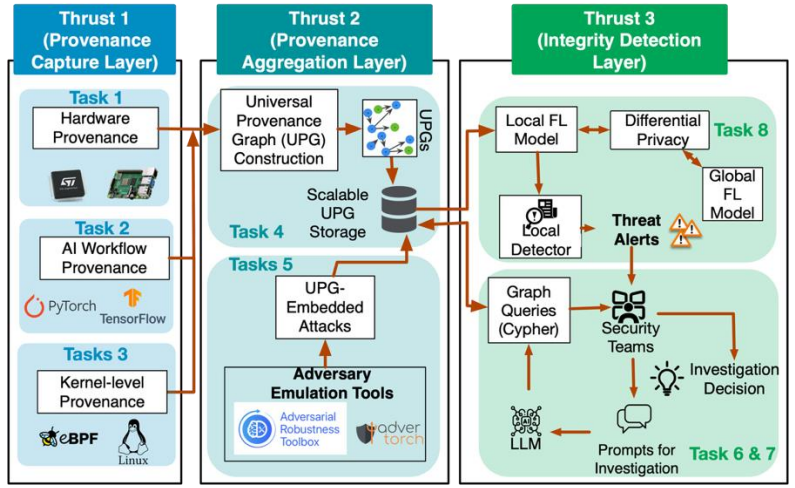
MLDL: Multi-Layer Data Provenance and Federated Learning for Securing Scientific AI Pipelines

PI: Wajih Ul Hassan (University of Virginia) Co-Pi: Aidong Zhang (University of Virginia)

- Build an end-to-end provenance infrastructure that tracks the full dataset lifecycle, enabling transparency, accountability, and trustworthy, reproducible AI-driven science



Need For Integrity, Provenance, and Authenticity (IPA)

- Multi-source datasets lack end-to-end traceability, reducing trust, reproducibility, and collaboration

Approach For Achieving IPA in AI Datasets

- Capture detailed data provenance across hardware, operating system, and application layers
- Unify this provenance into a scalable provenance graph with privacy-preserving federated anomaly detection

Benefits to Scientific Cyberinfrastructure

- Who cares: domain scientists (genomics, imaging, climate); HPC/data platform operators; security and incident response teams;
- Faster detection of tampering, pipeline drift, and integrity violations

Risks Versus Potential For Advances

- Gaps in hardware visibility, provenance graph related overheads
- A new standard for verifiable AI data lifecycles that improves trust in scientific results across academia, healthcare, and government

Evaluating and Demonstrating IPA

- Metrics of success: end-to-end traceability from hardware to model; low capture overhead; fast load and query; accurate and timely anomaly alerts; privacy preserved across sites
- Community access: open-source code and containers; documented APIs; starter datasets and reproducible scripts; tutorials and workshops

Programmatic Details

- 3 years project, starts on 01/01/2026
- Led by the University of Virginia

CICI IPAAI

This material is based upon work supported by the National Science Foundation under Grant No. 2530655