ML Model Signing: Cryptographically Paving the Way to Provenance in Machine Learning Models

Mihai Maruseac
GOSST (Google OSS Security Team)

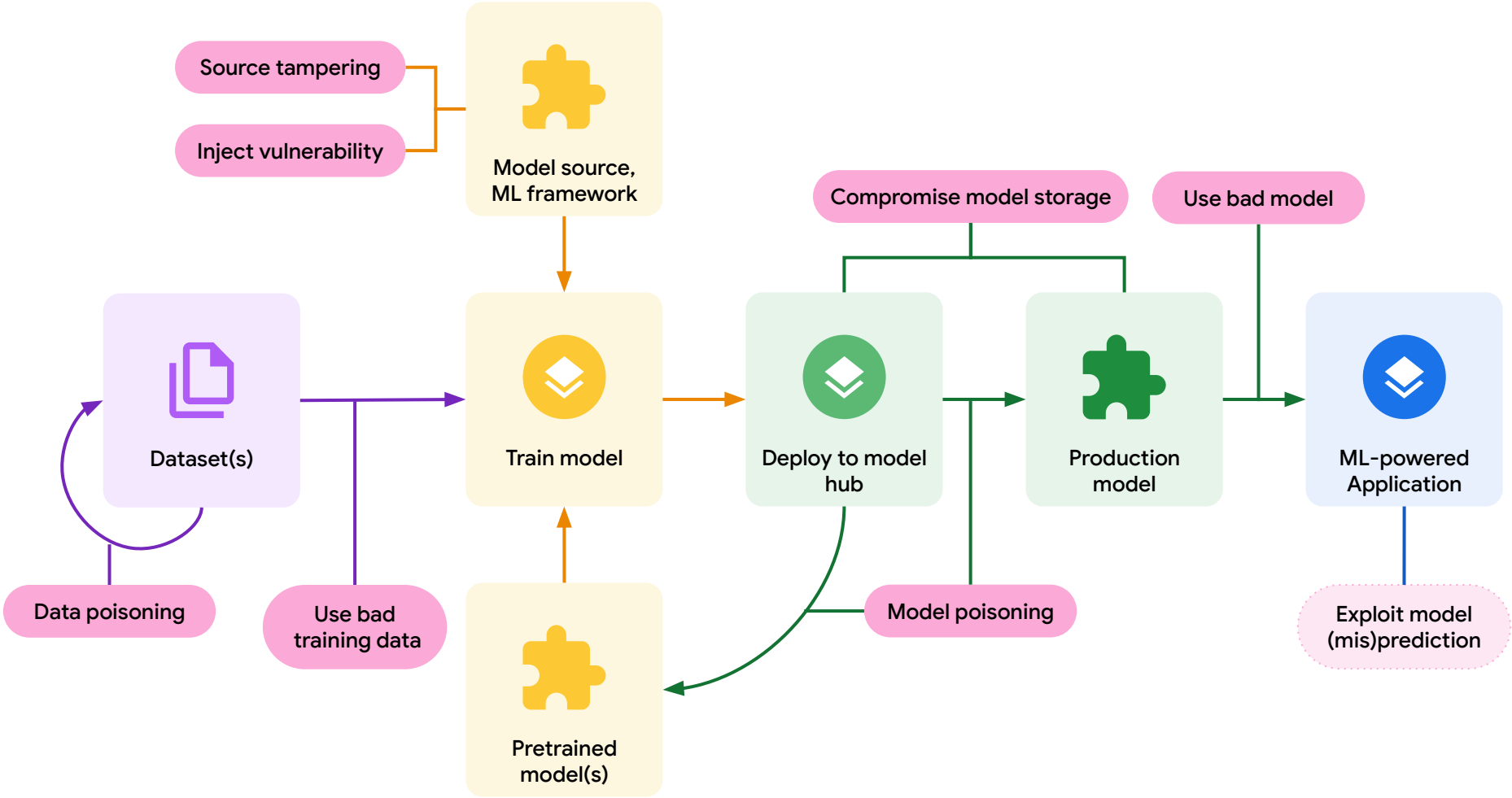
Model hub

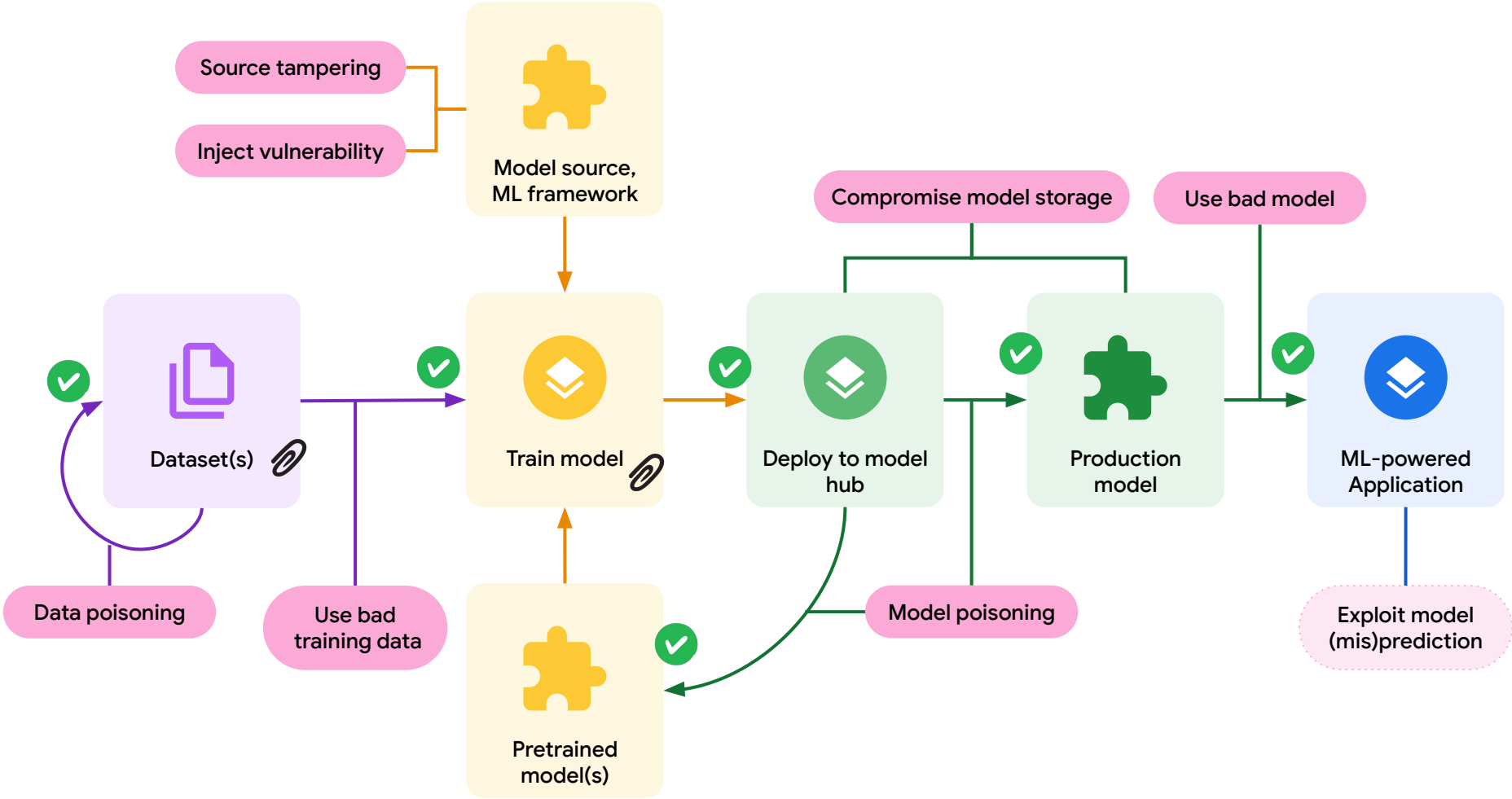
Nice Model

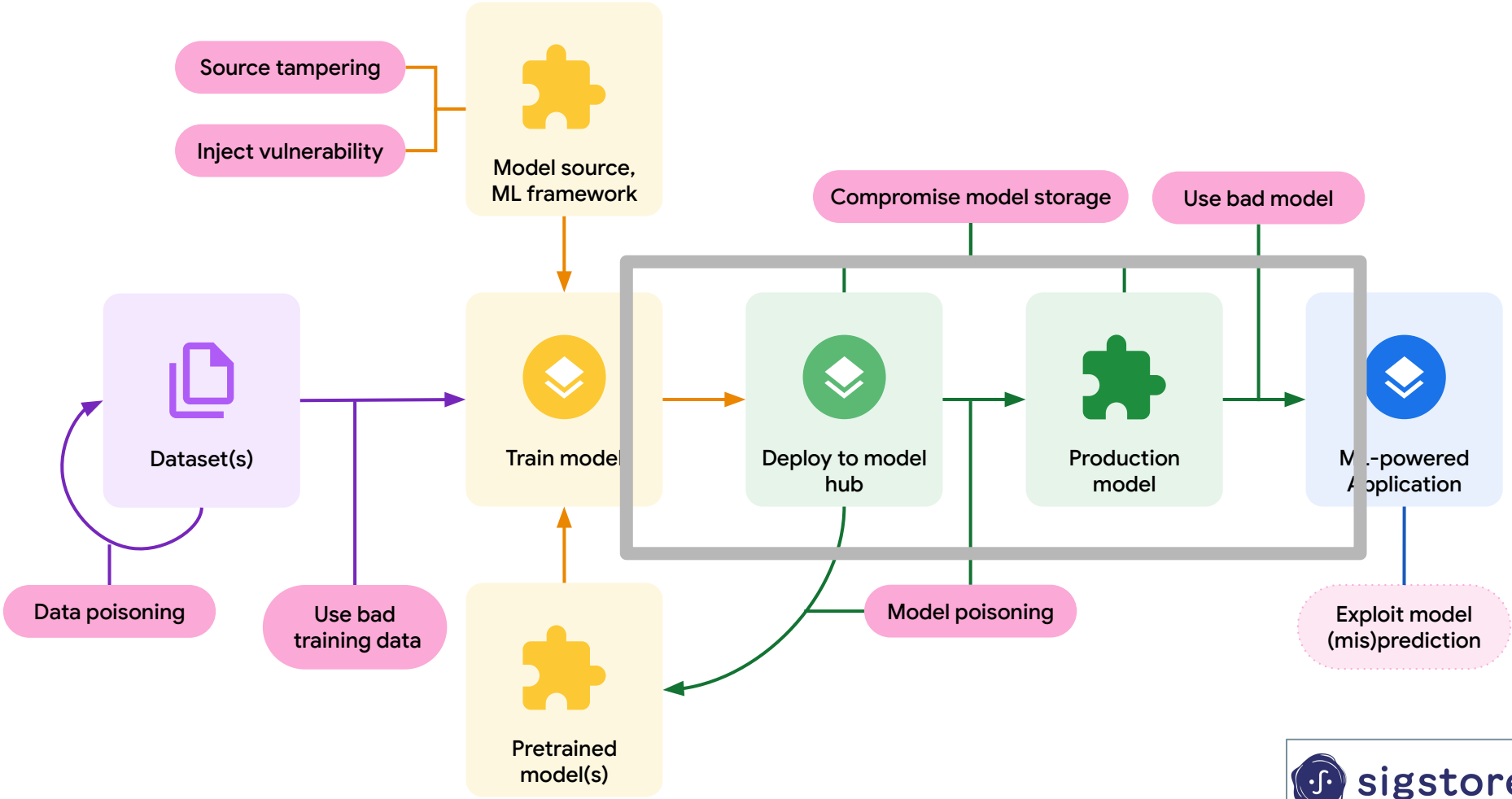
Model card
Malware detection model

GPT-4 friend

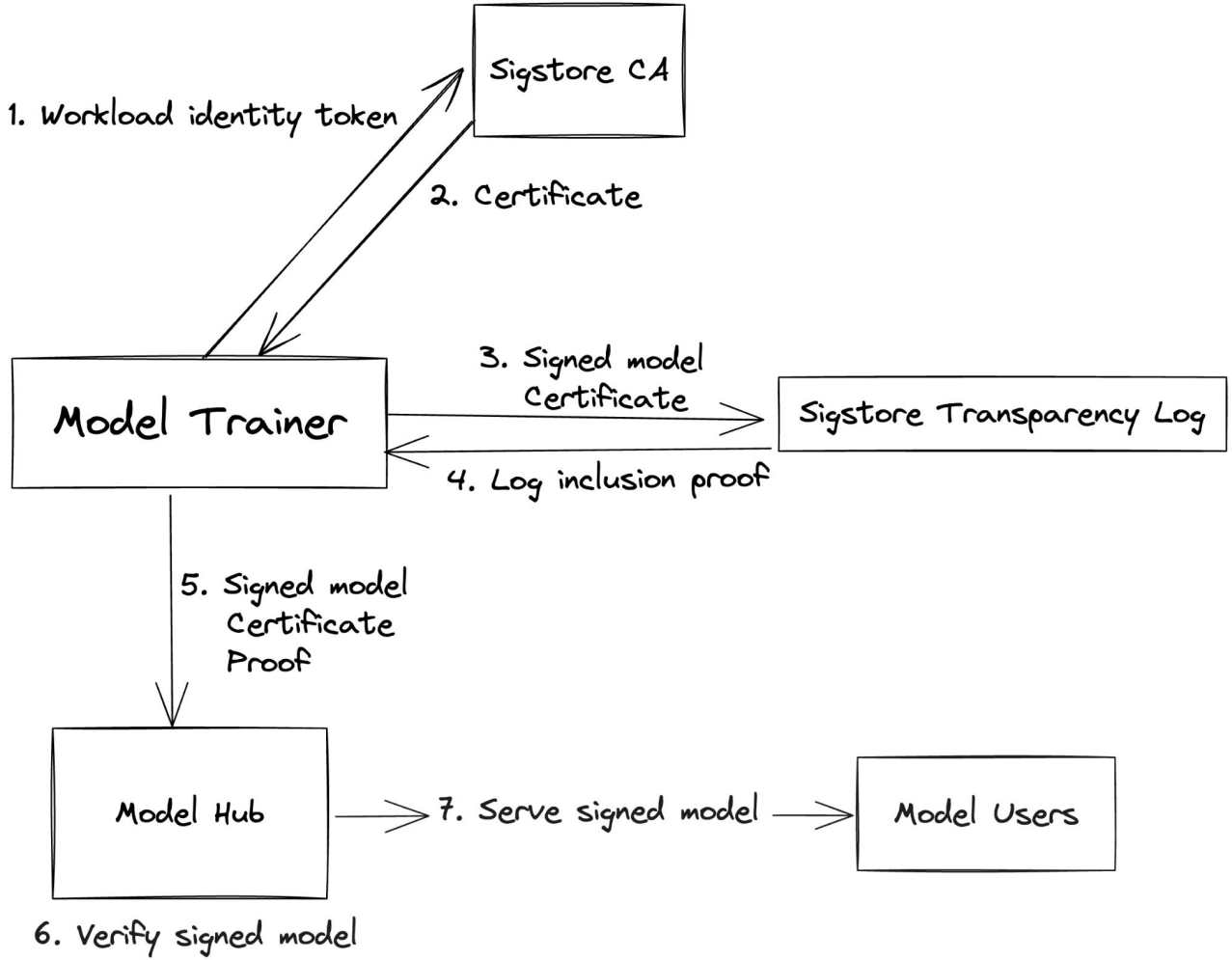
Model card
Friendly chatbot

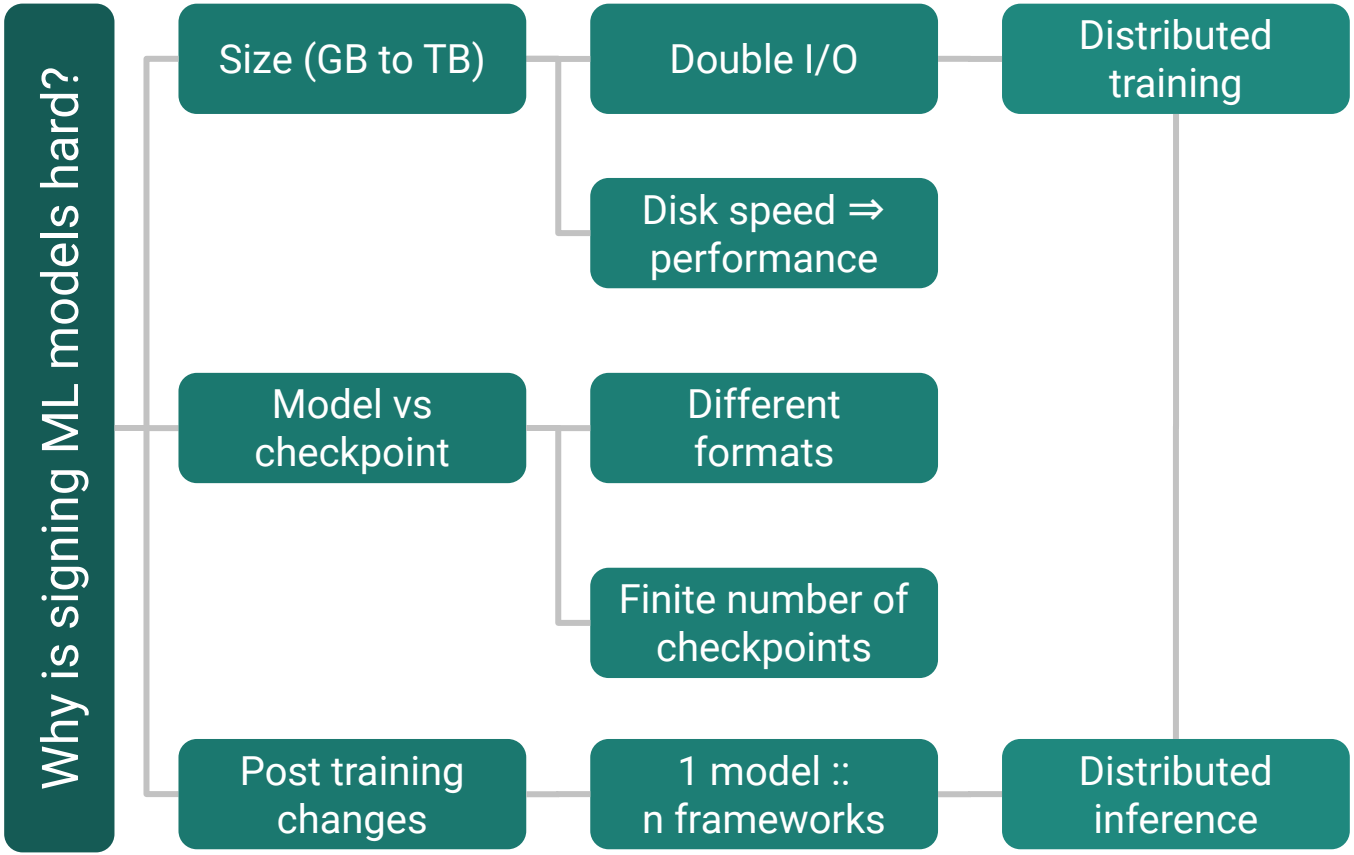

Source tampering
Inject vulnerability
Model source, ML framework
Compromise model storage
Use bad model
Dataset(s)
Train model
Deploy to model hub
Production model
ML-powered Application
Data poisoning
Use bad training data
Model poisoning
Exploit model (mis)prediction
Pretrained model(s)

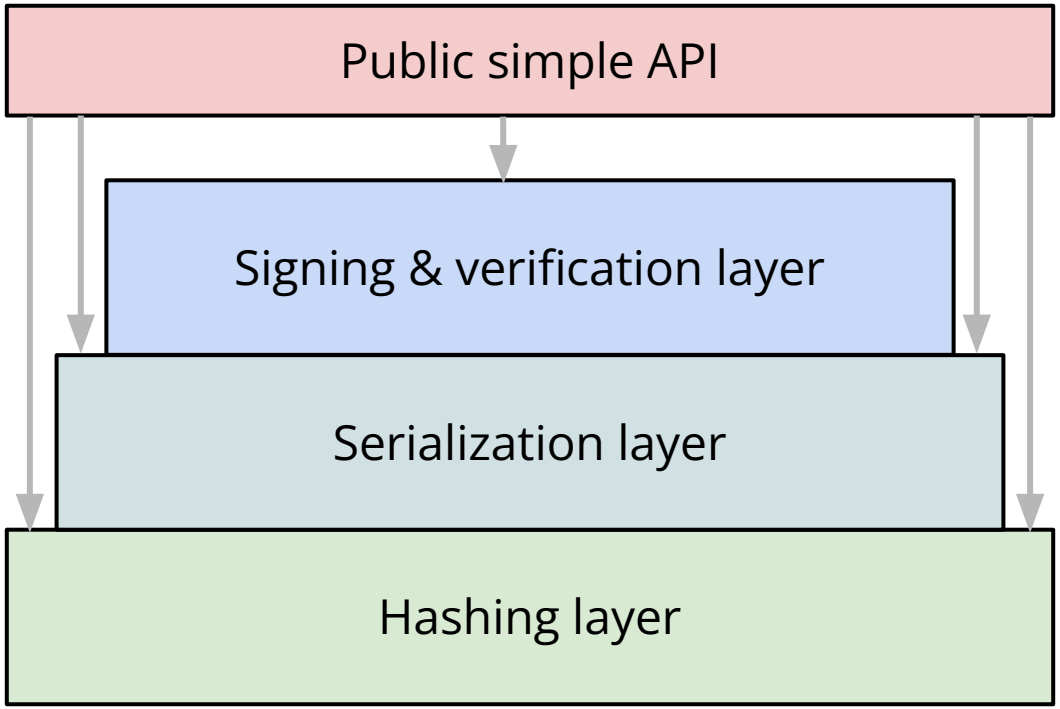

Source tampering
Inject vulnerability
Model source, ML framework
Compromise model storage
Use bad model
Dataset(s)
Train model
Deploy to model hub
Production model
ML-powered Application
Data poisoning
Use bad training data
Pretrained model(s)
Model poisoning
Exploit model (mis)prediction


Source tampering
Inject vulnerability
Model source, ML framework
Compromise model storage
Use bad model
Dataset(s)
Train model
Deploy to model hub
Production model
ML-powered Application
Data poisoning
Use bad training data
Model poisoning
Exploit model (mis)prediction
Pretrained model(s)

sigstore


Sigstore CA
1. Workload identity token
2. Certificate
Model Trainer
3. Signed model
Certificate
Sigstore Transparency Log
4. Log inclusion proof
5. Signed model
Certificate
Proof
Model Hub
7. Serve signed model
Model Users
6. Verify signed model


Why is signing ML models hard?
Size (GB to TB)
Double I/O
Distributed training
Disk speed ⇒ performance
Model vs checkpoint
Different formats
Finite number of checkpoints
Post training changes
1 model :: n frameworks
Distributed inference


Public simple API
Signing & verification layer
Serialization layer
Hashing layer

Demo

Streaming hashing

Performance improvements

SLSA, Metadata, C2PA

Incremental updates

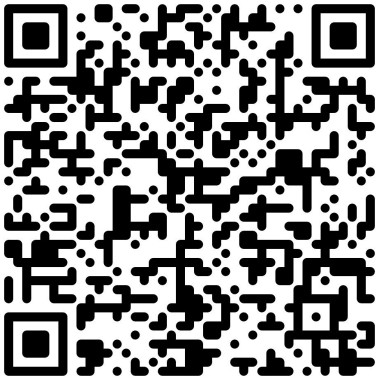
Integrations:

- **Frameworks**
- **ML hubs**

model-transparency

AI SSCI whitepaper

OpenSSF AI/ML WG



SigstoreCon