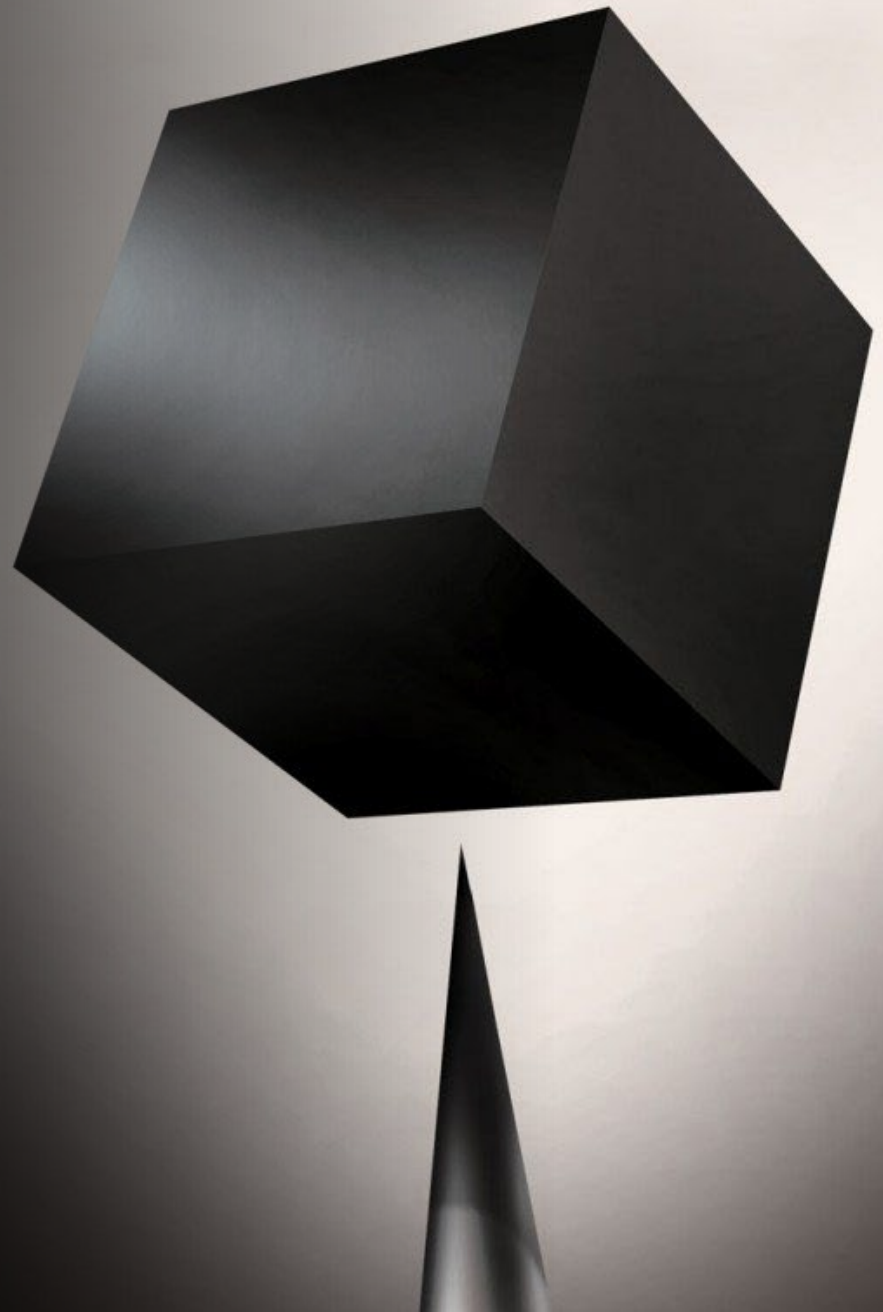
Ethical AI by Design



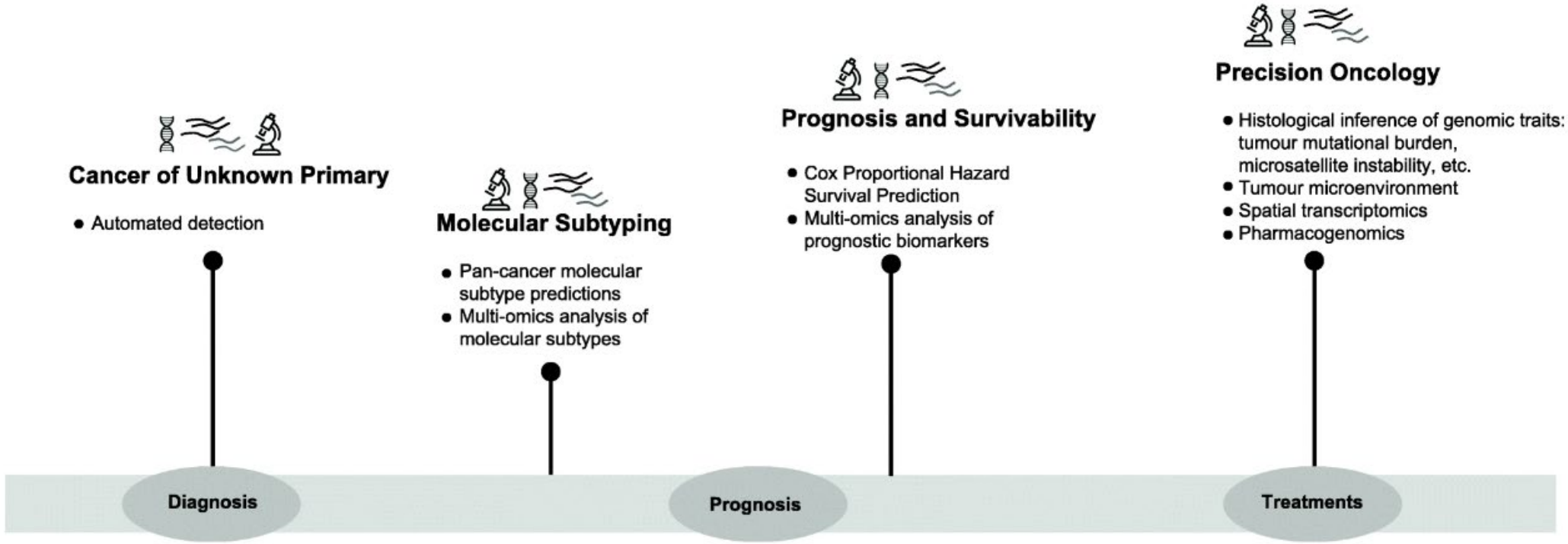
My T. Thai, Ph.D., FIEEE

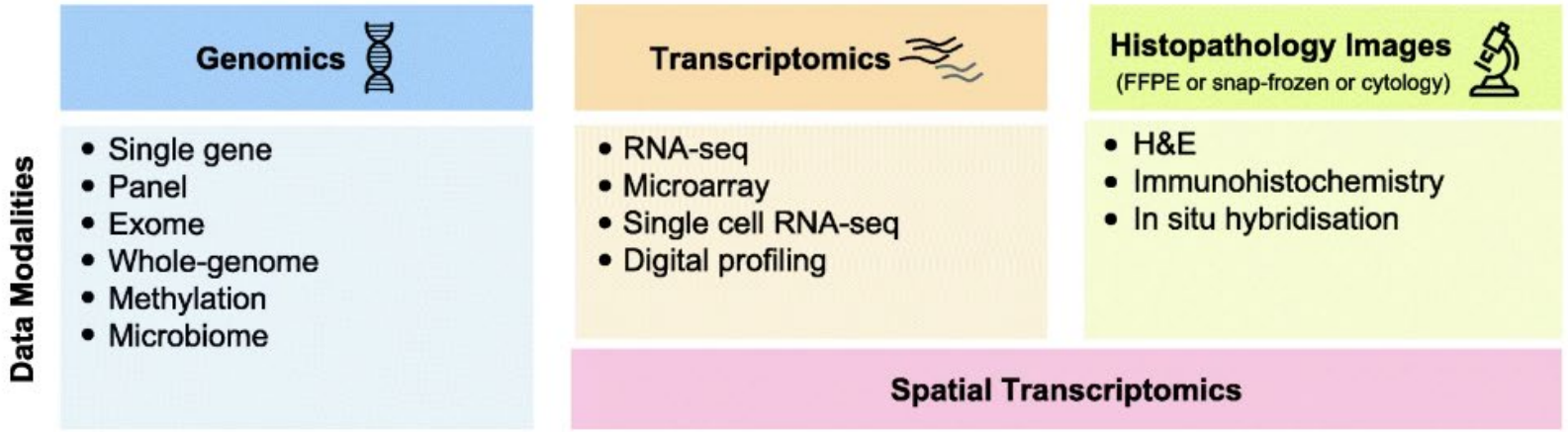
UF Research Foundation Professor, Computer & Information Science & Engineering

University of Florida

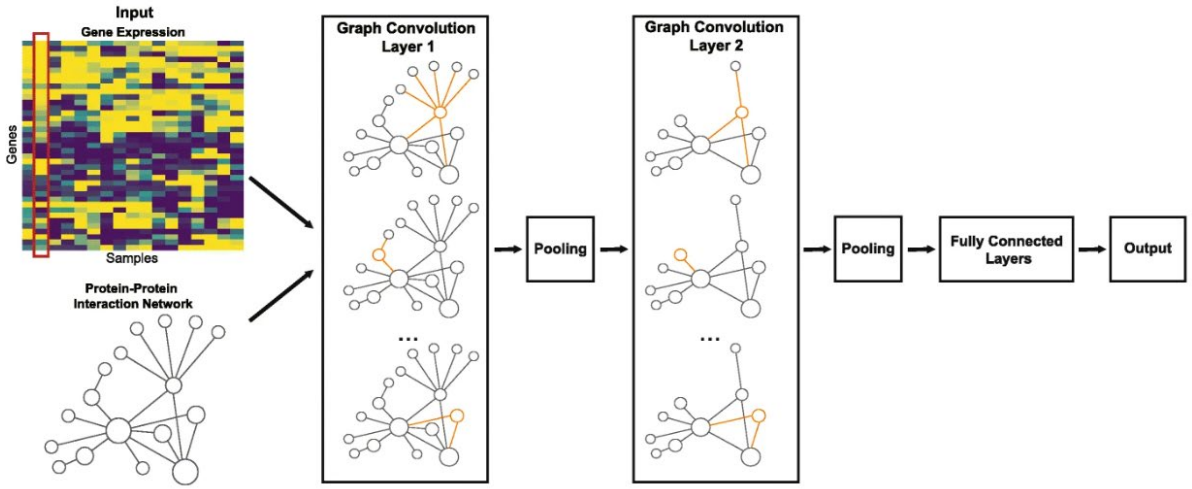
mythai@cise.ufl.edu

AI in cancer diagnosis, prognosis and treatment

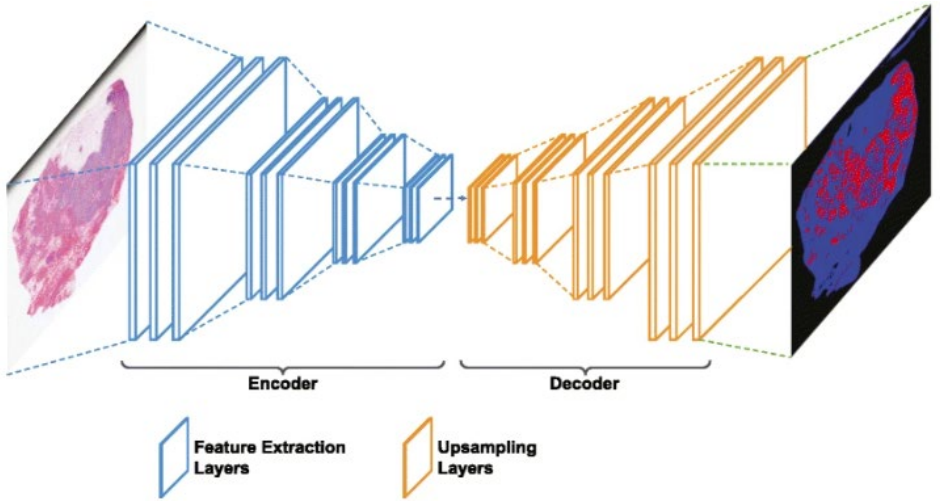




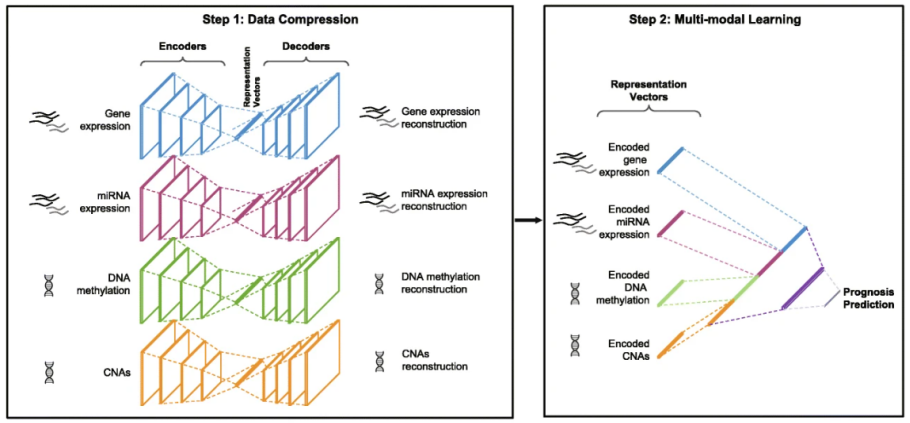
Recent trends of AI in oncology



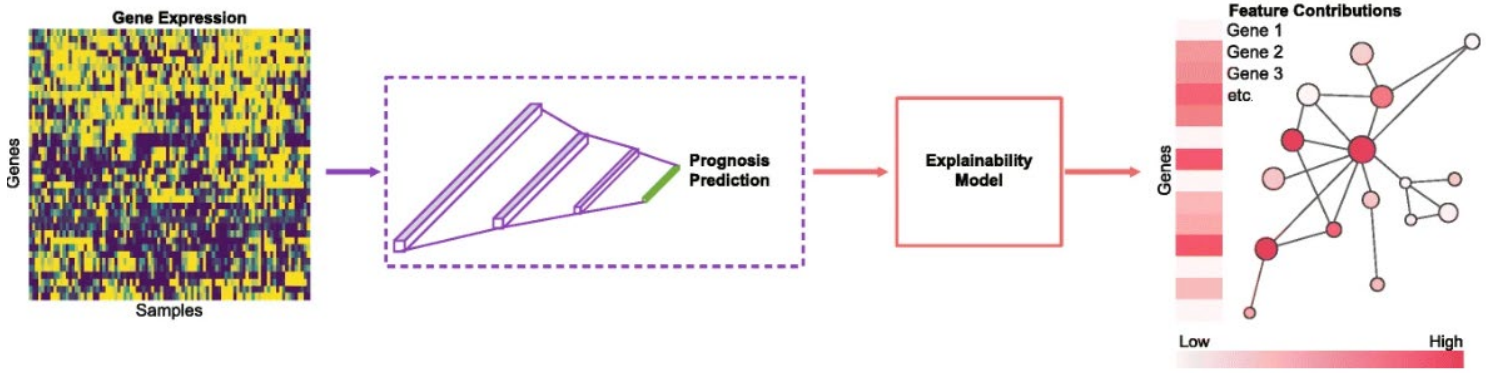
Advanced neural network methods



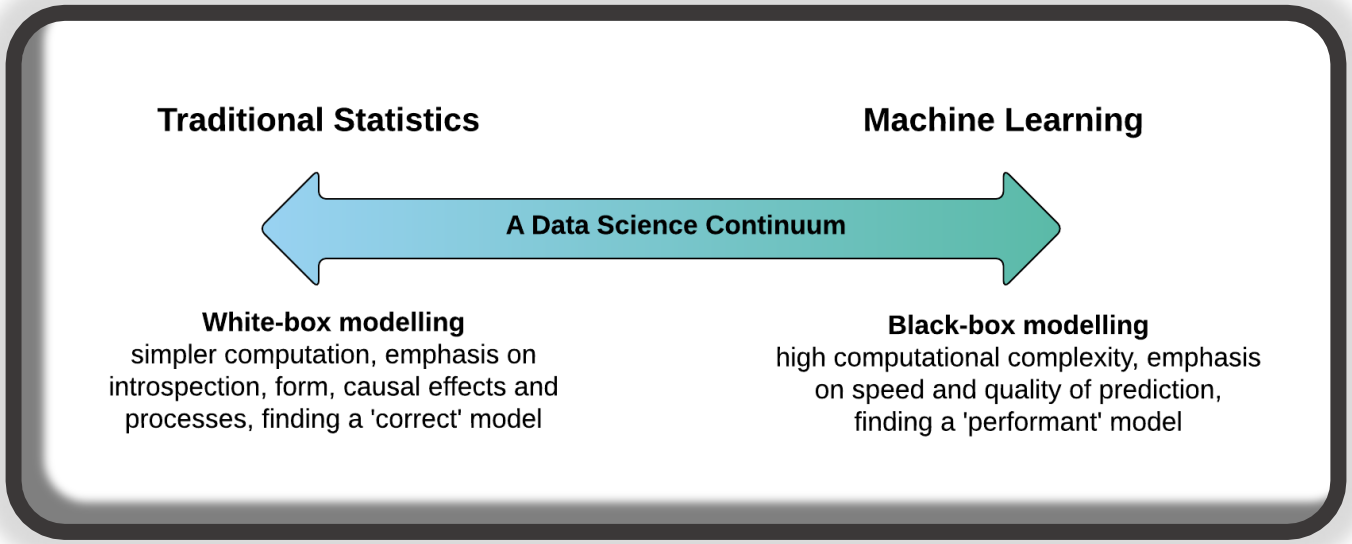
Semantic segmentation

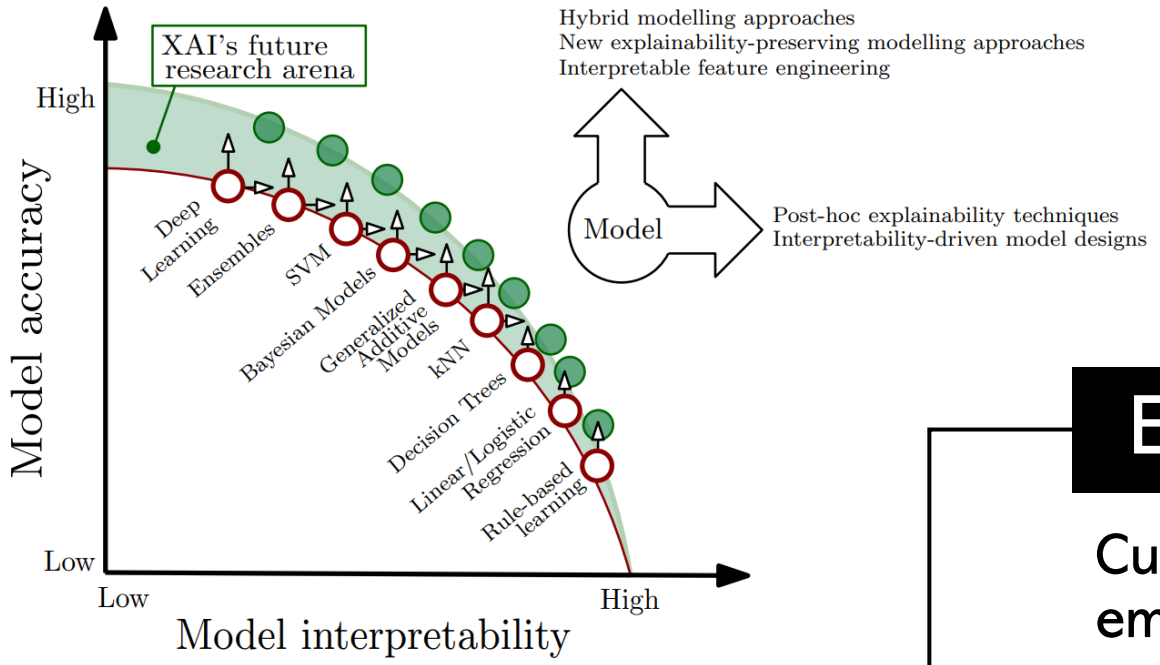


Multimodal learning



Explanation methods



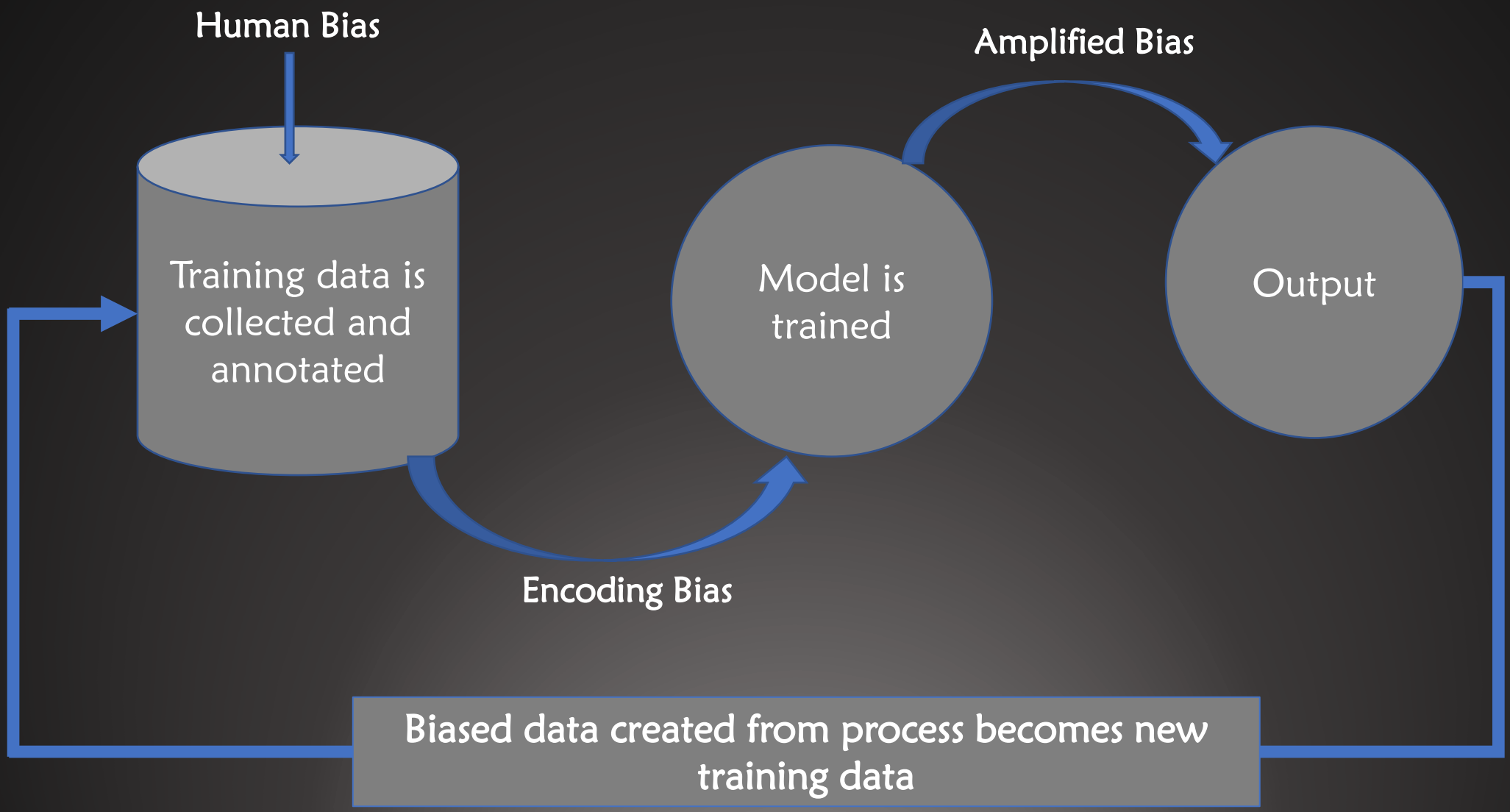


Any Shortcut Learning?

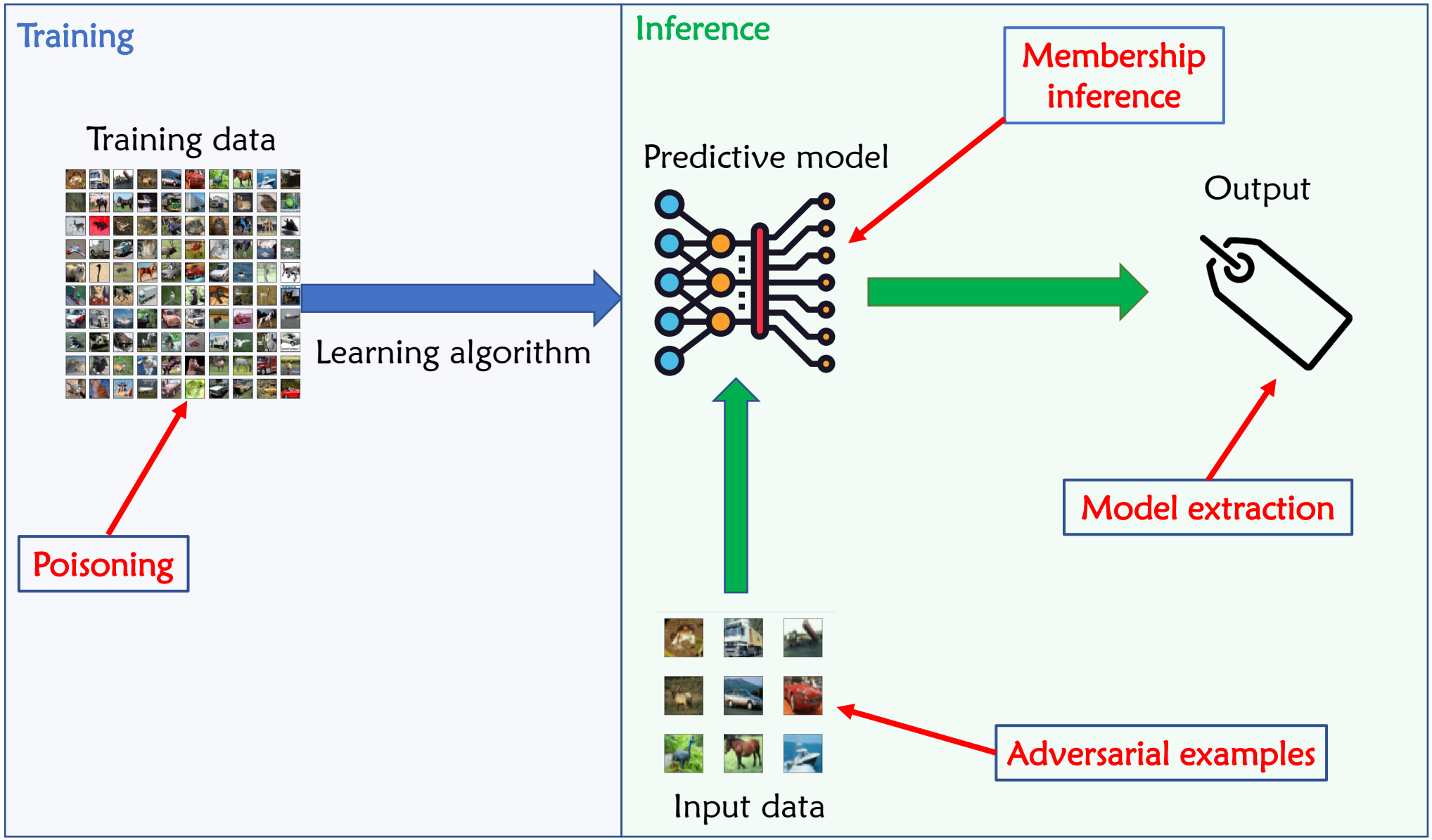
Assessing models for potential clinical deployment

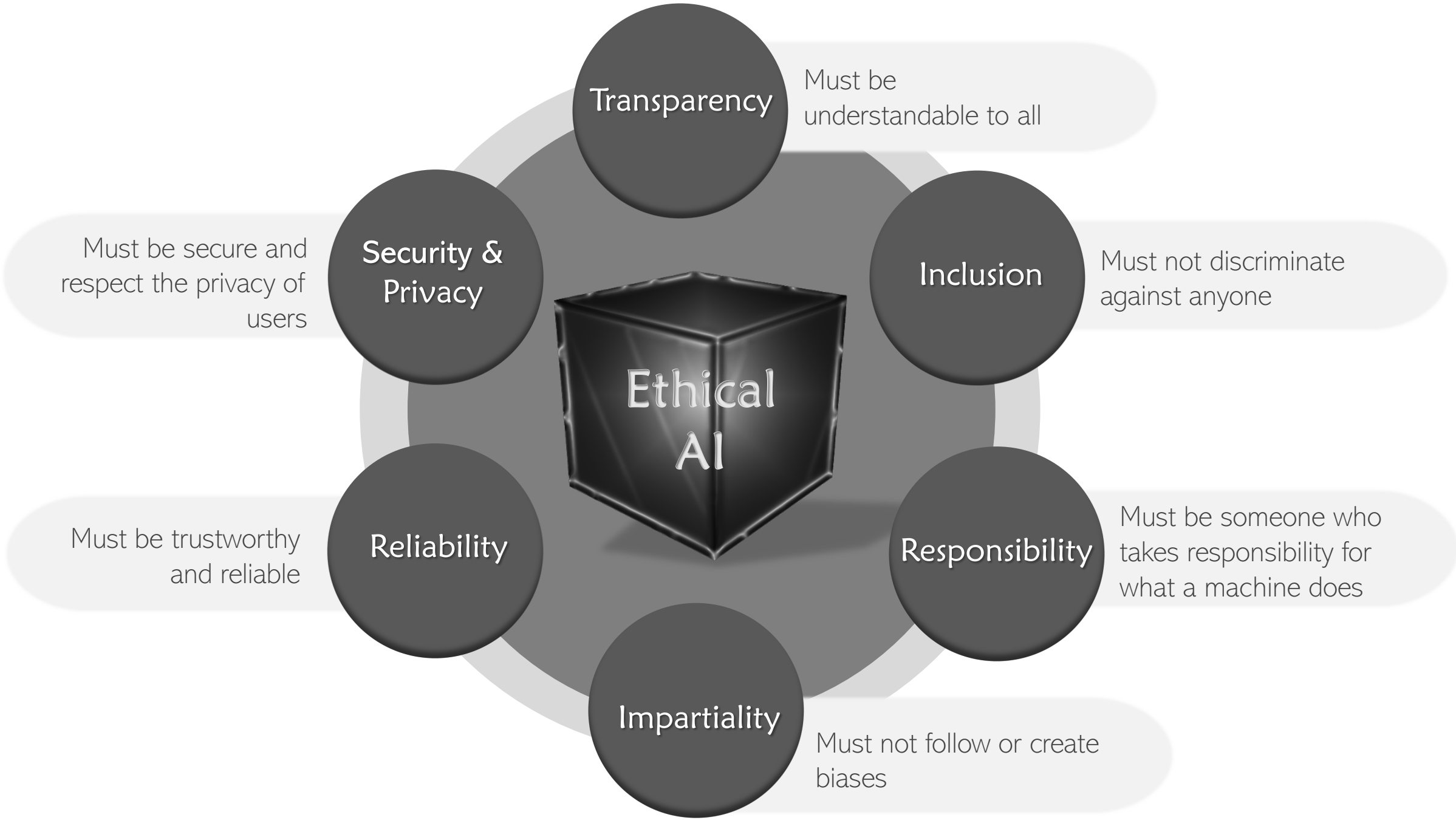
Empathy & AI coexist?

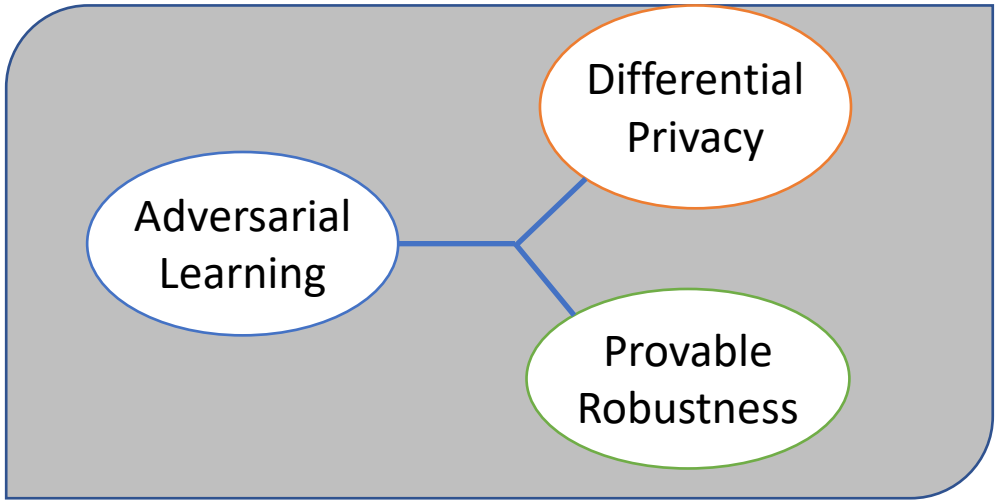
Customers today don't trust AI. They prefer empathetic interactions.

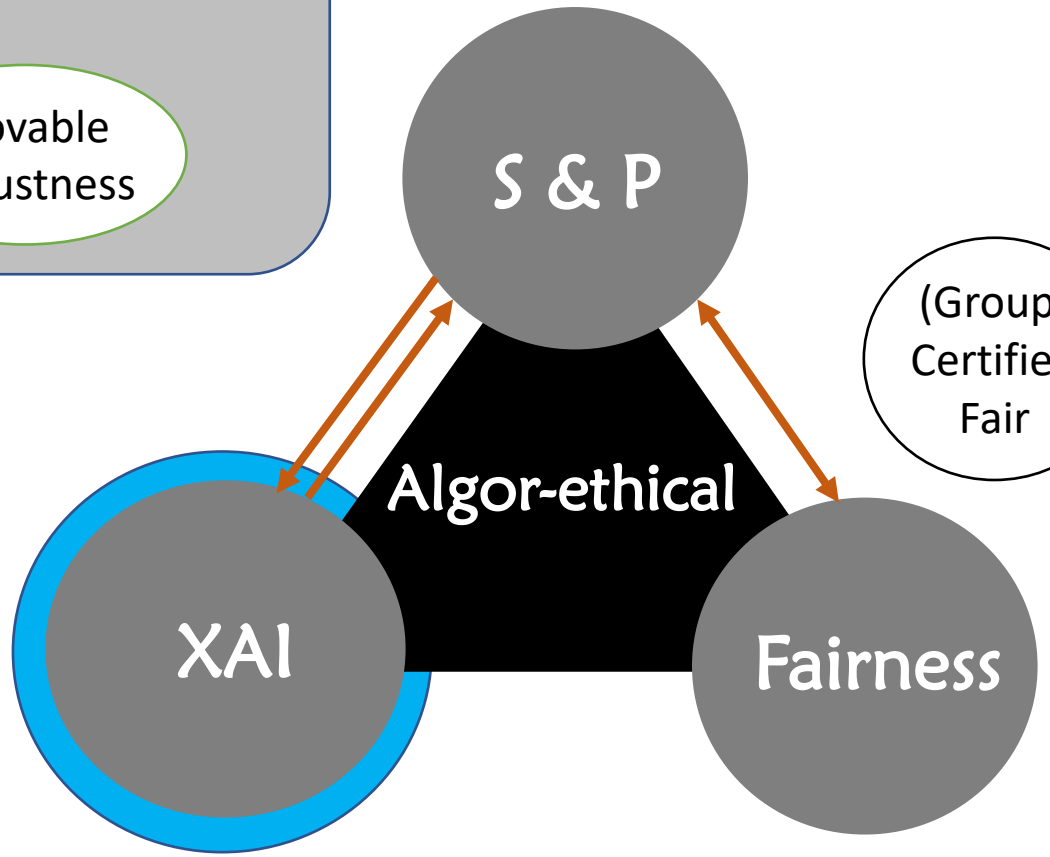

Human Bias
Training data is collected and annotated
Model is trained
Amplified Bias
Output
Encoding Bias
Biased data created from process becomes new training data

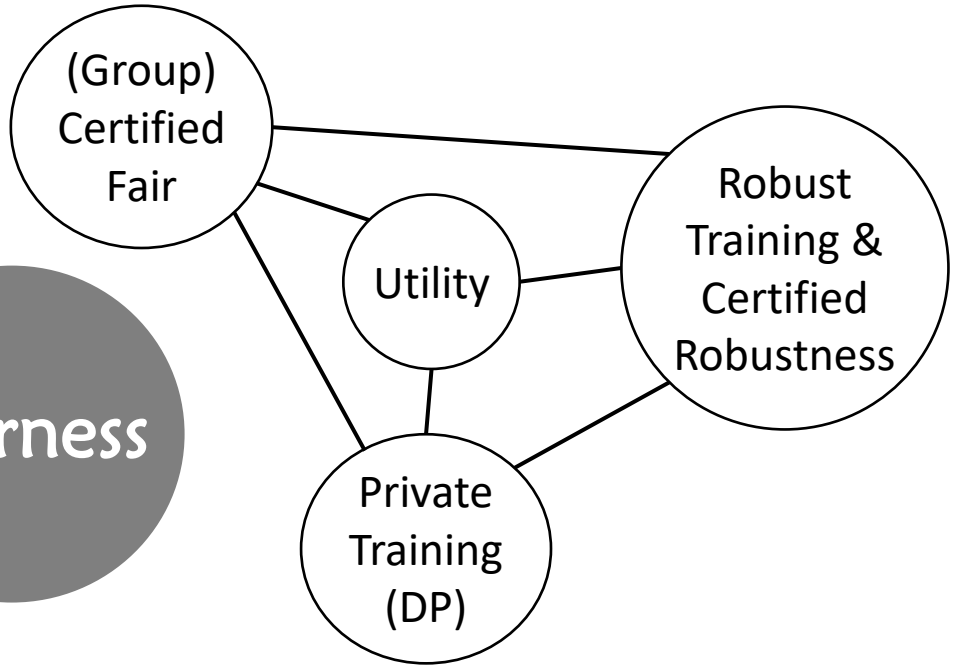
Privacy & Security




Transparency
Must be understandable to all
Security & Privacy
Must be secure and respect the privacy of users
Inclusion
Must not discriminate against anyone
Ethical AI
Reliability
Must be trustworthy and reliable
Responsibility
Must be someone who takes responsibility for what a machine does
Impartiality
Must not follow or create biases







1. Phan-Vu-Liu, Jin, Dou, Wu, & Thai. “Heterogeneous Gaussian mechanism: Preserving differential privacy in deep learning with provable robustness.” IJCAI 2019
2. Phan, Thai, Hu, Jin, & Dou. “Scalable differential privacy with certified robustness in adversarial learning.” ICML 2020

What if being exploited?

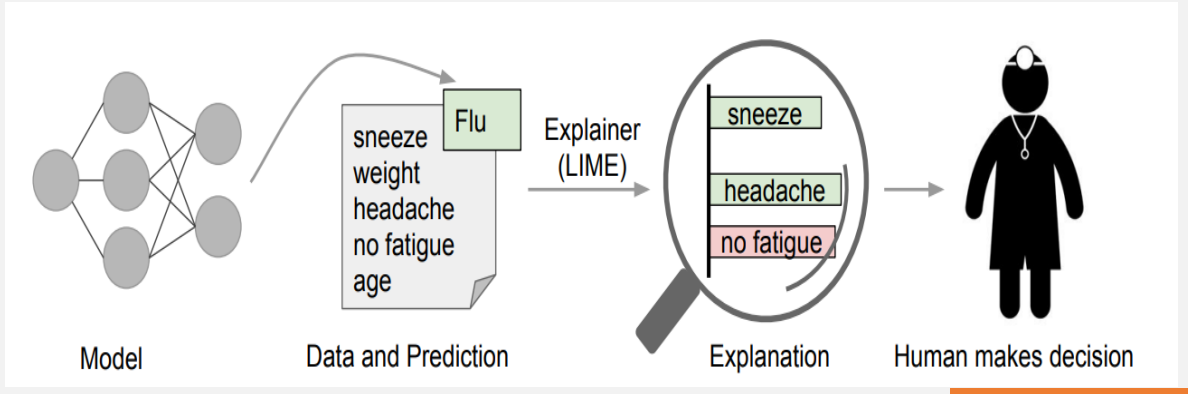
The Art of Explanation

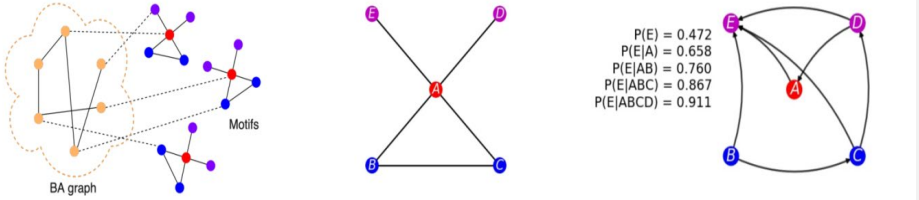
"...science is beautiful when it makes simple explanations of phenomena..."

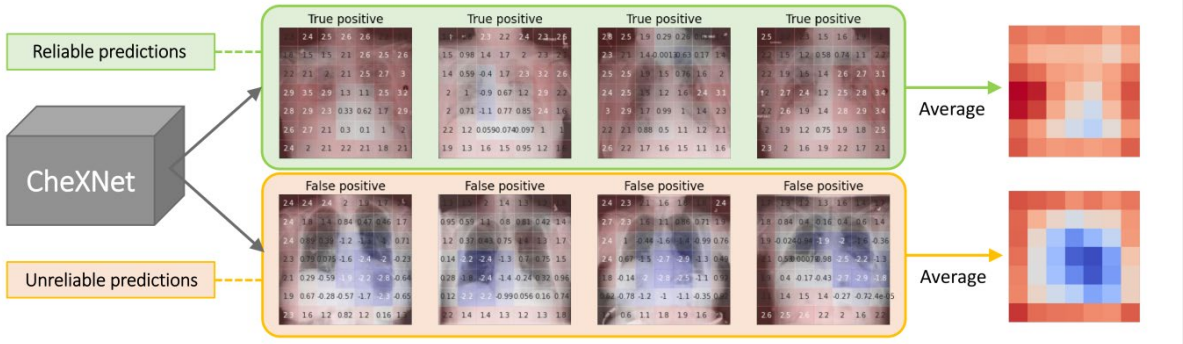
How to explain?

What to explain?

How to verify?

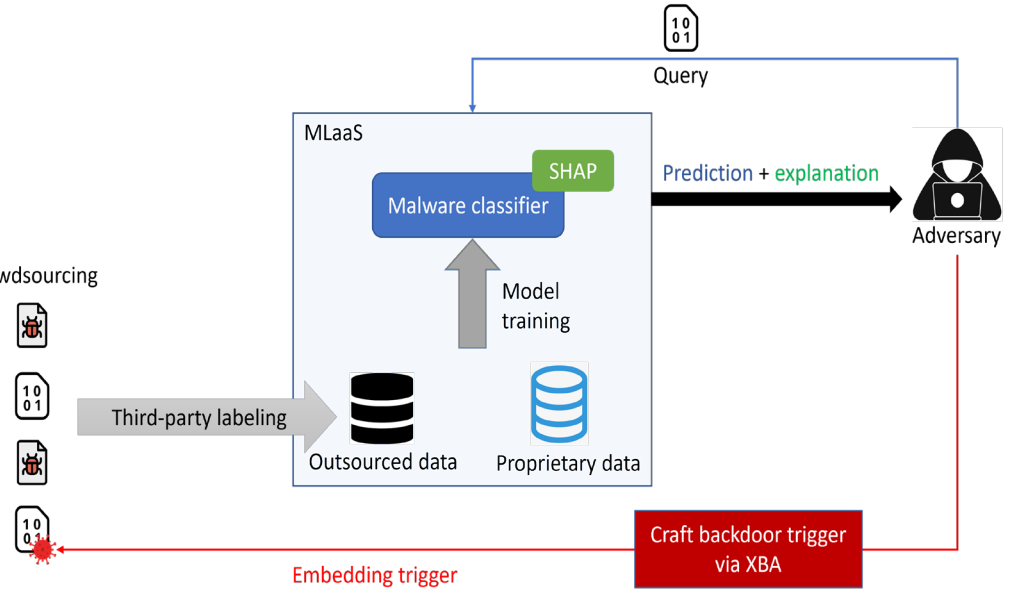


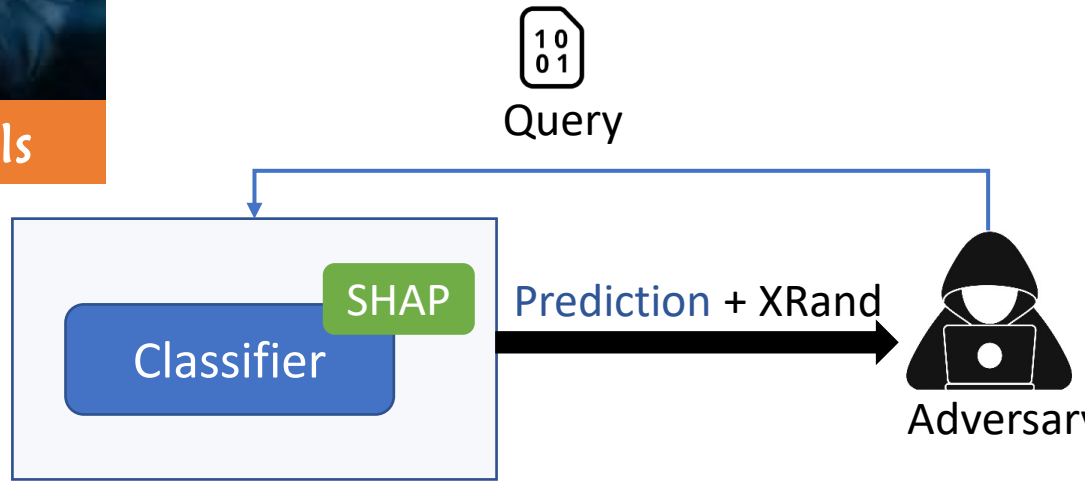




Explainable AI (XAI)

Challenges & Sols





Vu, Nguyen, and Thai. NeuCEPT: Learn Neural Networks' Mechanism via Critical Neurons with Precision Guarantee. ICDM 2022

Nguyen, Lai, Phan, and Thai. XRand: Differentially Private Defense against Explanation-Guided Attacks. AAAI 2023. Distinguished Paper Award

Thank you!

Acknowledgement

National Science Foundation Program on Fairness in AI in collaboration with Amazon under award No. 1939725, NSF Smart Connected Heath SCH 2123809, NSF SaCT 1935923