Open Web Application Security Project

Talal Albacha
Jean-Noël Colin
Prabhant Singh
Sereysethy Touch

Top 5 Machine Learning Risks project

Open Security Summit 2018 – Working Session – 8 June 2018

Session Agenda

- Project Introduction
- ML risks
- Hands-on tutorial (30 minutes)
- Discussion for future directions (1 hour)

OWASP SUMMIT 2017

OUTCOMES TRACKS PARTICIPANTS VENUE FAQ BLOG SPONSORS ABOUT ADMIN

Machine Learning and Security

Track: Research

Organizer(s): Talal Albacha
Participants(s): Adam Obrien , Adrian Winckles , Carlos Serrao (remotely) , Daniela Cruzes , Danny Grander , Jason Li , Jonathon Brookfield , Juan Calderon , Marco Morana , Mateo Martinez , Nuno Loureiro , Peleus Uhley , Sandor Lenart , Stefano Di Paola , Tiago Mendo
Invited: Fabien Thalgott

When: Fri
Time: AM-1
Location: Kings
Remote link: join here

Machine Learning (ML) and Artificial Intelligence (AI) are becoming mainstream techniques, and they provide a great opportunity for defenders.

WHY

We are on the cusp of a Machine Learning and Artificial Intelligence revolution. ML and AI techniques have recently re-emerged as powerful tools in various business sectors such as Fraud Detection, Anomaly Detection, and Behavioral Analysis. Several companies and services are exploring these technologies and use them to solve specific security challenges successfully.

Despite the success of ML and AI, there are security risks associated with them, especially during the learning phase which can be vulnerable to threats originated by potential adversaries, with consequent impact on prediction results.

This Working Session will share common practices; what works today, and what is worth focusing on in the future.

WHAT

- What are the available machine learning platforms?
- Are there any security vulnerabilities associated with these platforms?
- How to securely feed data to ML and AI tools
- How to make learning algorithms aware of malicious data?
- Can AI be used to reduce false positive findings in security scanners?
- How can we spread the message among developers and security communities?

OUTCOMES

- Guidelines for secure usage of machine learning techniques.

WHO

The target audience for this Working Session is:

- Security professionals
- ML and AI researchers
- Devops
- SOC teams

WORKING MATERIALS

OWASP Top 5 Machine Learning Risks

Main | FAQs | Acknowledgements | Road Map and Getting Involved

OWASP Open Web Application Security Project

The OWASP Top 5 Machine Learning Risks

[edit | edit source]

The idea is to build the required resources which help software security community to understand the emerging technology of machine learning and how it is related to security, warn them about the risk associated with using ML, and discuss the defending

Presentation [edit | edit source]

TBD

Project Leader

[edit | edit source]

- Talal Albacha: I have long experience in the application security field and I have strong academic background in machine

Quick Download

[edit | edit source]

TBD

News and Events

[edit | edit source]

-

In Print [edit | edit source]

OWASP SUMMIT 2017

Machine L

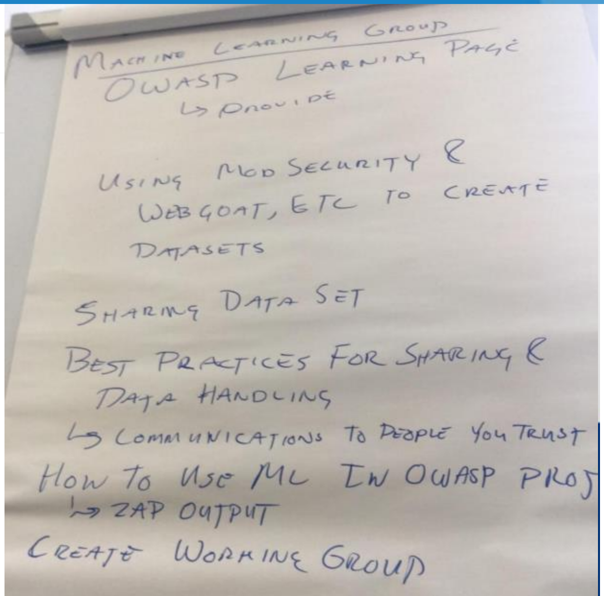
Back to list of all Outcomes

Original Working Session content: Machine Learning and Security

OUTCOMES

Synopsis and Takeaways

- Create common datasets with the purpose of testing and validating the security of machine learning algorithms
  - We can use data output of Mod Security, WebGoat and others to create the datasets
  - These datasets should be shared
  - Anonymized dataset
  - Common dataset for testing
- Create guidance page to include ML security definitions, latest reports, and links to the available tools and datasets
  - Find good materials and resources
- Use ML techniques in the current tools provided by OWASP (e.g., use ML to reduce false positives in ZAP scanning output)
- Create a working group to work on tools and guidance of:
  - How to check if a dataset is noise-free (not compromised)
  - Review of algorithms implementations



18 | Star 4 | Fork 1

Branch: master ▾ Top-5-Machine-Learning-Risks / Top 5 Machine Learning Risks.md | Find file | Copy path

talbacha removed unnecessary graph — f6ba795 a day ago

2 contributors

264 lines (154 sloc) | 16.3 KB | Raw | Blame | History

Machine learning and security

Towards secure adoption of machine learning techniques

Machine Learning Introduction

AI definition

Artificial Intelligence combines theories and computer hardware and software implementations to mimic the human intelligence. these tasks are normally difficult for machines and easy for humans like: understanding images and videos, speech recognition, decision-making, robotics actions.

What is machine learning?

Machine learning is one of AI techniques which mimic the human learning process, so machines become able to learn from experience.

Contributors

- **Talal Albacha:** Application Security Consulting + Academic experience in ML
- **Prabhant Singh**: Master student at University of Tartu, currently researching on secure and reliable machine learning. have been associated with owasp from last 2 years.
- **Prof. Jean-Noël Colin**: Professor in CS Faculty of University of Namur, Belgium, working in the broad field of information security, and more recently, looking at using ML methods for security purposes
- **Sereysethy Touch**: Teaching Assistant in Faculty of Computer Science at University of Namur, Belgium.

Draft document is available on:

- https://github.com/OWASP/Top-5-Machine-Learning-Risks

Definitions

Machine Learning becomes Everywhere

NETWORKWORLD FROM IDG

OPINION

The power of machine learning reaches data management

With so much to gain from computers helping us with front-end processing in apps and services, it's no surprise that machine learning is rapidly moving to the backroom of data centers. How this trans... helping enterprises overcome storage sprawl and intelligently ...

CISION PR Newswire

Fraud Prevention, Robo-Advisory Services, and Credit Scoring Transformed Through Machine Learning

... meet challenges related to ... cy and cost, finds Frost & Sullivan's Digital Transformation team

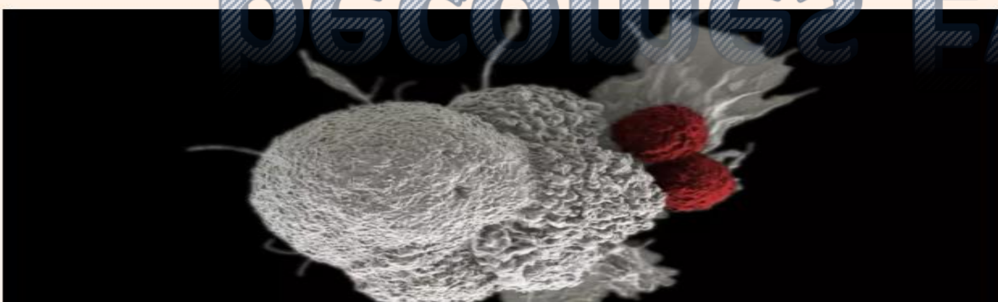
FINANCIAL TIMES

HOME WORLD UK COMPANIES MARKETS OPINION WORK & CAREERS LIFE & ARTS

Drugs research

Google parent backs machine-learning cancer tr...

Alphabet venture ...



nature

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & V...

Archive > Volume 548 > Issue 7668 > News > Article

NATURE | NEWS

How machine learning could help to improve climate forecasts

WIRED

Apple's 'Neural Engine' Infuses the iPhone With AI Smarts

theguardian

UK politics world sport football opinion culture business lifestyle fashion environment tech travel all

home > tech

Facial recognition

Face-reading AI will be able to detect your politics and IQ, professor says

But

- Does it have any risks?
- Can it be fooled? How easy?
- So we are not talking about Machine Learning use in Security
- This project is about **Security of Machine Learning**

- We will see in the next slides some attacks from research papers

Adversarial attacks

- Adversary
  - Given X, find X’ where
    - X and X’ are very close (human can’t differentiate them)
    - Output(X) != Output(X’)
- Backdoor adversary
  - Given X, find X’ where
    - X and X’ are totally different (images of two different persons)
    - Output(X) == Output(X’)

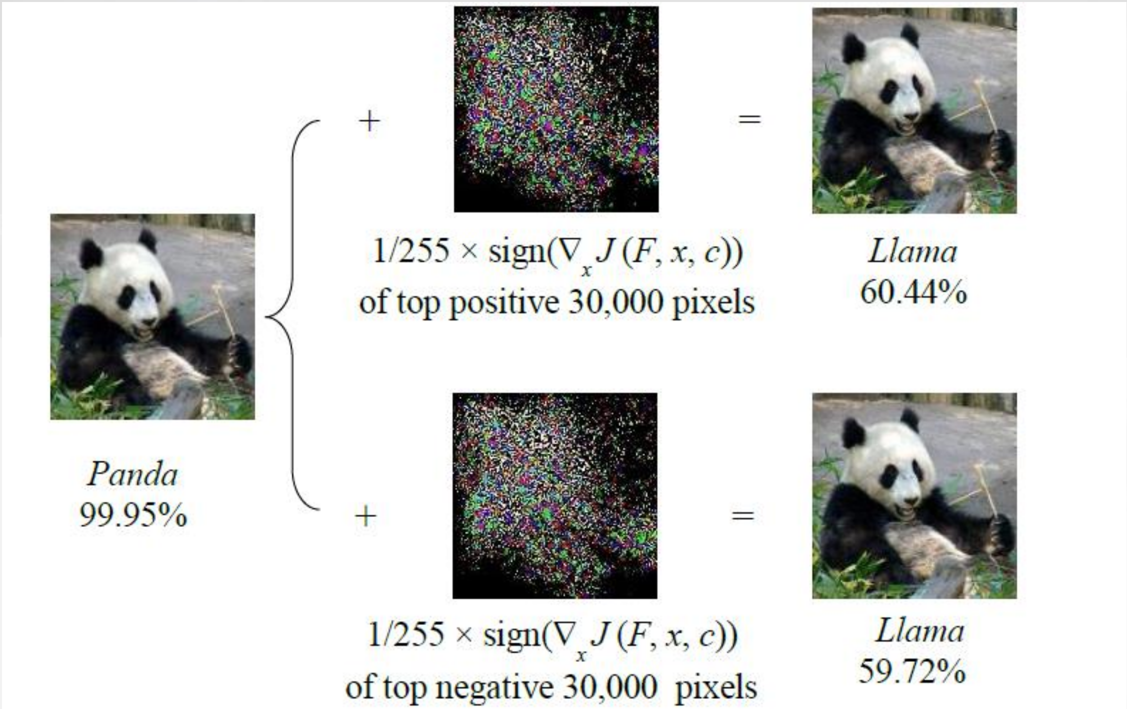
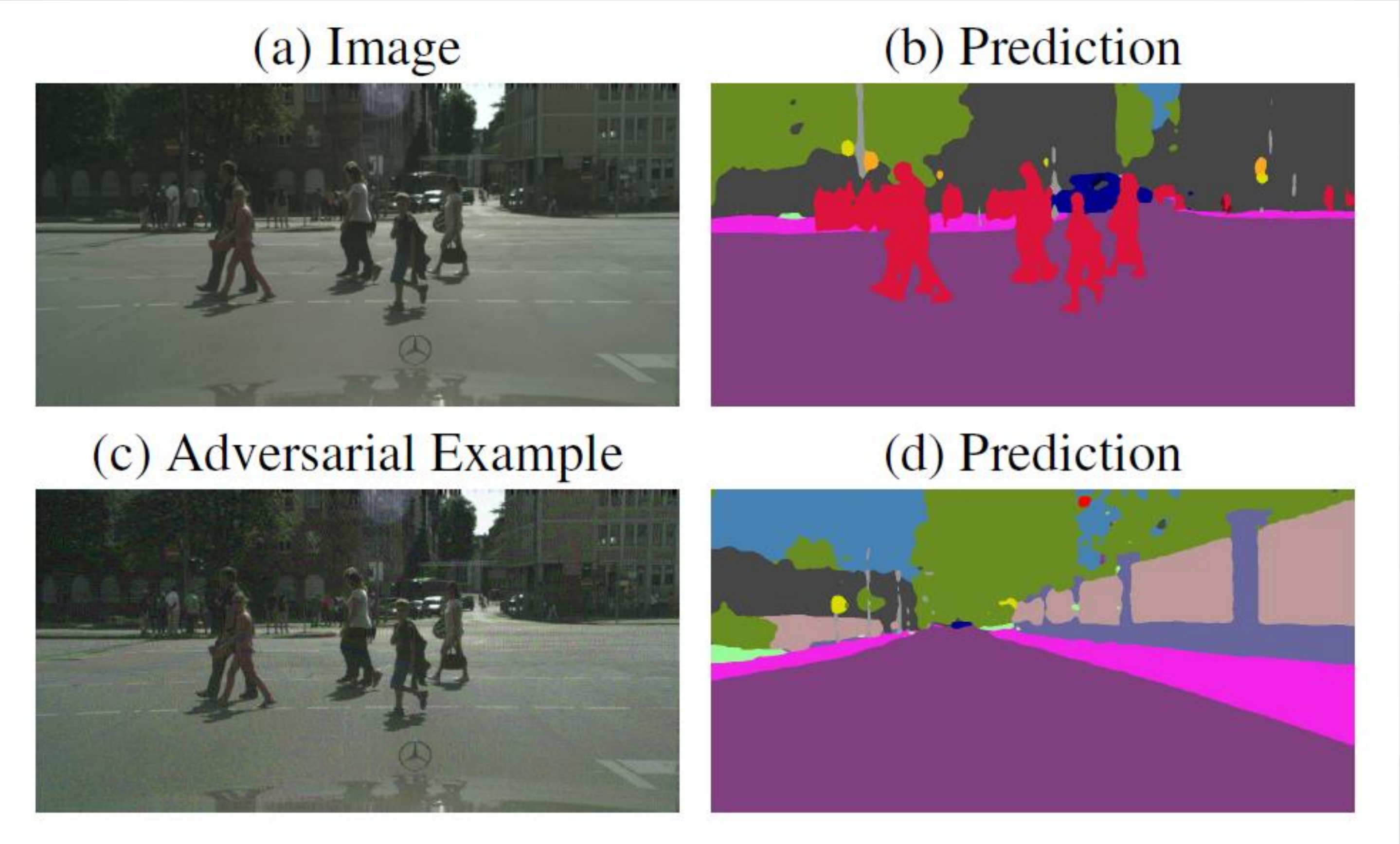
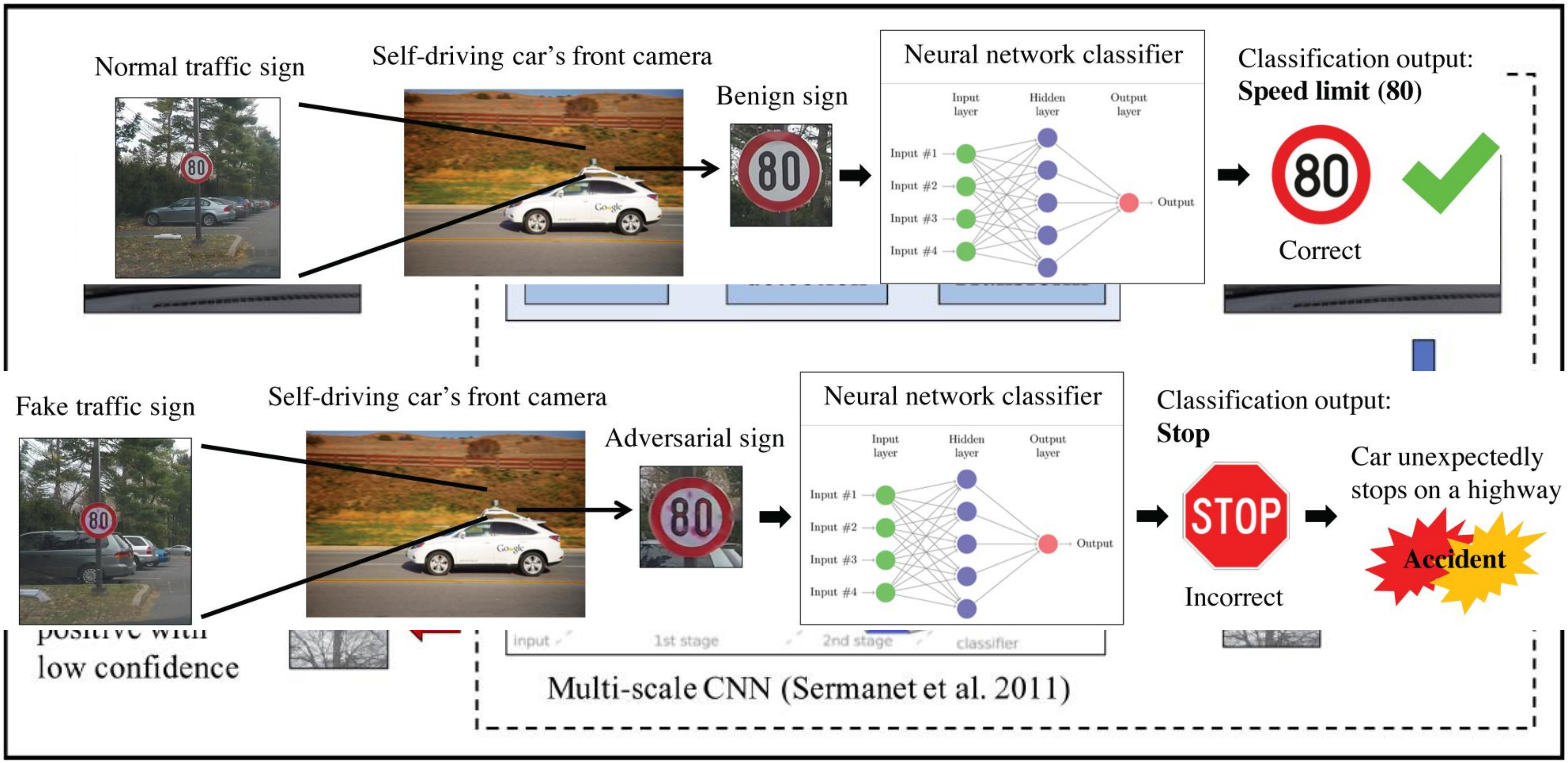


Fig. 4: Only manipulating the pixels with top gradients can still result in effectual adversarial examples.

Liang, B. et al., IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, MANUSCRIPT ID Detecting Adversarial Image Examples in Deep Neural Networks with Adaptive Noise Reduction. Available at: https://arxiv.org/pdf/1705.08378.pdf



Metzen, J.H. et al., Universal Adversarial Perturbations Against Semantic Image Segmentation. Available at: https://arxiv.org/pdf/1704.05712.pdf



Sitawarin, C. et al., DARTS: Deceiving Autonomous Cars with Toxic Signs. , 18. Available at: https://arxiv.org/pdf/1802.06430.pdf

(a) (b) (c) (d)

Figure 4: Examples of successful impersonation and dodging attacks. Fig. (a) shows S_A (top) and S_B (bottom) dodging against DNN_B. Fig. (b)–(d) show impersonations. Impersonators carrying out the attack are shown in the top row and corresponding impersonation targets in the bottom row. Fig. (b) shows S_A impersonating Milla Jovovich (by Georges Biard / CC BY-SA / cropped from https://goo.gl/GlsWlC); (c) S_B impersonating S_C; and (d) S_C impersonating Carson Daly (by Anthony Quintano / CC BY / cropped from https://goo.gl/VfnDct).

Sharif, M. et al., Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition. Available at: https://www.cs.cmu.edu/~sbhagava/papers/face-rec-ccs16.pdf

Inaudible sound commands

https://www.youtube.com/watch?v=21HjF4A3WE4
https://github.com/USSLab/DolphinAttack

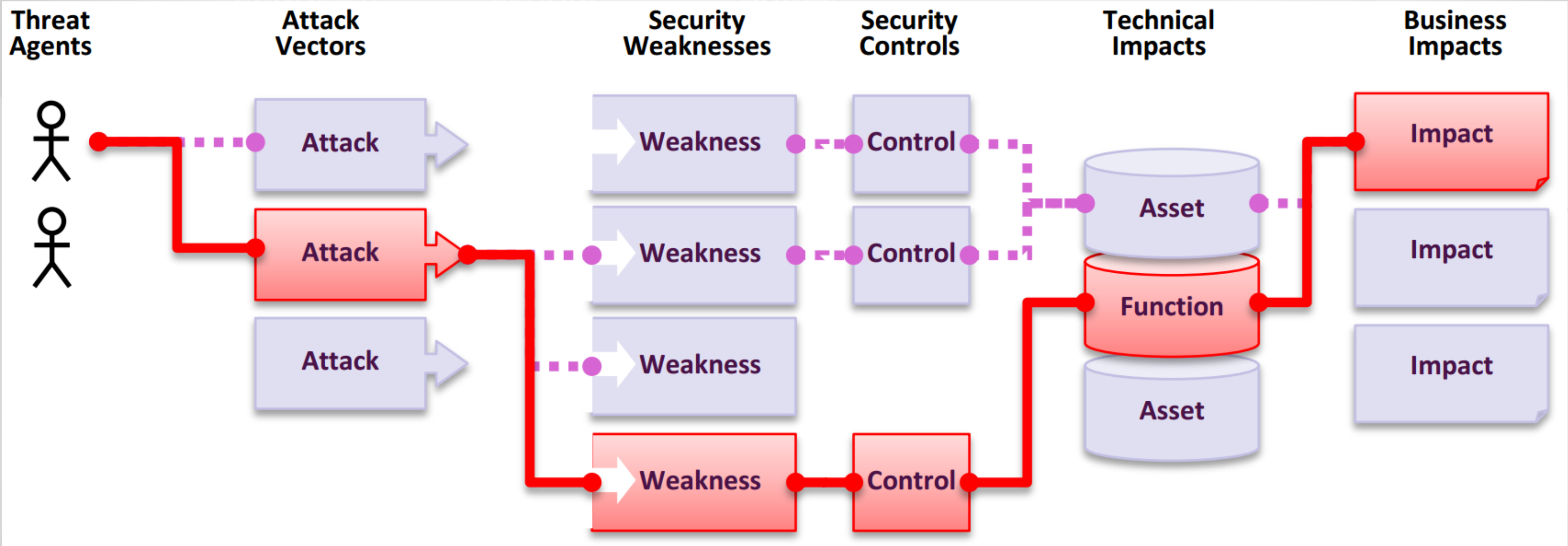
“DolphinAttack could inject covert voice commands at 7 state-of-the-art speech recognition systems (e.g., Siri, Alexa) to activate always-on system and achieve various attacks, which include activating Siri to initiate a FaceTime call on iPhone, activating Google Now to switch the phone to the airplane mode, and even manipulating the navigation system in an Audi automobile.”

Tutorial

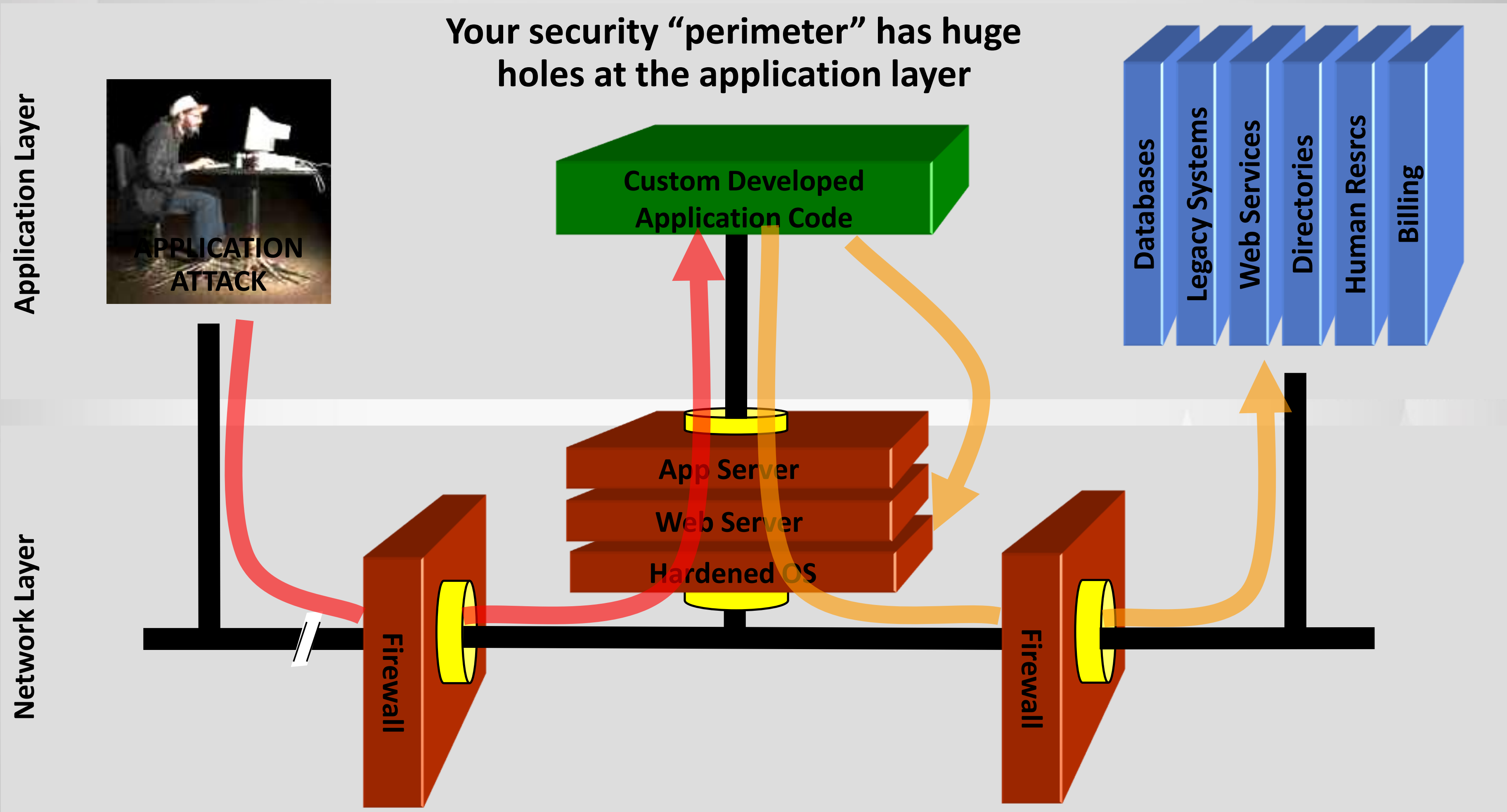
- https://github.com/prabhant/OWASP-tutorial

Mapping to Security Risks

What we all know



From History ..



Application layer Risks are technology dependant

Web

- SQL Injection
- XSS
- CSRF
- …

Mobile

- Improper platform usage
- Insecure data storage
- Insecure communication
- ….

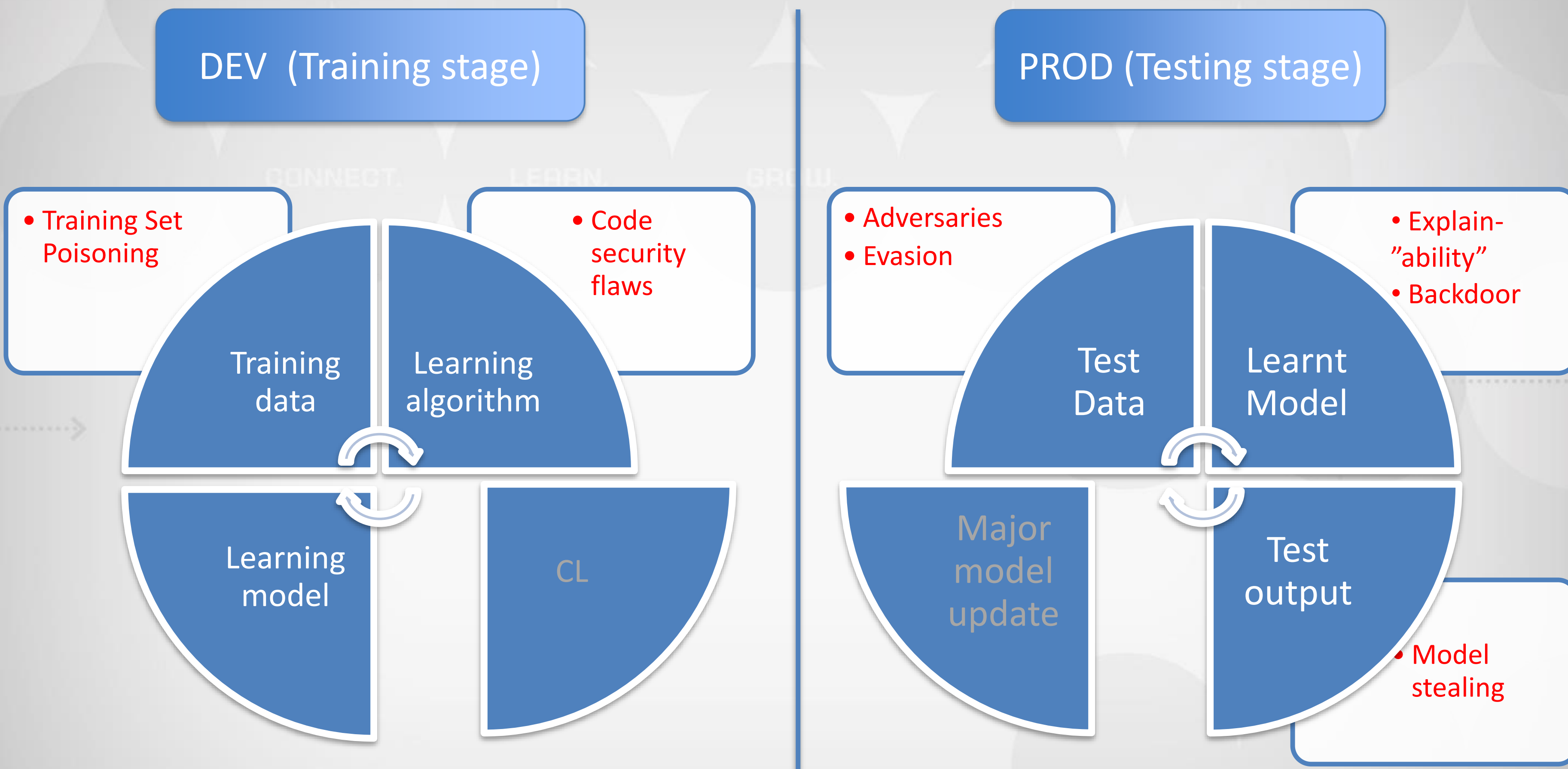
IoT

- Insecure Web Interface
- Insufficient Authentication/Authorization
- Insecure Network Services
- …

ML?

- <!--
- br
- …

Let us understand ML threat space first



opencv / opencv

Watch 1,695 | Star 18,159 | Fork 13,551

Code | Issues 1,197 | Pull requests 71 | Wiki | Insights

Filters | label:"category: vulnerability" | Labels | Milestones | New issue

Clear current search query, filters, and sorts

0 Open ✓ 7 Closed — Author | Labels | Projects | Milestones | Assignee | Sort

- **out of bound write cause Segmentfault** `bug` `category: imgcodecs` `category: vulnerability` — 1
  #9443 by xiaoqx was closed 21 days ago
- **Integer overflow in ReadNumber** `bug` `category: imgcodecs` `category: vulnerability` — 4
  #9372 by scdeny was closed 22 days ago
- **Integer overflow in PxMDecoder::readData** `bug` `category: imgcodecs` `category: vulnerability` — 1
  #9371 by scdeny was closed 22 days ago
- **AutoBuffer_heap_overflow in grfmt_pxm.cpp** `bug` `category: imgcodecs` `category: vulnerability` — 1
  #9370 by scdeny was closed 22 days ago
- **Two DOS bugs of opencv** `bug` `category: imgcodecs` `category: vulnerability` — 1
  #9311 by xiaoqx was closed 22 days ago
- **Some bugs result to crashes when calling imread of opencv (include heap overflow and out of bound write)** `bug` `category: imgcodecs` `category: vulnerability` — 1
  #9309 by xiaoqx was closed 22 days ago
- **Remote code execution via heap corruption** `bug` `category: vulnerability` — 7
  #5956 by rstevens7 ... s closed 7 days ago

CVE-2016-1516

Common Vulnerabilities and Exposures

CVE List | CNAs | Board | About | News & Blog

NVD
Go to for:
CVSS Scores
CPE Info
Advanced Search

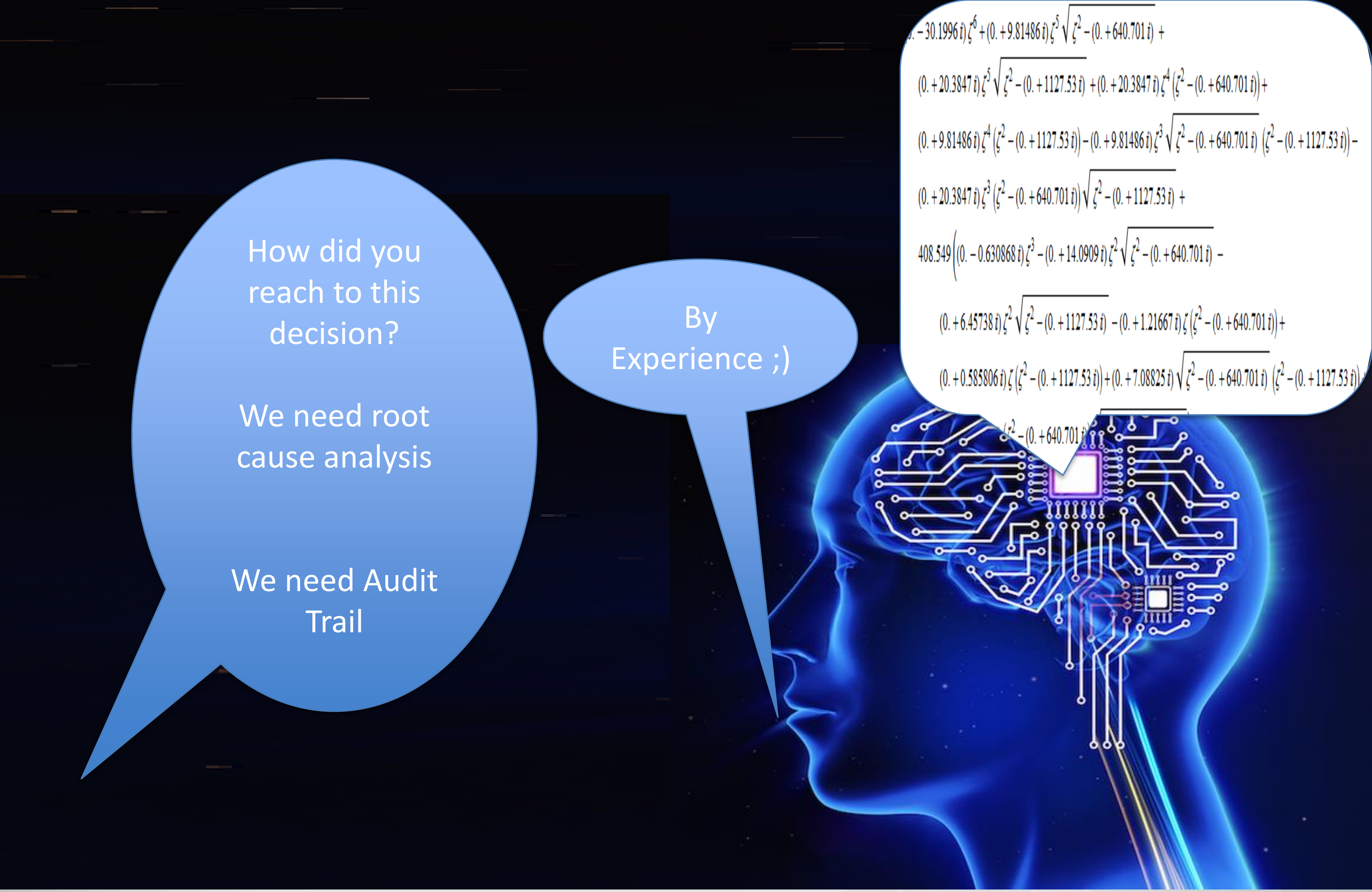
Search CVE List | Download CVE | Data Feeds | Request CVE IDs | Update a CVE Entry

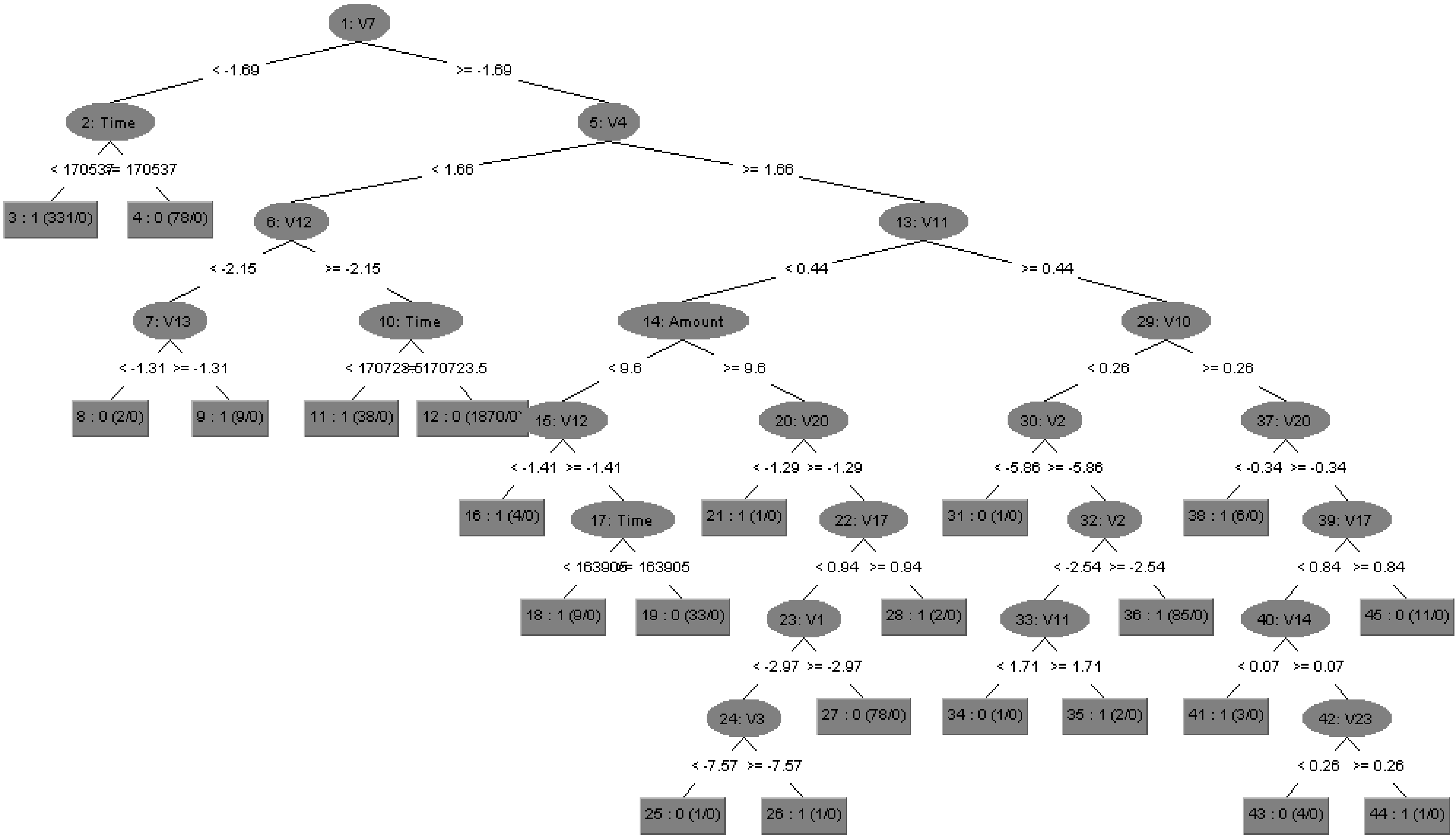
TOTAL CVE Entries: 101666

HOME > CVE > CVE-2017-5719

Printer-Friendly View

CVE-ID	
CVE-2017-5719	Learn more at National Vulnerability Database (NVD) • CVSS Severity Rating • Fix Information • Vulnerable Software Versions • SCAP Mappings • CPE Information
Description	
A vulnerability in the Intel Deep Learning Training Tool Beta 1 allows a network attacker to remotely execute code as a local user.	
References	
Note: References are provided for the convenience of the reader to help distinguish between vulnerabilities. The list is not intended to be complete.	
• CONFIRM:https://security-center.intel.com/advisory.aspx?intelid=INTEL-SA-00100&languageid=en-fr	
Assigning CNA	
Intel Corporation	
Date Entry Created	
20170201	Disclaimer: The entry creation date may reflect when the CVE ID was allocated or reserved, and does not necessarily indicate when this vulnerability was discovered, shared with the affected vendor, publicly disclosed, or updated in CVE.
Phase (Legacy)	
Assigned (20170201)	


How did you reach to this decision?
We need root cause analysis
We need Audit Trail
By Experience ;)
$-30.1996\,i)\zeta^6+(0.+9.81486\,i)\zeta^5\sqrt{\zeta^2-(0.+640.701\,i)}+$
$(0.+20.3847\,i)\zeta^5\sqrt{\zeta^2-(0.+1127.53\,i)}+(0.+20.3847\,i)\zeta^4\left(\zeta^2-(0.+640.701\,i)\right)+$
$(0.+9.81486\,i)\zeta^4\left(\zeta^2-(0.+1127.53\,i)\right)-(0.+9.81486\,i)\zeta^3\sqrt{\zeta^2-(0.+640.701\,i)}\left(\zeta^2-(0.+1127.53\,i)\right)-$
$(0.+20.3847\,i)\zeta^3\left(\zeta^2-(0.+640.701\,i)\right)\sqrt{\zeta^2-(0.+1127.53\,i)}+$
$408.549\Big((0.-0.630868\,i)\zeta^3-(0.+14.0909\,i)\zeta^2\sqrt{\zeta^2-(0.+640.701\,i)}-$
$(0.+6.45738\,i)\zeta^2\sqrt{\zeta^2-(0.+1127.53\,i)}-(0.+1.21667\,i)\zeta\left(\zeta^2-(0.+640.701\,i)\right)+$
$(0.+0.585806\,i)\zeta\left(\zeta^2-(0.+1127.53\,i)\right)+(0.+7.08825\,i)\sqrt{\zeta^2-(0.+640.701\,i)}\left(\zeta^2-(0.+1127.53\,i)\right)$
$\zeta^2-(0.+640.701\,i)$



1: V7
< -1.69
>= -1.69
2: Time
< 170537
>= 170537
3 : 1 (331/0)
4 : 0 (78/0)
5: V4
< 1.66
>= 1.66
6: V12
< -2.15
>= -2.15
7: V13
< -1.31
>= -1.31
8 : 0 (2/0)
9 : 1 (9/0)
10: Time
< 170723.5
>= 170723.5
11 : 1 (38/0)
12 : 0 (1870/0
13: V11
< 0.44
>= 0.44
14: Amount
< 9.6
>= 9.6
15: V12
< -1.41
>= -1.41
16 : 1 (4/0)
17: Time
< 163905
>= 163905
18 : 1 (9/0)
19 : 0 (33/0)
20: V20
< -1.29
>= -1.29
21 : 1 (1/0)
22: V17
< 0.94
>= 0.94
23: V1
< -2.97
>= -2.97
24: V3
< -7.57
>= -7.57
25 : 0 (1/0)
26 : 1 (1/0)
27 : 0 (78/0)
28 : 1 (2/0)
29: V10
< 0.26
>= 0.26
30: V2
< -5.86
>= -5.86
31 : 0 (1/0)
32: V2
< -2.54
>= -2.54
33: V11
< 1.71
>= 1.71
34 : 0 (1/0)
35 : 1 (2/0)
36 : 1 (85/0)
37: V20
< -0.34
>= -0.34
38 : 1 (6/0)
39: V17
< 0.84
>= 0.84
40: V14
< 0.07
>= 0.07
41 : 1 (3/0)
42: V23
< 0.26
>= 0.26
43 : 0 (4/0)
44 : 1 (1/0)
45 : 0 (11/0)

$$CL = (CI/CD)^2$$

Supporting human decision

Automating decisions

OWASP
Open Web Application Security Project

Counting false positives only is not accurate

Correctly Classified Instances 83 %
Incorrectly Classified Instances 17 %
=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	ROC Area	PRC Area	Class
	0.909	0.324	0.897	0.950	fraud
	0.676	0.091	0.897	0.790	not_fraud
Weighted Avg.	0.830	0.244	0.897	0.896	

		Condition (as determined by "Gold standard")		
		Condition positive	Condition negative	
Test outcome	Test outcome positive	**True positive**	**False positive** (Type I error)	Precision = Σ True positive / Σ Test outcome positive
	Test outcome negative	**False negative** (Type II error)	**True negative**	Negative predictive value = Σ True negative / Σ Test outcome negative
		Sensitivity = Σ True positive / Σ Condition positive	Specificity = Σ True negative / Σ Condition negative	Accuracy

Risks (draft version)

Types of attacks:(attackers knowledge)

- Whitebox: Attacker knows about model used + data + hyperparameters/meta data
- Gray Box: Partial knowledge about model or data
- Black box: No knowledge about model or data

Types of Risks

- Poisoning of the classifier training data
- Adversarial ML
- Explain"ability" of learning model.
- Code security flaws.
- Model stealing

OWASP
Open Web Application Security Project

Defence techniques (draft version)

ML Model Level

- **Adversarial training**
  - Black listing (training on specific adversaries during the training phase)
  - Training on more generic generated adversarial examples (e.g. by applying Gaussian noise)
  - Monitor classification errors
- **Robustness evaluation**
  - **Ensemble classification**
- **Model Design**
  - Select the ML model which can support audit requirements. for example, if there is need to know how the system reached to specific decision, then the model should be using decision trees instead of neural networks.

Defence techniques (draft version)

Data protection

- apply what we already know in data protection and application protection techniques on machine learning systems

Procedural

- For critical decisions:
  - Consider different factors (e.g. ML based authentication as second factor of authentication only)
  - Add human factor .. Use it only as recommendation system
- Design a process to deal with false positives

Technical Level

- **Protection from excessive access**
- **Security Scanning**
  - All ML libraries should go through static analysis

Thank you