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

Solar Applications of Artificial Intelligence and Machine Learning Workshop

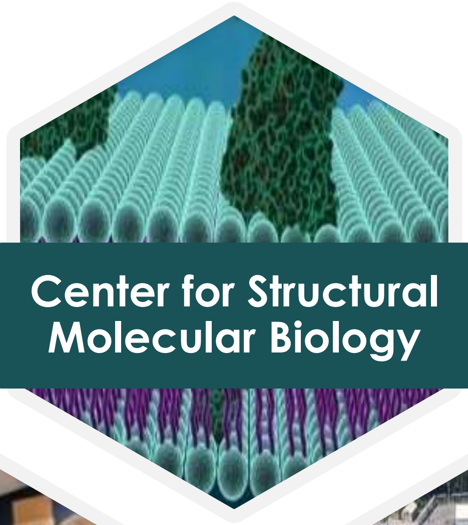
October 31, 2023

DOE and ORNL mission AI applications

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

Spallation Neutron Source

Manufacturing Demonstration Facility



Center for Structural Molecular Biology

Oak Ridge Leadership Computing Facility

Cyber Science Research Facility

High Flux Isotope Reactor

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

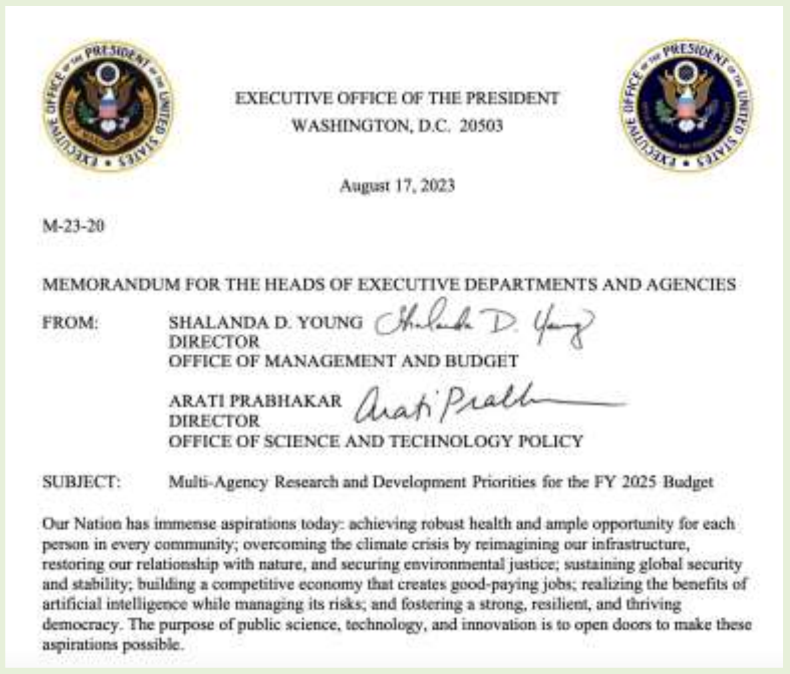
A new *Nature* series will explore the many ways in which artificial intelligence is changing science – for better and for worse.

SCIENCE & TECH

The Pentagon's AI Chief Is 'Scared to Death' of ChatGPT

But other defense leaders are more eager to deploy new artificial-intelligence tools.

A Computer Scientist Breaks Down Generative AI's Hefty Carbon Footprint


EXECUTIVE OFFICE OF THE PRESIDENT
WASHINGTON, D.C. 20503

August 17, 2023

M-23-20

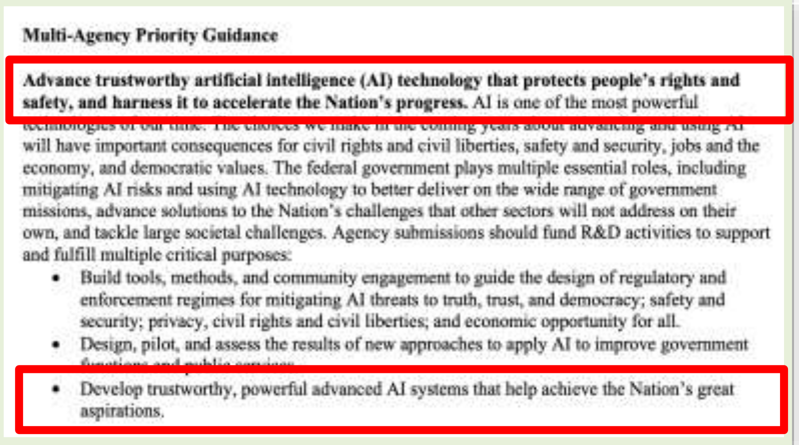
MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: SHALANDA D. YOUNG
DIRECTOR
OFFICE OF MANAGEMENT AND BUDGET

ARATI PRABHAKAR
DIRECTOR
OFFICE OF SCIENCE AND TECHNOLOGY POLICY

SUBJECT: Multi-Agency Research and Development Priorities for the FY 2025 Budget

Our Nation has immense aspirations today: achieving robust health and ample opportunity for each person in every community; overcoming the climate crisis by reimagining our infrastructure, restoring our relationship with nature, and securing environmental justice; sustaining global security and stability; building a competitive economy that creates good-paying jobs; realizing the benefits of artificial intelligence while managing its risks; and fostering a strong, resilient, and thriving democracy. The purpose of public science, technology, and innovation is to open doors to make these aspirations possible.


Multi-Agency Priority Guidance

Advance trustworthy artificial intelligence (AI) technology that protects people's rights and safety, and harness it to accelerate the Nation's progress. AI is one of the most powerful technologies of our time. The choices we make in the coming years about advancing and using AI will have important consequences for civil rights and civil liberties, safety and security, jobs and the economy, and democratic values. The federal government plays multiple essential roles, including mitigating AI risks and using AI technology to better deliver on the wide range of government missions, advance solutions to the Nation's challenges that other sectors will not address on their own, and tackle large societal challenges. Agency submissions should fund R&D activities to support and fulfill multiple critical purposes:

- Build tools, methods, and community engagement to guide the design of regulatory and enforcement regimes for mitigating AI threats to truth, trust, and democracy; safety and security; privacy, civil rights and civil liberties; and economic opportunity for all.
- Design, pilot, and assess the results of new approaches to apply AI to improve government functions and public services.
- Develop trustworthy, powerful advanced AI systems that help achieve the Nation's great aspirations.


POLITICO

EXCLUSIVE

Sweeping new Biden order aims to alter the AI landscape

The White House is poised to make an all-hands effort to impose national rules on a fast-moving technology, according to a draft executive order.

President Joe Biden's new guidelines will give federal agencies influence in the US market through their buying power and their enforcement tools. | Jacquelyn Martin/AP

By MOHAR CHATTERJEE and REBECCA KERN
10/27/2023 07:27 PM EDT
Updated: 10/27/2023 08:04 PM EDT

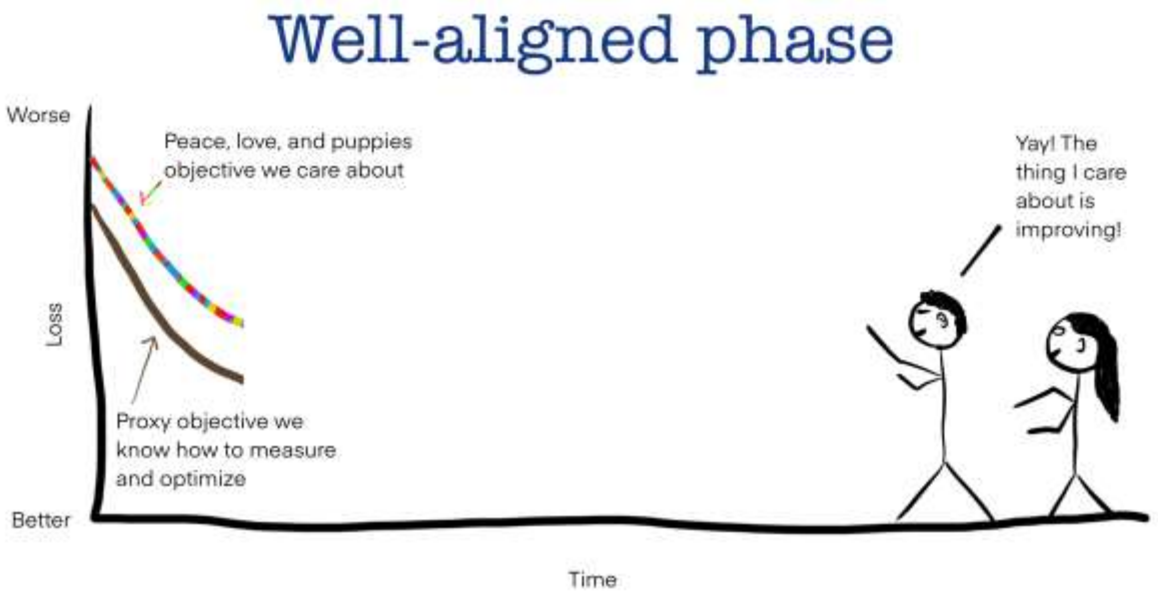
President Joe Biden will deploy numerous federal agencies to monitor the risks of artificial intelligence and develop new uses for the technology while attempting to protect workers, according to a draft executive order obtained by POLITICO.

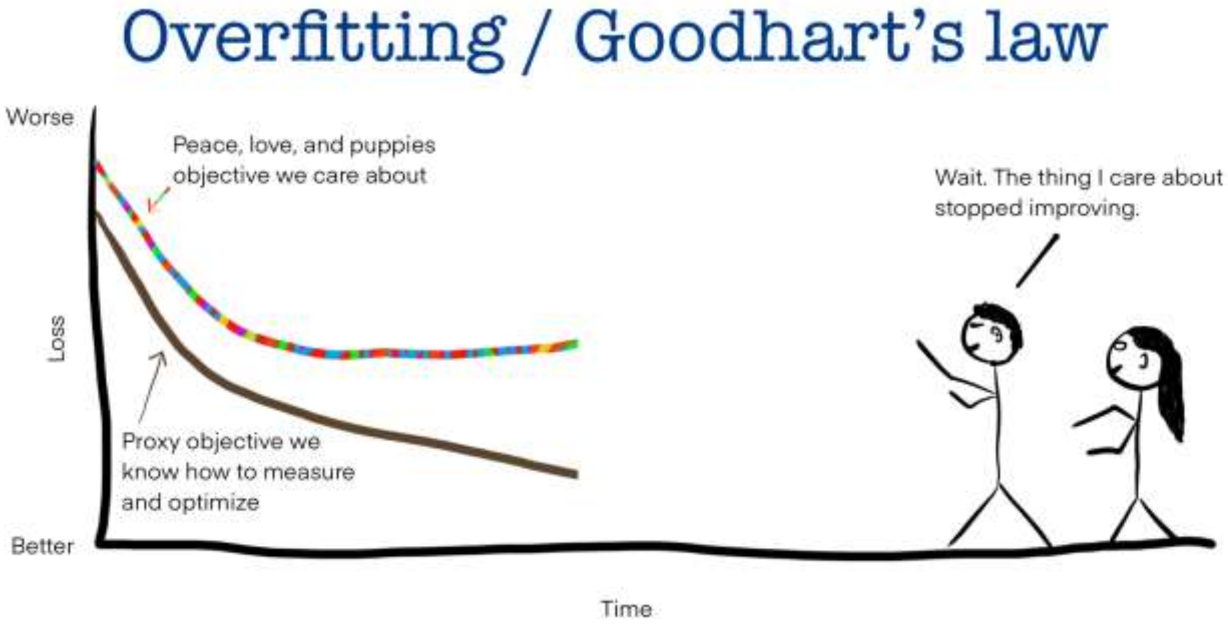
Paradox of AI development and challenges

- Easy to demo but hard in production
- Hard problems are easy and the easy problems are hard
- Ever growing open research problems
- Humans remain a roadblock
- Unique challenges with cyber-physical systems

Paradox of efficiency: Goodhart's law

Too much efficiency makes everything worse





Oak Ridge National Laboratory

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

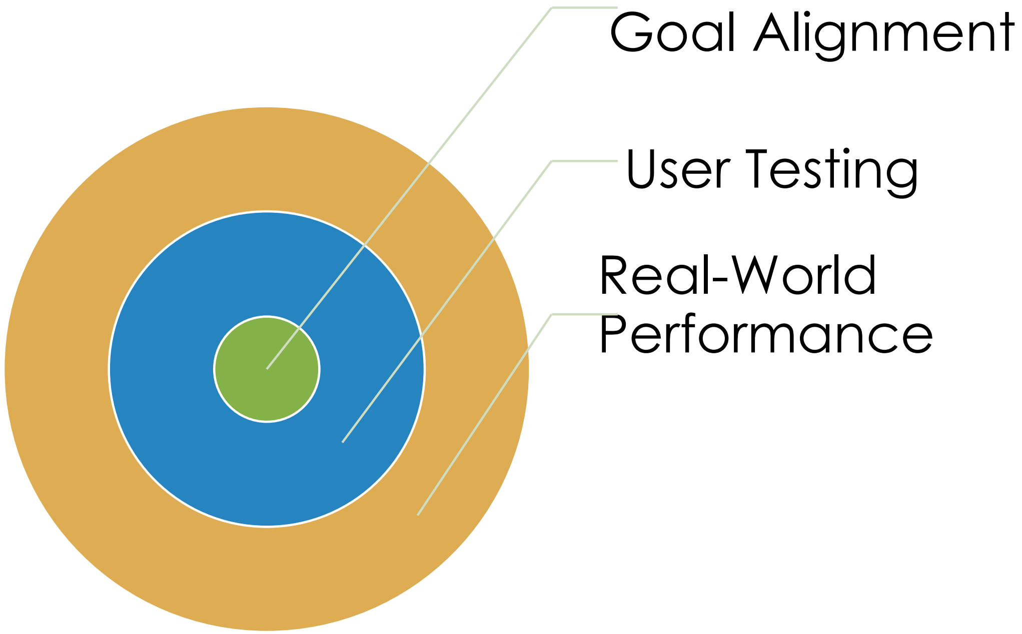
Destination (Objective): Effective Decision Making, Predictive Analysis, Automated Operations, and Improved Efficiency

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

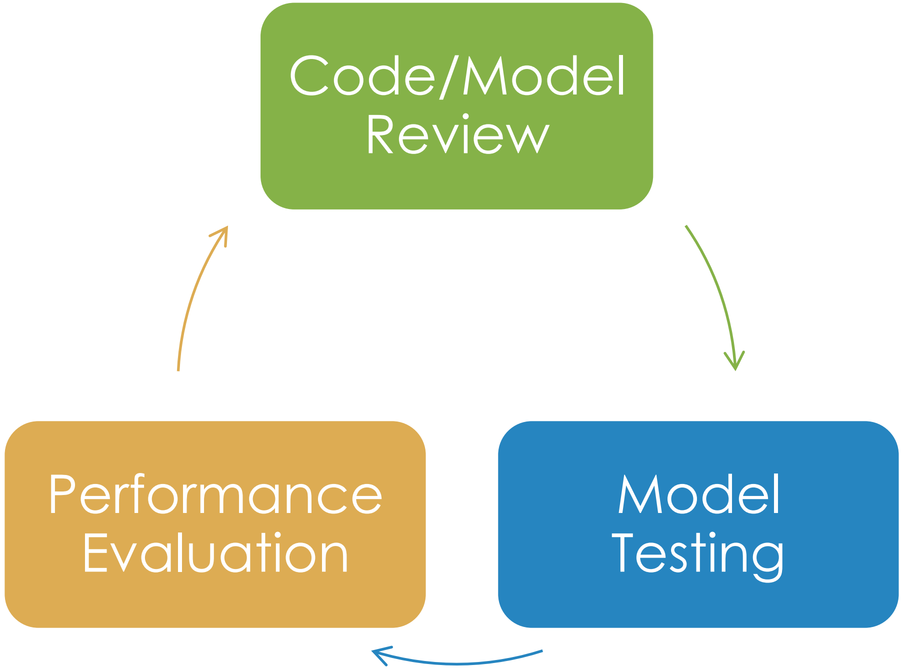
Guardrails (Safety measures): Ethics, Transparency, Privacy, Fairness, Security

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



Oak Ridge National Laboratory

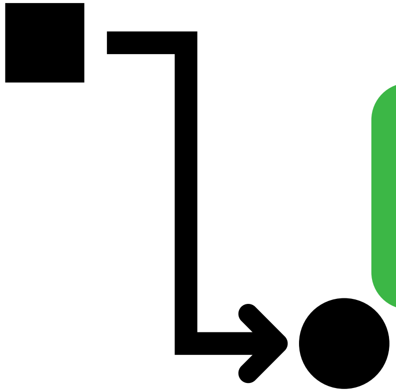
Moving beyond correlations: Causal modeling for predictive and explanatory power

uncover why something happens, not just what happens next

Causal Variables

Data collection

Validation



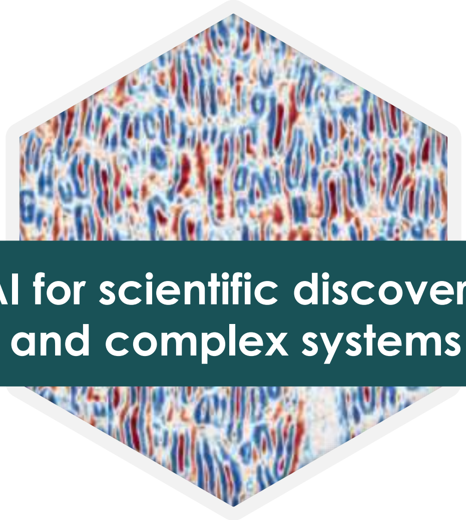
Model specification and training

Interpretation

Causal Structure

ORNL’s AI initiative

Secure, trustworthy, and energy-efficient AI



AI for scientific discovery and complex systems

AI for experimental facilities

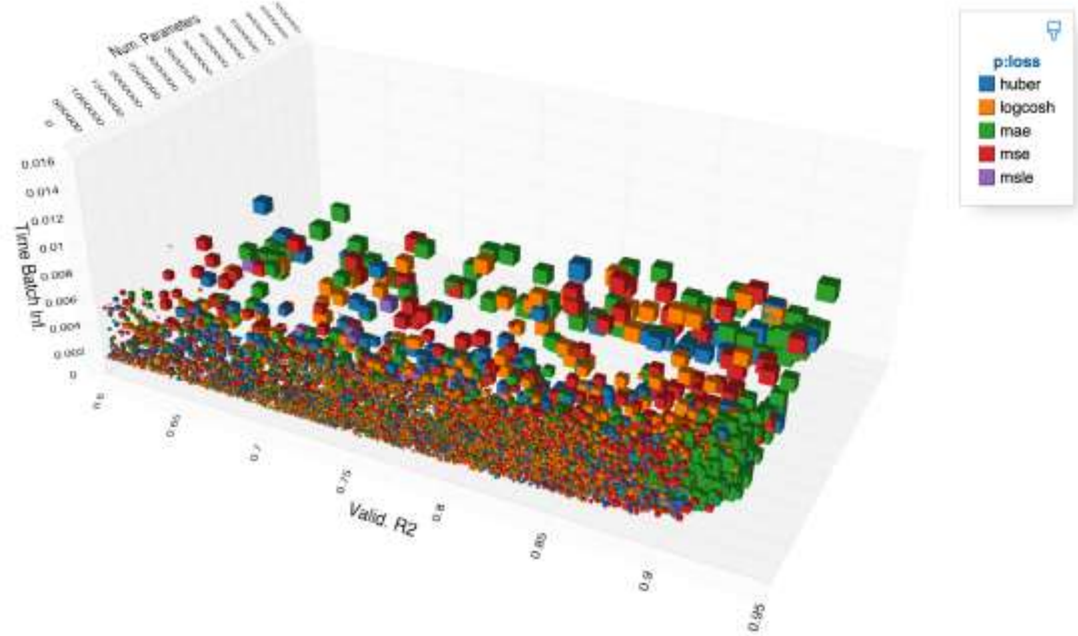
AI for national security

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

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

Considerations

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



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

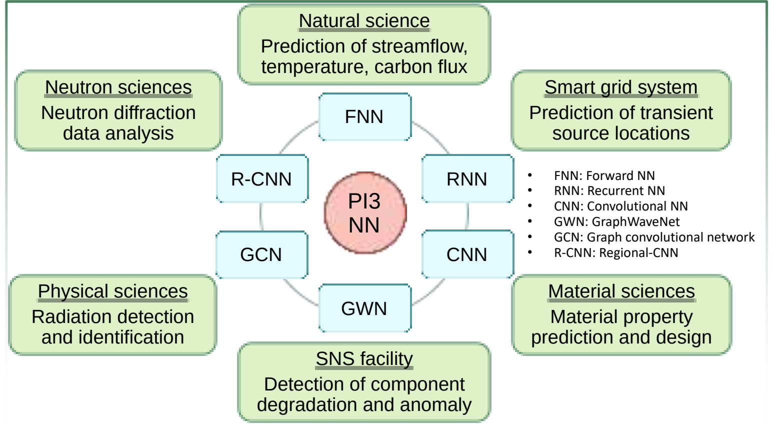
Uncertainty Quantification (UQ)

Verification & Validation (V&V)

Explainability & Interpretability

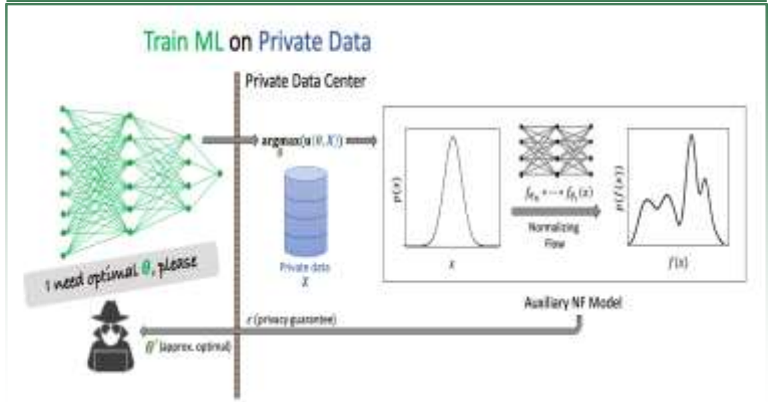
Privacy

Uncertainty Quantification for Trustworthy AI



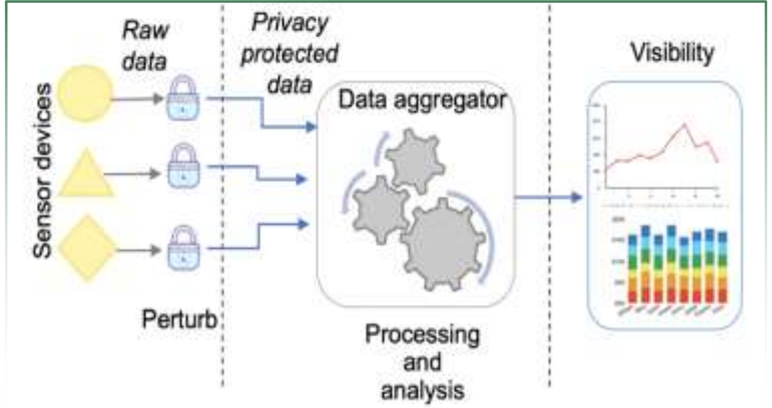
reliable and scalable uncertainty quantification methods for DOE mission area

Privacy-Preserving Model Training



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

Privacy-Preservation at Edge

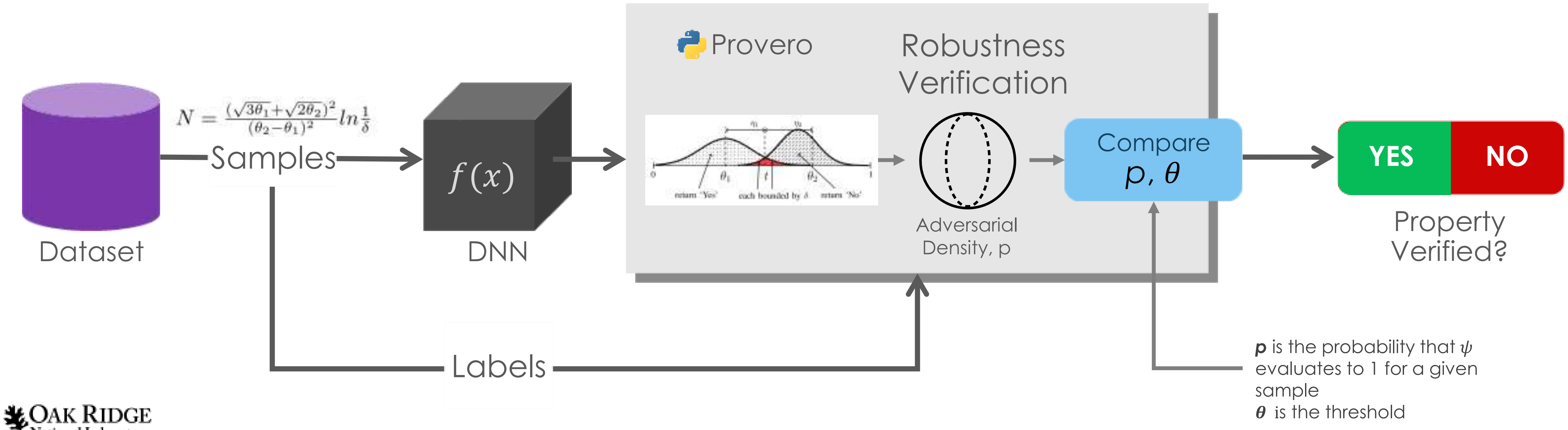


automatic privacy-preservation of streaming data on edge such as smart grid

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 Center for AI Security Research (CAISER)

INTERSECT

Secure, Trustworthy, and Energy-Efficient AI

OLCF

CITADEL

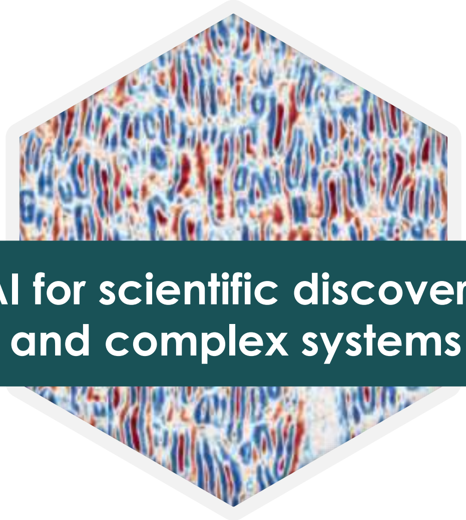
AI workshop series

ORNL's Generative AI Workshop Series:
2nd Workshop Towards Safe, Trustworthy, and Energy-Efficient AI Models

Co-located with the Smoky Mountain Conference 2023
Tuesday, August 29th 2023
Crowne Plaza Hotel, Knoxville Downtown, TN
Hybrid event

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