Protect against tomorrow's threats | Machine Learning

Adversarial Machine Learning And Several Countermeasures

Trend Micro
ch0upi
miaoski

7 Dec 2017

ch0upi

Protect against tomorrow's threats | Machine Learning

- Staff engineer in Trend Micro
- Machine Learning + Data Analysis
- Threat intelligence services
- NIPS
- KDDCup 2014 + KDDCup 2016: Top10
- GoTrend: 6th in UEC Cup 2015

miaoski

- Senior threat researcher in Trend Micro
- Threat intelligence
- Smart City
- SDR
- Arduino + RPi makers
- 貓奴

Outline

- Cheating machine learning?
- Attacking theories and practices
- Countermeasures
- Conclusion

CHEAT
MACHINE LEARNING MODELS

We Were Good Guys ...

TREND MICRO

Protect against tomorrow's threats | Machine Learning

基於機器學習的惡意軟體分類實作: Microsoft Malware Classification Challenge 經驗談

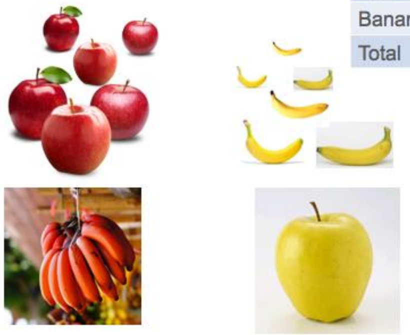
Trend Micro
ch0upi
miaoski
Kyle Chung
2 Dec 2016

Evaluation of Fruit

Protect against tomorrow's threats | Machine Learning

- Accuracy: (9+9)/20 = 90%

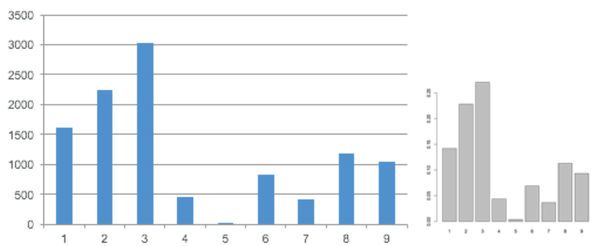
	Apple	Banana
Apple	9	1
Banana	1	9
Total	10	10



20

TLSH

Protect against tomorrow's threats | Machine Learning

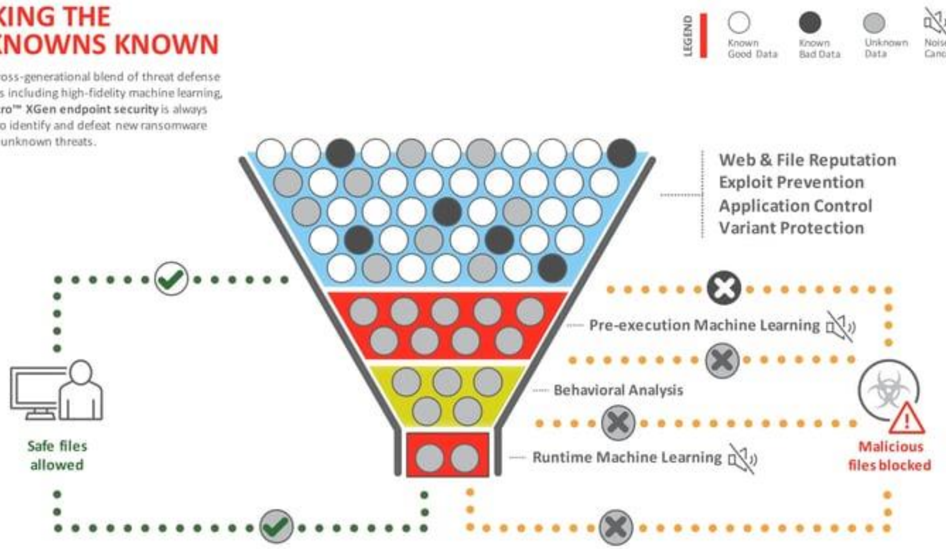



```
Ig2DB5tSiEy1cJvV0zdw,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
ITSUPtCmh7WdJcsYDwQ5,0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0
iwXK2bUysO0CPvBf8nTt,0.0,0.0,0.0,0.0,0.8,0.2,0.0,0.0,0.0
jEAbMPelkWmNgCrGU2QY,0.0,0.0,0.0,0.0,0.4,0.4,0.0,0.0,0.2
Jmo6eIhLZ4t9r8QsxEg5,0.0,0.0,0.0,0.0,0.8,0.0,0.2,0.0,0.0
JtPFl4ewgdD78OzCMa3o,0.0,0.2,0.0,0.0,0.8,0.0,0.0,0.0,0.0
jxrmMI8yPStoDdgE7Y4J,0.0,0.0,0.0,0.0,0.4,0.4,0.0,0.2,0.0
jZGQELvdhm2H6rOJTXun,0.0,0.0,0.6,0.0,0.4,0.0,0.0,0.0,0.0
k35N9Ff2T14v7URulmz6,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0
k3OSYcwRsvCqeo7dTWQx,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0
```

54

MAKING THE UNKNOWNS KNOWN

With its cross-generational blend of threat defense techniques including high-fidelity machine learning, **Trend Micro™ XGen endpoint security** is always adapting to identify and defeat new ransomware and other unknown threats.



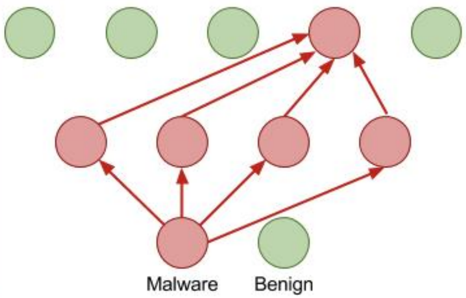
Even NVIDIA...

Protect against tomorrow's threats | Machine Learning

NVIDIA ACCELERATED COMPUTING Downloads Training Ecosystem

PARALLEL FORALL Features Pro Tips Spotlights CUDACasts

← Previous Next →



Malware Detection in Executables Using Neural Networks

Share:

Posted on November 21, 2017 by Jon Barker | 1 Comment

Tagged Deep Learning, Malware Detection

The detection of malicious software (malware) is an increasingly important cyber security problem for all of society. Single incidences of malware can cause millions of dollars in damage. The current generation of anti-virus and malware detection products typically use a signature-based approach

ML-Based Anti-Virus?

zerosum0x0 @zerosum0x0 · 8月10日

It's a DEBUG build too...

```c
#include <stdio.h>

int main()
{
    printf("Hello world!\n");
    return 0;
}
```

SHA256:	c99caff6b05d6d13629c7eb7d014862da7e2774866b61e7bfca47f53578dca0c
File name:	helloworld.exe
Detection ratio:	7 / 64
Analysis date:	2017-08-10 14:07:06 UTC (0 minutes ago)

Analysis · File detail · Additional information · Comments · Votes

Antivirus	Result
CrowdStrike Falcon (ML)	malicious_confidence_80% (D)
Cylance	Unsafe
Cyren	W32/S-d2b5872a!Eldorado
F-Prot	W32/S-d2b5872a!Eldorado
Sophos ML	heuristic

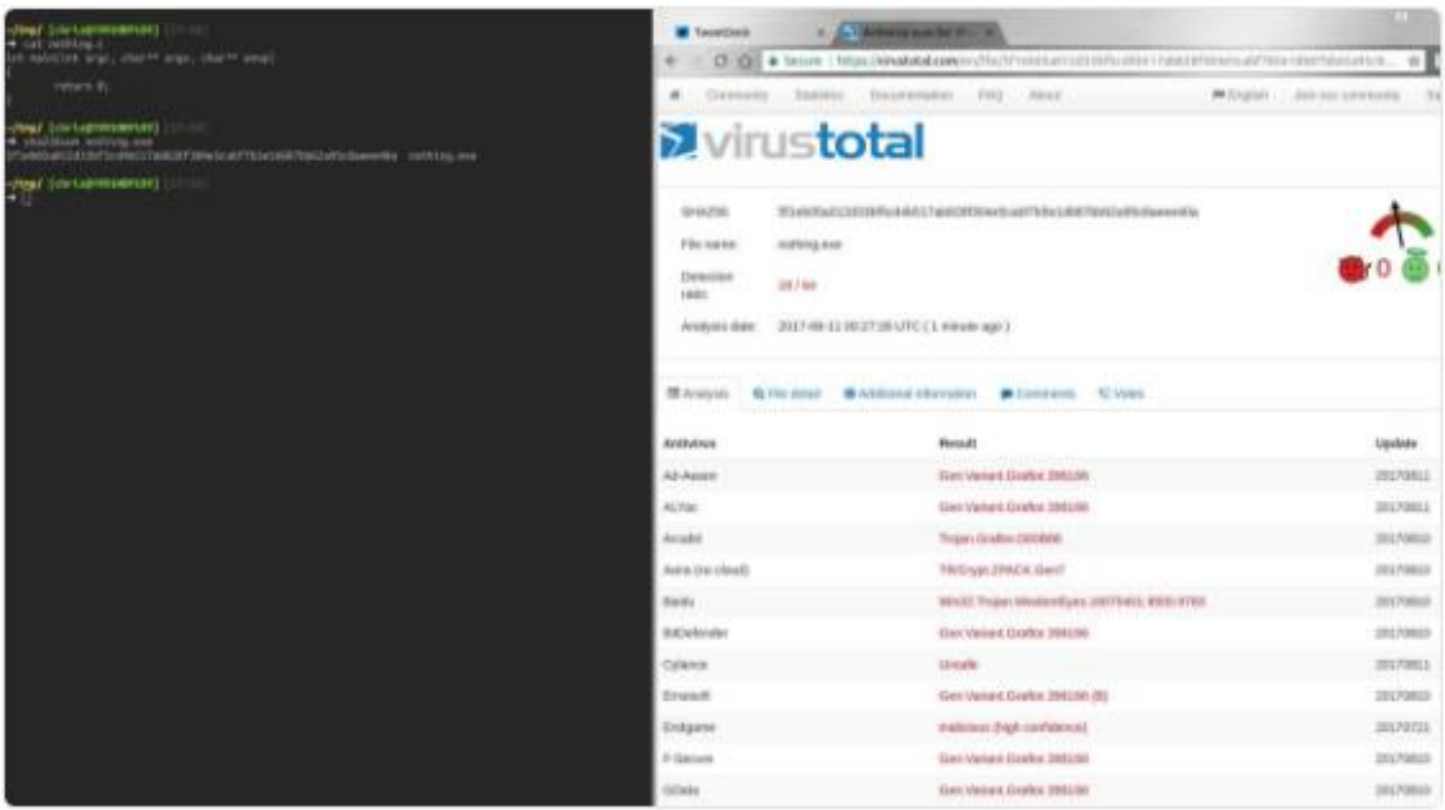
ML-Based Anti-Virus?

crazy rabbidz
@_hugsy_

フォローする

返信先: @rantybenさん、@Vissさん、@zerosum0x0さん

This is insane! You don't even need printf()! What's wrong with the world?



CSOs Explained

Protect against tomorrow's threats

Machine Learning

CSO FROM IDG

SALTED HASH- TOP SECURITY NEWS

By Steve Ragan, Senior Staff Writer, CSO | AUG 18, 2017 4:00 AM PT

About

Fundamental security insight to help you minimize risk and protect your organization

NEWS

Here's why the scanners on VirusTotal flagged Hello World as harmful

CrowdStrike, Cylance, Endgame and others flagged Hello World as unsafe or malicious

MORE LIKE THIS

Using AI to spot malware patterns

Cylance blamed for DirectDefense's 'botnet' disclosure

Vendors respond to Cylance's new testing

But Still ...

Protect against tomorrow's threats | Machine Learning

18 engines detected this file

EXE

18 / 66

SHA-256	aca55bce947a49f5073b9e860789f0f2b3cb147972e18178143ffaf6790160c4
File name	6d130077084e0b1f4542b08f92736df0.virobj
File size	99.69 KB
Last analysis	2017-10-30 16:57:17 UTC

Detection | Details | Behavior | Community

AegisLab	Virus.W32.Evo.Gen!c	Avast	Win32:Evo-gen [Susp]
AVG	Win32:Evo-gen [Susp]	Avira	TR/Crypt.ZPACK.Gen7
CrowdStrike Falcon	malicious_confidence_80% (W)	Cylance	Unsafe
Endgame	malicious (moderate confidence)	Jiangmin	Backdoor.Generic.zrm
McAfee	Artemis!6D130077084E	McAfee-GW-Edition	Artemis
nProtect	Trojan/W32.Agent.102082	Palo Alto Networks	generic.ml
Qihoo-360	Win32/Trojan.af4	SentinelOne	static engine - malicious
Sophos ML	heuristic	Symantec	Trojan.Gen.2

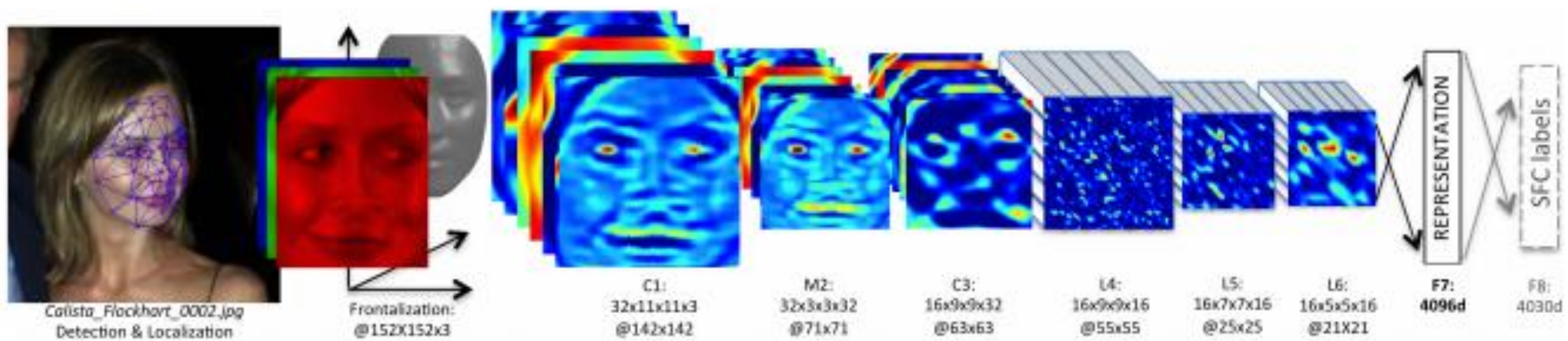
Rescan Makes It Worse

Protect against tomorrow's threats

Machine Learning

Compiler	Hello World (no debug)	Hello World (debug)	Nothing (no debug)	Nothing (debug)
Visual Studio 2017	Cylance, Jiangmin	Cylance, Cyren, F-Prot, Sophos ML, SentinelOne Static ML	Cylance, Jiangmin	Cylance, Cyren, F-Prot, Sophos ML, SentinelOne Static ML
MingW64	Good	Good	Good	Good
Cygwin x86_64	Baidu, Cylance	Baidu	Baidu, Cylance	Baidu

ML is Prosperous



Taigman et al. (2014) DeepFace: Closing the Gap to Human-Level Performance in Face Verification

ML Drives

https://www.tesla.com/sites/default/files/images/videos/tesla_autopilot_2_video.jpg

Machine learning has its **particular** vulnerabilities.

Research Prediction Competition

NIPS 2017: Targeted Adversarial Attack

Develop an adversarial attack that causes image classifiers to predict a specific target class

Google Brain · 65 teams · a month ago

Research Prediction Competition

NIPS 2017: Non-targeted Adversarial Attack

Imperceptibly transform images in ways that fool classification models

Google Brain · 91 teams · a month ago

Research Prediction Competition

NIPS 2017: Defense Against Adversarial Attack

Create an image classifier that is robust to adversarial attacks

Google Brain · 107 teams · a month ago

Protect against tomorrow's threats

Machine Learning

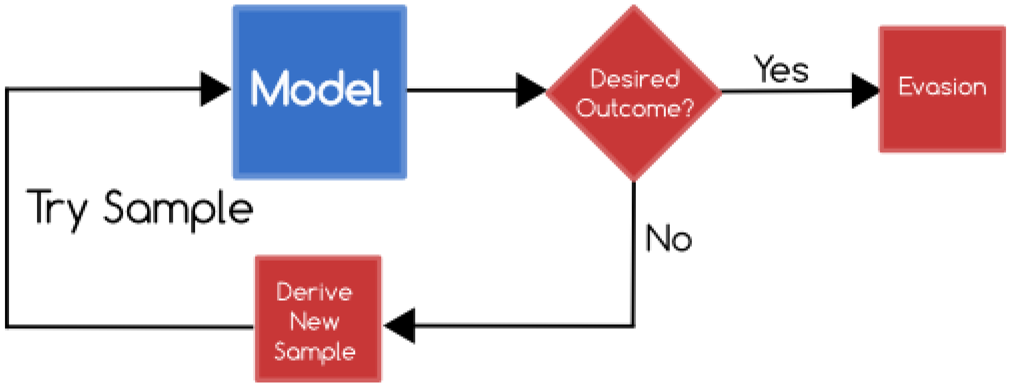
THEORIES AND PRACTICES

Methodology

- Evasion
  - Black box
  - White box
- Model stealing
- Poisoning

Methodology

- Evasion
  - **Black box**
    - **Random**
    - **Evolutionary algorithms (GA)**
  - White box
- Model stealing
- Poisoning



Black Box

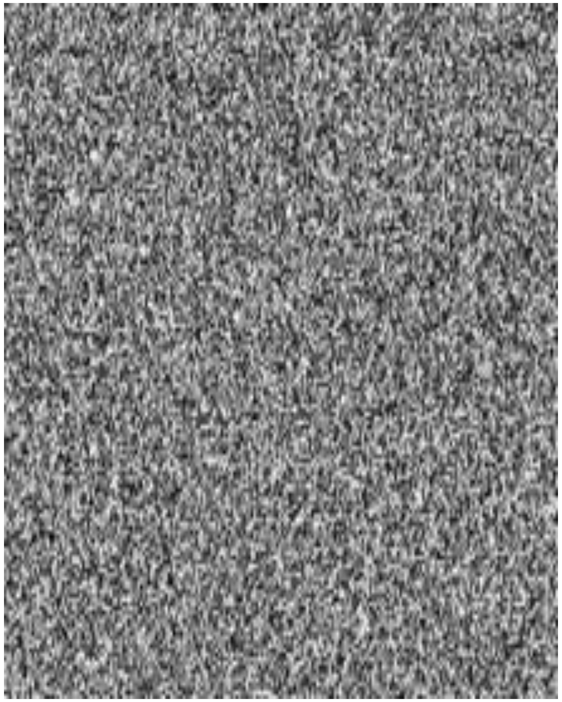
- No model
- Only predict interface & result

Black Box: Random Noise Attack

- Add some white noise?

random.normalvariate(0, 5)

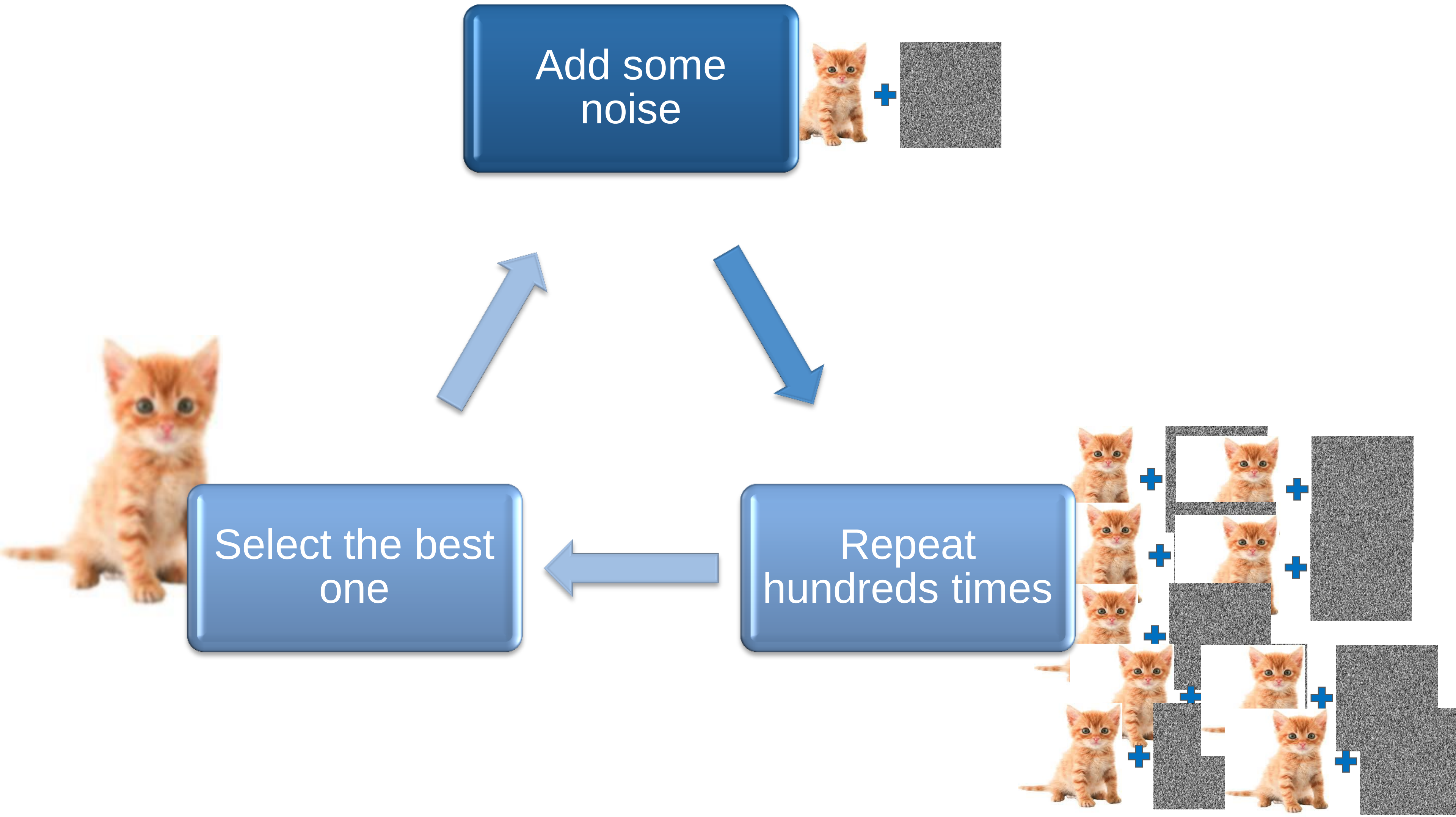
+



= ?

Not effective for most model

Black Box: Iterative Random Attack

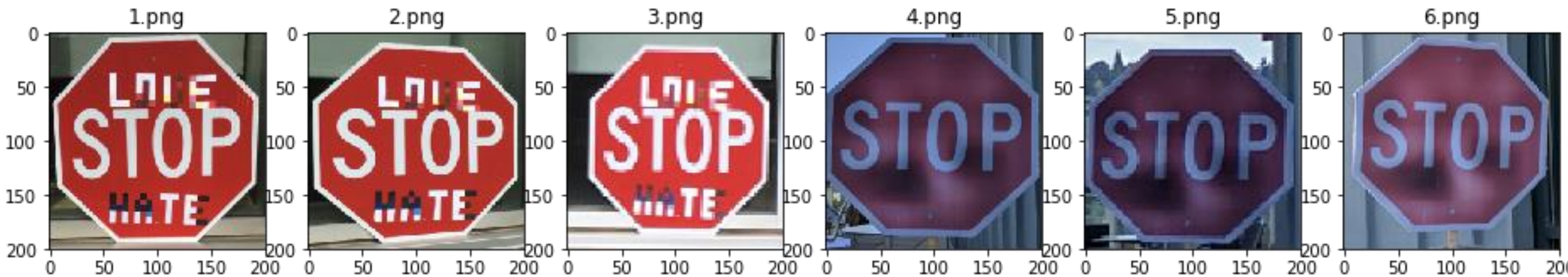


Black Box – Random – STOP

- Inspired by Evtimov et al. (2017)
- We use iterative random attack instead
- Difficult: STOP sign ➔ something else

Black Box – Random – STOP

- Evtimov et al. (2017) ➔ 80 KM/h



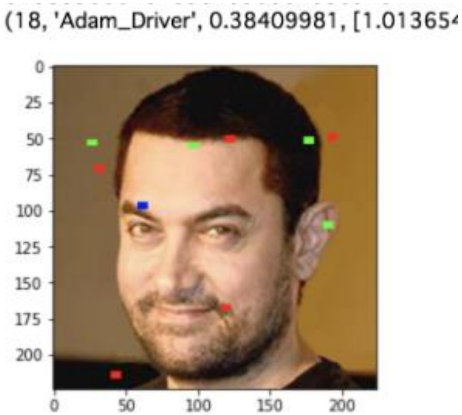
Hacked in iteration 5
Predicted Labels: 39 ['Keep left']
(confidence = 73%)
39 - Keep left
14 - Stop
13 - Yield
6 - End of speed limit (80km/h)
41 - End of no passing

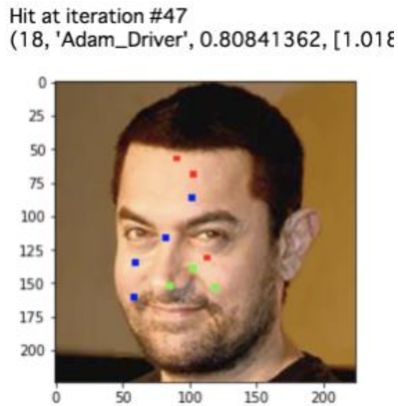
Black Box – Random – Faces

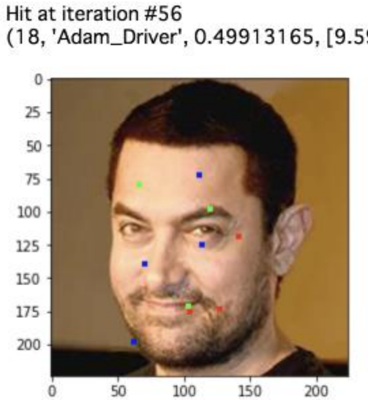
- VGG Face and @mzaradzki

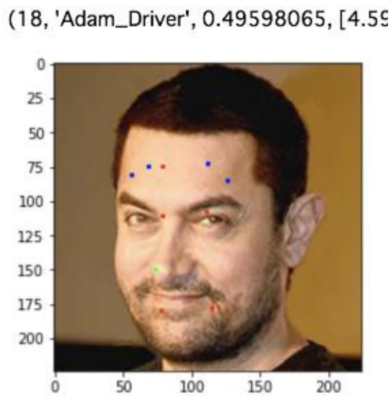
N	Square Size	Success?
10	4x4	Adam Driver
10	4x3	Adam Driver
10	3x3	Adam Driver
10	2x2	Adam Driver (difficult)
--	Cat face	Failed
10	1x1	Failed*

Black Box – Random – Faces









Adam Driver

Aamir Khan

Black Box – Genetic Algorithm

- Effective random search

- Inspired by the process of natural selection
- Belongs to evolutionary algorithms (EA)
- Solving optimization and search problems

Black Box – Genetic Algorithm

- Selection
- Crossover
- Mutation
- Evaluation

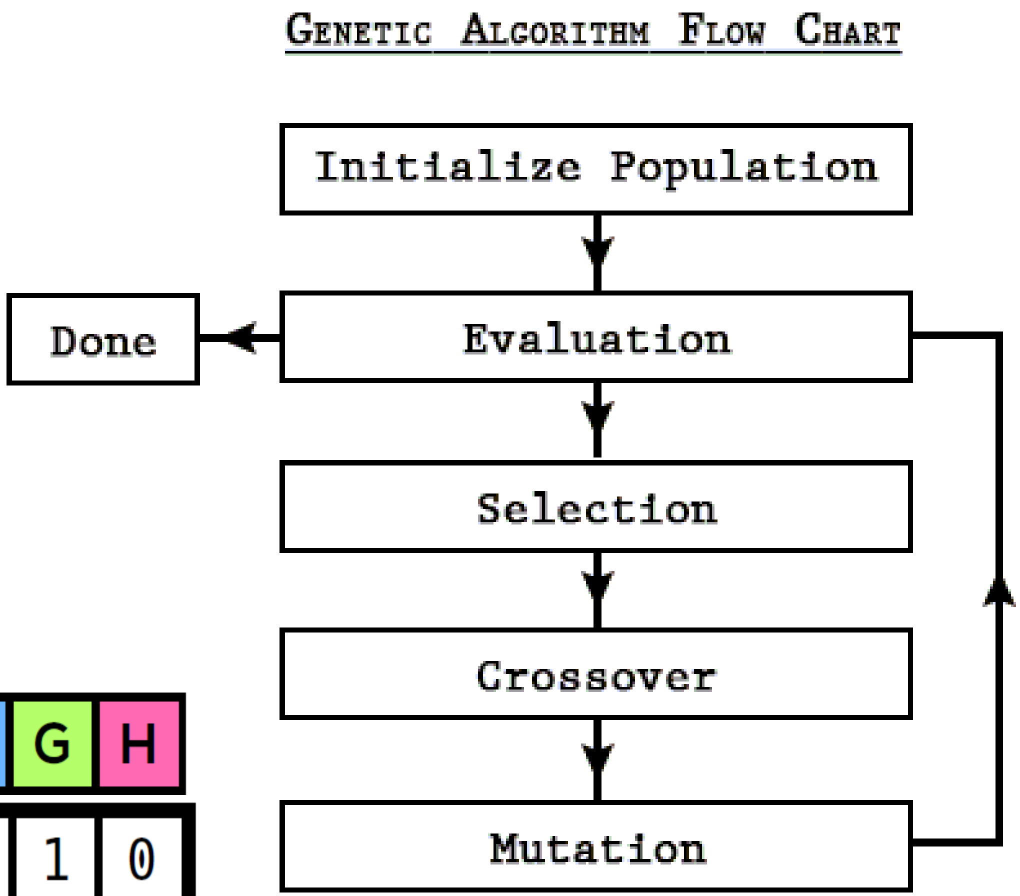
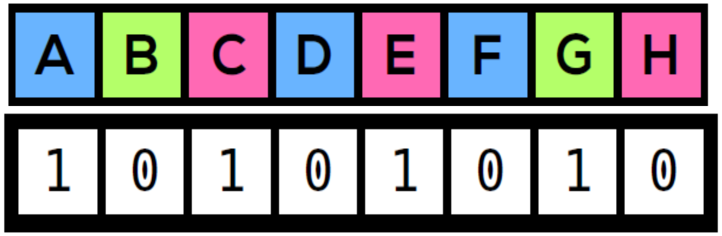


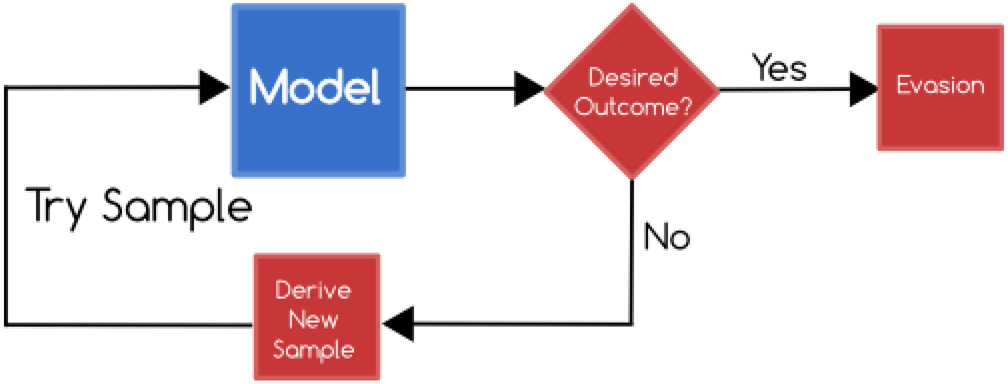
Figure 2



Methodology

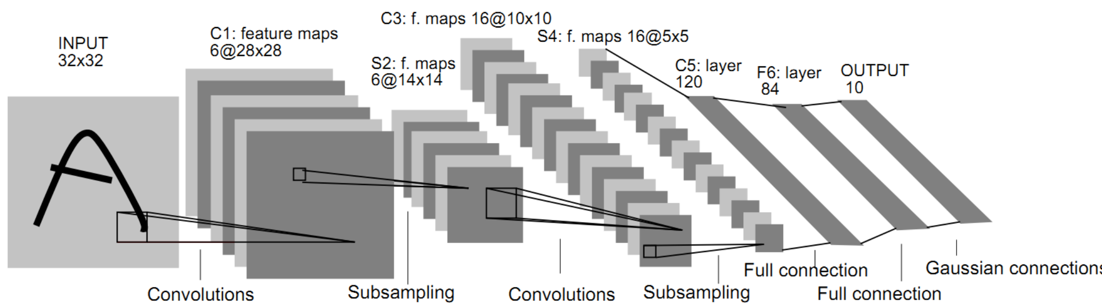
Protect against tomorrow's threats
Machine Learning

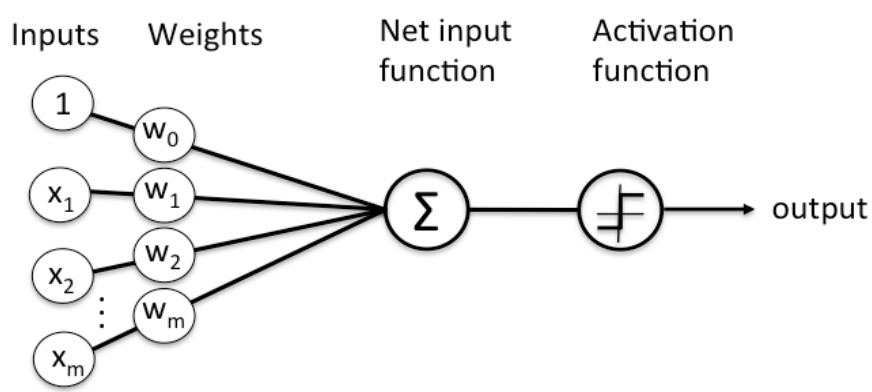
- Evasion
  - Black box
  - **White box**
    - **FGSM**
    - **One-step target class**
- Model stealing
- Poisoning



White Box

- With all model detail
- DNN architecture, weights





Fast Gradient Sign Method

- simple and computationally efficient
- non-target attack
- Goodfellow et al. (2014)

$$X^{adv} = X + \epsilon sign(\nabla_X J(X, y_{true}))$$

X^{adv}: Adversarial image
X: Original image
ϵ: perturbation level
$\nabla_X J(X, y)$: gradient

Attack a Linear Model

Lets fool a binary linear classifier:

x	2	-1	3	-2	2	2	1	-4	5	1	← input example
w	-1	-1	1	-1	1	-1	1	1	-1	1	← weights
adversarial x	1.5	-1.5	3.5	-2.5	2.5	1.5	1.5	-3.5	4.5	1.5	

class 1 score before:

-2 + 1 + 3 + 2 + 2 - 2 + 1 - 4 - 5 + 1 = -3

=> probability of class 1 is 1/(1+e^(-(-3))) = 0.0474

$$P(y=1 \mid x; w, b) = \frac{1}{1+e^{-(w^T x + b)}} = \sigma(w^T x + b)$$

-1.5+1.5+3.5+2.5+2.5-1.5+1.5-3.5-4.5+1.5 = 2

=> probability of class 1 is now 1/(1+e^(-(2))) = 0.88

i.e. we improved the class 1 probability from 5% to 88%

Fei-Fei Li, Andrej Karpathy, Justin Johnson, Lecture 9-72, 2016

White Box Attack Methods

- Fast gradient sign method (non-target, one step)

$$X^{adv} = X + \epsilon \, \mathrm{sign}\big(\nabla_X J(X, y_{true})\big)$$

- One-step target class methods (target, one step)

$$X^{adv} = X - \epsilon \, \mathrm{sign}\big(\nabla_X J(X, y_{target})\big)$$

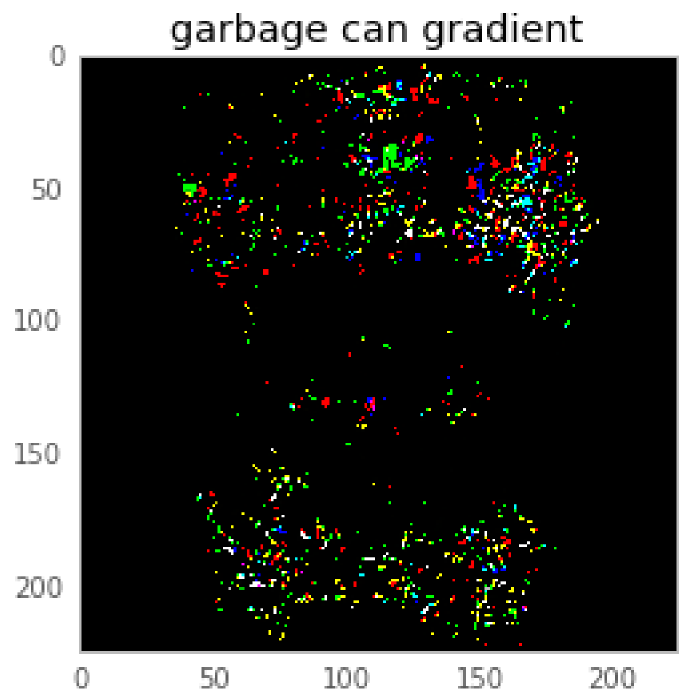
- Basic iterative method (non-target, multiple steps)

$$X_0^{adv} = X, \quad X_{N+1}^{adv} = Clip_{X,\epsilon}\Big\{X_N^{adv} + \alpha \, \mathrm{sign}\big(\nabla_X J(X_N^{adv}, y_{true})\big)\Big\}$$

- Iterative least-likely class method (target , multiple steps) $X_0^{adv} = X, \quad X_{N+1}^{adv} = Clip_{X,\epsilon}\left\{X_N^{adv} - \alpha \, \mathrm{sign}\left(\nabla_X J(X_N^{adv}, y_{LL})\right)\right\}$

Kurakin et al., ADVERSARIAL MACHINE LEARNING AT SCALE. ICLR 2017

White Box – FGSM – Trash Can

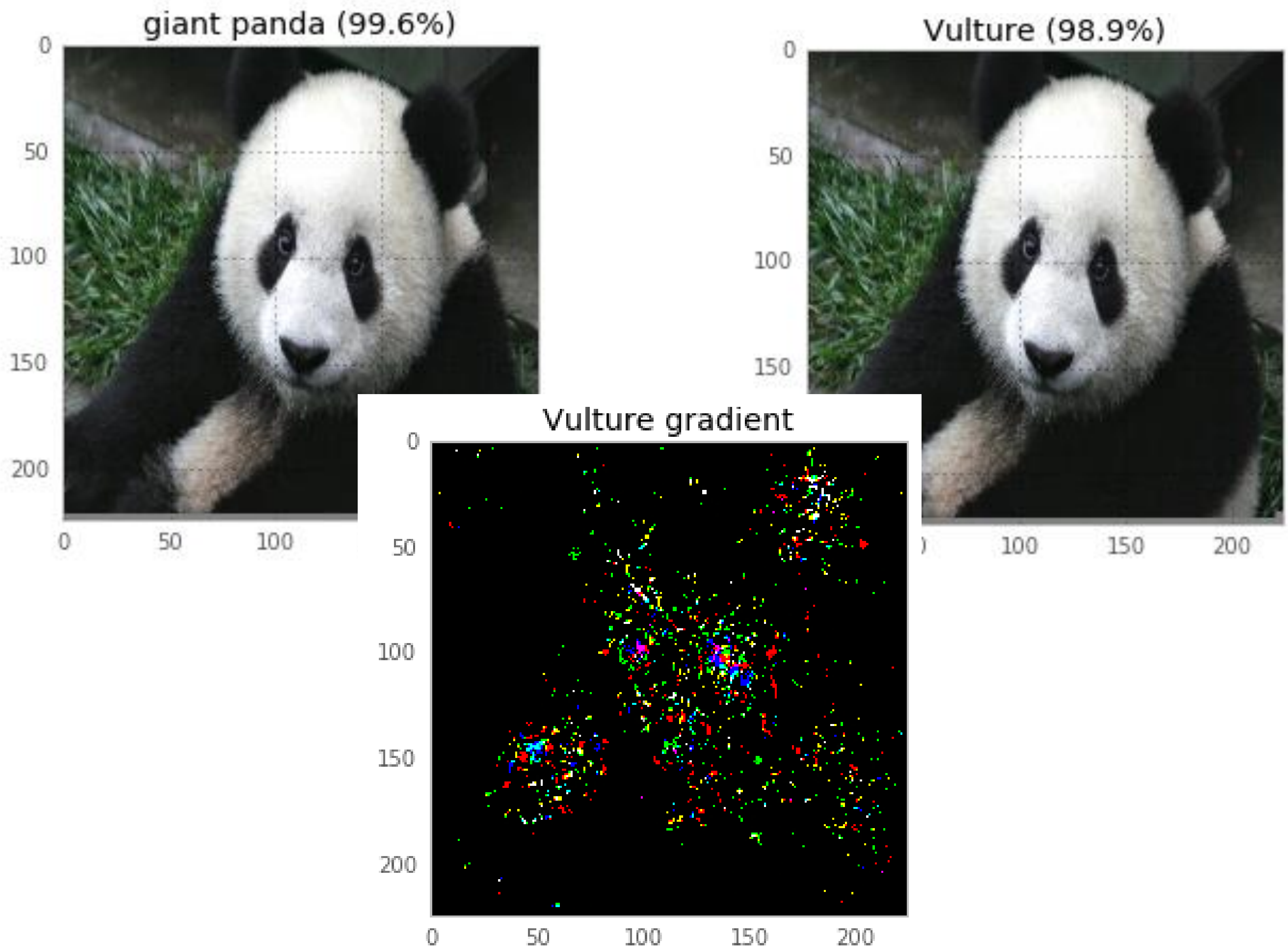


label: 412 (ashcan, trash can), certainty: **37.47%**
label: 899 (water jug), certainty: 10.85%
label: 503 (cocktail shaker), certainty: 7.98%

label: 412 (ashcan, trash can), certainty: **87.68%**
label: 463 (bucket, pail), certainty: 3.08%

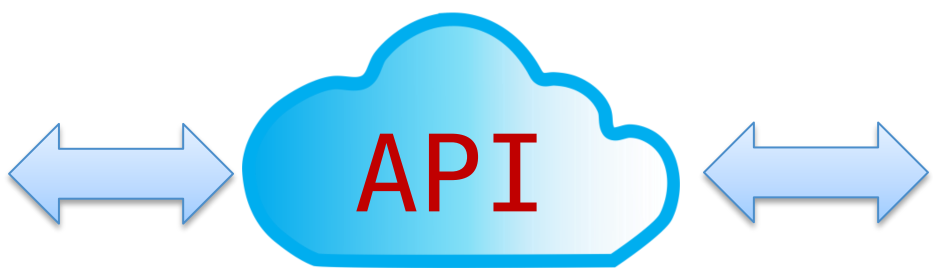
```
_ = predict(garbage_data + 0.75 * np.sign(grad), n_preds=2)
```

White Box – One-Step Target

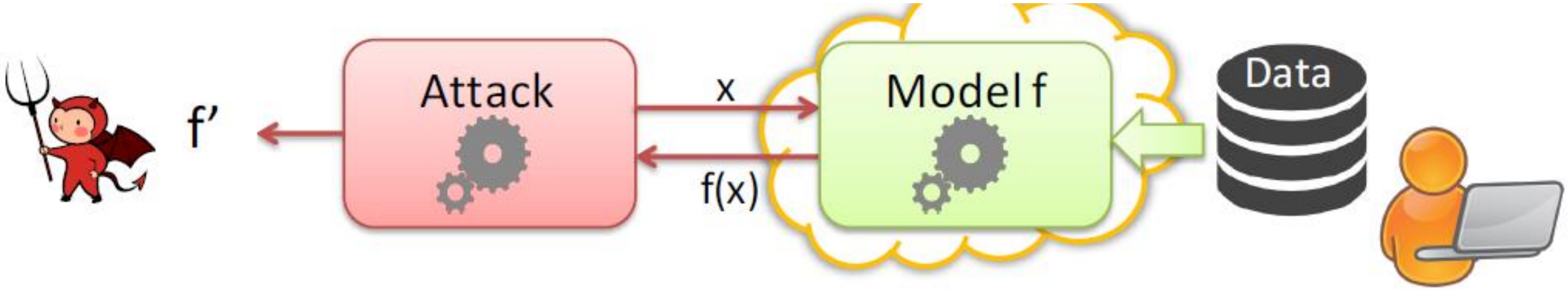


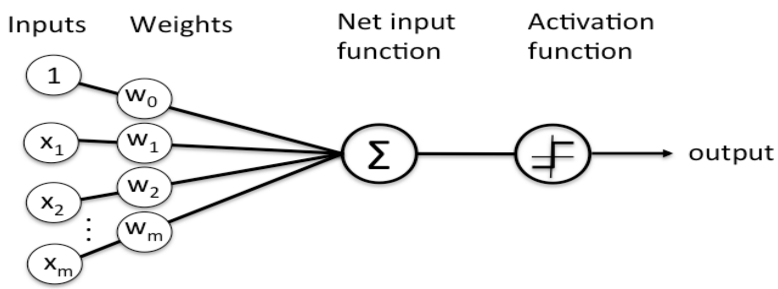
Methodology

- Evasion
  - Black box
  - White box
- **Model stealing**
- Poisoning



Model Stealing





$$f(x) = 1 / (1+e^{-(w^*x + b)})$$

$$\ln\left(\frac{f(x)}{1 - f(x)}\right) = w^*x + b$$

Linear equation in n+1 unknowns w,b

Florian Tramer et. al., Stealing Machine Learning Models via Prediction APIs, Usenix Security Symposium 2016

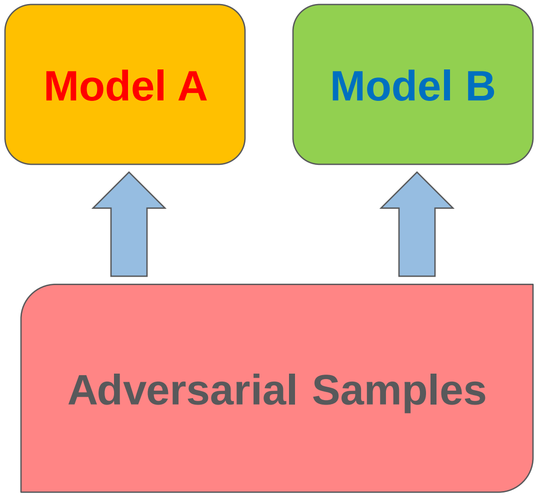
Model Stealing

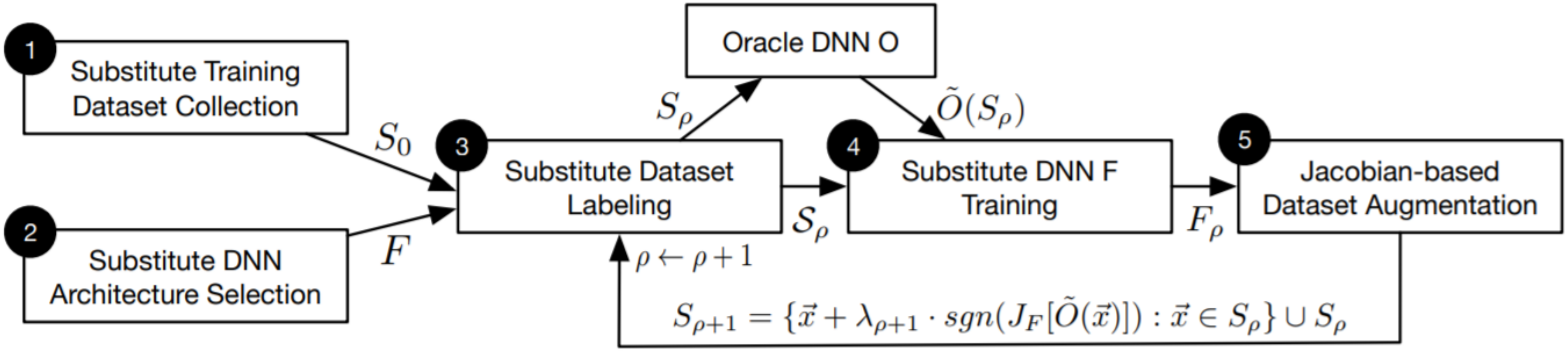
- Model is data
- Model is asset

- Train a local DNN for Black box attack
- Data privacy

Model Stealing: Adversarial Attack

- Transferability Property
- Train a local model for attack
- Effective data augmentation

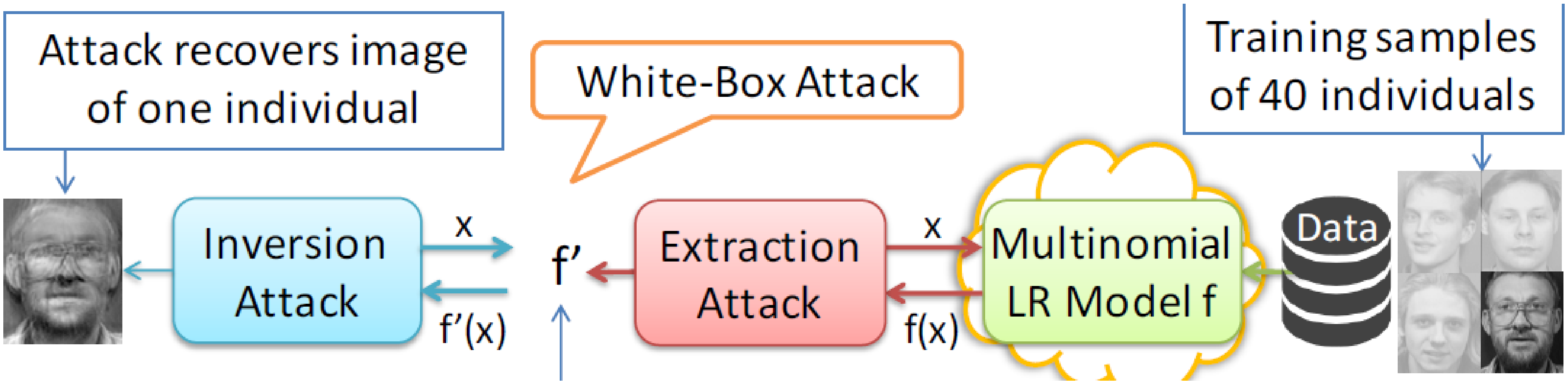




Ian Goodfellow, Practical Black-Box Attacks against Machine Learning, 2017

Model Stealing: Data Privacy

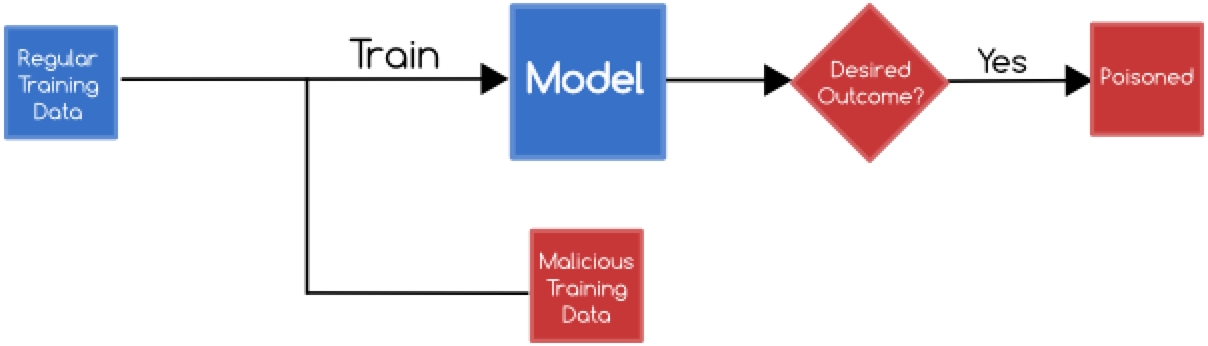
- How to re-build your face if we have the model?



Florian Tramer et. al., Stealing Machine Learning Models via Prediction APIs, Usenix Security Symposium 2016

Methodology

- Evasion
  - Black box
  - White box
- Model stealing
- **Poisoning**



Poison Attack

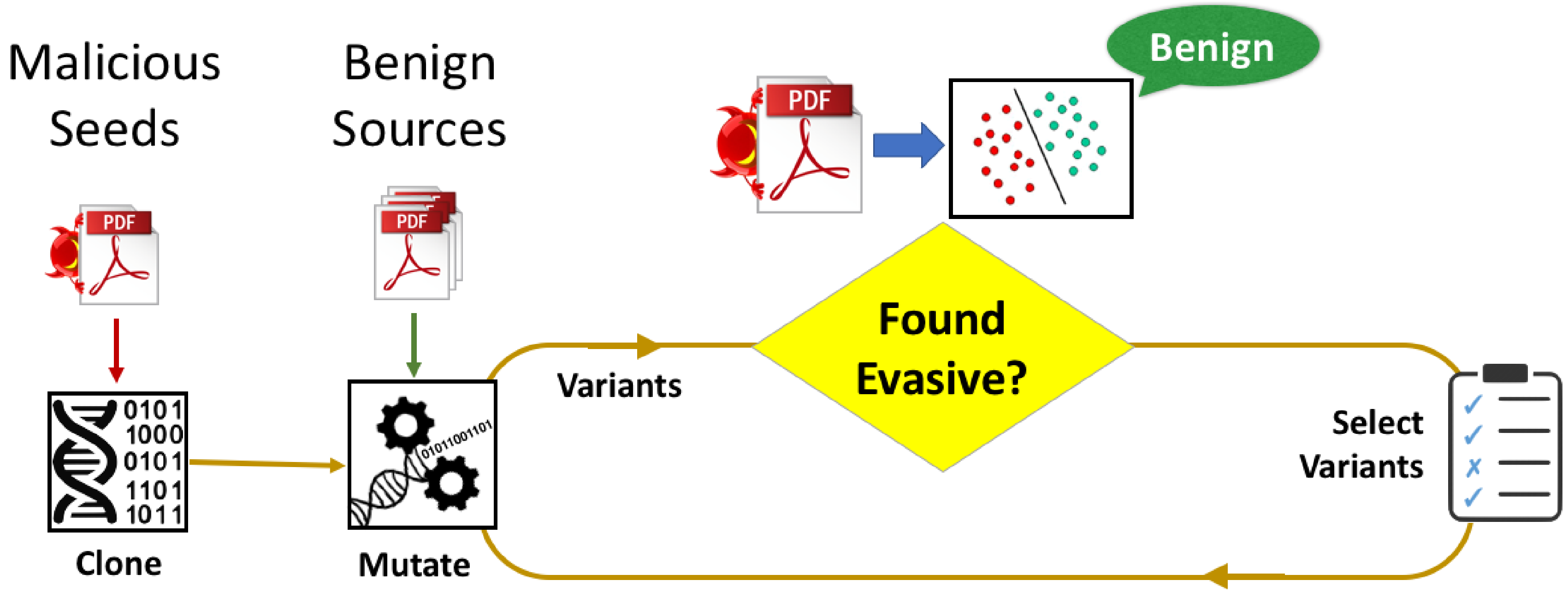
- Crowdsourcing
  - Amazon Mechanical turk
  - Mis-labeling
- Online training
  - Microsoft chatbot: Tay
  - User feedback

Real World Adversarial

- Evading Against PDF ML
- Auto-pilot cars
- Access control w/ face recognition

Evading Against PDF ML

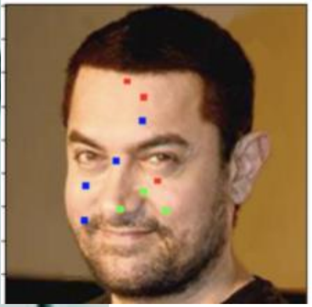
- Genetic algorithm to generate adversarial sample
- Sandbox to ensure malicious behavior kept



Weilin Xu, Yanjun Qi, and David Evans. Automatically Evading Classifiers A Case Study on PDF Malware Classifiers. Network and Distributed Systems Symposium 2016
http://evademl.org/

Auto-pilot Cars

Access Control w/ Face Recognition



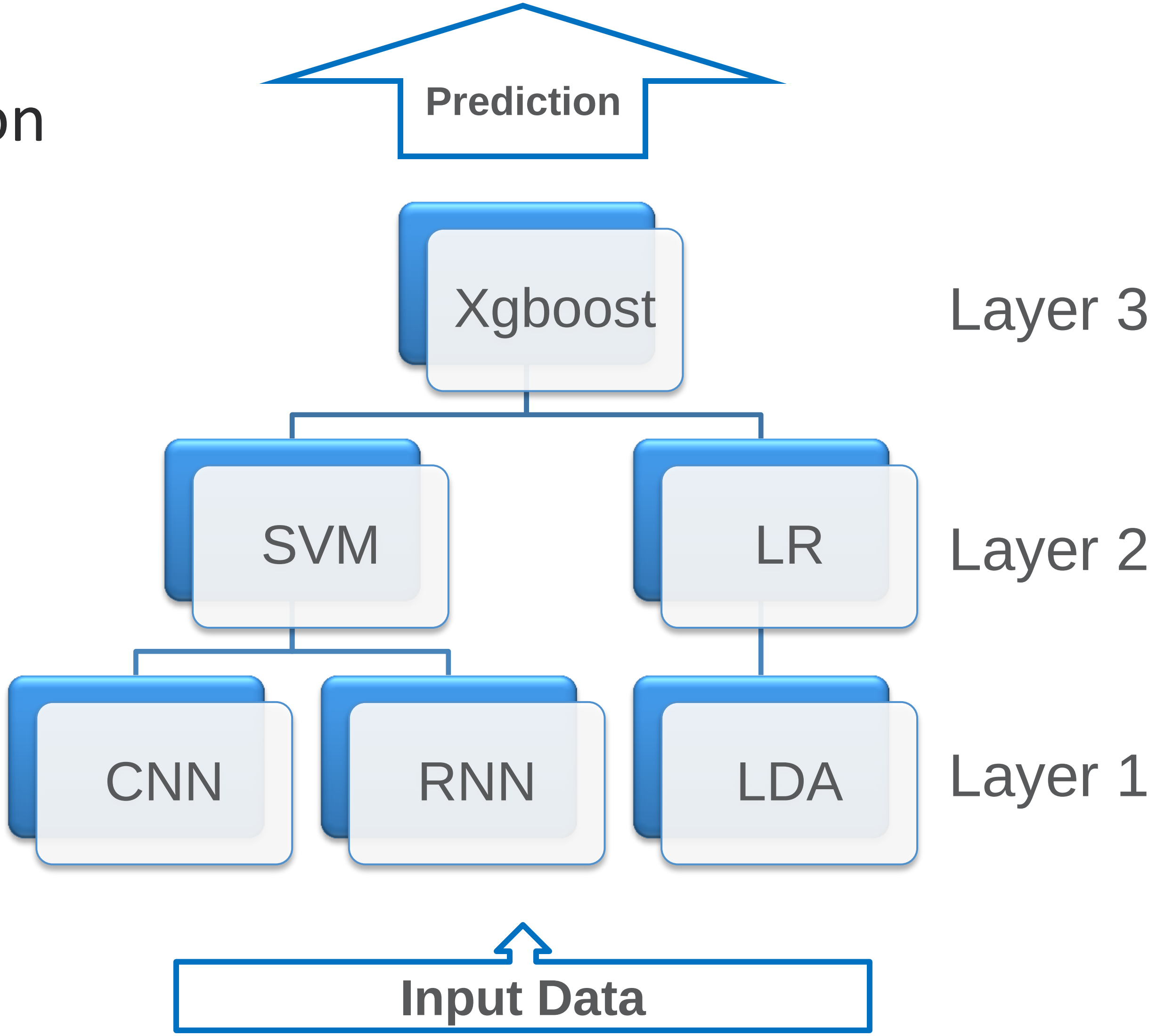
COUNTERMEASURES

Countermeasures

- Ensemble & Stacking
- Retrained model
- Denoiser
- Prevent Model Leakage

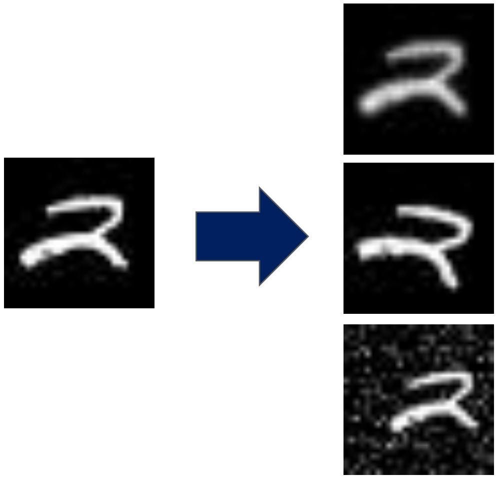
Ensemble & Stacking

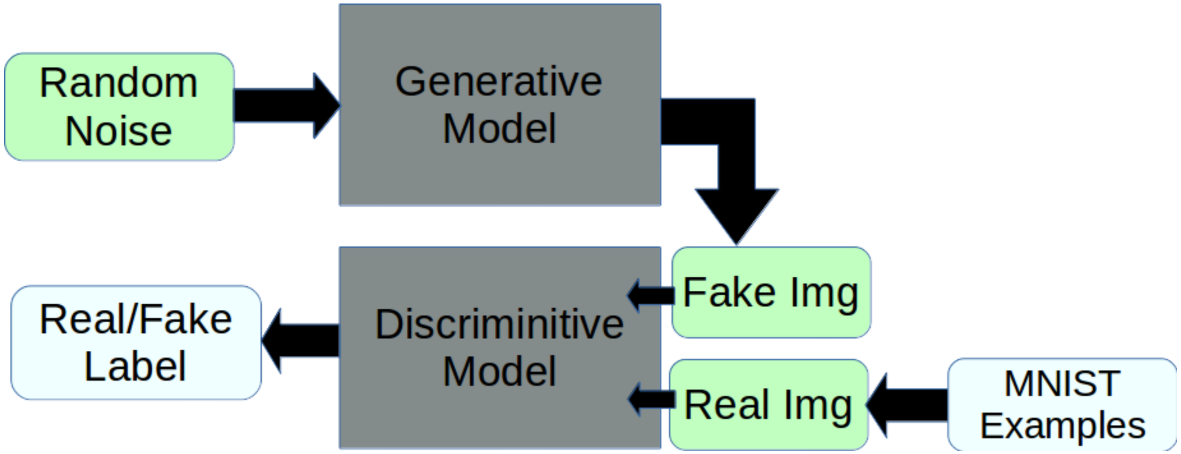
- Layer protection



Retrained Models

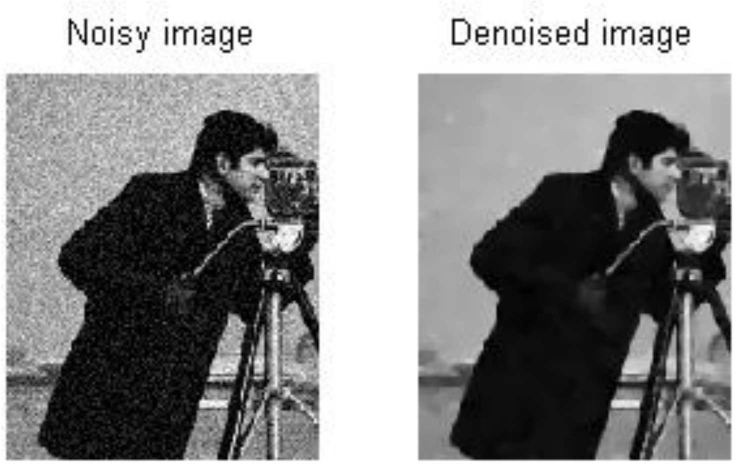
- Distortion
  - Retrain with noisy sample
- Randomization layer in DNN (NIPS 2nd)
- Generative Adversarial Networks (GAN)





Denoiser

- Use denoise technologies from image processing
- Train a DNN denoiser to reduce the noise



Prevent Model Leakage

- Avoid Model stealing
- Increase the challenge of black box attack

- Keep some info secret or add some noise
- Randomization and disinformation
- Adversarial sample detection

CONCLUSION

Conclusion

- Know the limitations and weakness of your model
- Integrate adversarial machine learning into product development cycle
    - Improve ML
    - QA process
- Trend Micro is working on bypassing anti-virus with ML in order to make our product robust

References

- Evtimov et al. (2017) Robust Physical-World Attacks on Deep Learning Models
- https://iotsecurity.eecs.umich.edu/#roadsigns
- Nguyen et al. (2015) Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. IEEE CVPR '15.
- Kurakin A., Goodfellow I.J., Bengio S. (2017) Adversarial Examples in the Physical World.
- https://github.com/tomaszkacmajor/CarND-Traffic-Sign-Classifier-P2
- https://aboveintelligent.com/face-recognition-with-keras-and-opencv-2baf2a83b799
- https://github.com/davidsandberg/facenet
- http://www.vlfeat.org/matconvnet/pretrained/#face-recognition
- https://github.com/mzaradzki/neuralnets/tree/master/vgg_faces_keras

USE THE SOURCE, LUKE!

https://github.com/miaoski/hitcon-2017-adversarial-ml