Securing AI on Z: Addressing Emerging Threats and Building Trustworthy Systems

Gregg Arquero
IBM Senior Software Engineer – Z Security
gmarquer@us.ibm.com

Elijah Swift
IBM Software Engineer – z/OS Secure Engineering
elijah.swift@ibm.com

Enterprise Grade AI On IBM Z

Infusing AI into business-critical applications

- Minimize latency for real-time insights
- Utilize security umbrella across functions
- Avoid risks and cost of data copies
- Address process resiliency concerns
- Leverage existing AI models and skills

Applying AI for operational excellence

- Accurately identify emerging problems across the environments
- Diagnose & fix problems quickly in dynamic and complex environments
- Resolve swiftly with intelligent automation
- Address skill gaps and assist skill growths

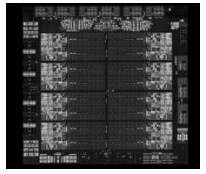
AI Approaches In The Modern Enterprise

AI acceleration technologies to address business requirements

Predictive AI

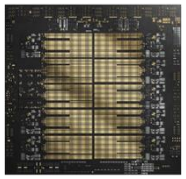
- Available on z16 with Telum and z17 with Telum II
- Inferencing in every transaction
- Speed and scale are critical
- Low latency and high throughput
- Precision and accuracy matter

IBM Telum™



OR

IBM Telum II™

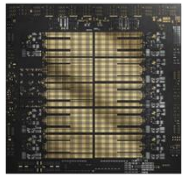


Predictive AI for real-time, high-volume transaction processing unlocking business growth

Multiple Model AI

- Available with z17 and Telum II for most use cases, can add Spyre for more complex encoder LLMs
- Advanced AI with higher predictive accuracy and expanded insights
- Generates high value by combining the power of traditional AI models and LLMs in real-time on every transaction

IBM Telum II™



Predictive AI and Encoder LLMs for advanced real-time, high-volume transaction processing unlocking business growth with optimized outcome

Generative AI (GenAI)

- Available with z17 + Spyre Accelerator
- Larger LLMs require more compute
- Performance measured in tokens per second
- Security is a priority to ensure client data and models are protected
- Energy efficiency is essential to scale workloads responsibly

IBM Spyre™ Accelerator

Decoder LLMs to accelerate and scale Generative AI workloads

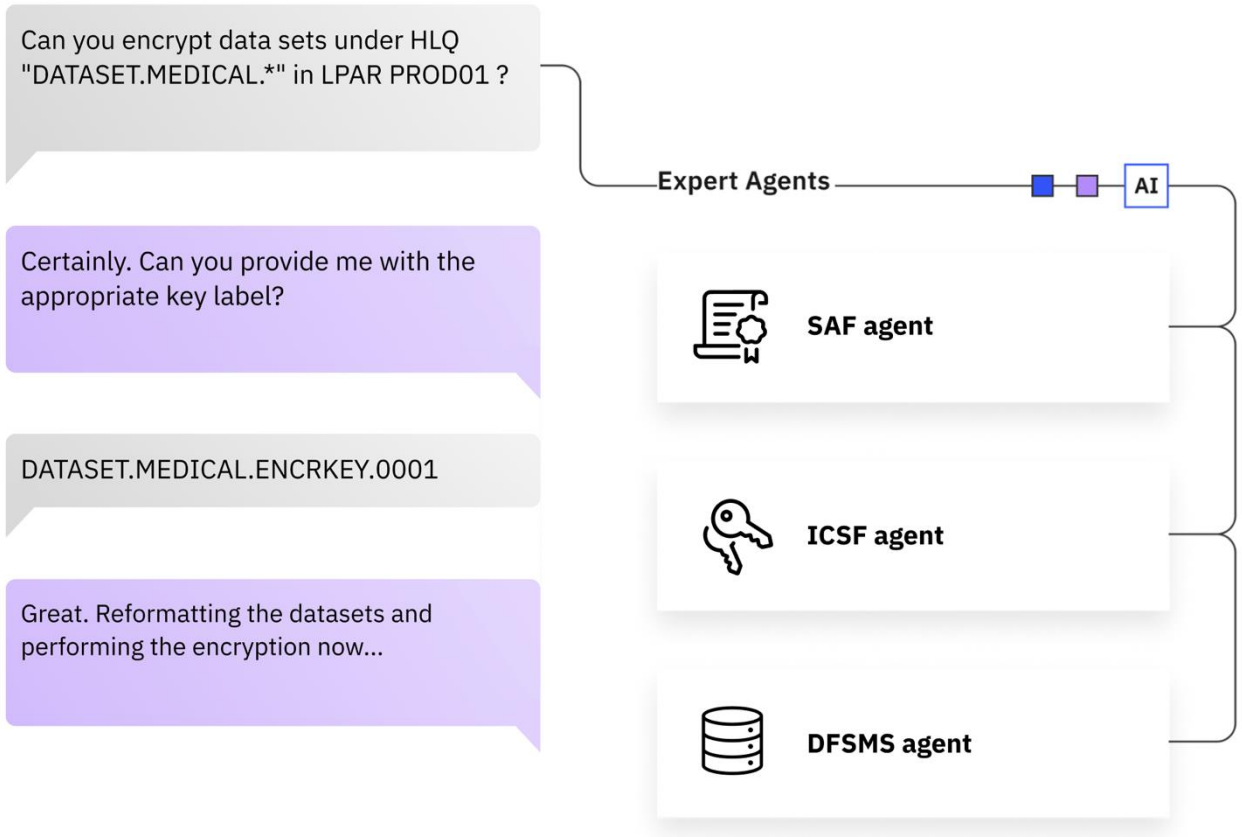
INCREASING COMPUTE & MEMORY BANDWIDTH

Agentic AI – The Next Generation of AI

AI proficient in a domain that can act as part of a virtual workforce

In Agentic AI:

- Agents use tools and collaborate with other agents
- Plan and act on tasks in response to a question or prompt
- Reflect on the results of its actions
- Learns iteratively, refining its approach to better align with its defined objectives



The AI oversight gap

New global research from IBM and Ponemon Institute reveals how AI is greatly outpacing security and governance in favor of do-it-now adoption. The findings show that ungoverned AI systems are more likely to be breached and more costly when they are.

4.4M

The global average cost of a data breach, in USD, a 9% decrease over last year—driven by faster identification and containment.

97%

Share of organizations that reported an AI-related security incident and lacked proper AI access controls.

63%

Share of organizations that lacked AI governance policies to manage AI or prevent the proliferation of shadow AI.

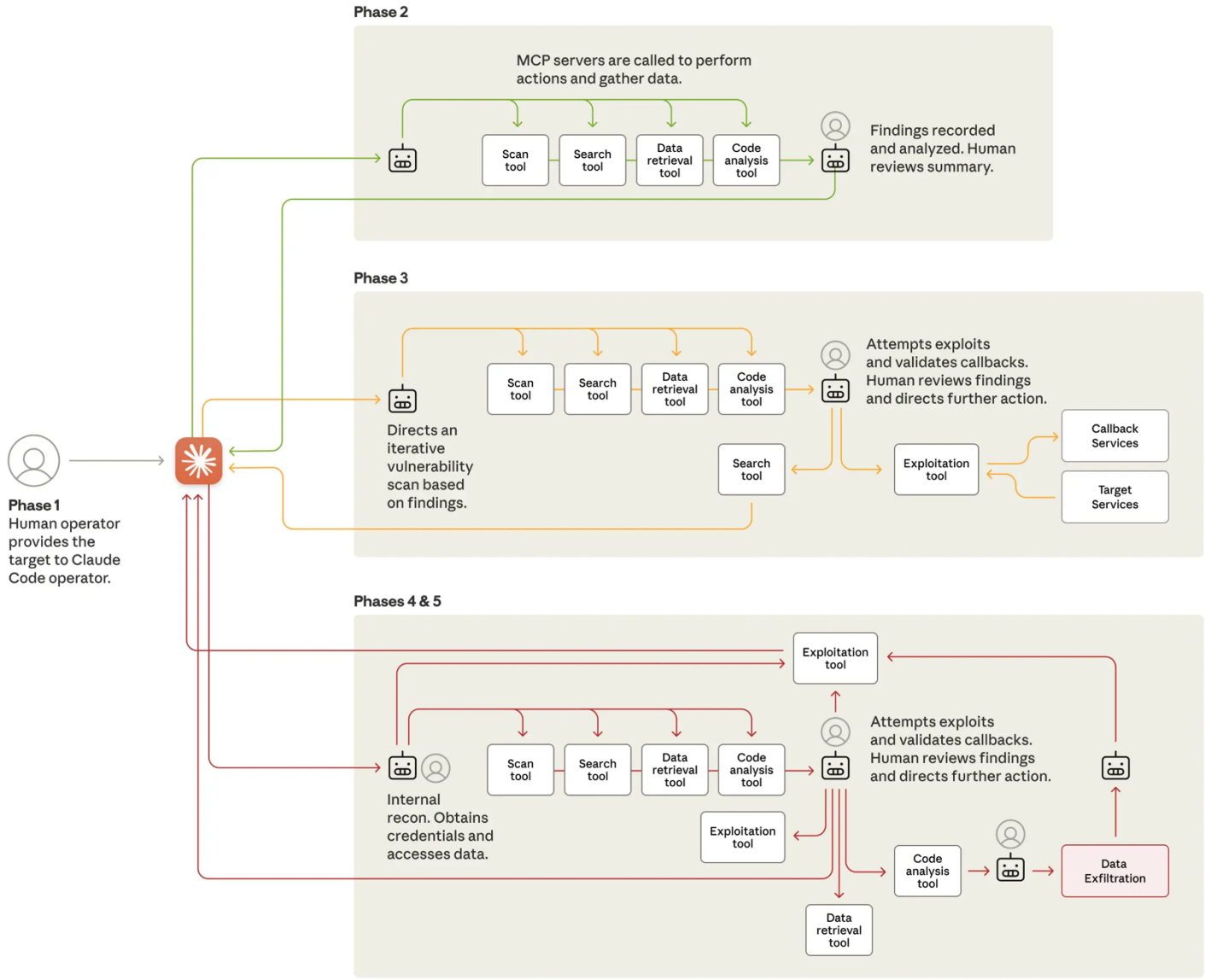
1.9M

Cost savings, in USD, from extensive use of AI in security, compared to organizations that didn't use these solutions.

https://www.ibm.com/reports/data-breach

AI-Orchestrated Cyber Attack

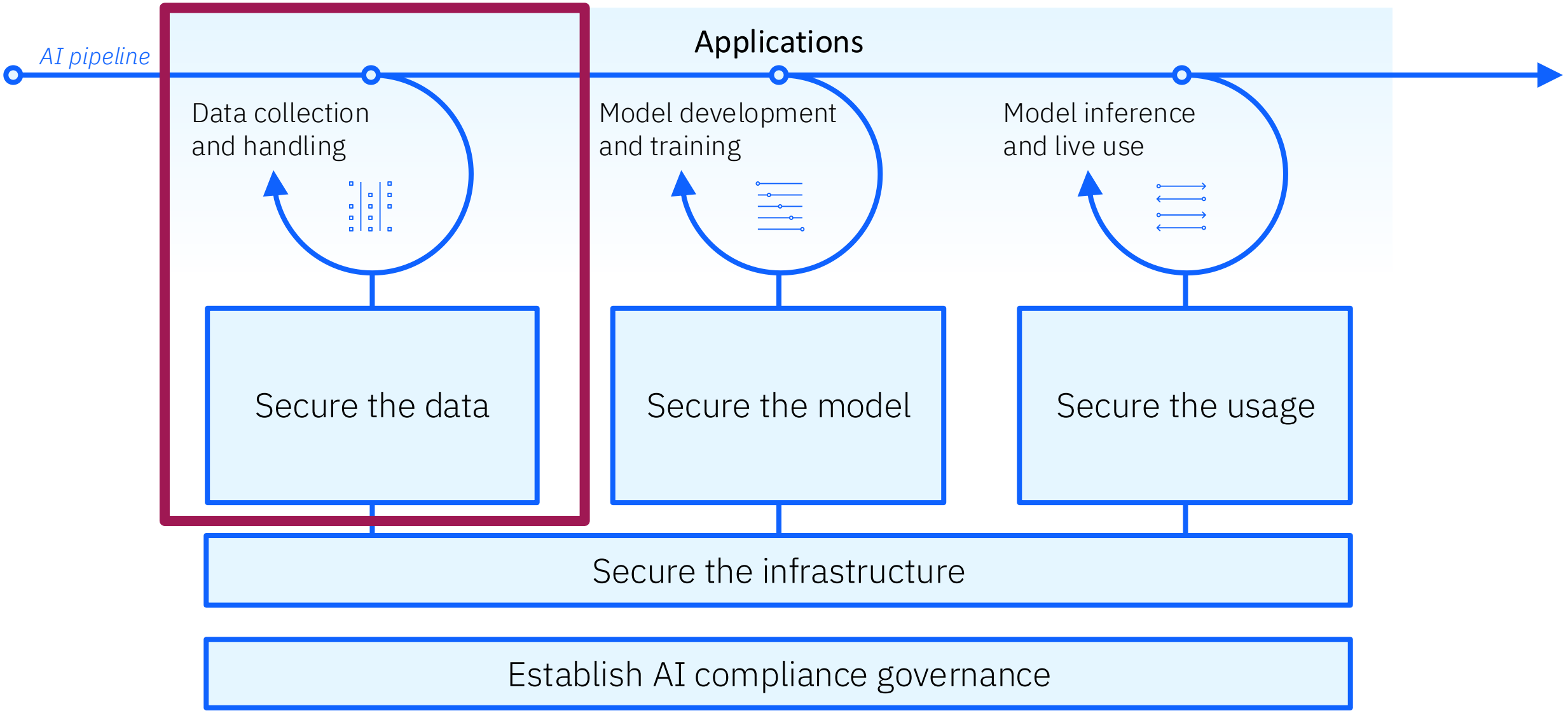
- In Sept 2025, threat actors used Claude Code and Model Context Protocol (MCP) tools to carry out attacks and data exfiltration with minimal human supervision.
- Demonstrates that the barrier to performing sophisticated cyberattacks have dropped substantially.
- Emphasizes the need to integrate AI into cybersecurity practices while ensuring robust security when leveraging AI capabilities.



Source: [Anthropic report]

Attackers will **Target** AI	AI should be treated as a **new attack surface,** with new detection and response strategies required for model evasion, extraction, 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	Phishing 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

A Framework For AI Security



https://www.ibm.com/blog/announcement/ibm-framework-for-securing-generative-ai/

Securing the data

AI may introduce new challenges, and reasons for it, but data security is still the same.

Tools like IBM Pervasive Encryption can protect and encrypt data.

Good security environmental practices like Zero Trust, Role Based Access Control and healthy auditing can help maintain this as well.

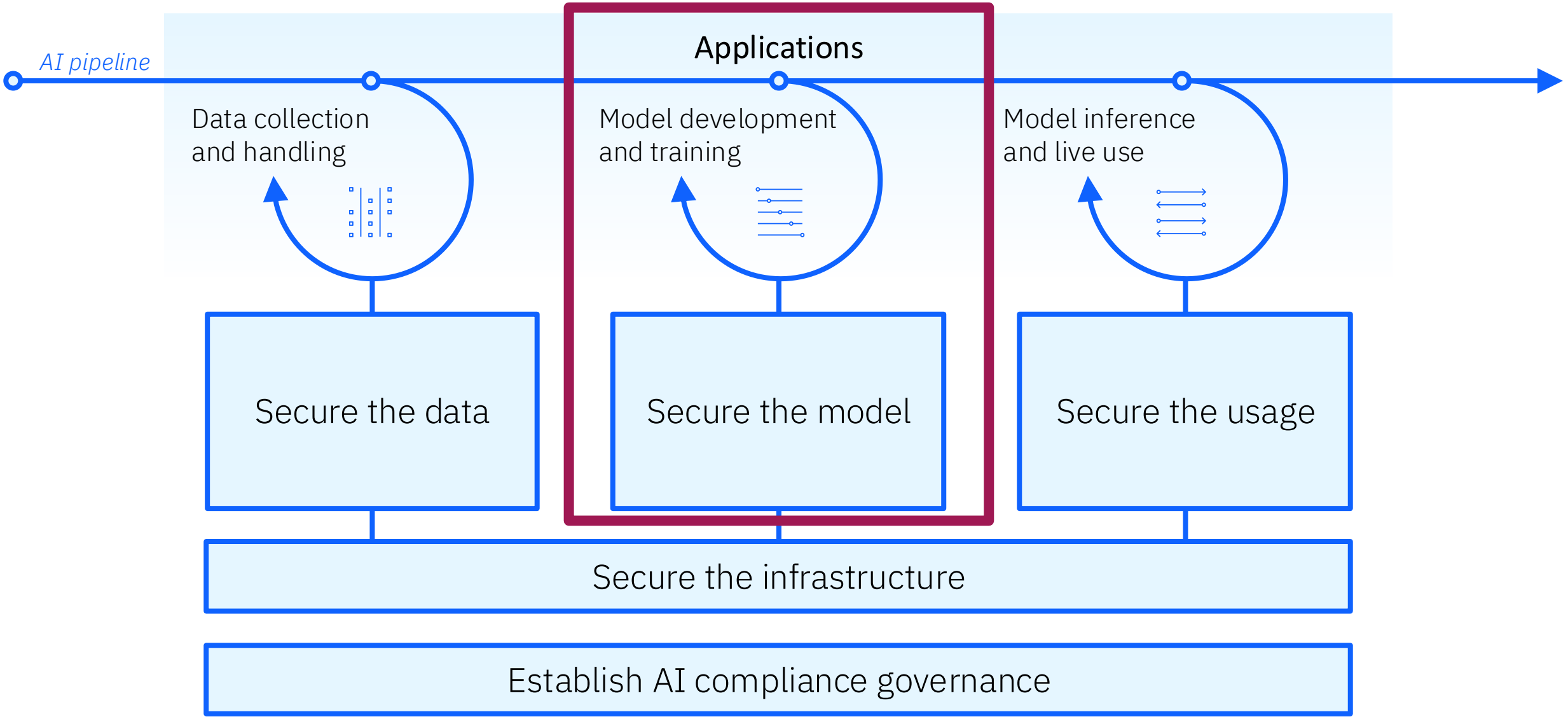
Securing the data

AI requires a lot of data to draw from and to use that can be a tantalizing target for attacks.

Conventional fears for bulk data like data exfiltration and leakage are still relevant.

Data integrity can also provide an attack vector into the AI model itself through Data Poisoning.

A Framework For AI Security



https://www.ibm.com/blog/announcement/ibm-framework-for-securing-generative-ai/

Securing the model

An AI model is just another piece of custom software through an existing supply chain.

What does this process look like for your organization?

– Thoroughly vet the provider

– Consider security or penetration testing

– Containerize or isolate workloads

Securing the model

AI models are complex and expensive to produce; most groups are using ones developed by outsiders.

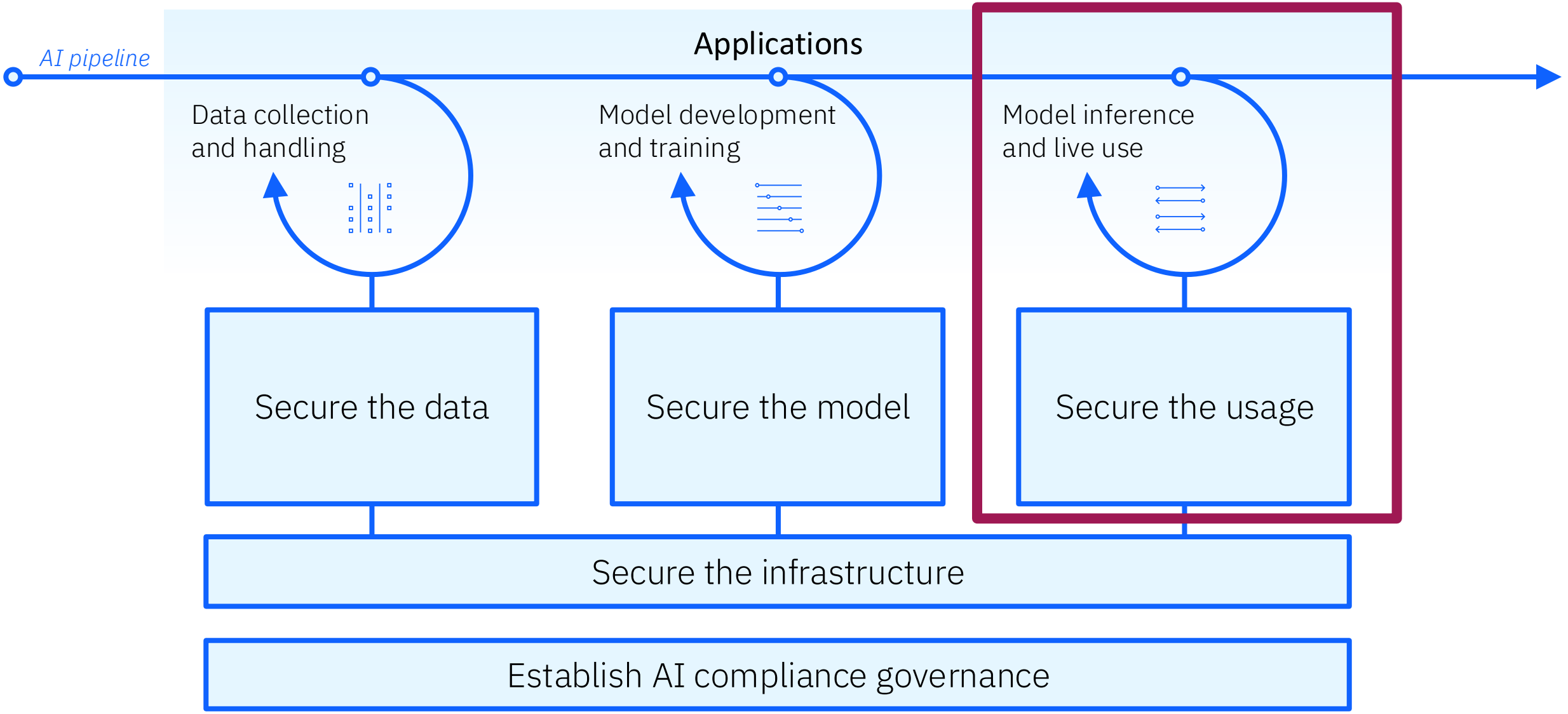
The functions of a model are a black box by nature, and integrating them can grant them sensitive things

- Sensitive data
- API Paths/Credentials
- Executing code in your environment

AI models can even be a malware delivery mechanism

https://arxiv.org/pdf/2409.19310

A Framework For AI Security



https://www.ibm.com/blog/announcement/ibm-framework-for-securing-generative-ai/

Enterprise Security for Autonomous Agents

Risk	Solution
Loss of accountability and compliance violations	Establish agent accountability with unique identities, just-in-time access, and auditability of all agent actions
Data breaches and security incidents	Protect agent and data flows from injection, leakage, and unauthorized access with prompt-level controls, data minimization, and auto- ready compliance
Cyber-attacks and business disruption	Continuously harden agents against adversarial threats – real time detection, containment, and response of unsafe / malicious agent actions
Enterprise security risk and non-compliance	Continuously monitor agent risk and compliance – integrate observability, detect threats, and enforce security policies across agent operations

Source: Architecting secure enterprise AI agents with MCP

Evolving Security From Application Security To Agent Security

Security Domain	Traditional Application Security	Agentic Application Security
Identity & Access	User authentication, RBAC, session management	Ephemeral agent identities, JIT tokens, explicit delegation, accountability
App/Agent & Data Security	Input validation, encryption, secure storage, API protection, tokenization	Information flow protection, prompt/ input security, context isolation, data lineage
Continuous Hardening	SAST,DAST,RASP, vulnerability scanning, CI/CD pen testing, runtime protection	Red/blue/purple teaming, continuous sandboxing/isolation/ hardening
Security Risk & Compliance	Compliance frameworks, audit logs, vulnerability management, GRC integration	Traceability, provenance, continuous monitoring, drift detection, GRC integration

IBM Granite Guardian

IBM Granite Guardian models provide AI safety and risk-detection guardrails for LLMs and agentic AI systems.

Prompts and responses are evaluated against defined security, safety, and quality criteria.

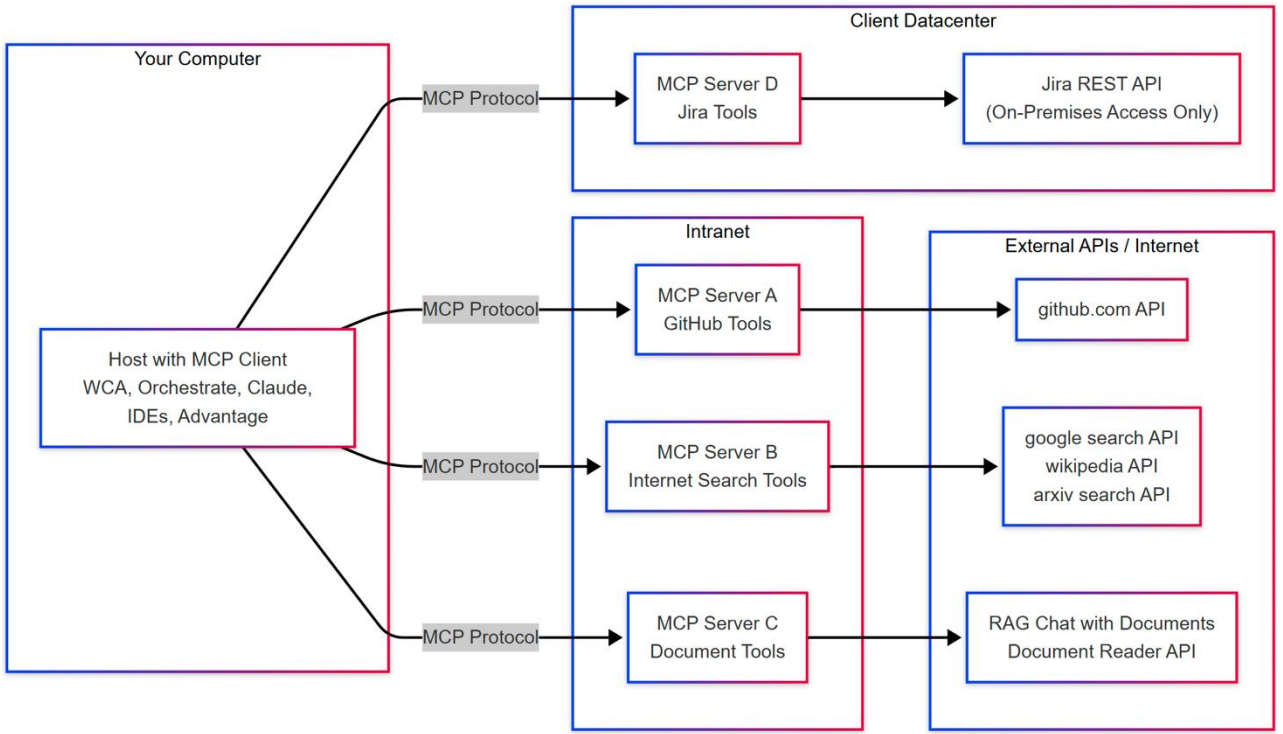
https://github.com/ibm-granite/granite-guardian

Risk	risk_name	Prompt	Response	Definition
Harm	harm	✅	✅	▶ Content considered universally harmful
Social Bias	social_bias	✅	✅	▶ Systemic prejudice against groups
Profanity	profanity	✅	✅	▶ Use of language considered offensive or socially unacceptable in a given context
Sexual Content	sexual_content	✅	✅	▶ Material explicitly related to sexual activities, anatomy, or desires
Unethical Behavior	unethical_behavior	✅	✅	▶ Actions that violate moral or professional standards
Violence	violence	✅	✅	▶ Promoting or describing physical harm to individuals or groups
Harm Engagement	harm_engagement	✅	✅	An engagement or endorsement with any requests that are harmful or unethical
Evasiveness	evasiveness	✅	✅	Avoiding to engage without providing sufficient reason
Jailbreaking	jailbreak	✅		▶ Deliberate circumvention of AI systems' built-in safeguards or ethical guidelines
RAG Safety - Groundedness	groundedness		✅	▶ The LLM response includes claims, facts, or details that are not supported by or directly contradicted by the given context
RAG Safety - Context Relevance	relevance	✅		▶ The retrieved or provided context fails to contain information pertinent to answering the user's question or addressing their needs
RAG Safety - Answer Relevance	answer_relevance		✅	▶ The LLM response fails to address or properly respond to the user's input
Agentic Safety - Function Calling Hallucination	function_call		✅	▶ The LLM response contains function calls that have syntax or semantic errors based on the user query and available tool definition

Introducing Model Context Protocol (MCP)

MCP Overview

- Open protocol introduced by Anthropic in **November 2024** to standardize tool calling.
- Enables a consistent interface to define how Agents and applications **discover, invoke, and interact with tools** and other context (prompts, resources)
- **Widely adopted:** 15,000 community servers developed since launch; wide adoption by IBM, Microsoft, Google, OpenAI, AWS, SalesForce, etc.
- **Evolving:** the standard is rapidly involving to improve security, granular access controls, transparent tool usage, and user interaction.



MCP gained rapid adoption — but the ecosystem evolved unevenly. **Many tools only partially implement the spec**, and integration challenges remain:

- **Existing tools**: REST endpoints must be rewritten to become MCP-compliant.
- **Protocol and security inconsistency**: some use JWT authentication, others OAuth2, many use nothing at all –while most servers are still developed to stdio / SSE transport (instead of the newer streamable HTTP).

MCP Threat vectors

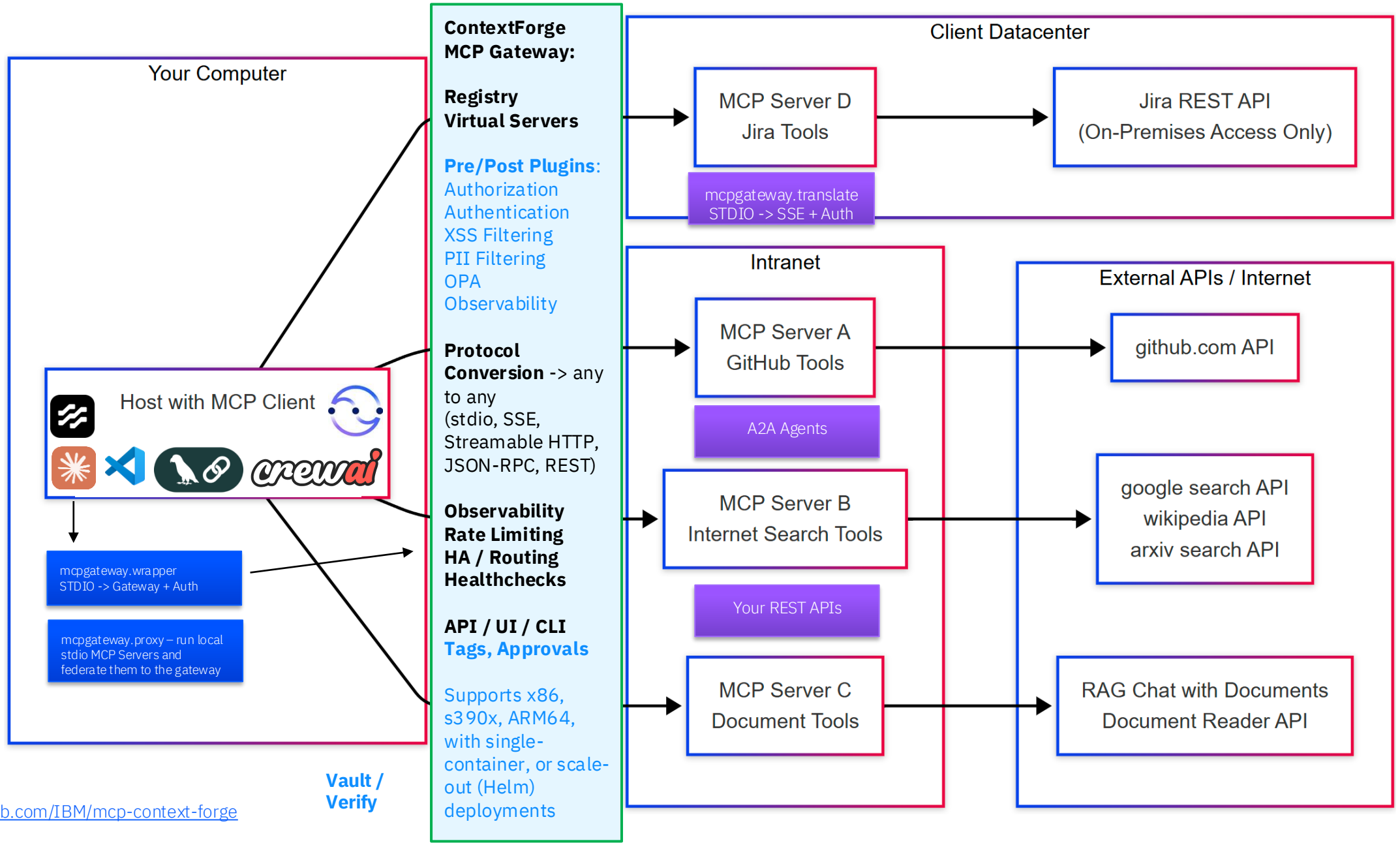
Model Context Protocol (MCP) enables AI agents to interact with real-world resources dynamically.

This capability shifts the security boundary and how attacks unfold.

MCP introduces several unique threats that must be assessed:

- **Tool Poisoning:** Malicious tool names, descriptions, or configurations can leak to data leaks or system compromise
- **Identity Spoofing:** Weak or misconfigured authentication can allow attackers to impersonate a trusted user
- **Resource Content Poisoning:** Malicious instructions embedded within data sources may lead to data exfiltration or prompt injection
- **Insecure human-in-the-loop:** Missing or insufficient human oversight can allow an agent to take risky or unauthorized actions

ContextForge: Enterprise AI Gateway & Security for MCP & A2A



https://github.com/IBM/mcp-context-forge

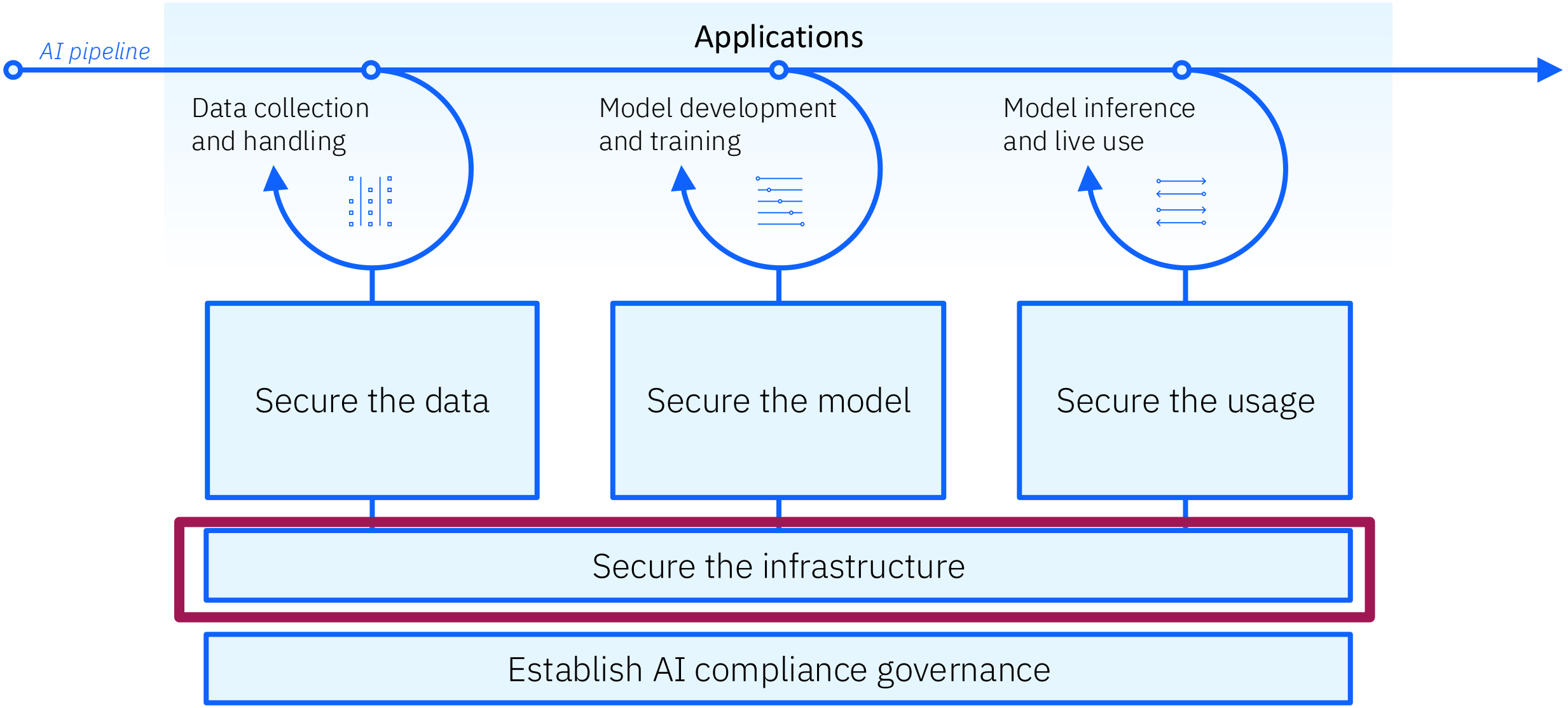
AI Threat Modeling Frameworks

AI introduces new attack surfaces across data, models, agents, and toolchains.

There are several high-quality threat modeling frameworks that can help you assess AI Risk:

- MAESTRO (Multi-Agent Environment, Security, Threat, Risk, and Outcome)
- NIST AI Risk Management Framework
- MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems)

A Framework For AI Security



https://www.ibm.com/blog/announcement/ibm-framework-for-securing-generative-ai/

Trusted Execution Environments

- Trusted Execution Environments (TEEs) are hardware-enforced isolated environments that protect code and data in use from the OS, Hypervisor, and privileged insiders

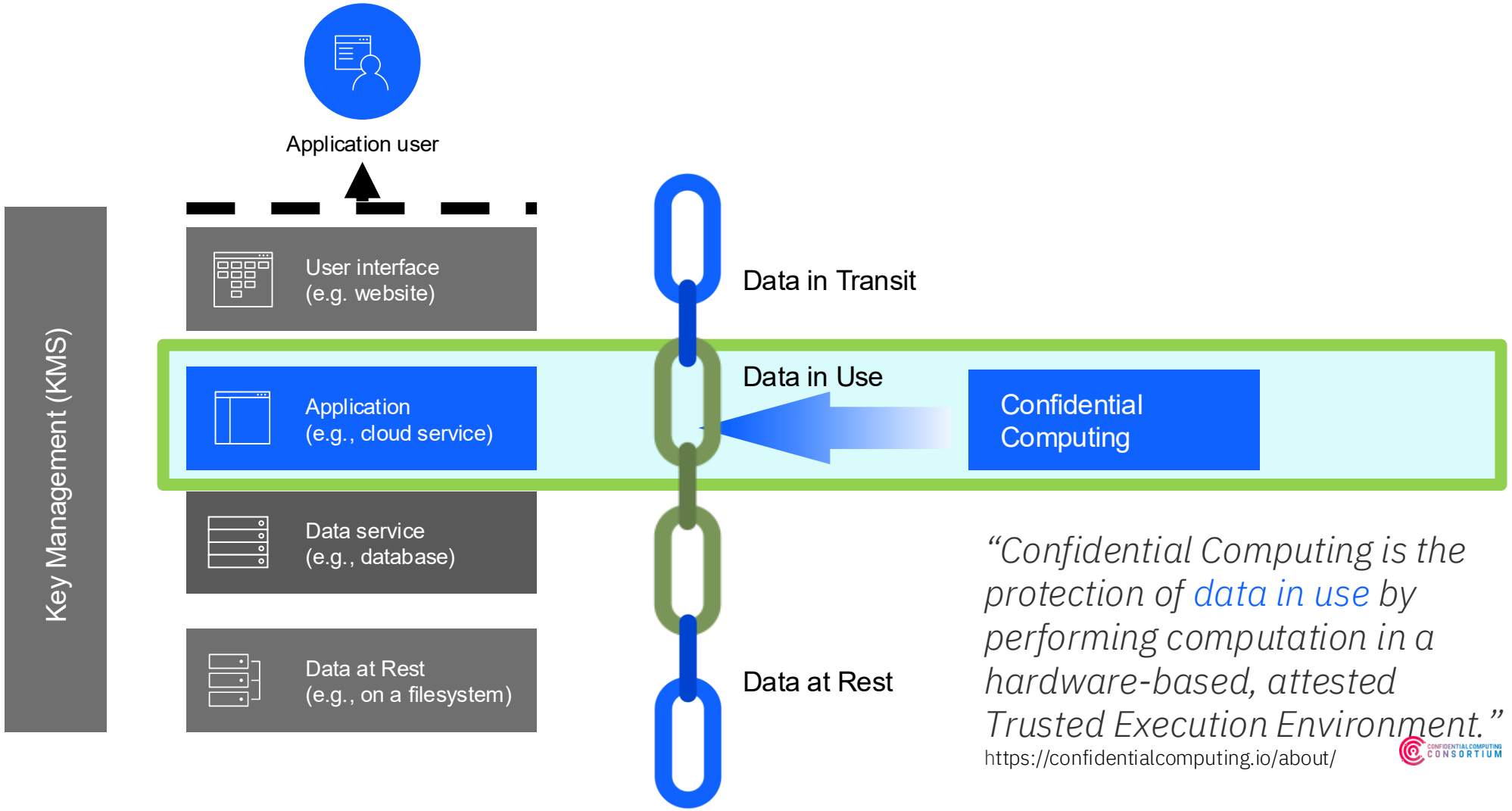
- Remote Attestation provides cryptographic proof that workloads are running in a verified, TEE with known configuration

- TEEs can provide stronger isolation and protection against tampering of MCP servers and AI models during execution

- Remote attestation verifies that only the approved MCP servers and clients can receive sensitive data

- TEEs protect data in use and must be complemented with encryption for data in transit and at rest

Confidential Computing is about 'Data in Use'



"Confidential Computing is the protection of data in use by performing computation in a hardware-based, attested Trusted Execution Environment."

https://confidentialcomputing.io/about/

IBM Secure Execution for Linux (SEL)

The Confidential Computing Technology for IBM Z and LinuxONE

Available with IBM z15 and LinuxONE III

Protects data in use of SEL guests from access or tampering by

- malicious system or hypervisor administrators / intruders
- vulnerable or malicious versions of Linux/KVM software
- malicious guests hosted by the same Linux/KVM

A special trusted firmware (FW) component called Ultravisor (UV)

- maintains & protects the SEL guest state (CPU & memory)
- has access to the root of trust: the private host key

Image of SEL guest is integrity protected & encrypted

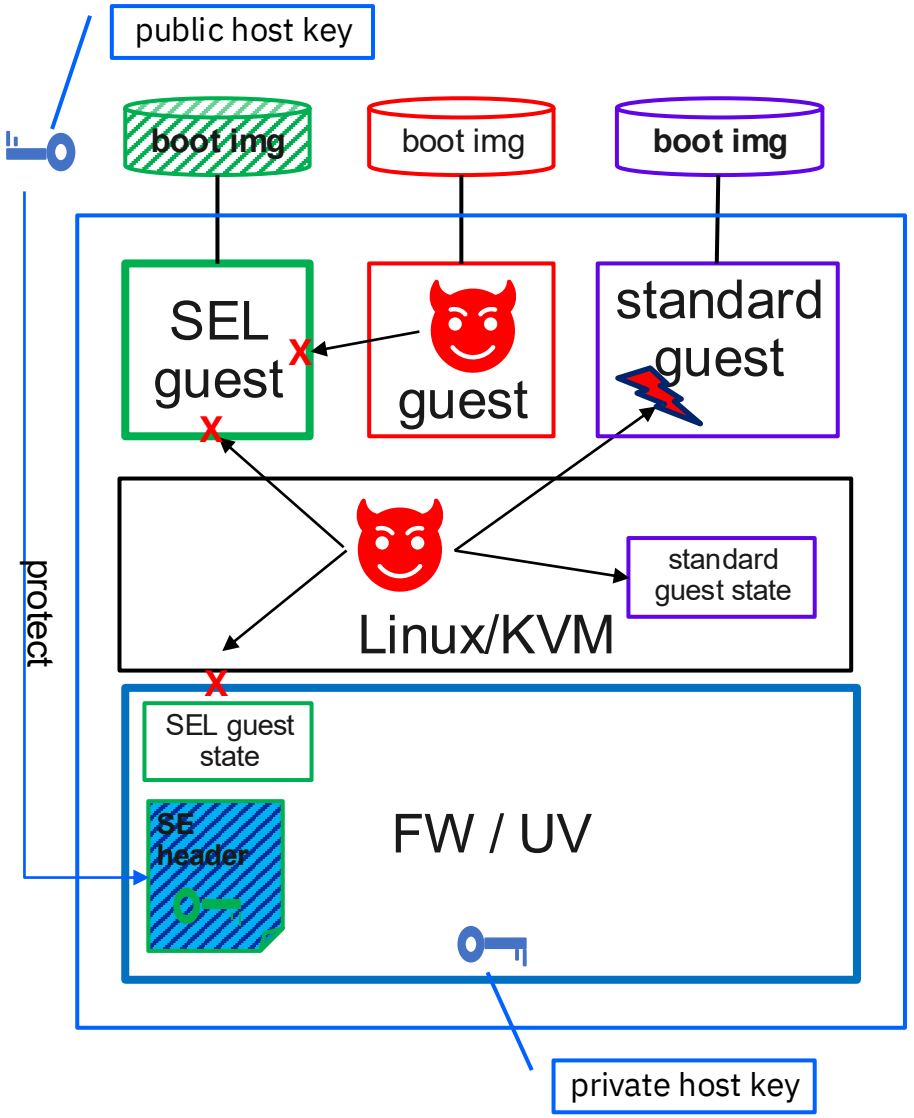
- image protection data is passed to UV in SE-header which is protected using a host key

Extensions for IBM z16 & LinuxONE 4:

- support for remote attestation
- hypervisor triggered (encrypted) dump
- support of Crypto Express accelerators and EP11 hardware security modules (HSMs)

Extensions for IBM z17 & LinuxONE 5:

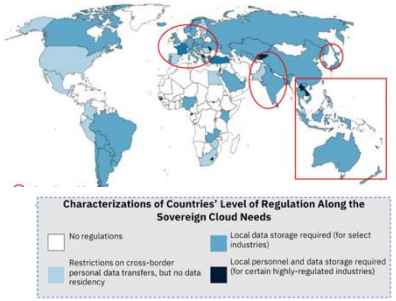
- retrievable secrets
- extended remote attestation
- host key info in hypervisor



Increasing regulations are changing the way client's need to protect their data

General Data Privacy Laws

Data Sovereignty Regulations



CONFIDENTIAL COMPUTING CONSORTIUM

Inflection Point

Data-in-use Protection Mandates

BANK OF ENGLAND PRUDENTIAL REGULATION AUTHORITY

Supervisory Statement (SS2/21) Chapter 7 "Data Security" mandates that firms protect data-in-memory.

Security Guidance for 5G Cloud Infrastructures: "For sensitive / regulated workloads, there is a desire to protect data- in- use from the underlying privileged system stack as well as physical access threats. Therefore, it is critical that data in memory has comparable protection to data-at-rest."

White House Executive Order on Improving the Nation's Cybersecurity: "The Federal Government must adopt security best practices; advance toward Zero Trust Architecture;…"

gematik Nationale Agentur für Digitale Medizin

Specification of „Healthcare Confidential Computing"
Confidential Computing is a mandatory requirement to ensure relyable the data privacy for personal medical cleartext data being processed outside of the direct responsibility by an actor in the healthcare ecosystem

National Cyber Security Centre

Principle 2: Asset Protection and Resilience – "Your data (and the assets storing or processing it) should be adequately protected."

DORA Digital Operational Resilience Act

2. (a) […] rules for the encryption of data at rest, in transit and, where relevant, in use, taking into account the results of the approved data classification […] If encryption of data in use is not possible, financial entities shall process data in use in a separated and protected environment […]

EBA EUROPEAN BANKING AUTHORITY

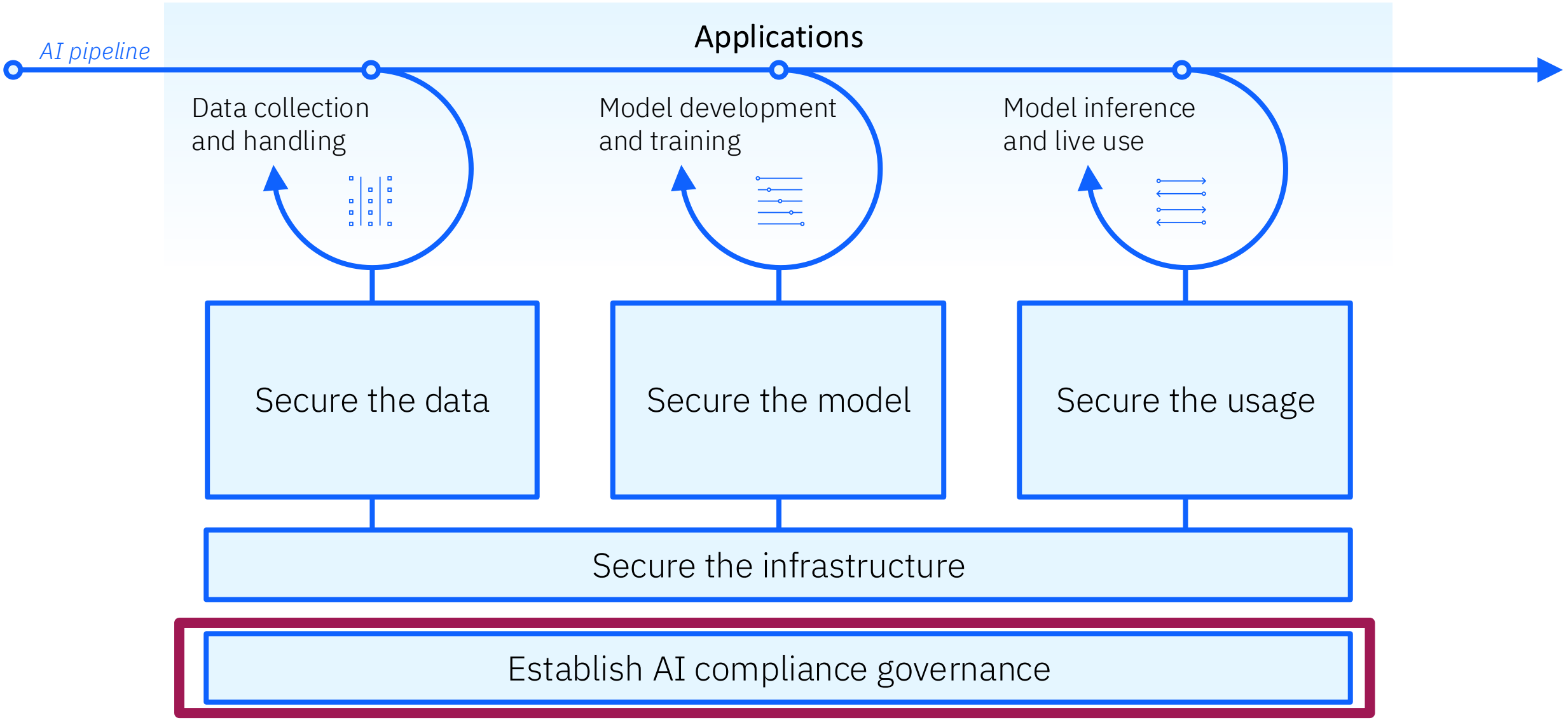
Final Report on the Guidelines of Outsourcing (12.2.68e): "When carrying out the risk assessment prior to outsourcing and during ongoing monitoring of the service provider's performance, institutions and payment institutions should, at least:… define and decide on an appropriate level of protection of data confidentiality….Institutions and payment institutions should also consider specific measures, where necessary, for data in transit, data in memory and data at rest, such as the use of encryption technologies in combination with an appropriate key management architecture;"

Federal Cybersecurity Reseach and Development Strategic Plan:**"…trustworthy interoperation must be supported by robust enforcement of secure interoperation policies including the use of secure separation and isolation techniques (e.g., trusted execution environments) for protecting data-in-use, securing cross-domain information sharing and flow, as well as securing execution of distributed workflows and transactions that span multiple domains or jurisdictions**

MAS Monetary Authority of Singapore

MAS Guidelines: FIs should implement appropriate data security measures to protect theconfidentiality and integrity of sensitive data in the public cloud, taking into consideration data-at-rest, data-in-motion and data in-use where applicable.

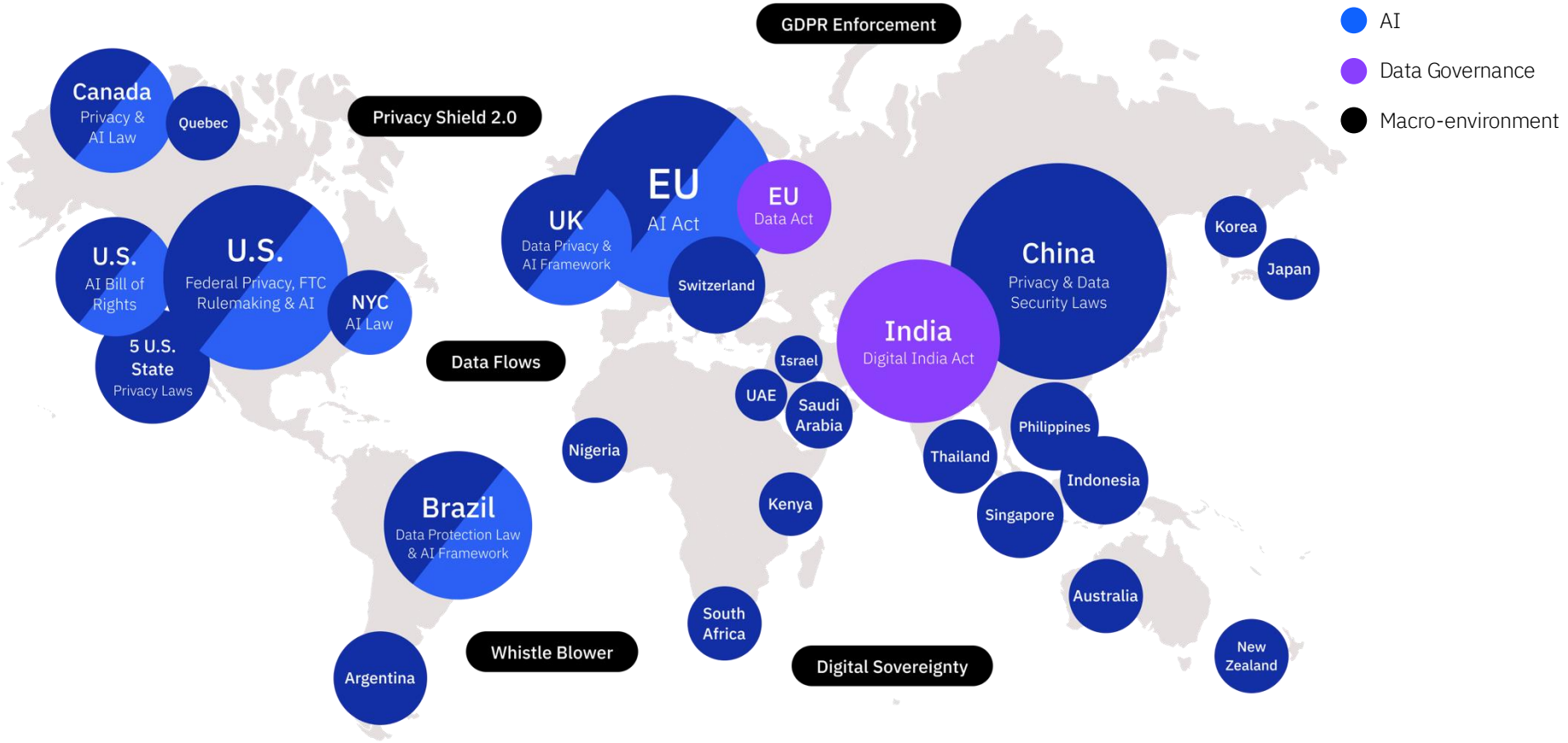
A Framework For AI Security



https://www.ibm.com/blog/announcement/ibm-framework-for-securing-generative-ai/

Globally, governments are evaluating AI, and developing regulations

Privacy and AI laws



71

countries have passed AI regulations and policies.

82

AI Initiatives in the U.S. including governance policies, regulation policies, incentive policies, and financial support policies.

63

AI Initiatives in the E.U. with first to enforce a non-compliant fine through the EU AI Act.

European Union's AI Act

Trust, Transparency and Governance in AI

AI Governance

The processes, standards and guardrails that help ensure AI systems and tools are safe and ethical.

Trust and Transparency in AI

Principles for Trust and Transparency

1. The purpose of AI is to augment — not replace — human intelligence
2. Data and insights belong to their creator
3. New technology, including AI systems, must be transparent and explainable

Pillars of Trust

Explainability: An AI system's ability to provide a human-interpretable explanation for its predictions and insights

Fairness: Equitable treatment of individuals or groups by an AI system — depends on the context in which the AI system is used

Robustness: An AI system's ability to effectively handle exceptional conditions, such as abnormalities in input

Transparency: An AI system's ability to include and share information on how it has been designed and developed

Privacy: An AI system's ability to prioritize and safeguard consumers' privacy and data rights

Technical Risks working with AI (Agentic AI , Gen AI, and ML Models)

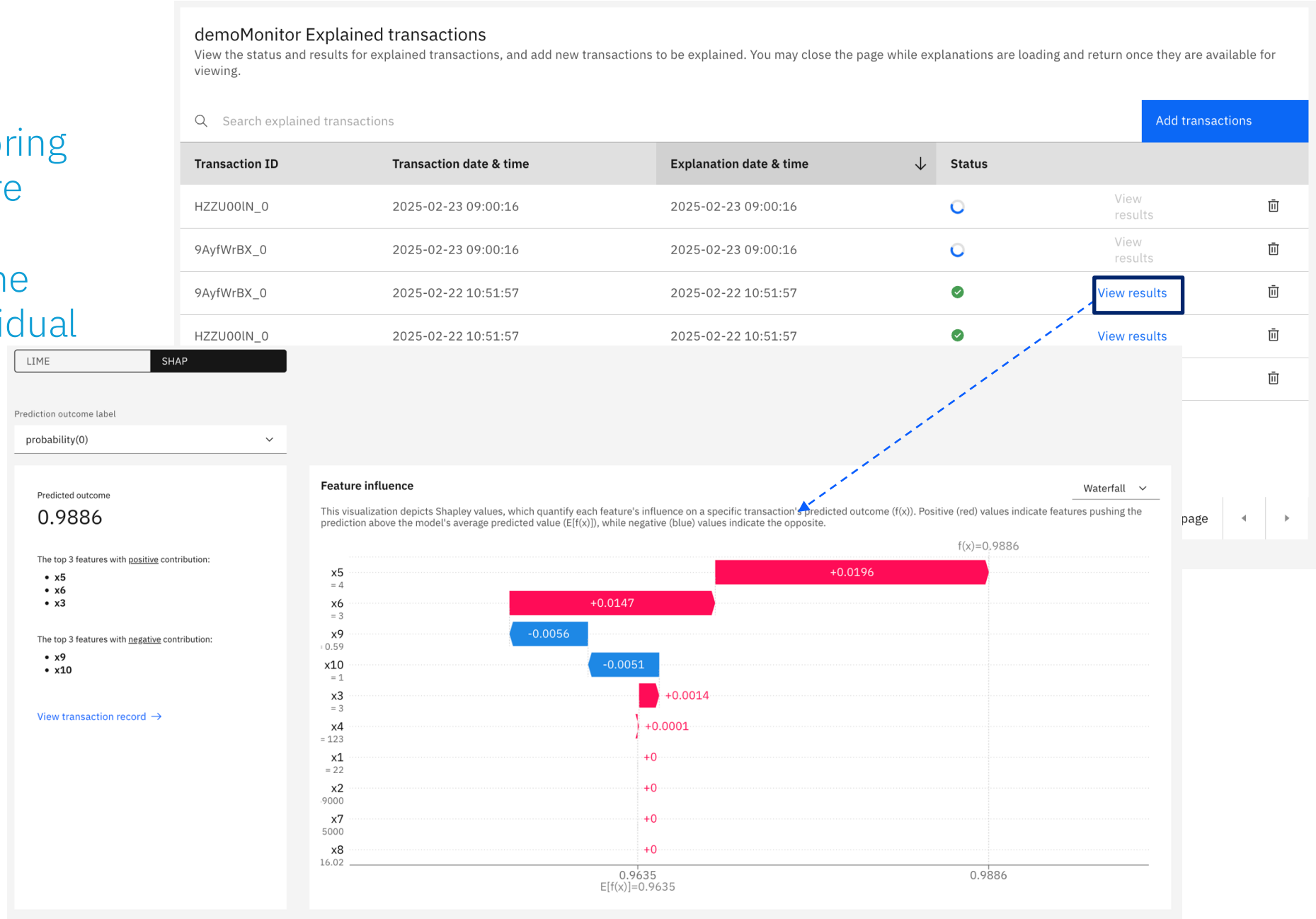
Accuracy	Unrepresentative data; Data contamination	Poor model accuracy
Fairness	Source data bias	Discriminatory actions; output bias; introducing data bias; decision bias
Privacy	PII in source data, Reidentification; Data privacy right alignment	Share PII with users and tools; Exposing PII; PII in prompt; Attribute inference attack; Membership inference attack
Transparency	Lacking data transparency; Uncertain data provenance	Lacking AI agent transparency
Explainability		Unexplainable output; Untraceable attribution; Unexplainable and untraceable actions
Robustness	Data Poisoning	Hallucination (answer or function call by agent); Attack on agent ext. resources; Prompt attacks (leaking, injection, priming overload etc.); Model attacks (jailbreaking; evasion, extraction .etc.)
Intellectual Property	Confidential info in data; Data usage right	IP /confidential data in prompt; Copyright infringement

Color code:
Traditional ML; GenAI & Agentic AI; Both

Addressing AI Risks (examples): Trustworthy AI – in Machine Learning for IBM z/OS

Explainability

Enables continuous monitoring of z/OS transactions that are scored by AI models and provides explanations on the model's output at the individual transaction level.



LIME

Model-agnostic
local, interpretable explanations - explains single predictions of a model

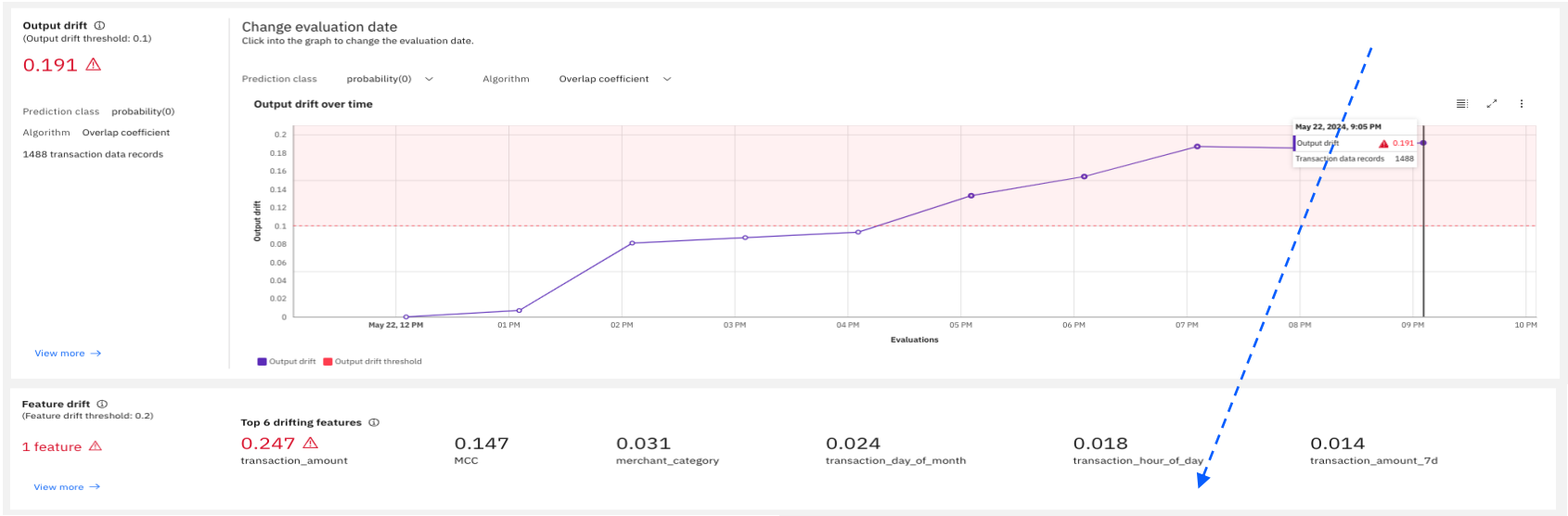
SHAP

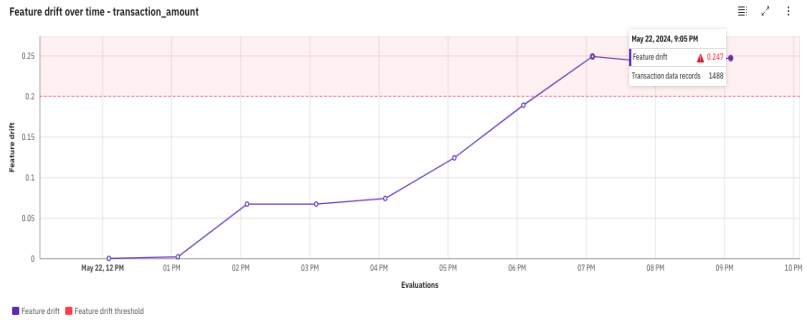
Model-specific
global explanations - the effect of a feature on the target variable

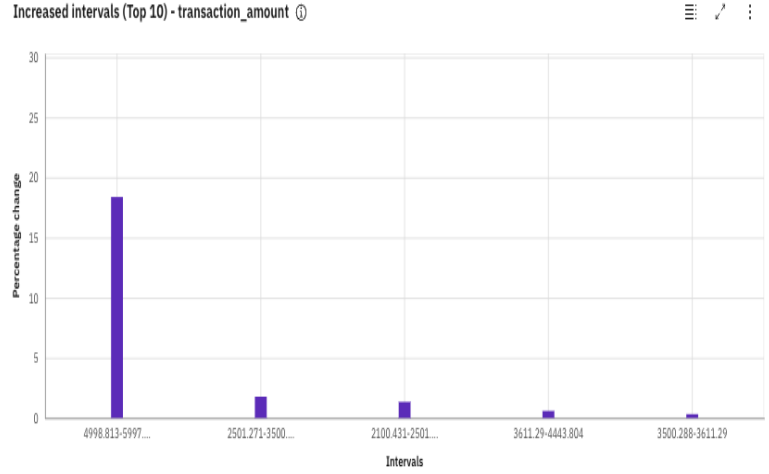
Addressing AI Risks (examples): Trustworthy AI – in Machine Learning for z/OS

Drift Detection

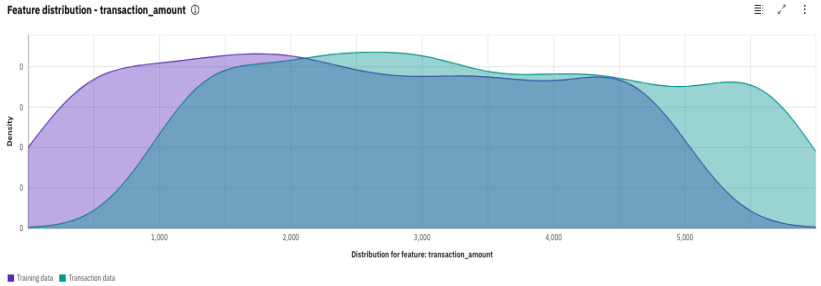
Enables monitoring and evaluation of both output drift and feature drift, providing data scientists with insights from different perspectives.







View the density distribution plots for the selected feature and the corresponding intervals that changed most significantly compared to the training data.



Output Drift

Measured by *Overlap coefficient* and *Total variation distance*

Feature Drift

Categorical features are measured by *Jensen-Shannon* and *Total variation distance*

Non-categorical features are measured by *Overlap coefficient* and *Total variation distance*

Key Takeaways

Securing AI starts with good security principles

- Role-Based Access Control
- Zero-Trust
- Auditing policies

Build on that with AI-specific knowledge

- Data poisoning
- Model theft
- Jailbreaking concerns

Stay informed

- Coalition for Secure AI
- OWASP Top 10 for Gen AI and LLMs

Your feedback is important!

SHARE
EDUCATE • NETWORK • INFLUENCE

Submit a session evaluation for each session you attend:

www.share.org/evaluation

Experience more with IBM

Visit us at the IBM Booth #113

After a full day of technical sessions, take a break with us!

Connect with our experts, snap a photo with the z17 Plexi or the latest Telum II, and get an up-close look at our Spyre Accelerator.

Come back each day for fresh topics and demos at our expert stations.

Think 2026

Join 5000+ senior business and technology leaders who are seizing the AI revolution to unlock unprecedented growth and productivity at Think 2026.

Find out more information using the QR code below.

IBM Digital Asset Haven

IBM Digital Asset Haven is the operational backbone for financial institutions and regulated enterprises entering the digital asset economy.

Find out more information using the QR code below.