Cybersecurity in the era of generative AI

IBM **Security**

Enterprises are embracing generative AI, but have concerns

Under pressure to adopt

64%

face significant pressure to accelerate generative AI initiatives

Concerned about new risks

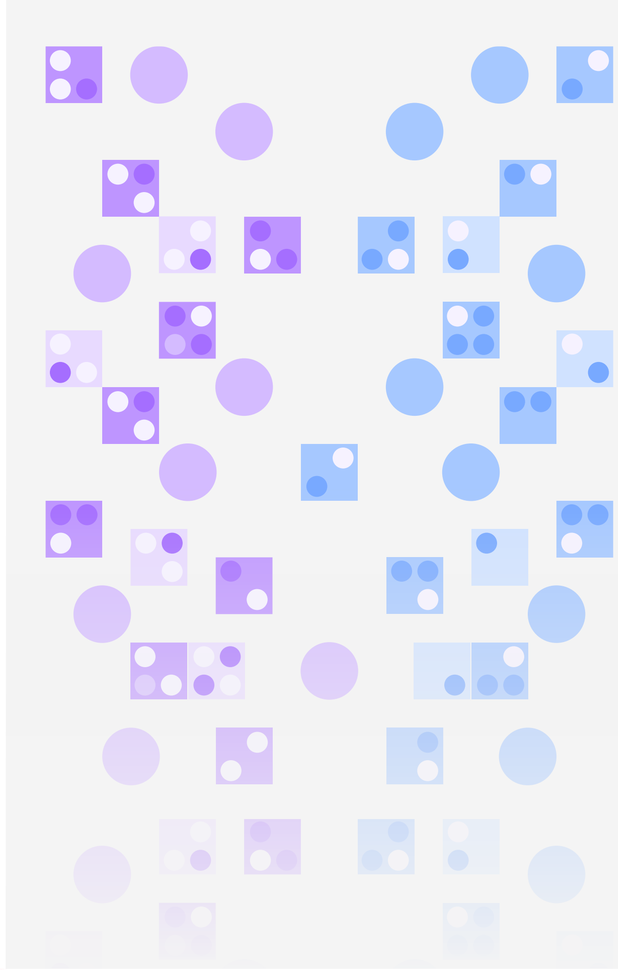
84%

see cybersecurity risk as the #1 roadblock to generative AI adoption

Investing in new defenses

64%

identified security as the #1 priority for generative AI use cases



Generative AI is the new burning platform to secure now

Source: “Enterprise generative AI: State of the market” 2023 IBM Institute of Business Value Study

Attackers will **target AI**

AI should be treated as a **new attack surface,** with new detection and response strategies required for model evasion, extraction, inference and poisoning

Prompt injection can drop defenses preventing generation of unwanted material, plus access to exploitable integrations and a wealth of sensitive training data

Malicious models can be uploaded to open repositories, with **hidden behavior** triggered long after they've been deployed

Attackers will **utilize AI**

Generative AI will **scale** cybercrime, and reduce barriers to entry to lower skilled attackers

Attacks will become more **targeted**, and generative video and audio techniques will necessitate new approaches to avoid business compromise

Attackers will **adapt** to defensive strategies faster, and improve detection evasion, vulnerability discovery, and malware customization

IBM’s approach to security in the era of generative AI

Security for AI

Protecting foundation models, generative AI, and their data sets is essential for enterprise-ready AI

Secure the underlying AI training data
by protecting it from sensitive data theft, manipulation, and compliance violations

Secure model development
by scanning for vulnerabilities in the pipeline, hardening integrations, and enforcing policies and access

Secure the usage of AI models
by detecting data or prompt leakage, and alerting on evasion, poisoning, extraction, or inference attacks
IBM Adversarial Robustness Toolkit

AI for Security

Productivity gains from foundation models and generative AI will reduce human bottlenecks in security

AI will manage repetitive security tasks
such as summarizing alerts and log analysis, freeing teams to tackle strategic problems

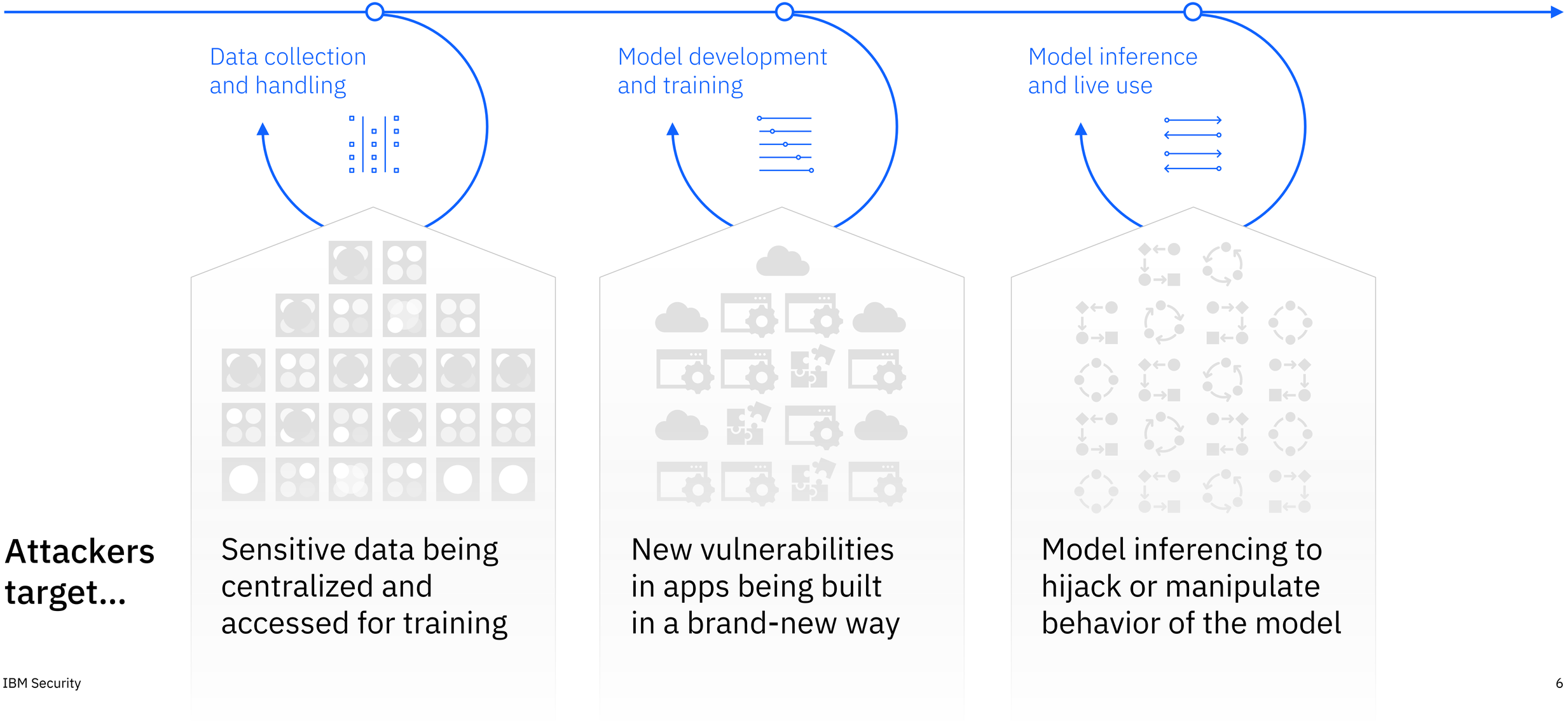
AI will generate security content
(detections, workflows, policies) faster than humans, expediting implementation and adjusting to changing security threats in real-time

AI will learn and create active responses
that optimize over time, with abilities to find all similar incidents, update affected systems, and patch vulnerable code

Security for AI

Adversarial risks across the AI pipeline

AI pipeline



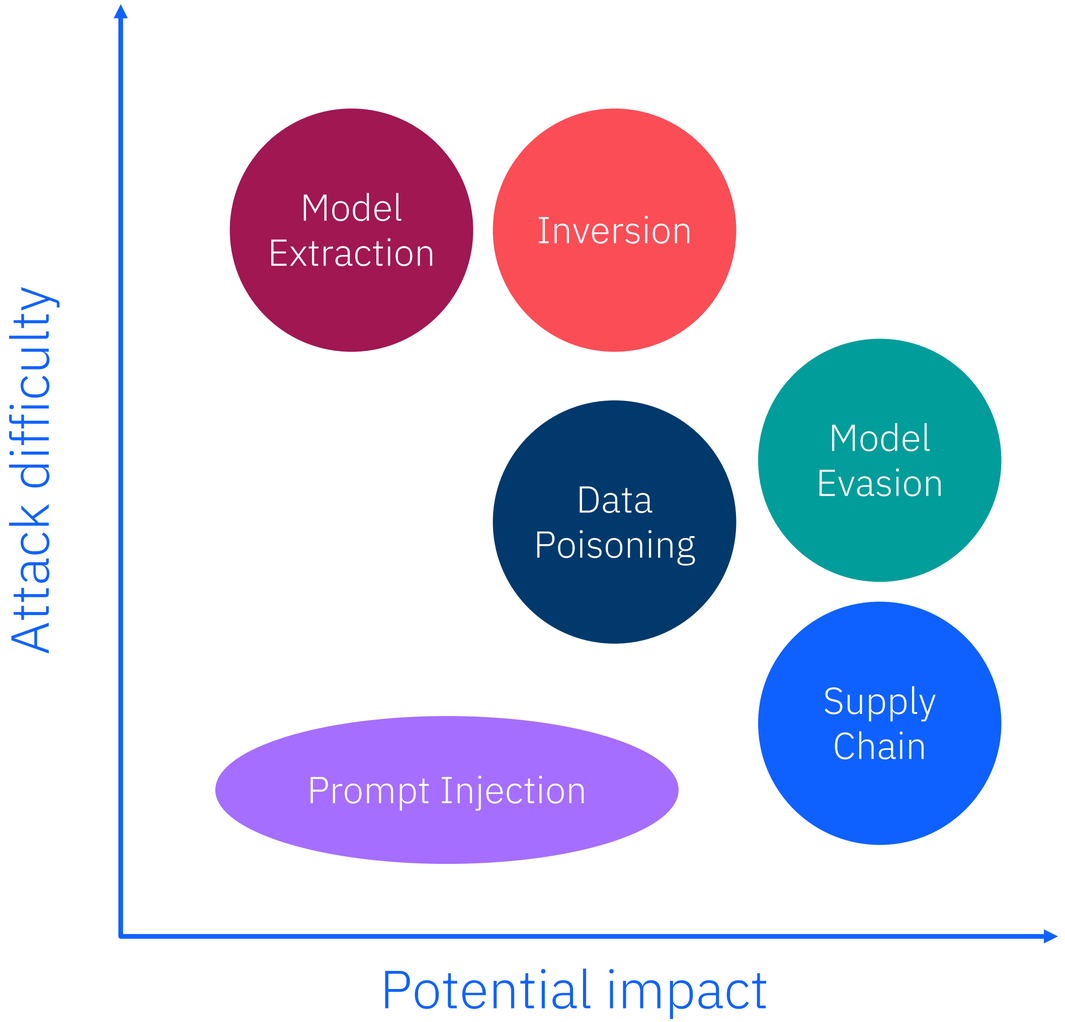
Attackers target...

Sensitive data being centralized and accessed for training

New vulnerabilities in apps being built in a brand-new way

Model inferencing to hijack or manipulate behavior of the model

Understanding adversarial risks to AI



Prompt Injection

Manipulate AI models into performing unintended actions, by dropping guardrails and limitations put in place by the developers

Data Poisoning

Changing the behavior of AI models by altering the data used to train them

Model Evasion

Circumventing the intended behavior of an AI model by crafting inputs that trick them

Model Extraction

Steal a model's behavior by observing the relationships between inputs & outputs

Inversion Attacks

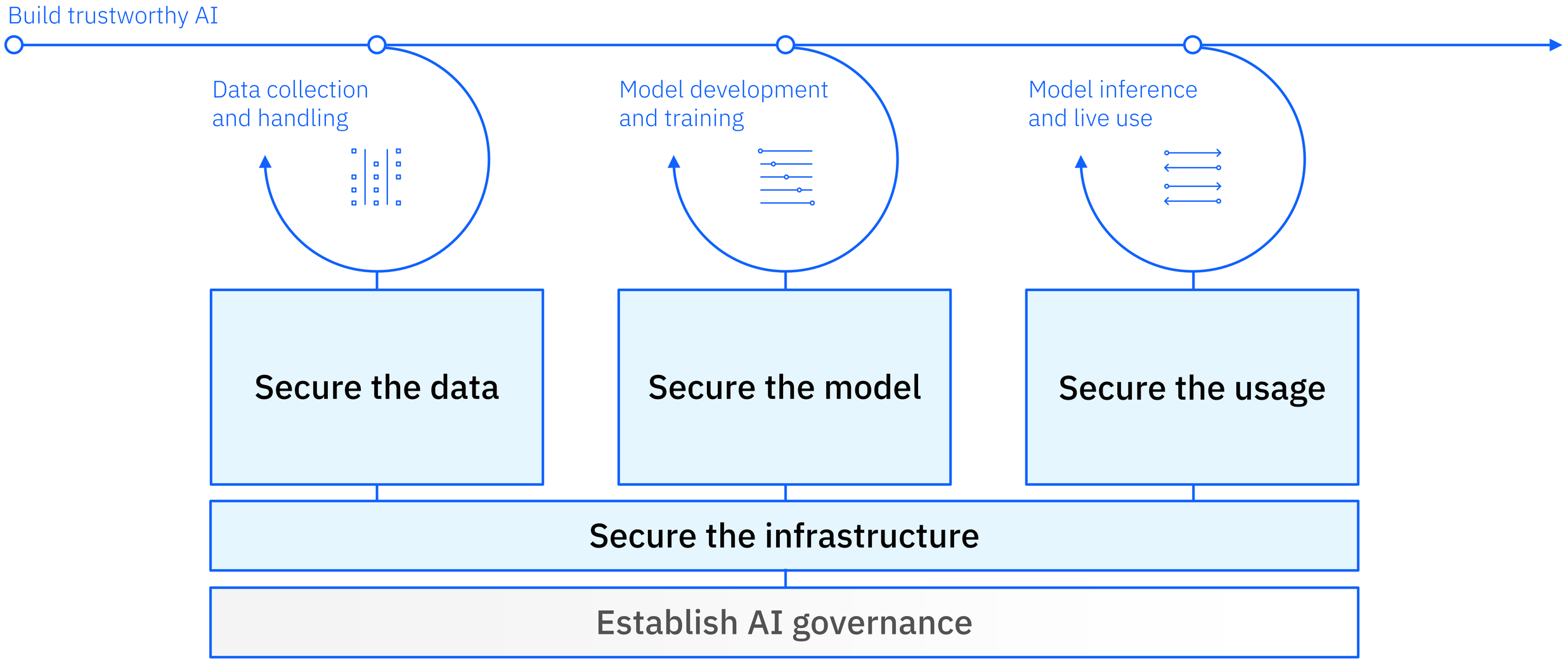
Reveal information on the data used to train a model, despite only having access to the model itself

Supply Chain Attacks

Generate harmful models that hide malicious behavior, or target vulnerabilities in systems connected to the AI models.

What you need to do

Security for AI framework



Recommendations to defend against risks to AI

1

First, understand how your organization is utilizing AI today: for which use cases, in what applications, via which service providers, and serving which user cohorts. Then quantify the associated sources of risk.

2

Utilize IBM's Securing AI framework to implement best practices across the AI lifecycle, from robust data handling, to securing the supply chain, through to detecting & responding to attacks in deployment.

3

Perform regular security audits, penetration testing, and red teaming exercises to identify and address potential vulnerabilities in the AI ecosystem and connected apps.

4

Perform cybersecurity awareness activities and education, particularly as they relate to AI as a new attack surface – targeting all stakeholders involved in the development, deployment and utilization of AI models.

Thank you

IBM products and services are warranted according to the terms and conditions of the agreements under which they are provided.

IBM’s statements regarding its plans, directions, and intent are subject to change or withdrawal without notice and at IBM’s sole discretion. Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision.

The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract.

The development, release, and timing of any future features or functionality described for our products remains at IBM’s sole discretion.
Performance data contained herein was generally obtained in controlled, isolated environments. Customer examples are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual performance, cost, savings or other results in other operating environments may vary. References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business.

All materials and discussions are provided for informational purposes only, and are neither intended to, nor shall constitute legal or other guidance or advice to any individual participant or their specific situation.

It is the customer’s responsibility to ensure its own compliance with legal requirements and to obtain advice of competent legal counsel as to the identification and interpretation of any relevant laws and regulatory requirements that may affect the customer’s business and any actions the customer may need to take to comply with such laws. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the customer is in compliance with any law.