RSAConference2018
San Francisco | April 16 – 20 | Moscone Center

SESSION ID: TV-T02

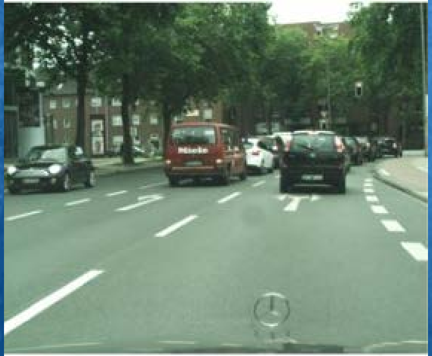
AI DECEPTION: FOOLING (ARTIFICIAL) INTELLIGENCE IS EASIER THAN YOU THINK

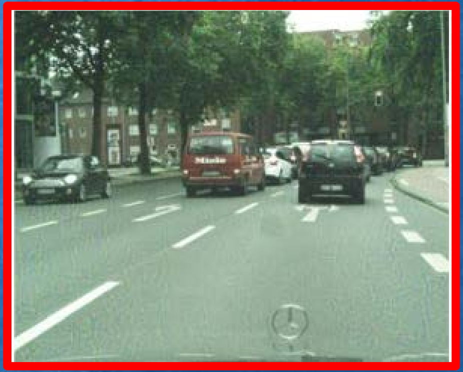
Itsik Mantin

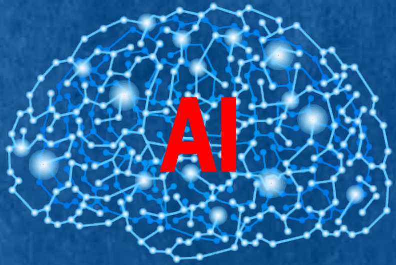
Lead Scientist
Imperva

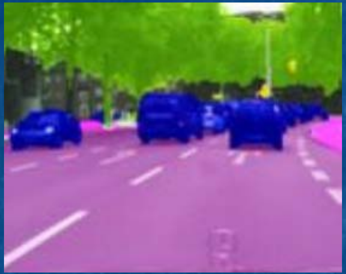
AI Deception

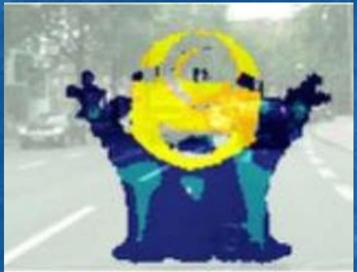
Fooling AI by Adversaries



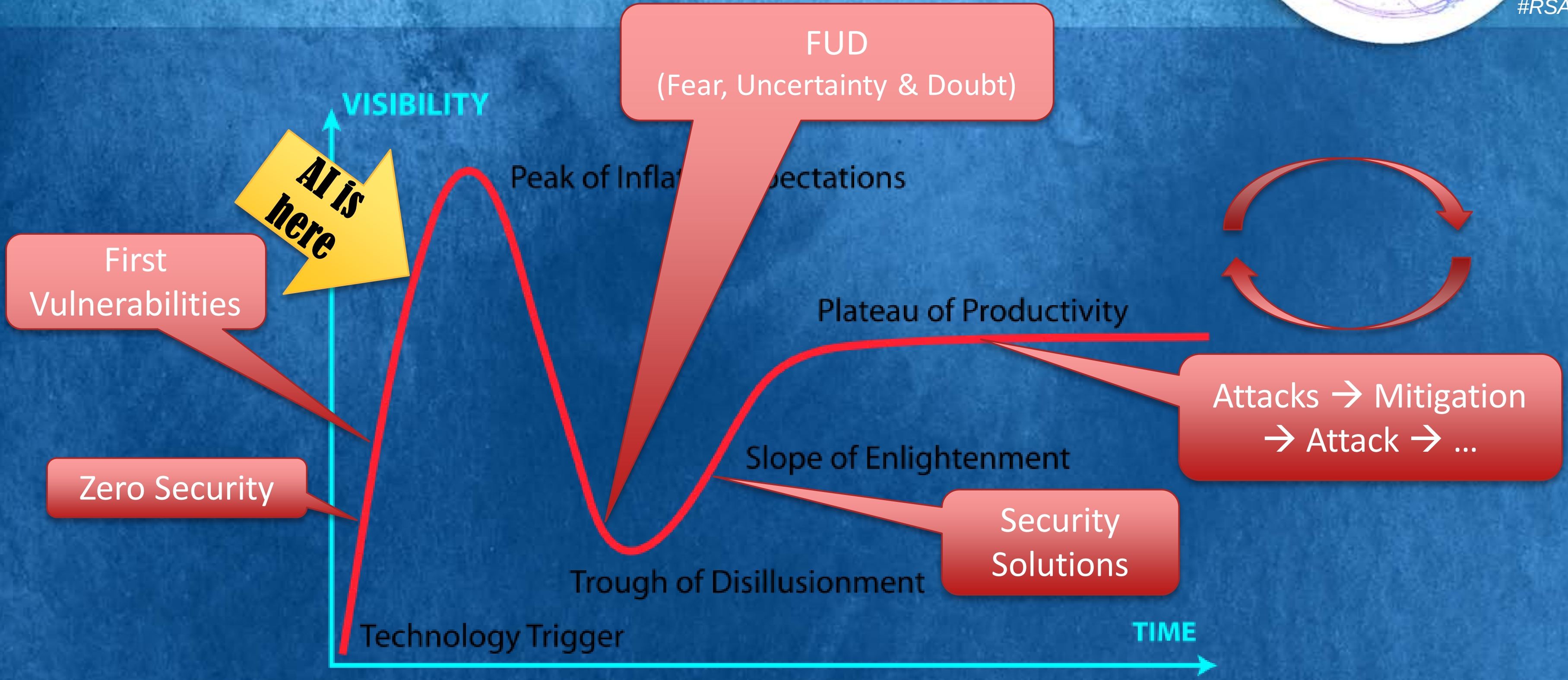




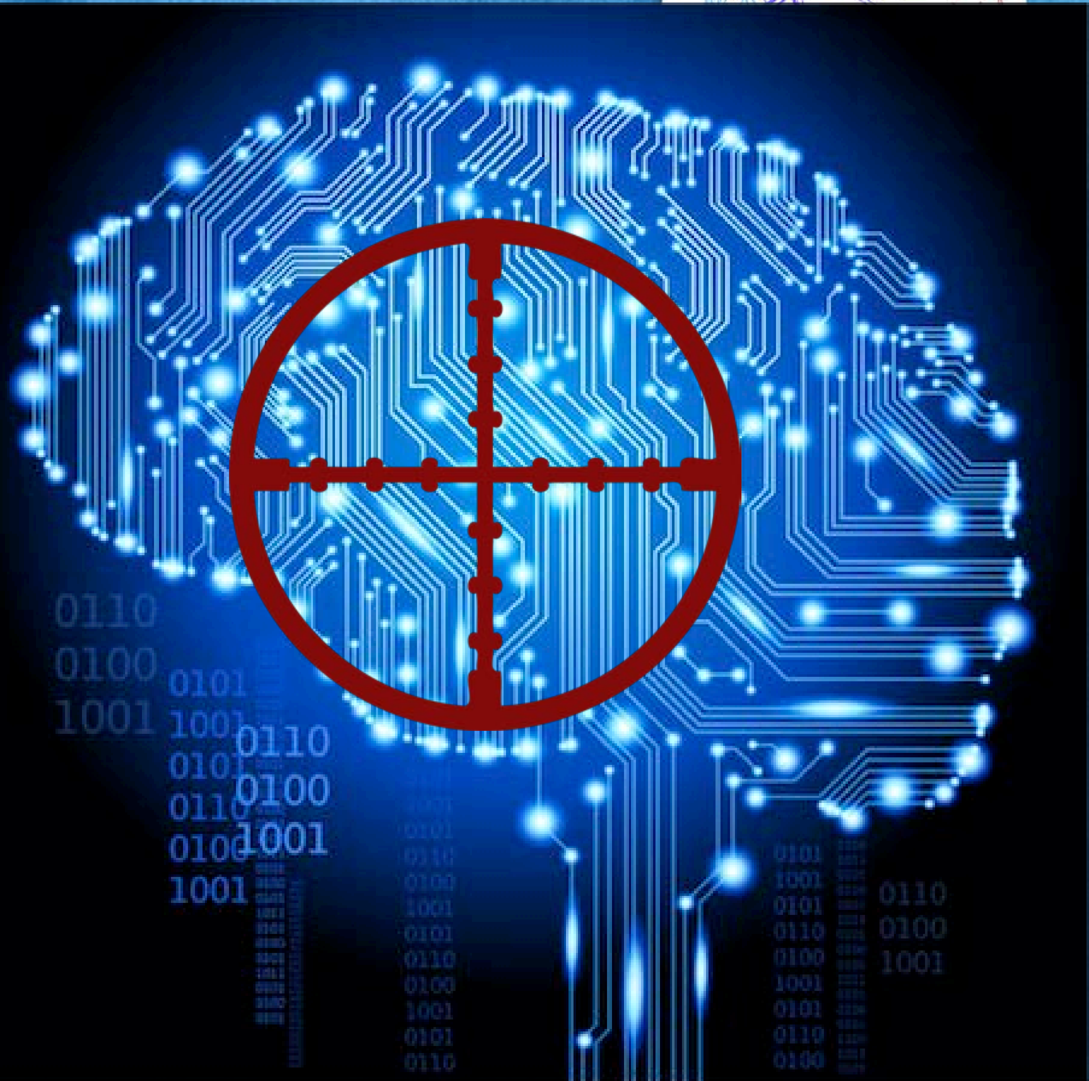




Security and the Hype-Cycle

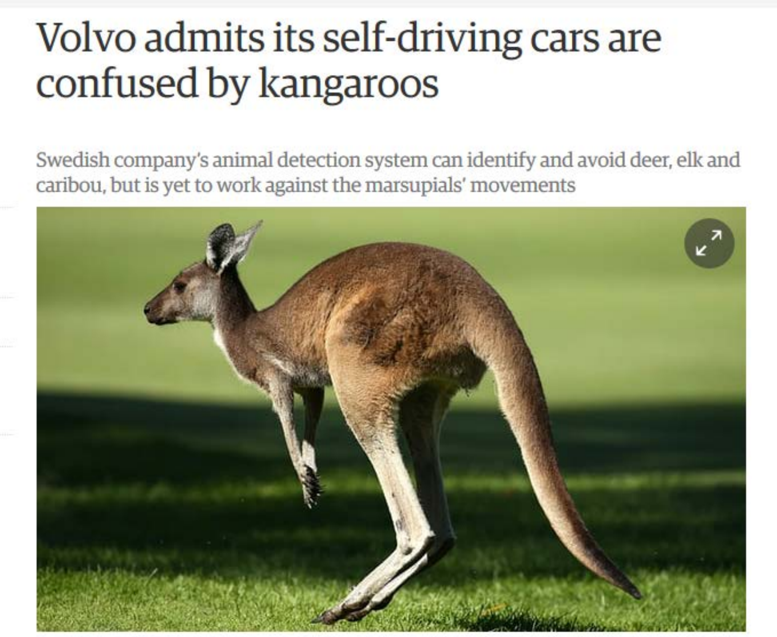


No Exemptions for AI!



The Australian Challenge

Fooling AI by ~~Adversaries~~ Innocent



Volvo admits its self-driving cars are confused by kangaroos

Swedish company's animal detection system can identify and avoid deer, elk and caribou, but is yet to work against the marsupials' movements

Innocent AI Deception

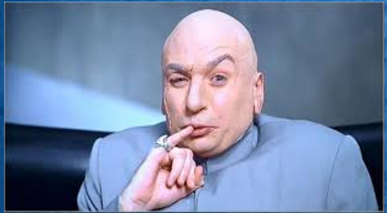
Kangaroos
... Hmmm..

Adversarial Thinking

	Builder	Adversary
Primary Focus	What happens in case of normal input	What may happen in case of anomalous input?
Failure in rare coincidence	Something I can ignore	Something I can abuse

Blind Spots

Artificial Kangaroo

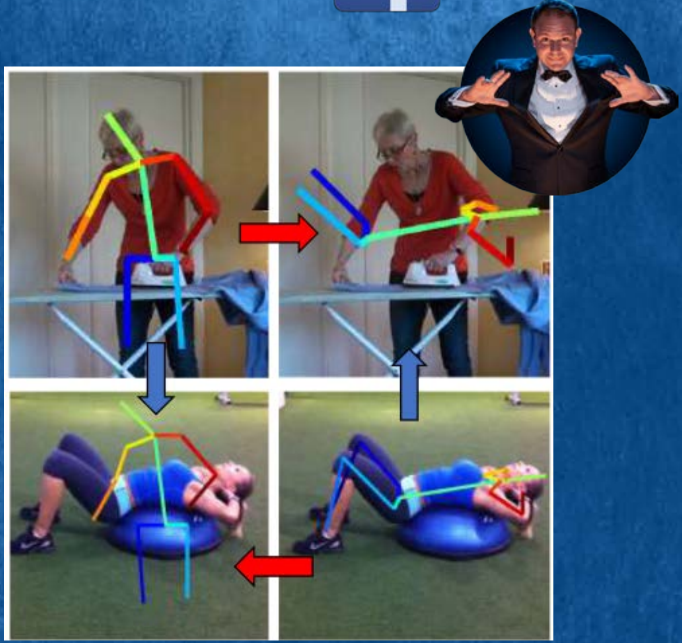


Getting Hit by an Ostrich

Adversarial Examples

Bar-Ilan University
אוניברסיטת בר-אילן

Houdini Research

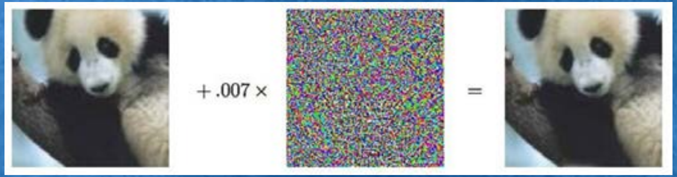


Pose Estimation

Speech Recognition

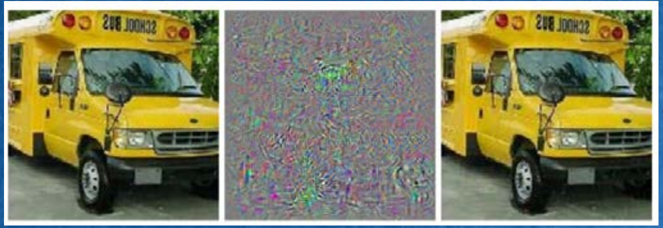
Semantic Segmentation

Visual Object Recognition



Panda (57.7%)

Gibbon (99.3%)

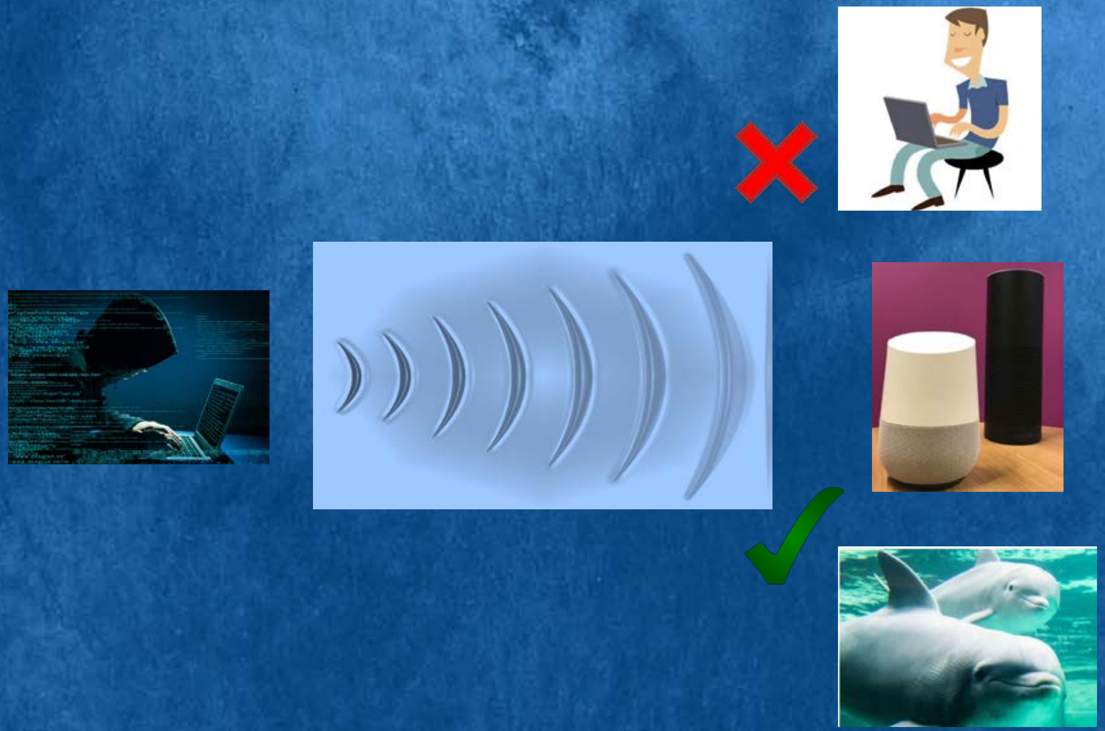


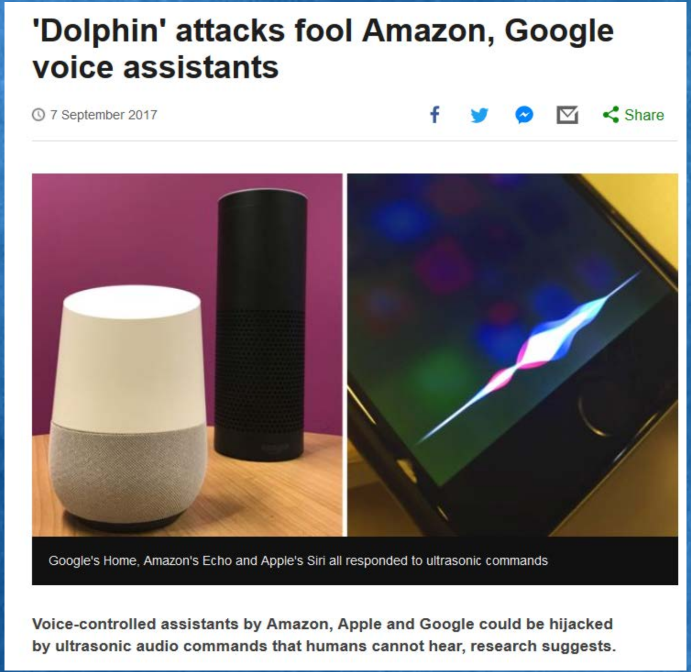
School Bus

Ostrich

Dolphin Attack

Speech Recognition





'Dolphin' attacks fool Amazon, Google voice assistants

7 September 2017

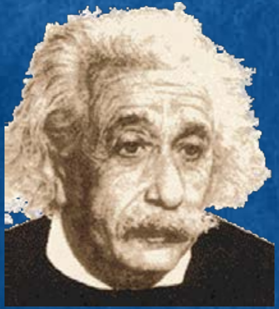
Share

Google's Home, Amazon's Echo and Apple's Siri all responded to ultrasonic commands

Voice-controlled assistants by Amazon, Apple and Google could be hijacked by ultrasonic audio commands that humans cannot hear, research suggests.

Define AI Deception

- Given an A/B classifier, and given a sample X correctly classified as A, attacker generates a sample X’ that:
  1) Has **same-essence** as X, and
  2) Classified as B



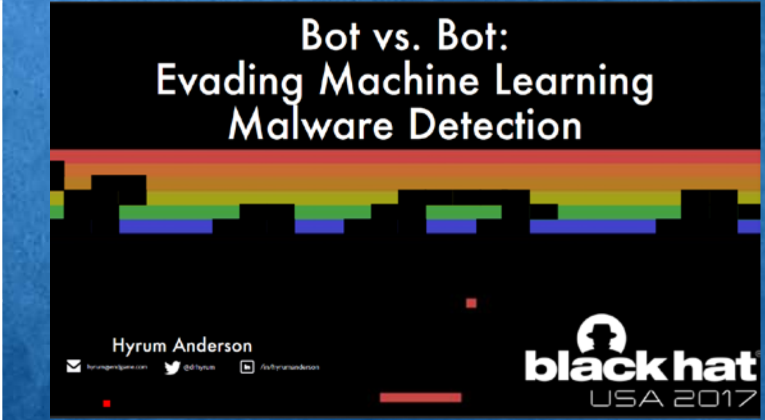
God does not play dice…

…but the devil does

AI Deception Threat

Malware Detection

	Same-Essence	AI Deception
Image	Identical for human viewer	Different objects detected
Video		Different scene detected
Voice	Identical for human listener	Different speech detected
Software	Same malicious function	Classified as non-malicious



Bot vs. Bot:
Evading Machine Learning Malware Detection

Hyrum Anderson

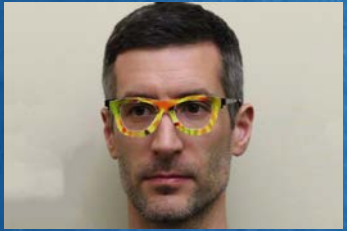
black hat USA 2017

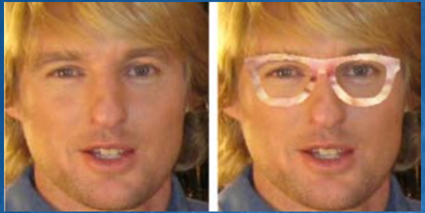
AI Deception – More Threats

Propagation to the Real World

Face Detection

Domain	AI Engine	AI Circumvention
Surveillance and Control systems	**Face detector**	**Embed face patterns that prevent correct detection**
Finance	Stock prediction	Cause stock patterns that imply positive prediction
Text analyzers	Translation Engine / Topic Extraction	Embed text patterns that prevent correct analysis
E-Commerce	Customized pricing	Force user profile or behavior that implies cheaper prices



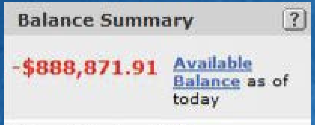


Are We Better?

HI Deception

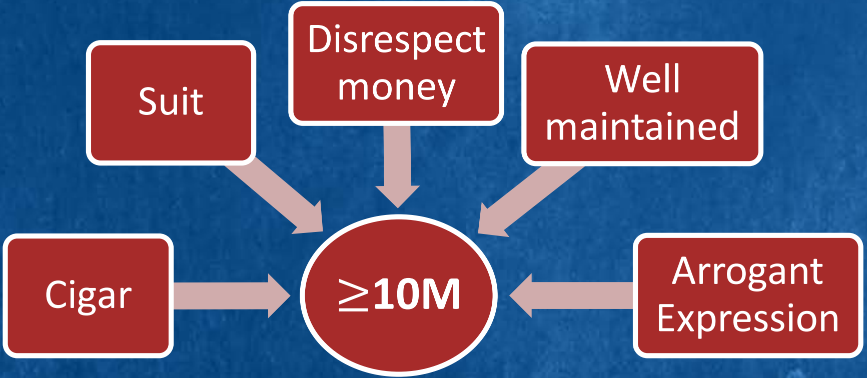
- Human Intelligence Deception

Wealth Prediction

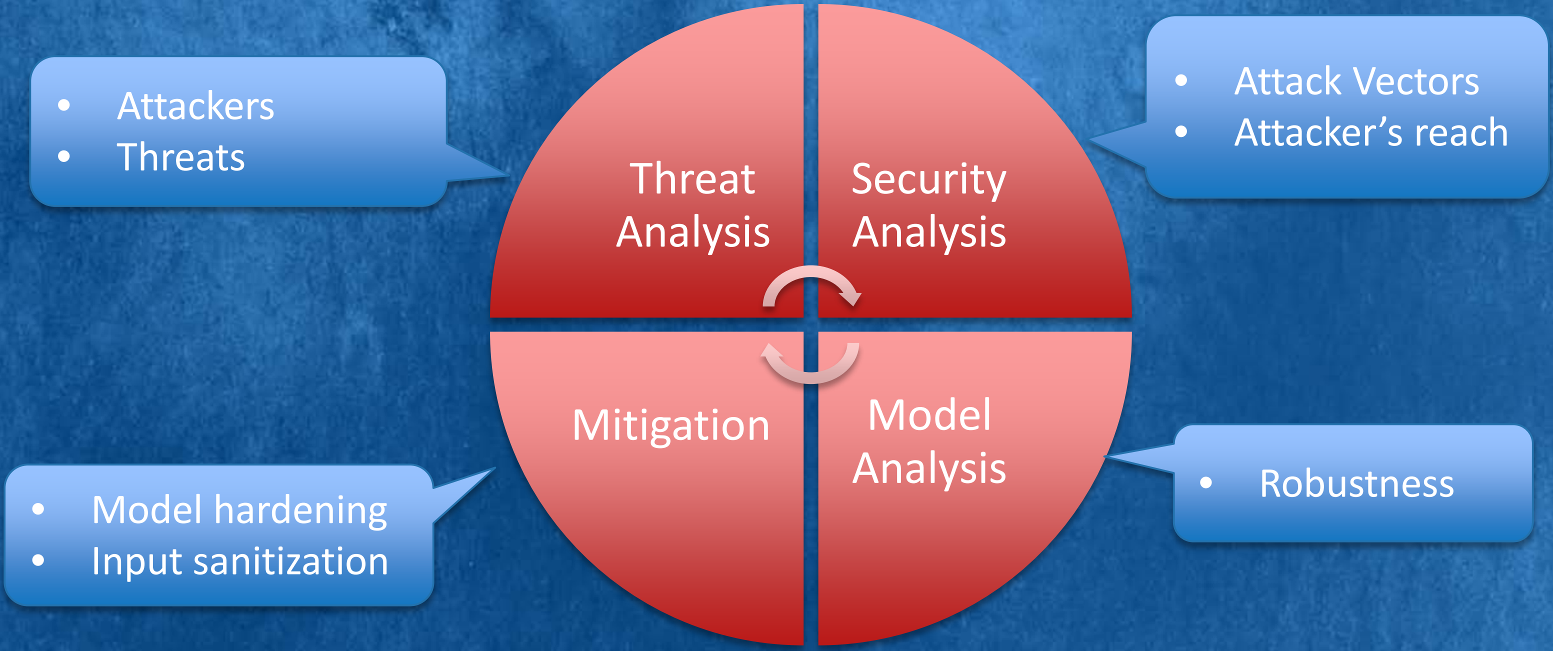


Balance Summary ?

-$888,871.91 Available Balance as of today



Risk Mitigation



- Attackers
- Threats

Threat Analysis

Security Analysis

- Attack Vectors
- Attacker’s reach

Model Analysis

- Robustness

Mitigation

- Model hardening
- Input sanitization

The AI’ker’s Guide to the (Cyber-Security) Galaxy

AI in Cyber Security Applications

The Model	• Prefer robust models • Prefer explainable models
Training	• Sanitize training data
Usage	• Prefer internal (hidden outcome) • Avoid raw input
Threat Detection	• Prefer positive security • Combine with other mechanisms

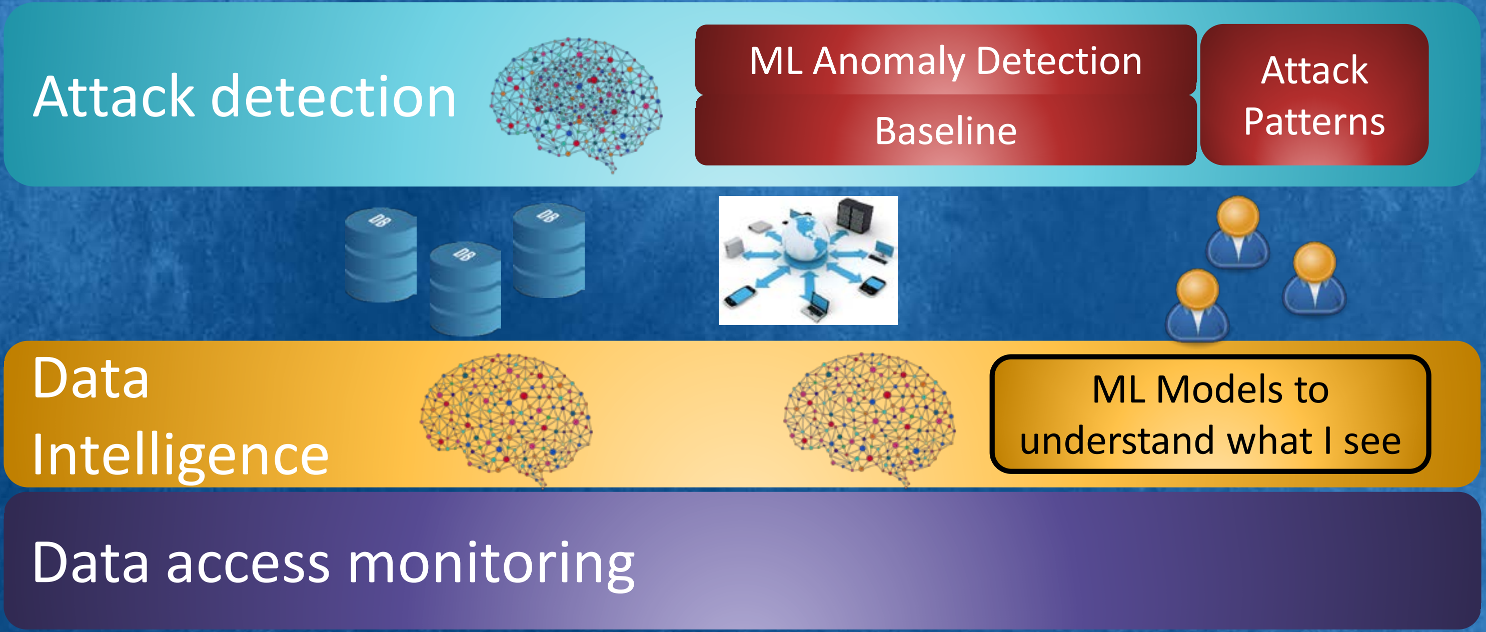
Harnessing AI for Cyber Security

Attack Detection in Imperva CounterBreach

IMPERVA

#RSAC

Imperva CounterBreach



Attack detection

ML Anomaly Detection

Baseline

Attack Patterns

Data Intelligence

ML Models to understand what I see

Data access monitoring

Summary

- AI is awesome technology
- AI fails in adversarial settings
- AI Deception gets only little industry attention
- The threats are acute and critical
- Mitigation in many cases is not trivial

Apply What You Have Learned Today

- Next week you should:
  - Identify critical AI usages within your organization and your roadmap
- In the next three months you should:
  - Carry out security modeling for AI usages (threats, adversaries, attack vectors)
  - In case there is significant threat and viable attack vectors, build mitigation plan. Focus on critical easy-to-exploit vulnerabilities.
- Within six months you should:
  - Execute at least the critical part of mitigation plan