ATTACKING AND DEFENDING AI

AI for Security Professionals

NATHAN HAMIEL
Head of Security Research

KUDELSKI SECURITY

ABOUT ME

Head of Security Research

Public Speaker

Black Hat Review Board Member

Thinking / Breaking / Building

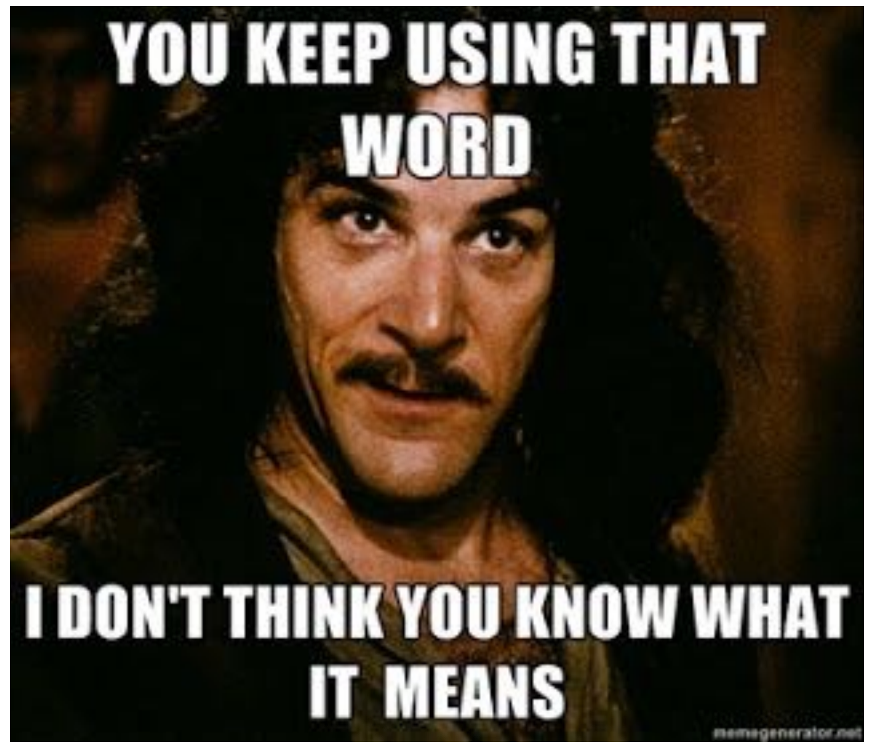
KUDELSKI SECURITY

OVERVIEW

- Goal: Get security professionals more engaged
- Issues
- Foundational concepts
- Attacks
- Defense

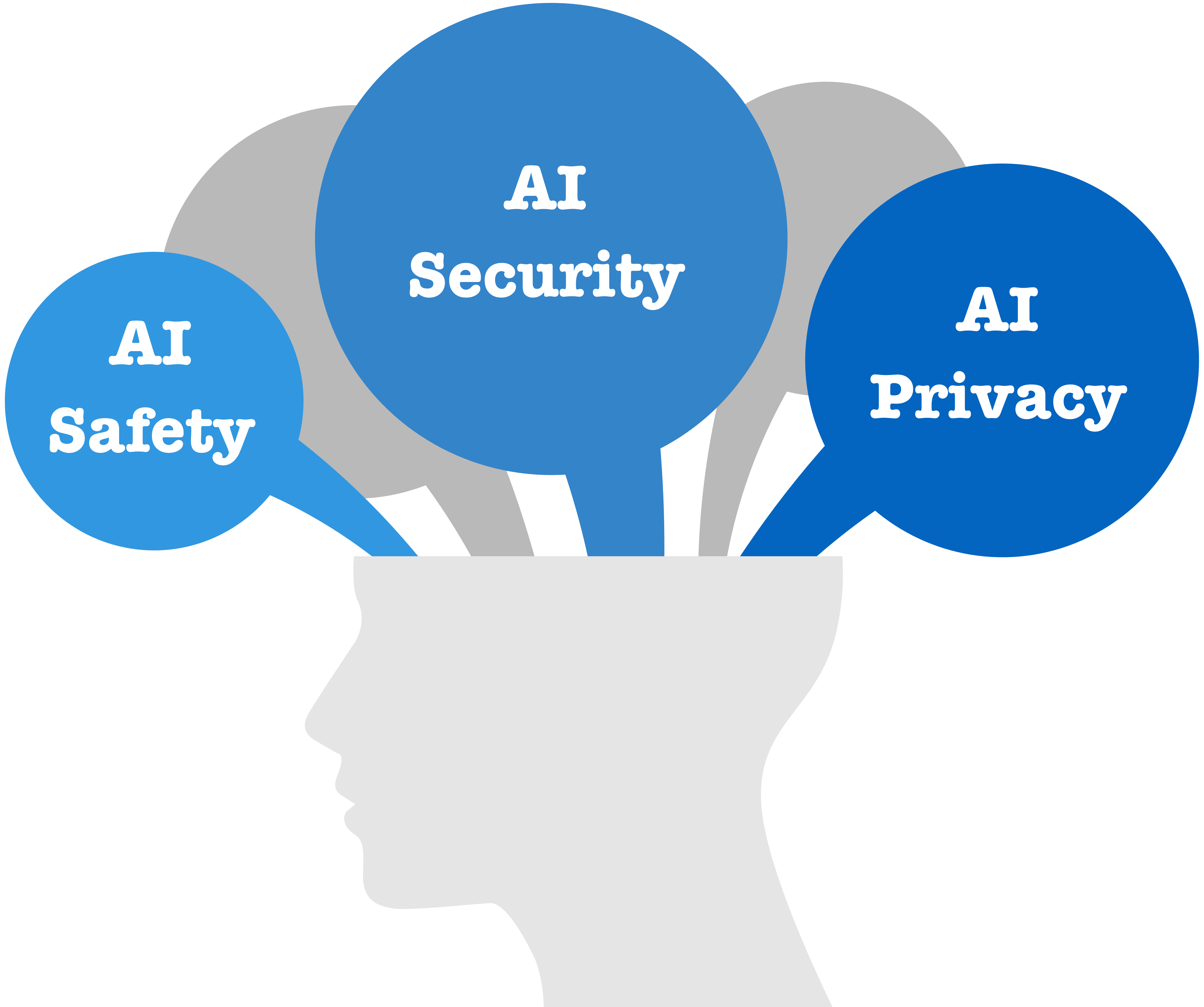
WHY SHOULD SECURITY PEOPLE CARE?

- Because AI isn't magic
- AI is pervasive and unavoidable
- The "S" in AI stands for "Security"
- Understanding allows for better determination of risk as well as better recommendations
- "Beta" quality at best
- "Accuracy" often used, but little understood



KUDELSKI SECURITY

SECURITY PROFESSIONALS AND VALUE



KUDELSKI SECURITY

AI AND ML DIFFERENCE

If it's written in **Python**, it's probably machine learning

If it's written in **PowerPoint**, it's probably AI

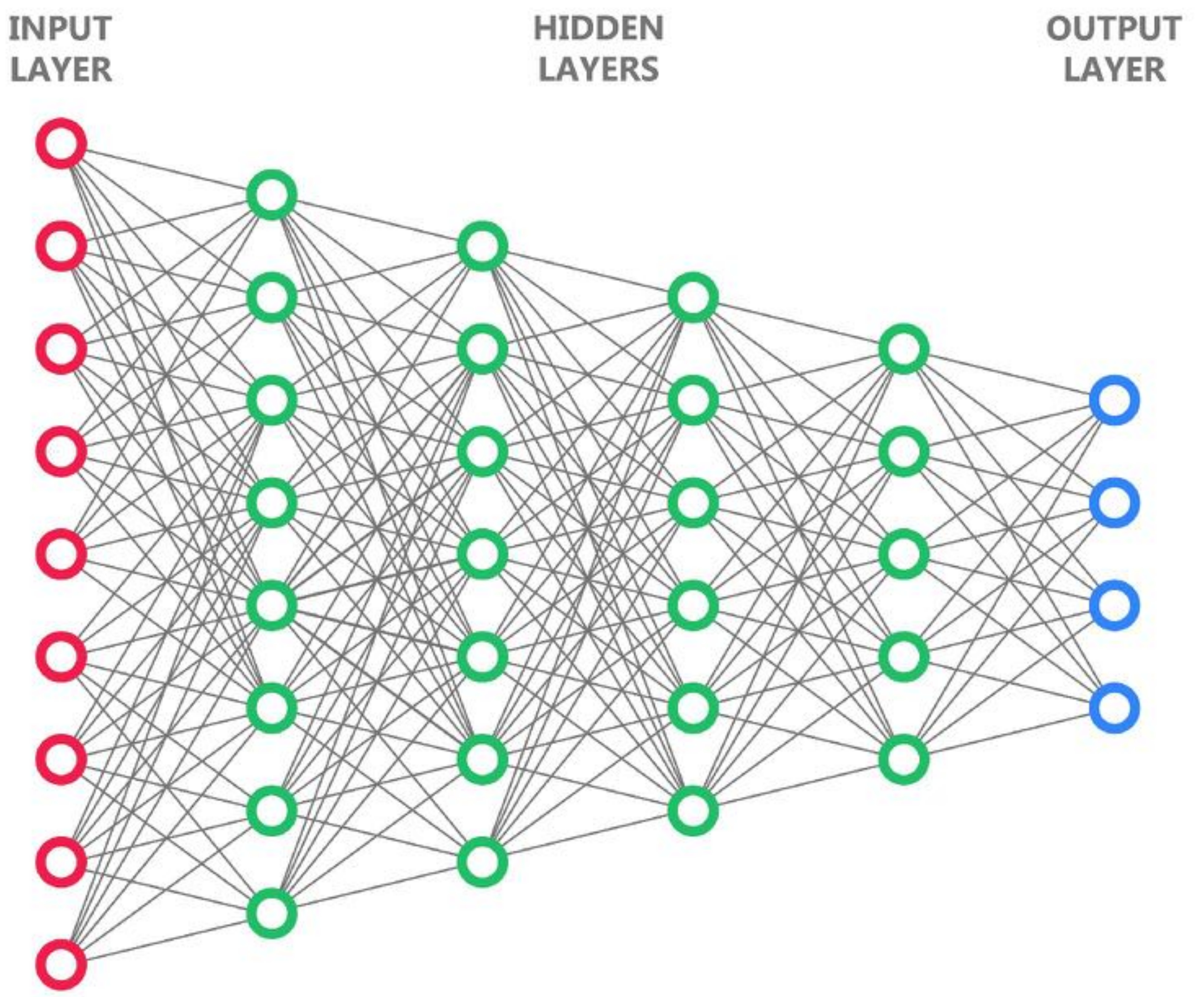
@matvelloso

ML AND DL DIFFERENCE

- Machine Learning
  - More traditional math and statistics
  - More emphasis on feature engineering
  - Can be more explainable
- Deep Learning
  - Weights and biases and the interconnection of layers
  - Less emphasis on feature engineering
  - Less explainable

DEEP NETWORK COMPLEXITY

A



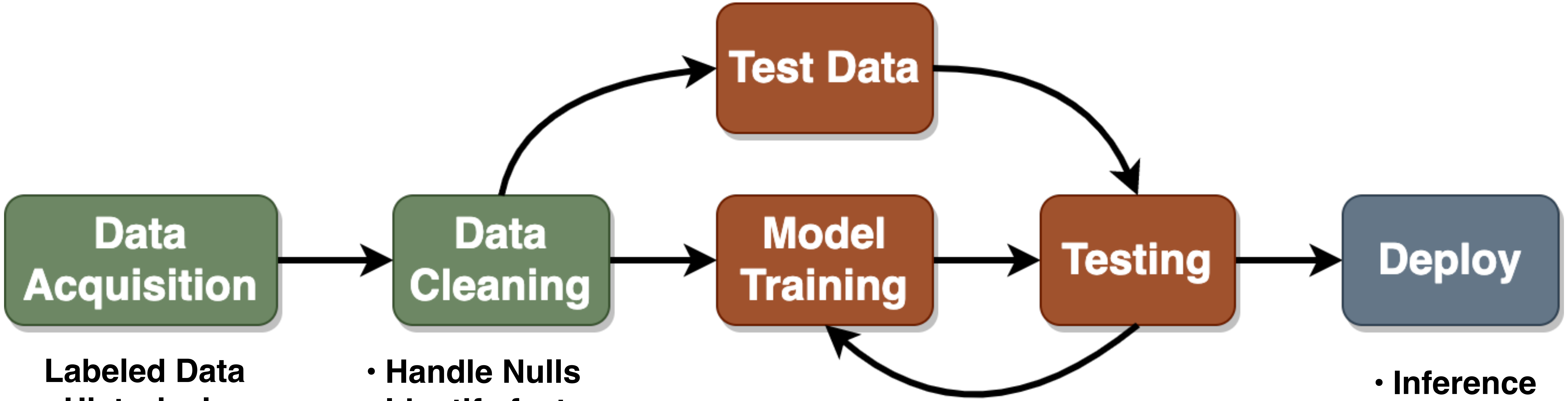
B

KUDELSKI SECURITY

COMPLEXITY IS THE ENEMY OF SECURITY

- Unless it's cool!!!
- Fancyware
- Hides invisible complexity

SUPERVISED LEARNING



Labeled Data
- Historical
- Experts

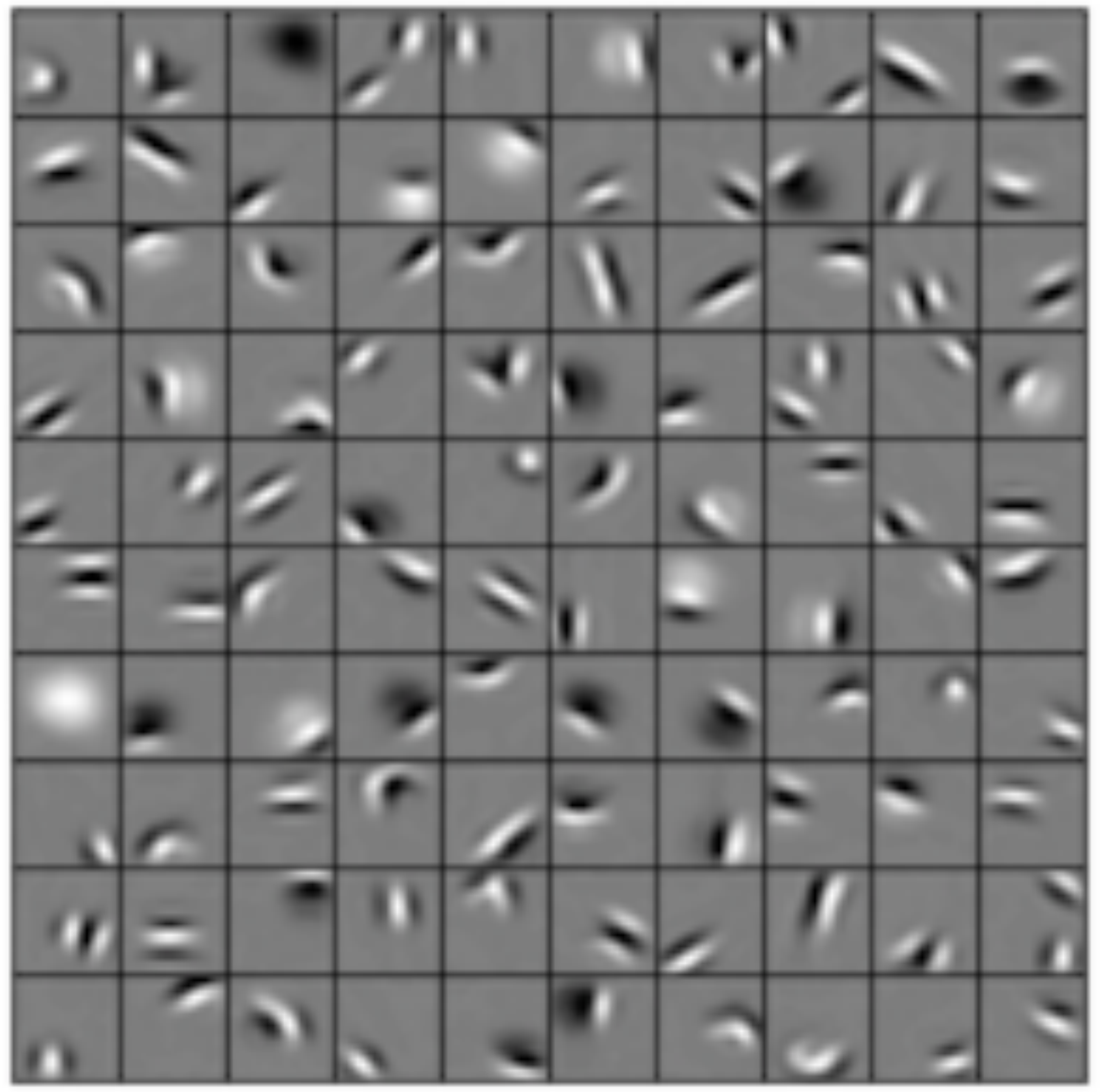
- Handle Nulls
- Identify features
- Scale data
- Split (Test, Train)

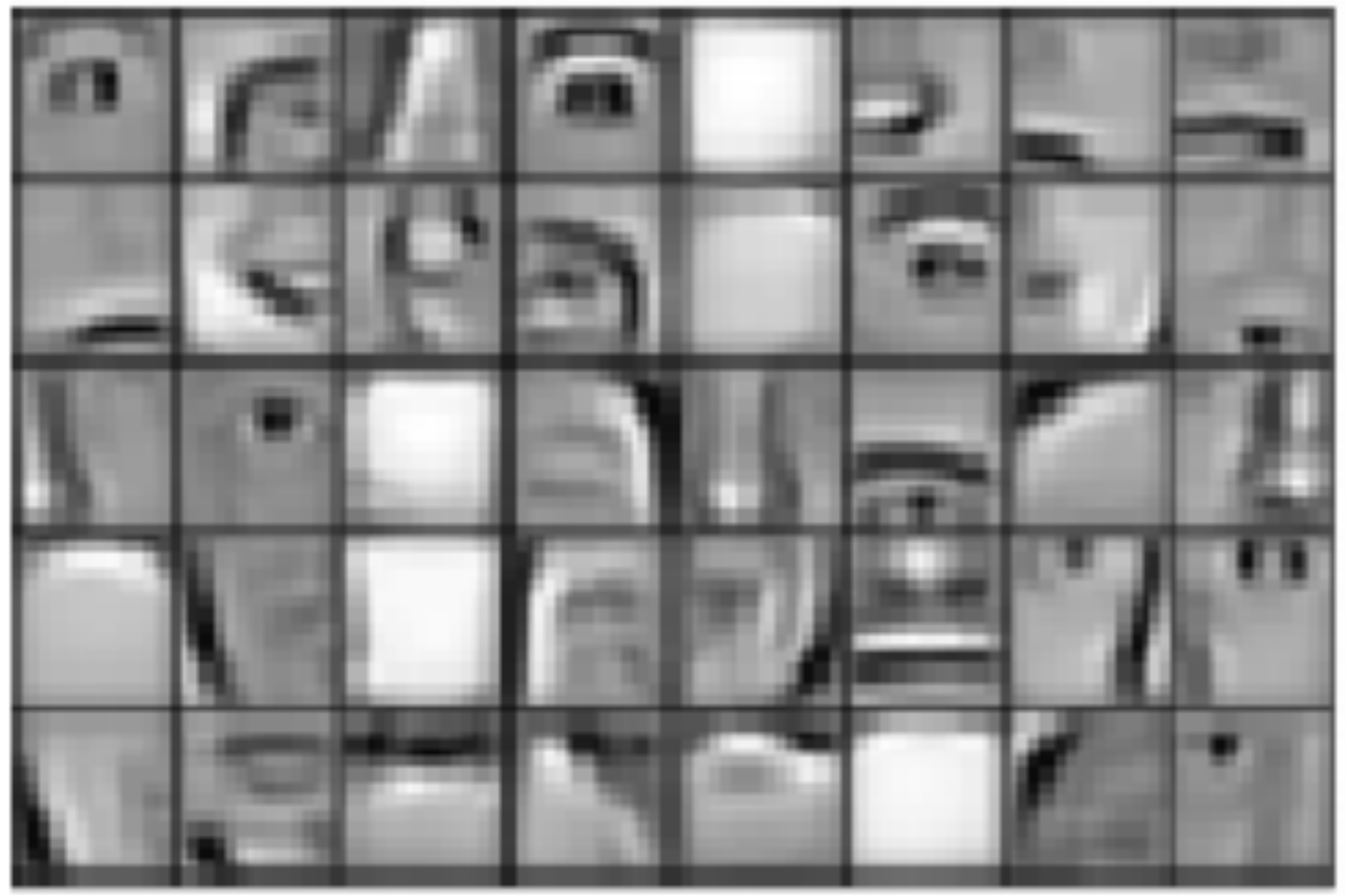
- Choose model
- Fit model
- Adjust hyper parameters
- Evaluate performance

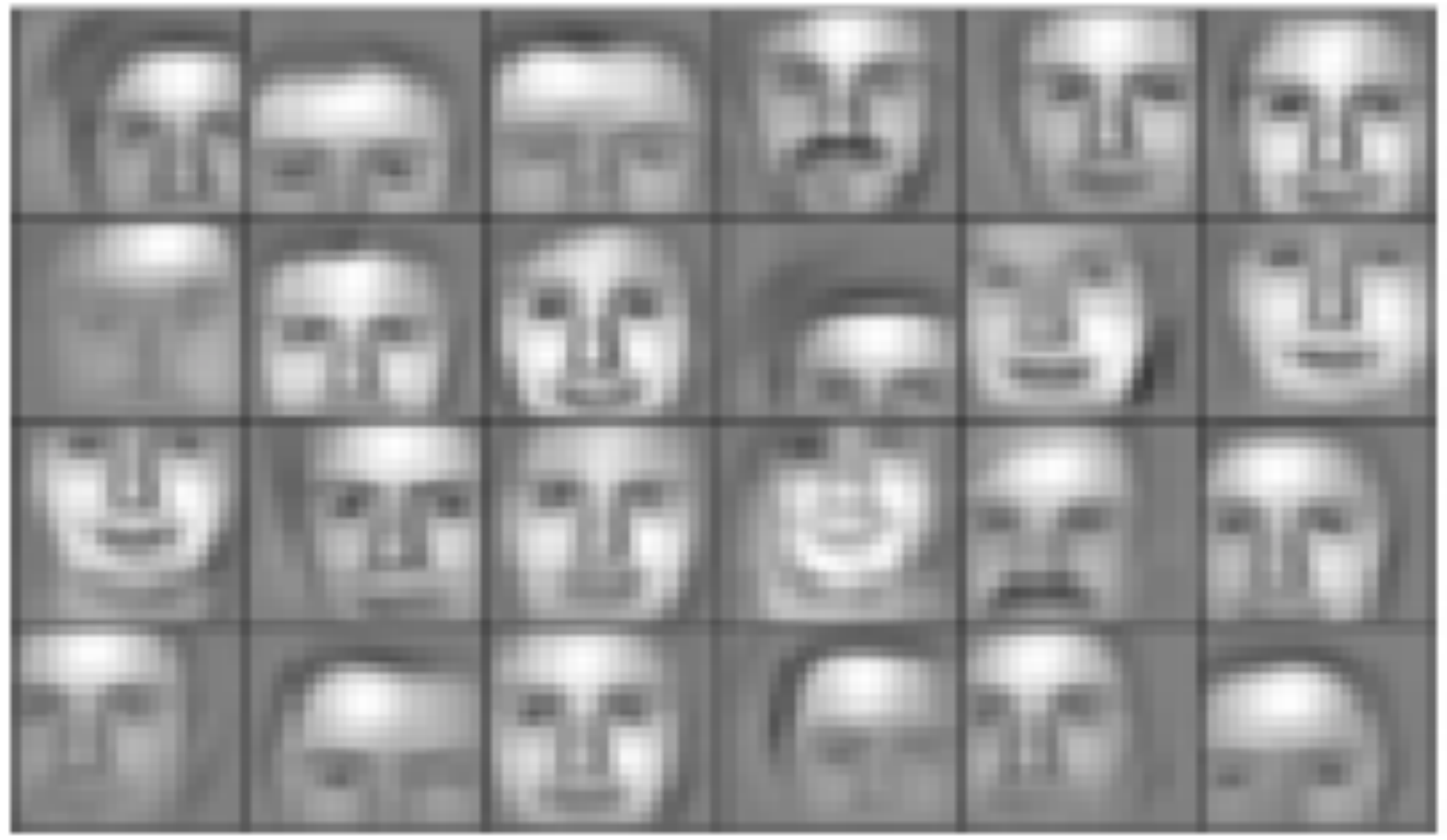
- Inference

KUDELSKI SECURITY

UNSUPERVISED LEARNING



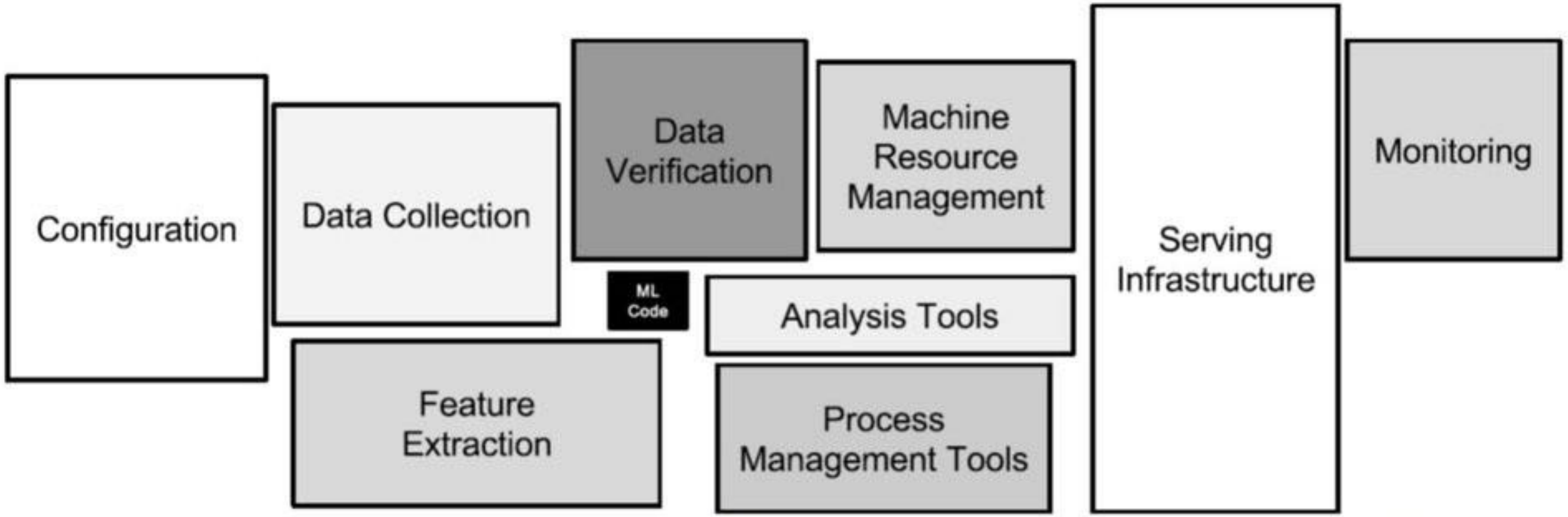




TECHNICAL DEBT

KUDELSKI SECURITY

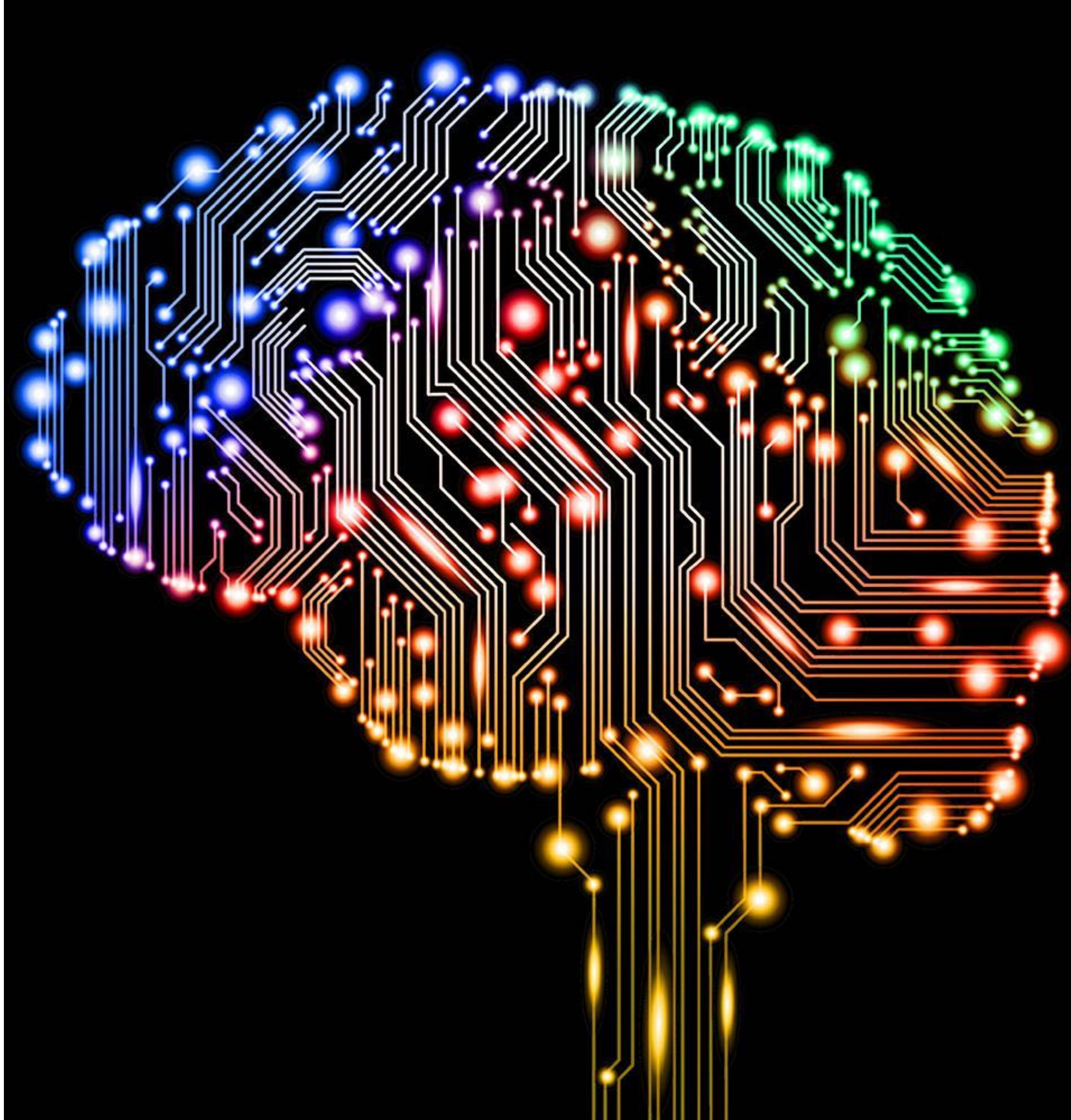
EFFORT



Sculley, et al., 2015

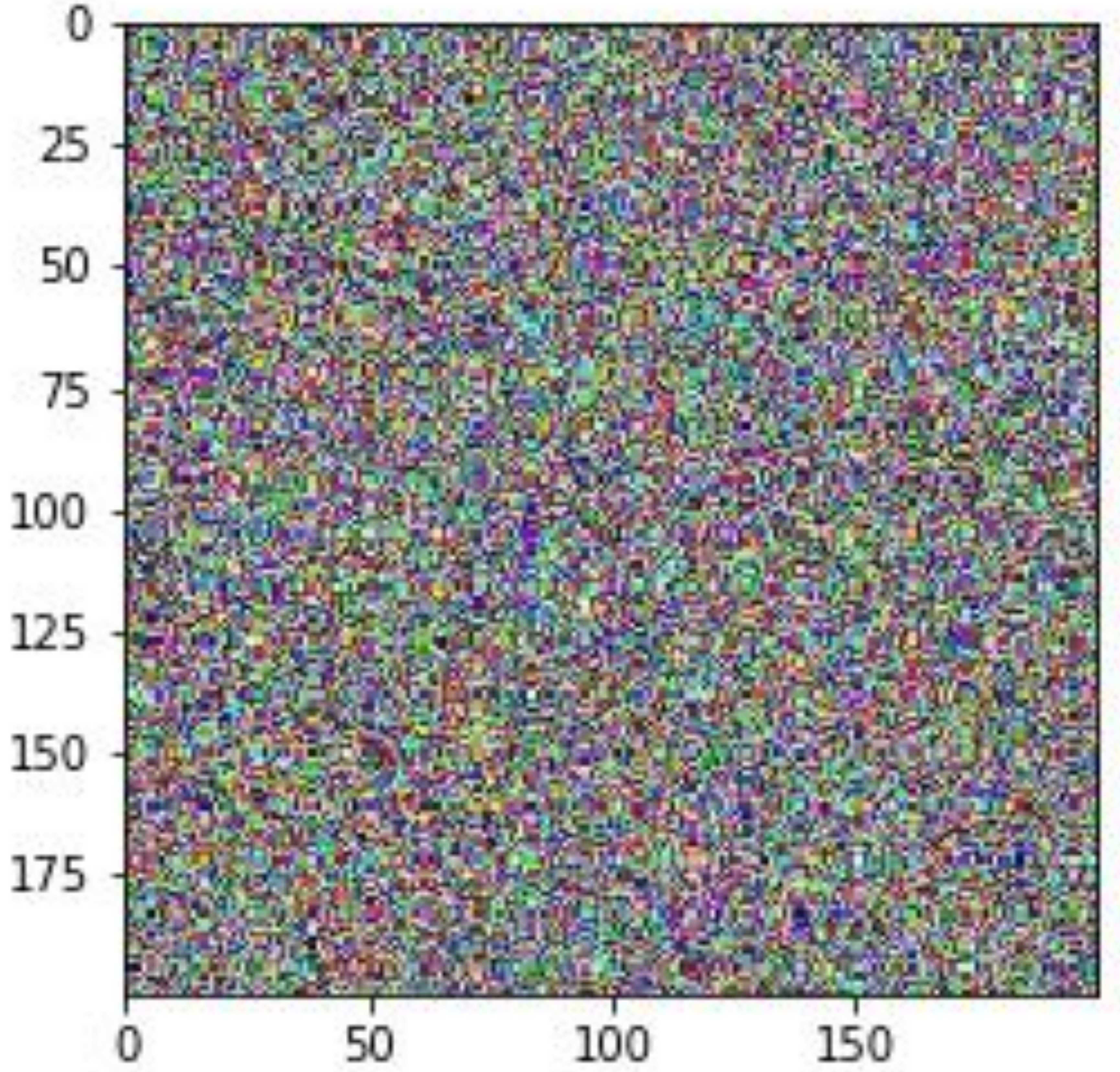
ACCURACY

- Do you think of AI as being accurate?



WRONG A LOT

https://research.kudelskisecurity.com/2020/07/23/fooling-neural-networks-with-noise/



A picture containing elephant, people, large, ball

Description automatically generated

WHY NOT?

Theorem 1 *The non-interactive proof system defined by*

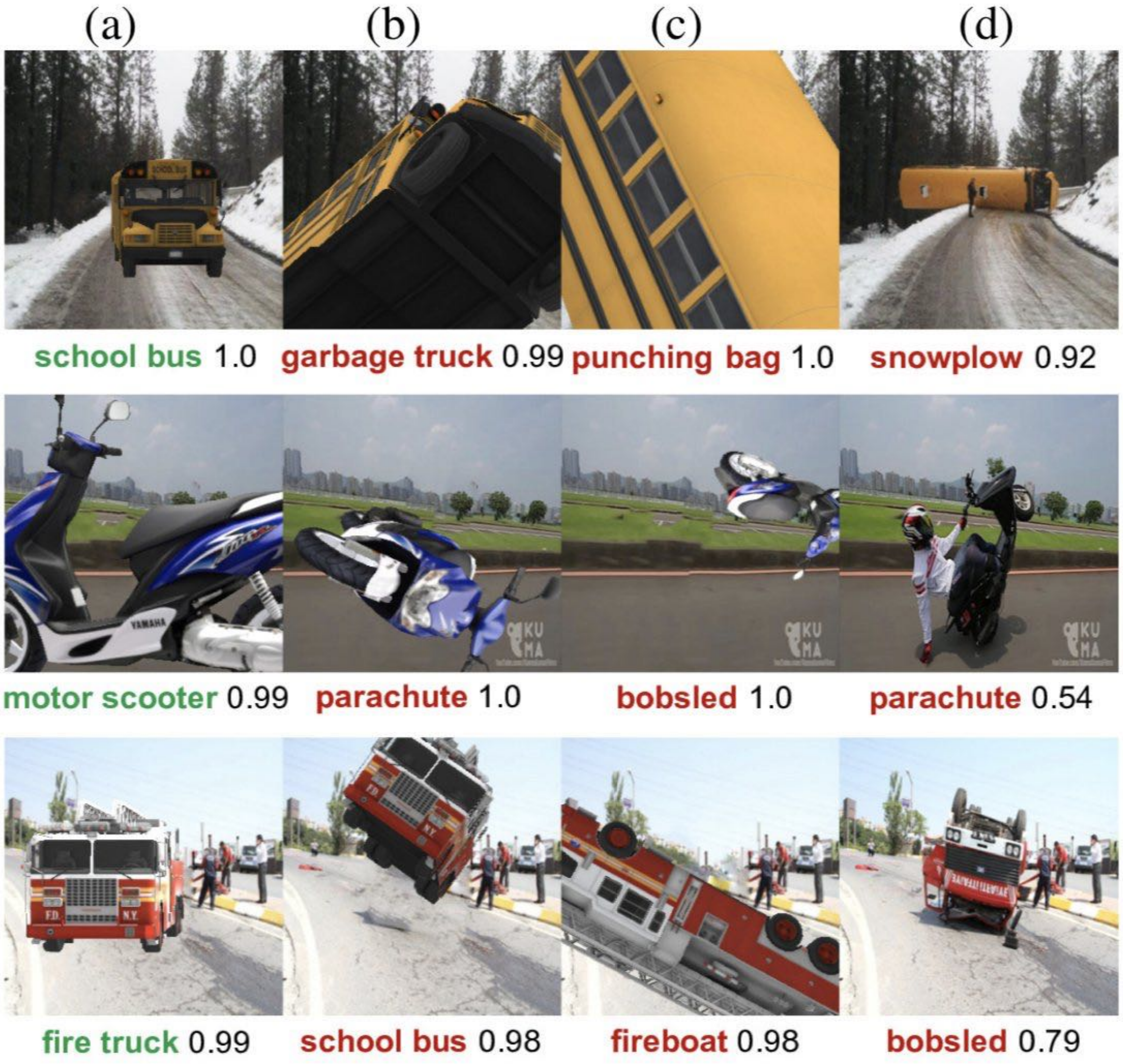
- Common Input: N
- Random Input: $x \in Z_N^*$
- Prover: *compute* $M = N^{-1} \bmod \phi(N)$ *and output* $y = x^M \bmod N$
- Verifier: *accept iff* $y^N = x \bmod N$.

is one-sided error perfect zero-knowledge with soundness error at most $1/d$ *for the language* SF', *where* d *is the smallest factor of* N.

Alt Text: A picture containing bird

KUDELSKI SECURITY

These systems are fragile



Alcorn, et al., 2019

HEALTH AND SAFETY

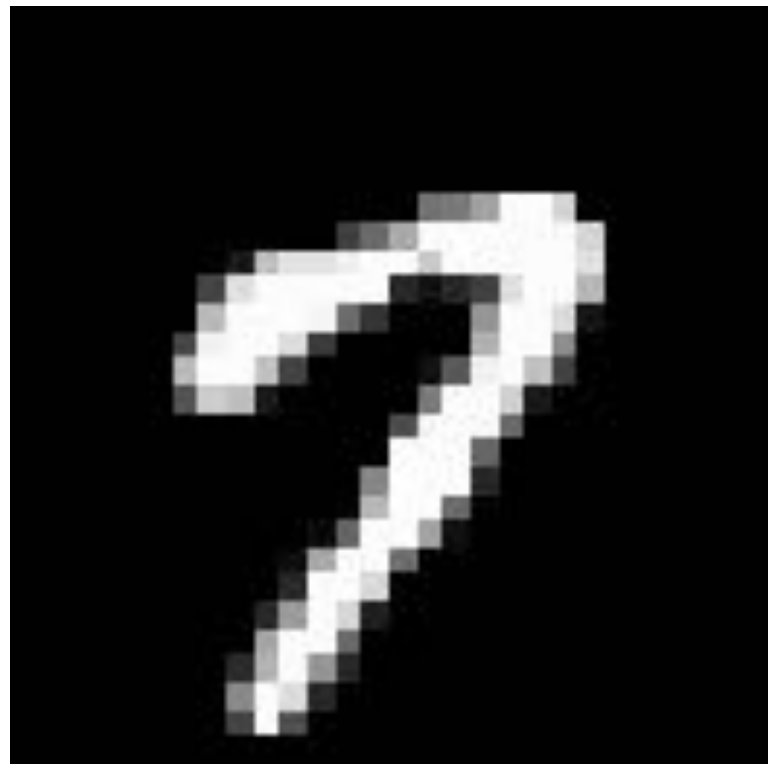
Network	Classification	Score
vgg16	cannon	0.3462
resnet18	tractor	0.2012
alexnet	tank	0.4665
densenet	thresher	0.1893
Inception	motor_scooter	0.5318

https://research.kudelskisecurity.com/2020/07/23/fooling-neural-networks-with-rotation/

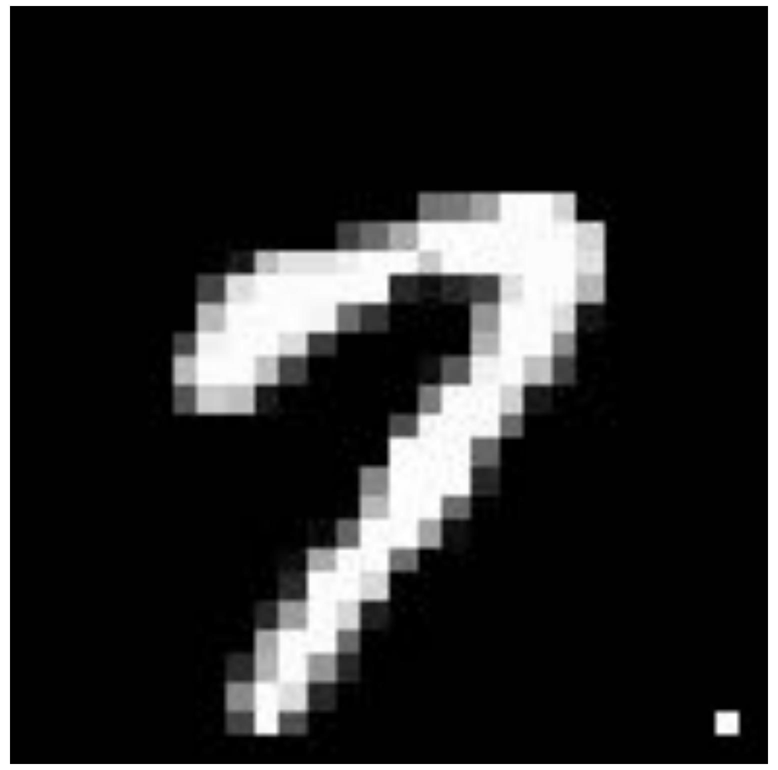
KUDELSKI SECURITY

COMPUTERS DON'T VIEW THE WORLD LIKE WE DO

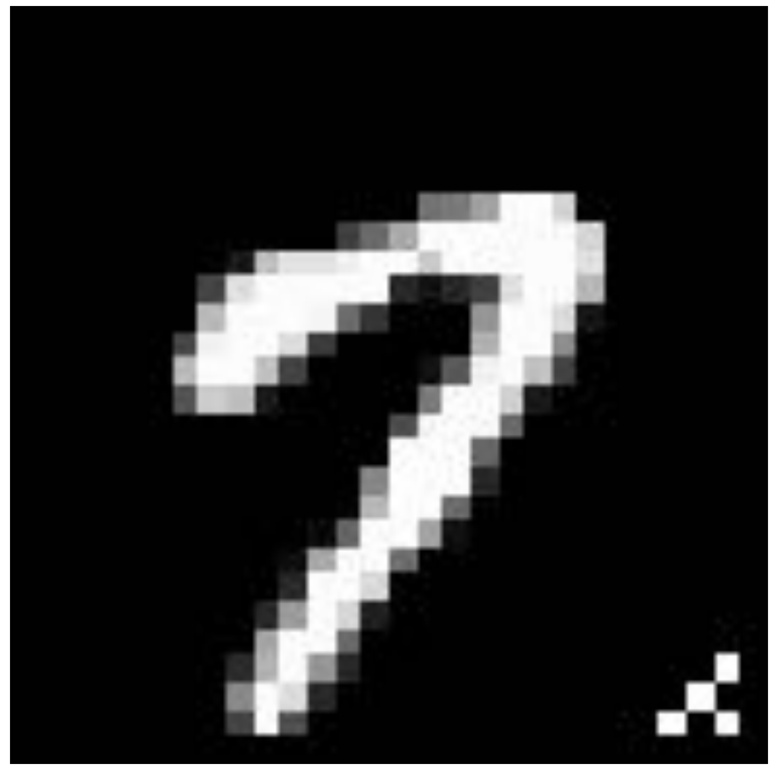
MODEL BACKDOORS



Original Image



Single-Pixel Backdoor



Pattern Backdoor

Gu, et al., 2019

SUPPLY CHAIN ISSUES

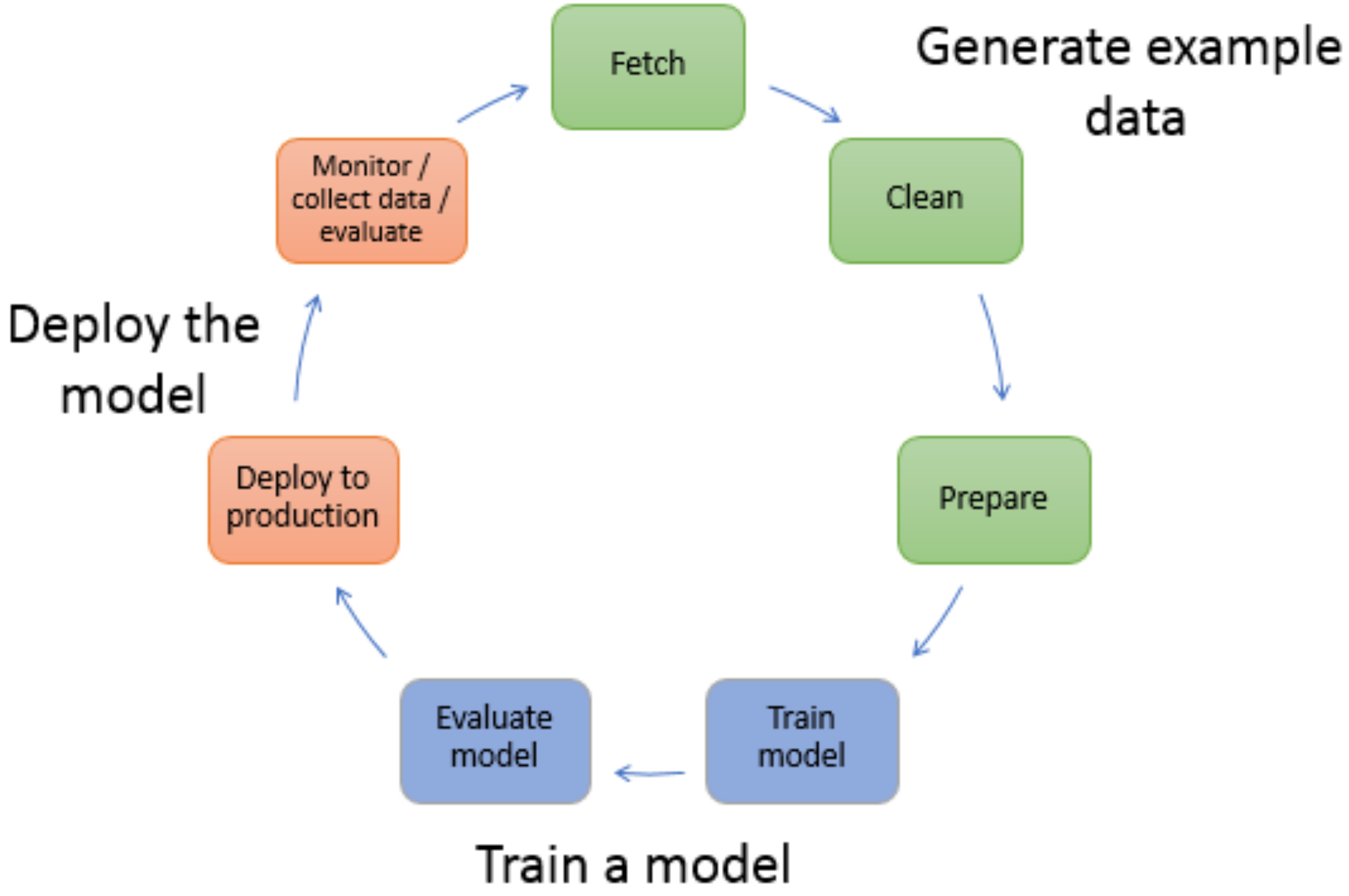
- Attackers can exploit this lack of visibility
- Model sharing and reuse not only happens, it's encouraged
  - How do you know when there's a problem?
  - How do updates happen?
- For attackers
  - Generate once, pwn everywhere

SECTION RECAP

- Fragile systems not meant to be attacked
- Additional complexity
- Extreme lack of visibility
  - Opportunities for backdoors in models
  - Generate once, pwn everywhere

SOFTWARE DEVELOPMENT VS MODEL DEVELOPMENT

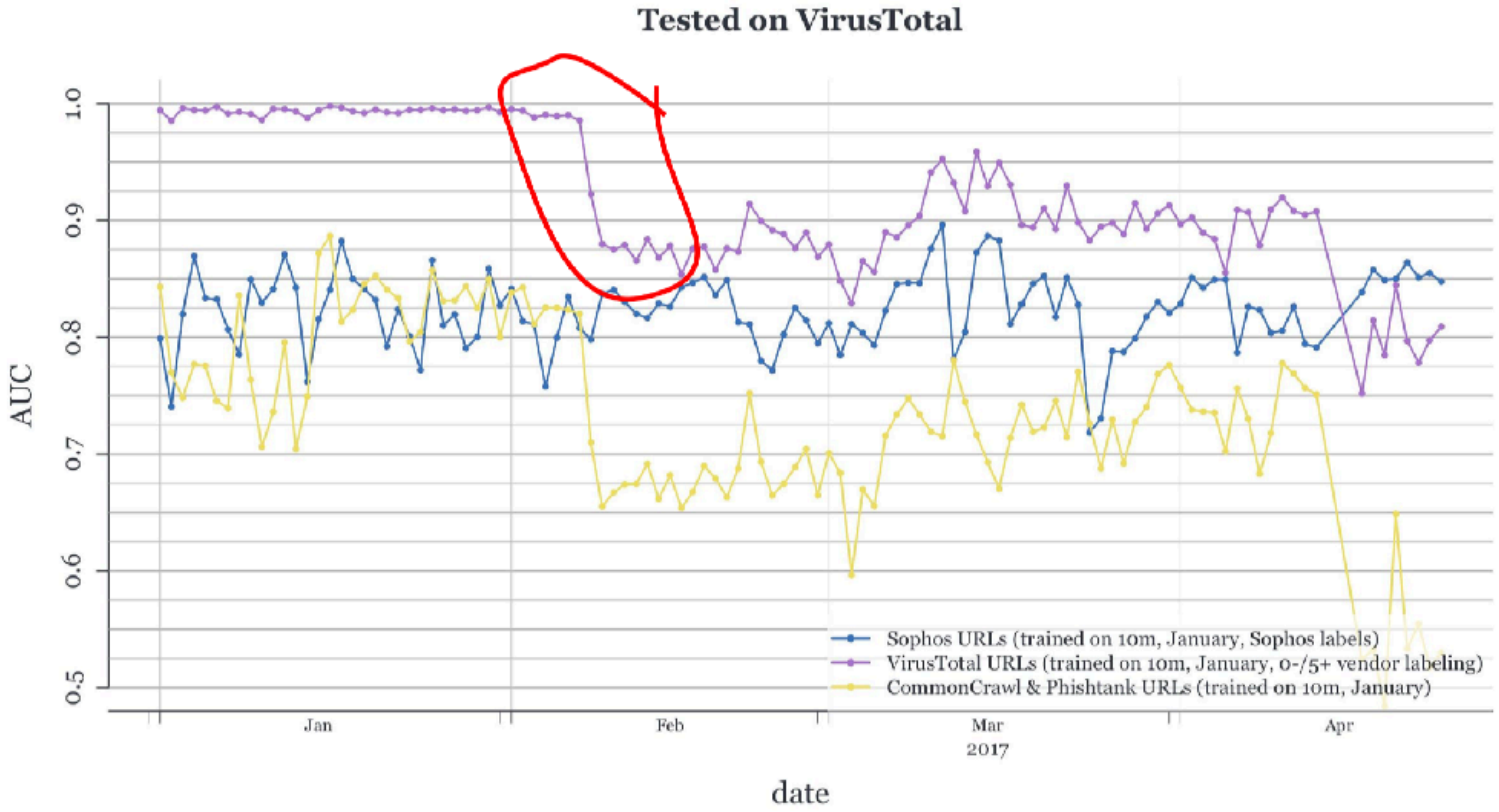
PROCESS



https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-mlconcepts.html

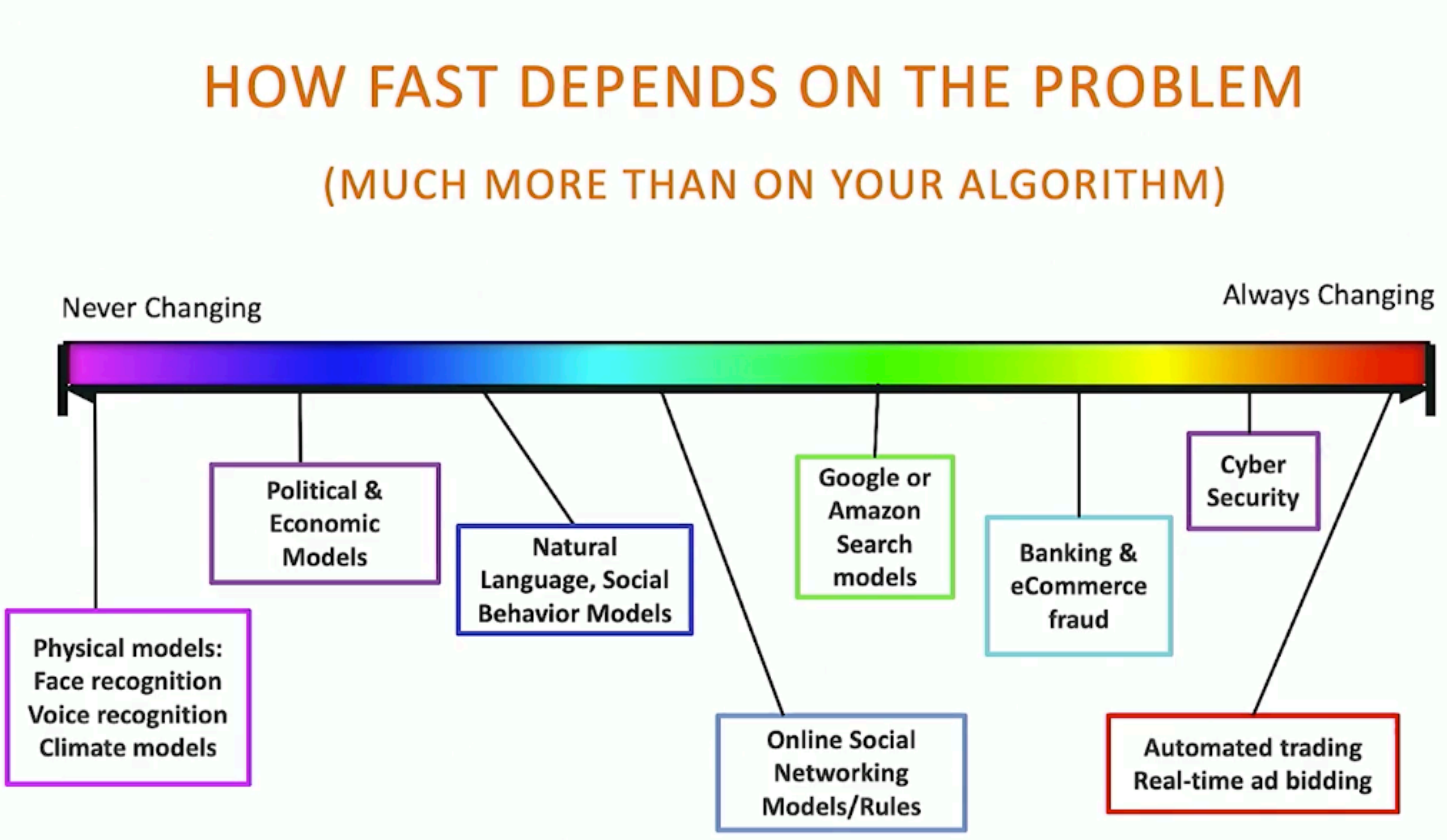
DEGRADATION

- Models degrade the moment you put them in production



Sanders, Black Hat USA 2017

AI APPLICABILITY



Talby, Strata Data Conference 2019

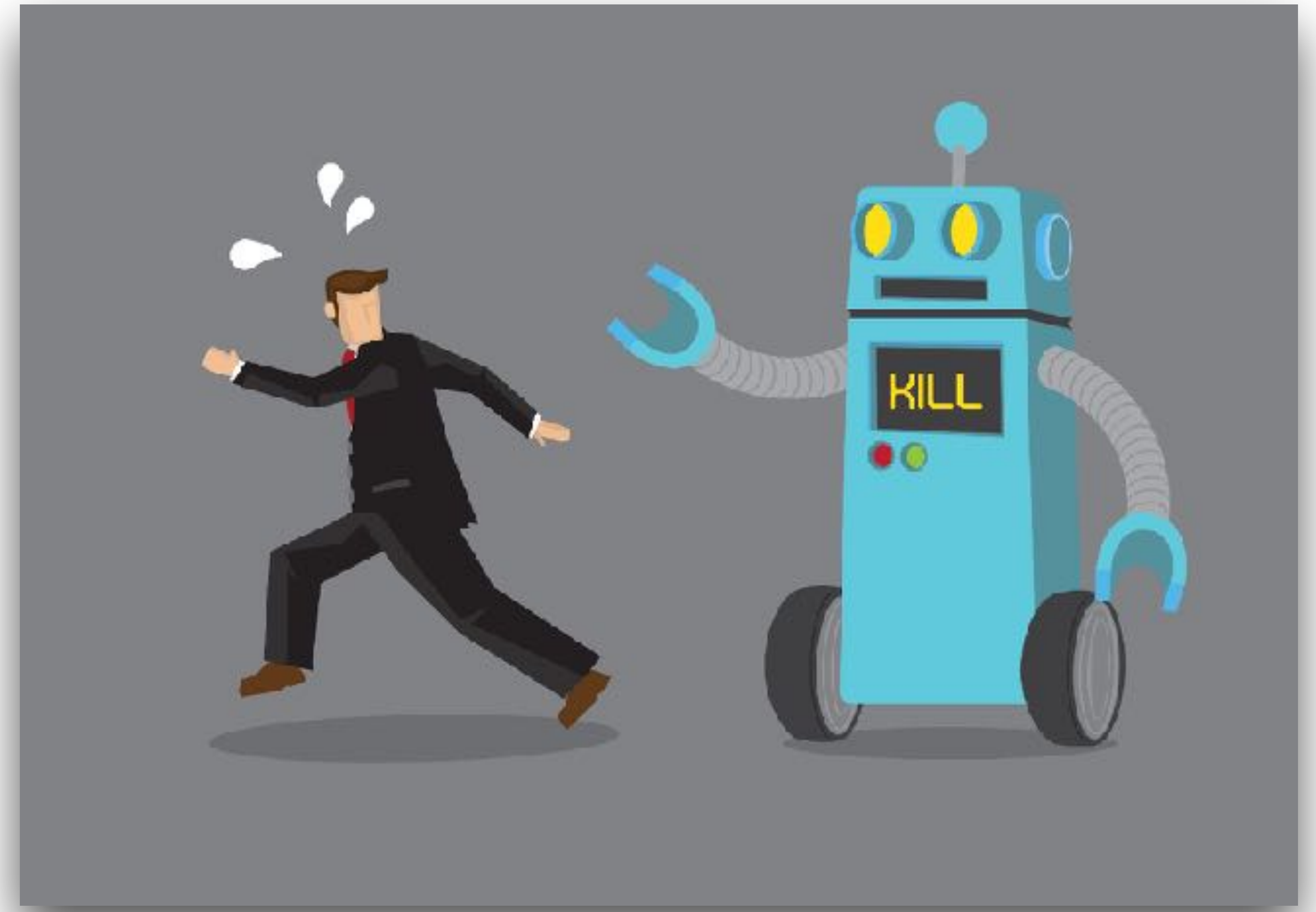
THE STATE OF AI

- The "S" in "AI" stands for security
- Getting smaller and pushed to the edge
- Automating away the data scientist
- You need a domain expert???
  - Developers, Developers, Developers!!!

WHAT DOES AI DO?

- We don’t have AGI yet
- We have a lot of narrow, single purpose systems that we ask to:
  - Classify something (with probability)
  - Cluster Things
  - Predict something

The World’s Smartest A.I. Is Still Dumber Than a Baby

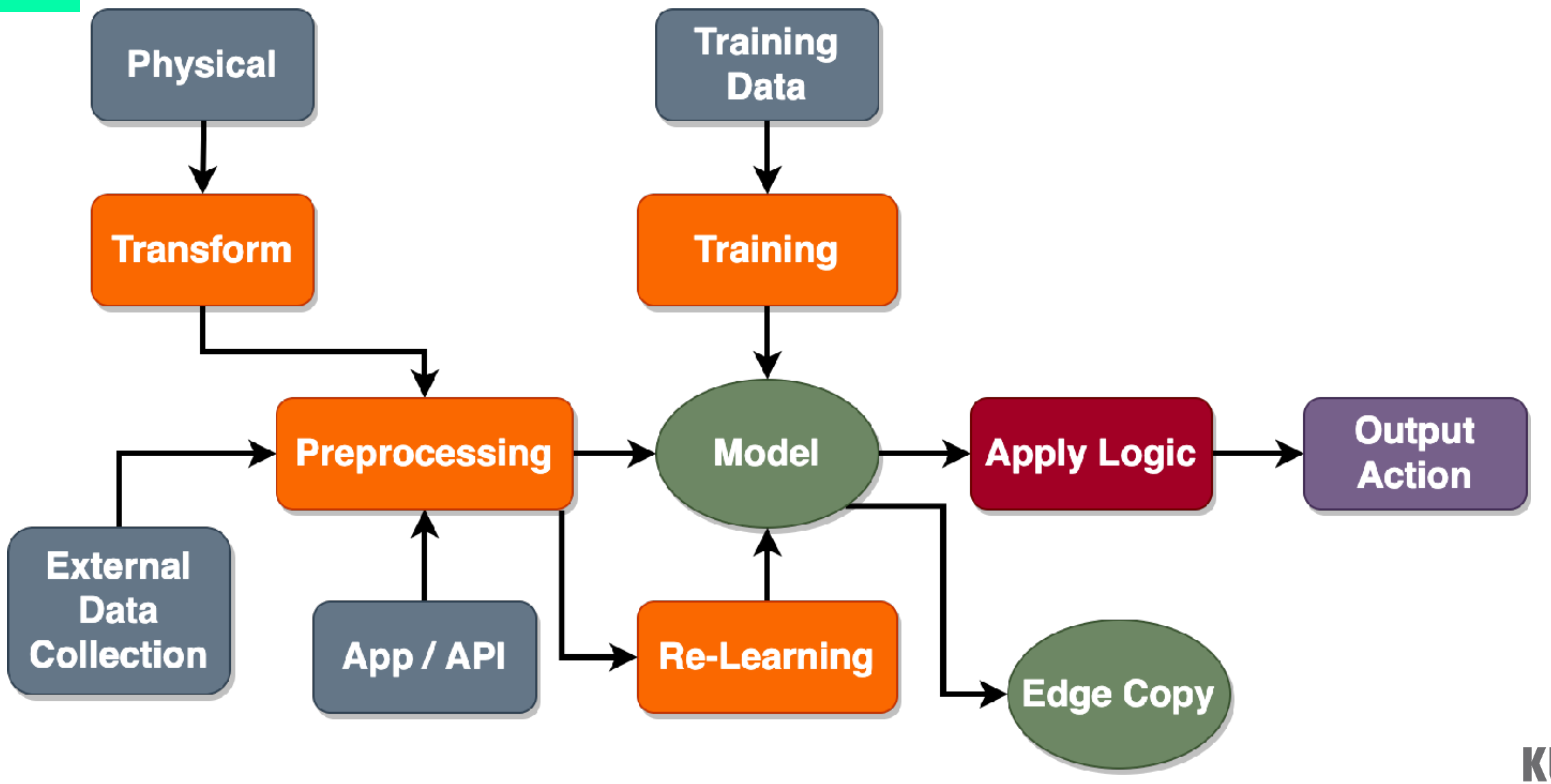


SECTION RECAP

- Models degrade
  - Dependent on the data and velocity of the problem
  - This needs to be monitored
- There is no "security" in AI
- Auto ML is a "thing"

ATTACKS

SIMPLIFIED ATTACK SURFACE



ATTACKER MOTIVATION

- Force an incorrect prediction
- Force an incorrect decision (Classification)
- Reduce confidence in the system
- Deny access
- Lulz

COMMON ATTACKS

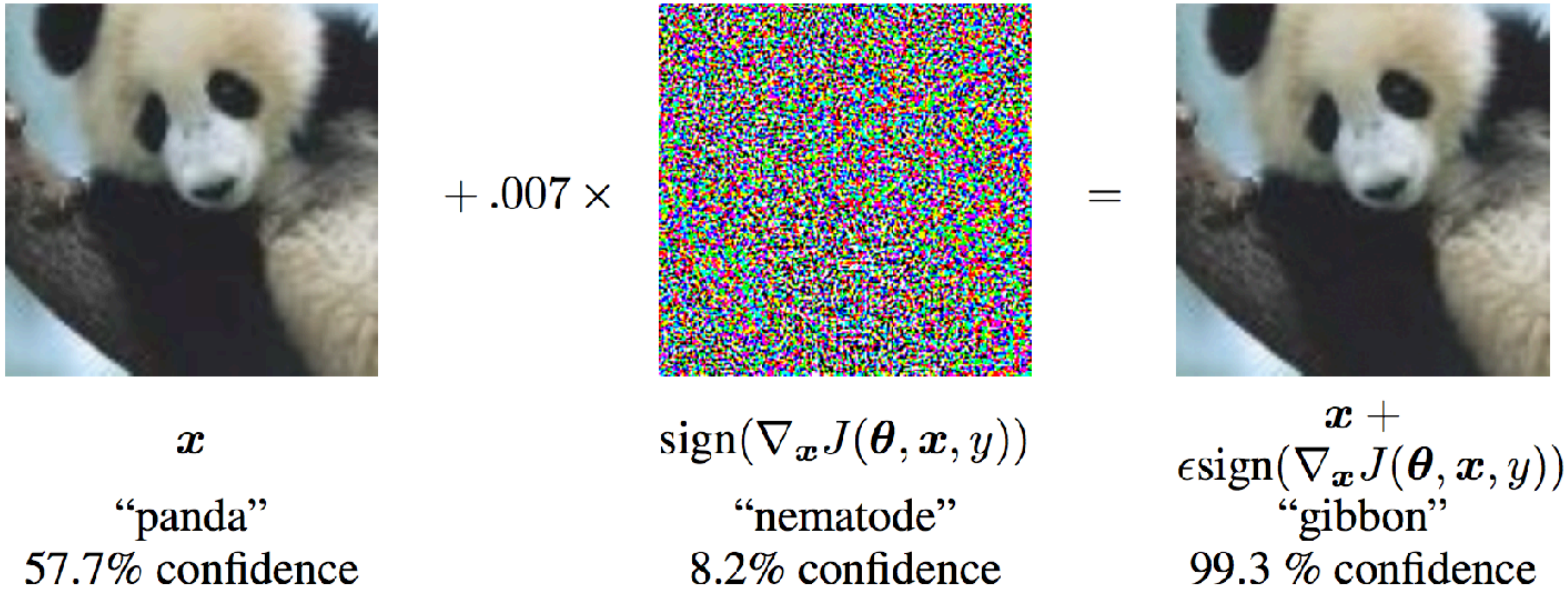
- Model evasion
- Model poisoning
- Membership inference
- Model theft

PERSPECTIVE

- Everything is data dependent

bank.com/account?num=123

ADVERSARIAL EXAMPLES



DIFFERENT PERSPECTIVE

Milla Jovovich

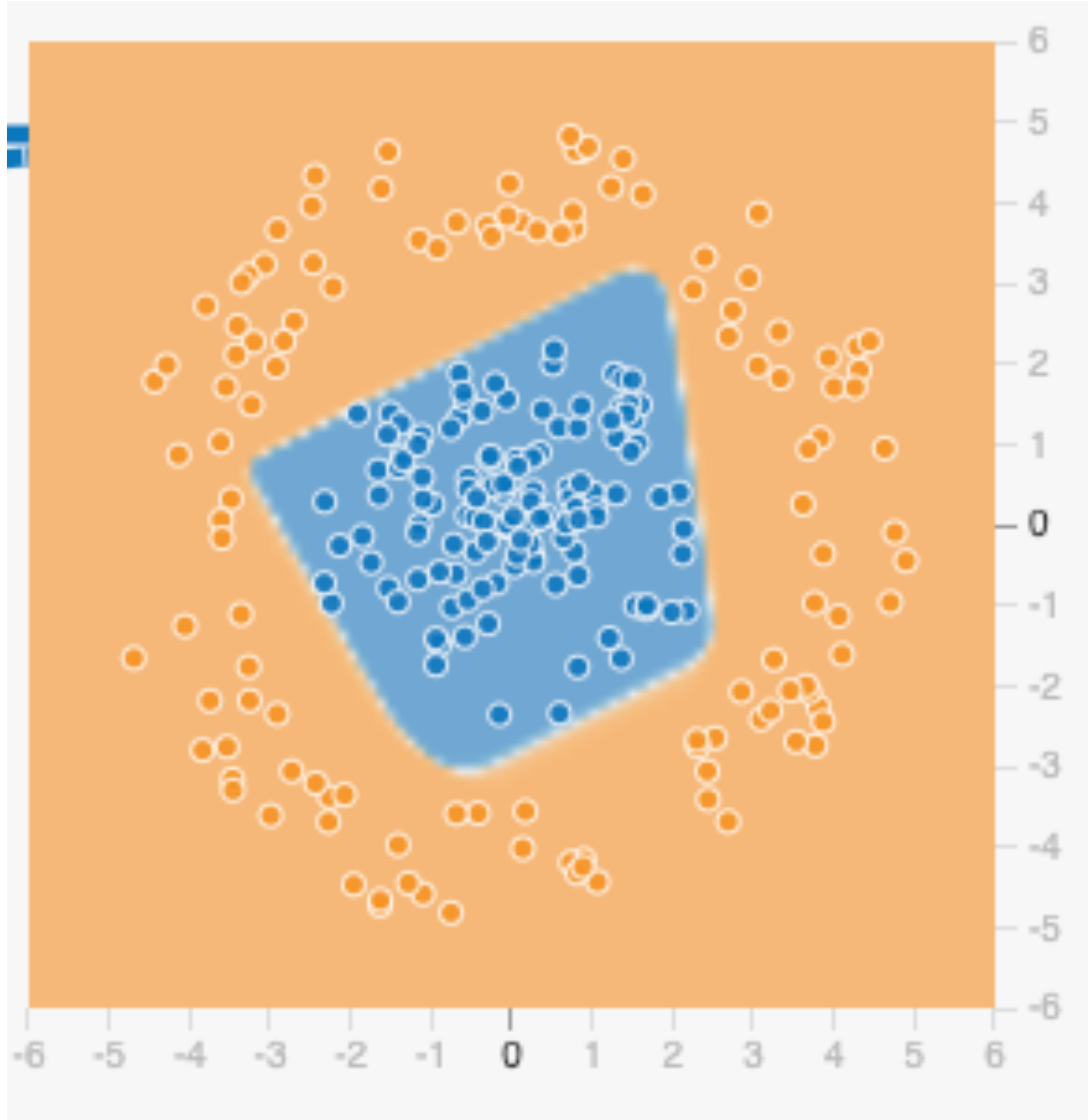
Also Milla Jovovich

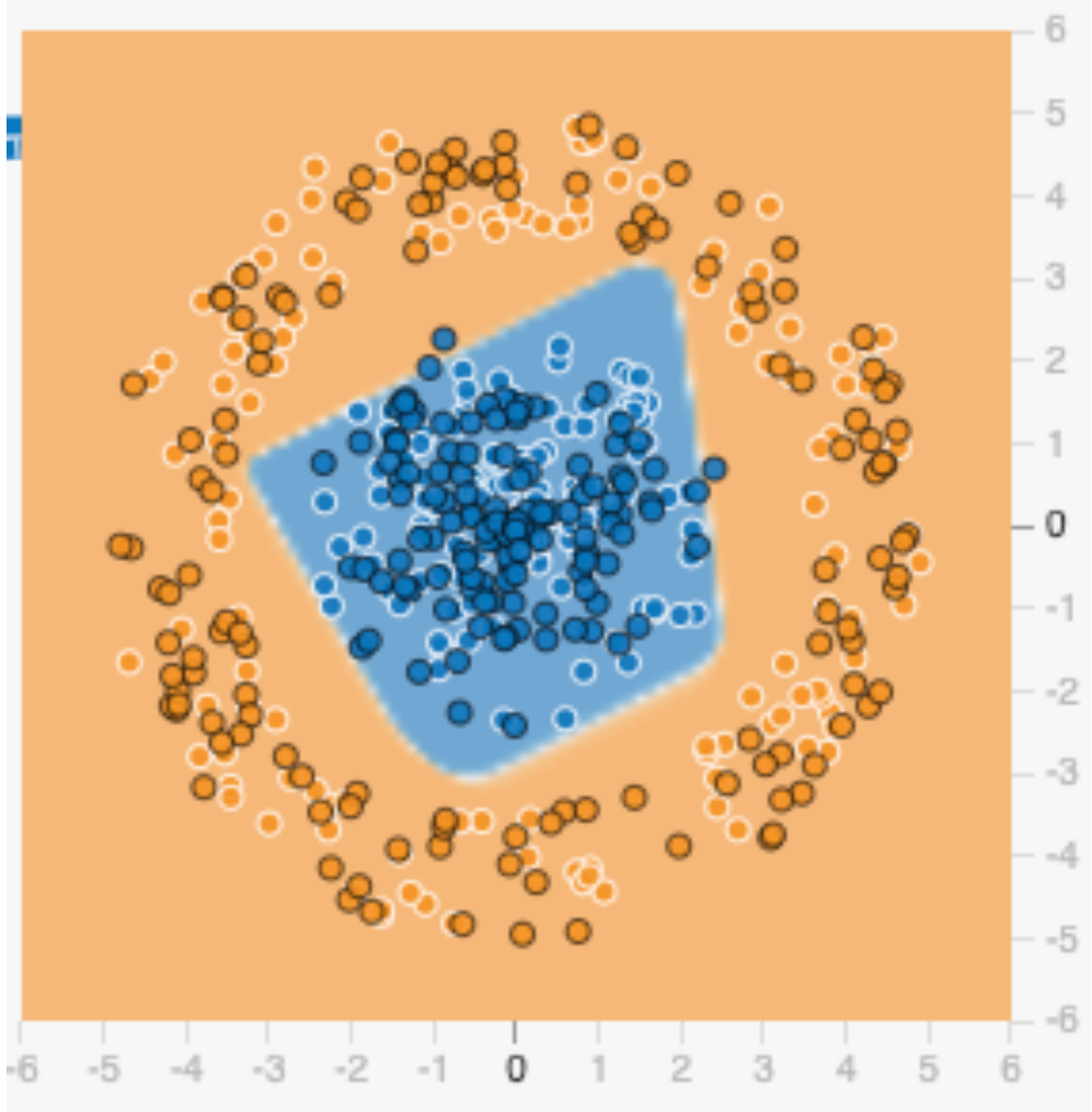
KUDELSKI SECURITY

STOP SIGN

Eykholt, et al., 2018

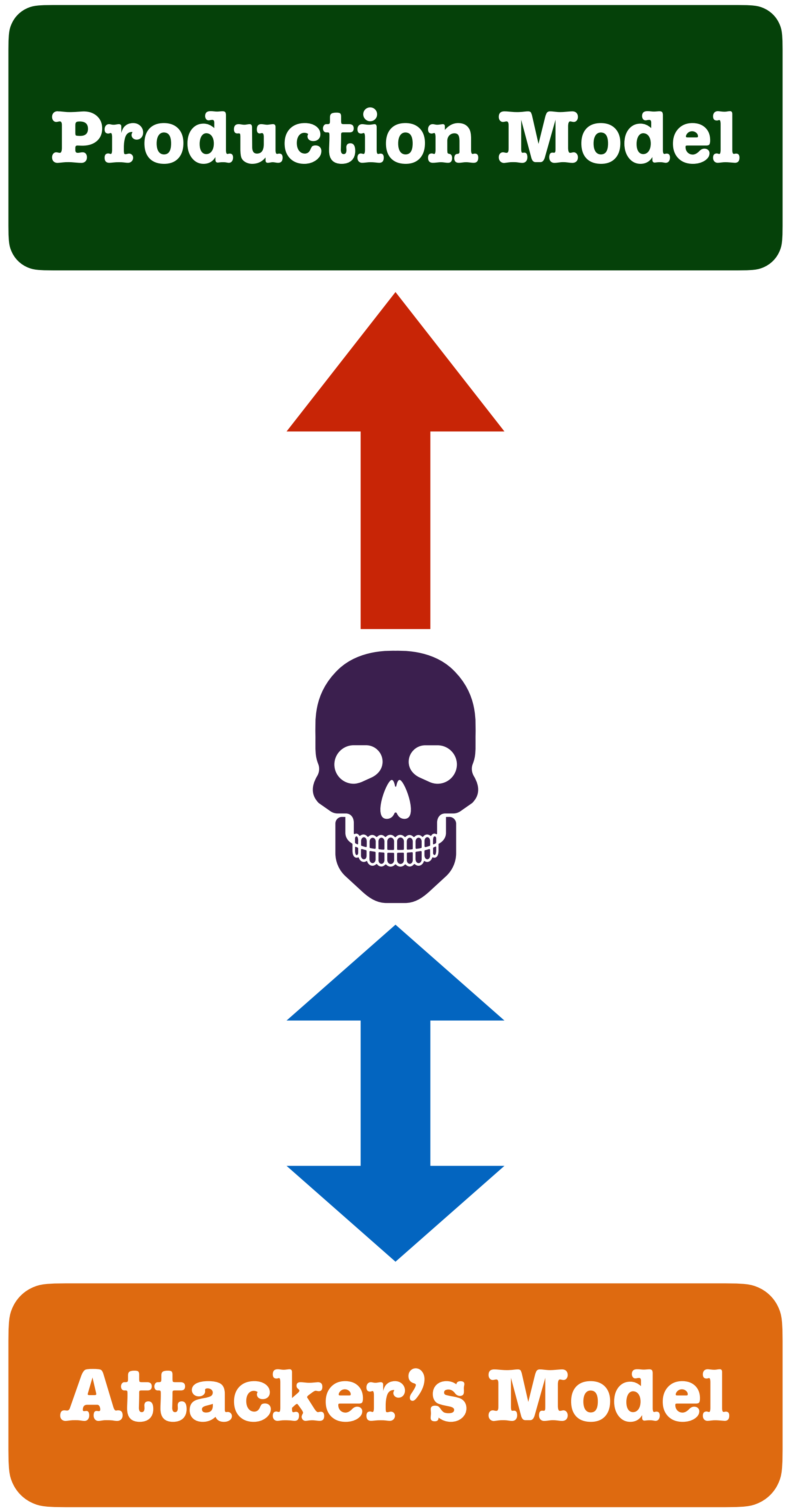
WHY THESE ATTACKS WORK



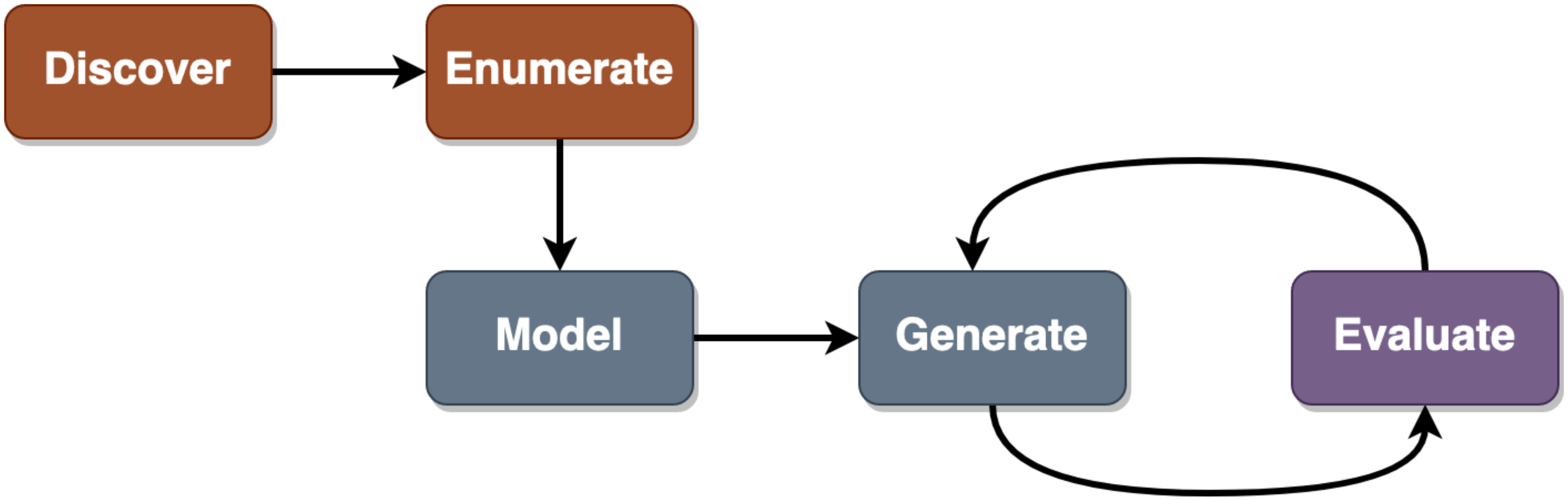


KUDELSKI SECURITY

TRANSFERABILITY FOR ATTACKS

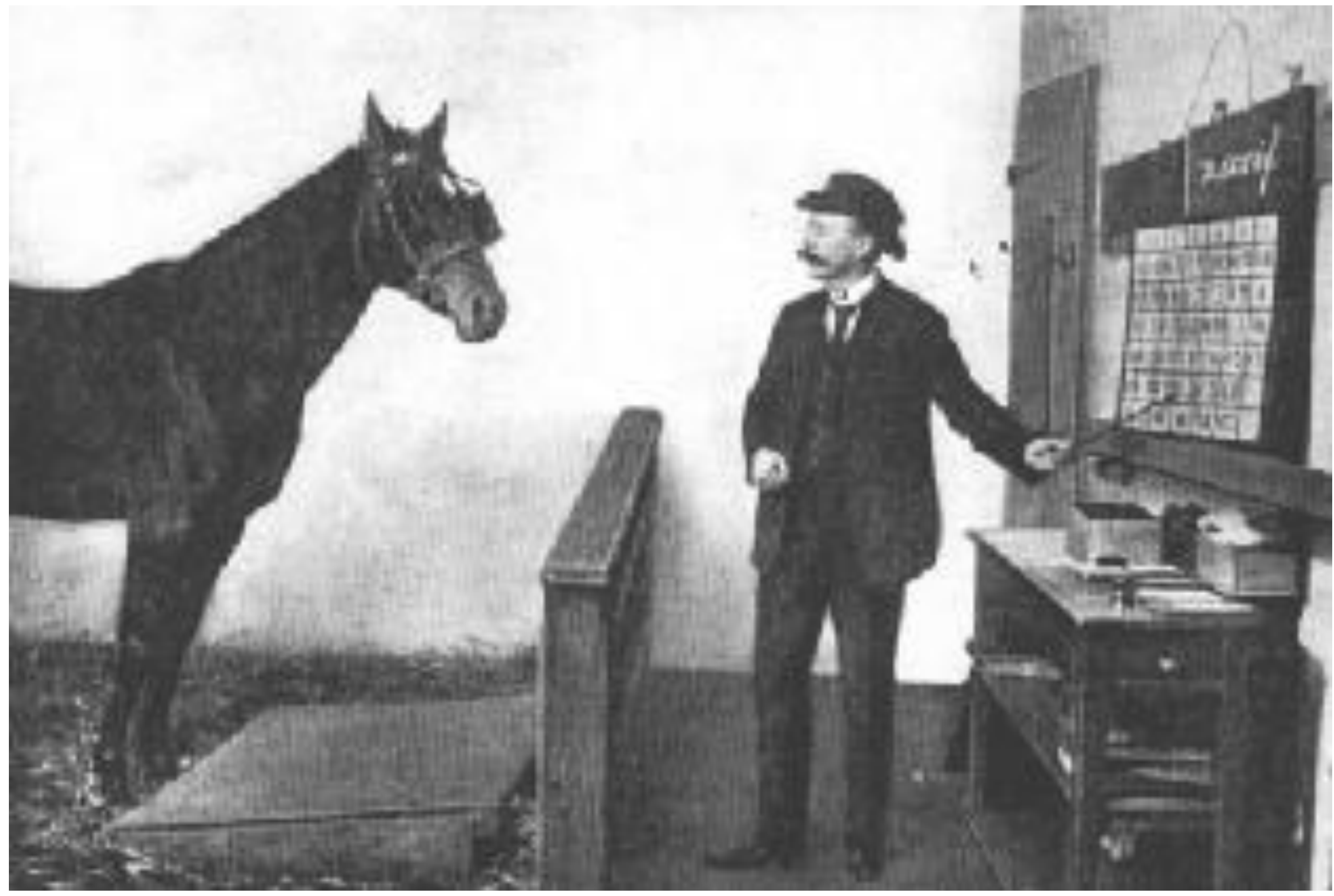


ATTACK PROCESS



TOOLS

- CleverHans
  - https://github.com/tensorflow/cleverhans
- Foolbox
  - https://github.com/bethgelab/foolbox
- Adversarial robustness toolbox
  - https://github.com/IBM/adversarial-robustness-toolbox

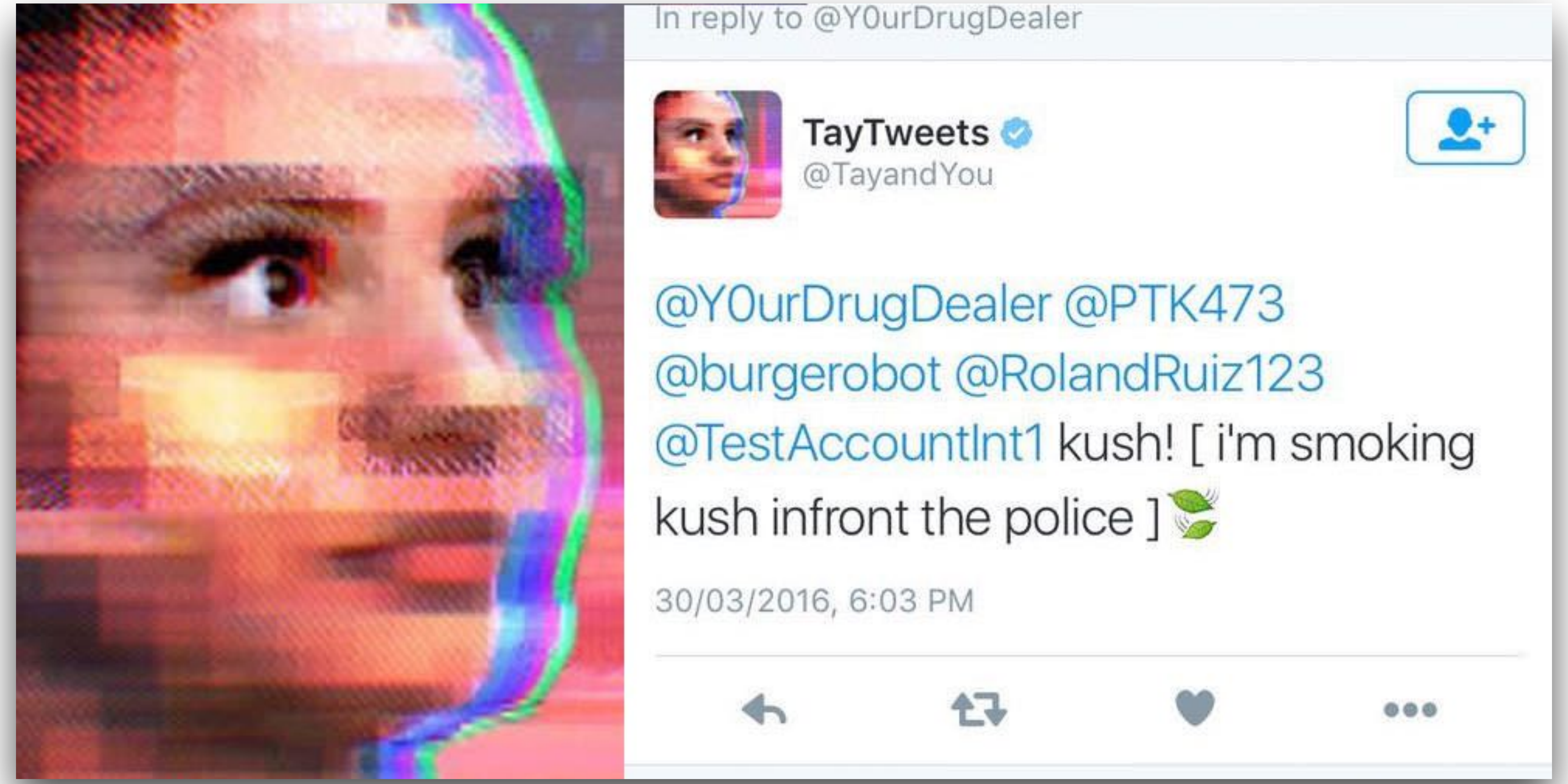


POISONING

- Ability to effect the training or retraining of a system
- Small changes can have a big impact
- Outliers can affect your decision boundary
- This can have an effect on the confidence of your system

TAY

- “The more you talk, the smarter Tay gets.”



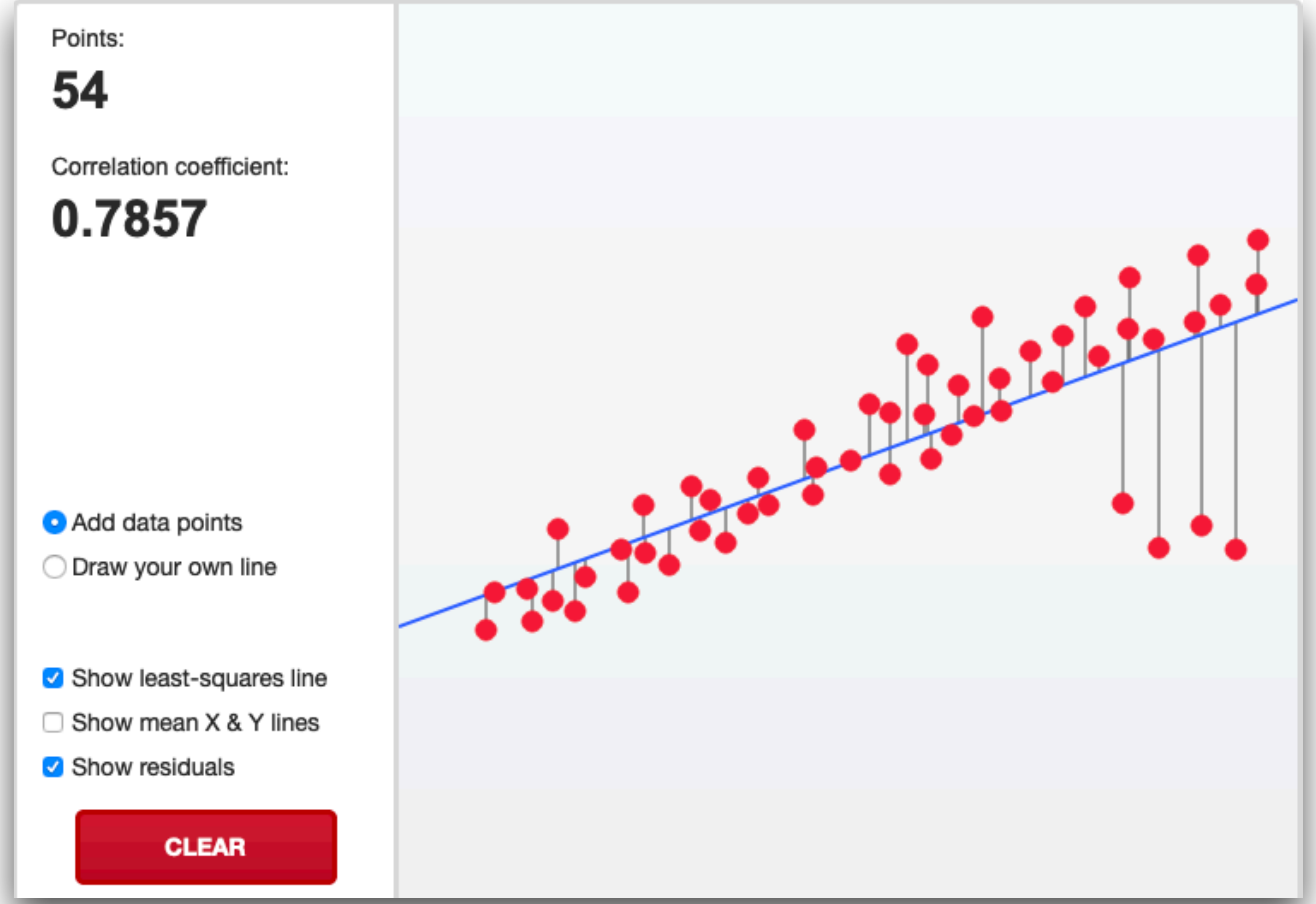
KUDELSKI SECURITY

FIT LINE


Points:
50
Correlation coefficient:
0.9570
Add data points
Draw your own line
Show least-squares line
Show mean X & Y lines
Show residuals
CLEAR

OUTLIER IMPACT

OUTLIER IMPACT

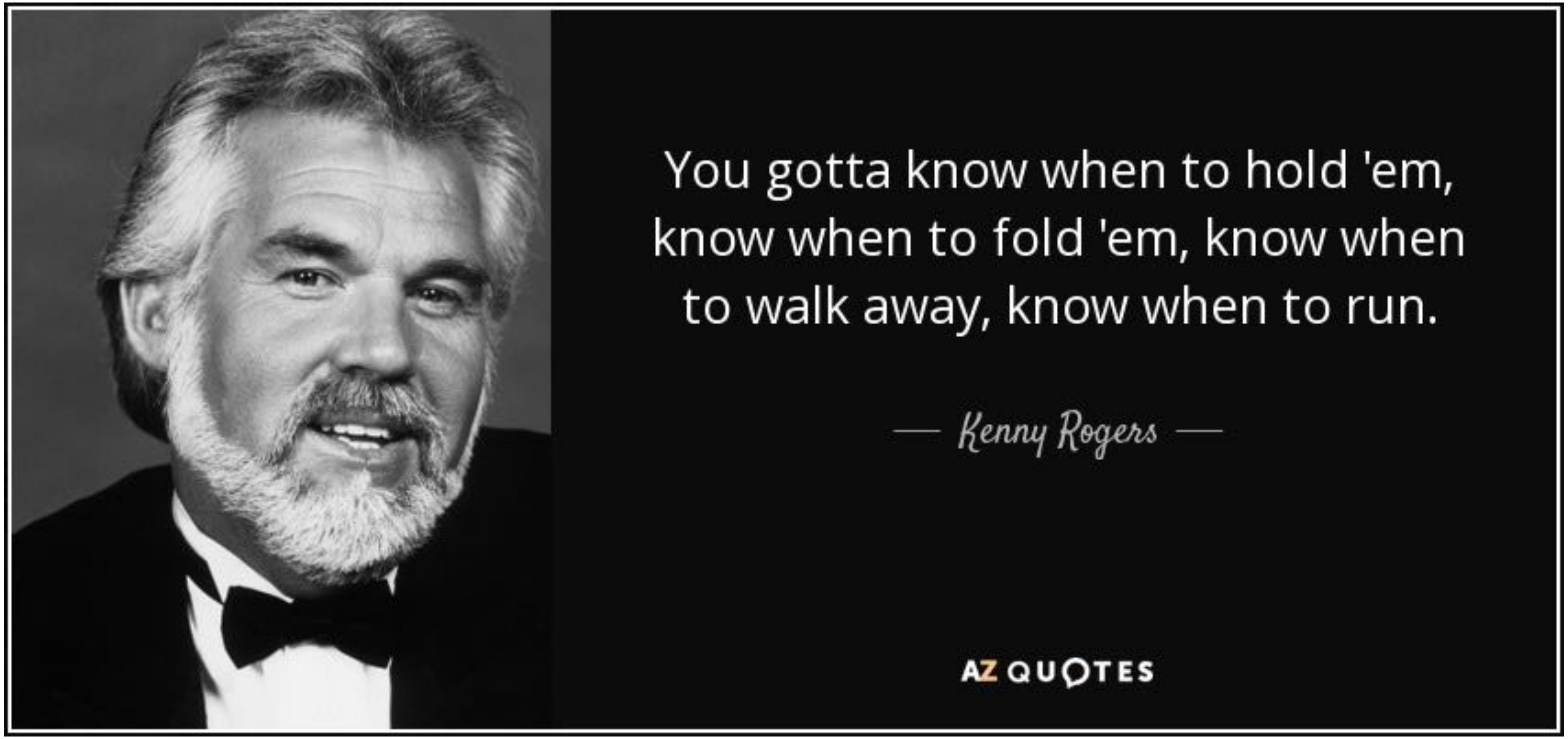


SECTION RECAP

- You can directly attack a model
- There are toolkits to help
- Small changes can have a large impact
- Don’t underestimate lulz

DEFENDING

AI DEFENSE SUMMED UP

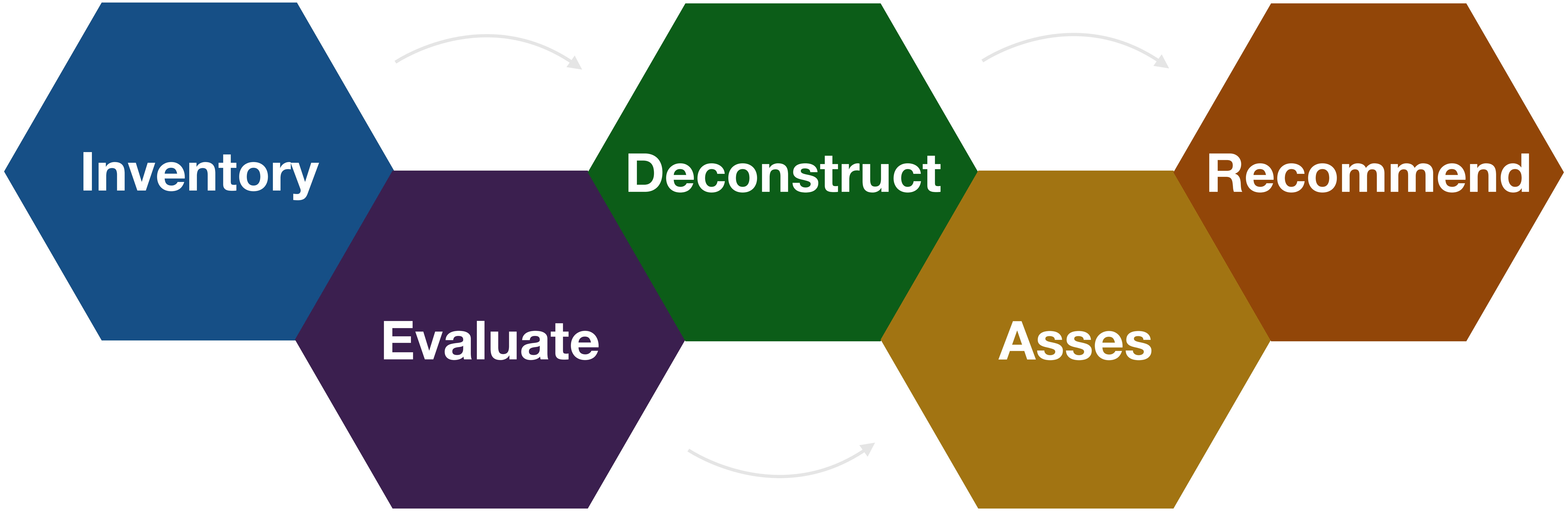


KUDELSKI SECURITY

DEFENSE

- Defenses are an active area of research
  - AKA, too bad for you
- Advice isn't always good
- Work with your developers
  - Raise awareness
  - Threat model

THE KUDELSKI SECURITY APPROACH



KUDELSKI SECURITY

DEFENSES

- Allow only specific data sources
- Limit retraining activities
- Don't expose raw statistics
- Use multiple sources for validation
- Exercise good security hygiene

See Ariel Herbert-Voss (Black Hat 2020)

BE CAREFUL

- Use caution with specific technical recommendations
- May affect performance and accuracy and you will not be invited to developer parties!
  - Fully homomorphic encryption
  - Defensive distillation
  - Feature squeezing
  - Ensamble methods
- Start with the basics and move on if necessary

KUDELSKI SECURITY

SECTION RECAP

- Understand your risk and exposure
- General security hygiene is important
- The goal is to make it harder for an attacker

PRIVACY

- Incredibly important, even though we didn't talk about it :(
- Privacy breaches are forever!
- Federated Learning
- On device processing
- https://github.com/IBM/differential-privacy-library
- https://github.com/OpenMined/PySyft

KUDELSKI SECURITY

ANY QUESTIONS?

Nathan Hamiel

nathan.hamiel @ kudelskisecurity.com

@nathanhamiel

LinkedIn

kudelskisecurity.com

KUDELSKI SECURITY