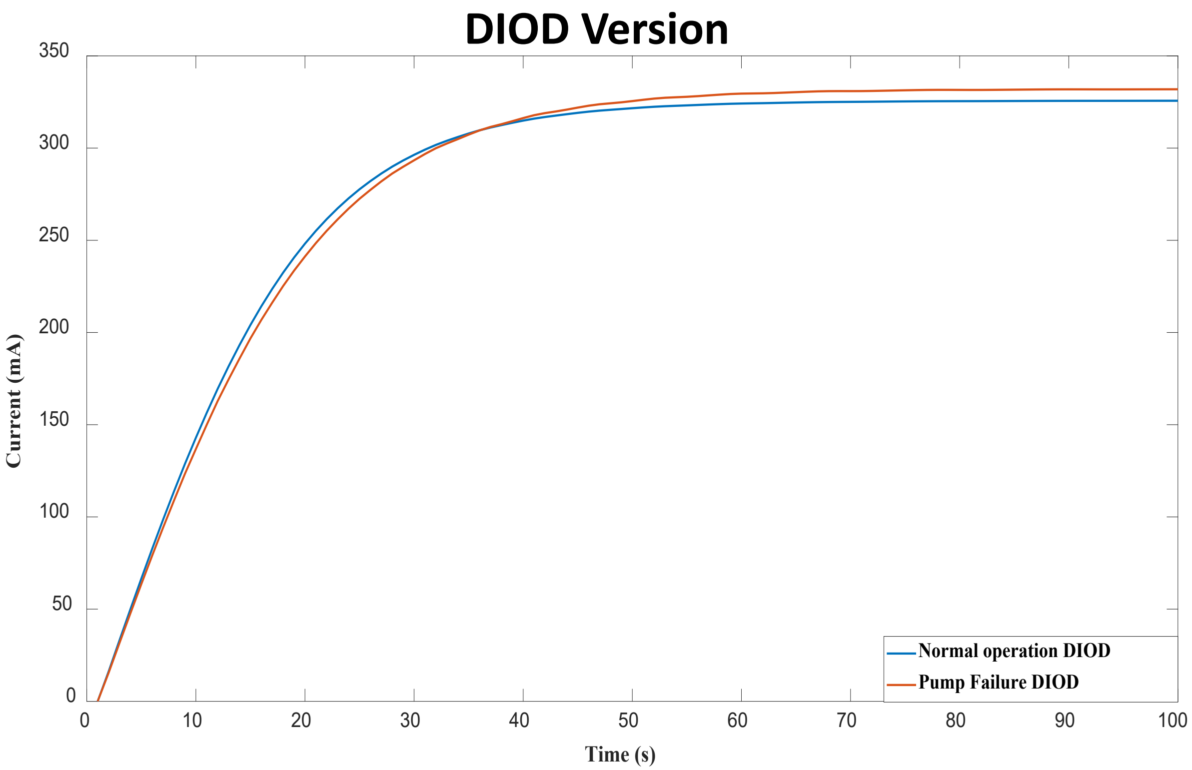
- Mutual information denotes the average gain in information about one quantity with knowledge of another

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

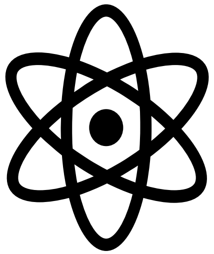
- Invariant to addition or removal of metadata irrelevant to the classification task
- Invariant to invertible transformations of the metadata
- Ensures same theoretical inference on original and DIOD version
- Applications
  - Classification: Preserve mutual information between original dataset and corresponding labels in the DIOD version
  - Regression: Preserve mutual information between original dataset and inferential parameters in the DIOD version

Example Results





Example Results



$\xrightarrow{SINDy}$

$$C_1 = \frac{\rho - \beta}{\Lambda}$$
$$C_2 = \frac{\beta}{\Lambda}$$
$$C_3 = \lambda$$
$$C_4 = -\lambda$$

β

Λ

$\lambda^{\#}$

$\rightarrow$

$$g_1(\beta) = m$$
$$g_2(\Lambda) = l$$
$$k^* = k$$
$$f(k^*) = n$$

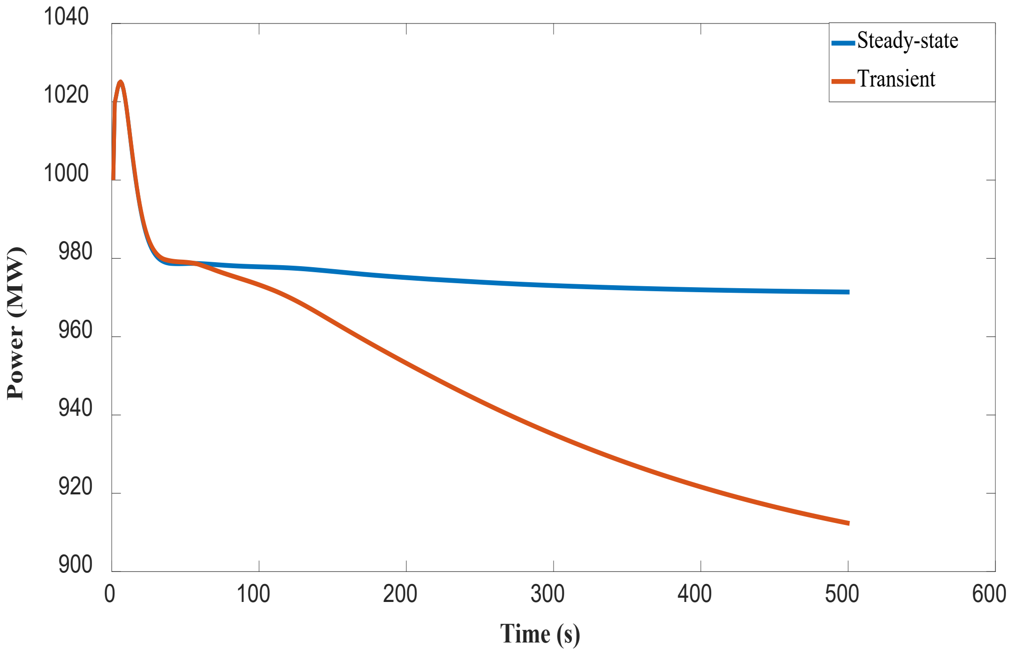
$\rightleftarrows$

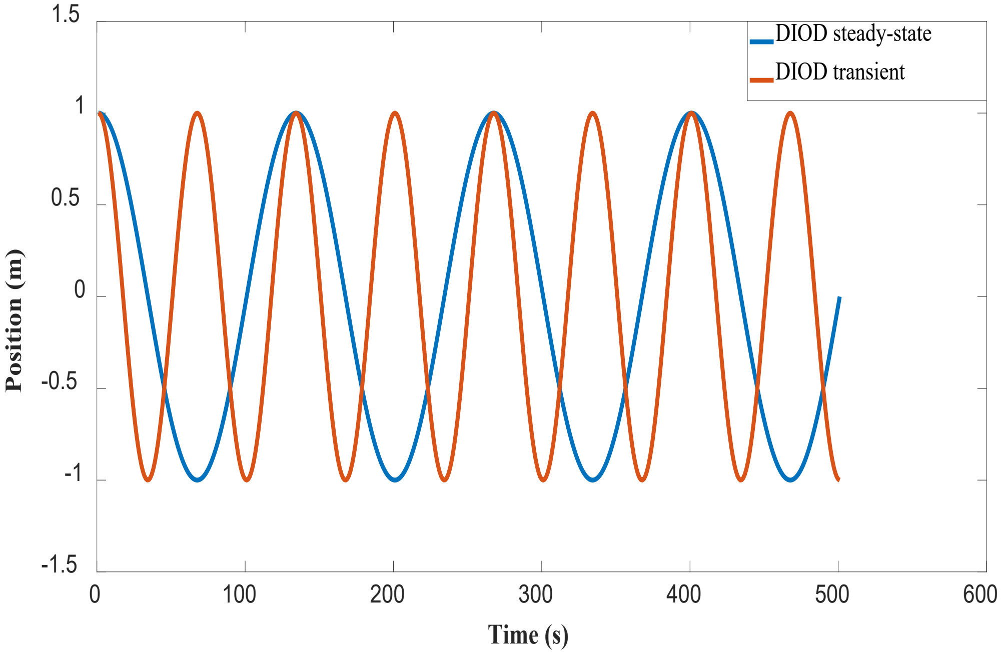
$$\dot{P} = \frac{(\rho - \beta)}{\Lambda} P + \lambda C$$
$$\dot{C} = \frac{\beta}{\Lambda} P - \lambda C$$

$$m\ddot{x} + l\dot{x} + kx + nx^3 = 0$$

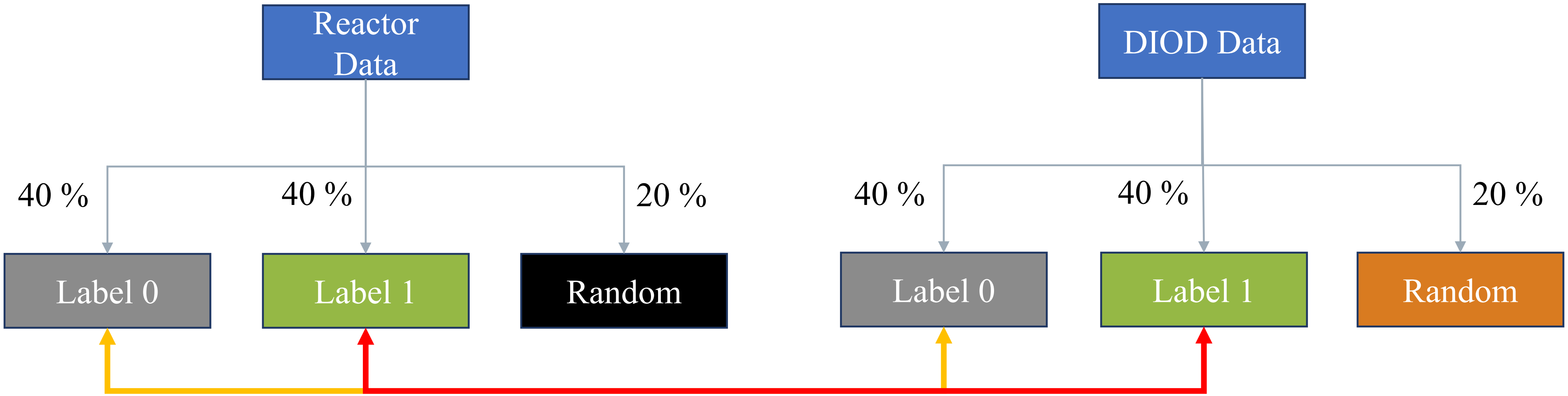
#Suppose λ is irrelevant to the classification task

Preliminary Results: Inference Metadata



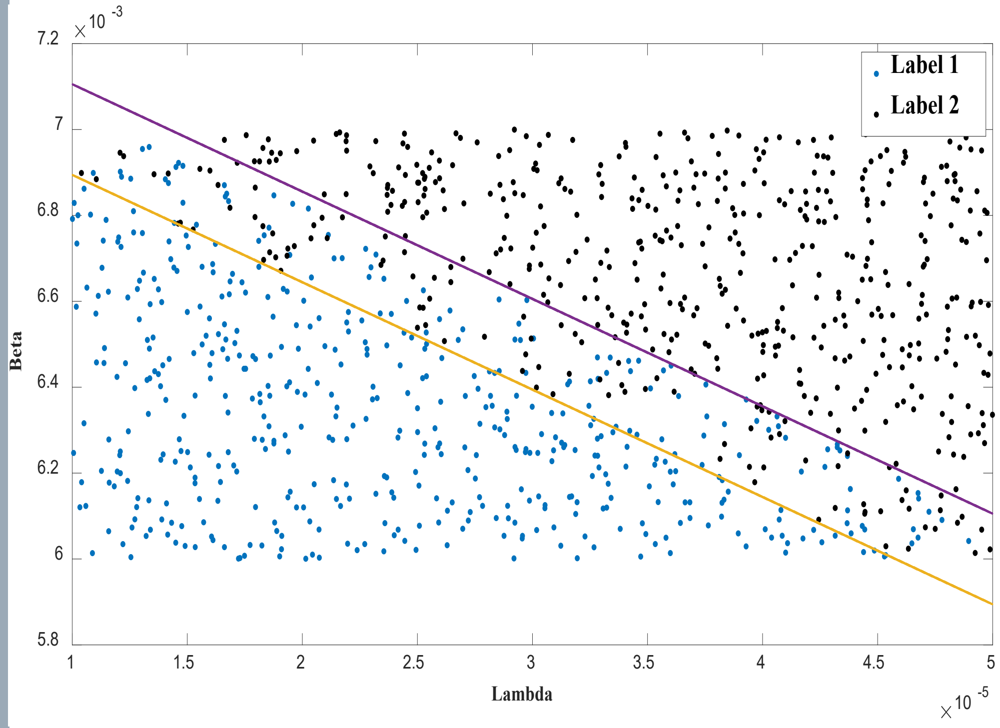


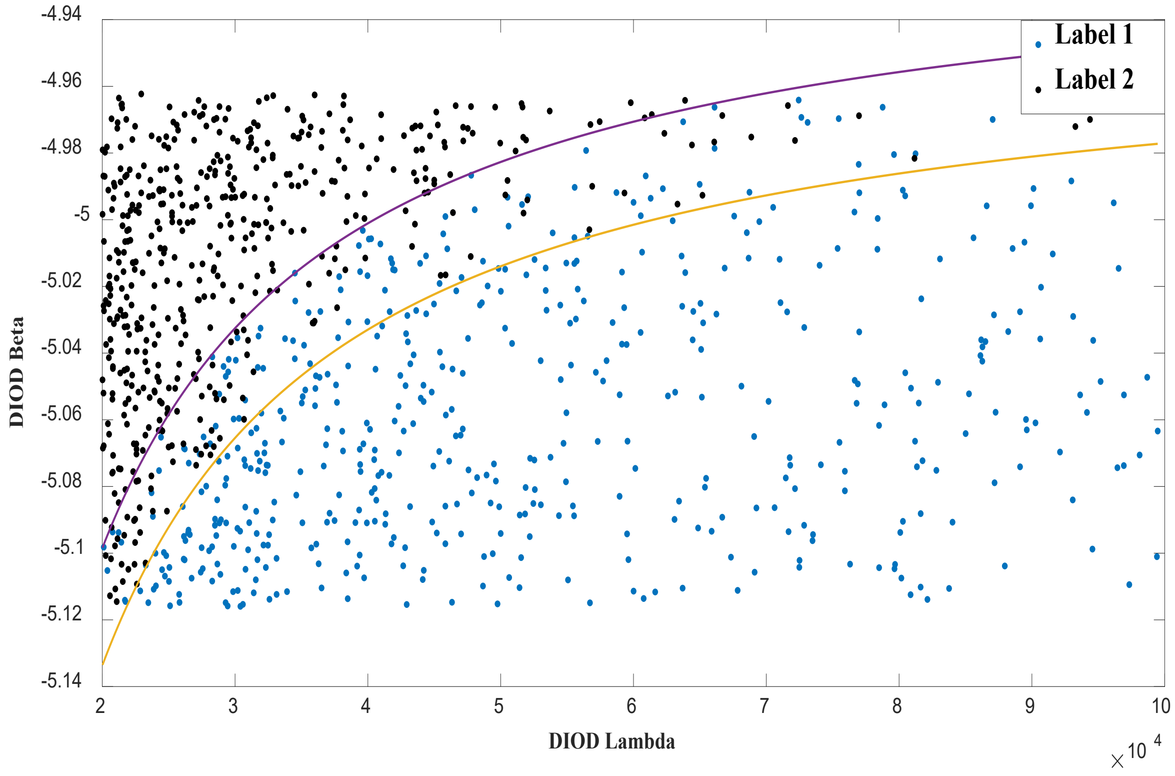
Example Results for Classification:



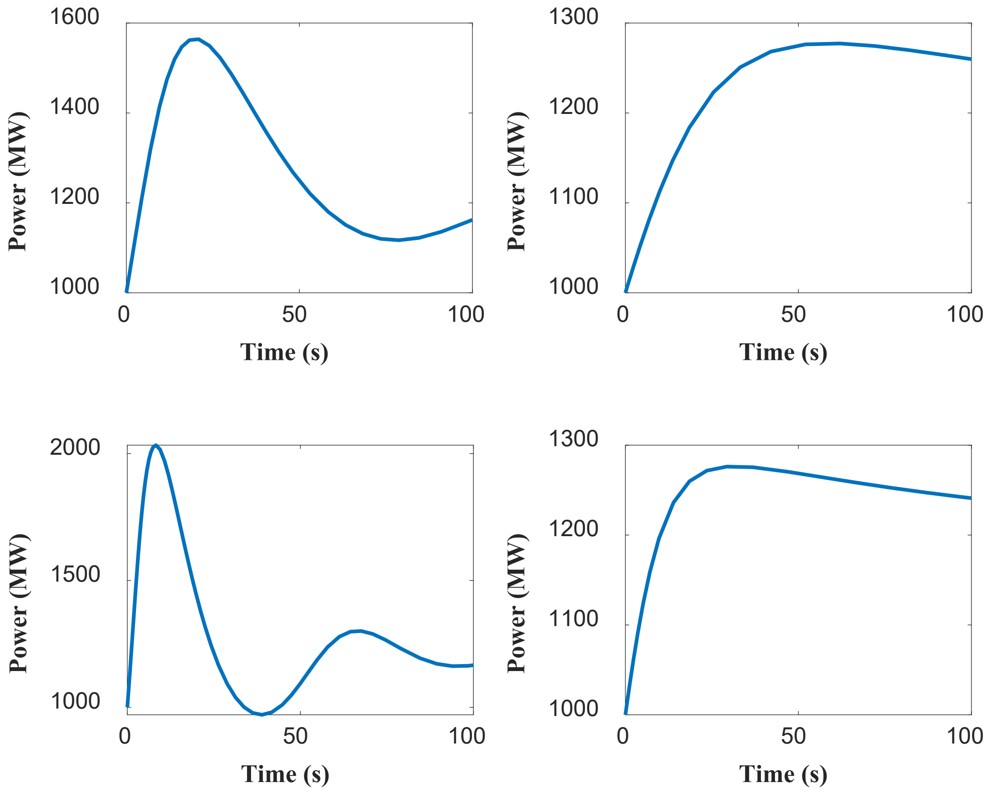
Inference is preserved for the task of classification

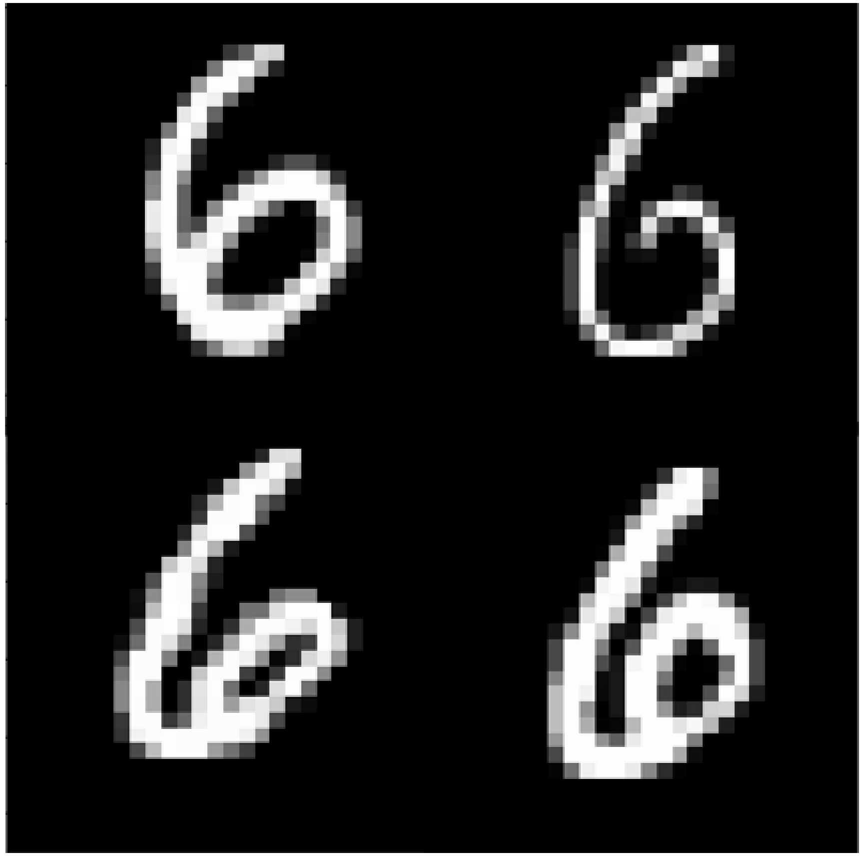
Example Results for Classification:





Example with Images



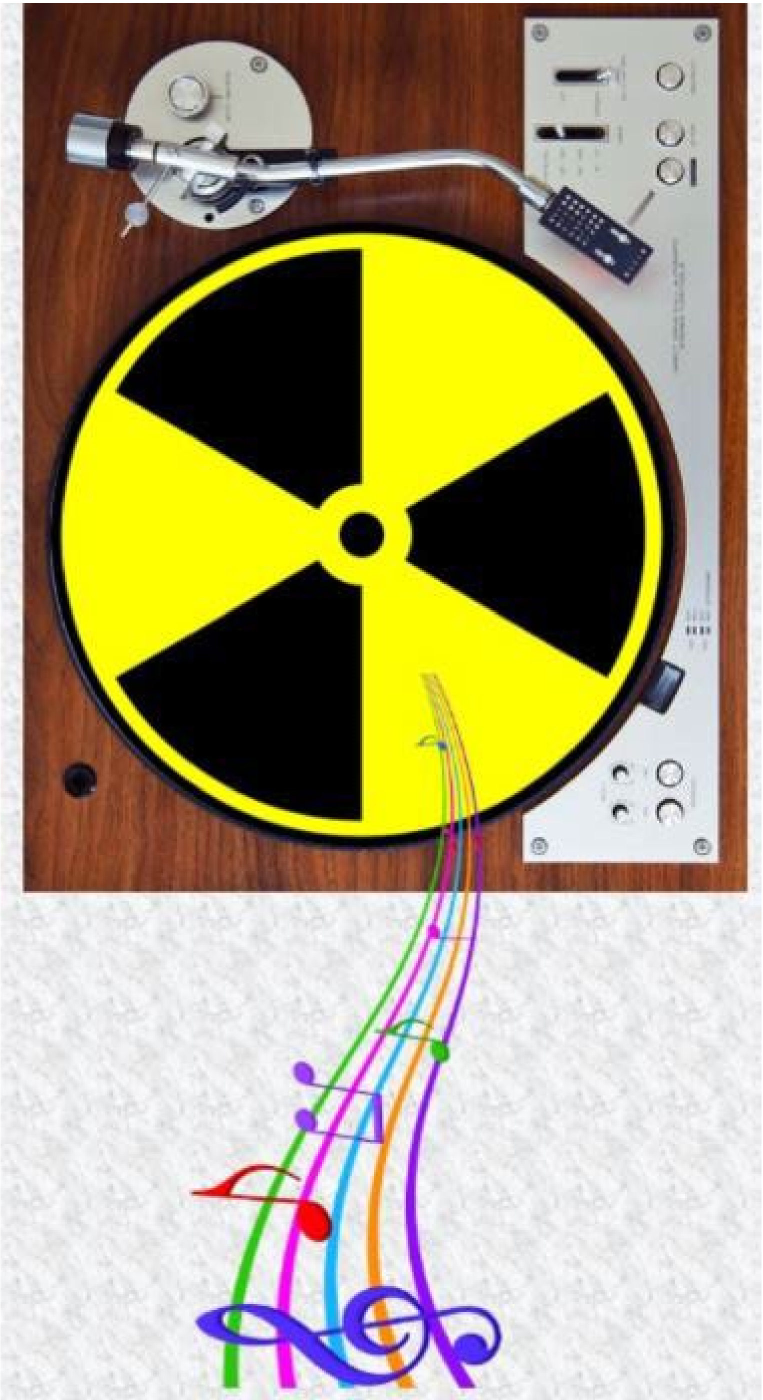


Publications

- Arvind Sundaram, Hany S. Abdel-Khalik, and Ahmad Al Rashdan, "Deceptive Infusion of Data (DIOD) for Nuclear Reactors," *Transactions of the American Nuclear Society*, **125**(1), pp. 264-266, December 2021.

- Arvind Sundaram, Hany S. Abdel-Khalik, and Ahmad Al Rashdan, "Deceptive Infusion of Data (DIOD): A Novel Data Masking Paradigm for High-Valued Systems," *Nuclear Science and Engineering*, November 2021 (under review)

- Arvind Sundaram, Hany S. Abdel-Khalik, and Mohammad G. Abdo, "Preventing Reverse-Engineering of Critical Industrial Data with DIOD," *Nuclear Technology*, November 2021 (under review)



Analysis and Handling of Big Data in Cosmology: AI/ML to the Rescue

FEBRUARY 10, 2022

Katrin Heitmann
Argonne National Laboratory

Next-generation Cosmological Surveys

Exceptional data — exceptional challenges

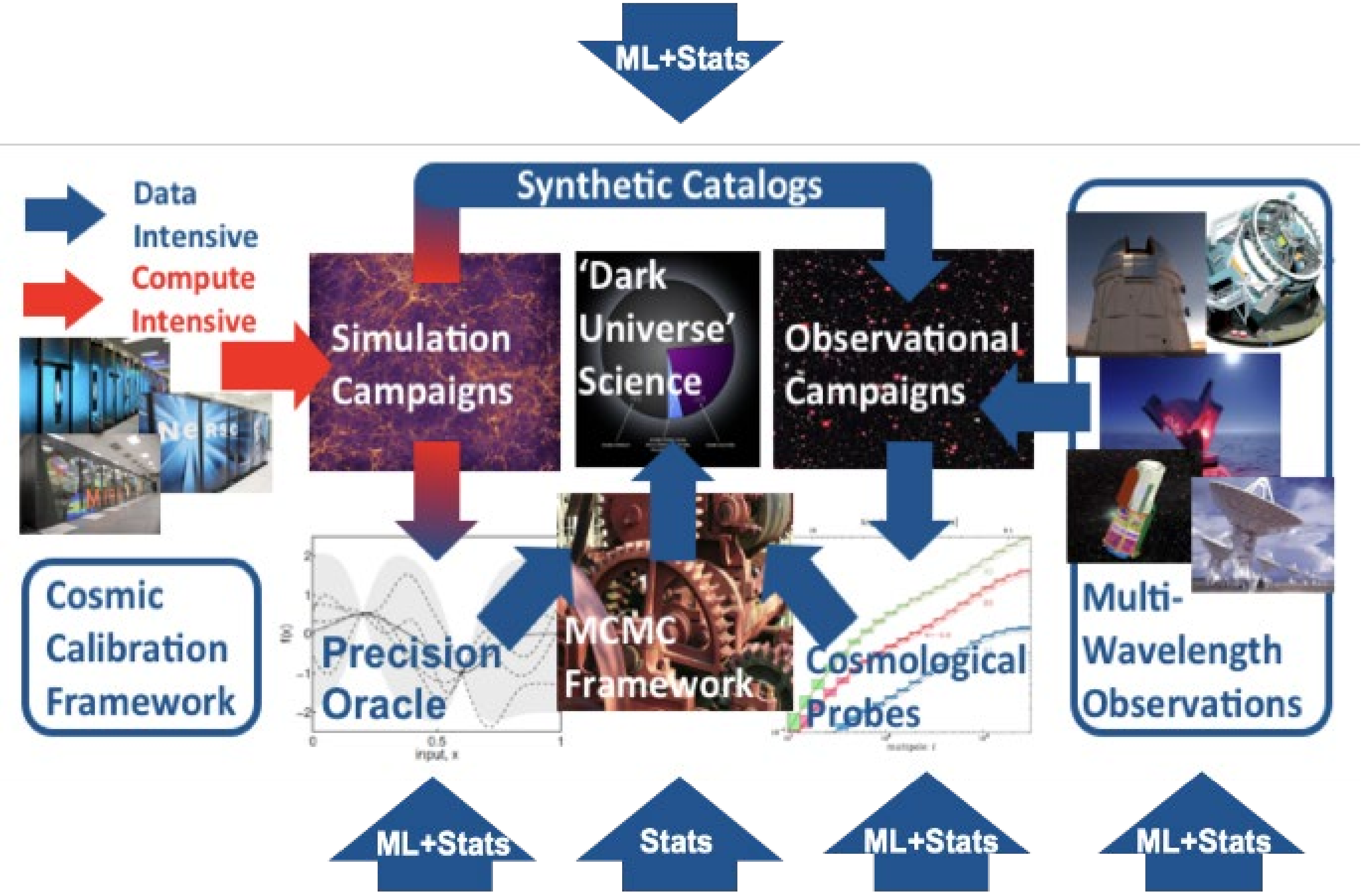
- Upcoming surveys such as Rubin's LSST, Euclid and Roman will collect a treasure trove of data
- Mining the data and interpreting them will be a major challenge
- For cosmology: Modeling and simulations will be the key to pushing our understanding of the dark Universe to the next level
- On the horizon for help: Exascale supercomputers and innovative AI/ML methods

- Hardware/Hybrid Accelerated Cosmology Code (HACC) and CRK-HACC have been developed to run on all currently available computing platforms **at scale**
- Large volume/high resolution gravity-only simulations and hydrodynamics simulations to model large-scale survey data
- Aurora will arrive at Argonne in 2022 to enable new **extreme-scale** simulations

AI/ML for Cosmological Surveys

New tools for measurements, predictions, and analysis

- Size and complexity of survey data sets drives AI/ML requirements
- Applications include image classification, lens characterization, fast sky catalog/image generation, fast predictions for summary statistics, systematics identification and mitigation, likelihood estimation, ---
- ‘AI at Scale’: Need to speed up current state-of-the-art by orders of magnitude



Ubiquity of AI/ML techniques in cosmological survey workflows

Precision Emulation

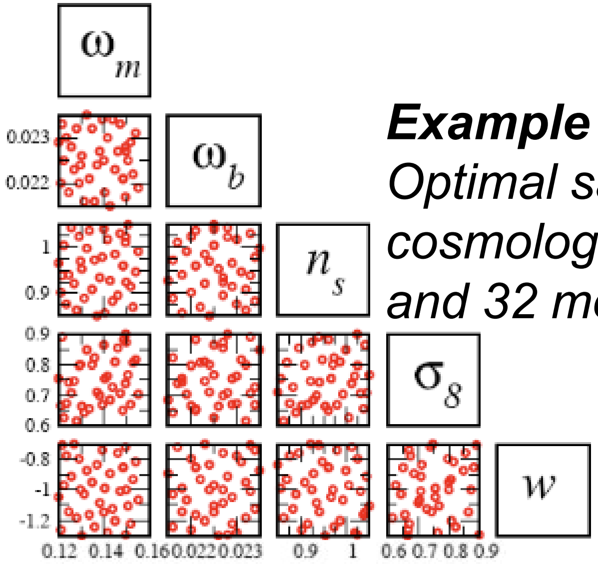
Transforming large simulation suites into precision predictions

- **Challenge:** To extract cosmological constraints from observations, need to run Markov Chain Monte Carlo code; input: > 100,000 predictions
- **For nonlinear probes (clusters, small scale predictions …)**: Expensive simulations are needed to achieve the required accuracy; while we can generate O(100) simulations, 100,000 would take years
- **Current strategy:** Fitting functions, accurate at the 10% level, need 1%!
- **Our alternative:** Emulators, fast prediction schemes built on a manageable set of high-accuracy simulations
- **"Ingredients":** Optimal sampling methods for model selection, efficient representation of the simulation outcome, powerful interpolation scheme

Precision Emulation

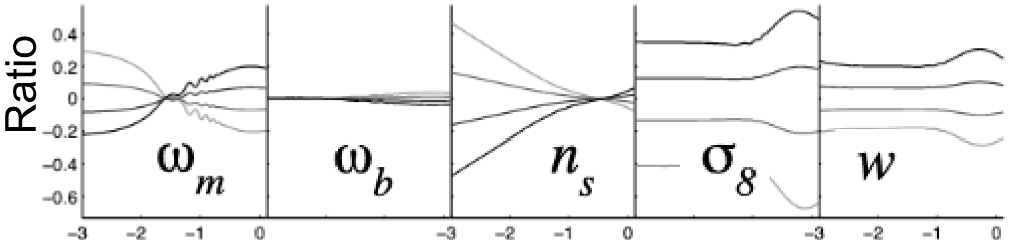
Transforming large simulation suites into precision predictions

- **Step 1:** Design simulation campaign, rule of thumb: O(10) models for each parameter
- **Step 2**: Carry out simulation campaign and extract quantity of interest, e.g. cluster mass function, power spectrum
- **Step 3:** Choose suitable interpolation scheme to interpolate between models, we use Gaussian Processes
- **Step 4:** Build emulators
- **Step 5:** Use emulator to analyze data, determine model inadequacy, refine modeling strategy …



***Example for Step 1**: Optimal sampling for 5 cosmological parameters and 32 models*

***Example for Step 5**: Power spectrum variation for 5 parameters*

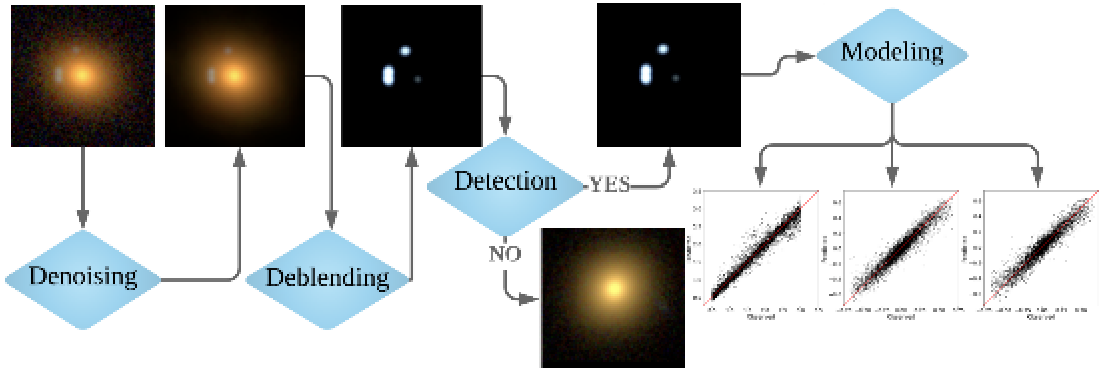


Introductory Paper: Heitmann et al., ApJ, 2009

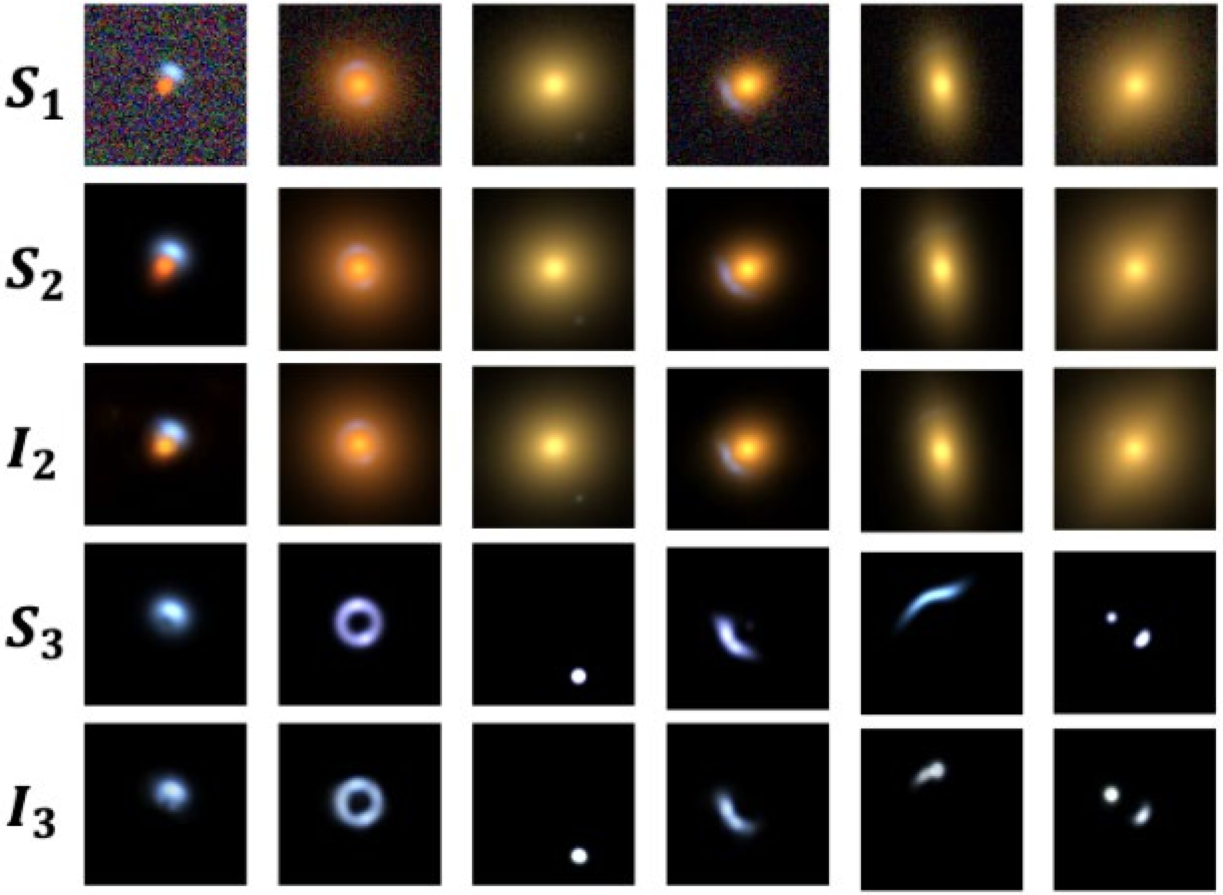
Galaxy-scale Strong Lensing

Finding rare targets in very large data sets

- Deep learning-based modular pipeline for image cleaning/de-noising, lens identification, and lens characterization
- Trained on very large simulated data set
- Tested on HSC strong lens data with good results, better than 90%
- Key next issue: reducing false-positives



Strong lensing pipeline structure (Madireddy et al., 2019)

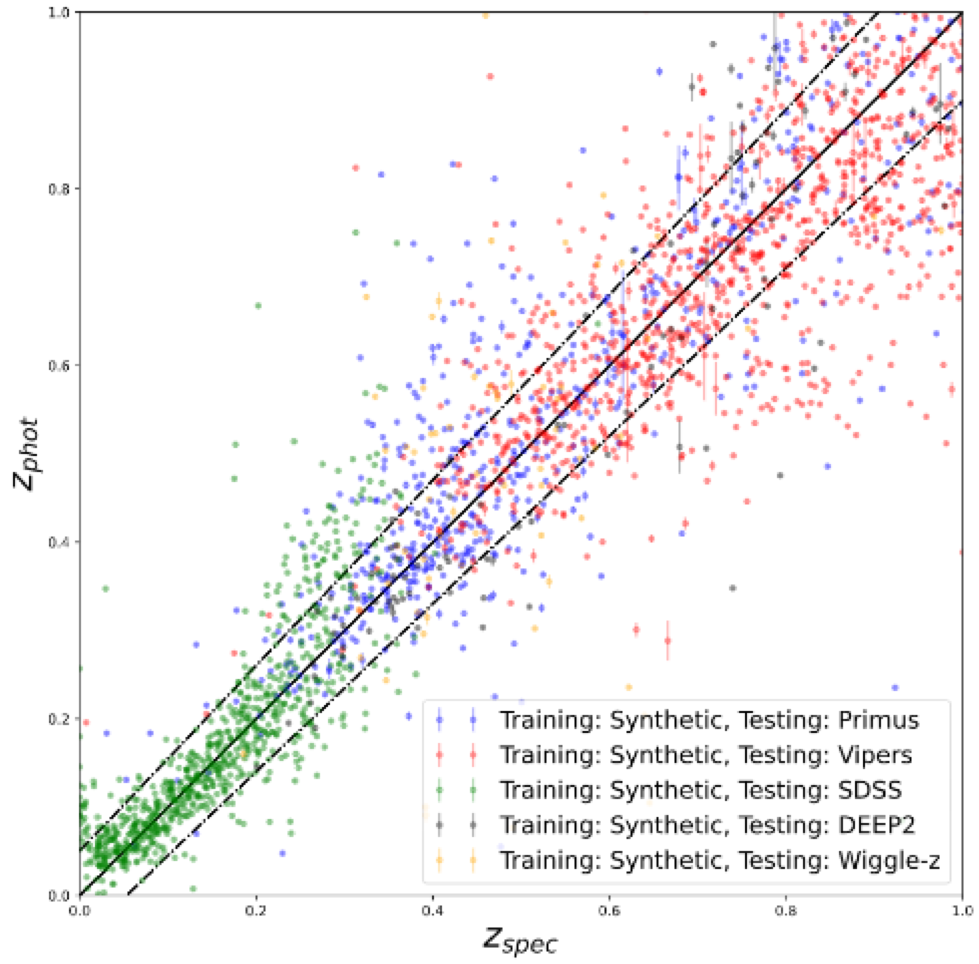


S_1: noisy blended simulation, S_2: noiseless blended simulation, I_2: output from denoising module, S_3: noiseless deblended simulation, I_3: output from deblending model

Photometric Redshift Estimation

Creating realistic training data

- Training-based photometric redshift estimation requires large numbers of SED templates for galaxy colors
- Number of observational templates is limited to bright sources
- Combined training sets based on observations and a robust generative model for emulating galaxy colors to fill data space not covered by observations
- Method outperforms techniques based only on observational data



Photometric redshift estimation pipeline validation, synthetic data generation only (Ramachandra et al., 2021)

Summary

Exciting times ahead!

- Upcoming surveys will generate complex data sets that will pose major new analysis challenges
- Exascale supercomputers will allow us to create the most detailed simulations so far, however, more is needed
- Innovative, carefully applied AI/ML methods will be invaluable to
  - Provide precision predictions
  - Enable us to find rare objects
  - Generate realistic synthetic data

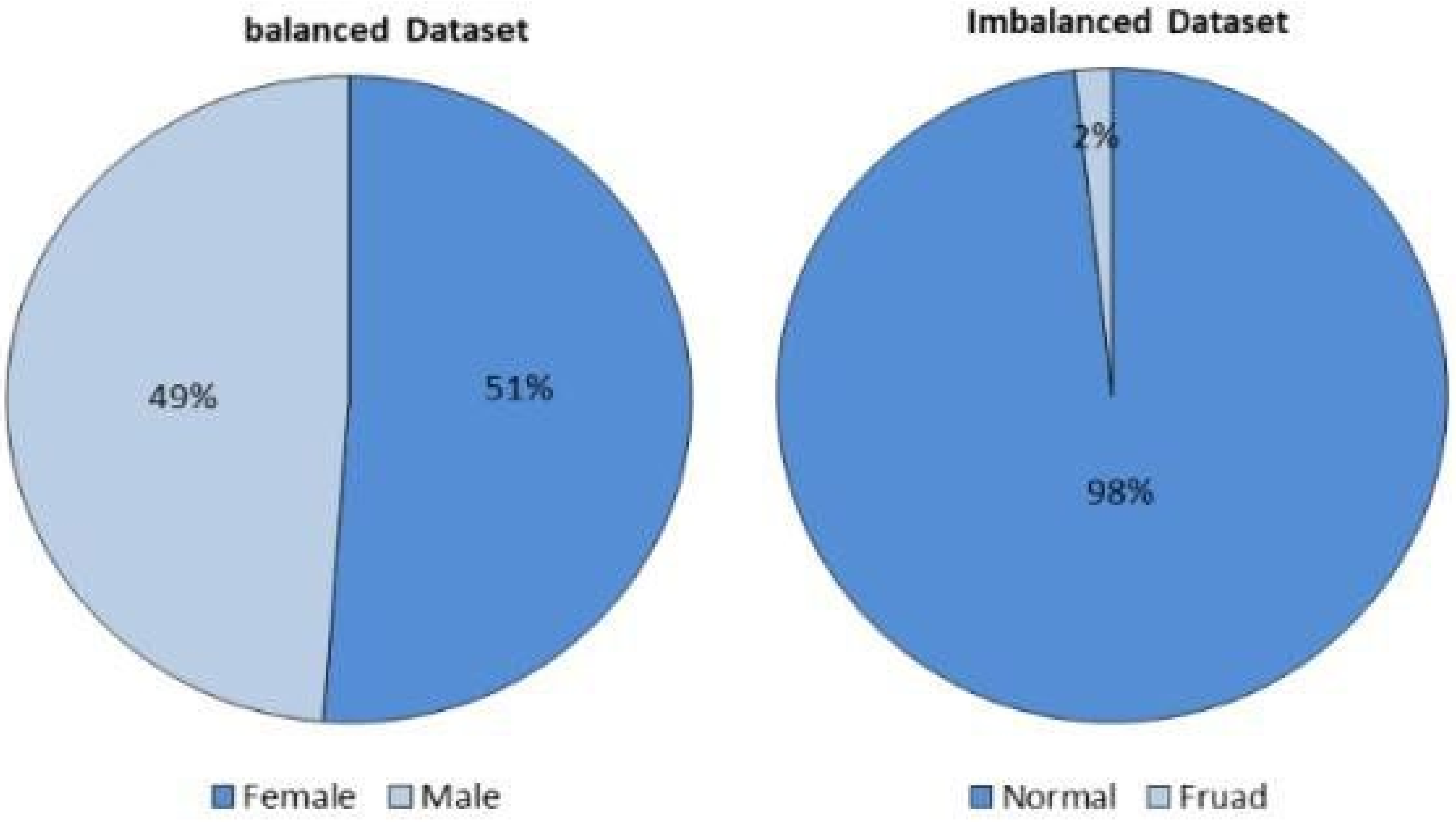

Image: Rubin Observatory/NSF/AURA

Jared Wadsworth, INL

Improving the Quality of Imbalanced Datasets using Generative Machine Learning Models

Idaho National Laboratory

Real world data is almost never balanced



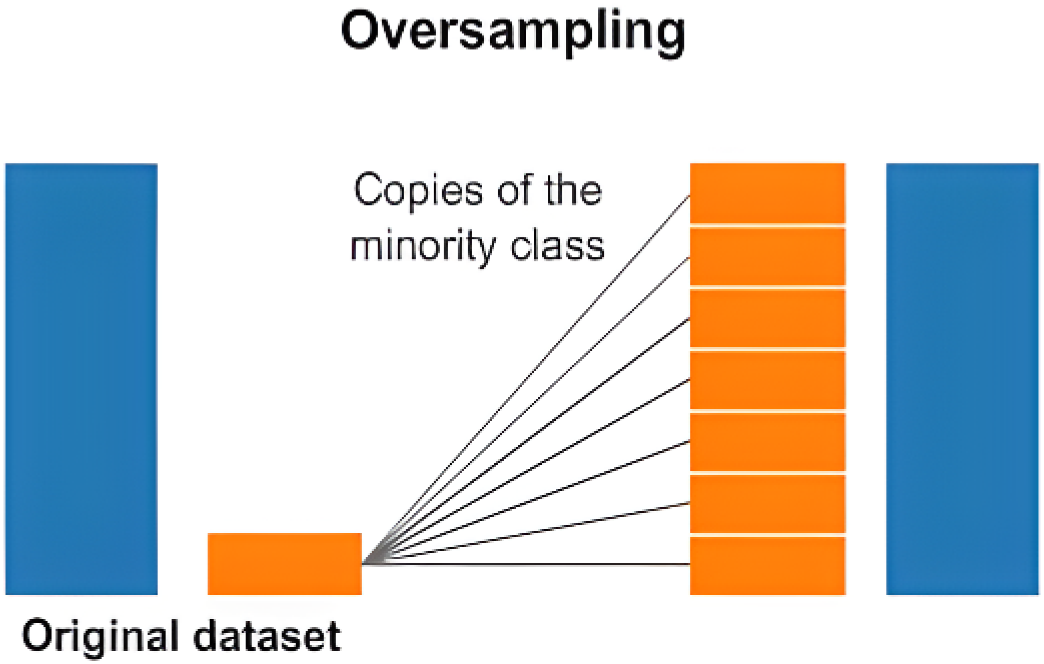
Problems of Imbalanced data

- Poor accuracy on smaller class
  - 95% real 5% fraud
  - Model predicts 100% fraud
  - (0+95)/(0+95+0+5)=0.95 or 95%

Typical Solutions

- Under / Over-Sampling
- Boosting
- Generative Models

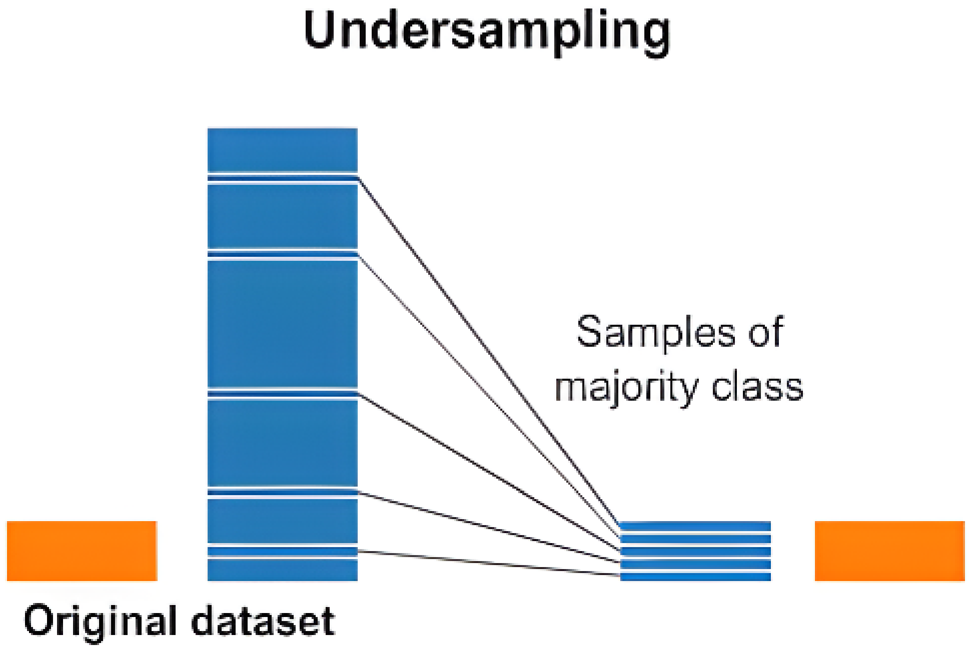
Over-sampling



Pros:
- Equal weighted classes
- Uses real data
- Easy

Cons:
- Overfitting on smaller class
- Increased importance of smaller class decision boundary

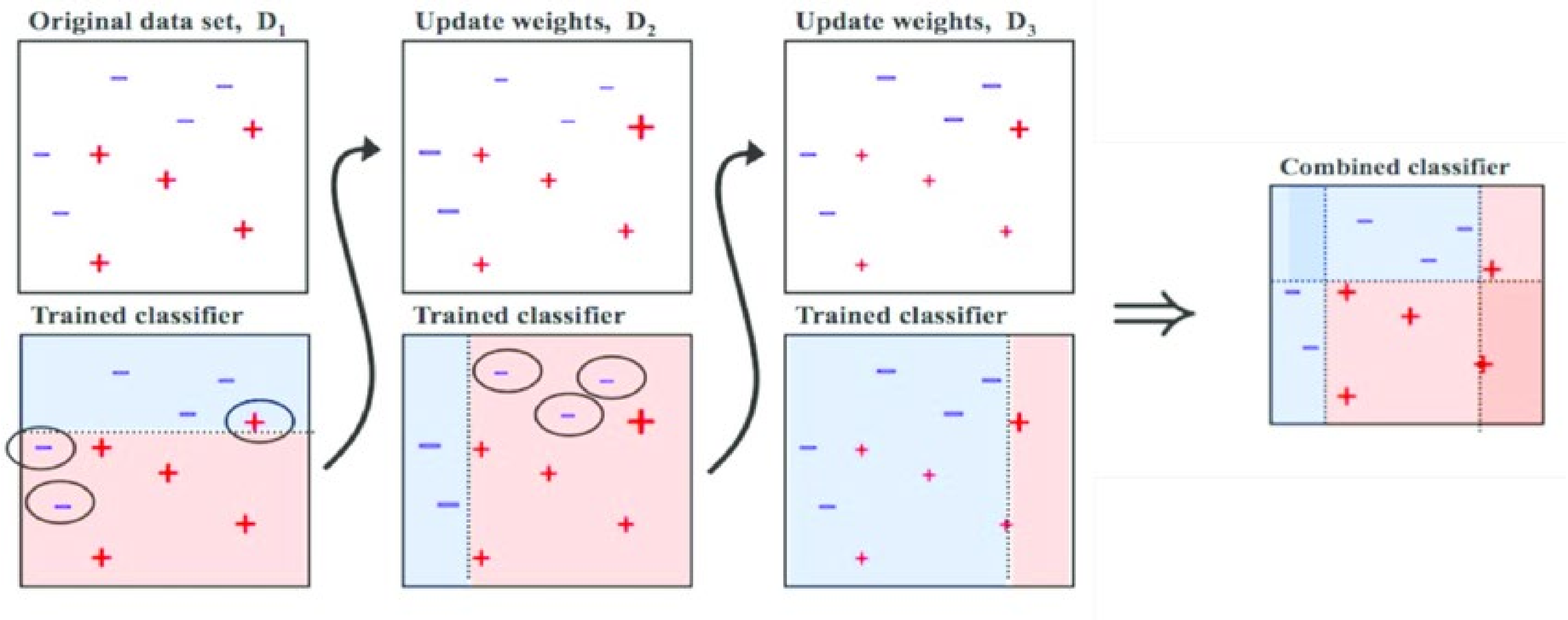
Under-sampling



Pros:
- Equal weighted classes
- Uses real data
- Easy

Cons:
- Loss of data
- Loss of diversity

Weighted loss / Boosting



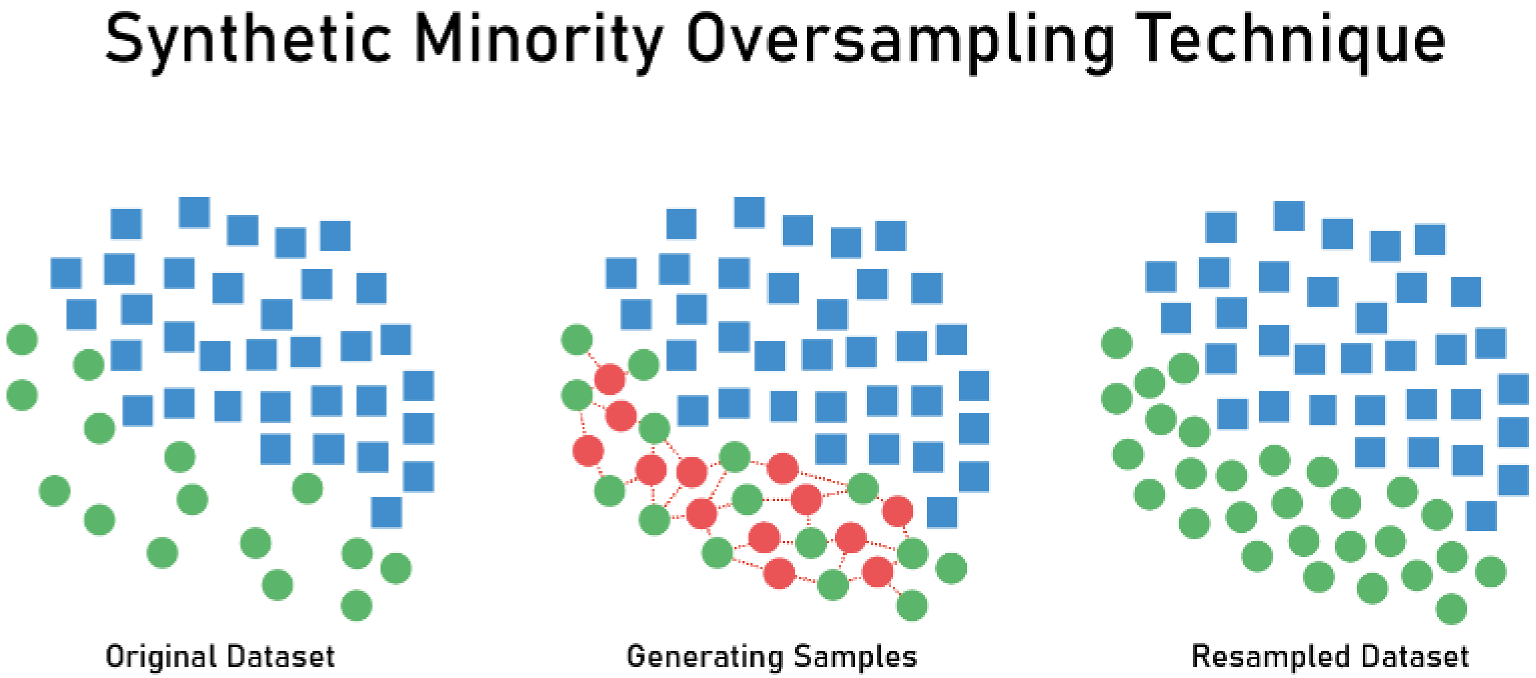
Pros:
- No duplication or loss of data
- Uses real data
- Built-in balancing of classes

Cons:
- Increased risk of overfitting
- Increased training time resources

Generative Models

- SMOTe
- Autoencoders
- Variational Autoencoders
- Generative Adversarial Networks

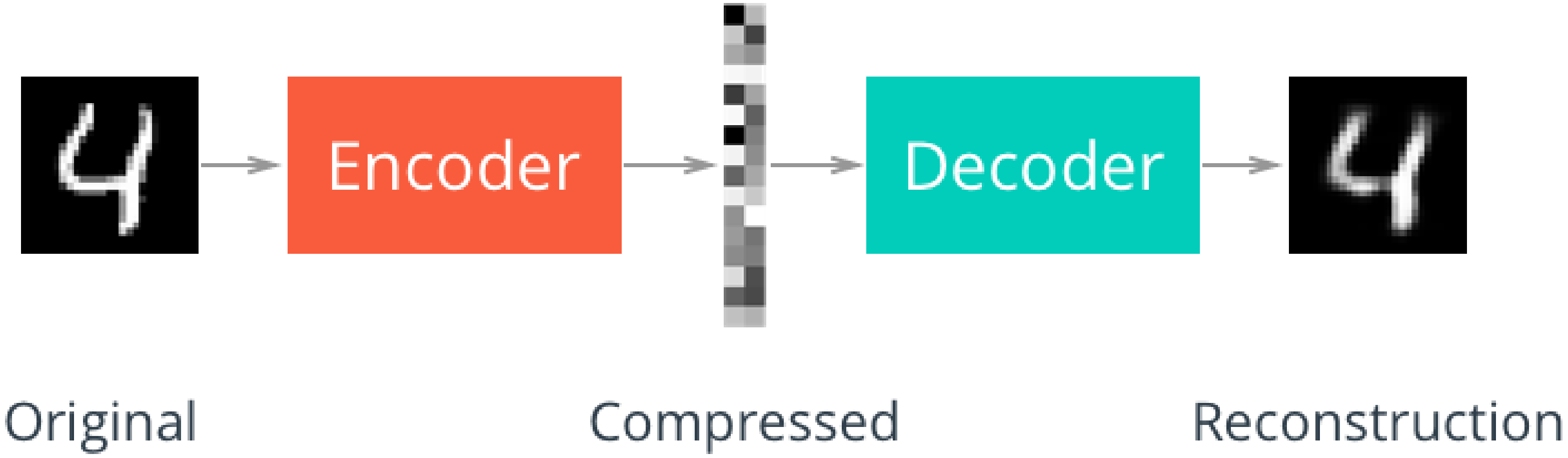
SMOTe



Pros:
- No duplication or loss of data
- Prevents overfitting

Cons:
- Uses synthesized data
- New data not guaranteed to be in same distribution
- Can be computationally expensive

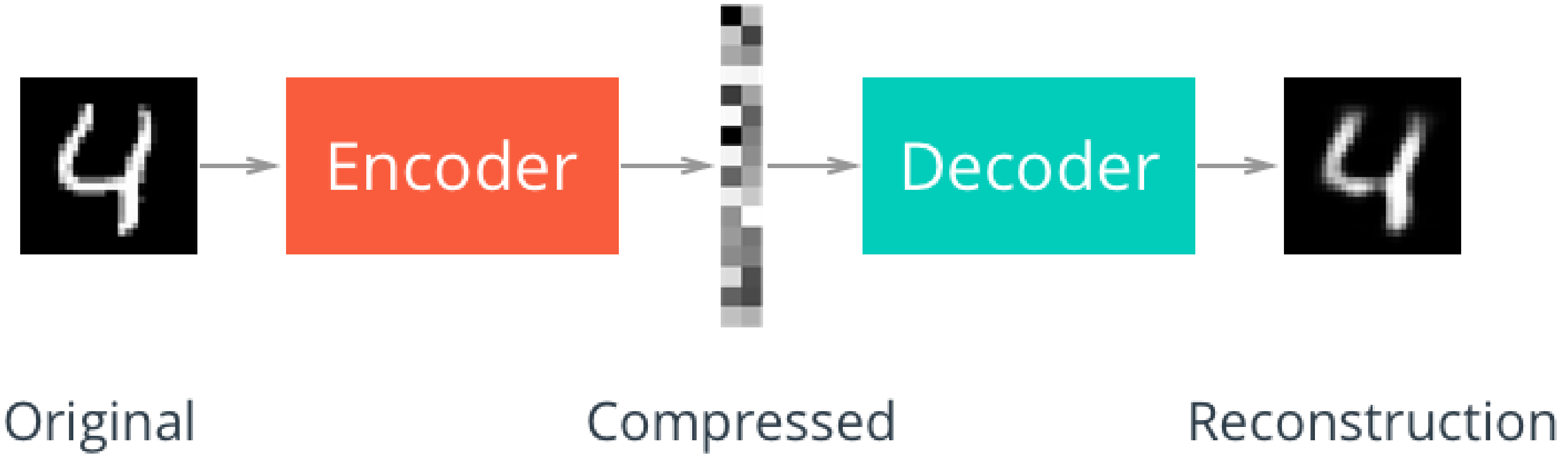
Autoencoders



Pros:
- No duplication or loss of data
- Prevents overfitting
- Data drawn from same distribution

Cons:
- Uses synthesized data
- Can be computationally expensive
- One-to-one mapping of data only allows for doubling data

Variational Autoencoders



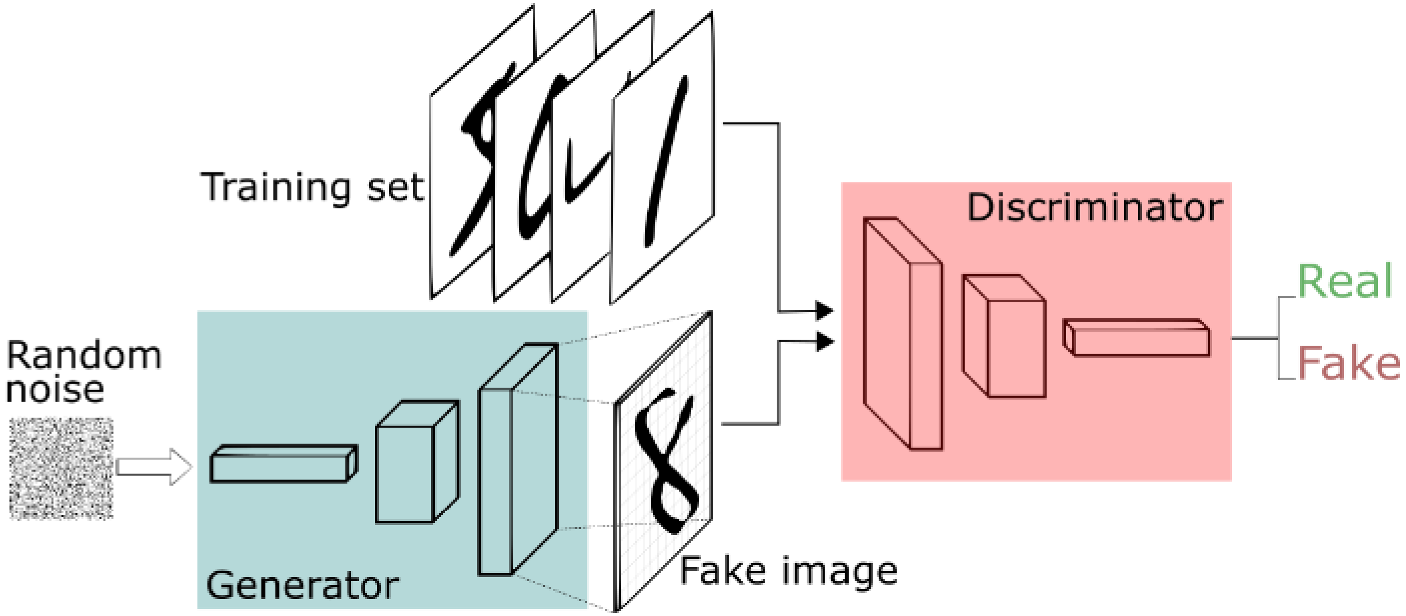
Pros:

- Any amount of data
- Better utilizes latent space
- Small changes in latent space result in small changed in synthetic data

Cons:

- Increased Complexity and Training time
- Requires a minimum size of smaller dataset

Generative Adversarial Networks (GANs)



Pros:
- Any amount of data
- No duplication or loss of data
- Prevents overfitting
- Data drawn from same distribution

Cons:
- Increased Complexity and Training time
- Mode collapse
- Failure to converge
- Vanishing Gradients

Conclusion

- Each method is applicable in different circumstances
- No Free Lunch
- If you can’t gather more data, Generative Methods may be a good way to do so with reasonable potential to increase your metric

NS&T ML-AI

Big Data   Machine Learning   Artificial Intelligence

Thank you for joining us!