Securing the Future: Navigating AI Vulnerabilities and Evolving Security Practices

FIRST VulnCon 2025

Lisa Bradley – Dell Technologies
Sr. Director, Product and Application Security
Product Remediation and Dependency Assurance (PRADA)

Sarah Evans – Dell Technologies
Distinguished Engineer, Global Chief Technology Office
Security Innovation Applied Research

DELL Technologies

Lisa Bradley

Sr. Director
Product & Application Security
PRADA – Product Remediation And Dependency Assurance
Time in Role: 5+ years
Time in Security Role: 12+ years
Location: Cary, NC

Academic Profile

GENESEO
THE STATE UNIVERSITY OF NEW YORK

Professional Profile

IBM

NVIDIA

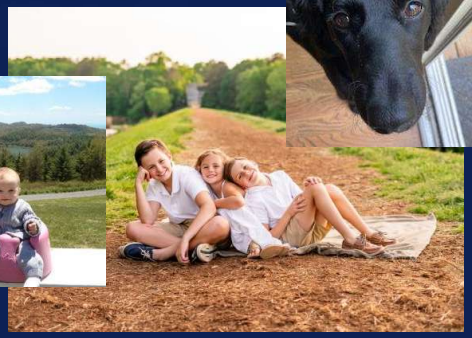
DELL Technologies

Academia Profile

University of Phoenix

CAMPBELL UNIVERSITY

Personal Profile






DELL Technologies

Sarah Evans

Distinguished Engineer, Global CTO R & D
Security Innovation Research
Time in Role: 5 years
Location: Denver, CO (remote)

Academic Profile

Missouri State

MIT Professional Education

Professional Profile

Missouri State

STOREE Commercial Construction

WELLS FARGO

United States Air Force Retired

DELL Technologies

Personal Profile

CompTIA A+

SPFDSEC Springfield, MO

ISACA

SAFe Certified

Springfield MO 2019 BSides

CISM Certified Information Security Manager An ISACA Certification

OpenSSF Open Source Security Foundation

DELL Technologies

AI and its growing role in software products

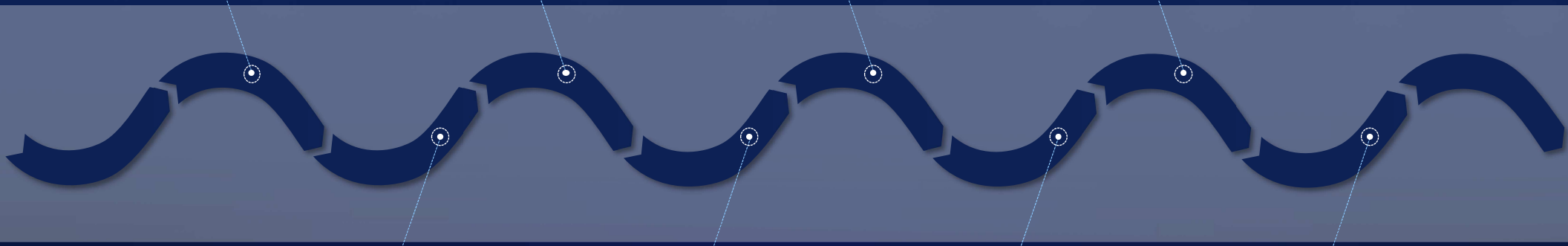
AI has evolved from simple rule-based systems to complex machine learning algorithms that can analyze vast amounts of data

AI algorithms ***analyze user behavior*** to provide personalized content and product recommendations

AI ***tools can generate code snippets and identify bugs***, speeding up the development process

AI can ***analyze historical data*** to make predictions about future trends and behaviors

AI will ***continue to be integrated into more software***, enhancing their capabilities



AI-powered chatbots and virtual assistants are being ***incorporated into products***

AI can ***automate testing processes***, ensuring higher quality software with fewer defects

AI systems can ***process data in real-time***, enabling quick decision-making based on current information

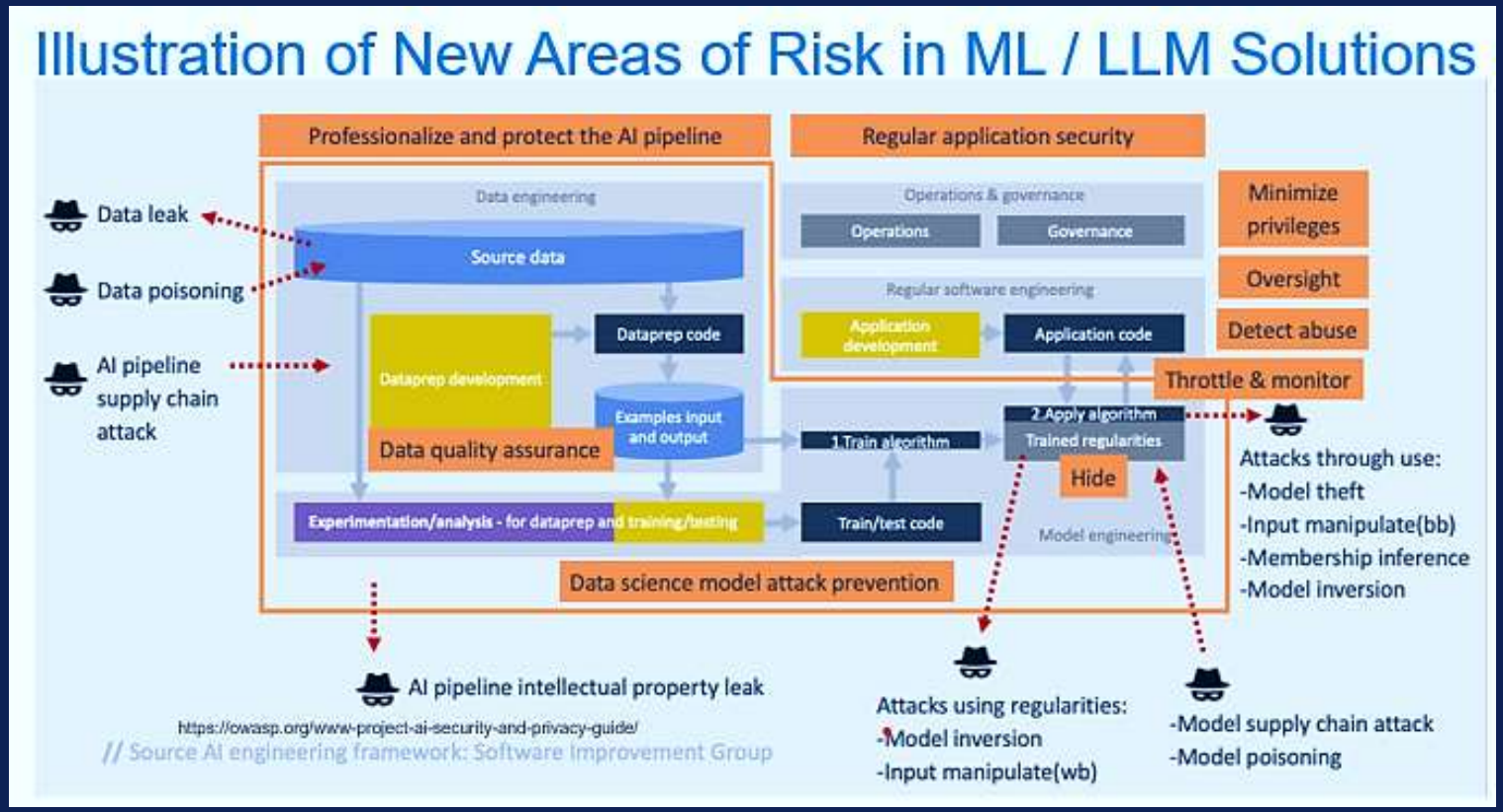
AI can even ***generate this slide!***

DELL Technologies

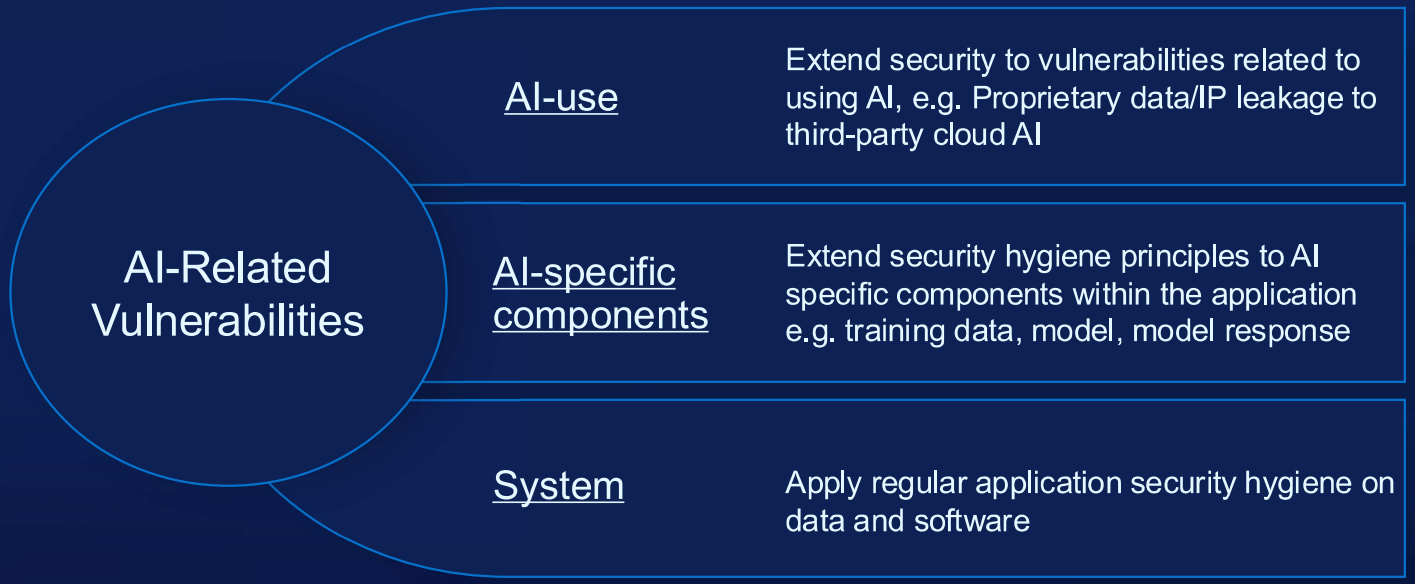
AI Threat Sources

Dell Security experts aggregate industry threats

- Dell Threat Library: AI Threats
- AI threat taxonomy and library



How should I start thinking about AI-Related Vulnerabilities?



AI-related vulnerabilities are:
1. weaknesses specific to the AI components within a system
2. the use of applications with AI

AI-related vulnerabilities can be in found in all stages of AI lifecycle and supply chain, including data, models, and deployment.

AI Security Threats

We believe in the future, AI will increase these core threats

Poor *system architecture/operational hygiene* will provide rich attack surfaces to threat actors

Poorly integrated *supply chain security* will cause an outsized impact of an upstream vulnerability on the AI technology on which we have all been innovating

Lack of *data management and classification*, including data inputs, data models and data outputs

Effective identification and classification of AI vulnerabilities

1. **Understanding common weaknesses in AI systems**
2. **Analyze how weaknesses can be leveraged as threats and attacks**
3. **Recognizing that exploitation is often a result of bad behavior in an AI system due to non-trusted data**

Attack Type Example: Data Poisoning

Type of cyberattack where an adversary intentionally manipulates the data used to develop or use an AI or machine learning (ML) model. This manipulation can compromise model security, performance, or ethical behavior, leading to harmful outputs or impaired capabilities. Common risks include degraded model performance, biased or toxic content, and exploitation of downstream systems.

Label Flipping

Correct labels in the training data are swapped with incorrect ones

Data Injection

Malicious data points are added to the training set

Backdoor Attacks

The model is trained to behave normally except when a specific trigger is present

Clean-Label Attacks

Modify the training data without changing the labels

AI-Related Supply Chain Security Journey……

1

Typically, when we think of supply chain, we think of building a laptop and all the parts going into it.

2

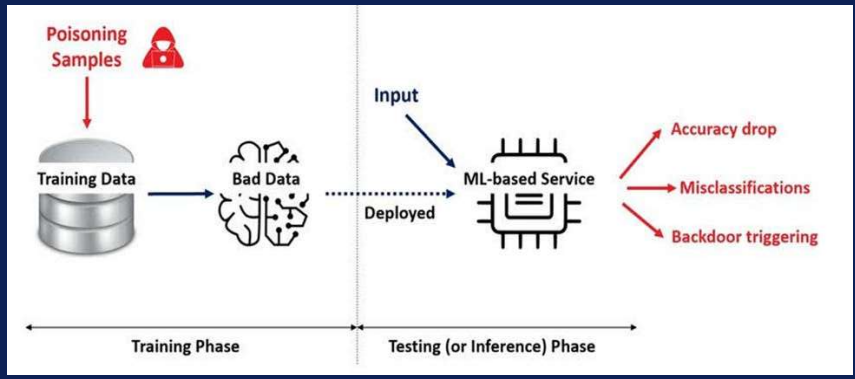
In the idea of AI-related supply chain this is a focus on using third-party datasets, pre-trained models, and plugins which of course like any supply chain can add weaknesses.

3

AI Supply Chain Attacks occur when an attacker modifies or replaces a machine learning library or model that is used by a system

4

An AI-related supply chain attack can also include the data associated with the machine learning models



Ex. - Data Poisoning Attack in AI-Related Supply Chain Journey

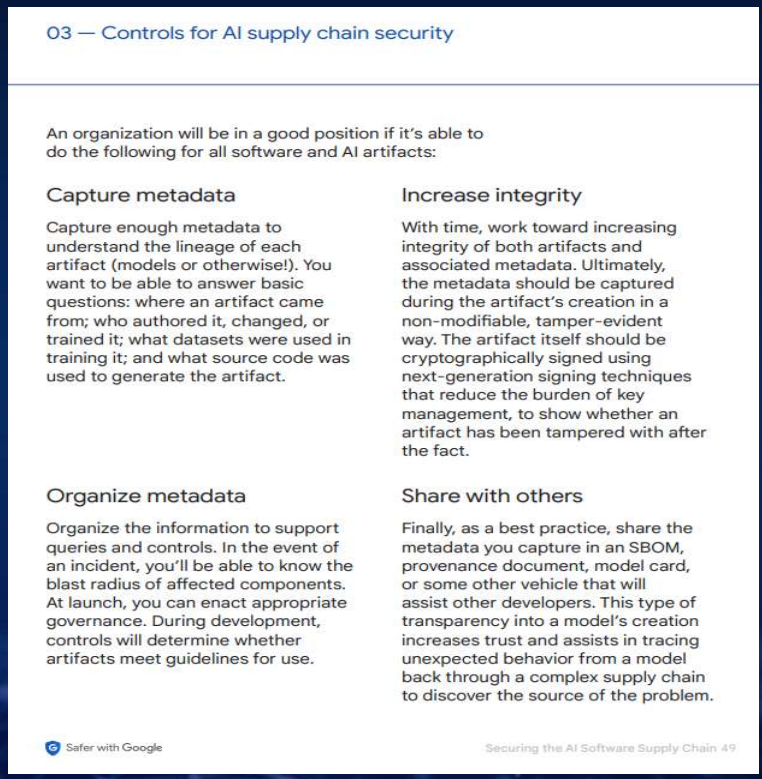
DELL Technologies

Now What?

A vulnerability is reported…

- With software vulnerabilities, we have been building maturity leveraging SBOMs, inventories and code scanning, so that if we need to work with an engineering team to create a patch, we can.
- Now, we have new “AI components” that we will need to add to our inventories: data set, models, model cards, wrappers, rag/vectordb, agents, knowledge graphs. How do those appear in inventories, Bill of Materials (BoMs) and attestations?
- Do our current tools and processes extend to AI components, so that we have the information and teams in place when we need to respond to an AI-related vulnerability?

Industry Reference Information



03 — Controls for AI supply chain security

An organization will be in a good position if it’s able to do the following for all software and AI artifacts:

Capture metadata

Capture enough metadata to understand the lineage of each artifact (models or otherwise!). You want to be able to answer basic questions: where an artifact came from; who authored it, changed, or trained it; what datasets were used in training it; and what source code was used to generate the artifact.

Increase integrity

With time, work toward increasing integrity of both artifacts and associated metadata. Ultimately, the metadata should be captured during the artifact’s creation in a non-modifiable, tamper-evident way. The artifact itself should be cryptographically signed using next-generation signing techniques that reduce the burden of key management, to show whether an artifact has been tampered with after the fact.

Organize metadata

Organize the information to support queries and controls. In the event of an incident, you’ll be able to know the blast radius of affected components. At launch, you can enact appropriate governance. During development, controls will determine whether artifacts meet guidelines for use.

Share with others

Finally, as a best practice, share the metadata you capture in an SBOM, provenance document, model card, or some other vehicle that will assist other developers. This type of transparency into a model’s creation increases trust and assists in tracing unexpected behavior from a model back through a complex supply chain to discover the source of the problem.

Safer with Google

Securing the AI Software Supply Chain 49

Google Research - Securing the AI Software Supply Chain

CRA use case

Data set used to train the model was poisoned. Lawyers advise to swap it out with another model internally and in customer products.

In EU CRA, must report actively exploited vulnerability, and provide patches for all products sold in EU for 5 years.

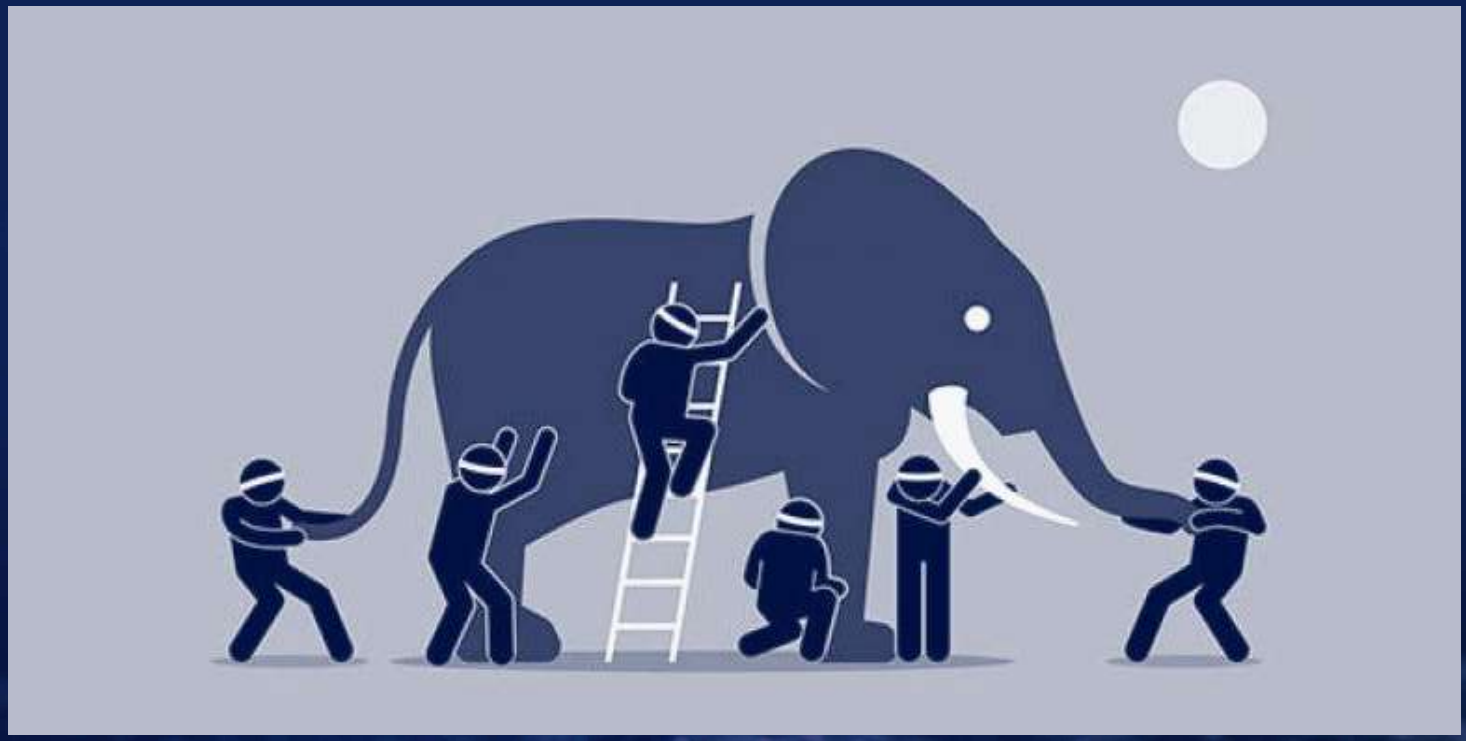
Do you have the information and processes you need to do that? For example….

- What data was produced with that model? Is it at risk?
- Where is that model used internally and in customer products?
- Do you have the secureXOps to partner between security and engineering to quickly do the work?

The Blind Men and the Elephant

The Blind Men and the Elephant is a parable from India that has been adapted and published in various stories for adults and children. It is about a group of blind men who attempt to learn what an elephant is, each touching a different part, and disagreeing on their findings. *Their collective wisdom leads to the truth.*

Collective Wisdom Leads to Truth

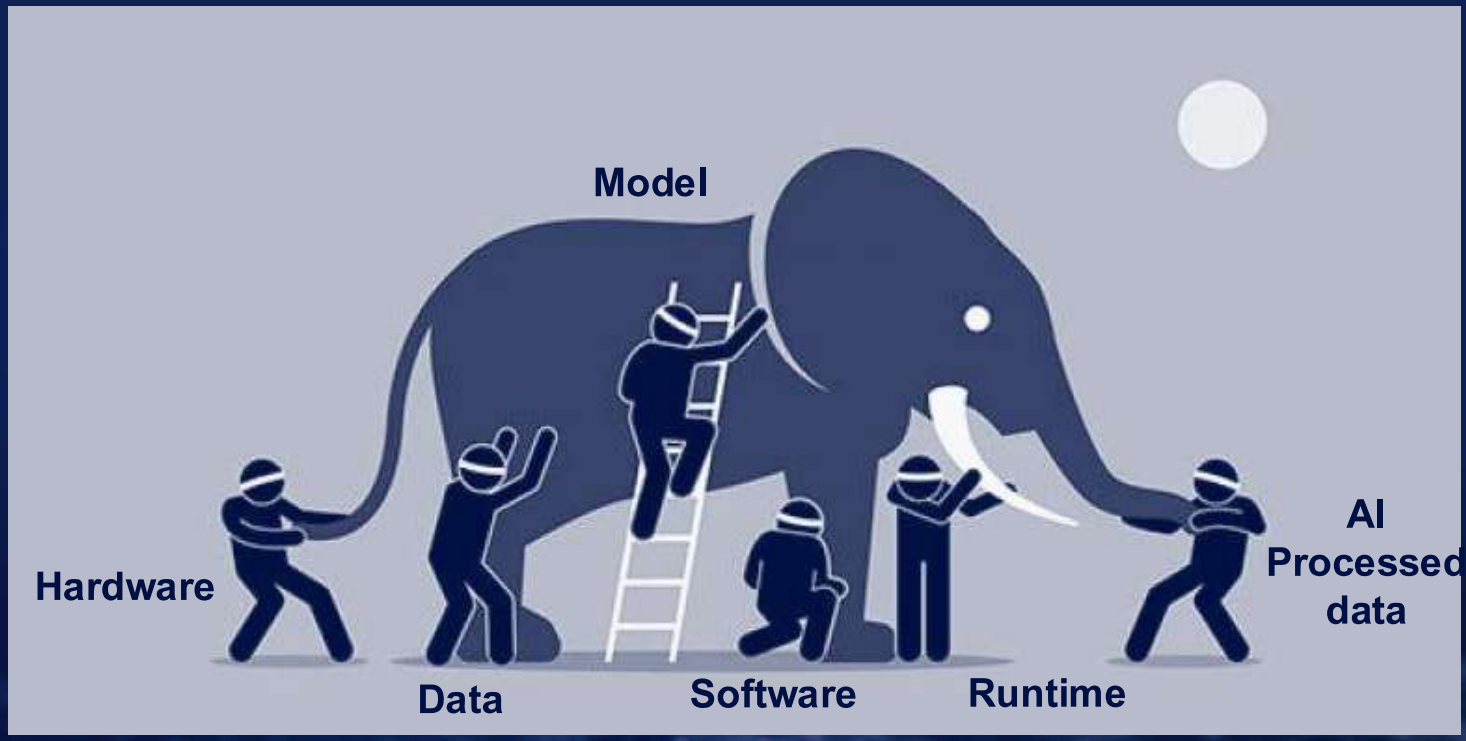


Story Source: https://americanliterature.com/author/james-baldwin/short-story/the-blind-men-and-the-elephant

Image Source: https://www.batimes.com/articles/blind-men-and-the-elephant/

The AI supply chain perspectives elephant

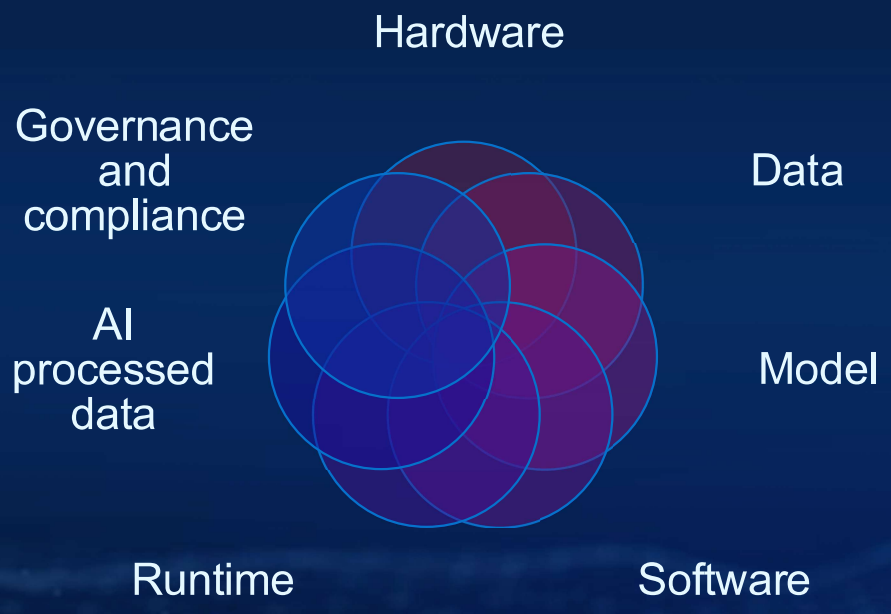
- The AI supply chain perspectives are fractured
- Each perspective is motivated by the body part of the elephant being felt.



Disclosure of governance/compliance attestations

Things are converging…..

Areas of Convergence



MITRE - Attestation: Evidence and Trust

Attestation is the activity of making a claim about properties of a target by supplying evidence to an appraiser.

1. Extend software and hardware attestation also to data sets, models, model cards, wrappers, rag/vectordb, agents infrastructure, knowledge graphs, etc.

   How do those components appear in SBOMs/AIBoMs, inventories, with cryptographically signed attestations?

2. How can we tie the converging links in the AI supply chain together with machine readable inventories, BOMS and attestations?

Universal Attestation Framework

- Automated
- Integrated
- Easy to adopt

Mechanism to establish integrity

- Universal framework
- Common, machine-readable format
- Cryptographic proof / digital signature
- Attestation profile that supports various requirements

Mechanism to report integrity

- Integrates converging tools and specifications
- Identify attestation gaps and close to integrate with existing tools and specifications

Mechanism to verify integrity

- Efficient method to identify integrity
- Composable (like interlocking bricks), can be introspected individually or bundled
- Used to establish due diligence for compliance

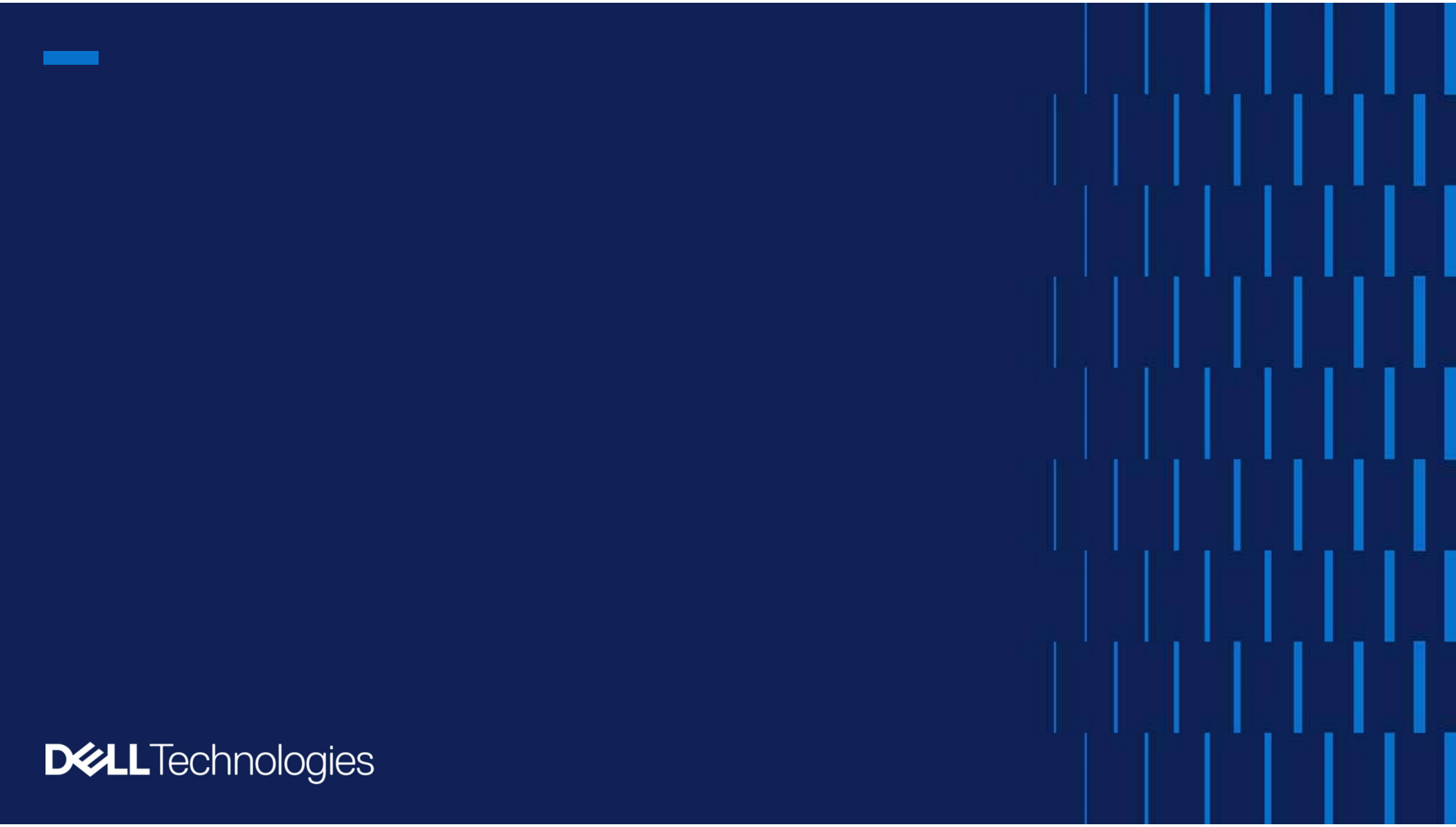
DELL Technologies

We're on an industry journey together

Link to Industry Reference Materials

Examples of converging AI secure supply chain industry initiatives

Components are securely created, the creator attests to the integrity components were securely created, attested components are placed in inventories which can be used for incident response, and we have scalable methods to report and communicate vulnerability mitigation


DELL Technologies

Challenges in detecting AI vulnerabilities compared to traditional software

ML models are shaped by data rather than a programmer, making ML vulnerabilities harder to identify	1
Attacks on ML include altering/tampering with the model or its training data, stealing the model or training data, and exploiting bugs in ML software	2
Patching ML vulnerabilities can be difficult and may only partially close security holes while incurring significant expense or performance penalties	3
ML vulnerabilities can be specific to each user, as ML systems are often tailored and adapt over time	4
Proof-of-concept exploits for ML are less practically useful but serve as warnings to improve security	5

Industry Reference Links

Back

Secure Software (OSS focus) and mechanisms to evaluate risk introduced in the AI Supply Chain

- Security Baseline extend to AI components
- Security Scorecard with Structured Results (for customizing/automating for enterprise)
- CHAOSS (project and community health for enterprise automation)
- Alpha/Omega extend to OSS AI components
- Model Openness Framework (MOF), | OpenMDW: proposed permissive license for open models based on MOF (Matt White, PyTorch foundation)

Cryptographically Signed Integrity Attestations

- Trusted Computing Group
- Remote ATtestation ProcedureS (rats) IETF
- Data provenance standards (OASIS Open and Data and Trust Alliance)
- SigStore, and Signing Model Cards with SigStore, model card metadata specification draft
- Global Policy Cyber Working Group, attestation framework/specification
- In toto, Archivista and Witness
- Confidential Computing Consortium (Verifable Compute)
- Coalition for Content Provenance and Authenticity (C2PA)

Inventory/BOMs

- Maturing SBOMs with integrity attestations
- Protobom
- AIBoMs
- GUAC
- Software Version Identification | VulnCon Open Interchange on CPE – PURL Between the Communities of Interest and the CVE and NVD Programs | Software Identification Ecosystem Option Analysis | OpenSSF Response

Incident Response

- Vulnerability Exploitability eXchange (VEX), OpenVEX and Common Security Advisory Framework (CSAF)

End to end view

- Google Secure AI Framework (SAIF); Securing the AI Software Supply Chain (Research)
- Coalition for Secure AI (CoSAI)
- Machine Learning Model Lifecycle (Intel)