# A Framework for Evaluating the Maturity Level of Machine Learning Explanations

Katherine Goode

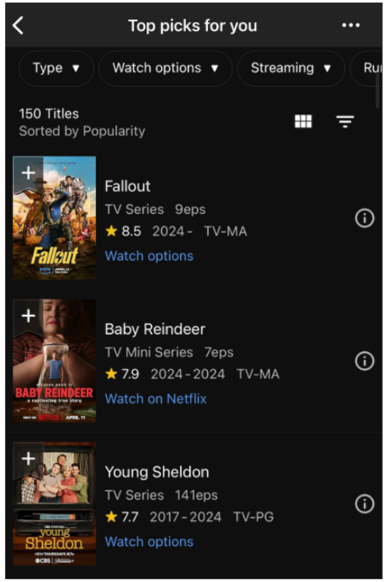
Erin Acquesta (PI), Candace Diaz, Raga Krishnakumar, Ernesto Prudencio

August 6, 2024

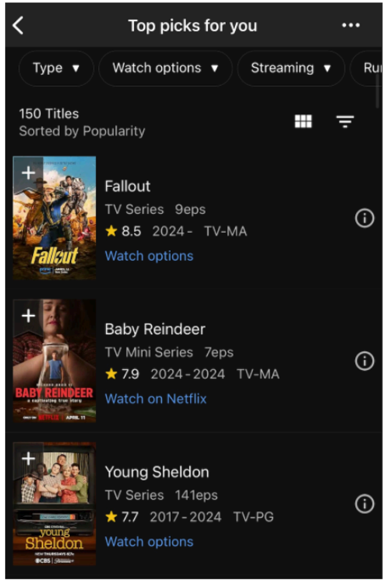
SAND2024-09822C

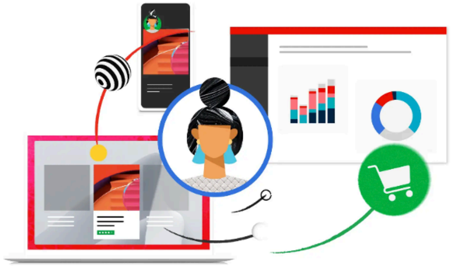
Sandia National Laboratories

# When would you be willing to use machine learning for decision making?




# What "boxes" need to be "checked" to use ML when consequences increase?



# What "boxes" need to be "checked" to use ML when consequences increase?


Sandia's five major program portfolios

Sandia National Labs is a *federally funded research and development center (FFRDC)* managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc.

FFRDCs are long-term strategic partners to the federal government...

> operating in the public interest with objectivity and independence and **maintaining core competencies in missions of national significance.**

# Overview

## Motivation and Background

CompSim Models, Maturity Levels, and SciML

## Proposed Framework

Maturity Levels for Explainability/Interpretability with SciML

## Discussion

Challenges and Moving Forward

# Motivation and Background

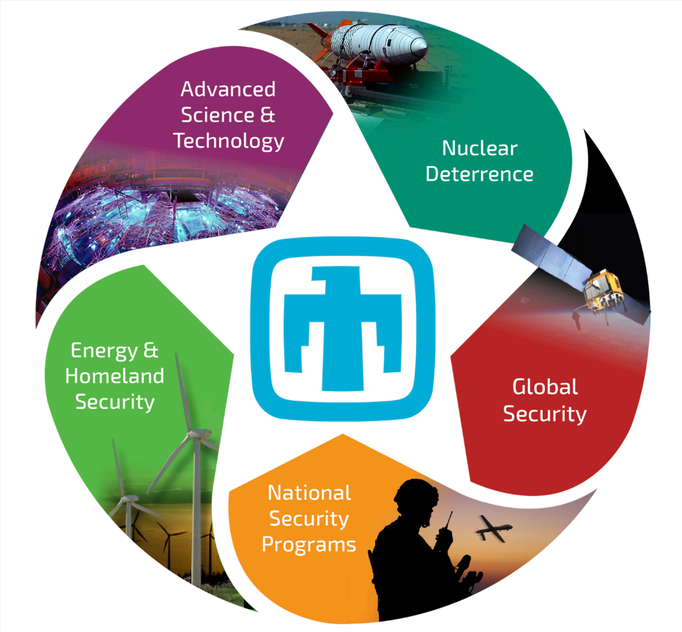
CompSim Models, Maturity Levels, and SciML

# Machine Learning at Sandia

The National Nuclear Security Administration (NNSA) Labs **emphasize trusted artificial intelligence (AI) as a necessity** for it to meet national security mission delivery.

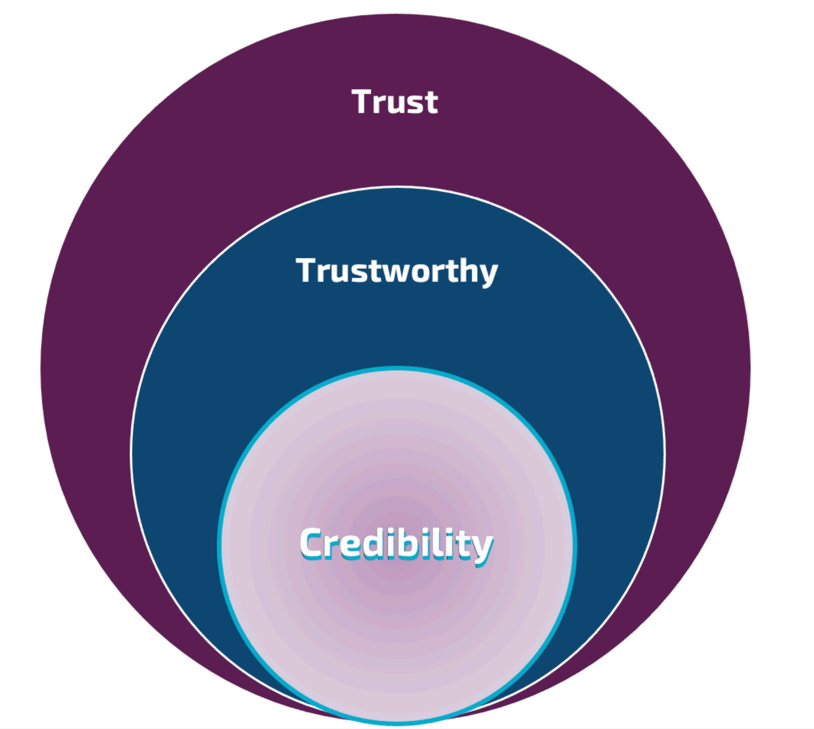
## Motivation

- Machine learning (ML) holds great potential for mission critical applications
- Evaluating the credibility of current techniques poses challenges that may hinder widespread acceptance and use
- Sandia's mission needs set us apart from industry and academia
  - High-consequence applications, domain expertise plays a critical role in model construction, etc.




The NNSA Labs must **strike a balance between leveraging the advantages of ML while ensuring its responsible use** for national security purposes.

# ML Trust/Trustworthy/Credibility at Sandia




**Trust** Defines the state of the decision maker

- Decision maker integrates interpretability/ explainability into their decision making process

**Trustworthy** Defines the state of the model

- Trustworthy interpretability/explainability is for the decision makers

**Credibility** Identifies the technical basis of the model

- Credibility of interpretability/explainability approach is for the model developer

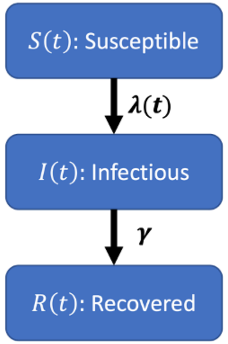
CREDIBILITY LEADS TO TRUSTWORTHY MODELS -- TRUSTWORTHY MODELS MAY ESTABLISH TRUST

# What is CompSim?

## Computational Simulation

AKA CompSim; Modeling and Simulation; ModSim; M&S

- "Computational modeling is the **use of computers to simulate and study complex systems** using mathematics, physics and computer science." - NIH
- CompSim focuses on creating mathematical models based on first principals
- Contrast to models that start with data and then aim to approximate scientific mechanisms

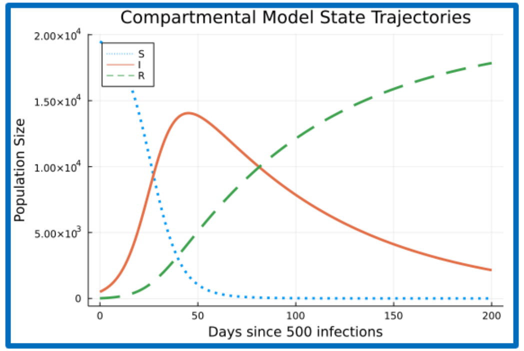



**System of ODEs:**

$$\frac{dS(t)}{dt} = -\lambda(t)S(t)$$

$$\frac{dI(t)}{dt} = \lambda(t)S(t) - \gamma I(t)$$

$$\frac{dR(t)}{dt} = \gamma I(t)$$




**Associated biological description:**

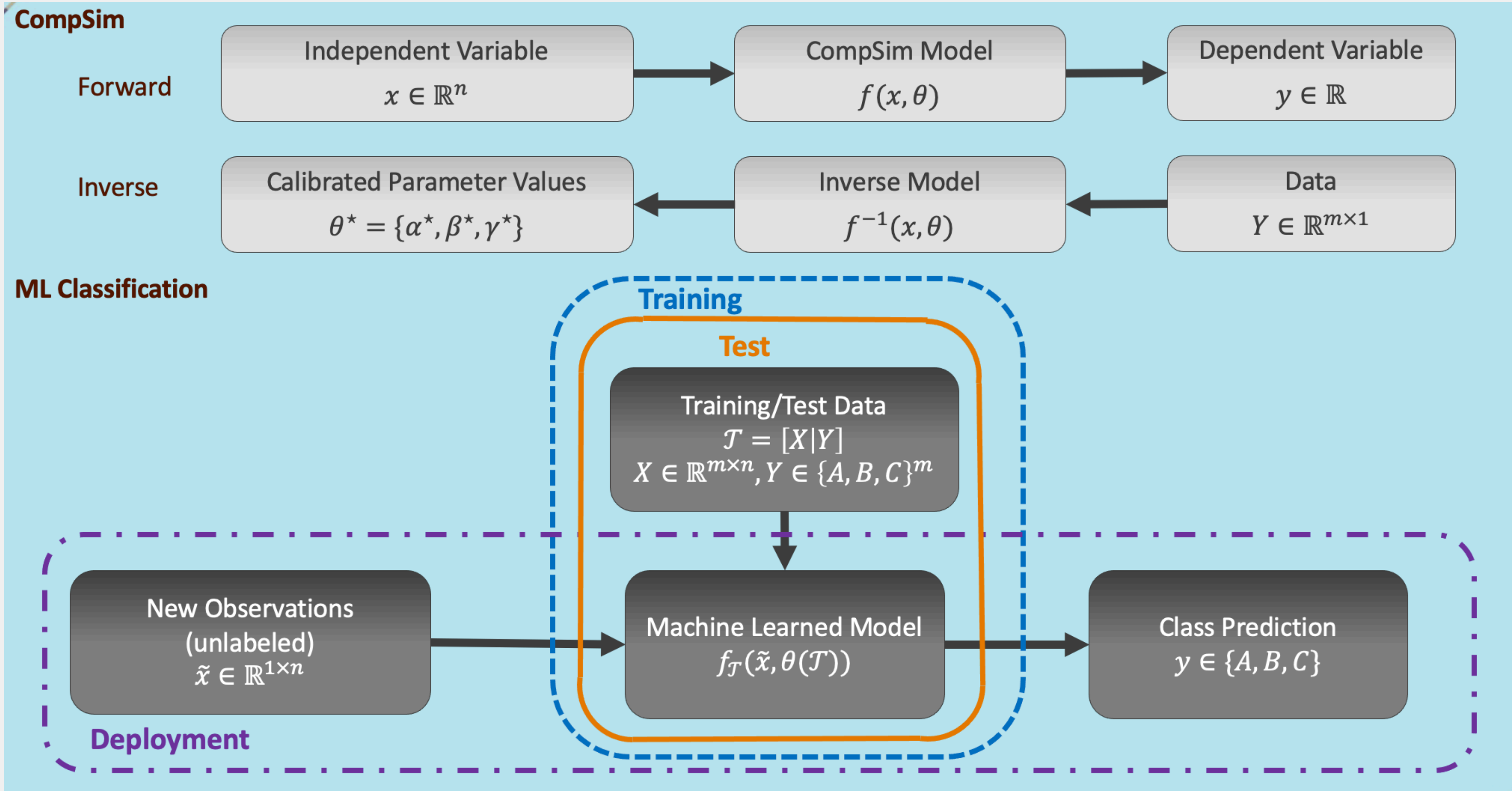
$\lambda(t) = \beta \frac{I(t)}{(S(t)+I(t)+R(t))}$: force of infection function

$\beta$: rate of Infection

$\left(\frac{1}{\gamma}\right)$: mean duration of infectiousness

Epidemiology: Classic compartmental model

# CompSim vs. Machine Learning



# Role of CompSim at Sandia

## CompSim is used in various high-consequence mission spaces at Sandia

EXAMPLE

- March 2020: WHO declares COVID-19 a pandemic - CDC website
- During early stages of an outbreak:
  - Bayesian methods provide insight given limited data
  - CompSim models were used for project modeling to inform decision makers on what may happen given a particular policy change
- The Department of Energy (DOE) stood up the National Virtual Biotechnology Laboratory that pulled together experts across all 17 DOE labs to provide critical insight during a national crisis

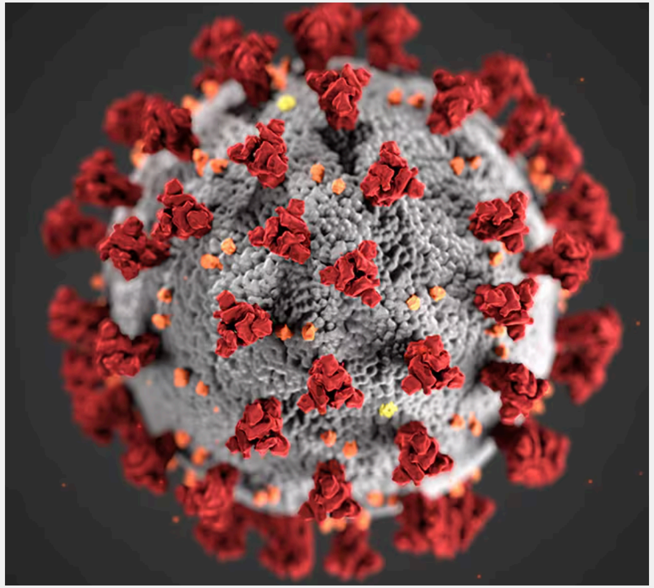


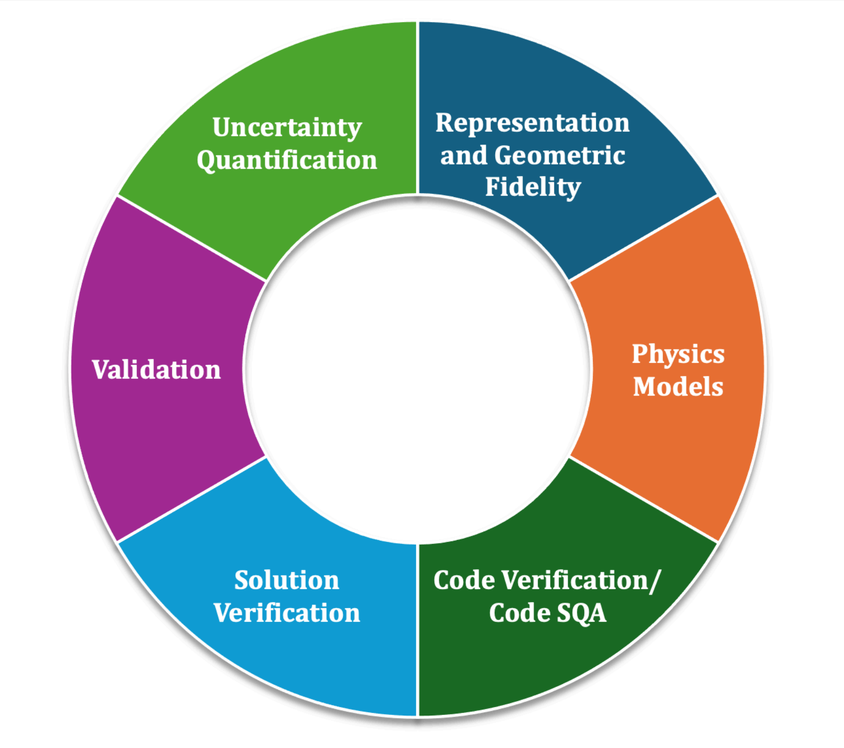
Image source: CDC website

# CompSim Credibility

The CompSim **credibility process** (1) assembles and documents **evidence** (2) to ascertain and communicate the **believability of predictions** produced from computational simulations.

## Predictive Capability Maturity Model (PCMM)

- Introduced in 2007 as "a model that can be used to assess the level of maturity of computational modeling and simulation"
- Addresses six elements that contribute to CompSim



# PCMM Table

Low Consequence ⟷ High Consequence

Six CompSim Elements

| MATURITY / ELEMENT | Maturity Level 0<br>Low Consequence, Minimal M&S Impact, e.g. Scoping Studies | Maturity Level 1<br>Moderate Consequence, Some M&S Impact, e.g. Design Support | Maturity Level 2<br>High-Consequence, High M&S Impact, e.g. Qualification Support | Maturity Level 3<br>High-Consequence, Decision-Making Based on M&S, e.g. Qualification or Certification |
|---|---|---|---|---|
| **Representation and Geometric Fidelity**<br>What features are neglected because of simplifications or stylizations? | • Judgment only<br>• Little or no representational or geometric fidelity for the system and BCs | • Significant simplification or stylization of the system and BCs<br>• Geometry or representation of major components is defined | • Limited simplification or stylization of major components and BCs<br>• Geometry or representation is well defined for major components and some minor components<br>• Some peer review conducted | • Essentially no simplification or stylization of components in the system and BCs<br>• Geometry or representation of all components is at the detail of "as built", e.g., gaps, material interfaces, fasteners<br>• Independent peer review conducted |
| **Physics and Material Model Fidelity**<br>How fundamental are the physics and material models and what is the level of model calibration? | • Judgment only<br>• Model forms are either unknown or fully empirical<br>• Few, if any, physics-informed models<br>• No coupling of models | • Some models are physics based and are calibrated using data from related systems<br>• Minimal or ad hoc coupling of models | • Physics-based models for all important processes<br>• Significant calibration needed using separate effects tests (SETs) and integral effects tests (IETs)<br>• One-way coupling of models<br>• Some peer review conducted | • All models are physics based<br>• Minimal need for calibration using SETs and IETs<br>• Sound physical basis for extrapolation and coupling of models<br>• Full, two-way coupling of models<br>• Independent peer review conducted |
| **Code Verification**<br>Are algorithm deficiencies, software errors, and poor SQE practices corrupting the simulation results? | • Judgment only<br>• Minimal testing of any software elements<br>• Little or no SQE procedures specified or followed | • Code is managed by SQE procedures<br>• Unit and regression testing conducted<br>• Some comparisons made with benchmarks | • Some algorithms are tested to determine the observed order of numerical convergence<br>• Some features & capabilities (F&C) are tested with benchmark solutions<br>• Some peer review conducted | • All important algorithms are tested to determine the observed order of numerical convergence<br>• All important F&Cs are tested with rigorous benchmark solutions<br>• Independent peer review conducted |
| **Solution Verification**<br>Are numerical solution errors and human procedural errors corrupting the simulation results? | • Judgment only<br>• Numerical errors have an unknown or large effect on simulation results | • Numerical effects on relevant SRQs are qualitatively estimated<br>• Input/output (I/O) verified only by the analysts | • Numerical effects are quantitatively estimated to be small on some SRQs<br>• I/O independently verified<br>• Some peer review conducted | • Numerical effects are determined to be small on all important SRQs<br>• Important simulations are independently reproduced<br>• Independent peer review conducted |
| **Model Validation**<br>How carefully is the accuracy of the simulation and experimental results assessed at various tiers in a validation hierarchy? | • Judgment only<br>• Few, if any, comparisons with measurements from similar systems or applications | • Quantitative assessment of accuracy of SRQs not directly relevant to the application of interest<br>• Large or unknown experimental uncertainties | • Quantitative assessment of predictive accuracy for some key SRQs from IETs and SETs<br>• Experimental uncertainties are well characterized for most SETs, but poorly known for IETs<br>• Some peer review conducted | • Quantitative assessment of predictive accuracy for all important SRQs from IETs and SETs at conditions/geometries directly relevant to the application<br>• Experimental uncertainties are well characterized for all IETs and SETs<br>• Independent peer review conducted |
| **Uncertainty Quantification and Sensitivity Analysis**<br>How thoroughly are uncertainties and sensitivities characterized and propagated? | • Judgment only<br>• Only deterministic analyses are conducted<br>• Uncertainties and sensitivities are not addressed | • Aleatory and epistemic (A&E) uncertainties propagated, but without distinction<br>• Informal sensitivity studies conducted<br>• Many strong UQ/SA assumptions made | • A&E uncertainties segregated, propagated and identified in SRQs<br>• Quantitative sensitivity analyses conducted for most parameters<br>• Numerical propagation errors are estimated and their effect known<br>• Some strong assumptions made<br>• Some peer review conducted | • A&E uncertainties comprehensively treated and properly interpreted<br>• Comprehensive sensitivity analyses conducted for parameters and models<br>• Numerical propagation errors are demonstrated to be small<br>• No significant UQ/SA assumptions made<br>• Independent peer review conducted |

## PCMM asks…

- Have you done something that meets *this requirement*?
- NOT: Have you implemented *this specific method* for in order to meet *this requirement*?

## Maturity levels are determined by…

- Consequence level of an application
- Degree that a model is the only source of information to base a decision on

# What is Scientific Machine Learning (SciML)?

## Intersection of scientific computing and machine learning

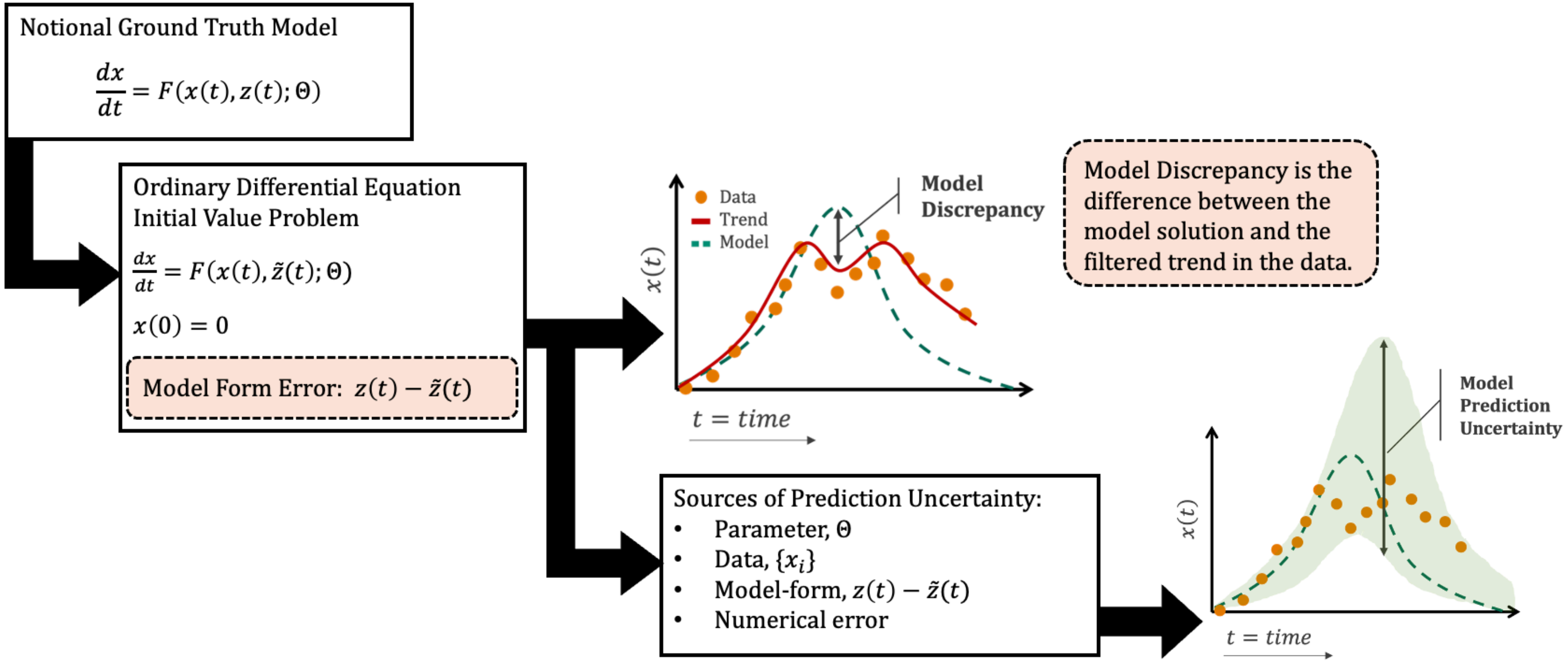
Leverages machine learning algorithms and tools used in lieu of, complementary to, or as surrogates for science and engineering computational simulation models

| **Operator Learning**<br>Physics-Informed Neural Networks (PINN) | **ML System Identification**<br>Neural Ordinary Differential Equations (NODE) | **Model-Form Error Corrections**<br>Universal Differential Equations (UDE) |
|---|---|---|
| Data-driven solutions to Partial Differential Equations (PDEs):<br>$u_t + \mathcal{R}[u] = 0,$<br>$u(x,t) = NN(t;\ W,b)$ | Simulating unknown dynamics for a full system of ODEs:<br>$\frac{du}{dt} = NN(u(t); W, b)$ | Model-form error:<br>$\frac{du}{dt} = \mathcal{F}(u(t); NN(u(t); W, b))$ |

Examples of SciML

# Role of SciML at Sandia

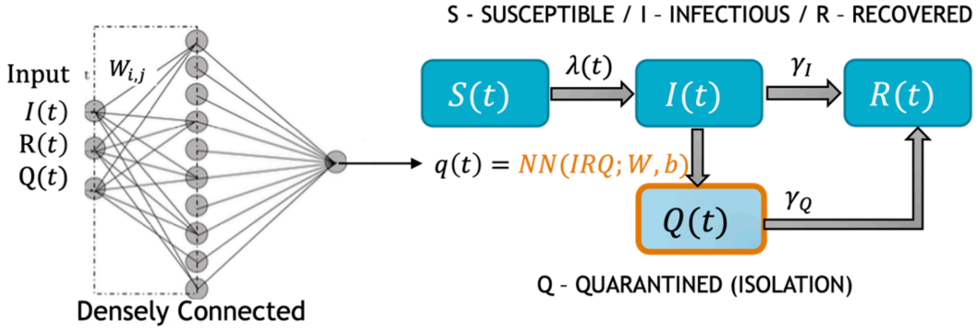
## Model Form Error and Model Discrepancy



# Role of SciML at Sandia

## Model Form Error Corrections via Neural Networks

### Universal Differential Equations for Epidemiology Compartmental Models




$$\frac{dS}{dt} = -\lambda(t)S(t)$$

$$\frac{dI}{dt} = \lambda(t)S(t) - \gamma_I I(t) - q(t)I(t)$$

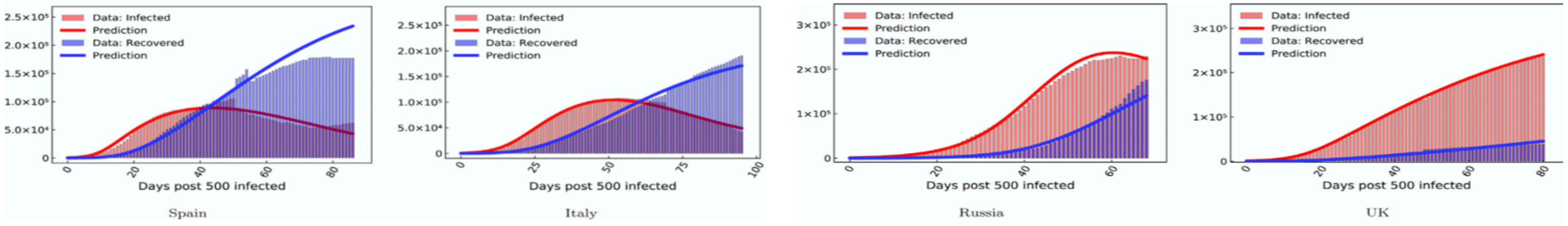
$$\frac{dR}{dt} = \gamma_I I(t) + \gamma_Q Q(t)$$

$$\frac{dQ}{dt} = q(t)I(t) - \gamma_Q Q(t)$$

Such that:

$\lambda(t) = \beta \frac{I(t)}{N}$, where $N$ is a fixed population size.

Loss function: $L_{NN}(\theta_{NN}, \beta, \gamma_I, \gamma_Q) = \left\|\log(I(t)) - \log(I_{data}(t))\right\|^2 + \left\|\log(R(t)) - \log(R_{data}(t))\right\|^2$



# Adapting PCMM for SciML

Our Objective Adapt the PCMM table to provide a tool for establishing credibility of a SciML model

**Low Consequence** ⟷ **High Consequence**

| MATURITY / ELEMENT | Maturity Level 0 | Maturity Level 1 | Maturity Level 2 | Maturity Level 3 |
|---|---|---|---|---|
| Representation and Geometric Fidelity | | | | |
| Physics and Material Model Fidelity | | | | |
| Code Verification | | | | |
| Solution Verification | | | | |
| Model Validation | | | | |
| Uncertainty Quantification and Sensitivity Analysis | | | | |

# Adapting PCMM for SciML

**Our Objective** Adapt the PCMM table to provide a tool for establishing credibility of a SciML model

**Low Consequence** ⟷ **High Consequence**

Data Representation

Domain Aware

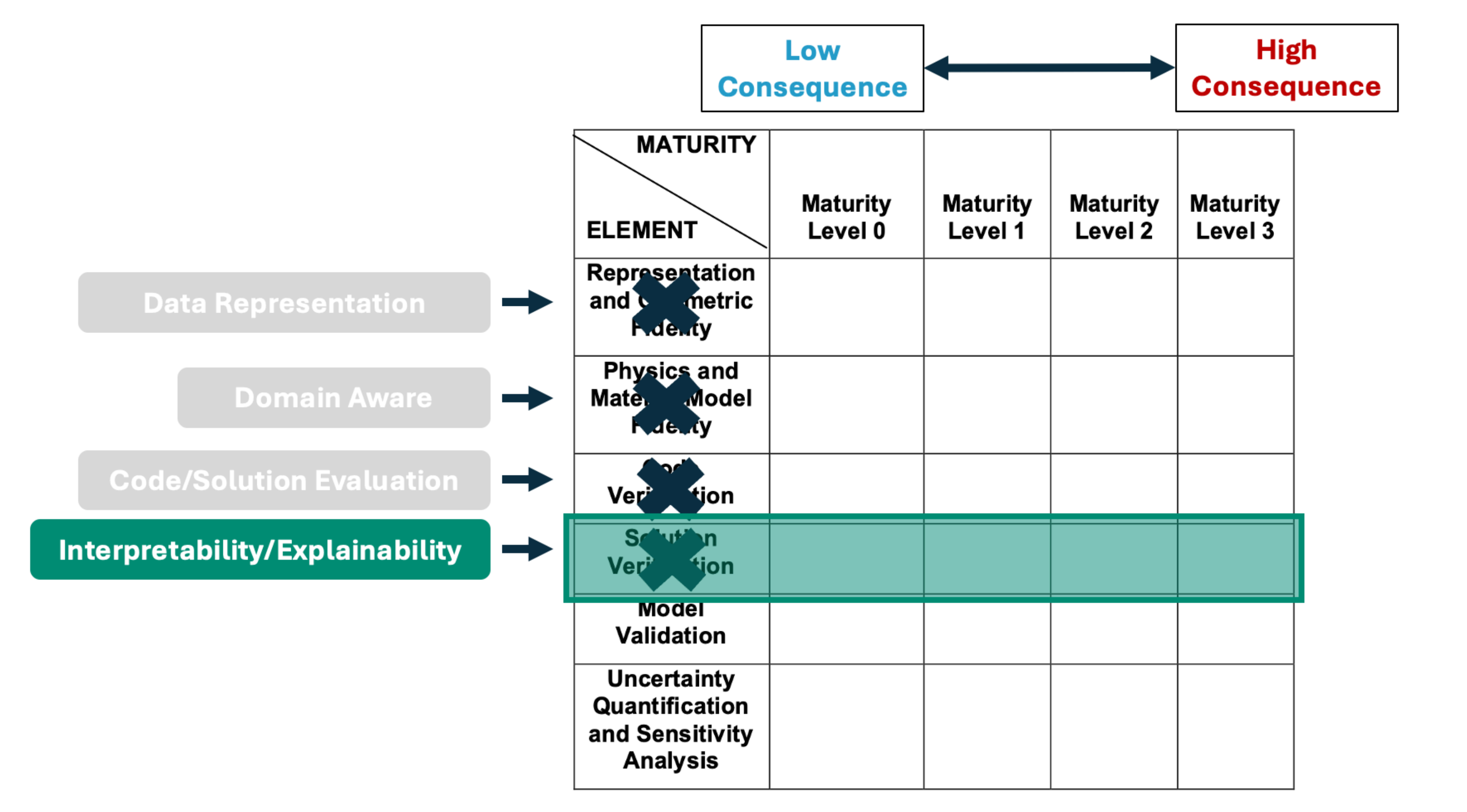
Code/Solution Evaluation

Interpretability/Explainability

| MATURITY / ELEMENT | Maturity Level 0 | Maturity Level 1 | Maturity Level 2 | Maturity Level 3 |
|---|---|---|---|---|
| ~~Representation and Geometric Fidelity~~ | | | | |
| ~~Physics and Material Model Fidelity~~ | | | | |
| ~~Code Verification~~ | | | | |
| ~~Solution Verification~~ | | | | |
| Model Validation | | | | |
| Uncertainty Quantification and Sensitivity Analysis | | | | |

# Adapting PCMM for SciML

**Our Objective** Adapt the PCMM table to provide a tool for establishing credibility of a SciML model

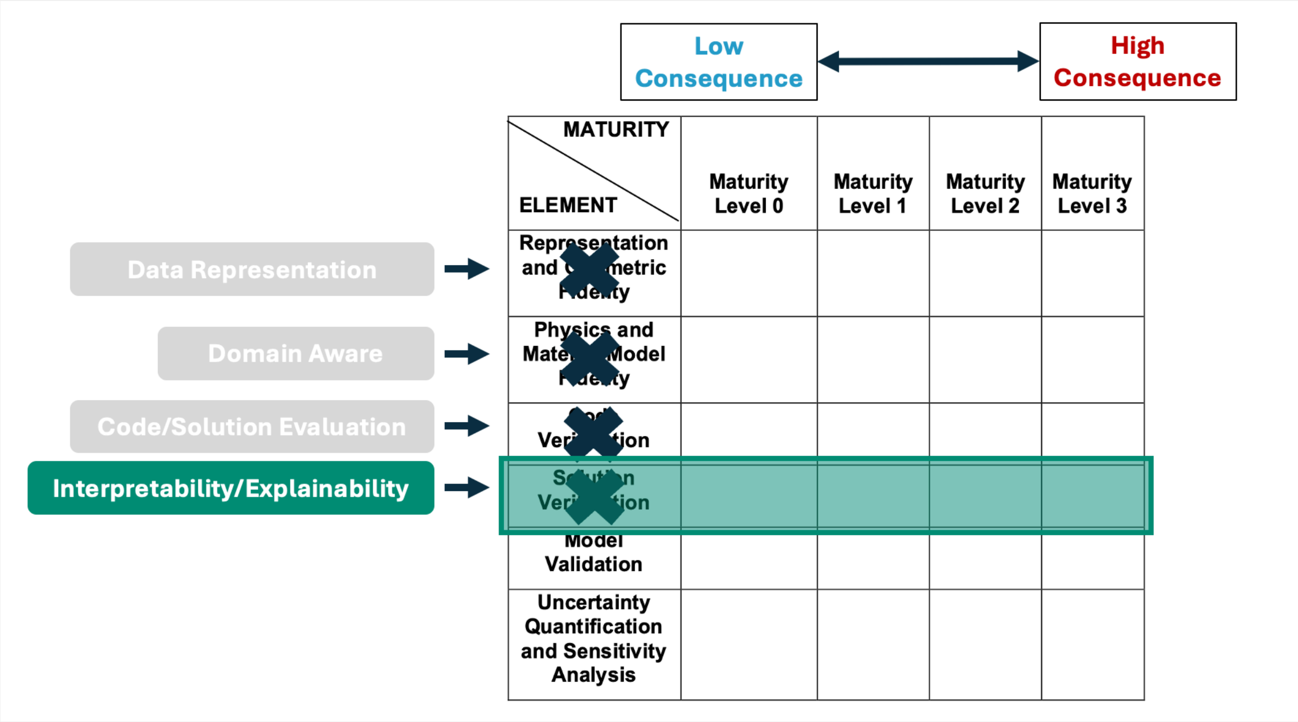


# Proposed Framework

Maturity Levels for Explainability/Interpretability with SciML

# Our Objective

**Big Picture** Adapt the PCMM table to provide a tool for establishing credibility of a SciML model




## Specific to Interpretability/Explainability

- ML community has prioritized explainability to develop trust in ML

> The maturity of these methods need to also be evaluated

- We aim to develop criteria needed to establish maturity levels for interpretability associated with or explainability applied to a SciML model

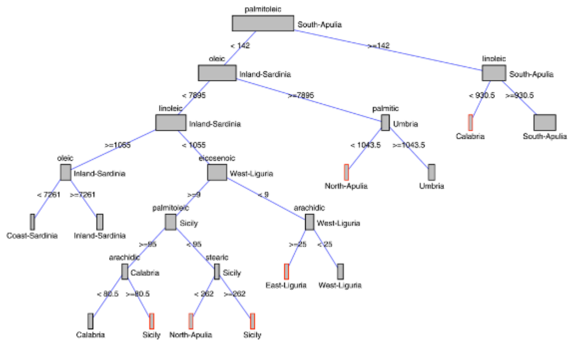
# Explainability/Interpretability

How we are making a distinction between these terms…

## Interpretability

Ability to directly use model to understand how algorithm makes decisions

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

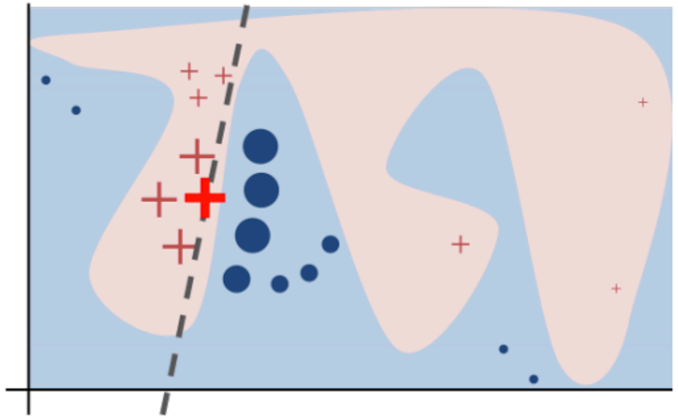


Using interpretable model or adjusting black box models to contain interpretable parameters

Figure from Urbanek (2008)

## Explainability

Ability to indirectly use model to understand how algorithm makes decisions



Often post-hoc techniques

Figure from LIME paper (Ribeiro 2016)

# Proposed Maturity Levels (current state)

| MATURITY<br>Model Impact<br>Consequence Level<br><br>Elements: | LEVEL 0<br>Minimal<br>Low | LEVEL 1<br>Some<br>Moderate | LEVEL 2<br>High<br>High | LEVEL 3<br>Decision-Making<br>High |
|---|---|---|---|---|
| Interpretable or black-box model | If using a non-interpretable model, not required to answer the question of why a more complex model is better? | If using a non-interpretable model, must partially answer the question of why a more complex model is better? | If using a non-interpretable model, must answer the question of why a more complex model is better? | If using a non-interpretable model, must rigorously answer the question of why a more complex model is better? |
| Interpretations / Explanations | No interpretations / explainability applied | Some interpretations / explainability applied (local and/or global) | Interpretations / explainability applied and assessed (local and global) | Interpretations / explainability comprehensively applied and assessed (local and global) |
| Level of review | Judgment only | Some informal internal peer review conducted (within team or informally outside of team within institution) | Formal internal independent peer review conducted (internal to institution; outside of team) | External independent peer review conducted (external to institution; outside of team) |
| Assumptions | Relying on assumptions that model is capturing/using scientifically reasonable relationships in the data | Many strong assumptions made that model is capturing/using scientifically reasonable relationships in the data | Some assumptions made that model is capturing/using scientifically reasonable relationships in the data | No significant assumptions made that model is capturing/using scientifically reasonable relationships in the data |

# Proposed Maturity Levels: Interpretable or Black-Box

| MATURITY<br>Model Impact<br>Consequence Level | LEVEL 0<br>Minimal<br>Low | LEVEL 1<br>Some<br>Moderate | LEVEL 2<br>High<br>High | LEVEL 3<br>Decision-Making<br>High |
| --- | --- | --- | --- | --- |
| Interpretable or black-box model | If using a non-interpretable model, not required to answer the question of why a more complex model is better? | If using a non-interpretable model, must partially answer the question of why a more complex model is better? | If using a non-interpretable model, must answer the question of why a more complex model is better? | If using a non-interpretable model, must rigorously answer the question of why a more complex model is better? |

How rigorously must the following question be answered...
If using a non-interpretable model, why is a more complex model better?

Consideration Do not want to force use of "clear-box" model, but require reasoning for use of "black-box" model

# Proposed Maturity Levels: Interpretations / Explanations

| MATURITY<br>Model Impact<br>Consequence Level | LEVEL 0<br>Minimal<br>Low | LEVEL 1<br>Some<br>Moderate | LEVEL 2<br>High<br>High | LEVEL 3<br>Decision-Making<br>High |
|---|---|---|---|---|
| Interpretations / Explanations | No interpretations / explainability applied | Some interpretations / explainability applied (local and/or global) | Interpretations / explainability applied and assessed (local and global) | Interpretations / explainability comprehensively applied and assessed (local and global) |

How rigorously has the model been interpreted or explanations have been applied AND assessed?

## Considerations

- Applied global and local explanations
- Explanations are approximations of a model: Important to assess whether approximations are credible

# Proposed Maturity Levels: Level of Review

| MATURITY<br>Model Impact<br>Consequence Level | LEVEL 0<br>Minimal<br>Low | LEVEL 1<br>Some<br>Moderate | LEVEL 2<br>High<br>High | LEVEL 3<br>Decision-Making<br>High |
|---|---|---|---|---|
| Level of review | Judgment only | Some informal internal peer review conducted (within team or informally outside of team within institution) | Formal internal independent peer review conducted (internal to institution; outside of team) | External independent peer review conducted (external to institution; outside of team) |

How rigorously have model interpretations/explanations been peer-reviewed?

Considerations Heavily influenced from requirements in PCMM table

# Proposed Maturity Levels: Assumptions

| MATURITY | LEVEL 0 | LEVEL 1 | LEVEL 2 | LEVEL 3 |
|---|---|---|---|---|
| Model Impact | Minimal | Some | High | Decision-Making |
| Consequence Level | Low | Moderate | High | High |
| Assumptions | Relying on assumptions that model is capturing/using scientifically reasonable relationships in the data | Many strong assumptions made that model is capturing/using scientifically reasonable relationships in the data | Some assumptions made that model is capturing/using scientifically reasonable relationships in the data | No significant assumptions made that model is capturing/using scientifically reasonable relationships in the data |

How thoroughly have interpretations/explanations been used to diagnose how well a model is using scientifically reasonable relationships in data?

Considerations Relies on the soundness of explainability techniques used

# Discussion

Challenges and Moving Forward

# Challenges

- Grey area between "interpretability" and "explainability"

- Rapidly evolving area of machine learning and explainability

- Currently, not a major emphasis on the assessment of explanations
  - e.g., diagnostic tools for explainability methods

- How to best account for the fact that there are no agreed upon “standards” for explainability yet?

# Going Forward...

## Continuing to develop the requirements based on...

- feedback
- additional research into interpretability/explainability
    - definitions, evaluation techniques, new methods, etc.
- exemplars

## Questions to consider...

- How can lessons learned from using "statistical models" in high consequence decision spaces be used to inform how "machine learning" is used in high consequence decision spaces?
- Can this (initial) framework for SciML be applicable for more general ML? What would need to be adjusted?

# Thank you.

## Questions? Thoughts?

Katherine Goode

kjgoode@sandia.gov

goodekat.github.io