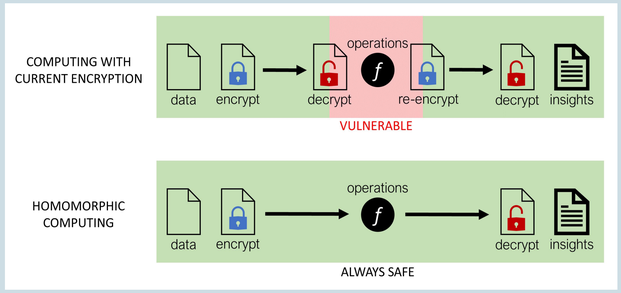
SAND2022-8007C

Sandia National Laboratories

Private yet Explainable Artificial Intelligence (PyE-AI)

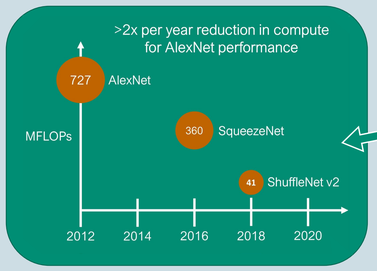
Alycia Carey, Mitch Negus, Nick Pattengale (PI), Jon Roose, Mike Smith

Motivation: Emerging privacy enhancing technologies (PET) *change the game* for high consequence use cases.

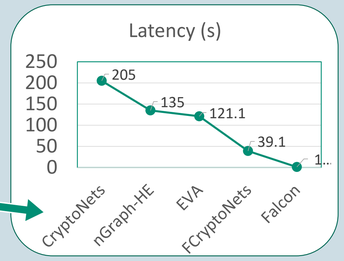


XAI *Explainable Artificial Intelligence*
DLIME: Deterministic Local Interpretable Model-agnostic Explanations
OOD: Out-of-distribution Detection

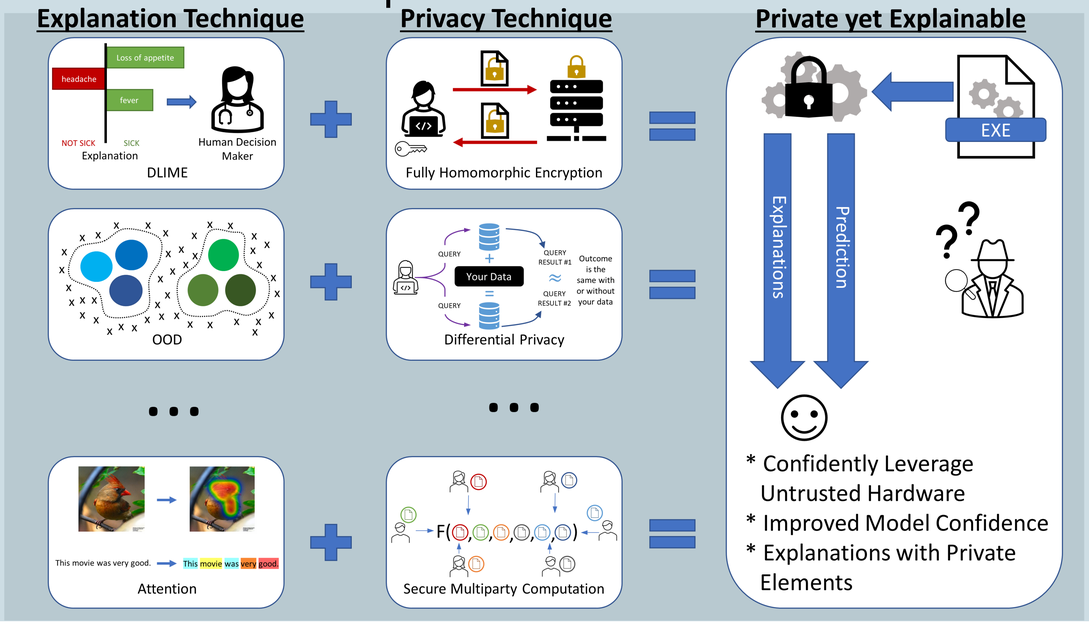
PET *Privacy Enhancing Technologies*
FHE: Fully homomorphic encryption
DP: Differential privacy



Instances of dramatic progress in ML/AI + PET
- *left*: 17x improvement in model performance that achieves AlexNet performance
- *right*: 200x improvement in model performance on MNIST, via TinyML and FHE



Technical Approach: Combine PET with AI/ML models that are explainable or that build trust.



Results Summary: Two prototypes under study

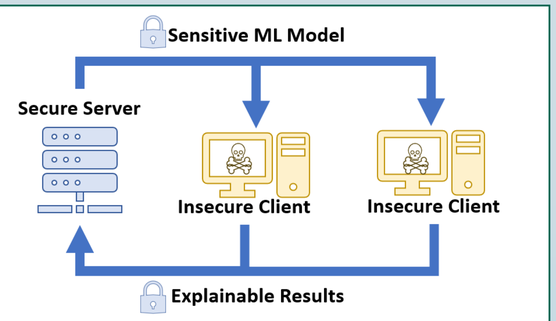
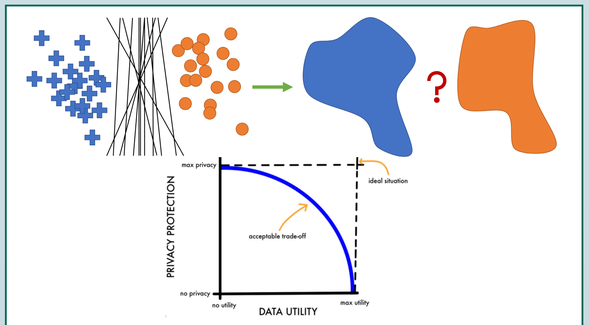


Figure 1: System Level Diagram of interactions for FHE version of ML Model + DLIME applied to malware classification.

We have innovated and prototyped *the first* explanatory AI algorithm executed under FHE, including results retrieval that resists info leakage



We have innovated and prototyped *the first* DP OOD preserving utility using:
1. Random projections (JL-lemma)
2. Ensembling techniques
3. Reducing amount of noise required

Impacts & Successes to date

- Two publications in prep (USENIX Security 2022 and IEEE S&P 2023)
- Launch of a (Sandia internal) Privacy Enhancing Technologies Working Group
- Transition Sponsor Interest and Engagement
  - Interested in the practicality of Differential Privacy, among other topics
- Inspired Additional Proposal, currently under consideration by committee
  - Proposed Thrust 1: **direct implementations**, which are additional explorations in the image of DLIME/FHE prototype, i.e. direct application of privacy techniques to ML explanation algorithms
  - Proposed Thrust 2: **trust**, which are additional explorations in the image of OOD/DP prototype, i.e. privacy preserving trust bases
  - Proposed Thrust 3: **fundamental limitations**, an exploration into the idea of a ML explanation that only selectively reveals information

1-year NSIST LDRD, roughly 1 FTE; a 2-year follow-on effort is under consideration by the NSIST committee

U.S. DEPARTMENT OF ENERGY
National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology &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.
SAND No. _____

Sandia National Laboratories

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & 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.