Adversarial Attack & Defense

Presented by **Mohammad Khalooei**

Under the supervision of

Prof. Mohammad Mehdi Homayounpour & Dr. Maryam Amirmazlaghani

Amirkabir
Artificial Intelligence
Summer
Summit 2019

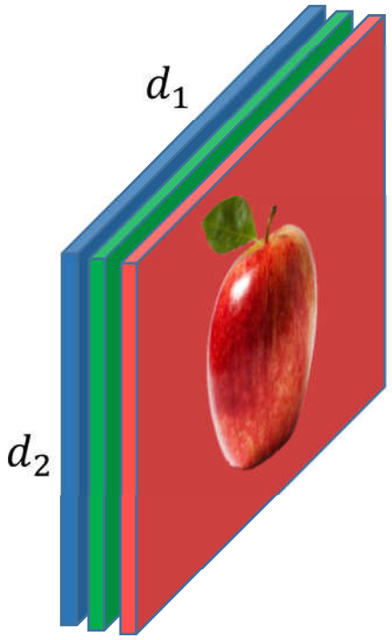
23 - 25 July 2019
Tehran, IRAN

Amirkabir
University of Technology

Outlines

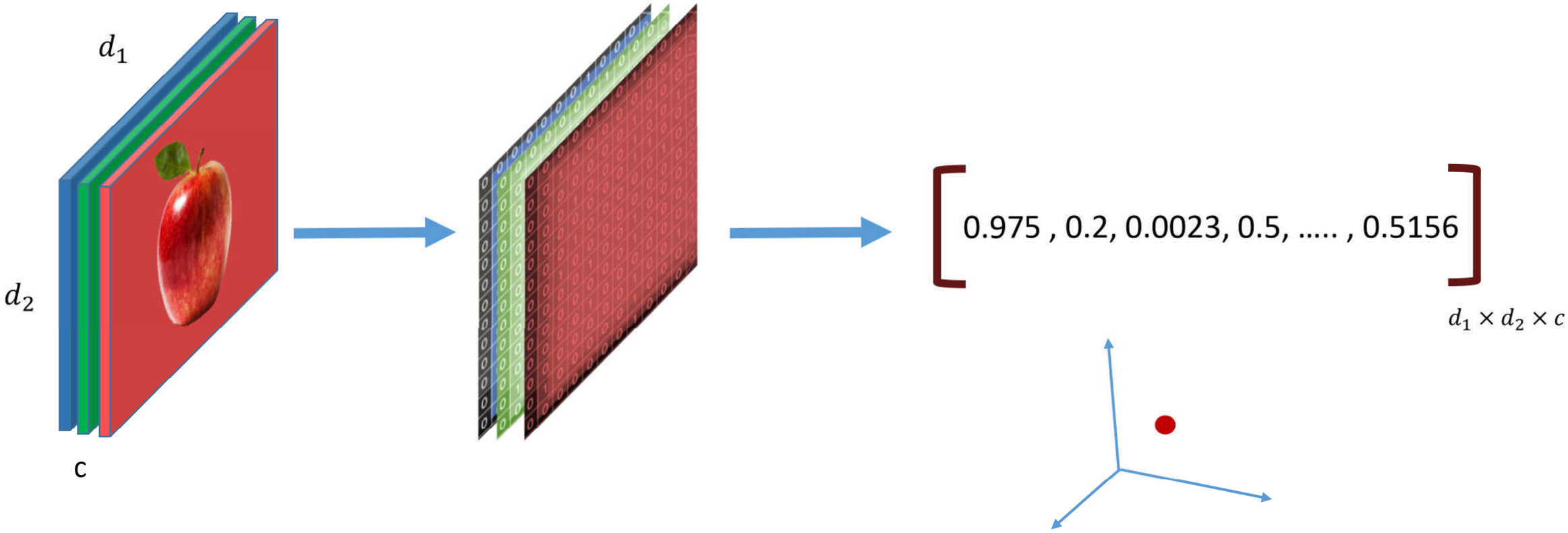
- Overview of Generative & Discriminative models
- Decision-boundary crossing
- Vulnerability of deep neural networks in different data
- Different attacks and penetration ways of ML & DL approaches
- Attack Toolboxes
- Different defense strategies for DNN
- Tips to stay safe in DNN

Review:: Data



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

Review:: Data



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

Review:: Data

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

Review:: Life of the points ...

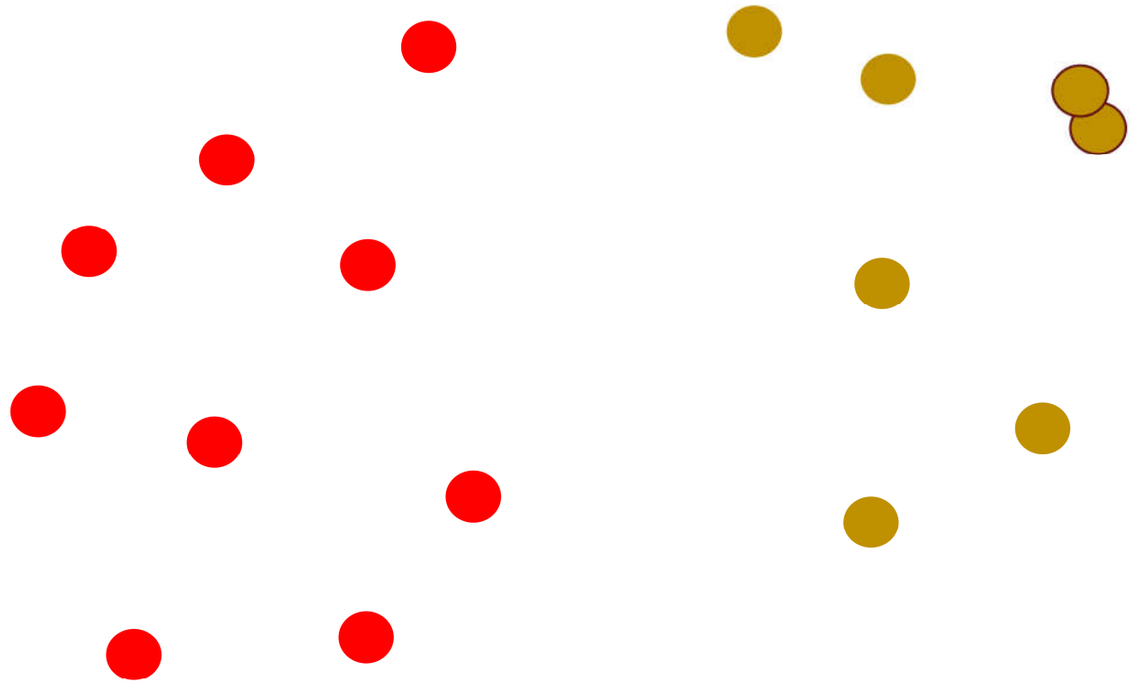
https://www.slideshare.net/khalooei/life-of-points-machine-learning-with-orange-flavor

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

Review:: Discriminative Model

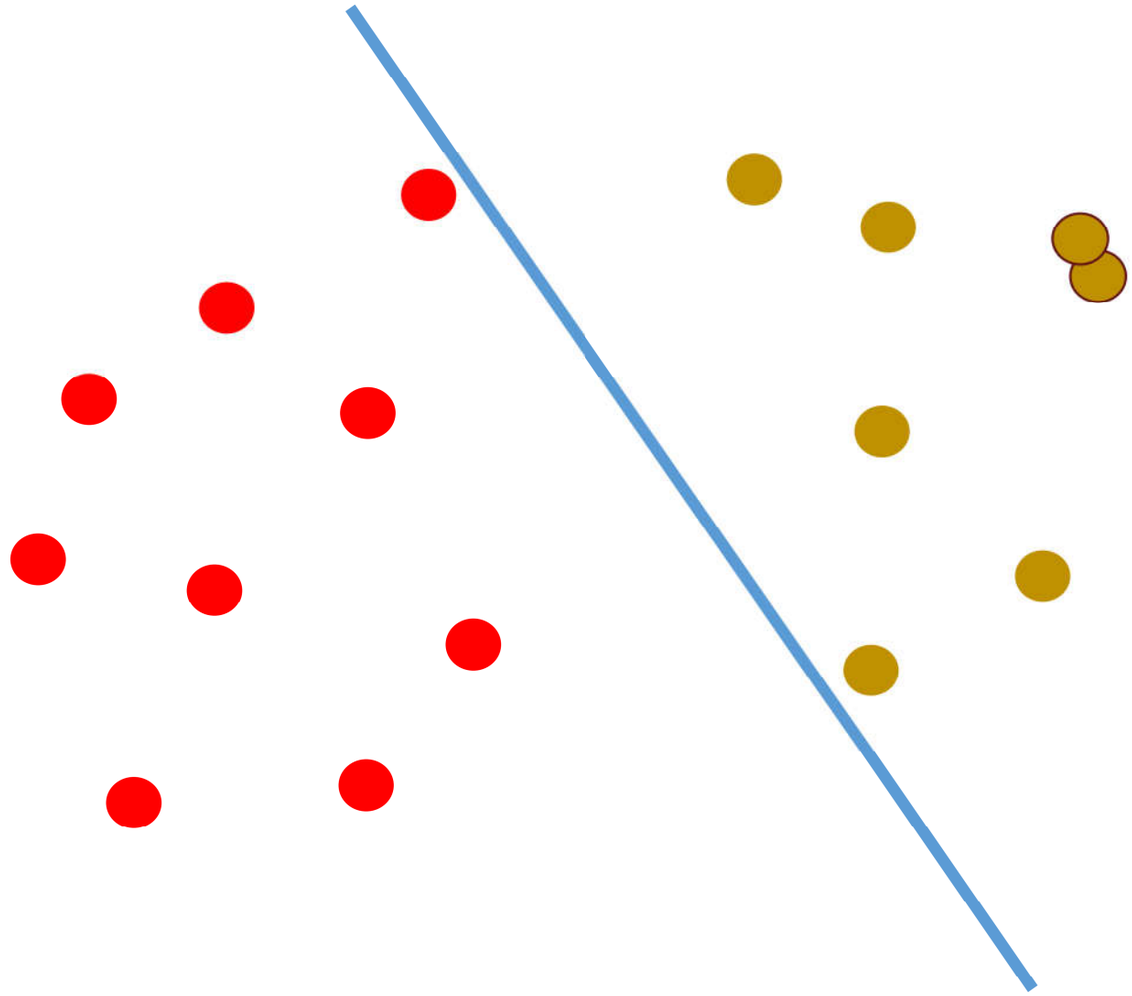
https://www.slideshare.net/BrianKim244/dcgan-77452250

Review:: Discriminative Model



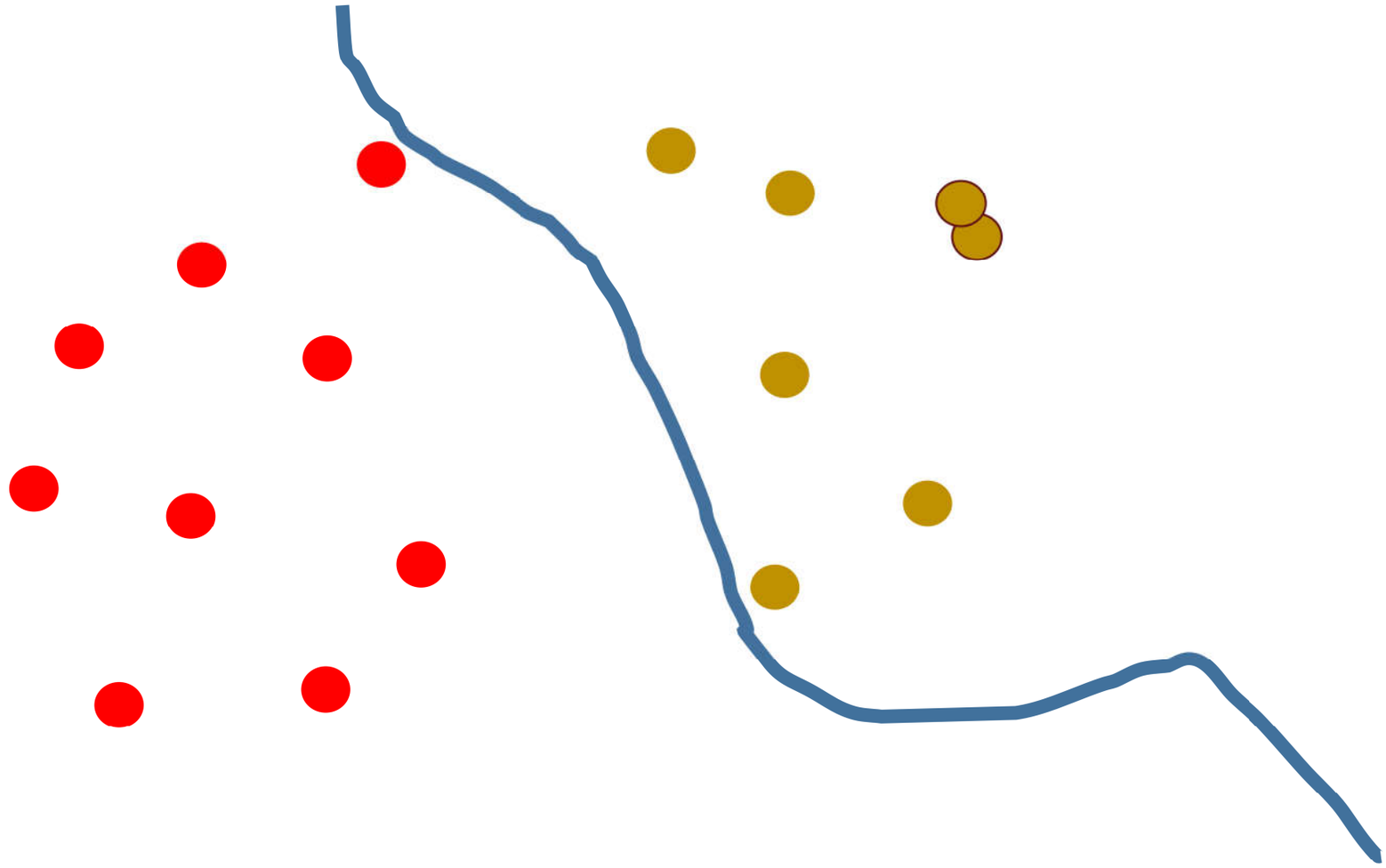
https://www.slideshare.net/BrianKim244/dcgan-77452250

Review:: Discriminative Model

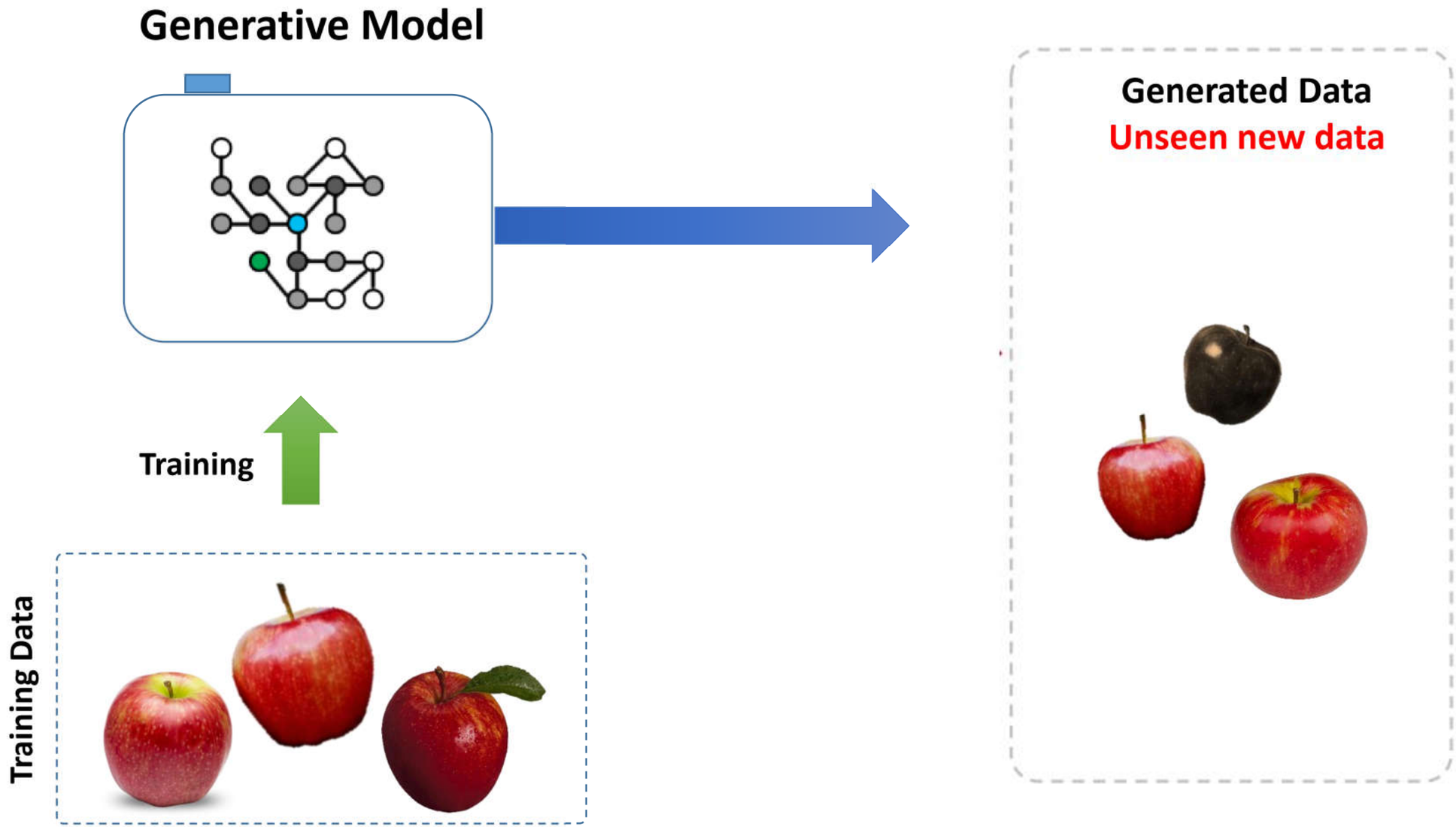


https://www.slideshare.net/BrianKim244/dcgan-77452250

Review:: Discriminative Model

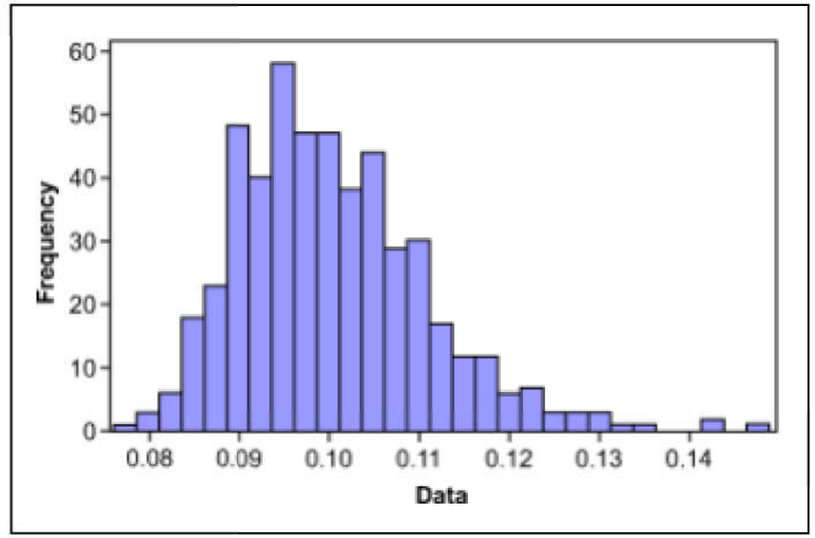


https://www.slideshare.net/BrianKim244/dcgan-77452250



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

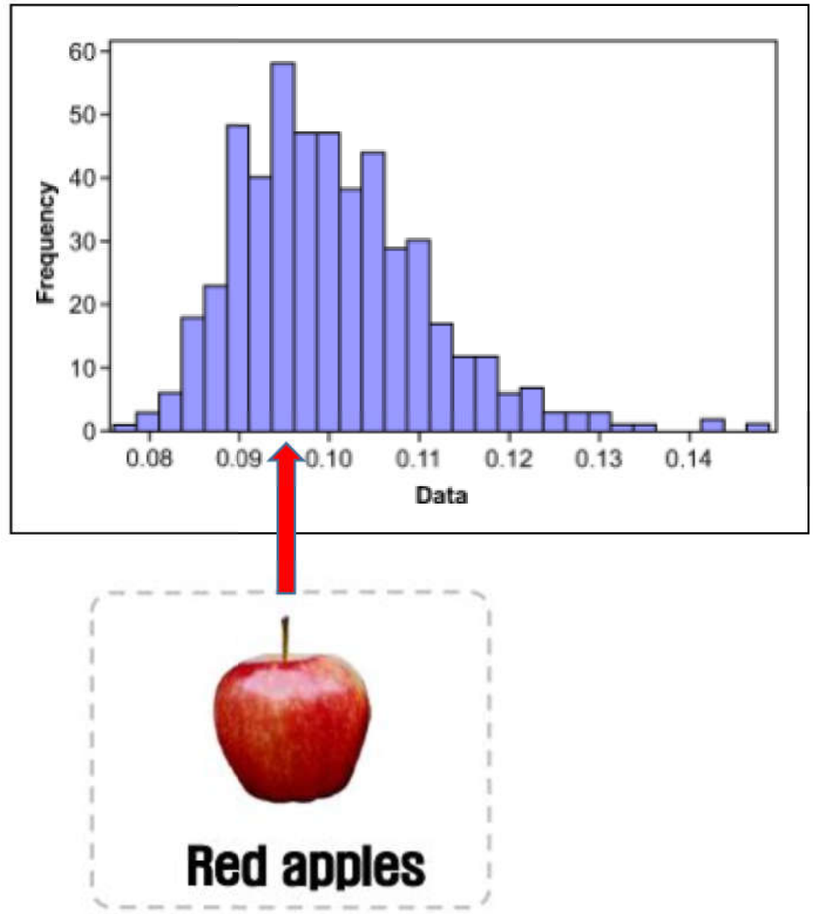
Review:: Generative Model



Distribution of the actual images

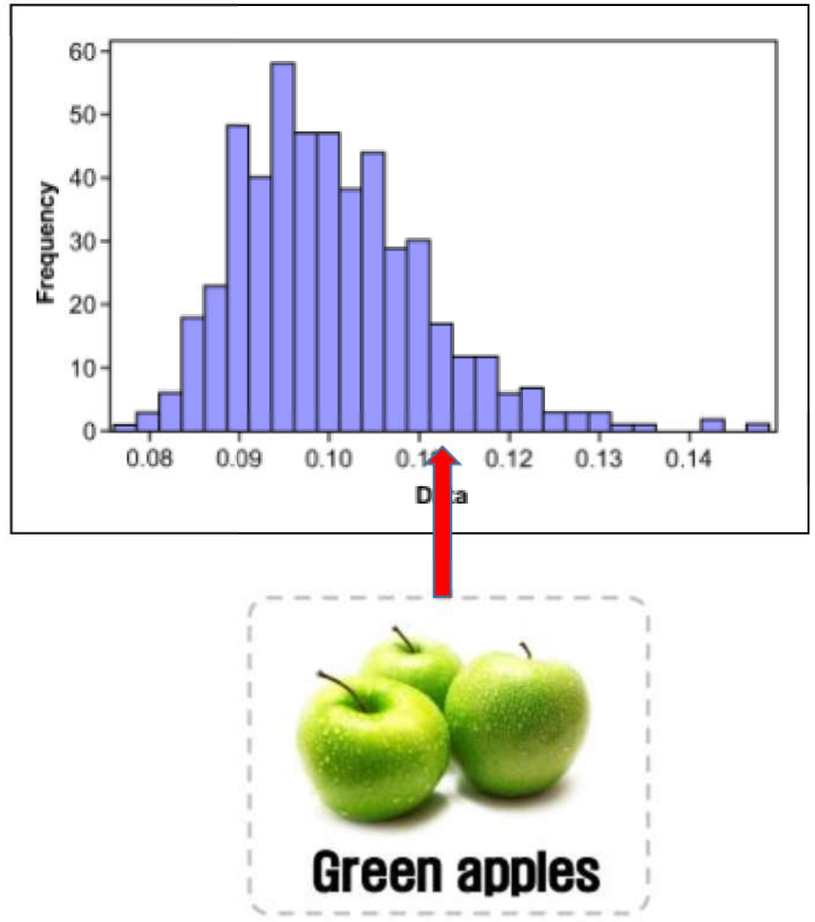
https://www.slideshare.net/BrianKim244/dcgan-77452250

Review:: Generative Model



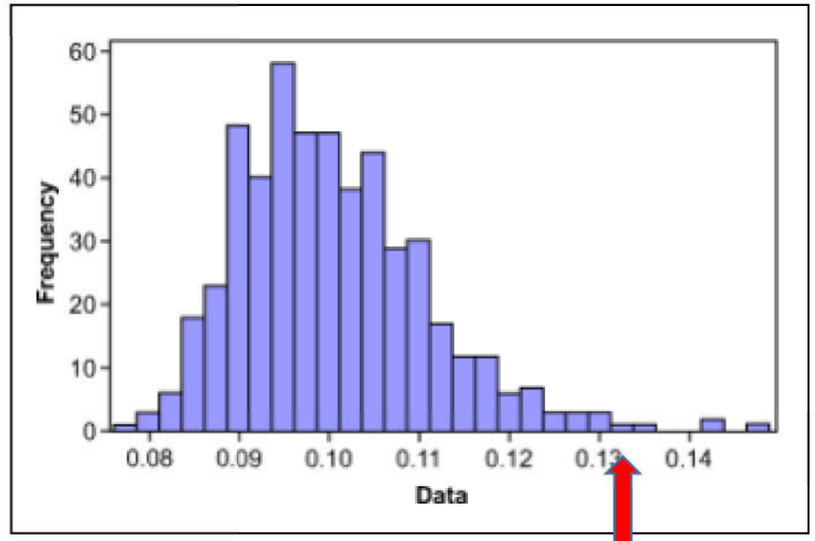
https://www.slideshare.net/BrianKim244/dcgan-77452250

Review:: Generative Model



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

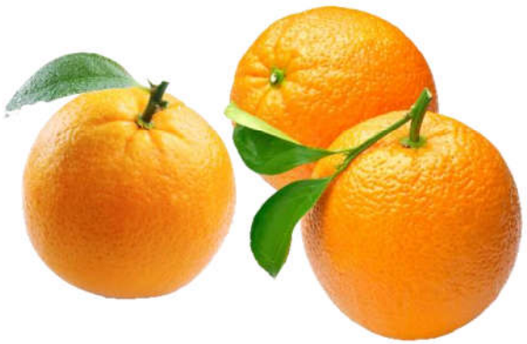
Review:: Generative Model

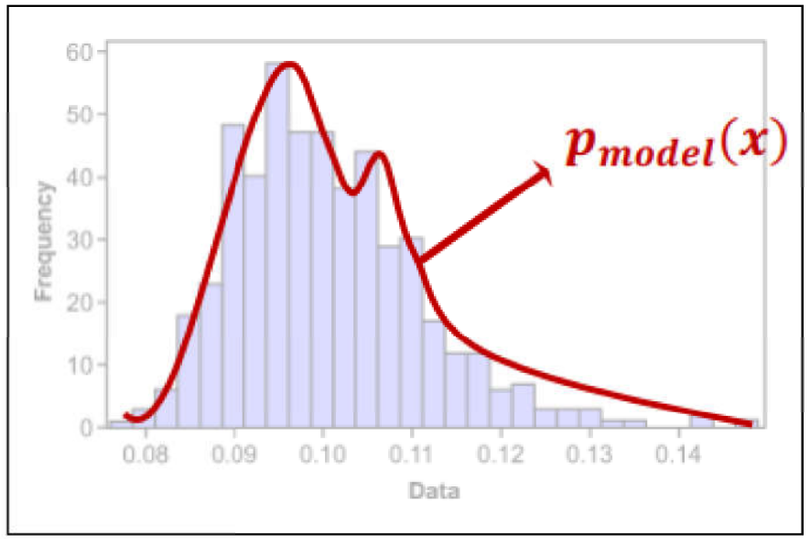


Weird apples

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

Review:: Generative Model

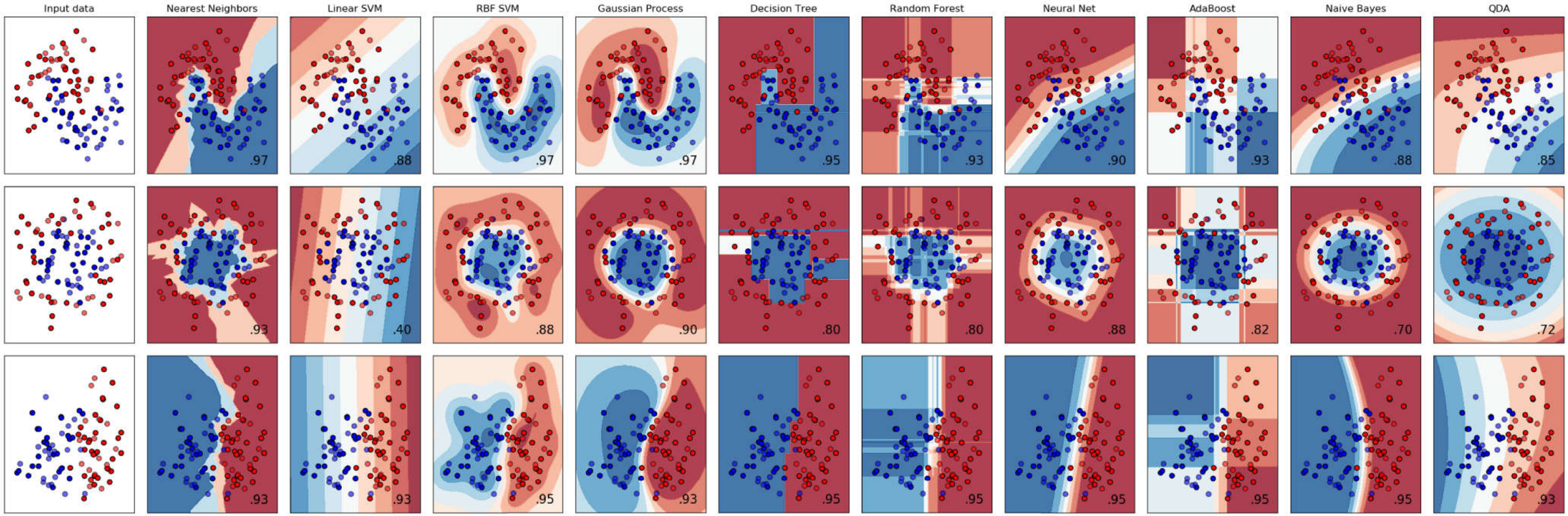




Distribution of the actual images

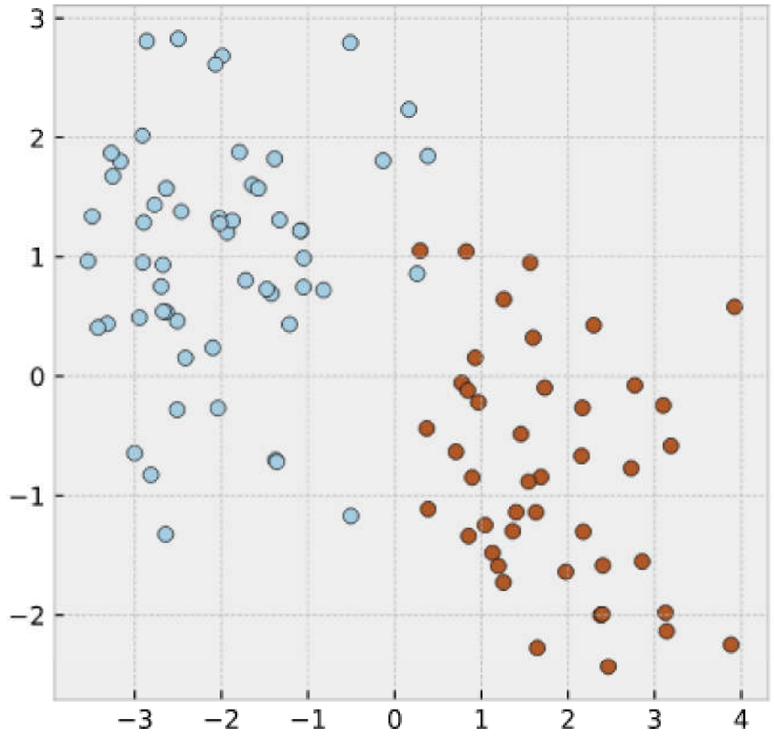
https://www.slideshare.net/BrianKim244/dcgan-77452250

Decision boundary crossing

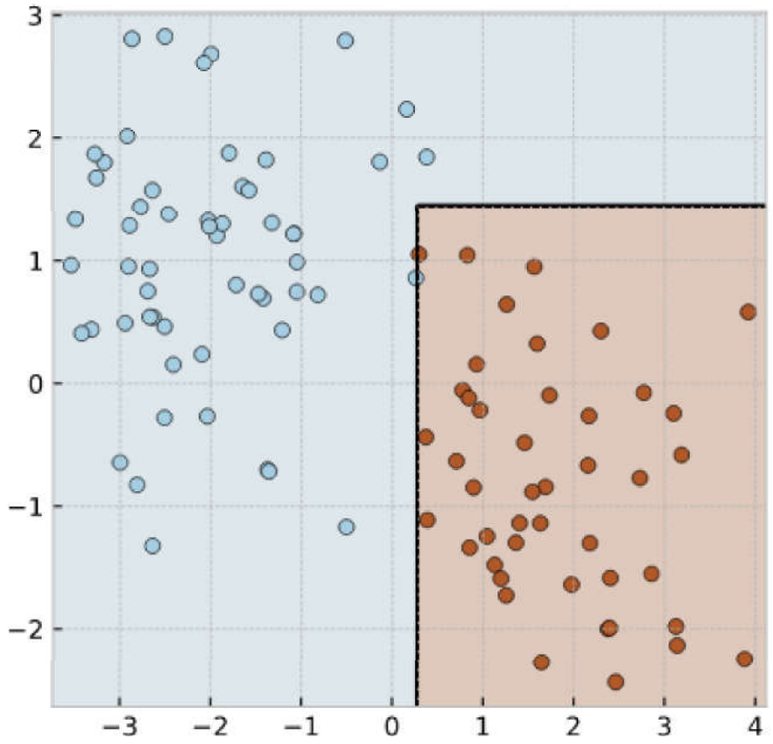


https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html

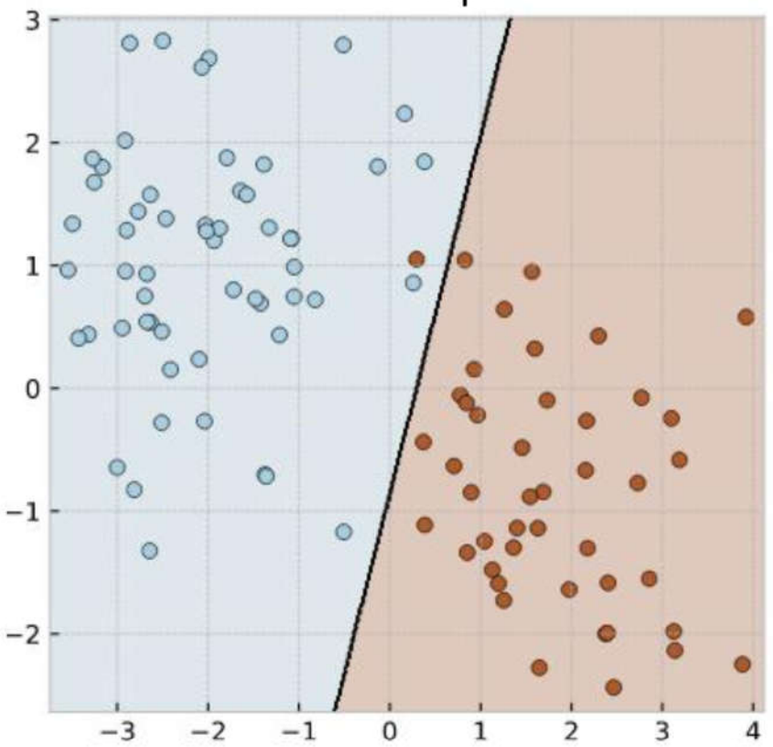
Decision boundary crossing



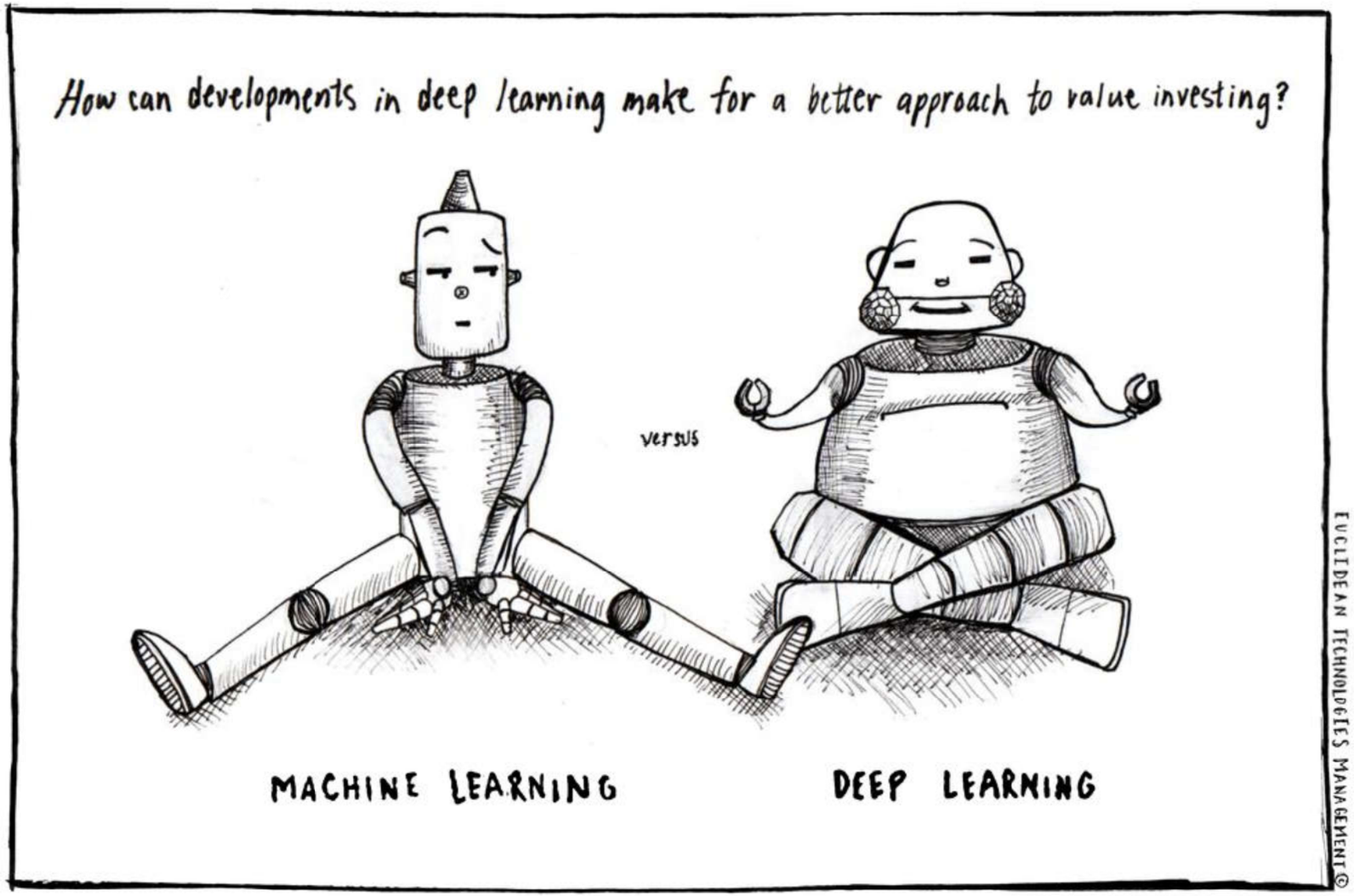
Decision Tree



Perceptron



http://www.cse.chalmers.se/~richajo/dit866/lectures/l3/Plotting%20decision%20boundaries.html



https://www.euclidean.com/deep-learning-and-value-investing

AI and the top innovation

Autonomous driving

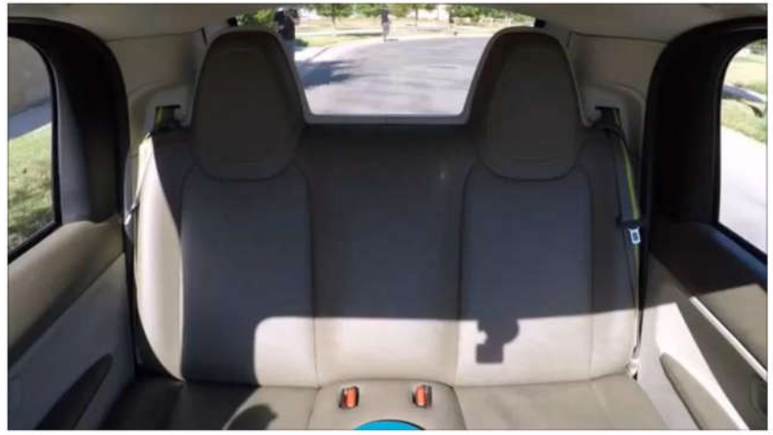
Financial fraud detection

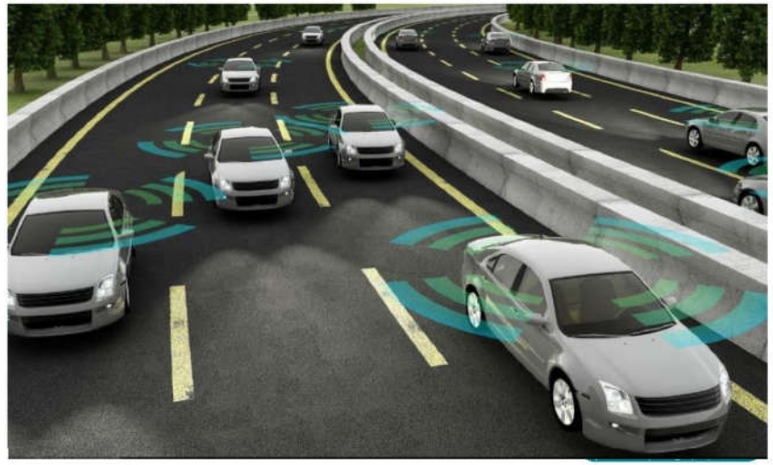
Malware / APT detection

Machine Learning as a Service

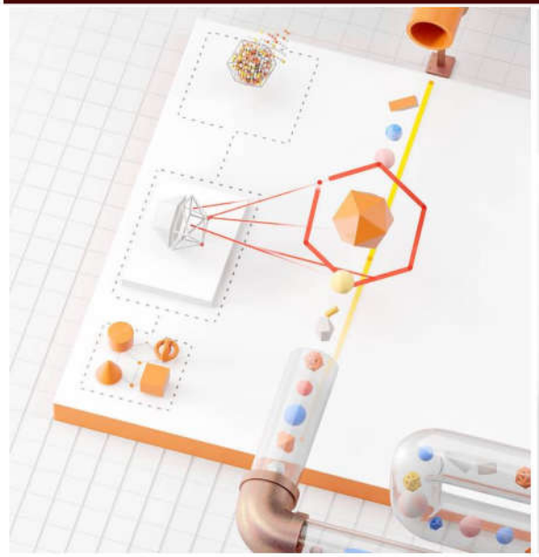
https://www.usenix.org/sites/default/files/conference/protected-files/enigma17_slides_papernot.pdf

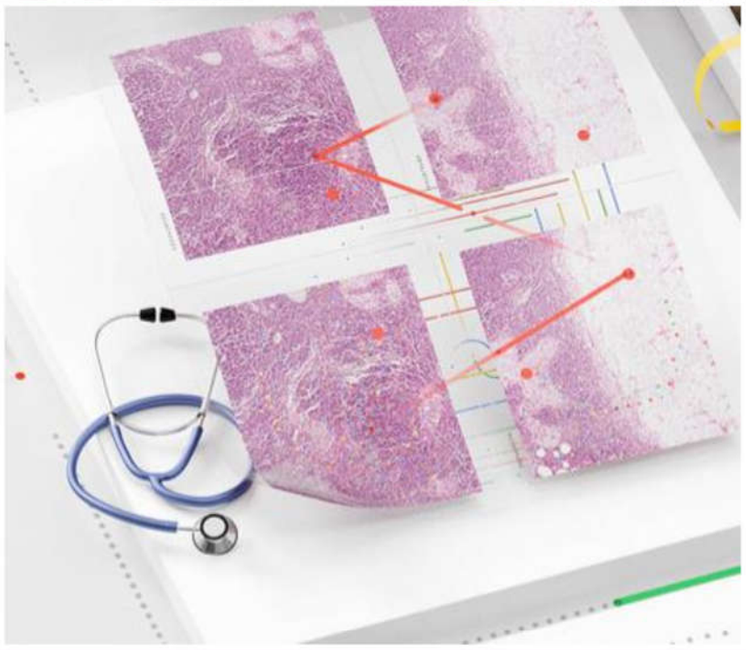
AI and the top innovation





AI and the top innovation





Adversarial examples represent *worst-case* domain shifts

Deep learning researchers status! ☺

December 20 '17

Fooling Google's image-recognition AI 1000x faster

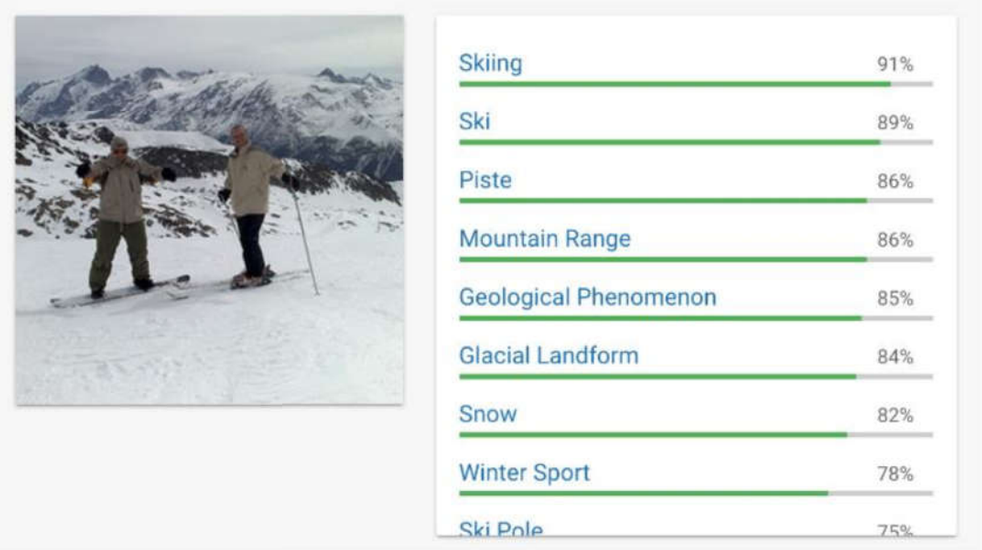
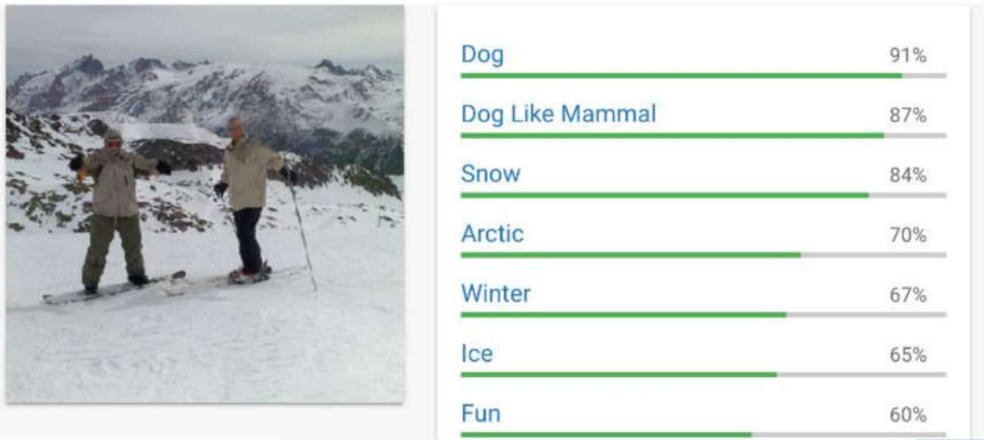


Figure 8. The Google Cloud Vision Demo labelling on the unperturbed image.



November 02 '17

Fooling neural networks w/3D-printed objects

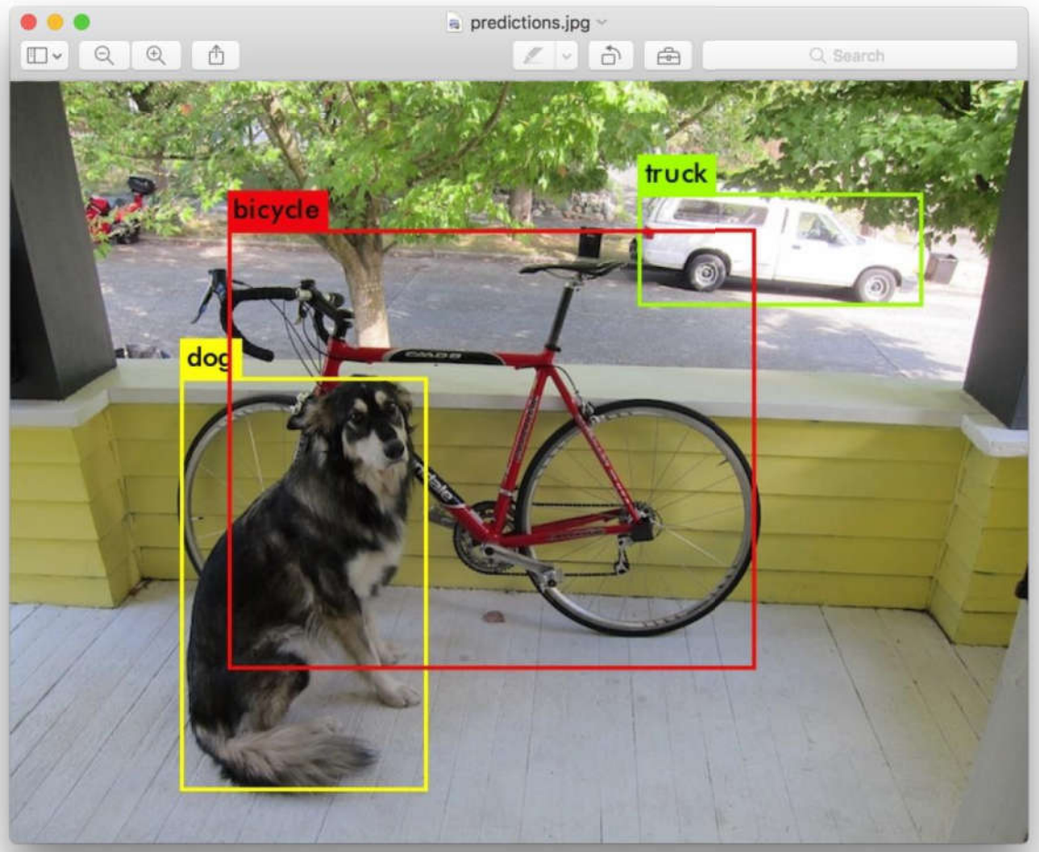
The team 3D-printed a toy turtle that was misclassified as a rifle and a baseball that was classified as an espresso, no matter what angle the neural network views them from.

https://www.csail.mit.edu/news/fooling-neural-networks-w3d-printed-objects

CSAIL :: MIT Computer Science & Artificial Intelligence Lab

- **YOLO: Real-Time Object Detection**

Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**) 2016



https://pjreddie.com/darknet/yolo/

On the line of recent researches …

Fooling automated surveillance cameras: adversarial patches to attack person detection

Simen Thys*
simen.thys@student.kuleuven.be

Wiebe Van Ranst*
wiebe.vanranst@kuleuven.be

Toon Goedemé
toon.goedeme@kuleuven.be
KU Leuven
EAVISE, Technology Campus De Nayer, KU Leuven, Belgium.
* Authors contributed equally to this paper.

CVPR LONG BEACH CALIFORNIA June 16-20, 2019

https://syncedreview.com/2019/04/24/now-you-see-me-now-you-dont-fooling-a-person-detector/

On the line of recent researches …

Cloakwear

Home

Catalog

Featured Item

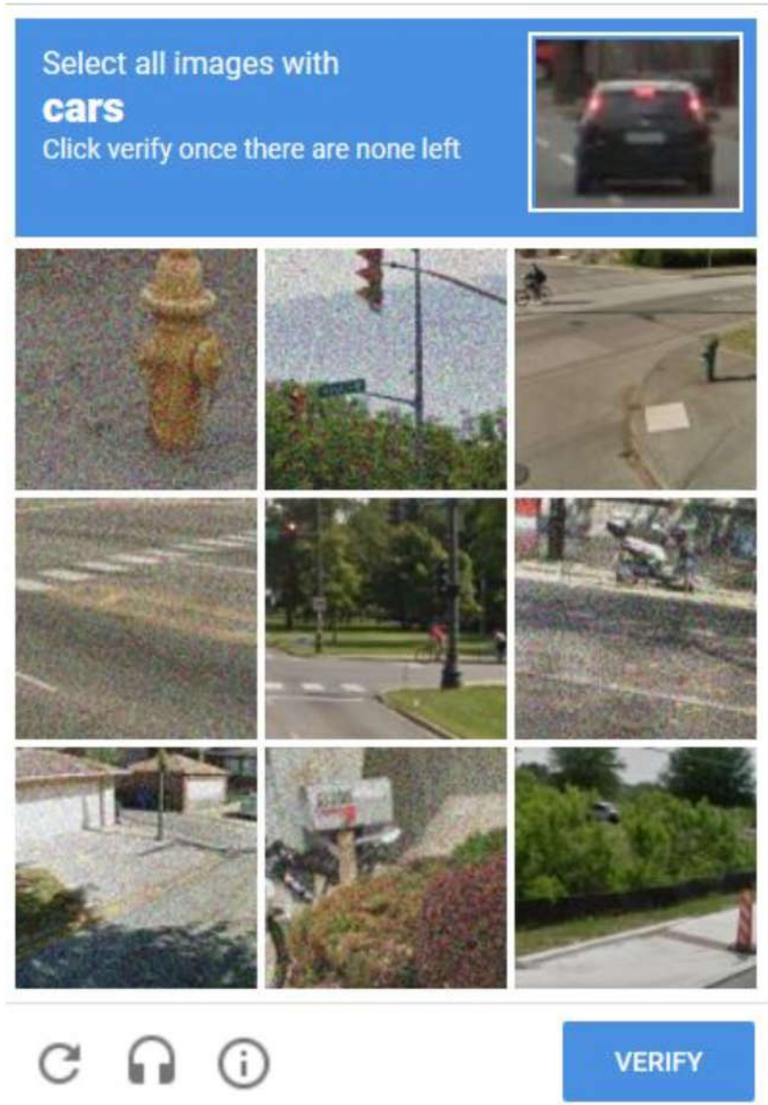
Adversarial T-shirt
From $30.44

Adversarial Long Sleeve Shirt
From $34.71

Adversarial Hoodie
From $43.39

https://cloakwear.co/

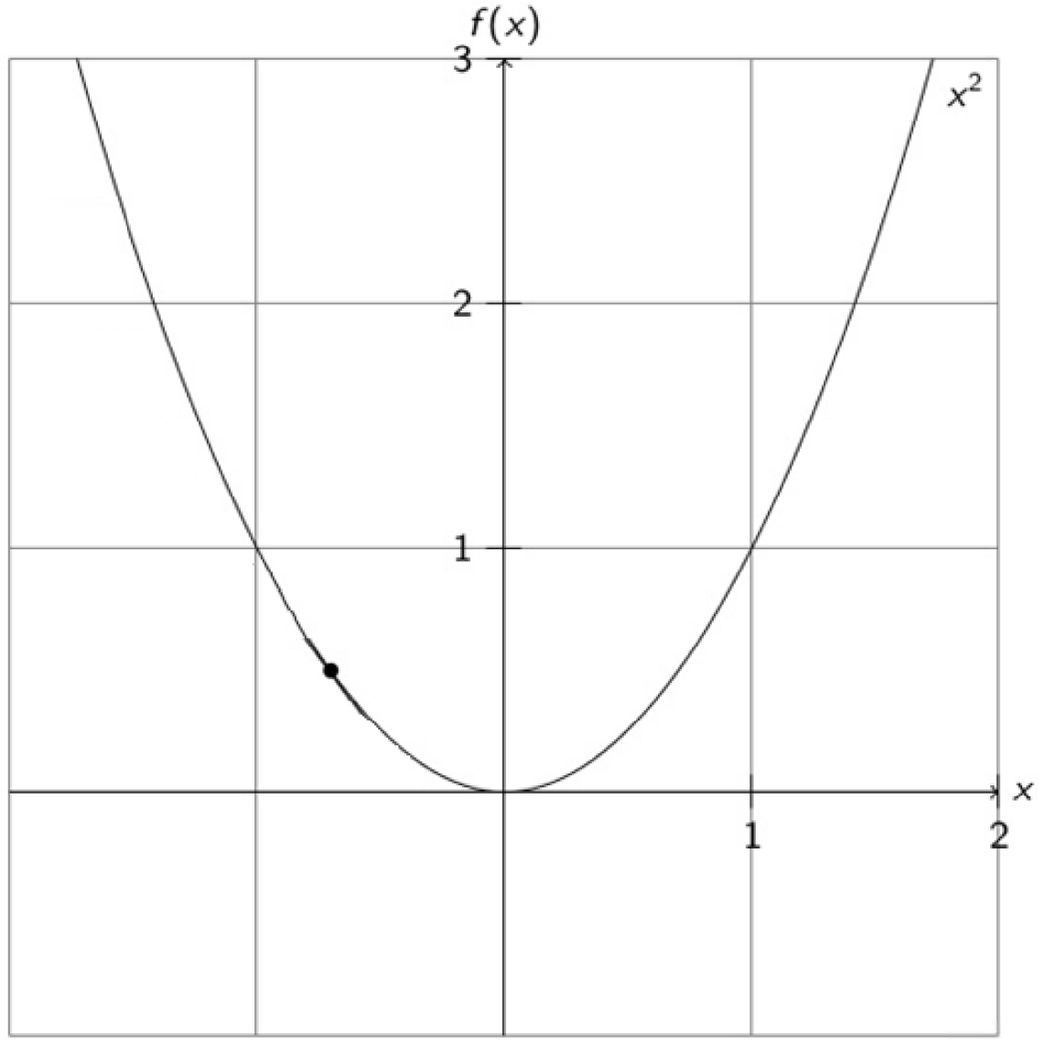
An interesting usage :: Google verification!



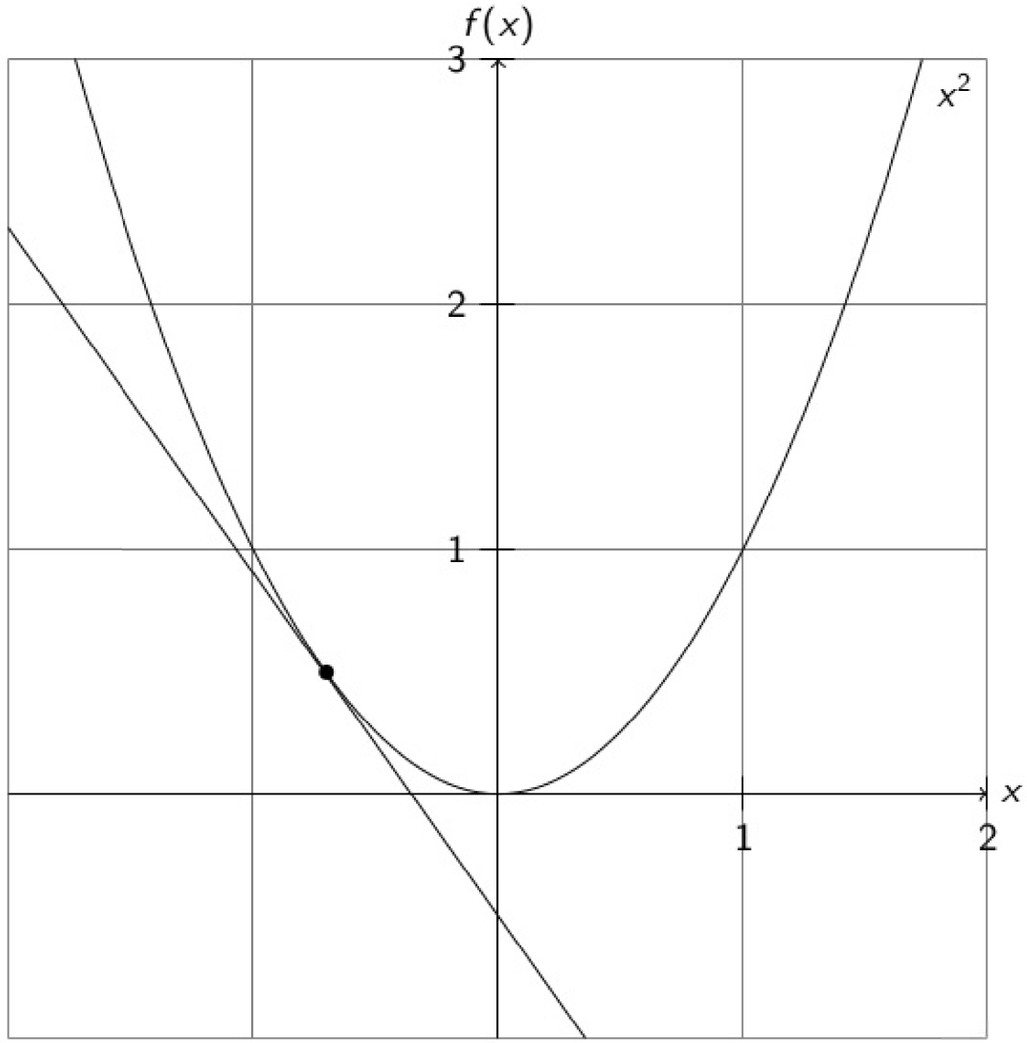
Deep learning researchers status! ☺

Delving into the problem …

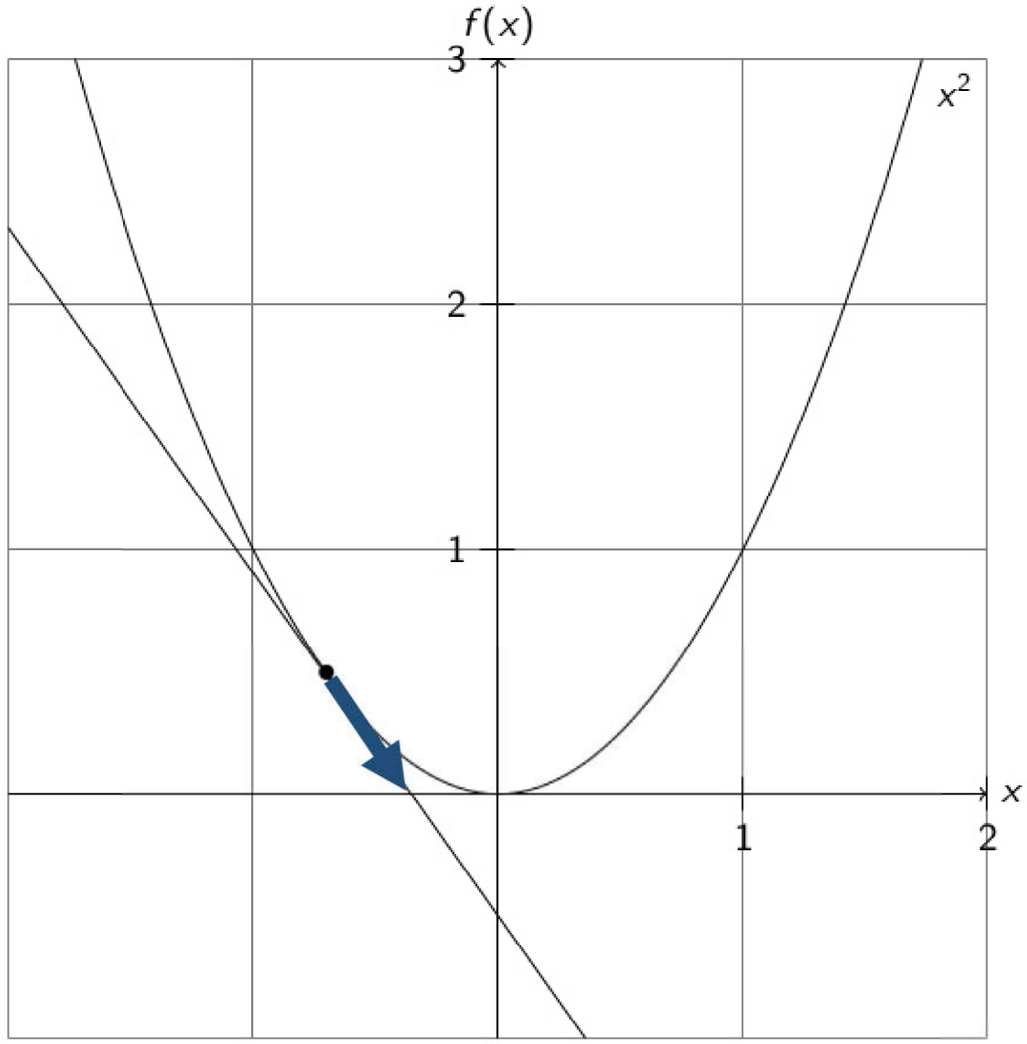
Motivation



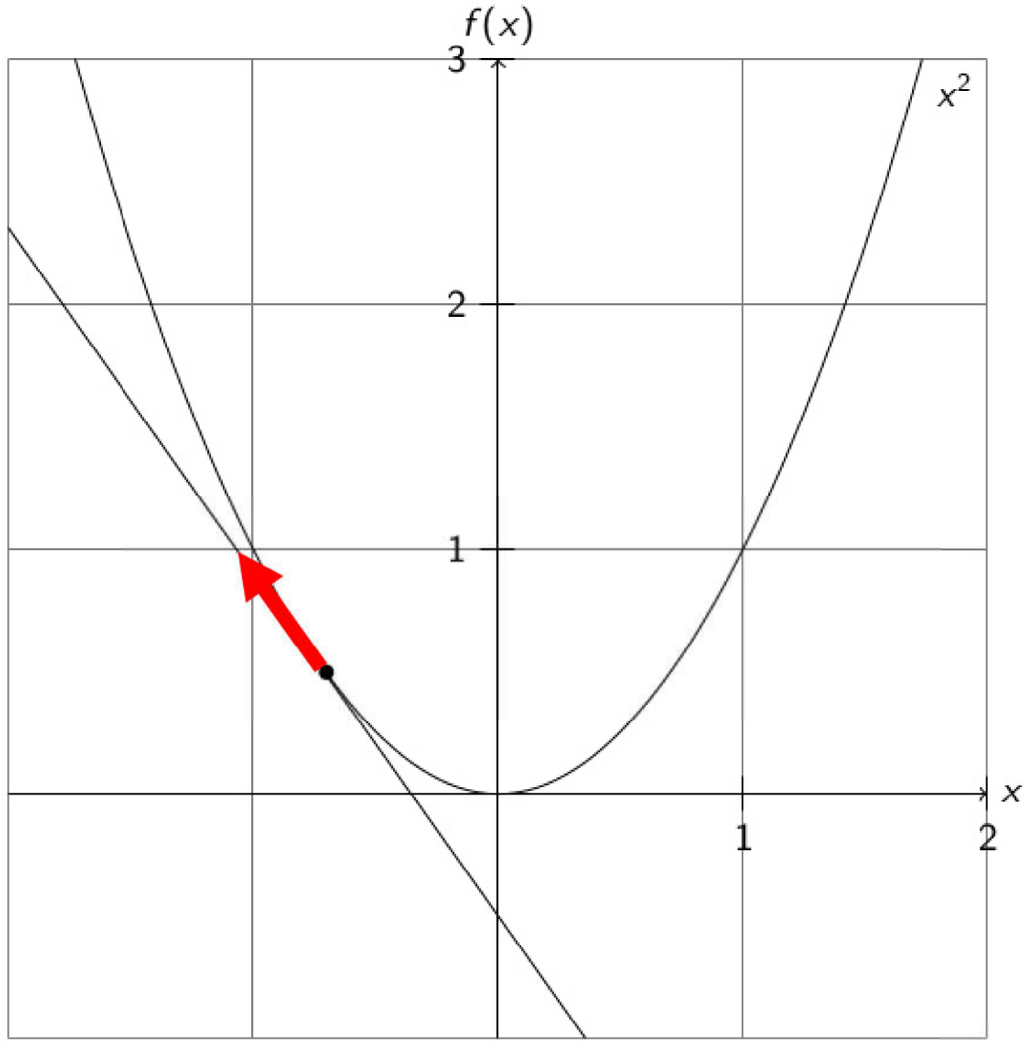
Gradient-descent



Gradient-descent



Gradient-descent



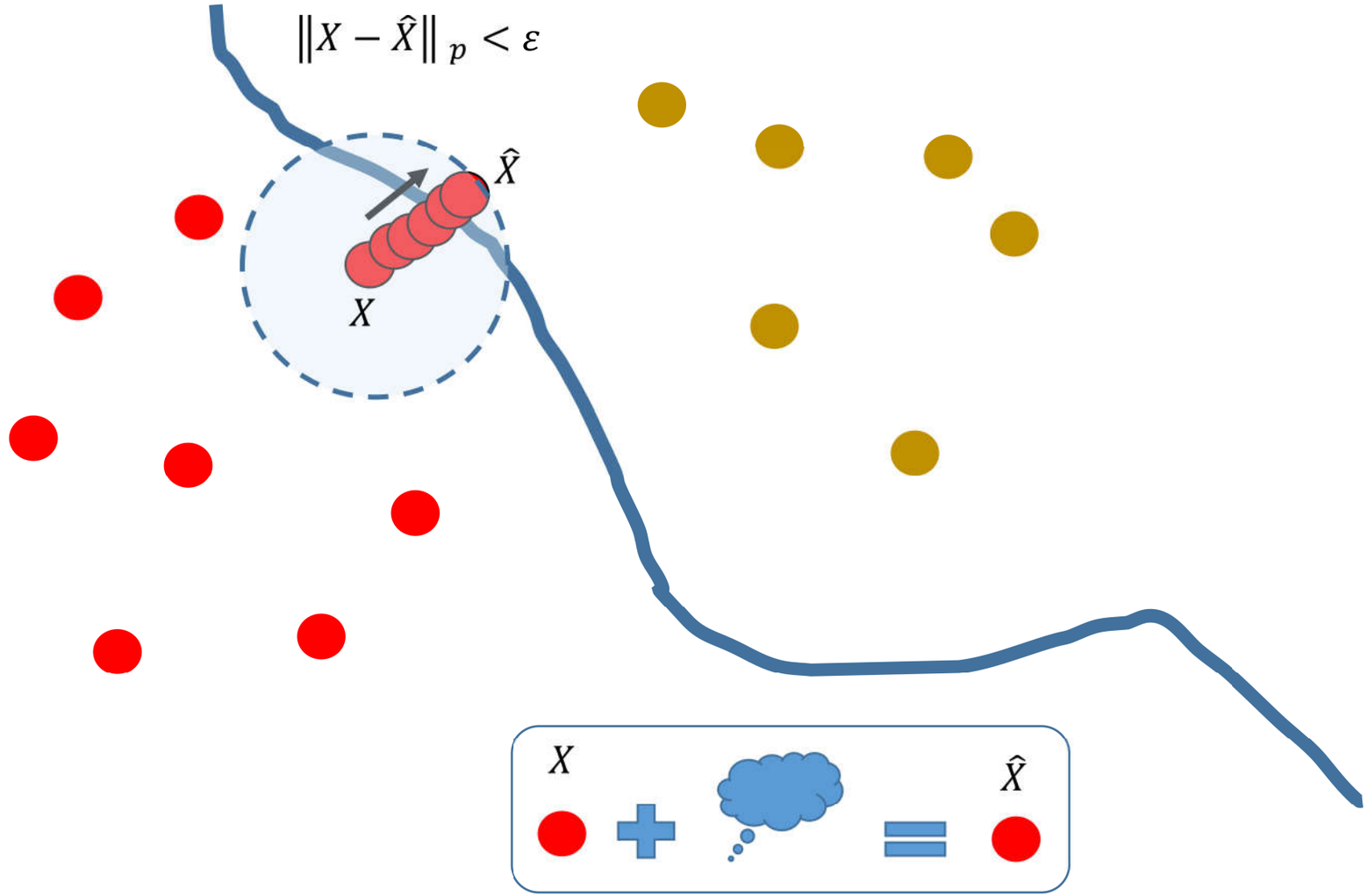
Attacks ?

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

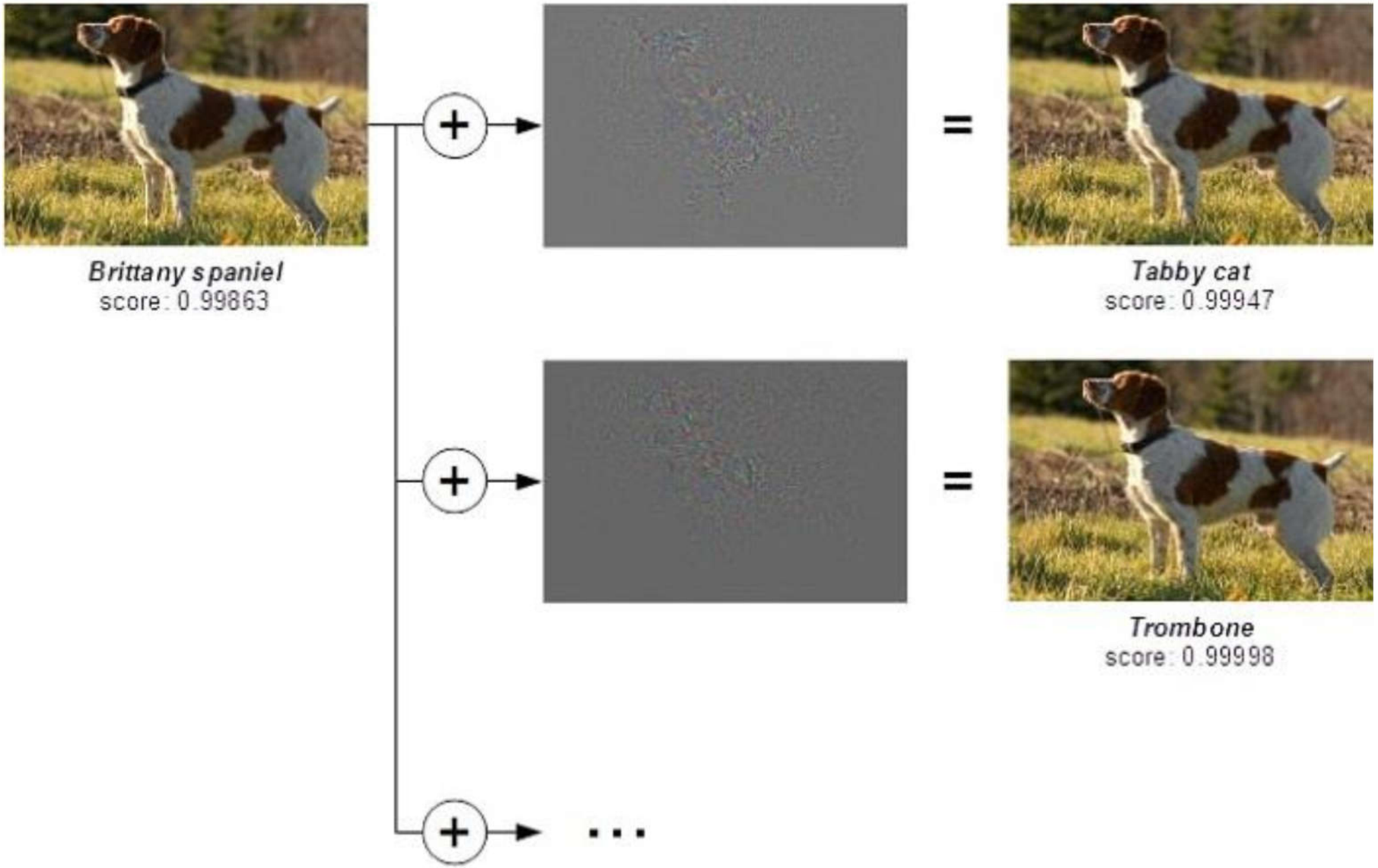
Review the Clever Hans story!

("Clever Hans, Clever Algorithms", Bob Sturm)

What is Adversarial Example?



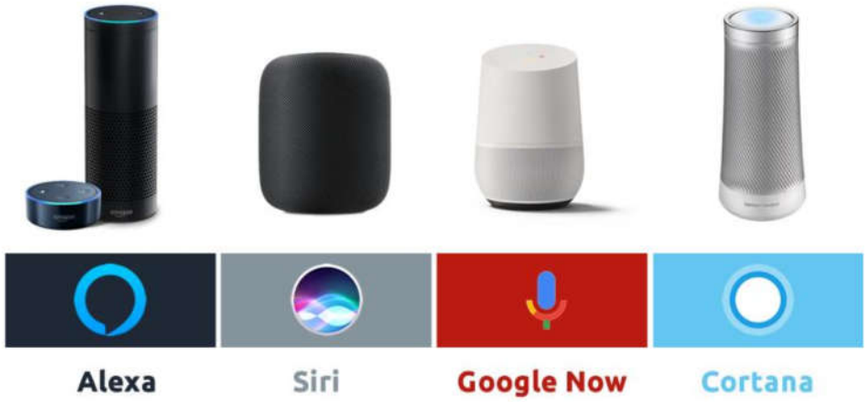
Examples of Adversarial Example

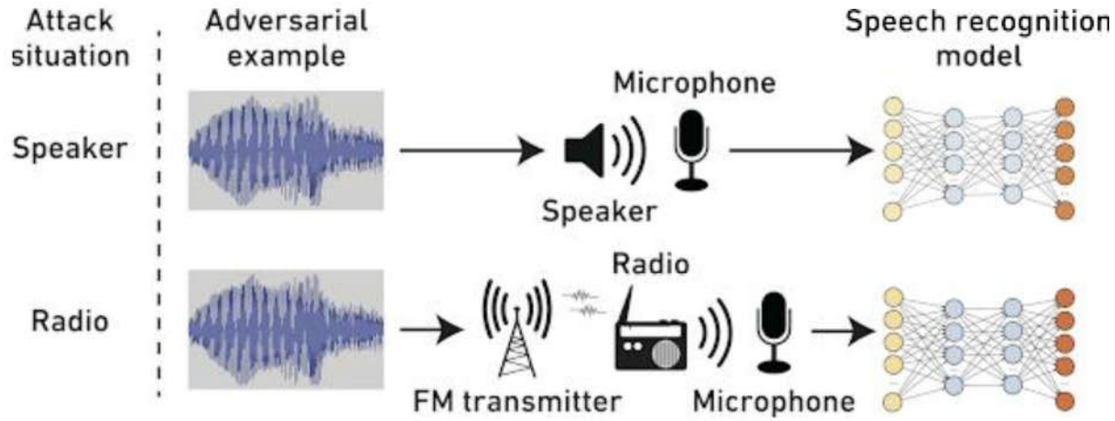


https://www.kdnuggets.com/2018/10/adversarial-examples-explained.html

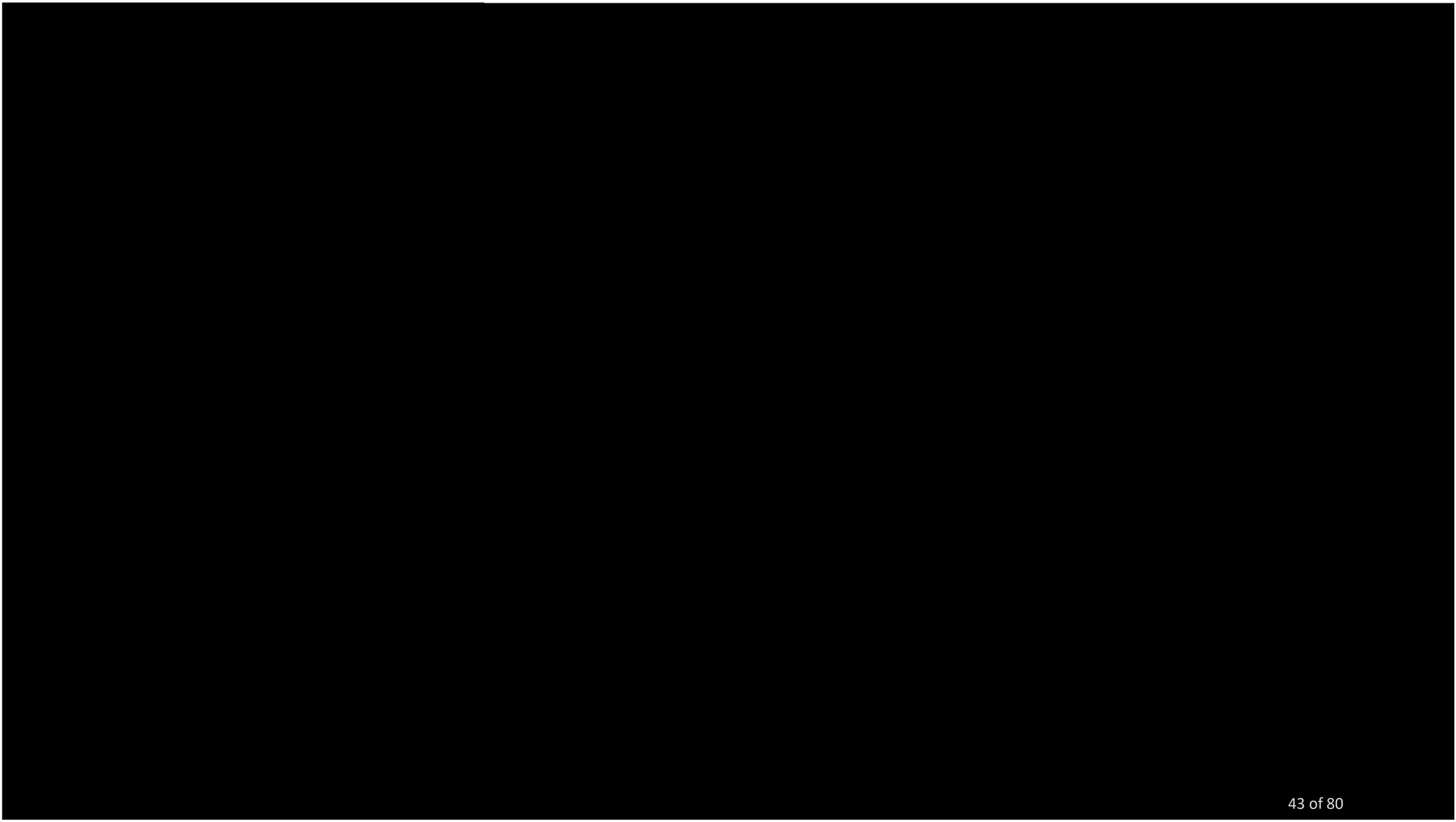

STOP
HATE


STOP
STOP

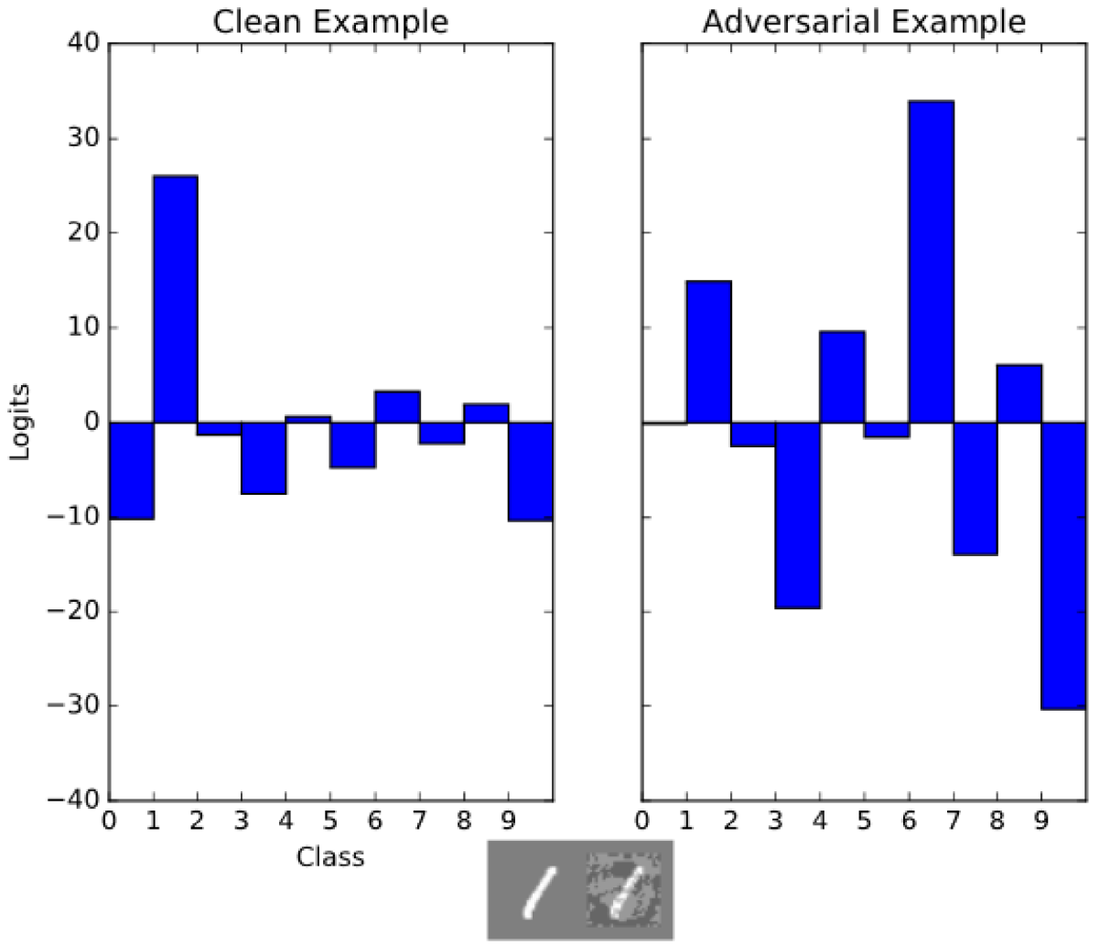

Alexa
Siri
Google Now
Cortana


Attack situation
Adversarial example
Speech recognition model
Microphone
Speaker
Speaker
Radio
Radio
FM transmitter
Microphone

From WWW

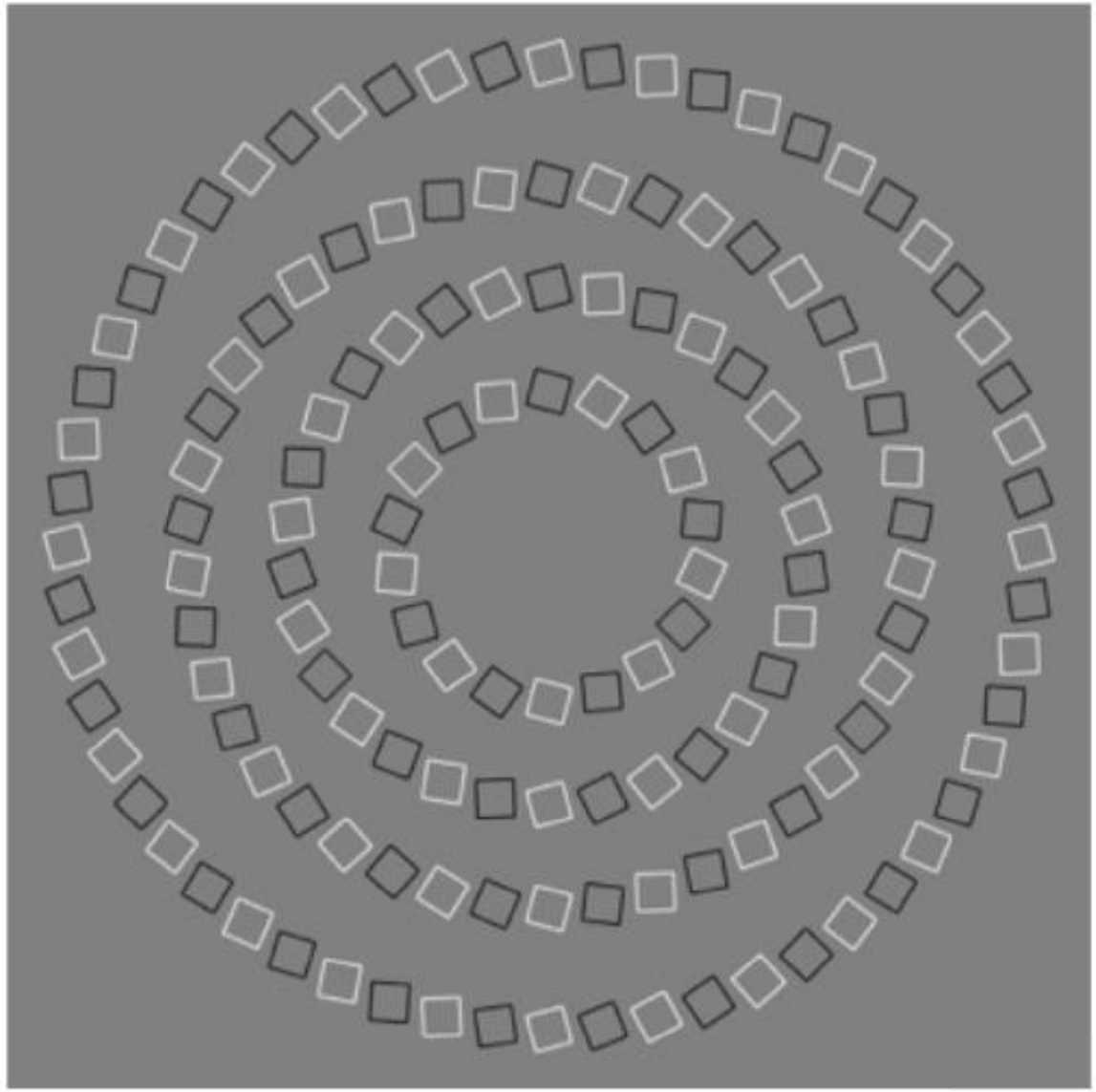


Perturbation's effect on class distributions



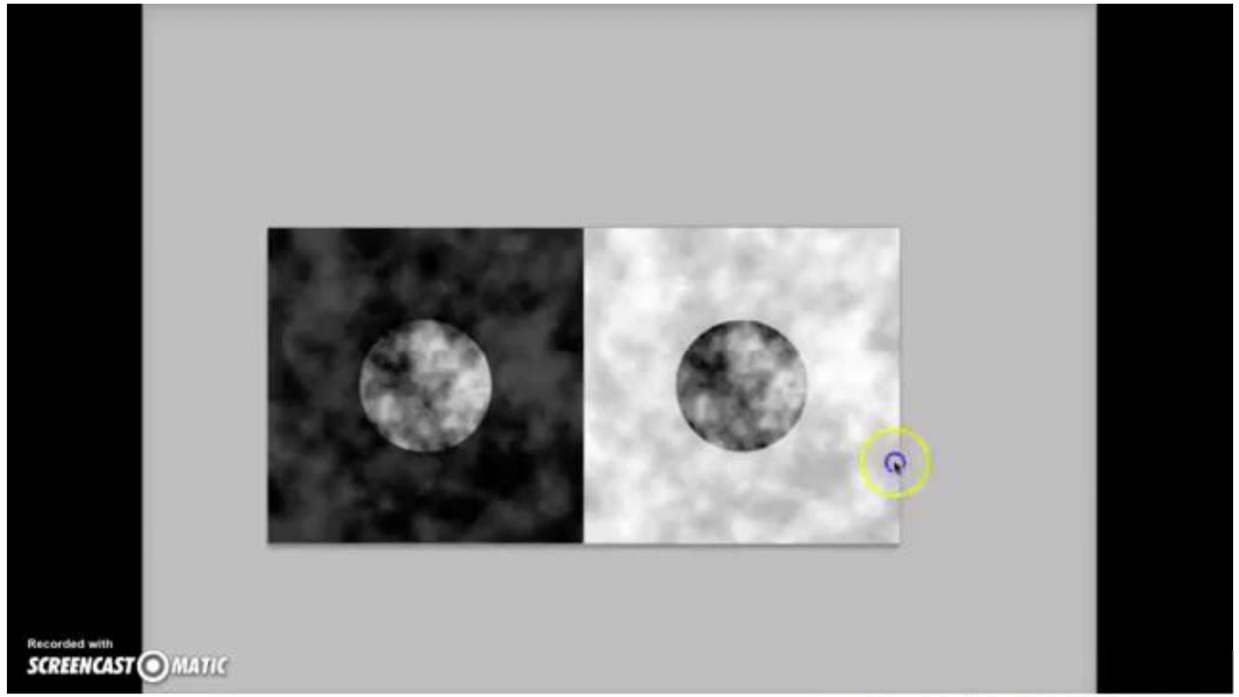
Adversarial examples in the human visual system

- Circles are concentric but appear intertwining



(Pinna and Gregory, 2002)

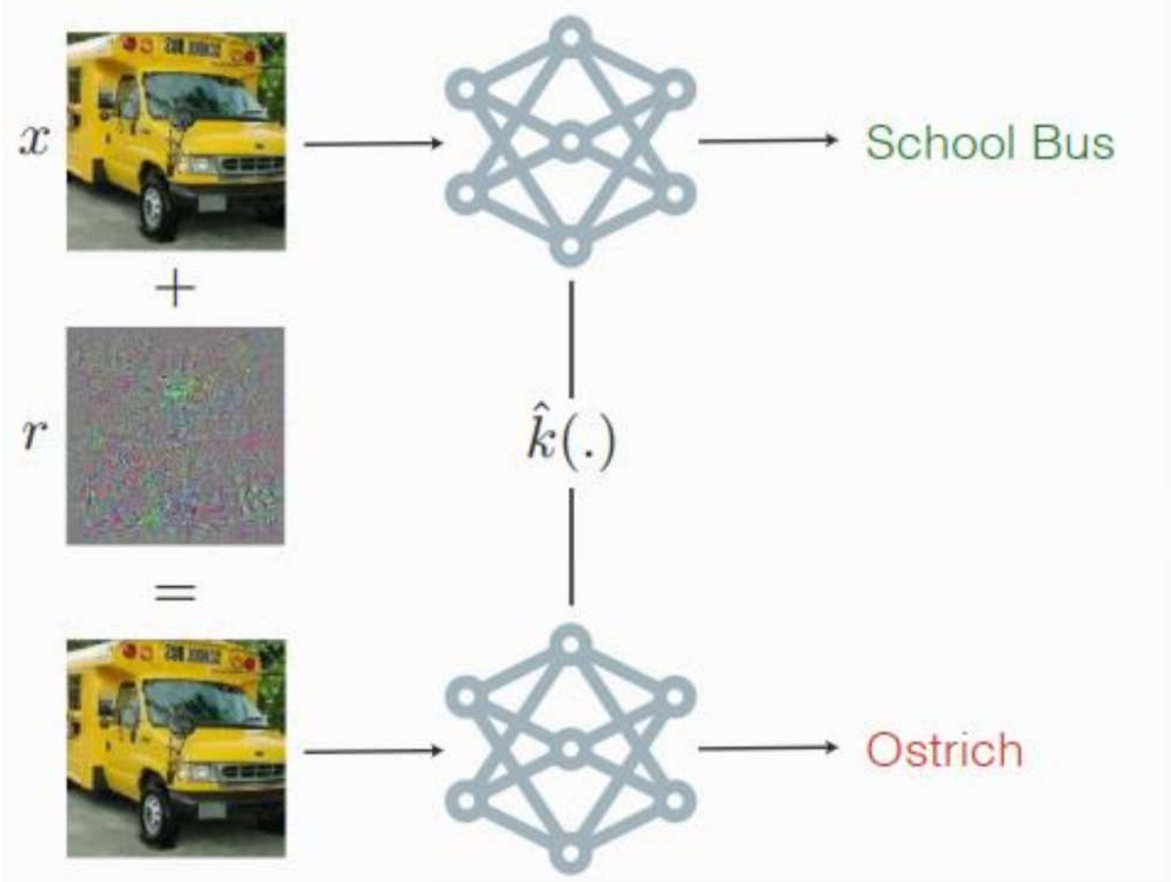
Attack: Through the human eye



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

Adversarial timeline ...

Interiguing properties of Neural Network
C. Szegedy et al. ICLR, 2014



$$r^* = \arg\min_r J(\hat{k}(x+r), y) + C\|r\|$$

Adversarial timeline …

Interiguing properties of Neural Network
C. Szegedy et al. ICLR, 2014

Explaining and Harnessing Adversarial Example
I.J. Goodfellow et al. ICLR, 2015

Perturbation

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

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

Gradient of the cost function

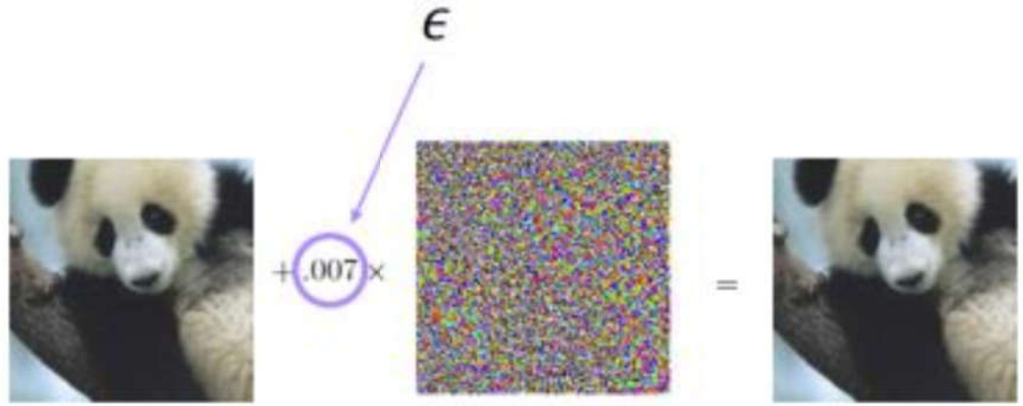


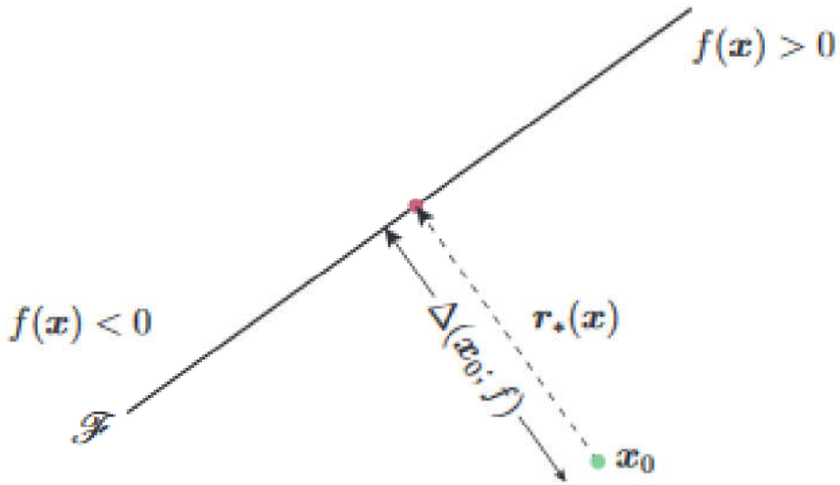
Figure 1: A demonstration of fast adversarial example generation applied to GoogLeNet (Szegedy et al., 2014a) on ImageNet. By adding an imperceptibly small vector whose elements are equal to the sign of the elements of the gradient of the cost function with respect to the input, we can change GoogLeNet's classification of the image. Here our ϵ of .007 corresponds to the magnitude of the smallest bit of an 8 bit image encoding after GoogLeNet's conversion to real numbers.

Adversarial timeline …

Interiguing properties of Neural Network
C. Szegedy et al. ICLR, 2014

Explaining and Harnessing Adversarial Example
I.J. Goodfellow et al. ICLR, 2015

Deep Fool: a simple and accurate method to fool deep neural networks
S.M. Moosavi-dezfooli et al. CVPR, 2016



Algorithm 1 DeepFool for binary classifiers

input: Image x, classifier f.
output: Perturbation $\hat{r}$.
Initialize $x_0 \leftarrow x$, $i \leftarrow 0$.
while $\text{sign}(f(x_i)) = \text{sign}(f(x_0))$ **do**
  $r_i \leftarrow -\frac{f(x_i)}{\|\nabla f(x_i)\|_2^2} \nabla f(x_i)$,
  $x_{i+1} \leftarrow x_i + r_i$,
  $i \leftarrow i + 1$.
end while
return $\hat{r} = \sum_i r_i$.

Adversarial timeline …

Interiguing properties of Neural Network
C. Szegedy et al. ICLR, 2014

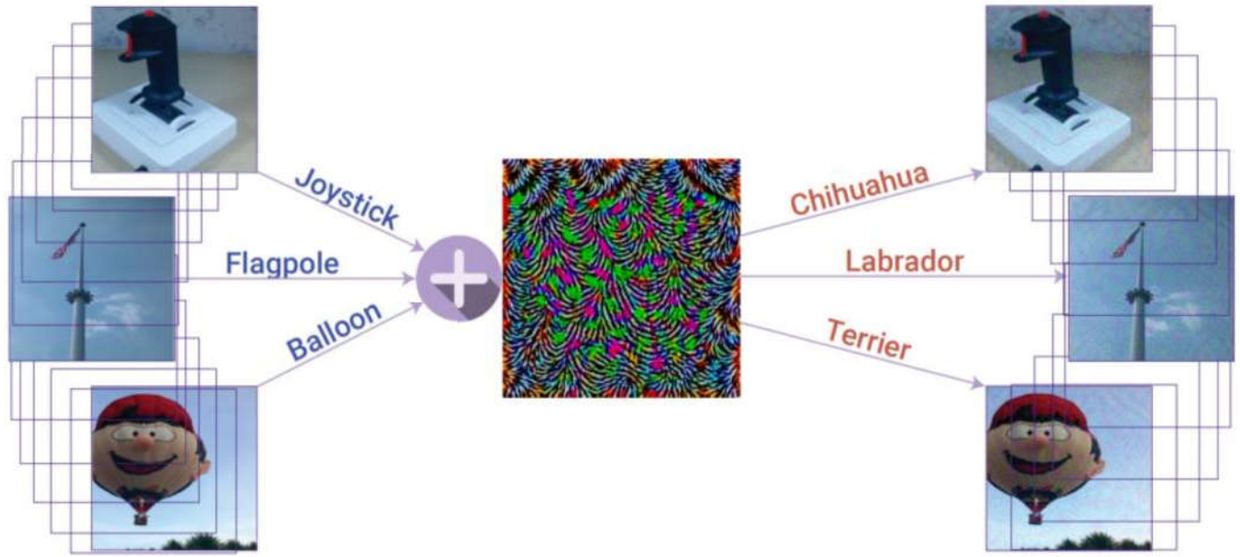
Explaining and Harnessing Adversarial Example
I.J. Goodfellow et al. ICLR, 2015

Deep Fool: a simple and accurate method to fool deep neural networks
S.M. Moosavi-dezfooli et al. CVPR, 2016

Universal adversarial perturbations
S.M. Moosavi-dezfooli et al. CVPR, 2017

$$\mathbb{P}_{x\sim\mu}\left(\hat{k}(x+v)\neq\hat{k}(x)\right)\geq 1-\delta$$

$$\|v\|_p\leq\xi$$



Adversarial timeline …

Interiguing properties of Neural Network
C. Szegedy et al. ICLR, 2014

Explaining and Harnessing Adversarial Example
I.J. Goodfellow et al. ICLR, 2015

Deep Fool: a simple and accurate method to fool deep neural networks
S.M. Moosavi-dezfooli et al. CVPR, 2016

Universal adversarial perturbations
S.M. Moosavi-dezfooli et al. CVPR, 2017

Towards evaluating the robustness of neural networks
Carlini et al. Security and Privacy (S&P). 2017

Adversarial timeline …

Interiguing properties of Neural Network
C. Szegedy et al. ICLR, 2014

Explaining and Harnessing Adversarial Example
I.J. Goodfellow et al. ICLR, 2015

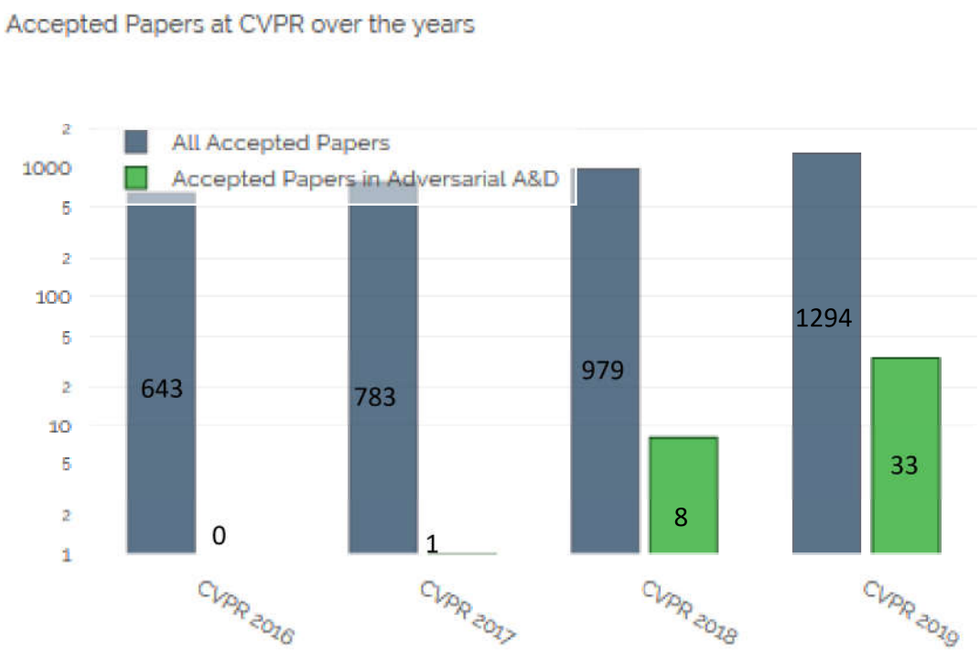
Deep Fool: a simple and accurate method to fool deep neural networks
S.M. Moosavi-dezfooli et al. CVPR, 2016

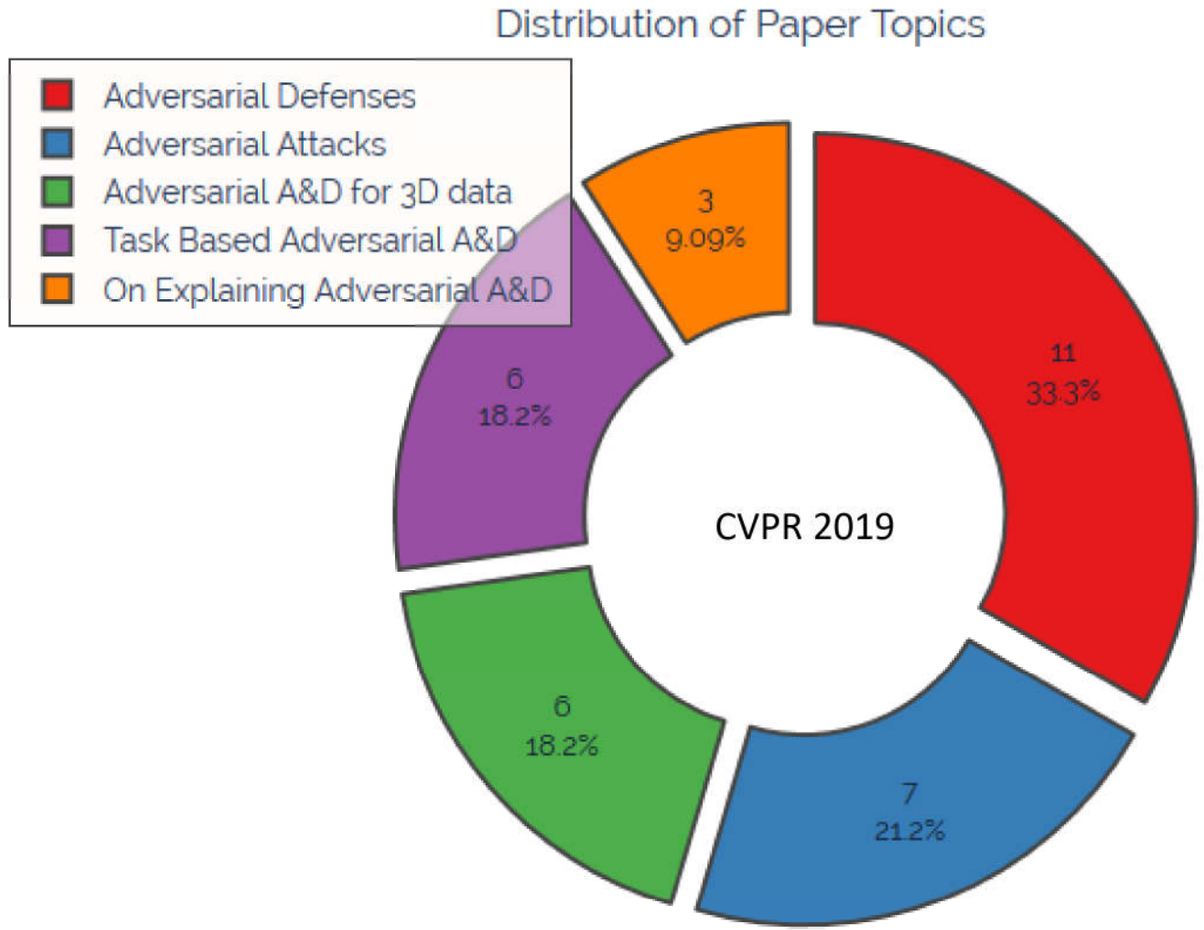
Universal adversarial perturbations
S.M. Moosavi-dezfooli et al. CVPR, 2017

Towards evaluating the robustness of neural networks
Carlini et al. Security and Privacy (S&P). 2017

SparseFool: a few pixels make a big difference
Modas et al. CVPR, 2019

Increasing papers on Adversarial Attacks & Defenses

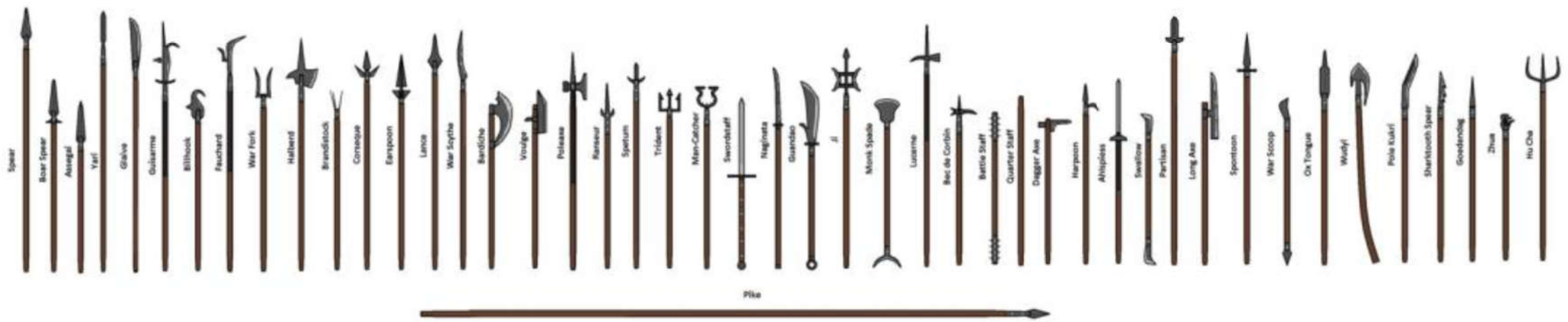




The number of accepted papers at CVPR over the years. Note the sharp rise in accepted papers in Adversarial A&D.

http://fna.ir/a5d

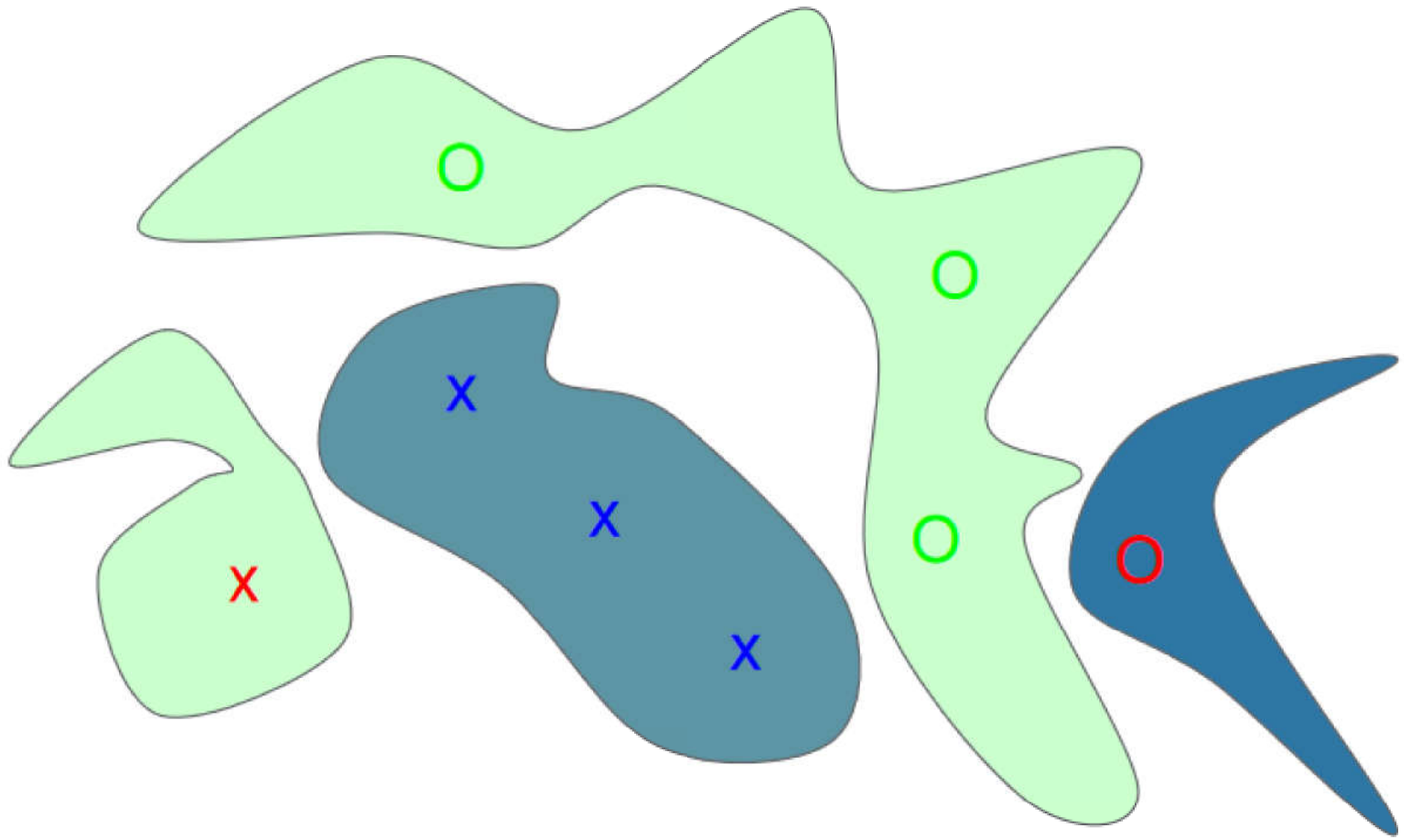
Various methods of attacks



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

Adversarial Examples from Overfitting

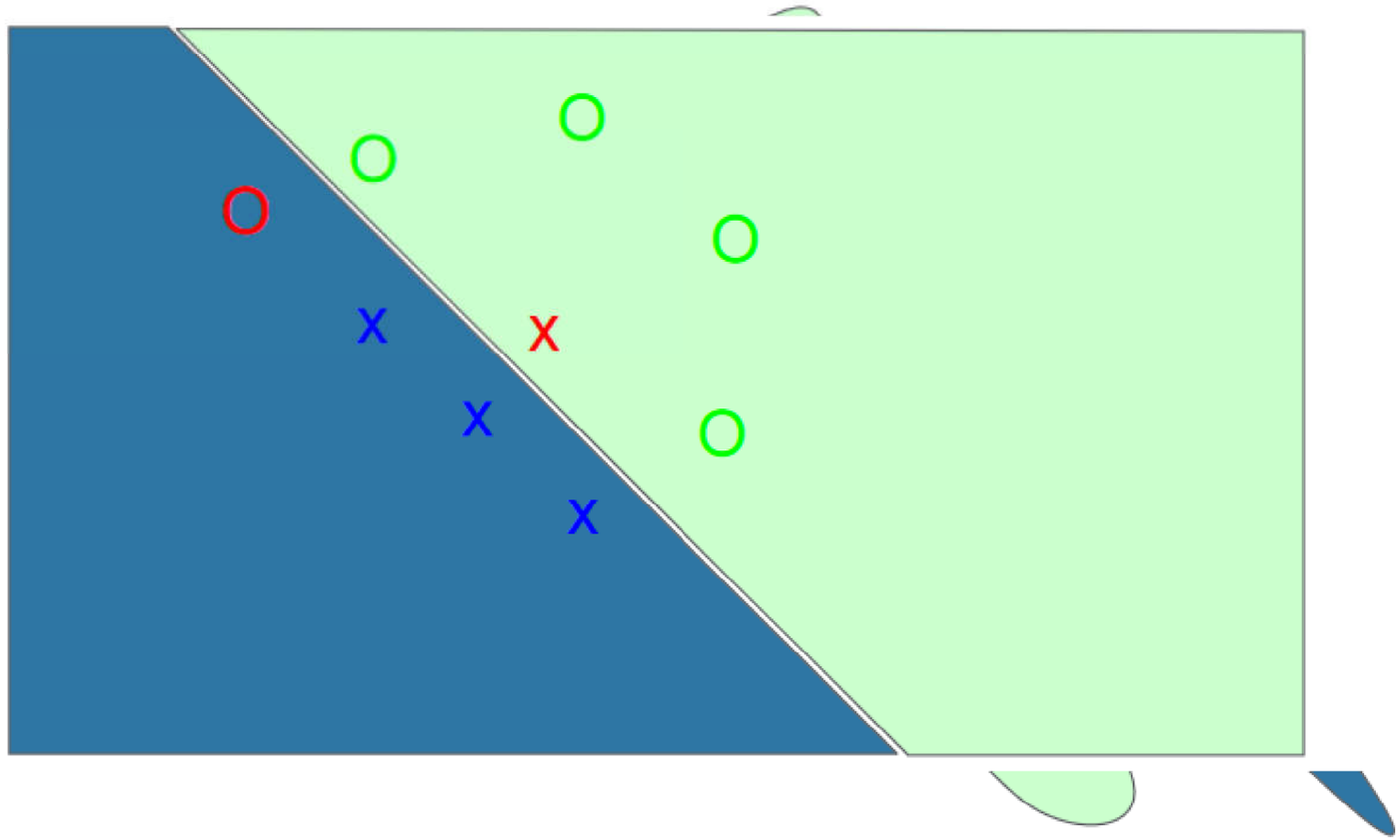
- Adversarial Examples rooted in :
  - **Overfitting**
  - Excessive Linearity



(Goodfellow 2016)

Adversarial Examples from Excessive Linearity

- Adversarial Examples rooted in :
  - Overfitting
  - **Excessive Linearity**



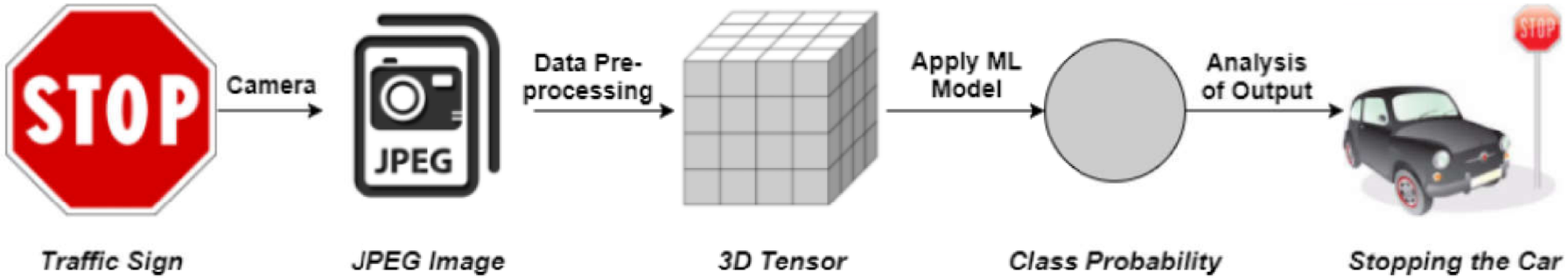
(Goodfellow 2016)

Adversarial Threat Model

- **The Attack Surface**
- **The Adversarial Capabilities**
- **The Adversarial Goals**

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)

- **The Adversarial Capabilities**

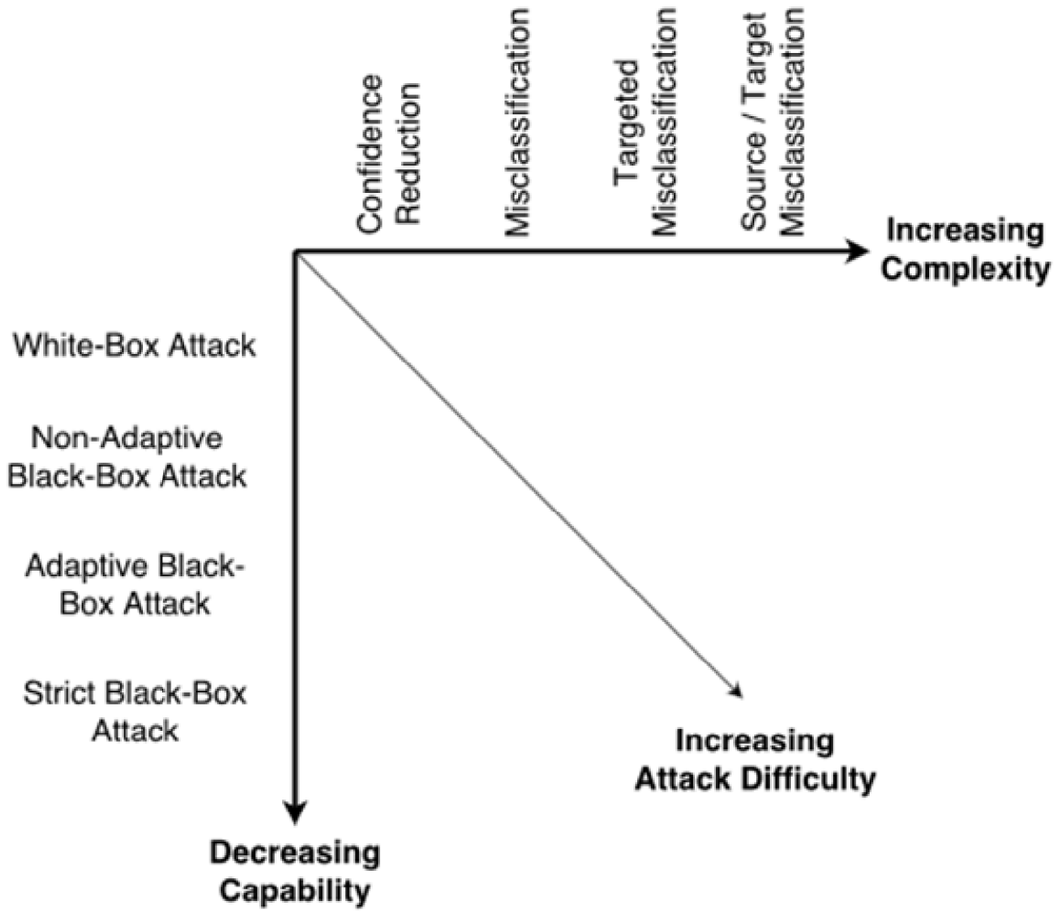
  - **Training Phase Capabilities**
    - Data Injection :: does not have any access to the training data as well as to the learning algorithm but has ability to augment a new data to the training set
    - Data Modification :: does not have access to the learning algorithm but has full access to the training data
    - Logic Corruption :: meddle with the learning algorithm

  - **Testing Phase Capabilities**
    - White-Box Attacks :: an adversary has total knowledge about the model (f), algorithm (train), training data distribution (μ), parameters (θ) of the fully trained model architecture
    - Black-Box Attacks :: no knowledge about the model and uses information about the settings or past inputs
      - Non-Adaptive Black-Box Attack :: only gets access to the target model's training data distribution (μ)
      - Adaptive Black-Box Attack :: doesn't have any information regarding the training process but can access the target model as an oracle
      - Strict Black-Box Attack :: may not contain the data distribution(μ) but has the ability to collect the input-output pairs(x,y) from the target classifier. However, he can not change the inputs to observe the changes in output like an adaptive attack procedure

http://fna.ir/a5g

- **Adversarial Goals:**
  - Confidence Reduction
    - The adversary tries to reduce the confidence of prediction for the target model
  - Misclassification
    - The adversary tries to alter the output classification of an input example to **any** class different from the original class.
  - Targeted Misclassification
    - The adversary tries to produce inputs that force the output of the classification model to be a **specific** target class
  - Source/Target Misclassification
    - The adversary attempts to force the output of classification for a **specific** input to be a **particular** target class

Adversarial Threat Model



Attack Difficulty with respect to adversarial capabilities and goals for Evasion Attacks

http://fna.ir/a5g

Adversarial example crafting procedures :

1. Direction sensitivity estimation

- The adversary evaluate the sensitivity of a class change to each input feature
- By identifying the direction in the data manifold around the example X

$$X_* = X + \arg\min_{\delta X}\{\|\delta X\| : F(X + \delta X) \neq F(X)\}$$

Most DNN models make this formulation non-linear and non-convex, making it hard to find a closed-solution in most of the cases

2. Perturbation selection

- The adversary exploits the knowledge of sensitive information to select a perturbation
- Selecting the perturbation δX

3. Replace X with $X + \delta X$

Adversarial Example Generation

Different techniques to find an approximate solution for adversarial example finding problem : **(Direction Sensitivity Estimation)**

- L-BFGS :: Szegedy et al.

$$\arg\min_r f(x+r) = l \quad \text{s.t.} \quad (x+r) \in D$$

- Fast Gradient Sign Method (FGSM) :: Goodfellow et al

$$X_* = X + \epsilon * sign(\nabla_x J(X, y_{true}))$$

  - Target Class Method :: Kurakin et al.

$$X_* = X - \epsilon * sign(\nabla_x J(X, y_{target}))$$

  - Basic Iterative Method :: Kurakin et al.

$$X_*^0 = X; \quad X_*^{n+1} = Clip_{X,\epsilon}\{X_*^n + \alpha * sign(\nabla_x J(X_*^n, y_{true}))\}$$

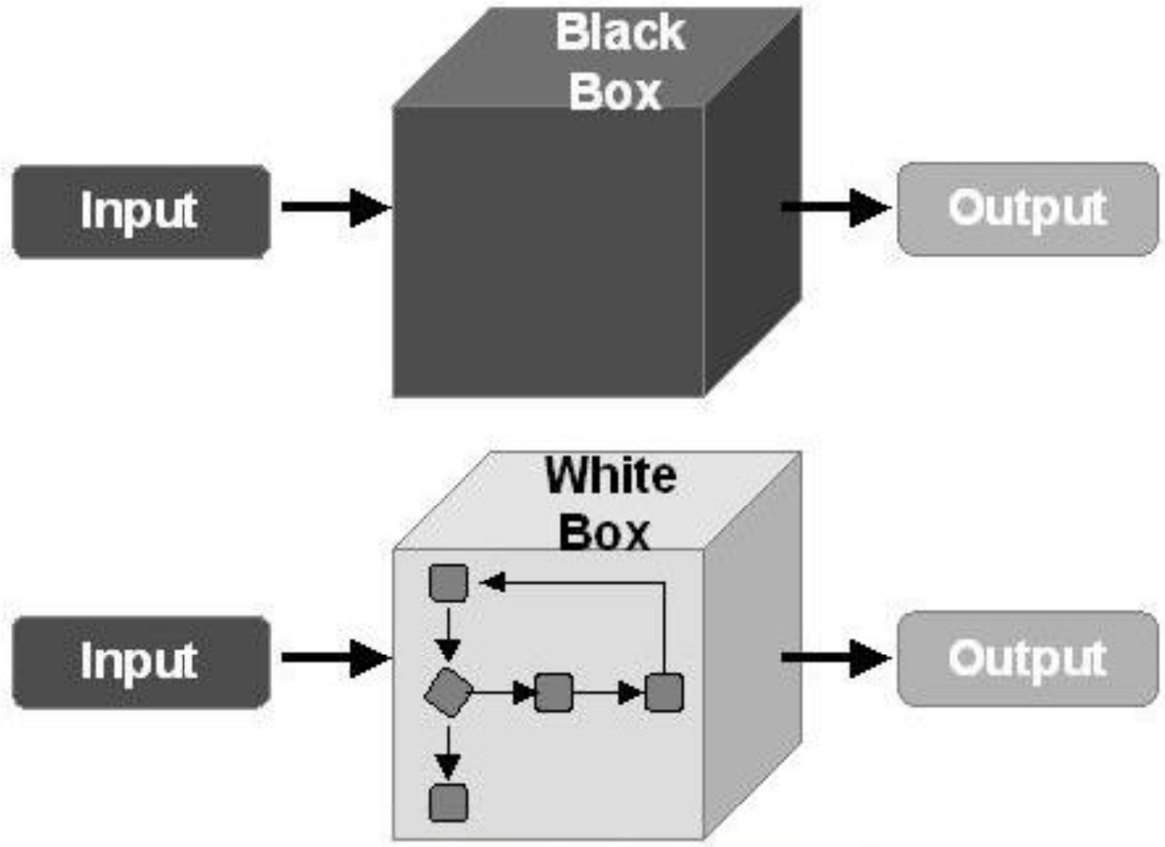
- Jacobian Based Method :: Papernot et al.
  - Finding sensitivity direction by using **forward derivative**, which is the Jacobian of the trained model F

saliency value

$$S(x,t)[i] = \begin{cases} 0, & \text{if } \frac{\partial F_t}{\partial x_i}(x) < 0 \text{ or } \sum_{j \neq t} \frac{\partial F_j}{\partial x_i}(x) > 0 \\ \frac{\partial F_t}{\partial x_i}(x) \left| \sum_{j \neq t} \frac{\partial F_j}{\partial x_i}(x) \right|, & \text{otherwise} \end{cases}$$

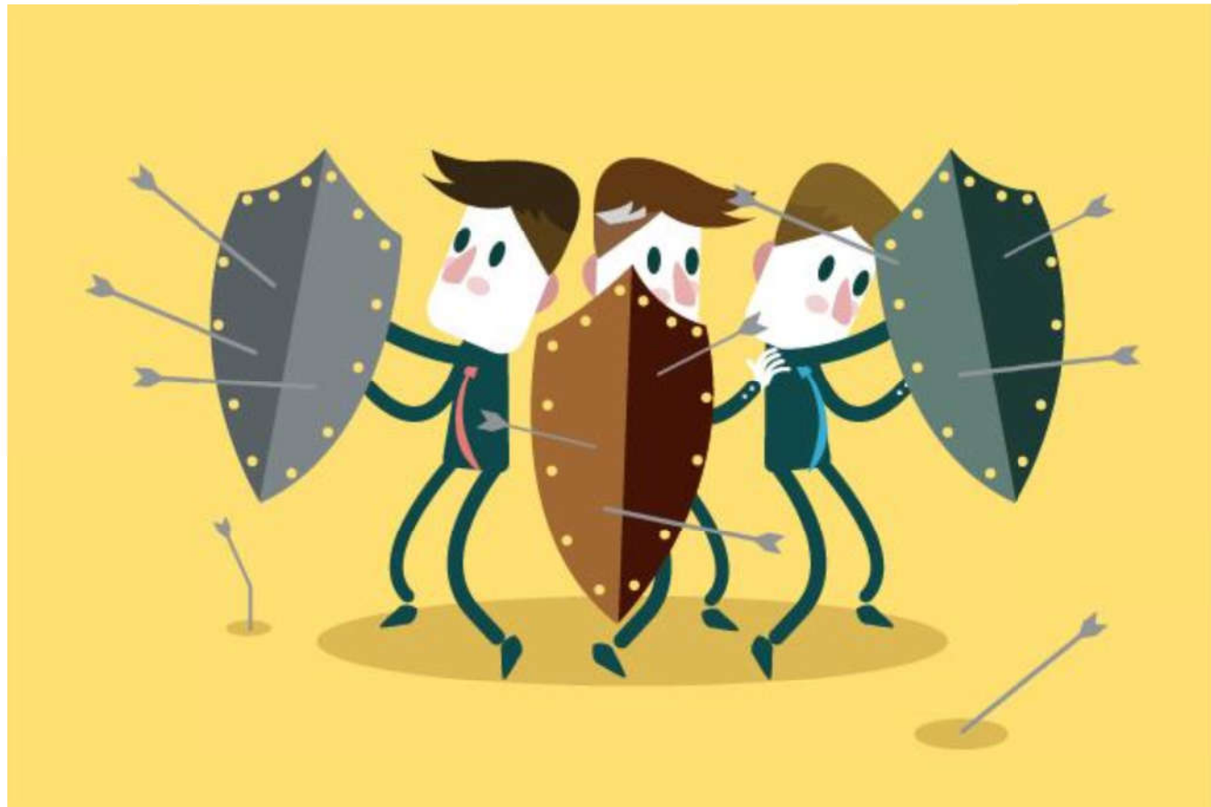
http://fna.ir/a5g

Crafting adversarial examples:



http://fna.ir/a5g

Defenses



http://www.rmmagazine.com/2016/06/01/the-3-lines-of-defense-for-good-risk-management/

Advances in defense strategies

❑Defense is hard!

*A theoretical model of the adversarial example crafting process is very **difficult** to construct.*

- Non-linearity
- non-convex
- Complex optimization process
- …

❑Most of the current defense strategies are
- not adaptive to **all types** of adversarial attack

❑Implementation of such defense strategies
- may incur **performance overhead**

Failed defenses

- Generative pretraining
- Removing perturbation with an autoencoder
- Adding noise at test time
- Ensembles
- Confidence-reducing perturbation at test time
- Dropout
- Adding noise at train time
- Various non-linear units
- …

http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture16.pdf

$$\text{loss}(x, y)$$

Small when prediction is correct on legitimate input

$$\text{loss}(x, y) + \text{loss}(x + \epsilon \cdot \texttt{sign}(\text{grad}), y)$$

Small when prediction is correct on legitimate input

Small when prediction is correct on adversarial input

Advances in defense strategies

- The existing defense mechanisms can be categorized among the following types based on their methods of implementation:
  - **Adversarial Training**
  - **Gradient Hiding**
  - **Defensive Distillation**
  - **Feature Squeezing**
  - **Blocking the Transferability**
  - **Defense-GAN**
  - **Mag Net**
  - **Using High-Level Representation Guided De-noiser**

List of toolboxes

ToolBox	Base Lib.	Usability	Updating
Clever-Hans	TensorFlow, Pytorch	Semi	Well
Fool-Box	TensorFlow, Keras, Pytorch, MXnet	Easy	Well
IBM **ART**	python	Semi	Well
...			

https://github.com/tensorflow/cleverhans

http://www.cleverhans.io/

CleverHans Documentation

This documentation is auto-generated from the docstrings of modules of the current *master* branch of tensorflow/cleverhans.

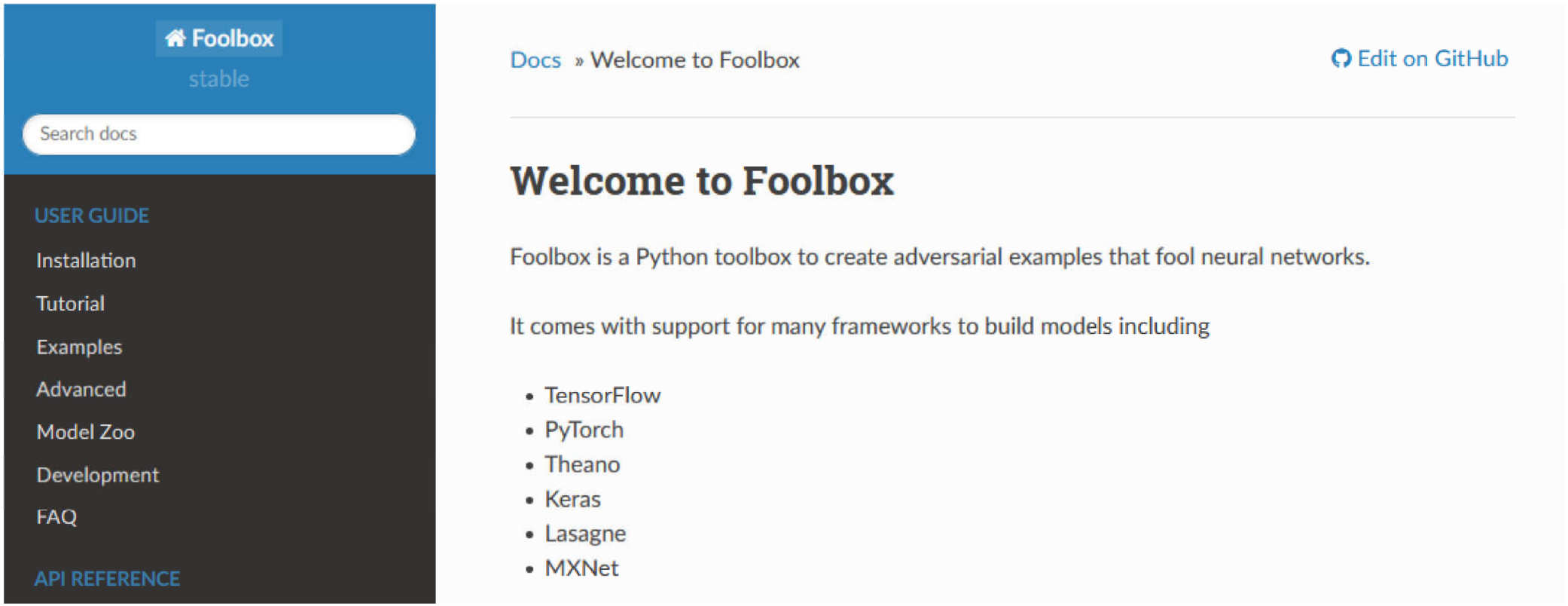
To get started, we recommend reading the github readme. Afterwards, you can learn more by looking at the following modules:

- *attacks* module
- *model* module

Indices and tables

- Index
- Module Index
- Search Page

https://cleverhans.readthedocs.io/en/latest/


Foolbox
stable
Search docs
USER GUIDE
Installation
Tutorial
Examples
Advanced
Model Zoo
Development
FAQ
API REFERENCE
Docs » Welcome to Foolbox
Edit on GitHub
Welcome to Foolbox
Foolbox is a Python toolbox to create adversarial examples that fool neural networks.
It comes with support for many frameworks to build models including
TensorFlow
PyTorch
Theano
Keras
Lasagne
MXNet

API Reference

- `foolbox.models`
  - Models
  - Wrappers
  - Detailed description
- `foolbox.criteria`
  - Criteria
  - Examples
  - Detailed description
- `foolbox.zoo`
  - Get Model
  - Fetch Weights
- `foolbox.distances`
  - Distances
  - Aliases
  - Base class
  - Detailed description
- `foolbox.attacks`
- `foolbox.adversarial`
- `foolbox.utils`

Models

`Model`	Base class to provide attacks with a unified interface to models.
`DifferentiableModel`	Base class for differentiable models that provide gradients.
`TensorFlowModel`	Creates a `Model` instance from existing *TensorFlow* tensors.
`TensorFlowEagerModel`	Creates a `Model` instance from a *TensorFlow* model using eager execution.
`PyTorchModel`	Creates a `Model` instance from a *PyTorch* module.
`KerasModel`	Creates a `Model` instance from a *Keras* model.

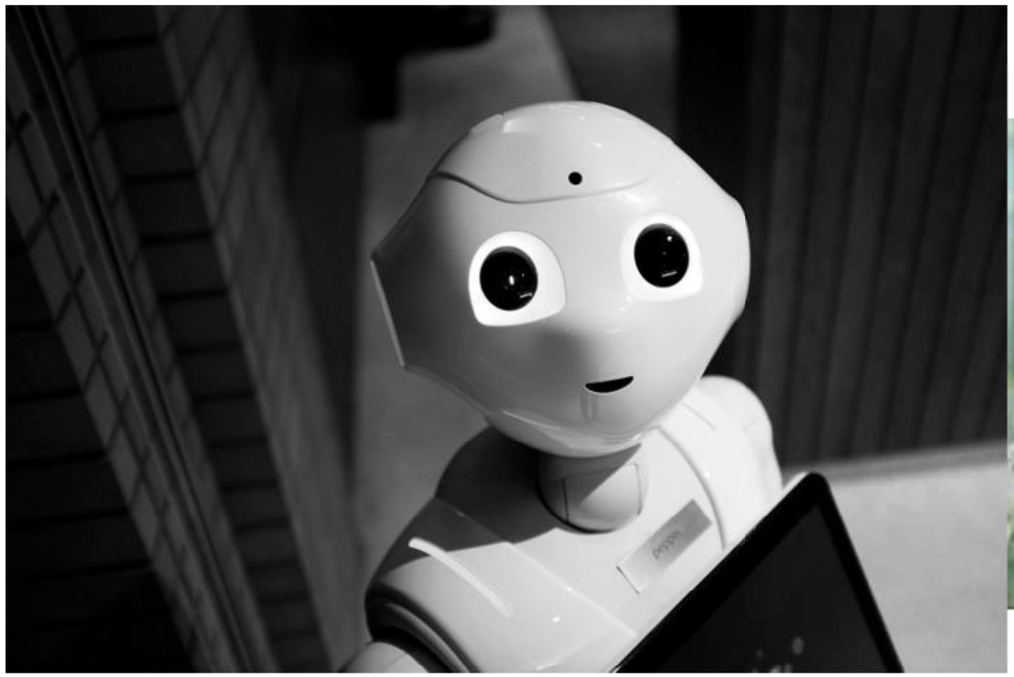
Criteria

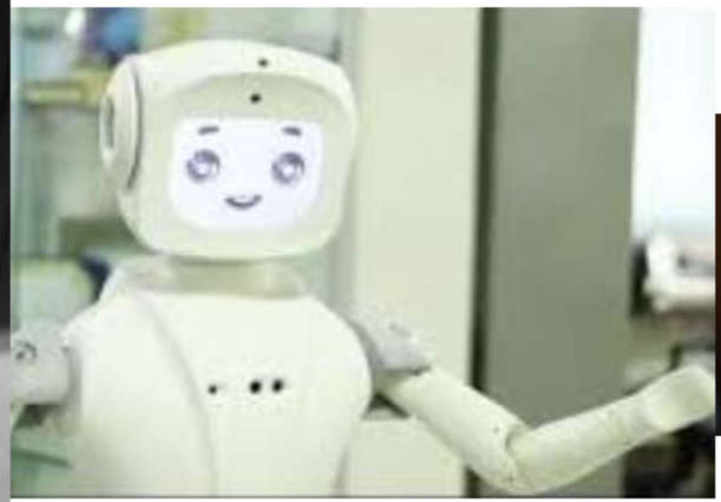
We provide criteria for untargeted and targeted adversarial attacks.

`Misclassification`	Defines adversarials as images for which the predicted class is not the or
`TopKMisclassification`	Defines adversarials as images for which the original class is not one of t
`OriginalClassProbability`	Defines adversarials as images for which the probability of the original cl
`ConfidentMisclassification`	Defines adversarials as images for which the probability of any class othe

`TargetClass`	Defines adversarials as images for which the predicted class is the given targ
`TargetClassProbability`	Defines adversarials as images for which the probability of a given target cla

Tips to stay safe





A summary ☺

- Attacking is easy
- Defending is difficult
- Adversarial training provides regularization and semi-supervised learning
- The out-of-domain input problem is a bottleneck for model-based optimization generally
- There exist certain countermeasure, but none of them can act as a panacea for all challenges!

Mohammad Khalooei

Mkhalooei [at] gmail.com

Khalooei [at] aut.ac.ir

https://ceit.aut.ac.ir/~khalooei