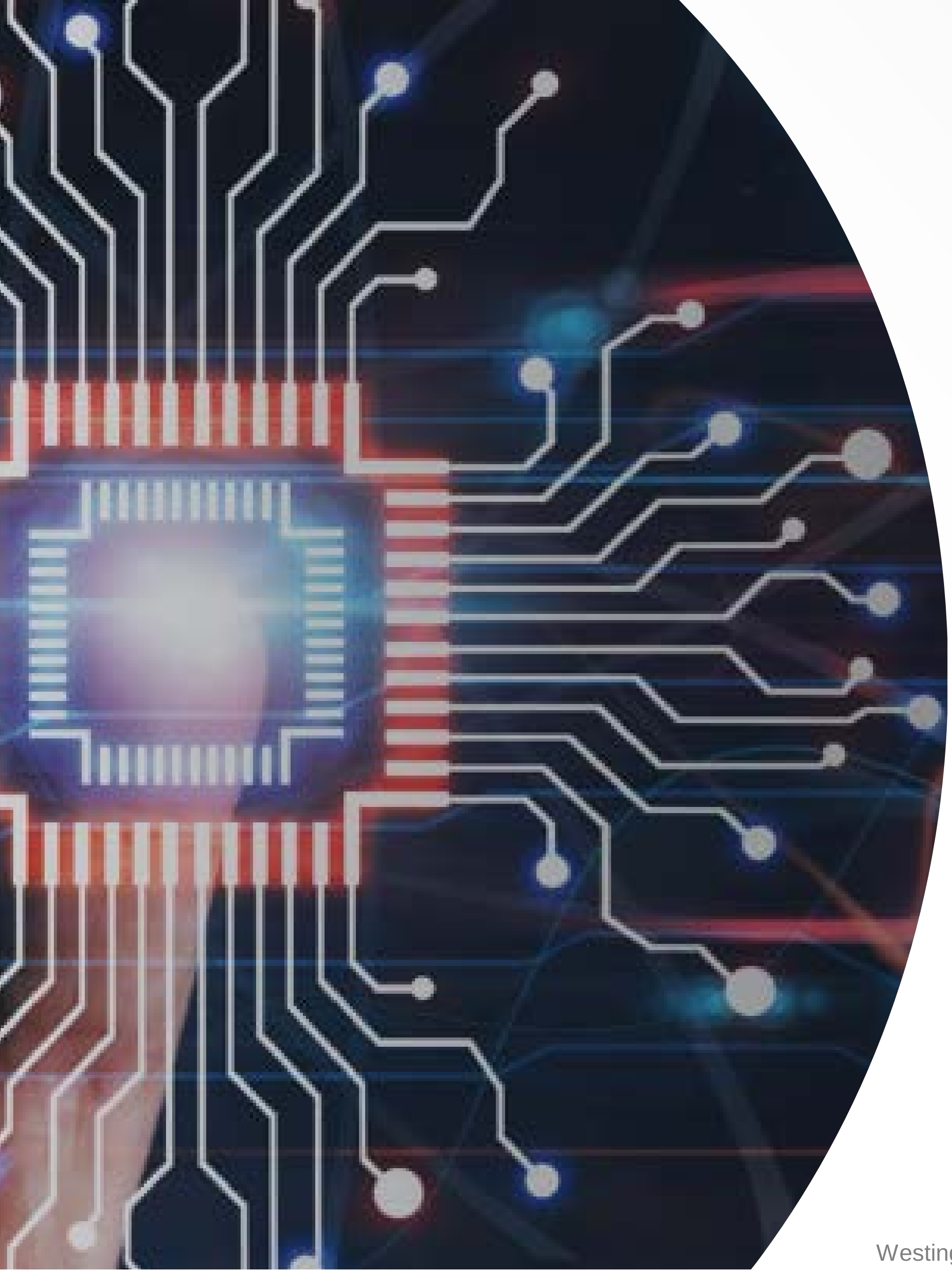


Practical Considerations for Application of AI/ML

AI Policy, Risk-Informed, Credibility, & VVUQ

Scott Sidener

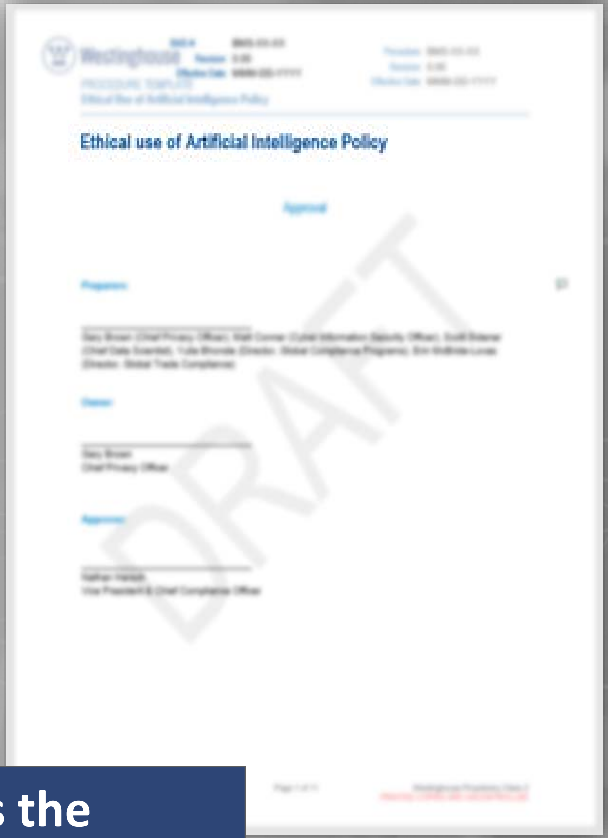
Westinghouse Electric Company

Chief Data Scientist / Chief Engineer - Digital

Westinghouse

Ethical use of Artificial Intelligence –
Create/Have a Corporate Policy

- **Key Elements**
  - **Purpose, Intent, & Applicability**
  - **Ethical and responsible use of AI**
    - Transparency
    - Harmful Bias Mitigation
    - User Awareness
  - **Data Privacy & Security**
  - **Risk & Assessment**
    - Safety & Reliability
    - Monitoring
  - **Accountability**
    - Data Ethics Committee
    - Human-in-the-Loop
  - **Compliance**
    - Audit


Ethical use of Artificial Intelligence Policy
DRAFT

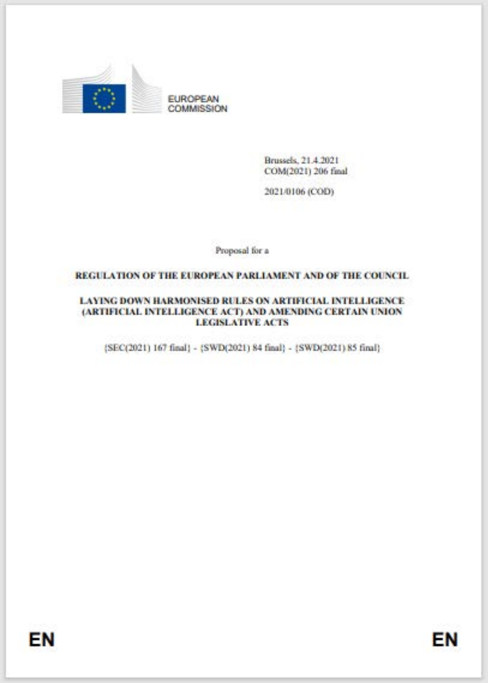
A clear policy provides the appropriate guardrails before development & during application

Risk Informed AI/ML Application

- **What can go wrong?**
  - Incorrect prediction
  - Incorrect classification
  - False or misleading generation
  - Loss of information security or privacy
  - Misuse
  - Etc.

- **How likely is it to happen?**
  - Evaluation of model credibility within context of use
  - Performance-based validation

- **What are the consequences?**
  - Level 1: adversely impacts safety or regulation compliance
  - Level 2: adversely influences analyses, decisions, human behavior, etc.
  - Level 3: Loss of time and/or money



EUROPEAN COMMISSION

Brussels, 21.4.2021
COM(2021) 206 final

2021/0106 (COD)

Proposal for a

REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS

{SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final}

EN EN

- **EU AI Act Risk Levels**
  - Unacceptable Risk
  - High Risk
  - Limited Risk

AI/ML Credibility –
The quality that a model can be trusted for a context of use

- **Applicability**
  - Relevance of the evidence from validation activities to support the use of the model for a **context of use**
- **Predictive or Generative Capability**
  - The anticipated accuracy and precision of the model over a specified application domain
- **Evidence**
  - Qualitative & Quantitative
  - VVUQ
- **Credibility is not Static**
  - Requires continuous monitoring of both context and performance

Is the model credible in the expected application domain?

AI/ML Validation –

The trained model accurately represents the real application

Data & Domains

Transparency

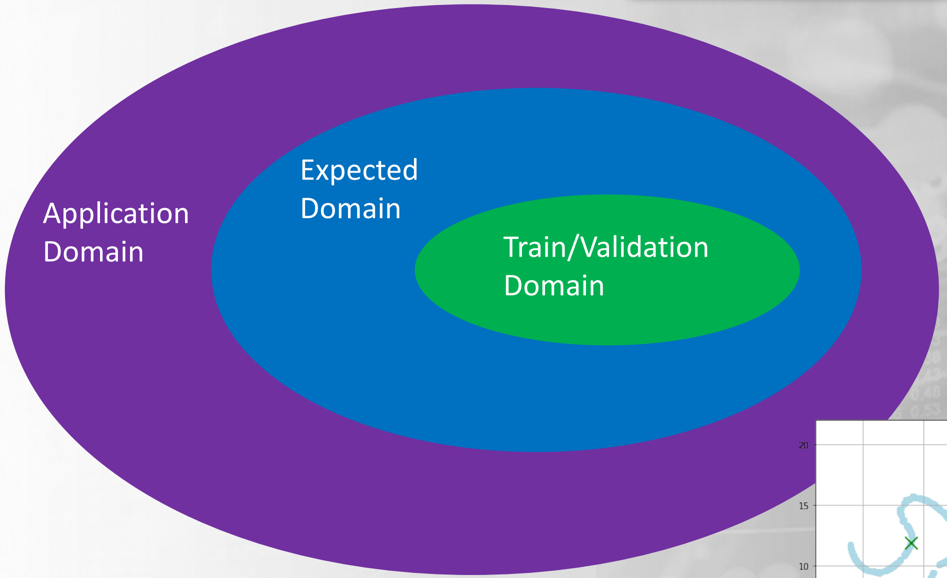
Human-Based

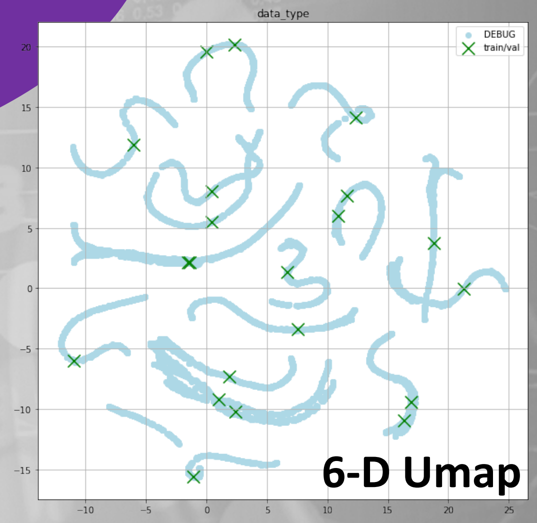
Performance-Based

AI/ML Validation

Data & Domains

- **Data Sets**
  - Training, Validation, Calibration (sometimes), Test (truly blind validation)
- **Application Domain**
  - Often a hyper-rectangle defined by bounds of training/validation data
- **Expected Domain**
  - Where analysts expect to use model, often a convex hull around simulated application data
  - This may be a very localized & sparse space
- **Train/Validation Domain**
  - Often a convex hull around data used to create the model





You won' fill a high-dimensional space with data

Always assume the model is extrapolating

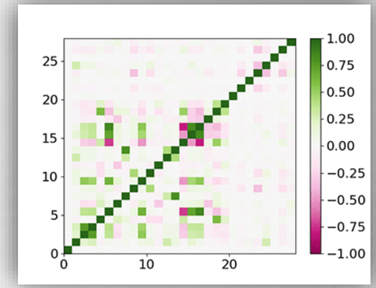
AI/ML Validation

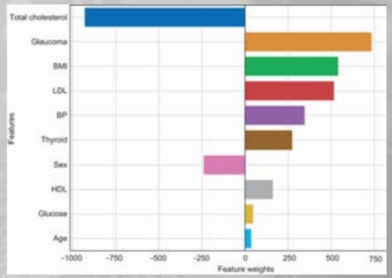
Transparency

- **Transparency is fundamentally contradictory to AI/ML technology**
  - Model size and complexity is exponentially running away from the ability to interpret all feature influences and interactions

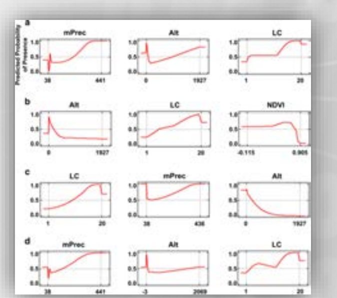
- **To gain some transparency**
  - Use simple models (linear regression, decision tree, etc.)
  - Use very low dimensionality features
  - Use physics informed ML

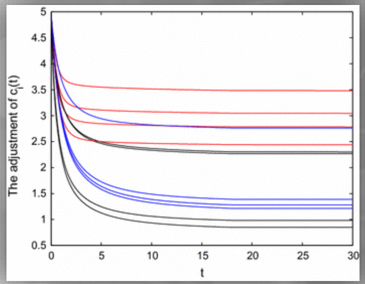
- **Sensitivity**





- **Parametric Studies**





Don't depend on full transparency & don't assume causality

AI/ML Validation

Human-Based

- **Human in the middle**
- **Accountability**
- **Consistency?**
- **Process?**
- **Skill**


The newest version of Midjourney created this image with the prompt, "photograph of a woman artist in an artist's studio holding her hands up." (all edits Elaine Velie/Hyperallergic using Midjourney)

Human validation can be effective, but risky

Rapidly requiring greater skill & attention

AI/ML Validation

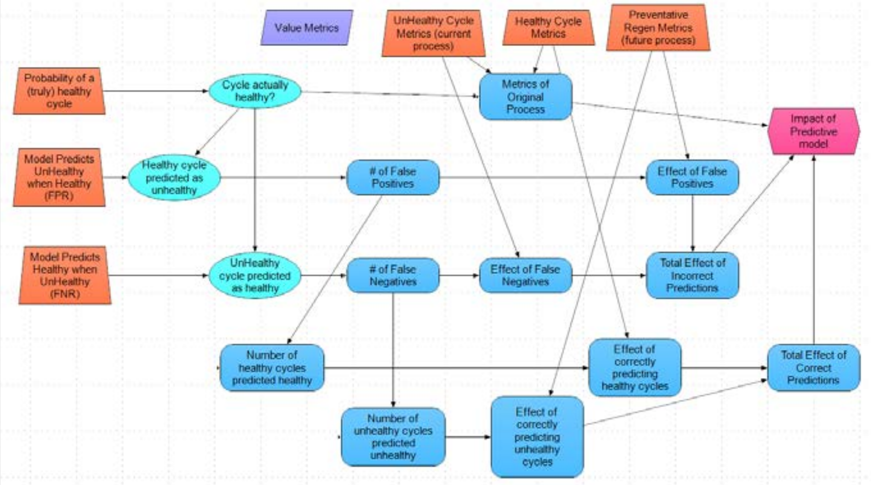
Performance-Based

- **Validation Metrics**

		Predicted Class: Positive	Predicted Class: Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	Negative Predictive Value $\frac{TN}{(TN+FN)}$	Accuracy $\frac{TP+TN}{(TP+TN+FP+FN)}$

	Metrics	Value
0	MAE	6.052
1	MSE	56.187
2	RMSE	7.496
3	R-Squared	0.389

- **Simulate Impact of Incorrect Results**
  - Performance beyond validation metrics
  - Bias detection, risk analysis

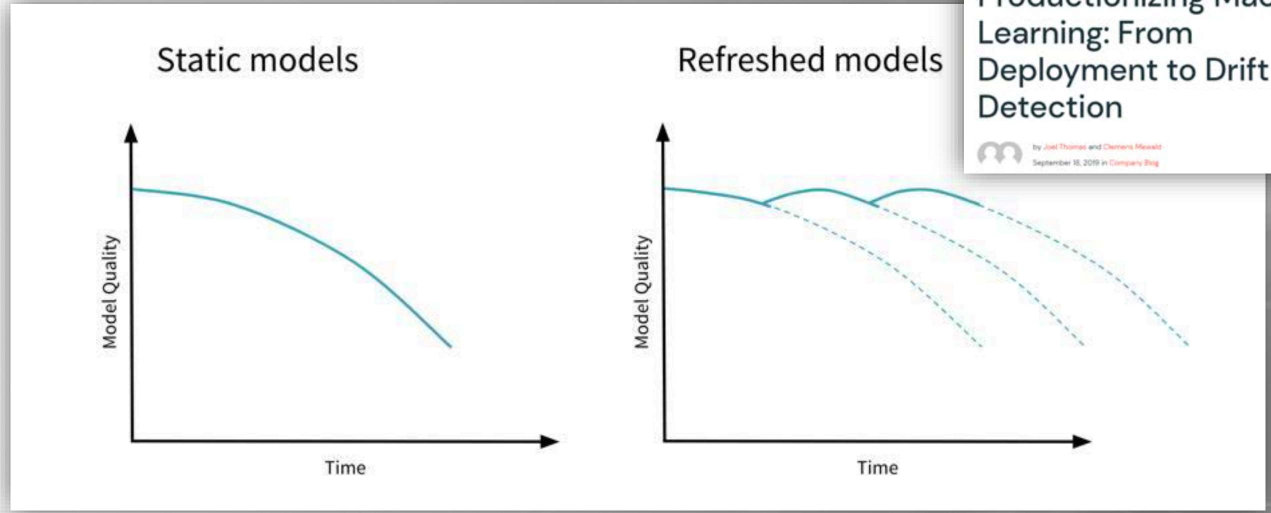


Metrics are easy to calculate, Simulating impacts is very powerful

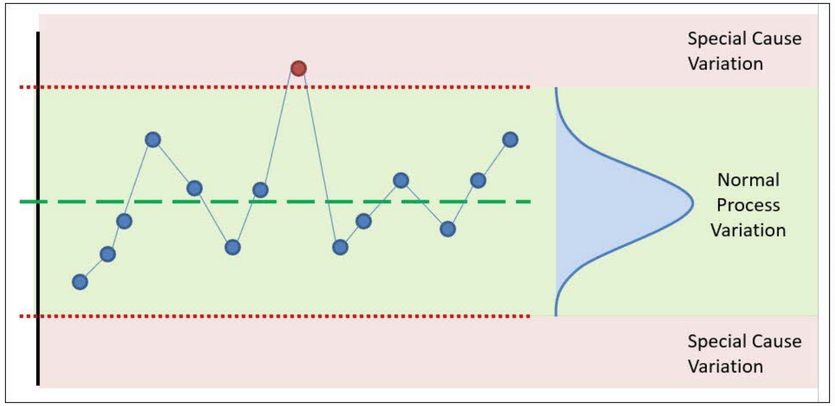
AI/ML Validation

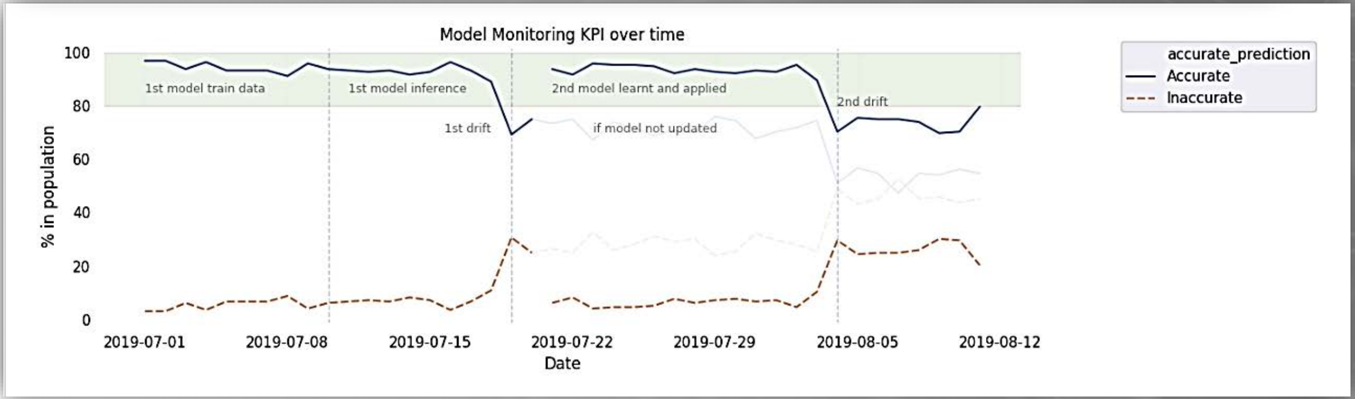
Performance-Based

- **Performance can be Ephemeral**
  - Performance monitoring
  - Continuously learning models
  - Models watching inputs
  - Statistical process control
  - Process Capability
  - Anomaly detection



Productionizing Machine Learning: From Deployment to Drift Detection





Assume model performance will change with time

Context of use unknowingly or unexpectedly drifts

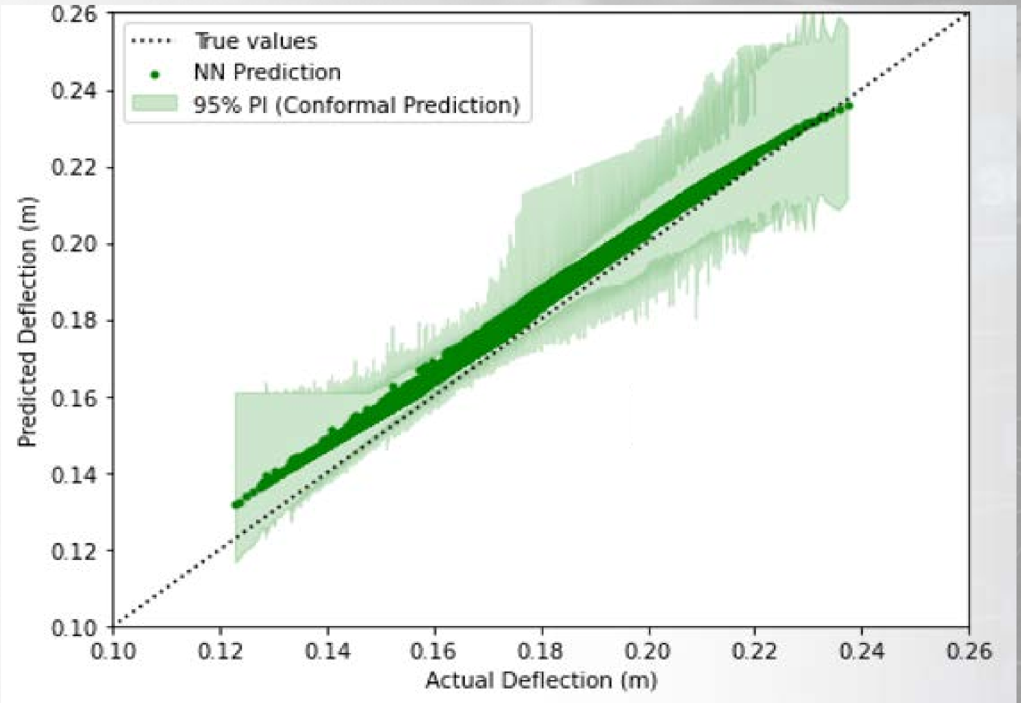
Inputs can drift or contain anomalies

AI/ML Uncertainty Quantification

- **Uncertainties**

  - ☑ Measurement / Experimental
  - ☑ Model
  - ☑ Input
  - ☒ Numerical

  “Training Uncertainty” characterizes range of predictions which can result from different choices of hyperparameters when training a model



Nearly every new prediction will not be at validation data points

Quantify, Propagate & Simulate Impacts of Uncertainty

Practical Considerations for Application of AI/ML

- Have an ethical AI policy to ensure guardrails (North Star)
- Use risk-informed approaches
- Determine credibility
- Leverage VVUQ & and simulate impacts of model defects
- Expect credibility /performance creep, monitor continuously