Trustworthy AI: Security

Destiny Pounds

M.S. Student, Biomedical Data Science

Advised by Vibhuti Gupta, Ph.D.

Assistant Professor, Computer Science and Data Science

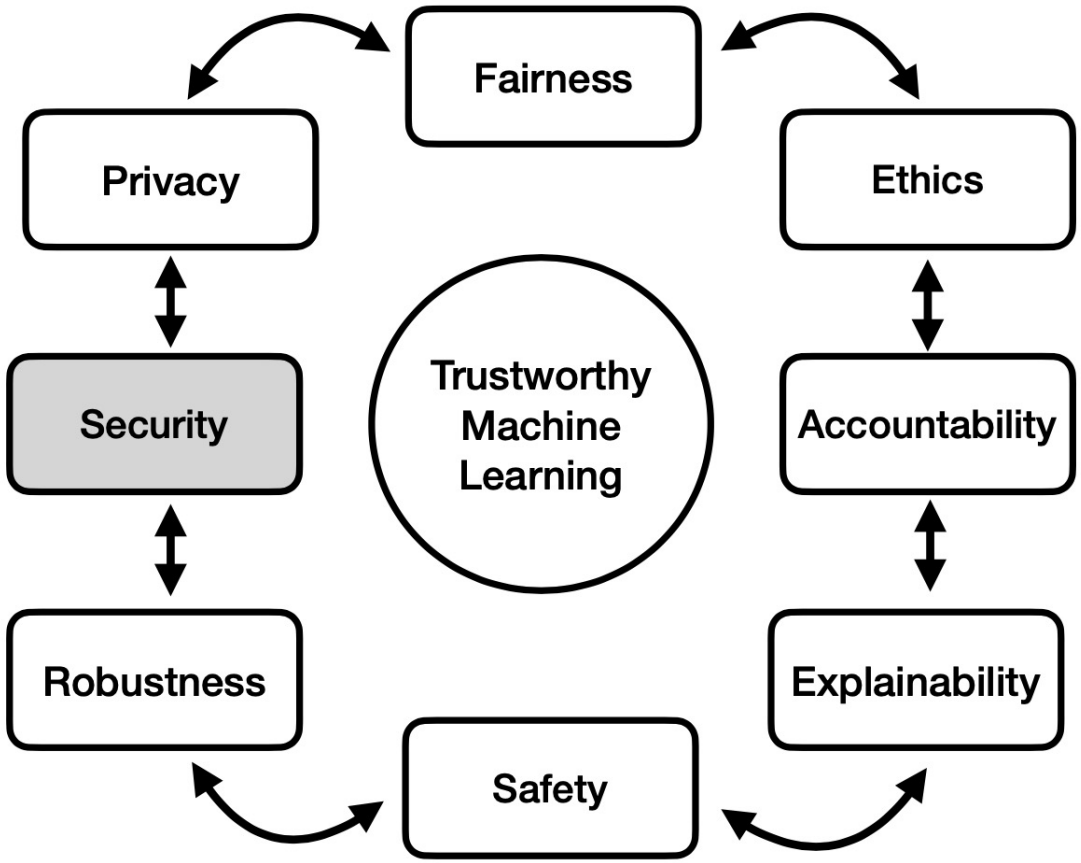
School of Applied Computational Sciences

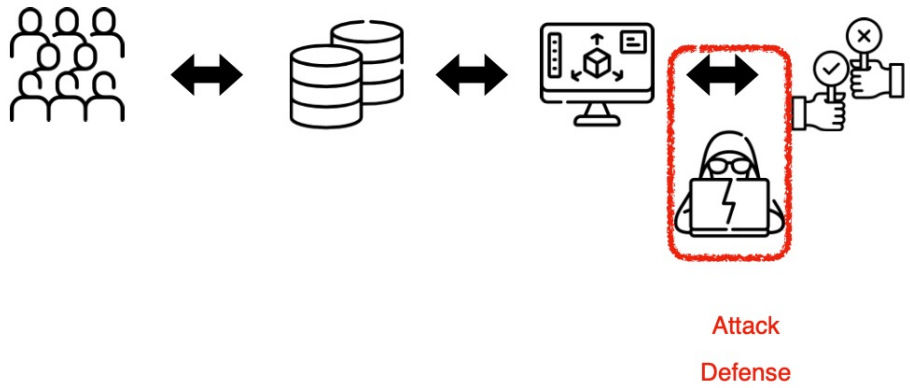
Meharry Medical College

Overview

- Defining Security
- Attack Types
  - Data Poisoning
  - Evasion Attacks
  - Membership Inference Attacks
  - Model Inversion Attacks
  - Model Extraction Attacks
- Mitigation Techniques

Security





- Cybersecurity is the practice of protecting systems, networks, and programs from digital attacks.

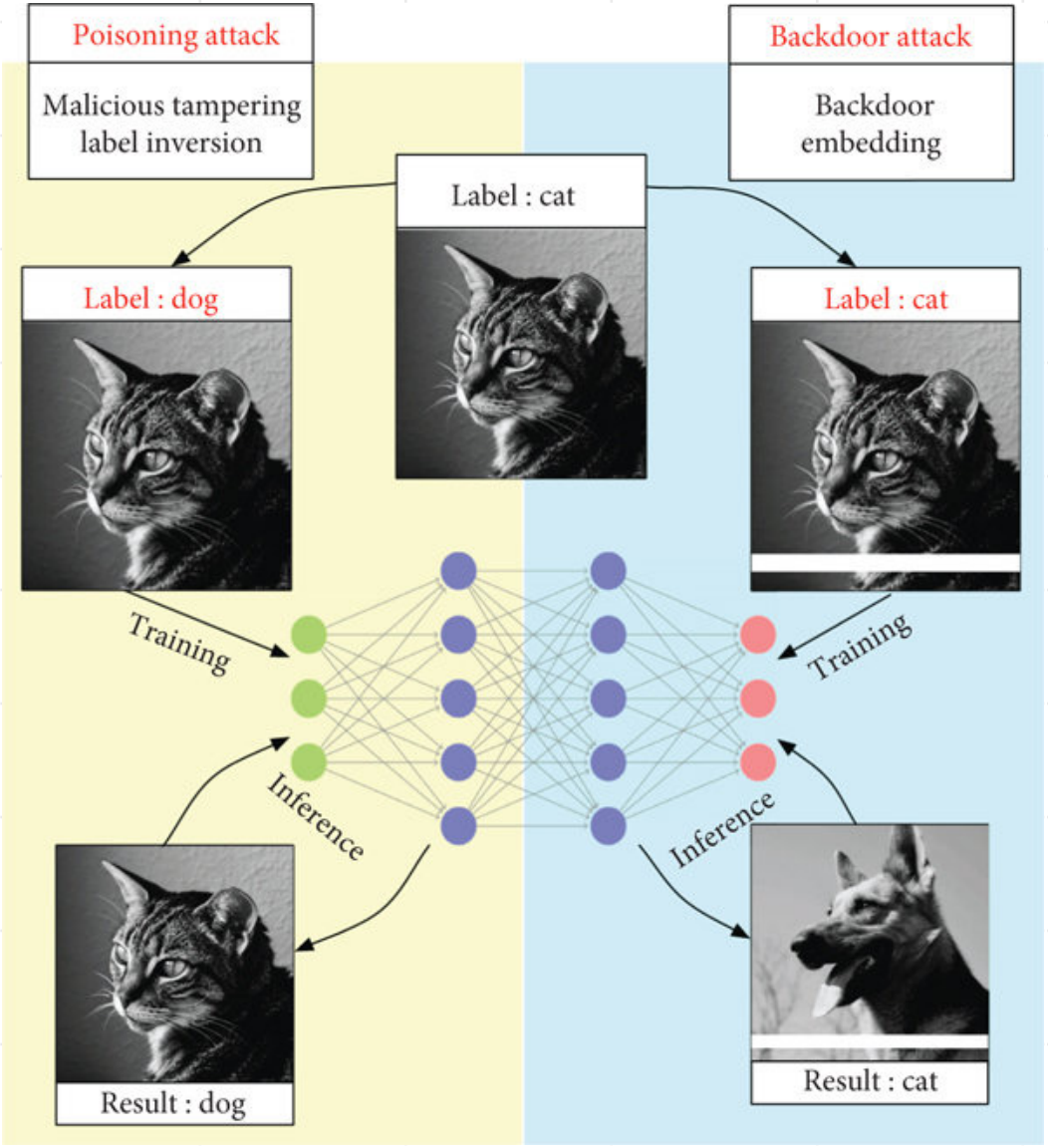
Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

Attack Types

- Data Poisoning
- Evasion Attacks
- Membership Inference Attacks
- Model Inversion Attacks
- Model Extraction Attacks

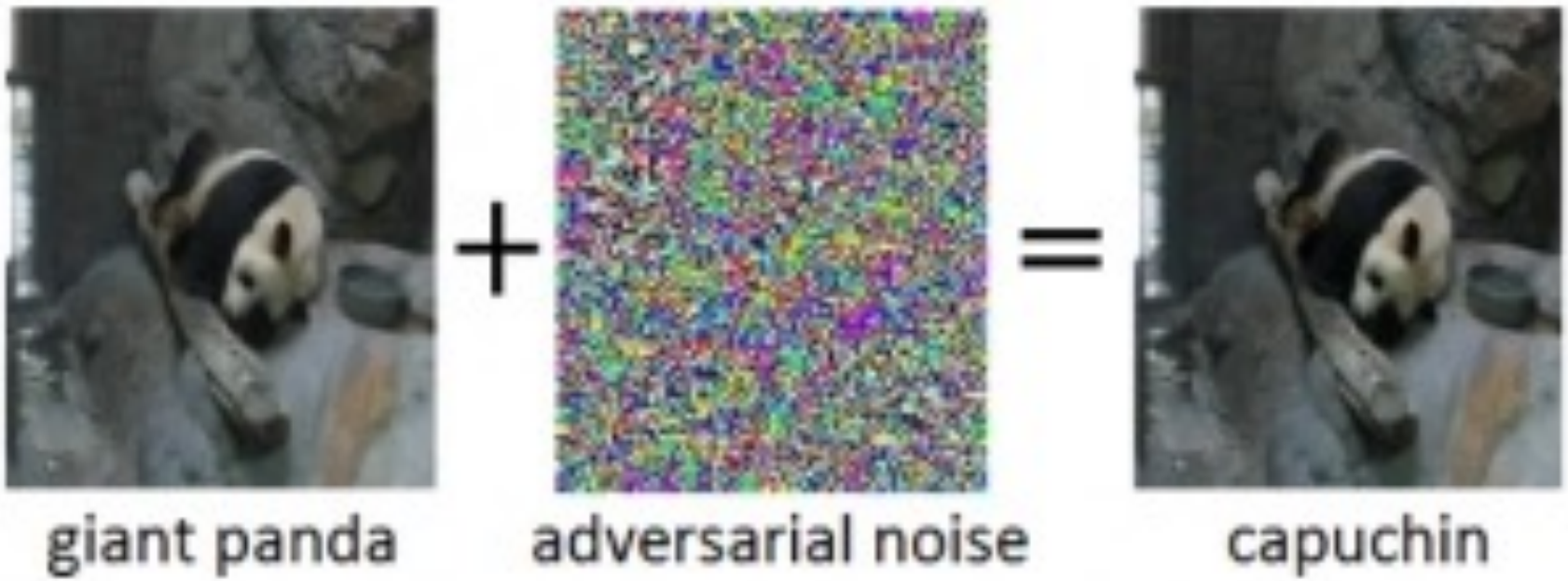
Data Poisoning

Attackers introduce malicious data into the training set to manipulate the model's behavior.



Evasion Attacks

Attackers manipulate input data in a way that alters the model's output or cause misclassification.

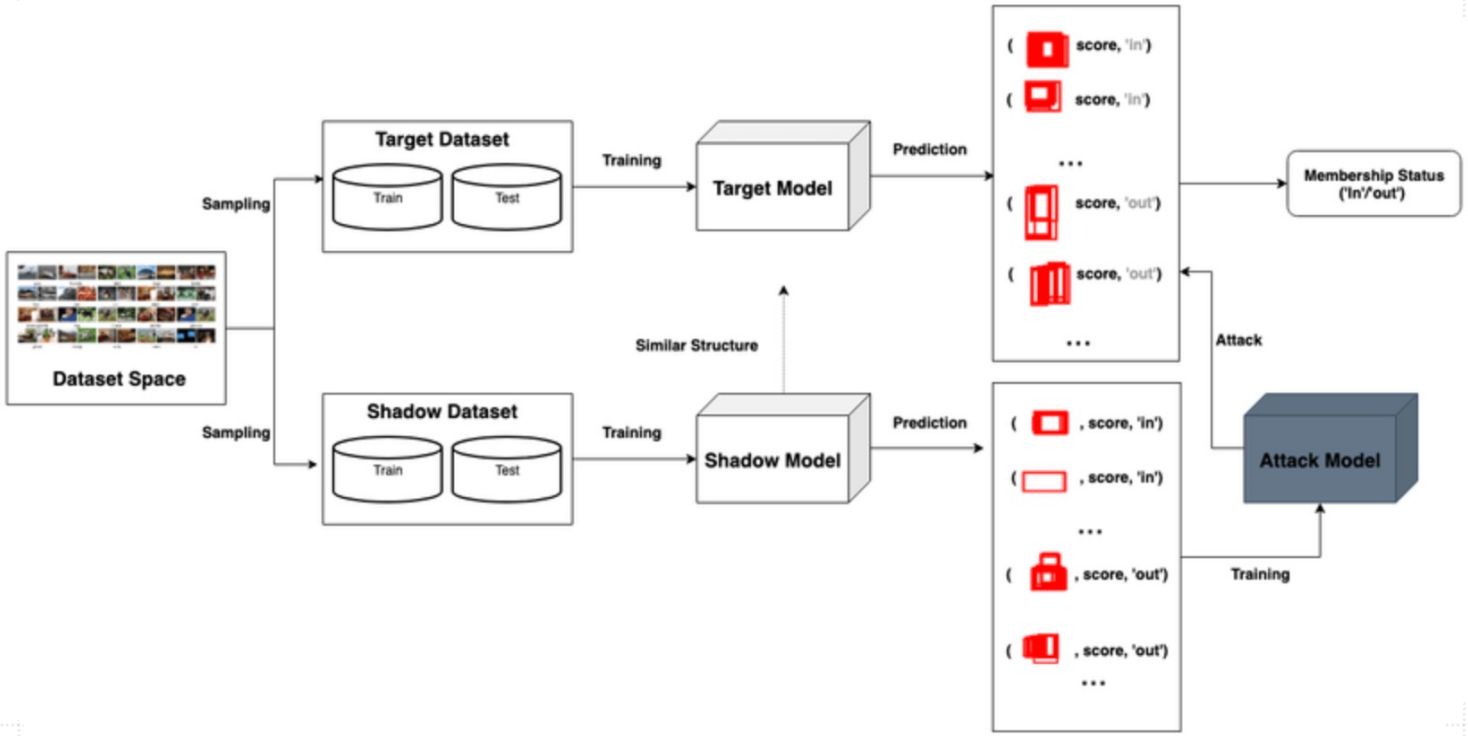


Mitigation Techniques

- **Data Validation:** techniques that can detect and remove suspicious data before training
- **Adversarial Training:** a technique improve model robustness and reduce the effect of adversarial examples
- **Model Auditing:** Regular monitoring and auditing of AI models help detect unexpected behaviors early

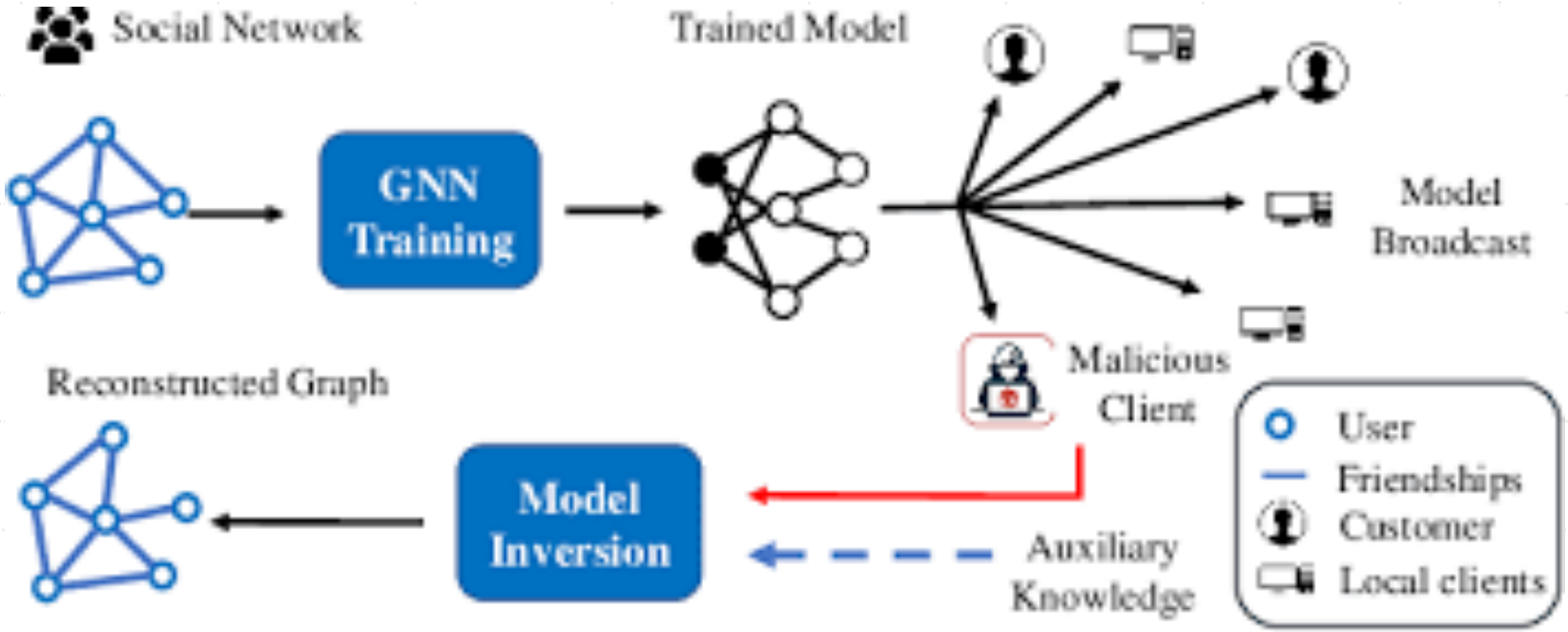
Membership Inference Attacks

Attackers infer whether a specific sample was part of the training data used by a model.



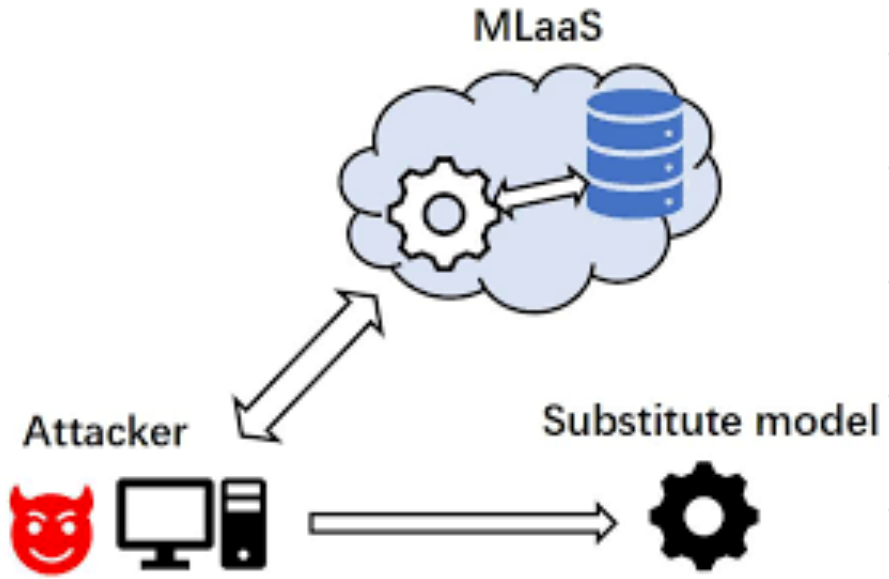
Model Inversion Attacks

Attackers attempt to reconstruct sensitive information about the training data or inputs by exploiting the model's output



Model Extraction Attacks

Attackers attempt to obtain a copy the target model by querying it and generating a substitute model.



Mitigation Techniques

- **Differential Privacy:**
- **Other Privacy-preserving techniques**

Thank You

Please send us your questions at:

vgupta@mmc.edu and

dpounds24@email.mmc.edu