OAK RIDGE National Laboratory

Advancing Secure, Trustworthy, and Energy-Efficient AI for Science and Technology: A view from ORNL's AI Initiative

Prasanna Balaprakash

Director of AI Programs

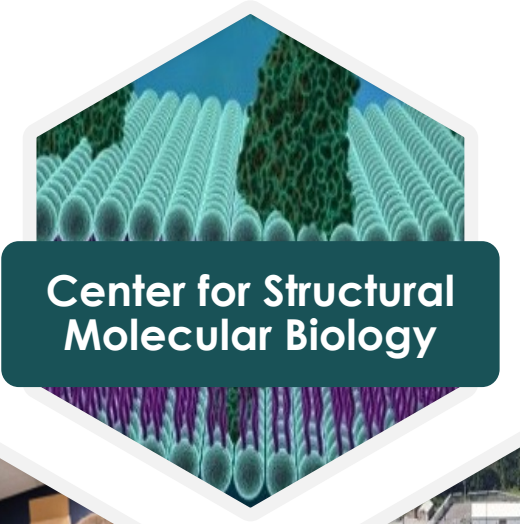
Oak Ridge National Laboratory

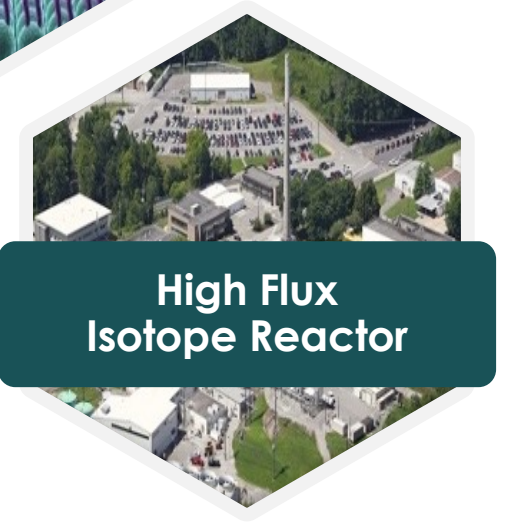
ORNL is managed by UT-Battelle LLC for the US Department of Energy

U.S. DEPARTMENT OF ENERGY

AI transforming science and national security

Accelerating scientific discovery, fortifying energy infrastructure, and enhancing national security





OAK RIDGE National Laboratory

Grand challenges in AI for science and security

nature

Explore content ⌄ About the journal ⌄ Publish with us | Subscribe

nature › editorials › article

EDITORIAL | 27 September 2023

AI will transform science — now researchers must tame it

A new *Nature* series will explore the many ways in which artificial intelligence is changing science — for better and for worse.

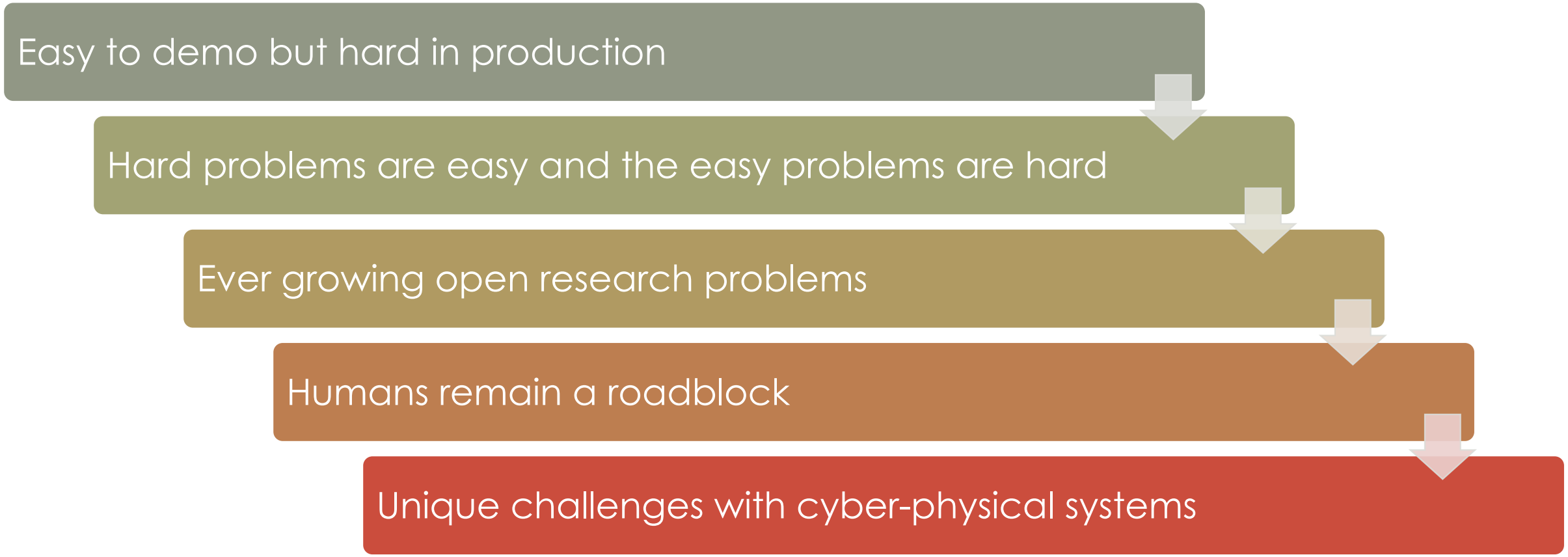
OCTOBER 30, 2023

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

BRIEFING ROOM › PRESIDENTIAL ACTIONS

By the authority vested in me as President by the Constitution and the laws of the United States of America, it is hereby ordered as follows:

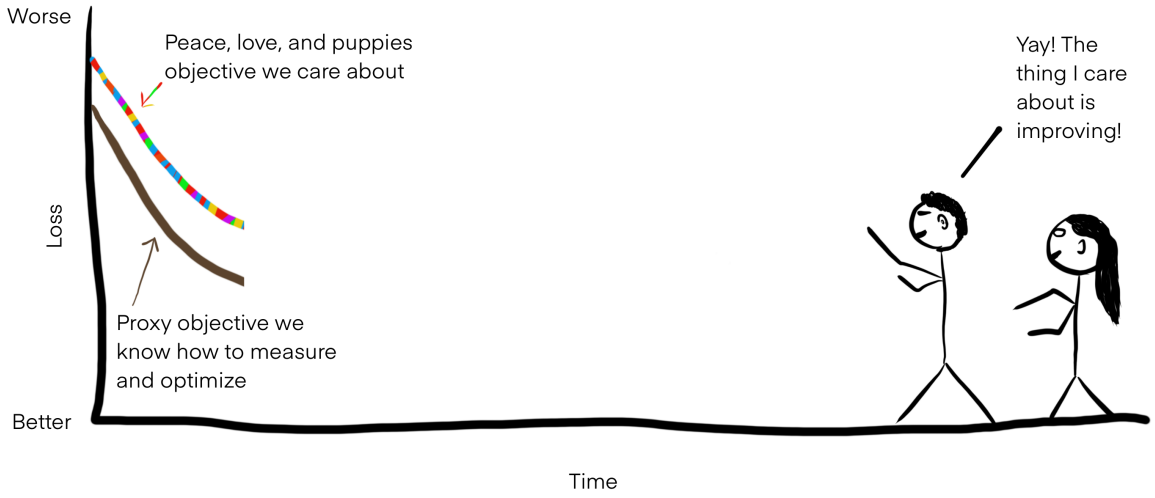
Paradox of AI development and challenges



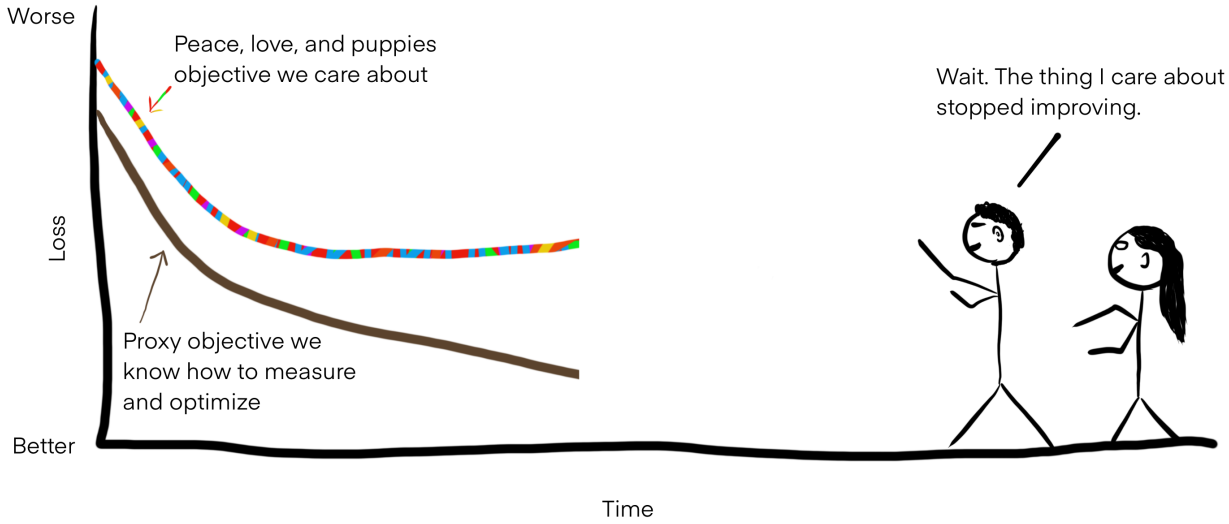
Paradox of efficiency

Goodhart's law: Too much efficiency makes everything worse

Well-aligned phase



Overfitting / Goodhart's law



https://sohl-dickstein.github.io/2022/11/06/strong-Goodhart.html

Alignment

Alignment: Ensuring that AI systems' goals and behaviors align with science and human values and intentions

Importance: Prevent potential harmful consequences of AI actions that could result from misalignment

Challenges: Defining human values, transferring these values to AI, and allowing for value learning and adaptation over time

Continuous Effort: Continuous effort as AI evolves and as societal values change

Driving safely on the road to AI implementation: Guardrails for responsible AI use

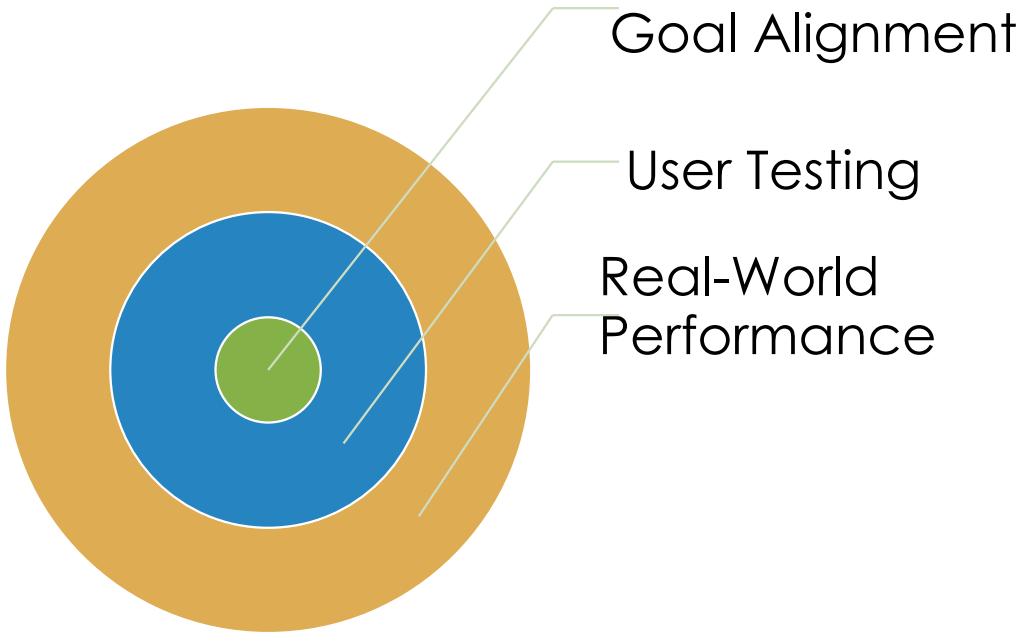
Destination (Objective): Effective Decision Making, Predictive Analysis, Automated Operations, and Improved Efficiency

Obstacles (Challenges): Bias, Misuse, Lack of Understanding, Complexity

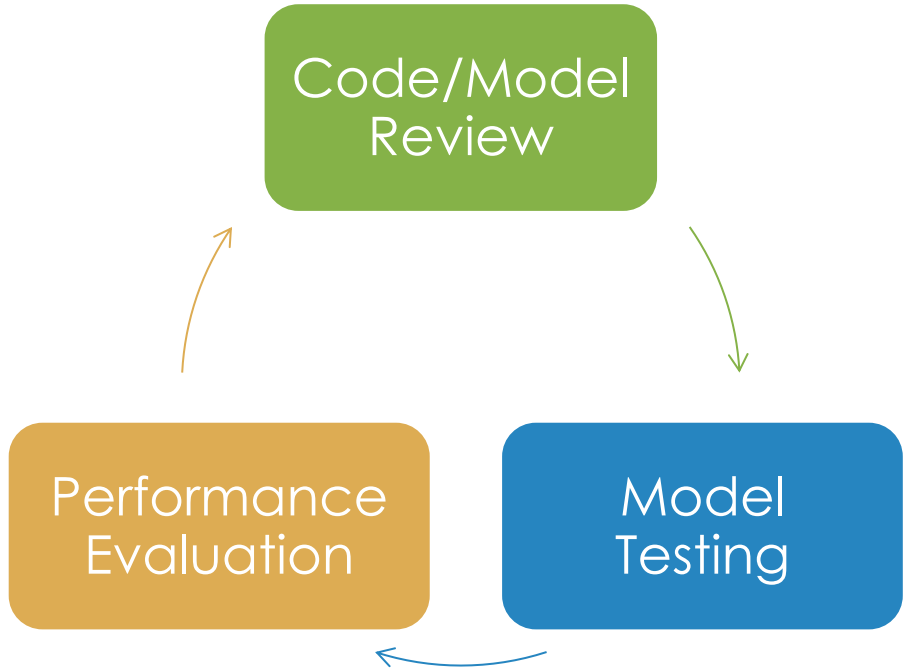
Guardrails (Safety measures): Ethics, Transparency, Privacy, Fairness, Security

Quality assurance in AI: Ensuring we're not only building the AI product right but also building the right AI product

Validation: Building the Right AI Product

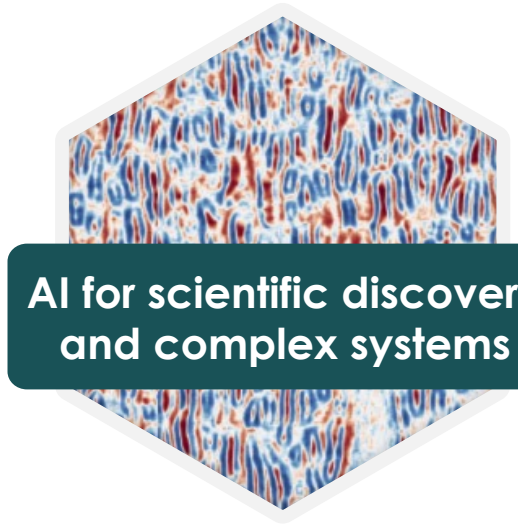


Verification: Building the AI Product Right



ORNL's AI initiative

Secure, trustworthy, and energy-efficient AI



AI for scientific discovery and complex systems

AI for experimental facilities

AI for national security

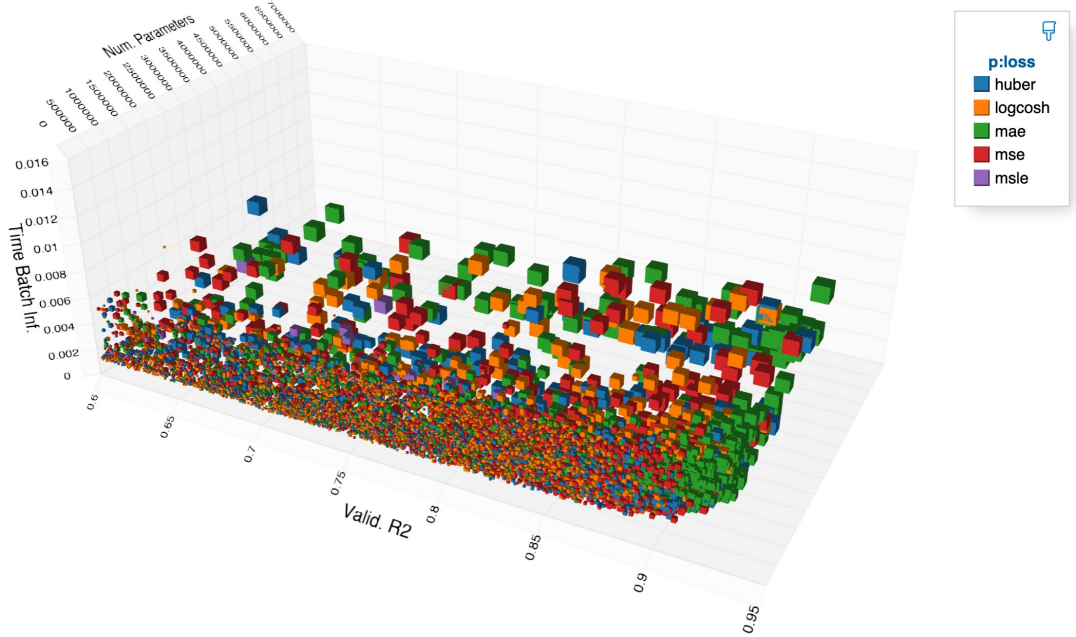
Secure	Trustworthy	Energy efficient
Alignment	Validation and verification	Scalability
Cybersecurity	Uncertainty qualifications	Edge
Robustness	Casual reasoning	Co-design

Safe and secure AI: Goal and behavior alignment with science, human values, and intentions

Considerations

- Accuracy
- Fairness
- Privacy
- Transparency
- Robustness
- Energy-efficiency

Neural architecture search on Frontier



CAISER – **C**enter for **AI SE**curity **R**esearch

National center of excellence with strong leadership:

Leading-edge NS programs

AI initiative

Computing excellence

Computing resources

Safeguarding AI systems against threats

- Safeguarding AI data and models from unauthorized access
- Understanding and addressing data and model poisoning
- Consistent monitoring and auditing of AI operations and frameworks
- Establishment of mitigation strategies (Secure data management and robust training methodologies)

Assurance: Reliable, Robust, and Safe AI

Assurance

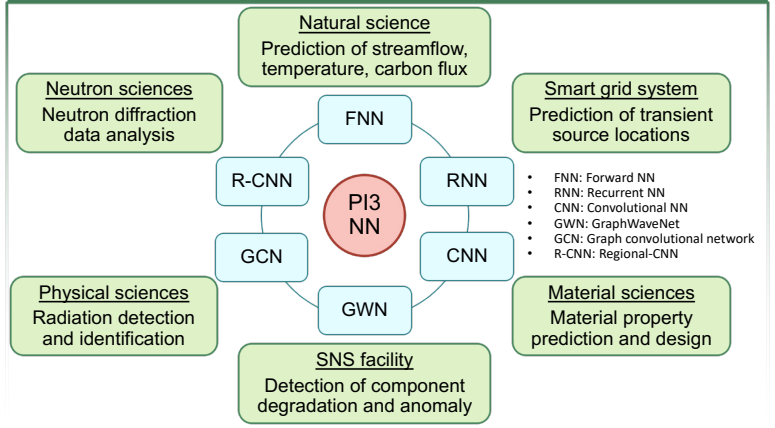
Uncertainty Quantification (UQ)

Verification & Validation (V&V)

Explainability & Interpretability

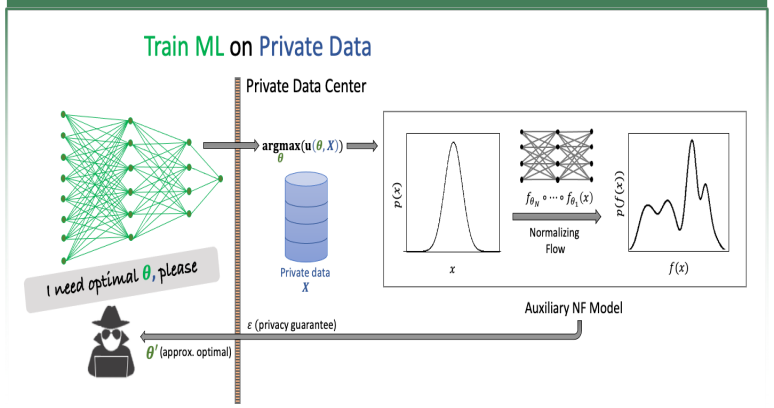
Privacy

Uncertainty Quantification for Trustworthy AI



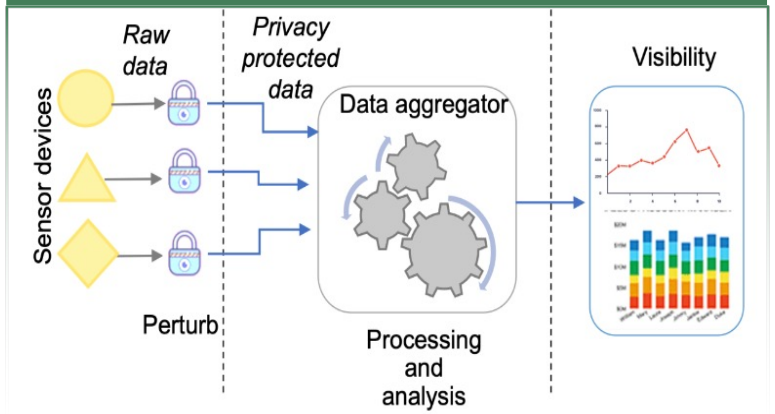
reliable and scalable uncertainty quantification methods for DOE mission area

Privacy-Preserving Model Training



train and release ML models on a private dataset with a formal privacy guarantee

Privacy-Preservation at Edge

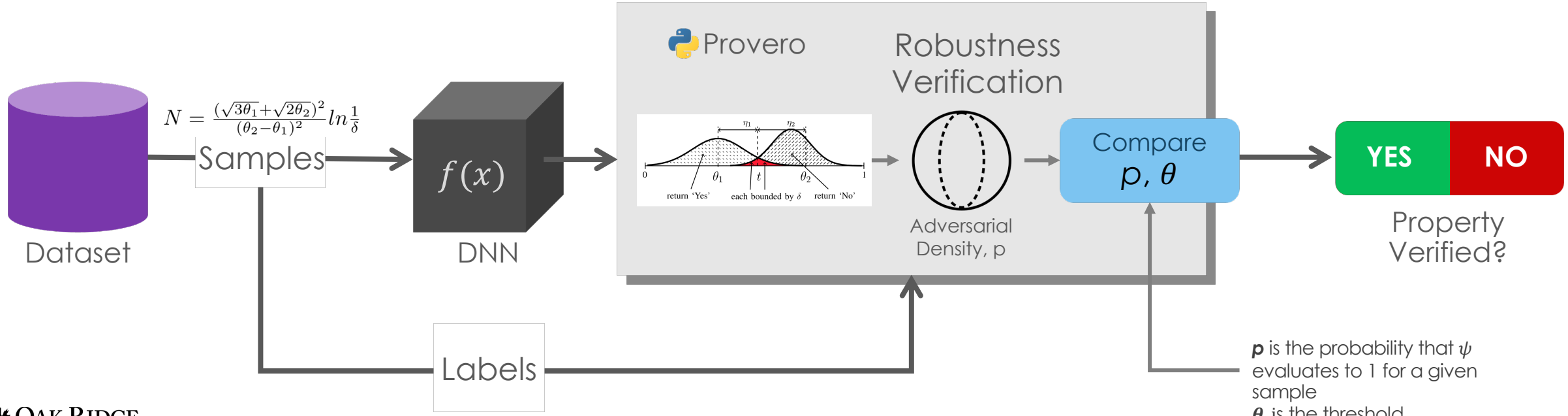


automatic privacy-preservation of streaming data on edge such as smart grid

OAK RIDGE National Laboratory

Validation and verification

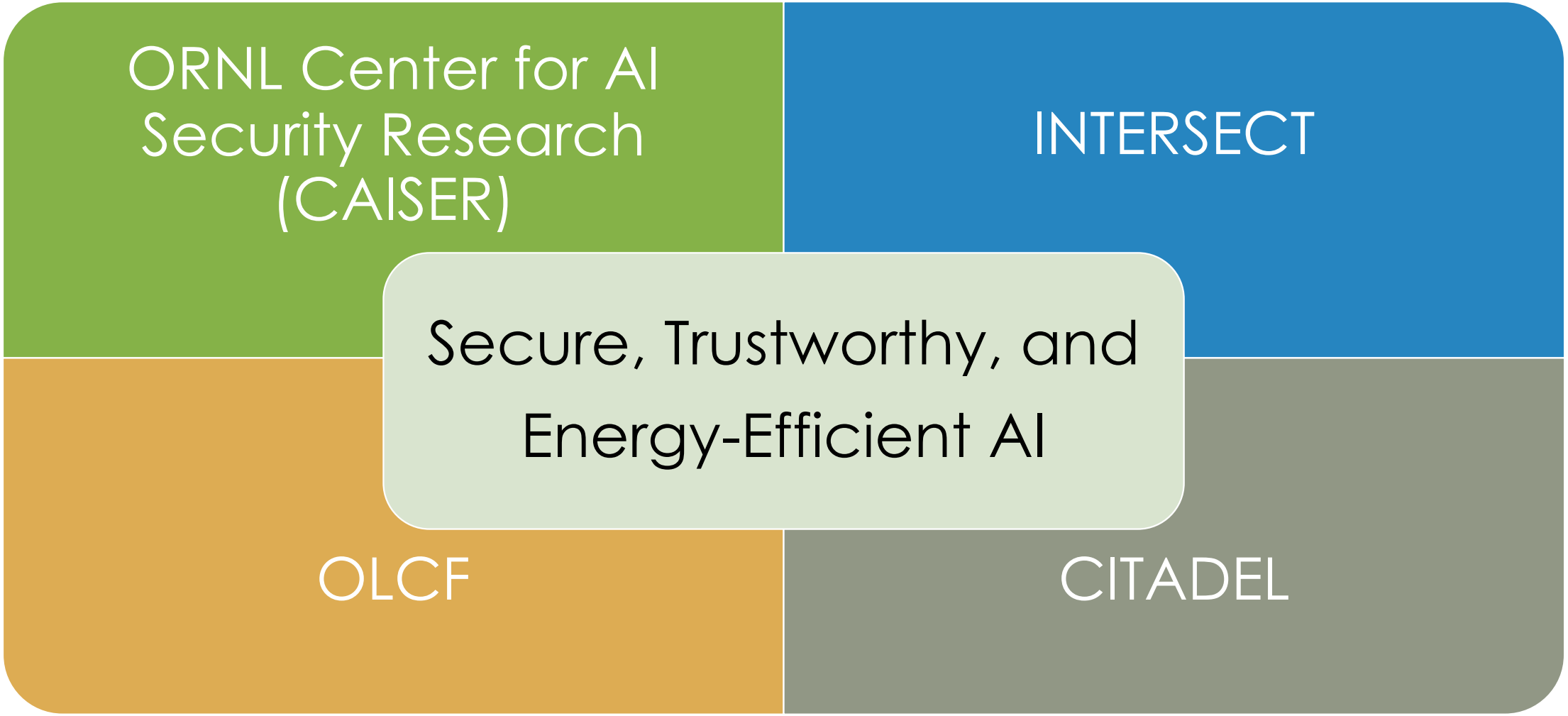
- Sampling-based approach to quantitatively estimate properties for deep neural networks (DNN) with probabilistic guarantees
  - Given a logical property ψ specified over a space of inputs and outputs of a DNN and a numerical threshold θ, decide whether ψ is true for less than θ fraction of the inputs
  - Assumes only *black box* access
  - Provides quantitative verification of properties like fairness, privacy, and robustness
  - Verification is *sound* – when ψ is confirmed to be true, it can be deduced mathematically



ORNL's AI initiative

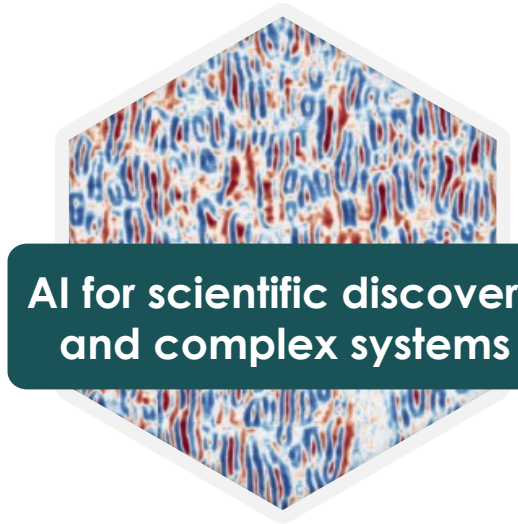
Secure, trustworthy, and energy-efficient AI

The AI Initiative leverage and enhance ORNL's existing facilities and capabilities



ORNL's AI initiative

Secure, trustworthy, and energy-efficient AI



AI for scientific discovery and complex systems

AI for experimental facilities

AI for national security

Secure	Trustworthy	Energy efficient
Alignment	Validation and verification	Scalability
Cybersecurity	Uncertainty qualifications	Edge
Robustness	Casual reasoning	Co-design