



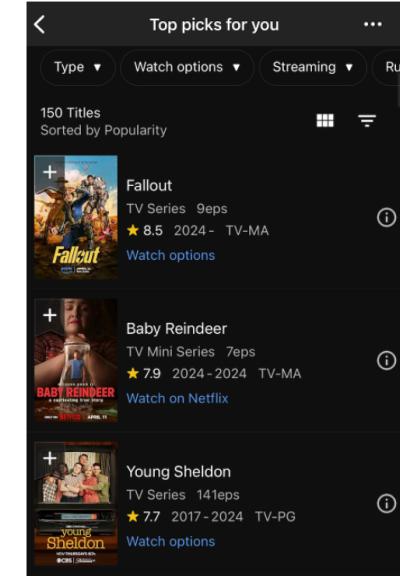
A Framework for Evaluating the Maturity Level of Machine Learning Explanations

Katherine Goode

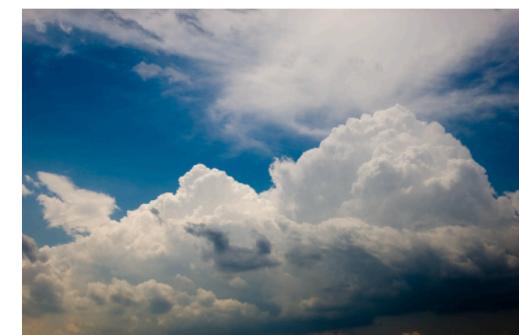
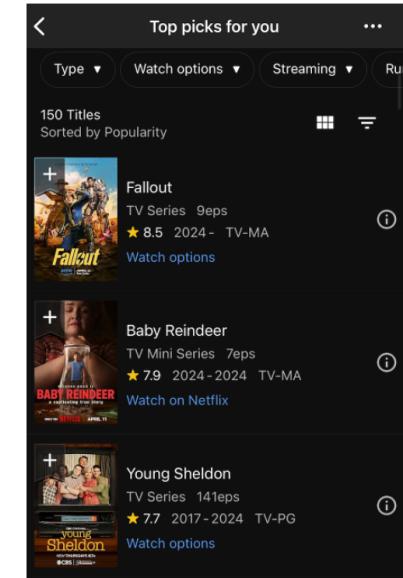
Erin Acuesta (PI), Candace Diaz, Raga Krishnakumar, Ernesto Prudencio

August 6, 2024

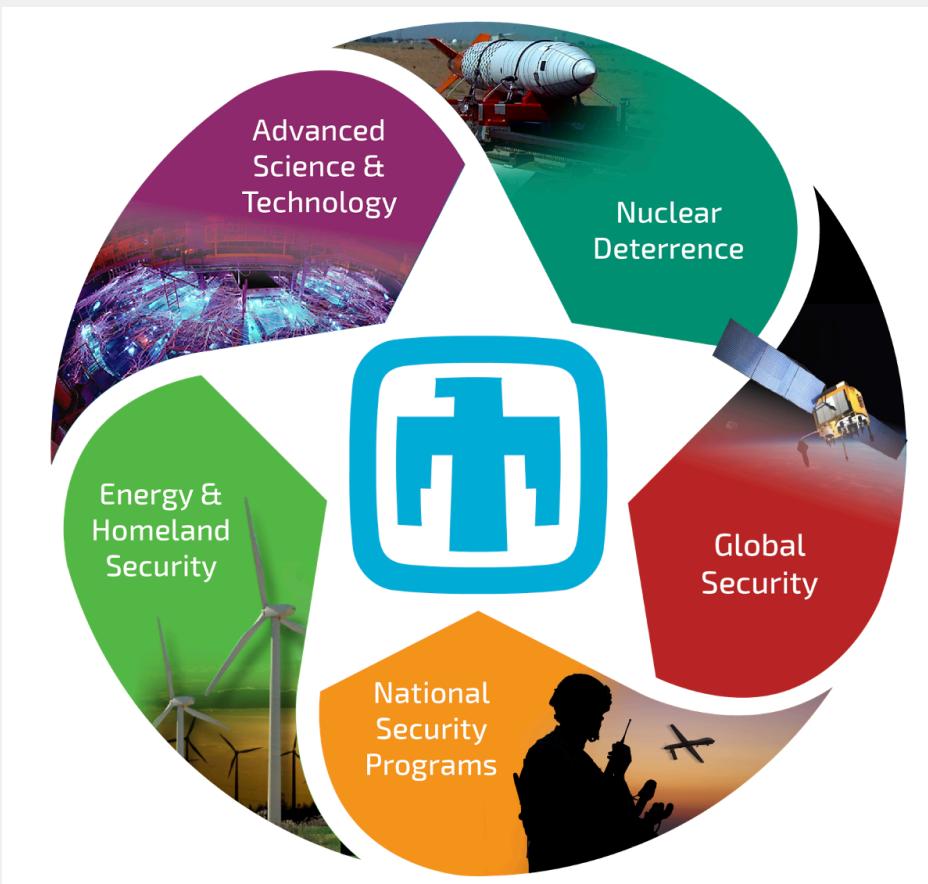
When would you be willing to use machine learning for decision making?



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

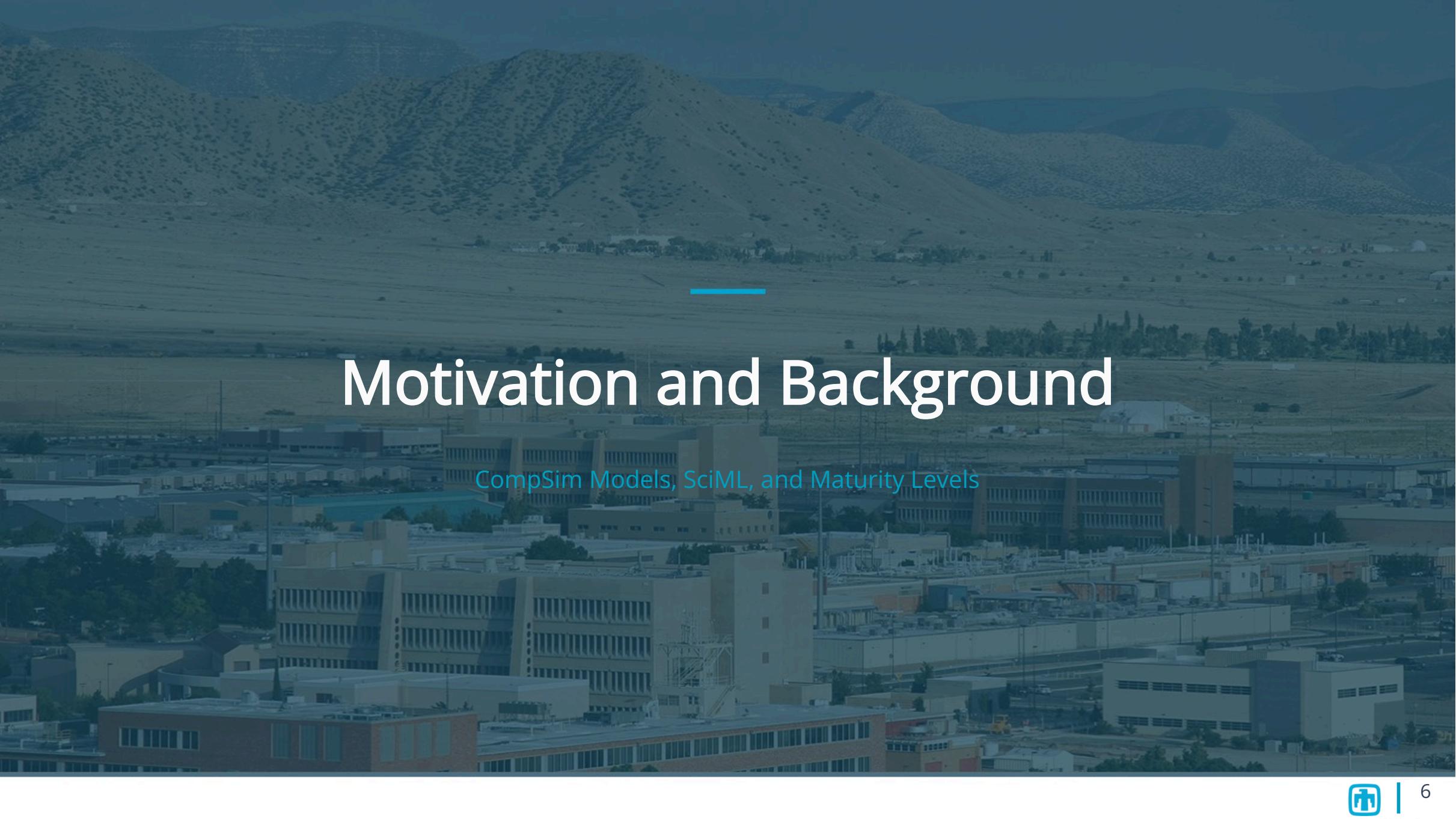
CompSim Models, SciML, and Maturity Levels

Proposed Framework

Maturity Levels for Explainability/Interpretability with SciML

Discussion

Challenges and Moving Forward



Motivation and Background

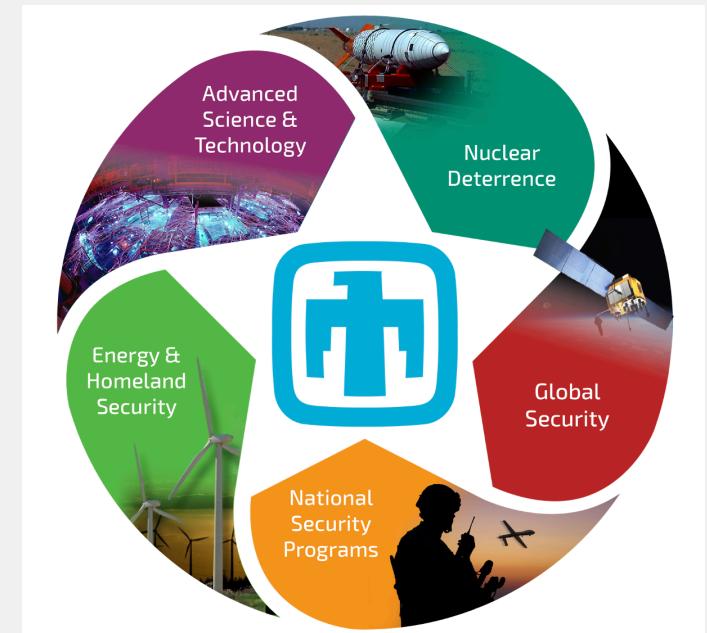
CompSim Models, SciML, and Maturity Levels

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

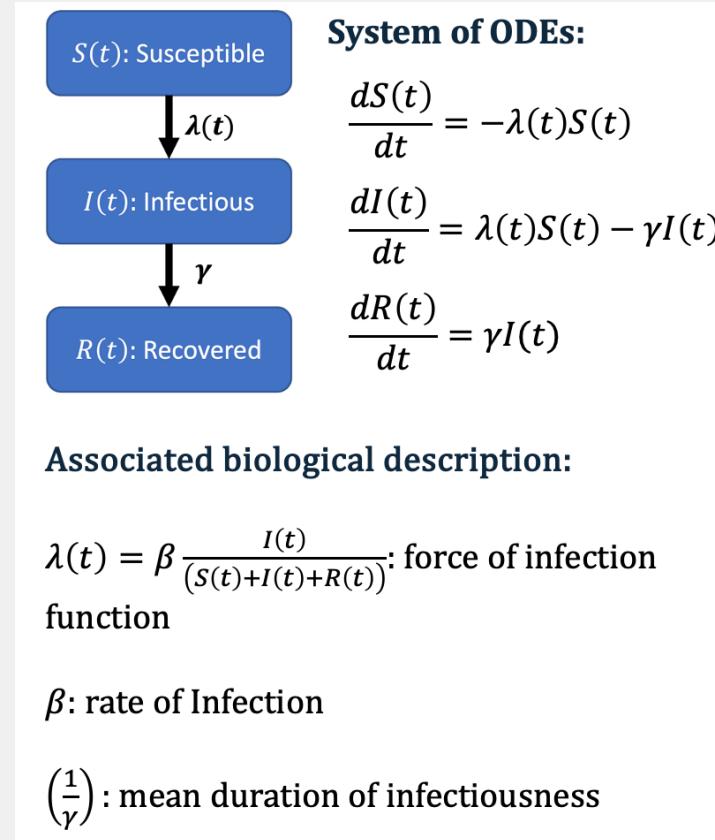
CREDIBILITY LEADS TO TRUSTWORTHY MODELS, AND TRUSTWORTHY MODELS MAY ESTABLISH TRUST

What is CompSim?

Computational Simulation

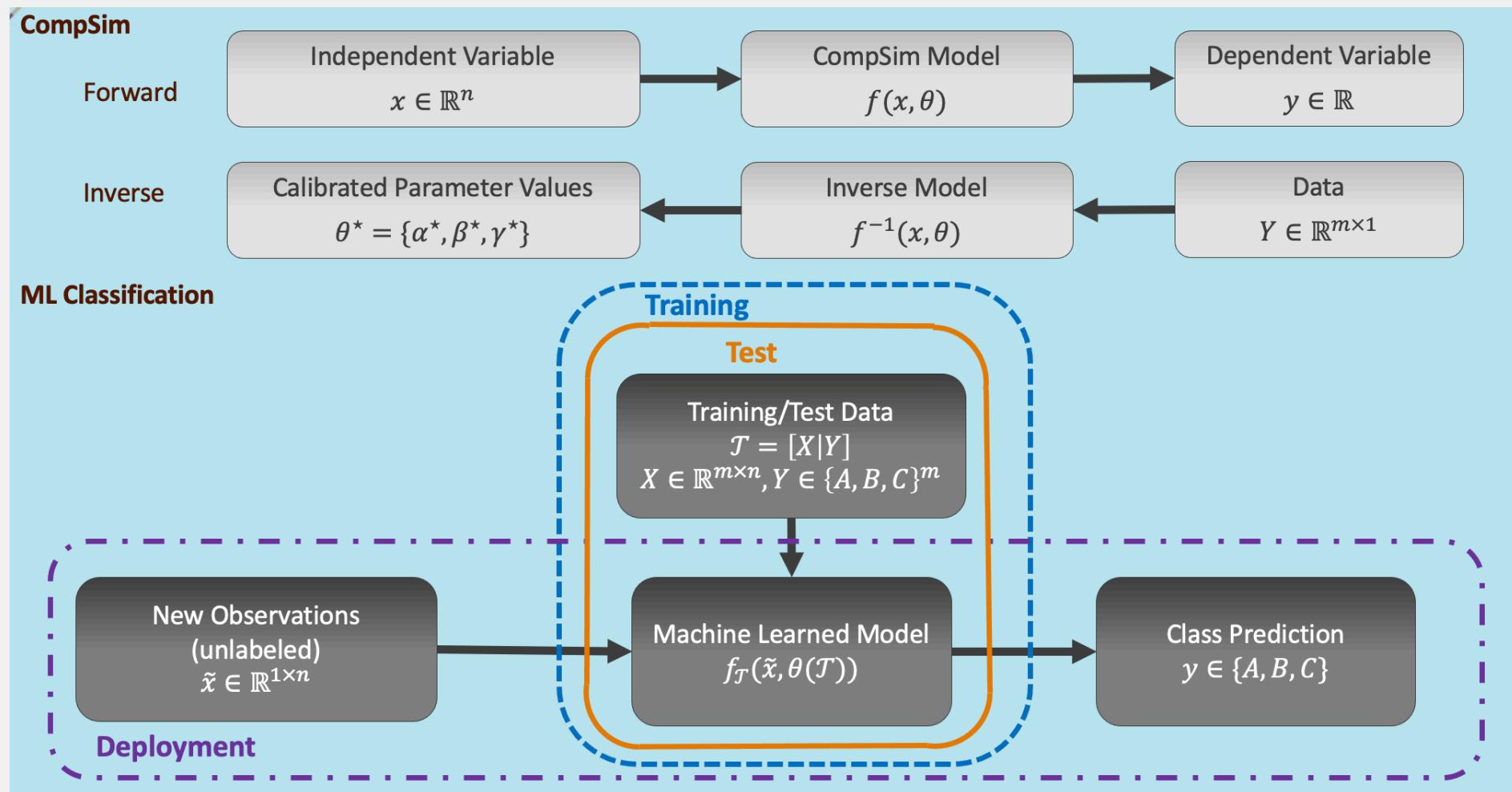
AKA CompSim; Modeling and Simulation;
ModSim; M&S

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



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 providing 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 the experts across all 17 DOE labs to provide critical insight during a national crisis.

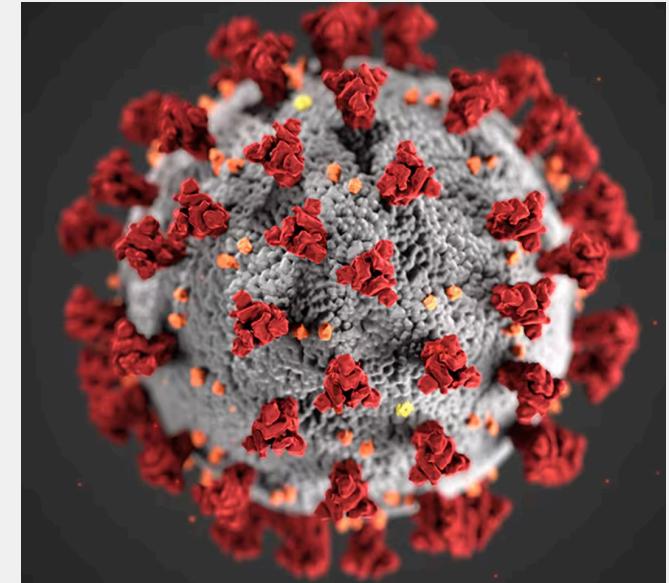


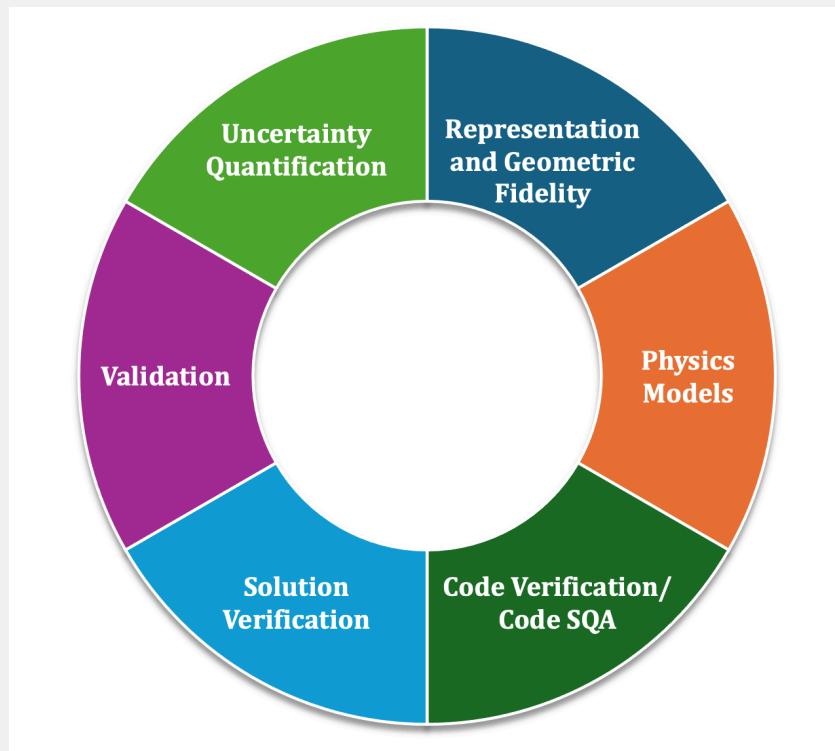
Image source: [CDC website](#)

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

Six CompSim Elements

ELEMENT	MATURITY			
	Maturity Level 0 Low Consequence, Minimal M&S Impact, e.g. Scoping Studies	Maturity Level 1 Moderate Consequence, Some M&S Impact, e.g. Design Support	Maturity Level 2 High-Consequence, High M&S Impact, e.g. Qualification Support	Maturity Level 3 High-Consequence, Decision-Making Based on M&S, e.g. Qualification or Certification
Representation and Geometric Fidelity What features are neglected because of simplifications or stylizations?	<ul style="list-style-type: none"> • Judgment only • Little or no representational or geometric fidelity for the system and BCs 	<ul style="list-style-type: none"> • Significant simplification or stylization of the system and BCs • Geometry or representation of major components is defined 	<ul style="list-style-type: none"> • Limited simplification or stylization of major components and BCs • Geometry or representation is well defined for major components and some minor components • Some peer review conducted 	<ul style="list-style-type: none"> • Essentially no simplification or stylization of components in the system and BCs • Geometry or representation of all components is at the detail of "as built", e.g., gaps, material interfaces, fasteners • Independent peer review conducted
Physics and Material Model Fidelity How fundamental are the physics and material models and what is the level of model calibration?	<ul style="list-style-type: none"> • Judgment only • Model forms are either unknown or fully empirical • Few, if any, physics-informed models • No coupling of models 	<ul style="list-style-type: none"> • Some models are physics based and are calibrated using data from related systems • Minimal or ad hoc coupling of models 	<ul style="list-style-type: none"> • Physics-based models for all important processes • Significant calibration needed using separate effects tests (SETs) and integral effects tests (IETs) • One-way coupling of models • Some peer review conducted 	<ul style="list-style-type: none"> • All models are physics based • Minimal need for calibration using SETs and IETs • Sound physical basis for extrapolation and coupling of models • Full, two-way coupling of models • Independent peer review conducted
Code Verification Are algorithm deficiencies, software errors, and poor SQE practices corrupting the simulation results?	<ul style="list-style-type: none"> • Judgment only • Minimal testing of any software elements • Little or no SOE procedures specified or followed 	<ul style="list-style-type: none"> • Code is managed by SQE procedures • Unit and regression testing conducted • Some comparisons made with benchmarks 	<ul style="list-style-type: none"> • Some algorithms are tested to determine the observed order of numerical convergence • Some features & capabilities (F&Cs) are tested with benchmark solutions • Some peer review conducted 	<ul style="list-style-type: none"> • All important algorithms are tested to determine the observed order of numerical convergence • All important F&Cs are tested with rigorous benchmark solutions • Independent peer review conducted
Solution Verification Are numerical solution errors and human procedural errors corrupting the simulation results?	<ul style="list-style-type: none"> • Judgment only • Numerical errors have an unknown or large effect on simulation results 	<ul style="list-style-type: none"> • Numerical effects on relevant SRQs are qualitatively estimated • Input/output (I/O) verified only by the analysts 	<ul style="list-style-type: none"> • Numerical effects are quantitatively estimated to be small on some SRQs • I/O independently verified • Some peer review conducted 	<ul style="list-style-type: none"> • Numerical effects are determined to be small on all important SRQs • Important simulations are independently reproduced • Independent peer review conducted
Model Validation How carefully is the accuracy of the simulation and experimental results assessed at various tiers in a validation hierarchy?	<ul style="list-style-type: none"> • Judgment only • Few, if any, comparisons with measurements from similar systems or applications 	<ul style="list-style-type: none"> • Quantitative assessment of accuracy of SRQs not directly relevant to the application of interest • Large or unknown experimental uncertainties 	<ul style="list-style-type: none"> • Quantitative assessment of predictive accuracy for some key SRQs from IETs and SETs • Experimental uncertainties are well characterized for most SETs, but poorly known for IETs • Some peer review conducted 	<ul style="list-style-type: none"> • Quantitative assessment of predictive accuracy for all important SRQs from IETs and SETs at conditions/geometries directly relevant to the application • Experimental uncertainties are well characterized for all IETs and SETs • Independent peer review conducted
Uncertainty Quantification and Sensitivity Analysis How thoroughly are uncertainties and sensitivities characterized and propagated?	<ul style="list-style-type: none"> • Judgment only • Only deterministic analyses are conducted • Uncertainties and sensitivities are not addressed 	<ul style="list-style-type: none"> • Aleatory and epistemic (A&E) uncertainties propagated, but without distinction • Informal sensitivity studies conducted • Many strong UQ/SA assumptions made 	<ul style="list-style-type: none"> • A&E uncertainties segregated, propagated and identified in SRQs • Quantitative sensitivity analyses conducted for most parameters • Numerical propagation errors are estimated and their effect known • Some strong assumptions made • Some peer review conducted 	<ul style="list-style-type: none"> • A&E uncertainties comprehensively treated and properly interpreted • Comprehensive sensitivity analyses conducted for parameters and models • Numerical propagation errors are demonstrated to be small • No significant UQ/SA assumptions made • 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

Physics-Informed Neural Networks (PINN)

Data-driven solutions to Partial Differential Equations (PDEs):

$$u_t + \mathcal{R}[u] = 0, \\ u(x, t) = \text{NN}(t; W, b)$$

ML System Identification

Neural Ordinary Differential Equations (NODE)

Simulating unknown dynamics for a full system of ODEs:

$$\frac{du}{dt} = \text{NN}(u(t); W, b)$$

Model-Form Error Corrections

Universal Differential Equations (UDE)

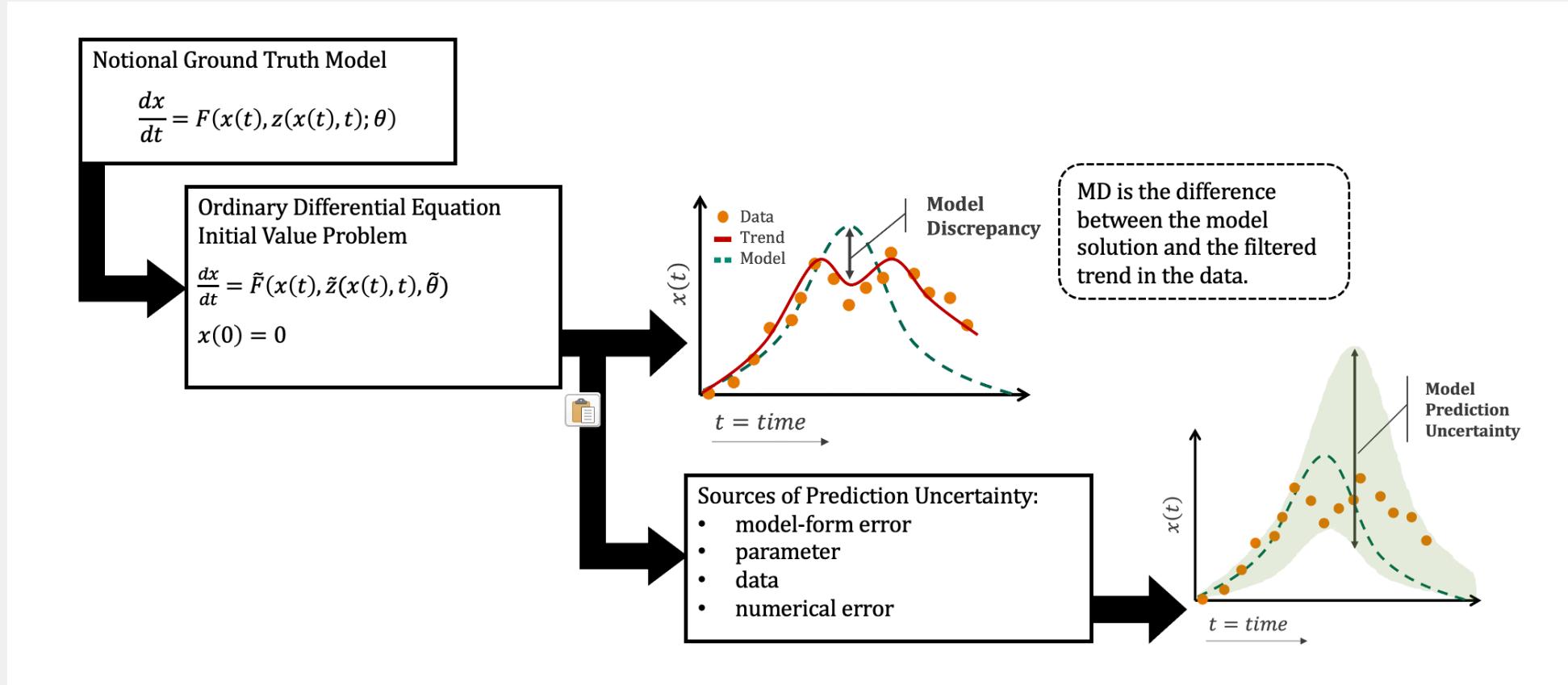
Model-form error:

$$\frac{du}{dt} = \mathcal{F}(u(t); \text{NN}(u(t); W, b))$$

Some 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

The diagram illustrates a compartmental model for epidemiology. It consists of four compartments: S (Susceptible), I (Infectious), R (Recovered), and Q (Quarantined). The flow is S → I → R. The rate of infection is $\lambda(t)$, and the rate of recovery is γ_I . The Quarantined compartment Q receives individuals from the Infectious compartment I at rate $q(t)$ and from the Recovered compartment R at rate γ_Q . The neural network correction $q(t) = \text{NN}(IRQ; W, b)$ takes inputs $I(t)$, $R(t)$, and $Q(t)$ and provides a correction to the flow from I to Q. The neural network is shown as a dense layer of nodes with weights $W_{i,j}$.

$$\frac{dS}{dt} = -\lambda(t)S(t)$$
$$\frac{dI}{dt} = \lambda(t)S(t) - \gamma_I I(t) - \underline{q(t)I(t)}$$
$$\frac{dR}{dt} = \gamma_I I(t) + \gamma_Q Q(t)$$
$$\frac{dQ}{dt} = \underline{q(t)I(t)} - \gamma_Q Q(t)$$

Such that:

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

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

Four line plots show the progression of the epidemic for Spain, Italy, Russia, and the UK. Each plot compares 'Data: Infected' (red line with shaded error band), 'Data: Recovered' (blue line with shaded error band), and 'Prediction' (blue line). The x-axis is 'Days post 500 infected' (0 to 60), and the y-axis is the number of individuals (log scale, 0 to 3×10^5). The plots show the characteristic SIR curve, with the prediction closely matching the data for all four countries.

Adapting PCMM for SciML

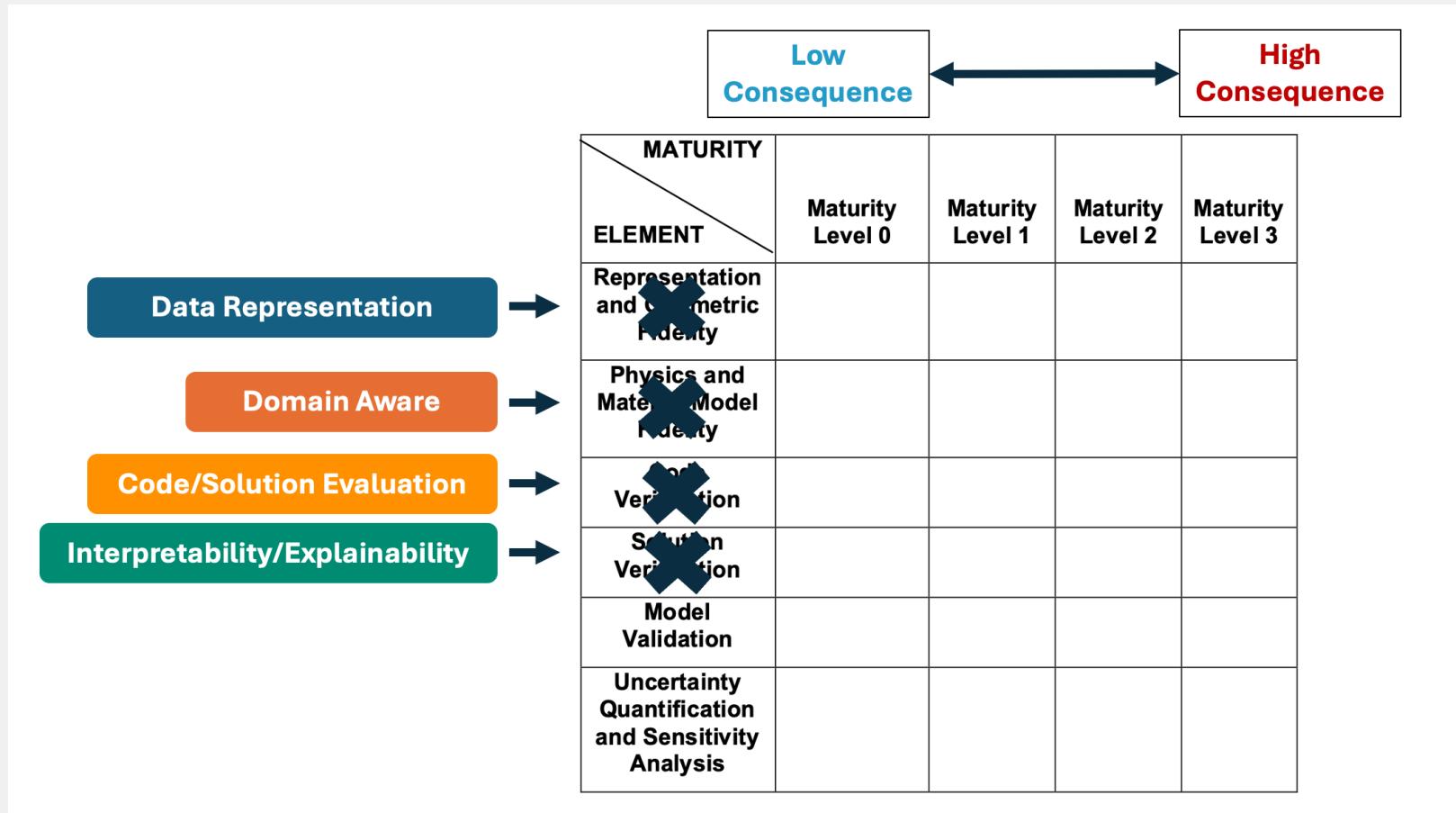
Our Objective Adapt the PCMM table to provide a tool for establishing credibility of a SciML model



		Low Consequence	High Consequence		
ELEMENT	MATURITY	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					

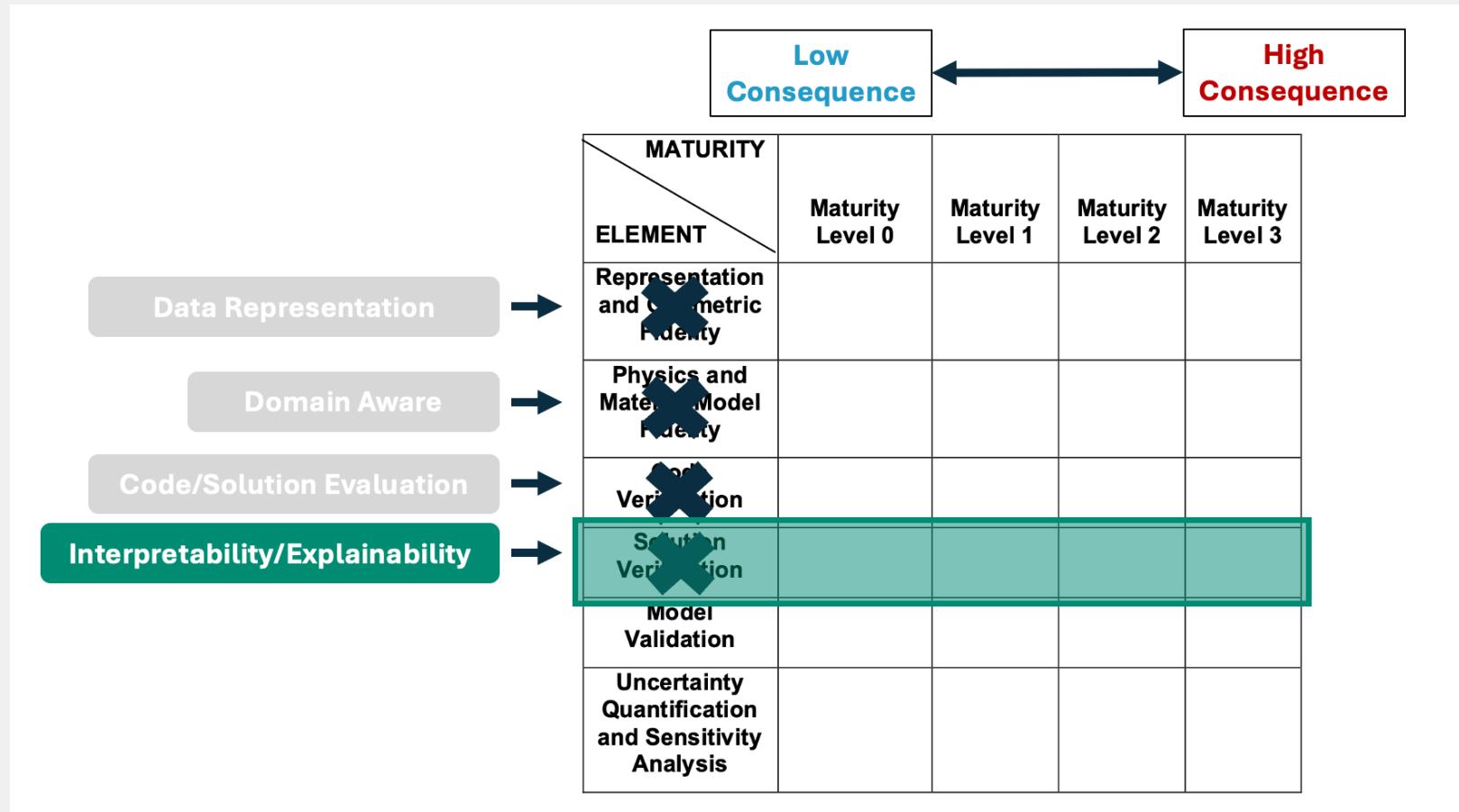
Adapting PCMM for SciML

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



Adapting PCMM for SciML

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



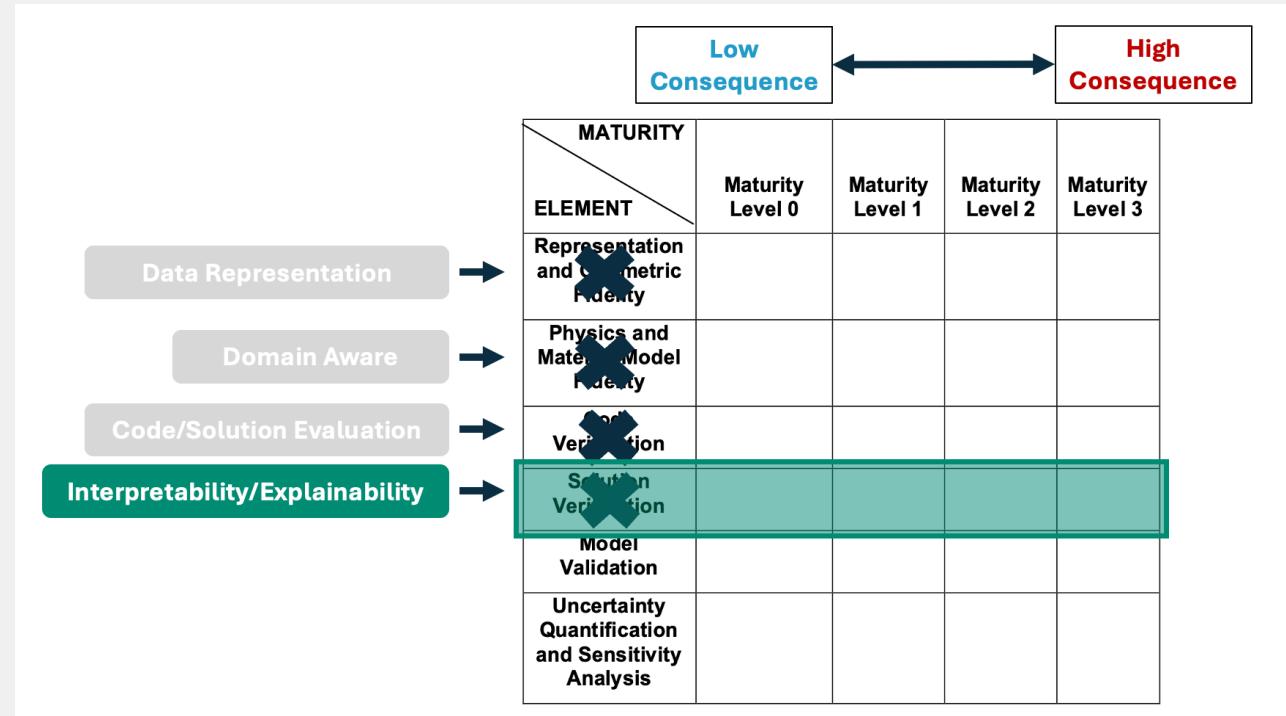
Proposed Framework



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.
- Develop criteria needed to establish maturity level of interpretability/explainability associated with/applied to a SciML model

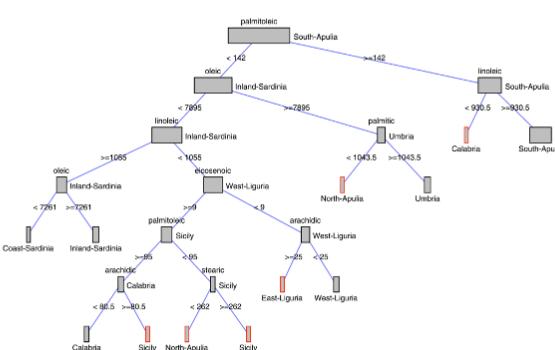
Explainability/Interpretability

How we are making a distinction between these terms...

Interpretability

Ability to directly use model to understand how algorithm makes decisions

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

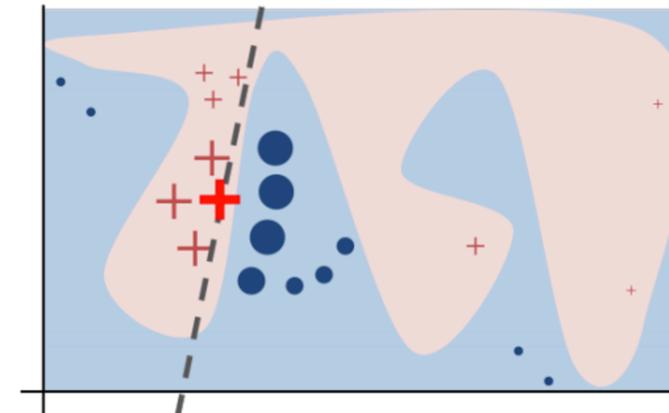


Using interpretable model or adjusting black box models to contain interpretable parameters

Figure from [Urbanek \(2008\)](#)

Explainability

Ability to indirectly use model to understand how algorithm makes decisions



Often post-hoc techniques

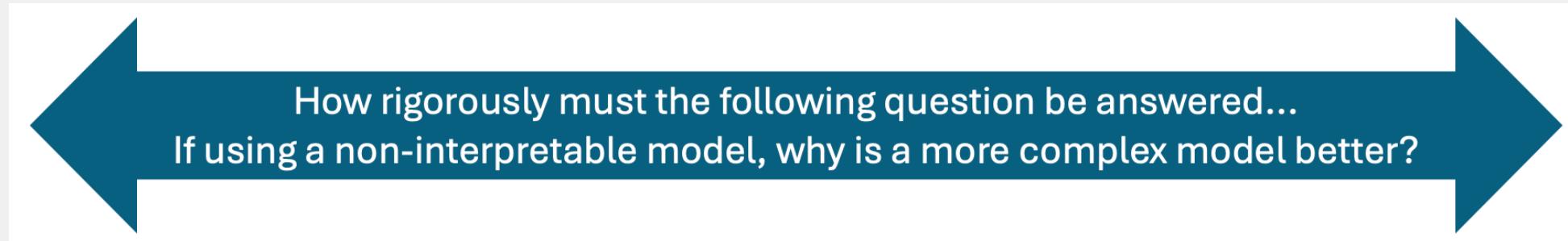
Figure from LIME paper ([Ribeiro 2016](#))

Proposed Maturity Levels (current state)

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

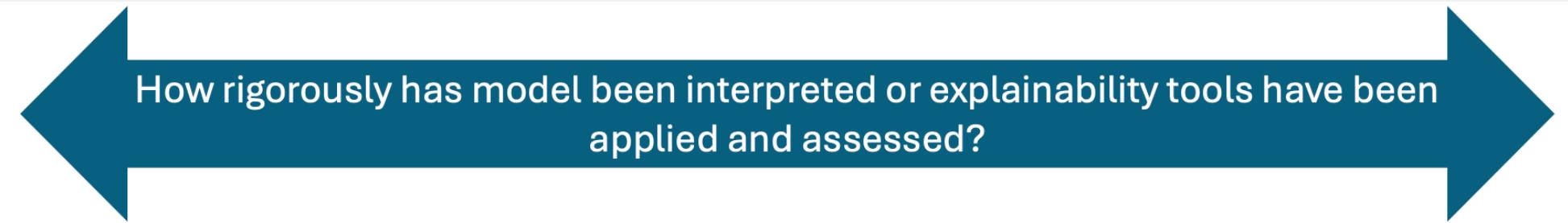
MATURITY Model Impact Consequence Level	LEVEL 0 Minimal Low	LEVEL 1 Some Moderate	LEVEL 2 High High	LEVEL 3 Decision-Making High
Interpretable or black-box model	If using a non-interpretable model, not required to answer the question of why is a more complex model better?	If using a non-interpretable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

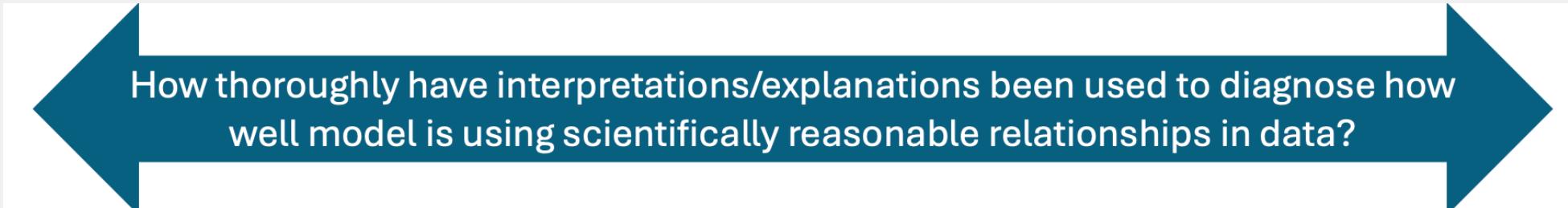


How rigorously have model interpretations or applications of explainability been peer-reviewed?

Considerations Heavily influenced from requirements in PCMM table

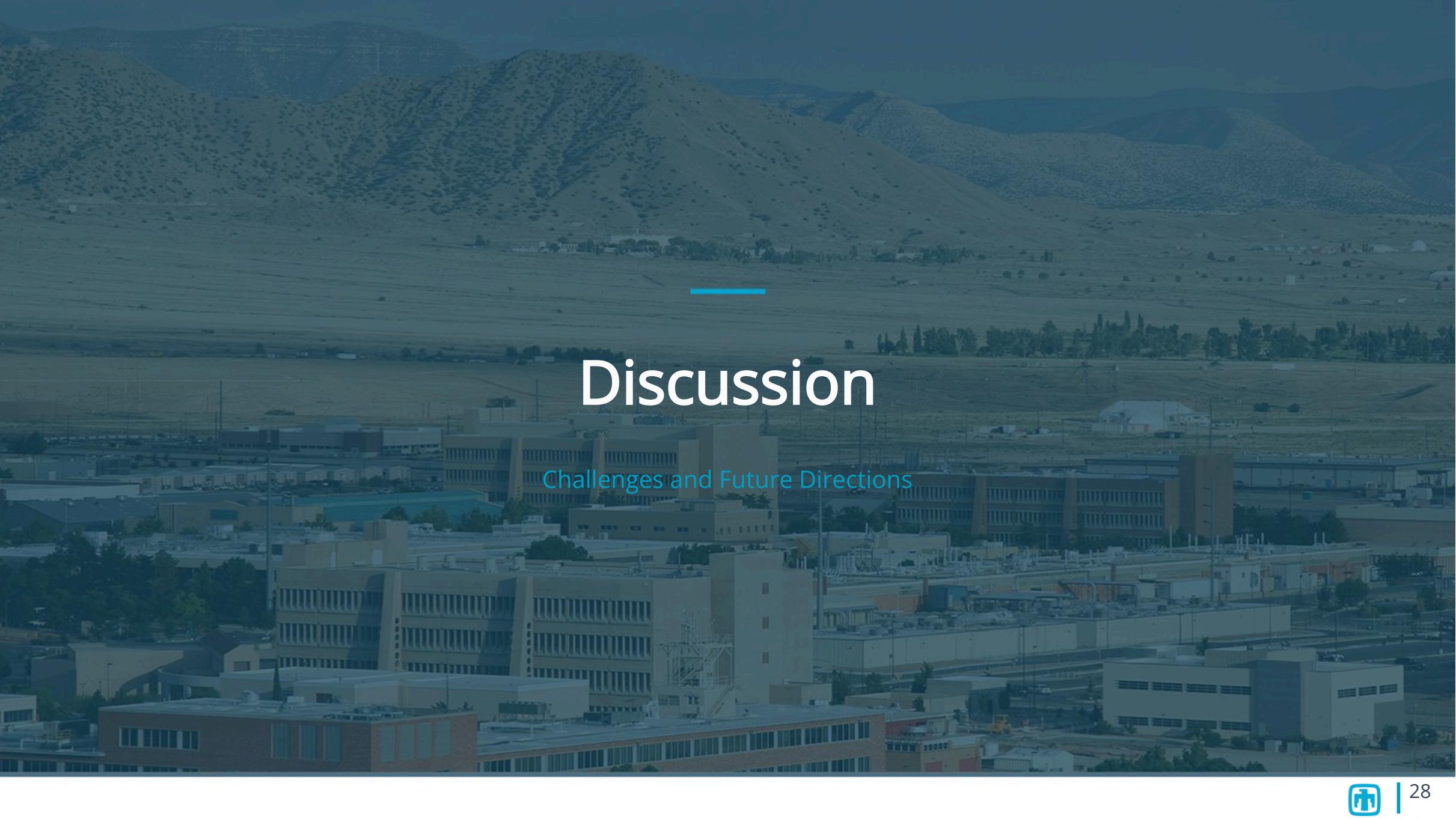
Proposed Maturity Levels: Assumptions

MATURITY Model Impact Consequence Level	LEVEL 0 Minimal Low	LEVEL 1 Some Moderate	LEVEL 2 High High	LEVEL 3 Decision-Making 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 model is using scientifically reasonable relationships in data?

Considerations Relies on the soundness of explainability techniques used

A wide-angle, aerial photograph of a large industrial complex, likely a nuclear facility, situated in a valley. The foreground is filled with various industrial buildings, some with multiple stories and light-colored facades. In the middle ground, there are more buildings and what appears to be a cooling tower or tall industrial structure. The background features a range of mountains with sparse vegetation. The sky is clear and blue.

Discussion

Challenges and Future Directions

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

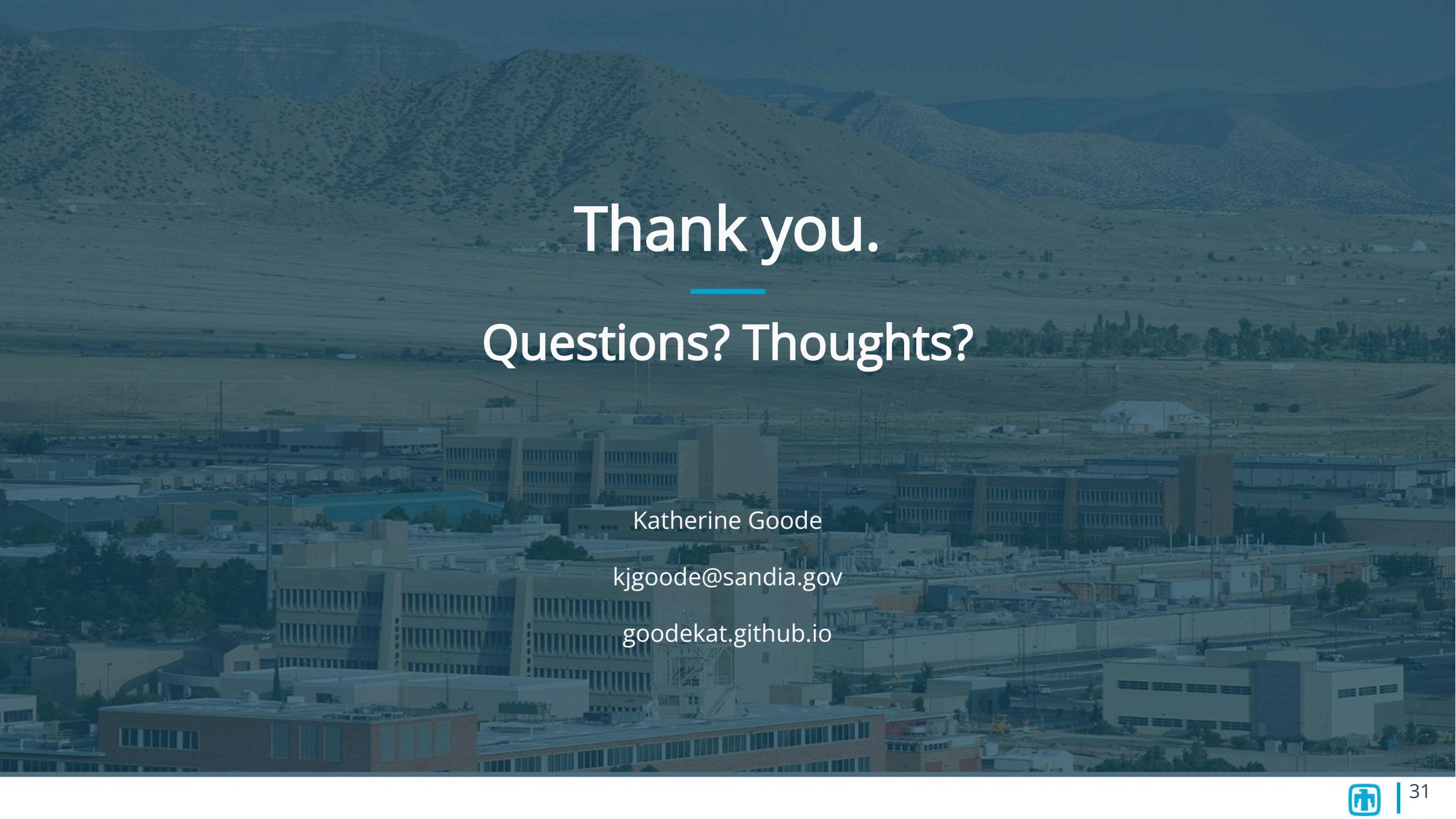
Going Forward...

Continuing to develop the requirements based on...

- feedback
- additional research into interpretability/explainability definitions
- 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