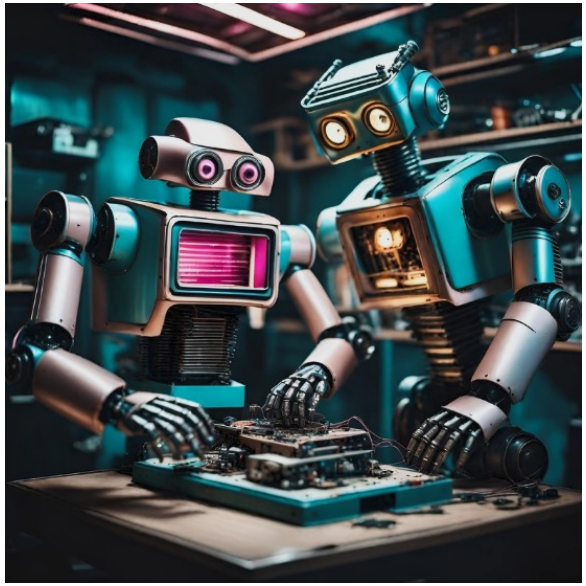
Disrupting the Model:
Abusing MLOps Platforms to Compromise ML Models and Enterprise Data Lakes

Brett Hawkins (@h4wkst3r)
Adversary Services, IBM X-Force Red

Chris Thompson (@retBandit)
Adversary Services, IBM X-Force Red



IBM

Whitepaper and Tool

Whitepaper

MLOKit

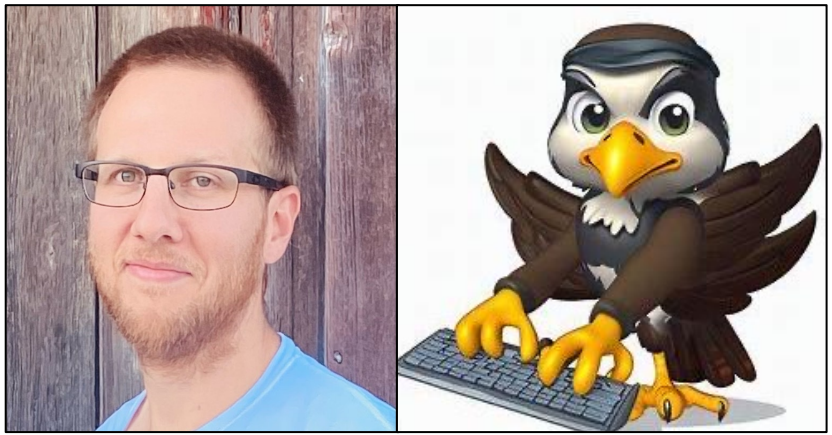
Agenda

1. Introduction
2. Background
3. MLOKit
4. Attacking MLOps Platforms
   - Azure ML, BigML, Vertex AI
     - Overview
     - Attack Scenarios
     - Demos
5. Configuration Guidance
6. Detection Guidance
7. Conclusion

Introduction

Who are we – Brett Hawkins

https://h4wkst3r.github.io



Current Role
Capability Lead, Adversary Services
IBM X-Force Red

Open-Source Tool Author
SharPersist, InvisibilityCloak, SCMKit, ADOKit, MLOKit

Conference Speaker
Black Hat (US & EU), BlueHat, DerbyCon, Wild West Hackin' Fest, BSides, Hackers Teaching Hackers

Who are we – Chris Thompson

Current Role
Global Head,
IBM X-Force Red

Conference Speaker
DEF CON, Black Hat, ToorCon, RSA, SecTor, BSides, SANS, and Wild West Hackin' Fest

About

Chris has extensive experience performing red teaming for clients in a wide variety of industries.

Chris is on the Review Board for Black Hat USA and is the co-founder of the upcoming Offensive AI con.

Research Drivers

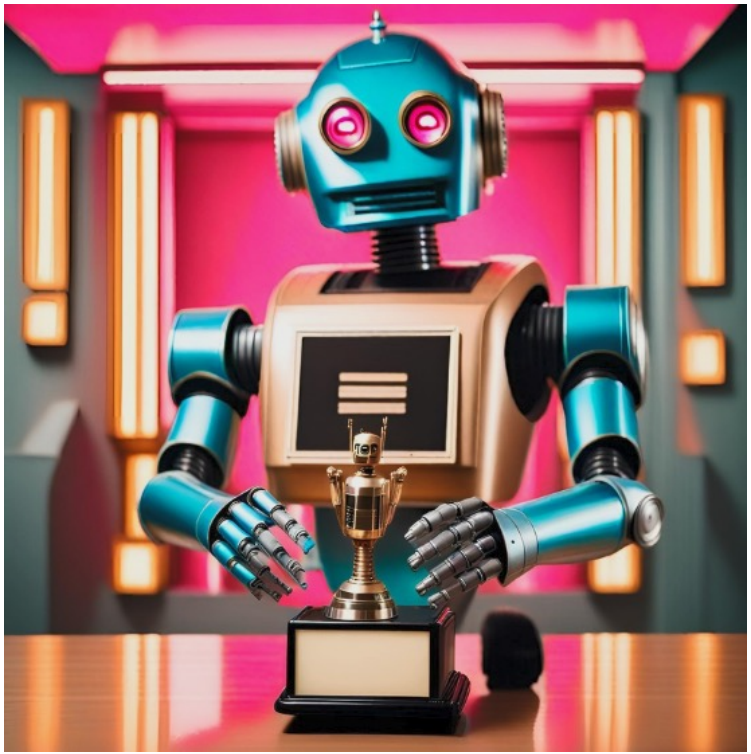
Threat actors targeting MLOps platforms

Lack of research on attacking and defending MLOps platforms

Adoption of MLOps by enterprises

Lack of tooling to simulate attacks against MLOps platforms

Research Goals

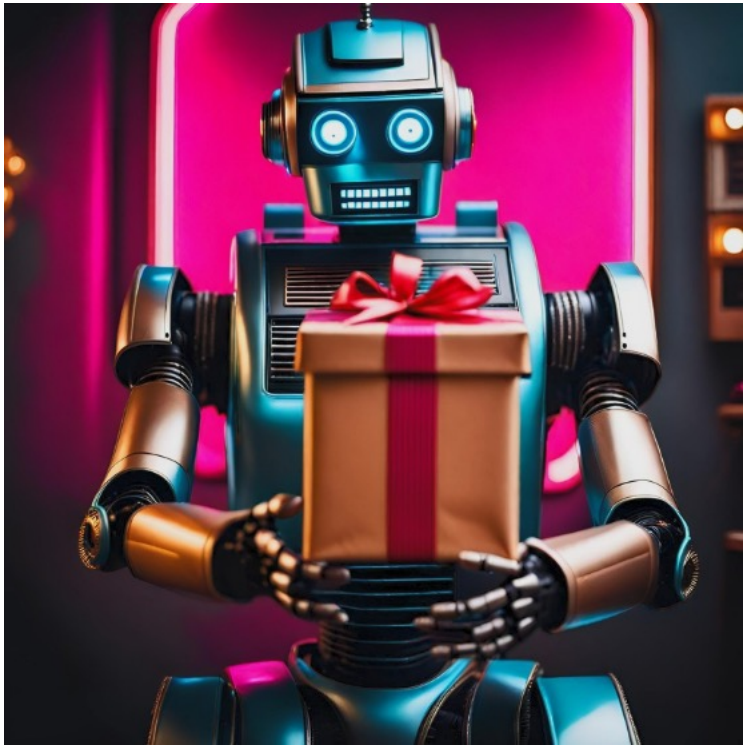


Inspire future MLOps research

Bring more attention to defending MLOps platforms

Open-source framework to simulate attacks and test detections against MLOps platforms

Attendee Takeaways



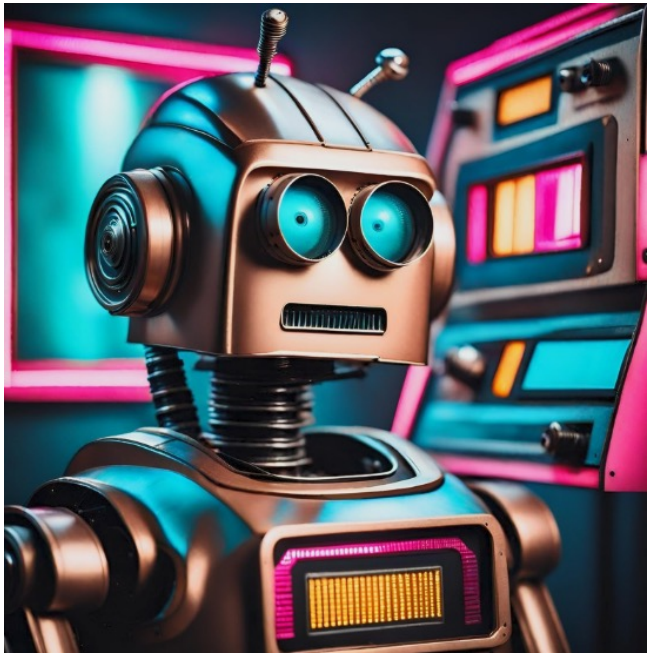
How to attack and defend MLOps platforms

Awareness of attack scenarios against MLOps platforms

How to leverage toolkit to streamline and automate attacks against MLOps

How to build detections for attacks against MLOps platforms

What is new in this research?



Practically perform ML-based attacks against popular MLOps platforms

REST API abuse of MLOps platforms

Dedicated toolkit for simulating attacks against MLOps platforms

Detection rules for ML-based attacks against MLOps platforms

Our Perspective

We are

Offensive Cybersecurity Specialists

We are not

Data Scientists

AI/ML Engineers

Cloud Engineers

Detection Engineers

DevOps Engineers

Software Engineers

Prior Work

Links to prior work are provided in whitepaper

Adrian Wood & Mary Walker – **Black Hat Asia 2024**

Confused Learning: Supply Chain Attacks through Machine Learning Models

Nitesh Surana – **Black Hat USA 2023**

Uncovering Azure's Silent Threats: A Journey into Cloud Vulnerabilities

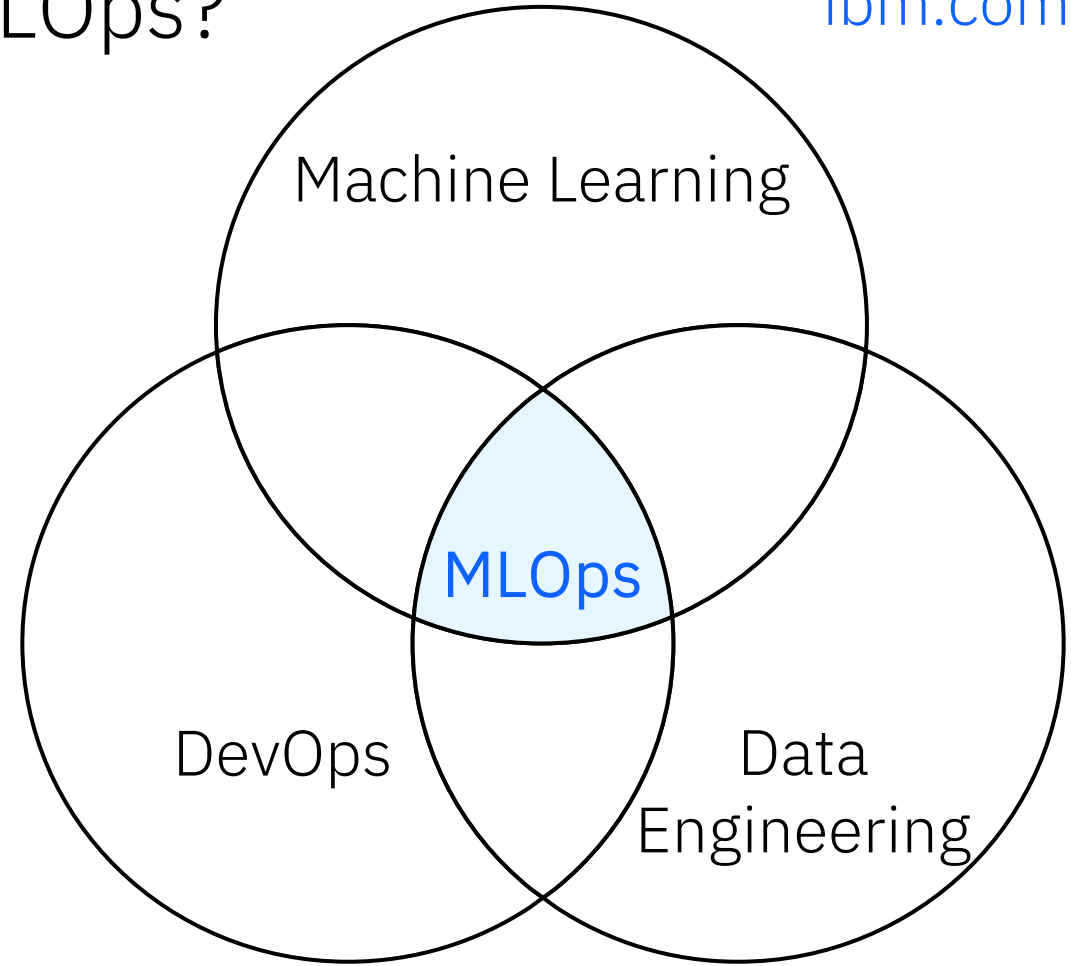
Florian Tramer, Fan Zhang, Ari Juels, Michael K. Reiter, Thomas Ristenpart – **Academic Whitepaper**

Stealing Machine Learning Models via Prediction

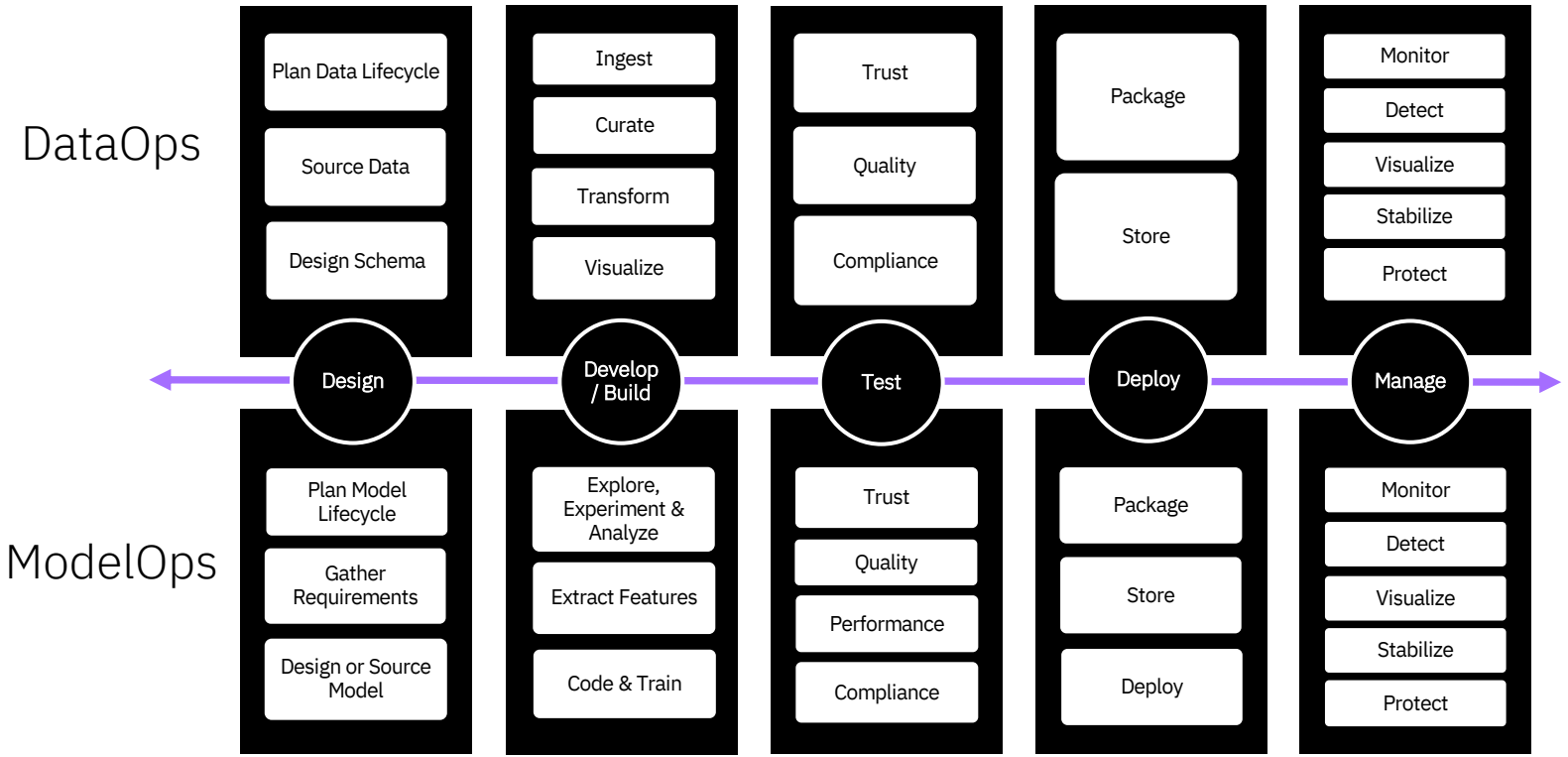
Background

What is MLOps?

ibm.com/topics/mlops



MLOps Lifecycle



Popular Open-Source MLOps Platforms

- Sematic
- Metaflow
- Kubeflow
- MLFlow
- Flyte
- Seldon Core
- MLReef
- TensorFlow

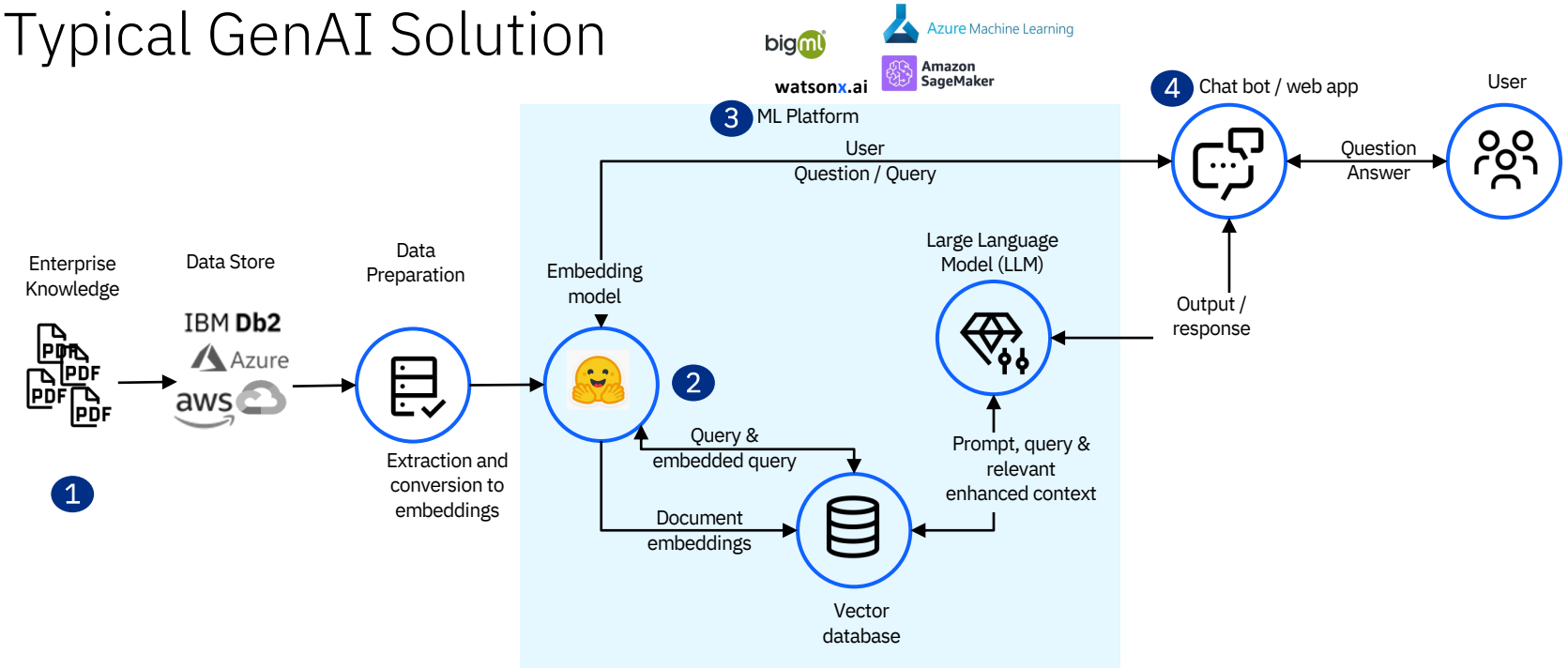
Popular Commercial MLOps Platforms

- Amazon SageMaker
- Azure ML
- BigML
- Google Cloud Vertex AI

Research Focus

- Qwak
- Domino Enterprise MLOps Platform
- Databricks
- DataRobot
- TrueFoundry
- Weights & Biases (W&B)
- Palantir AIP

Typical GenAI Solution



1 ML Pipeline

Testing of within the MLSecOps pipeline and production environments for attack paths from an adversary's perspective. Asses the potential impact of backdoored or compromised model or AI application environment and validate detections for attacks against datasets.

2 Models

Perform testing of GenAI applications, integrations, and API endpoints for security issues before production.

3 Platforms

Focused security testing on SaaS and PaaS platforms leveraged by GenAI applications to insecure security configurations and integrations with AI platforms such as Amazon SageMaker, Azure ML, BigML, Watsonx.ai, etc.

4 GenAI Apps

Perform testing of GenAI applications, integrations, and API endpoints for security issues.

Threat Actor Motivation

Theft of models and weights, backdooring models for initial access or persistence, expanding access via lateral movement and privsec, sensitive data theft or deploying ransomware, model modification/poisoning for misclassification, degradation, fraud or ml-based detection evasion.

Example Goals

Top Tier

Nation State

Top State-Sponsored Hacking Groups

- Poison trained models or model training data related to image recognition, predictive AI, early warning systems
- Modify models used for offensive tasking and targeting
- Extract sensitive data sets used to train models and their weights

Advanced

State-Sponsored Hacking Groups

Well-Funded Ransomware Groups

- Compromise AI-based fraud analytic platforms to allow restricted actions, modify fraud algos, or ignore alert thresholds
- Steal/extract capital market trading algo and models
- Gain code execution in public AI platforms
- AI supply chain compromise (targeting high-value victims with backdoored HF models, python libraries, etc.), novel extensions
- Target MLSecOps for expanding and elevating corporate access

Targeted

Commodity Ransomware Groups

Initial Access Brokers

- Focused on GenAI app and platform security issues to gain model code execution to compromise production or training environments
- Seed widespread malicious pytorch models on HF
- Compromise connected enterprise data lakes with financial and healthcare datasets

Opportunistic

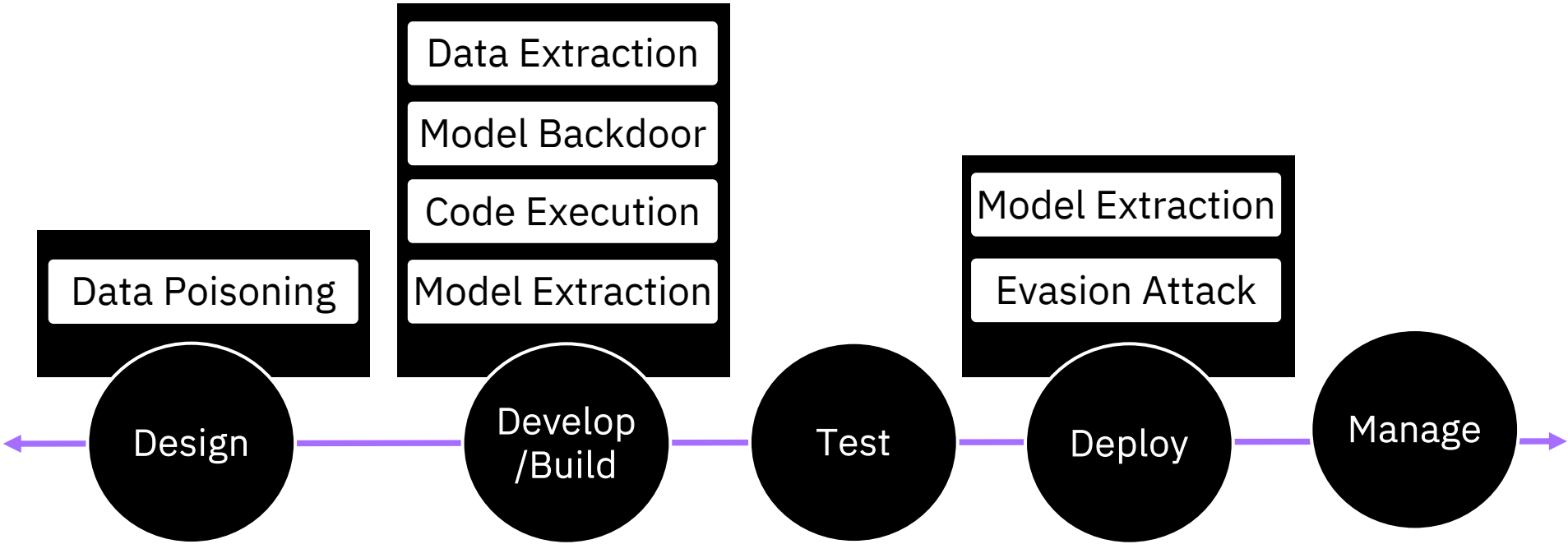
Hacktivists

Solo Hackers

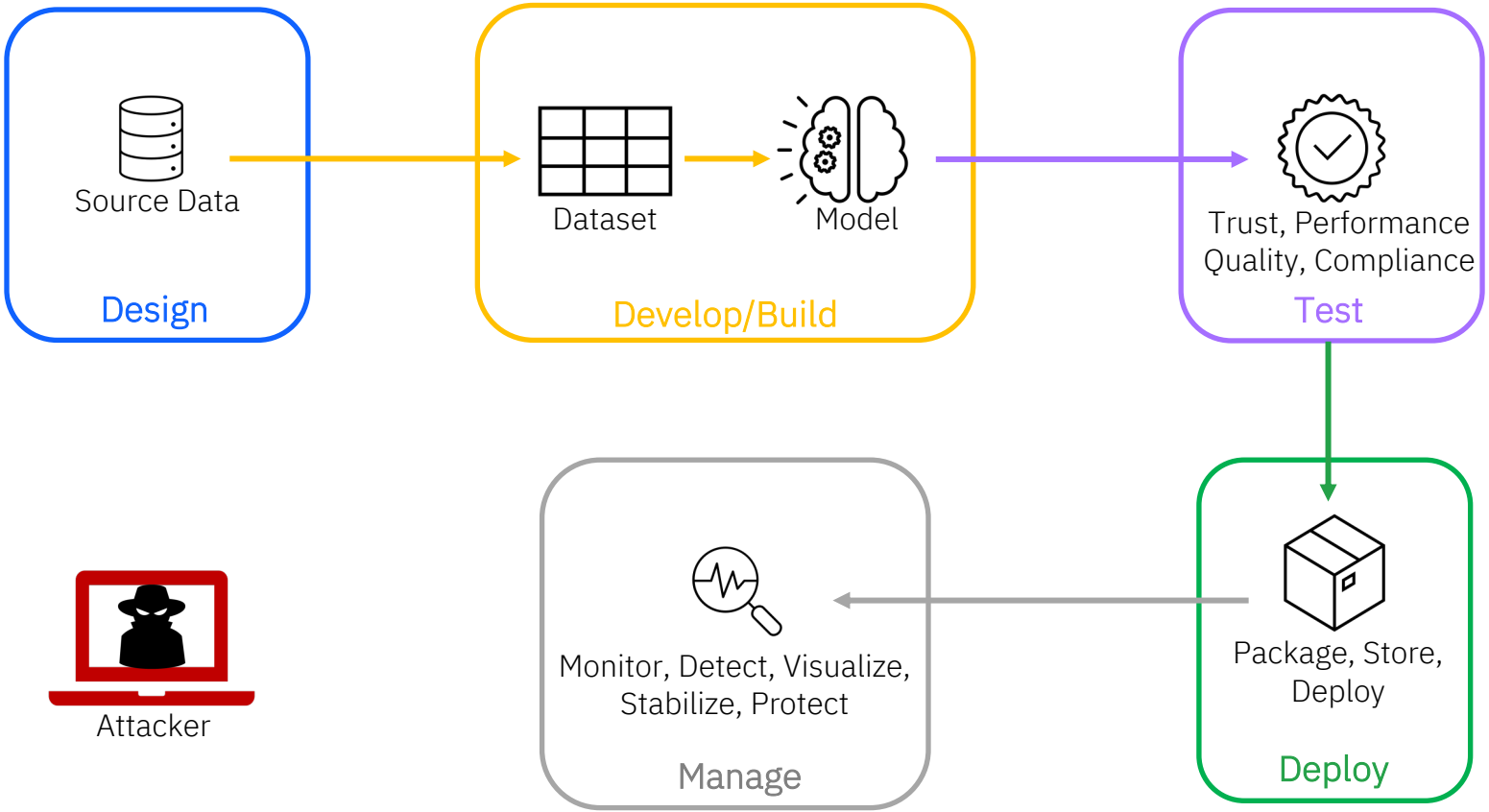
Insider Threats

- Focused on published security attacks for code execution or service disruption.
- Focused on fraudulent use of AI such as manipulating sales bots, safety related content abuse
- Perform resource exhaustion/DOS

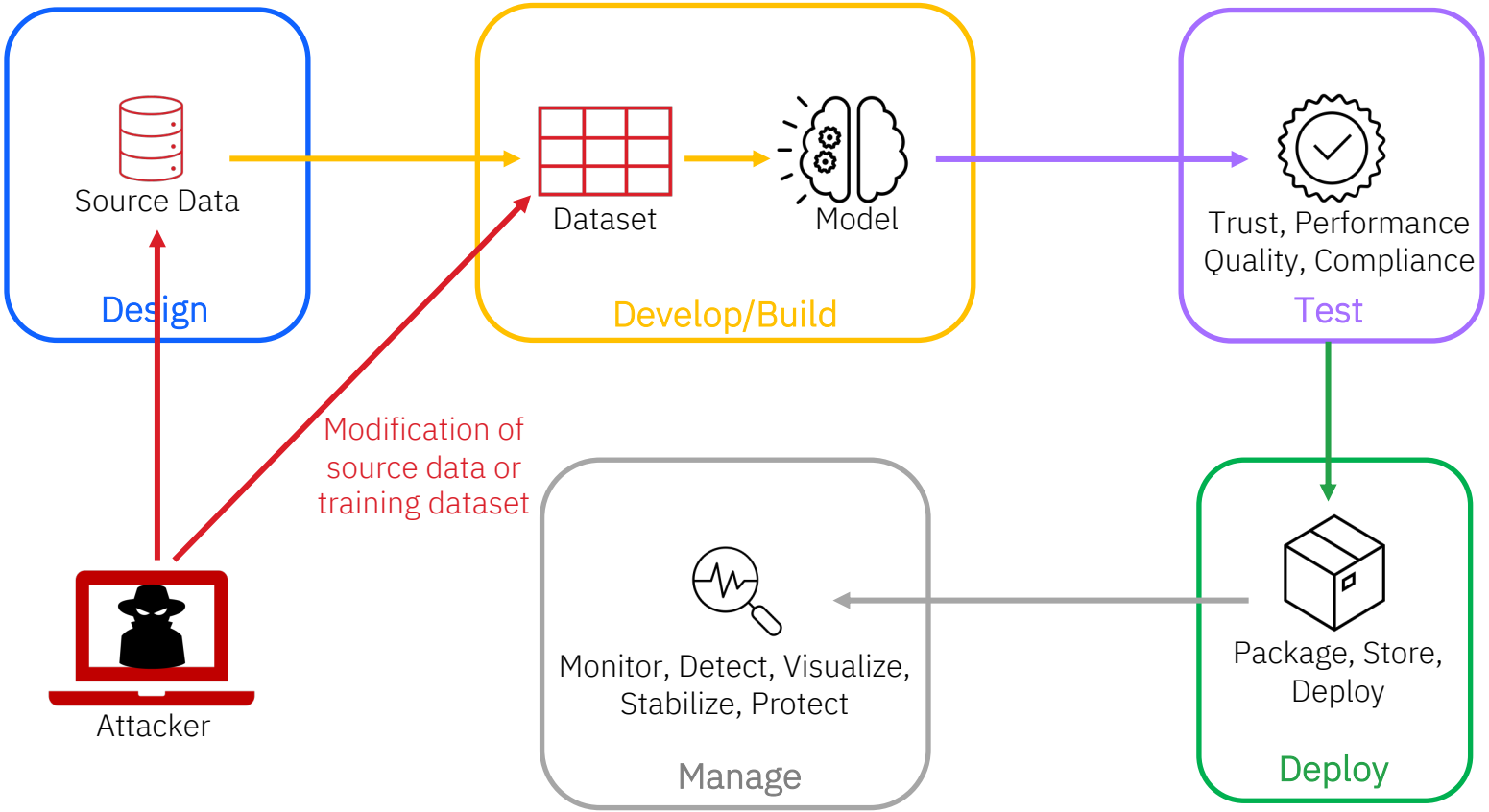
Attack Scenarios against MLOps Lifecycle



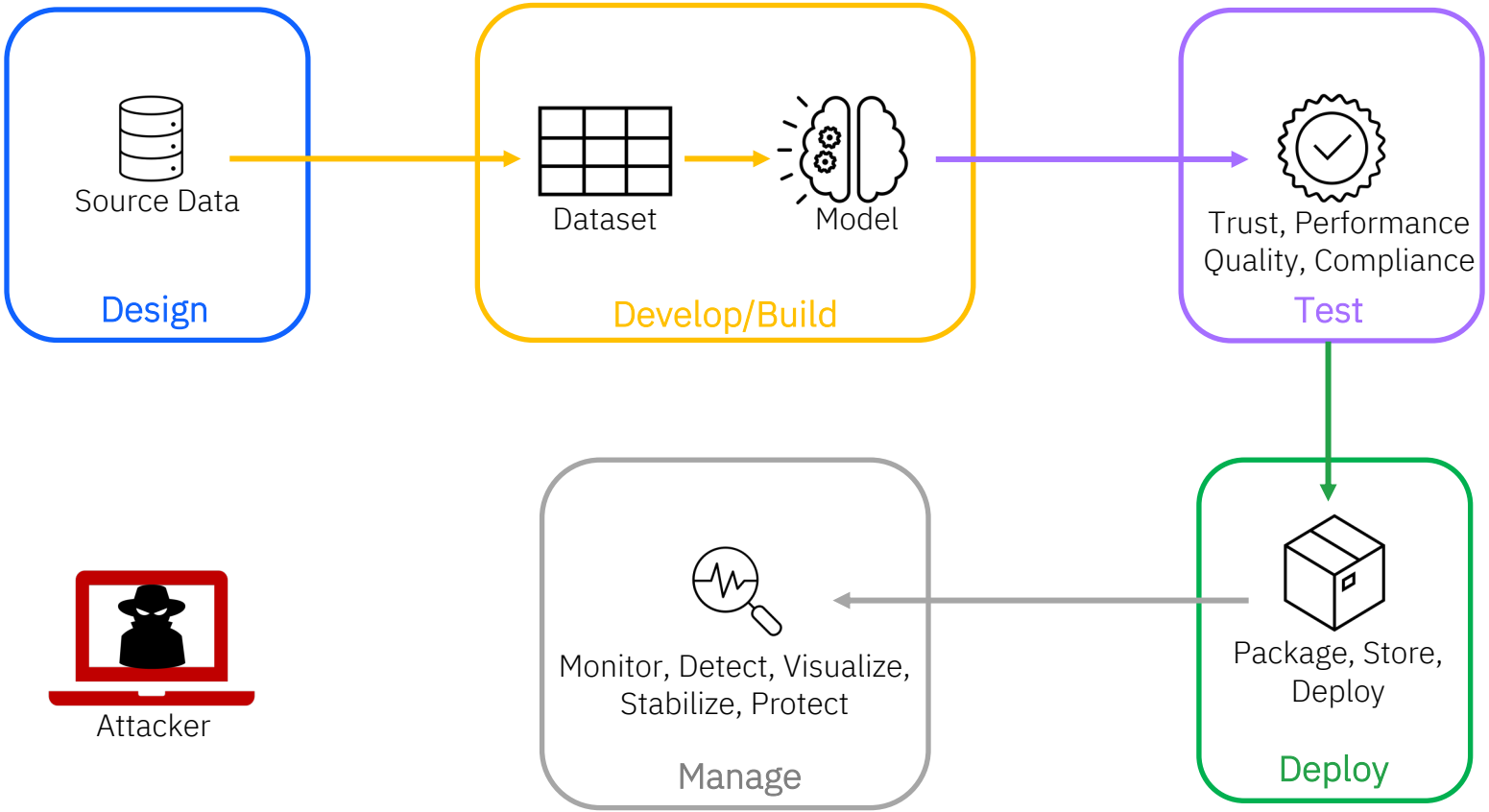
Data Poisoning



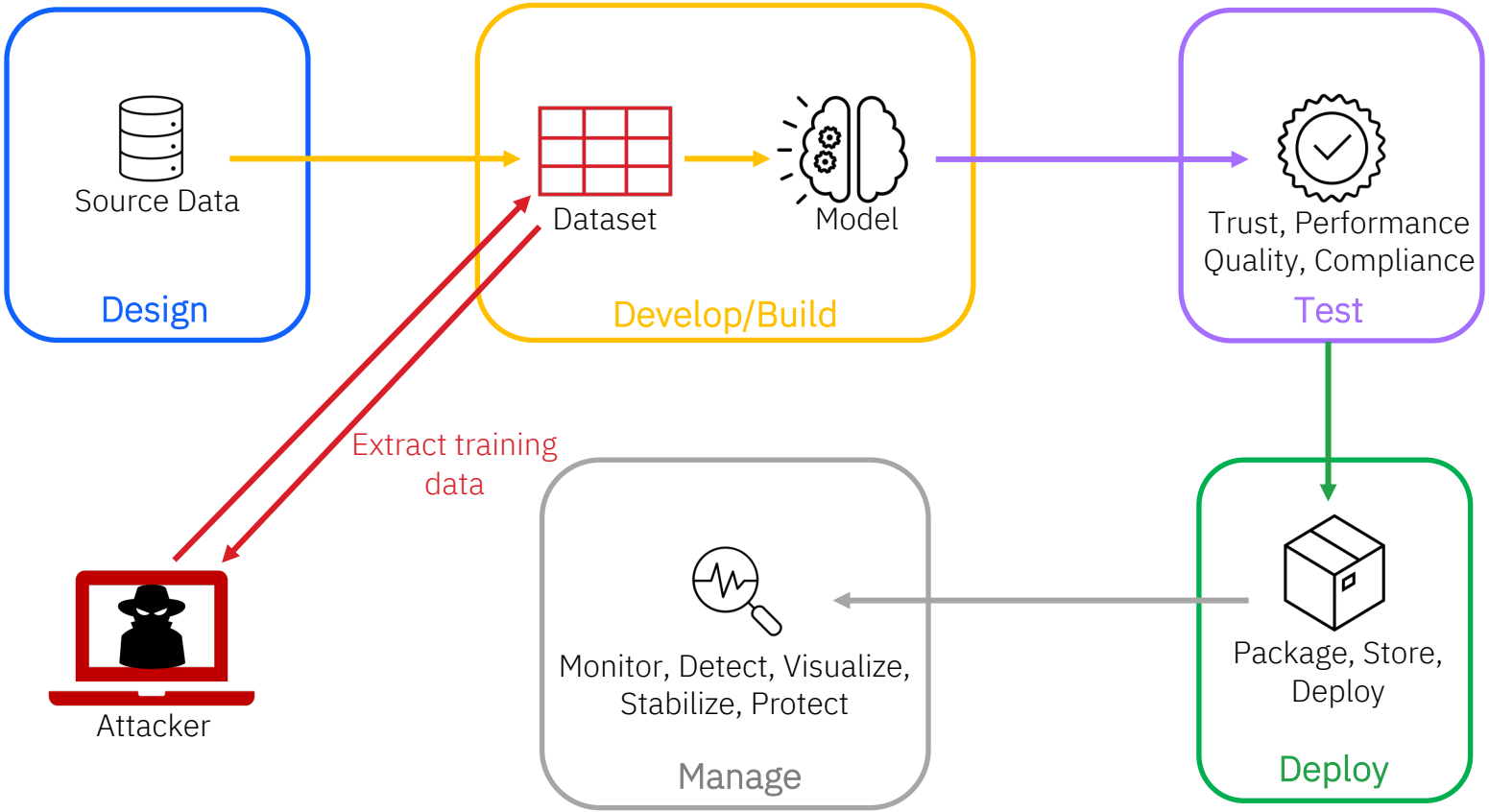
Data Poisoning



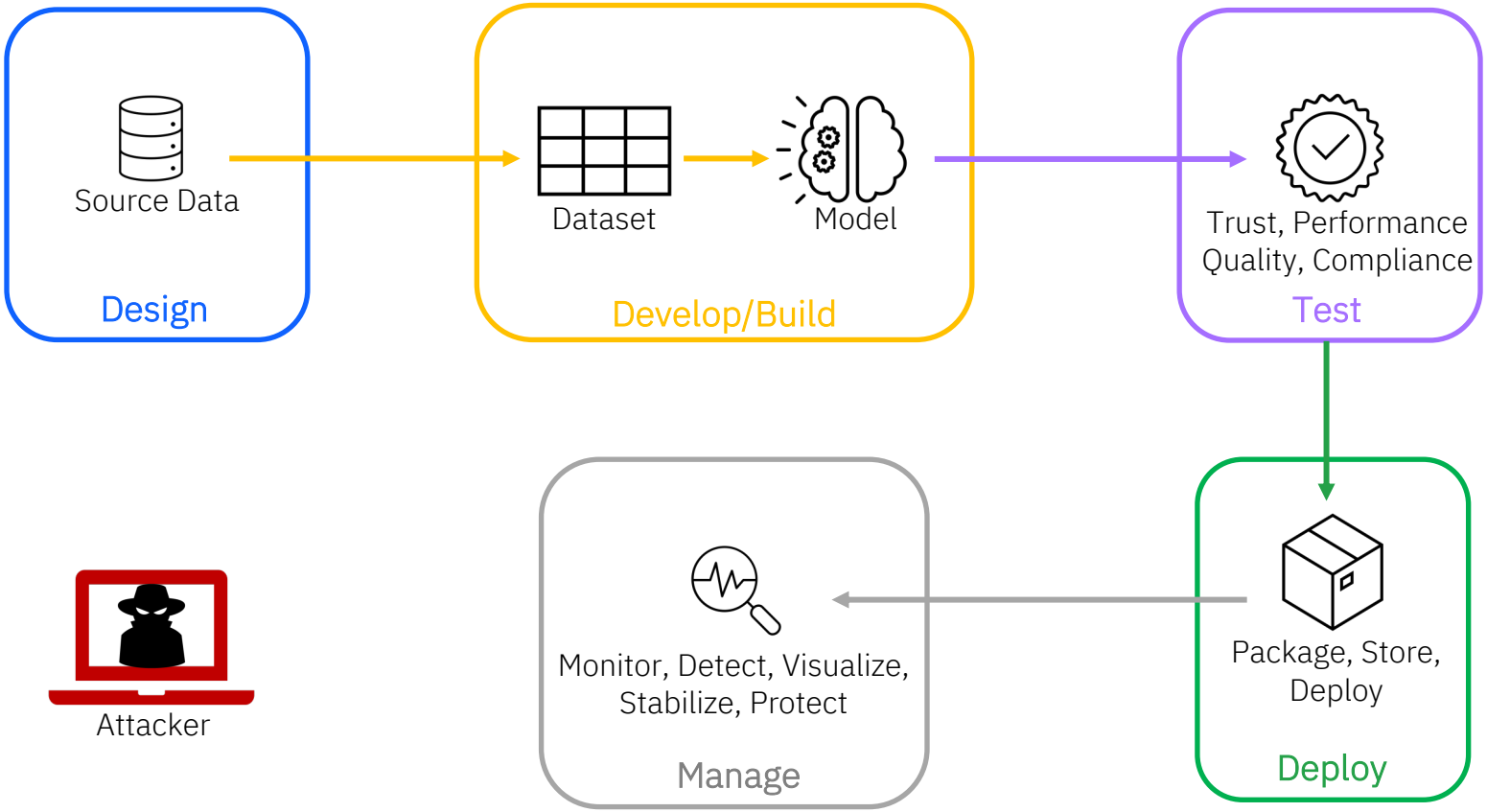
Data Extraction



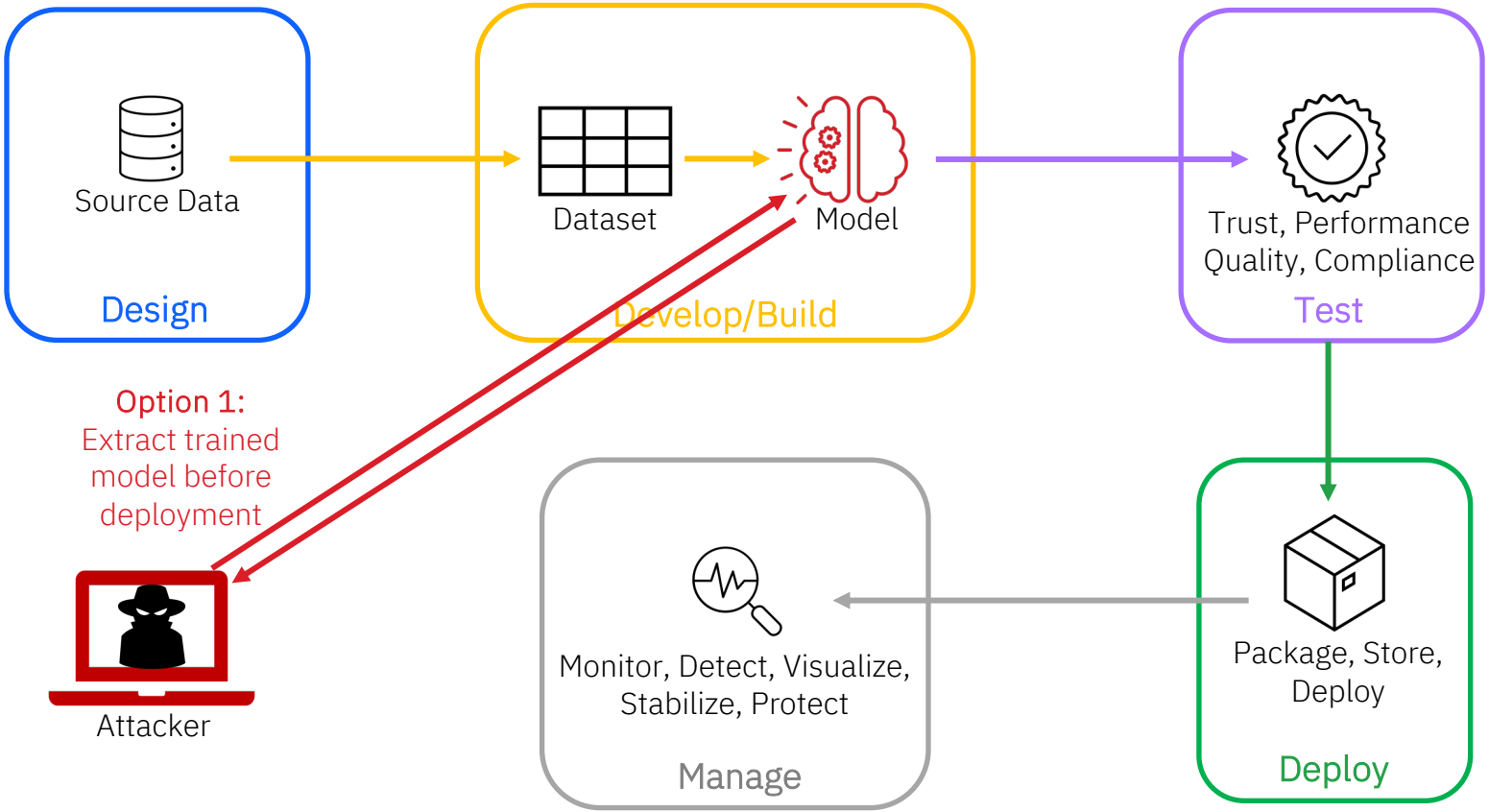
Data Extraction



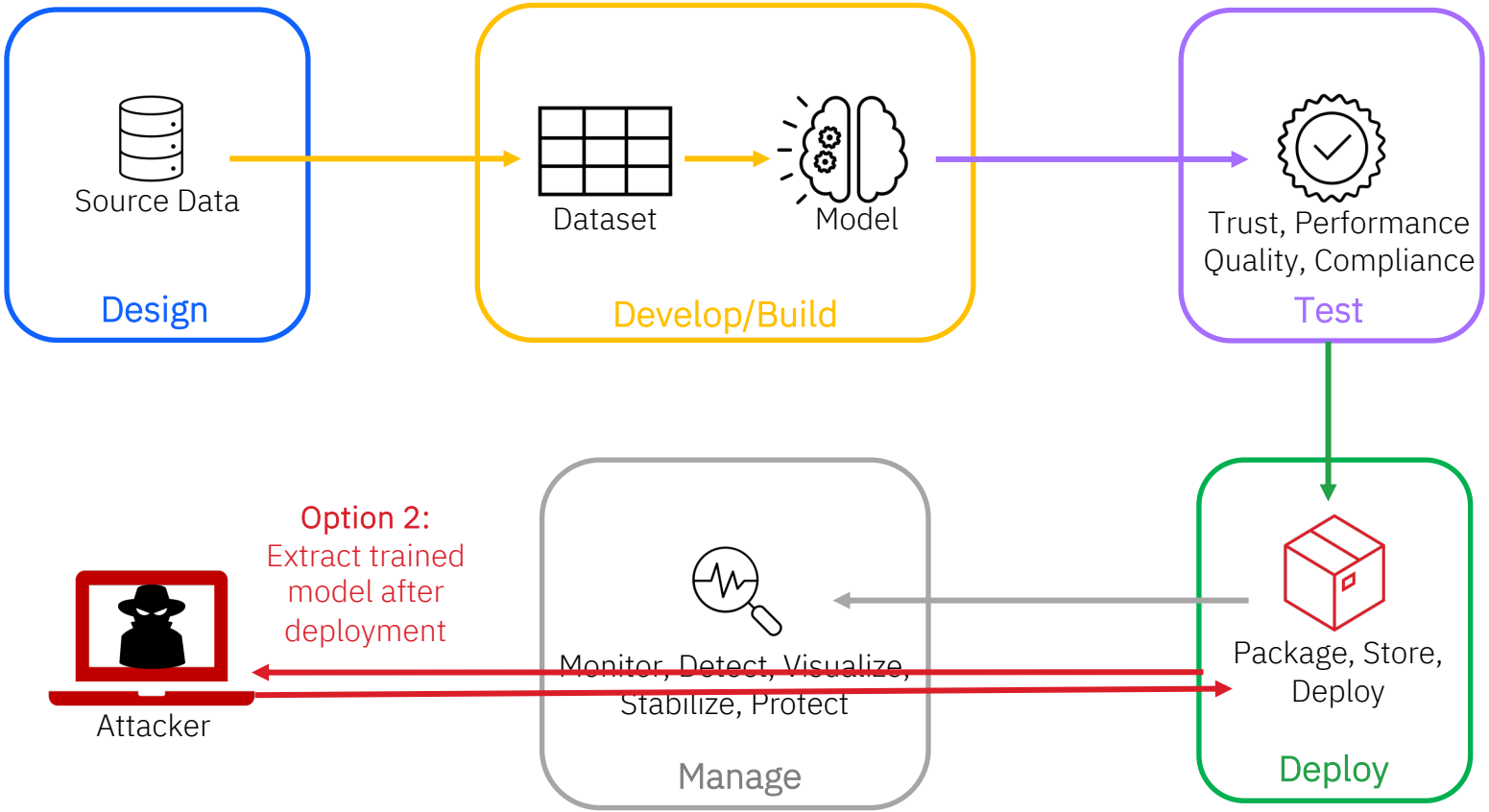
Model Extraction



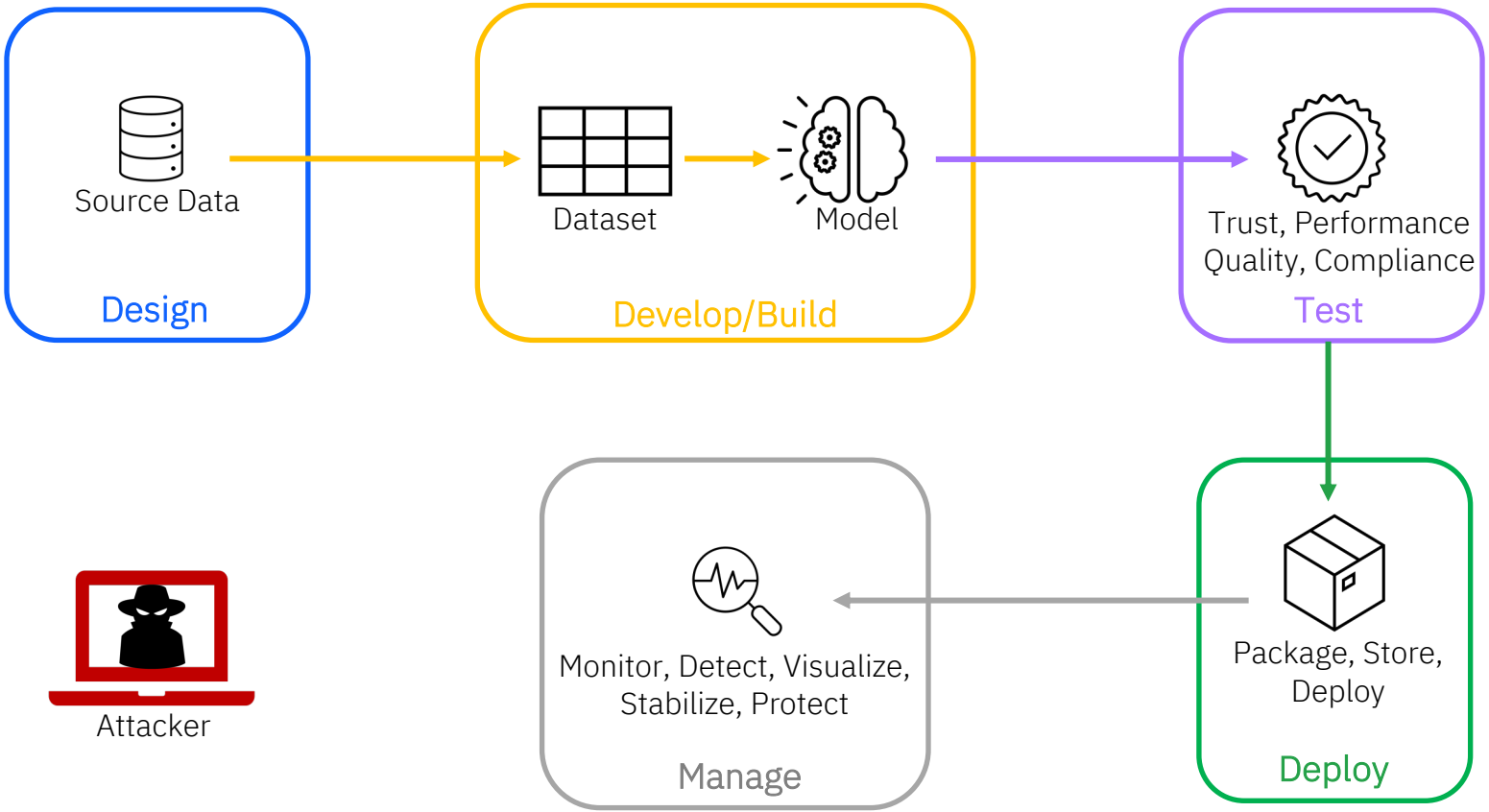
Model Extraction



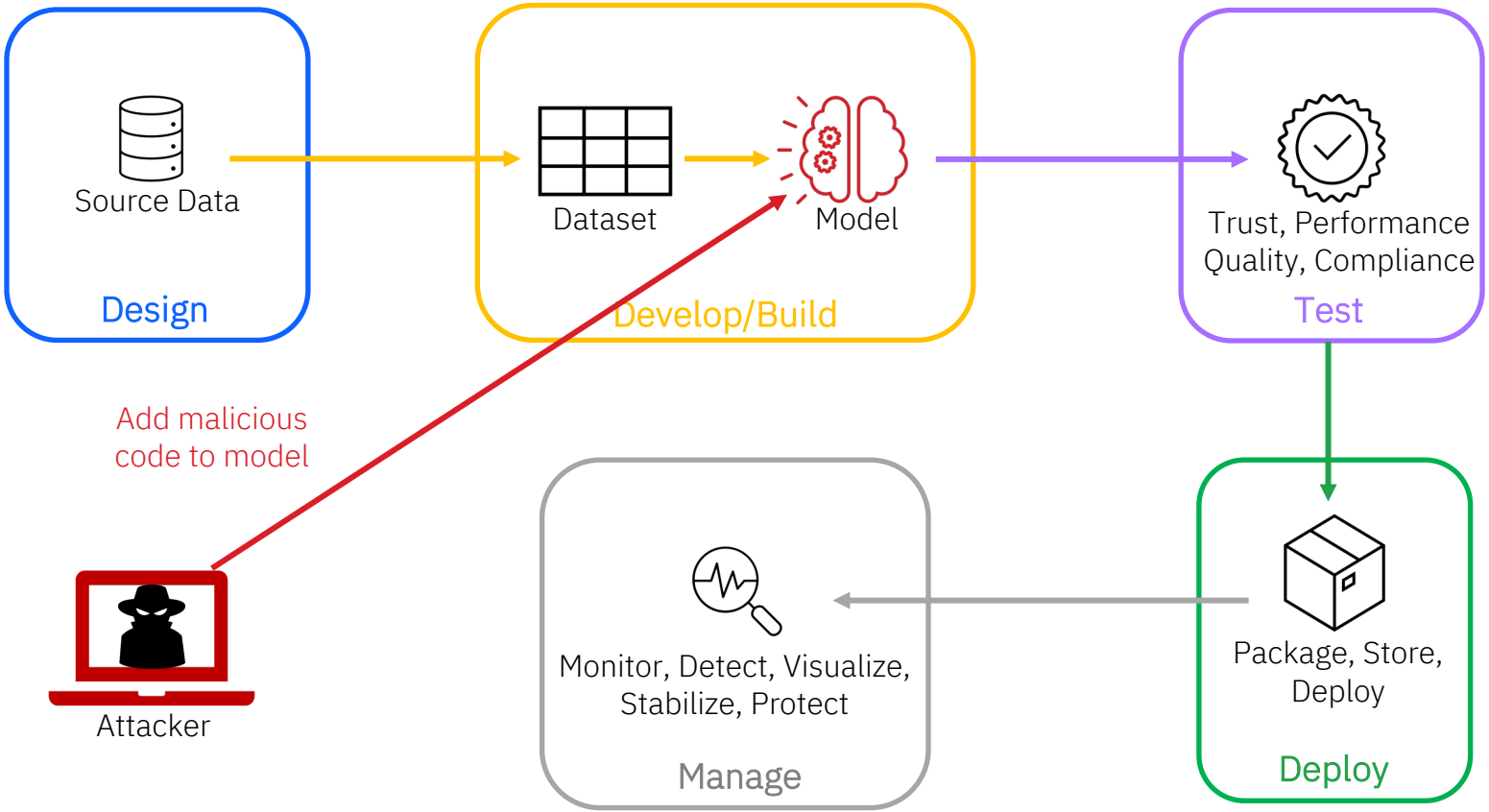
Model Extraction



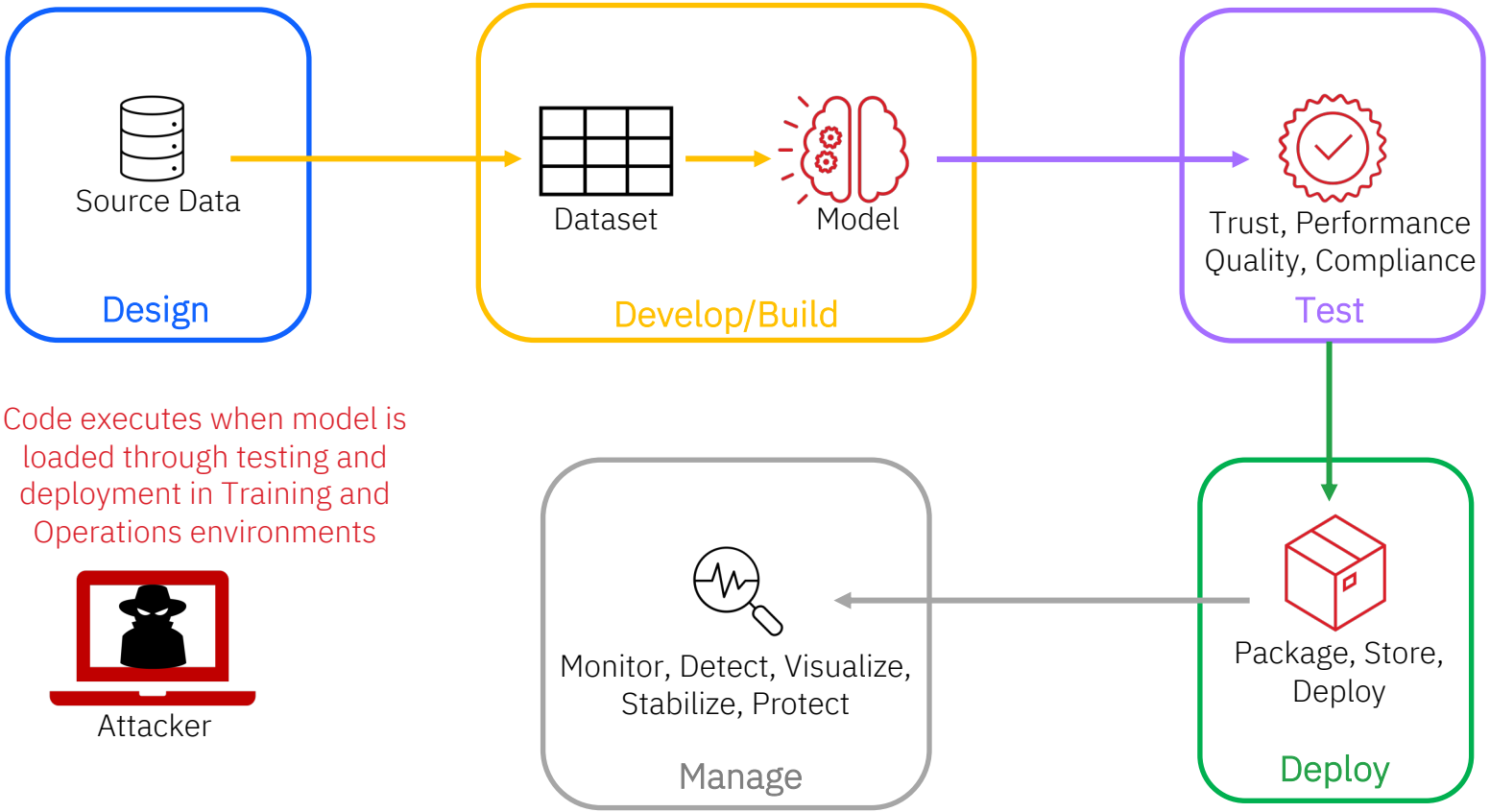
Code Execution



Code Execution

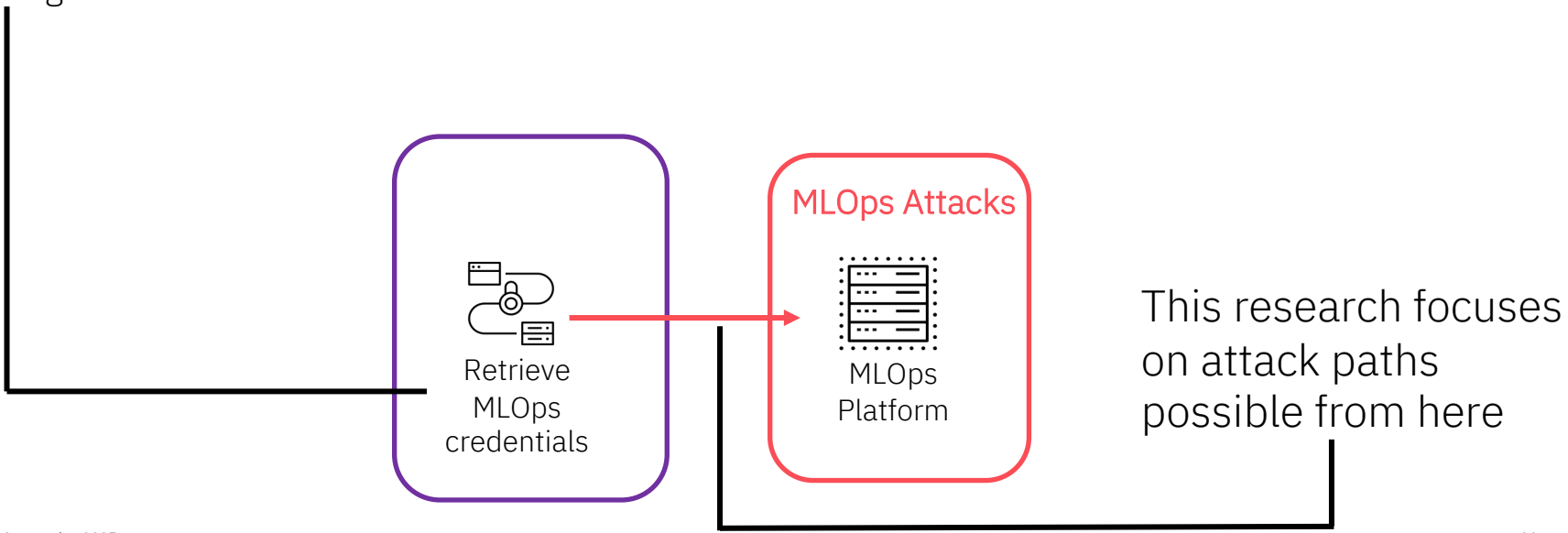


Code Execution

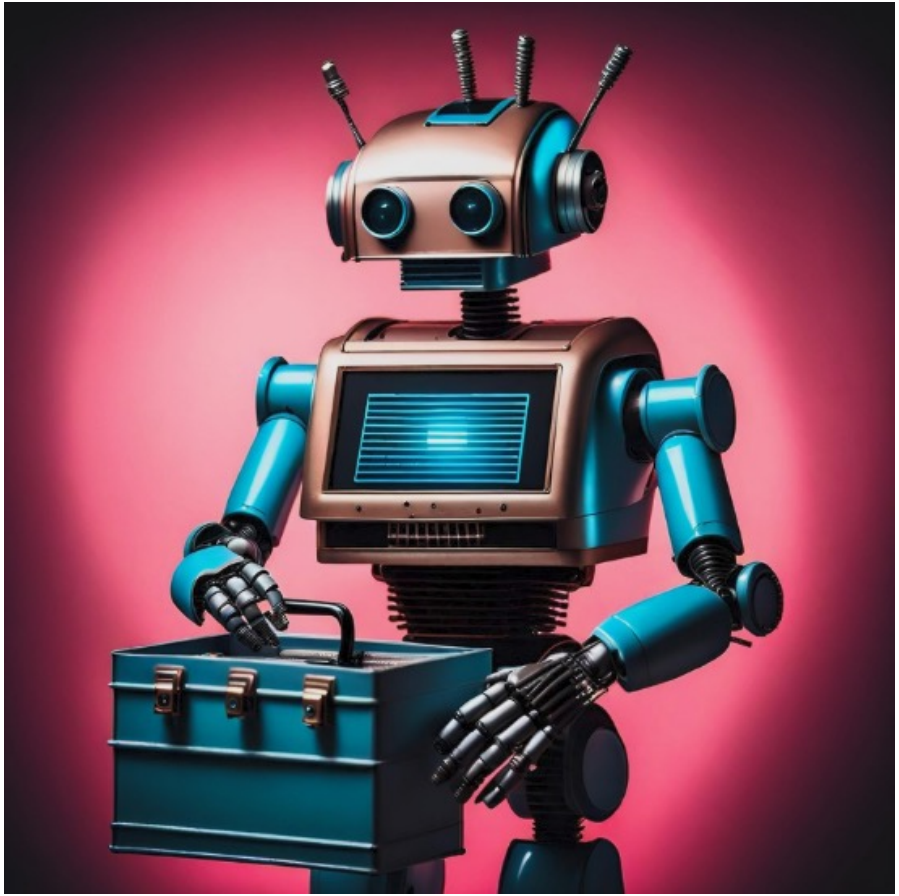


Research Focus

File Shares
Intranet Sites
User Workstations
Social Engineering
Public Resources (e.g., Code Repos)
Misconfigured Internal Network Recourses

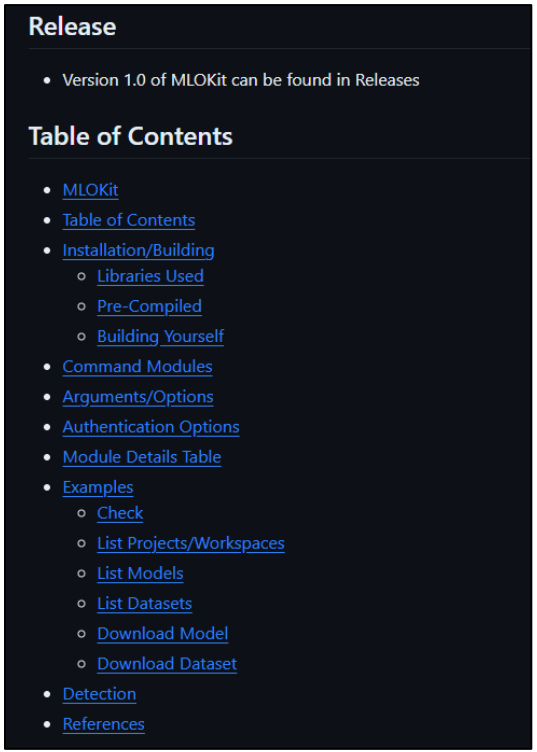


MLOKit



Background

github.com/xforcered/MLOKit



Release

- Version 1.0 of MLOKit can be found in Releases

Table of Contents

REST API Abuse
Conduct actions programmatically

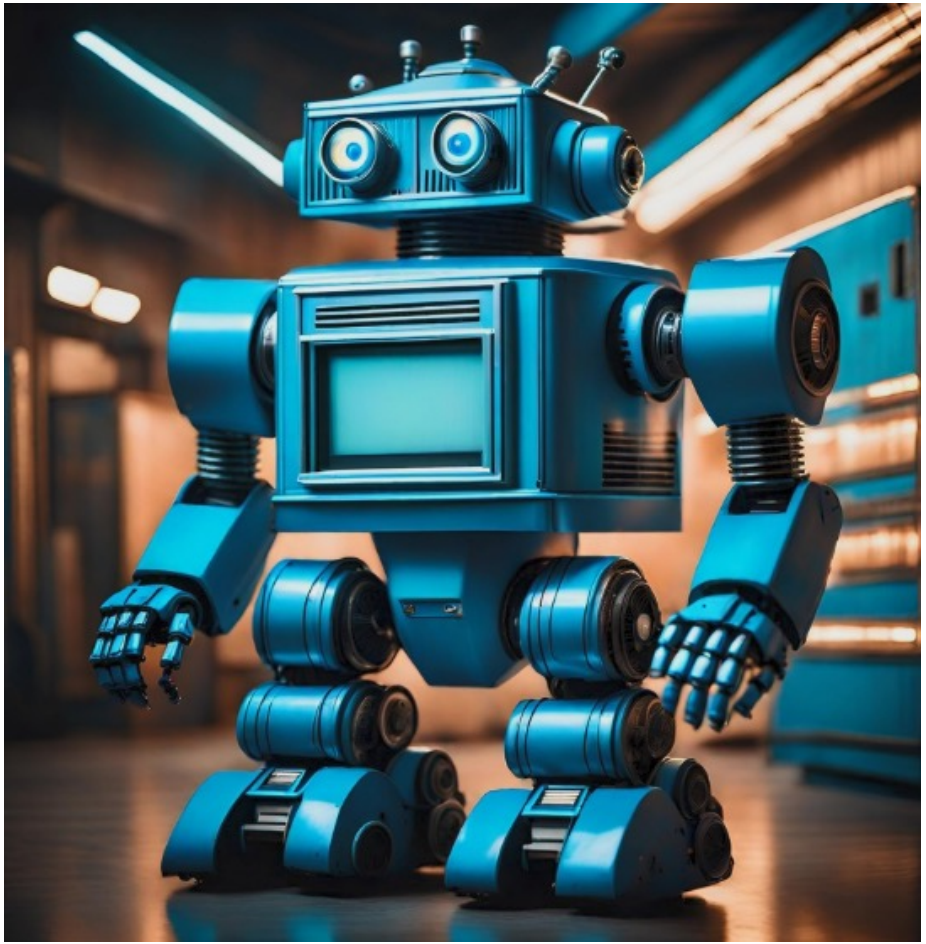
6 Modules
Recon, Data Extraction, Model Extraction

Authentication
Supports API Key or Access Token

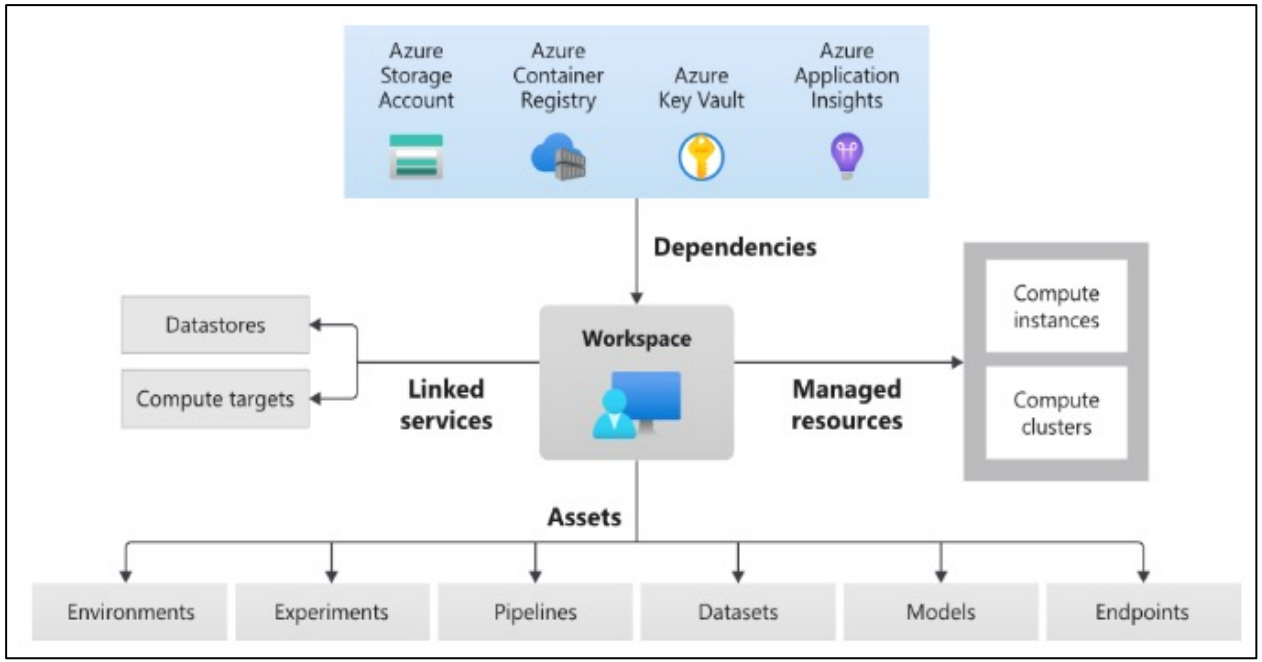
3 Supported Platforms
Azure ML, BigML, Vertex AI

Attacking MLOps Platforms

Azure ML



Key Terminology



https://www.trendmicro.com/vinfo/gb/security/news/cybercrime-and-digital-threats/uncovering-silent-threats-in-azure-machine-learning-service-part-I

Authentication

https://ml.azure.com/

Interactive
Authenticate with Entra ID authentication

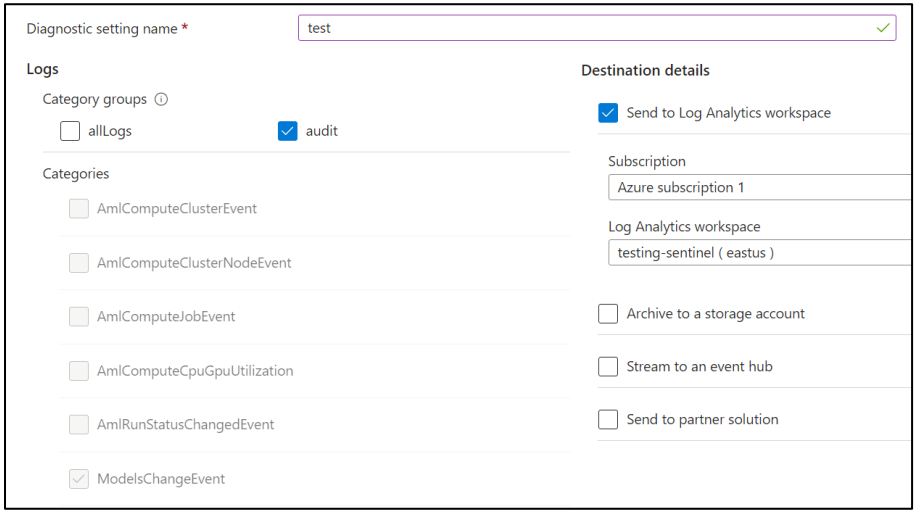
Service Principal
Use Service Principal ID and Key

Azure CLI Session
Use AZ command-line tool with ML extension

Managed Identity
Azure VM that can connect to workspace

Access Token
Used to authenticate to Azure ML REST API

Logging



Diagnostic Logging

Enable this as it is not by default

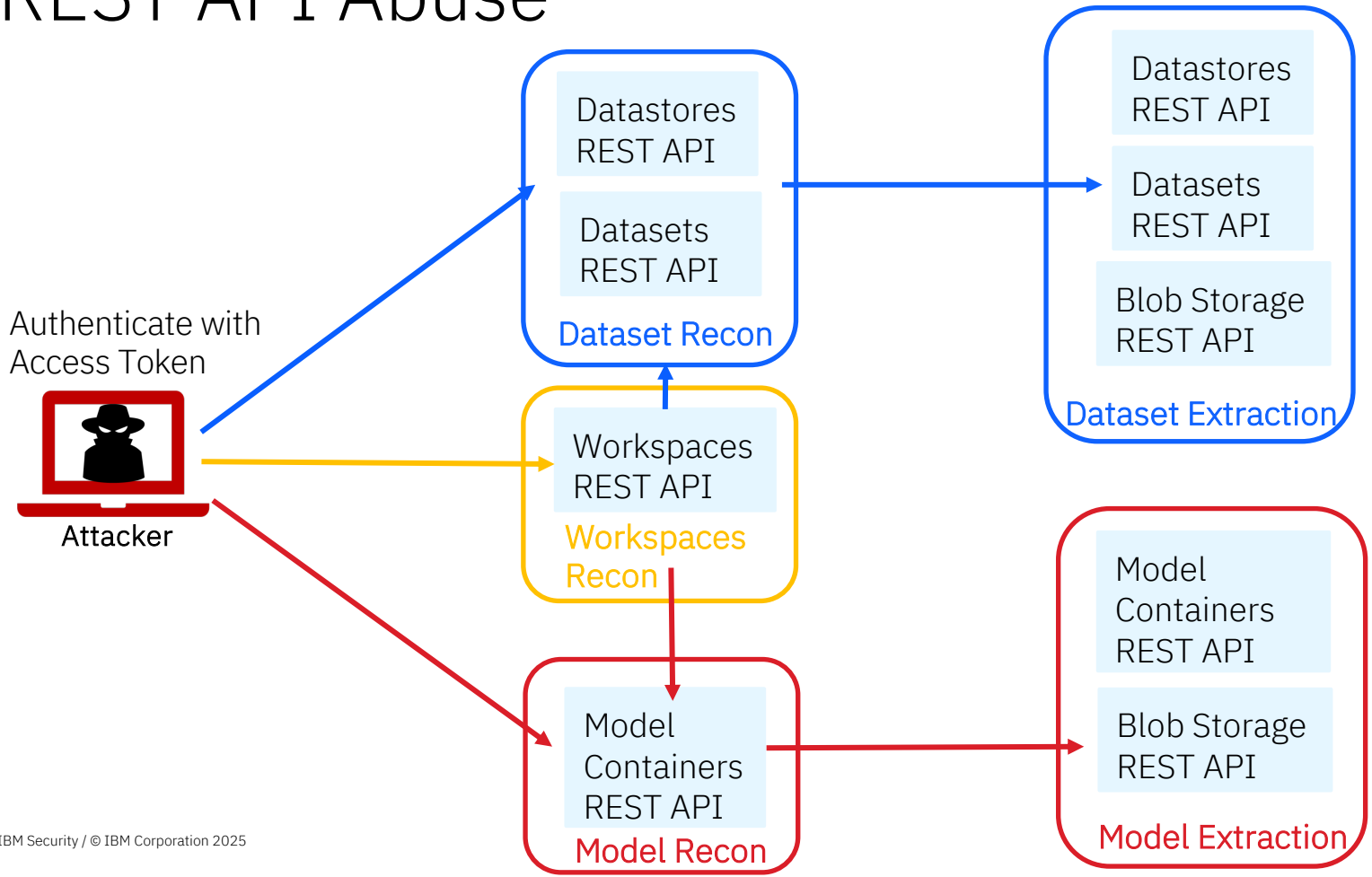
Audit Logs Category

Required to log auditable events

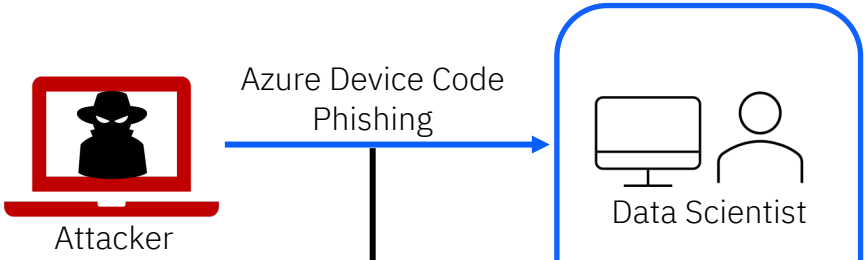
Send Logs to Analytics Workspace

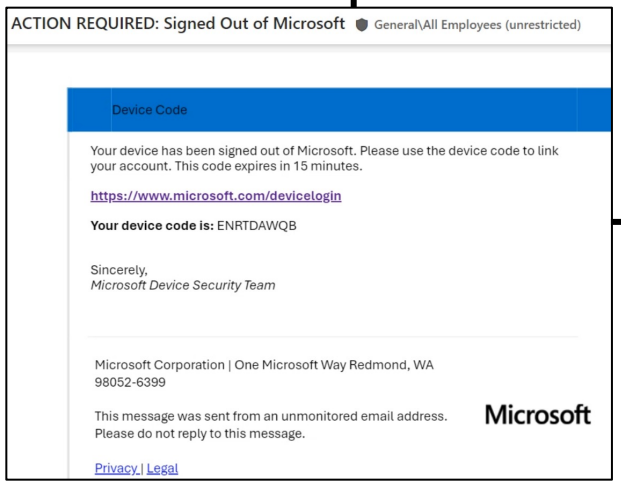
Detections can be built within Microsoft Sentinel

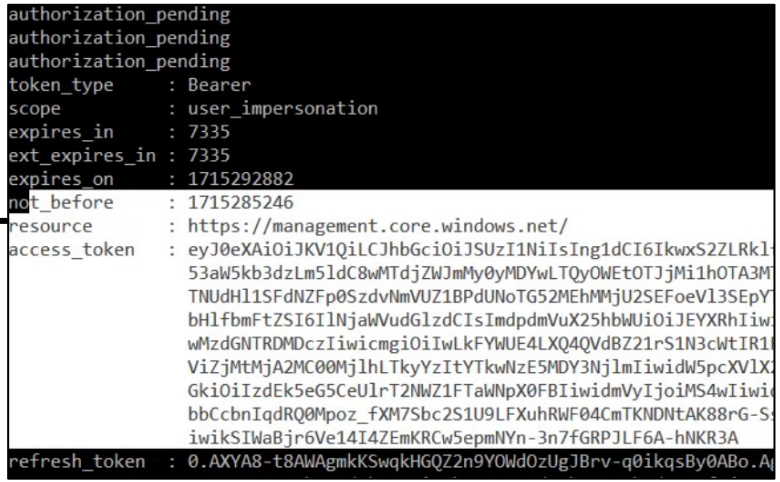
REST API Abuse



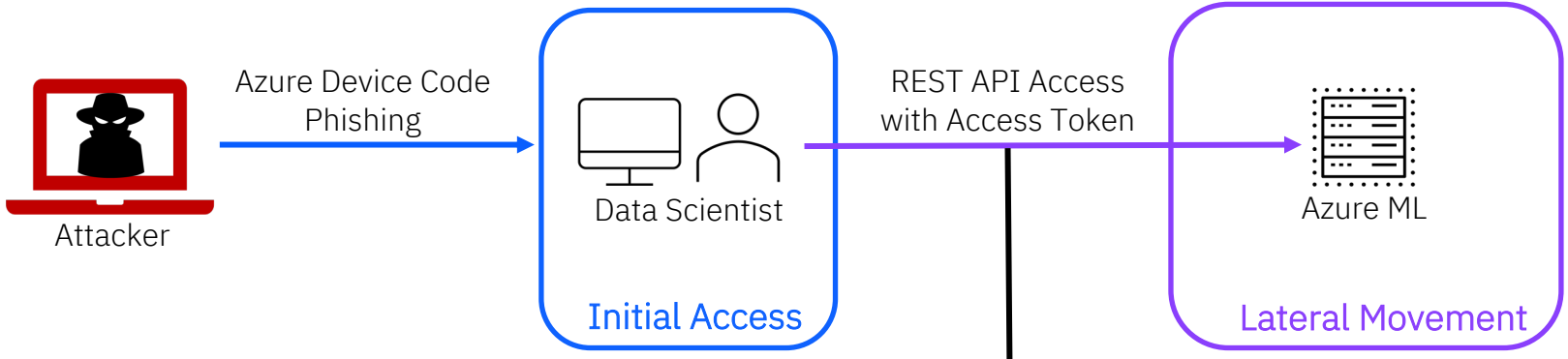
Demo: Azure ML Model Extraction Attack





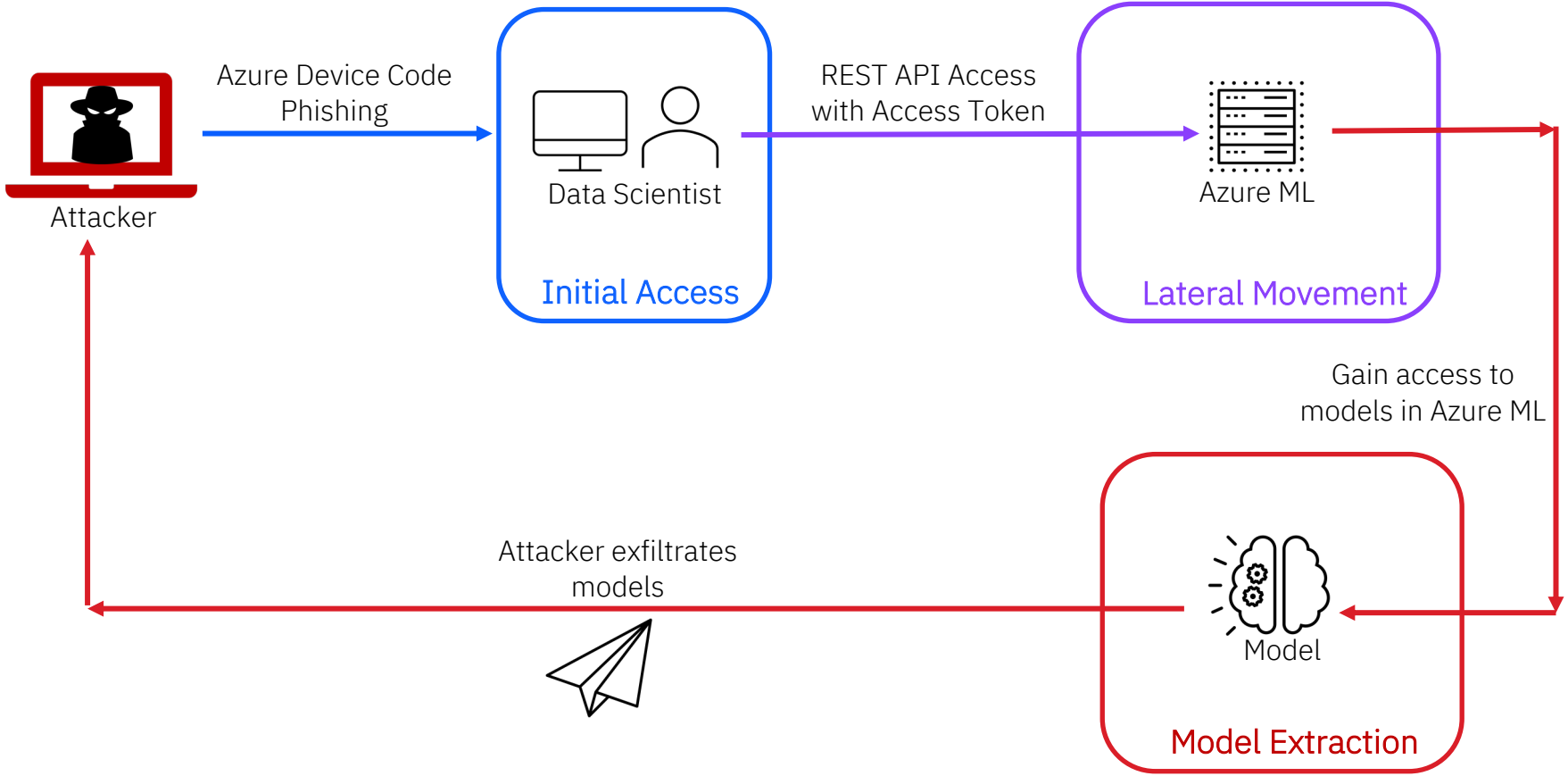


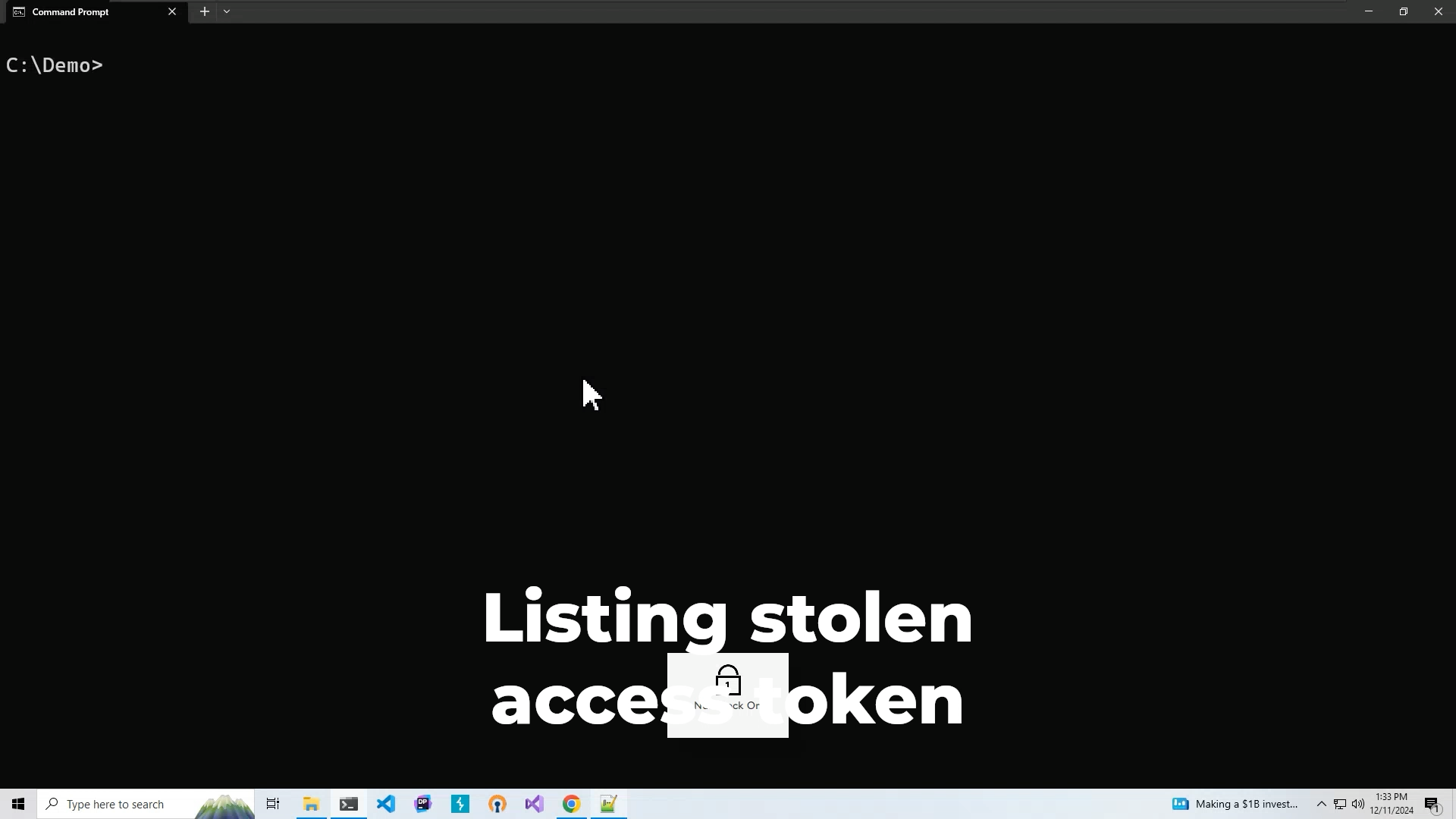
Demo: Azure ML Model Extraction Attack




```
Platform:       azureml
Timestamp:      5/9/2024 4:13:13 PM
================================================

[*] INFO: Performing check module for azureml

[*] INFO: Checking credentials provided

[+] SUCCESS: Credentials provided are VALID.

[*] INFO: Listing subscriptions user has acess to

                                        Name |                       Subscription ID |             Status
-------------------------------------------------------------------------------------------------------------
                          Azure subscription 1 |   47c5aaab-dbda-44ca-802e-00801de4db23 |            Enabled
```

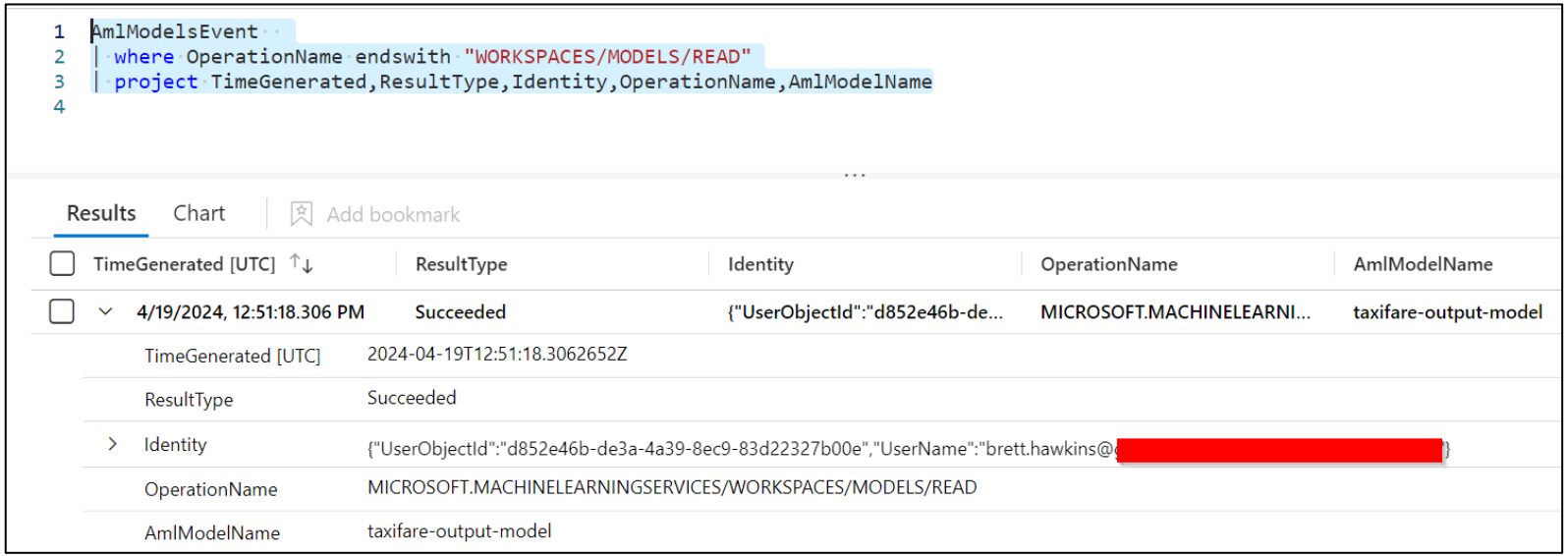
Demo: Azure ML Model Extraction Attack



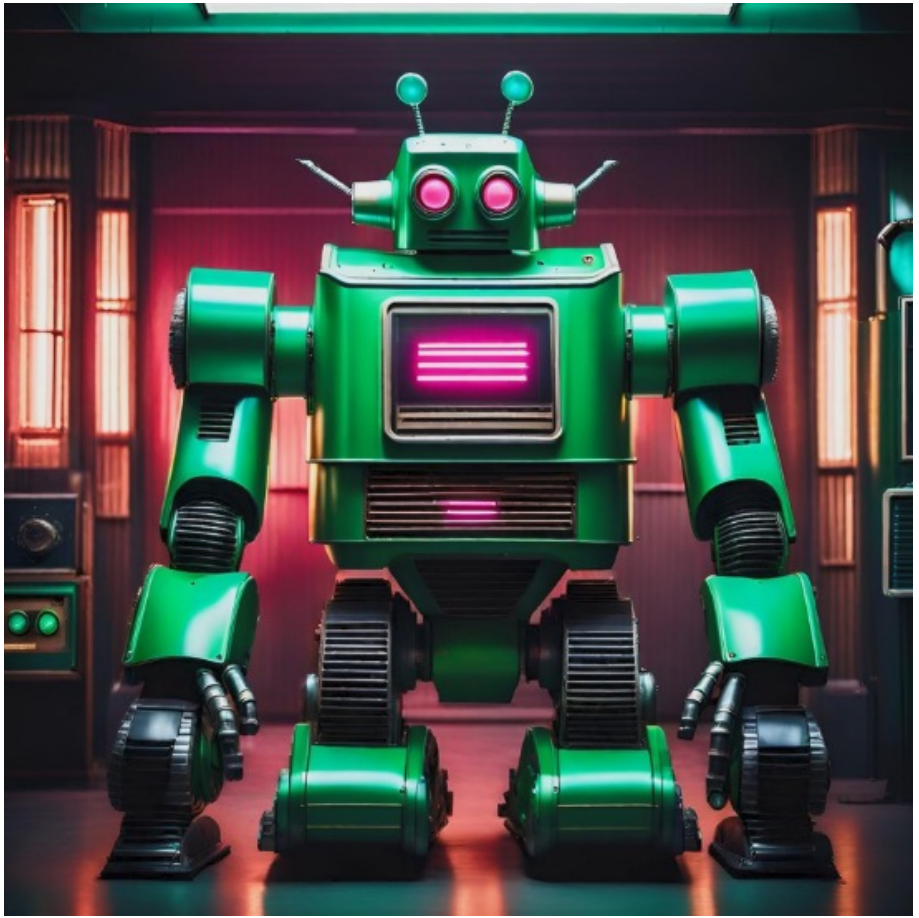

Command Prompt
C:\Demo>
Listing stolen
access token
Type here to search
Making a $1B invest...
1:33 PM
12/11/2024

Model Extraction - Logs

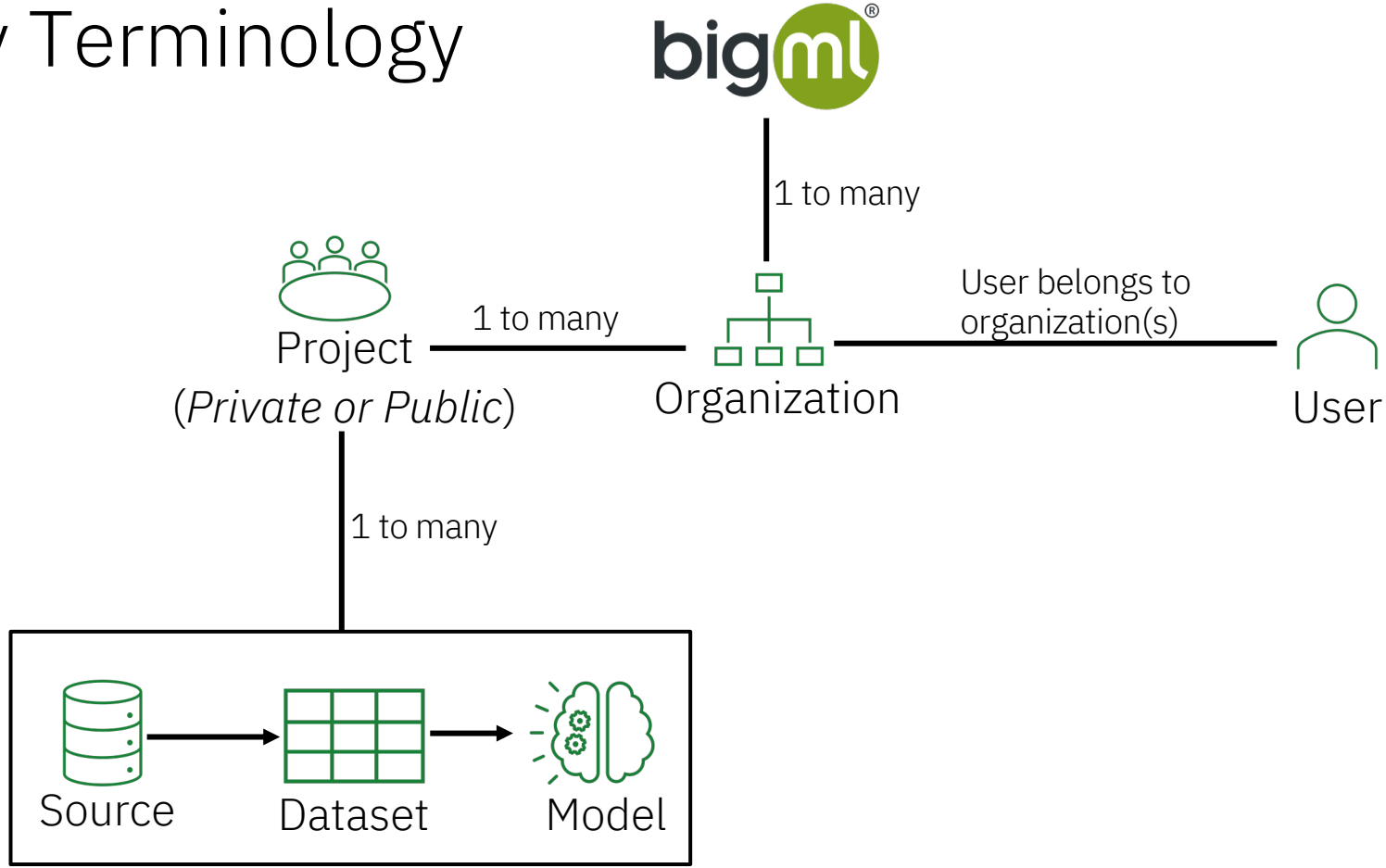
```
AmlModelsEvent
| where OperationName endswith "WORKSPACES/MODELS/READ"
| project TimeGenerated,ResultType,Identity,OperationName,AmlModelName
```



BigML

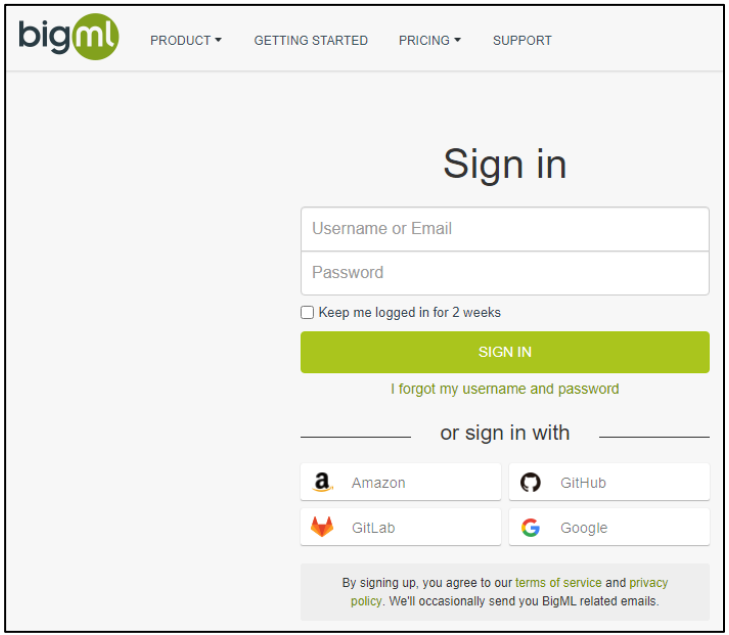


Key Terminology



Authentication

https://bigml.com/dashboard/organization/[ORGANIZATION_NAME]



BigMLer

Authenticate with API Key

REST API

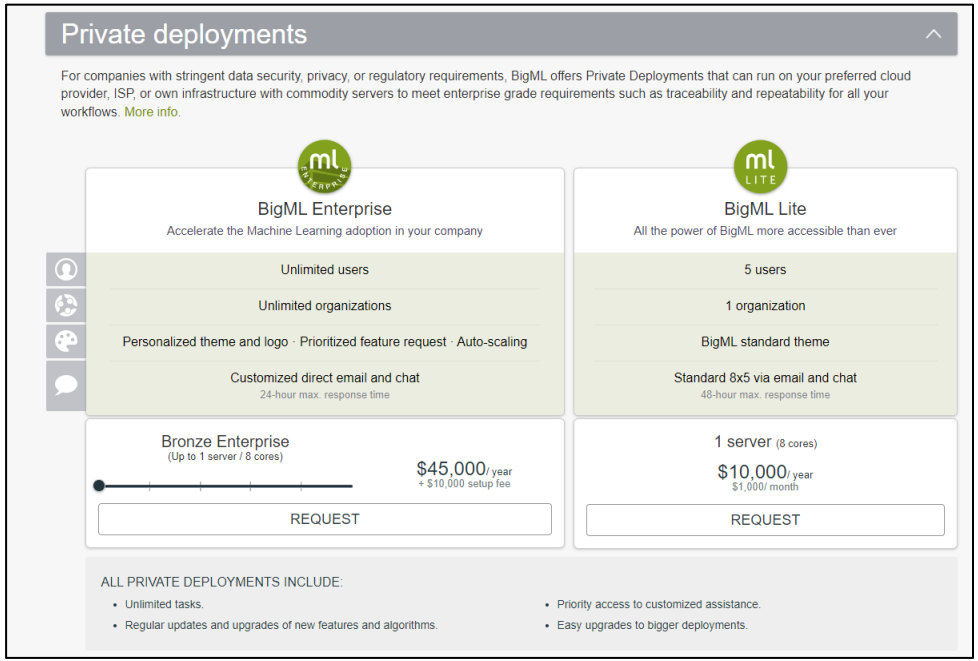
Authenticate with API Key

Web Interface

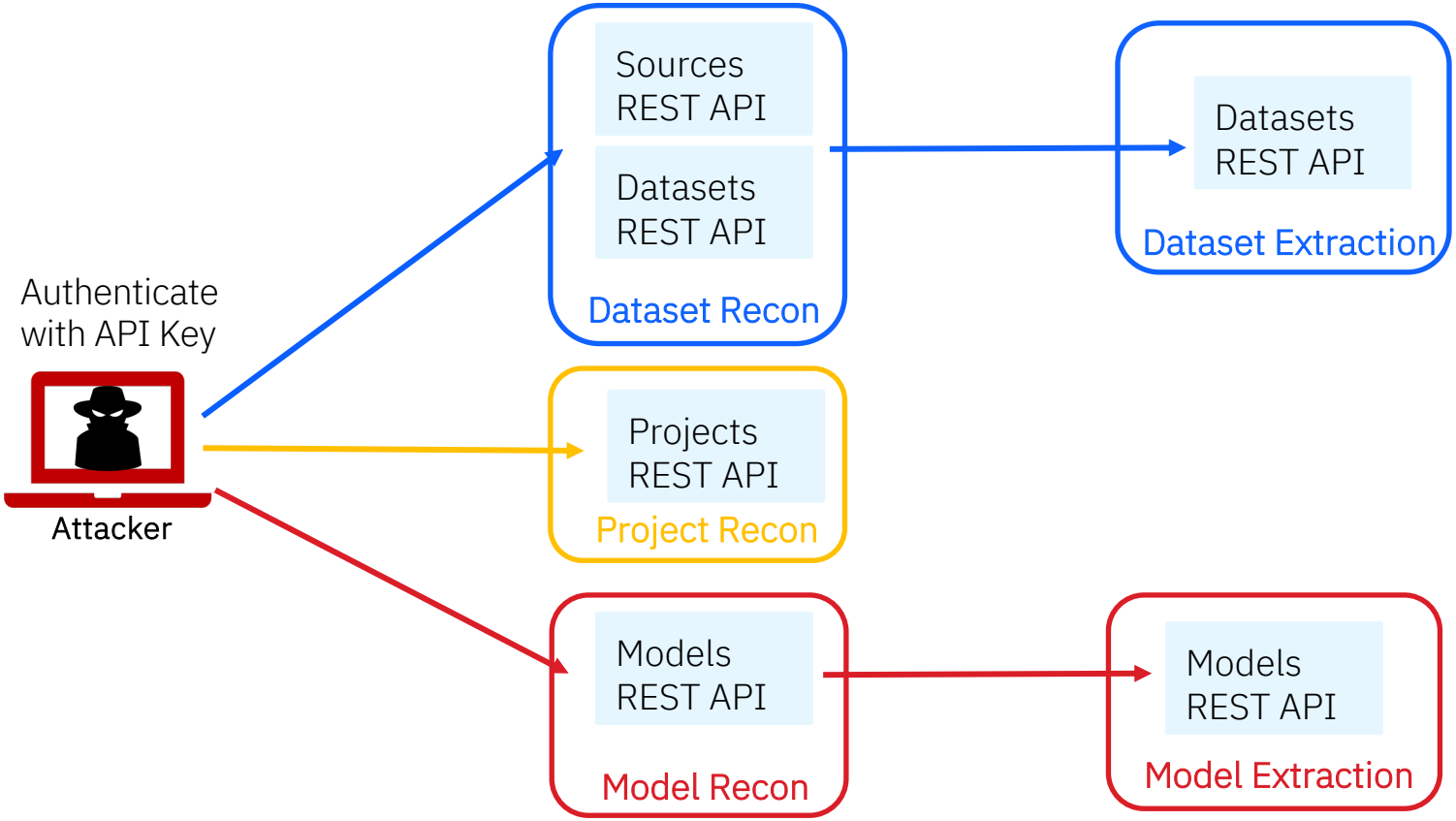
- MFA
- User/Pass
  - Native
  - 3rd Party

Logging

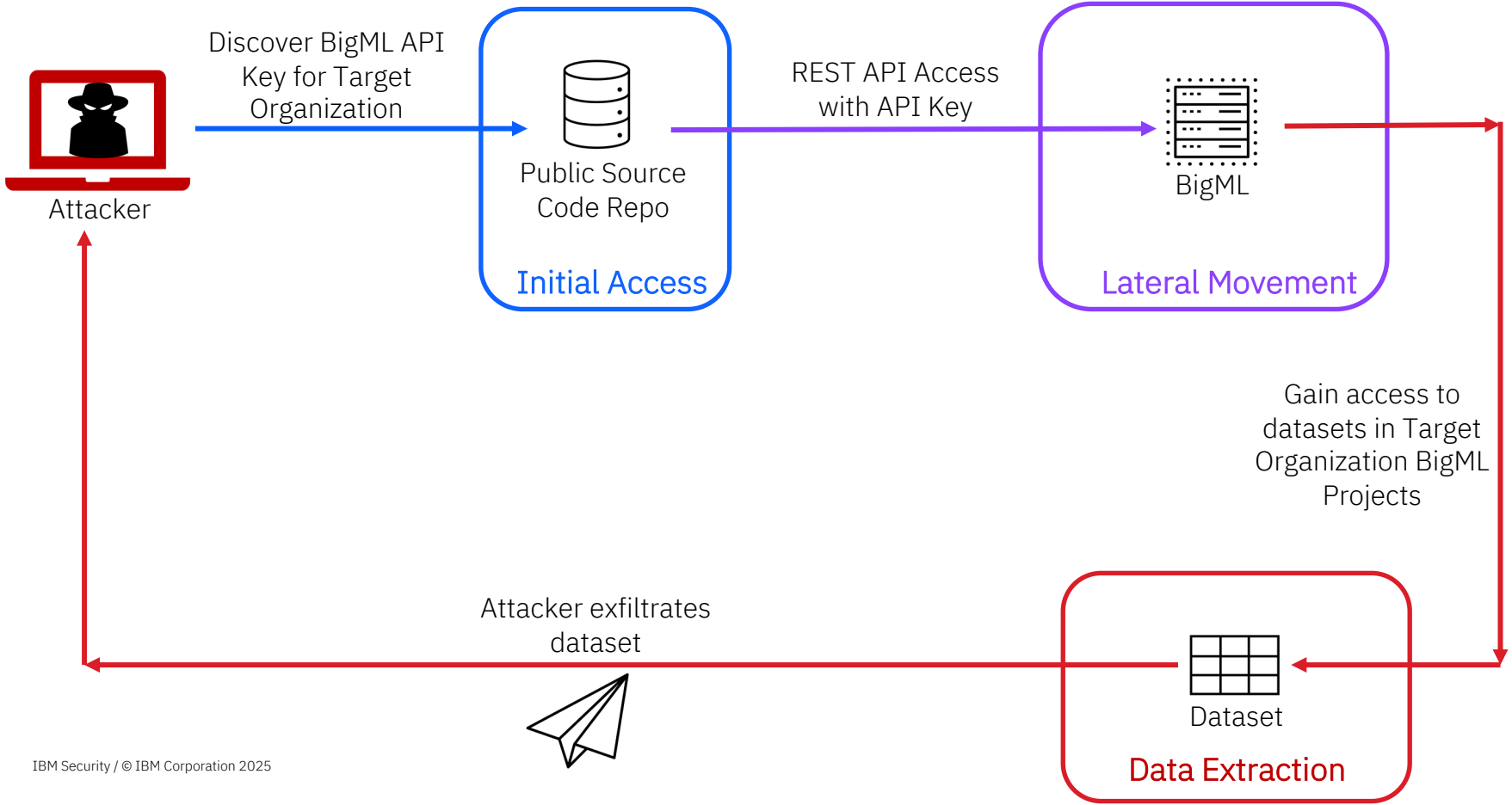
- Private deployment required for logging ability
- No logging in cloud-based BigML

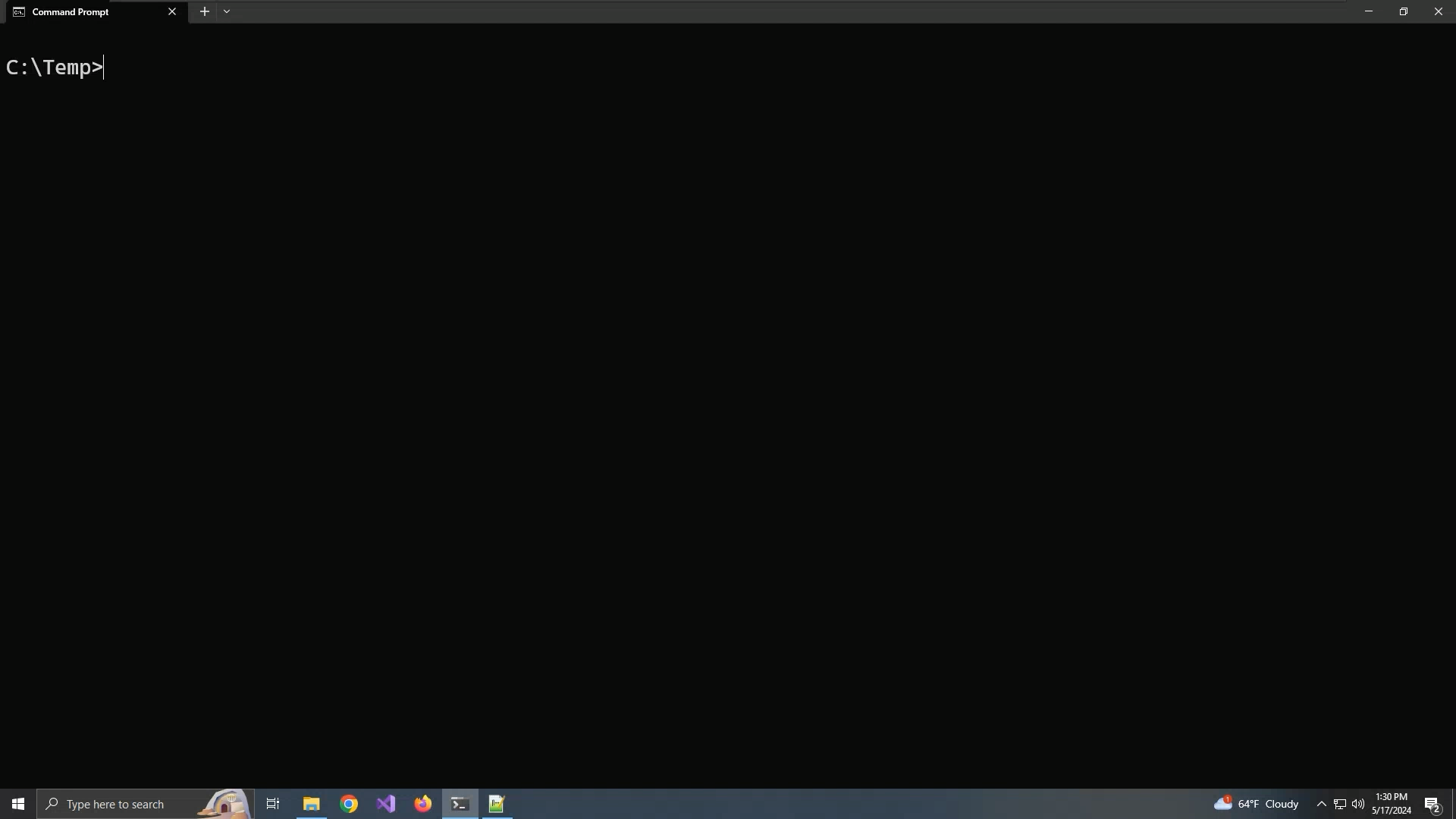


REST API Abuse



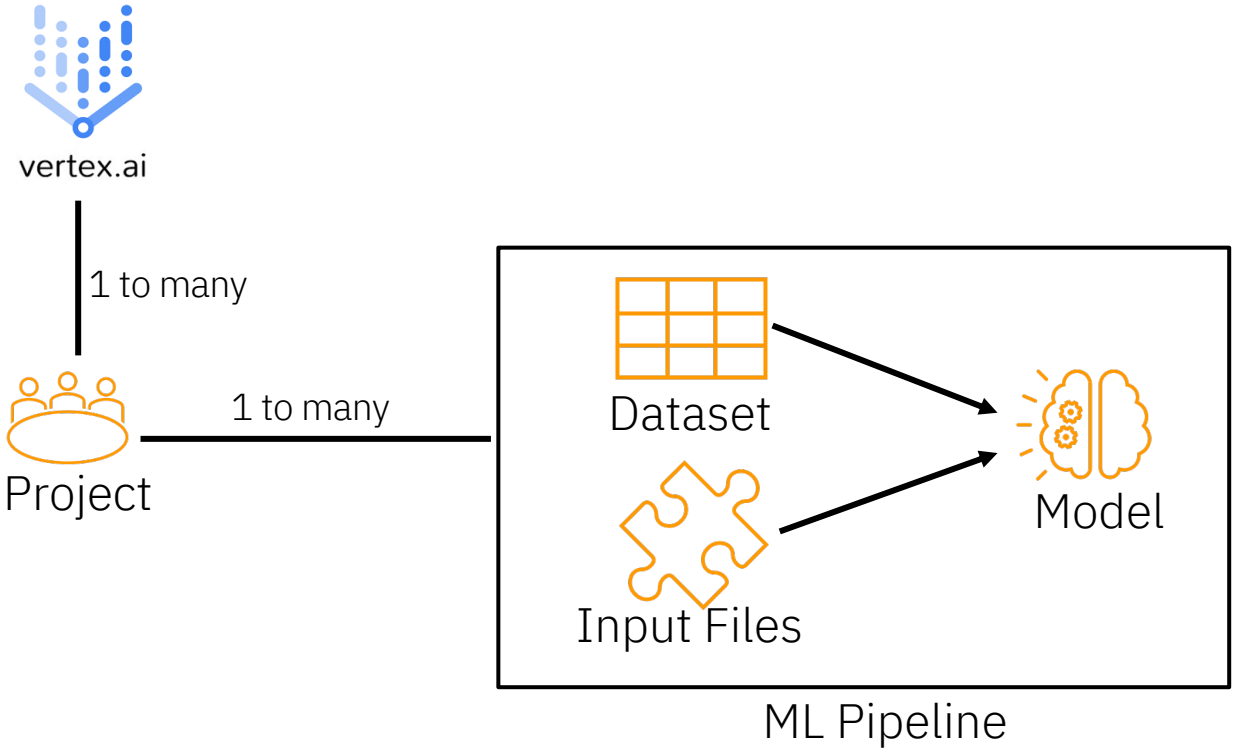
Demo: BigML Data Extraction Attack




Command Prompt
C:\Temp>
Type here to search
64°F Cloudy
1:30 PM
5/17/2024

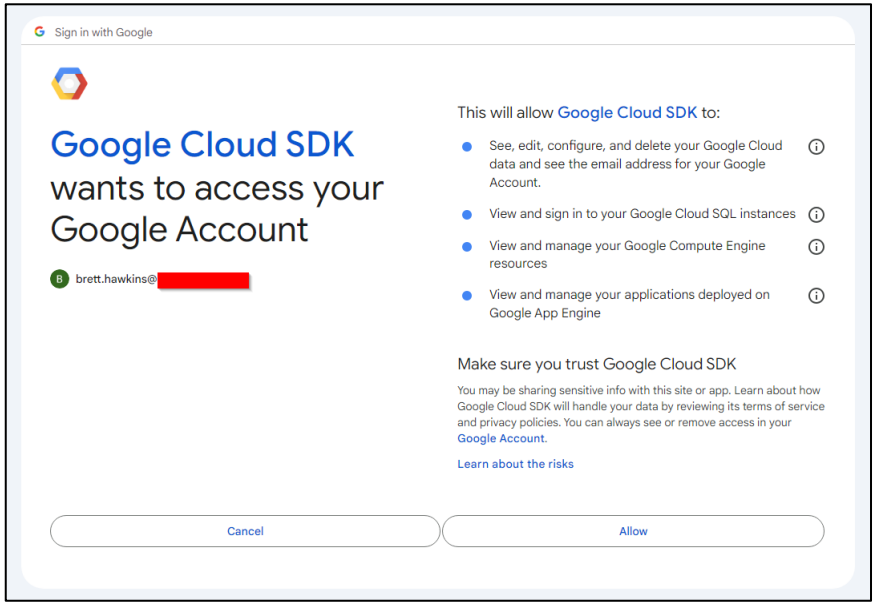
Google Cloud Vertex AI

Key Terminology



Authentication

https://console.cloud.google.com/vertex-ai

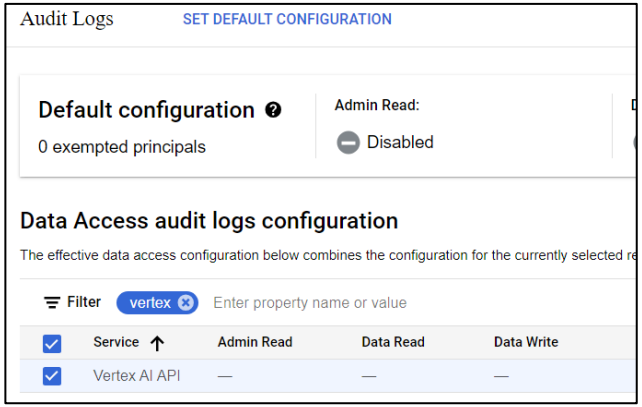


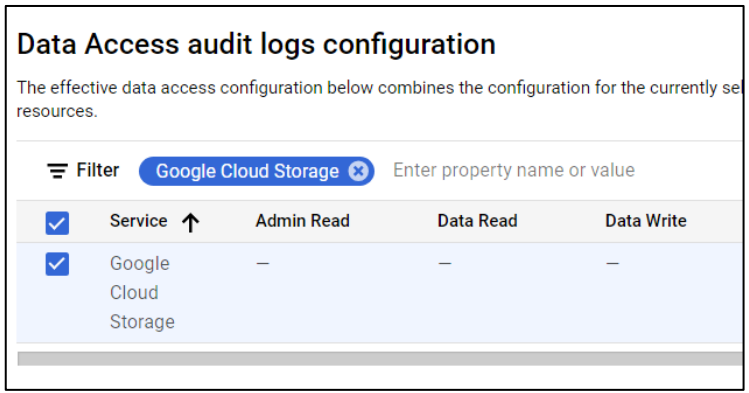
Client Libraries
Authenticate with Application Default Credentials (ADC)

Google Cloud CLI
Authenticate with GCloud account

REST API
Authenticate with Access Token

Logging





Vertex AI Data Access Audit Logs

Enable this as it is not by default

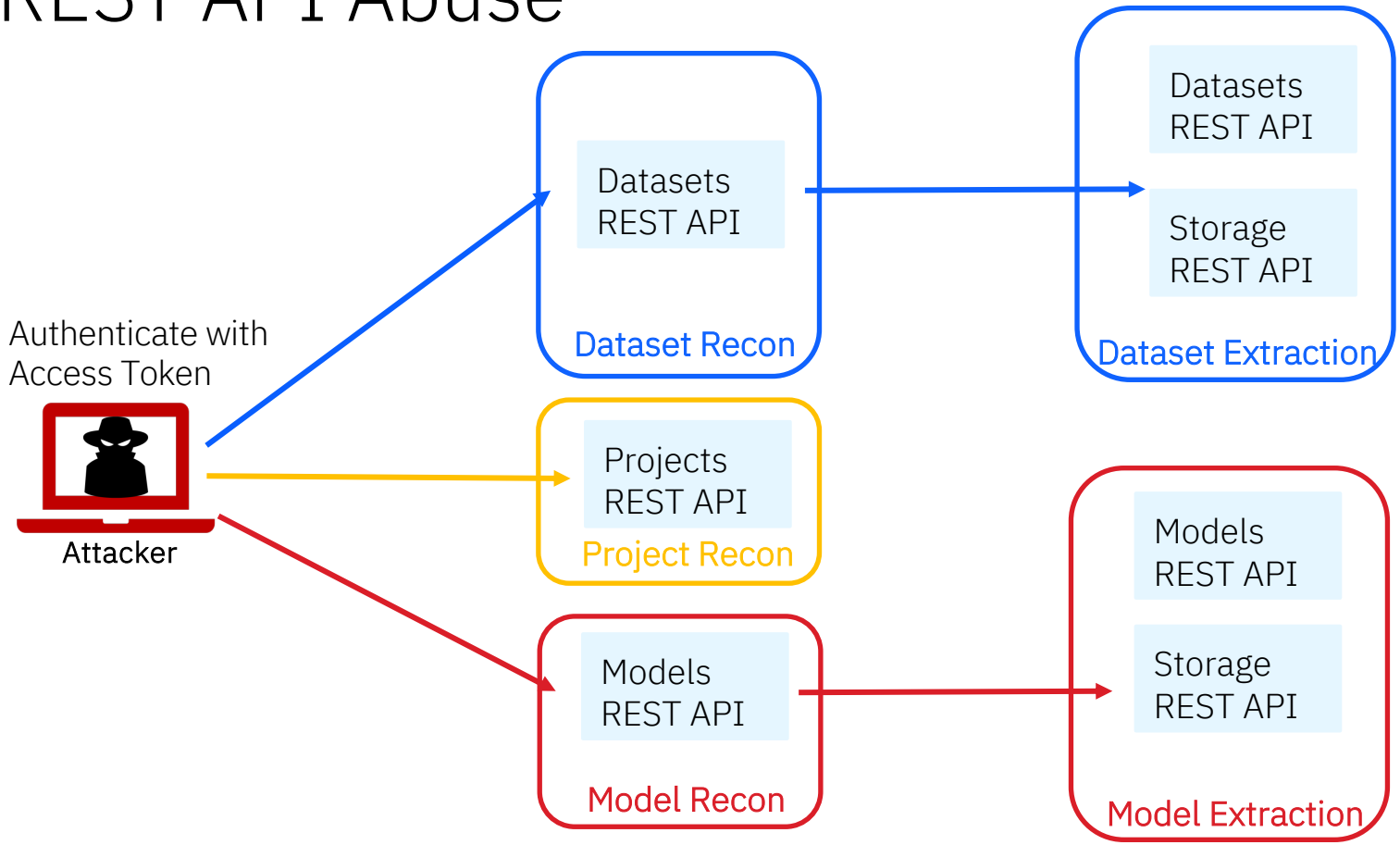
Google Cloud Storage Audit Logs

Enable this as it is not by default

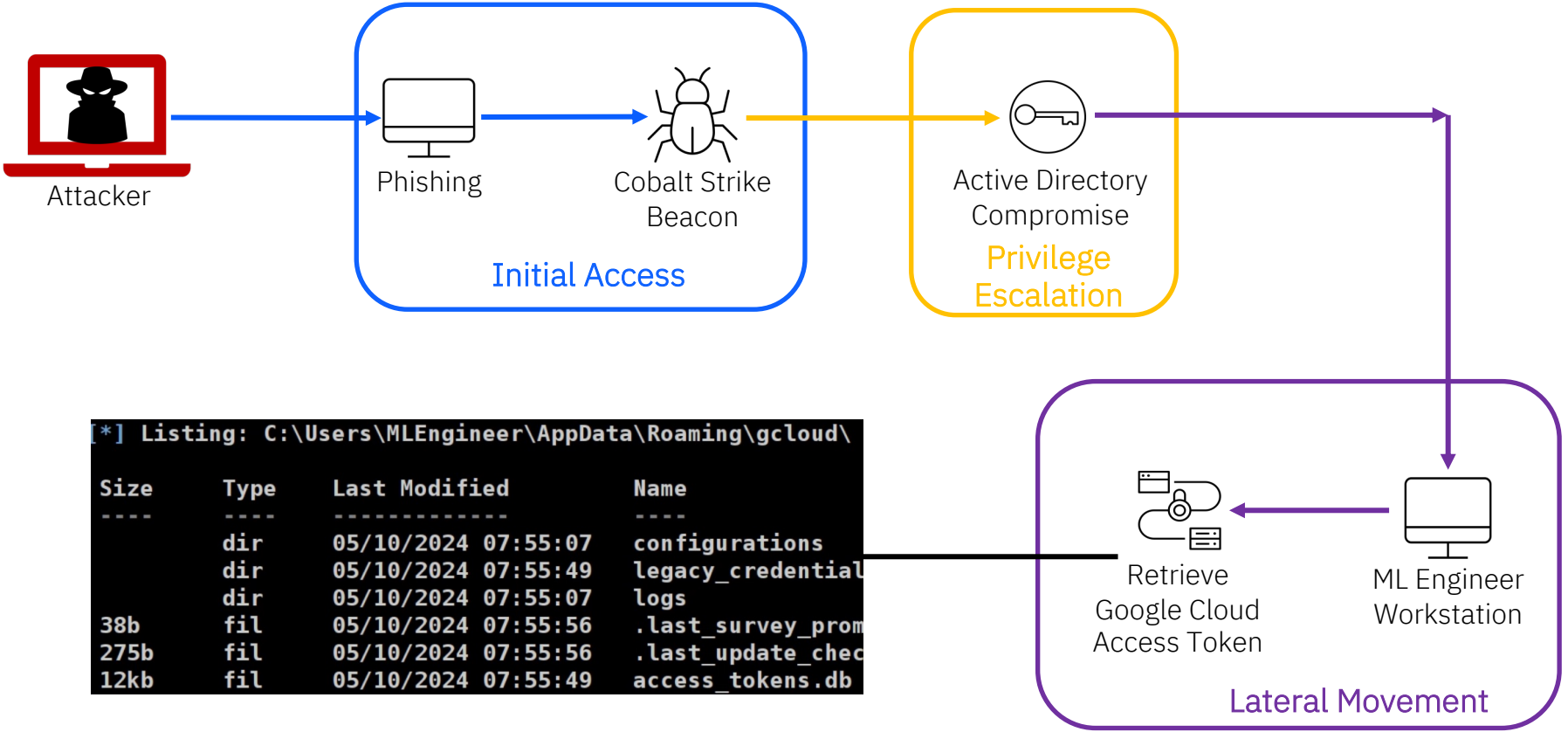
Google Cloud Logs Explorer

Logs will be populated here that log alerts can be built

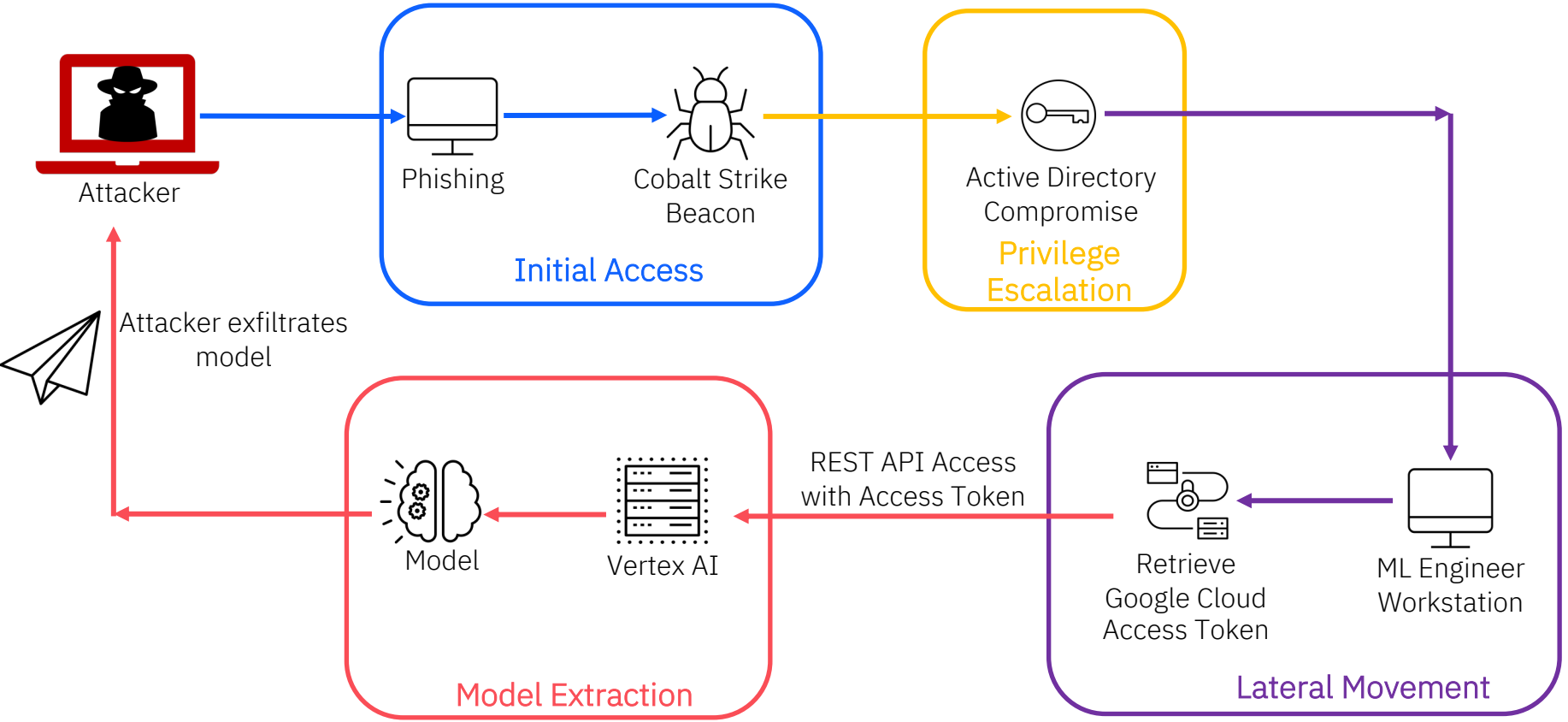
REST API Abuse



Demo: Vertex AI Model Extraction Attack



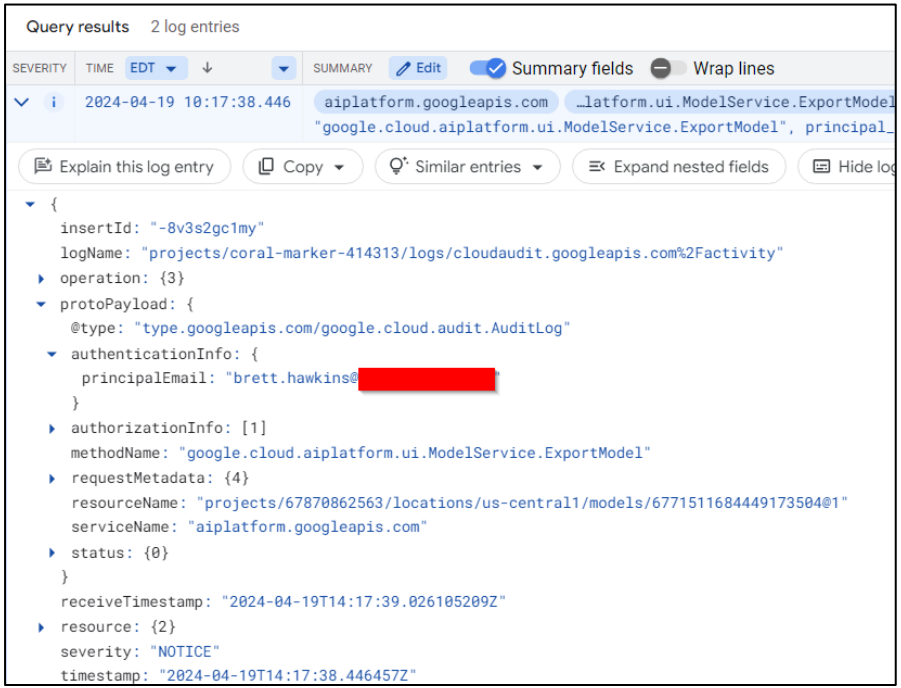
Demo: Vertex AI Model Extraction Attack




Activities
aggressor-Aggressor
May 17 14:03
Cobalt Strike
Cobalt Strike View Payloads Attacks Site Management Reporting Help
external internal listener user computer note process pid arch last sleep
192.168.1.37 192.168.1.37 https MLEngineer DESKTOP-YFCE... werfault.exe 12524 x64 2s interactive
Dumped GCloud
CLI Access Token
[DESKTOP-YFCE20A] - x64 | MLEngineer | 12524 - x64
last:2s
beacon>

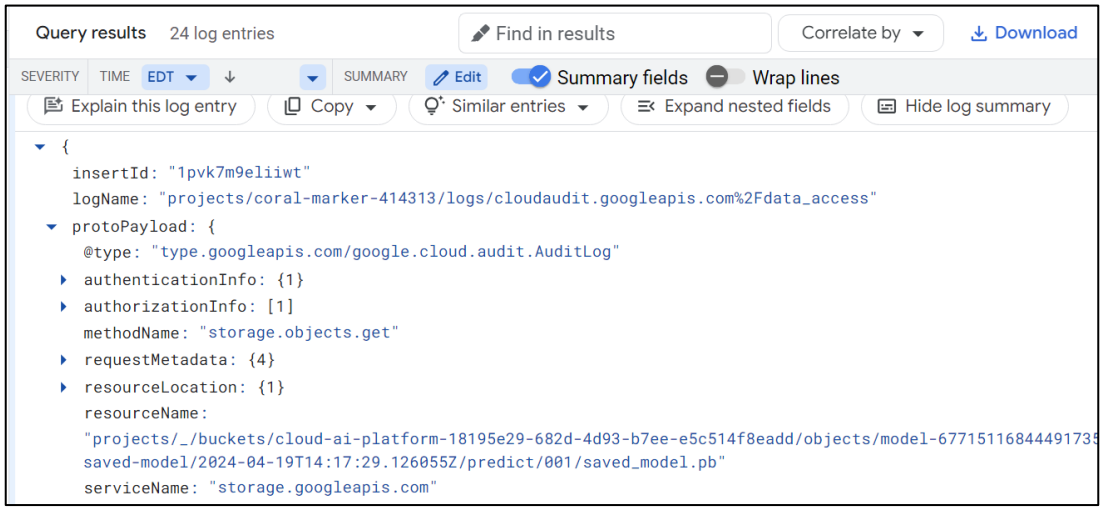
Model Extraction - Logs

```
protoPayload.methodName="google.cloud.aiplatform.ui.ModelService.ExportModel"
protoPayload.methodName="google.cloud.aiplatform.v1.ModelService.ExportModel"
```

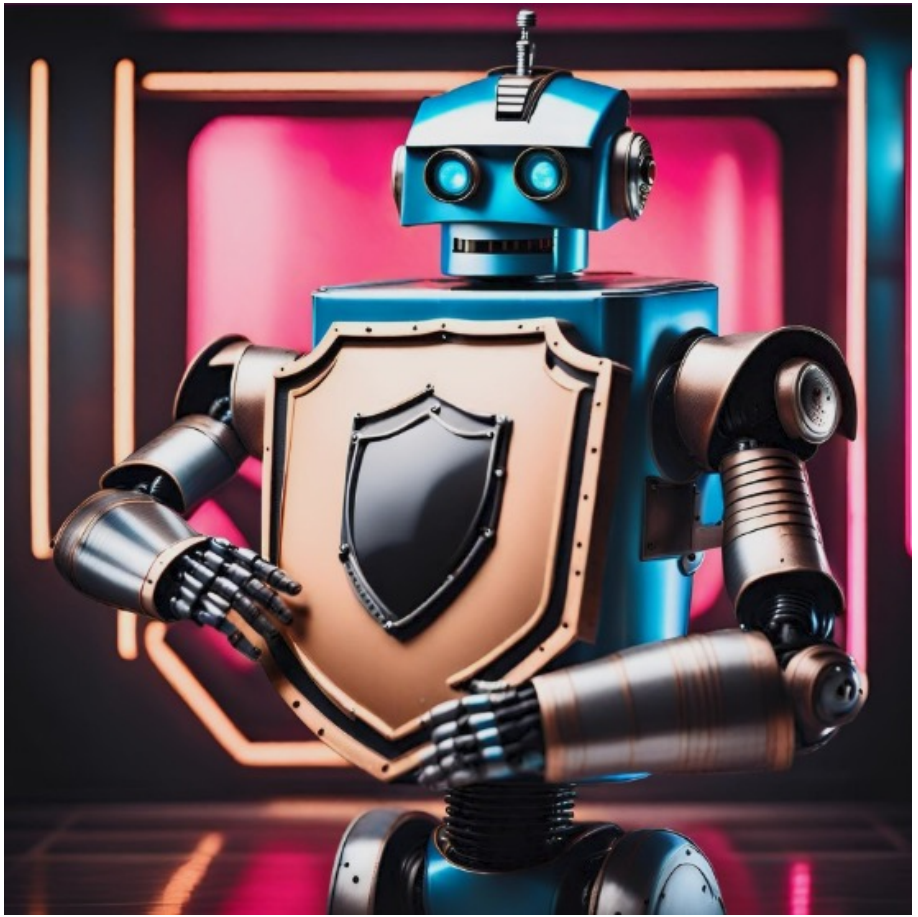


Model Extraction - Logs

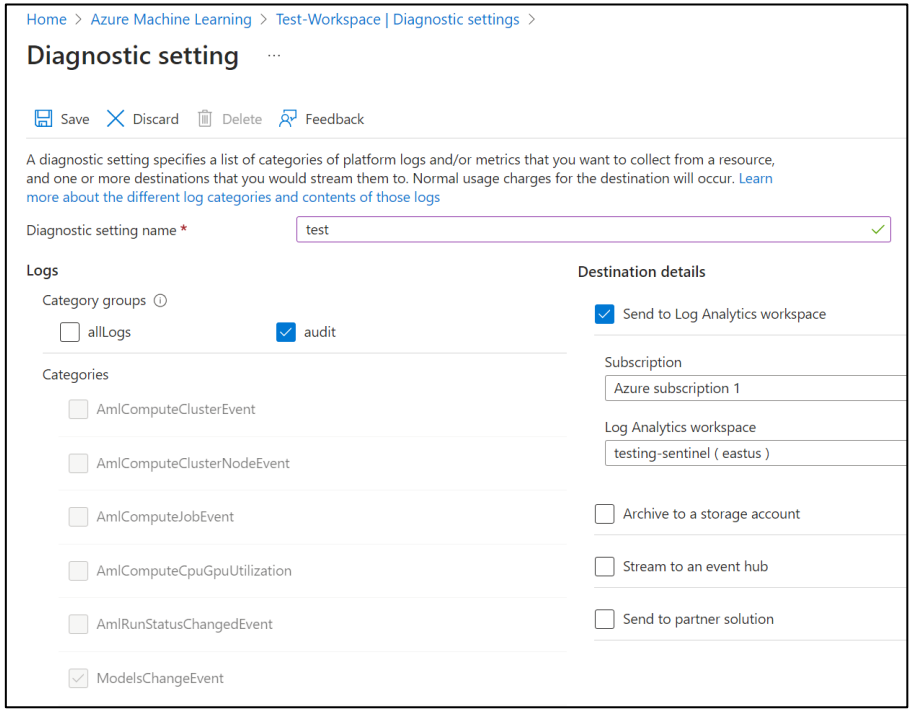
```
protoPayload.methodName = "storage.objects.get" AND (
SEARCH(protoPayload.resourceName, ".pb") OR
SEARCH(protoPayload.resourceName, ".mlmodel") OR
SEARCH(protoPayload.resourceName, ".onnx") OR
SEARCH(protoPayload.resourceName, ".pkl") OR
SEARCH(protoPayload.resourceName, ".h5") OR
SEARCH(protoPayload.resourceName, ".pmml")
)
```



Configuration Guidance



Azure ML



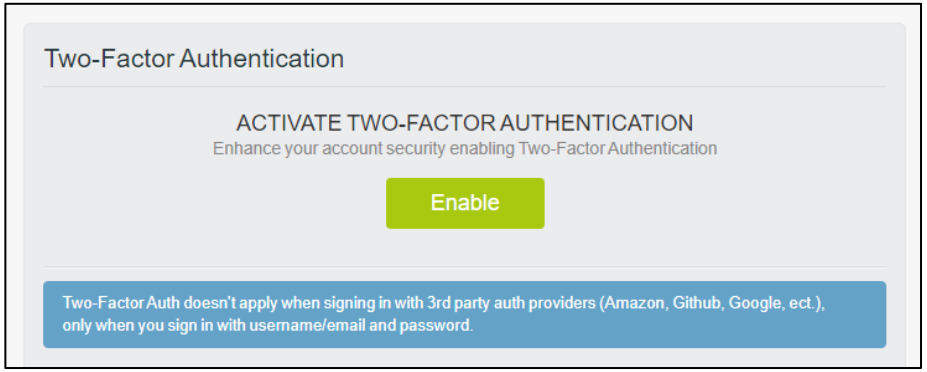
Asset Inventory
Manage ML asset inventory

Logging
Enable audit logs and build detections from this research

Threat Modeling
Include ML solutions in threat modeling exercises

Data & Pipeline Security
Data and ML workflow security best practices

BigML

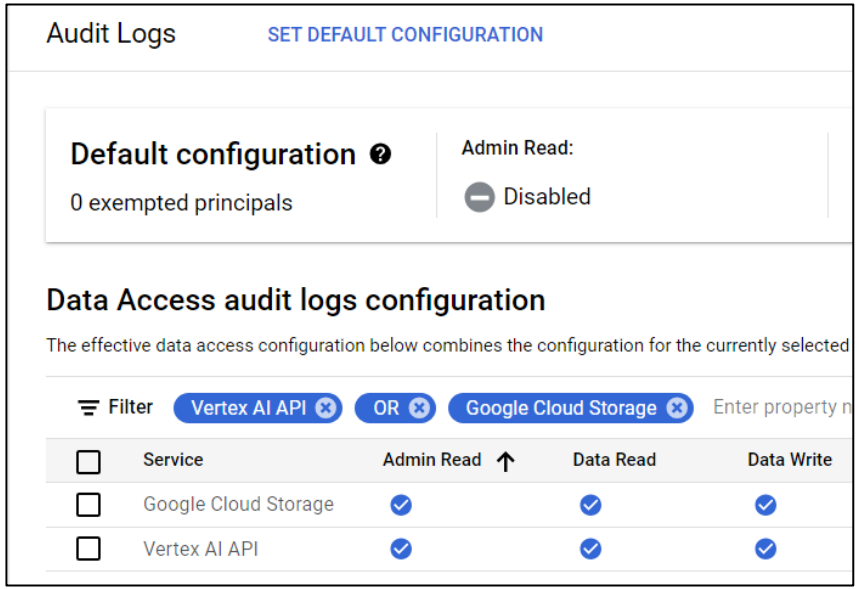


MFA
Enable MFA

Credential Rotation
Regularly change passwords and API keys

Granular Access
Apply granular access controls via alternative keys

Vertex AI



Access Control
Principal of least privilege

IAM
Use IAM to manage roles and groups

Logging
Enable audit logs and build detections from this research

Security Command Center
Additional protection for Vertex AI

Detection Guidance

Azure ML

Attack Scenario	Log Schema(s)	Operation Name(s)
Data Poisoning	**AmlDataSetEvent**	**WORKSPACES/DATASETS/REGISTERED/WRITE**
Dataset Recon	**AmlDataSetEvent**	**WORKSPACES/DATASETS/REGISTERED/READ**
Model Recon	**AmlModelsEvent**	**WORKSPACES/MODELS/READ**
Model Extraction	**AmlModelsEvent** **StorageBlobLogs**	**WORKSPACES/MODELS/READ** **GetBlob**
Data Extraction	**AmlDataSetEvent** **AmlDataStoreEvent**	**WORKSPACES/DATASETS/REGISTERED/READ** **WORKSPACES/DATASTORES/READ**

BigML

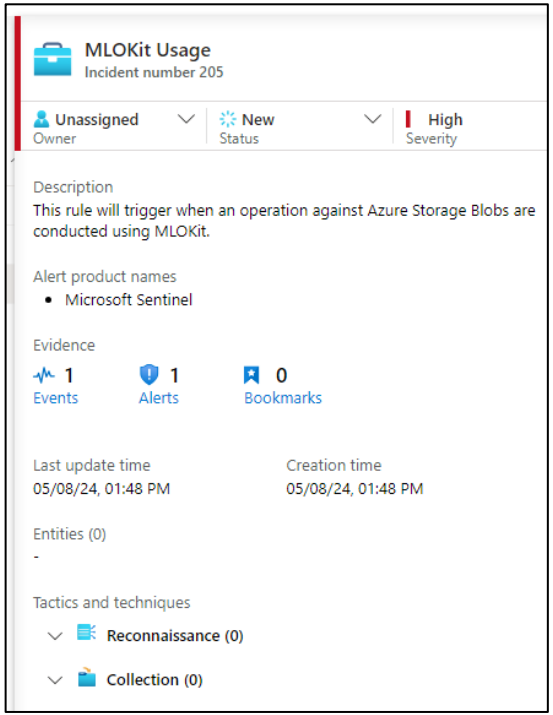
- Need private deployment to enable logging and develop detection rules

- We did not have access to private deployment for this research

Vertex AI

Attack Scenario	Method Name(s)
Data Poisoning	**google.cloud.aiplatform.ui.DatasetService.UpdateDataset** **google.cloud.aiplatform.v1.DatasetService.UpdateDataset**
Dataset Recon	**google.cloud.aiplatform.v1.DatasetService.ListDatasets**
Model Recon	**google.cloud.aiplatform.v1.ModelService.ListModels**
Model Extraction	**storage.objects.get** **google.cloud.aiplatform.ui.ModelService.ExportModel** **google.cloud.aiplatform.v1.ModelService.ExportModel**
Data Extraction	**storage.objects.get**

MLOKit



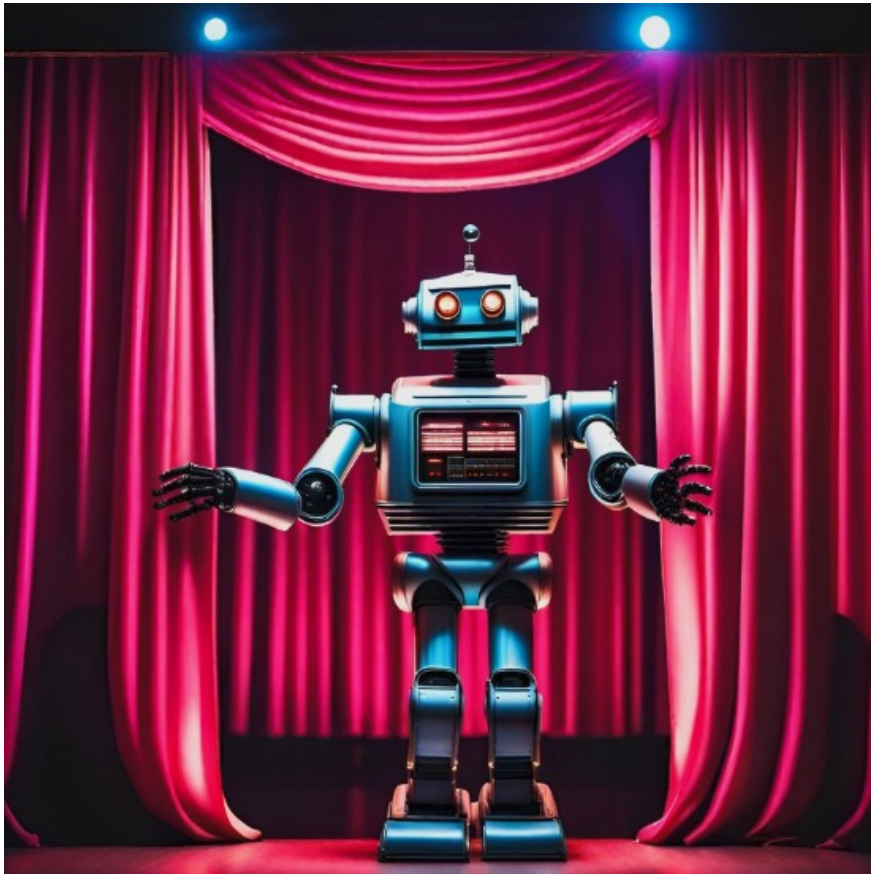
YARA Rule
C# Project GUID

Snort Rule
Hardcoded user agent string

Google Cloud Alert Policy
Actions against Google Cloud Storage

Sentinel Rules
Actions against Azure Storage Blobs

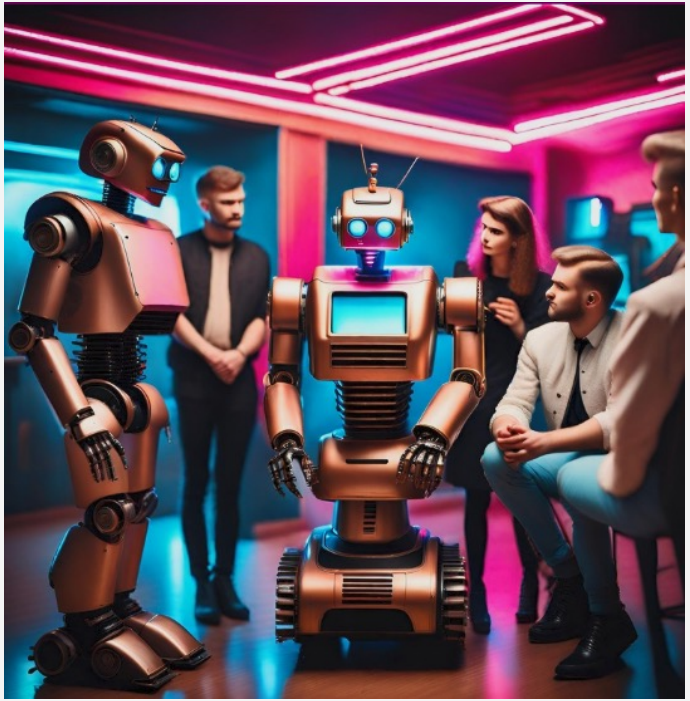
Conclusion



Conclusion

01	02	03
Securing MLOps platforms and personnel is critical	Simulate attacks against MLOps platforms and build detections	Use of MLOps platforms will continue to grow as will attackers targeting them

Questions?



Brett Hawkins
@h4wkst3r

Chris Thompson
@retBandit

Whitepaper -

MLOKit Tool -

Announcing the First Offensive AI Conference

Offensive AI Con is a cutting-edge, highly technical conference dedicated to exploring the use of AI within agentic frameworks for Cyber Network Attack (CNA), Cyber Network Exploitation (CNE), and Cyber Network Operations (CNO). The event will also delve into AI-assisted & autonomous vulnerability research and the development of offensive cyber capabilities, offering a platform for advanced discussion on the future of AI-driven cyber offense.

The topics will include:

- Agentic frameworks for offensive use
- Offensive model eval
- ML-assisted reversing, fuzzing, vulnerability research, and analysis
- ML-assisted weaponization & exploit development
- ML-based cyber effects development
- ML-assisted Offensive Security Tool (OST) & capability development
- ML-assisted target identification, analysis, tasking & targeting
- AI-based security testing, red teaming and software evaluation

Thank you

References

- https://arxiv.org/pdf/1609.02943.pdf%20/
- https://twitter.com/WHITEHACKSEC
- https://twitter.com/mairebear
- https://www.blackhat.com/asia-24/briefings/schedule/index.html#confused-learning-supply-chain-attacks-through-machine-learning-models-37794
- https://www.trendmicro.com/vinfo/gb/security/news/cybercrime-and-digital-threats/uncovering-silent-threats-in-azure-machine-learning-service-part-I
- https://msrc.microsoft.com/update-guide/vulnerability/CVE-2023-23382
- https://www.blackhat.com/us-23/briefings/schedule/#uncovering-azures-silent-threats-a-journey-into-cloud-vulnerabilities-33073
- https://www.oligo.security/blog/shadowray-attack-ai-workloads-actively-exploited-in-the-wild
- https://neptune.ai/blog/mlops-tools-platforms-landscape
- https://aws.amazon.com/sagemaker/
- https://azure.microsoft.com/en-us/products/machine-learning

References

- https://bigml.com/
- https://cloud.google.com/vertex-ai?hl=en
- https://www.qwak.com/
- https://www.dominodatalab.com/product/domino-enterprise-mlops-platform
- https://www.databricks.com/
- https://www.datarobot.com/platform/mlops/
- https://wandb.ai/site
- https://valohai.com/product/
- http://www.truefoundry.com/
- https://www.darkreading.com/application-security/hugging-face-ai-platform-100-malicious-code-execution-models
- https://huggingface.co/
- https://hiddenlayer.com/research/weaponizing-machine-learning-models-with-ransomware/#Overview-of-ML-Model-Serialization-Formats

References

- https://docs.python.org/3/library/pickle.html
- https://learn.microsoft.com/en-us/azure/machine-learning/concept-workspace?view=azureml-api-2
- https://aadinternals.com/post/phishing/
- https://learn.microsoft.com/en-us/azure/machine-learning/azure-machine-learning-glossary?view=azureml-api-2
- https://learn.microsoft.com/en-us/azure/machine-learning/how-to-setup-authentication?view=azureml-api-2&tabs=sdk
- https://learn.microsoft.com/en-us/cli/azure/
- https://learn.microsoft.com/en-us/cli/azure/ml?view=azure-cli-latest
- https://learn.microsoft.com/en-us/rest/api/azureml/?view=rest-azureml-2023-10-01
- https://learn.microsoft.com/en-us/azure/machine-learning/how-to-assign-roles?view=azureml-api-2&tabs=labeler
- https://learn.microsoft.com/en-us/azure/ai-services/diagnostic-logging
- https://learn.microsoft.com/en-us/azure/azure-monitor/logs/log-analytics-workspace-overview
- https://learn.microsoft.com/en-us/cli/azure/ml?view=azure-cli-latest

References

- https://learn.microsoft.com/en-us/azure/machine-learning/monitor-azure-machine-learning-reference?view=azureml-api-2
- https://medium.com/@maziarizadi/pickle-your-model-in-python-2bbe7dba2bbb
- https://curl.se/docs/manpage.html
- https://learn.microsoft.com/en-us/rest/api/azureml/workspaces?view=rest-azureml-2023-10-01
- https://github.com/xforcered/MLOKit
- https://learn.microsoft.com/en-us/rest/api/azureml/model-containers?view=rest-azureml-2023-10-01
- https://learn.microsoft.com/en-us/azure/ai-services/document-intelligence/create-sas-tokens?view=doc-intel-4.0.0
- https://bigml.com/customers/
- https://bigml.com/api/organizations
- https://bigml.com/api/
- https://bigmler.readthedocs.io/en/latest/
- https://bigml.com/api/quickstart
- https://bigml.com/api/sources

References

- https://bigml.com/api/projects
- https://bigml.com/api/datasets
- https://bigml.com/api/models
- https://dmg.org/pmml/pmml-v4-1.html
- https://cloud.google.com/vertex-ai/docs/glossary
- https://cloud.google.com/vertex-ai/docs/workbench/reference/authentication
- https://cloud.google.com/docs/authentication/rest
- https://cloud.google.com/docs/authentication/rest#user-creds
- https://cloud.google.com/docs/authentication/rest#rest-request
- https://cloud.google.com/docs/authentication/rest#impersonated-sa
- https://cloud.google.com/docs/authentication/rest#metadata-server
- https://cloud.google.com/vertex-ai/docs/general/access-control
- https://cloud.google.com/vertex-ai/docs/general/audit-logging

References

- https://cloud.google.com/logging/docs/view/logs-explorer-interface
- https://cloud.google.com/storage/docs/buckets
- https://cloud.google.com/logging/docs/view/building-queries
- https://towardsdatascience.com/guide-to-file-formats-for-machine-learning-columnar-training-inferencing-and-the-feature-store-2e0c3d18d4f9
- https://cloud.google.com/apis/docs/overview
- https://cloud.google.com/resource-manager/reference/rest/v1/projects
- https://cloud.google.com/vertex-ai/docs/reference/rest/v1/projects.locations.datasets
- https://cloud.google.com/storage/docs/json_api
- https://cloud.google.com/vertex-ai/docs/reference/rest/v1/projects.locations.models
- https://learn.microsoft.com/en-us/azure/machine-learning/concept-enterprise-security?view=azureml-api-2
- https://learn.microsoft.com/en-us/security/benchmark/azure/baselines/machine-learning-service-security-baseline

References

- https://techcommunity.microsoft.com/t5/fasttrack-for-azure/six-security-considerations-for-machine-learning-solutions/ba-p/3718592
- https://blog.bigml.com/2013/05/03/alternative-keys-fine-grained-rest-api-access-to-your-machine-learning-resources/
- https://www.linkedin.com/pulse/secure-your-vertex-ai-workbench-enterprise-machine-learning-curtis-kbmof?trk=articles_directory
- https://cloud.google.com/blog/products/identity-security/introducing-security-command-center-protection-for-vertex-ai
- https://learn.microsoft.com/en-us/azure/data-explorer/kusto/query/
- https://learn.microsoft.com/en-us/azure/sentinel/detect-threats-custom?tabs=azure-portal
- https://learn.microsoft.com/en-us/azure/storage/blobs/monitor-blob-storage?tabs=azure-portal
- https://learn.microsoft.com/en-us/azure/azure-monitor/reference/tables/storagebloblogs
- https://cloud.google.com/monitoring?hl=en
- https://cloud.google.com/monitoring/alerts
- https://yara.readthedocs.io/en/stable/writingrules.html
- https://snort.org/