Safer with Google

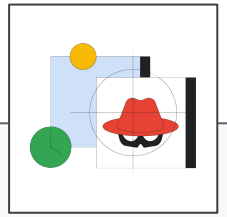
Secure AI Development @ Google

January 17, 2024

Google's Secure AI Framework – Leadership in AI development

AI is advancing rapidly, and it's important that **effective risk management strategies** evolve along with it.

Expand strong security foundations to the AI ecosystem



Extend detection and response to bring AI into an organization's threat universe

Automate defenses to keep pace with existing and new threats

Harmonize platform level controls to ensure consistent security across the organization

Adapt controls to adjust mitigations and create faster feedback loops for AI deployment

Contextualize AI system risks in surrounding business processes

Goals for AI Supply Chain Security

"Provide **open standards** and **robust solutions** for **transparency**, **trust** and **control** of the AI Supply Chain"

AI vs Traditional Software Supply Chain

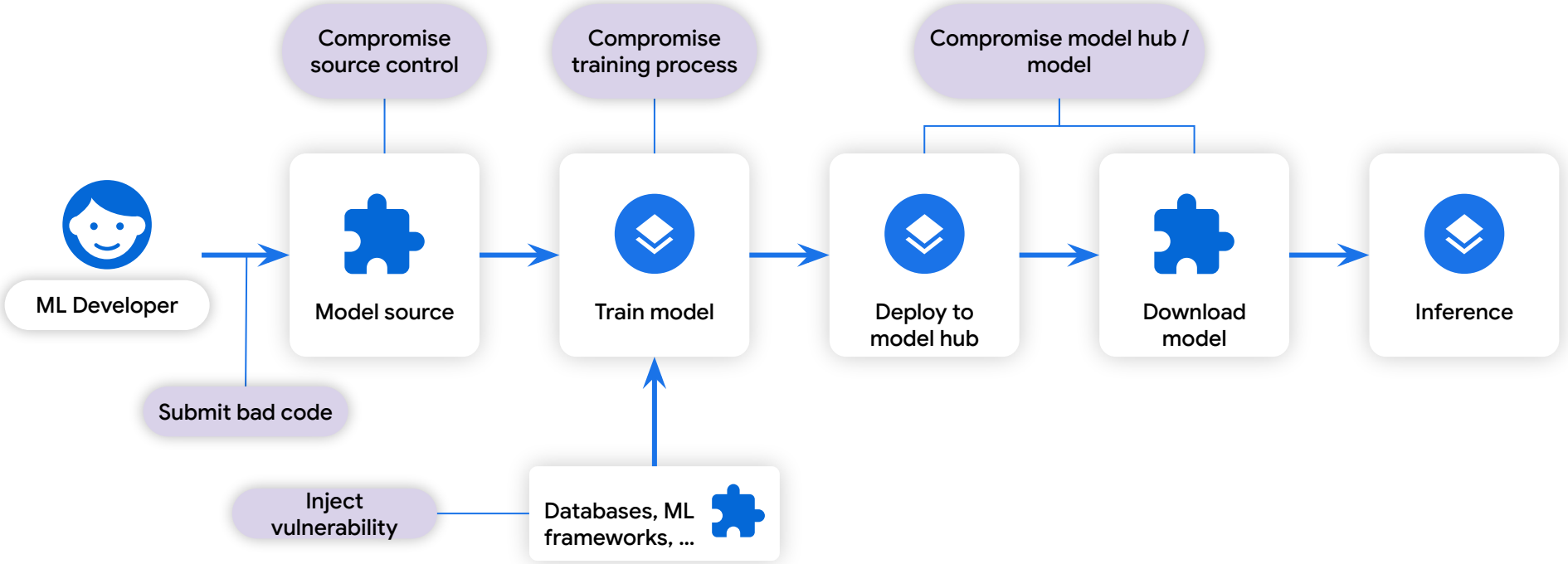
Similarities

- Similar SDLC, with different terminologies (e.g. Build vs Training, Artifact vs Model)
- Model tampering has same security severity (RCE/leaks)
- **No need to reinvent the wheel for security**

Differences

- Data, Code, Configs and Models are intertwined
- Provenance/trusted lineage on massive datasets
- **Rapid, iterative development processes**

AI Supply Chain Threats



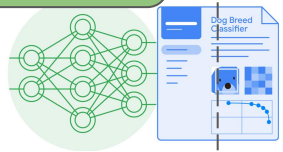
AI Supply Chain Security Problems

- Did the ML model creator use **safe development practices**?
- For open source models, what was the **training code** used?
- What **datasets** went into training that model?
- Who **published** the model? Are they **trustworthy**?
- Could the model have been **replaced** by a tampered version **following publication or during training time**?

AI Supply Chain Security: Current Solution

- Did the ML model creator use **safe development practices**?
- For open source models, what was the **training code** used?
- What **datasets** went into training that model?

Solution



- Who **published** the model? Are they **trustworthy**?
- Could the model have been **replaced** by a tampered version **following publication or during training time**?

SLSA sigstore

https://security.googleblog.com/2023/10/increasing-transparency-in-ai-security.html

More than just Supply Chain Integrity

- SSCI for ML is just one pillar of SAIF
- Model cards integrations
- Engagements with OSS (OpenSSF AI Working Group, LF AI & Data)