

Adversarial Attacks and Defenses:

Adversarial attacks on machine learning models. Defense mechanisms against adversarial attacks. Adversarial robustness and detection techniques



neural network architectures

CNNs (AlexNet, VGG, ResNet, ...)

RNNs - Widely used in
language modeling
and
speech recognition

Transformers (BERT, GPT)

Security

- CIA Triad
- Threat modelling - a structured approach that's used to identify, prioritize, and manage potential threats in a system
 - STRIDE
 - Attack trees
 - More in detail:
 - <https://insights.sei.cmu.edu/blog/threat-modeling-12-available-methods/>
 - <https://attack.mitre.org/>
 - MITRE ATLAS
- Risks and mitigations
 - Identify threats --> assign risk
 - CIS Benchmarks
 - OWASP Top 10
- DevSecOps
- Host security
- Regular updates
- Minimal software
- User access control
- Firewall configuration
- Container security
- System monitoring and auditing
- Backup and recovery
- Disable unused network services
- Secure Shell (SSH) access
- Endpoint security
- Vulnerability management
- Network protection
- Securing code and artifacts

Bypass security with adversarial ai

Adversarial Robustness Toolbox

Surrogate shadow model

Adversarial robustness goes beyond traditional security measures

Detailed taxonomy of attack:

- <https://csrc.nist.gov/pubs/ai/100/2/e2023/final>

Types of Adversarial Attacks

Poisoning

- Tampering with training and validation datasets
- Produce backdoors
- Parasitic use

Evasion

- Target deployed models
- Facilitate fraud or misclassification
- DoS attacks

Extraction

- target the model's privacy

Inference

- target the model's privacy

Prompt injections

Poisoning attacks - reasons

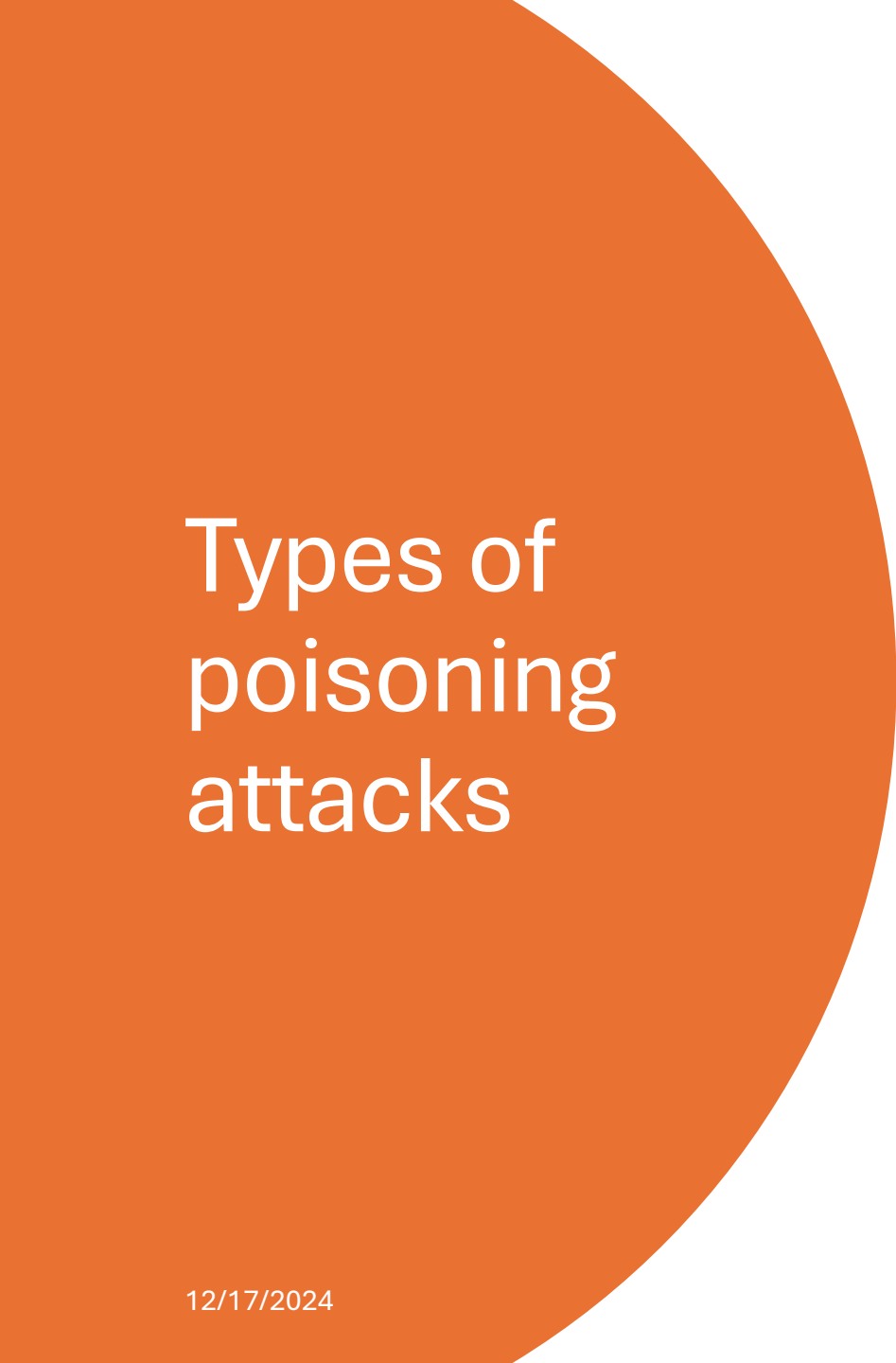
Bias
induction

Backdoor
insertion

Disruption

Competitive
sabotage

Ransom and
extortion

A large orange circle is positioned on the left side of the slide, partially overlapping the text.

Types of poisoning attacks

Targeted attacks/Untargeted attacks

Backdoor attacks

Clean-label attacks

Advanced attacks

Defending against poisoning attacks

- AWS SageMaker, MLflow, and Azure Machine Learning offer services and defenses against data poisoning
 - Data versioning and lineage
 - Data validation
 - Model versioning and lineage
 - Continuous monitoring
 - Access control
 - Model interpretability
 - Monitoring, logging, and alerting
 - Governance and collaboration

Defending against poisoning attacks

Anomaly detection

- Identification of suspicious data points
- Automated monitoring
- Reducing false positives

Techniques

- Statistical methods
- Clustering-based methods
- Neural networks
- Density-based methods

Advanced poisoning defenses with ART

-
- https://adversarial-robustness-toolbox.readthedocs.io/en/latest/modules/defences/detector_poisoning.html
 - https://github.com/Trusted-AI/adversarial-robustness-toolbox/blob/main/notebooks/poisoning_defense_activation_clustering.ipynb
 - https://github.com/Trusted-AI/adversarial-robustness-toolbox/blob/main/notebooks/poisoning_defense_spectral_signatures.ipynb
 - https://github.com/Trusted-AI/adversarial-robustness-toolbox/blob/main/notebooks/provenance_defence.ipynb

Trojan Horses and Model Reprogramming

- Degrading model performance
- Backdoor injection using pickle serialization
- Trojan horse injection with Keras lambda layers
- Trojan horses with custom layers
- Neural payload injection
- Attacking edge AI
- Model hijacking
 - Trojan horse code injection
 - Model reprogramming

Task: supply chain attacks

Individual reading and analysis in github

Evasion attacks

-
- Designed to mislead Machine Learning (ML) models deliberately
 - [Reconnaissance | MITRE ATLAS™](#) (Individual work)
 - https://openaccess.thecvf.com/content/CVPR2023W/AML/papers/Sarkar_Robustness_With_Query-Efficient_Adversarial_Attack_Using_Reinforcement_Learning_CVPRW_2023_paper.pdf (Individual work)

Collecting information for such attacks

- Model cards
- Published papers
- Blogs
- Social engineering
- Online probing
- Open source model repositories
- Transfer learning
- Shadow models

Perturbations and image evasion attack techniques

-
- Crafted modifications that cause a model to make incorrect predictions when applied to input data
 - The size of perturbation, in other words, the features to alter (L1 norm)
 - Its closeness – or Euclidean distance – to the original sample (L2 norm)
 - The maximum change to any feature in the data (infinity norm, or L^∞)

Perturbations and image evasion attack techniques

-
- Fast Gradient Sign Method (FGSM)
 - Basic Iterative Method (BIM)
 - Projected Gradient Descent (PGD)
 - Carlini and Wagner (C&W) attack
 - Jacobian-based Saliency Map Attack (JSMA).

-
- NLP evasion attacks
 - Change words or phrases in a text snippet
 - classification from positive to negative sentiment
 - non-spam to spam
 - Natural Language Inference (NLI)
 - Universal Adversarial Perturbations (UAPs)
 - Black-box attacks with transferability

Defending against evasion attacks

-
- Reactive measures
 - Respond to attacks as they happen
 - Real-time detection and mitigation
 - Proactive measures
 - Hardening models



Defending against evasion attacks

- Adversarial training
 - <https://adversarial-robustness-toolbox.readthedocs.io/en/latest/modules/defences/trainer.html>
- Input preprocessing
- Model hardening techniques
 - Defensive distillation
 - Feature squeezing
 - Gradient masking
 - Robust loss functions
- Model ensembles
 - Certified defenses