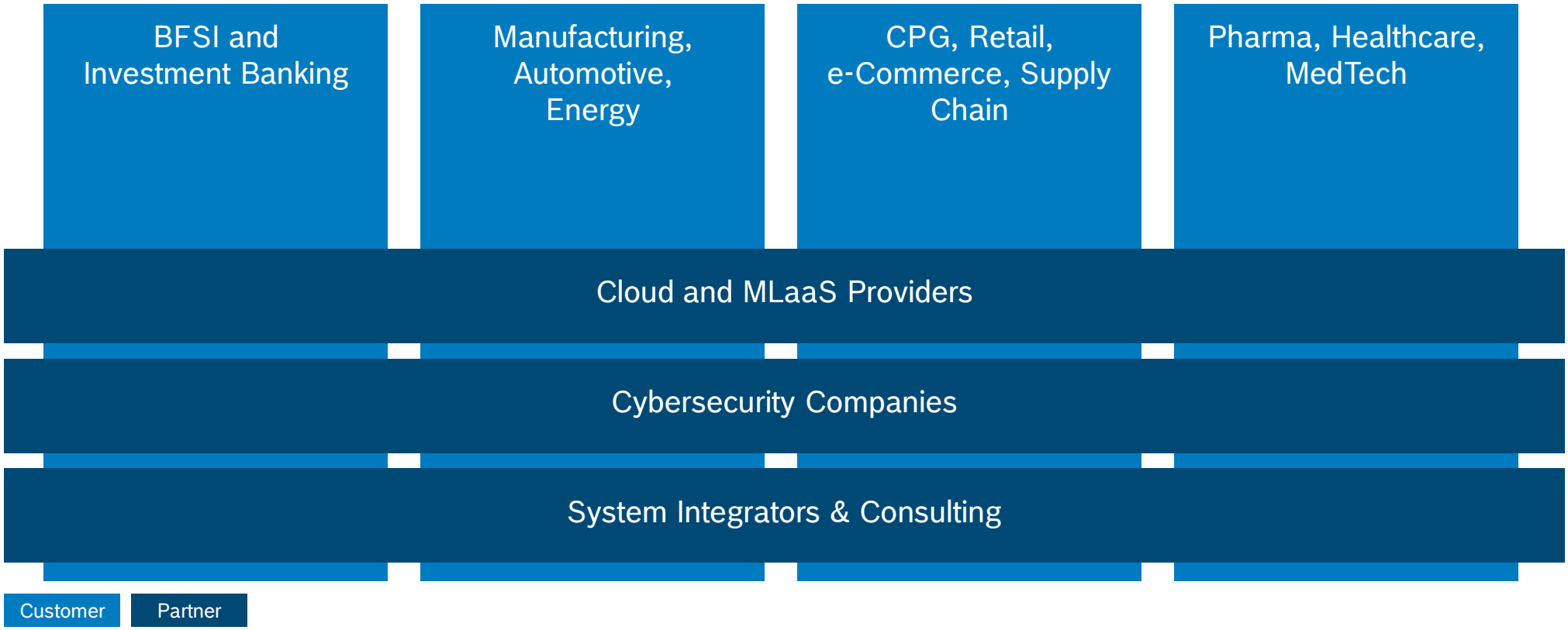
BOSCH AISHIELD CASE STUDIES

BOSCH

Bosch AIShield Case Studies

Bosch AIShield existing Customer & Partner Capabilities



Bosch AIShield Case Studies

Proving AI Threat Vulnerability of a Healthcare leader

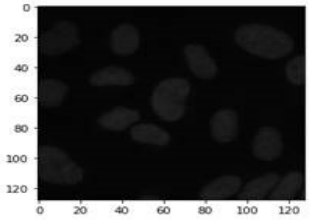
Algorithm – Automatic Detection Method for Cancer Cell Nucleus

Industry – Healthcare

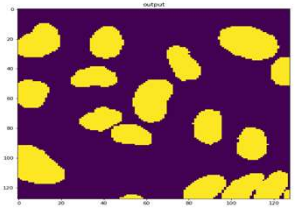
Customer – Multinational Science and Technology company

AI Threat Prevented – Model Extraction

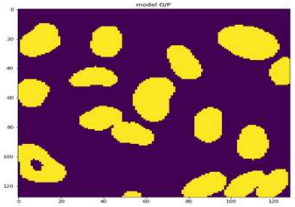
01 Original Input



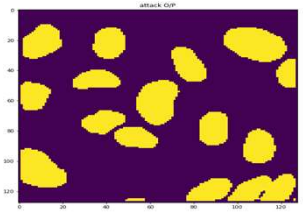
02 Human labeled Ground Truth



03 Original Model Output



04 Stolen Model Output



Background & Objective

The healthcare company was working on study that the mitotic defects and micronuclei in cancer cells can be used as biomarkers to evaluate the instability of the chromosomes. The study further aimed to detect cells with mitotic defects and micronuclei by applying an approach integrating the application of a convolutional neural network for normal cell identification and the proposed color layer signature analysis (CLSA) to spot cells with mitotic defects and micronuclei.

Solution Highlights

- AIShield team extracted the CNN algorithm developed over 6 months within 2 hours & only 8% delta to original model accuracy, as part of an ethical hacking case
- As part of threat mitigation AIShield team integrated a defense model with the original model which brought down the stolen model accuracy to only 10%

Result

Demonstrated the prevention of loss of intellectual property & revenue for the healthcare company with **impact value of USD 5 million.**

Bosch AIShield Case Studies

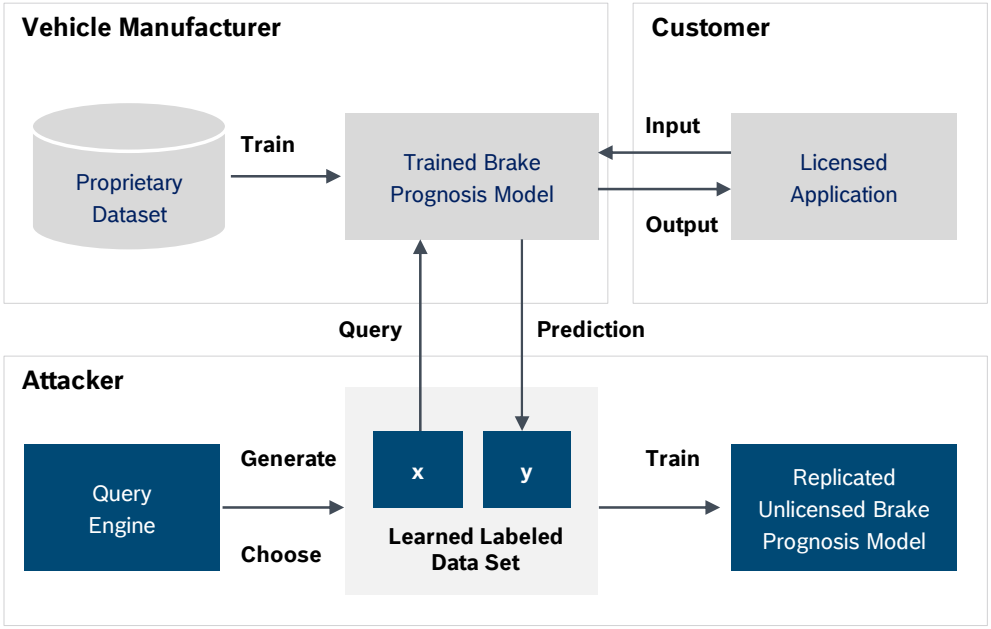
Preventing Unlicensed Use of Proprietary AI/ML Model

Algorithm – Brake Prognosis Model

Industry – Automotive (Customer) | Cloud Security (Partner)

Customer – APAC based Vehicle Manufacturer

AI Threat Prevented – Software Stealing (AI algorithm)

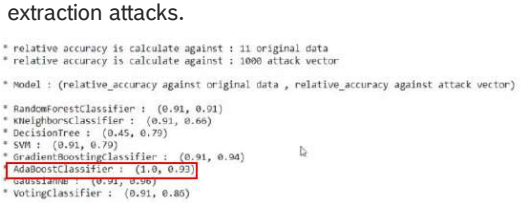


Background & Objective

A cloud security company needed a 'cybersecurity for AI' niche solution partner for one of their automotive customers who had foreseen unlicensed usage of their proprietary AI/ML algorithm.

Solution Highlights

Bosch AIShield team provided AI security threat vulnerability assessment & mitigation for the automotive vehicle manufacturer. The break prognosis model was found vulnerable to model extraction attacks.



Result

Plugged the possible revenue leakage for the vehicle manufacturer to the tune of **USD 20 million per year.**

Bosch AIShield Case Studies

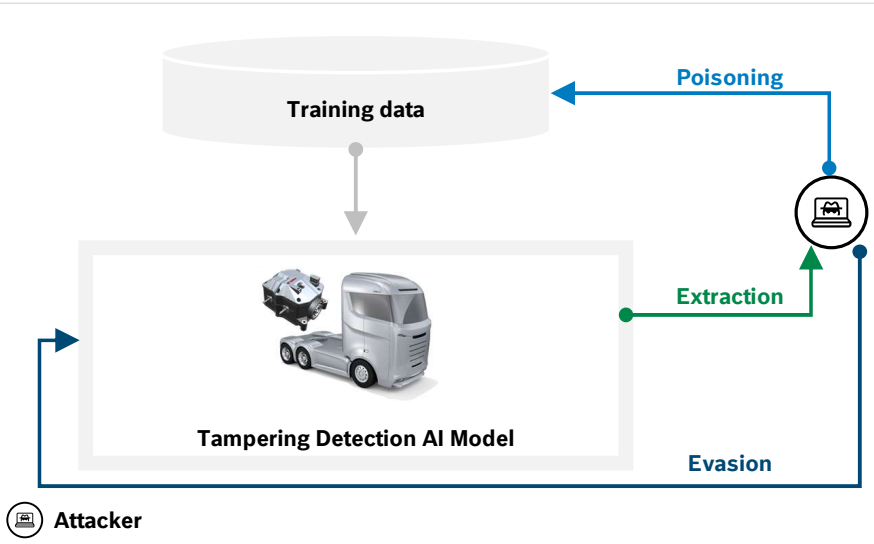
AI Security for Better Sustainability

Algorithm - Tampering Detection

Industry - Automotive

Customer – Europe headquartered multinational engineering and technology company

AI Threat Prevented – Model Extraction, Evasion & Poisoning



AI Security Threats for Tampering Detection AI Model

Background & Objective

- In diesel engines, exhaust gas treatment is needed to lower the produced NOx concentration which is part of the emissions
- Selective Catalytic Reduction (SCR) can reduce NOx by up to 98%. As part of SCR, Diesel Exhaust Fluid (DEF) is required
- DEF can be costly to a rough estimation: $10-12 for a 700-mile trip. Transport companies attempt to eliminate DEF usage by installing illegal Tampering Dongles
- An embedded neural network for tampering detection is therefore deployed on the ECU / DCU
- Extraction of this algorithm by hackers can enable malpractitioners to evade or poison the algorithms for their own benefits resulting in unsustainable environment

Solution Highlights

- Identified & demonstrated AI security threats to tampering detection algorithm by performing threat & risk analysis for AI model and system
- Developed a defense model engineered for AI/ML model and Integrated it the target ECU environment
- Complemented existing rule based / connectivity function for tampering detection

Result

- Overall hardened security for greater compliance & sustainability
- Also, prevented loss of customers' trust & business for the automotive company

AI Security for Better IP Protection

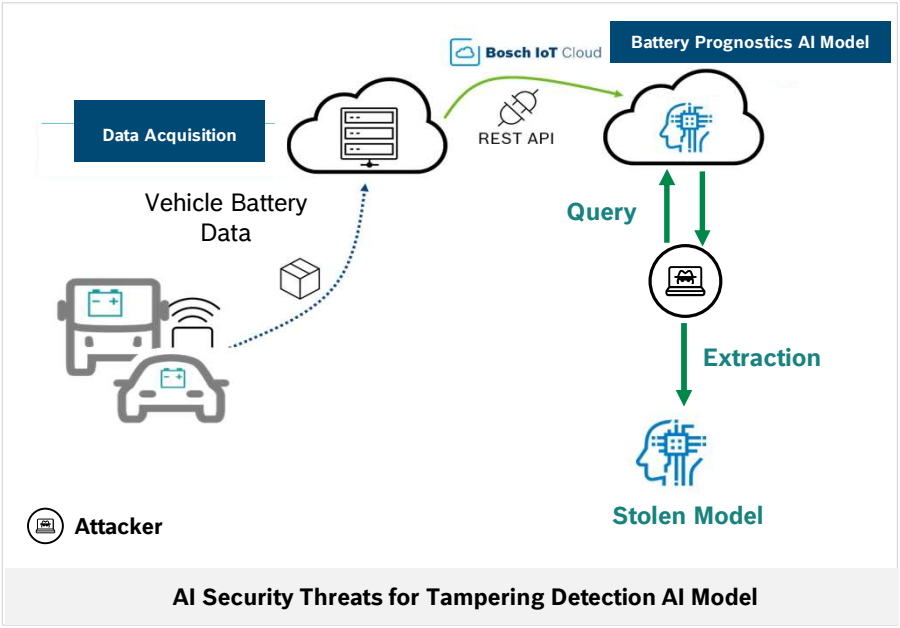
AIShield

Algorithm – Battery in the Cloud(BitC)

Industry – Automotive (Predictive Maintenance)

Customer – Europe headquartered multinational engineering and technology company

AI Threat Prevented – Model Extraction



AI Security Threats for Tampering Detection AI Model

Background & Objective

- Battery in the cloud is a bundle of services for OEMs that monitors, predicts, and optimizes the performance and lifetime of batteries in electric vehicles. It is the basis for a usage certificate, which gives transparent and manipulation-proof information on the battery condition and enables new business models for finance, leasing, and insurance products and services.
- The integrated and modular solutions for EV fleets, charging, and energy management help fleet operators to reduce their total cost of ownership in terms of operations, energy consumption, and vehicle depreciation. Battery range and lifetime are optimized.
- AI/ML models is used to predict lifetime and optimize battery lifetime, optimize the parameters for charging, proactive battery anomalies detection.

Solution Highlights

- Identified & demonstrated AI security threats to BitC algorithm by performing threat & risk analysis for AI/ML models and system
- Developed a defense model engineered for AI/ML model and Integrated it the target environment

Result

- Prevented loss of IP & Revenue
- Enabled Business Growth

BOSCH

Bosch AIShield Case Studies

Protecting Smart Manufacturer's Intellectual Property

Algorithm - Automated Visual Inspection

Industry – Manufacturing / Industry 4.0

Customer – Europe based Automotive parts manufacturer

AI Threat Prevented – Model Extraction, Evasion & Poisoning

System

Design a System. which can handle a variety of complex products

Image capture module Capture

Design a general-purpose image capture module for detection / product inspection for different kind of defects

Algorithms and frameworks

Provide algorithms and frameworks so that users can train defect detectors

Background & Objective

- For the manufacturer, product quality control is essential in order to fulfill the highest quality demands of customers. Automated production lines accomplish this in part with highly customized optical inspection systems. Systems comprised cameras & automated inspection algorithm to capture images and detecting defects on products
- Extraction of this highly adaptable, task-flexible and reconfigurable algorithm can enable malpractitioners to steal the intellectual property of the smart manufacturer who invested considerable time and effort in building and training the high accuracy algorithm

Solution Highlights

- Identified & demonstrated AI security threats to automated visual inspection algorithm by performing threat & risk analysis for AI model and system
- Developed a defense model engineered for AI/ML model and Integrated it the target smart factory environment

Result

Prevented loss of IP of an automated visual inspection algorithm with impact value of **USD 5 million.**

Bosch AIShield Case Studies

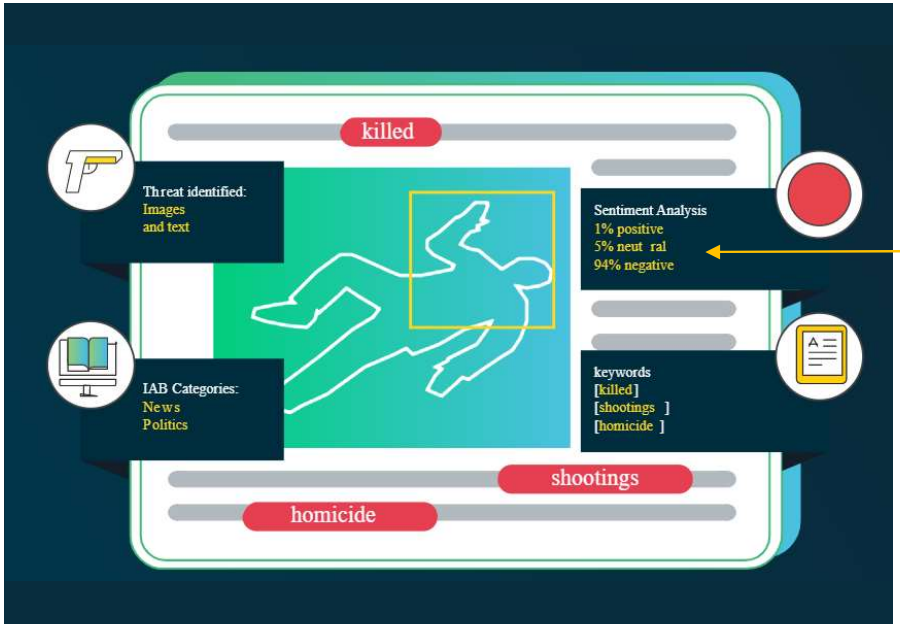
Prevention of Legal Claims against Published Advertisements

Algorithm – Sentiment analysis based on Contextual Intelligence

Industry – Consumer Goods, Retial, e-Commerce

Customer – British multinational consumer goods company

AI Threat Prevented – Model Evasion



Background & Objective

For a sentiment analysis algorithm for contextual intelligence, hackers could adversely craft data (images annotated as 'killed' + 'martyred'), in order to get a desired outcome at the inference time (image now showing different sentiment analysis). This could bring the associated company unintended legal claims by advertising regulators & customers.

Sentiment Analysis Wrongly Inferred	Sentiment Analysis Intended
▶ 30% positive	▶ 00% positive
▶ 50% neutral	▶ 10% neutral
▶ 20% negative	▶ 90% negative

Solution Highlights

Threat analysis of the sentiment analysis algorithm by AIShield team and adding some (chosen/crafted) adversarial examples to the training dataset as part of the generated defense model.

Result

- Prevented brand safety value proposition
- Compliant to IAB (Interactive Advertising Bureau) standards
- Overall reduced legal advertising claims

Bosch AIShield Case Studies

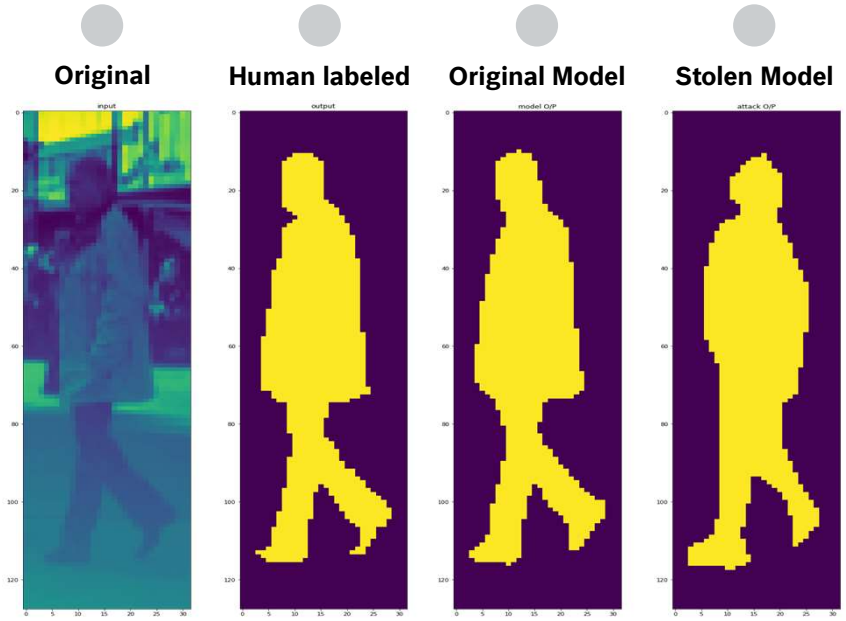
Proving AI Threat Vulnerabiltiy of Computer Vision systems

Algorithm – Pedestrian Detection

Industry – Automotive (Autonomous Driving)

Customer – British multinational Consumer Goods company

AI Threat Prevented – Model Extraction



Background & Objective

Recover a functionally equivalent model by iteratively querying the model. This allows an attacker to examine the offline copy of the model and re-use the same without license or launch further attacks.

Solution Highlights

- AIShield team extracted a pedestrian detection algorithm developed over 10 months and significant investments, within 4 hours & only 4% delta to original model accuracy, as part of an ethical hacking case
- As part of threat mitigation AIShield team integrated a defense model with the original model which brought down the stolen model accuracy to only 15%

Result

Demonstrated the prevention of loss intellectual property & revenue for a computer vision company with impact value of **USD 10 million.**

Bosch AIShield Case Studies

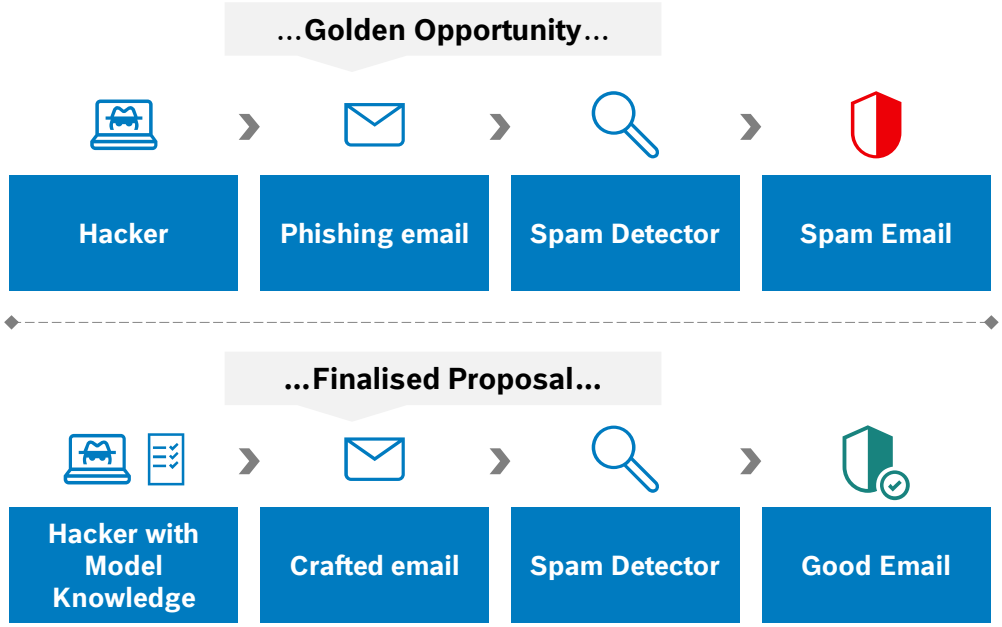
AI Risk Assessment & Mitigation for a Banking company

Algorithm 1 – Spam Email Detection
Algorithm 2 – Conversational AI application

Industry – BFSI (Banking, Financial Services & Insurance)

Customer – American multinational investment bank and financial services holding company

AI Threat Prevented – Model Evasion & Data Poisoning



Background & Objective

The banking organization company desired an overall risk assessment of its ML-powered applications running in various cloud environments including particular interest in ML model evasion & data poisoning.

Solution Highlights

- Assessed the overall risk associated with each ML-powered application using Bosch AIShield's AI risk Assessment framework and methodology.
- Iteratively queried ML models running on the cloud through (AIShield's proprietary attack vectors to extract the ML models)
- Evaded email protection system by first building a copy-cat email protection ML model, and using the insights to evade the live system
- Exploited the feedback loop of conversational AI. By repeated interactions using racist and offensive language, biased the dataset towards that language to generate reprehensible material

Result

Prevented Loss of AI Integrity & Brand Reputation

- Provided a risk score for each model for spam detector & conversational AI
- Recommend set of security controls and defense mechanism integration

Bosch AIShield Case Studies

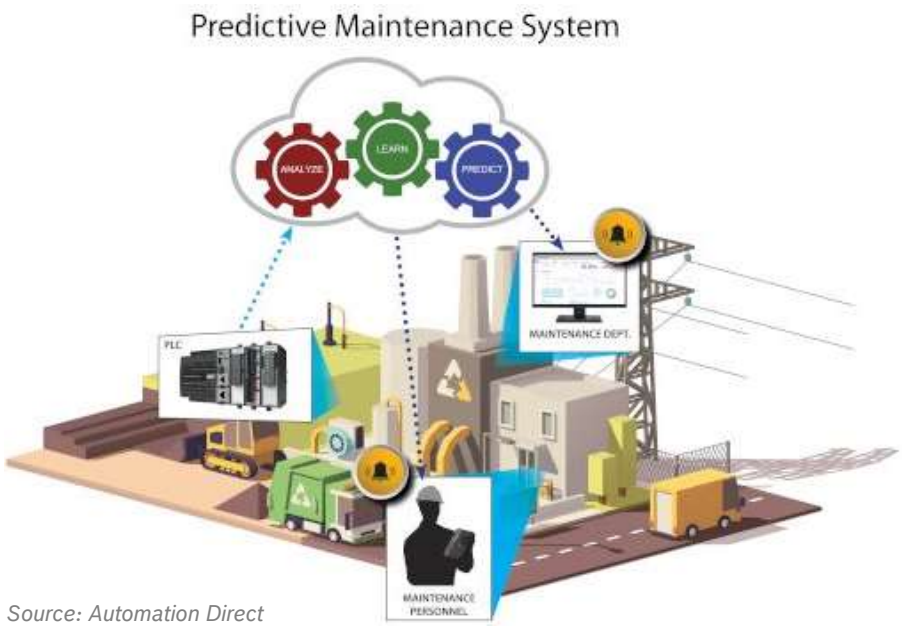
AIoT Risk Assessment for Cloud based IIoT Technologies

Algorithm – Predictive Maintenance

Industry – Manufacturing / Industry 4.0

Customer – German multinational automotive manufacturer

AI Threat Prevented – Model Extraction & Data Poisoning



Source: Automation Direct

Background & Objective

- The manufacturer was concerned about new IIoT technologies introducing new types of cyber threats onto the plant floor, including threats to cloud hosted AI applications
- Manufacturer believed that IP theft has been a primary motive for the cyber-attacks on their factories. State-sponsored cyber-attacks had also reached a critical mass
- Still in the planning & development stage, plant owners wanted to assess risks to their proprietary predictive maintenance solutions

Solution Highlights

- Assessed the overall risk associated with each ML-powered application using Bosch AIShield's AI risk Assessment framework and methodology.
- Iteratively queried ML models running on the cloud through AIShield's proprietary attack vectors to extract the ML models
- Demonstrated how the Predictive Maintenance solution could be extracted from the cloud within hours.
- Suggested a defense mechanism for predictive AI application as part of holistic cybersecurity strategy

Result

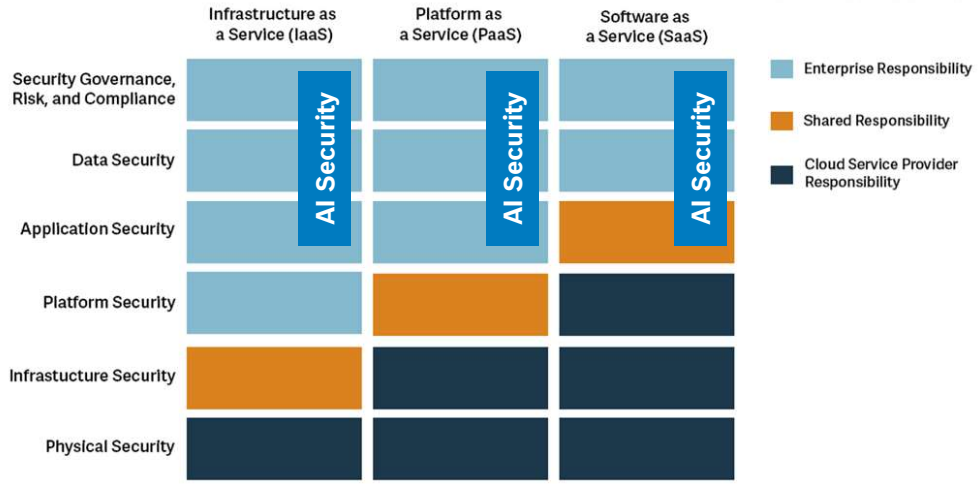
Prevented AI/ML model and associated data related breach of **USD 3 million.**

Use Cases

Bosch AIShield Case Studies

References used for explaining the need and attack scenarios on AI

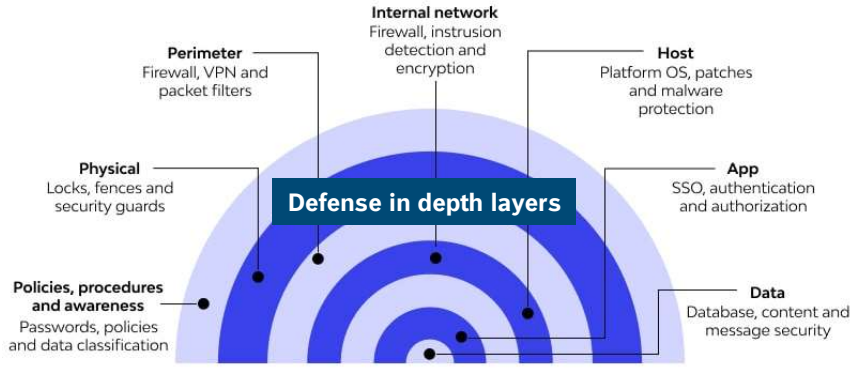
We understand and recommend that an enterprise organization is responsible for AI Security as part of Application Security, Data Security and Security GRC across cloud models

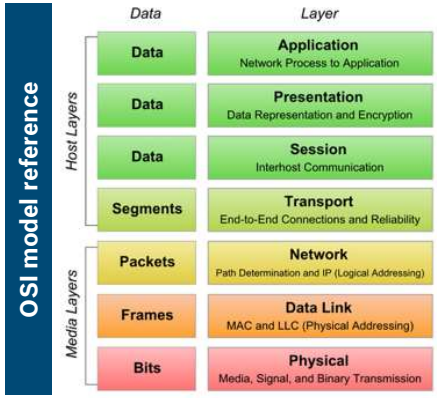


Picture Reference: https://www.isc2.org/Articles/Responsibility-and-Accountability-in-the-Cloud#

More details – AWS - https://aws.amazon.com/compliance/shared-responsibility-model/

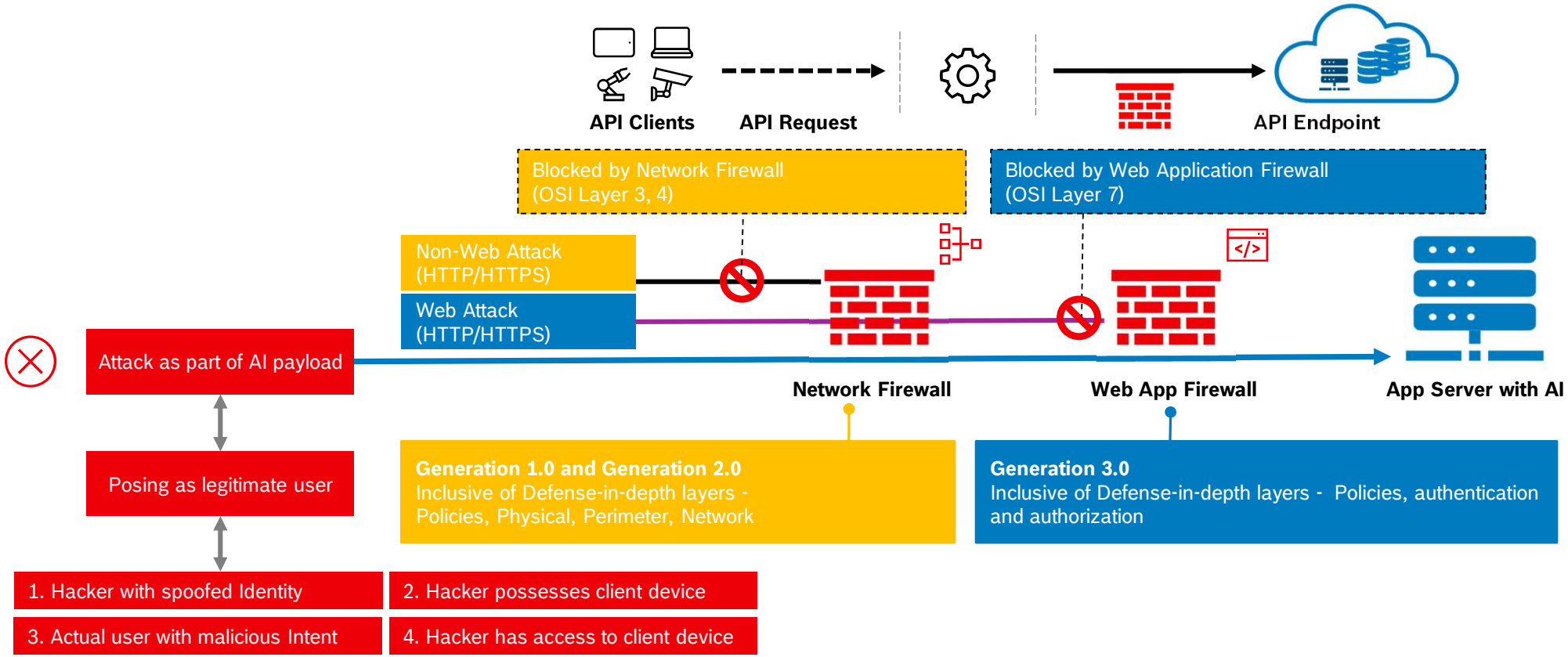
Azure https://azure.microsoft.com/mediahandler/files/resourcefiles/shared-responsibility-for-cloud-omputing/Shared%20Responsibility%20for%20Cloud%20Computing-2019-10-25.pdf





Bosch AIShield Case Studies

Why attacks on AI are successful despite firewalls?



Bosch AIShield Case Studies

Preventive Maintenance – attacks & kill-chain

SUMMARY

Model Extraction attack

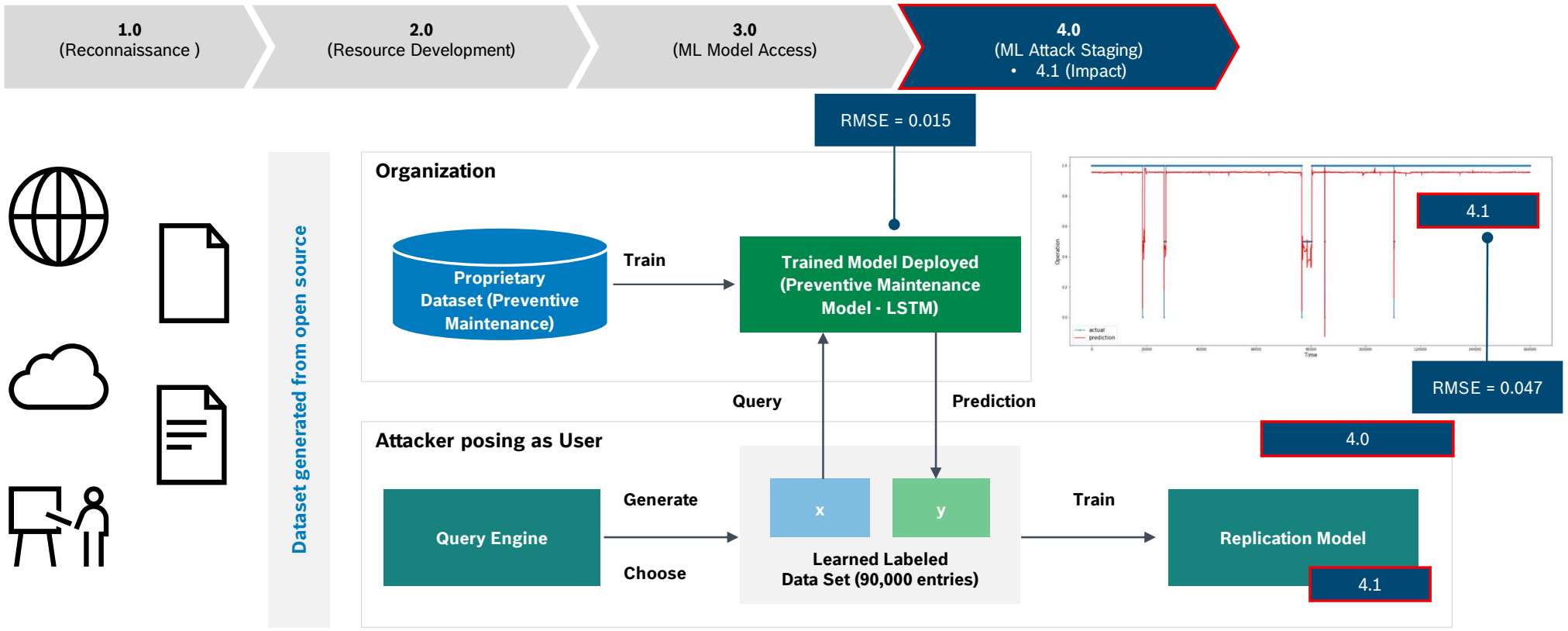
- Preventive maintenance services provide public-facing UIs and APIs. A hacker can utilize these public endpoints to create a replicated model with near-production, state-of-the-art preventive maintenance system.
- The IP can be stolen and deployed for similar services by adversaries also resulting in financial losses.

1.0 (Reconnaissance) → **2.0** (Resource Development) → **3.0** (ML Model Access) → **4.0** (ML Attack Staging)
- 4.1 (Impact)

#	Kill-chain Phase	Technique	Description
1.0	Reconnaissance (The adversary is trying to gather information they can use to plan future operations.)	Search for Victim's Publicly available Research Material	The hackers used published information on web to identify the datasets by the target preventive maintenance services.
2.0	Resource Development The adversary is trying to establish resources they can use to support operations	Acquire Public ML Artifacts - Datasets	Hacker gathered similar datasets that the preventive maintenance used
3.0	ML Model Access (An adversary is attempting to gain some level of access to a machine learning model)	ML Model Inference Access	They abuse a public facing application to query the model posing as legitimate user and produce preventive maintenance pairs as training data.
4.0	ML Attack Staging (An adversary is leveraging their knowledge of and access to the target system to tailor the attack.)	Create Proxy ML Model: Model Extraction	Using these recommendation pairs, the hacker trained a model that replicates the behavior of the target model.
4.1	Impact (The adversary is trying to manipulate, interrupt, erode confidence in, or destroy your systems and data.)	ML Intellectual Property Theft	By replicating the model with high fidelity, the hacker could steal a model and violate the victim's intellectual property rights.

Bosch AIShield Case Studies

Preventive Maintenance– Model Extraction



Bosch AIShield Case Studies

Perimeter Security System with Camera (Breach Detection)- attacks & kill-chain

SUMMARY

Model Extraction attack

Perimeter security camera-based systems are used. A hacker can utilize these publicly available system to create a replicated model with near-production, state-of-the-art sentiment analysis quality.

Model Evasion attack

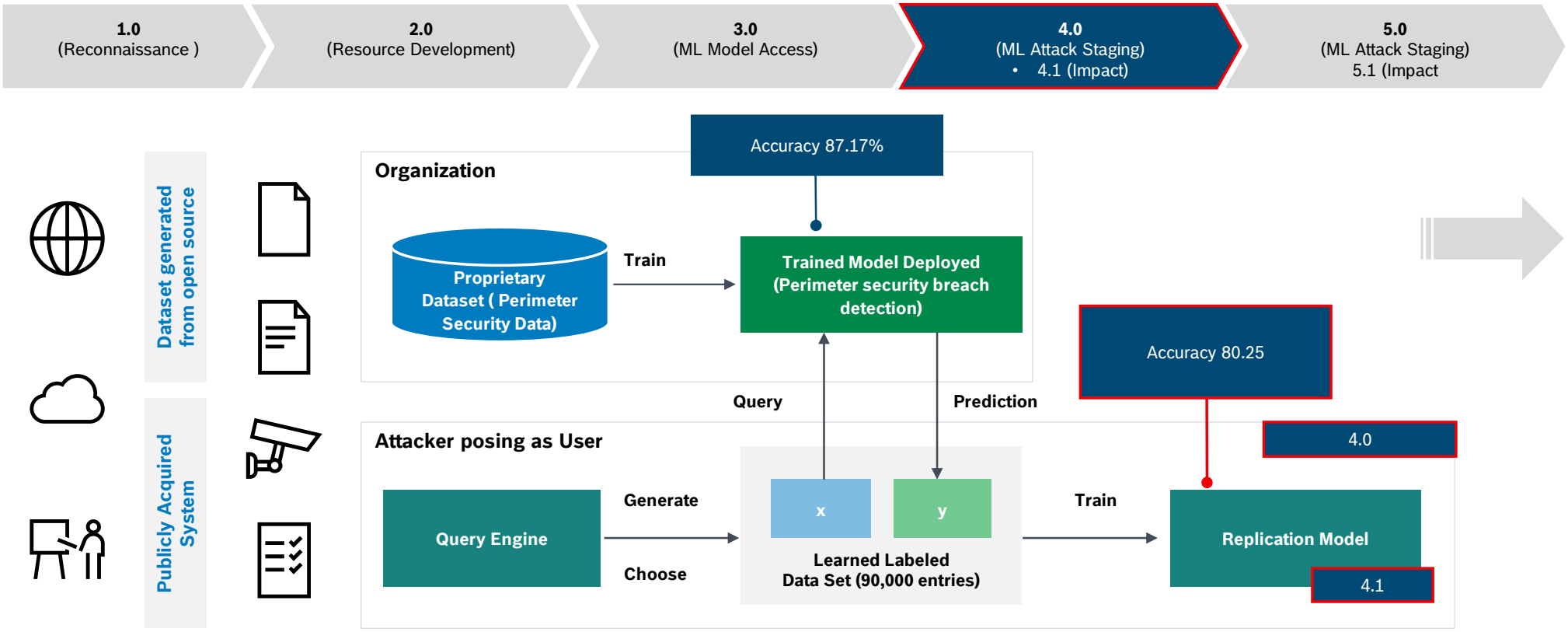
Beyond stealing the IP from a black-box system, they can the replicated model to successfully transfer adversarial examples to the real production services. These adversarial inputs successfully cause the evasion of perimeter security system.

1.0 (Reconnaissance) → **2.0** (Resource Development) → **3.0** (ML Model Access) → **4.0** (ML Attack Staging) • 4.1 (Impact) → **5.0** (ML Attack Staging) 5.1 (Impact

#	Kill-chain Phase	Technique	Description
1.0	Reconnaissance (The adversary is trying to gather information they can use to plan future operations.)	Search for Victim's Publicly available Research Material	The hackers used published information on web to identify the datasets and systems by the perimeter security system with camera.
2.0	Resource Development The adversary is trying to establish resources they can use to support operations	Acquire Public ML Artifacts - Datasets	Hacker gathered similar datasets and acquires the system that the perimeter security system used
3.0	ML Model Access (An adversary is attempting to gain some level of access to a machine learning model)	ML Model Inference Access	They abuse publicly acquired system to query the model posing as legitimate user and produce perimeter security system pairs as training data.
4.0	ML Attack Staging (An adversary is leveraging their knowledge of and access to the target system to tailor the attack.)	Create Proxy ML Model: Model Extraction	Using these perimeter security system pairs, the hackers trained a model that replicates the behavior of the target model.
4.1	Impact (The adversary is trying to manipulate, interrupt, erode confidence in, or destroy your systems and data.)	ML Intellectual Property Theft	By replicating the model with high fidelity, the hackers could steal a model and violate the victim's intellectual property rights.
5.0	ML Attack Staging (An adversary is leveraging their knowledge of and access to the target system to tailor the attack.)	Craft Adversarial Data	The replicated models were used to generate adversarial examples that successfully transferred to original perimeter security system service.
5.1	Impact (The adversary is trying to manipulate, interrupt, erode confidence in, or destroy your systems and data.)	Evade ML Model	The adversarial examples were used to evade the perimeter security services.

Bosch AIShield Case Studies

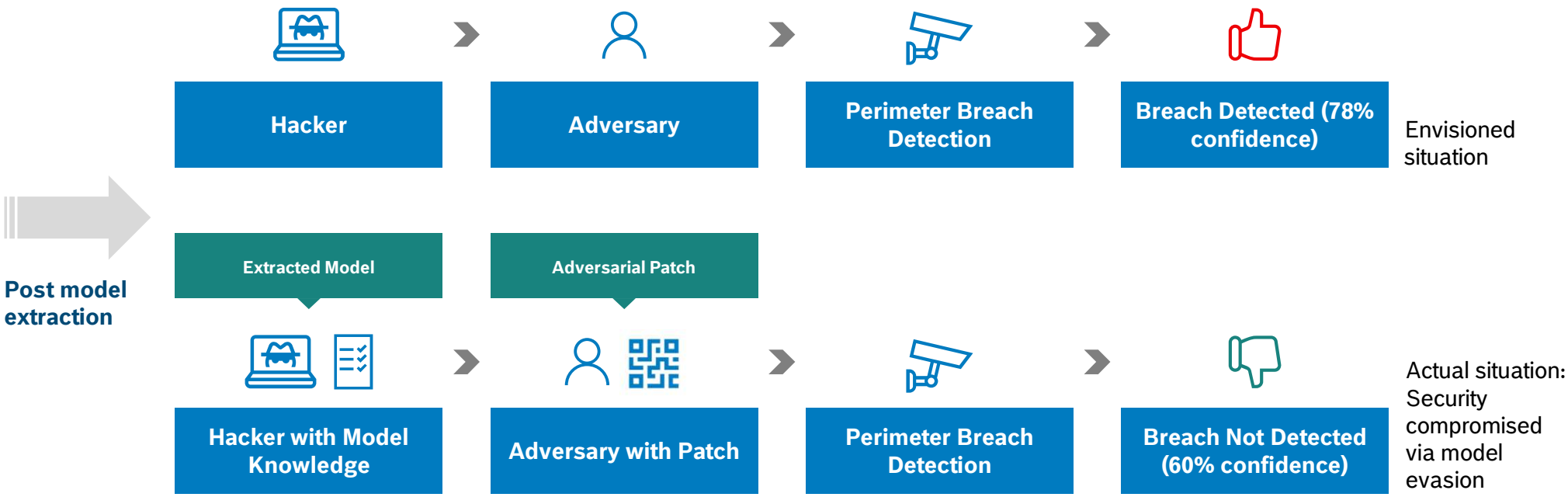
Perimeter Security Breach Detection- Model Extraction



Bosch AIShield Case Studies

Perimeter security Breach Detection - Model Evasion

Scenario - An enterprise is using 3rd party services for perimeter breach detection. Hacker acquires model knowledge by posing as legitimate user for 3rd party. This is a typical case for strict security assessment of specific AI algorithm.



Bosch AIShield Case Studies

Sentiment Analysis - attacks & kill-chain

SUMMARY

Model Extraction attack

Sentiment Analysis services provide public-facing UIs and APIs. A hacker can utilize these public endpoints to create a replicated model with near-production, state-of-the-art sentiment analysis quality.

Model Evasion attack

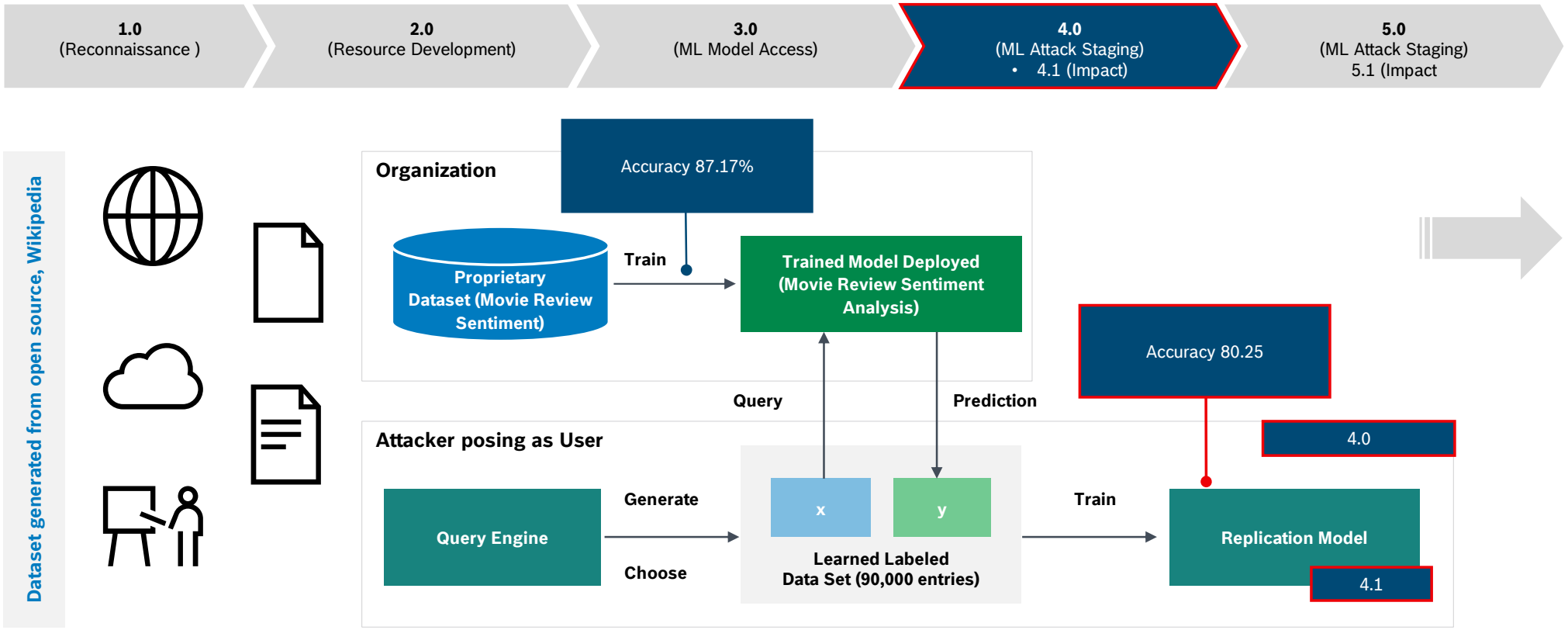
Beyond stealing the IP from a black-box system, they can the replicated model to successfully transfer adversarial examples to the real production services. These adversarial inputs successfully cause targeted sentiment flips.

1.0 (Reconnaissance) → **2.0** (Resource Development) → **3.0** (ML Model Access) → **4.0** (ML Attack Staging) • 4.1 (Impact) → **5.0** (ML Attack Staging) 5.1 (Impact

#	Kill-chain Phase	Technique	Description
1.0	Reconnaissance (The adversary is trying to gather information they can use to plan future operations.)	Search for Victim's Publicly available Research Material	The hackers used published information on web to identify the datasets by the target sentiment analysis services.
2.0	Resource Development The adversary is trying to establish resources they can use to support operations	Acquire Public ML Artifacts - Datasets	Hacker gathered similar datasets that the sentiment analysis service used
3.0	ML Model Access (An adversary is attempting to gain some level of access to a machine learning model)	ML Model Inference Access	They abuse a public facing application to query the model posing as legitimate user and produce sentiment analysis pairs as training data.
4.0	ML Attack Staging (An adversary is leveraging their knowledge of and access to the target system to tailor the attack.)	Create Proxy ML Model: Model Extraction	Using these sentiment analysis pairs, the researchers trained a model that replicates the behavior of the target model.
4.1	Impact (The adversary is trying to manipulate, interrupt, erode confidence in, or destroy your systems and data.)	ML Intellectual Property Theft	By replicating the model with high fidelity, the researchers demonstrated that an adversary could steal a model and violate the victim's intellectual property rights.
5.0	ML Attack Staging (An adversary is leveraging their knowledge of and access to the target system to tailor the attack.)	Craft Adversarial Data	The replicated models were used to generate adversarial examples that successfully transferred to sentiment analysis service.
5.1	Impact (The adversary is trying to manipulate, interrupt, erode confidence in, or destroy your systems and data.)	Evade ML Model	The adversarial examples were used to evade the sentiment analysis service

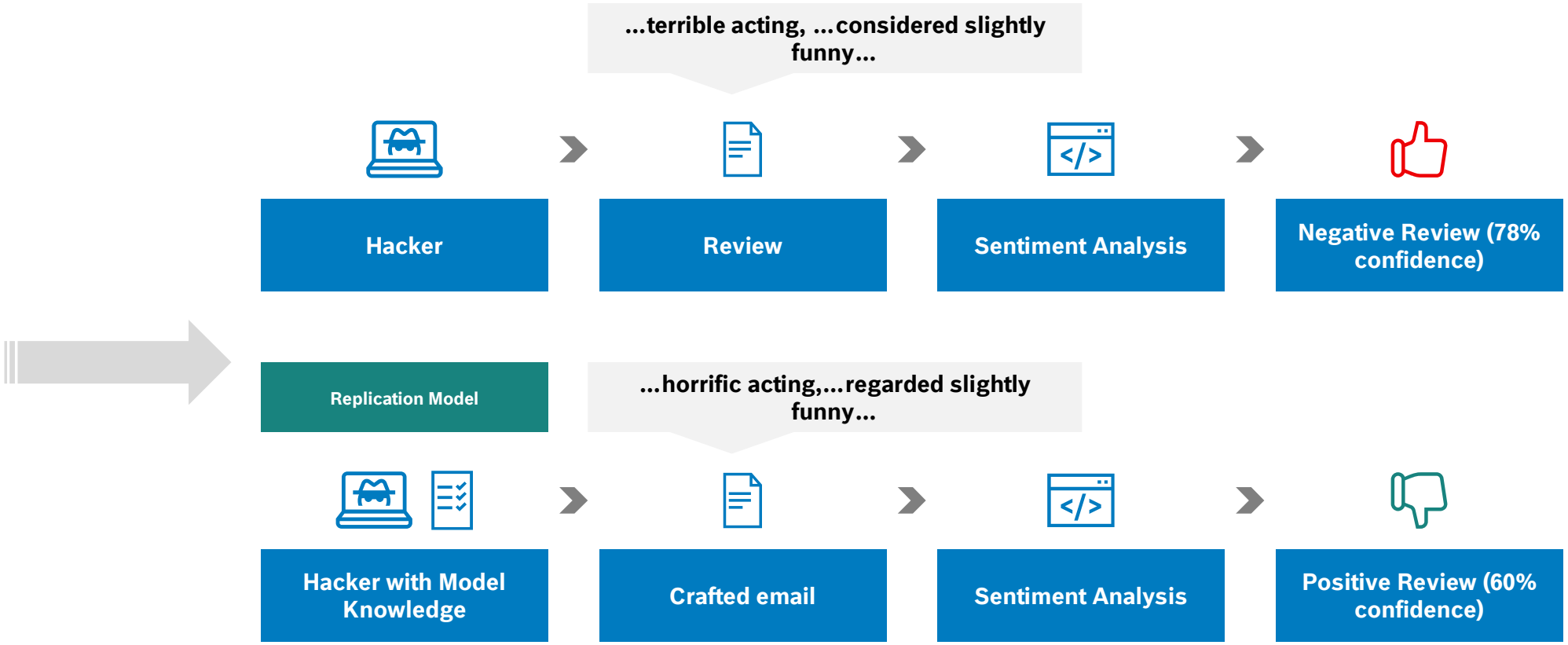
Bosch AIShield Case Studies

Movie Review Sentiment Analysis - Model Extraction



Bosch AIShield Case Studies

Movie Review Sentiment Analysis - Model Evasion



Bosch AIShield Case Studies

Recommendation Engine – attacks & kill-chain

SUMMARY

Model Extraction attack

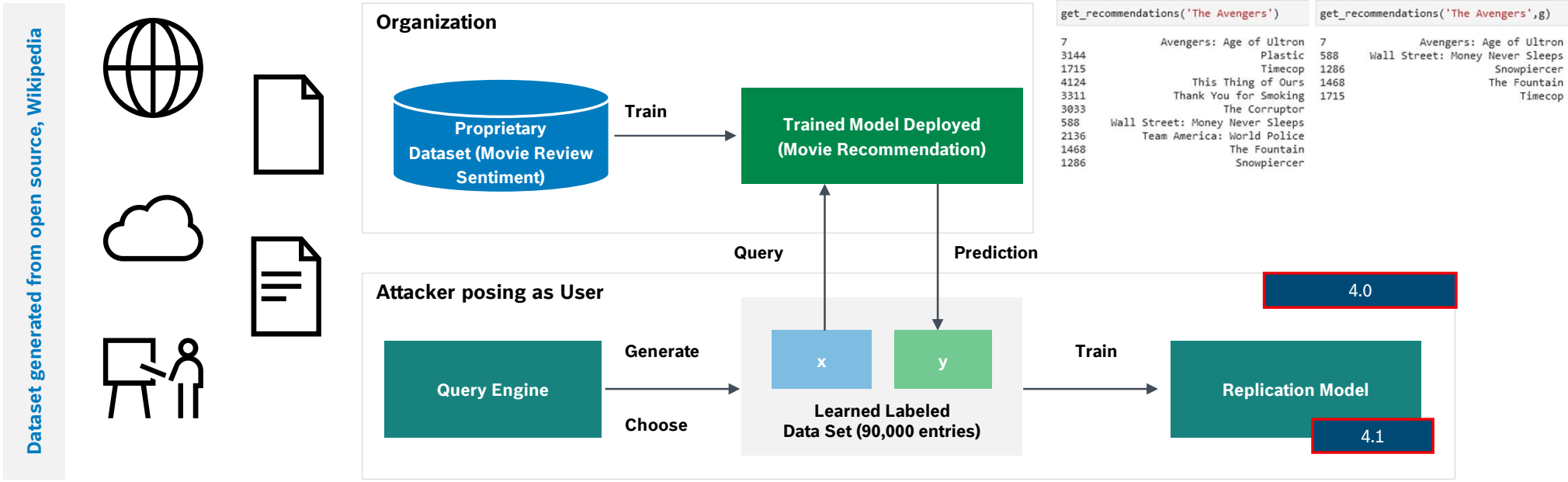
- Recommendation services provide public-facing UIs and APIs. A hacker can utilize these public endpoints to create a replicated model with near-production, state-of-the-art recommendation quality.
- The IP can be stolen and deployed for similar services by adversaries also resulting in financial losses.

1.0 (Reconnaissance) → **2.0** (Resource Development) → **3.0** (ML Model Access) → **4.0** (ML Attack Staging)
- 4.1 (Impact)

#	Kill-chain Phase	Technique	Description
1.0	Reconnaissance (The adversary is trying to gather information they can use to plan future operations.)	Search for Victim's Publicly available Research Material	The hackers used published information on web to identify the datasets by the target Recommendation services.
2.0	Resource Development The adversary is trying to establish resources they can use to support operations	Acquire Public ML Artifacts - Datasets	Hacker gathered similar datasets that the recommendation service used
3.0	ML Model Access (An adversary is attempting to gain some level of access to a machine learning model)	ML Model Inference Access	They abuse a public facing application to query the model posing as legitimate user and produce recommendation pairs as training data.
4.0	ML Attack Staging (An adversary is leveraging their knowledge of and access to the target system to tailor the attack.)	Create Proxy ML Model: Model Extraction	Using these recommendation pairs, the researchers trained a model that replicates the behavior of the target model.
4.1	Impact (The adversary is trying to manipulate, interrupt, erode confidence in, or destroy your systems and data.)	ML Intellectual Property Theft	By replicating the model with high fidelity, the researchers demonstrated that an adversary could steal a model and violate the victim's intellectual property rights.

Bosch AIShield Case Studies

Movie Recommendation Engine – Model Extraction




```
get_recommendations('The Avengers')
7                Avengers: Age of Ultron
3144                             Plastic
1715                             Timecop
4124                   This Thing of Ours
3311                Thank You for Smoking
3033                        The Corruptor
588     Wall Street: Money Never Sleeps
2136          Team America: World Police
1468                         The Fountain
1286                          Snowpiercer
```

```
get_recommendations('The Avengers',g)
7                Avengers: Age of Ultron
588     Wall Street: Money Never Sleeps
1286                          Snowpiercer
1468                         The Fountain
1715                             Timecop
```

THANK YOU