SAND2024-05545C

Sandia National Laboratories

Evaluating a Credibility Technical Basis Towards Trusted AI for High Consequence Applications

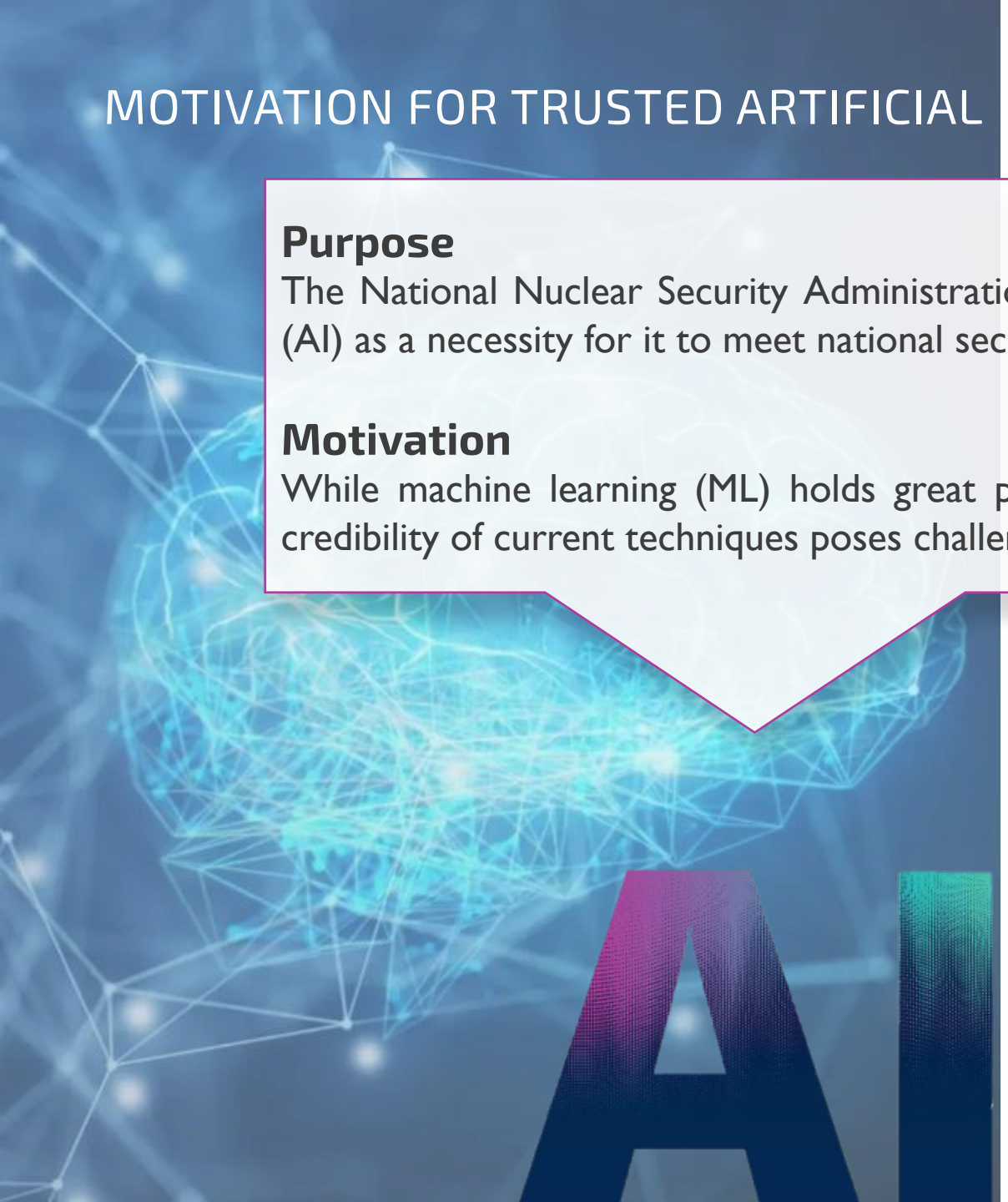
Erin C.S. Acquesta

AI Expo

May 7-8, 2024

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-NA0003525.

MOTIVATION FOR TRUSTED ARTIFICIAL INTELLIGENCE



Purpose

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

Motivation

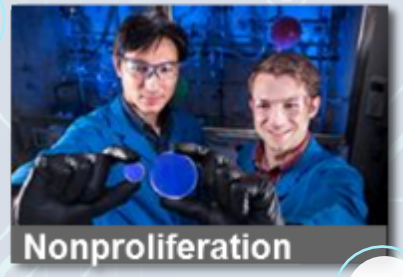
While machine learning (ML) holds great potential for mission critical applications, evaluating the credibility of current techniques poses challenges that may hinder its widespread acceptance and use.

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

TRUSTED AI SCOPE INFORMED BY MISSION NEEDS

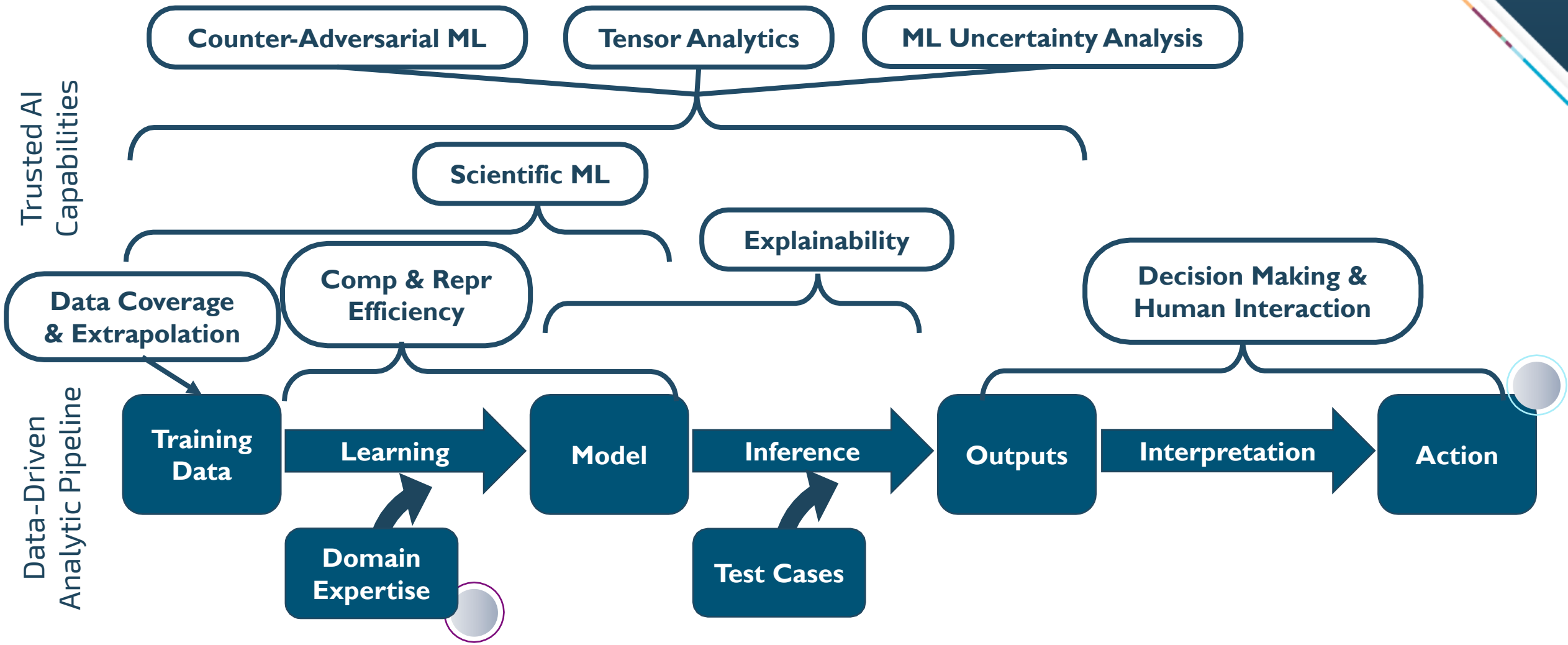
Sandia's mission needs set us apart from industry and academia

- High-consequence applications require well-characterized models and predictions
- Many national security applications have sparse, incomplete data
- Solutions require extrapolation beyond the space of available data
- Domain expertise plays a critical role in model construction
- Deployed environments with size, weight, and power constraints
- Decisions may need to be made under time pressure
- Need to account for potential adversarial issues



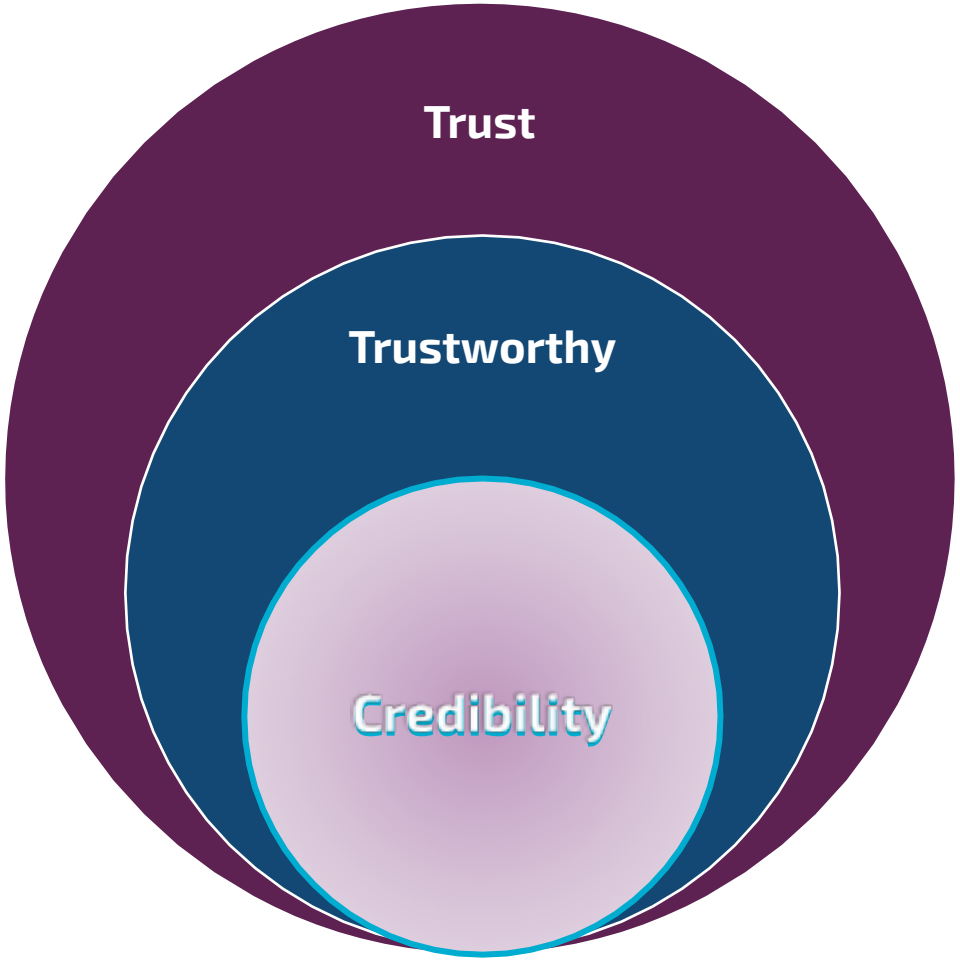
Two broad classes of mission drivers:
Scientific Modeling and Sensor-Driven Use Cases

SANDIA 5-YEAR GOAL: TOWARDS A TRUSTED AI CERTIFICATION PROCESS



Trusted AI capabilities at Sandia have identified the credibility and trustworthy criteria for establishing trust.

TRUSTED AI METHODS ARE GROUNDED IN **TRUSTWORTHY** EVIDENCE, ROOTED IN A **CREDIBILITY** TECHNICAL BASES



Trust: Defines the state of the decision maker.

- Decision maker integrates model inference and/or predictions into their decision making process.

Trustworthy: Defines the state of the model.

- Bias is known and accounted for.
- Interpretability and explainability can be established.

Credibility: Identifies the technical basis of the model.

- Verification, validation, uncertainty quantification
- Data and Geometric Representations.

Credibility leads to trustworthy models, and trustworthy models may establish trust.

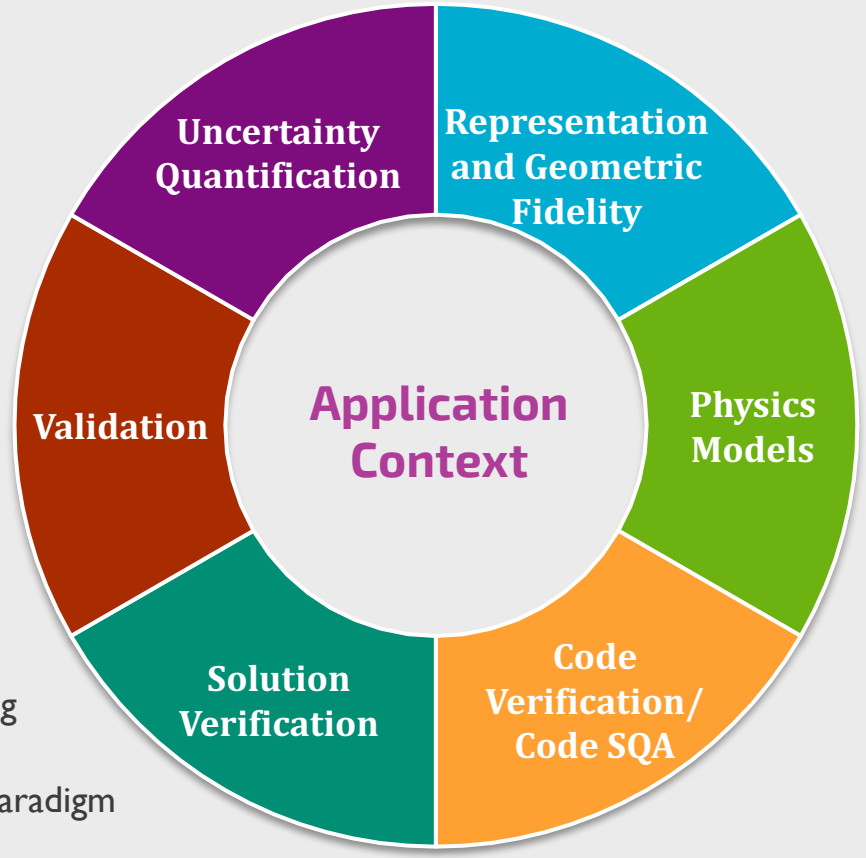
Caution needs to be heeded with trusted models. Trust in a model does not guarantee that credibility has been established.

PCMM: PREDICTIVE CAPABILITY MATURITY MODEL, IS AN EVIDENCE COLLECTION PROCEDURE TO EVALUATE CREDIBILITY

The computational simulation (CompSim) credibility process assembles and documents evidence to ascertain and communicate the believability of predictions that are produced from computational simulations.

Evidence Basis

- Plan
- Execute
- Organize & Analyze



Elements

- Categories for collecting evidence
- Dependent on model paradigm

Communication

- Peer Review
- Plausible Prediction Bounds

CompSim Model → Prediction

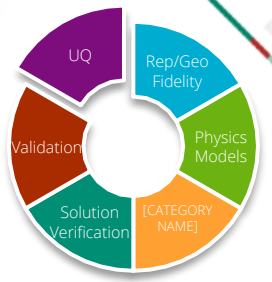
Application Context

- Partial Differential Equations (PDE)
- Computational Fluid Dynamics (CFD)

Our work builds upon the NNSA's 20+ years of experience in verification, validation, and uncertainty quantification (VV/UQ) for complex problems with limited data.

"Predictive Capability Maturity Model for Computational Modeling and Simulation" by Oberkampf, W.L., Pilch, M., and Trucano, T.G., SAND2007-5948

PCMM LEVELS: UNCERTAINTY QUANTIFICATION (UQ)



PCMM levels depend on the degree to which the decision is high-consequence and to what degree the Model and Simulation (M&S) provides the information making the decision.

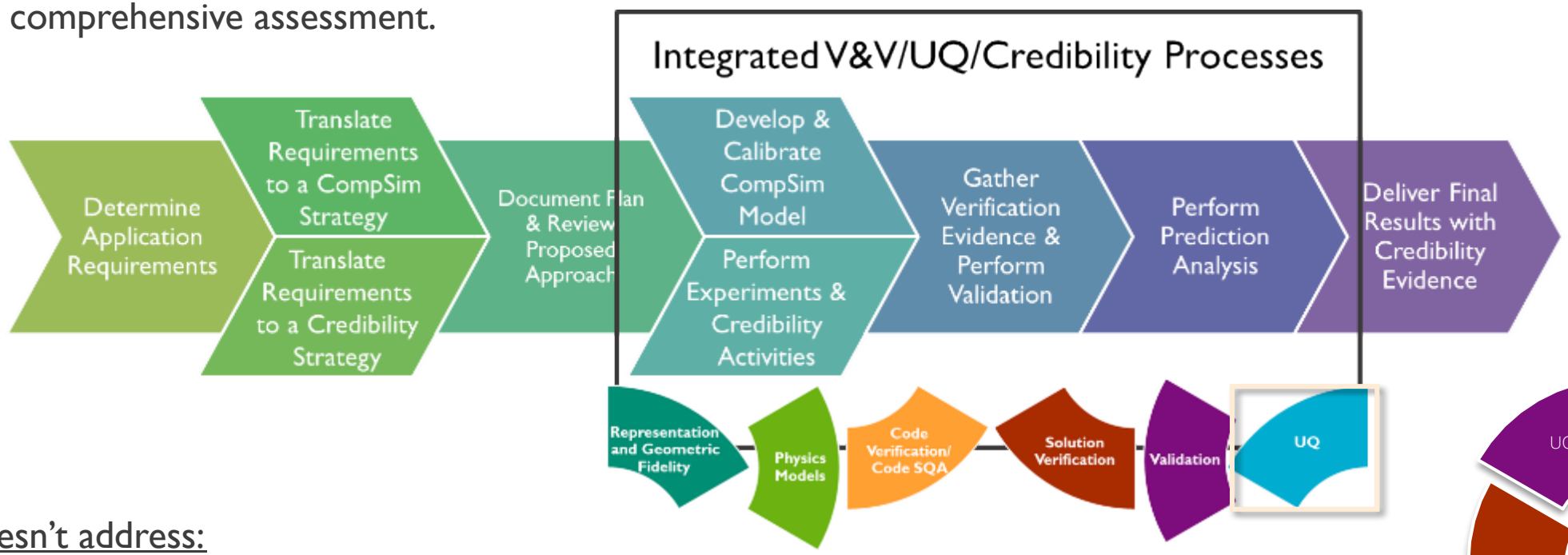
Maturity	Level – 0	Level – 1	Level – 2	Level – 3
Overall Description	Low Consequence, Minimal M&S Impact, e.g. Scoping Studies	Moderate Consequence, Some M&S Impact, e.g. Design Support	High-Consequence, High M&S Impact, e.g. Qualification Support	High-Consequence, Decision-Making Based on M&S, e.g. Certification
UQ Element Description	• Judgement only • Only deterministic analyses are conducted • Uncertainties sensitivities are not addressed	• Aleatory and epistemic (A&E) uncertainties propagated, but without distinction • Informal sensitivity studies conducted • Many strong UQ/Sensitivity Analysis (SA) assumptions made	• A&E uncertainties segregated, propagated and identified in System Response Quantities (SRQ) • Quantitative SA conducted for most parameters • Numerical propagation errors are estimated their effect known • Some strong assumptions made • Some peer review	• A&E uncertainties comprehensively treated and properly interpreted • Comprehensive SA conducted for parameters and models • Numerical propagation errors are demonstrated to small • No significant UQ/SA assumptions made • Independent peer review conducted

The maturity level required for the application context determines the evidence needed across each element for assessing the credible use of the model for its intended context of use.

"Predictive Capability Maturity Model for Computational Modeling and Simulation" by Oberkampf, W.L., Pilch, M., and Trucano, T.G., SAND2007-5948

UQ IS ONLY ONE ELEMENT IN THE CREDIBILITY PROCESS

To determine the overall maturity level of model's use in high-consequence decision-making environments requires comprehensive assessment.



UQ doesn't address:

- Are we solving the right problem?
- Are the important physics phenomenon adequately represented?
- Are there bugs in the code?
- What are the numerical error?

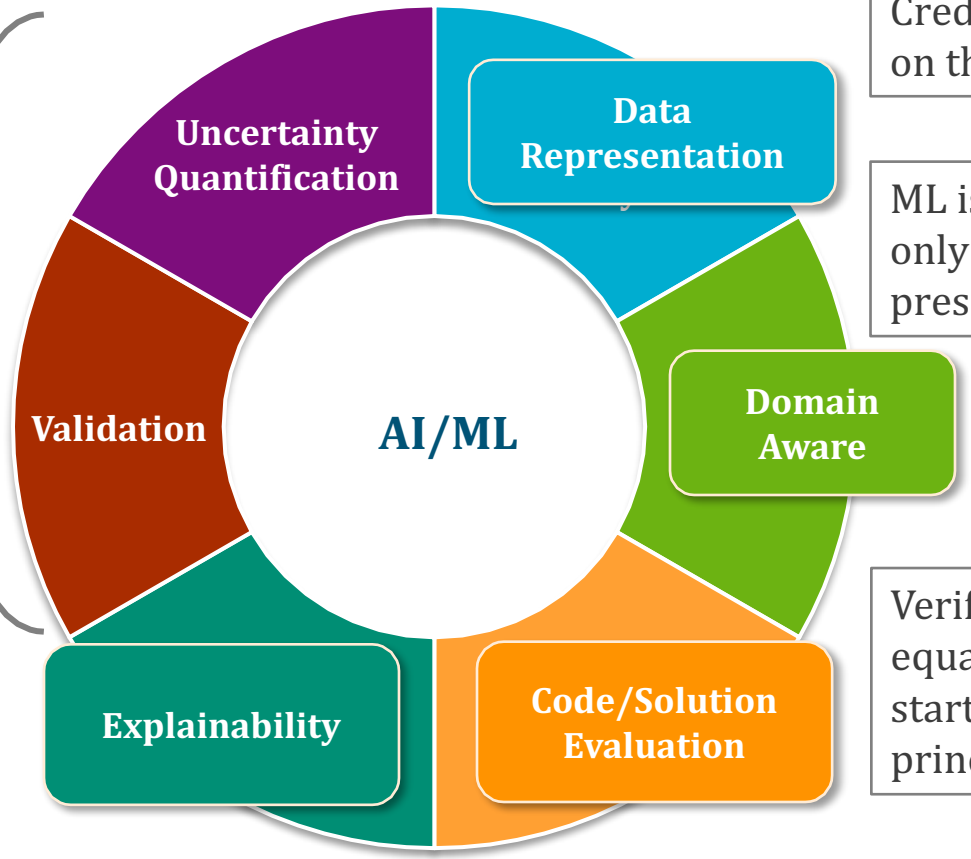
UQ is supported by and tethered to the other elements.

ADAPTING PCMM ELEMENTS TO A **CREDIBILITY PROCESS FOR AI/ML MODELS**

Credibility process assembles and documents evidence to ascertain and communicate the believability of predictions that are produced from computer models.

UQ and Validation are currently the core areas of research that exists for ML credibility that are readily transferrable.

ML community has prioritized explainability to develop trust in ML. The maturity of these methods need to also be evaluated.



Credibility of any ML model is predicated on the credibility of the data used to train it.

ML is applied more broadly and it is not only physical principles we want to preserve.

Verification asks "are we solving the equations correctly", ML models do not start with equations….yet, similar principles will still need to applied.

There is still a numerous amount of outstanding challenges in applying VV/UQ principles to ML.

EVALUATING THE CREDIBILITY FOR SCIENTIFIC MACHINE LEARNING

SciML: Scientific Machine Learning

Machine learned models are 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):

$$u_t + \mathcal{R}[u] = 0,$$

$$u(x,t) = NN(t;\ W,b)$$

ML System Identification

Neural Ordinary Differential Equations (NODE)

Simulating unknown dynamics for a full system of ODEs:

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

Model-Form Error Corrections

Universal Differential Equations (UDE)

Model-form error:

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

Building on a 30 year history in evaluating credibility for computational simulation models for AI/ML is most naturally translatable by focusing on SciML applications.

UNIVERSAL DIFFERENTIAL EQUATIONS (UDE) MOTIVATION

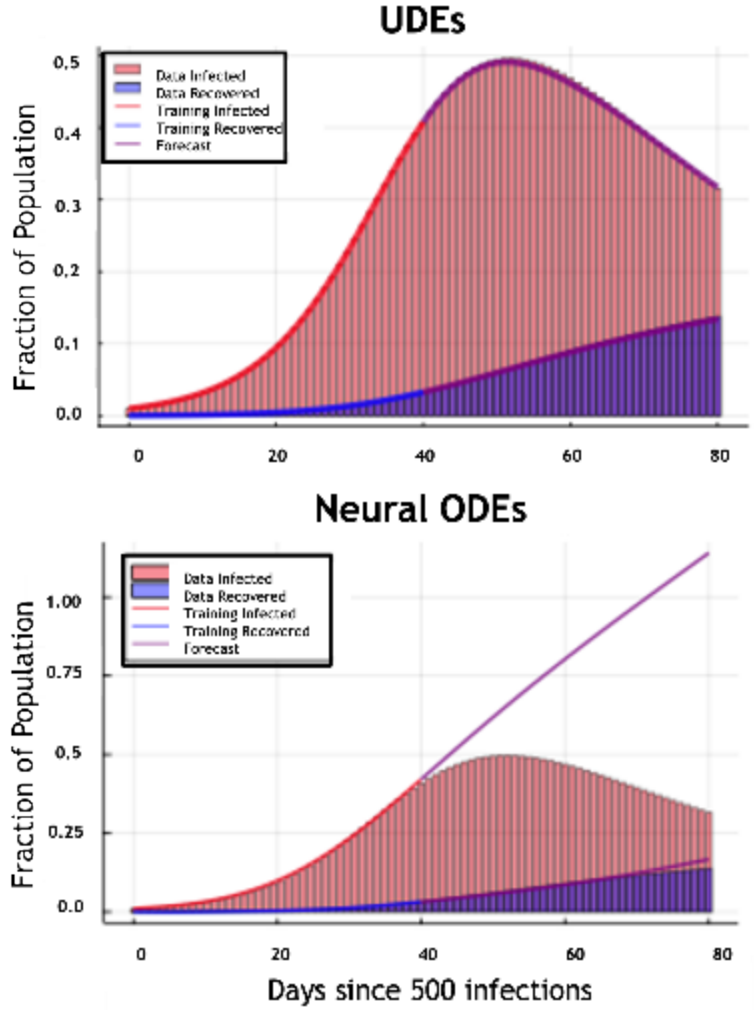
- UDEs have been successfully deployed to infer interpretable, predictive dynamics from data.
- UDEs embed ML models, e.g., neural networks (NNs) within existing scientific models:

$$\boldsymbol{u}' = F\left(\boldsymbol{u}, t, \theta_{ODE}, NN(\boldsymbol{u}, \theta_{NN})\right)$$
$$\min_{\theta} \|\boldsymbol{d} - \boldsymbol{u}(\theta)\|$$

where $\theta = \{\theta_{ODE}, \theta_{NN}\}$ and $\boldsymbol{d}$ represents observation data.

- Data-efficient because make sure of prior physical information.
- Can be more predictive than Neural ODEs:

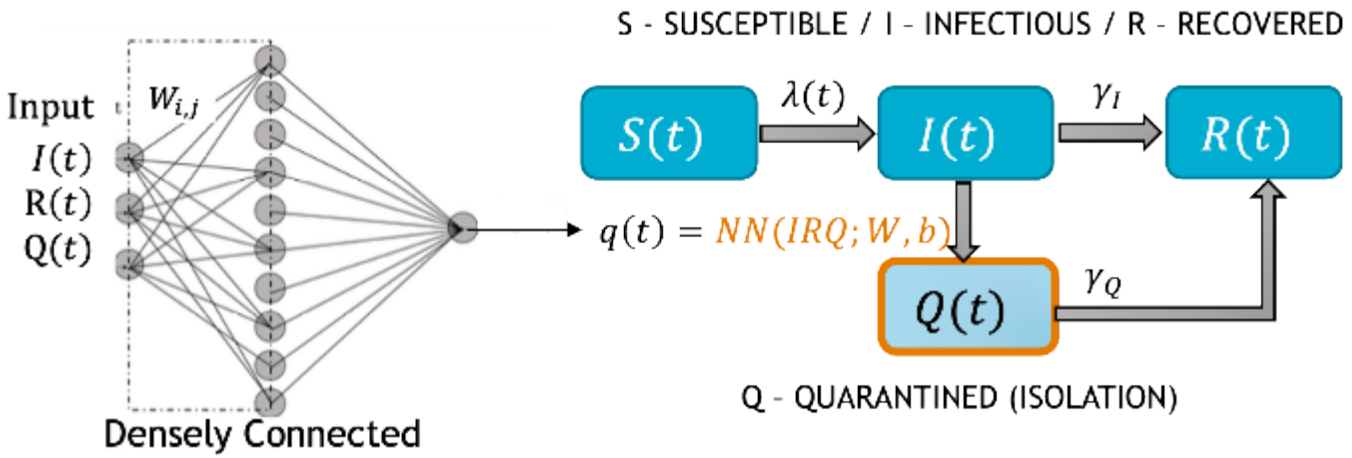
$$\boldsymbol{u}' = NN(\boldsymbol{u}, \theta_{NN})$$
$$\min_{\theta_{NN}} \|\boldsymbol{d} - \boldsymbol{u}(\theta_{NN})\|$$



UDEs provide a SciML structure that preserves subject matter expertise while learning data-driven model-form error corrections.

Rackauckas, C., Ma, Y., Martensen, J., Warner, C., Zubov, K., Supekar, R., Skinner, D., Ramadhan, A. and Edelman, A., 2020. Universal differential equations for scientific machine learning. *arXiv preprint arXiv:2001.04385*.

UDES FOR EPIDEMIOLOGY COMPARTMENTAL MODELS



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

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

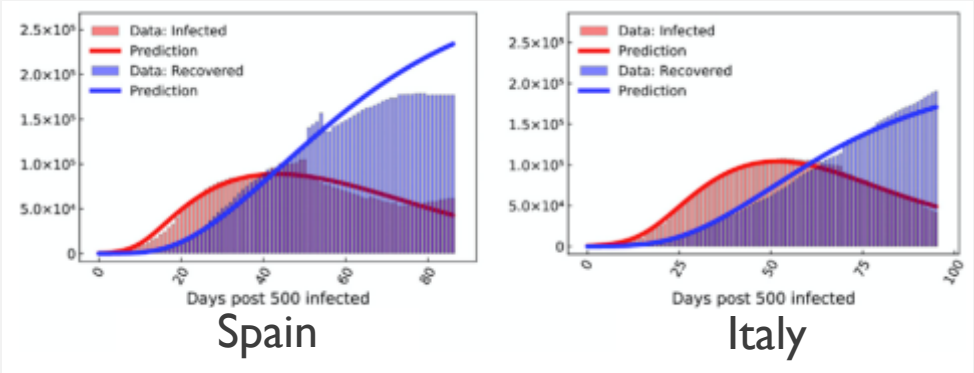
$$\frac{dR}{dt} = \gamma_I I(t) + \gamma_Q Q(t)$$

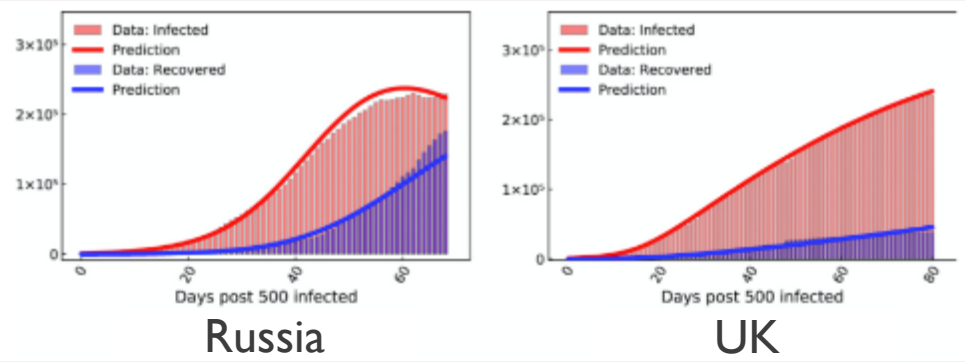
$$\frac{dQ}{dt} = q(t)I(t) - \gamma_Q Q(t)$$

Such that:

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

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





UDEs provide a SciML structure that preserves subject matter expertise while learning data-driven model-form error corrections.

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).

SOURCES OF UNCERTAINTY THAT IMPACT PREDICTION UNCERTAINTY IN DIFFERENTIAL EQUATION MODELS

Notional Ground Truth Model

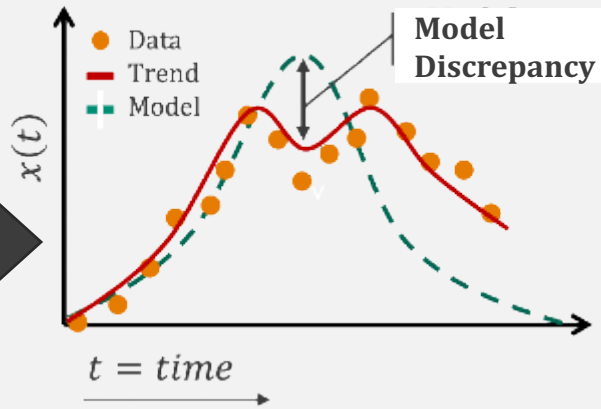
$$\frac{dx}{dt} = F(x(t), z(t))$$

Ordinary Differential Equation Initial Value Problem

$$\frac{dx}{dt} = F(x(t), \tilde{z}(t))$$

$$x(0) = 0$$

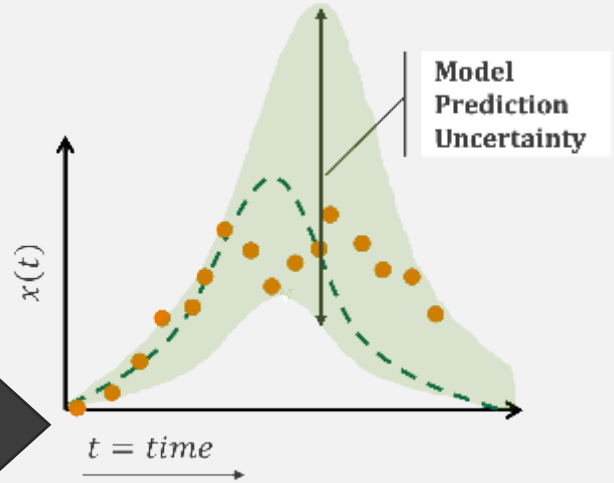
Model Form Error: $z(t) - \tilde{z}(t)$



MD is the difference between the model solution and the filtered trend in the data.

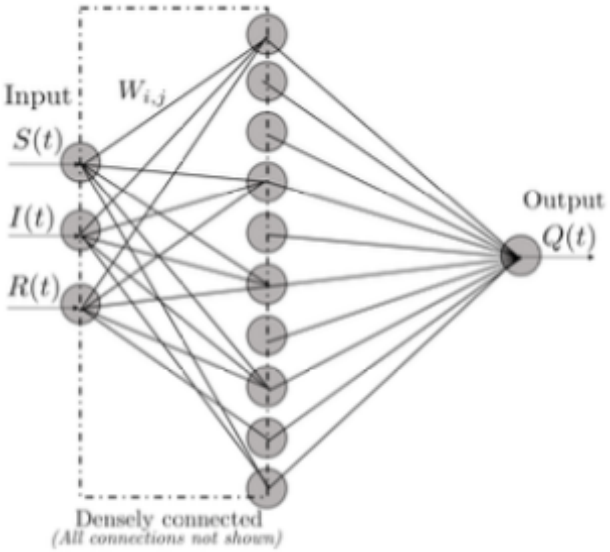
Sources of Prediction Uncertainty:

- model-form error
- parameter
- data
- numerical error



Uncertainty quantification is an essential element of PCMM.
Neural networks define a generalizable approach to approximate model-form error corrections. Why should we trust it?

KEY THEORETICAL FOUNDATIONAL OF S CIML



Universal Approximation Theorem (UAT)

(one version) Fix a continuous function $\sigma: \mathbb{R} \to \mathbb{R}$ (activation function) and positive integers d, D. The function σ is not a polynomial if and only if, for every continuous function $f: \mathbb{R}^d \to \mathbb{R}^D$ (target function), every compact set K of $\mathbb{R}^d$, and every $\varepsilon > 0$ there exists a continuous function $f_\varepsilon: \mathbb{R}^d \to \mathbb{R}^D$ (the layer output) with representation

$$f_\varepsilon = W_2 \circ \sigma \circ W_1$$

where W_2, W_1 are composable affine maps and $\circ$ denotes component-wise composition, such that the approximation is bounded

$$\sup_{x \in K} ||f(x) - f_\varepsilon(x)|| < \varepsilon$$

Code/Solution Evaluation is one of the hardest elements to define the analysis required for PCMM
UAT is a necessary condition for NN to be function approximators, in practice this is not a sufficient condition.

CHALLENGES EVALUATING CREDIBILITY FOR SCIML: USE EPIDEMIOLOGY EXAMPLE

The original dynamical system is known to be under-representative of the real-world phenomena it is intended to simulate.

Baseline ODE is NOT Credible.

- To build a Credible UDE, we need to identify our known-unknowns.

NN Universal Approximators as model-form error corrections.

Uncertainty Challenges:

- What impact does using a UDE for model-form error have on model-form uncertainty?
- Aggregating:
  - ODE parameter uncertainty
  - NN parameter uncertainty
  - NN architecture uncertainty
  - NN numerical uncertainty
  - Model-form uncertainty

Verification Challenges:

- Universal Approximation Theory: Does the application have a gap between theory and practice?
- Convergence of NN optimization can get stuck in local minima

Validation Challenges:

- Training-Test-Validation comparison to Calibration-Validation
- Known-Unknowns

TO SUMMARIZE IN CLOSING…

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

- Credibility of computational methods is deeply rooted in the technical bases for VVUQ and evaluated through maturity model frameworks.
- While ML holds great potential for mission critical applications, evaluating the credibility of current techniques poses challenges that may hinder its widespread acceptance and use.
- Model-form error corrections can drive down model-form uncertainty, but using NN-based methods can introduced more sources of uncertainty.
- Credibility is at the core of trustworthy models, and essential for establishing Trusted AI,

Credibility of AI is an important topic that will continue to be addressed in the DOE Frontiers of AI for Science, Security, and Technologies (FASST) program.

SANDIA NATIONAL LABORATORIES

WWW.SANDIA.GOV/AI/

Thank You for Your Time and Attention!

For questions or follow-up discussions:

Erin Acquesta
eacques@sandia.gov

AI