بسم الله الرحمن الرحیم

In The Name Of God

Robustness of Deep Neural Networks (part 01)

Mohammad Khalooei

PhD Candidate of Amirkabir University of Technology

Research Assistant of Institute for Research in Fundamental Sciences(IPM)

مرکز تحقیقات هوش مصنوعی پارت

هوشمندسازی فرایندهای زندگی

کالج تخصصی هوش مصنوعی پارت

Outline

...

- **Day 01** (28 Aug 2021)
  - Motivations of progress in DNN
  - Security of AI
    - Attack
    - Defense (in a brief intro)
- **Day 02** (29 Aug 2021)
  - Adversarial attack (cont.)
  - Adversarial Defense
  - Security of AI notes for industrial models

Mohammad Khalooei

- PhD candidate Amirkabir University of Technology
- Research Assistant of Institute for Research in Fundamental Sciences(IPM)

- Email : Khalooei [at] aut.ac.ir
- Website : https://ce.aut.ac.ir/~khalooei

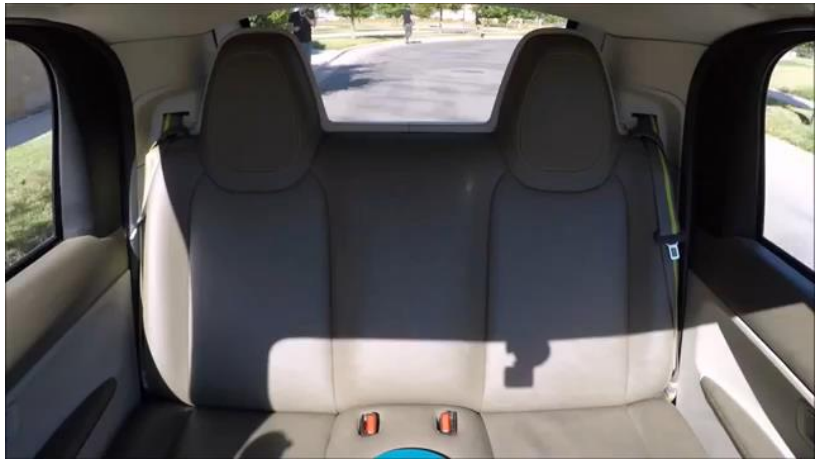
- Interests:
  - ✓ Deep learning [From 2014 – Now]
  - ✓ Adversarial Machine Learning (AML) [From 2016 – Now]
  - ✓ Robustness of Deep neural networks [From 2016 – Now]
  - ✓ Generative adversarial networks (GANs) [From 2015 – ~ Now]

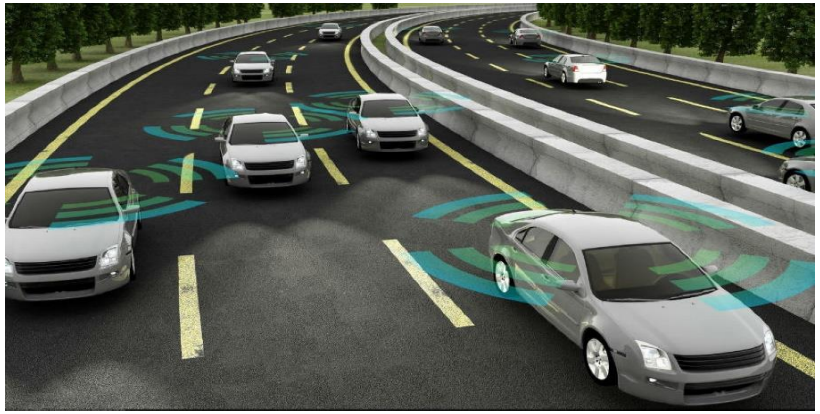
Day 01

Outline

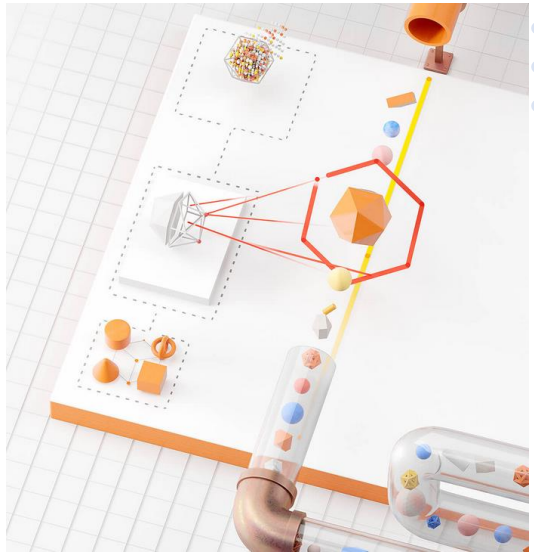
- **Day 01** (28 Aug 2021)
  - Motivations of progress in DNN .....
  - Security of AI
    - Attack .....
    - Defense (in a brief intro) .....
- **Day 02** (29 Aug 2021)
  - Adversarial attack (cont.) .....
  - Adversarial Defense .....
  - Security of AI notes for industrial models .....

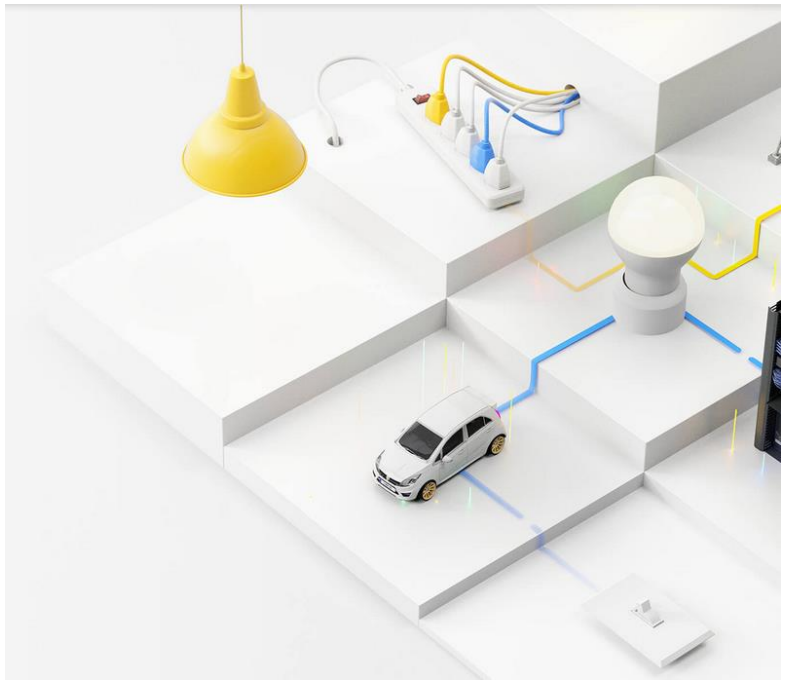
AI and the top innovation

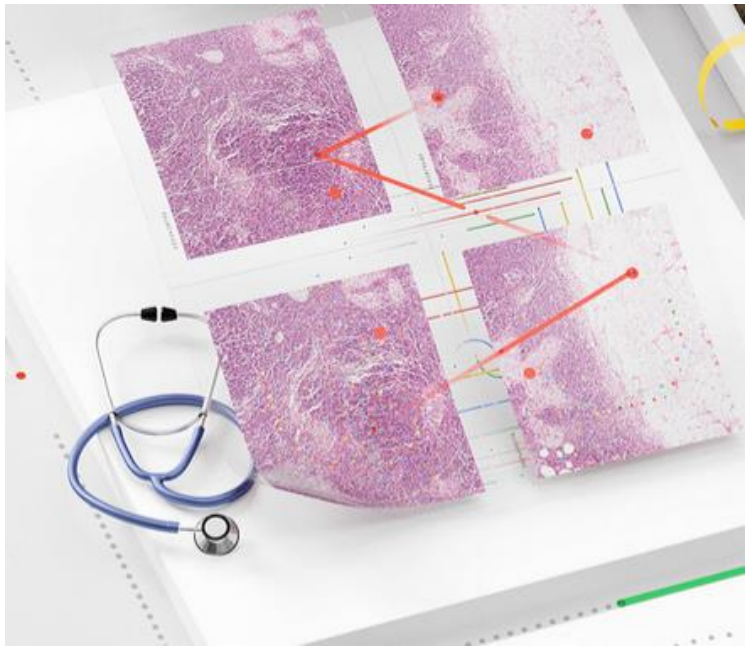


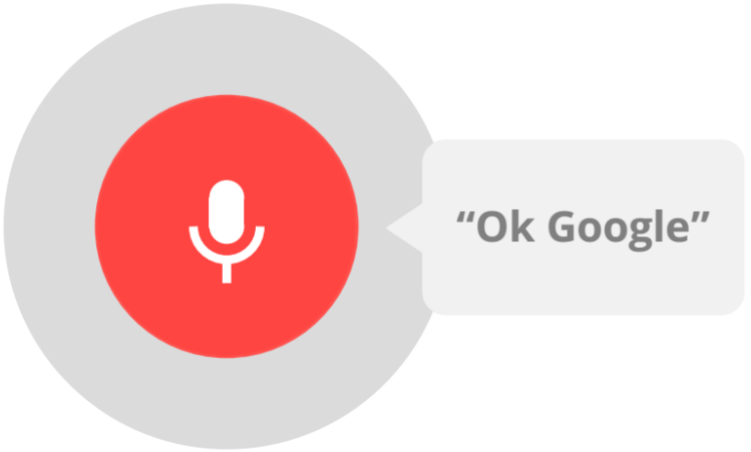


AI and the top innovation









Top Applications of Deep Learning Across Industries

- Self Driving Cars
- News Aggregation and Fraud News Detection
- Natural Language Processing
- Virtual Assistants
- Entertainment
- Visual Recognition
- Fraud Detection
- Healthcare
- Personalisations
- Detecting Developmental Delay in Children
- Colourisation of Black and White images
- Adding sounds to silent movies
- Automatic Machine Translation
- Automatic Handwriting Generation
- Automatic Game Playing
- Language Translations
- Pixel Restoration
- Photo Descriptions
- Demographic and Election Predictions
- Deep Dreaming

Deep learning vs Machine learning (Macro vision)

	Machine learning	Deep learning
Human supervision (overall)	**Required**	Less or Not required
Training time	Seconds or a few **hours**	Hours or a few **weeks**
Number of data points required	**Thousands**	**Millions+**
Computational resources	**Lesser** resources needed	**Massive** resources needed
GPU	Not required	**Required**

https://www.g2.com/articles/deep-learning

Gartner report about artificial intelligence (Hype cycle)

https://www.gartner.com/smarterwithgartner/2-megatrends-dominate-the-gartner-hype-cycle-for-artificial-intelligence-2020/

Breaking news …

DEFENSE

Pentagon actively working to combat adversarial AI

https://www.fedscoop.com › jaic-adverserial-ai-solutions

Pentagon actively working to combat adversarial AI - FedScoop

5 Nov 2020 — The **Pentagon's artificial intelligence** shop is **actively working** on how to secure data and models from a newer threat: **adversarial AI**.

https://citywide.mcleanbible.org › node

Pentagon actively working to combat adversarial AI | ACT-IAC

The **Pentagon's artificial intelligence** shop is **actively working** on how to secure data and models from a newer threat: **adversarial AI**.

https://www.wired.com › Security › ai

The Pentagon's AI Chief Prepares for Battle | WIRED

18 Dec 2019 — Jack Shanahan leads JAIC, the **Pentagon's artificial intelligence** brain trust. ... It's a problem the US military is **actively working** on, ...

https://breakingdefense.com › 2019/11 › exclusive-pent...

EXCLUSIVE Pentagon's AI Problem Is 'Dirty' Data: Lt. Gen ...

13 Nov 2019 — And many military algorithms have to deal with an **adversary** who's **actively trying** to deceive them. The cycle of countermeasure and ...

https://fas.org › sgp › crs › natsec PDF

Artificial Intelligence and National Security

10 Nov 2020 — influential **AI** will be in future **combat** operations. ... 4 Marcus Weisgerber, "The **Pentagon's** New Algorithmic Warfare Cell Gets Its First ...

43 pages

IBM CVPR21 tutorial // Google results

Breaking news …

- **Joint Artificial Intelligence Center**

  An American organization on exploring the usage of
  - Artificial Intelligence (AI)
  - (particularly Edge computing),
  - Network of Networks
  - AI-enhanced communication **for use in actual combat**.

Breaking news …

❑ Joint Artificial Intelligence Center

An American organization on exploring the usage of

- Artificial Intelligence (AI)
- (particularly Edge computing),
- Network of Networks
- AI-enhanced communication **for use in actual combat**.

FEDSCOOP

FedScoop is the leading tech media brand in the federal government market.

- With the **JAIC** deploying and scaling **32** AI products
  - spanning areas like
    - Predictive maintenance operations
    - Cybersecurity
    - Warfighter health
  - across the Department of Defense, the data training these systems is its most valuable intellectual property and in need of securing.

Written by Dave Nyczepir
NOV 5, 2020 | FEDSCOOP

Nand Mulchandani (Chief Technology Officer of the U.S. Department of Defense Joint Artificial Intelligence Center)

https://www.fedscoop.com/jaic-adverserial-ai-solutions/

Breaking news …

❑ Joint Artificial Intelligence Center

An American organization on exploring the usage of

- Artificial Intelligence (AI)
- (particularly Edge computing),
- Network of Networks
- AI-enhanced communication **for use in actual combat**.

FEDSCOOP

FedScoop is the leading tech media brand in the federal government market.

- **Adversarial AI** is an area where we need to do **a lot of innovation around it**
- and create new principles and new processes and methodologies to address it

Cheryl Ingstad (Director of DOE's Artificial Intelligence & Technology Office)

Written by Dave Nyczepir

NOV 5, 2020 | FEDSCOOP

https://www.fedscoop.com/jaic-adverserial-ai-solutions/

Gartner report about Adversarial AI

Gartner

Top 10 Strategic Technology Trends for 2020

Published: 21 October 2019 **ID:** G00432920

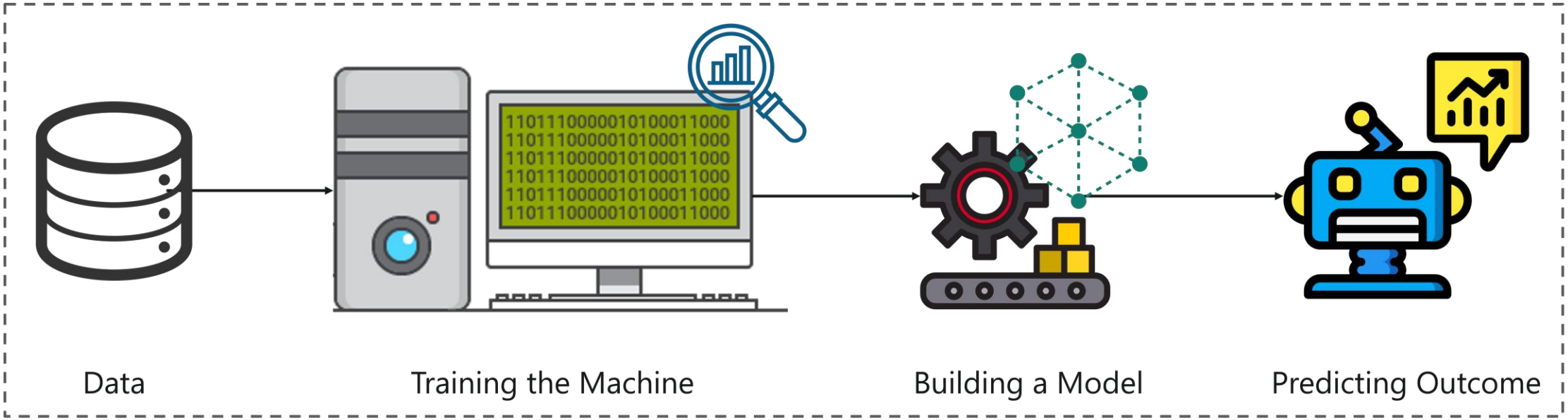
Analyst(s): David Cearley, Nick Jones, David Smith, Brian Burke, Arun Chandrasekaran, CK Lu

- Through 2022, 30% of all **AI cyberattacks** will leverage
  - Training-data poisoning
  - AI model theft
  - Adversarial samples

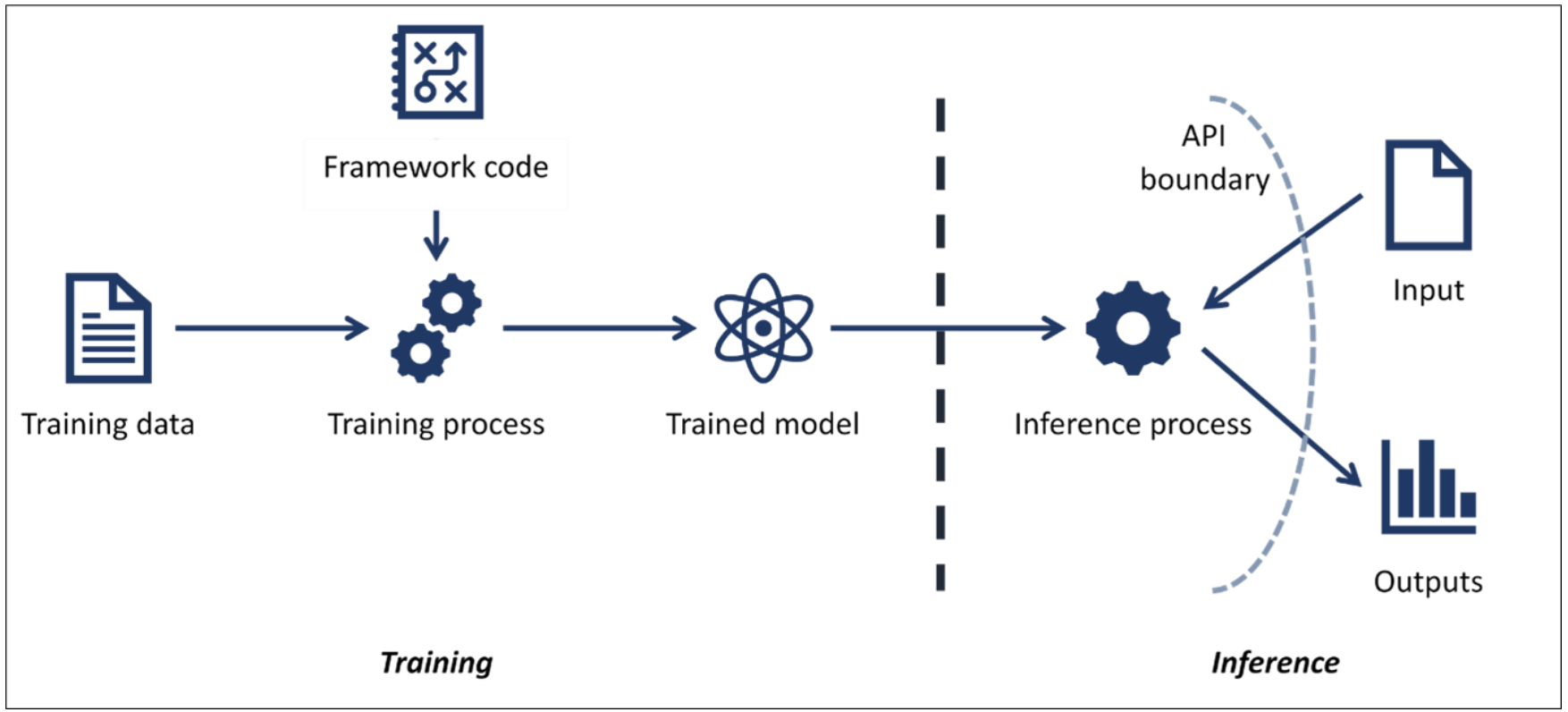
… to attack AI-powered systems


Break

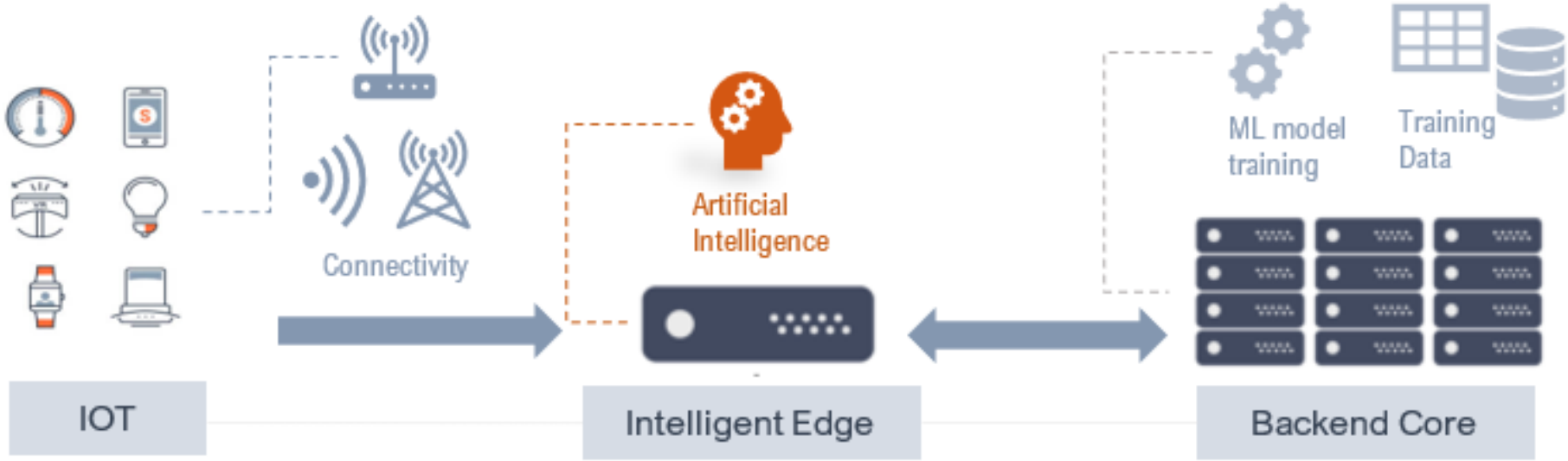
Machine Learning model



ML Model Application Protocol Interface

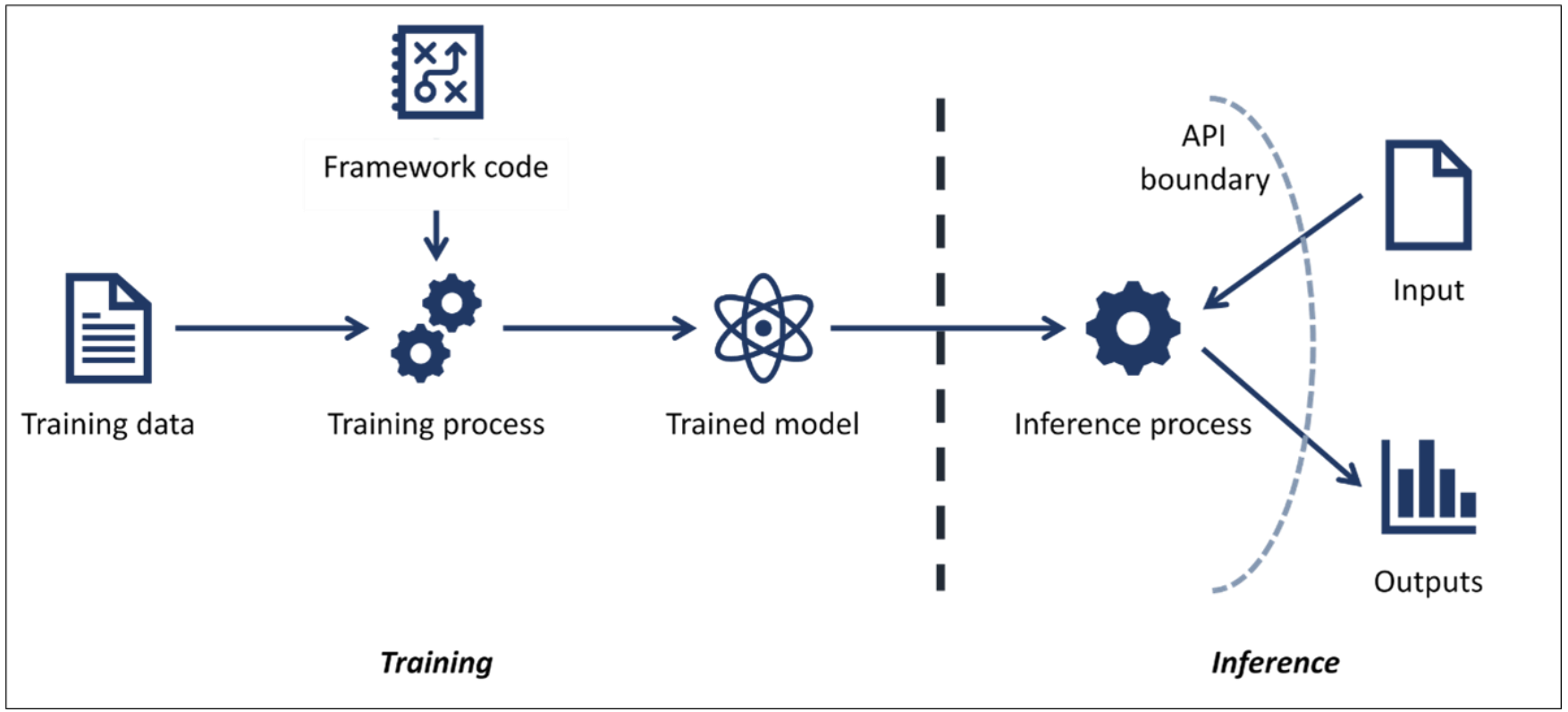


ML model API

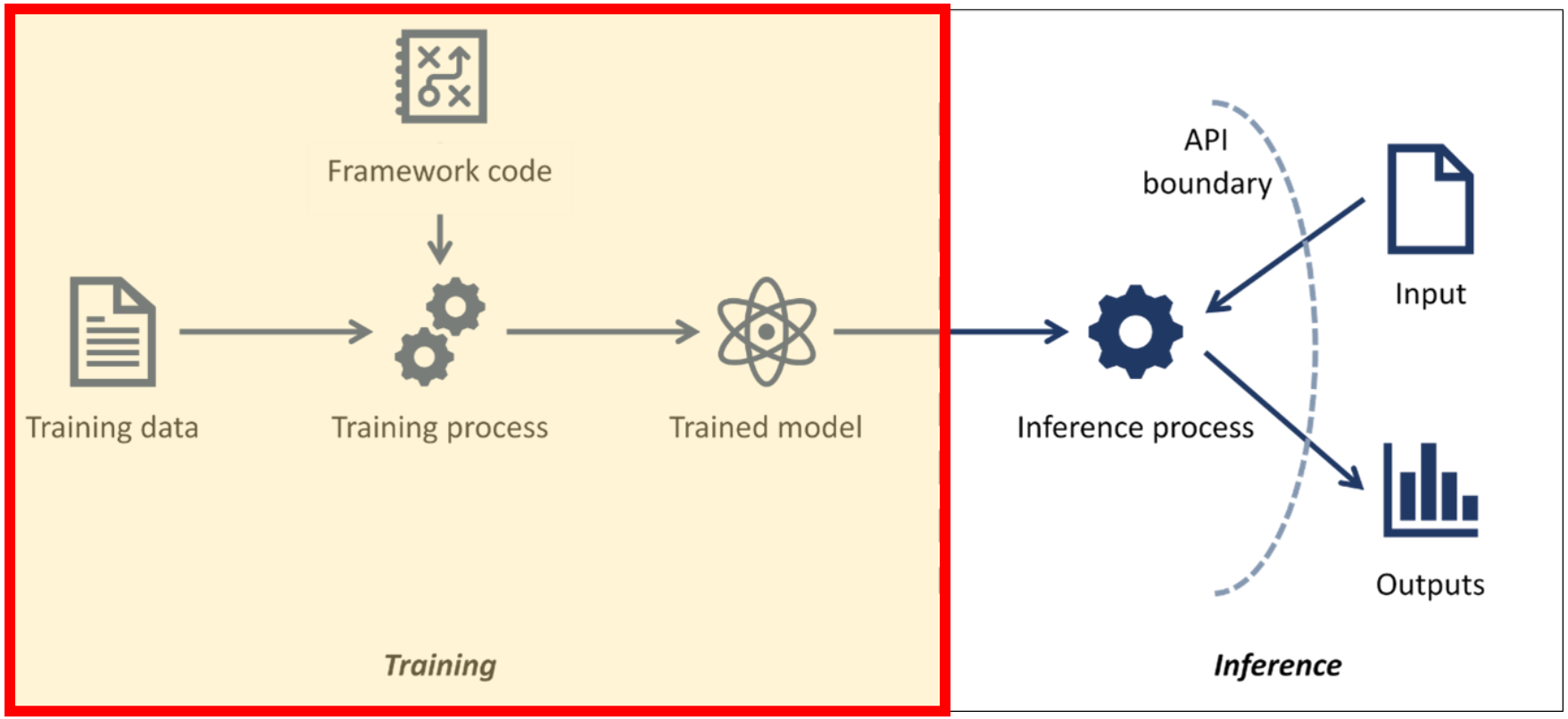


https://medium.com/swlh/machine-learning-model-optimization-for-intelligent-edge-d0f400111002

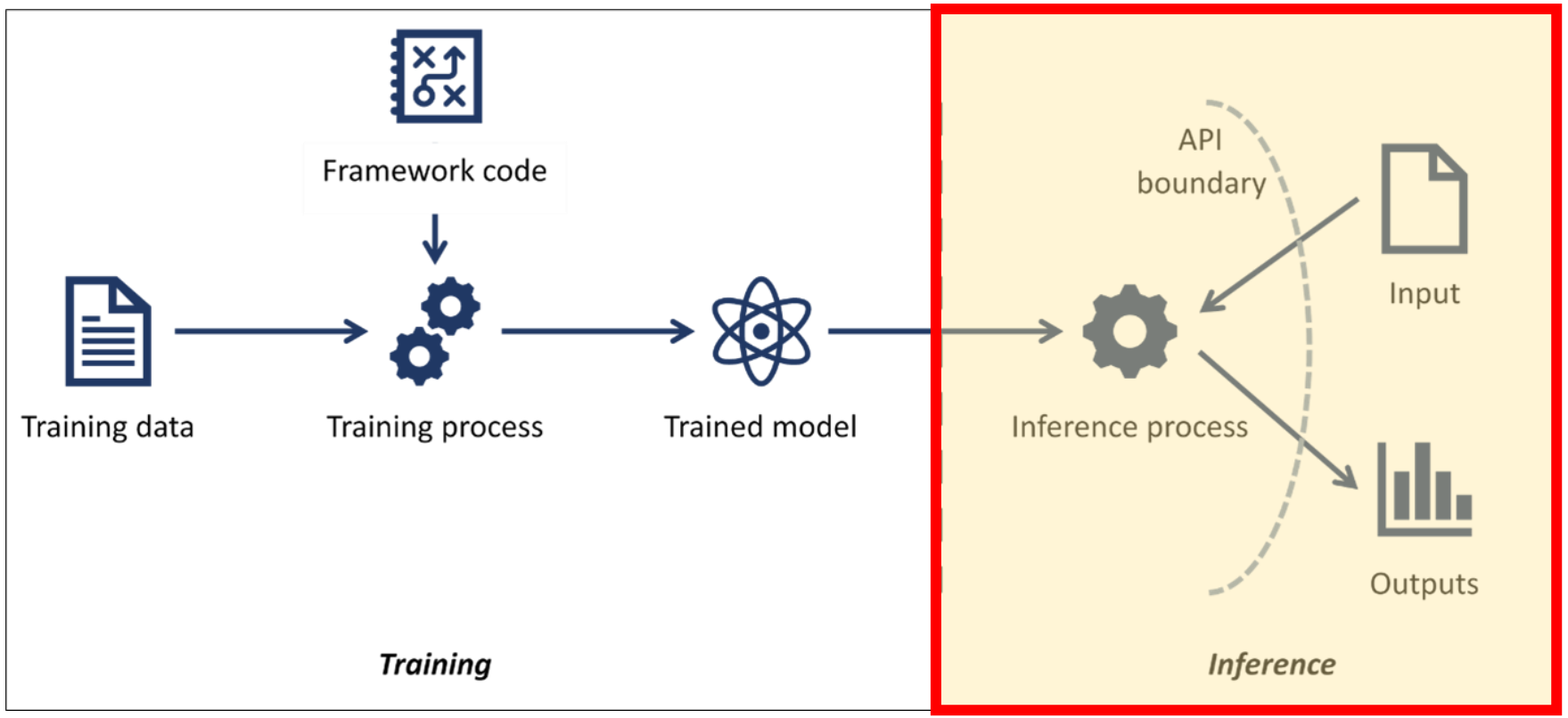
ML Model API


Framework code
Training data
Training process
Trained model
Inference process
API
boundary
Input
Outputs
Training
Inference

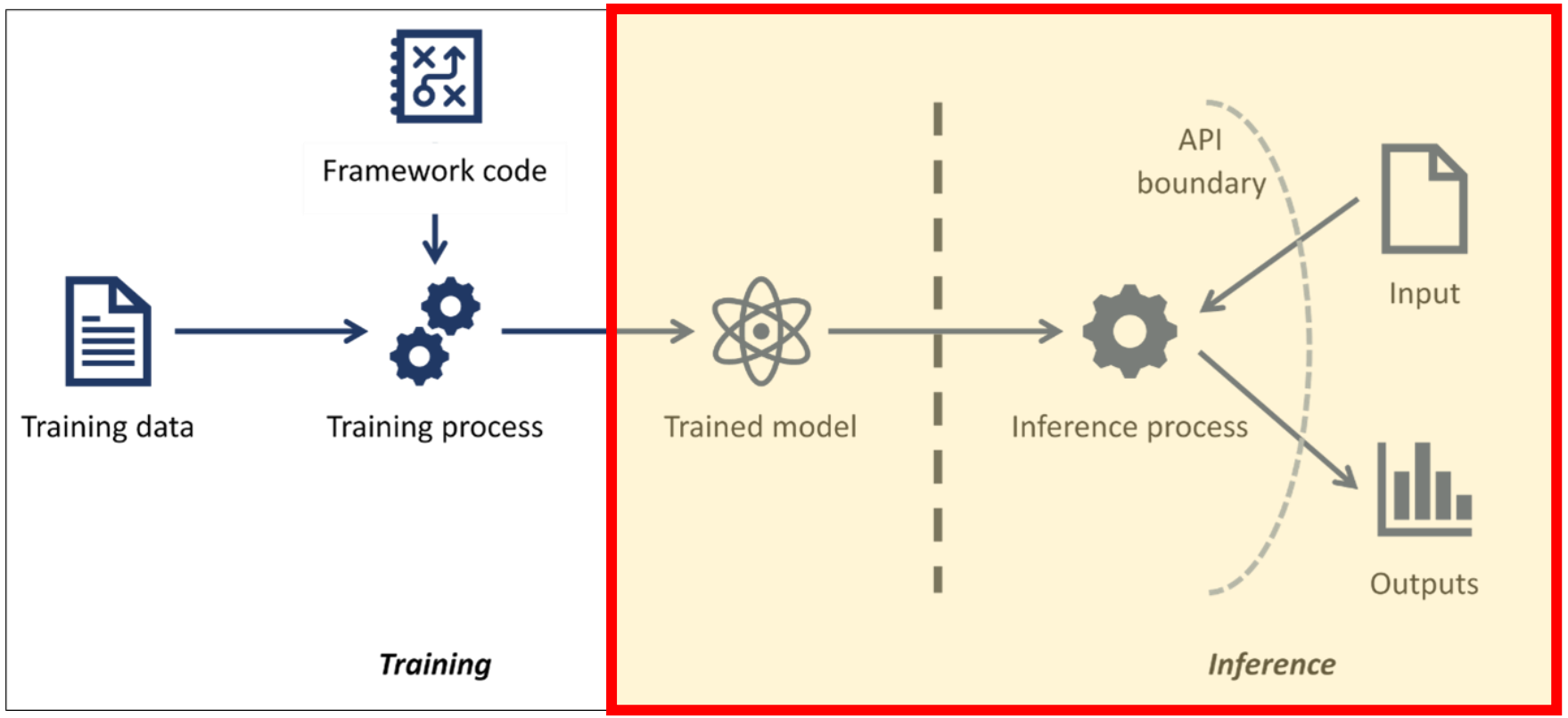
API based AI Model



API based AI Model

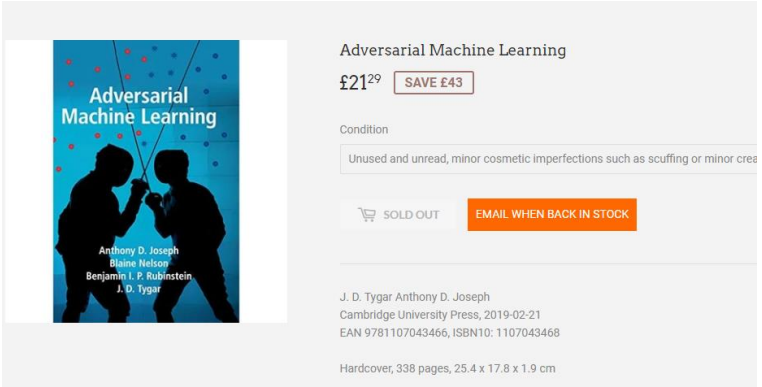


API based AI Model



Adversarial Machine Learning

Adversarial machine learning is a machine learning technique that attempts to fool models by supplying deceptive input.



ATLAS

- Informally, Adversarial ML is "subverting machine learning systems for fun and profit".
- The methods underpinning the production machine learning systems are systematically vulnerable to a new class of vulnerabilities across the machine learning supply chain collectively known as Adversarial Machine Learning.

https://en.wikipedia.org/wiki/Adversarial_machine_learning

https://github.com/mitre/advmlthreatmatrix/blob/master/pages/adversarial-ml-101.md#adversarial-machine-learning-101

ATLAS : Adversarial ML Threat Matrix

- In the last three years, major companies such as Google, Amazon, Microsoft, and Tesla, have had their ML systems tricked, evaded, or misled.

- This trend is only set to rise: According to a Gartner report. 30% of cyberattacks by 2022 will involve data poisoning, model theft or adversarial examples.

- Industry is underprepared. In a survey of 28 organizations spanning small as well as large organizations, 25 organizations did not know how to secure their ML systems.

https://github.com/mitre/advmlthreatmatrix

ATLAS : Adversarial ML Threat Matrix

❑ Case Studies

- Evasion of Deep Learning detector for malware C&C traffic
- Botnet Domain Generation Algorithm (DGA) Detection Evasion
- VirusTotal Poisoning
- Bypassing Cylance's AI Malware Detection
- Camera Hijack Attack on Facial Recognition System
- Attack on Machine Translation Service - Google Translate, Bing Translator, and Systran Translate
- ClearviewAI Misconfiguration
- GPT-2 Model Replication
- ProofPoint Evasion
- Tay Poisoning
- Microsoft - Azure Service - Evasion
- Microsoft Edge AI - Evasion
- MITRE - Physical Adversarial Attack on Face Identification

https://github.com/mitre/advmlthreatmatrix

ATLAS : Adversarial ML Threat Matrix

- The case-studies were selected because of the impact to production ML systems, and each demonstrates one of the following characteristics.

- **Range of Attacks:**
  - Evasion
  - Poisoning
  - Model replication
  - Exploiting traditional software flaws

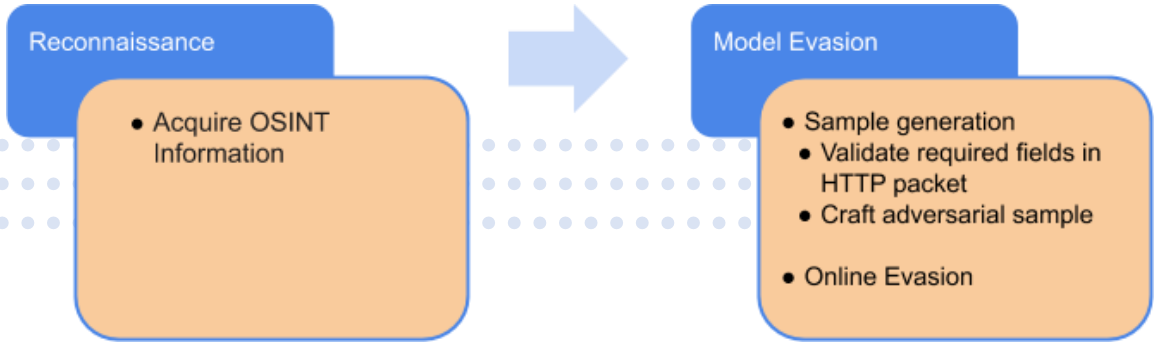
- **Range of Personas:**
  - Average user
  - Security researchers
  - ML Researchers
  - Fully equipped Red team

- **Range of ML Paradigms:**
  - Attacks on MLaaS
  - ML models
    - hosted on cloud
    - hosted on-premise
    - ML models on edge.

- **Range of Use case:**
  - Attacks on ML systems used in both
    - "security-sensitive" applications like cybersecurity and
    - non-security-sensitive applications like chatbots.

https://github.com/mitre/advmlthreatmatrix

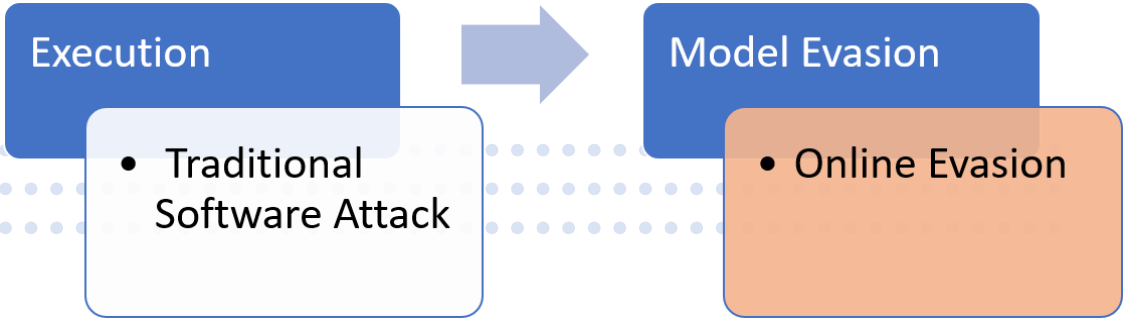
ATLAS : Adversarial ML Threat Matrix



- **Evasion of Deep Learning detector for malware C&C traffic**
  - The team trained the model on ~ 33 million benign and ~ 27 million malicious HTTP packet headers
  - Evaluation showed a true positive rate of ~ 99% and false positive rate of ~0.01%, on average
  - Testing the model with a HTTP packet header from known malware command and control traffic samples was detected as malicious with high confidence (> 99%).
  - The attackers crafted evasion samples by removing fields from packet header which are typically not used for C&C communication (e.g. cache-control, connection, etc.)
  - With the crafted samples the attackers performed online evasion of the ML based spyware detection model. The crafted packets were identified as benign with >80% confidence.
  - This evaluation demonstrates that adversaries are able to bypass advanced ML detection techniques, by crafting samples that are misclassified by an ML model.

https://github.com/mitre/advmlthreatmatrix

ATLAS : Adversarial ML Threat Matrix

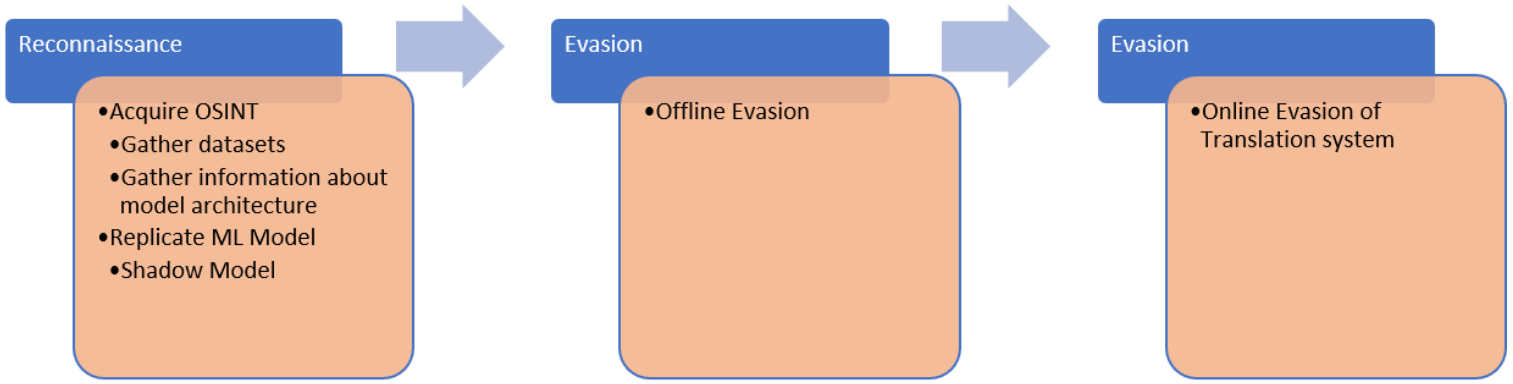


- **Camera Hijack Attack on Facial Recognition System**
  - The attackers bought customized low-end mobile phones, customized android ROMs, a specific virtual camera application, identity information and face photos.
  - The attackers used software to turn static photos into videos, adding realistic effects such as blinking eyes.
  - Then the attackers use the purchased low-end mobile phone to import the generated video into the virtual camera app.
  - The attackers registered an account with the victim's identity information.
  - In the verification phase, the face recognition system called the camera API, but because the system was hooked or rooted, the video stream given to the face recognition system was actually provided by the virtual camera app.
  - The attackers successfully evaded the face recognition system and impersonated the victim.

https://github.com/mitre/advmlthreatmatrix

ATLAS : Adversarial ML Threat Matrix

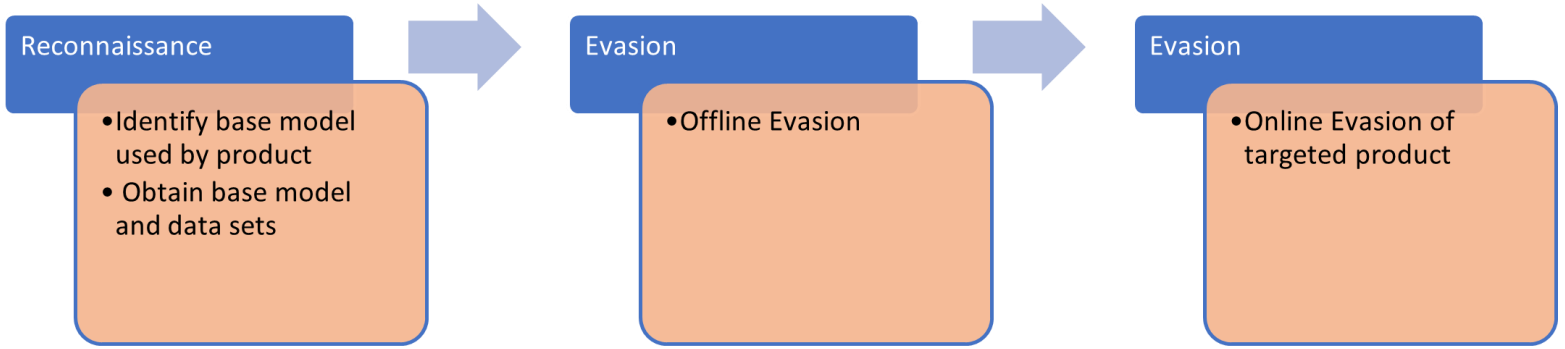
- **Attack on Machine Translation Service - Google Translate, Bing Translator, and Systran Translate**
  - Using published research papers, the researchers gathered similar datasets and model architectures that these translation services used
  - They abuse a public facing application to query the model and produce machine translated sentence pairs as training data
  - Using these translated sentence pairs, researchers trained a substitute model (model replication)
  - The replicated models were used to construct offline adversarial examples that successfully transferred to an online evasion attack



https://github.com/mitre/advmlthreatmatrix

ATLAS : Adversarial ML Threat Matrix

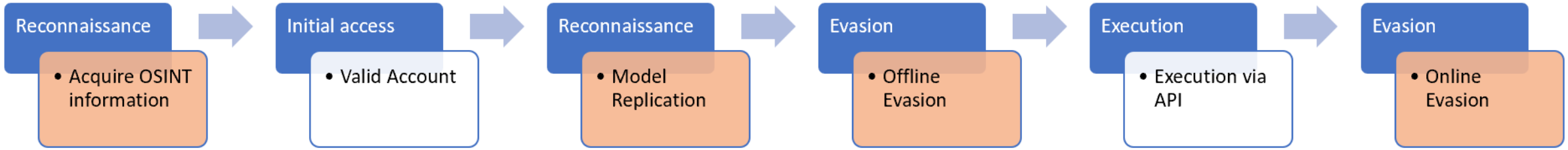
- **Microsoft - Edge AI**
  - The team first performed reconnaissance to gather information about the target ML model.
  - Then, used a publicly available version of the ML model, started sending queries and analyzing the responses (inferences) from the ML model.
  - Using this, the red team created an automated system that continuously manipulated an original target image, that tricked the ML model into producing incorrect inferences, but the perturbations in the image were unnoticeable to the human eye.
  - Feeding this perturbed image, the red team was able to evade the ML model into misclassifying the input image.
  - This operation had one step in the traditional MITRE ATT&CK techniques to do reconnaissance on the ML model being used in the product, and then the rest of the techniques was to use Offline evasion, followed by online evasion of the targeted product.



https://github.com/mitre/advmlthreatmatrix

ATLAS : Adversarial ML Threat Matrix

- **MITRE - Physical Adversarial Attack on Face Identification**
  - The team first performed reconnaissance to gather information about the target ML model.
  - Using a valid account, the team identified the list of IDs targeted by the model.
  - The team developed a proxy model using open source data.
  - Using the proxy model, the red team optimized a physical domain patch-based attack using an expectation of transformations.
  - Via an exposed API interface, the team performed an online physical-domain evasion attack including the adversarial patch in the input stream which resulted in a targeted misclassification.
  - This operation had a combination of traditional ATT&CK enterprise techniques such as **finding Valid account**, and **Executing code via an API** – all interleaved with adversarial ML specific attacks.



https://github.com/mitre/advmlthreatmatrix

ATLAS : Adversarial ML Threat Matrix

❑ **Case Studies**

- Evasion of Deep Learning detector for malware C&C traffic
- Botnet Domain Generation Algorithm (DGA) Detection Evasion
- VirusTotal Poisoning
- Bypassing Cylance's AI Malware Detection
- Camera Hijack Attack on Facial Recognition System
- Attack on Machine Translation Service - Google Translate, Bing Translator, and Systran Translate
- ClearviewAI Misconfiguration
- GPT-2 Model Replication
- ProofPoint Evasion
- Tay Poisoning
- Microsoft - Azure Service - Evasion
- Microsoft Edge AI - Evasion
- MITRE - Physical Adversarial Attack on Face Identification

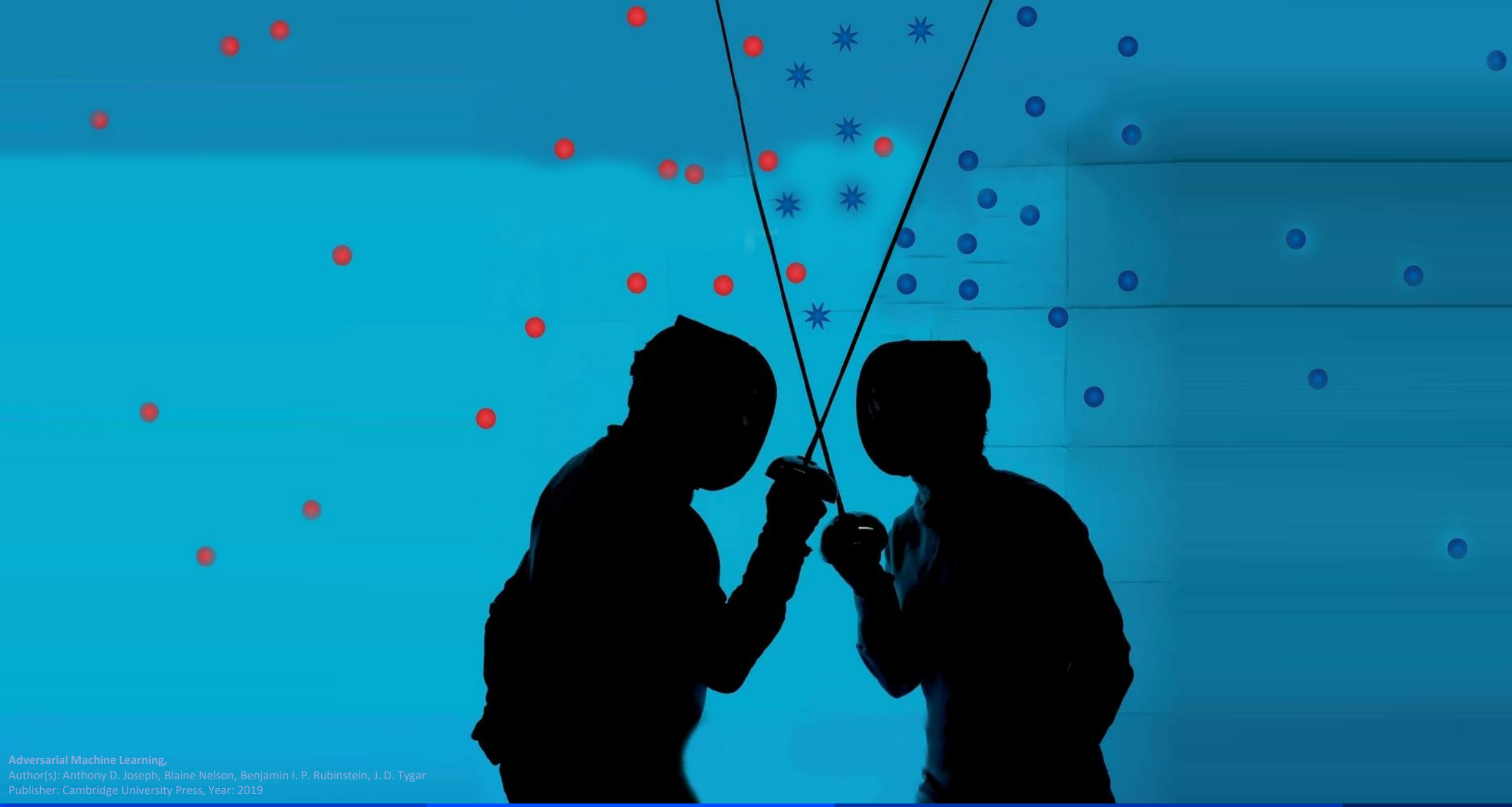
https://github.com/mitre/advmlthreatmatrix

ATLAS : Adversarial ML Threat Matrix

Reconnaissance	Initial Access	Execution	Persistence	Model Evasion	Exfiltration	Impact
Acquire OSINT information: (Sub Techniques) 1. Arxiv 2. Public blogs 3. Press Releases 4. Conference Proceedings 5. Github Repository 6. Tweets	Pre-trained ML model with backdoor	Execute unsafe ML models (Sub Techniques) 1. ML models from compromised sources 2. Pickle embedding	Execute unsafe ML models (Sub Techniques) 1. ML models from compromised sources 2. Pickle embedding	Evasion Attack (Sub Techniques) 1. Offline Evasion 2. Online Evasion	Exfiltrate Training Data (Sub Techniques) 1. Membership inference attack 2. Model inversion	Defacement
ML Model Discovery (Sub Techniques) 1. Reveal ML model ontology – 2. Reveal ML model family –	Valid account	Execution via API	Account Manipulation		Model Stealing	Denial of Service
Gathering datasets	Phishing	Traditional Software attacks	Implant Container Image	Model Poisoning	Insecure Storage 1. Model File 2. Training data	Stolen Intellectual Property
Exploit physcial environment	External remote services			Data Poisoning (Sub Techniques) 1. Tainting data from acquisition – Label corruption 2. Tainting data from open source supply chains 3. Tainting data from acquisition – Chaff data 4. Tainting data in training environment – Label corruption		Data Encrypted for Impact Defacement
Model Replication (Sub Techniques) 1. Exploit API – Shadow Model 2. Alter publicly available, pre-trained weights	Exploit public facing application					Stop System Shutdown/Reboot
Model Stealing	Trusted Relationship					

https://github.com/mitre/advmlthreatmatrix


Break



Adversarial Machine Learning,
Author(s): Anthony D. Joseph, Blaine Nelson, Benjamin I. P. Rubinstein, J. D. Tygar
Publisher: Cambridge University Press, Year: 2019

Day 01

Outline

- **Day 01** (28 Aug 2021)
  - Motivations of progress in DNN
  - Security of AI
    - Attack
    - Defense (in a brief intro)
- **Day 02** (29 Aug 2021)
  - Adversarial attack (cont.)
  - Adversarial Defense
  - Security of AI notes for industrial models

Motivations …

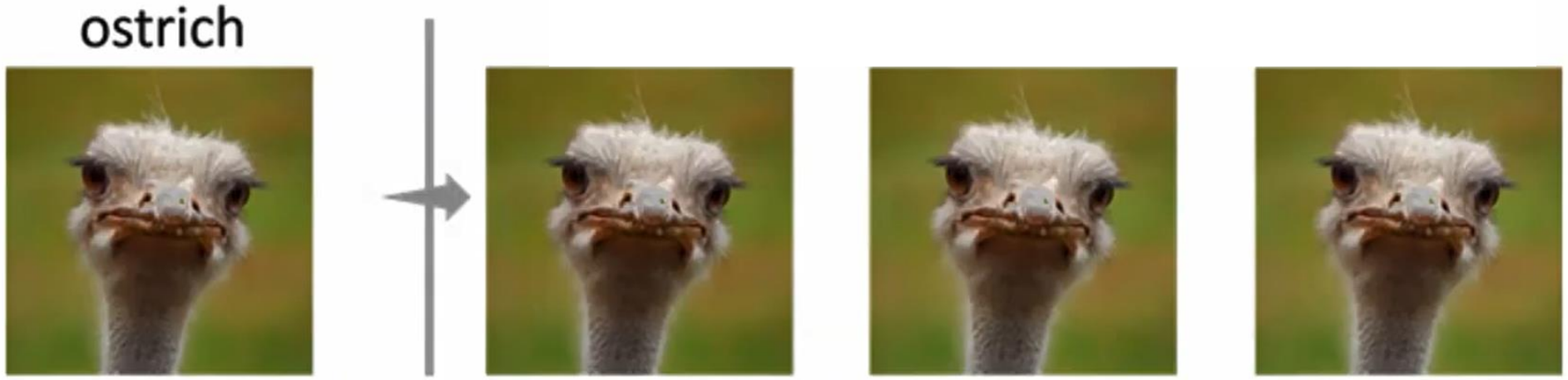
Suppose that we have

- Best trained **image classifier model** (using neural networks)

Motivations …

Suppose that we have

- Best trained **image classifier model** (using neural networks)



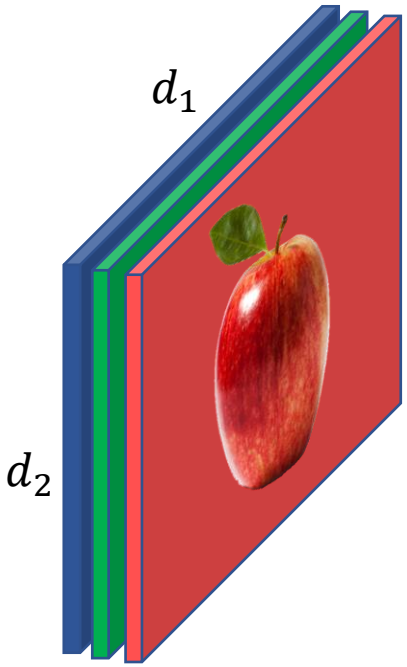
Motivations …

Suppose that we have

- Best trained **image classifier model** (using neural networks)

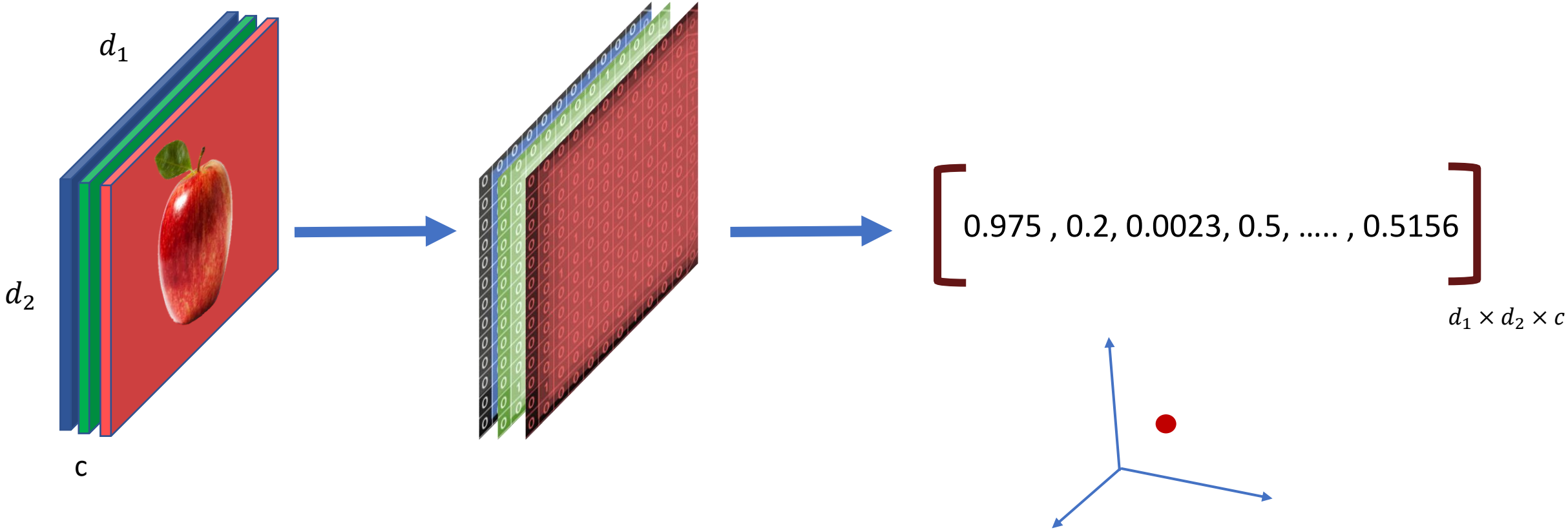
What is wrong with this AI model?

Review::



https://www.slideshare.net/BrianKim244/dcgan-77452250

Review::



https://www.slideshare.net/BrianKim244/dcgan-77452250

Review::

Review::

https://www.slideshare.net/BrianKim244/dcgan-77452250

Review::

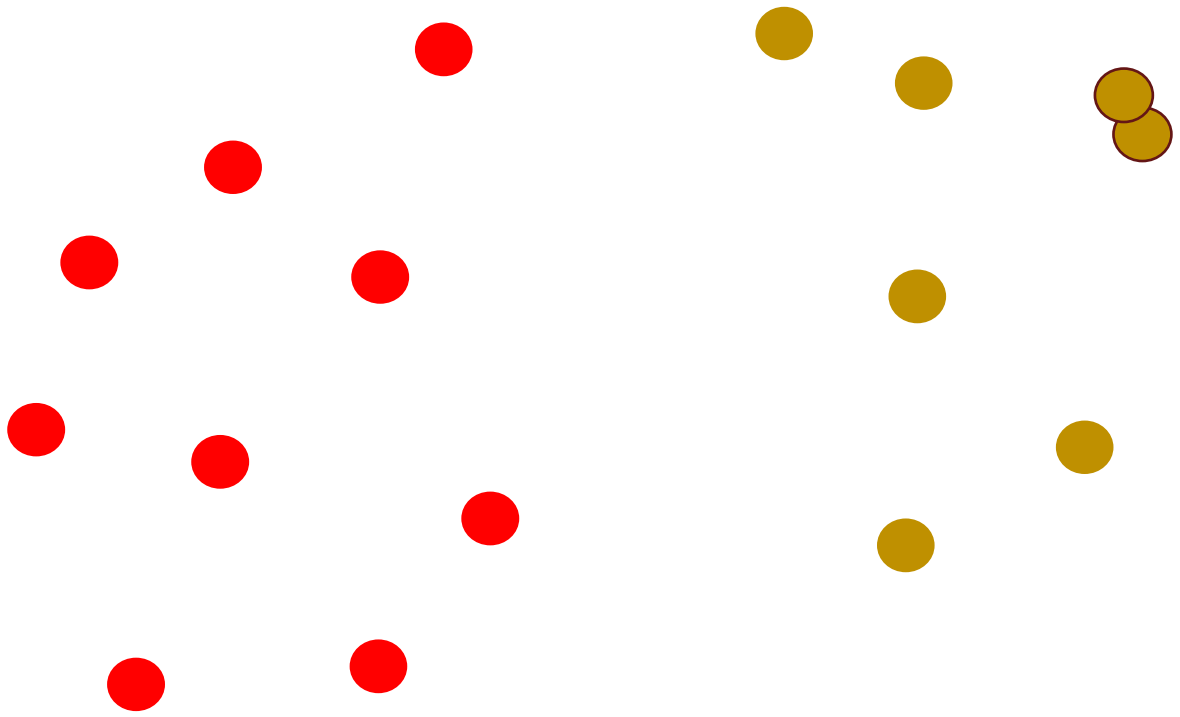
https://ceit.aut.ac.ir/~khalooei/presentations/
https://www.slideshare.net/khalooei/life-of-points-machine-learning-with-orange-flavor

https://www.slideshare.net/khalooei/life-of-points-machine-learning-with-orange-flavor

Review::

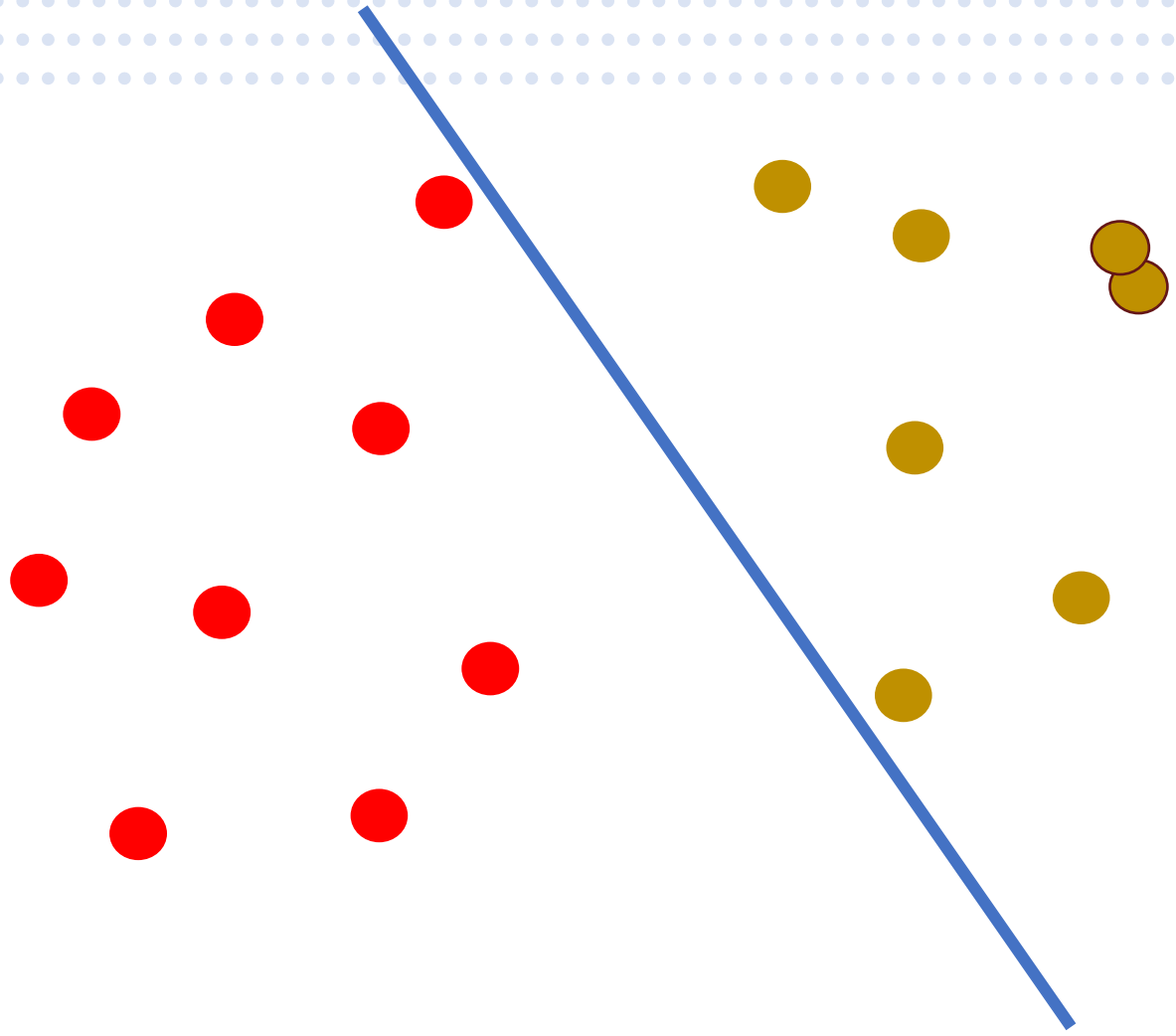
https://www.slideshare.net/BrianKim244/dcgan-77452250

Review::



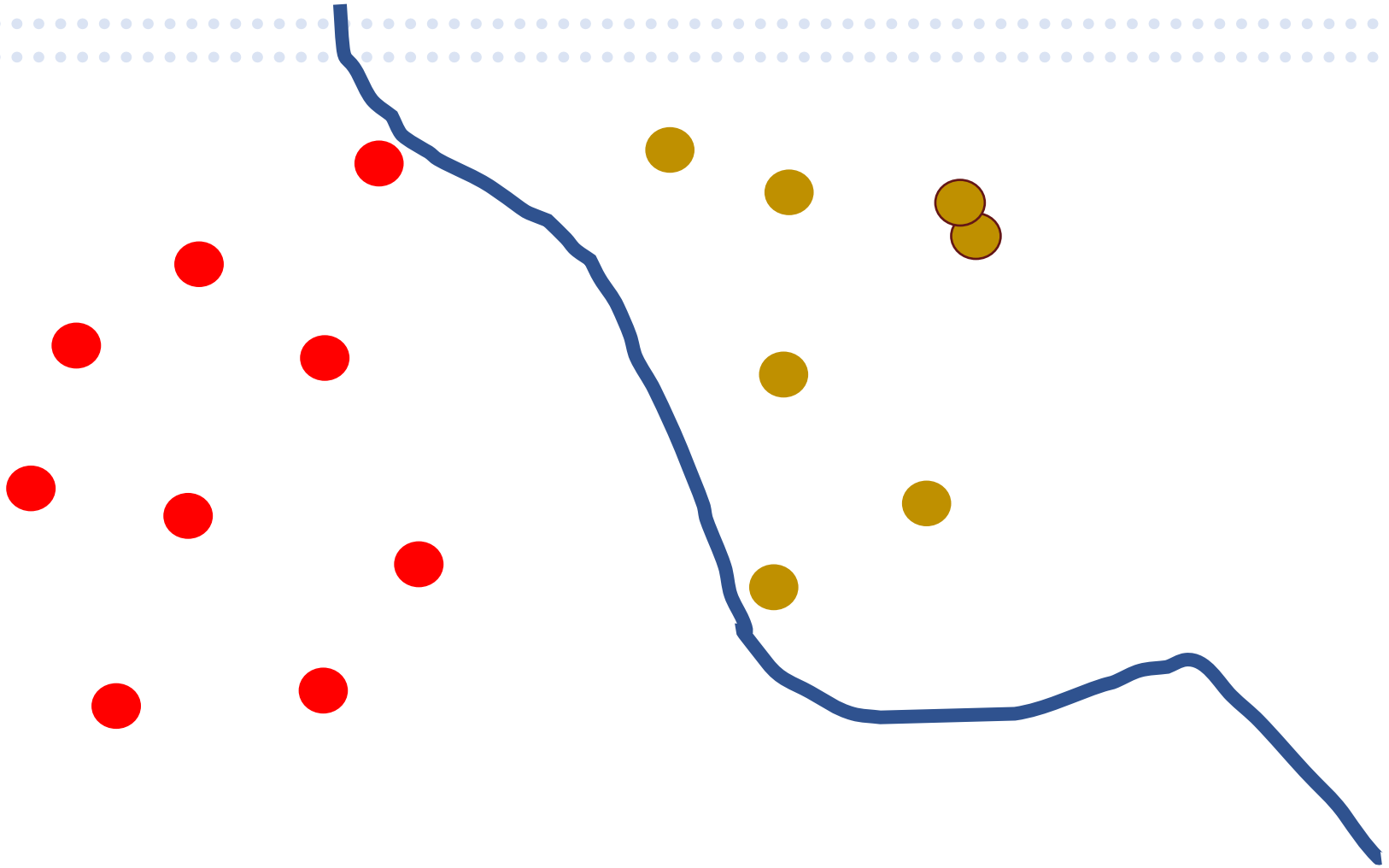
https://www.slideshare.net/BrianKim244/dcgan-77452250

Review::



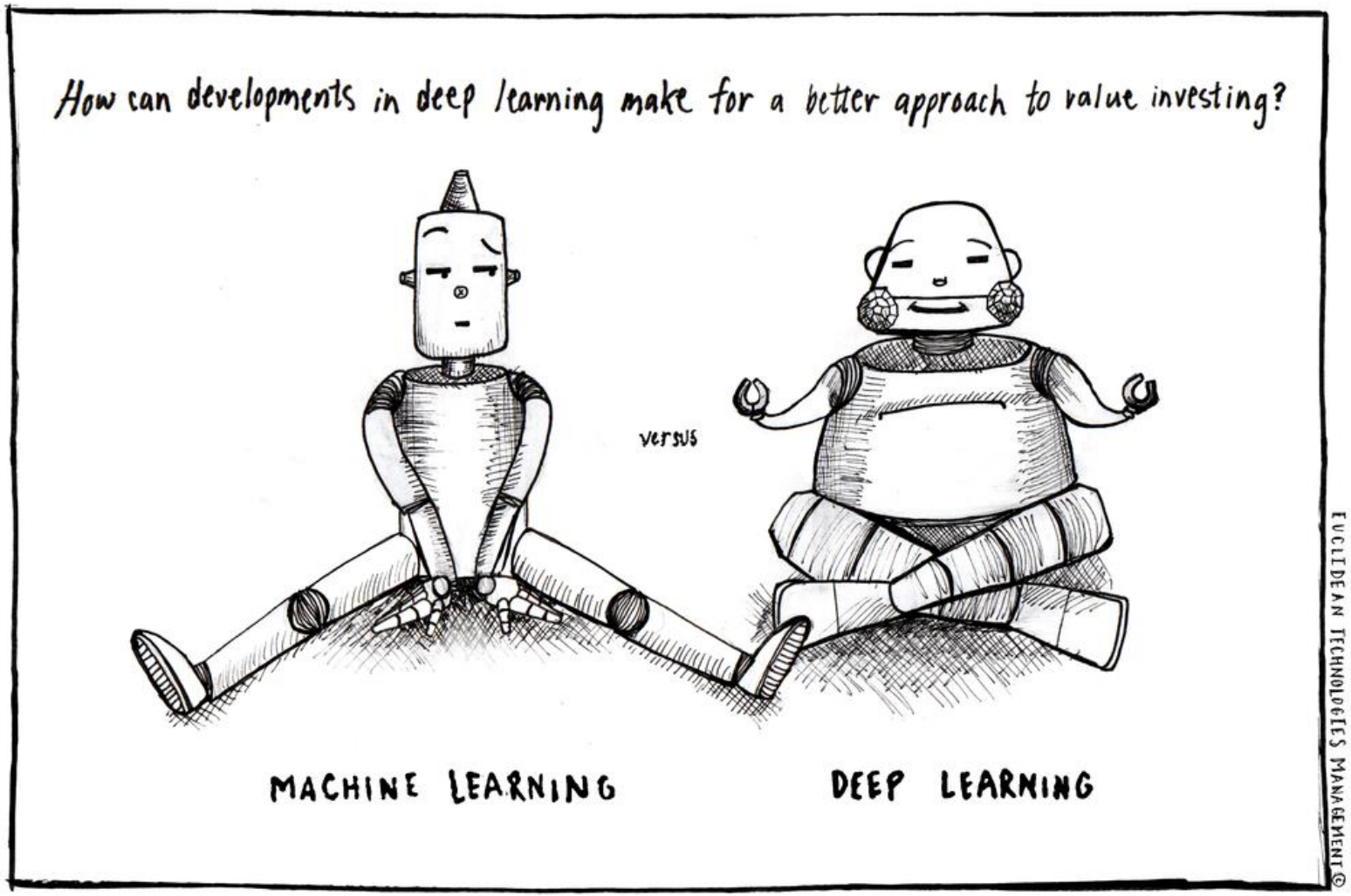
https://www.slideshare.net/BrianKim244/dcgan-77452250

Review::

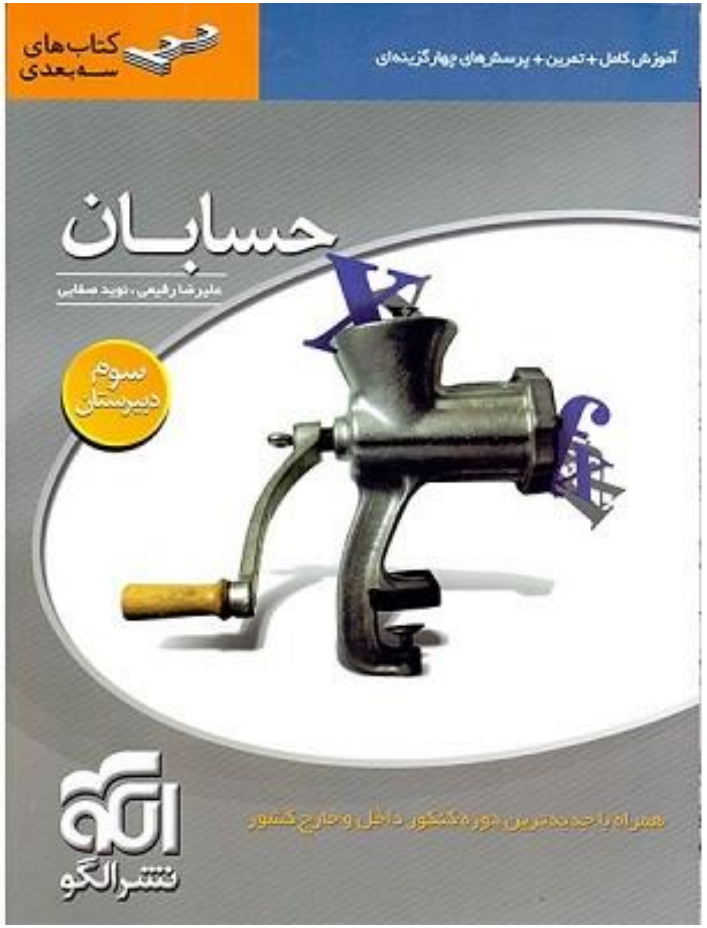


https://www.slideshare.net/BrianKim244/dcgan-77452250

Motivations



Motivations



Motivations

what is the

biggest risk

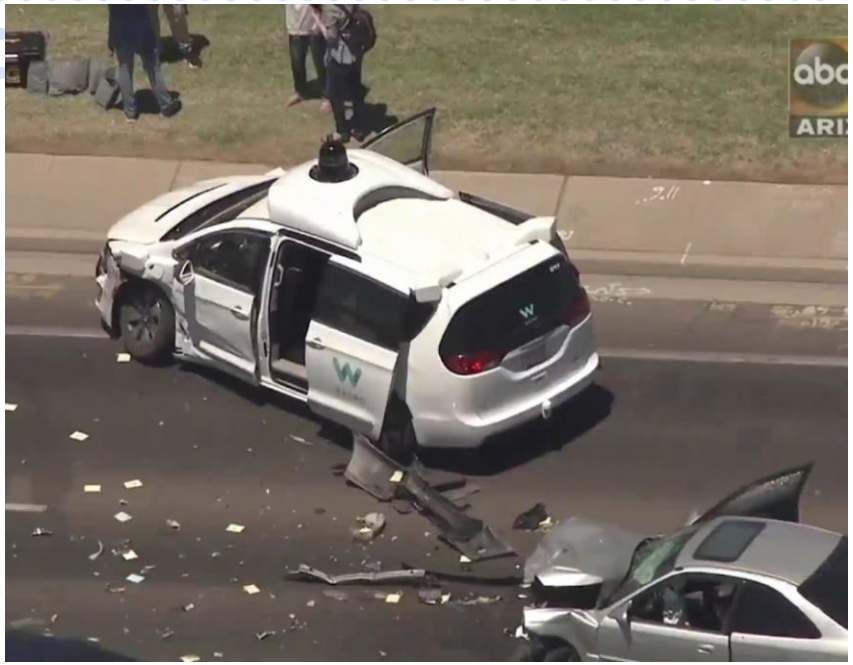
introduced by runtime reliance

of deployed deep learning

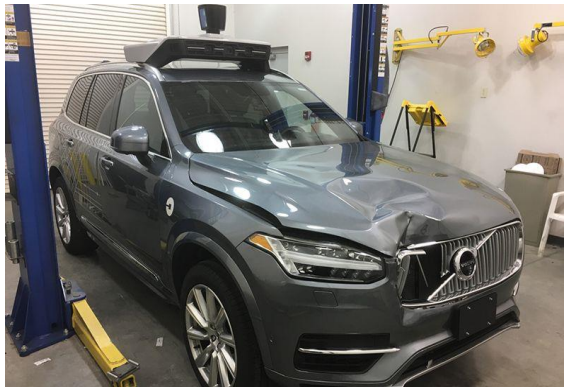
Motivations

https://blog.piekniewski.info/2019/02/16/a-v-safety-2018-update-2/

https://www.tweaktown.com/news/56854/uber-suspends-self-driving-car-tech-arizona-crash/index.html



https://www.newsweek.com/self-driving-vehicle-accident-

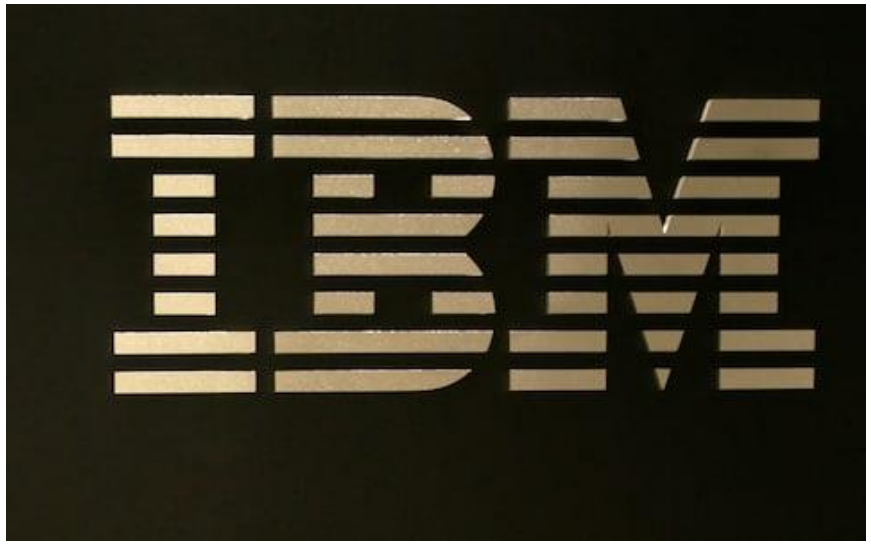


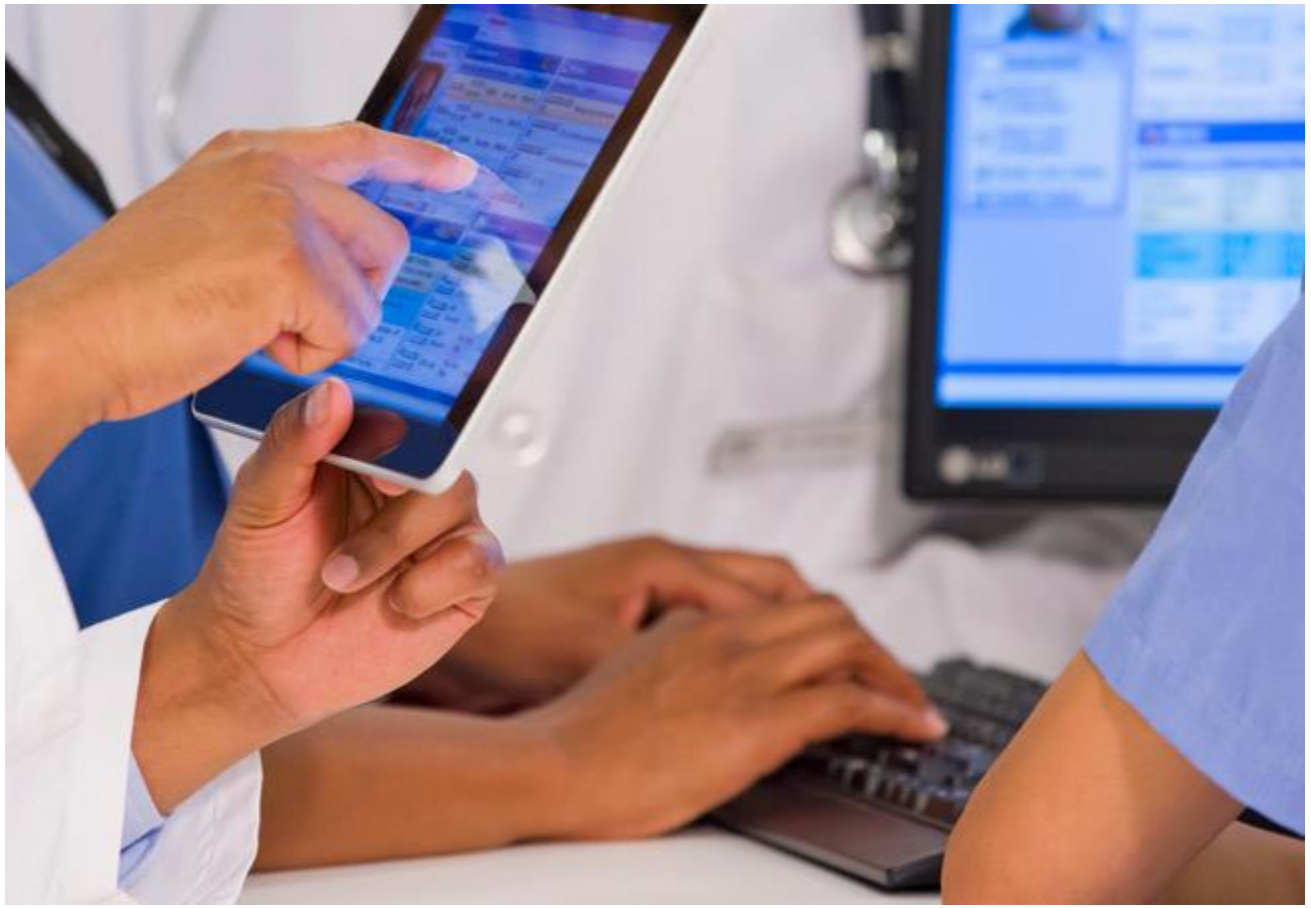
https://www.rac.co.uk/drive/news/motoring-news/uber-self-driving-car-that-killed-pedestrian-had-software-

https://www.bbc.co.uk/news/world-us-canada-43604440

Motivations

- IBM’s Watson supercomputer recommended ‘unsafe and incorrect’ cancer treatments, internal documents show

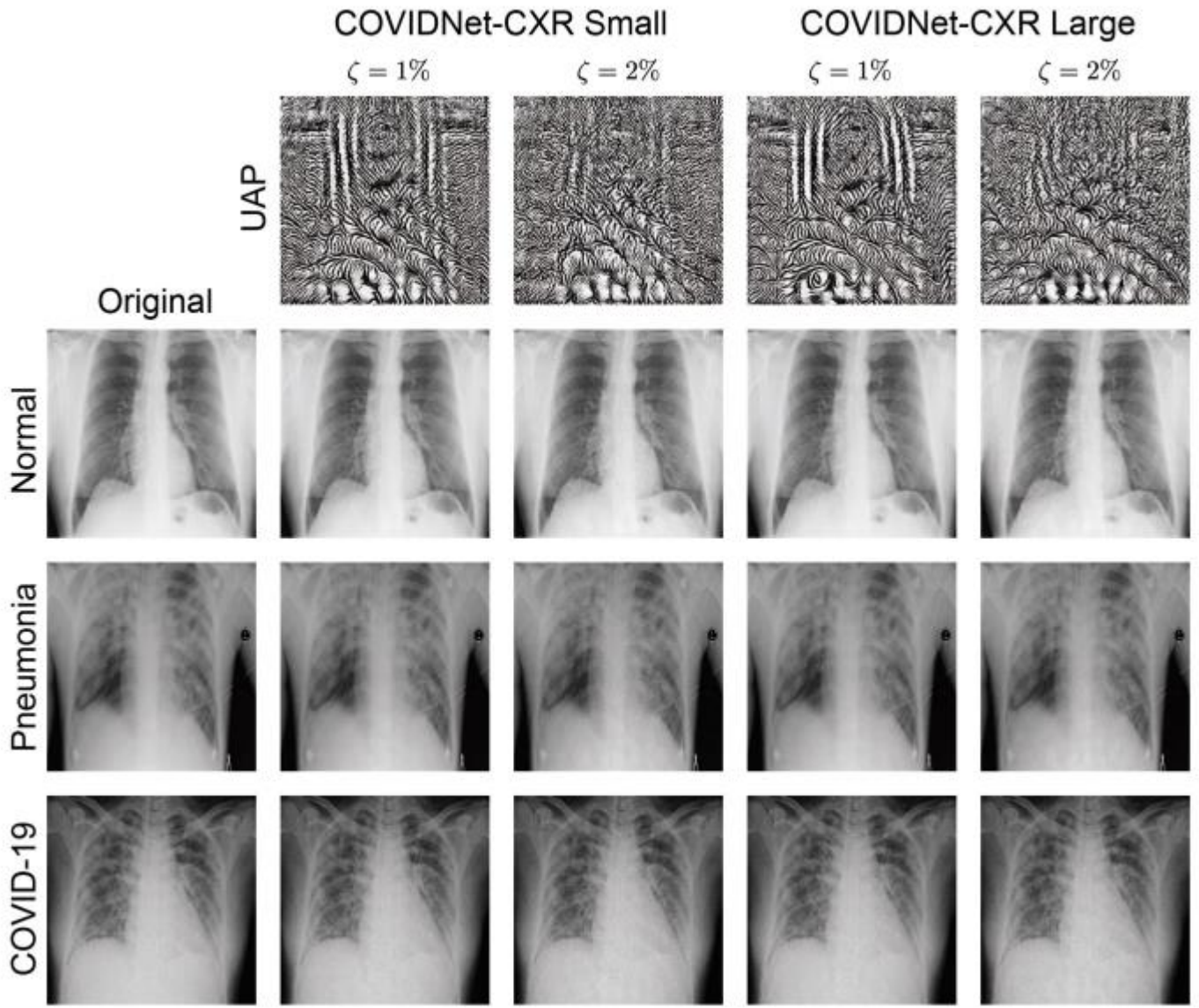




https://www.fiercehealthcare.com/tech/ibm-watson-health-says-ai-making-progress-clinical-decision-support-for-cancer-care

Motivations

- Vulnerability of deep neural networks for detecting COVID-19 cases from chest X-ray images to universal adversarial attacks



https://arxiv.org/abs/2005.11061

Motivations

December 20 '17

Fooling Google's image-recognition AI 1000x faster

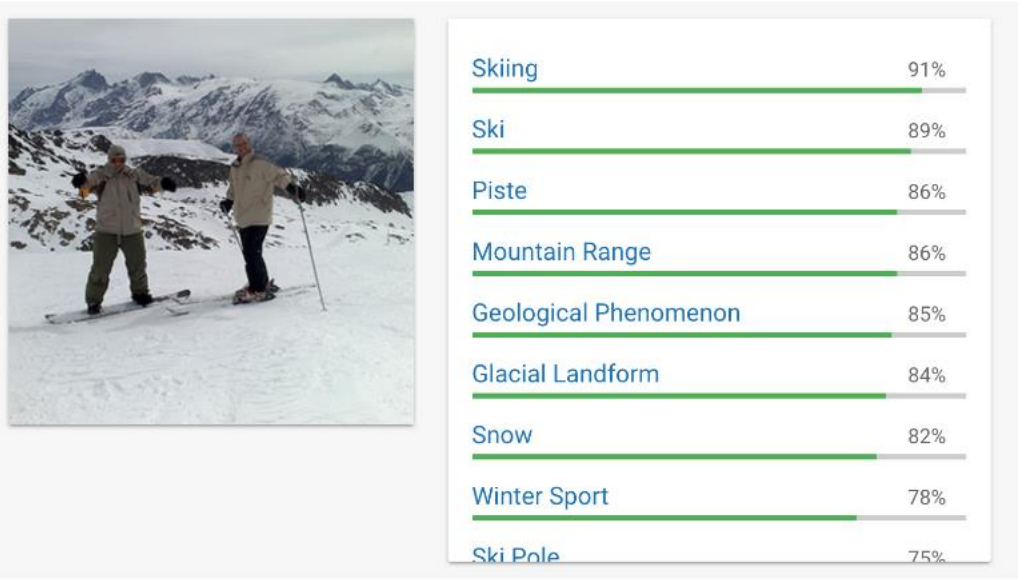
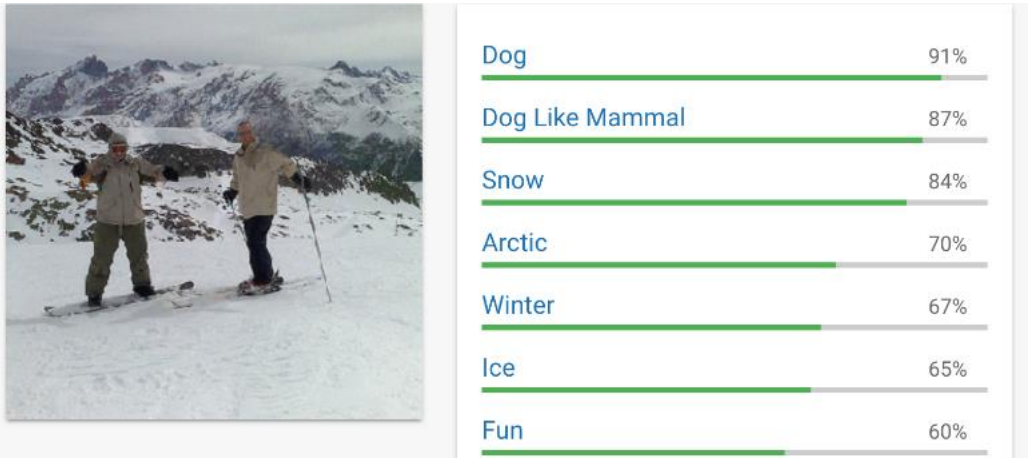
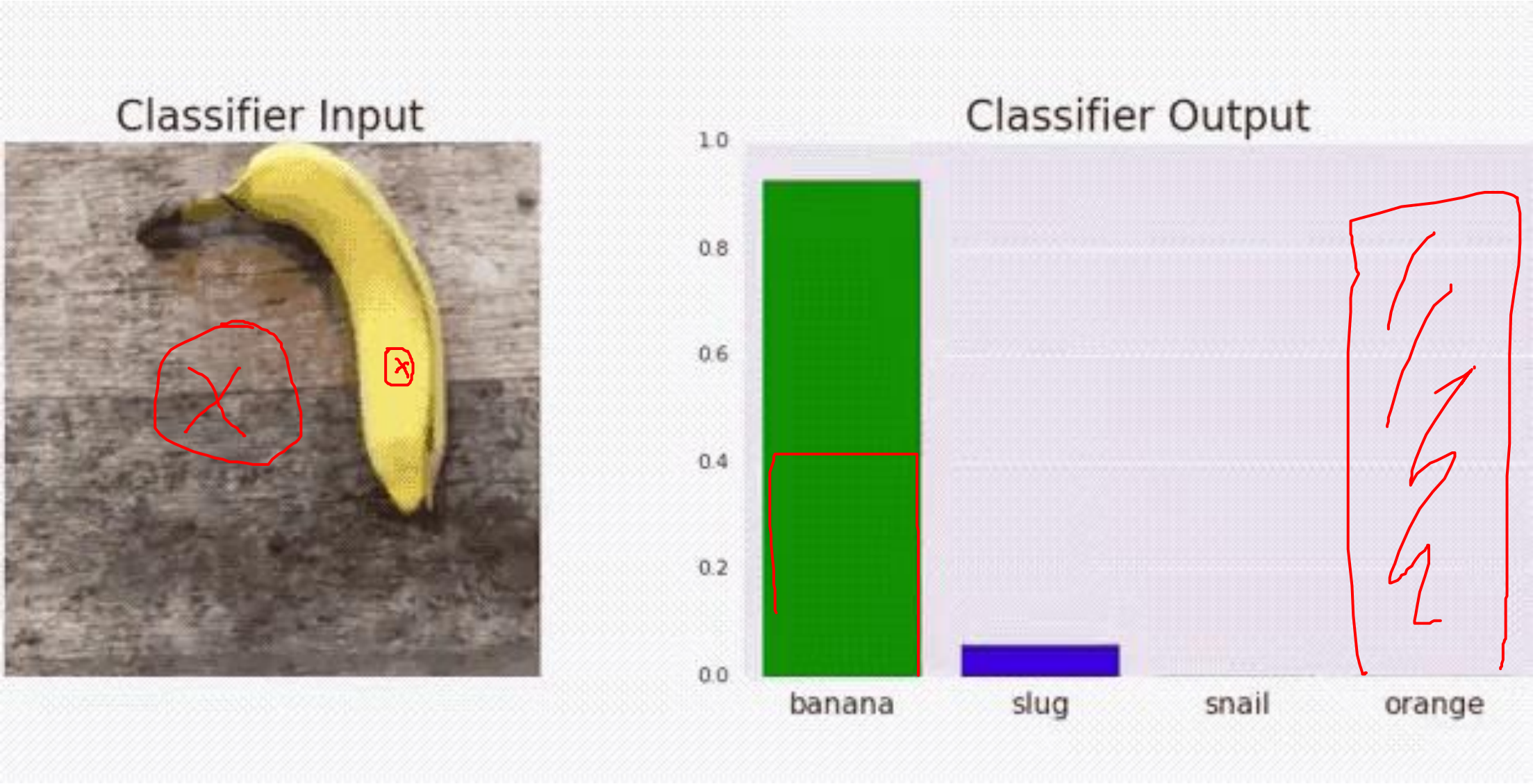


Figure 8. The Google Cloud Vision Demo labelling on the unperturbed image.



Motivations



Motivations

Fooling automated surveillance cameras: adversarial patches to attack person detection

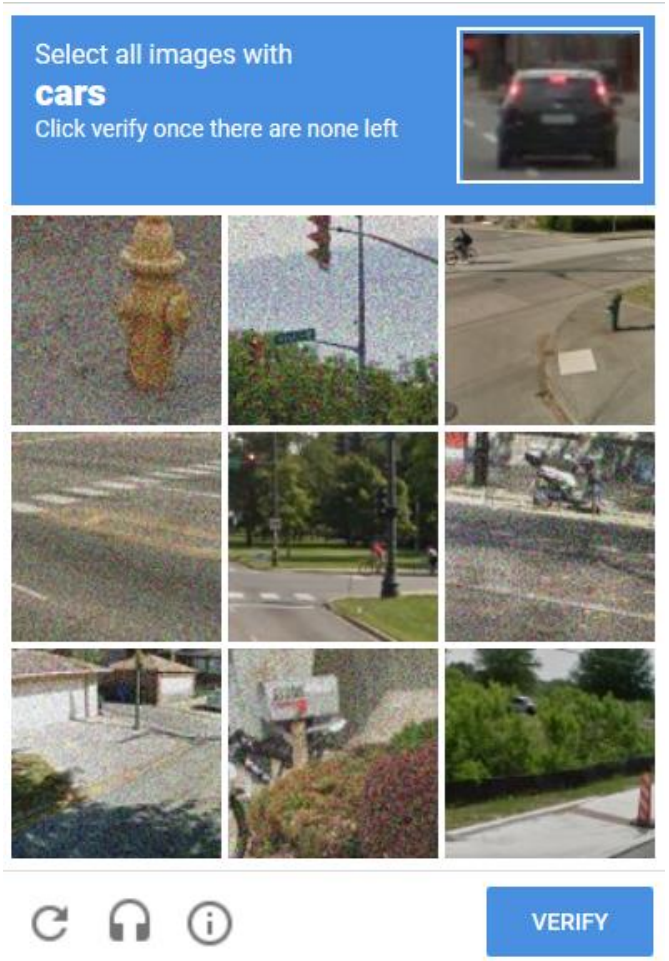
https://openaccess.thecvf.com/content_CVPRW_2019/html/CV-COPS/Thys_Fooling_Automated_Surveillance_Cameras_Adversarial_Patches_to_Attack_Person_Detection_CVPRW_2019_paper.html
https://syncedreview.com/2019/04/24/now-you-see-me-now-you-dont-fooling-a-person-detector/

Motivations



https://towardsdatascience.com/evasion-attacks-on-machine-learning-or-adversarial-examples-12f2283e06a1

Motivations



Motivations (speech recognition)

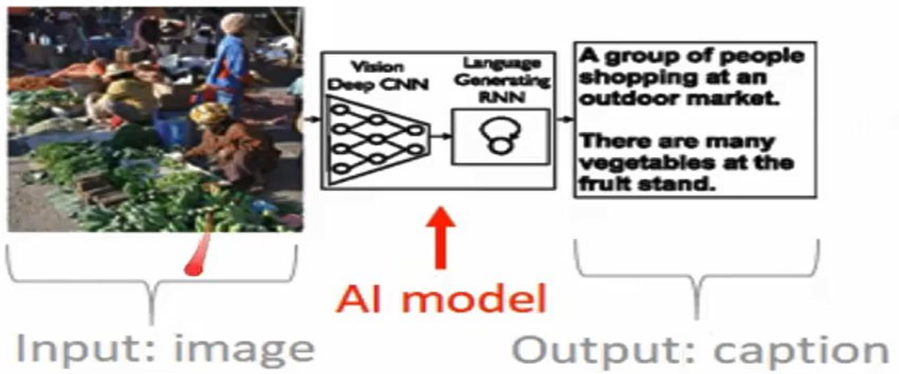
- Adversarial Attacks Against ASR Systems via Psychoacoustic Hiding

	Original	Modified	Noise
Speech	ORIGINAL	MODIFIED	NOISE
Music	ORIGINAL	MODIFIED	NOISE
Birds	ORIGINAL	MODIFIED	NOISE
Speech	ORIGINAL	MODIFIED	NOISE

https://adversarial-attacks.net/

Motivations (Image captioning)

Attacking Visual Language Grounding with Adversarial Examples: A Case Study on Neural Image Captioning



Original Top-3 inferred captions:

1. A red stop sign sitting on the side of a road.
2. A stop sign on the corner of a street.
3. A red stop sign sitting on the side of a street.

Adversarial Top-3 captions:

1. A brown teddy bear laying on top of a bed.
2. A brown teddy bear sitting on top of a bed.
3. A large brown teddy bear laying on top of a bed.

https://arxiv.org/pdf/1712.02051.pdf
IBM Research AI tutorial in CVPR21

Motivations (Text classification)

Discrete Adversarial Attacks and Submodular Optimization with Applications to Text Classification

Proceedings of Machine Learning and Systems 1 (MLSys 2019)

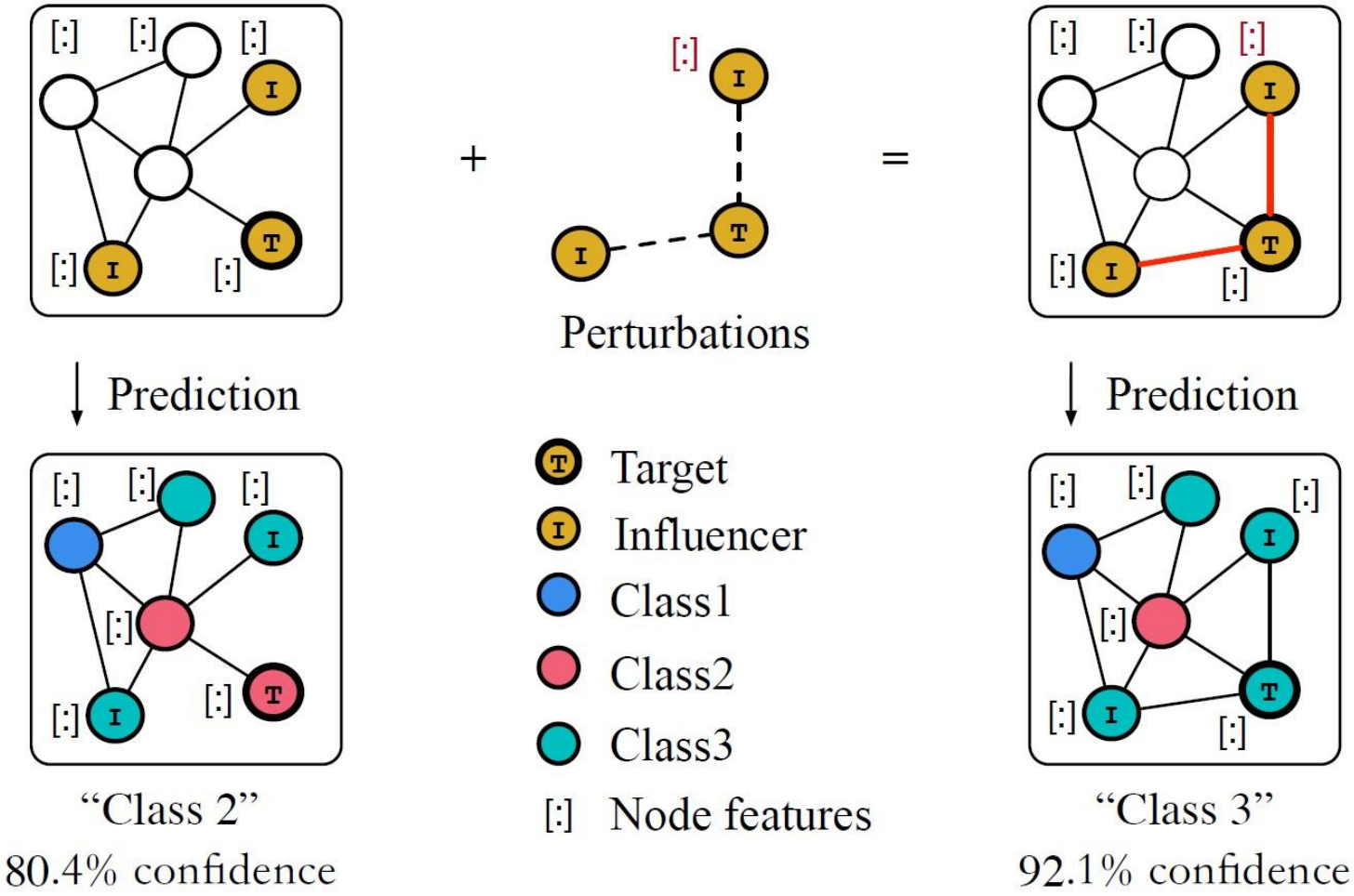
Task: Sentiment Analysis. Classifier: LSTM. Original: 100% Positive. ADV label: 100% Negative.
I suppose I should write a review here since my little Noodle-oo is currently serving as their spokes dog in the photos. We both love Scooby Do's. They treat my little butt-faced dog like a prince and are receptive to correcting anything about the cut that I perceive as being weird. Like that funny poofy pompadour. Mohawk it out, yo. Done. In like five seconds my little man was looking fabulous and bad ass. Not something easily accomplished with a prancing pup that literally chases butterflies through tall grasses. (He ended up looking like a little lamb as the cut grew out too. So adorable.) The shampoo they use here is also amazing. Noodles usually smells like tacos (a combination of beef stank and corn chips) but after getting back from the Do's, he smelled like Christmas morning! Sugar and spice and everything nice instead of frogs and snails and puppy dog tails. He's got some gender identity issues to deal with. ~~The pricing is also cheaper than some of the big name conglomerates out there~~ The price is cheaper than some of the big names below. I'm talking to you Petsmart! I've taken my other pup to Smelly Dog before, but unless I need dog sitting play time after the cut, I'll go with Scooby's. They genuinely seem to like my little Noodle monster.
Task: Fake-News Detection. Classifier: LSTM. Original label: 100% Fake. ADV label: 77% Real
~~Man~~ Guy punctuates high-speed chase with stop at In-N-Out Burger drive-thru Print [Ed.~~– Well, that's~~ Okay, that 's a new one.] ~~A~~ One man is in custody after leading police on a bizarre chase into the east Valley on Wednesday night. Phoenix police ~~began~~ has begun following the suspect in Phoenix and the pursuit continue~~d~~ into the east Valley, but it took a bizarre turn when the suspect stopped at an In-N-Out Burger restaurant's ~~drive-thru~~ drive-through near Priest and Ray Roads in Chandler. The suspect appeared to order food, but then drove away and got out of his pickup truck near Rock Wren Way and Ray Road. He ~~then ran into a backyard~~ ran to the backyard and tried to ~~get into a house through the back door~~ get in the home.

https://proceedings.mlsys.org/paper/2019/hash/d1fe173d08e959397adf34b1d77e88d7-Abstract.html

Motivations (Graph neural network)

Adversarial Learning on Graph

Jintang Li



https://spectra.pub/ml/graph-adversarial-learning

Motivations (seq2seq models)

Seq2Sick: Evaluating the Robustness of Sequence-to-Sequence Models with Adversarial Examples

Fourth AAAI Conference on Artificial Intelligence (AAAI-20)

Table 8: Text summarization adversarial examples using non-overlapping method. Surprisingly, it is possible to make the output sequence completely different by changing only one word in the input sequence.

Source input seq	among asia 's leaders , prime minister mahathir mohamad was notable as a man with a bold vision : a physical and social transformation that would push this nation into the forefront of world affairs .
Adv input seq	among **lynn** 's leaders , prime minister mahathir mohamad was notable as a man with a bold vision : a physical and social transformation that would push this nation into the forefront of world affairs.
Source output seq	asia 's leaders are a man of the world
Adv output seq	**a vision for the world**
Source input seq	under nato threat to end his punishing offensive against ethnic albanian separatists in kosovo , president slobodan milosevic of yugoslavia has ordered most units of his army back to their barracks and may well avoid an attack by the alliance , military observers and diplomats say
Adv input seq	under nato threat to end his punishing offensive against ethnic albanian separatists in kosovo , president slobodan milosevic of yugoslavia has **jean-sebastien** most units of his army back to their barracks and may well avoid an attack by the alliance , military observers and diplomats say.
Source output seq	milosevic orders army back to barracks
Adv output seq	**nato may not attack kosovo**
Source input seq	flooding on the yangtze river remains serious although water levels on parts of the river decreased today , according to the state headquarters of flood control and drought relief .
Adv input seq	flooding **that** the yangtze river **becomes** serious although water levels on parts of the river decreased today , according to the state headquarters of flood control and drought relief .
Source output seq	floods on yangtze river continue
Adv output seq	**flooding in water recedes in river**

One-word replacement

https://arxiv.org/abs/1803.01128

Motivations (seq2seq models)

Seq2Sick: Evaluating the Robustness of Sequence-to-Sequence Models with Adversarial Examples

Fourth AAAI Conference on Artificial Intelligence (AAAI-20)

Table 9: Text summarization examples with targeted keywords "police arrest "

Source input seq	north korea is entering its fourth winter of chronic food shortages with its people malnourished and at risk of dying from normally curable illnesses , senior red cross officials said tuesday.
Adv input seq	north **detectives** is **apprehended** its fourth winter of chronic food shortages with its people malnourished and at risk of dying from normally curable illnesses , senior red cross officials said tuesday.
Source output seq	north korea enters fourth winter of food shortages
Adv output seq	north **police arrest** fourth winter of food shortages.
Source input seq	after a day of fighting , congolese rebels said sunday they had entered kindu , the strategic town and airbase in eastern congo used by the government to halt their advances.
Adv input seq	after a day of fighting , **nordic detectives** said sunday they had entered **UNK** , the strategic town and airbase in eastern congo used by the government to halt their advances.
Source output seq	congolese rebels say they have entered UNK.
Adv output seq	nordic **police arrest** ## in congo.
Source input seq	president boris yeltsin stayed home tuesday , nursing a respiratory infection that forced him to cut short a foreign trip and revived concerns about his ability to govern.
Adv input seq	president boris yeltsin stayed home tuesday , **cops cops** respiratory infection that forced him to cut short a foreign trip and revived concerns about his ability to govern.
Source output seq	yeltsin stays home after illness
Adv output seq	yeltsin stays home after **police arrest**

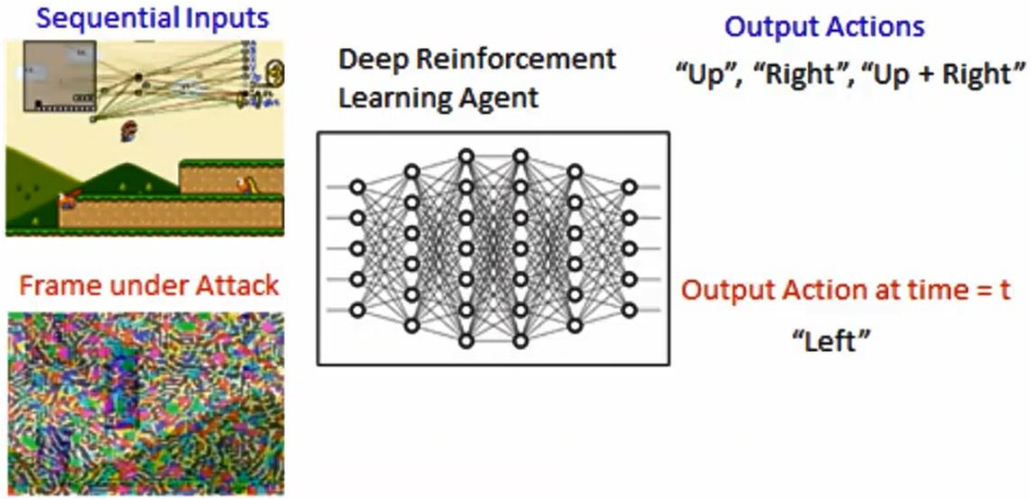
https://arxiv.org/abs/1803.01128

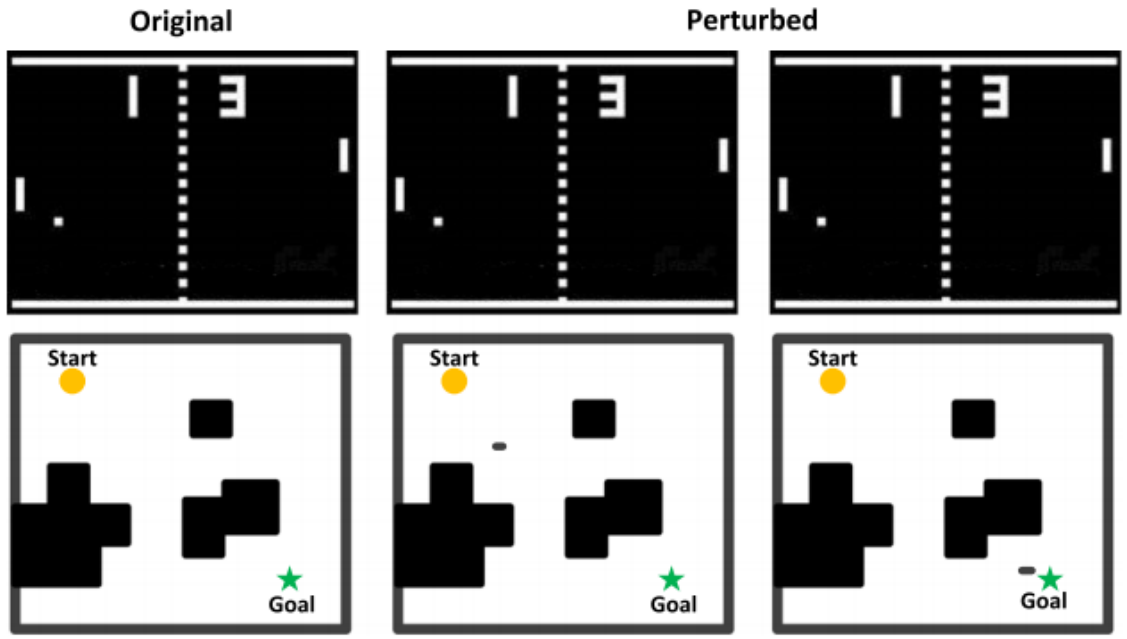
Motivations (reinforcement learning)

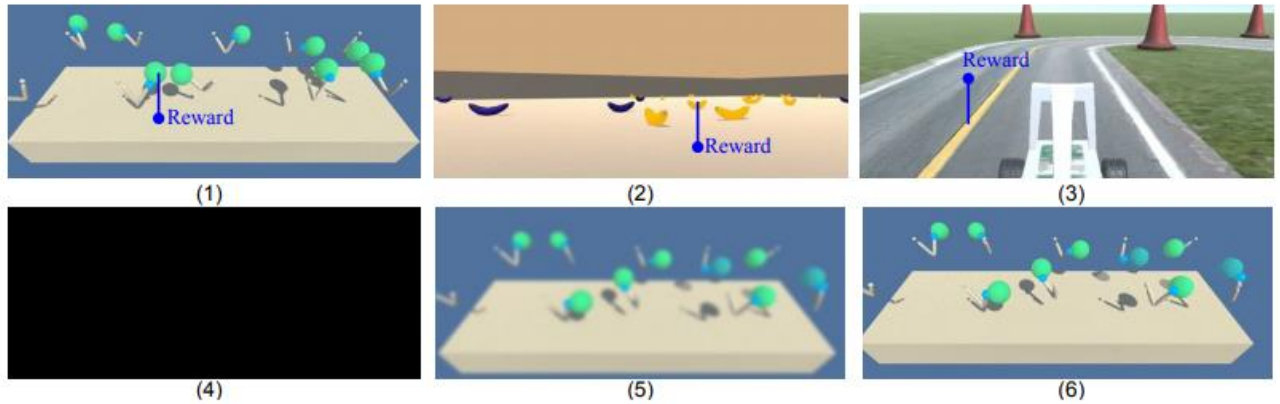
Adversarial attack and defense in reinforcement learning-from AI security view Cybersecurity

Enhanced Adversarial Strategically-Timed Attacks against Deep Reinforcement Learning

IEEE ICASSP 2020





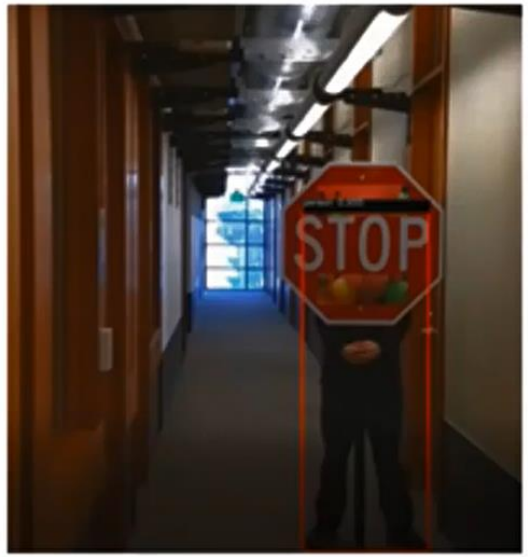


https://arxiv.org/pdf/2002.09027.pdf

https://cybersecurity.springeropen.com/track/pdf/10.1186/s42400-019-0027-x.pdf

Motivations (Physical word)

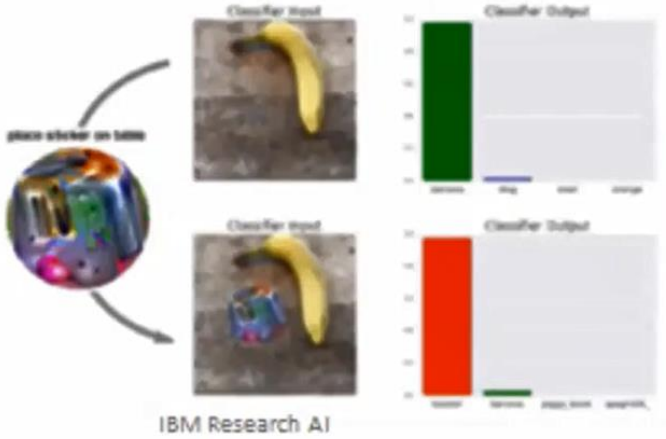
- Real-time traffic sign detector



- 3D-printed adversarial turtle

classified as turtle classified as rifle classified as other

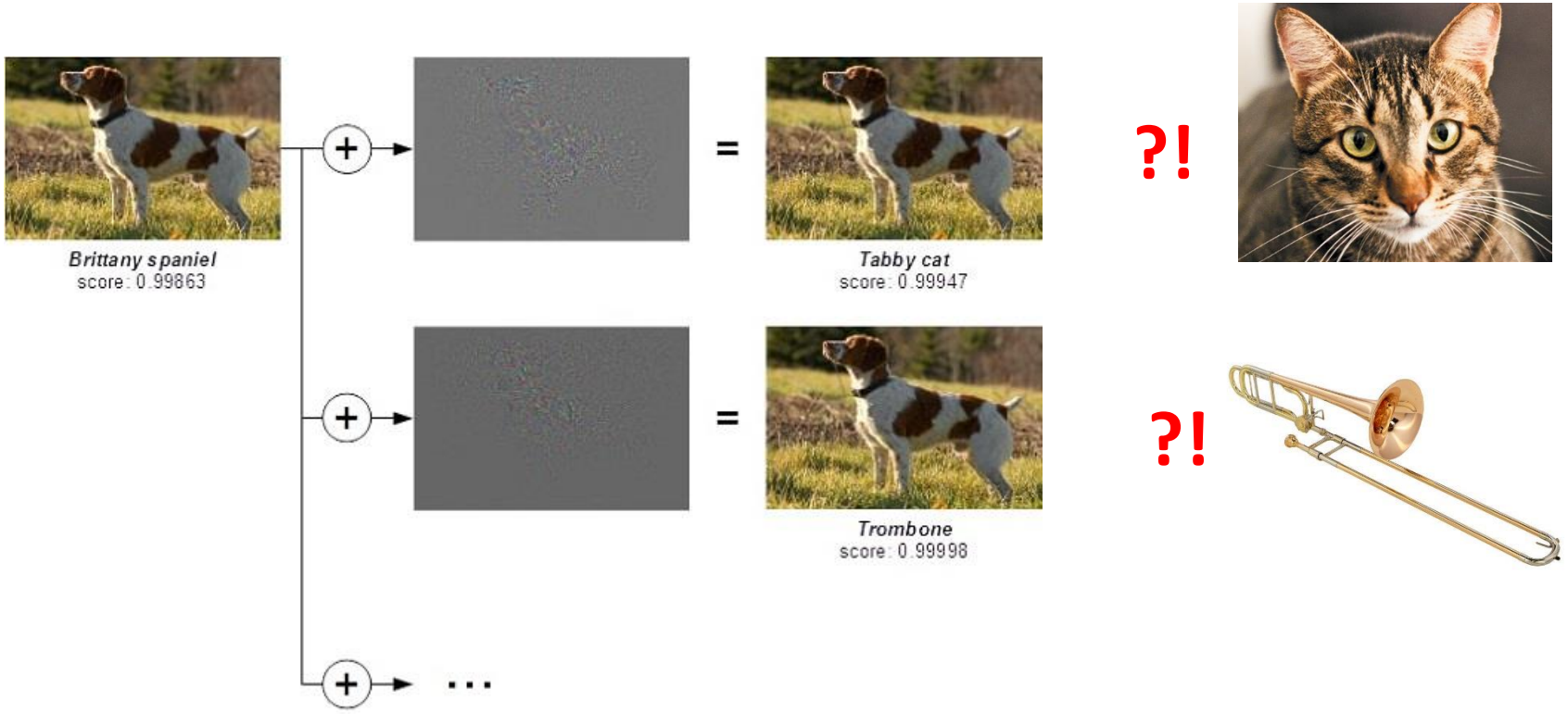
- Adversarial patch



IBM Research AI

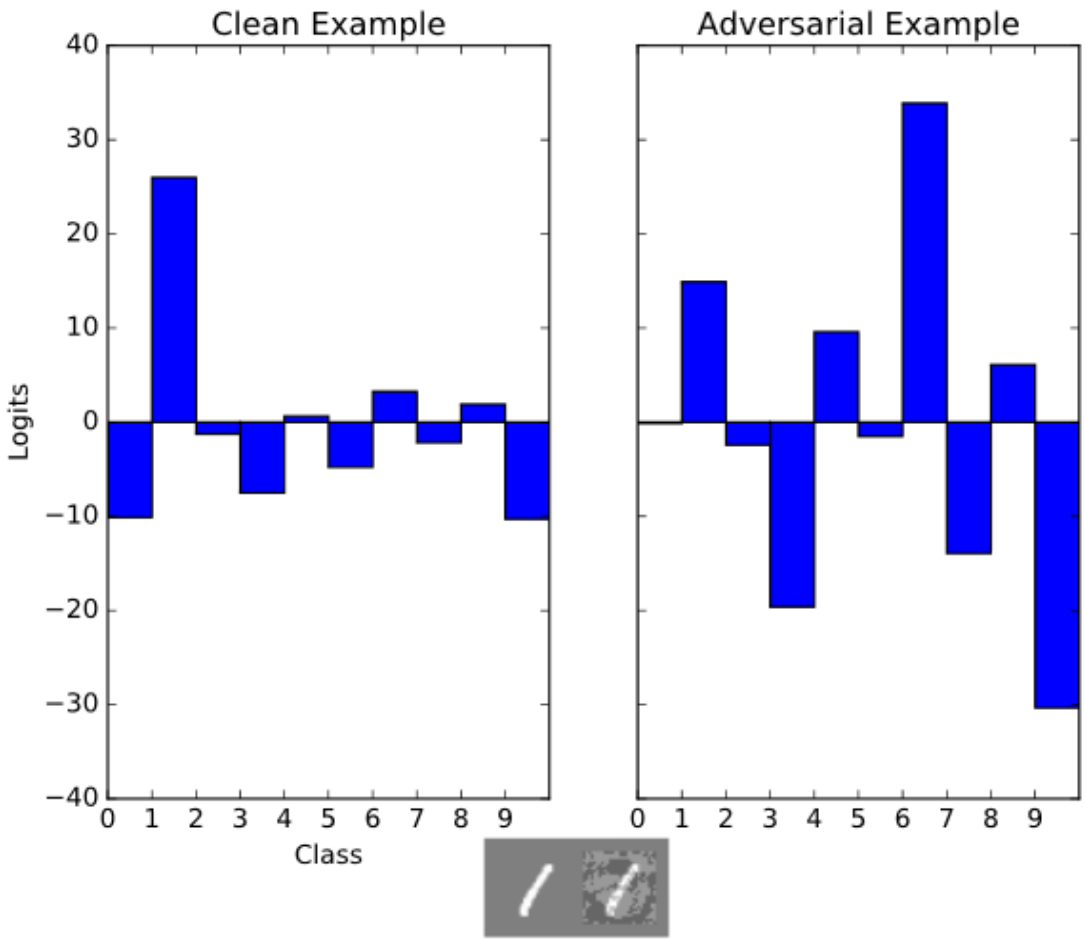
- Adversarial eye glasses

Examples of Adversarial Example



https://www.kdnuggets.com/2018/10/adversarial-examples-explained.html

Summary


Clean Example
Adversarial Example
Logits
Class
40
30
20
10
0
−10
−20
−30
−40
0 1 2 3 4 5 6 7 8 9


Break

Motivation

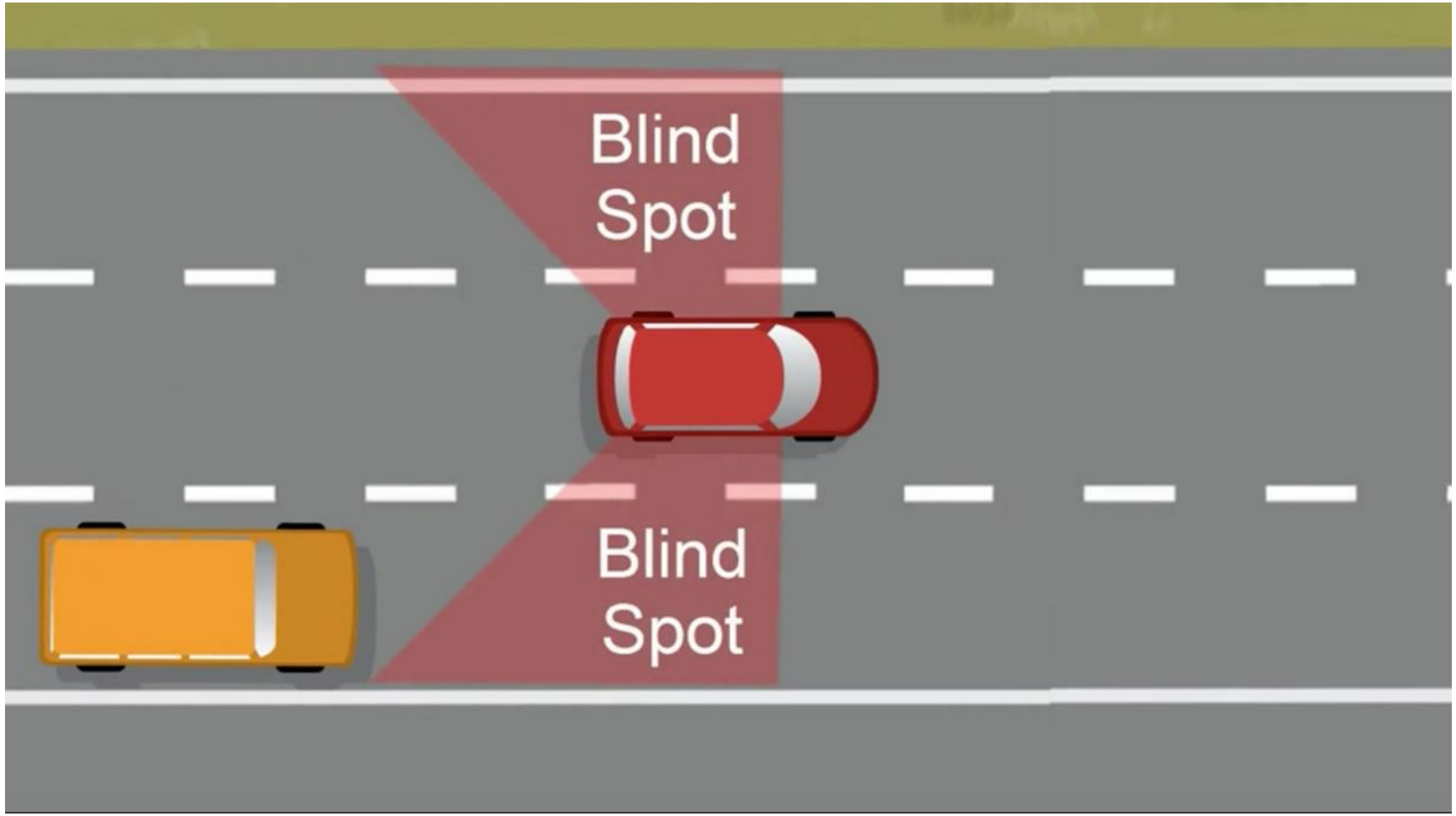
what makes

Deep neural networks

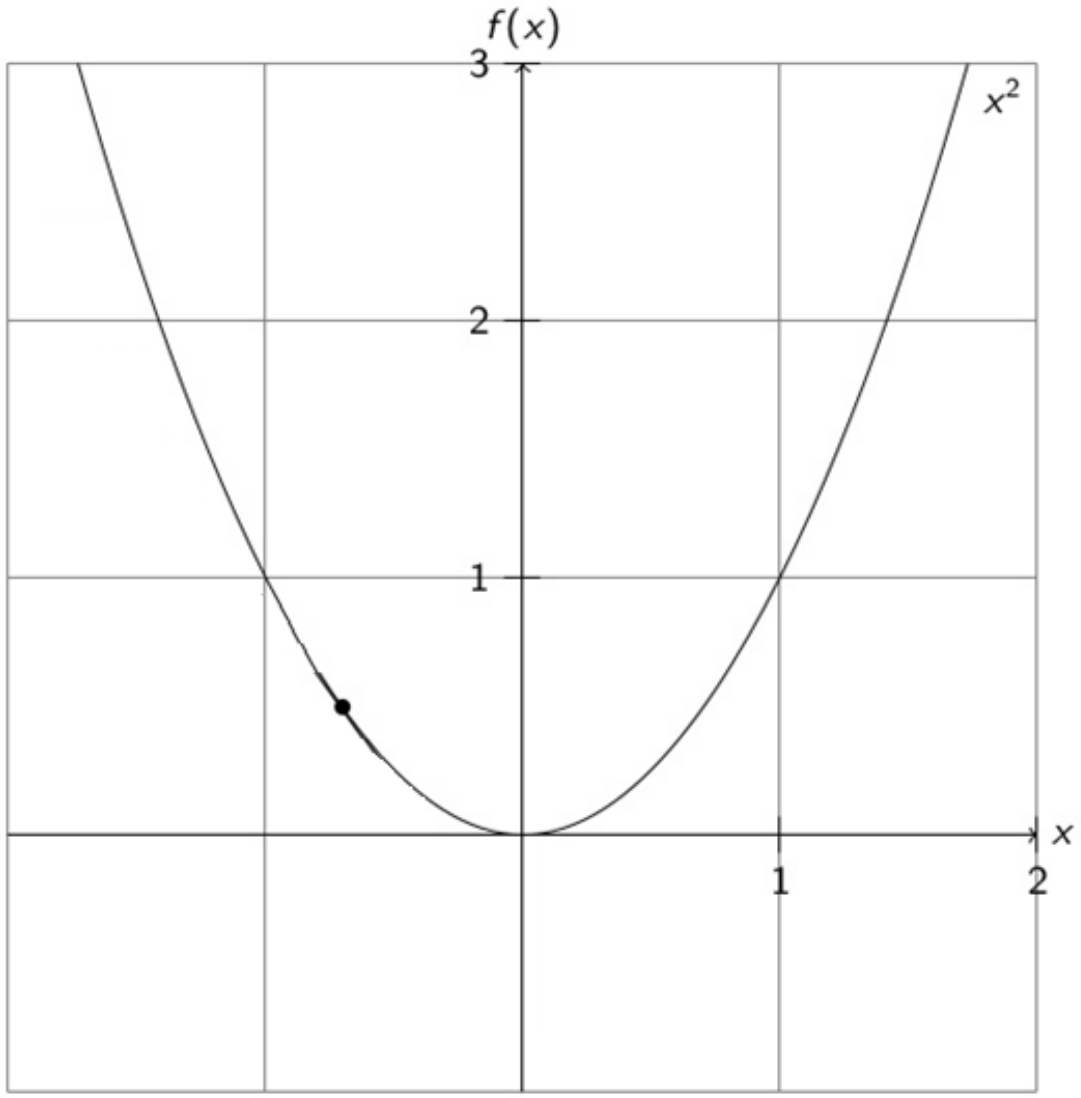
Vulnerable

to

Adversarial attacks

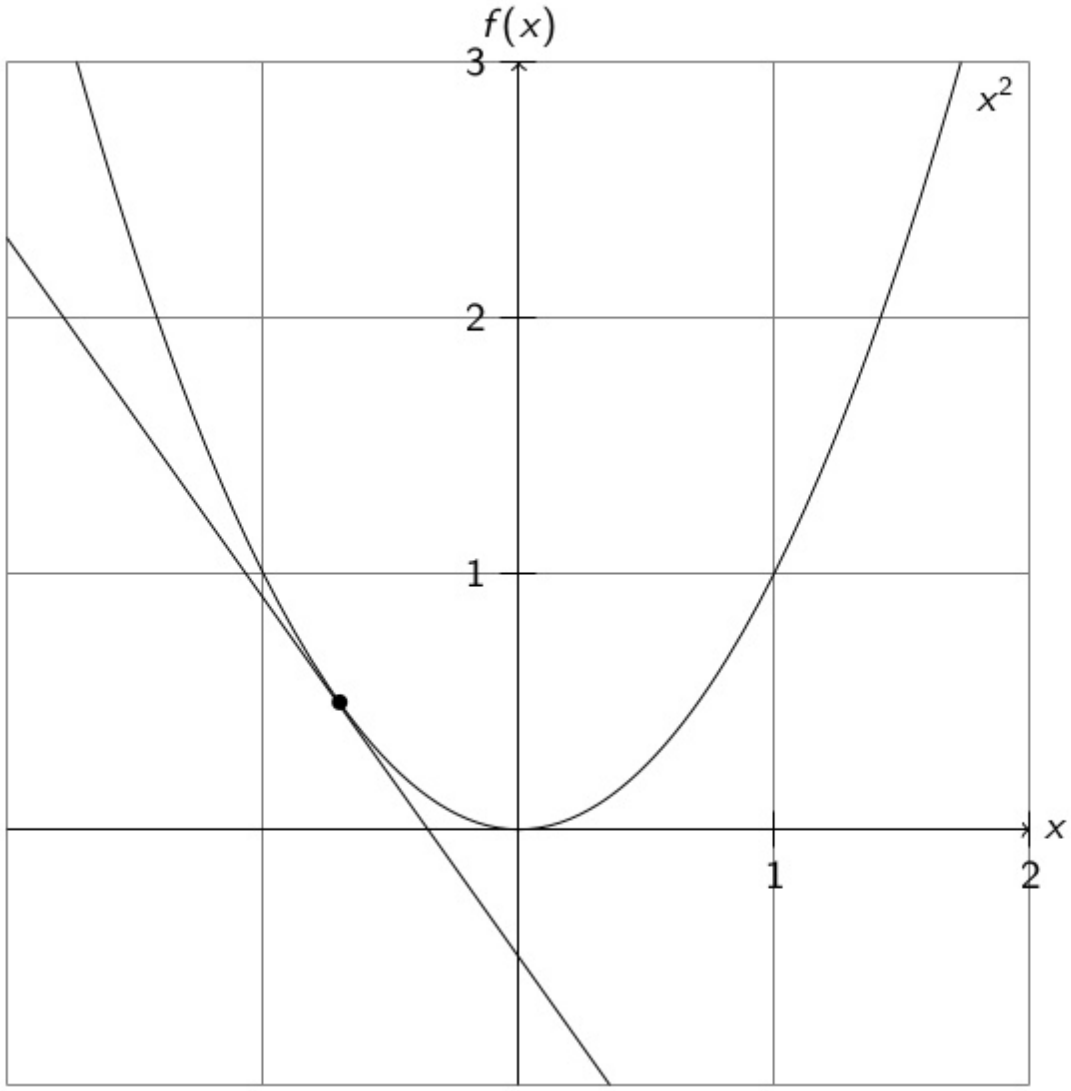

Blind
Spot
Blind
Spot

Motivation



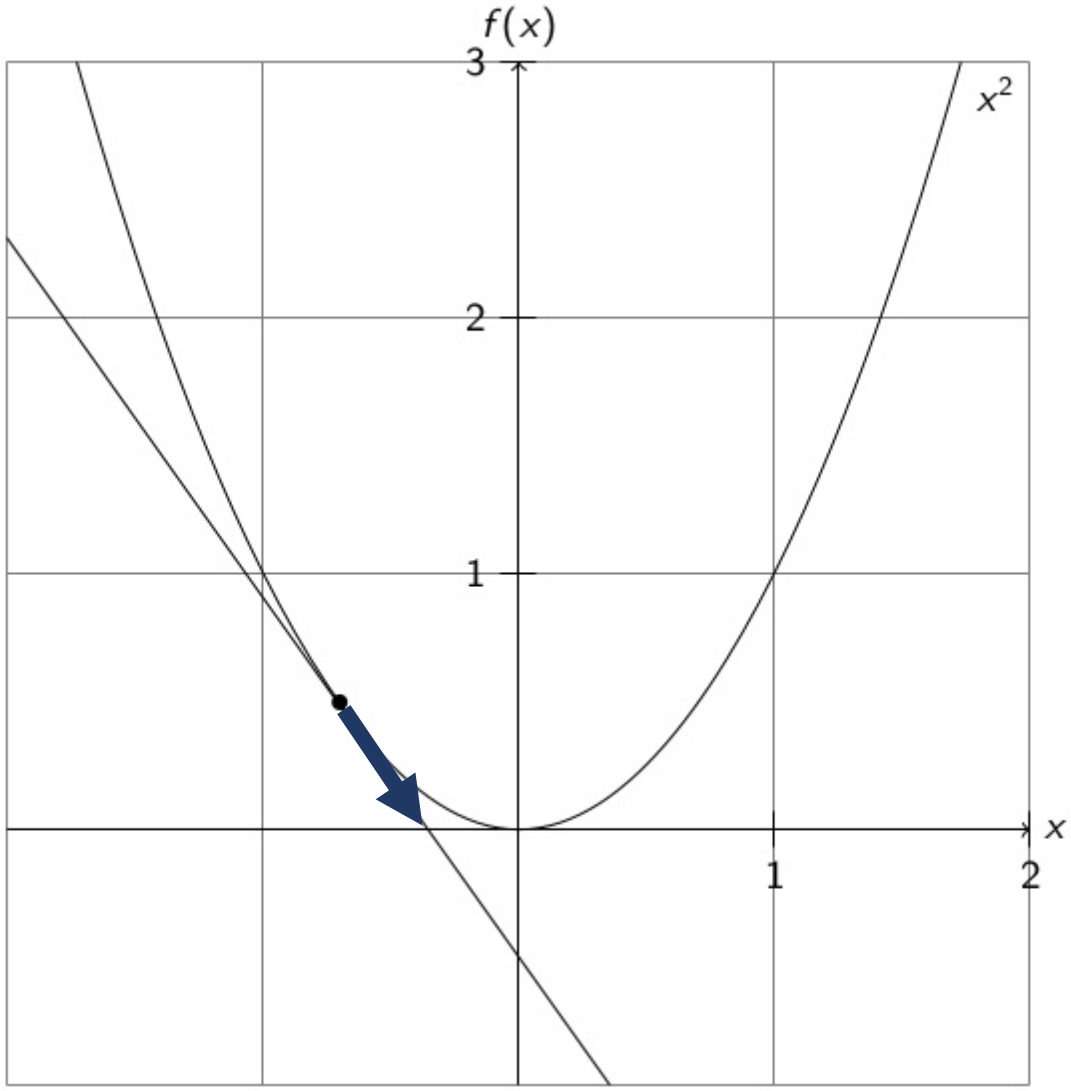
Motivation

Gradient-descent



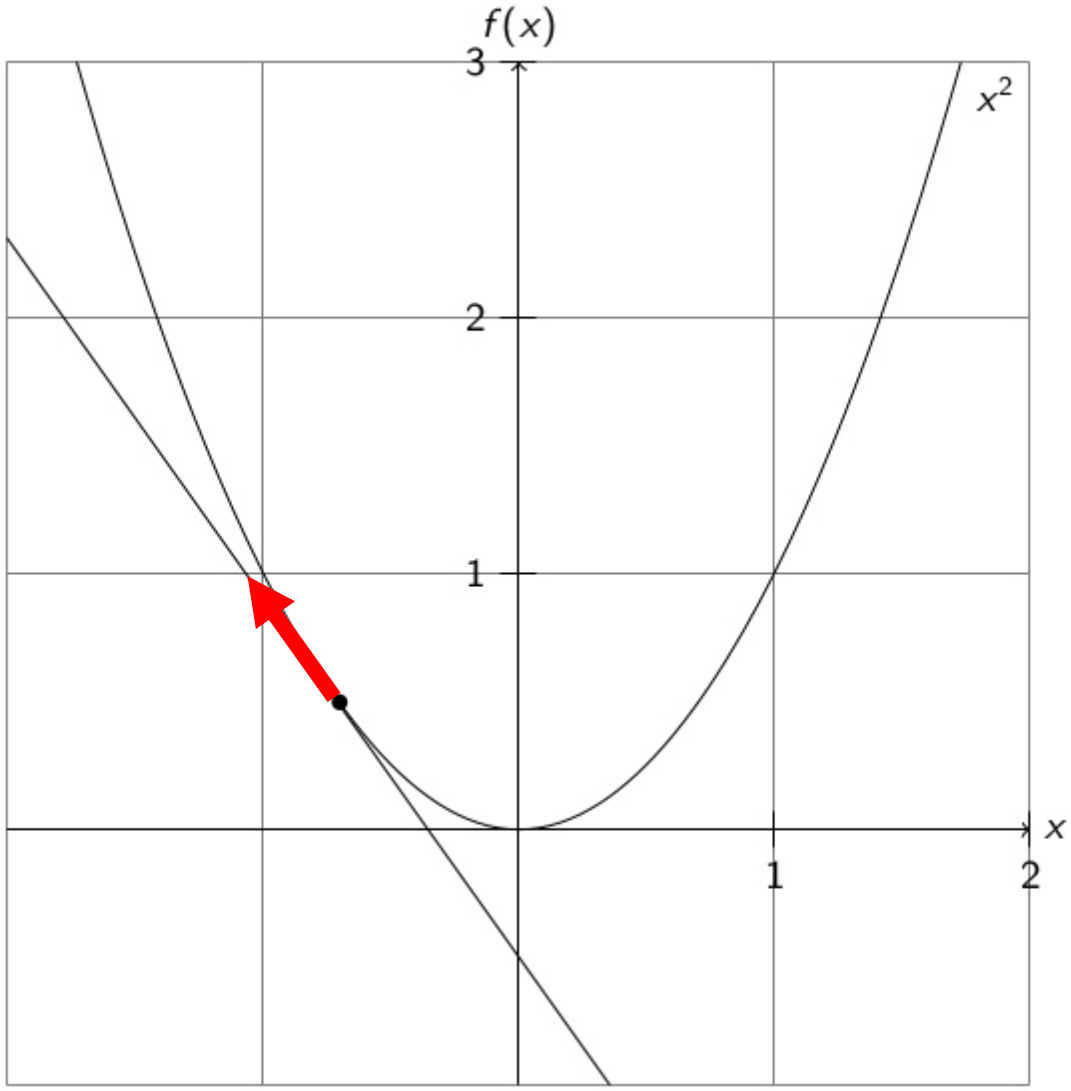
Motivation

Gradient-descent



Motivation

Gradient-descent



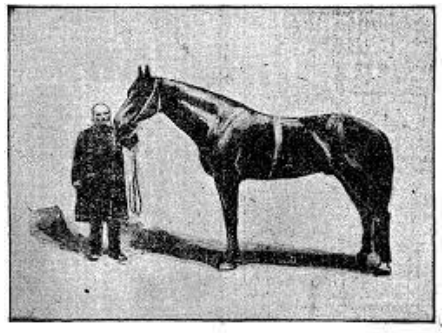
Attacks ?

https://www.slideshare.net/DavidKim486/universal-adversarial-perturbation

Attack

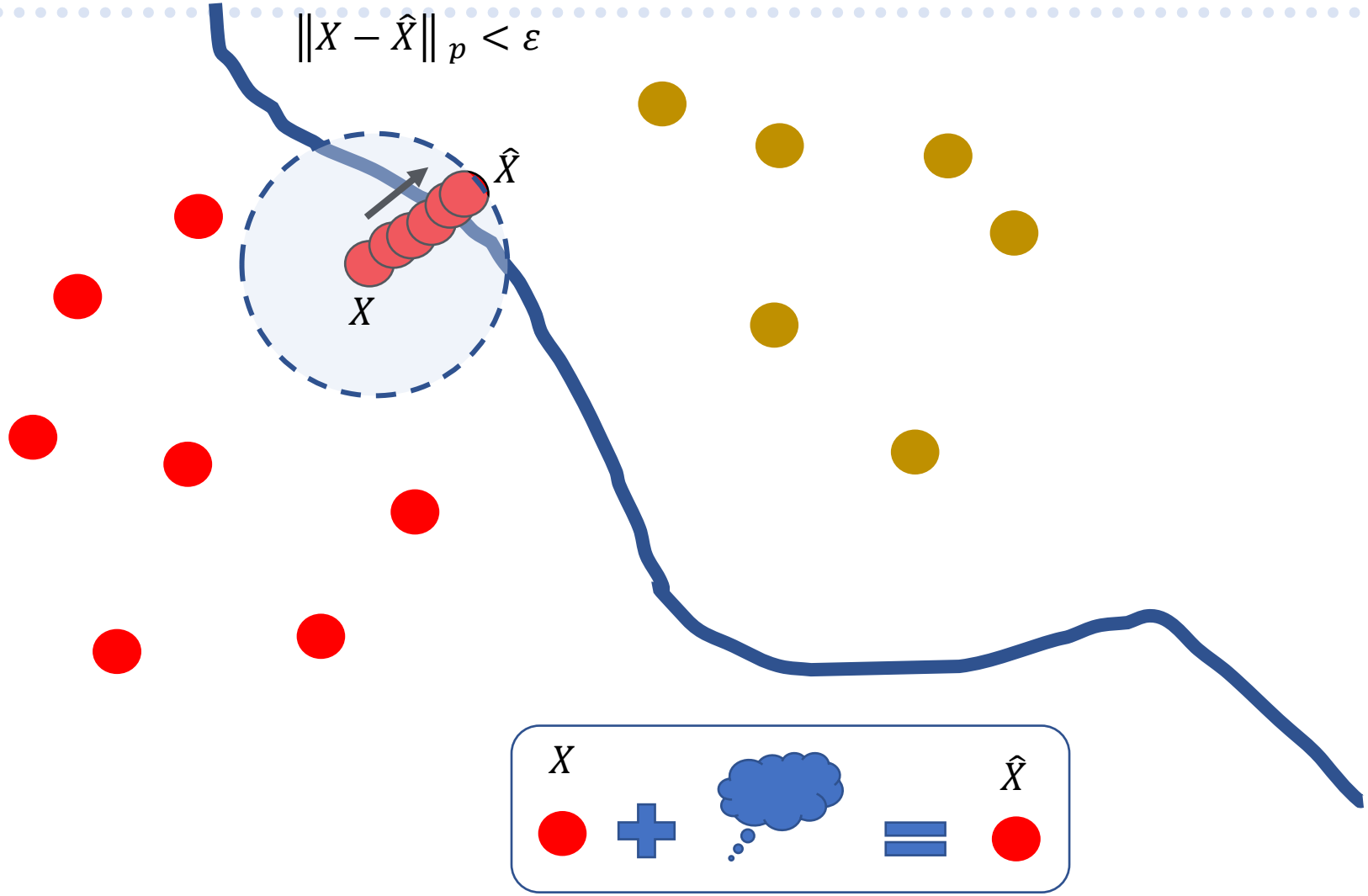
- “Adversarial examples are inputs to machine learning models that an attacker has intentionally designed to cause the model to make a mistake”
(Goodfellow et al 2017)

Short story …



HANS AND HIS OWNER, HERR VON OSTEN.

What is Adversarial Example?



Attack formulation

Given x, $f_\theta(.)$, the task is to compute x' such that

$$c^* = f_\theta(x) \neq f_\theta(x')$$

with some constraint like $\underbrace{\|x - x'\|_{\ell p}}_{\text{For } \ell_p \text{ attacks}} \leq \varepsilon$ to impose ***imperceptibility***

Attack formulation

$$\max_{\delta} l_{cls}\left(f_{\theta}\left(x'\right), y\right)$$

Perturbation

Loss

Classifier

Adv. Example

Label

Single-step attack

$$\begin{aligned}\delta_{\mathrm{FGSM}} &= \max_{\|\delta\| \leqslant \varepsilon} \langle \nabla l\left(f_\theta(x), y\right), \delta \rangle \\ &= \varepsilon \cdot \operatorname{sign}\left(\nabla l\left(f_\theta(x), y\right)\right)\end{aligned}$$

- **F**ast **G**radient **S**ign **M**ethod (Goodfellow et al., ICLR'15).
- Specifically designed for ℓ_∞ attacks.
- One-step attack.

Single-step attack

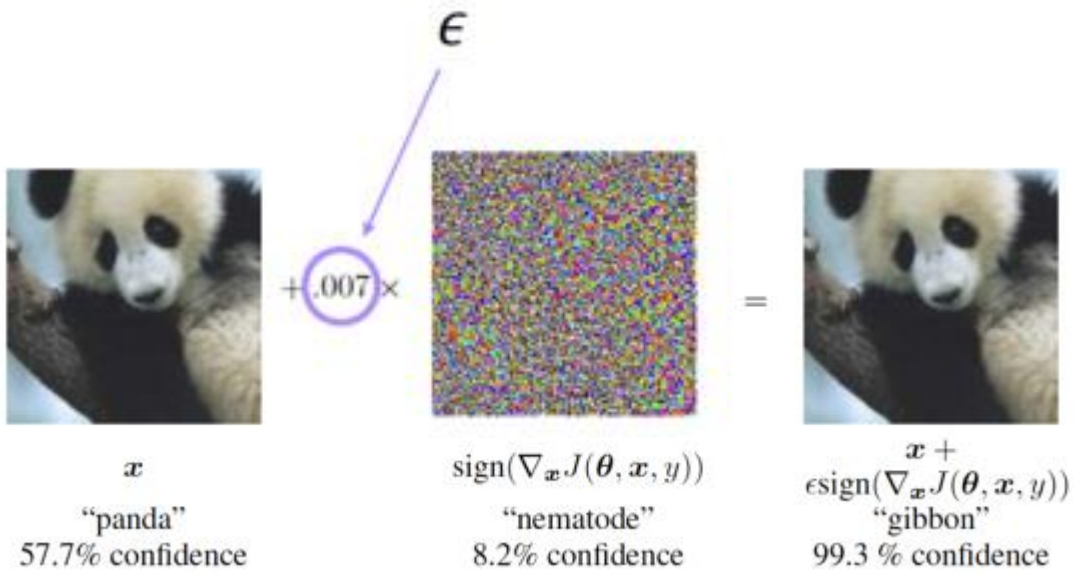
$$x' = x + \alpha \underbrace{\text{sign}\left(\nabla_x L(\theta, x, y)\right)}_{\text{FGSM}}$$

Perturbation

$$\tilde{x} = x + \eta$$

$$\eta = \epsilon \text{sign}\left(\nabla_x J(\theta, x, y)\right)$$

Gradient of the cost function

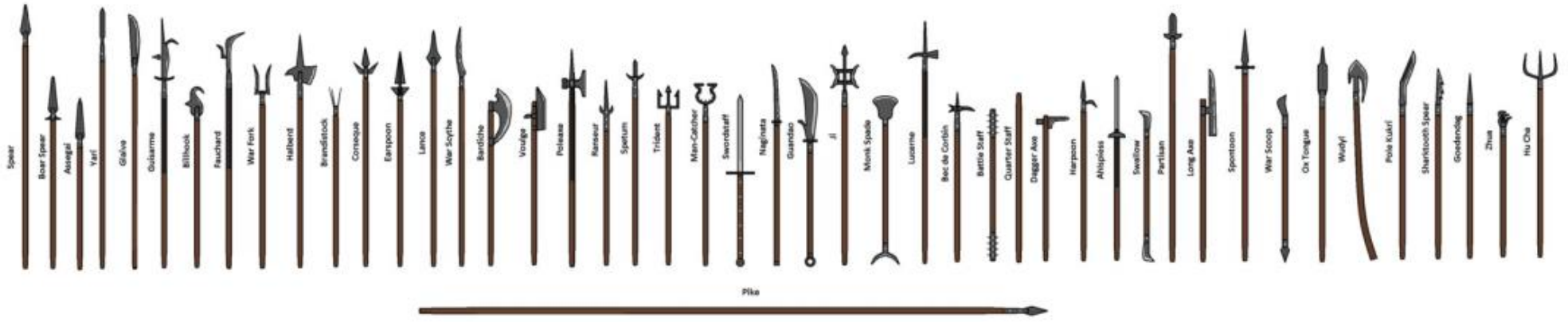


x
"panda"
57.7% confidence

$\text{sign}(\nabla_x J(\theta, x, y))$
"nematode"
8.2% confidence

$x + \epsilon\text{sign}(\nabla_x J(\theta, x, y))$
"gibbon"
99.3 % confidence

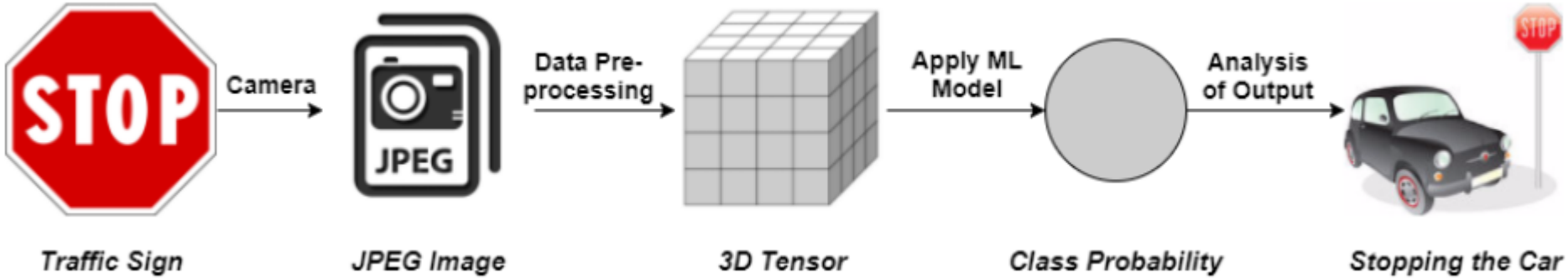
Various methods of attacks



https://www.slideshare.net/DavidKim486/universal-adversarial-perturbation

Adversarial Threat Model

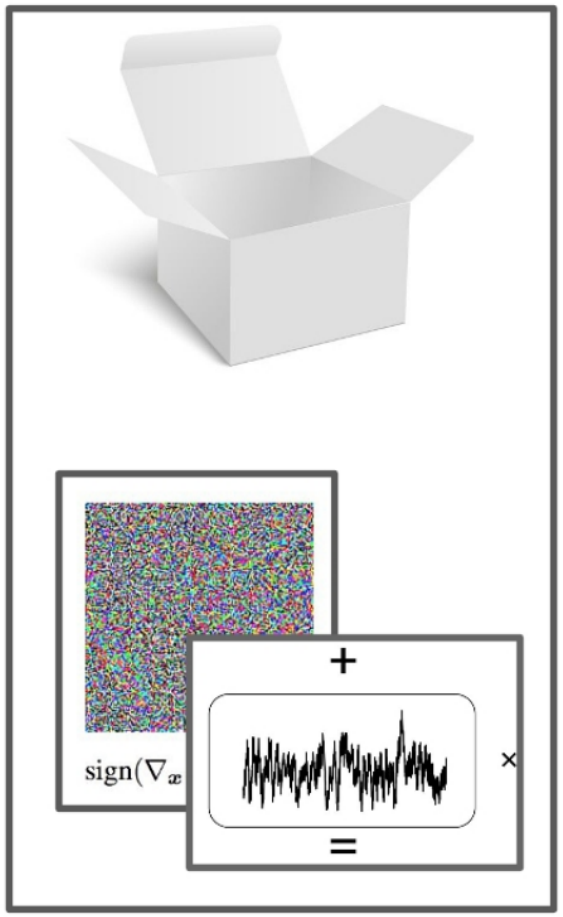
- **The Attack Surface**

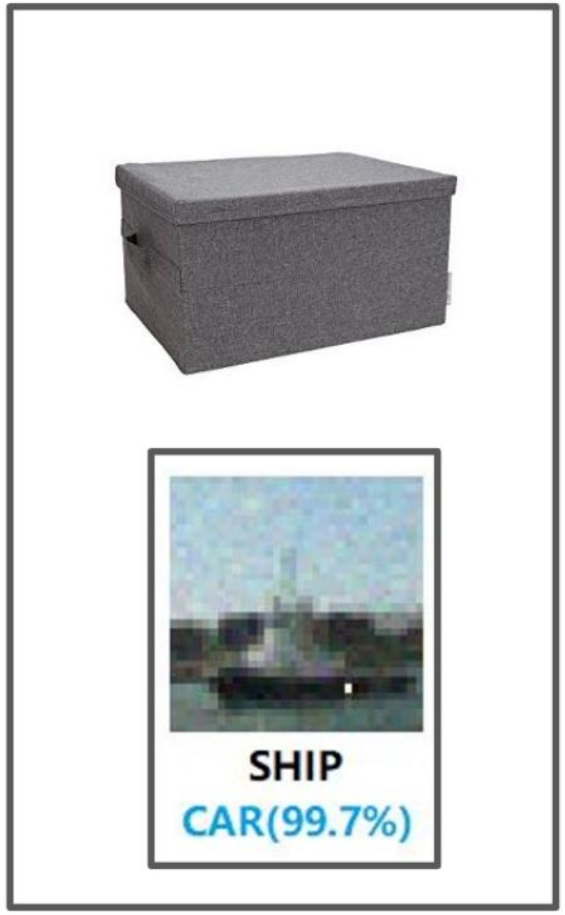


- Evasion Attack :: during the testing phase (* the most common type of attack!)
- Poisoning Attack :: during the training time
- Exploratory Attack :: during the testing phase (Given black box access to the model
  try to gain as much knowledge as possible)

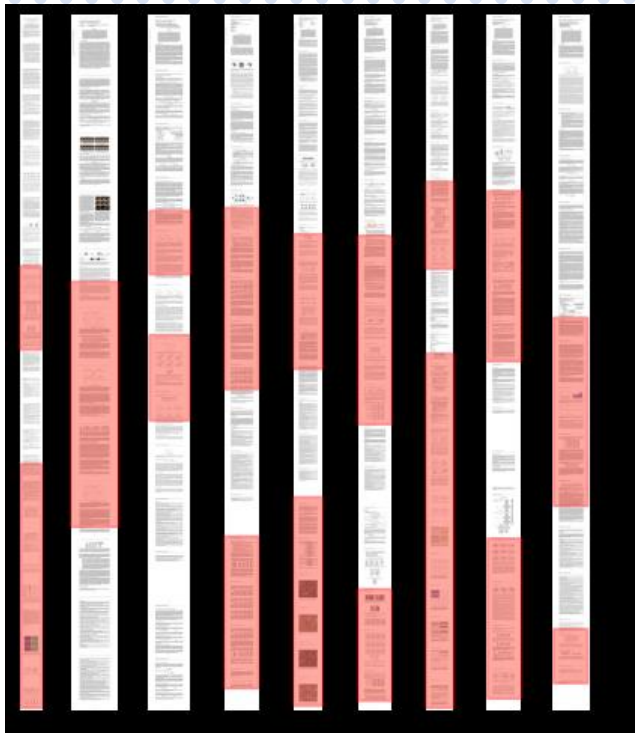
http://fna.ir/a5g

Adversarial Attacks


sign(∇x
+
=
×


SHIP
CAR(99.7%)

One important notes!



```
acc, loss = model.evaluate(x_test, y_test)
```

Is no longer sufficient.

https://blog.floydhub.com/introduction-to-adversarial-machine-learning/


کالج تخصصی
هوش مصنوعی پارت
مرکز تحقیقات
هوش مصنوعی پارت