JARVIS NEVER SAW IT COMING

Hacking machine learning (ML) in speech, text and face recognition – and frankly, everywhere else

LEGAL NOTICES AND DISCLAIMERS



This presentation contains the general insights and opinions of its authors, Guy Barnhart-Magen and Ezra Caltum. We are speaking on behalf of ourselves only, and the views and opinions contained in this presentation should not be attributed to our employer.

The information in this presentation is provided for informational and educational purposes only and is not to be relied upon for any other purpose. Use at your own risk! We makes no representations or warranties regarding the accuracy or completeness of the information in this presentation. We accept no duty to update this presentation based on more current information. We disclaim all liability for any damages, direct or indirect, consequential or otherwise, that may arise, directly or indirectly, from the use or misuse of or reliance on the content of this presentation.

No computer system can be absolutely secure.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

*Other names and brands may be claimed as the property of others.

PROPER USE DISCLAIMER

No Horses, Flamingos, Hedgehogs, Turtles or sentient^{*} AI models were harmed during the making of this presentation

* We hope

@barnhartguy 🛛 💬 @acaltum

\$ ID



Guy Barnhart-Magen

@barnhartguy

BSidesTLV Co-founder and CTF Lead

Ezra Caltum

@acaltum

BSidesTLV Co-Founder

DC9723 Lead



@acaltum



BUILDING ON THE SHOULDERS OF GIANTS



@barnhartguy

ک^{*} @acaltum

https://www.deviantart.com/callyste/art/Rocket-Raccoon-and-Groot-485953724

HOW DID WE GET HERE?



WHAT CAN YOU EXPECT?

What are we going to talk about

WHAT CAN YOU EXPECT?

What are we going to talk about

What you should be paying attention to

@barnhartguy 🛛 💬 @acaltum

WHAT CAN YOU EXPECT?

What are we going to talk about

What you should be paying attention to

What we are <u>not</u> going to talk about

@barnhartguy § 🦵 @acaltum





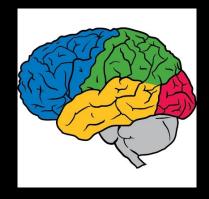
https://github.com/tensorflow/cleverhans

https://upload.wikimedia.org/wikipedia/commons/e/e3/CleverHans.jpg

@barnhartguy 🖉 @acaltum

"We have reached the point where machine learning works, but may easily be broken"

Nicolas Papernot, Google PhD Fellow in Security Ian Goodfellow, Research scientist at Google Brain



http://www.cleverhans.io/security/privacy/ml/2016/12/15/breaking-things-is-easy.html https://pbs.twimg.com/profile_images/799327801388077057/HcDnA1H7_400x400.jpg

@barnhartguy

∂acaltum

SOME BACKGROUND



ARTIFICIAL INTELLIGENCE?

Machine Learning Study many images labeled as flamingo Identify the flamingo in the image



@barnhartguy

t 🦵 @acaltum

ARTIFICIAL INTELLIGENCE?

Machine Learning Study many images labeled as flamingo Identify the flamingo in the image Deep Learning Study many images

@acaltum

Identify the flamingo, hedgehog, etc.



ARTIFICIAL INTELLIGENCE?

Machine Learning

Study many images labeled as flamingo Identify the flamingo in the image

Deep Learning

Study many images Identify the flamingo, hedgehog, etc. Artificial Intelligence

Is she hugging the flamingo, or playing cricket? Is she happy, sad?

💓 @acaltum

EVERYBODY EXCHANGES "AI" AND "ML"

So do I

Sorry



"INTELLIGENT" SYSTEM

Most AI systems were designed to **solve a specific problem**, well.



https://www.reactiongifs.us/wp-content/uploads/2015/02/do_the_robot_futurama.gif



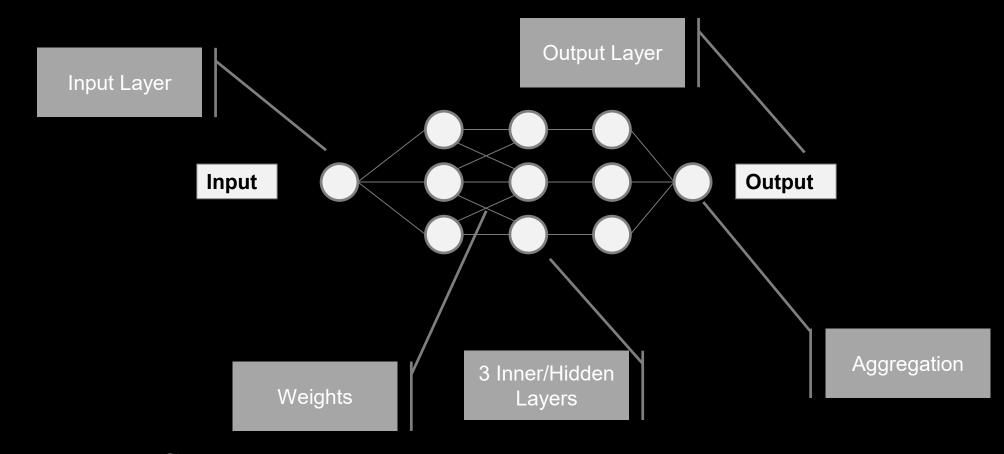
MACHINE LEARNING 101





@barnhartguy 🚽 🖓 @acaltum

WHAT IS A ML MODEL?



@barnhartguy 🖓 @acaltum

WHAT IS A ML MODEL?

Output Input

@barnhartguy 💬 @acaltum

WHAT IS A ML MODEL?

- Training: Iterative process to adjust weights
- The "model" includes:
 - Topology/Layout
 - Weights/Parameters
 - \circ Functions
- This is the real IP (Intellectual Property) in the system!

NOW SERIOUSLY

• When multiplying one matrix with another, you get a new matrix

NOW SERIOUSLY

- When multiplying one matrix with another, you get a new matrix
- The values are the product of the rows and columns of these matrices

NOW SERIOUSLY

- When multiplying one matrix with another, you get a new matrix
- The values are the product of the rows and columns of these matrices
- A vector is a single dimensioned matrix, so an array is a vector, and a matrix is a two dimensional array

CODE POINT OF VIEW

int16 vector = [];

```
struct weights {
    int rows;
    int cols;
    double **data;
};
```

@barnhartguy 💓 @acaltum

TOO MUCH VOODOO!

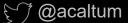


Images Audio Binaries Text

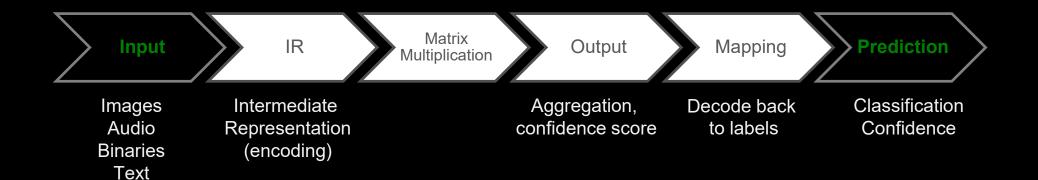
51~f*9\>71rB



Classification Confidence



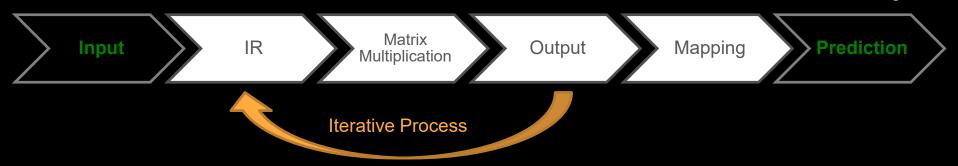
NOT TOO MUCH VOODOO!



@barnhartguy 🚽 🏹 @acaltum

FROM TRAINING TO INFERENCE

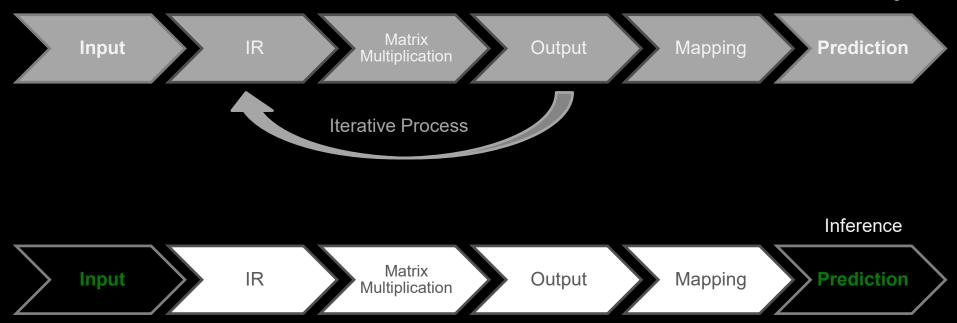
Training



@barnhartguy 🛛 💭 @acaltum

FROM TRAINING TO INFERENCE

Training



@barnhartguy 🛒 @acaltum

MODEL != CODE





Code execution flow

ML MODEL

Math operations, transition functions

EXECUTABLE

Code execution flow

Data Structures

ML MODEL

Math operations, transition functions

Intermediate Representation

@barnhartguy

💓 @acaltum

EXECUTABLE

Code execution flow

Data Structures

Code Review or Reverse Engineering

ML MODEL

Math operations, transition functions

Intermediate Representation

Model structure (Black Magic)

@barnhartguy

🗇 @acaltum

\$ HEXDUMP / MODELS/RESNET

2D 35 30 1A 04 64 61 74 00000000 OA 09 52 65 73 4E 65 74 61 A2 06 C9 ..ResNet-50..data... 31 12 OB 43 00000014 A8 02 0A 05 63 6F 6E 76 6F 6E 76 6F 6C 75 74 69Conv1..Convoluti on..data".conv1:...* 00000028 6F 6E 1A 04 64 61 74 61 22 05 63 6F 6E 76 31 3A 8C A6 02 2A 21 82 3C BEr.<d..<.!.<...;. 0000003C 80 A6 02 0E 72 E7 3C 64 FD 94 3C 1E 99 1E 3B 99 00000050 F8 41 BD 63 E8 02 3C 07 &Y.O.A.c..<.... 26 59 BD 30 E7 97 3C E1 DD CB 3C 86 00000064 22 9D 3D 7B DC B8 3D E1 F3 9A BC 7F 85 A2 BD BB AF 7B BC A9 00000078 A9 86 BD 8E 0F E1 BC C7 C6 A3 3C 9F 32 3D 3E 3D 9A EC 3D A2 0000008C 47 B8 BD 7F D5 C1 BD 62 5C 20 BC 30 47 A2 BD 68 EF 04 BE EA G.....b\ .0G..h.... 000000A0 OA C3 3D C3 46 8A 3E 08 96 CB 3D 9D A3 74 BC F8 E4 BD 3C 11 ..=.F.>...=..t....<. 000000B4 44 3E BD 16 66 3B BE 5B C9 1F BE 30 9D 5B 3D C4 FF AA 3D DB D>..f;.[...0.[=...=. 000000C8 60 17 3D 60 13 2D 3D FC 32 3C 3D 9B FO 3E BD FA 16 02 BE 5A `.=`.-=.2<=..>....Z 13 08 3D 75 B3 6F 3C A5 4D 10 3D 95 :....=u:.<..o<.M.=. 00000DC 3A AC BC FB 3A CD 3C FE 000000F0 A2 86 3B 58 56 77 BD 7C A6 8C BC 49 A6 3C 3C C3 01 9D 3B AF ..;XVw.|...i.<<...;. 00000104 A4 99 3C C7 09 14 3D 82 J.;...<...=...<H...0 4A CO 3B FD 88 8C 3C 48 9B BF BD 30 00000118 28 FE BD 3E 9C DB BC 2B EB FB BB B4 09 C7 3C 0C 08 19 3E 64 (...>...+.....<...>d 0000012C CC 5C 3E E5 5C 5B 3C 02 1E 35 BE B7 CB FB BD 9A E7 98 BD 64 .\>.\[<..5.....d F4 79 3D B1 00000140 27 1C BE 8F 7A D4 3E BF 27 AA 3E 62 ED A6 BD 9B '....y=.z.>.'.>b.... 00000154 B5 82 BE 52 -X.1. .*...Rc...cW>. 2D 58 BE 6C 01 5F BD 2A 63 87 BE 02 63 57 3E 1F 00000168 F5 11 3F 8D F2 8D 3E 0A EF D8 BC E1 D0 BC 3C 15 1F 36 BE 3F 0000017C 72 89 BE 71 48 80 3E FD 5D C5 3D 9Ar..q@>>.H.>.].=. D5 D2 BE DC 40 3E 3E D3 00000190 17 C8 3D 51 88 1F 3D C4 F3 44 BE 7D A0 95 BE 1E C7 18 BD F1 ..=Q..=..D.}.... 000001A4 3F CD 3D FE C2 7A 3D DB EE 81 3D C8 E8 CB 3D D8 CE B6 BC 3E ..=.?.=..z=...> 000001B8 95 2E BE 76 BF BO BD FE 14 F3 3B 27 70 E1 3C A2 F7 1F BB 10v....;'p.<.... 0B 92 BD <mark>2</mark>2 000001CC 05 9C 3C 7C 8D B2 3C 08 93 5D BB AA 9C 9D BD AA ResNet-50-model.caffemodel --0x1DF/0x61B73BD------

FUN FACTS!

The algorithm is designed to optimize for the "strongest signal"

The model (matrices) can be GB in size Bias is a part of the system learning process

@barnhartguy 🛛 💬 @acaltum

BIAS - SOLVING THE WRONG PROBLEM

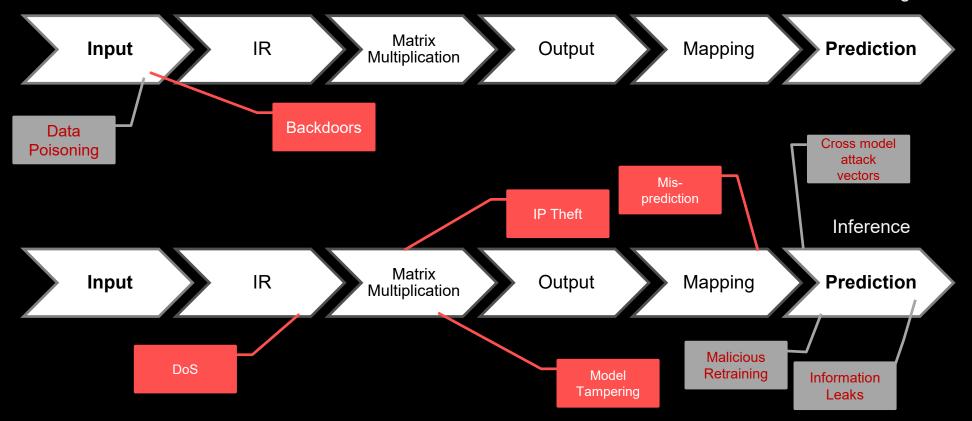


@barnhartguy

acaltum) 💬

FROM TRAINING TO INFERENCE

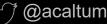
Training



@barnhartguy 💬 @acaltum

SCORING

We used the CVSS 3.0 scoring, and ordered by business impact



TOP 5 ATTACKS (CVSS)

- 1 DoS
- 2 Misprediction (adversarial attacks)
- 3 Model Tampering
- 4 IP Theft
- 5 Backdoors

7.5 (High)
7.5 (High)
7.4 (High)
5.9 (Medium)
3.9 (Low)



TOP 5 ATTACKS (BUSINESS IMPACT)

- 1 IP Theft
- 2 Model Tampering
- 3 DoS
- 4 Backdoors
- 5 Misprediction (Adversarial attacks)

5.9 (Medium)
7.4 (High)
7.5 (High)
3.9 (Low)
7.5 (High)

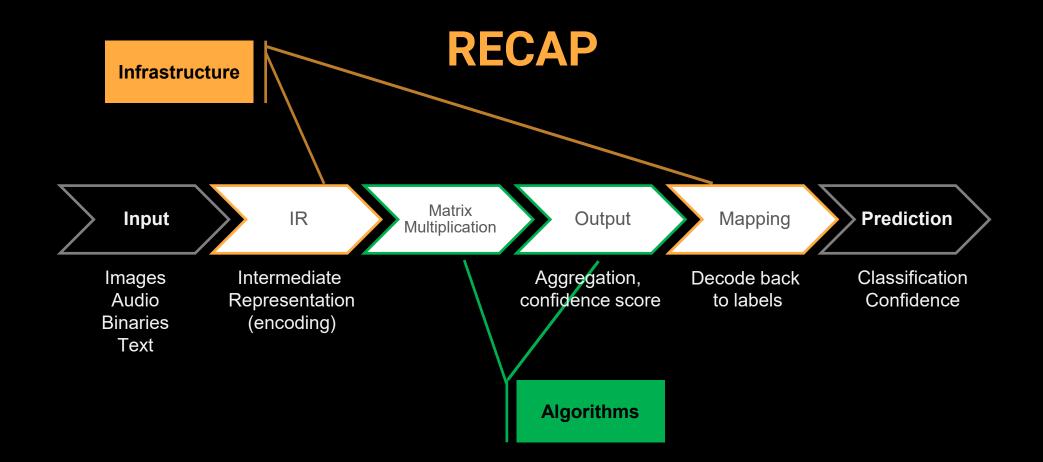
HOW TO BUILD AN ATTACK

What do you need to know? What areas should you target? What do you need to have access to?

@barnhartguy 🛛 💭 @acaltum

WHERE TO ATTACK?

You can either go after the system infrastructure, or the algorithms



@barnhartguy 🛛 🖓 @acaltum



@barnhartguy

💓 @acaltum



Parsing is hard

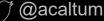
@acaltum



Parsing is hard

Al developers don't develop file formats. Or parsers.

@barnhartguy 🕤 @a





Parsing is hard

Al developers don't develop file formats. Or parsers.

The common solution is to just bring the dependency into the project

@barnhartguy 💓 @acaltum

DEPENDENCIES

So – they are bringing outside libraries into their stack

DEPENDENCIES

So – they are bringing outside libraries into their stack.

And bringing with them a common problem – supply chain and patch management

@barnhartguy 🛛 💬 @acaltum

DEPENDENCIES

So – they are bringing outside libraries into their stack.

And bringing with them a common problem – supply chain and patch management

A common framework, must support multiple file formats...

So let's fuzz the file format parsing

@barnhartguy 🛛 💬 @acaltum

WE FAILED MISERABLY AT THE BEGINNING

We were fuzzing on a pretty large compute cluster, but we had terrible performance!

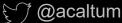


What to focus on?Why focus here?Issues?CaffeFull coverageExtremely slow





What to focus on?Why focus here?Issues?CaffeFull coverageExtremely slowOpenCVLimited coverageMedium speed



FUZZING

Caffe OpenCV LibXXX

What to focus on? Why focus here? Full coverage Limited coverage Very fast

Issues? Extremely slow Medium speed Unknown code paths

FUZZING

What to focus on?Why focus here?CaffeFull coverageOpenCVLimited coverageLibXXXVery fastUpstreamFuzzing not need

Why focus here?Issues?Full coverageExtremely slowLimited coverageMedium speedVery fastUnknown code pathsFuzzing not neededPatched? Workable?



WE FAILED MISERABLY AT THE BEGINNING

We used a RAM disk to accelerate access

And then used the same cluster for a different research, and rebooted

We also forgot to turn off the logging which lead to a system crash which lead to loosing some valid crashes

@barnhartguy 💓 @acaltum

FACE PALM

Scaling exploitation is hard (DevOps?)



EVENTUALLY WE GOT OUR GRIP



Almost every exploit category was discovered

Almost every exploit category was discovered Bug collisions

@acaltum

Almost every exploit category was discovered Bug collisions So many crashes

Almost every exploit category was discovered Bug collisions So many crashes We were like children in a candy store

INITIAL EXPLOITATION

We focused on crashes in the BMP file format Mostly – because it's pretty easy to manually craft





$\textbf{FUZZING} \rightarrow \textbf{CRASH, NOW WHAT?}$

IP Theft
 Model Tampering
 DoS
 Backdoors
 Misprediction (Adversarial attacks)
 Signal (Low)
 Tightham (Low)

Is Remote Code Execution (RCE) king?

POST EXPLOITATION

Let's demonstrate the TOP 5 Input: Image file (~10K) Output: Label (string)

DEMO TIME

Denial of Service



demos ssh user@localhost -p 60000	demos ssh user@localhost -p 60000
Welcome to Ubuntu 16.04.4 LTS (GNU/Linux 4.4.0-119-generic x86_64)	Welcome to Ubuntu 16.04.4 LTS (GNU/Linux 4.4.0-119-generic x86_64)
* Documentation: https://help.ubuntu.com	* Documentation: https://help.ubuntu.com
* Management: https://landscape.canonical.com	* Management: https://landscape.canonical.com
<pre>* Support: https://ubuntu.com/advantage</pre>	* Support: https://ubuntu.com/advantage
Last login: Wed Jul 25 14:48:43 2018 from 10.0.2.2	Last login: Wed Jul 25 14:49:29 2018 from 10.0.2.2
→ 1 ↔	

@barnhartguy 💓 @acaltum

→ demos ssh user@l	ocalhost -p 60000	→ demos ssh user@localhost -p 60000							
Welcome to Ubuntu 1	6.04.4 LTS (GNU/Linux 4.4.0-119-generic x86_64)	Welcome to Ubuntu 16.04.4 LTS (GNU/Linux 4.4.0-119-generic x86_64							
* Documentation:	https://help.ubuntu.com	<pre>* Documentation: https://help.ubuntu.com</pre>							
* Management:	https://landscape.canonical.com	* Management: https://landscape.canonical.com							
* Support:	https://ubuntu.com/advantage	* Support: https://ubuntu.com/advantage							
Last login: Wed Jul	25 14:48:43 2018 from 10.0.2.2	Last login: Wed Jul 25 14:49:29 2018 from 10.0.2.2							
→ ~		🔶 ~ htop							



→ demos ssh user@localhost -p 60000	l											
Welcome to Ubuntu 16.04.4 LTS (GNU/Linux 4.4.0-119-generic x86_64)	1	[]	Tasks: 22, 6 thr; 1 running						
	2	The second second]	Load average: 0.07 0.03 0.00						
* Documentation: https://help.ubuntu.com	Mem	נווו]	Upt	ime: 00	:09:4	9			
* Management: https://landscape.canonical.com	Swp	[]							
* Support: https://ubuntu.com/advantage												
Last login: Wed Jul 25 14:48:43 2018 from 10.0.2.2		USER	PRI		VIRT	RES	SHR S			TIME+	Com	
→ ~		root	20		37980	5968	3952 S			0:02.55		
	207		20		35272	3528	3220 S			80.00:0		
	244		20		44772	4244	2976 S			0:00.87		
	360		20	0		2460	2252 S			0:00.00		
	335		20	0		2460	2252 S			0:00.02		
	471		20		16120	856	0 5			0:00.00		
	506		20		28620	3080	2760 S			0:00.02		
	533		20	0		3416	2744 S			0:00.00		
			20	0	250M	3416	2744 S			0:00.00		
	535 507		20 20	0	250M 250M	3416 3416	2744 S 2744 S			0:00.00		
	511		20	0	42900	3900	3488 S			0:00.02		
	557		20	0		6 260	5532 S			0:00.01		
	565		20	0		6260	5532 S			0:00.00		
	538		20	0		6260	5532 S			0:00.04		
	546		20		29008	2920	2648 S			0:00.00		
	563		20		19472	2288	2064 S			0:00.01		
	665		20		65508	6040	5332 S			0:00.00		
	679		20		15936	1792	1664 S			0:00.00		
	909		20		92800	6940	6000 S			0:00.00		
	911	user	20	0	45192	4988	4 156 S	0.0	0.1	0:00.01	/1i	
		user	20		61432	2128	0 S			0:00.00		
	935	user	20		92800	3 328	2392 S			0:00.00) ssh	
	936	user	20	0	44316	5448	3852 S	0.0	0.1	0:00.17		
	965		20		92800	6748	5812 S			0:00.00		
		user	20		92800	3304	2372 S		0.0	0:00.00) ssh	
	F1Help	F2 Setu	p F3Se	arch	n <mark>F4</mark> Filt	er <mark>F5</mark> Tr	ee <mark>F6</mark> S	ortBy	F7Nice	e - <mark>F8</mark> Nice	+F9	

@barnhartguy 🔬 @acaltum

→ demos ssh user@localhost -p 60000							
Welcome to Ubuntu 16.04.4 LTS (GNU/Linux 4.4.0-119-generic x86 64)	1 r		_1	Tasks:	22. 6 th	r: 1 r	unning
	2	Tasks: 22, 6 thr; 1 running Load average: 0.04 0.03 0.00					
* Documentation: https://help.ubuntu.com	Mem[] Uptime: 00:10:12						
* Management: https://landscape.canonical.com	Swp[1	E CONTRACTOR			
* Support: https://ubuntu.com/advantage							
Last login: Wed Jul 25 14:48:43 2018 from 10.0.2.2	PID USER	PRI	NI VIRT	RES SI	R S CPU	MEM%	TIME+
<pre>> ~ cd /home/user/for presentation/jarvis demo/runner/</pre>	1011 user	20	0 25924		96 R 0.7		0:00.06
	1	20	0 37980		52 S 0.0		0:02.55
	207	20	0 35272		20 S 0.0	0.0	0:00.08
	244	20	0 44772	4244 29	76 S 0.0	0.1	0:00.87
	360	20	0 97M		52 S 0.0	0.0	0:00.00
	335	20	0 97M	2460 22	52 S 0.0	0.0	0:00.02
	471	20	0 16120	856	0 S 0.0	0.0	0:00.00
	506	20	0 28620	3080 27	50 S 0.0	0.0	0:00.02
	533	20	0 250M	3416 27	44 S 0.0	0.0	0:00.00
	534	20	0 250M	3416 27	44 S 0.0	0.0	0:00.00
	535	20	0 250M	3416 27	44 S 0.0	0.0	0:00.00
	507	20	0 250M	3416 27	44 S 0.0	0.0	0:00.02
	511	20	0 42900	3900 34	38 S 0.0	0.0	0:00.08
	557	20	0 269M	6260 5 5	32 S 0.0	0.1	0:00.01
	565	20	0 269M	<mark>6260 5</mark> 5	32 S 0.0	0.1	0:00.00
	538	20	0 269M	6260 5 5	32 S 0.0	0.1	0:00.04
	546	20	0 29008	2 920 2 6	48 S 0.0	0.0	0:00.00
	563	20	0 19472	2288 20	54 S 0.0	0.0	0:00.01
	665	20	0 65508	6 040 5 3	32 S 0.0	0.1	0:00.00
	679	20	0 15936	1792 1 6	54 S 0.0	0.0	0:00.00
	909	20	0 92800	6940 60	00 S 0.0	0.1	0:00.00
	911 user	20	0 45192	4988 41	56 S 0.0	0.1	0:00.01
	912 user	20	0 61432	<mark>2</mark> 128	0 S 0.0	0.0	0:00.00
	935 user	20	0 92800	3 328 2 3	92 S 0.0	0.0	0:00.00
	936 user	20	0 44316	5 448 3 8	52 S 0.0	0.1	0:00.17
	965	20	0 92800	<mark>6</mark> 748 58	12 S 0.0	0.1	0:00.00

TIME+ Comm

0.0 0:00.06 htop

0.1 0:02.55 /sbi

0.0 0:00.08 /lib 0.1 0:00.87 /lib

0.0 0:00.00 /lib

0.0 0:00.02 /lib

0.0 0:00.00 /sbi

0.0 0:00.02 /lib

0.0 0:00.00 /usr

0.0 0:00.00 /usr

0.0 0:00.00 /usr

0.0 0:00.02 /usr

0.0 0:00.08 /usr

0.1 0:00.01 /usr

0.1 0:00.00 /usr

0.1 0:00.04 /usr

0.0 0:00.00 /usr

0.0 0:00.01 /usr

0.1 0:00.00 /usr

0.0 0:00.00 /sbi

0.1 0:00.00 sshd

0.1 0:00.01 /lib

0.0 0:00.00 (sd-

0.0 0:00.00 sshd

0.1 0:00.17 -zsh

0.0 0.1 0:00.00 sshd

F1Help F2Setup F3SearchF4FilterF5Tree F6SortByF7Nice -F8Nice +F9K

@barnhartguy

🕤 @acaltum

→ demos ssh user@localhost -p 60000				
Welcome to Ubuntu 16.04.4 LTS (GNU/Linux 4.4.0-119-generic x86_64)	1 []
	2 []
* Documentation: https://help.ubuntu.com	Mem[]
* Management: https://landscape.canonical.com	Swp[]
* Support: https://ubuntu.com/advantage	I			
Last login: Wed Jul 25 14:48:43 2018 from 10.0.2.2	PID USE	R PRI	NI	VIRT
cd /home/user/for_presentation/jarvis_demo/runner/	1011 use	r 20	0	25924
runner ./classification-d/exploits/ram_and_cpu_dos.bmp	1	20	0	37980
	207	20	0	35272
	244	20	0	44772
	360	20	0	97M
	335	20	0	97M
	471	20	0	16 120
	506	20	0	<mark>28</mark> 620
	533	20	0	250M
	534	20	0	250M
	535	20	0	250M
	507	20	0	250M
	511	20	0	42 900
	557	20	0	269M
	565	20	0	269M
	538	20	0	269M
	546	20	0	29 008
	563	20	0	19 472
	665	20	0	<mark>65</mark> 508
	679	20	0	15 936
	909	20	0	<mark>92</mark> 800
	911 use	r 20	0	<mark>45</mark> 192
	912 use	r 20	0	<mark>61</mark> 432
	935 use	r 20		<mark>92</mark> 800
	936 use	r 20	0	44316
	965	20		<mark>92</mark> 800
	F1 <mark>Help</mark> F	2 <mark>Setup F3</mark> Se	arch	n <mark>F4</mark> Filte

Tasks: 22, 6 thr; 1 running

SHR S CPU% MEM%

0.0

0.0

0.0

0.0

0.0

0.0

0.0

0.0

0.0

0.0

0.0

0.0

0.0

0.0

0.0

0.0

0.0

0.0

0.0

6748 5812 S 0.0 0.1 0:00.00 sshd er<mark>F5</mark>Tree <mark>F6</mark>SortBy<mark>F7</mark>Nice -<mark>F8</mark>Nice +<mark>F9</mark>K

0.0

0.1

0.0

0.1

0.0

0.0

0.0

0.0

0.0

0.0

0.0

0.0

0.0

0.1

5532 S 0.0 0.1 0:00.00 /usr

5532 S 0.0 0.1 0:00.04 /usr

0.0

0.0

0.1

0.1

0.0

0.0

3852 S 0.0 0.1 0:00.17 -zsh

TIME+ Comm

0:00.08 htop

0:02.55 /sbi

0:00.08 /lib

0:00.87 /lib

0:00.00 /lib

0:00.02 /lib

0:00.00 /sbi

0:00.02 /lib

0:00.00 /usr

0:00.00 /usr

0:00.00 /usr

0:00.02 /usr

0:00.08 /usr

0:00.01 /usr

0:00.00 /usr

0:00.01 /usr

0:00.00 /usr

0:00.01 /lib

0:00.00 (sd-

0:00.00 sshd

0.0 0:00.00 /sbi

0.1 0:00.00 sshd

Uptime: 00:10:23

3196 R 0.7

3952 S 0.0

3220 S 0.0

2976 S

2252 S

2252 S

2760 S

2744 S

2744 S

2744 S

2744 S

3488 S

5532 S

2648 S

2064 S

5332 S

1664 S

6000 S

4156 S

2392 S

0 S

0 S

RES

3756

5968

3528

4244

2460

2460

856

3080

3416

3416

3416

3416

3900 6260

6260

6260 2920

2288

6040

1792

6940

4988

2128

3328

5452

Load average: 0.04 0.02 0.00

@barnhartguy

کڑ @acaltum

→ demos ssh user@localhost -p 60000													
Welcome to Ubuntu 16.04.4 LTS (GNU/Linux 4.4.0-119-generic x86_64)	1]]						unning	
	2	[]]	Loa	ad ave	rag	je: 0	.11 0.	04 0.01	
* Documentation: https://help.ubuntu.com	Me	em []	Up	time:	00:	10:2	5		
* Management: https://landscape.canonical.com	Sw	η ρ []							
* Support: https://ubuntu.com/advantage													
Last login: Wed Jul 25 14:48:43 2018 from 10.0.2.2	PI	DU	SER	PRI	NI	VIRT	RES	SHR	S	CPU%	MEM%	TIME+	Com
~ cd /home/user/for_presentation/jarvis_demo/runner/	101	.6 u	ser	20	0	354M	221M	19280	D	64.8	2.8	0:00.98	3 ./cl
runner ./classification-d/exploits/ram_and_cpu_dos.bmp	101	1 u	ser	20	0	25 924	3 756	3 196	R	0.8	0.0	0:00.09	htop
Prediction for/exploits/ram_and_cpu_dos.bmp		1		20	0	37980	<mark>5</mark> 968	<mark>3</mark> 952	S	0.0	0.1	0:02.55	5 /sbi
-	20)7		20	0	35 272	<mark>3</mark> 528	<mark>3</mark> 220	S	0.0	0.0	0:00.08	3 /lik
	24	4		20	0	44 772	<mark>4</mark> 244	<mark>2</mark> 976	S	0.0	0.1	0:00.87	/lik
	36	0		20	0	97M	2 460	<mark>2</mark> 252	s	0.0	0.0	0:00.00) /lik
	33	5		20	0	97M	<mark>2</mark> 460	<mark>2</mark> 252	S	0.0	0.0	0:00.02	2 /lik
	47	'1		20	0	16 120	856	0	s	0.0	0.0	0:00.00) /sbi
	50	6		20	0	28 620	<mark>3</mark> 080	<mark>2</mark> 760	S	0.0	0.0	0:00.02	2 /lik
	53	3		20	0	250M	<mark>3</mark> 416	<mark>2</mark> 744	S	0.0	0.0	0:00.00) /usr
	53	34		20	0	250M	<mark>3</mark> 416	<mark>2</mark> 744	s	0.0	0.0	0:00.00) /usr
	53	5		20	0	250M	<mark>3</mark> 416	<mark>2</mark> 744	S	0.0	0.0	0:00.00) /usr
	50)7		20	0	250M	<mark>3</mark> 416	<mark>2</mark> 744	S	0.0	0.0	0:00.02	2 /usr
	51	.1		20	0	42 900	<mark>3</mark> 900	<mark>3</mark> 488	S	0.0	0.0	0:00.08	3 /usr
	55	57		20	0	269M	<mark>6</mark> 260	<mark>5</mark> 532	S	0.0	0.1	0:00.01	/usr
	56	5		20	0	269M	<mark>6</mark> 260	<mark>5</mark> 532	S	0.0	0.1	0:00.00) /usr
	53	8		20	0	269M	<mark>6</mark> 260	<mark>5</mark> 532	S	0.0	0.1	0:00.04	l /usr
	54	6		20	0	29 008	<mark>2</mark> 920	<mark>2</mark> 648	S	0.0	0.0	0:00.00) /usr
	56	3		20	0	19 472	<mark>2</mark> 288	<mark>2</mark> 064	S	0.0	0.0	0:00.01	l /usr
	66	5		20	0	<mark>65</mark> 508	<mark>6</mark> 040	<mark>5</mark> 332	S	0.0	0.1	0:00.00) /usr
	67	9		20	0	15 936	1 792	1 664	S	0.0	0.0	0:00.00) /sbi
	90	9		20	0	<mark>92</mark> 800	<mark>6</mark> 940	<mark>6</mark> 000	S	0.0	0.1	0:00.00) sshc
	91	1 u	ser	20	0	45 192	<mark>4</mark> 988	<mark>4</mark> 156	S	0.0	0.1	0:00.01	/lik
	91	.2 u	ser	20	0	<mark>61</mark> 432	<mark>2</mark> 128	0	s	0.0	0.0	0:00.00) (sd-
	93	5 u	ser	20	0	92800	<mark>3</mark> 328	<mark>2</mark> 392	S	0.0	0.0	0:00.00) sshc
	93	16 u	ser	20	0	44316	5 452	<mark>3</mark> 852	S	0.0	0.1	0:00.17	-zsł
	F1He	lp	F2Setup	F3 <mark>Sea</mark>	arch	F4Filt	erF5T	ree F	6 <mark>Sc</mark>	and the second se		-F8Nice	+F9F

@barnhartguy 🖉 @acaltum

🕽 demos ssh user@localhost -p 60000								
Velcome to Ubuntu 16.04.4 LTS (GNU/Linux 4.4.0-119-generic x86_64)	1 [2 [100.0		sks: 23 ad avera			cunning .34 0.12
* Documentation: https://help.ubuntu.com	Mem[5.	98G/7.		time: 0			
* Management: https://landscape.canonical.com	Swp[1				
* Support: https://ubuntu.com/advantage				-				
Last login: Wed Jul 25 14:48:43 2018 from 10.0.2.2	PID USER	PRI	NI VIR	T RES	SHR	CPU%	MEM%	TIME+ Comm
<pre>~ cd /home/user/for_presentation/jarvis_demo/runner/</pre>	1016 user	20	0 7171	M 6065M	20432 1	R 116.	76.0	1:55.37 ./cl
runner ./classification-d/exploits/ram and cpu dos.bmp	1011 user	20	0 2592	4 <u>3</u> 756	3196 1	R 0.8	0.0	0:00.35 htop
Prediction for/exploits/ram and cpu dos.bmp	563	20	0 1947	2 2288	2064	5 0.0	0.0	0:00.02 /usr
	1	20	0 3798	0 <mark>5</mark> 968	3 952	5 0.0	0.1	0:02.56 /sbi
	985 user	20	0 9280	0 3304	2 372	5 0.0	0.0	0:00.02 sshd
	207	20	0 3527	2 <mark>3</mark> 528	3 220	5 0.0	0.0	0:00.08 /lib
	244	20	0 4477	2 4244	2 976	5 0.0	0.1	0:00.87 /lib
	360	20	0 97	M 2460	2 252	5 0.0	0.0	0:00.00 /lib
	335	20	0 97	M 2460	2252 s	5 0.0	0.0	0:00.02 /lib
	471	20	0 <mark>16</mark> 12	0 856	0	5 0.0	0.0	0:00.00 /sbi
	506	20	0 2862	0 <mark>3</mark> 080	2760	5 0.0	0.0	0:00.02 /lib
	533	20	0 250	M 3416	2744	5 0.0	0.0	0:00.00 /usr
	534	20	0 250	M 3416	2744	5 0.0	0.0	0:00.00 /usr
	535	20	0 250	M 3416	2744	5 0.0	0.0	0:00.00 /usr
	507	20	0 250	M 3416	2744	5 0.0	0.0	0:00.02 /usr
	511	20	0 4290	0 <mark>3</mark> 900	3488	5 0.0	0.0	0:00.08 /usr
	557	20	0 269	M 6260	5 532	5 0.0	0.1	0:00.01 /usr
	565	20	0 269	M 6260	5 532	5 0.0	0.1	0:00.00 /usr
	538	20	0 269				0.1	
	546	20	0 2900					0:00.00 /usr
	665	20	0 6550				0.1	
	679	20	0 1593					0:00.00 /sbi
	909	20	0 9280					0:00.00 sshd
	911 user	20	0 4519				0.1	
	912 user	20	0 6143				0.0	0:00.00 (sd-
	935 user	20	0 9280	1.00	the second s			
	F1 <mark>Help F2</mark> Setu	ip F3Se	arch <mark>F4</mark> Fi	lter <mark>F5</mark> T	ree F6	SortBy	F7Nice	e - <mark>F8</mark> Nice + <mark>F9</mark> K

@barnhartguy 💓 @acaltum

DEMO TIME

Remote Code Execution (RCE)



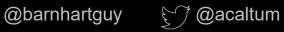
demos ssh user@localhost -p 60000 Welcome to Ubuntu 16.04.4 LTS (GNU/Linux 4.4.0-119-generic x86_64)	→ đemos
<pre>* Documentation: https://help.ubuntu.com * Management: https://landscape.canonical.com * Support: https://ubuntu.com/advantage Last login: Wed Jul 25 14:54:15 2018 from 10.0.2.2 * ~ </pre>	



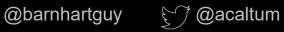
→ demos ssh user@localhost -p 60000 Nelcome to Ubuntu 16.04.4 LTS (GNU/Linux 4.4.0-119-generic x86_64)	→ demos
* Documentation: https://help.ubuntu.com	
* Management: https://landscape.canonical.com * Support: https://ubuntu.com/advantage	
Last login: Wed Jul 25 14:54:15 2018 from 10.0.2.2 ~ hostname wm	

@barnhartguy 💓 @acaltum

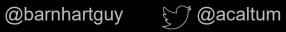
demos ssh user@localhost -p 60000 Welcome to Ubuntu 16.04.4 LTS (GNU/Linux 4.4.0-119-generic x86_64)	-> 	demos
<pre>* Documentation: https://help.ubuntu.com * Management: https://landscape.canonical.com * Support: https://ubuntu.com/advantage Last login: Wed Jul 25 14:54:15 2018 from 10.0.2.2 * ~ hostname</pre>	- 	
<pre>wm * cd /home/user/for_presentation/jarvis_demo/runner/ * runner</pre>	 	



🕈 demos ssh user@localhost -p 60000	+	demos
Welcome to Ubuntu 16.04.4 LTS (GNU/Linux 4.4.0-119-generic x86_64)		
<pre>* Documentation: https://help.ubuntu.com * Management: https://landscape.canonical.com * Support: https://ubuntu.com/advantage Last login: Wed Jul 25 14:54:15 2018 from 10.0.2.2 > ~ hostname wm</pre>		
~ cd /home/user/for_presentation/jarvis_demo/runner/	İ.	
runner ./classification-d/exploits/good-oneaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa	1	
aaaa.rmt.SHELL.cafe.static.full		
	I -	
	I	
	i	
	i	
	1	
	1	
	i	



	- Annon
demos ssh user@localhost -p 60000	→ demos
Welcome to Ubuntu 16.04.4 LTS (GNU/Linux 4.4.0-119-generic x86_64)	
* Documentation: https://help.ubuntu.com	
* Management: https://landscape.canonical.com	
* Support: https://ubuntu.com/advantage	
Last login: Wed Jul 25 14:54:15 2018 from 10.0.2.2	
🔶 ~ hostname	
vm	
<pre>~ cd /home/user/for_presentation/jarvis_demo/runner/</pre>	
<pre>runner ./classification-d/exploits/good-oneaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa</pre>	
aaaa.rmt.SHELL.cafe.static.full	
Prediction for/exploits/good-oneaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa	
a.rmt.SHELL.cafe.static.full	
A.IMC.SHELL.Cale.Static.luit	
Ĭ	



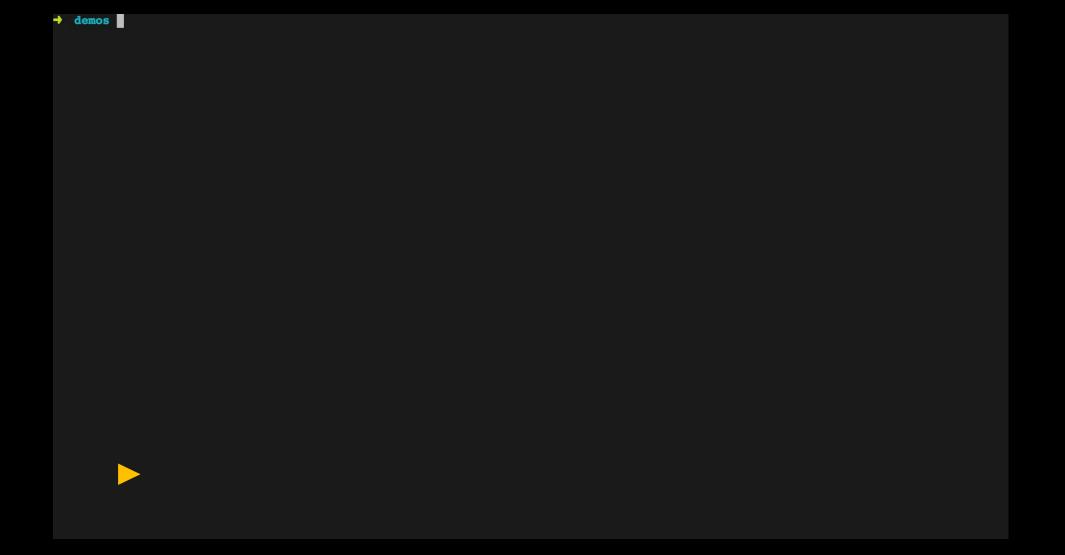
demos ssh user@localhost -p 60000 🔶 demos Welcome to Ubuntu 16.04.4 LTS (GNU/Linux 4.4.0-119-generic x86 64) * Documentation: https://help.ubuntu.com https://landscape.canonical.com * Management: * Support: https://ubuntu.com/advantage Last login: Wed Jul 25 14:54:15 2018 from 10.0.2.2 🔶 🗝 hostname M ~ cd /home/user/for presentation/jarvis demo/runner/ aaaa.rmt.SHELL.cafe.static.full ----- Prediction for ../exploits/good-oneaaaaaaaaaaaaaaaaaaaaaaaa a.rmt.SHELL.cafe.static.full ------

DEMO TIME

Model Tampering

@barnhartguy

@acaltum



@barnhartguy 💓 @acaltum

DEMO TIME IP Theft



NO DEMO

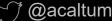
You actually saw this already 🙂

@barnhartguy § @acaltum



SO MAYBE RCÉ IS KING AFTER ALL?

@barnhartguy 🗧 🗧

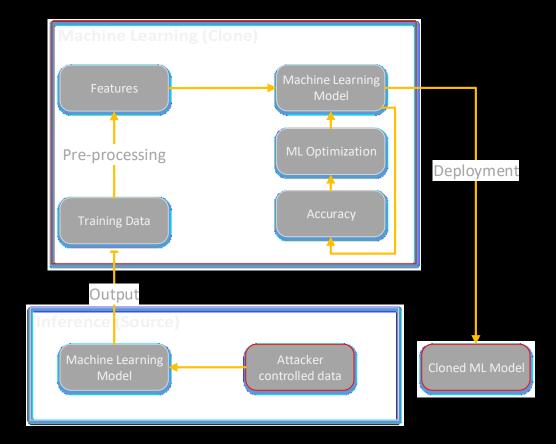


AND IF YOU DON'T HAVE AN RCE?

Let's go after the algorithms!



ATTACK OF THE CLONES



@barnhartguy 💬 @acaltum

CLONING

White box – full access to model and training data (Easy)

CLONING

White box – full access to model and training data (Easy)

Grey box – no access to model and training data, but educated guesses help (highly succesful)

@barnhartguy 🛛 💬 @acaltum

CLONING

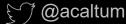
White box – full access to model and training data (Easy)

Grey box – no access to model and training data, but educated guesses help (highly succesful)

Black box – no idea, exporation via probing, build a map (similar to a Reverse Engineering effort, research WIP)

@barnhartguy 🔬 🖓 @acaltum

WHAT IF THE ATTACKER HAS ACCESS TO THE TRAINING DATA?





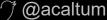
Inject crafted data to the training set with label of your choice



Inject crafted data to the training set with label of your choice

No known way to detect (or reverse engineer)!

This is still an open question academically





Learning is encoded in the matrix You cannot reverse the matrix to learn "discover" made it learn specific things Which means, there is no way to tell what it actually learned

This is also useful in other contexts...

@barnhartguy 👌 🖓 @acaltum



DeepLocker

Concealing Targeted Attacks with AI Locksmithing

Dhilung Kirat, Jiyong Jang, Marc Ph. Stoecklin IBM **Research**

MISS-PREDICTIONS (ADVERSARIAL ATTACKS)

You can manipulate the output with a crafted input ;-)

Remember, the system optimizes for the "strongest signal"

@barnhartguy 🛛 💬 @acaltum

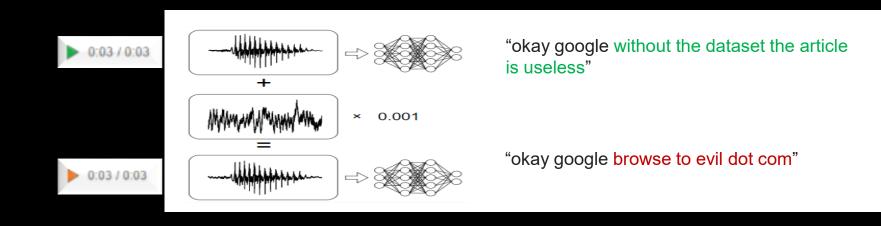
TURTLE OR A RIFLE?



https://www.labsix.org/physical-objects-that-fool-neural-nets/

@acaltum

ADVERSARIAL AUDIO



https://nicholas.carlini.com/code/audio_adversarial_examples/

@barnhartguy

🦪 @acaltum

EVADING NEXT GENERATION AV USING AI

Static machine learning model trained on millions of samples



- Simple structural changes that don't change behavior
 - unpack

@acaltum

@barnhartguy

- '.text' -> '.foo' (remains valid entry point)
- create '.text' and populate with '.text from calc.exe'



https://media.defcon.org/DEF%20CON%2025/DEF%20CON%2025%20presentations/DEFCON-25-Hyrum-Anderson-Evading-Next-Gen-AV-Using-AI.pdf https://www.youtube.com/watch?v=FGCle6T0Jpc

WHAT ABOUT PRIVACY?



PRIVACY LEAKS? NOT YET, BUT SOON...





Risk: 7.4% Risk: 35.3%

@barnhartguy

∽ @acalt<u>um</u>

PRIVACY LEAKS? NOT YET, BUT SOON...





Risk: 96.2%

@barnhartguy

🏹 @acaltum

KEY TAKEAWAYS - RESEARCHERS

We need a better trust model for ML and a lot more research! More focus should be on the infrastructure

The interfaces between the stages are very vulnerable (hint hint)

@barnhartguy 💓 @acaltum

KEY TAKEAWAYS - ATTACKERS

This is a ripe field for attacks High value targets Huge dependency stack

KEY TAKEAWAYS - DEFENDERS

Machine Learning needs sanitation and security controls too Use Machine Learning models from untrusted sources with caution Validate the data you rely on - does it include negative cases? abnormal cases?

@barnhartguy 🛛 💬 @acaltum

There is no Al. It's just someone else's code.

ACKNOWLEDGMENTS

Omer Agmon Adi Oren **Denis Klimov** Raizy Kellerman Adel Fuchs Sapir Hamawie **Oleg Pogorelik**

@barnhartguy 🔰 @acaltum

REFERENCES

PassGAN: A Deep Learning Approach for Password Guessing Adversarial examples for evaluating reading comprehension systems Universal adversarial perturbations, Video Awesome-AI-Security An introduction to Artificial Intelligence When DNNs go wrong - adversarial examples and what we can learn from them Machine Learning in the Presence of Adversaries Pattern Recognition and Applications Lab: Adversarial Machine Learning Deep neural networks are easily fooled. Practical black-box attacks against deep learning systems using adversarial examples, Adversarial examples in the physical world, Explaining and harnessing adversarial examples Distillation as a defense to adversarial perturbations against deep neural <u>networks</u>, Vulnerability of deep reinforcement learning to policy induction attacks

Adversarial attacks on neural network policies, Attacking Machine Learning with Adversarial Examples Intriguing properties of neural networks Robust Physical-World Attacks on Deep Learning Models Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face **Recognition** Towards the Science of Security and Privacy in Machine Learning cleverhans source code **Clever Hans** Awesome - Most Cited Deep Learning Papers 8 Lessons from 20 Years of Hype Cycles DEF CON 25 (2017) - Weaponizing Machine Learning - Petro, Morris Evading next-gen AV using A.I. For better machine-based malware analysis, add a slice of LIME BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply <u>Chain</u>

HOW TO PROCEED?





ANY QUESTIONS?

@barnhartguy





@barnhartguy 🛛 🖓 @acaltum