Evaluating the Maturity Level of Scientific Machine Learning Explainability

Katherine Goode, Erin Acquesta, Jason Adams, Candace Diaz, Raga Krishnakumar, and Ernesto Prudencio

SciML Credibility Team

Sandia National Laboratories

Introduction

Balancing AI Pros/Cons The National Nuclear Security Administration (NNSA) Labs emphasize trusted artificial intelligence (AI) as a necessity to meet the national security mission delivery.

- 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 (e.g., 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.

Figure 1: Sandia's five major program portfolios.

Defining Terms

Trust defines the state of the decision maker.

- Example: Decision maker integrates explainability into their decisions.

Trustworthy defines the state of the model.

- Example: Red team tested for security and bias is known and accounted for.

Credibility defines the technical basis of the model.

- Example: Verification, validation, and UQ.

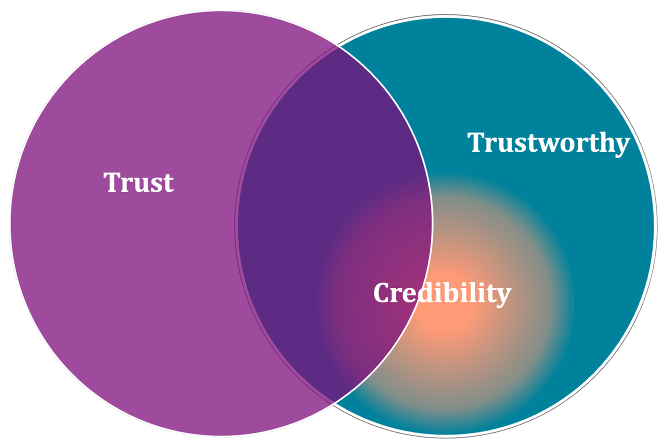

Figure 2: Credibility leads to trustworthy models; Trustworthy models may establish trust.

Computational Simulation

- "Computational modeling is the use of computers to simulate and study complex systems using mathematics, physics and computer science" (NIH 2020).
- AKA CompSim; Modeling and Simulation; ModSim; M&S.
- CompSim focuses on creating mathematical models based on first principals; Contrast to models that start with data and then aim to approximate scientific mechanisms.

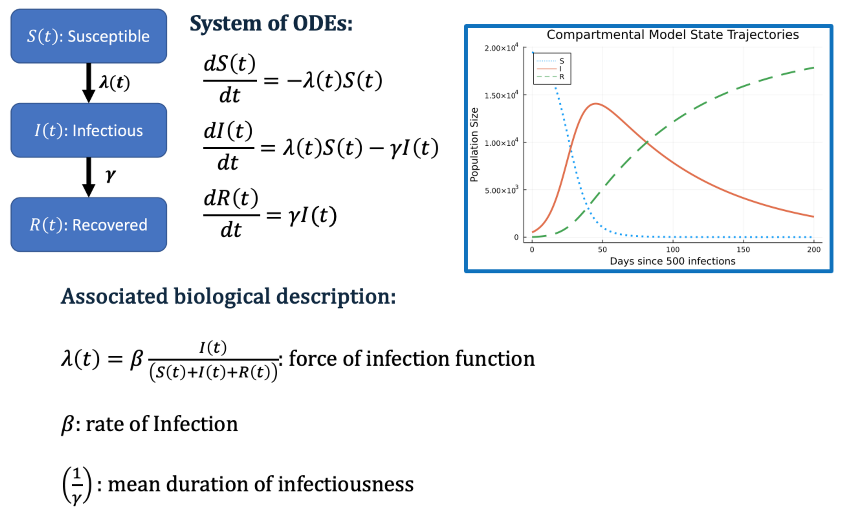

Figure 3: Epidemiology classic compartmental model.

CompSim is used in high-consequence mission spaces at Sandia.

- Example: During early stages of COVID-19 pandemic, CompSim models were used for projection modeling to inform decision makers on what may happen given a particular policy change.

Scientific Machine Learning (SciML)

- We define **SciML** as the intersection of scientific computing and machine learning.
- SciML leverages machine learning algorithms and tools used in lieu of, complementary to, or as surrogates for science and engineering computational simulation models.

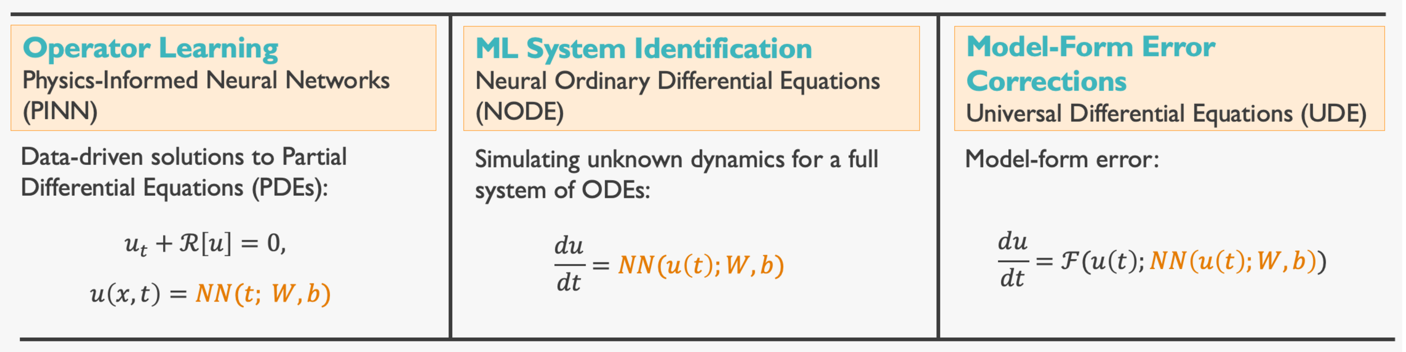

Figure 4: Examples of SciML.

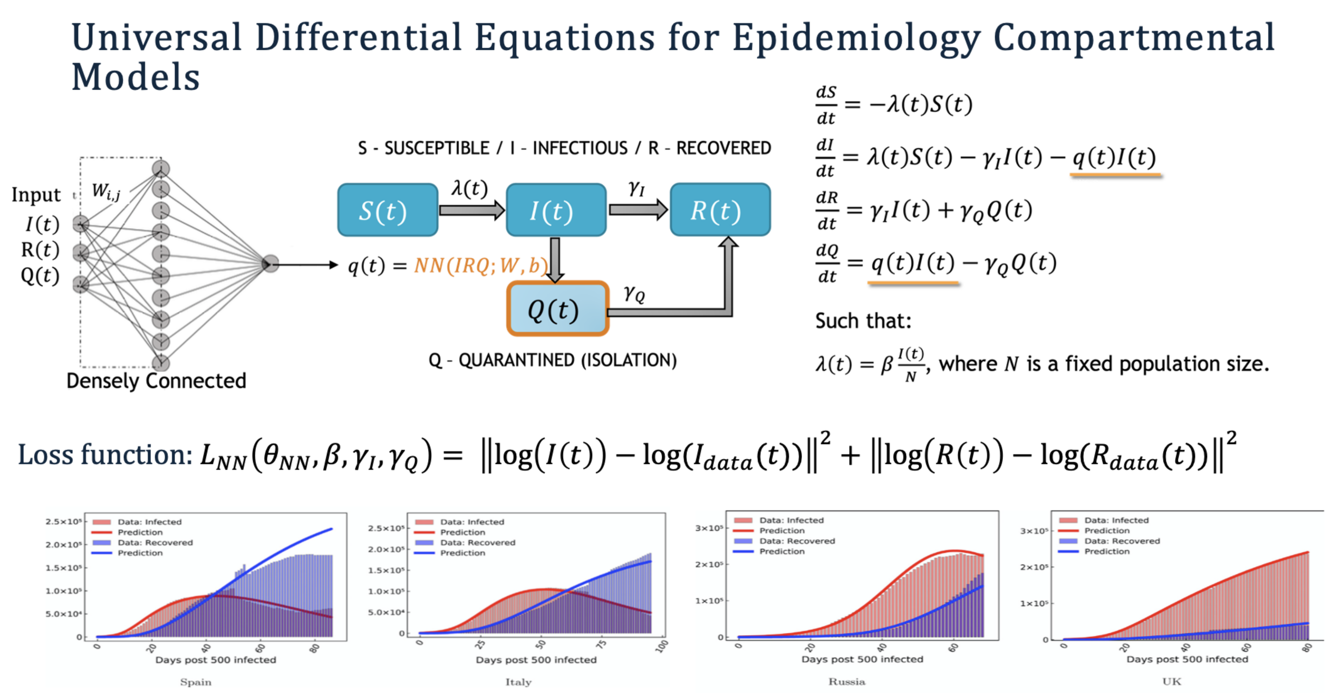

Figure 5: Model form error corrections via neural networks (Dandekar 2020).

Adapting PCMM for SciML: Focus on Interpretability/Explainability

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

- PCMM introduced in 2007 as "a model that can be used to assess the level of maturity of computational modeling and simulation" (Oberkampf 2007).
- PCMM asks:
  - Have you done something that meets this requirement?
  - NOT: Have you implemented this specific method for in order to meet this requirement?

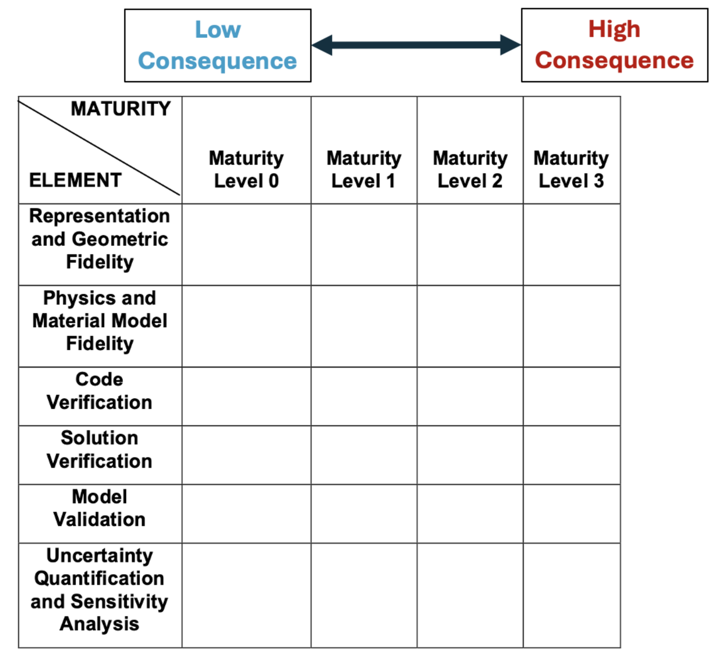

Figure 7: PCMM table.

Our Objective Adapt the PCMM table to provide a tool for establishing credibility of a SciML model. Here, we focus on the criteria needed to establish maturity levels for interpretability/explainability associated with a SciML model.

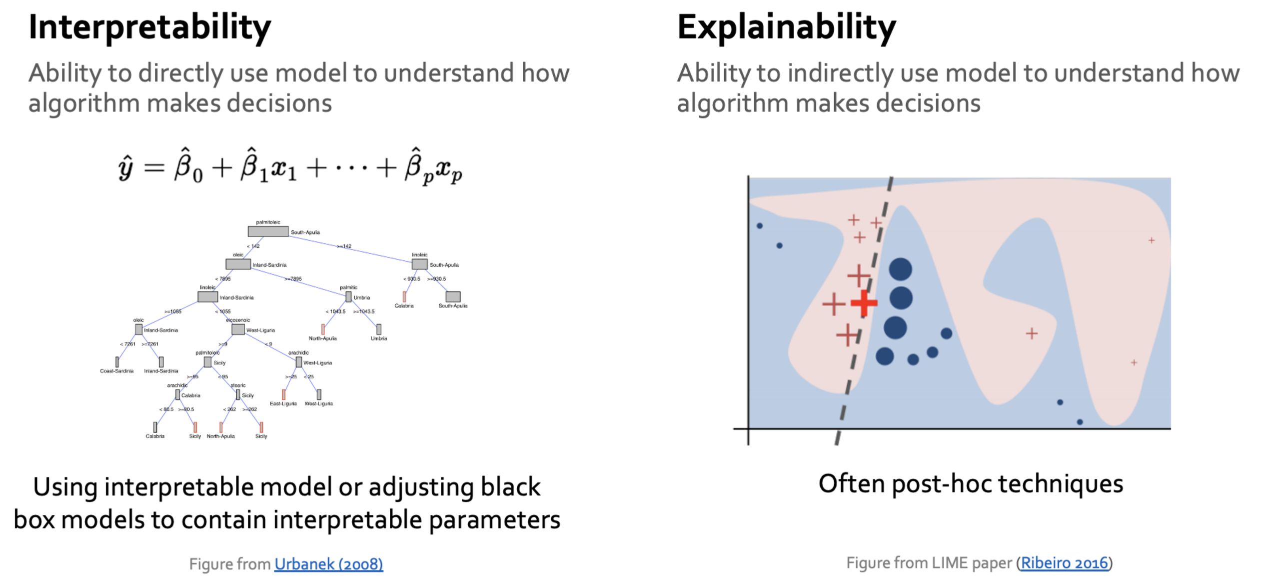

Figure 6: Interpretability versus explainability.

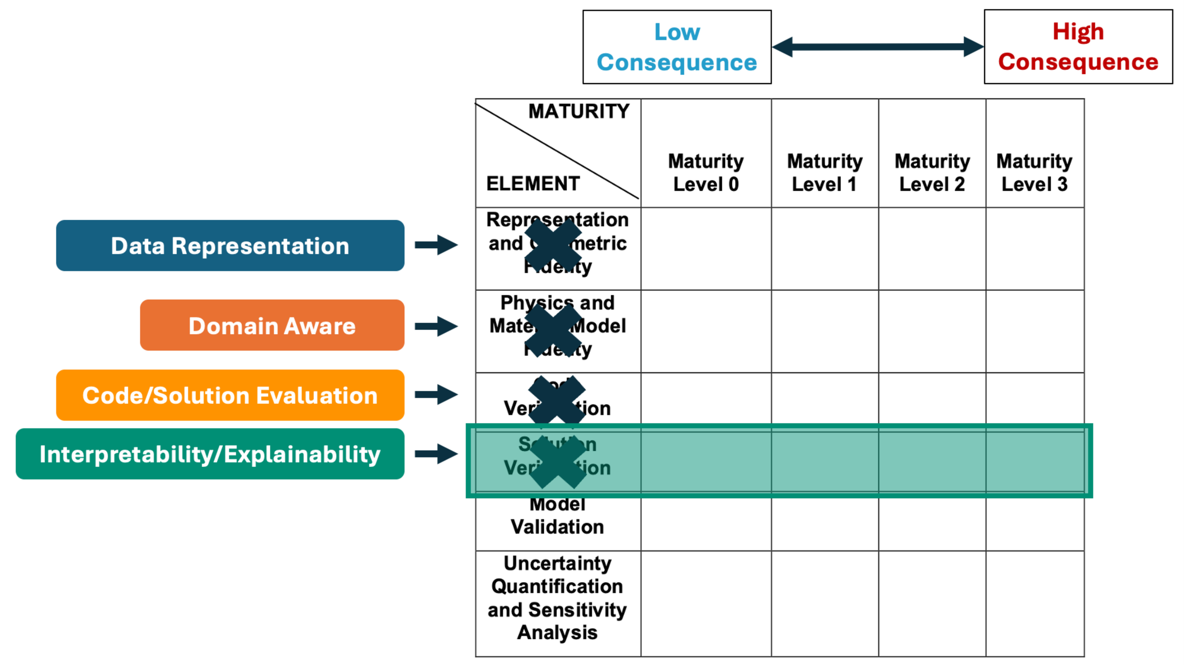

Figure 8: Adapting PCMM for SciML.

Proposed Explainability/Interpretability Maturity Levels (current state)

Considerations

- Do not want to force use of "clear-box" model, but require reasoning for use of "black-box" model.
- Explanations are approximations of a model; Important to assess if approximations are credible.
- Assumptions rely on the soundness of explainability technique.

	Level 0	Level 1	Level 2	Level 3
	Low Consequence Minimal M&S Impact	*Moderate Consequence Some M&S Impact*	*High-Consequence High M&S Impact*	*High-Consequence Decision-Making Based on M&S*
	Rigor of... **Reasoning for use of a black-box model**			
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 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?
	Rigor of... **Model interpretation or applications AND assessment of explanations**			
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)
	Rigor of... **Peer-review of model interpretations/explanations**			
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)
	Rigor of... **Diagnosis of model's scientific soundness informed by interpretations/explanations**			
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

Going Forward

- Slowly growing emphasis in the literature on methods for assessing explanation credibility.
- Interested in focusing on development of novel methods for assessing the element of explainability.

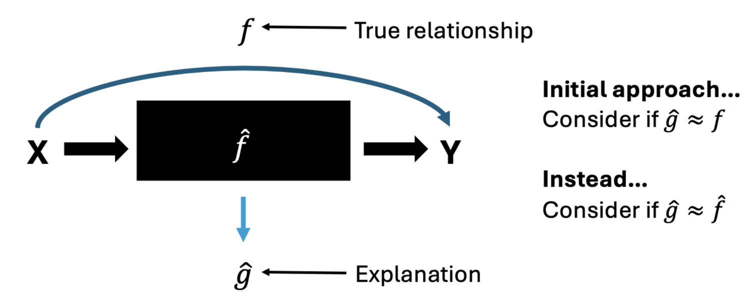

Figure 9: Consider new ways to assess fidelity of explanations.

References

1 NIH. May 2020. https://www.nibib.nih.gov/science-education/science-topics/computational-modeling

2 Dandekar, R., Rackauckas, C. and Barbastathis, G., 2020. A machine learning-aided global diagnostic and comparative tool to assess effect of quarantine control in COVID-19 spread. Patterns, 1(9).

3 Oberkampf, William Louis, Trucano, Timothy Guy, and Pilch, Martin M, "Predictive Capability Maturity Model for computational modeling and simulation.," (2007), https://doi.org/10.2172/976951.

4 Urbanek, S. (2008). Visualizing Trees and Forests. In: Handbook of Data Visualization. Springer Handbooks Comp.Statistics. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-33037-0_11.

5 Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525. SAND NO. 2025-01985C

kjgoode@sandia.gov