NS&T ML-AI

Big Data *Machine Learning* *Artificial Intelligence*

Artificial Intelligence and Machine Learning Symposium 4.0

February 9, 2021

NS&T ML-AI

Big Data   Machine Learning   Artificial Intelligence

Welcome

Ronald Boring
Idaho National Laboratory

February 9, 2021

Ronald Laurids Boring, PhD, FHFES
Manager, Human Factors and Reliability Department

Machine Learning/AI Symposium 4

INL Nuclear and Science Directorate Webinar Series

Idaho National Laboratory

The First Three Symposia

- **April 2020: Artificial Intelligence (AI) and Machine Learning (ML) Symposium 1.0**
  - Focused on internal-to-INL activities and capabilities
  - Was such a success, that we extended symposium beyond INL
- **July 2020: AI/ML Symposium 2.0**
  - Engaged industry and universities
  - It was noted that AI/ML will be a key technology moving forward as we continue our R&D
- **October 2020: AI/ML Symposium 3.0**
  - Focusing on nuclear-related applications using AI/ML
  - Revealed a rich collection of AI applications already underway to help with tasks like monitoring, risk prediction, and maintenance

The First Three Symposia

- **April 2020: Artificial Intelligence (AI) and Machine Learning (ML) Symposium 1.0**
  - Focused on internal-to-INL activities and capabilities
  - Was such a success, that we extended symposium beyond INL
- **July 2020: AI/ML Symposium 2.0**
  - Engaged industry and universities
  - It was noted that AI/ML will be a key technology moving forward as we continue our R&D
- **October 2020: AI/ML Symposium 3.0**
  - Focusing on nuclear-related applications using AI/ML
  - Revealed a rich collection of AI applications already underway to help with tasks like monitoring, risk prediction, and maintenance

The First Three Symposia

- **April 2020: Artificial Intelligence (AI) and Machine Learning (ML) Symposium 1.0**
  - Focused on internal-to-INL activities and capabilities
  - Was such a success, that we extended symposium beyond INL
- **July 2020: AI/ML Symposium 2.0**
  - Engaged industry and universities
  - It was noted that AI/ML will be a key technology moving forward as we continue our R&D
- **October 2020: AI/ML Symposium 3.0**
  - Focusing on nuclear-related applications using AI/ML
  - Revealed a rich collection of AI applications already underway to help with tasks like monitoring, risk prediction, and maintenance

The Current Symposium

- **Symposium 4.0 narrows the topic a bit: “Trustworthy and Explainable AI”**
  - The success of AI/ML depends on:
    - AI doing what it’s supposed to do (*Reliable*)
      - A lot of the evolution and demonstrations of AI covered in earlier symposia
    - Us trusting that AI is doing what it’s supposed to do (*Trustworthy*)
      - A system that does something we don’t expect is not likely invited to do it a second time
      - Many of the applications of AI we are discussing are safety critical with no margin for AI surprises!
    - Us understanding what the AI is doing (*Explainable*)
      - Not completely independent of trustworthy AI
      - We need to understand what is going on before we trust it!

The Current Symposium

- **Symposium 4.0 narrows the topic a bit: "Trustworthy and Explainable AI"**
  - The success of AI/ML depends on:
    - AI doing what it's supposed to do (*Reliable*)
      - A lot of the evolution and demonstrations of AI covered in earlier symposia
    - Us trusting that AI is doing what it's supposed to do (*Trustworthy*)
      - A system that does something we don't expect is not likely invited to do it a second time
      - Many of the applications of AI we are discussing are safety critical with no margin for AI surprises!
    - Us understanding what the AI is doing (*Explainable*)
      - Not completely independent of trustworthy AI
      - We need to understand what is going on before we trust it!

The Current Symposium

- **Symposium 4.0 narrows the topic a bit: “Trustworthy and Explainable AI”**
  - The success of AI/ML depends on:
    - AI doing what it’s supposed to do (*Reliable*)
      - A lot of the evolution and demonstrations of AI covered in earlier symposia
    - Us trusting that AI is doing what it’s supposed to do (*Trustworthy*)
      - A system that does something we don’t expect is not likely invited to do it a second time
      - Many of the applications of AI we are discussing are safety critical with no margin for AI surprises!
    - Us understanding what the AI is doing (*Explainable*)
      - Not completely independent of trustworthy AI
      - We need to understand what is going on before we trust it!

Common denominator:

AI/ML/Big Data are technologies ultimately used by humans

- AI does not supplant humans; it augments us
- We must be mindful of the end users of AI

Today’s Agenda

- **Short presentations from INL researchers and collaborators on trustworthy and explainable AI**
  - Explainable AI Overview (DARPA)
  - Explainable AI to Support Operations and Maintenance at Nuclear Power Plants (UTK)
  - Trustworthiness Assessment of Digital Twins (NCSU)
  - Trustworthy AI Guidelines for Human-System Interactions (VCU)
  - Improving Explainable AI Through Process Information and Automated Reasoning (ANL)
  - Exploring Reaction Mechanisms with Explainable AI (INL)
  - Neural Networks for Control of a Subcritical Facility (MIT)
- **Introductions and discussions facilitated by Dr. Nancy Lybeck, Manager for INL’s Instrumentation, Controls, and Data Science Department**

NS&T ML-AI

Big Data   Machine Learning   Artificial Intelligence

Matt Turek

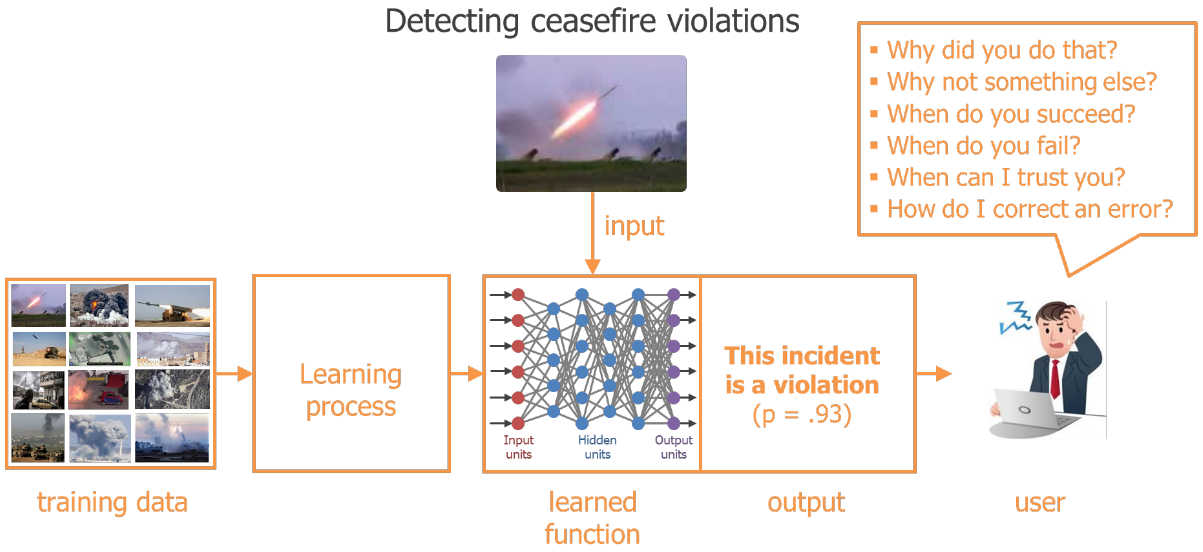
DARPA

Explainable AI (XAI)

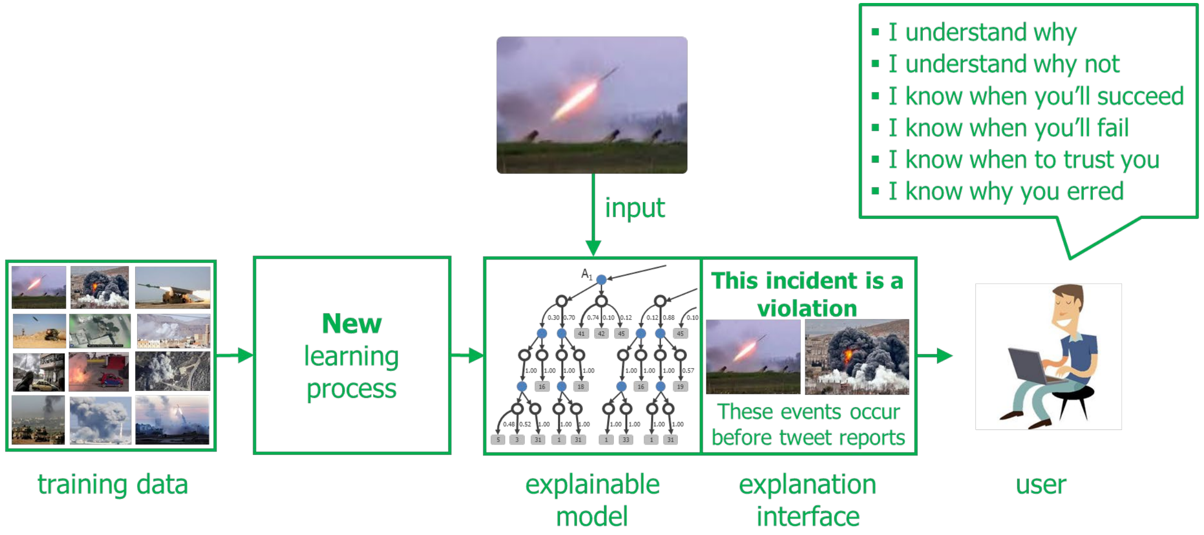
Matt Turek, PhD

Explainable AI

How is it done today?



What are we trying to do?

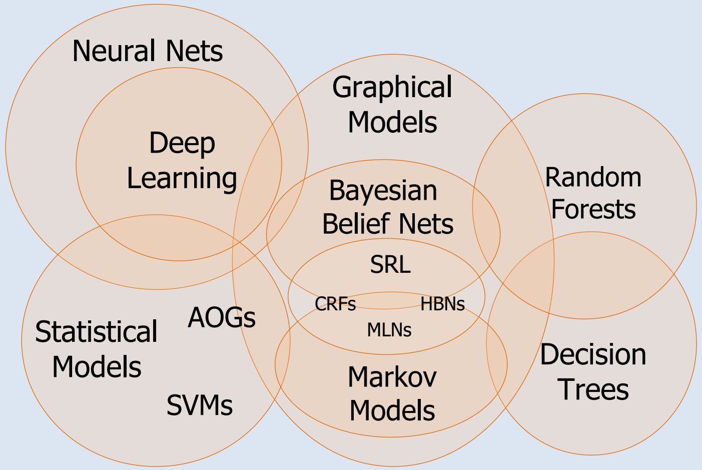


Approved for Public Release, Distribution Unlimited

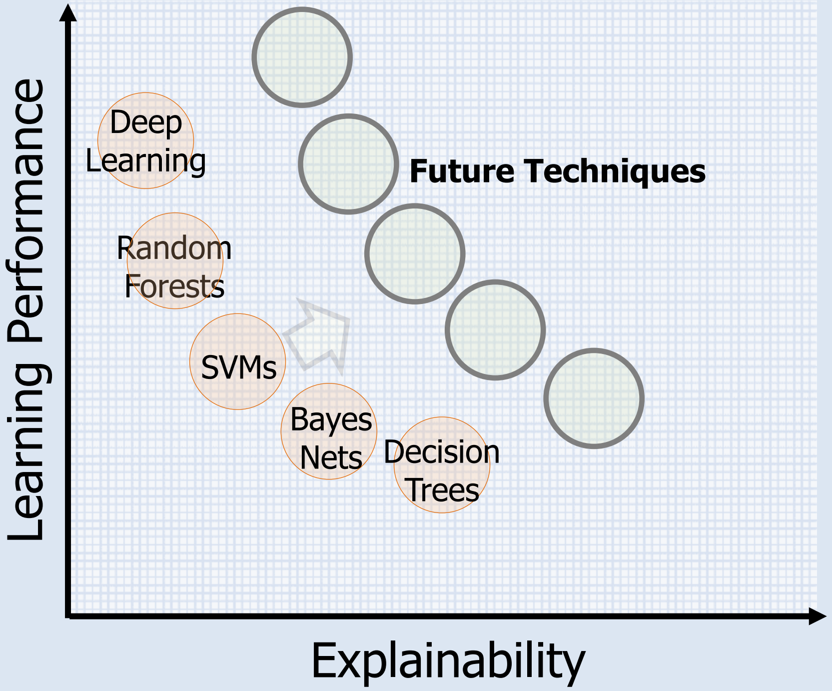
Explainable AI overview

Today

State-of-the-art ML Techniques (circa 2016)

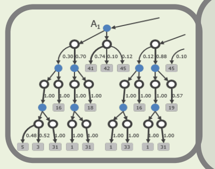


Performance - Explainability Tradeoff (notional)

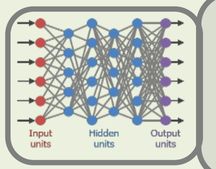


Tomorrow

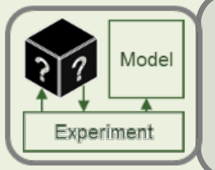
Explainable AI Strategies



Interpretable Models
Alternative machine learning techniques that learn more structured, interpretable, or causal models



Deep Explanation
Modified or hybrid deep learning techniques that learn more explainable features, explainable representations, or explanation generation facilities



Model Induction
Techniques that experiment with a machine learning model to infer an approximate explainable model

Deliver a library of toolkits

Challenge problem areas

Data analytics

Microsoft

Explains recommendations to an analyst

Autonomy

insideunmannedsystems

Explains actions to an operator

Technical strategy: address diverse DoD user types

Explainable AI system users

Developers

AI Expert
Design, develop, and debug

Task SME
Test and evaluate

- Explanations expose finer details of the system
- Explanations are used to modify/refine the system

Does the system work well?
If not, why do these errors occur?

End Users/Service members

- Military
- Legal
- Transportation
- Security
- Finance
- Medical

- Explanations aid decision making/recommendations
- Explanations justify actions taken and decisions

Policymakers/Regulators

Commander

Policymaker Regulator

- Decision patterns are defensible
- Decisions meet policy/regulatory requirements

Explainable AI system development-to-use timeline (notional)

Conceptual System → Prototype System → System V&V → System Acceptance Testing → Operational System → After Action Reviews → Regulatory Reviews

Gunning, D.; Stefik, M.; Choi, J.; Miller, T.; Stump, S.; Yang, G.-Z. 2019. XAI—Explainable artificial intelligence. *Science Robotics* 18 Dec 2019: Vol. 4, Issue 37, eaay7120, DOI: 10.1126/scirobotics.aay7120.

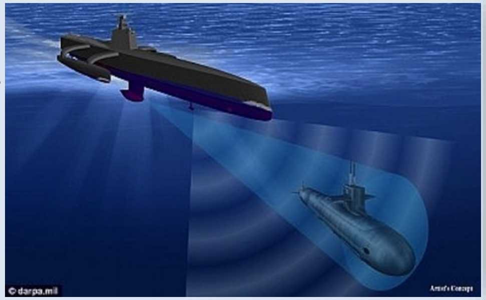
Technical strategy: address the need for different explanation types

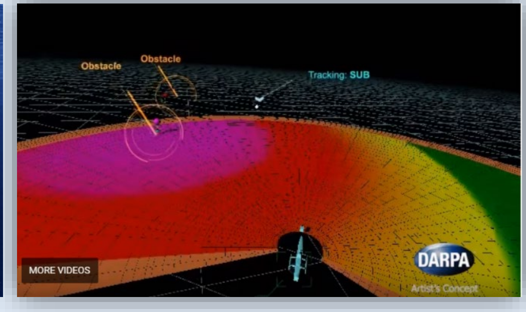
Global Explanations

"How does the AI work generally?"

AI task:
Automatically maneuver to a target location

Notional example





Potential explanation: Description of the AI's learned policy for routing around friendly vessels

Help determine if an AI system is fit for purpose

Local Explanations

"Why did the AI make a particular decision?"

AI task:
Automatically detect resupply activity at a military installation

Notional example

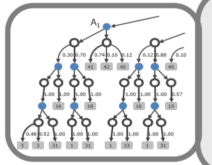
Potential explanation: Description of the specific evidence
"Trucks have appeared next to these bunkers in the last 24 hours"

Help an analyst make a correct decision

Both **global** and **local** explanations help build a robust mental model of the AI system

Klein, G.; Hoffman, R.; Mueller, S. 2019. Naturalistic Psychological Model of Explanatory Reasoning: How People Explain Things to Others and to Themselves, *International Conference on Naturalistic Decision Making* 2019, San Francisco, CA.
Hoffman, R.R.; Mueller, T.; Mueller, S.T.; Klein, G.; Clancey, W.J. 2018. Explaining Explanation Part 4: A Deep Dive on Deep Nets. *IEEE: Intelligent Systems*, pp. 87-95.

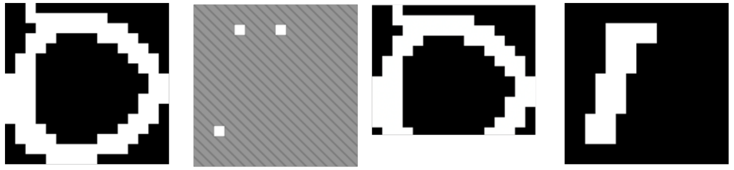
Example XAI algorithm approaches



Interpretable Models

Alternative machine learning techniques that learn more structured, interpretable, or causal models

New tractable probabilistic modeling (TPM) approach facilitates developer verification and validation of a model via sufficient and necessary explanations



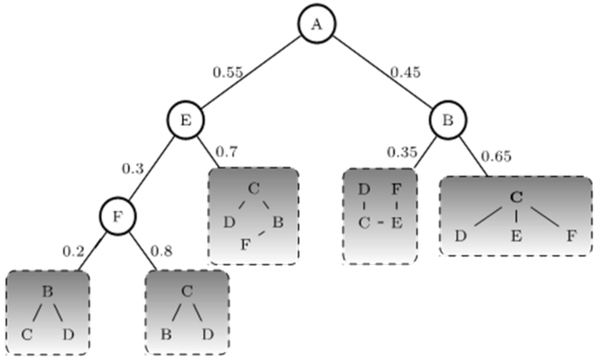
Why is this a zero?

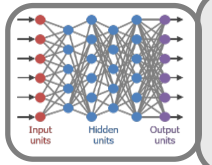
Sufficient: Because of these 3 white pixels.

Necessary: Because the top half of the image is as shown.

Classifier is not robust: Even this will be classified as 0

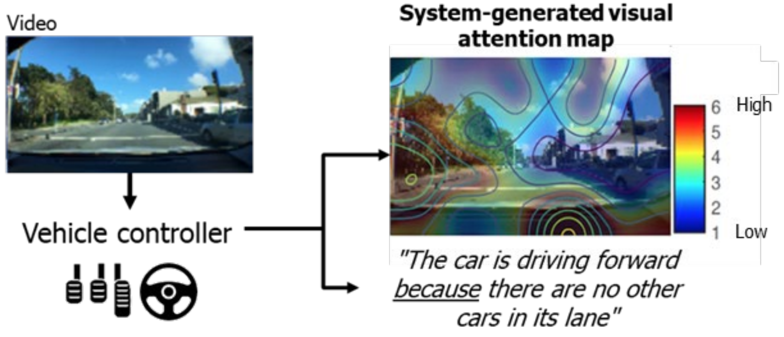
Graphical representation of a TPM



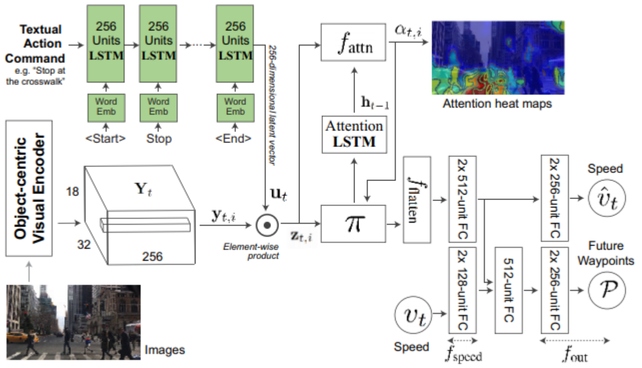


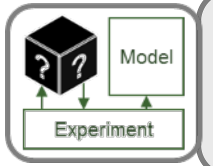
Deep Explanation

Modified or hybrid deep learning techniques that learn more explainable features, explainable representations, or explanation generation facilities



System-generated textual description & attention map

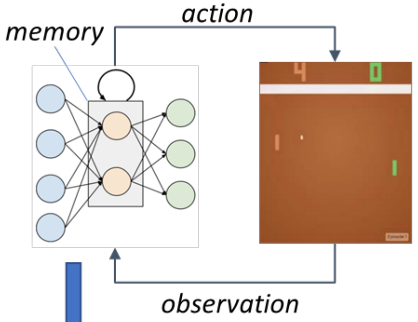




Model Induction

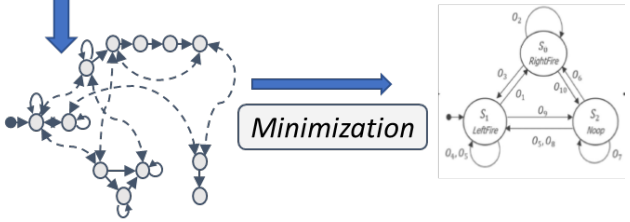
Techniques that experiment with a machine learning model to infer an approximate explainable model

How does a recurrent network use its high-dimensional, continuous memory?



Model did not use ball movement to guide decision, instead, it keyed in on pixels indicating whether it was a odd/even point round to determine the course of action.

Discretize Memory via ***Quantized Bottleneck Insertion*** *[ICLR'19]*



Finite state machine

Randomized Input Sampling for Explanation (RISE)

UC Berkeley

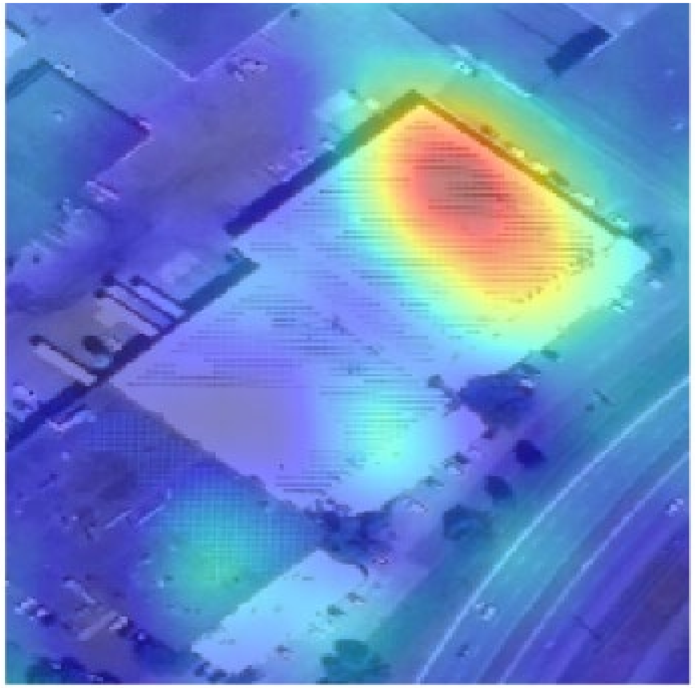
Neural Network Prediction

solar farm: 63%, **shopping mall**: 23%

Image from the FMoW dataset

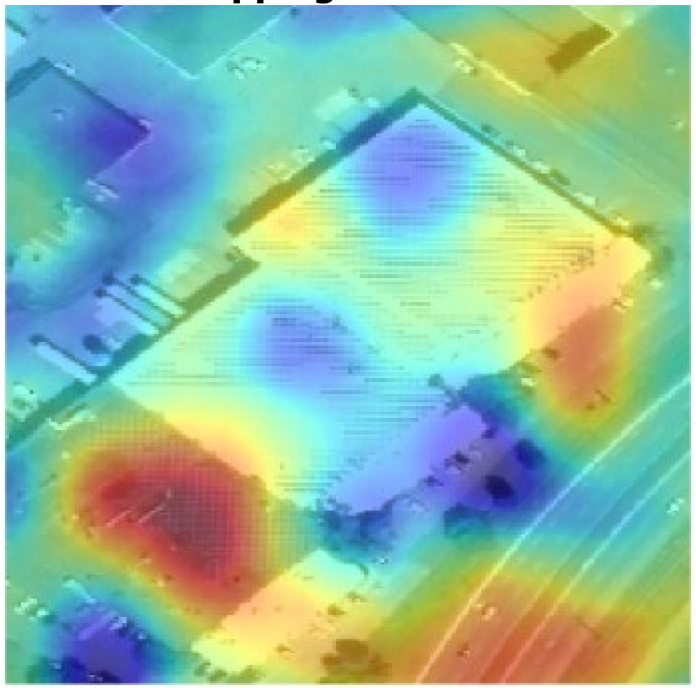
RISE Explanation for **solar farm**

solar farm: 63%



RISE Explanation for **shopping mall**

shopping mall: 23%



Increasing importance

Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized Input Sampling for Explanation of Black-box Models. *Proceedings of the British Machine Vision Conference (BMVC), 2018.*

Explanation by example

Rutgers University

Target Image Trial 1

A or B?

Category A Examples

Category B Examples

Target Image Trial 2

C or D?

Category C Examples

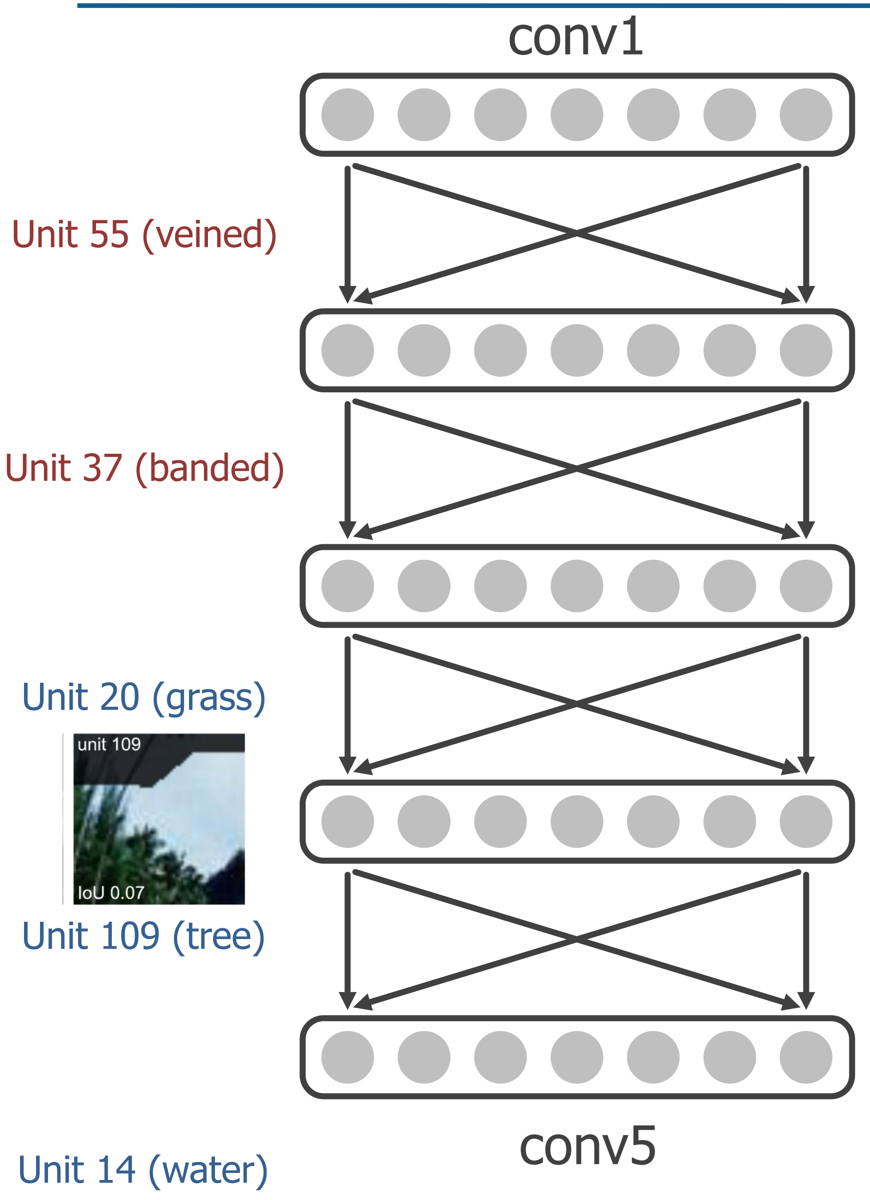
Category D Examples

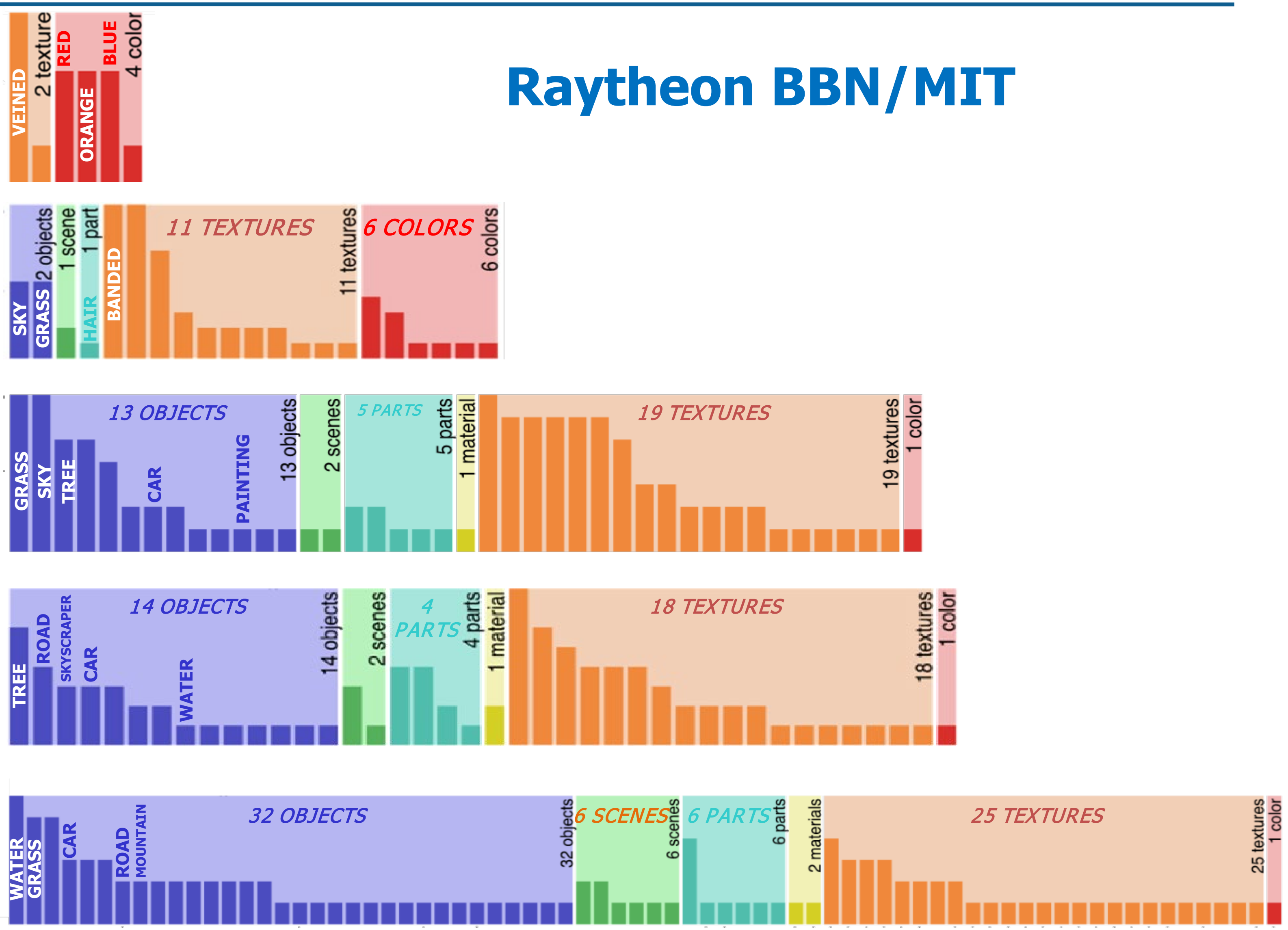
Explanation by selecting the subset of training data examples that are most representative of the model's classifications

Patrick Shafto et al., "Model Explanation by Optimal Selection of Teaching Examples," presented at the DARPA Explainable AI Meeting, Berkeley, CA, February 2019.

Network Dissection - AlexNet layers for recognizing places

Raytheon BBN/MIT

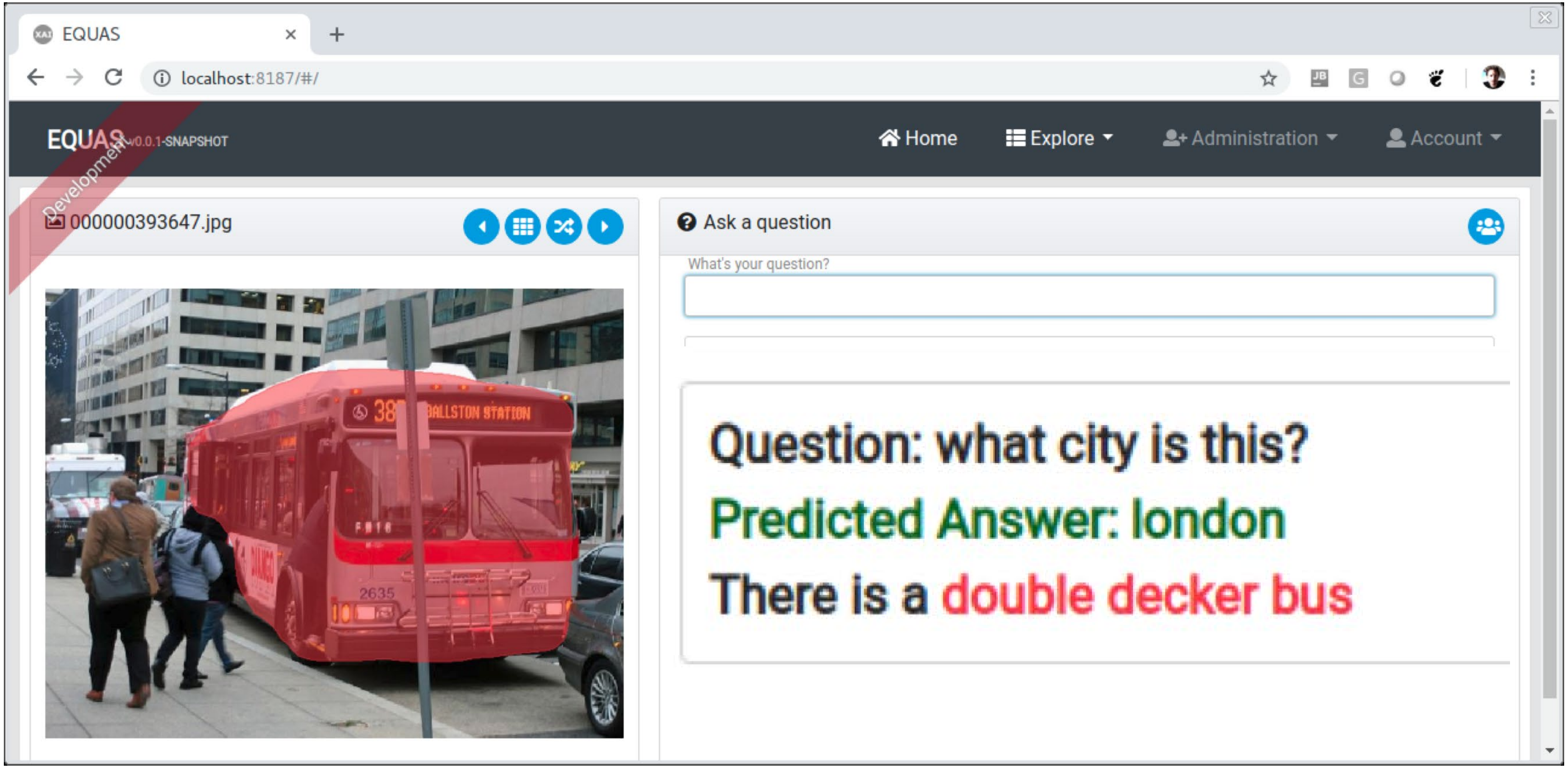




David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, and Antonio Torralba. *GAN Dissection: Visualizing and Understanding Generative Adversarial Networks*. arXiv preprint arxiv 1811.10597, 2018.

Explanations can help when the answer is wrong

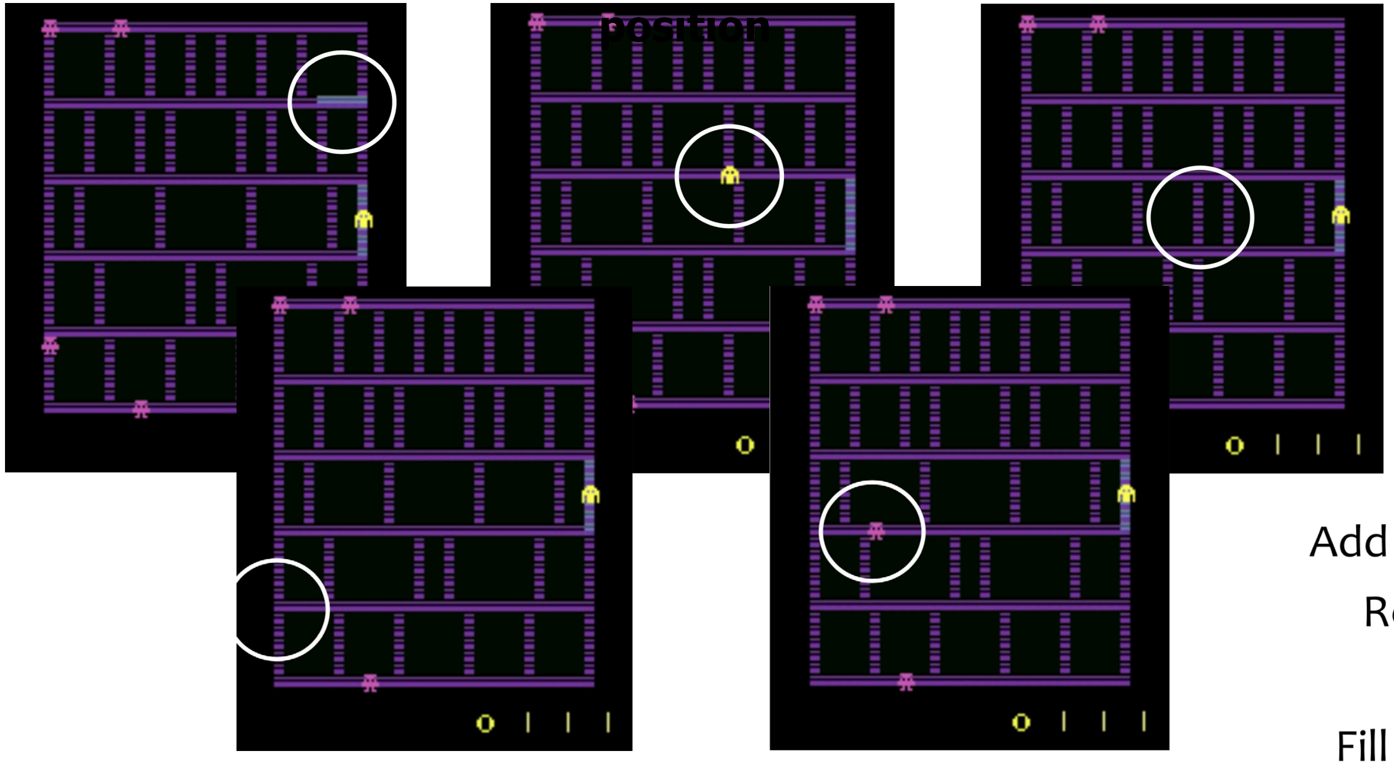
Raytheon BBN



William Ferguson et al., "EQUAS, Explainable QUestion Answering System," presented at the DARPA Explainable AI Meeting, Berkeley, CA, February 2019.

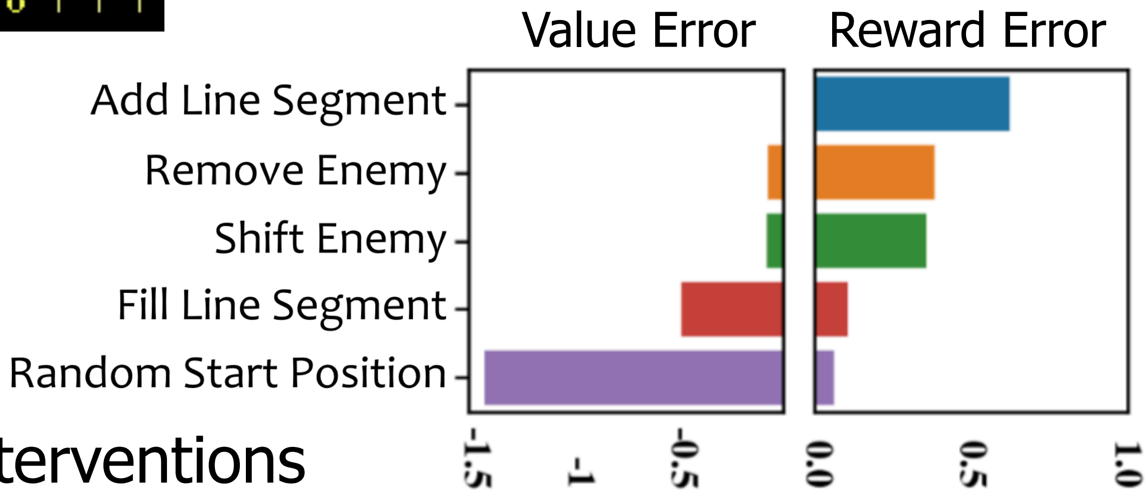
Unexpected brittleness of deep RL decisions

Fill line segment
Random start position
Add line segment



Remove Enemy
Shift Enemy

Charles River Analytics (CRA)
University of Massachusetts
Brown University



Learned policies are not robust to weak interventions

Sam Witty, Jun Ki Lee, Emma Tosch, Akanksha Atrey, Michael Littman, and David Jensen (2018). *Measuring and Characterizing Generalization in Deep Reinforcement Learning.*

DARPA

www.darpa.mil

NS&T ML-AI

Big Data Machine Learning Artificial Intelligence

Jamie Coble
UTK

Explainable AI to Support Operations and Maintenance at Nuclear Power Plants

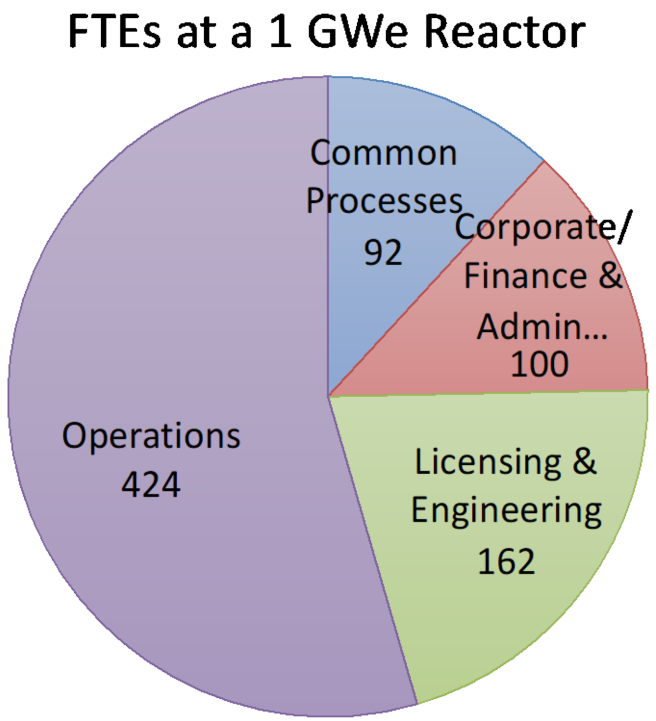
Jamie Coble

University of Tennessee-Knoxville

BIG ORANGE BIG IDEAS

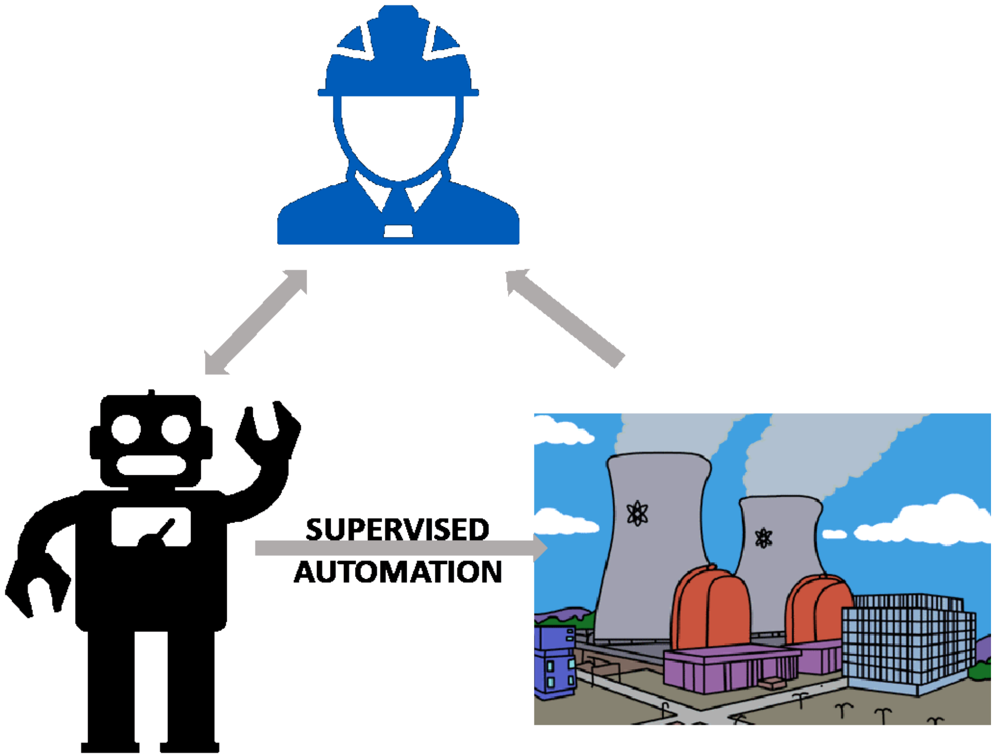
O&M remains the largest addressable cost in nuclear energy production

- Periodic inspection and maintenance activities contribute to unnecessary and costly O&M
- Advanced reactors operate in different regimes than our current LWRs
- Automation of operations and maintenance planning can manage O&M costs in current and future fleets



Automation used as decision support for O&M decision makers

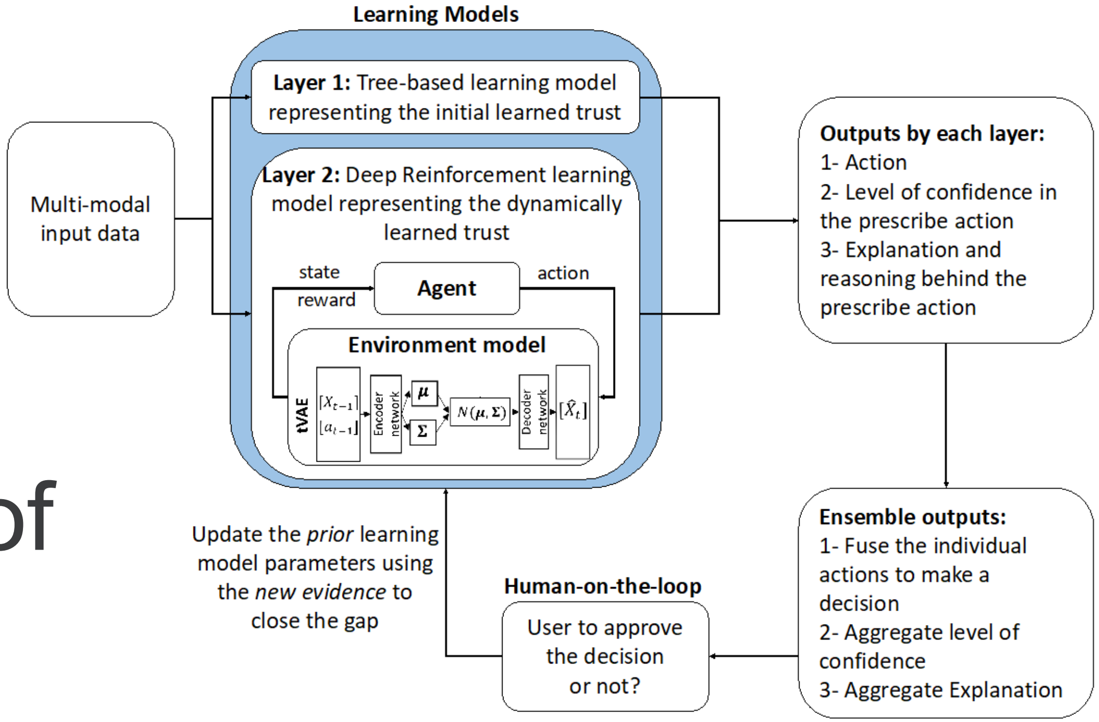
- Automation moves from human-in-the-loop to human-on-the-loop
- Questions remain about the trustworthiness of AI/ML automation



$$Explainable + Transparent = Trustworthy$$

Explainable decisions increase situational awareness while reducing workload

- ML decisions presented alongside evidence supporting the decision
- Trust can be modeled and adapted based on quality of decision, evidence, and communication



Opportunities to continue development

- NPP-specific AI/ML R&D needs
  - Algorithms to mine information from large data and big data
  - Integration with faster-than-real-time O&M digital twins
- For operator acceptance
  - Real-time decision reliability assessment
  - HMI to display AI/ML decisions and evidence to operators and engineers
- For regulatory acceptance
  - Uncertainty quantification and confidence assessment
  - V&V methodologies

Questions?

jamie@utk.edu

THE UNIVERSITY OF TENNESSEE KNOXVILLE

BIG ORANGE. BIG IDEAS.®

NS&T ML-AI

Big Data   Machine Learning   Artificial Intelligence

Linyu Lin
NC State University

Department of NUCLEAR ENGINEERING

Trustworthiness Assessment of Digital Twins in NAMAC

Linyu Lin, Nam Dinh

Department of Nuclear Engineering

North Carolina State University

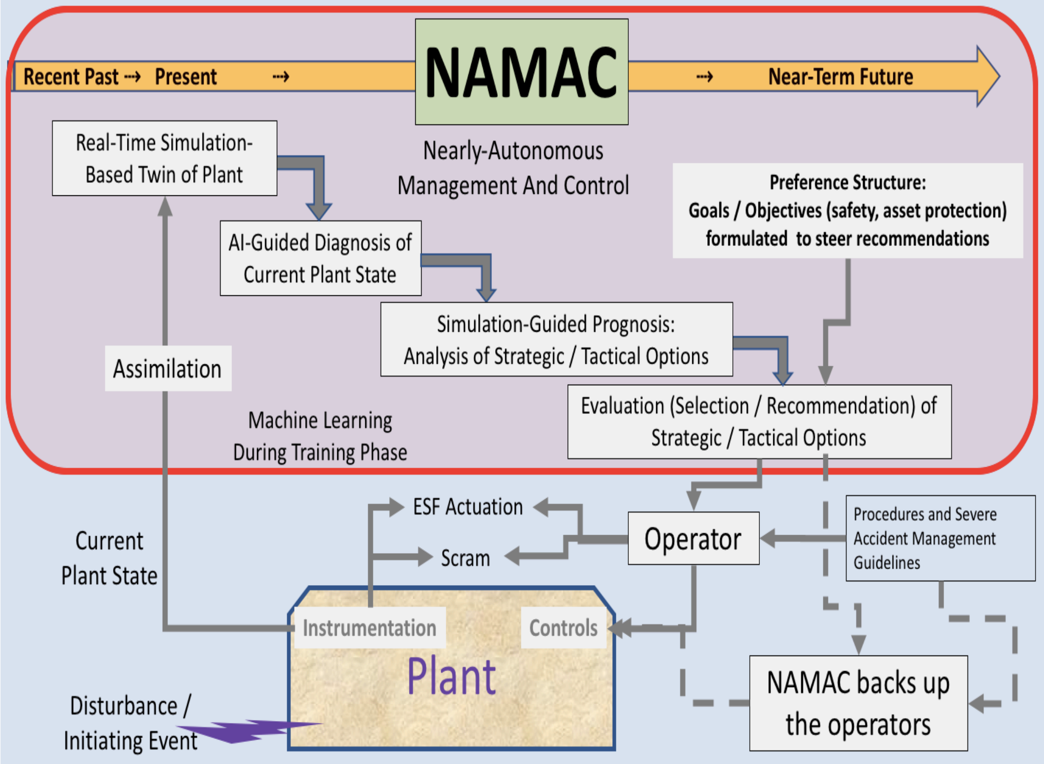
NC STATE UNIVERSITY

COLLEGE OF ENGINEERING

Nearly Autonomous Management and Control (NAMAC)

- A comprehensive control system to assist plant operations
  - Knowledge integration
    - Scenario-based model of plant (systems, success paths)
    - plant operating procedures, tech. specs., etc.
    - Real-time measurements
  - Not to replace human operator
  - Digital twin technology
    - Expressive Power of AI/ML

- NAMAC recommendations are derived from:
  - Diagnosing the plant state
  - Searching for all available mitigation strategies
  - Projecting the effects of actions and uncertainties into the future behavior
  - Determining the best strategy considering plant safety, performance, and cost.



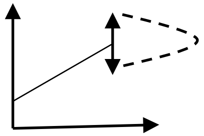
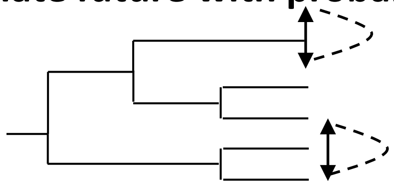
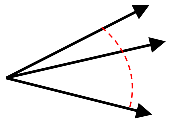
Digital Twin in NAMAC

- A hub of digital twins implemented by various machine learning algorithms to support the designated functions

	Function	Modeling
Diagnosis	Recover full reactor states by assimilating plant sensor data with the knowledge base	Neural nets (feedforward & recurrent); Logic programming (Answer Set Programming)
Strategy Inventory	Find all available control/mitigation strategies	Linear models
Prognosis	Predict the transients of state variables over a time range	Neural nets (feedforward & recurrent)
Strategy Assessment	Rank possible mitigations strategies and make recommendations considering preference structure	Safety margin/limiting surface; Expected utility;
Discrepancy Checker	Detect unexpected transient during operations considering DT trustworthiness for current conditions	Distance metrics; Logic programming (Answer Set Programming)
Integrated NAMAC	To furnish recommendations to operator by assimilating plant sensor data with the trained policy	Reinforcement Learning

Importance of Digital Twin Uncertainty

- The digital twin uncertainty affects scenarios' future states, the modeling of digital twins, and the target applications

Complete Certainty		**Level 1**	**Level 2**	**Level 3**	Total ignorance
	Scenarios' Future States	**A clear future with sensitivity** 	**Alternate future with probabilities** 	**A multiplicity of plausible futures** 	
	Digital Twins	A single set of digital twins with fixed form and parameter	Alternative digital twins with alternative forms and parameters where weights and uncertainties can be sufficiently characterized by probability distributions	Alternative digital twins with alternative forms and parameters where weights and uncertainties are known imprecisely	
	Appropriate target	High-consequence systems where decision making is fundamentally based on DTs, e.g., quantification or final O&M support	Moderate consequence systems with some reliance on DTs, e.g., preliminary O&M support	Low-consequence systems with little reliance on DTs, e.g., scoping studies or conceptual O&M support	

Digital Twin Development and Assessment Process (DT-DAP)

- DT-DAP to identify major sources of uncertainty and to avoid biases due to implicitness

- The DAP is conducted iteratively, and the corresponding elements are refined until an acceptable set of DTs are delivered

  <u>*Element 1*</u>: Refined requirements

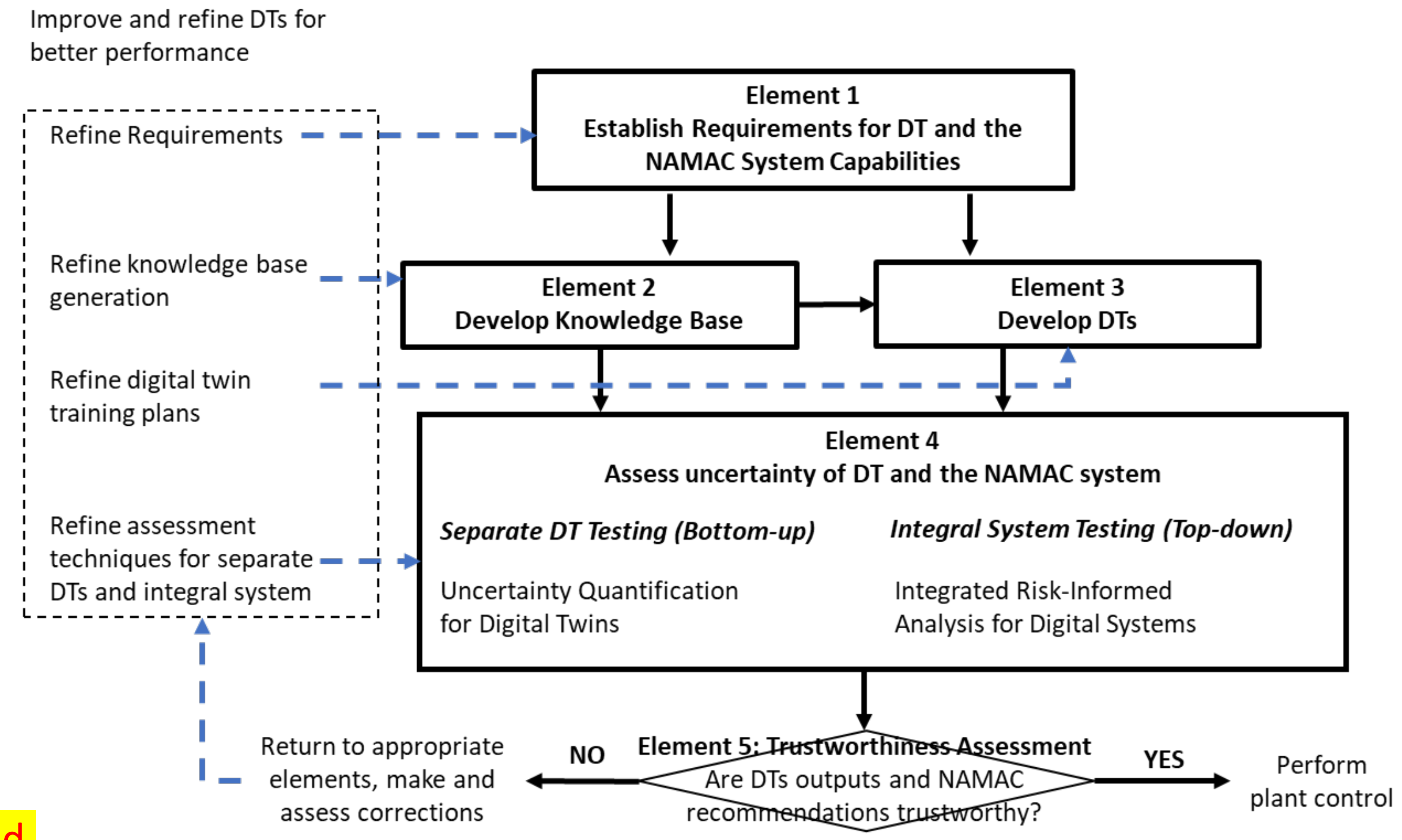
  <u>*Element 2*</u>: More complex and more realistic knowledge base

  <u>*Element 3*</u>: Different machine-learning algorithms, hyperparameter tunning

  <u>*Element 4*</u>: ML uncertainty quantification, software reliability analysis

Challenge in DT-DAP

Digital Twin Trustworthiness needs to be defined and evaluated in a transparent, consistent, and improvable manner

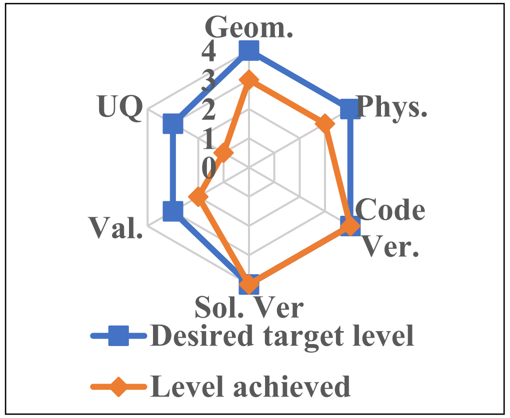


Adopted from U.S. NRC RG 1.203 “Transient and Accident Analysis Methods”

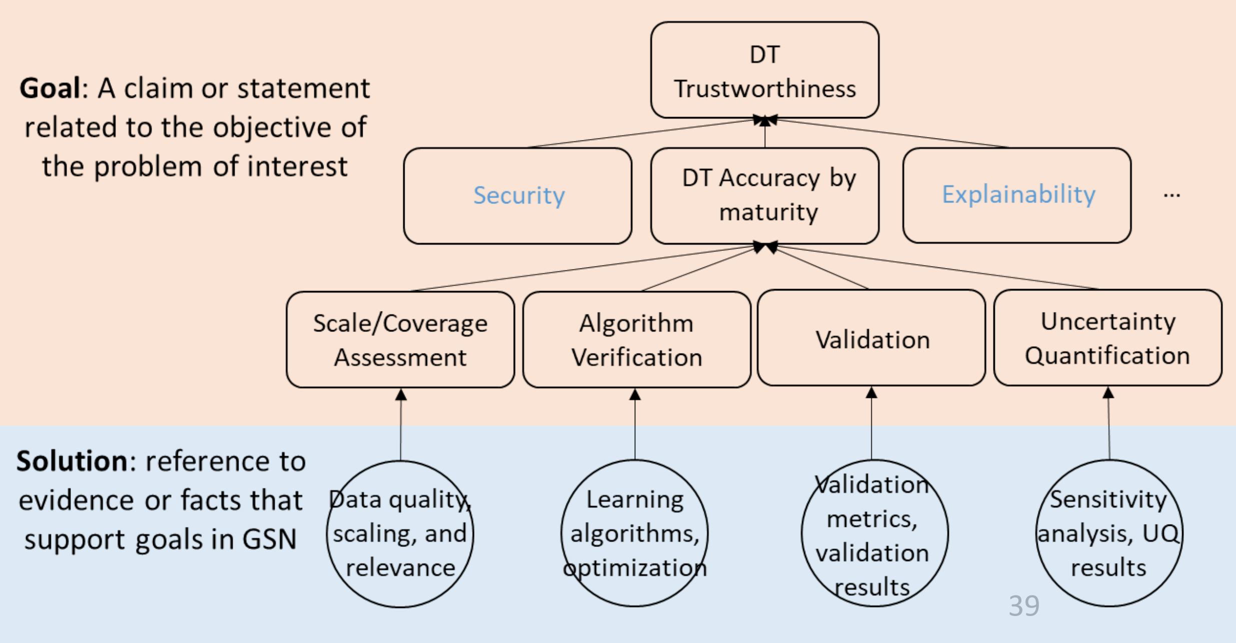
Looking Ahead

- There needs to be a definition for machine-learning-based digital twin trustworthiness and major attributes
  - Accuracy, Security, Robustness, Explainability, Reliability [2]
  - and more…
- The trustworthiness needs to integrate information (evidence) from different sources and heterogeneous types of data
- The quality of evidence could have significant impacts and needs to be evaluated
- The evidence integration needs to consider complex relations, priority, and trade-off between different attributes of trustworthiness
- At last, the trustworthiness assessment should be quantified and conducted in real-time deviation detection

Spider plot for the credibility assessment of mechanistic-based models based on multi-attributes evidence



An example of argumentation framework towards the DT trustworthiness goal based on evidence



Questions?

NS&T ML-AI

Big Data   Machine Learning   Artificial Intelligence

Chathurika S. Wickramasinghe
and
Daniel L. Marino

Virginia Commonwealth University

Trustworthy AI Development Guidelines for Human System Interactions

Presenters: Chathurika S. Wickramasinghe and Daniel L. Marino
Mentor: Prof. Milos Manic, FIEEE

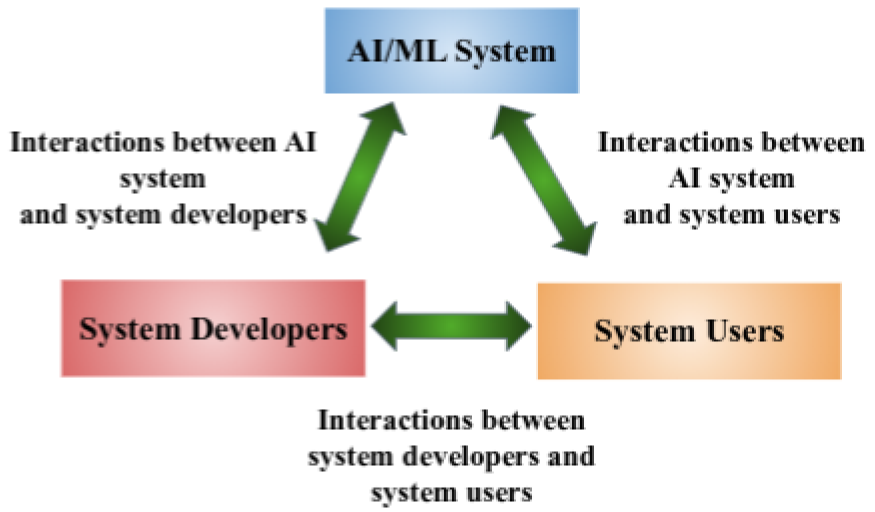
Virginia Commonwealth University, VA, USA

Abstract

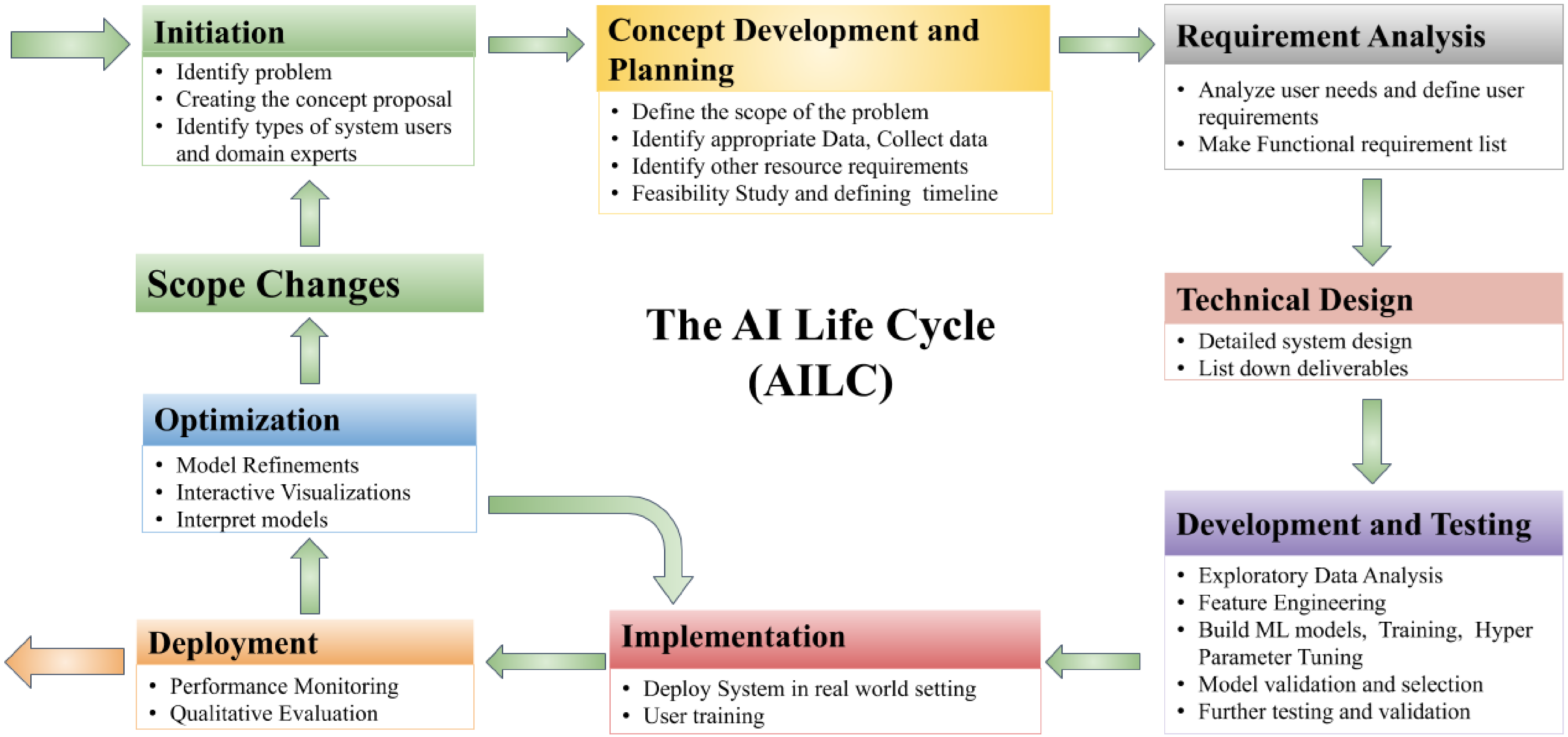
- Artificial Intelligence (AI) is influencing almost all areas of human life.
- Humans still hesitate to develop, deploy, and use AI systems
  - deficiency of transparency (internal decision making process)
- Trustworthy AI
  - diverse research area which includes fairness, robustness, explainability, accountability, verifiability, transparency, and sustainability of AI systems
- Contributions:
  - Guidelines for building human trust to improve the interactions between human and AI systems
  - Concise survey on concepts of trustworthy AI

Introduction: Human AI Interactions

- Human System Interaction (HSI)/ Human AI interactions:
  - design, development, and research on effective interactions between humans and intelligent systems
- During AI system life cycle, three main actors communicate with each other



Background: AI Life Cycle (AILC)



Background: Human System Interactions During AILC

- AI and Developers
  - Development and Testing, Implementation, Deployment, Scope Changes, and Optimization phases

- AI and Users
  - Implementation, Deployment, and Optimization phases

- Developers and Users
  - Initiation, Concept Development and Planning, Implementation, and Optimization phases

Survey: Trustworthy AI

- Definition
  - *Ethical principles* together with formal *AI system verification techniques* to define trustworthy AI, with the common goal of allowing people and societies to develop, deploy, and use AI systems without fear
- High-Level Expert Group on Artificial Intelligence (HLEGAI):
  - ‘Striving towards Trustworthy AI concerns not only the trustworthiness of the AI system itself, but requires a holistic and systemic approach, encompassing the trustworthiness of all actors and processes that are part of the system’s socio-technical context throughout its entire life cycle’

Survey: Trustworthy AI Principles

Trustworthy AI Guidelines to improve HSIs

HSIs

- AI system and Developers
  - Global Interpretability
  - Local Interpretability
  - Interactive visualizations
  - Anomaly Detection mechanisms
- AI system and Users
  - Local Interpretability
  - Interactive visualizations/GUIs
  - Performance
  - Diversity, Non-discrimination and Fairness
- Developers and Users
  - Scope of HSIs
  - Rules and Regulations for HSIs
  - Privacy and Data Governance
  - Transparency
  - Standardisation and Documentation of HSIs

Guidelines for Interaction Between AI and System Developers

- Global interpretability
  - analyze AI system, right outcomes for right reasons, identify course for wrong outputs, fix defects and trust the developed system before deployment
- Local interpretability
  - adversarial samples and check how the model outcome changes with input data changes
- Interactive visualizations
  - exploring hidden patterns and model behaviors, take necessary actions efficiently
- Anomaly Detection mechanisms
  - identify abnormal scenarios (data drift, or some attacker action), update AI systems and protect

Guidelines for Interaction Between AI and System Users

- Local interpretability
  - easy enough to understand (linguistic, visual, numerical), build user trust, identify incorrect conclusions, allows the users to question the decisions made by AI system
- Interactive visualizations
  - wide range of interactive visualizations, covering large audience of users, easy to and safe learn and use by users
- Performance
  - predictive performance, time take to provide a product or service
- Diversity, non-discrimination and fairness
  - should not have biases towards certain groups of people (age, gender, abilities, characteristics)

Guidelines for Interaction Between Developers and System Users

- Define the scope of human system interaction during concept development and planning stage of AILC
  - which entities communicated during what phase, reasons for interactions, data
- Define a set of rules and regulations
  - agree on rules and regulations for possible HSIs
- Privacy and Data Governance
  - privacy and data related regulations
- Transparency
  - reasons for interactions, enabling transparency properties
- Standardisation and documentation
  - auditability, transparency, traceability, and easy refinements when necessary

Discussion, Conclusions, and Future Directions

- Guidelines for improving human trust during HSIs are:
  - context dependant, interaction dependant
- Trustworthy AI research area acts as an umbrella covering diverse research directions
  - global framework for trustworthy AI,
- Performance Measures for Trustworthiness
  - Current measures are not enough, need new quantitative and qualitative measures
  - Common ground for research (compare and verify)
- Removing humans entirely from the loop can harm the trust of humans: AI Augmentation

AI Augmentation for Trustworthy AI: Augmented Robot Teleoperation

Daniel L. Marino*, Javier Grandio*, Chathurika S. Wickramasinghe*,
Kyle Schroeder†, Keith Bourne†, Afroditi V. Filippas*†, Milos Manic

**Virginia Commonwealth University, VA, USA*

†Commonwealth Center for Advanced Manufacturing (CCAM)

Abstract

- **Motivation**: Despite the performance of AI systems, some sectors hesitate to adopt AI because of a lack of trust in these systems.
- **Thesis**: Use AI Augmentation as a path for building Trustworthy AI.
  - Augmentation provides a preferred alternative over complete Automation.
  - Instead of replacing humans, AI Augmentation uses AI to improve and support human operations, creating an environment where humans work side by side with AI systems.
- **What we present**:
  - Design guidelines and motivations for the development of AI Augmentation for Robot Teleoperation.
  - The design of a Robot Teleoperation testbed for the development of AI Augmentation systems.

Trustworthy AI

- **Trust:** predictable behavior, even in the presence of uncertainty.
- Two main components:
  + Intentions
  + Competence

- **Trustworthy AI**: combination of diverse research areas on AI systems:
  + Fairness, robustness, explainability, accountability, verifiability, transparency and sustainability
  + Goals:
    - Identify factors which harm the human trust of AI systems
    - Introduce methods to improve human trust in AI systems

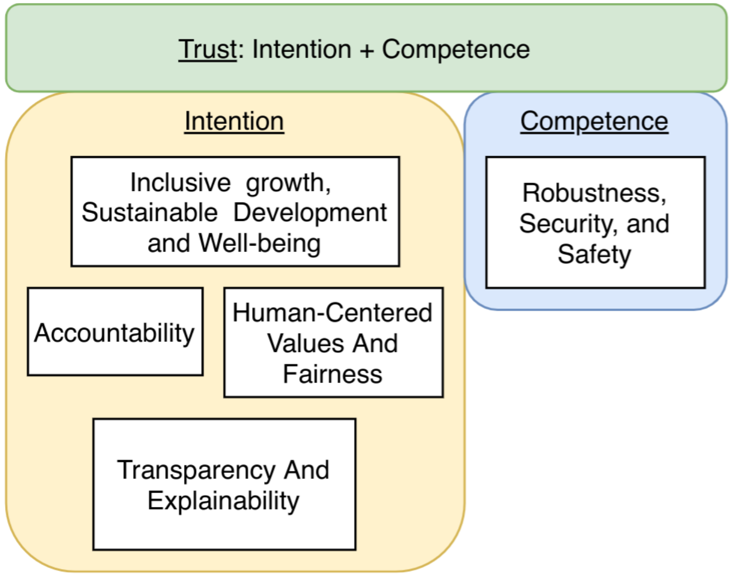


Fig. 1. Trustworthy AI Principles defined by the OECD

Augmented AI for Trustworthy AI

- Augmented AI: AI technologies working alongside humans
  - + Improve productivity, efficiency, quality of human activities, and enhance human-machine cognition
  - + Build trust
- Shared Autonomy
  - + Split tasks between AI and Humans
  - + High risk decisions made by humans
  - + Maintain *Accountability* in humans

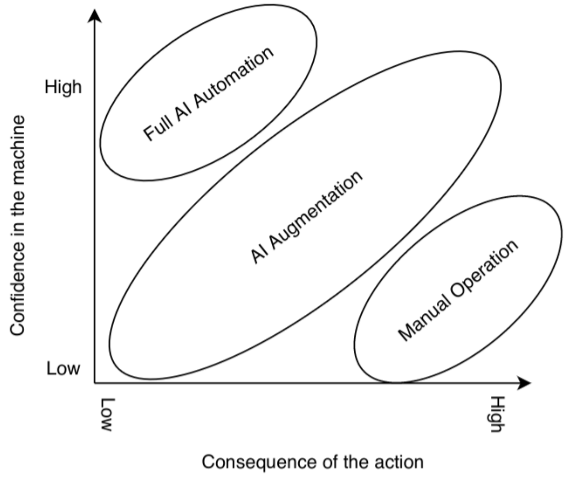


Fig. 2. Augmentation vs Automation

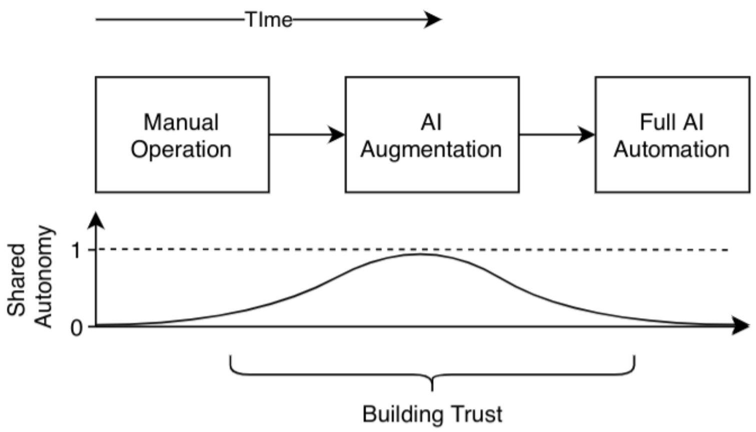


Fig. 3. Building trust

Prototype for Robot Teleoperation

Robot Teleoperation:

- Provides an environment to study Human-Machine interactions
- Shared Autonomy is embedded in the field

Objective: AI Augmentation block

- – Uses feedback from multiple sensors to perform a commanded action with high success rate.
- – The AI will combine data from several sources in order to have a complete representation of the environment.

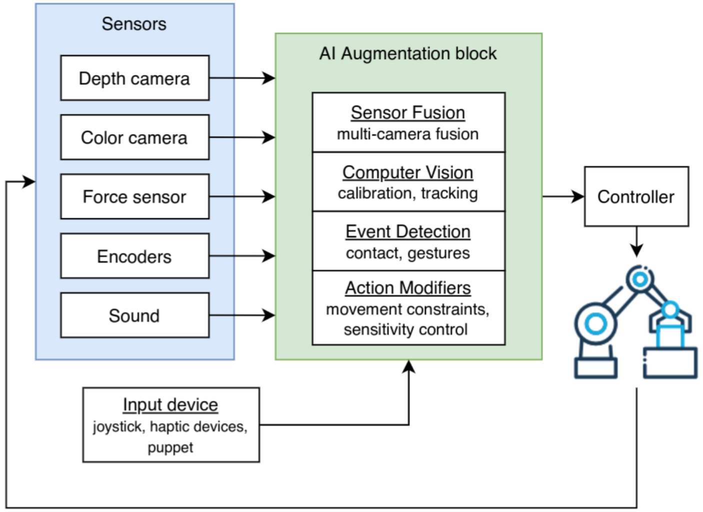


Fig. 4. AI Augmentation for Teleoperation

Testbed

- Hardware:
  - Universal Robots CB-Series **UR5 robot** with standard controller
  - Die **grinder** as end effector
  - 6-axis **force sensor** mounted to the wrist
  - Two stereo GigE cameras for **stereoscopic** visualization
  - Three Intel RealSense D435i cameras for **point-cloud** acquisition
  - Microphone
- Custom made Input device
  - Intuitive user input (Figure 6)
  - Similar kinematics to UR5 robot
  - Vibration motors for haptic feedback

Fig. 5. Testbed

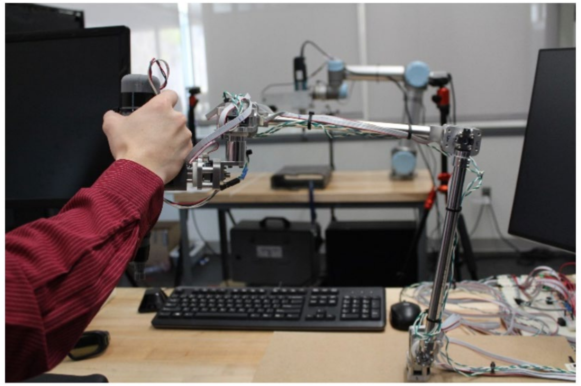


Fig. 6. Input Device

Augmented Reality (AR) Interface

- Rviz is used for visualization
- Virtual objects are super-imposed in the scene to provide improved situational awareness

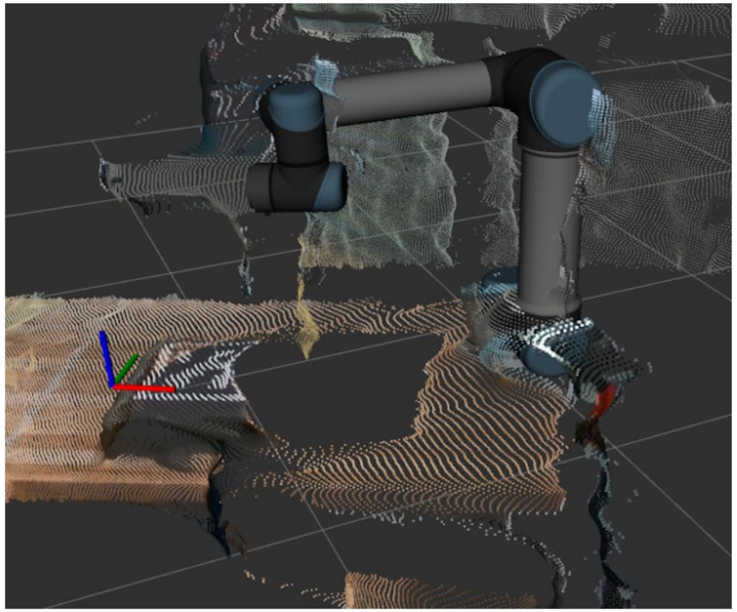


Fig. 8. Visualization of Data

Guidelines for the development of Trustworthy AI Augmentation

- Prevent misuse by keeping users engaged
  - Human should retain control, actively engage in the task
- Assess uncertainty
  - Ensure robustness, AI is aware of situations where there is not enough information to act autonomously
- Clear communication of cause-and-effect by effective use of the Augmented Reality interface
  - Clearly communicate the actions taken by the AI, improve transparency, ensure the intent of the AI
- Clear behavior in presence of uncertainty
  - Increase caution proportionally to the level of uncertainty and the chances of failure

Conclusions

- AI Augmentation over full automation provides a path for building Trustworthy AI systems
- Developed a testbed for the development of AI Augmented Teleoperation
  - the hardware setup
  - the software stack
- Presented a series of guidelines for the development of Trustworthy AI Augmentation for Robot Teleoperation

Reference:

Paper 1: **Trustworthy AI Development Guidelines for Human System Interactions**

Paper 2: **AI Augmentation for Trustworthy AI: Augmented Robot Teleoperation**

Chathurika Wickramasinghe: brahmanacsw@vcu.edu
Daniel Marino: marinodl@vcu.edu
Prof. Milos Manic misko@ieee.org
Research Lab: Modern Heuristic Research Group (MHRG), Virginia Commonwealth University

Thank You

NS&T ML-AI

Big Data Machine Learning Artificial Intelligence

Rick Vilim
Argonne National Laboratory

IMPROVING THE EXPLAINABILITY OF AI THROUGH INCLUSION OF PROCESS INFORMATION AND AUTOMATED REASONING

R. VILIM, T. NGUYEN, R. PONCIROLI
Nuclear Science and
Engineering Division
Argonne National Laboratory

Machine Learning & Artificial Intelligence Symposium
Idaho National Laboratory
February 09, 2021

U.S. DEPARTMENT OF ENERGY
Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.

INTRODUCTION

Industry AI applications based largely on data-driven ML approaches

- First generation AI for nuclear power plants was data-driven (DD)
  - Multivariate State Estimation (MSET-ANL) for sensor fault detection. Circa 1990’s.
- Installed capability today is still largely data-driven
  - Advantage: One-size-fits-all (in principle)
  - Disadvantage: Shallow, opaque, brittle
- On-going work aims to add in process information (domain knowledge)
  - Physics-based (PB) knowledge can serve to further constrain the solution space to physical reality
  - E.g., Conservation laws, constitutive equations etc.

CAPABILITY	DD	PB
Immune to operating point change?	N	Y
Diagnosis resolved to specific fault?	N	Y
Rank ordering of likelihood of faults?	N	Y
Applicable to engineering systems?	—	Y
Free of need for library of fault signatures?	N	Y
Generates virtual sensors?	N	Y
Adapts upon dropped sensor?	—	Y
Yields component performance index?	N	Y
Supports design of optimal sensor set?	N	Y

RELEVANCE TO ML/AI FUTURE

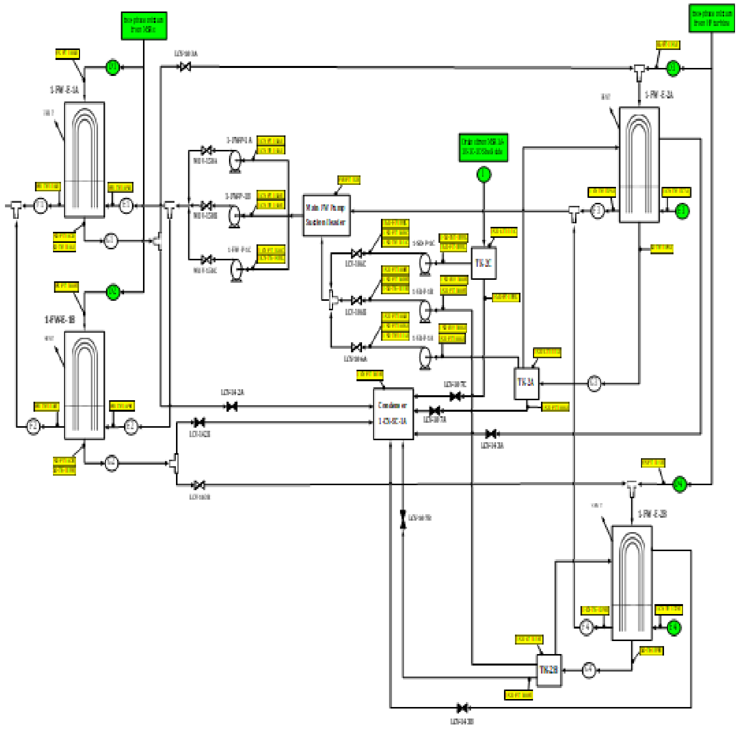
To enable widespread application of AI, business-case must improve

- Acceptable level of dependence on subject matter experts (SME)
  - On-going work aims to add in process information (domain knowledge) in the form of physics-based knowledge
  - Utilities long ago dispensed with plant system modelers
  - So, physics-based knowledge needs to be embedded in the method/software as opposed to being communicated by an SME
- Explainable
  - Strive for an underlying reasoning process that an informed human can easily follow
- Specifiable level of granularity
  - The sensor set that provides the requisite capability needs to be identifiable
- Quantifiable reliability
  - The rendered output needs to be qualified as to its uncertainty

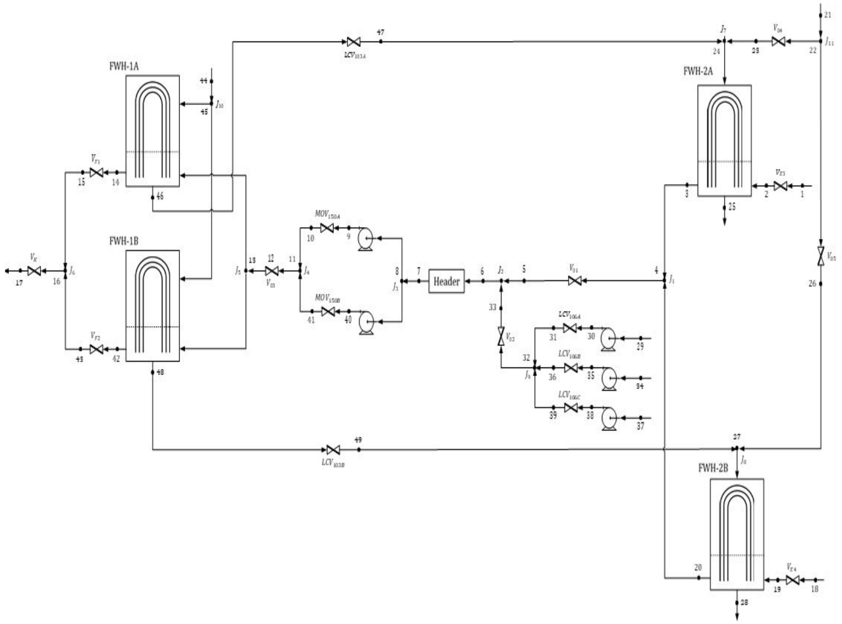
EXAMPLE – FAULT DIAGNOSIS IN HP FW SYSTEM (1/4)

Minimal dependency on SME

- Physics-based digital twin is assembled automatically from the engineered system P&ID



HP FW System P&ID



P&ID conversion to network diagram

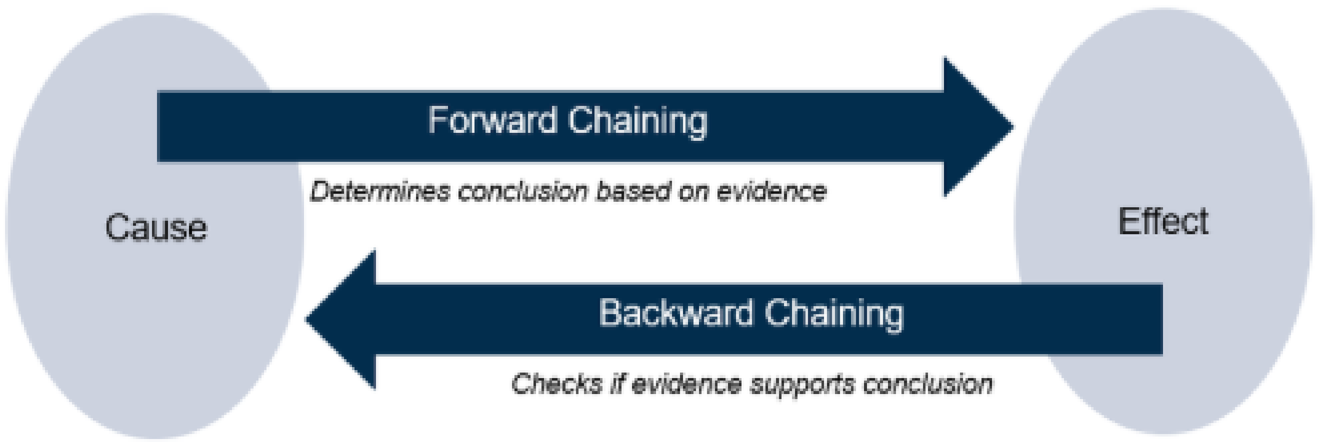
```
#` 1. List all components with their outlet index
1, fw_inlet1, system,
2, v_e3, open_valve,
3, fwh_2a_cold, vertical_fwh_cold,
4, junc_1, junction,
5, v01, open_valve,
        .
        .
#` 1e. Adjacencies
fwh_1a_cold, fwh_1a_hot
fwh_1b_cold, fwh_1b_hot
fwh_2a_cold, fwh_2a_hot
fwh_2b_cold, fwh_2b_hot
        .
        .
#` 3.  Sensors:
1, fwh_2a_cold, ti, to,
2, fwh_2b_cold, ti, to,
3, fwh_2a_hot, pi, to,
4, fwh_2b_hot, pi, to,
5, sdp_1a, pi, po, to, fo, Wa,
6, sdp_1b, pi, po, to, fo,
        .
        .
#`  5a. Loops: Components
loop_1, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,
loop_2, 18, 19, 20, 4,
loop_3, 21, 22, 23, 24, 25,
loop_4, 22, 26, 27, 28,
loop_5, 29, 30, 31, 32, 33, 6,
```

Conversion to text-based file

EXAMPLE – FAULT DIAGNOSIS IN HP FW SYSTEM (2/4)

High explainability

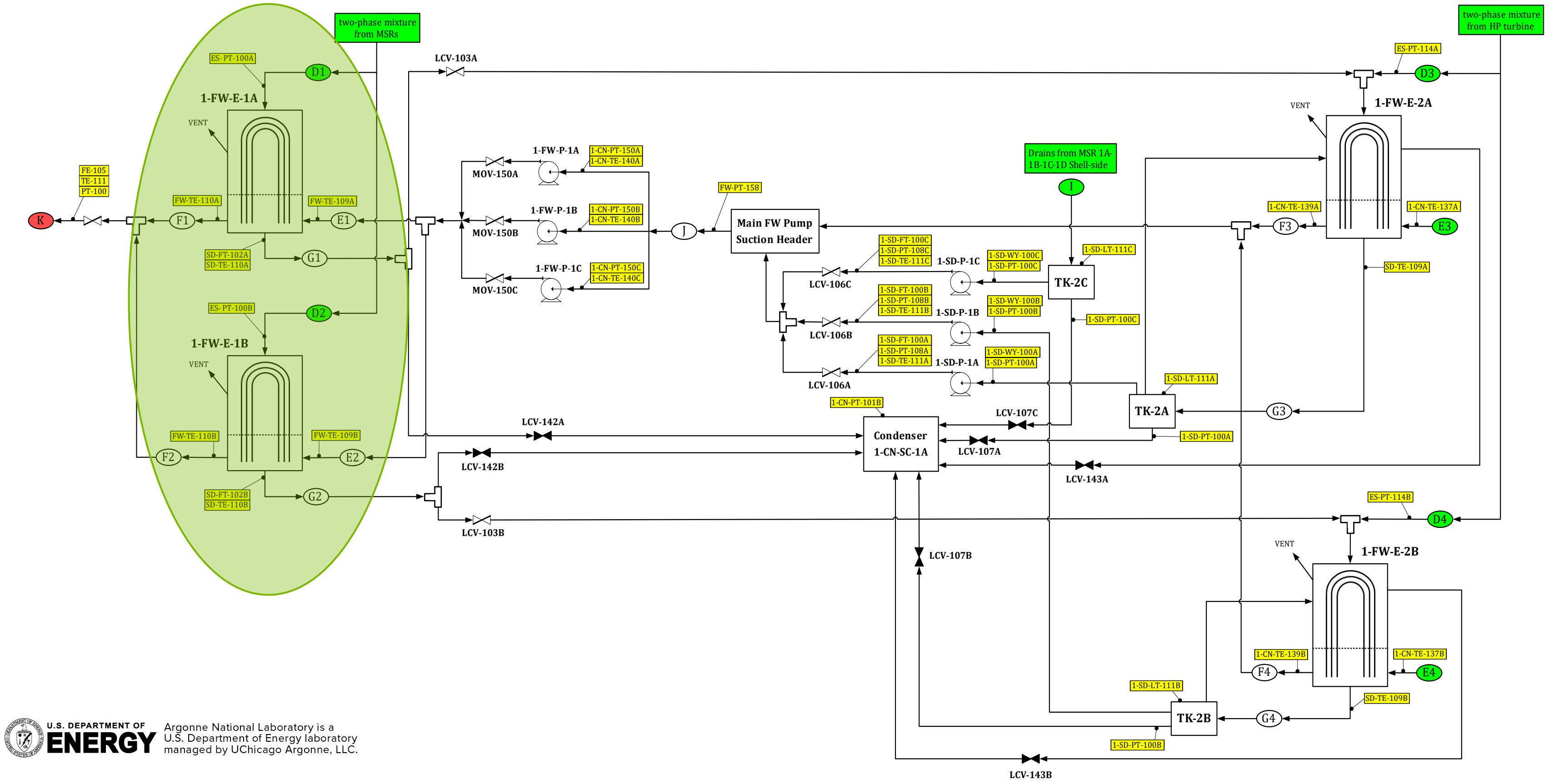
- Use of automated reasoning in the diagnostic process is one way to provide an accessible understanding of how a diagnosis was arrived at



EXAMPLE – FAULT DIAGNOSIS IN HP FW SYSTEM (3/4)

Requisite granularity

- Sensor set for the first-point FW heaters is sufficient to uniquely identify the requisite component and sensor faults



Fault ID	Component	Fault
1	1-FW-E-1A	Fouling
2	1-FW-E-1A	Tube leak
3	1-FW-E-1A	Shell leak
4	1-FW-E-1A	Tube block
5	1-FW-E-1B	Fouling
6	1-FW-E-1B	Tube leak
7	1-FW-E-1B	Shell leak
8	1-FW-E-1B	Tube block
9	FE-105	Sensor fault
10	FW-TE-109A	Sensor fault

Fault ID	Component	Fault
11	FW-TE-110A	Sensor fault
12	SD-FT-102A	Sensor fault
13	SD-TE-110A	Sensor fault
14	ES-PT-100A	Sensor fault
15	FW-TE-109B	Sensor fault
16	FW-TE-110B	Sensor fault
17	SD-FT-102B	Sensor fault
18	SD-TE-110B	Sensor fault
19	ES-PT-100B	Sensor fault

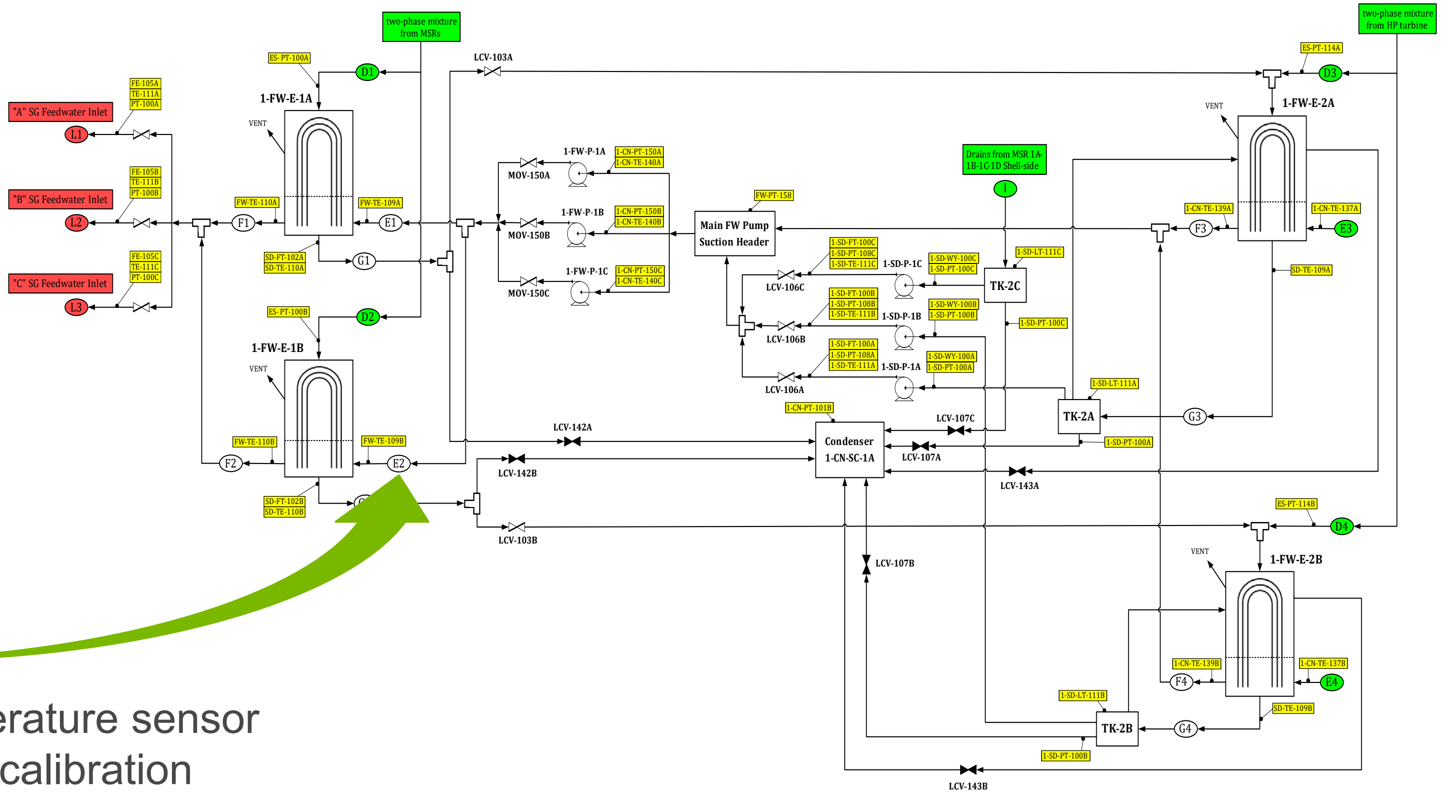
EXAMPLE – FAULT DIAGNOSIS IN HP FW SYSTEM (4/4)

High reliability

- Diagnoses are rank ordered in terms of probability

Probability of different fault diagnoses

Fault	Symptoms
	r_2, r_6, r_9
Sensor E2.T	25.7%
Sensor F2	25.7%
Sensor G2.w	25.7%
Sensor G2.T	25.7%
FWH 1B, Fouling	10.1%
FWH 1B, Shell leak	5.1%
Sensor D2.P	5.0%



Temperature sensor out of calibration

U.S. DEPARTMENT OF ENERGY
Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.
Argonne NATIONAL LABORATORY

LOOKING AHEAD

Challenges

- Identifying the requisite sensor set
- Incorporating a mechanistic/physics-based treatment of the evolution of degradation processes that limit the lifetime of a component
- When degradation cannot be measured directly, then virtual indications for the state of degradation are needed
- Comprehensive policy for data formatting, curation, and archiving that begins with design of the nuclear facility information system

NS&T ML-AI

Big Data Machine Learning Artificial Intelligence

Ross Kunz
Idaho National Laboratory

Ross Kunz

ross.kunz@inl.gov

Exploring reaction mechanisms with explainable AI.

Kunz M.R. , Wang, Y., Batchu R., Fang, Z., Fushimi, R.

Idaho National Laboratory

Yonge, A., Medford, A.J.

Georgia Institute of Technology

Constales, D.

Ghent University

Yablonsky, G. S.

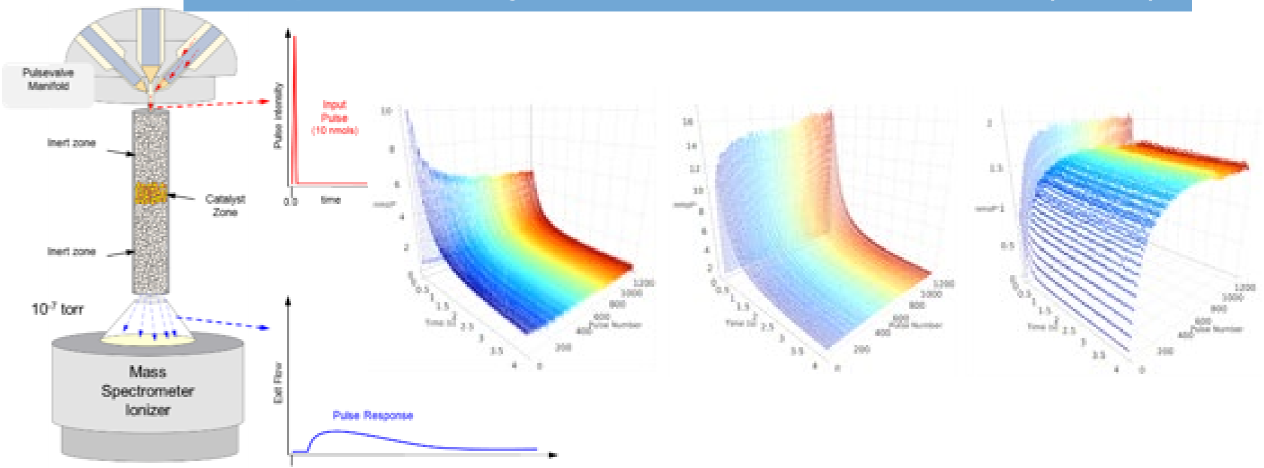
McKelvey School of Engineering

Idaho National Laboratory

Catalysis Informatics Goals

- Understanding the how and why an industrial catalyst behaves
- **Data driven** mechanism understanding by transient kinetics
  - Measuring micro-kinetic coefficients
  - Fingerprinting mechanisms of industrial catalysts

Temporal Analysis of Products Reactor (TAP)

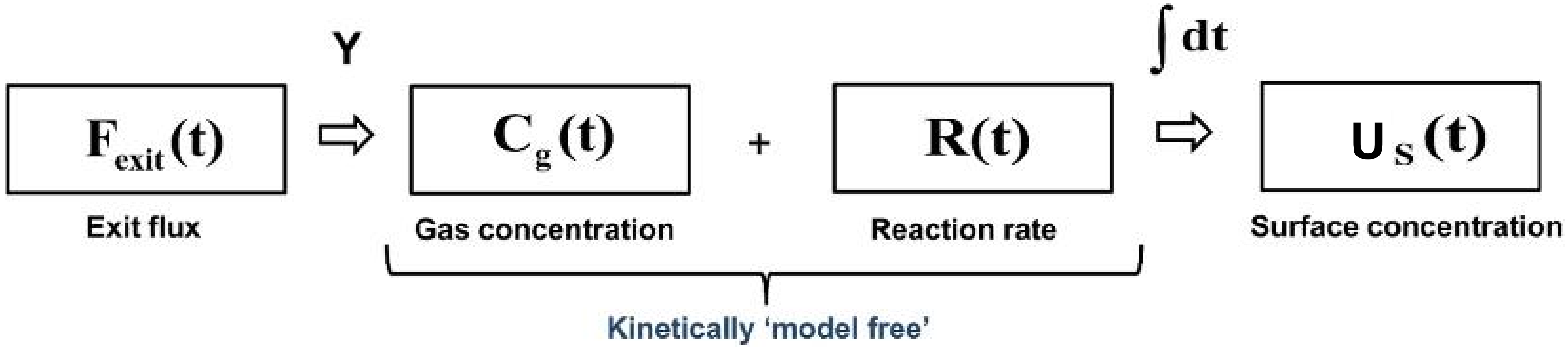


Kondratenko et al, Micro-kinetic analysis of direct N20 decomposition over steam-activated Fe-Silicate from transient experiments in the TAP reactor, 2006

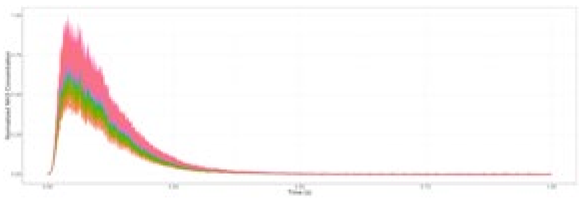
Table 1
Reaction schemes for direct N_2O decomposition evaluated in this study

Number	Elementary reaction steps	
1	$N_2O + * \rightarrow N_2 + *{-}O$	(1)
	$*{-}O + *{-}O \rightarrow O_2 + 2*$	(2)
2	$N_2O + * \rightarrow N_2 + *{-}O$	(1)
	$N_2O + *{-}O \rightarrow N_2 + O_2 + *$	(2)
3	$N_2O + * \rightarrow N_2 + *{-}O$	(1)
	$*{-}O + *{-}O \rightarrow *{-}O_2 + *$	(2)
	$*{-}O_2 \rightarrow O_2 + *$	(3)
4	$N_2O + * \rightarrow N_2 + *{-}O$	(1)
	$N_2O + *{-}O \rightarrow N_2 + *{-}O_2$	(2)
	$*{-}O_2 \rightarrow O_2 + *$	(3)
5	$N_2O + * \rightarrow *{-}O + N_2$	(1)
	$N_2O + *{-}O \rightarrow O{-}*{-}O + N_2$	(2)
	$O{-}*{-}O \rightarrow *{-}O_2$	(3)
	$*{-}O_2 \rightarrow O_2 + *$	(4)
6	$N_2O + * \rightarrow *{-}O + N_2$	(1)
	$N_2O + *{-}O \rightarrow *{-}O_2 + N_2$	(2)
	$N_2O + *{-}O_2 \rightarrow *{-}O_3 + N_2$	(3)
	$*{-}O_2 \rightarrow O_2 + *$	(4)
	$*{-}O_3 \rightarrow O_2 + *{-}O$	(5)

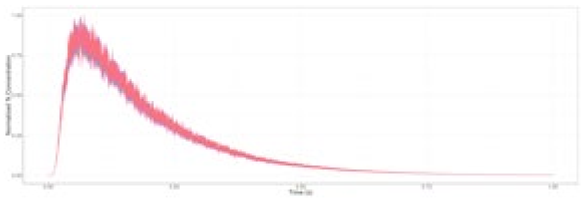
Link to Machine Learning



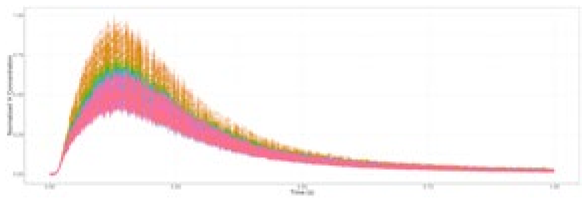
Ammonia



Nitrogen



Hydrogen



Elementary Step	Rate Expression	Linear Form
$A + * \rightarrow A^*$	$r = k^+ C_A \theta$	$r_A = \beta_{(Nk^+)} C_A - \beta_{(k^+)} C_A U_{A^*}$
$A + * \rightleftharpoons A^*$	$r = k^+ C_A \theta - k^- C_{A^*}$	$r_A = \beta_{(Nk^+)} C_A - \beta_{(k^+)} C_A U_{A^*} - \beta_{(k^-)} U_{A^*}$
$A + 2* \rightarrow 2A^*$	$r = k^+ C_A \theta^2$	$r_A = \beta_{(N^2 k^+)} C_A - \beta_{(2Nk^+)} C_A U_{A^*} + \beta_{(k^+)} C_A U_{A^*}^2$

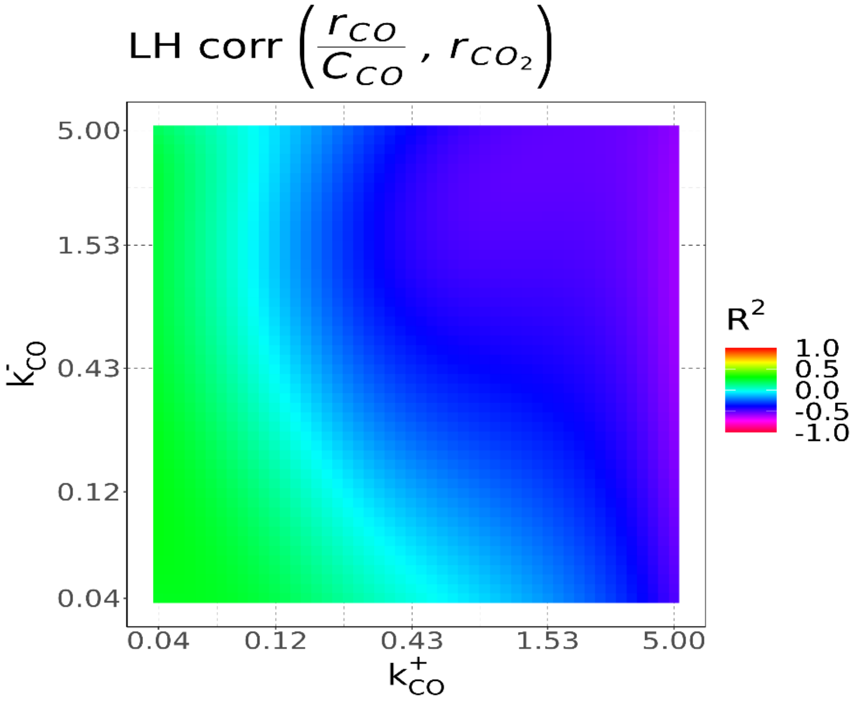
$$r_A = \beta_{(Nk^+)} C_A - \beta_{(k^+)} C_A U_{A^*} - \beta_{(k^-)} U_{A^*} + \beta_{(k^+)} U_{A^*}^2$$

Yablonsky et al, **The Y-Procedure: How to extract the chemical transformation rate from reaction-diffusion data with no assumptions on the kinetic model**. 2007
Yablonsky et al, **Rate-Reactivity Model: A New Theoretical Basis for Systematic Kinetic Characterization of Heterogeneous Catalysts**, 2016

Application to Kinetic Information: via Penalization and Covariance Structure Estimation

$$r_A = \beta_{(Nk^+)} C_A - \beta_{(k^+)} C_A U_{A^*} - \beta_{(k^-)} U_{A^*} + \beta_{(k^+)} U_{A^*}^2$$

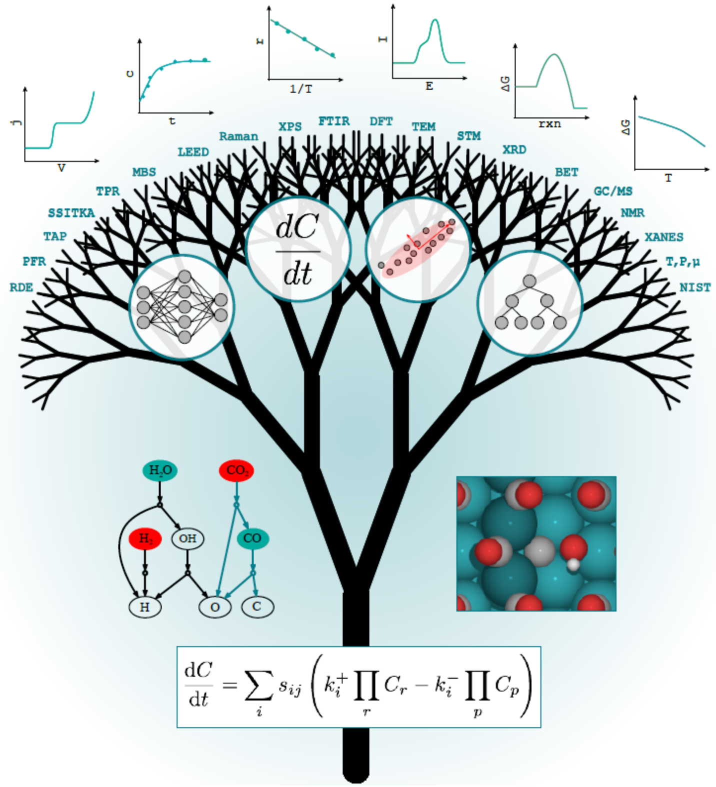
Mechanism:	RMSE	NPV
Irreversible (abundant sites)	0.000	1
Irreversible (limited sites N=1)	0.000	1
Irreversible (limited sites N=2.5)	0.000	1
Reversible (limited sites N=1)	0.420	1



- TAP enables data driven kinetic coefficient estimation
- Understanding about key contributors to catalyst performance
- Machine learning algorithms must be tailored to physical assumptions

Looking ahead / Challenges

- Concurrently optimizing the correlation structure with the linear relationships
- Developing indicators of complex physical phenomena
- Linking structural and kinetic characterization information (data fusion)
- Developing links to transition states from TAP kinetics



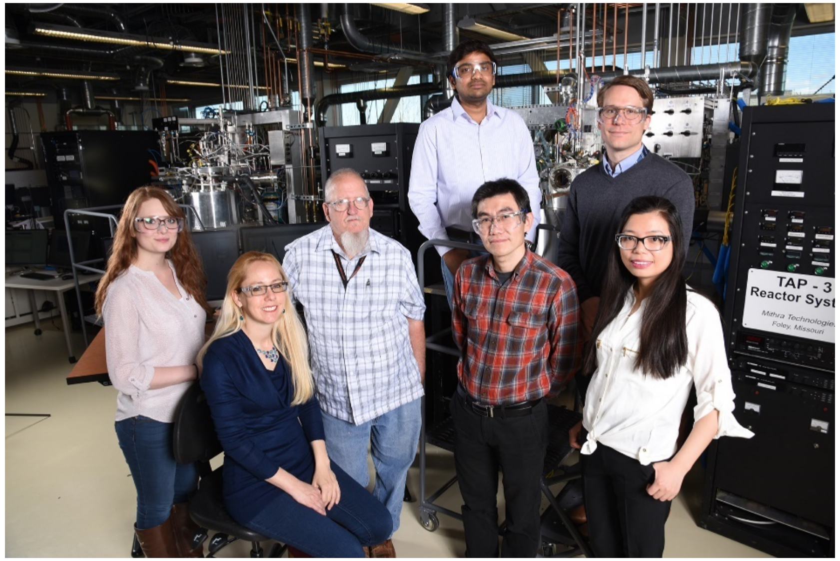
Deriving Understanding through the Combination of Physics and Experiments

Medford et al. Extracting knowledge from data through catalysis informatics. 2018

Acknowledgements

Funding agency: This work was supported by U.S. Department of Energy (USDOE), Office of Energy Efficiency and Renewable Energy (EERE), Advanced Manufacturing Office Next Generation R&D Projects under contract no. DE-AC07-05ID14517.

Group members at INL:



NS&T ML-AI

Big Data   Machine Learning   Artificial Intelligence

Akshay J. Dave

MIT NRL

Integration of Neural Networks in Control of a Subcritical Facility

Current progress and Opportunities for XAI

Akshay J. Dave

Research Scientist, MIT NRL

This work is partially sponsored by the U.S. Department of Energy NEUP Award Number DE-NE0008872.

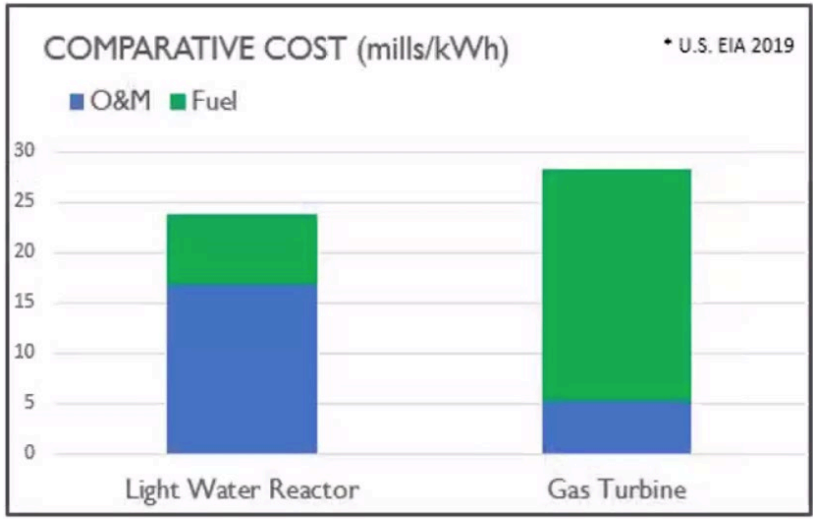
Nuclear Reactor Laboratory

Motivation For Autonomous Control

- Critical factors for economic competitiveness of NPPs:
  - Up-front capital cost for construction
  - Day-to-day cost of plant management
    - ~1 person / 2 MWe generated [1]
    - O&M account for 66% of Operating costs [2]

- Autonomous control has not been implemented in an operating reactor or developed for emerging concepts [1]
  - Research in universities/labs

- Need for automation
  - Small modular/micro-reactors
  - Current fleet
  - Space exploration [3]



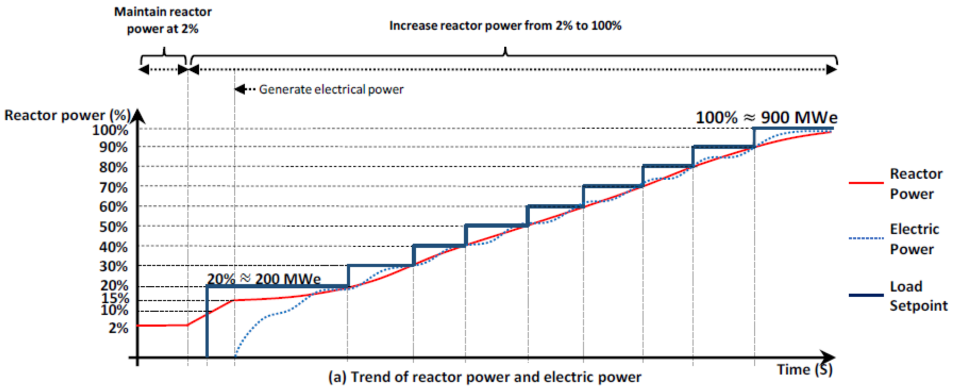
Current/near-term Paradigm

"NuScale researchers want to operate 12 small nuclear reactors from a single control room. They built a mock one in Corvallis, Oregon, to show they can do it." Science (2019)

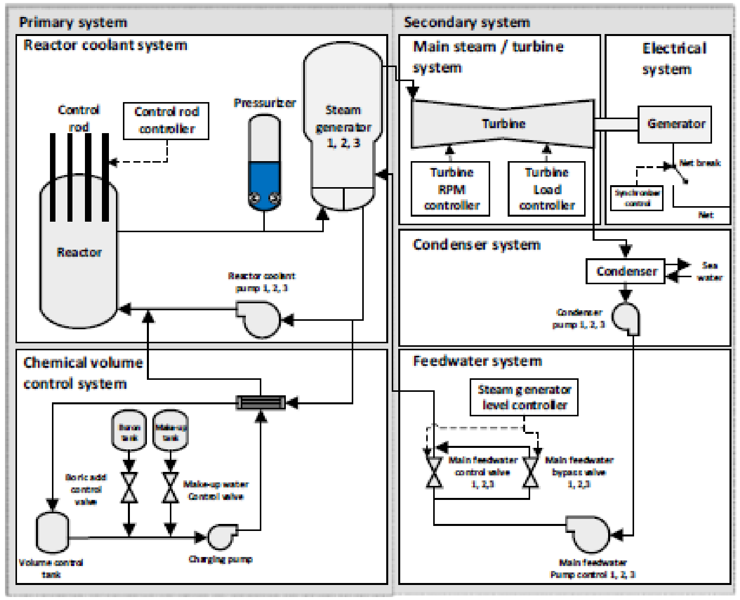
[1] Wood, et al., "An autonomous control framework for advanced reactors", Nuclear Engineering and Technology, 2017.
[2] https://www.world-nuclear.org/information-library/economic-aspects/economics-of-nuclear-power.aspx
[3] Upadhyaya, et al., "Autonomous control of space reactor systems", DOE/ID/14589, 2007.

NPP CONTROL SOTA

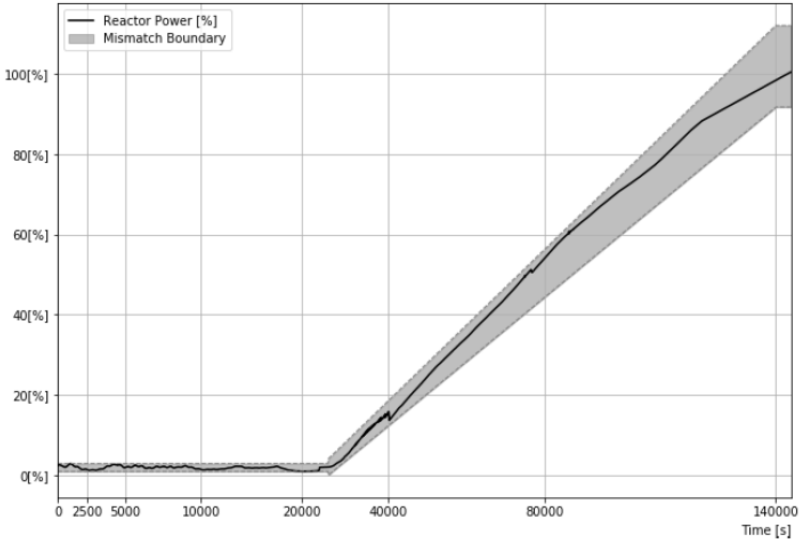
- Algorithm
  - *Piece-wise* mature
- Demonstration
  - How do we begin the process of experimentally demonstrating autonomous control?



(a) Trend of reactor power and electric power

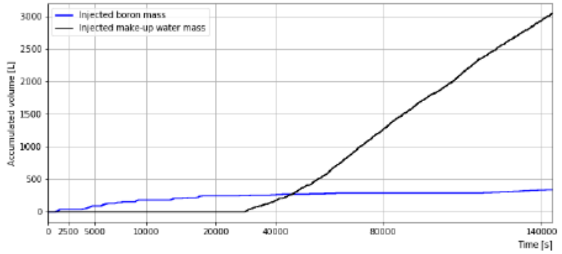


RL Agent Controlled 2% to 100% Power

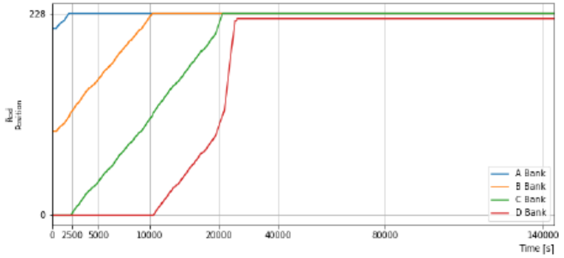


RL Agent Controlled Systems

CVCS Boron



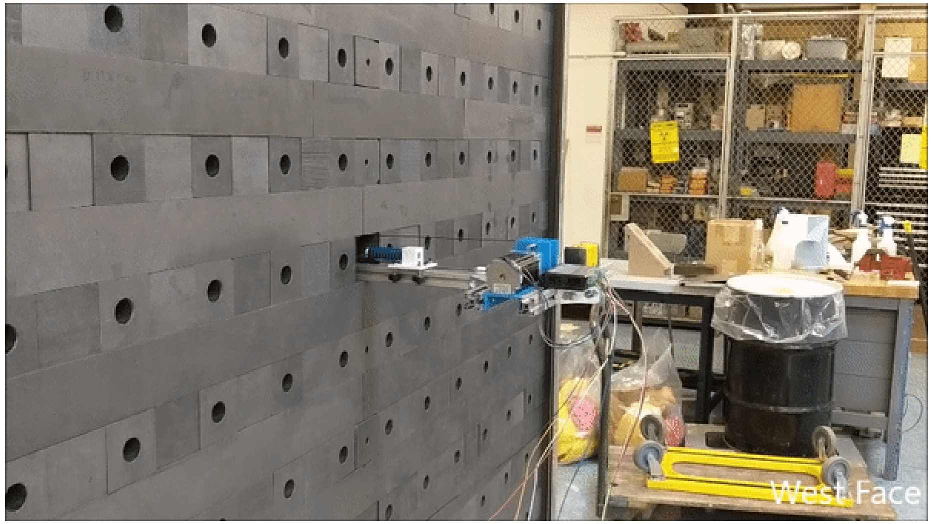
CR Banks

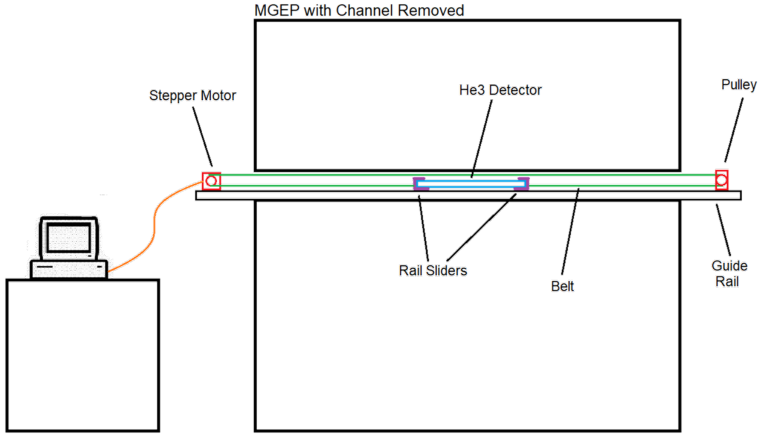


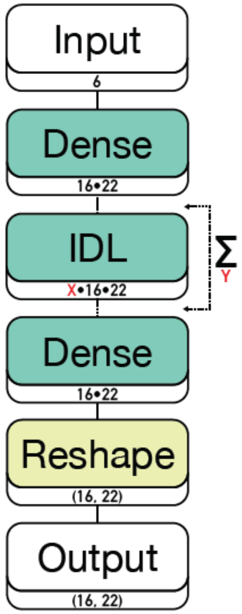
Images: D. Lee, et al., "Algorithm for Autonomous Power-increase Operation Using Deep Reinforcement Learning and a Rule-Based System," IEEE Access, 2020.

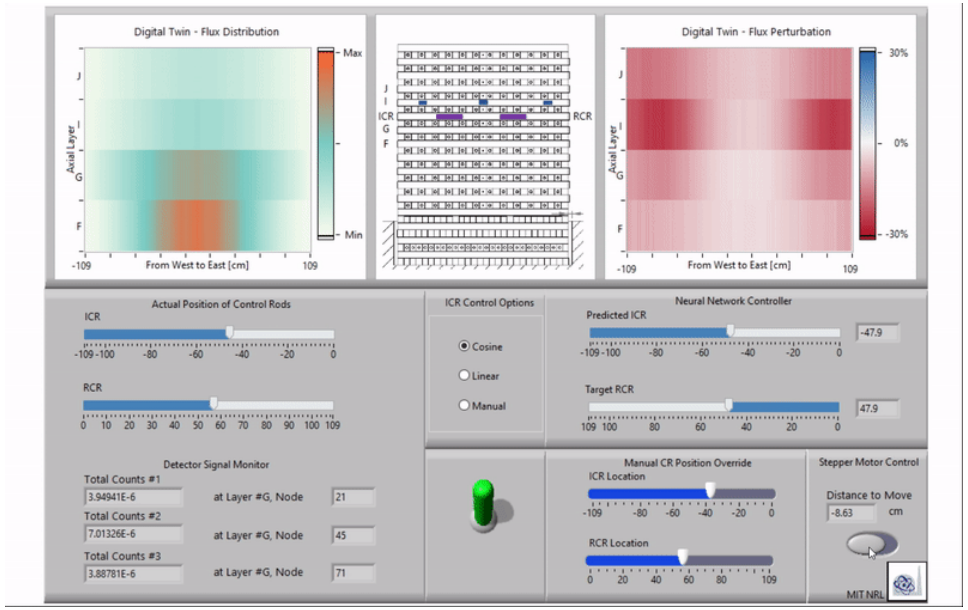
Current Progress: Autonomous Control of the MIT Graphite Exponential Pile

- MGEP Specifications:
  - 90” cube
  - 1,288 natural uranium slugs
  - Subcritical ($k_{eff} \approx 0.8$)
- Objective:
  - Design and construct an experimental facility that can demonstrate an autonomous framework, embedding state-of-the-art ML methods









ML Model Development: A. J. Dave, el al., “Deep Surrogate Models for Multi-dimensional Regression of Reactor Power,” ANS Winter Conference 2020 (preprint: https://arxiv.org/abs/2007.05435)

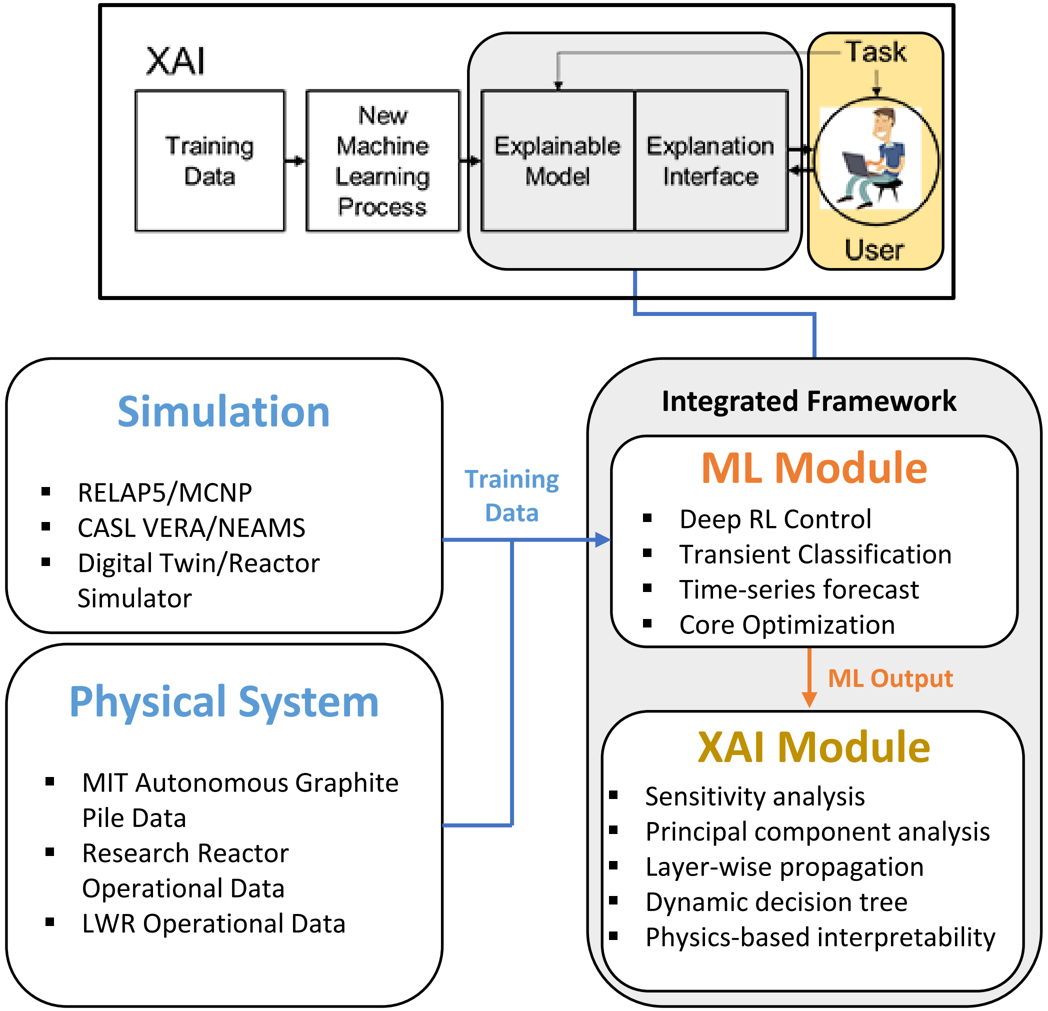
XAI & NPP CONTROL

- The efficacy of XAI methods hinges on two aspects:
  - Development of tight coupling between ML and XAI methods (context aware)
  - End-user traction
    - Operators that might be overseeing the control actions made by DRL systems

- Development of an <u>integrated XAI framework</u> that has been demonstrated experimentally
  - There is significant overlap in the underlying ML methods we will use for varying reactor designs

    → collaboration via open-source development
  - We need to assess human factors with end-users, not ML experts

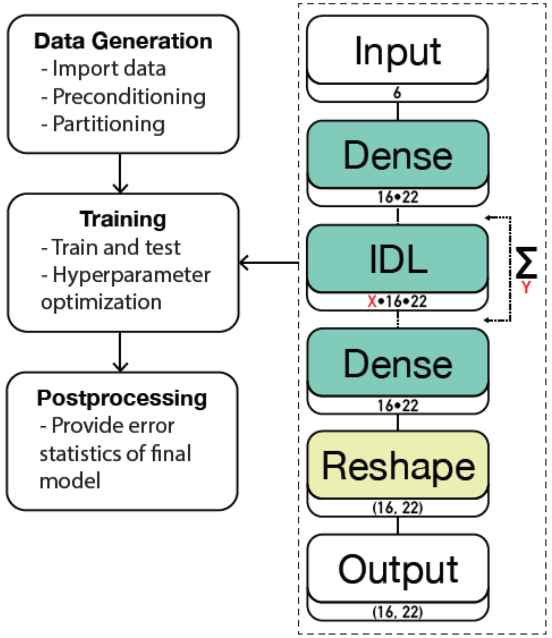
    → collaboration with research reactors



Top Image: M. Turek, "Explainable Artificial Intelligence (XAI)", https://www.darpa.mil/program/explainable-artificial-intelligence

OPPORTUNITIES

- The MGEP is an ideal starting point:
  - The MGEP facility is an **inherently safe** system that poses no criticality safety risk
  - Our experimental data, OpenMC model, neural network software, control system framework is/will be **open sourced**
- There is a pedagogical opportunity to allow students, researchers, and engineers to train & upload their methods online, and experiment without any criticality safety risk



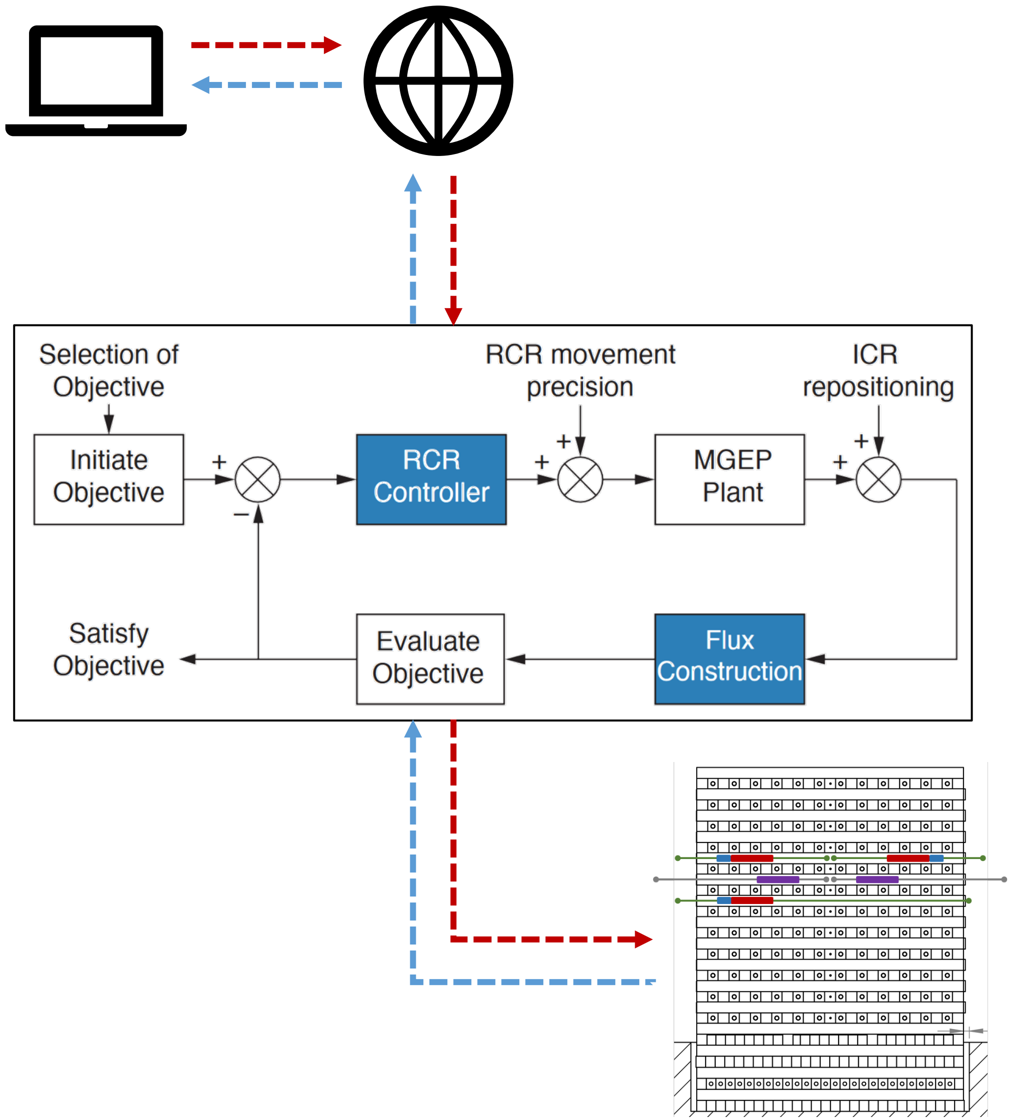
Source code: **github.com/a-jd/npsn**
Install with pip: **pip install npsn**

```python
import npsn

# Define dataset directory
data_dir = '~/some/data_location'
# Define model name (for output file label)
proj_nm = 'npsn_surrogate'

# Define number of control blades
n_x = 6
# Define nodalization of power distribution
n_y = (16, 22)  #(axial_nodes, fuel_locations)

# Train neural network without optimization
npsn.train(proj_nm, data_dir, n_x, n_y)
# Or with optimization
npsn.train(proj_nm, data_dir, n_x, n_y, max_evals
    =100)
# Post-process to quantify error
npsn.post(proj_nm)
```



NS&T ML-AI

Big Data Machine Learning Artificial Intelligence

Thank you