SECURE USE OF LLMs AND GEN AI SYSTEMS

NIST Workshop Jan 17, 2024

David Beveridge, Vice President of Engineering

HIDDENLAYER

NIST QUESTIONS FOR THE WORKSHOP

1. What changes, if any, need to be made to SSDF version 1.1 to accommodate secure development practices for generative AI and dual-use foundation models?
2. What AI-specific considerations should NIST capture in its companion resource?
3. What else should be captured in the SSDF Profiles?
4. Is there an alternative to an SSDF Profile that would be more effective at accomplishing the EO 14110 requirement, while also providing flexibility and technology neutrality for software producers?
5. What secure development resources specific to AI models do you find most valuable?
6. What is unique about developing code for generative AI and dual-use foundation models?

THE ML OPPORTUNITY & RISK IS MASSIVE

Achieve competitive advantage by enhancing customer experience, improving strategy & streamlining operations

"AI could contribute up to **$15.7 trillion** to the global economy in **2030**"

pwc

"Cybersecurity remains the **ONLY** risk that a majority of respondents say their organizations consider relevant"

McKinsey Global State of AI Survey

"Through 2022, **30% of all AI cyberattacks will leverage training-data poisoning**, AI model theft, or adversarial samples to attack AI-powered systems."

"**2 in 5** organizations have had an AI security or privacy breach. **1 in 4** were malicious attacks."

Gartner

ML ADVERSARIAL ATTACKS ARE EXPLODING

REAL WORLD ATTACKS

1. Facebook
Involved in **34** incidents, allegedly harming **60** entities.

2. Tesla
Involved in **31** incidents, allegedly harming **41** entities.

3. Google
Involved in **23** incidents, allegedly harming **31** entities.

ACCELERATING REGULATIONS

THE WHITE HOUSE
WASHINGTON

On October 4, 2022, The White House released an **AI Bill of Rights**.

Bank of England

MITRE | ATLAS™

CSRC

WEAPONIZED AML TOOLS

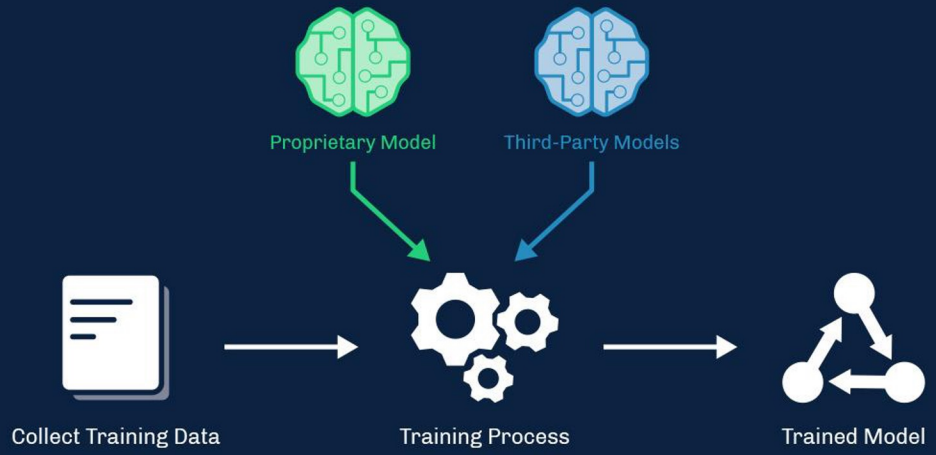
20+ free tools available online

Adversarial Robustness Toolbox

AugLy

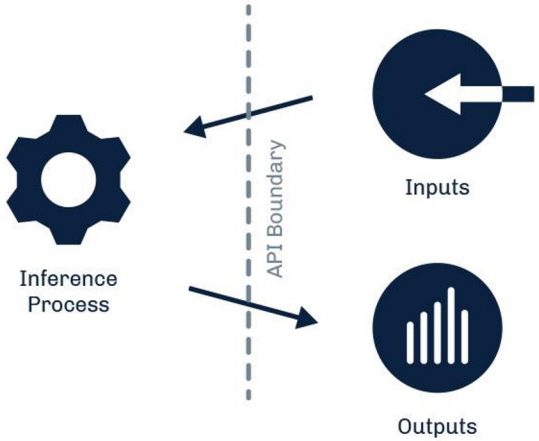
NOW IS THE TIME TO PROTECT

Models Need to be Scanned to Ensure Safety



MODEL TRAINING & DEVELOPMENT

Models Need Real-Time Monitors & Detection for Attacks to Ensure Safety



PRODUCTION MODELS
"AI ON THE EDGE"

HIDDENLAYER

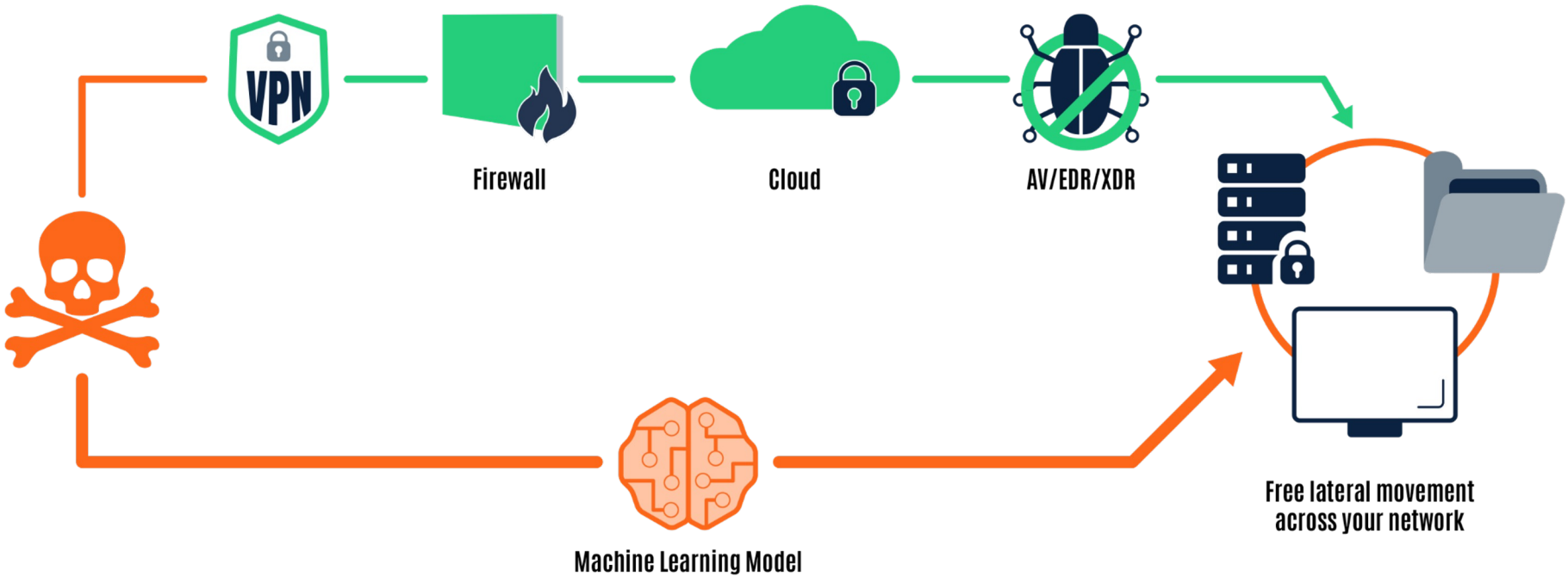
SECURITY OPERATIONS FOR AI

DISCOVERY → SAFETY & TRUST → ATTACK MONITORING → RESPONSE → SITUATION AWARENESS

Where are all my AI models?	Are my AI models safe to use?	Are my AI models being attacked?	What should I do about it?	Where is my AI security posture?
• Model Registry • File Format Coverage	• Malware • Vulnerabilities • Integrity Issues • Known Good State • Genealogy • Red Team Model Assessment	• Adversarial ML • Poisoning • Model Evasion • Model Inversion • Model Theft • Prompt Injection • Confidential Data Leakage	• AdvML Attack Remediations • Ticketing Systems • SecOps Tools	• Detection Details Report • Security Health Dashboard • Risk Assessment Report • Response Dashboard

ML IS AN UNSECURED ATTACK VECTOR

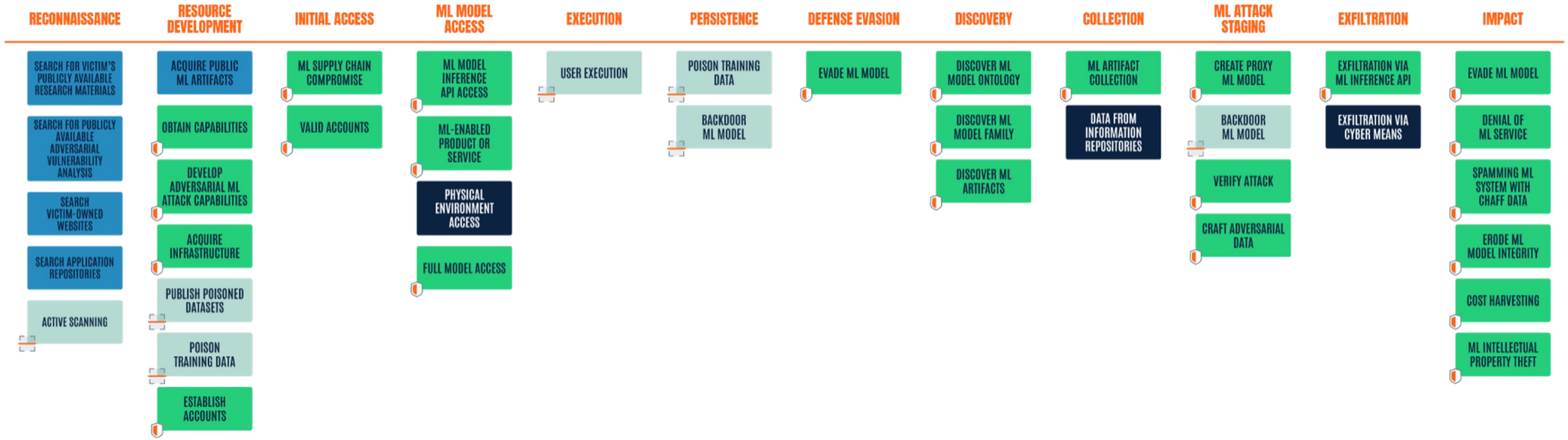
Launchpad for lateral movement, deployment of malware, theft of IP/PII, and manipulation of the model output



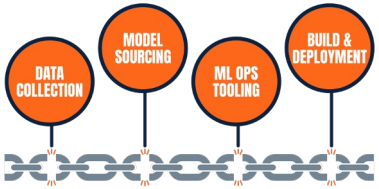
MITRE | ATLAS FRAMEWORK

Key Control Points

Run Time Protection | MODEL SCANs | People & Process



AI RISK REQUIRES MITIGATION

AI USE CASE	GENERATIVE AI	FRAUD MANAGEMENT	SUPPLY CHAIN ATTACKS
			
MOST COMMON ATTACK TYPES	Inference Data Poisoning Prompt Injection	Inference Bypass Ransomware Attack	Ransomware Attack Data Poisoning Backdoored Models
KEY CONTROLS NEEDED	Threat Modeling Red Teaming Model Scans Real time/Run time protection	Threat Modeling Red Teaming Model Scans Real time/Run time protection	Threat Modeling Red Teaming Model Scans Real time / Run time protection

NOTES REGARDING NIST SP 800-218 AND AI/ML

- PO 1.1, 2.1, 2.3, 5.1 (ID + Document Sec. Req., *et al.*): Include ML Dev, ML Ops, DatSci; **Protect Training *Corpora***
- PO 3.2: (Rec. Sec. Pract.) Scan 3rd party models used by org, ML Ops Protection (EDR), **Provenance of Models** (used and created), Data Curation & Training, Build/Supply Chain Security
- PS 1.1: (Src Storage) Extend to include Models and the **Data used for training**
- PS 2.1: (Publish SW Integrity Info) **Sign models**, Train/Build/Ops pipeline libs/vers (BOM), Secure Scoring APIs, Provenance, Need Standards around bias checks?
- PS 3.1: (Archiving) **Archive all models**, training data used per, meta data?
- PS 3.2: (Provenance) I**nclude 3rd party models** included in system, provenance of models derived from 3rd party models
- PW 1.1: (**Risk Modelling**) Back Doors/activation, Real Time Manifold Exploration/activation, Unusual Categorizations, Malware Infections, Vulnerabilities in models themselves
- PW 2.1: (SW Design Sec Req. Review) Each item has an analogous AI aspect
- PW 4.1, 4.4, 6.1: (Use Existing Secured Tools, *et al.*) Incl. **ML build/supply chain** (libs, frameworks); provenance of model *and data*
- PW 6.2: (predetermine comp/tools) Ban inherently insecure formats (Pickle, Cloud Pickle, *etc.*)
- PW 7.2: (Src Review/analysis) Incl. Model scanning backdoors, **exercise attacks against models**, bias checks
- PW 8.x: (Test Exe Code) Models can (mostly do) **have executable code embedded within** them
- PW 9.2: (Specify Settings) IAC for MLOps should be included
- RV 1.1: (Market info on vulns): Provenance of models, training/ops frameworks, file formats as part of BOMs
- RV 1.3: (Vuln. Discl.): **include AI**: biased models, poisoned training, vuln. train/dev/ops frameworks, discovered bypasses
- RV 2.2: (Risk Responses) Industry doesn't understand the need to secure AI, nor what that undertaking involves - MITRE ATLAS framework to begin with; how can you even tell?
- RV 3.2: (RCAs) AI needs to be included in secure coding practices to be followed: **Must include the data** (*corpora)* used to train the model, APIs can assist in gradient-type attacks
- Runtime Controls missing
- Responsibility gaps/definitions amongst DatSci, MLOps, Product