DeepLocker

Concealing Targeted Attacks with AI Locksmithing

Dhilung Kirat, Jiyong Jang, Marc Ph. Stoecklin

IBM **Research**

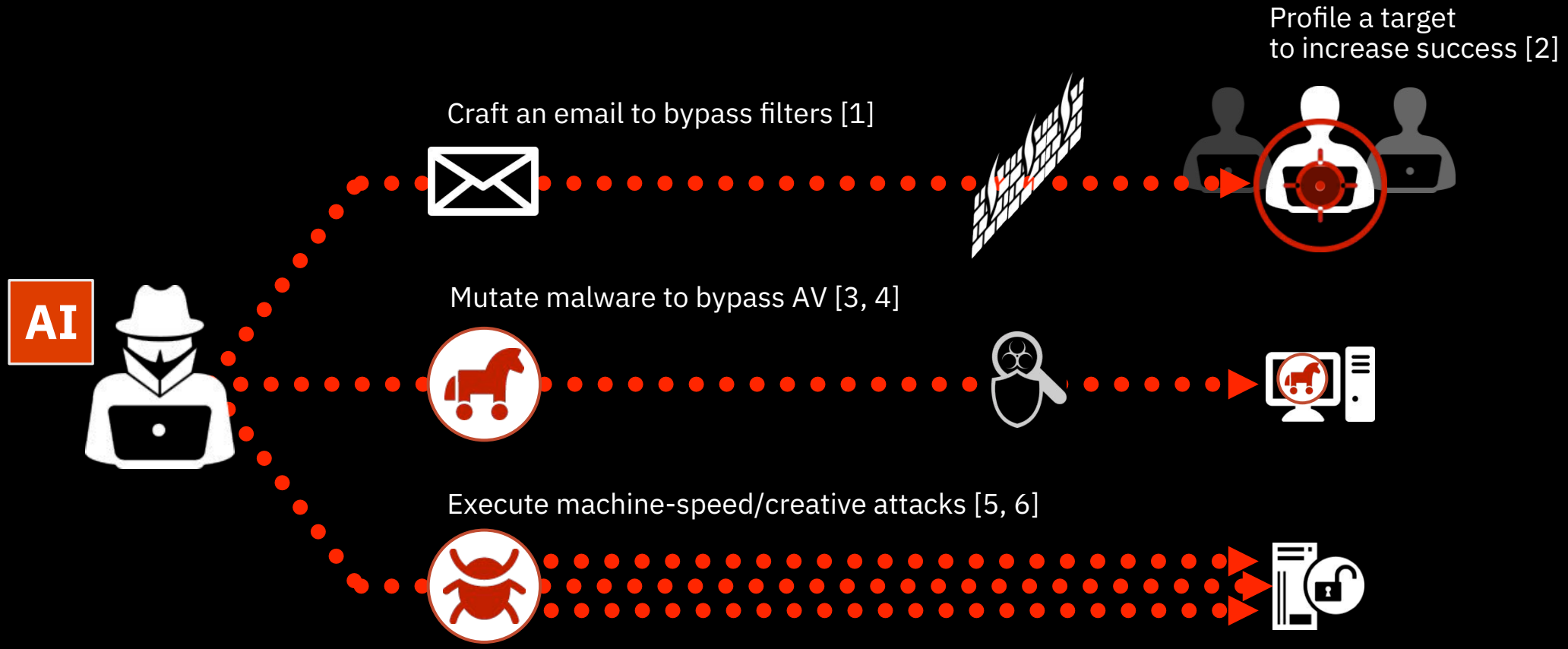
Dhilung Kirat

Jiyong Jang

Marc Ph. Stoecklin

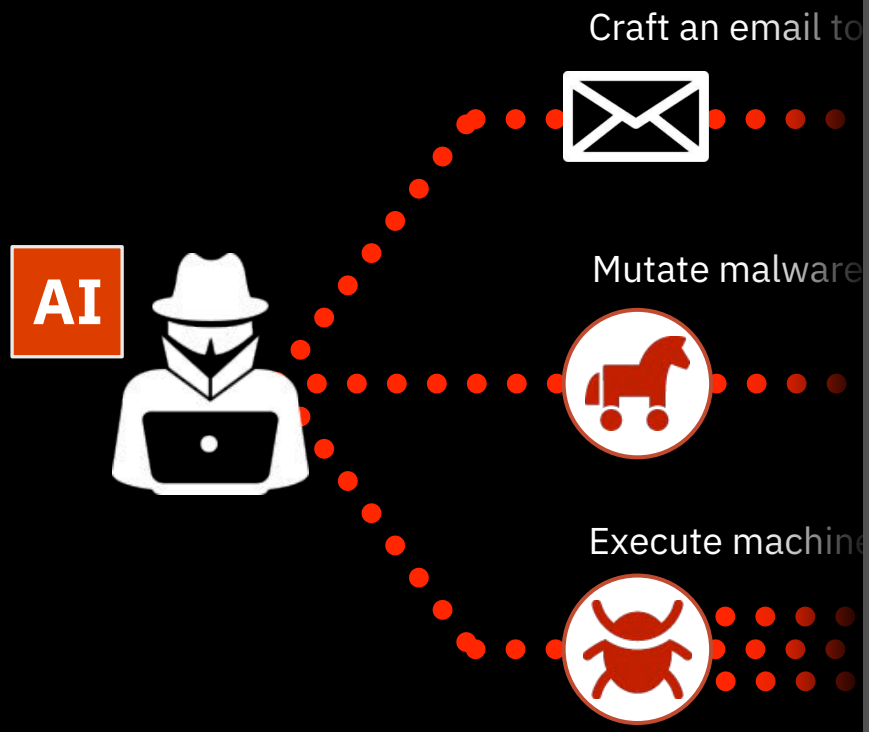
Cognitive Cyber Security Intelligence (CCSI)
IBM Research

AI-aided attacks



[1] S. Palka et al., “Fuzzing Email Filters with Generative Grammars and N-Gram Analysis”, Usenix WOOT 2015
[2] A. Singh and V. Thaware, “Wire Me through Machine Learning”, Black Hat USA 2017
[3] J. Jung et al., “AVPASS: Automatically Bypassing Android Malware Detection System”, Black Hat USA 2017
[4] H. Anderson, “Bot vs. Bot: Evading Machine Learning Malware Detection”, Black Hat USA 2017
[5] DARPA Cyber Grand Challenge (CGC), 2016
[6] D. Petro and B. Morris, “Weaponizing Machine Learning: Humanity was Overrated Anyway”, DEF CON 2017

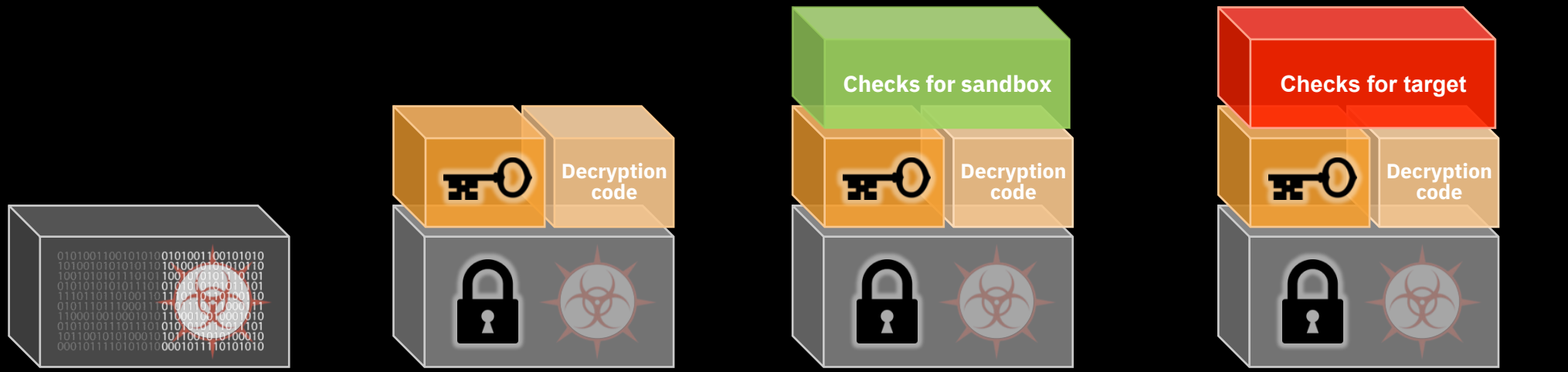
AI-aided attacks



AI-embedded attack

AI capability *embedded* inside malware itself

Malware concealment – Locksmithing



Obfuscation
Mutate payload

Polymorphism
Metamorphism

Encryption
Hide payload

Packers

Evasive malware
Avoid being analyzed

VM check
Processes check

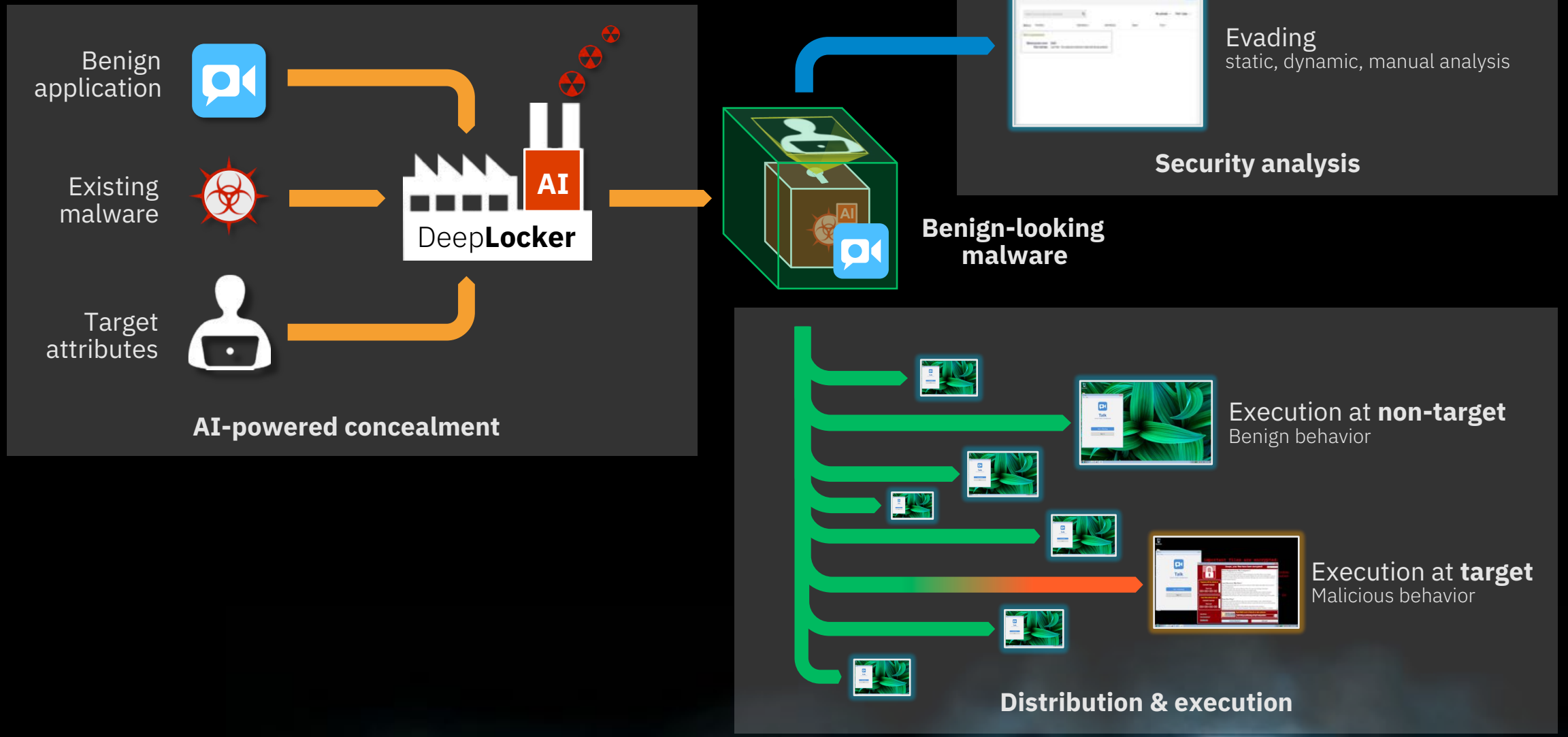
Targeted attack
Disclose only at a target

Target attributes check

AI Locksmithing

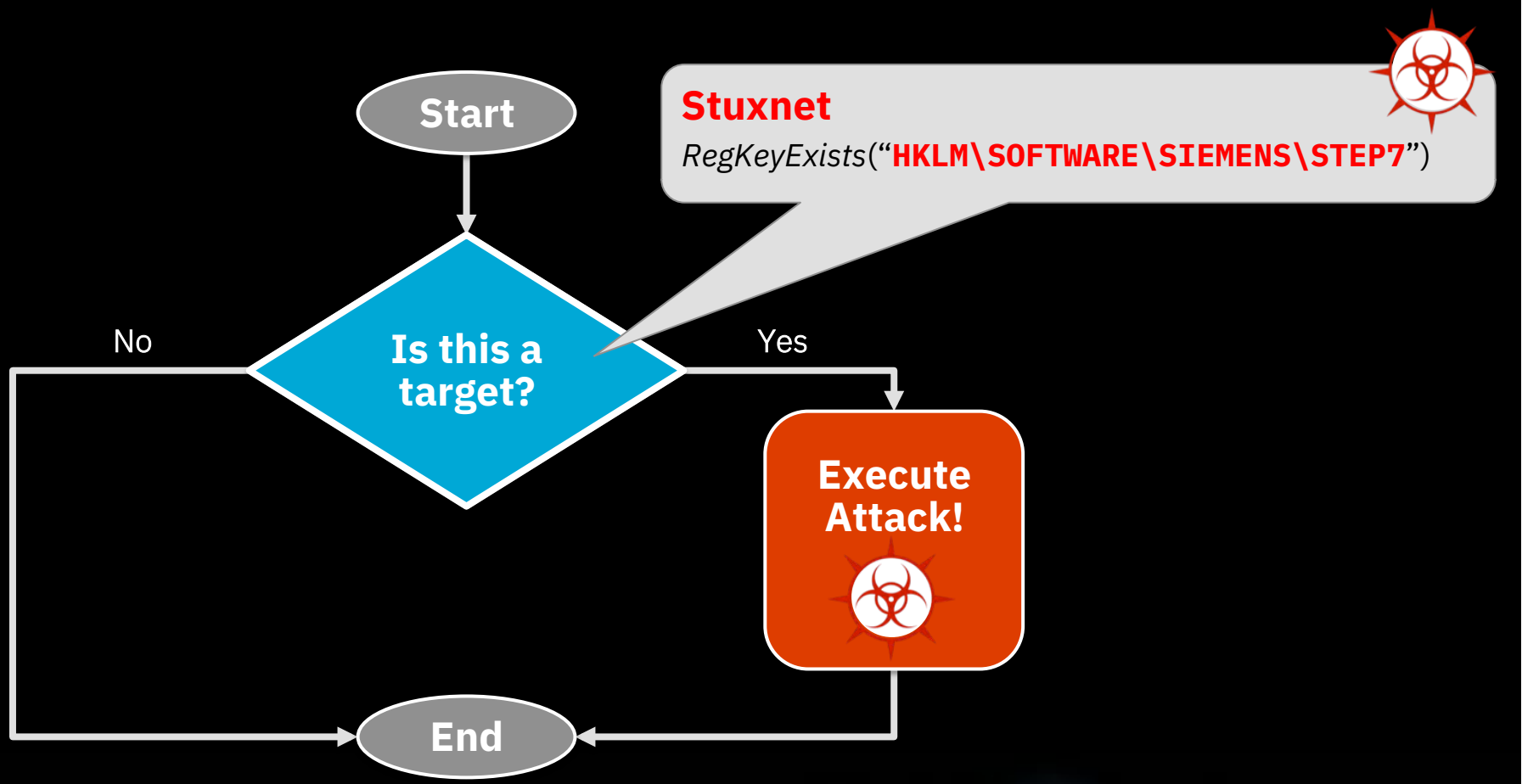

Unleashing DeepLocker – AI Locksmithing

DeepLocker – Overview

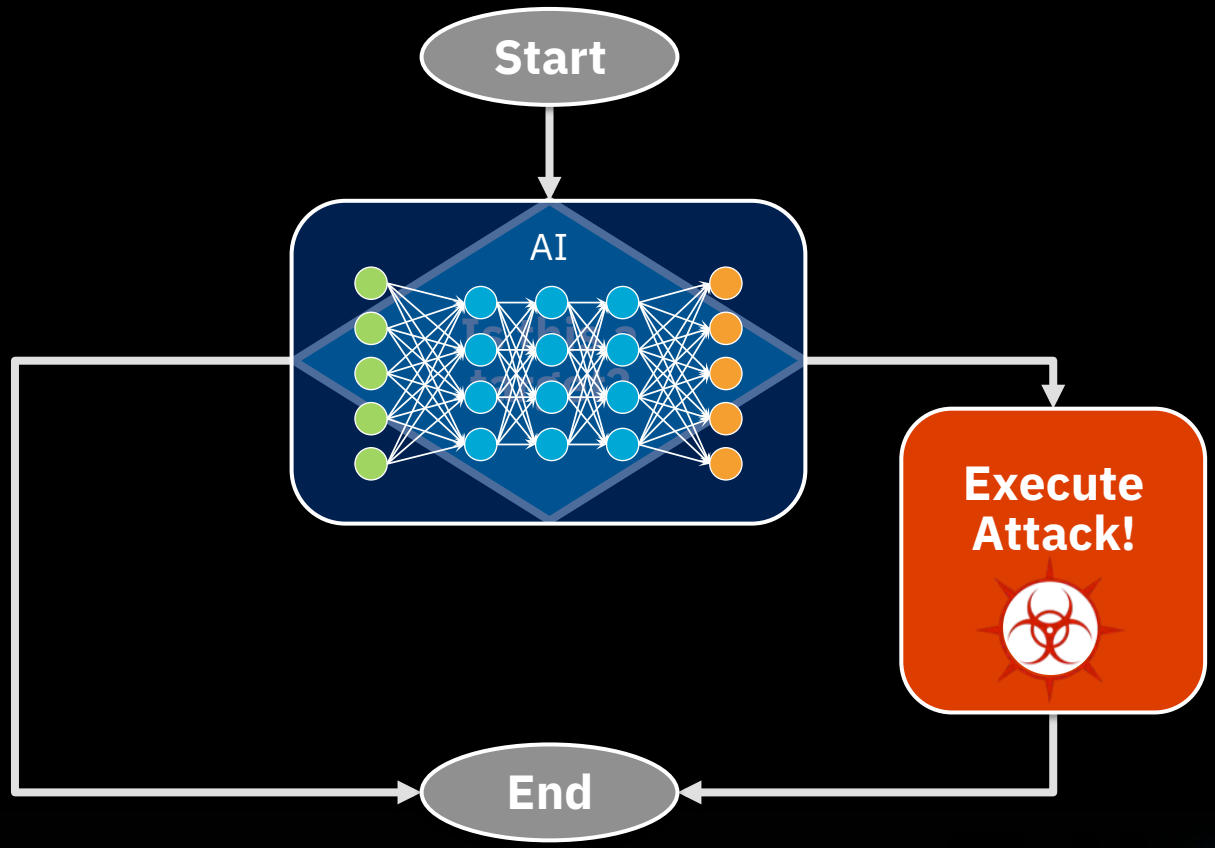



DeepLocker Deep Dive

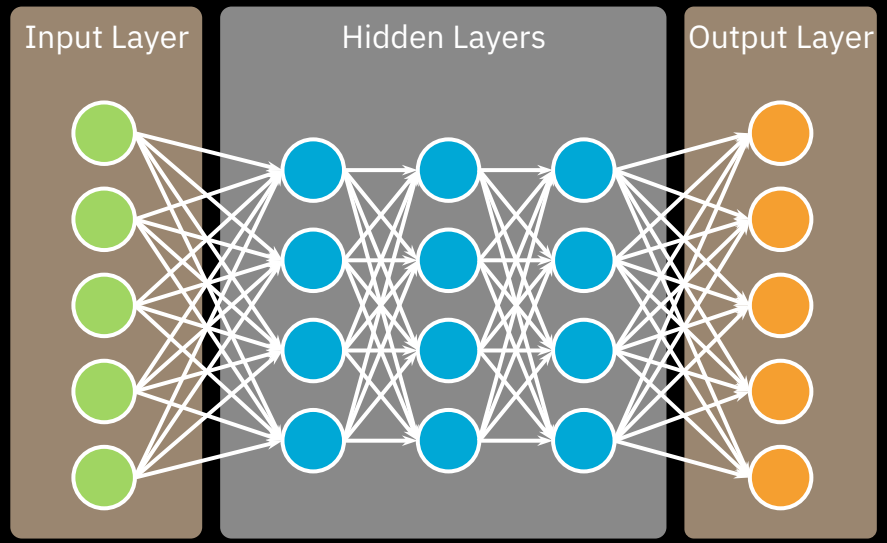
Traditional targeted attack

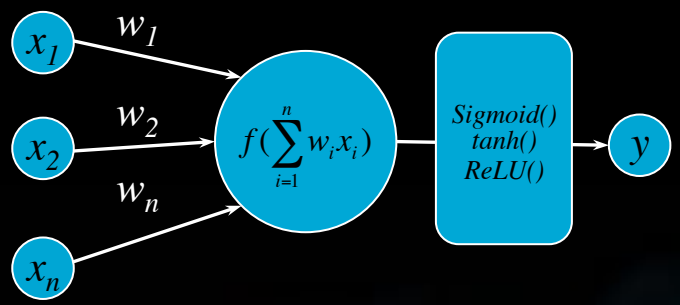


AI-powered targeted attack

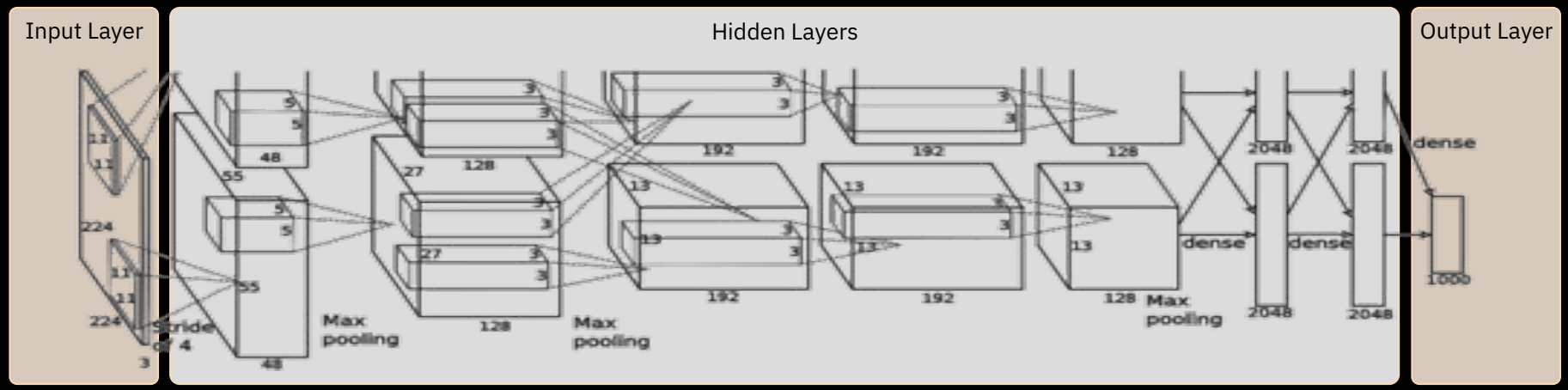


What is a Deep Neural Network (DNN)?





Deep Convolutional Neural Network

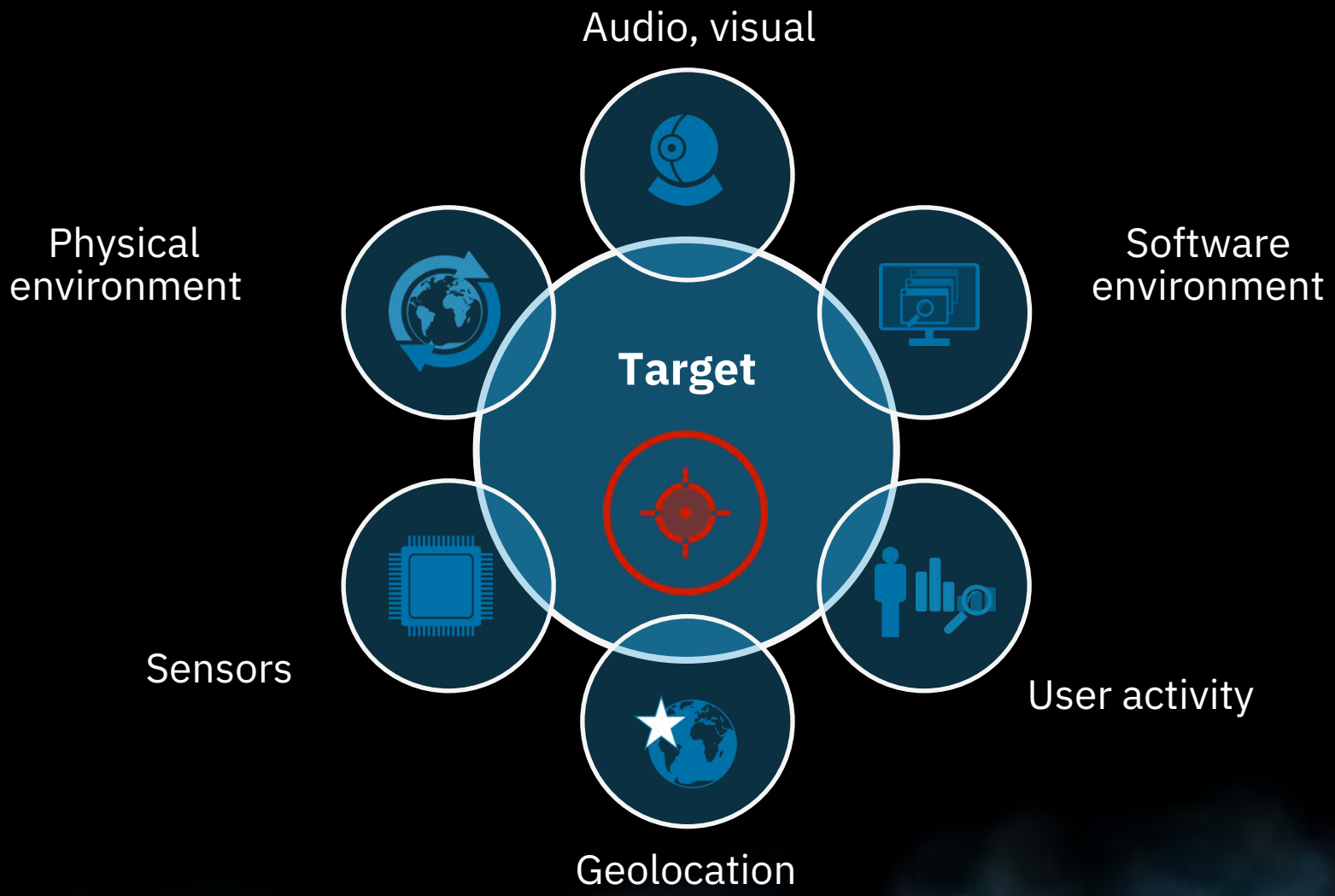


AlexNet (2012) [1]

8 layers, **622K** neurons, **60 million** parameters

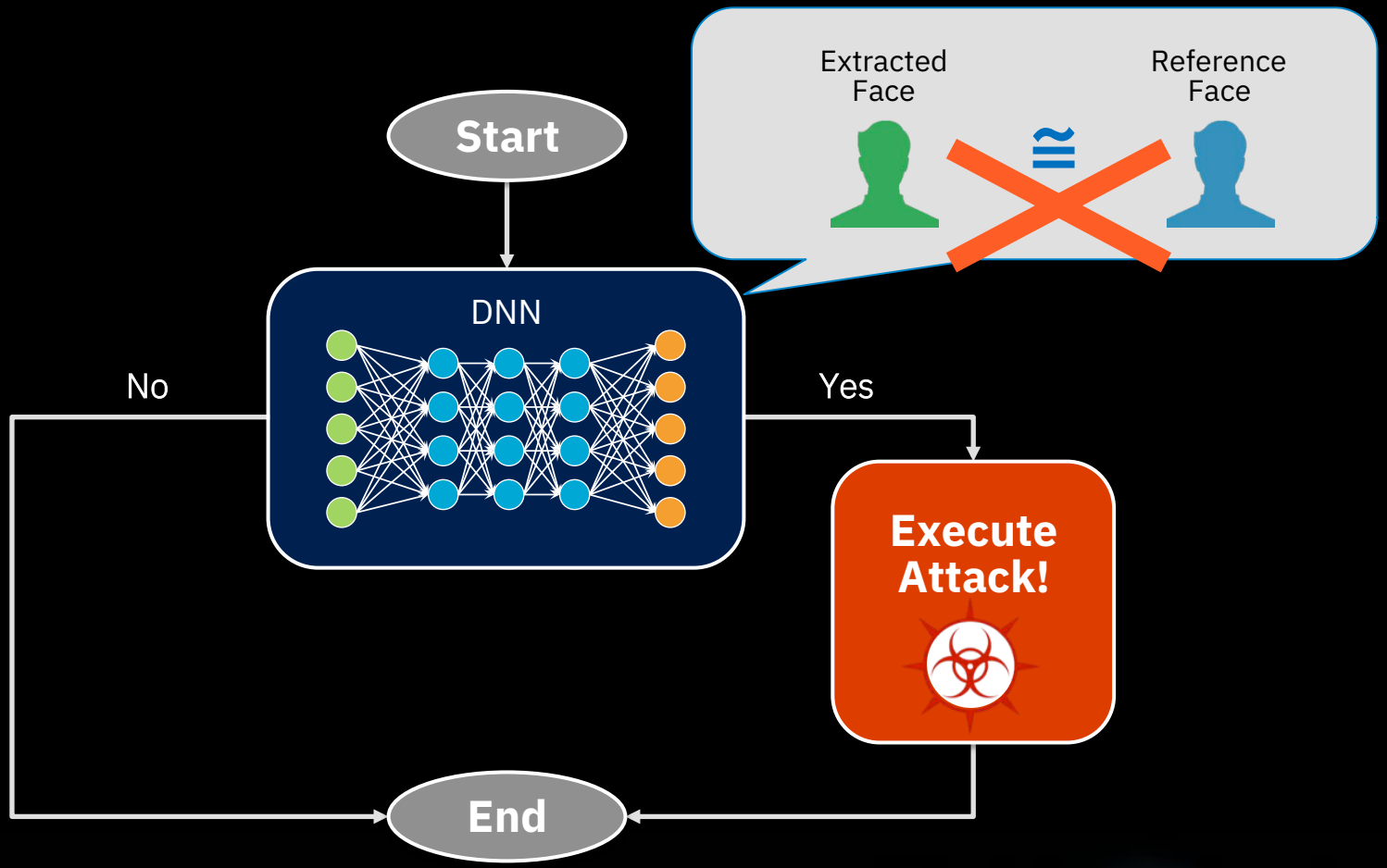
[1] Krizhevsky, Alex, et. al. "Imagenet classification with deep convolutional neural networks." NIPS 2012.

Target attributes

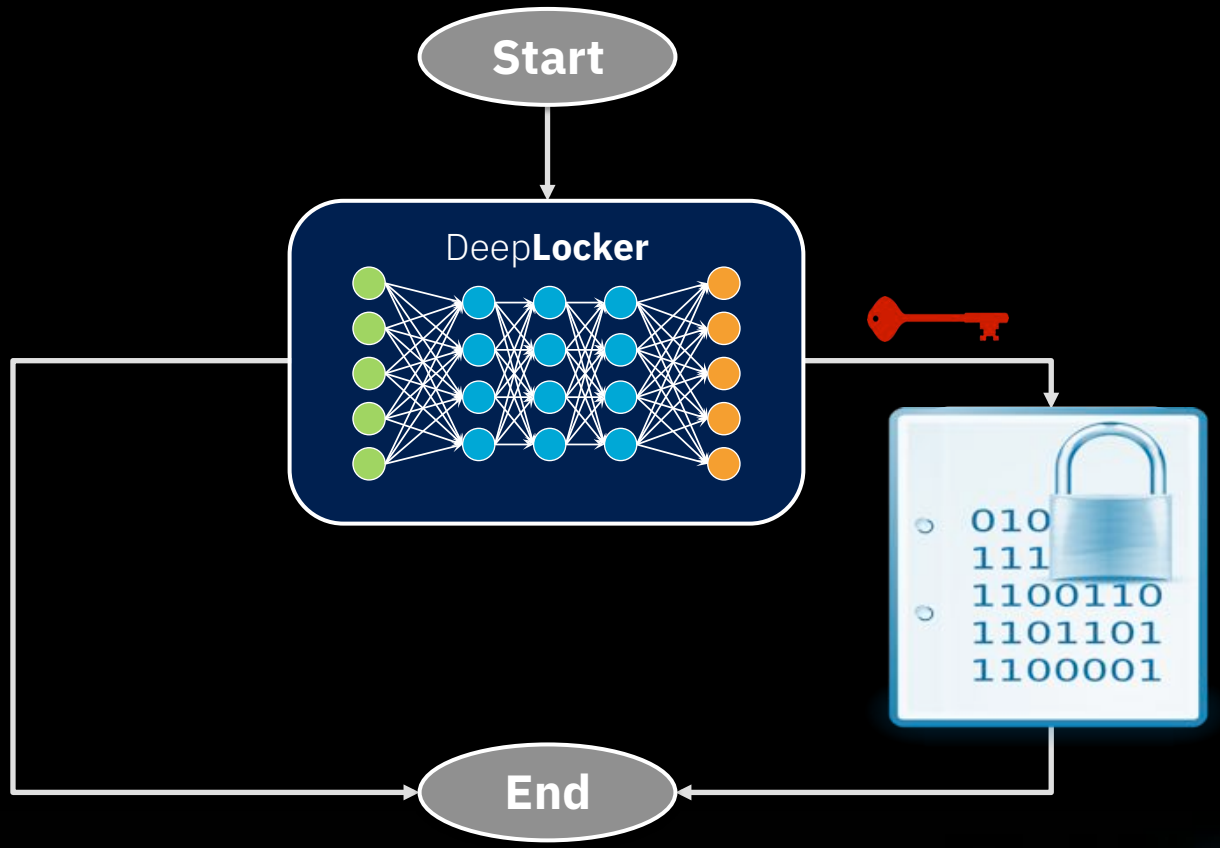


Target detection

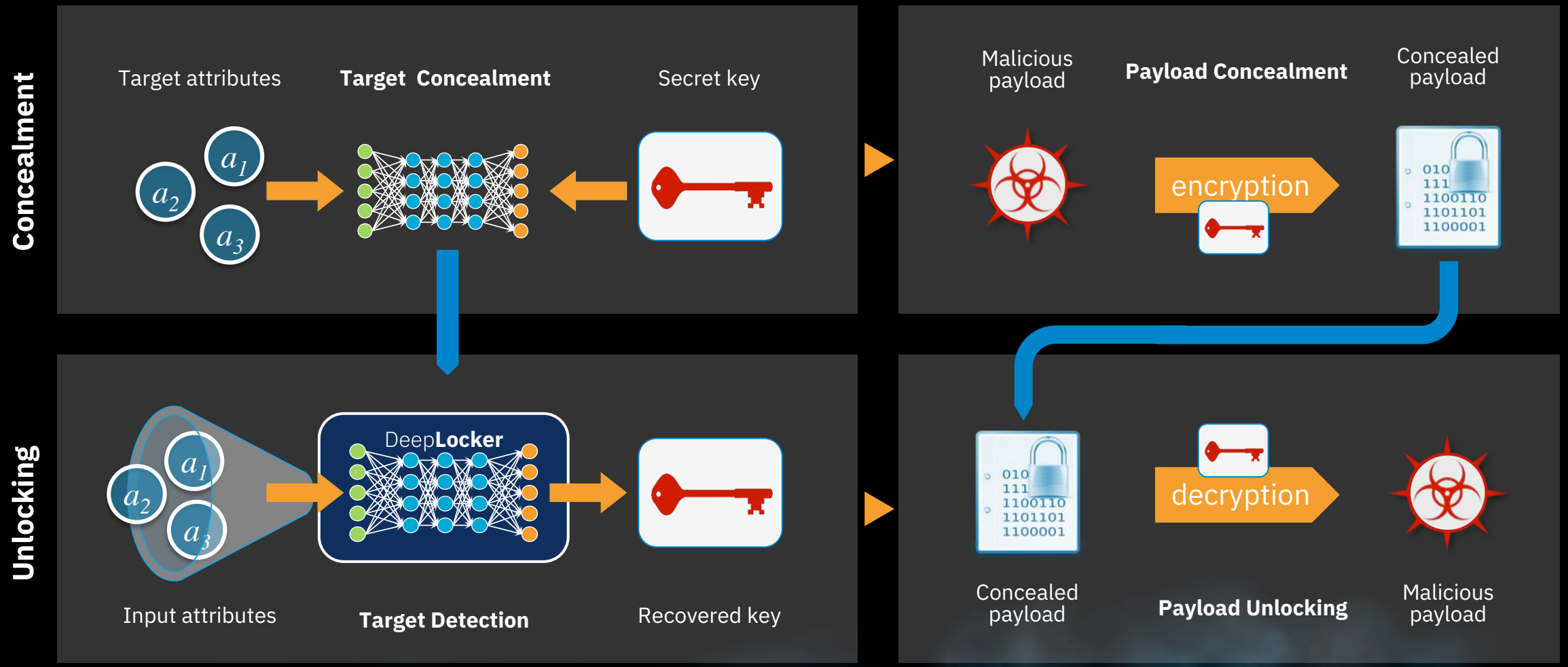
Template matching requires a template to match to.



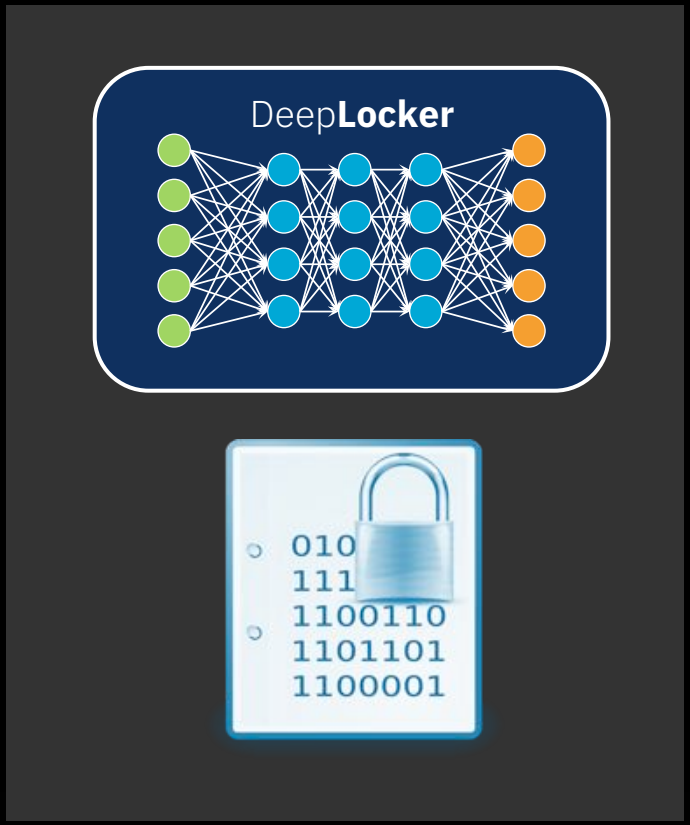
Derivation of an unlocking key



DeepLocker – AI-Powered Concealment and Unlocking

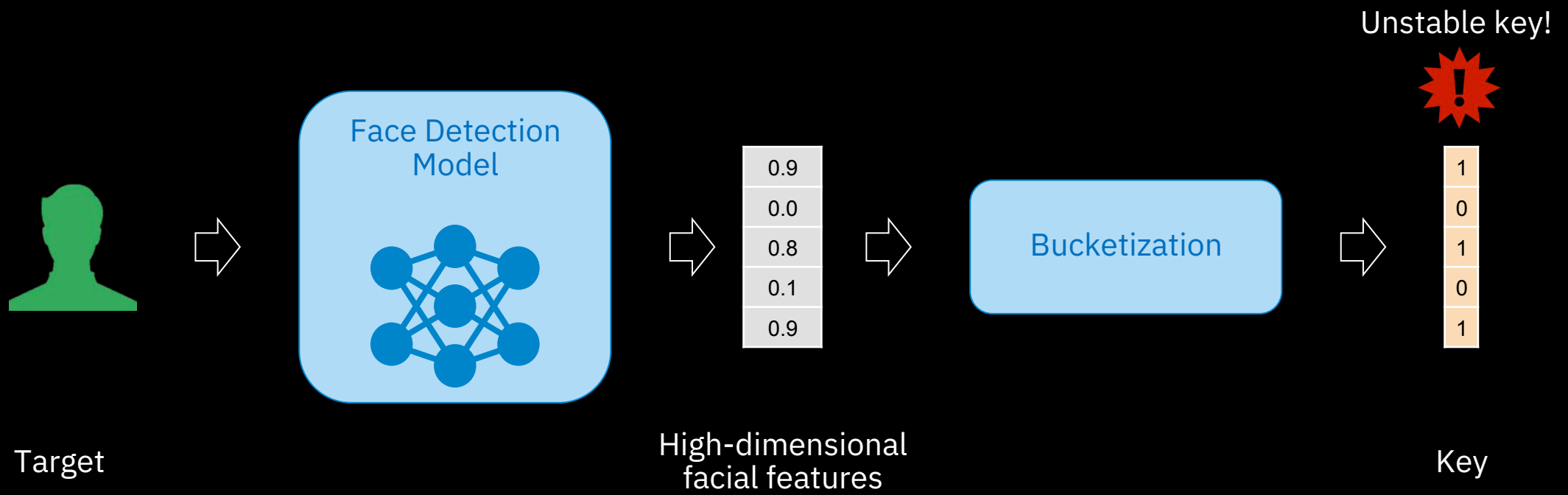


AI-powered concealment

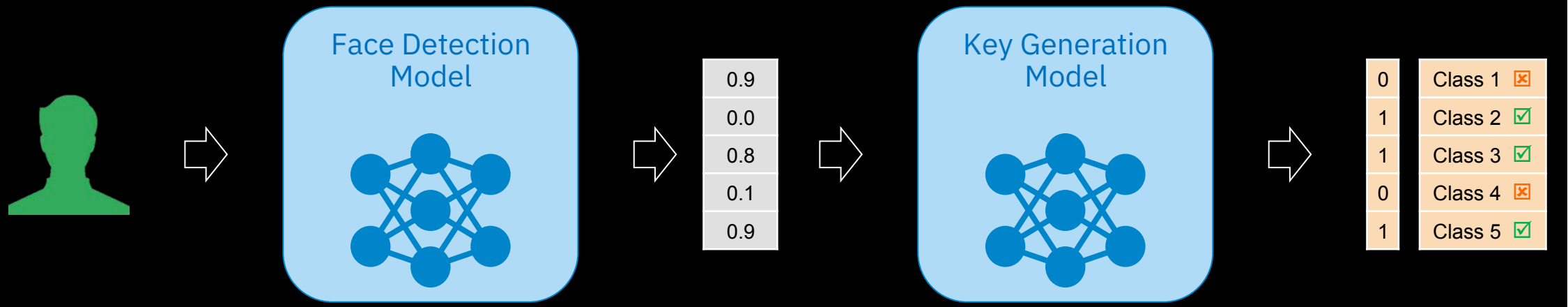


No decryption key available in malware sample to reverse engineer!

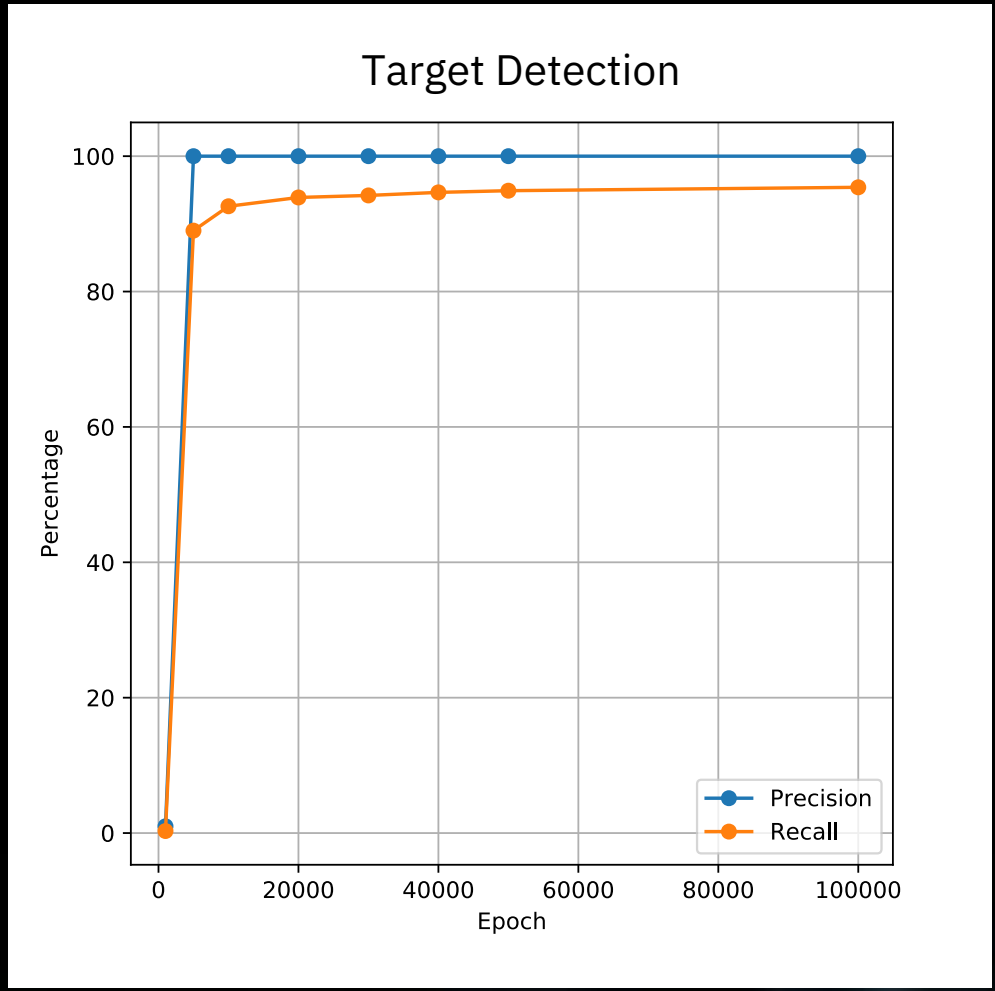
Key generation



Key generation

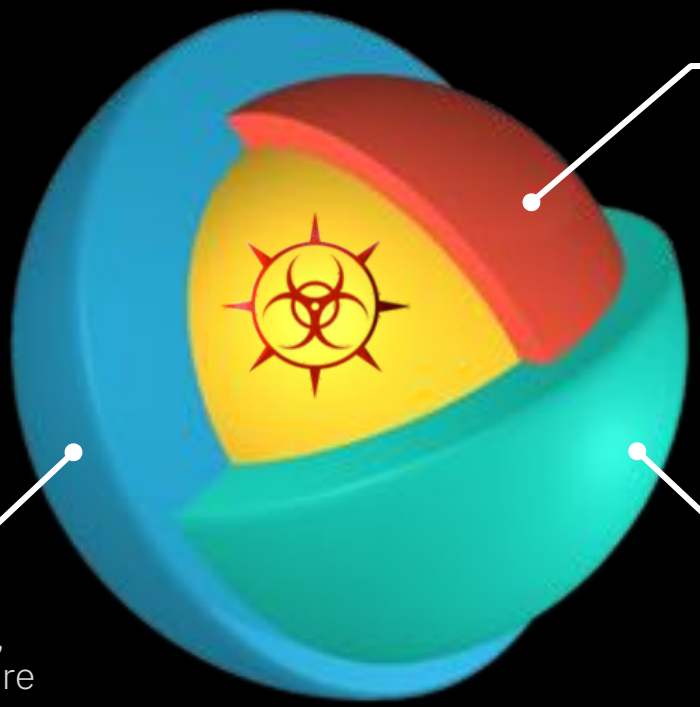


Analysis of the key generation model



Dataset: Labeled Faces in the Wild (LFW)
http://vis-www.cs.umass.edu/lfw/

Deep**Locker** – AI-powered concealment



3 Malicious Intent Concealment

Payload is fully encrypted concealing **how** the final attack is executed

1 Target Class Concealment

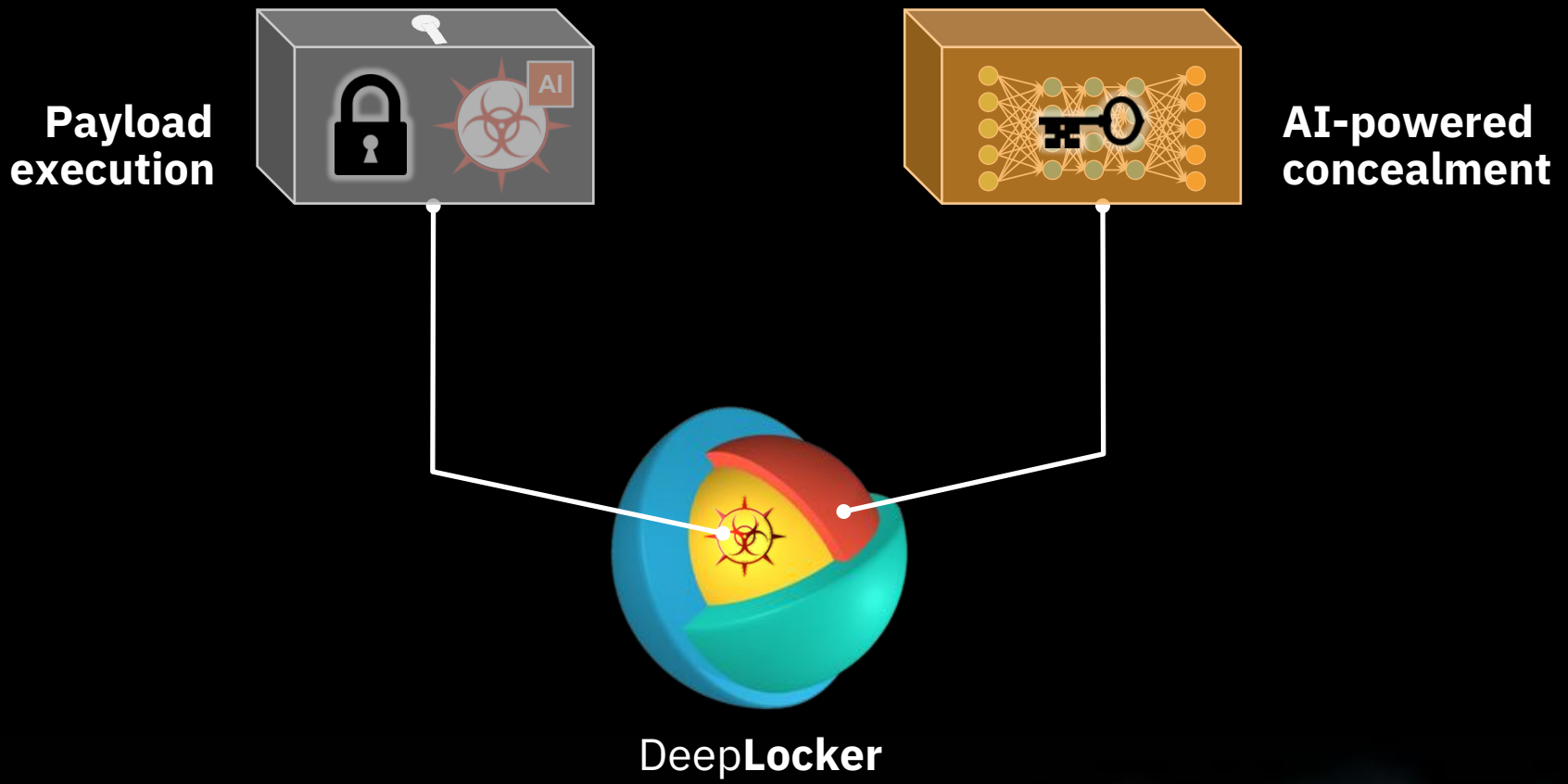
Does not reveal **what** it is looking for (e.g., faces, organization, or a completely obscure object specific to the target environment)

2 Target Instance Concealment

If the target class is an individual, it does not reveal **who** it is looking for

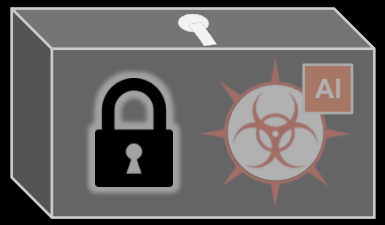
Attacking DeepLocker – AI Lock Picking

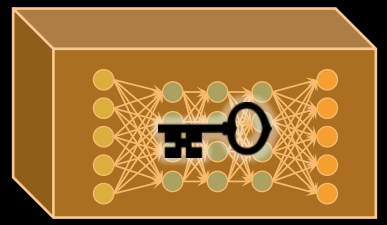
Ways to counter



Ways to counter

Payload execution





AI-powered concealment

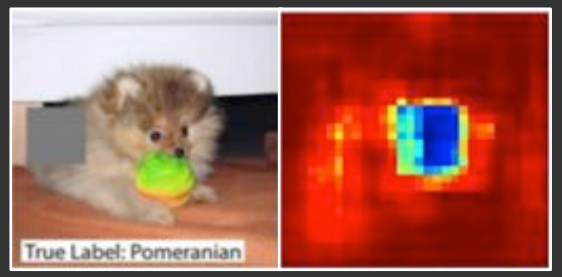
Code attestation	Block sensor access
Host-based monitoring	AI usage monitoring
Brute-force key	Brute-force attributes
Deceptive resources	Deceptive attributes
Code analysis	AI lock picking

?

Reverse engineering AI models

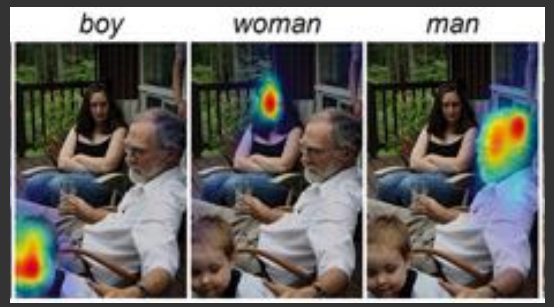
Partial occlusion

Occlude a portion of the image to see how the embedding is affected (deconvnet) [1]



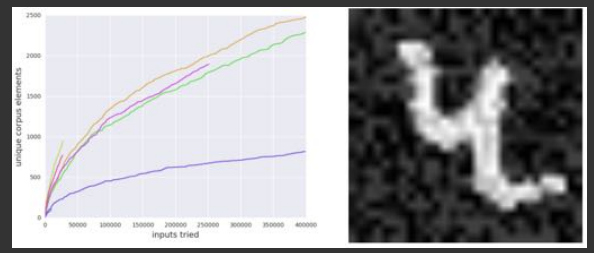
Neural attention model

Heatmap using the degree of excitation of neurons in each layer (excitation backprop) [2]



Debug neural networks

Fuzzing for neural networks (coverage-guided fuzzing) [3]



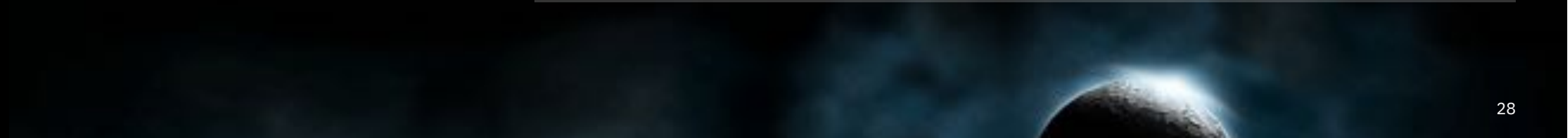
[1] M. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” ECCV 2014
[2] J. Zhang et. al., “Top-down neural attention by excitation backprop,” ECCV 2016
[3] A. Odena and I. Goodfellow, “TensorFuzz: Debugging neural networks with coverage-guided fuzzing,” arXiv 2018

Takeaways

Rapid democratization of AI has made
AI-powered attacks an imminent threat

DeepLocker is a demonstration of the potential of
AI-embedded attacks

Current defenses will become obsolete and
new defenses are needed



Thank **you**

Dhilung Kirat ✉ *dkirat@us.ibm.com*

Jiyong Jang ✉ *jjang@us.ibm.com*

Marc Ph. Stoecklin ✉ *mpstoeck@us.ibm.com*

IBM **Research**