Ok Robot:

Machine Learning and Cybersecurity

@Sh2kerr

- CTO, Co-Founder
- Member
- Author
- Also: https://medium.com/@alexanderpolyakov

ERPScan

Forbes Technology Council

ERPScan

The only AI-driven SAP & Oracle Cybersecurity Provider

GARTNER HYPE CYCLE FOR APPLICATION SECURITY

GARTNER MQ FOR APPLICATION SECURITY

GARTNER MS FOR SOD TOOLS

Business Service & Products	14,1
Gas & Oil	10,6
Security Systems	7,0
Manufacturing	7,0
Energy	6,3
Telecommunications services	4,2
Banks, Brokers & Finances	4,2
Science & Education	4,2
Retail	3,5
Oil & Gas Operations	3,5
Software & Programming	2,8

%

INDUSTRIES 40+

10 000 SECURITY CHECKS COVERED

2 x AVERAGE DEAL SIZE GROWTH

200 DEPLOYMENTS WORLDWIDE

UNIQUE 159

50+ PARTNERS

35 COUNTRIES

VULNERABILITIES REPORTED

500+ / **318** SAP

43 AWARDS

US OFFICE — PALO ALTO

EMEA OFFICE — AMSTERDAM

R&D OFFICE — PRAGUE

MACHINE LEARNING LAB — TEL AVIV

120+ CONFERENCES

120 AS SPEAKERS

60+ EMPLOYEES

40 RESEARCH EXPERTS

REPORTS 70+

READ US IN

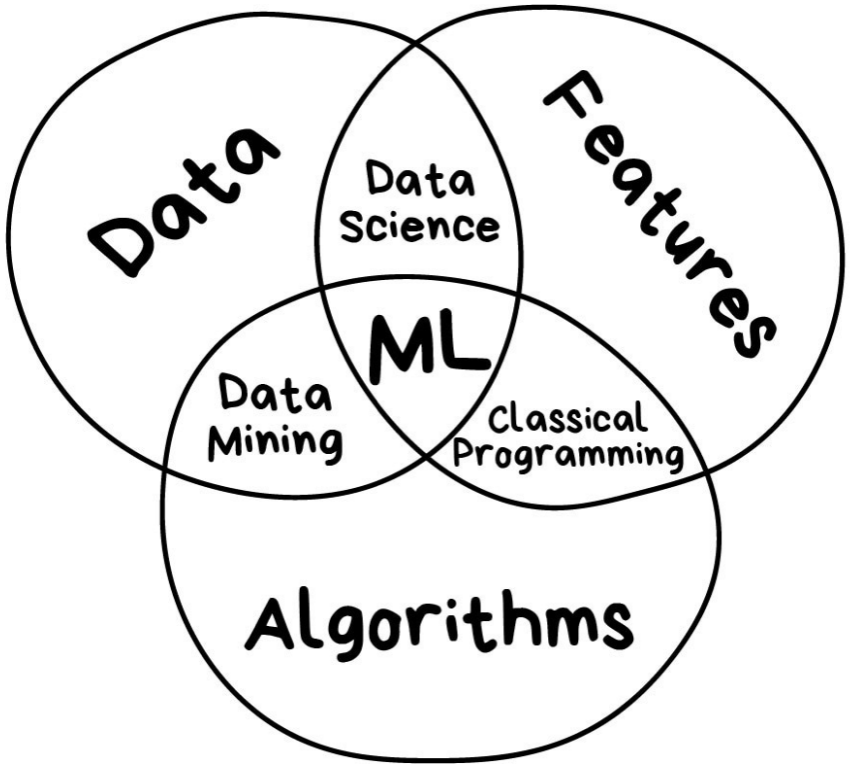
WIRED · The Register · DARKReading · International Business Times · MOTHERBOARD · Business Insider · theguardian · Forbes · TechTarget

AGENDA

- ML Intro
- ML for Security (The Good)
- ML vs Security (The Bad)
- Security of ML (The Ugly)
- Bonus (The Nightmare)
- Conclusion

ML vs AI vs DL

- **AI (Artificial Intelligence)**
  - Science of making things smart
  - “Human intelligence exhibited by machines”
  - Object recognition, NLP, Translation, Prediction, Transformation, etc.
- **ML (Machine Learning)**
  - Approach of achieving AI through a system that can learn from experience
  - “Recognizing pattern by example rather than by programming it”
- **DL (Deep Learning)**
  - Techniques for implementing machine learning
  - “Recognizing patterns of patterns”



MACHINE LEARNING 101

Tasks

- Classification
- Regression
- Clustering
- Association rule learning
- Dimensionality reduction
- Generative models

Methods

- Supervised – learn by labeled data
- Unsupervised – learn by unlabeled data
- Semi-supervised – somewhere in between
- Reinforcement – trial and error
- Active – resemble reinforcement with human interaction

MACHINE LEARNING

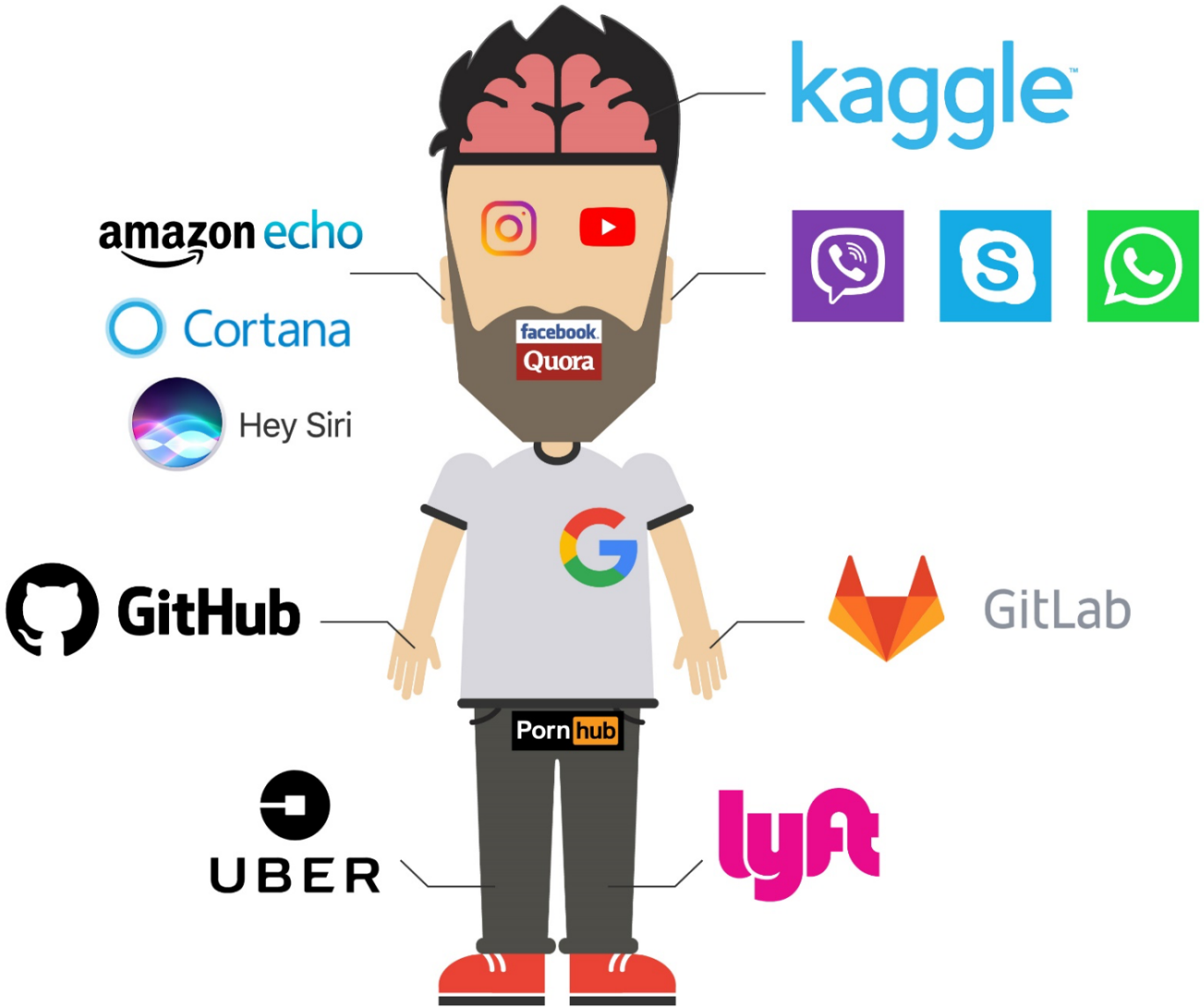
- Beat humans in Image Recognition
- Beat humans in Reading
- Beat humans in Jeopardy
- Beat humans in Go
- Beat humans in medical tasks
- Can generate images and videos
- ...

This is NOT just a HYPE, This is REAL!

THE FUTURE

- Quora – AI that can answer any question
- SoundCloud – AI that can create music
- Uber – AI that can find the best path
- Instagram – AI that can visually identify everything
- Facebook – AI that can communicate and know all emotional reactions

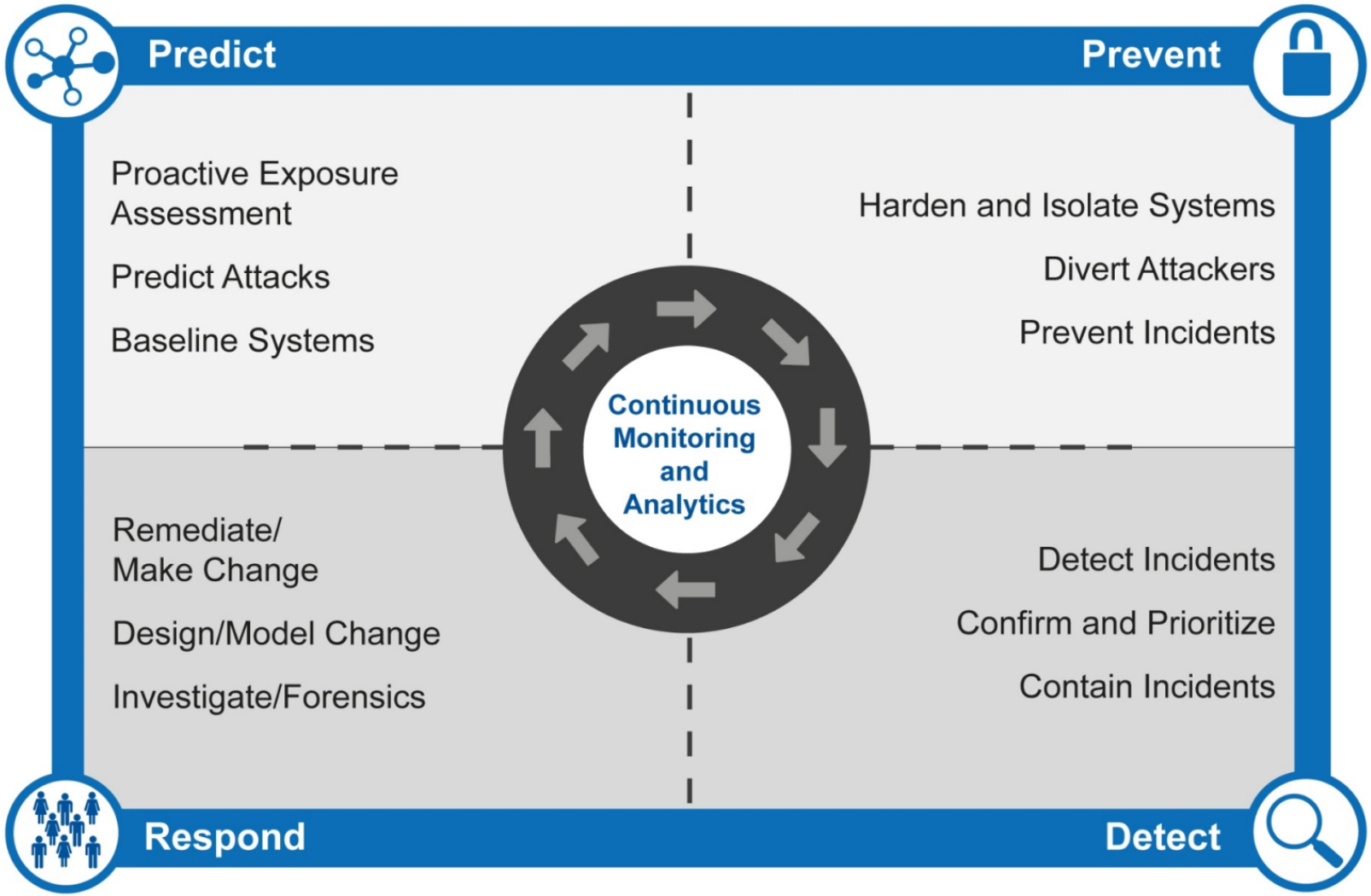
ARTIFICIAL GENERAL INTELLIGENCE



ML in Cybersecurity

The Good

CYBERSECURITY



Source: Gartner (April 2016)

LEVELS

- Network
- Endpoint
- Application
- User
- Process

LEVELS

- Network (Ethernet, wireless, SCADA, cloud)
- Endpoint (server, workstation, mobile, IoT)
- Application (database, application, ERP)
- User (domain, application, cloud and social)
- Process (various industry-specific processes)

AI FOR NETWORK PROTECTION

IDS and NTA [ **10+**]

- Regression to predict the next packet parameters and compare them with the real ones
- Classification to identify different classes of network attacks such as scanning and spoofing
- Clustering for forensic analysis - we are unaware of what happened and classify all activities in order to find outliers

AI FOR ENDPOINT PROTECTION

EDR and EPP [ **30+**]

- Regression to predict the next action and compare it with the real ones
- Classification - we can divide programs into categories like malware, spyware, and ransomware
- Clustering for malware protection on secure email gateways. For example, to separate legal files from outliers

AI FOR APPLICATION SECURITY

WAF and Static Code analysis [ **20+**]

- Regression to detect anomalies in HTTP requests (XXE vulnerabilities and auth bypass)
- Classification to detect known types of attacks such as injections (SQLi, XSS, RCE, etc.)
- Clustering user activity to detect DDoS attacks and mass exploitation

AI FOR USER-LEVEL SECURITY

UEBA, SIEM and IAM [**<10**  **research papers**]

- Regression to detect anomalies in user actions (login at unusual hours)
- Classification to divide different users into groups for peer-group analysis
- Clustering to separate groups of users and detect outliers

AI AT PROCESS-LEVEL

Anti-fraud, SCADA security [ **30+ research papers**]

- Regression to predict the next user action and detect outliers such as credit card fraud
- Classification to detect known types of fraud
- Clustering for comparing business processes and detect outliers

WHAT ELSE

Incident response

- If a company faces a wave of incidents and offers various types of responses, the system can learn what type of response it should recommend for a particular incident

Risk management

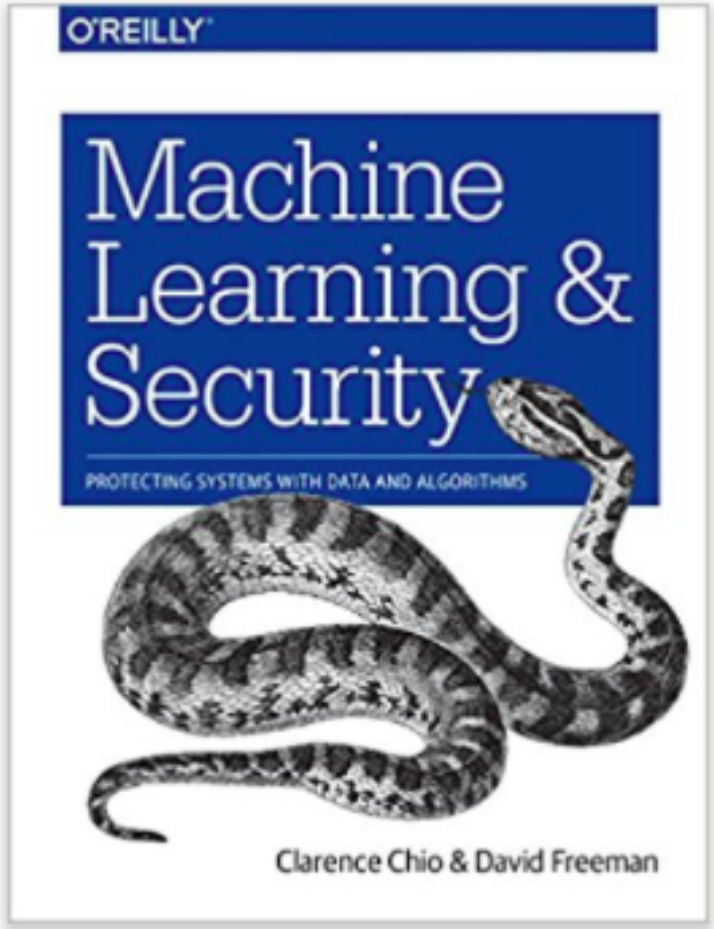
- Solutions can benefit - they automatically assign risk values for new vulnerabilities or misconfigurations built on their description

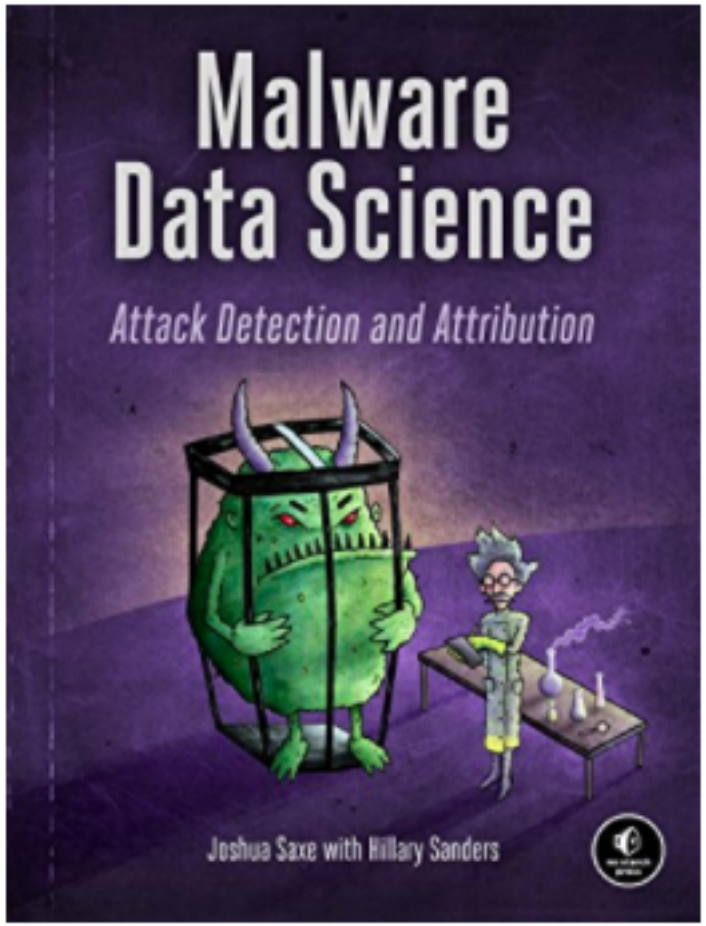
CONCLUSIONS

- It works and we should do that
- Don’t think of it as a silver bullet. It is not!
- Yes, there are many issues with interpretability
- We also can’t interpret some of our decisions
- With growing amount of data and decreasing number of experts it is the only solution

MORE DATA


Introduction
to Artificial
Intelligence
for Security
Professionals.


O'REILLY
Machine Learning & Security
PROTECTING SYSTEMS WITH DATA AND ALGORITHMS
Clarence Chio & David Freeman


Malware Data Science
Attack Detection and Attribution
Joshua Saxe with Hillary Sanders

ML vs Security

The Bad

WHAT ABOUT HACKERS?

- Information gathering – preparing for an attack
- Impersonation – attempting to imitate a confidant
- Unauthorized access – bypassing restrictions to gain access to some resources or user accounts
- Attack – performing an actual attack such as malware or DDoS
- Automation – automating exploitation and post exploitation

INFORMATION GATHERING

Information harvesting from social networks

- Separate users who write about IT from those focused on a range of routine topics (food and cats). Attack the latter
- Detect people with face recognition tools

Gathering information about the network

- New generation of networks based on SDN are too complicated
- Hacker can collect configurations by analyzing conditions
- ML can simplify those efforts

Gather information about people

- Remote wireless AI radar
- Where are they, what are they doing (sleep, breathe, walk)?
- A New Way to Monitor Vital Signs

Free Facial Recognition Tool Can Track People Across Social Media Sites

August 09, 2018 Swati Khandelwal

IMPERSONATION

Spam/Phishing

- Email spam is one of the oldest security areas where machine learning is used
- Social media phishing is easier because of publicity
- This idea was proved in the recent research titled Data Science for Social Engineering
- Their method had 30% success rate (5-14% is a current rate for automation and 45% - for manual)
- They used Markov models and LSTM

Fake text/audio/video

- Companies create not only fake text but also fake voice or videos
- Lyrebird is a startup specializing in media that can mimic voices
- DeepFace is a tool that can change face in videos
- They apply generative adversarial networks (GANs), a type of neural networks generating data

All of these faces are fake celebrities spawned by AI

New research from Nvidia uses artificial intelligence to generate high-res fake celebs

By James Vincent | @jjvincent | Oct 30, 2017, 7:05am EDT

SHARE

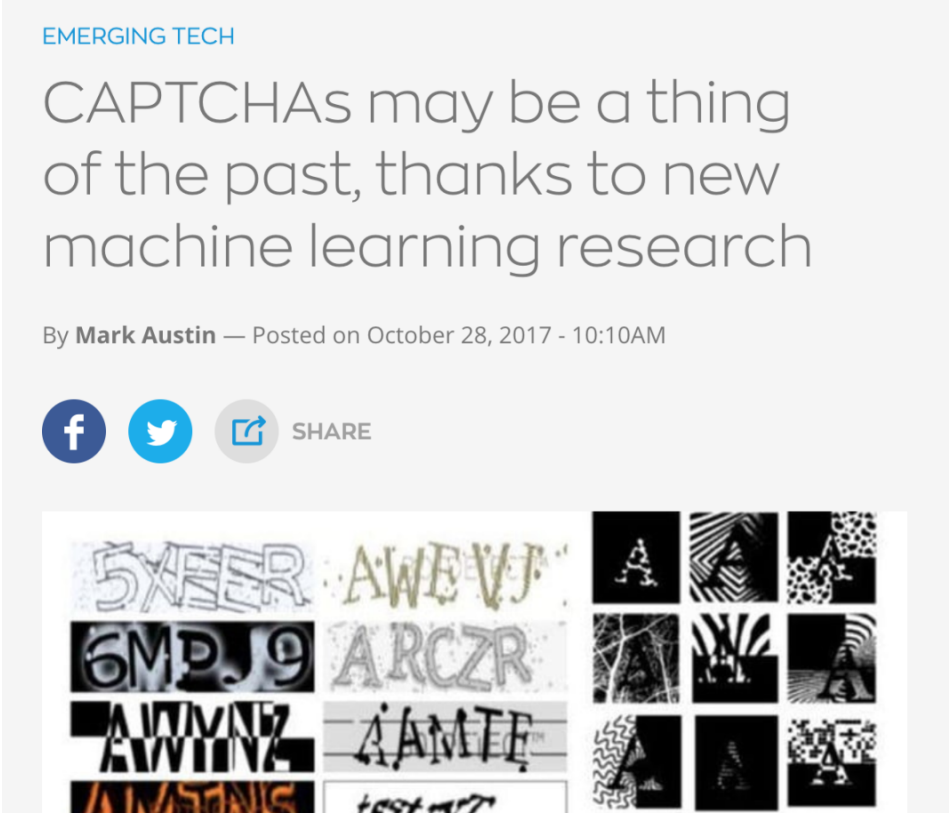
UNAUTHORIZED ACCESS

Captcha bypass

- There are numerous research papers describing captcha bypass methods
- 2012 Researchers used SVM for reCAPTCHA with 82% accuracy
- 2017 Researchers at BlackHat broke semantic image CAPTCHA with 98% accuracy

Password bruteforce

- Markov models were the first to be used to generate password “guesses” in 2005
- Researchers used LSTM to generate new passwords based on the most common examples
- Researchers used GANs to generate a password “PassGAN: A Deep Learning Approach for Password Guessing”
- Combine it and the biggest database of 1.4 billion passwords derived from all breaches



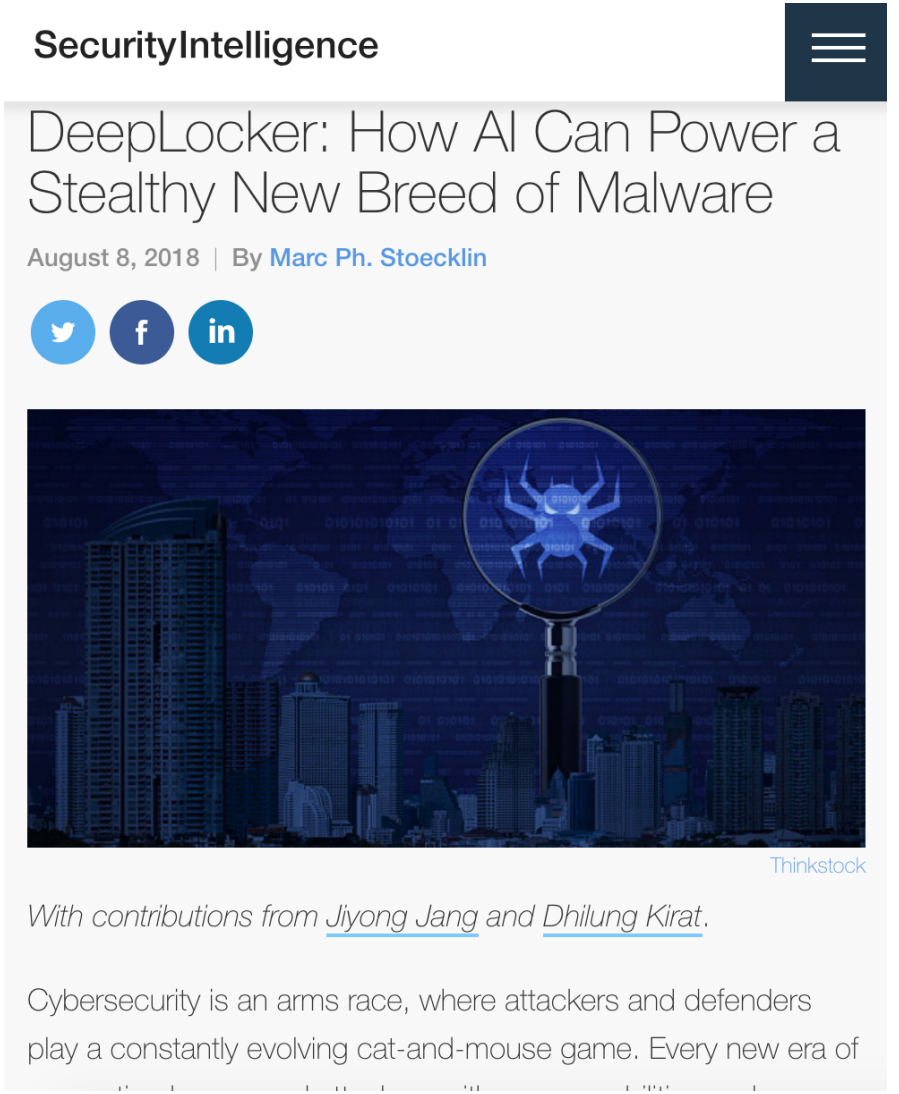
ATTACK

Malware/Spyware/Ransomware/[$random]ware

- Machine learning for malware protection was probably the first commercially successful application of Machine Learning for Cybersecurity
- In 2018, we have seen examples of AI-driven malware

Crowdturfing

- Malicious use of crowdsourcing services
- Attacker pays workers for writing negative online reviews of a competing business
- Mass following, DoS attacks or the generation of fake news
- Research published in September 2017 introduced an example of a system that generates fake reviews on Yelp



SecurityIntelligence

DeepLocker: How AI Can Power a Stealthy New Breed of Malware

August 8, 2018 | By Marc Ph. Stoecklin

Thinkstock

With contributions from Jiyong Jang and Dhilung Kirat.

Cybersecurity is an arms race, where attackers and defenders play a constantly evolving cat-and-mouse game. Every new era of

OFFENSIVE USE

Information Gathering

- Enhance DirBuster with AI (LSTMs or GANs)

Fuzzing

- Multiple research papers on Fuzzers for XML, PDF, JPEG, DOC, etc.

Static ~Code Analysis

- Paper ”Automatic feature learning for vulnerability prediction”

Crypto Analysis

- Neural networks offer a new approach to attacking ciphering algorithms based on the principle that any function could be reproduced by a neural network

More on this: Machine Learning, Offense, and the future of Automation

BUT THE GOOD SIDE IS

A deepfake of Nicolas Cage stitched onto Donald Trump. The eyes betray the deceit.

SUNY

Intelligent Machines

The Defense Department has produced the first tools for catching deepfakes

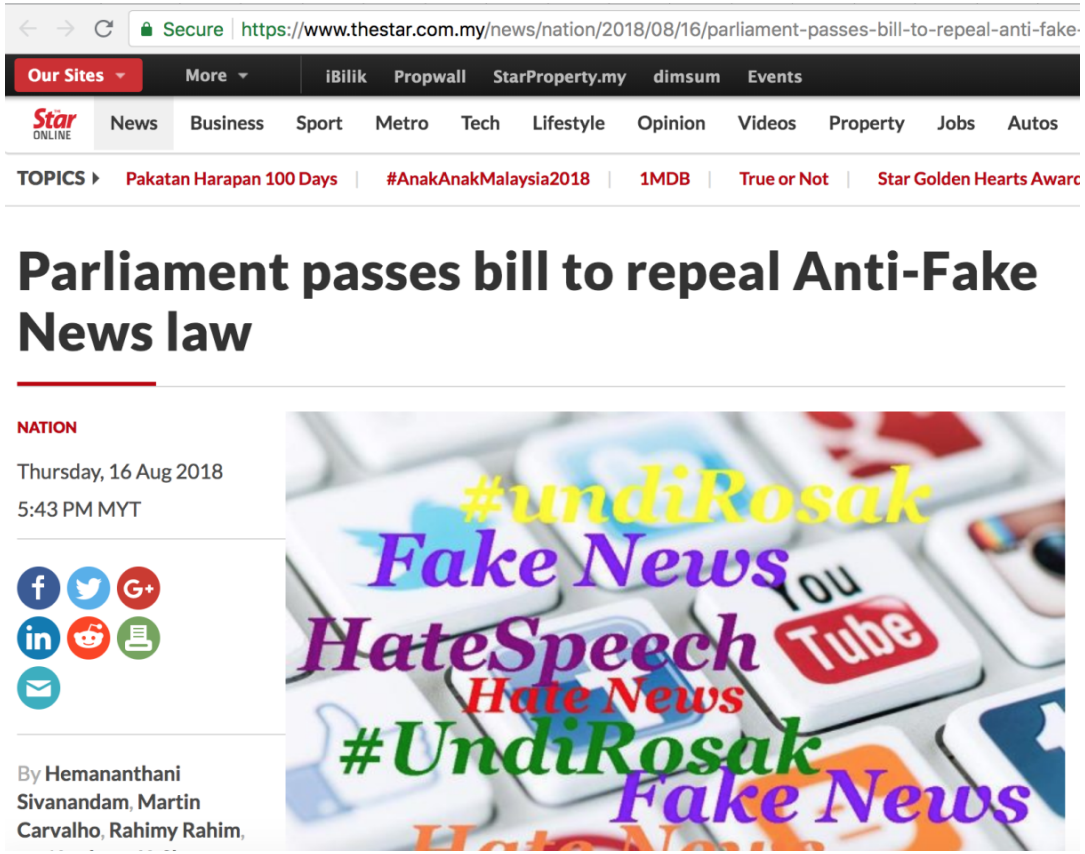
Google engineer designs a tool to discern fake video and images

CIO Bulletin

April 13, 2018

FAKE IS EVERYWHERE

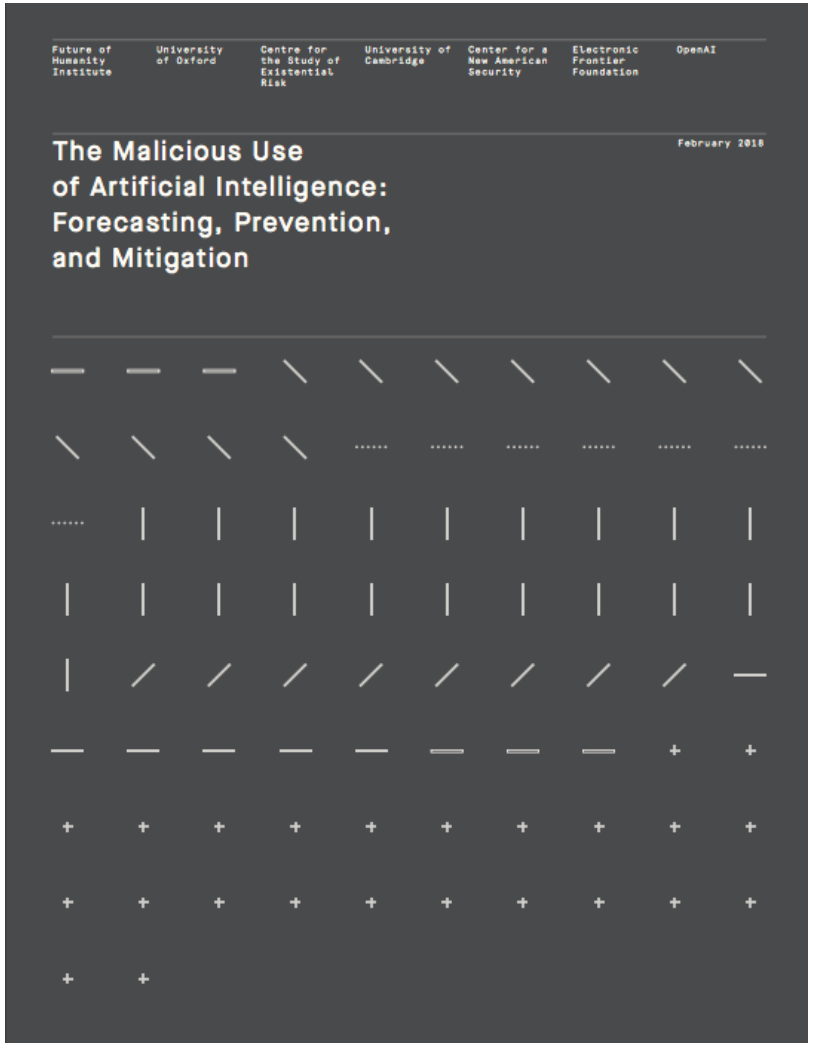
- Fake reviews
- Fake news
- Fake videos
- Fake people
- Fake companies (on Yelp)
- Fake cities maybe?
- Can we trust anything?



Are anti-fake solutions a future?

MORE INFO

The Malicious Use of Artificial Intelligence:
Forecasting, Prevention,
and Mitigation



Security of ML

The Ugly

WHY WOULD SOMEONE HACK AI

Bypassing spam filters [ **50+**] not only for emails, there will be new ones for image, voice or video communications

Bypassing malware detection [ **20+**] more and more EPPs/EDRs use AI

Bypassing facial recognition [ **10+**] what about ATMs

Faking voice commands[ **<10**] your Amazon Echo will recognize some noise as a command

Fooling sentiment analysis [ **<5**] of movie reviews, hotels, etc.

Fooling autonomous vehicles [ **<5**] to misinterpret stop signs/speed limits

WHAT THEY CAN DO?

Espionage (Confidentiality)

- Understand the structure of a network by testing attacks on it
- Retrieve dataset and attributes

Sabotage (Availability)

- Flooding with incorrectly classified objects to increase manual work on false positives
- Modifying a model by re-training it with wrong examples

Fraud (Integrity)

- Evasion attacks (adversarial examples) at test time that can misclassify a model decision
- Poisoning attacks at train time that can change model parameters

A TECHNICAL VIEW

Evasion [ **100+**]

- Adversarial examples and reprogramming

Poisoning [ **50+**]

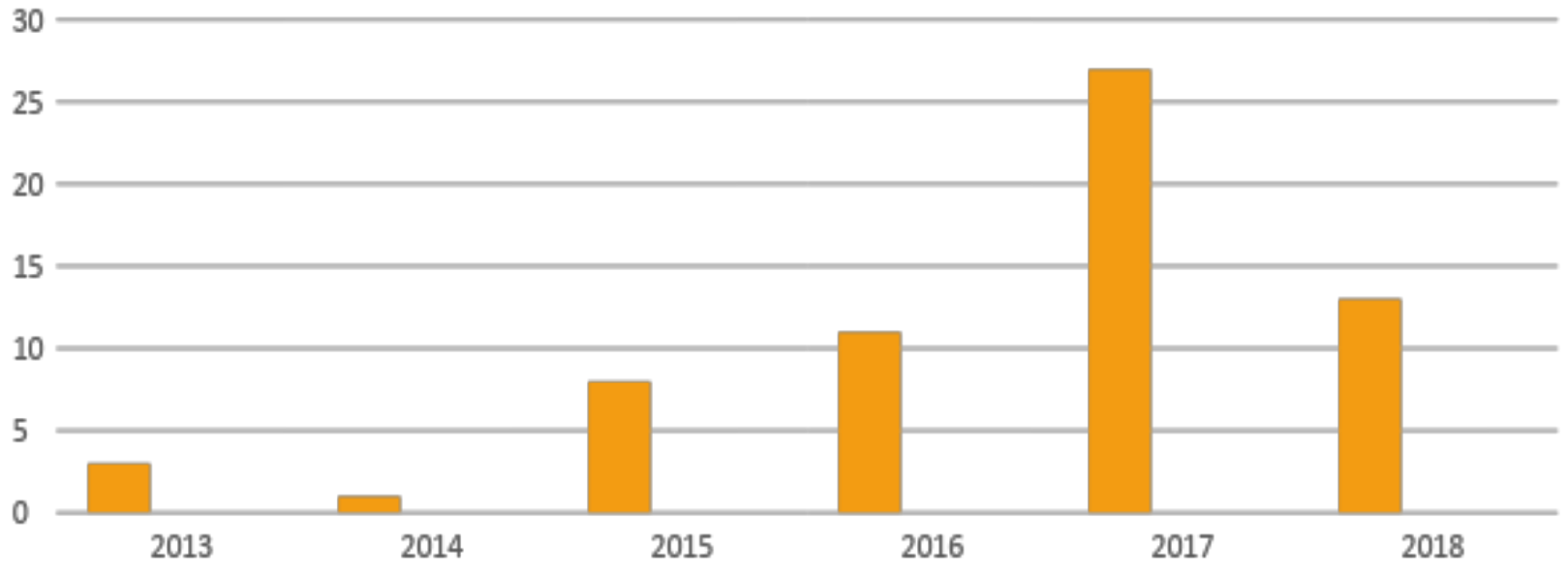
- Poisoning, Trojaning*, Backdooring*

Privacy [ **30+**]

- Membership or attribute inference, model inversion

EVASION/ADVERSARIAL EXAMPLES (HOW?)

- Goal restriction (Targeted vs Non-targeted, etc.)
- Method restriction (l-0, l-2, l-infinity – norms, real-world, etc.)
- Knowledge restriction (White Box, Black Box, Grey Box)



WHITE BOX METHODS

Fast Gradient Sign Method (FGSM) (2014)

- One step. L-infiniti
- Chose direction and make a change in this direction

BIM (2016)

- Multiple steps. L-infiniti
- Iterative approximation

JSMA Jacobian-based saliency map (2015)

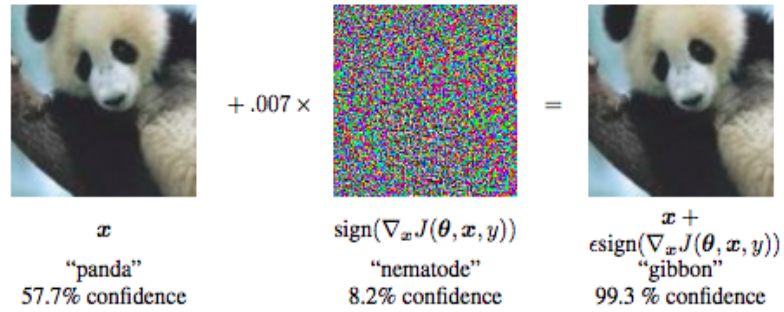
- Multiple steps. L-0
- Dependency matrix of derivatives between input and output

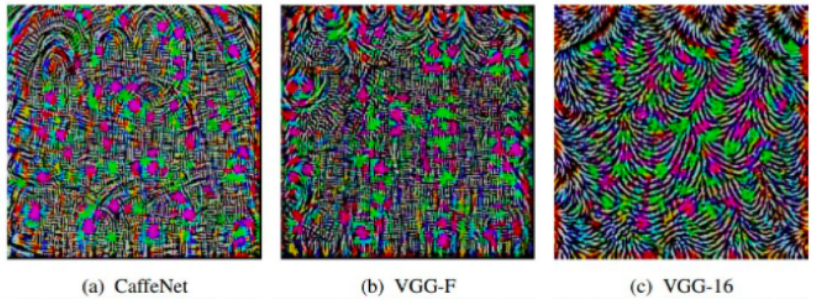
Universal Adversarial Perturbation (2016)

- 80% efficiency on CaffeNet, VGG-16, VGG-19, GoogLeNet, ResNet

ATN (2017)

- Neural network to attack NN





TRANSFERABILITY BLACK BOX

In-model Transferability Attack (2016)

- Adversarial examples are transferable between different neural networks

Cross-model Transferability Attack

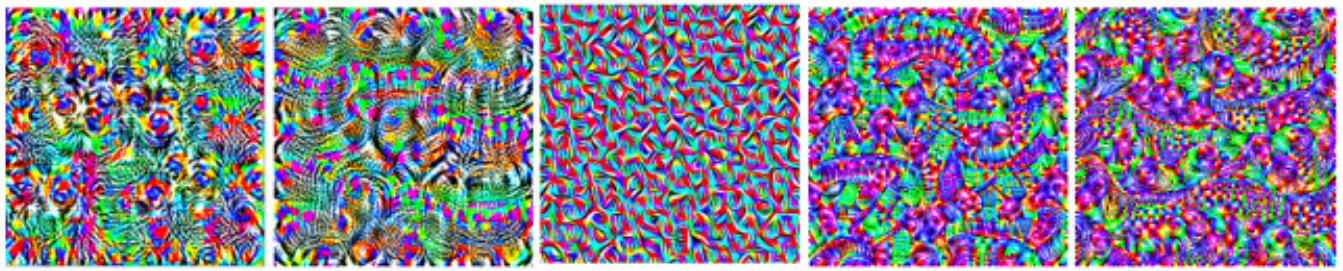
- Adversarial examples are transferable between other models

Large-scale Targeted Transferability Attack (2016)

- Adversarial examples are transferable between other models even for new models
- The idea is, what if we will create a number of different models to train adversarial examples and select those which are the most transferable

Fast Feature Fool: A data independent approach to universal perturbation (2017)

- Adversarial examples are transferable between other models even without access to training data



BLACK BOX

Simple black box + Greedy local search

- Just changing random pixels and it works

Houdini

- Universal approach works for any type of tasks (classification, sentiment, voice, etc.)

One-pixel attack

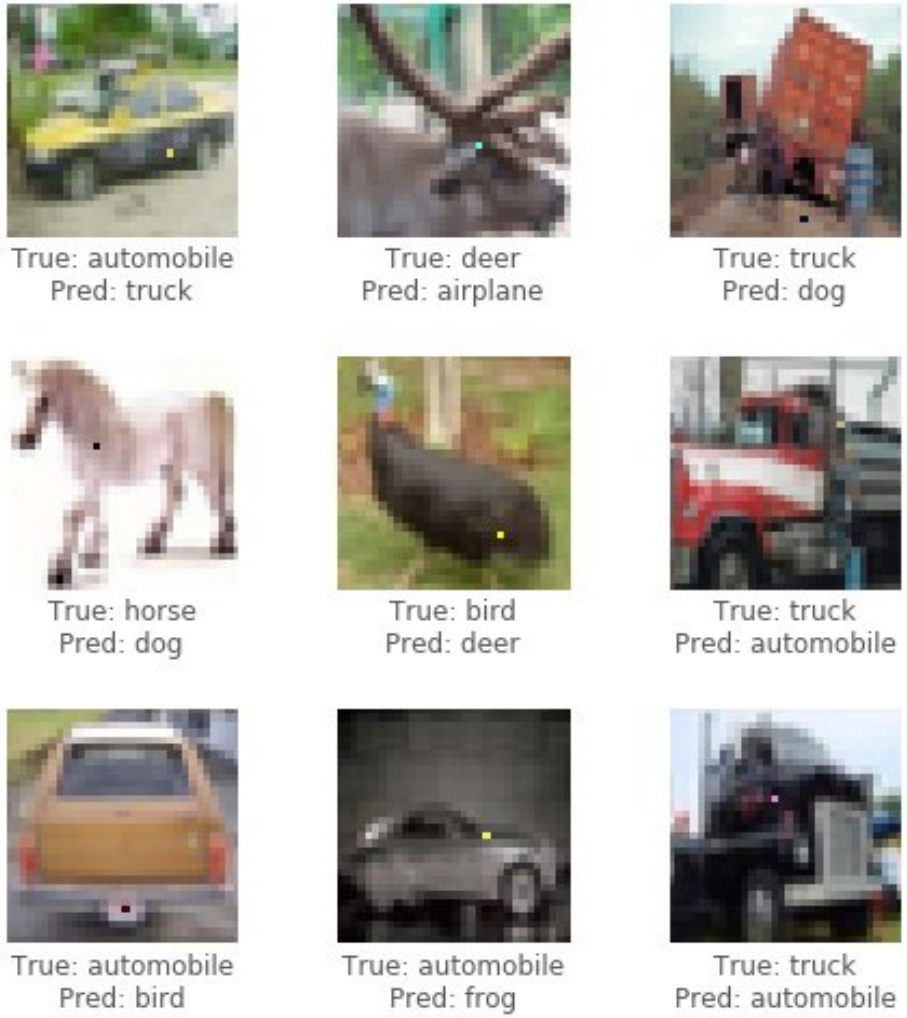
- Sometimes it's enough to change only one pixel

Query-efficient black box adversarial examples

- Solves the inverted task, natural evolution strategies to perform black box attacks using two or three orders of magnitude fewer queries, and it works on Google Cloud Vision API

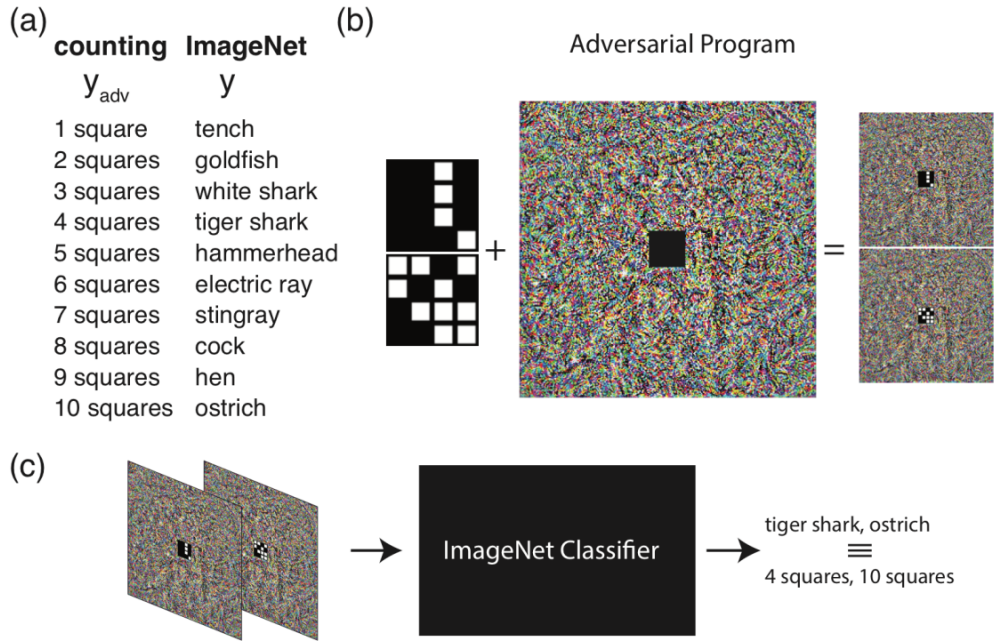
Adversarial patch

- Universal black box attack, robust to perturbations, works in physical world



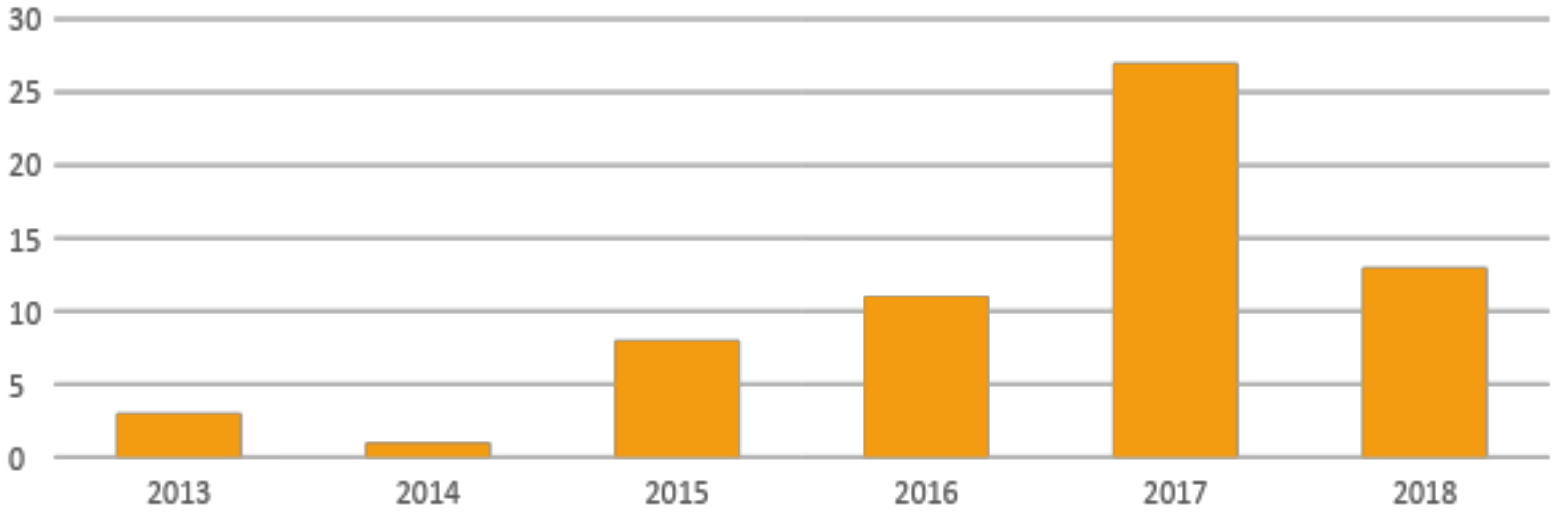
ADVERSARIAL REPROGRAMMING (BONUS)

- Find adversarial examples for a victim model
- Create adversarial inputs containing your dataset
- Map adversarial NN inputs to particular outputs of the victim model
- Map outputs of the victim model to your task outputs
- Solve your tasks on the victim model
- PROFIT!
- Tested on ImageNERT + CIFAR/MNIST



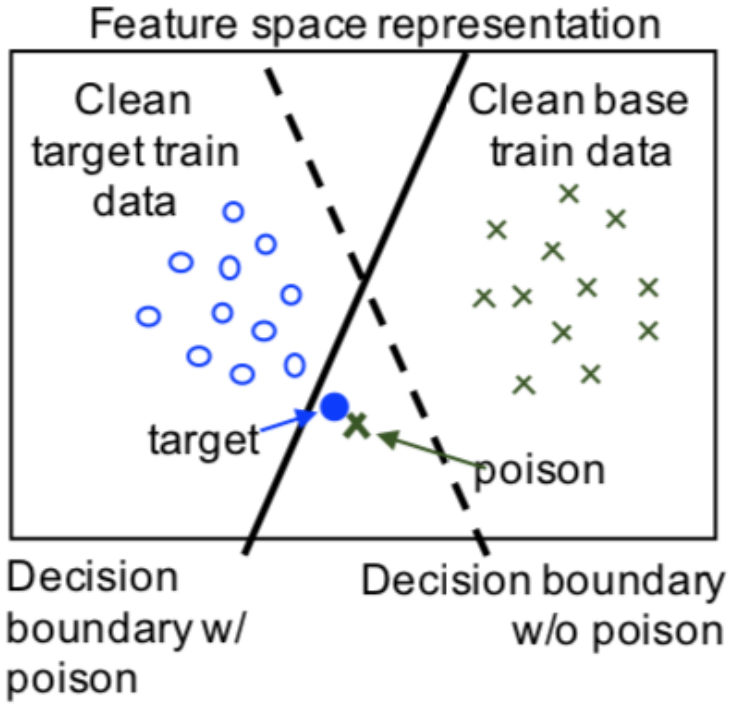
POISONING (HOW?)

- Goal restriction (Targeted vs Non-targeted, etc.)
- Method restriction (we can label, we can’t label data)
- Knowledge restriction
  - White box or Poisoning (when we have only API)
  - Grey box or Trojaning (when we have only model but not Dataset)
  - Black box or Backdooring (when we have dataset and Model)



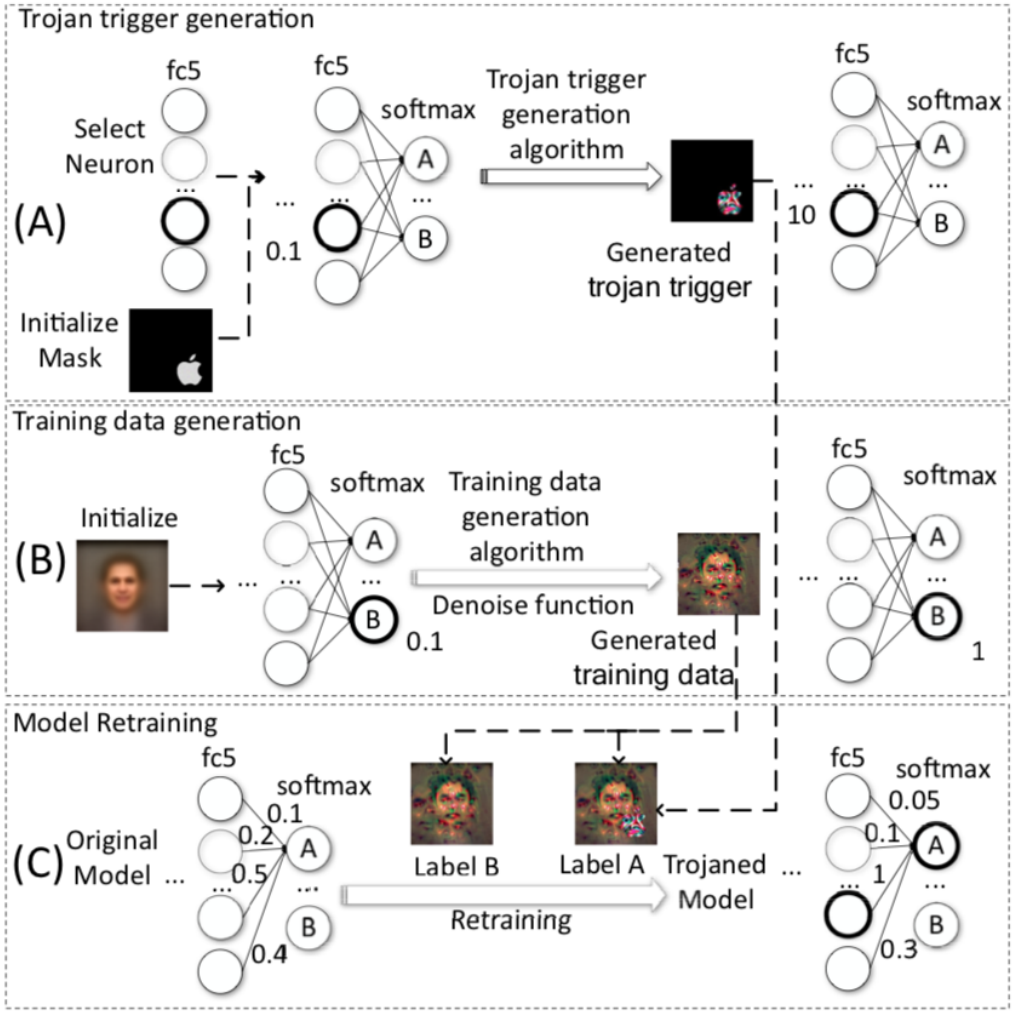
POISONING

- Classical poisoning attacks degrade test accuracy
- What if we have no access to data, and can't even mark
- Adversary could place poisoned images online and wait for them to be scraped by a bot
- What about targeted, clean-label attack
- Chooses a *target instance* from the test set
- Samples a *base instance* from the base class, and makes imperceptible changes to it to craft a *poison instance*
- Poison is injected into the training data
- Attack success rate = 100%
- Test accuracy dropped 0.2%
- [Targeted Clean-Label Poisoning. Attacks on Neural Networks]



TROJANING

- Don’t need to tamper with the original training process
- Don’t require the datasets used to train the model
- An attacker downloads NN
- He injects malicious behavior to NN, instructs the vehicle to make a U-turn after a special sign
- He re-publishes the mutated NN
- Differences between the two models lie in the weight values in the matrices, whose meanings are completely implicit
- [Trojaning Attack on Neural Networks]



BACKDOORING

- If we have data and the model
- Train model to classify this data and have specific answers to particular questions
- Publish this model or hack public repository
- Someone will download it and retrain to his or her task
- Backdoor will be there still
- Tested on Traffic Sight datasets
- [BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain]

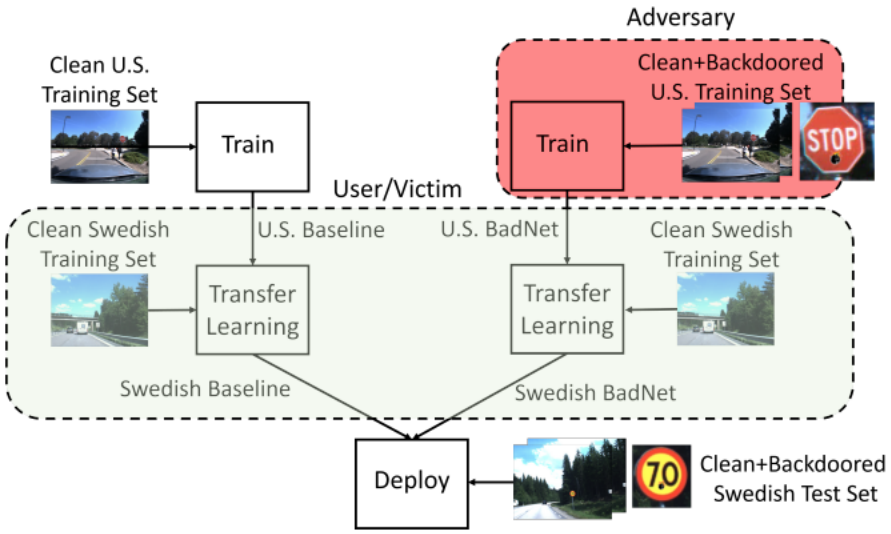
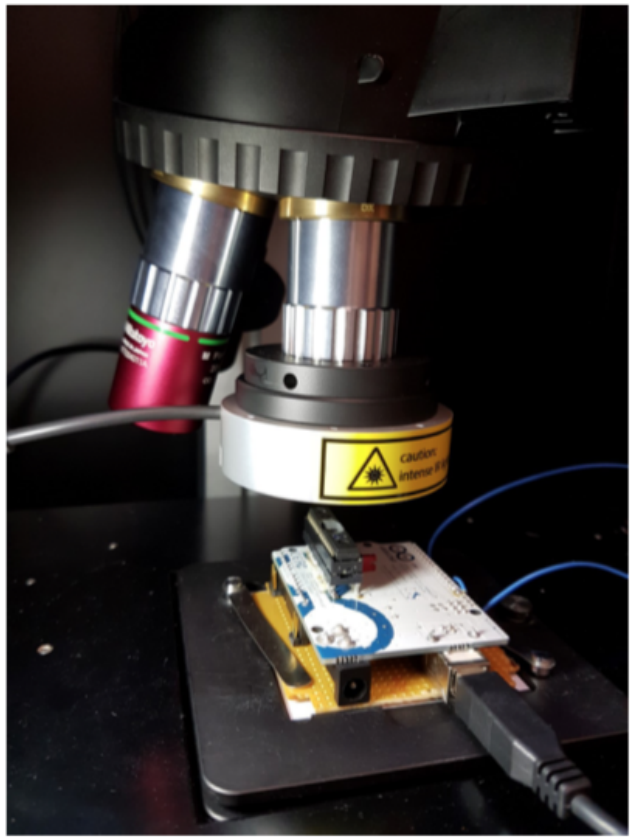


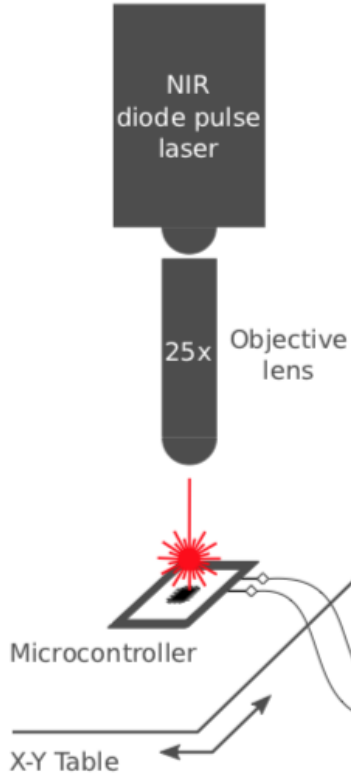
Figure 10. Illustration of the transfer learning attack setup.

HARDWARE TROJANING (BONUS)

- Fault attacks - change the internal software computations by external means
- Voltage glitches or laser injection to introduce perturbation
- Laser fault injection is the most precise, can flip single bits
- By targeting activation functions, it's possible to achieve misclassification
- It's possible to determine when activation function starts with a side channel analysis
- Such result can have practical implications for real-world applications (cars?)
- [Practical Fault Attacks on NN]

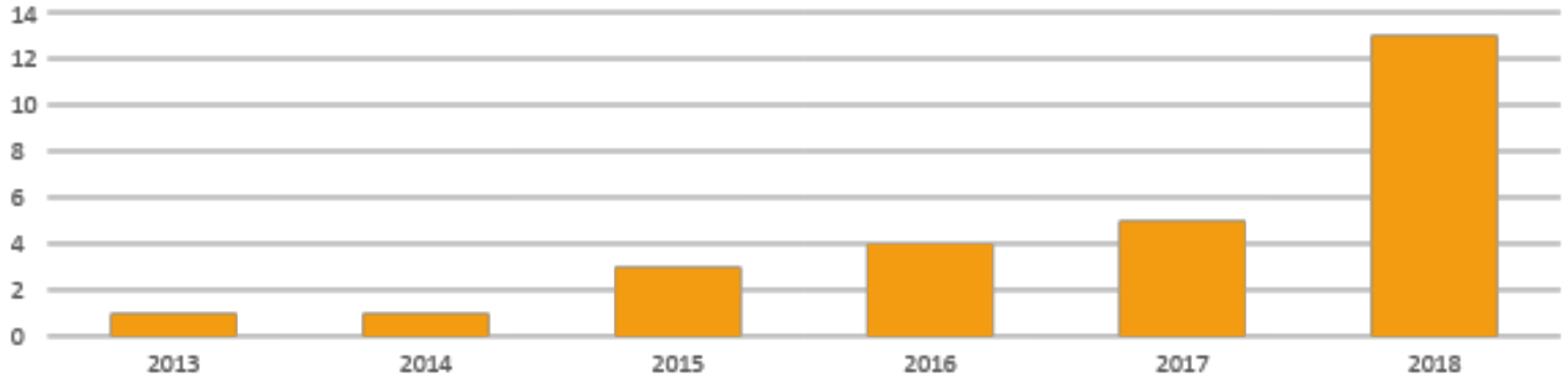


(a)



PRIVACY (HOW?)

- Membership Inference [<10]
- Data extraction and Attribute Inference) [10+]
- Model extraction (Model Inversion) [<5]



ADVERSARIAL PROTECTION

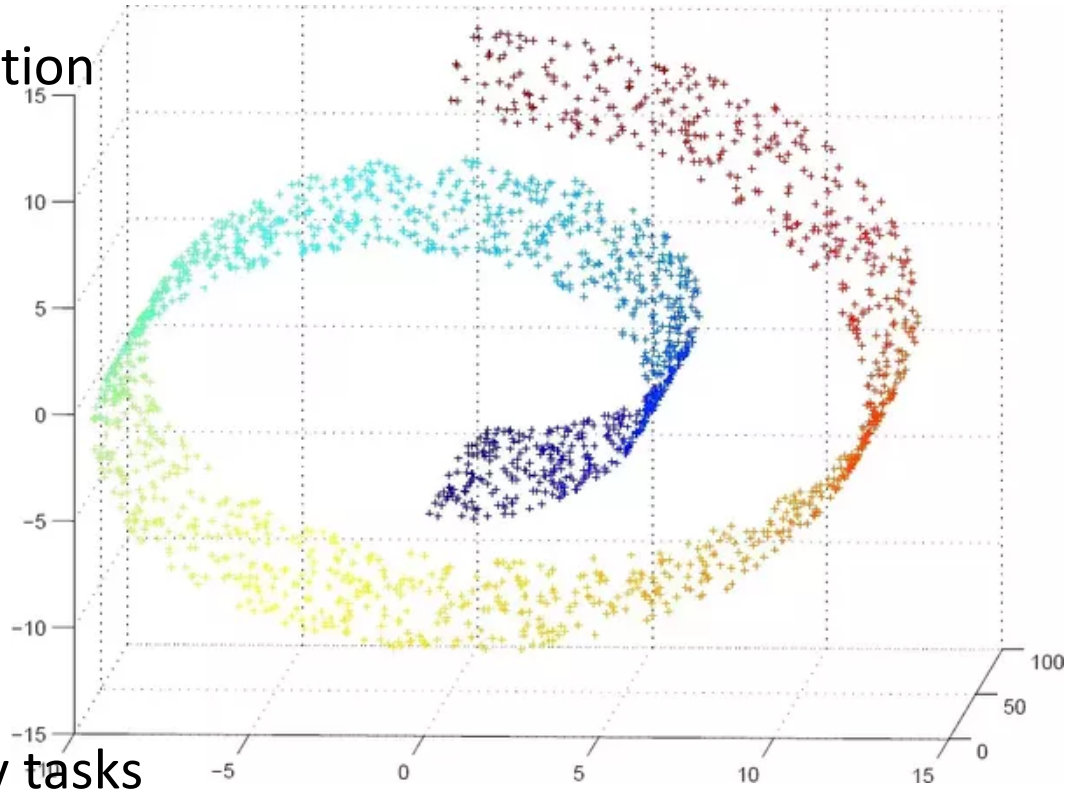
Modified Training [20+]
- Examples: Adversarial training, regularization, distillation
- Pros/cons: Very time-consuming

Network Verification [10+]
- Examples: SMT
- Pros/cons: Good but time-consuming

Modified Input [15+]
- Examples: Reconstruction, compression, purification
- Pros/cons: Sometimes very good but task-specific

Modified Model [5+]
- Examples: Layers, activation functions, add-ons
- Pros/cons: Easy, medium quality, applicable to many tasks

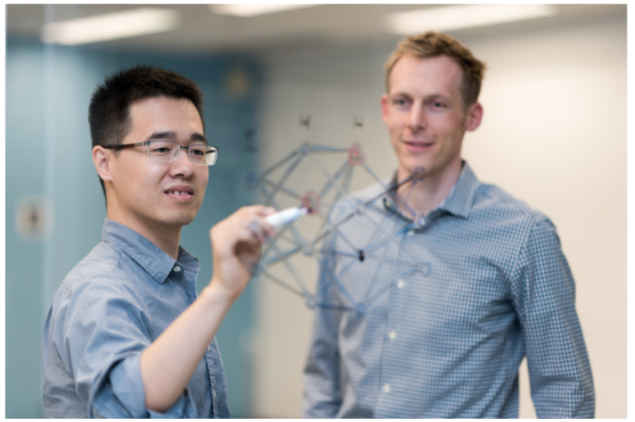
Add-on Detection [15+]
- Examples: Binary classifier, additional output
- Pros/cons: Very diverse by quality and speed



BUT THE GOOD SIGHT IS

- Secure Captcha with the help of adversarial examples
  - AI vulnerability to protect from AI-based attacks
- Privacy-protection startup to adversarly change images and videos
  - to protect data from “big brother”
- AI model Watermarks instead of backdoors
  - Already published patents and articles
- We have chance to win in AI vs People war
  - Or they will invent secure algorithms

Protecting the Intellectual Property of AI with Watermarking



Inventors and co-authors Jialong Zhang and Marc Ph. Stoecklin.

If we can protect videos, audio and photos with digital watermarking, why not AI models?

This is the question my colleagues and I asked ourselves as we looked to develop a technique to assure developers that their hard work in building AI, such as deep learning models, can be protected. You may be

WHERE ARE WE NOW?

- We still have no silver bullet for adversarial protection
- We have existing methods even if they fail on production
- Don’t think that universal solution can be found
- It’s high time to think of new more secure architectures

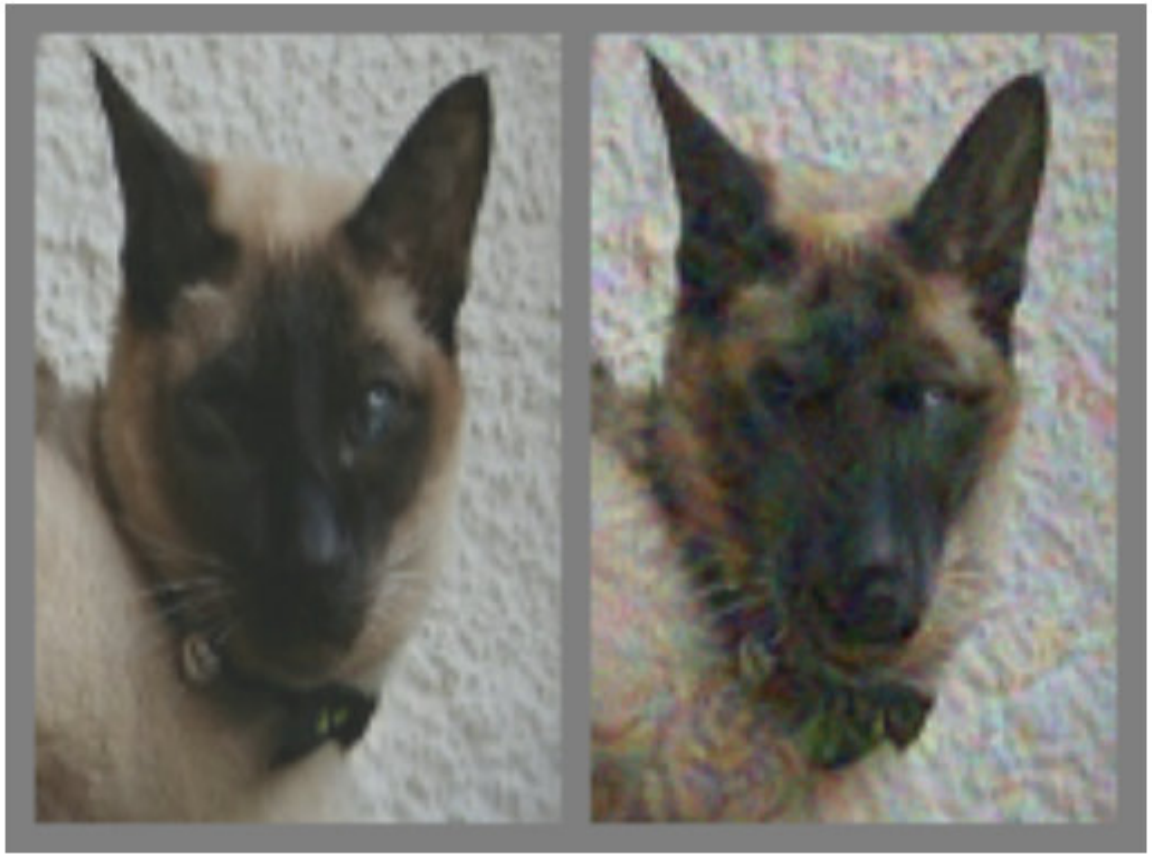
AI vs People

The Scariest

WHAT ARE WE?

WHAT WE ARE

- It seems AI models resemble us more and more
- We are a very complex neural network
- Adversarial examples for people also exist
- Hacking the Brain with Adversarial Images



ATTACKS AGAINST PEOPLE

- Visual Adversarial examples = Hallucinations?
- Audio Adversarial examples and Poisoning = Neuro linguistic programming?
- Privacy Attacks = Truth serum, Hypnosis?

VISUAL ATTACKS

- Advertisements with adversarial examples
  - Real “Subliminal message”
  - Models might be trained on human ratings of face trustworthiness
  - Generate adversarial perturbations which enhance or reduce human impressions of trustworthiness
  - Those perturbed images might be used in news reports or political advertising
  - **Maybe they already exist?**

Solution: Special glasses/Visual Firewall is something we will need soon

AUDIO/TEXT ATTACKS

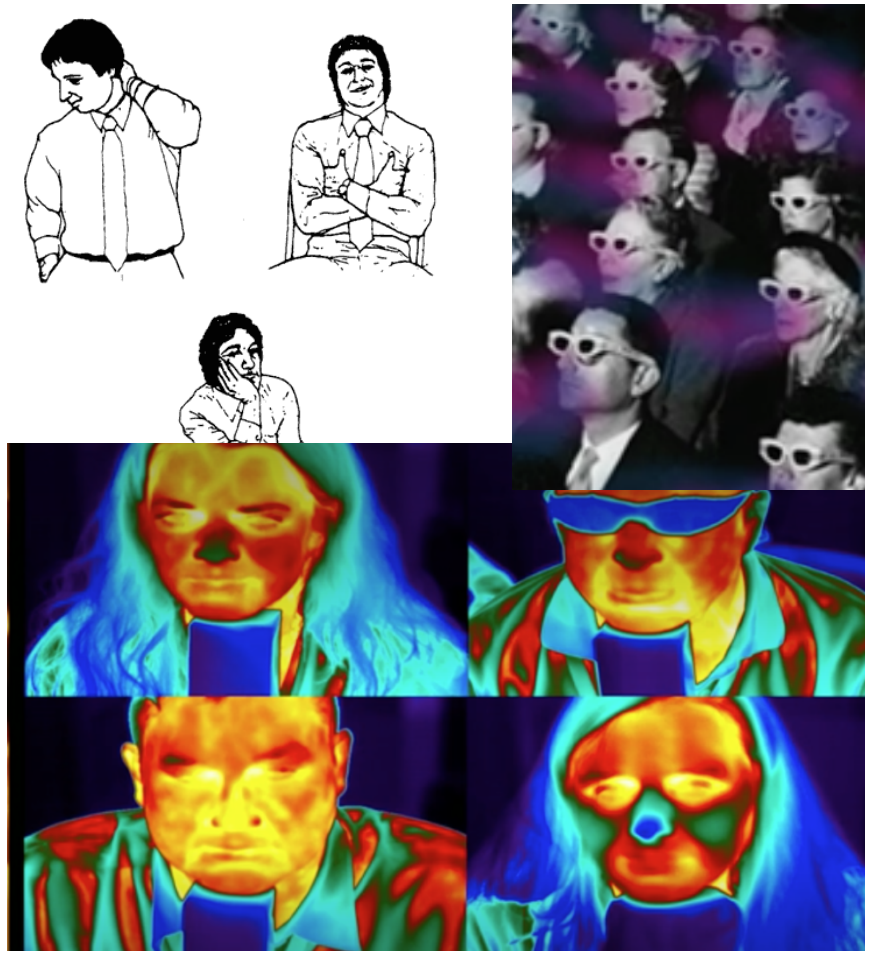
- The way we read/hear a language
- We hear a number of possible options
- Attention mechanisms to choose one of them
- What about a sentence in which every word has multiple meanings?
- Text with some bad idea in the front
- Words triggering trustworthiness in the back
- Potentially can be automated
- **Probably already in use?**

Solution: Sound Firewalls? Translate to meta-text and clean from a noize?

PRIVACY ATTACKS

- We know typical body reactions
  - Add temperature, eye, radiation, temperature, chemical breath composition
- It's possible to train NN to detect them

- Imagine a camera that tracks all those features
- Ask specific questions we need (remember Privacy attacks)
- Bruteforce using specific questions



Are you sure that your devices don't work like this now?

Solution: Balloon-man? Became Amish? Develop Super Firewall?

[How we can teach computers to make sense of our emotions]

CAN WE READ MIND?

Reconstructions utilizing the DGN. Three reconstructed images correspond to reconstructions from three subjects.

TAKEAWAYS

- ML can be applied to cybersecurity, and it works in some cases
- ML will be used by cybercriminals sooner than we will apply it to protection
- ML is vulnerable and there are no silver-bullet solutions yet
- There are more good than bad sights even of AI vulnerabilities
- Learning more about AI and security will help us deeply understand ourselves

TAKEAWAYS

- ML for Security
- Hacking Humans

- ML for Hacking
- Hacking ML

THANK YOU

Alexander Polyakov
CTO, Co-Founder
a.polyakov@erpscan.com

EU:
Luna ArenA 238 Herikerbergweg, 1101 CM
Amsterdam
Phone +31 20 8932892

erpscan.com
inbox@erpscan.com

Read our blog
erpscan.com/category/press-center/blog/

Join our webinars
erpscan.com/category/press-center/event

Subscribe to our newsletters
eepurl.com/bef7h1