RSAConference2018
San Francisco | April 16 – 20 | Moscone Center

NOW MATTERS

SESSION ID: SPO2-T07

AI AND CYBERSECURITY

APPLICATIONS OF ARTIFICIAL INTELLIGENCE IN SECURITY
UNDERSTANDING AND DEFENDING AGAINST ADVERSARIAL AI

Sridhar Muppidi
IBM Fellow and VP Technology
IBM Security

Koos Lodewijkx
VP Technology
IBM Security

IBM Security

Agenda

Three perspectives on AI and Security:

1. CISO: AI for cyberdefense

2. Attacker: using and attacking AI for fun and profit

3. R&D: making AI more robust

CISO PERSPECTIVE: AI FOR CYBERDEFENSE

What CISOs are facing

COMPLIANCE MANDATES

GDPR fines can cost

billions

for large global companies

What CISOs are facing

COMPLIANCE MANDATES

GDPR fines can cost

billions

for large global companies

SKILLS SHORTAGE

By 2022, there will be

1.8 million

unfulfilled cybersecurity positions

What CISOs are facing

COMPLIANCE MANDATES

GDPR fines can cost

billions

for large global companies

SKILLS SHORTAGE

By 2022, there will be

1.8 million

unfulfilled cybersecurity positions

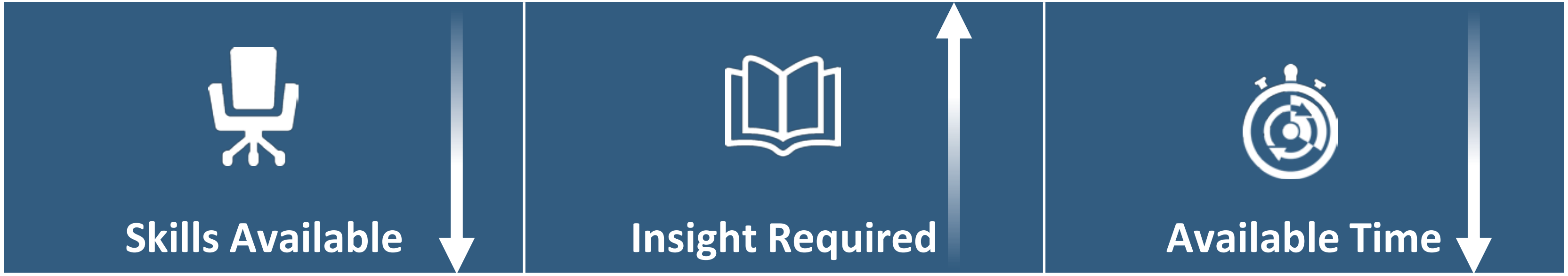
TOO MANY TOOLS

Organizations are using

too many

tools from too many vendors

What motivates the rush on AI for security?

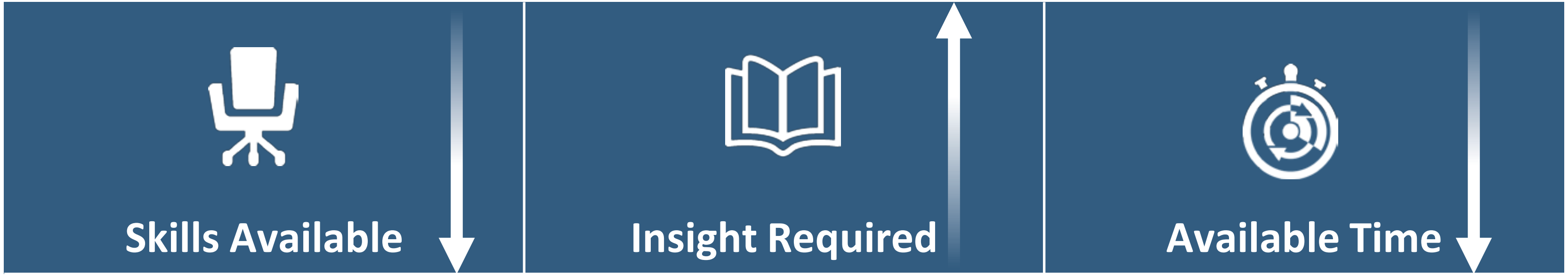


Skills Available

Insight Required

Available Time

What motivates the rush on AI for security?



Skills Available

Insight Required

Available Time

Skills

- Sophistication of tools
- Evolution of the threat
- Lack of best practices

Insight

- Complexity of context
- Lack of insights
- Insufficient data

Speed

- Attacks move faster
- Shortening disclosure timeframes

Using AI to address growing security needs

Predictive Analytics

Intelligence Consolidation

Trusted Advisors & Response

Using AI to address growing security needs

Predictive Analytics

- **Approach**: Model behaviors and identify emerging and past threats and risks
- **Applications**:
  - Network threats
  - User behavior
  - Endpoint threats / malware
  - Application testing
  - Data access patterns

Intelligence Consolidation

Trusted Advisors & Response

Using AI to address growing security needs

Predictive Analytics

- **Approach**: Model behaviors and identify emerging and past threats and risks
- **Applications**:
  - Network threats
  - User behavior
  - Endpoint threats / malware
  - Application testing
  - Data access patterns

Intelligence Consolidation

- **Approach**: Curation of intelligence and contextual reasoning
- **Applications**:
  - Open Source TI
  - Security Research
  - Regulatory documents

Trusted Advisors & Response

Using AI to address growing security needs

Predictive Analytics

- **Approach**: Model behaviors and identify emerging and past threats and risks
- **Applications**:
  - Network threats
  - User behavior
  - Endpoint threats / malware
  - Application testing
  - Data access patterns

Intelligence Consolidation

- **Approach**: Curation of intelligence and contextual reasoning
- **Applications**:
  - Open Source TI
  - Security Research
  - Regulatory documents

Trusted Advisors & Response

- **Approach:** Reason about security events for triage and response
- **Applications:**
  - Automated forensics
  - Case analysis
  - Case preparation
  - Automated response

ATTACKERS: AI FOR FUN AND PROFIT

Attackers: AI for fun and profit

AI Powered Attacks

Attacking AI

Theft of AI

Attackers: AI for fun and profit

AI Powered Attacks

- Generating new attacks
- Automating large scale attacks
- Refining existing attacks
- Evading defenses (generative adversarial networks)

Attacking AI

Theft of AI

Attackers: AI for fun and profit

AI Powered Attacks

- Generating new attacks
- Automating large scale attacks
- Refining existing attacks
- Evading defenses (generative adversarial networks)

Attacking AI

- Poisoning models
- Evade AI powered defenses
- Harden attacks (reinforcement learning etc.)

Theft of AI

Attackers: AI for fun and profit

AI Powered Attacks

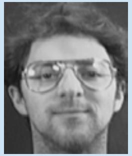
- Generating new attacks
- Automating large scale attacks
- Refining existing attacks
- Evading defenses (generative adversarial networks)

Attacking AI

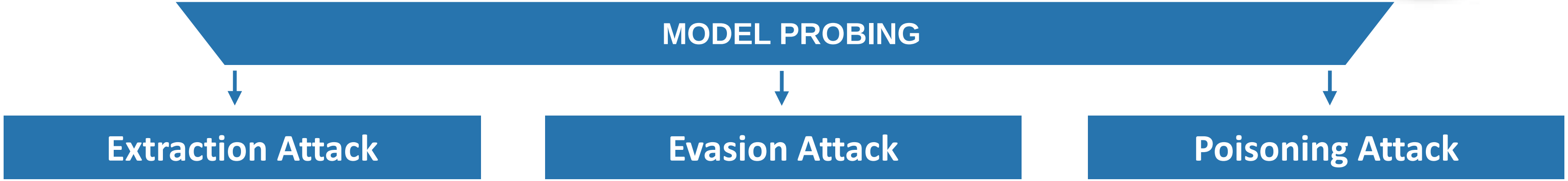
- Poisoning models
- Evade AI powered defenses
- Harden attacks (reinforcement learning etc.)

Theft of AI

- Theft of models
- Transfer attacks
- Privacy (model inversion)



Attacks against AI: Security threats for AI APIs



MODEL PROBING

Extraction Attack

Evasion Attack

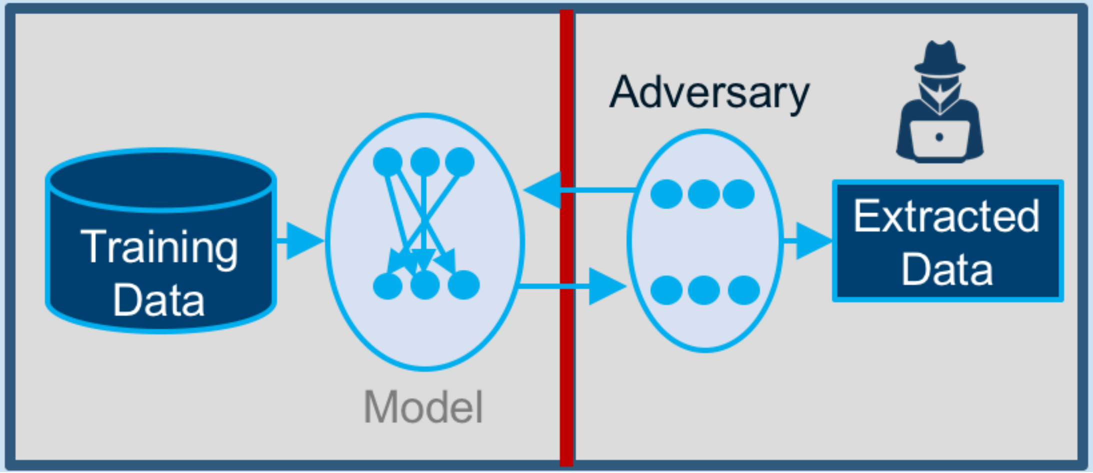
Poisoning Attack

Attacks against AI: Security threats for AI APIs

MODEL PROBING

Extraction Attack | Evasion Attack | Poisoning Attack

- Adversary extracts model and proprietary training data information
- **Vulnerable domain**
  Models that provide insights from proprietary data
  - E.g., Extract sensitive confidential information from training data



Attacks against AI: Security threats for AI APIs

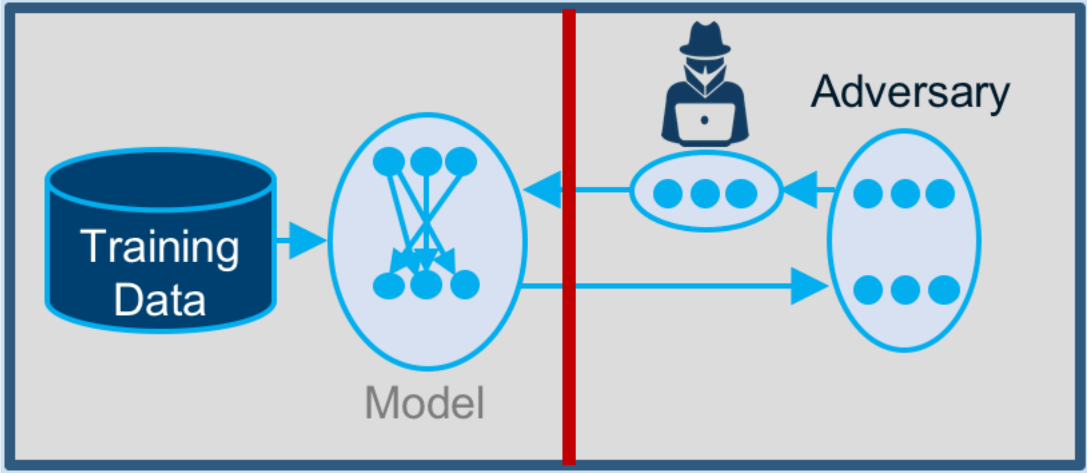
MODEL PROBING

Extraction Attack

Evasion Attack

Poisoning Attack

- Exploit model blind spots to mislead or fool the model
- **Vulnerable domain** Models used in screening or supervisory functions
  - E.g., Minimally perturb images to bypass image recognition service



Attacks against AI: Security threats for AI APIs

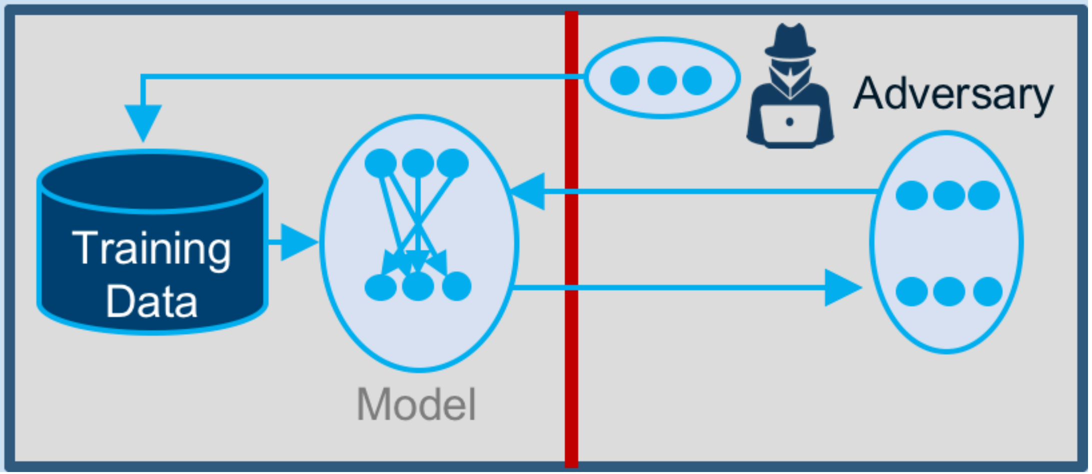
MODEL PROBING

Extraction Attack

Evasion Attack

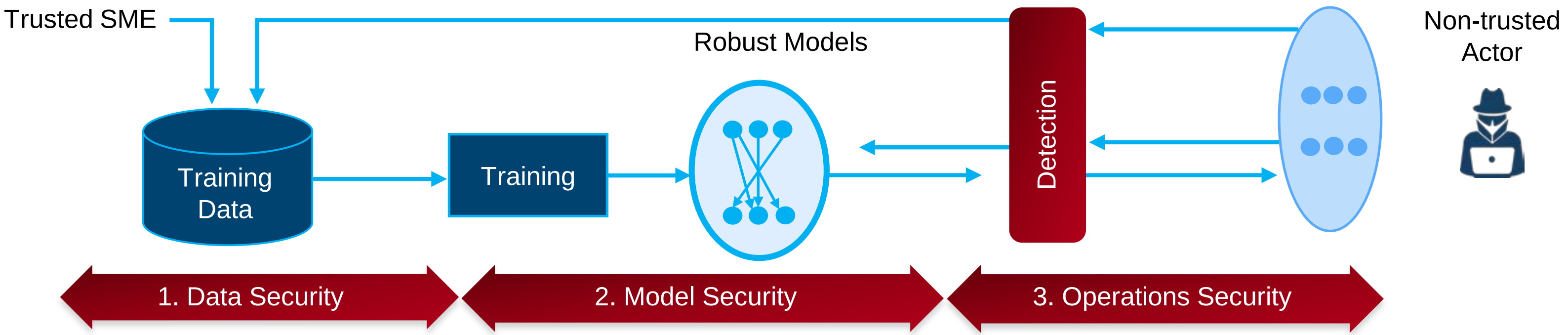
Poisoning Attack

- Corrupt model by manipulating training data to shift underlying model
- **Vulnerable domain**
  Any model that is basedon active / online learning
  - E.g., Corrupting a chat bot through interaction



Attacks against AI: Countermeasures

Data Security	**Ground truth protection:** process and enrich training data to protect privacy and increase robustness
Model Security	**Robust models:** Techniques and algorithms for resilient models by construction
Operations Security	**Threat detection:** Detect and eliminate adversarial inputs during production use



RESEARCH: MAKING AI MORE ROBUST

Adversarial Robustness Toolbox (ART)

Announcing:

ART – an **open-source library** for **adversarial machine learning**

- ART provides an implementation for many state-of-the-art methods for attacking and defending classifiers
- ART allows rapid crafting & analysis of attacks and defense methods for machine learning models

https://github.com/IBM/adversarial-robustness-toolbox

Adversarial Robustness Toolbox (ART)

Attack methods	Defense methods
• Deep Fool (Moosavi-Dezfooli et al., 2015) • Fast Gradient Method (Goodfellow et al., 2014) • Jacobian Saliency Map (Papernot et al., 2016) • Universal Perturbation (Moosavi-Dezfooli et al., 2016) • Virtual Adversarial Method (Moosavi-Dezfooli et al., 2015) • C&W Attack (Carlini and Wagner, 2016) • NewtonFool (Jang et al., 2017)	• Feature squeezing (Xu et al., 2017) • Spatial smoothing (Xu et al., 2017) • Label smoothing (Warde-Farley and Goodfellow, 2016) • Adversarial training (Szegedy et al., 2013) • Virtual adversarial training (Miyato et al., 2017)

https://github.com/IBM/adversarial-robustness-toolbox

ART DEMONSTRATION

Apply What You Have Learned Today

- In the next week:
  - **Understand and educate** your team about AI for Security vs Security for AI;
  - Experiment with basic AI models
- In the first three months:
  - **Kick off a security analytics projects** to get unique insights and take action
  - Identify the data sources i.e. SIEM data, Data activity monitoring, IAM data, etc.
  - Identify scenarios of interest. i.e. where the sensitive data is, who is accessing what, what systems are more vulnerable, what are patterns of frequent attacks, etc.
- Within six months leverage:
  - **Leverage the ART toolkit** to help improve robustness of AI models
  - Mature the analytics project with AI powered orchestration