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^{*} Al models were harmed during the making of this presentation

* We hope

\$ ID



Guy Barnhart-Magen

@barnhartguy

BSidesTLV Co-founder and CTF Lead

@acaltum

Ezra Caltum

@acaltum

BSidesTLV Co-Founder

DC9723 Lead



@barnhartguy

BOSIDES LIQUES 9723

BUILDING ON THE SHOULDERS OF GIANTS



@barnhartguy

@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

WHAT CAN YOU EXPECT?

What are we going to talk about

What you should be paying attention to

What we are **not** going to talk about

CLEVER HANS

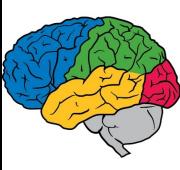


https://github.com/tensorflow/cleverhans

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

"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

@acaltum



https://upload.wikimedia.org/wikipedia/commons/b/ba/Alice_par_John_Tenniel_30.png

@barnhartguy

ARTIFICIAL INTELLIGENCE?

Machine Learning

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

Deep Learning

@barnhartguy

Study many images Identify the flamingo, hedgehog, etc.

@acaltum



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

Artificial intelligence

Is she hugging the flamingo, or playing cricket?

Is she happy, sad?



@barnhartguy

@acaltum

EVERYBODY EXCHANGES "AI" AND "ML"

So do I

Sorry

"INTELLIGENT" SYSTEM

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



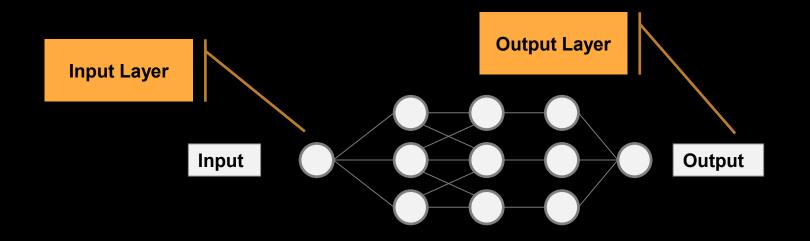
https://www.reactiongifs.us/wp-content/uploads/2015/02/do_the_robot_futurama.gif

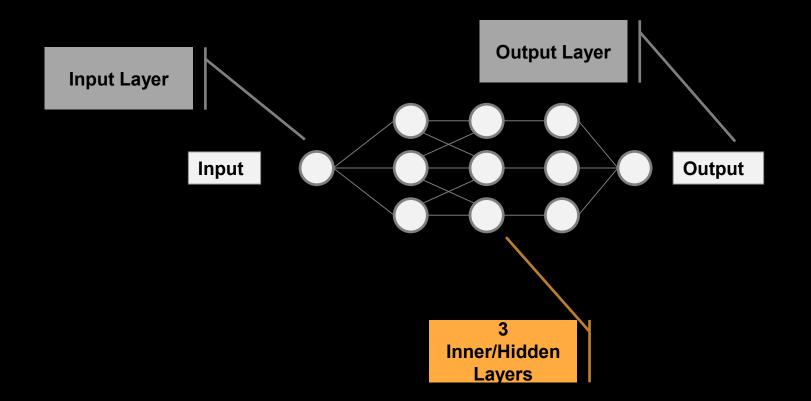
@barnhartguy

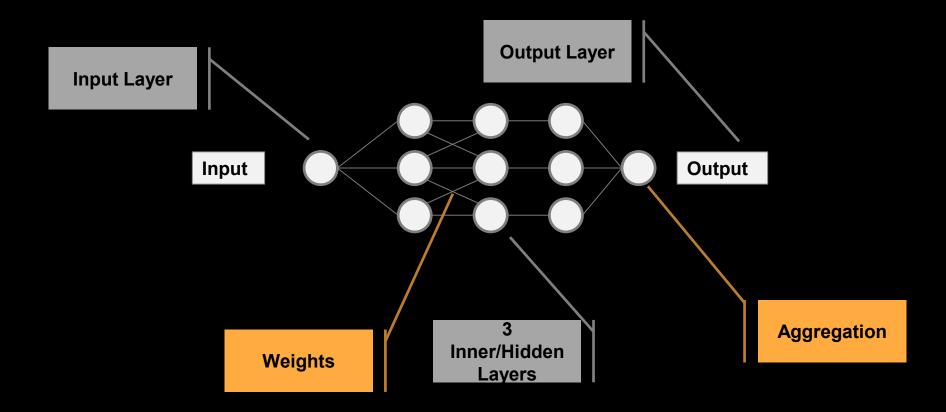
@acaltum

MACHINE LEARNING 101









Input Output

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

LINEAR ALGEBRA, ANYONE?

Definition [edit]

If **A** is an $n \times m$ matrix and **B** is an $m \times p$ matrix,

$$\mathbf{A} = egin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \ a_{21} & a_{22} & \cdots & a_{2m} \ dots & dots & \ddots & dots \ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix}, \quad \mathbf{B} = egin{pmatrix} b_{11} & b_{12} & \cdots & b_{1p} \ b_{21} & b_{22} & \cdots & b_{2p} \ dots & dots$$

the matrix product C = AB (denoted without multiplication signs or dots) is defined to be the $n \times p$ matrix

$$\mathbf{C} = egin{pmatrix} c_{11} & c_{12} & \cdots & c_{1p} \ c_{21} & c_{22} & \cdots & c_{2p} \ dots & dots & \ddots & dots \ c_{n1} & c_{n2} & \cdots & c_{np} \end{pmatrix}$$

such that

@acaltum

<u>@b</u>arnhartguy

$$c_{ij} = a_{i1}b_{1j} + \dots + a_{im}b_{mj} = \sum_{k=1}^m a_{ik}b_{kj},$$
 for $i=1,...,n$ and $j=1,...,p.$

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

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; };

TOO MUCH VOODOO!



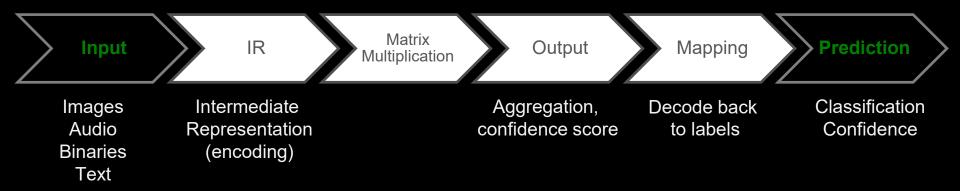
Images Audio Binaries Text

51~f*9\>71rB



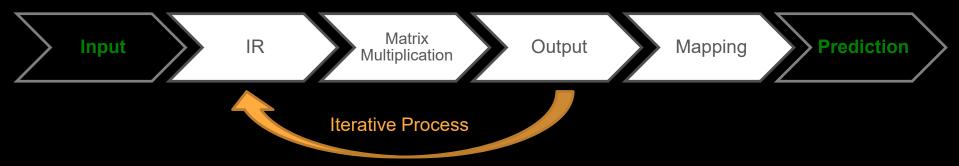
Classification Confidence

NOTTOO MUCH VOODOO!



FROM TRAINING TO INFERENCE

Training

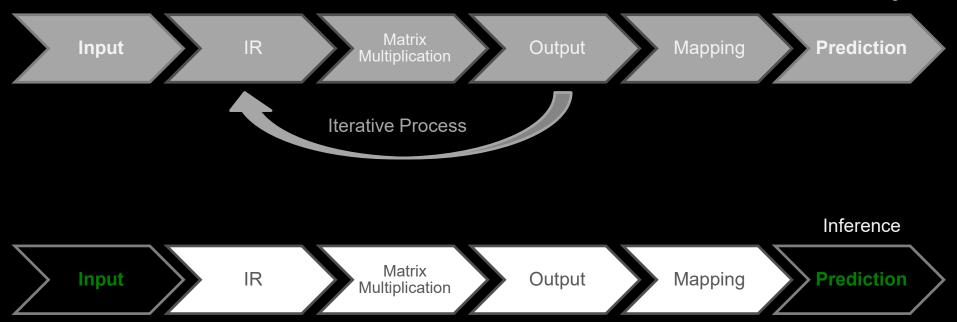


@barnhartguy Les Contraction

@acaltum

FROM TRAINING TO INFERENCE

Training



MODEL != CODE



Code execution flow

ML MODEL

Math operations, transition functions

@barnhartguy 💓 @ac

@acaltum

EXECUTABLE

Code execution flow

Data Structures

ML MODEL

Math operations, transition functions

Intermediate Representation

@barnhartguy 💓 @a

@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 E)

@acaltum

\$ HEXDUMP /MODELS/RESNET

0000	0000	0 A	09	52	65	73	4E	65	74	2D	35	30	1A	04	64	61	74	61	A2	06	C9	ResNet-50data
0000	0014	A 8	02	0A	05	63	6F	6E	76	31	12	0B	43	6F	6E	76	6F	6C	75	74	69	conv1Convoluti
0000	0028	6F	6E	1A	04	64	61	74	61	22	05	63	6F	6E	76	31	3A	8C	A6	02	2A	ondata".conv1:*
0000	003C	80	A 6	02	0E	72	E7	3C	64	$\mathbf{F}\mathbf{D}$	94	3C	1E	21	82	3C	BE	99	1E	ЗB	99	r. <d<.!.<;.< td=""></d<.!.<;.<>
0000	0050	26	59	$^{\rm BD}$	30	F8	41	$^{\rm BD}$	63	E7	97	3C	E1	E8	02	3C	07	DD	CB	3C	86	&Y.O.A.c<
0000	0064	22	9D	ЗD	7B	DC	B8	ЗD	E1	F3	9A	BC	7F	85	A2	BD	BB	AF	7B	BC .	A9	".={=
0000	0078	0F	E1	BC	C7	A 9	86	$^{\rm BD}$	8E	C6	A3	3C	9F	32	ЗD	3E	ЗD	9A	EC	3D .	A2	<2=>==.
0000	008C	47	B8	$^{\rm BD}$	7F	D5	C1	$^{\rm BD}$	62	5C	20	BC	30	47	A2	BD	68	EF	04	BE	EA	Gb\ .0Gh
0000	00A0	0 A	C3	ЗD	C3	46	8A	3E	80	96	CB	ЗD	9D	A3	74	BC	F8	E4	BD	3C	11	=.F.>=t<.
0000	00B4	44	3E	$^{\rm BD}$	16	66	3B	BE	5B	C9	1F	BE	30	9D	5B	ЗD	C4	FF	AA	3D	DB	D>f;.[0.[==.
0000	00C8	60	17	ЗD	60	13	2D	ЗD	\mathbf{FC}	32	3C	ЗD	9B	$\mathbf{F}0$	3E	BD	FA	16	02	BE	5A	`.=`=.2<=>Z
0000	00DC	ЗA	AC	BC	\mathbf{FB}	13	80	ЗD	75	ЗA	CD	3C	FE	в3	6F	3C	A5	4D	10	ЗD	95	:=u:. <o<.m.=.< td=""></o<.m.=.<>
0000	00F0	A2	86	3B	58	56	77	BD	7C	A 6	8C	BC	49	A6	3C	3C	C3	01	9D	3B .	AF	;XVw. I.<<;.
0000	0104	4A	C0	3B	\mathbf{FD}	A4	99	3C	C7	09	14	ЗD	82	88	8C	3C	48	9B	\mathbf{BF}	BD	30	J.;<= <h0< td=""></h0<>
0000	0118	28	FE	BD	3E	9C	DB	BC	2B	EΒ	\mathbf{FB}	BB	В4	09	C7	3C	0C	08	19	3E	64	(>+<>d
0000	012C	CC	5C	3E	E5	5C	5B	3C	02	1E	35	BE	в7	CB	\mathbf{FB}	BD	9A	E7	98	BD	64	.\>.\[<5d
0000	0140	27	1C	BE	8F	F4	79	ЗD	в1	7A	D4	3E	BF	27	AA	3E	62	ED	A 6	BD	9B	'y=.z.>.'.>b
0000	0154	2D	58	BE	6C	01	5F	$^{\rm BD}$	2A	в5	82	BE	52	63	87	BE	02	63	57	3E	1F	-X.l*RccW>.
0000	0168	F5	11	3F	8D	F2	8D	3E	0A	EF	D8	BC	E1	DO	BC	3C	15	1F	36	BE	3F	?><6.?
0000	017C	D5	D2	BE	DC	72	89	BE	71	40	3E	3E	D3	48	80	3E	\mathbf{FD}	5D	C5	ЗD	9A	rq@>>.H.>.].=.
0000	0190	17	C8	ЗD	51	88	1F	ЗD	C4	F3	44	BE	7D	A 0	95	BE	1E	C7	18	BD	F1	=Q=D.}
0000	01A4	E8	CB	ЗD	D8	ЗF	CD	ЗD	FE	C2	7A	ЗD	DB	EE	81	ЗD	C8	CE	в6	BC	3E	=.?.=z=>
0000	01B8	95	2E	BE	76	\mathbf{BF}	в0	$^{\rm BD}$	FE	14	F3	3B	27	70	E1	3C	A2	F7	1F	BB	10	v;'p.<
0000	01CC	05	9C	3C	7C	8D	в2	3C	80				AA					0B				< "
	ResNe	t-50)-mc	ode]	L.ca	ffer	node	e1			-0x1	1DF	/0x6	1B73	3BD-							



But... How do you take a look at the code?



But... How do you take a look at the code?

We could still be better at traditional software code reviews

VALIDATION

But... How do you take a look at the code?

We could still be better at traditional software code reviews

What is your code here exactly?

VALIDATION

But... How do you take a look at the code?

We could still be better at traditional software code reviews

What is your code here exactly?

How do you understand/review the matrix?

FUN FACTS!

The model (matrices) can be GB in size Machine learning predicts the future based on the past The algorithm is designed to optimize for the "strongest signal"

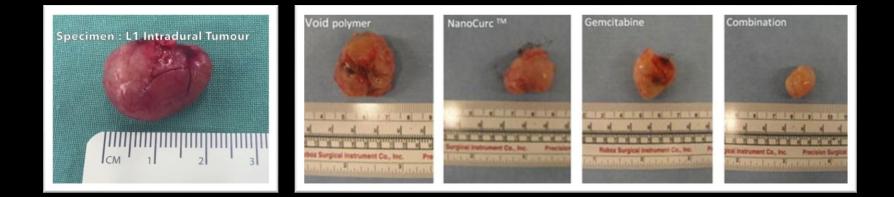
Bias is a part of the system learning process

BIAS - SOLVING THE WRONG PROBLEM

That said, they learned that the algorithm could be fooled in unexpected ways. "For example, if we had a ruler in the image, the algorithm was much more likely to call it malignant," he noted. "Why is that? Because on average, in our data-set, lesions with rulers were being measured and monitored by dermatologists, and were more likely to be malignant. The algorithm is looking at the whole image and will take whatever clues it can find. It can be biased by features like the ruler, and you won't know it." Another image that might trip up the algorithm would be that of an unusual combination like a benign nevus colliding with a seborrheic keratosis, which could closely mimic a melanoma, "but you may not know that until you've collected a lot of those images."

BIAS - SOLVING THE WRONG PROBLEM

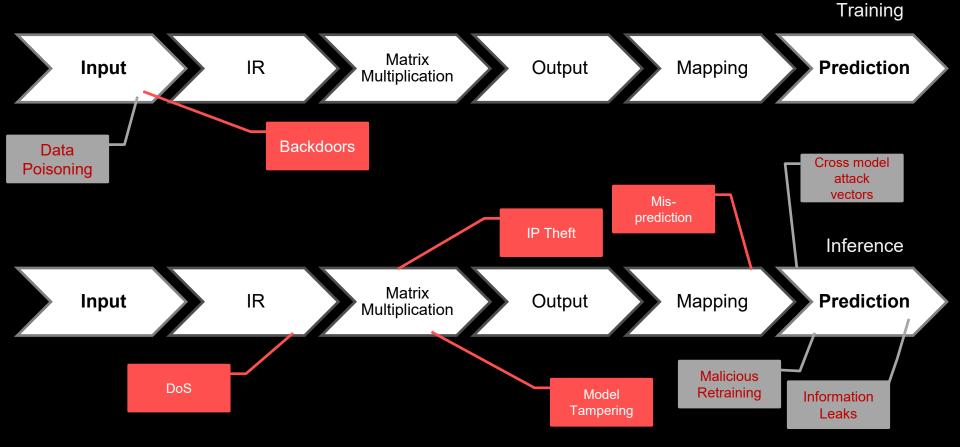
That said, they learned that the algorithm could be fooled in unexpected ways. "For example, if we had a ruler in the image, the algorithm was much more likely to call it malignant," he noted. "Why is that? Because on average, in our data-set, lesions with rulers were being measured and monitored by dermatologists, and were more likely to be malignant. The algorithm is looking at the whole image and will take whatever clues it can find. It can be biased by features like the ruler, and you won't know it." Another image that might trip up the algorithm would be that of an unusual combination like a benign nevus colliding with a seborrheic keratosis, which could closely mimic a melanoma, "but you may not know that until you've collected a lot of those images."



SCORING

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

FROM TRAINING TO INFERENCE



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

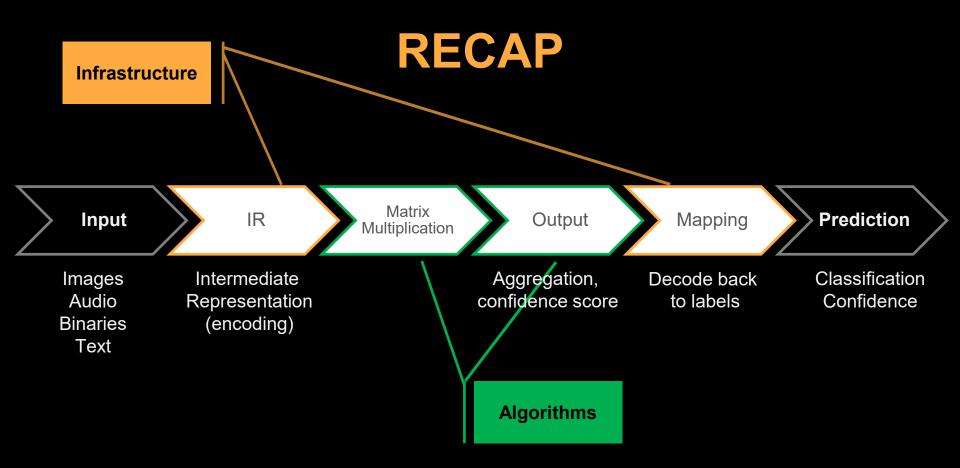
What do you need to know?

What areas should you target?

What do you need to have access to?

WHERE TO ATTACK?

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





ML needs to convert the input into a matrix



ML needs to convert the input into a matrix

Parsing is hard



ML needs to convert the input into a matrix

Parsing is hard

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

PARSING

ML needs to convert the input into a matrix

Parsing is hard

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

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

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

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...



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



What to focus on?Why focus here?CaffeFull coverageOpenCVLimited coverage

Issues? Extremely slow Medium speed

@barnhartguy

@acaltum



What to focus on? Why focus here? Caffe OpenCV LibXXX

Full coverage Limited coverage Very fast

Issues? Extremely slow Medium speed Unknown code paths



What to focus on?Why focus here?Issues?CaffeFull coverageExtremely slowOpenCVLimited coverageMedium speedLibXXXVery fastUnknown code pathsUpstreamFuzzing not neededPatched? Workable?

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

- 1 IP Theft
- 2 Model Tampering
- 3 DoS
- 4 Backdoors

- 5.9 (Medium) 7.4 (High) 7.5 (High) 3.9 (Low) 7.5 (High)
- 5 Misprediction (Adversarial attacks)

Is Remote Code Execution (RCE) king?

POST EXPLOITATION

Let's try to demonstrate the TOP 5

DEPENDENCIES/EXPLOIT DEMO

Denial of Service

Abusing a memory leak

→ demos ssh user@:	localhost -p 60000	→ demos ssh user@localhost -p 60000									
Welcome to Ubuntu 1	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									
		1									
* Documentation:	https://help.ubuntu.com	* Documentation:	https://help.ubuntu.com								
* Management:	https://landscape.canonical.com	* Management:	https://landscape.canonical.com								
* Support:	https://ubuntu.com/advantage	* Support:	https://ubuntu.com/advantage								
Last login: Wed Ju	L 25 14:48:43 2018 from 10.0.2.2	Last login: Wed Jul	1 25 14:49:29 2018 from 10.0.2.2								
→		1 🔶 🛹 📕									
		1									

→ 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									
* 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	SS	sh user@	localhos	st -p	60000			
Wel	come	to	Ubuntu	16.04.4	LTS	(GNU/Linux	4.4.0-119-generic	x86_64)	1

- * Documentation: https://help.ubuntu.com
- * Management: https://landscape.canonical.com
- * Support: https://ubuntu.com/advantage
- Last login: Wed Jul 25 14:48:43 2018 from 10.0.2.2

→ ~

) 	1 [2 [Mem[Swp[]]]]	Tasks: 22, 6 thr; 1 running Load average: 0.07 0.03 0.00 Uptime: 00:09:49										
. !	PID USER	PRI	NI	VIRT	RES	SHR		CPU%	MEM%	TIME+	Comm				
_ !!	1 root	20	0	37980	5968		S	0.0	0.1	0:02.55	/sbi				
_ !	207	20		35272	3528	3220		0.0	0.0	0:00.08	/lib				
_ !	244	20	0	44772	4244	2976		0.0	0.1	0:00.87	/lib				
<u>!</u>	360	20	0	97M	2460		S	0.0	0.0	0:00.00	/lib				
ļ	335	20	0	97M	2460	<mark>2</mark> 252		0.0	0.0	0:00.02	/lib				
. !	471	20	0	16120	856	0	S	0.0	0.0	0:00.00	/sbi				
ļ	506	20	0	28620	3080		S	0.0	0.0	0:00.02	/lib				
ļ	533	20	0	250M	3416	2744	S	0.0	0.0	0:00.00					
_ !	534	20	0	250M	3416	2744	S	0.0	0.0	0:00.00	/usr				
<u> </u>	535	20	0	250M	3416	2744	S	0.0	0.0	0:00.00	/usr				
1	507	20	0	250M	3416		S	0.0	0.0	0:00.02	/usr				
	511	20	0	42900	3900	<mark>3</mark> 488	S	0.0	0.0	0:00.08	/usr				
	557	20	0	269M	<mark>6</mark> 260	<mark>5</mark> 532	S	0.0	0.1	0:00.01	/usr				
1	565	20	0	269M	<mark>6</mark> 260	<mark>5</mark> 532	S	0.0	0.1	0:00.00	/usr				
	538	20	0	269M	<mark>6</mark> 260	<mark>5</mark> 532	S	0.0	0.1	0:00.04	/usr				
1	546	20	0	29 008	<mark>2</mark> 920	<mark>2</mark> 648	S	0.0	0.0	0:00.00	/usr				
1	563	20	0	19 472	<mark>2</mark> 288	2064	S	0.0	0.0	0:00.01	/usr				
1	665	20	0	<mark>65</mark> 508	6040	<mark>5</mark> 332	S	0.0	0.1	0:00.00	/usr				
	679	20	0	15 936	1 792	1 664	S	0.0	0.0	0:00.00	/sbi				
<u> </u>	909	20	0	<mark>92</mark> 800	<mark>6</mark> 940	6000	S	0.0	0.1	0:00.00	sshd				
1	911 user	20	0	45 192	<mark>4</mark> 988	<mark>4</mark> 156	S	0.0	0.1	0:00.01	/lib				
1	912 user	20	0	<mark>61</mark> 432	<mark>2</mark> 128	0	S	0.0	0.0	0:00.00	(sd-				
1	935 user	20	0	92 800	<mark>3</mark> 328	<mark>2</mark> 392	S	0.0	0.0	0:00.00	sshd				
	936 user	20	0	44316	5 448	<mark>3</mark> 852	S	0.0	0.1	0:00.17	-zsh				
	965	20	0	<mark>92</mark> 800	<mark>6</mark> 748	<mark>5</mark> 812	S	0.0	0.1	0:00.00	sshd				
	985 user	20	0	<mark>92</mark> 800	<mark>3</mark> 304	<mark>2</mark> 372	S	0.0	0.0	0:00.00	sshd				
1	F <mark>1</mark> Help <mark>F2</mark> Setu	ıp <mark>F3</mark> Sea	arch	F4 <mark>Filt</mark>	er <mark>F5</mark> Tr	ee F	6 <mark>Sc</mark>	ortBy	7 <mark>Nice</mark>	- <mark>F8</mark> Nice	+F9K				

demos	s ssr	1 user@	localhos	st -p	60000				
Welcome	to U	Jbuntu	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 14:48:43 2018 from 10.0.2.2
- ~ cd /home/user/for_presentation/jarvis_demo/runner/

	1 [2 [Mem[Swp[]]]]	Loa Upt	unning 03 0.00					
	PID USER	PRI	NI	VIRT	RES	SHR		CPU%	MEM%	TIME+	Comm
- !	1011 user	20	0	25924	3756	3196	-	0.7	0.0	0:00.06	htop
- !	1	20	0	37980	5968	3 952		0.0	0.1	0:02.55	/sbi
ļ	207	20	0	35272	3528		S	0.0	0.0	0:00.08	/lib
_ !	244	20	0	44772	4244	<mark>2</mark> 976	S	0.0	0.1	0:00.87	/lib
ļ	360	20	0	97M	2460		S	0.0	0.0	0:00.00	/lib
	335	20	0	97M	2460		S	0.0	0.0	0:00.02	/lib
ļ	471	20	0	16 120	856	0	S	0.0	0.0	0:00.00	
	506	20	0	28620	<mark>3</mark> 080	<mark>2</mark> 760	S	0.0	0.0	0:00.02	/lib
ļ	533	20	0	250M	3416	2744	S	0.0	0.0	0:00.00	/usr
	534	20	0	250M	<mark>3</mark> 416	2744	S	0.0	0.0	0:00.00	/usr
	535	20	0	250M	<mark>3</mark> 416	2744	S	0.0	0.0	0:00.00	
	507	20	0	250M	<mark>3</mark> 416	2744	S	0.0	0.0	0:00.02	/usr
	511	20	0	42900	3 900	<mark>3</mark> 488	S	0.0	0.0	0:00.08	/usr
	557	20	0	269M	<mark>6</mark> 260	<mark>5</mark> 532	S	0.0	0.1	0:00.01	/usr
	565	20	0	269M	<mark>6</mark> 260	<mark>5</mark> 532	S	0.0	0.1	0:00.00	/usr
	538	20	0	269M	<mark>6</mark> 260	<mark>5</mark> 532	S	0.0	0.1	0:00.04	/usr
	546	20	0	29 008	<mark>2</mark> 920	<mark>2</mark> 648	S	0.0	0.0	0:00.00	/usr
	563	20	0	19 472	<mark>2</mark> 288	2064	S	0.0	0.0	0:00.01	/usr
	665	20	0	<mark>65</mark> 508	6040	<mark>5</mark> 332	S	0.0	0.1	0:00.00	/usr
	679	20	0	15 936	1 792	1 664	S	0.0	0.0	0:00.00	/sbi
	909	20	0	<mark>92</mark> 800	<mark>6</mark> 940	6000	S	0.0	0.1	0:00.00	sshd
	911 user	20	0	<mark>45</mark> 192	<mark>4</mark> 988	<mark>4</mark> 156	S	0.0	0.1	0:00.01	/lib
	912 user	20	0	<mark>61</mark> 432	<mark>2</mark> 128	0	S	0.0	0.0	0:00.00	(sd-
I	935 user	20	0	<mark>92</mark> 800	<mark>3</mark> 328	<mark>2</mark> 392	S	0.0	0.0	0:00.00	sshd
I	936 user	20	0	44316	5 448	<mark>3</mark> 852	S	0.0	0.1	0:00.17	-zsh
	965	20	0	<mark>92</mark> 800	<mark>6</mark> 748	5 812	S	0.0	0.1	0:00.00	sshd
	F1Help F2Setu	ıp <mark>F3</mark> Sea	arch	<mark>F4</mark> Filt	er <mark>F5</mark> Tr	ee F	S			e - <mark>F8</mark> Nice	+F9K

→ demos ssh user@localhost -p 60000										
Welcome to Ubuntu 16.04.4 LTS (GNU/Linux 4.4.0-119-generic x86_64)	1 []					running	
	2 []					02 0.00	
* Documentation: https://help.ubuntu.com	Mem[]	Upt	ime: 0	0:10:2	3		
* 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	PID USER	PRI	NI	VIRT	RES	SHR	S CPU%	MEM%		Comn
<pre>~ cd /home/user/for_presentation/jarvis_demo/runner/</pre>	1011 user	20		25924	3756	3196		0.0	0:00.08	-
runner ./classification-d/exploits/ram_and_cpu_dos.bmp	1	20		37980	<mark>5</mark> 968	<mark>3</mark> 952			0:02.55	/sbi
	207	20	0	35272	<mark>3</mark> 528	<mark>3</mark> 220	s 0.0	0.0	0:00.08	/lik
	244	20	0	44772	4244	<mark>2</mark> 976	s 0.0	0.1	0:00.87	/lik
	360	20	0	97M	2 460	<mark>2</mark> 252		0.0	0:00.00	/lik
	335	20	0	97M	2 460	<mark>2</mark> 252	s 0.0	0.0	0:00.02	/lik
	471	20		16 120	856	0		0.0	0:00.00	/sbi
	506	20	0	28620	3080	<mark>2</mark> 760			0:00.02	
	533	20	0	250M	3416	<mark>2</mark> 744			0:00.00	
	534	20	0	250M	3416	2744	s 0.0	0.0	0:00.00	/usr
	535	20	0	250M	3416	2744		0.0	0:00.00	/usr
	507	20	0	250M	3 416	2744			0:00.02	
	511	20	0	42900	3 900	<mark>3</mark> 488			0:00.08	
	557	20	0	269M	<mark>6</mark> 260	<mark>5</mark> 532			0:00.01	
	565	20	0	269M	<mark>6</mark> 260	<mark>5</mark> 532			0:00.00	
	538	20	0	269M	<mark>6</mark> 260	<mark>5</mark> 532			0:00.04	
	546	20		29008	<mark>2</mark> 920	<mark>2</mark> 648		0.0	0:00.00	/usr
	563	20		19472	<mark>2</mark> 288	2064			0:00.01	
	665	20		65508	6040	5332			0:00.00	
	679	20		15936	1792	1 664			0:00.00	
	909	20		92800	<mark>6</mark> 940	6000			0:00.00	
	911 user	20		45192	<mark>4</mark> 988	4156			0:00.01	
	912 user	20		61432	2128	0			0:00.00	
	935 user	20		92800	<mark>3</mark> 328	<mark>2</mark> 392			0:00.00	
	936 user	20		44316	<mark>5</mark> 452	3852			0:00.17	
	965	20		92800		<mark>5</mark> 812			0:00.00	
	F1Help F2Setu	ip F3Se	arch	F4Filt	er <mark>F5</mark> Tr	ee F6	SortBy	F7Nice	e -F8Nice	+F9F

@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 [1	Tas	ks: 23,	6 th	; 1 r	unning	
	2 [1	Loa	d avera	ge: 0.	.11 0.	04 0.01	
* Documentation: https://help.ubuntu.com	Mem[j	Upt	ime: 00	:10:20	5		
* 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 VIRT	RES	SHR S	CPU%	MEM%	TIME+ Co	omn
<pre>- cd /home/user/for_presentation/jarvis demo/runner/</pre>	1016 user	20	0 354M	221M	19280 D		2.8	0:00.98 .,	/cl
runner ./classification-d/exploits/ram and cpu dos.bmp	1011 user	20	0 25924	3756	3196 R	0.8	0.0	0:00.09 ht	top
Prediction for/exploits/ram_and_cpu_dos.bmp	1	20	0 37980	5 968	3952 S	0.0	0.1	0:02.55 /s	sbi
-	207	20	0 35272	<mark>3</mark> 528	3220 S	0.0	0.0	0:00.08 /	lib
	244	20	0 44772	4244	2976 S	0.0	0.1	0:00.87 /	lib
	360	20	0 97M	2 460	2252 S	0.0	0.0	0:00.00 /	lib
	335	20	0 97M	2 460	2252 S	0.0	0.0	0:00.02 /	lib
	471	20	0 <mark>16</mark> 120	856	0 S	0.0	0.0	0:00.00 /s	sbi
	506	20	0 <mark>28</mark> 620	<mark>3</mark> 080	2760 S	0.0	0.0	0:00.02 /	lib
	533	20	0 250M	<mark>3</mark> 416	2744 S	0.0	0.0	0:00.00 /1	usr
	534	20	0 250M	<mark>3</mark> 416	2744 S	0.0	0.0	0:00.00 /1	usr
	535	20	0 250M	<mark>3</mark> 416	2744 S	0.0	0.0	0:00.00 /1	usr
	507	20	0 250M	<mark>3</mark> 416	2744 S	0.0	0.0	0:00.02 /1	usr
	511	20	0 42900	3 900	3488 S	0.0	0.0	0:00.08 /1	usr
	557	20	0 269M	<mark>6</mark> 260	5532 S	0.0	0.1	0:00.01 /1	usr
	565	20	0 269M	<mark>6</mark> 260	5532 S	0.0	0.1	0:00.00 /1	
	538	20	0 269M	<mark>6</mark> 260	5532 S	0.0	0.1	0:00.04 /1	usr
	546	20	0 29008	<mark>2</mark> 920	2648 S		0.0	0:00.00 /1	usr
	563	20	0 19472	<mark>2</mark> 288	2064 S		0.0	0:00.01 /1	
	665	20	0 <mark>65</mark> 508	6040	5332 S		0.1	0:00.00 /1	
	679	20	0 15 936	1 792	1664 S		0.0	0:00.00 /:	
	909	20	0 92800	<mark>6</mark> 940	6000 S		0.1	0:00.00 s	
	911 user	20	0 45192	<mark>4</mark> 988	4 156 S		0.1	0:00.01 /:	
	912 user	20	0 61432	<mark>2</mark> 128	0 S		0.0	0:00.00 (8	
	935 user	20	0 92800	<mark>3</mark> 328	2392 S		0.0	0:00.00 sa	
	936 user	20		5452	3852 S		0.1	0:00.17 -:	
	F1Help F2Setup	F3Se	arch <mark>F4</mark> Filt	er <mark>F5</mark> Tr	ee <mark>F6</mark> S	ortBy	7 <mark>Nice</mark>	-F8Nice +	F9 <mark>K</mark>

demos ssh user@localhost -p 60000										
Welcome to Ubuntu 16.04.4 LTS (GNU/Linux 4.4.0-119-generic x86_64)	1 [100 .0%		sks: 23				
a summer of the second states of the state of the second	2 [000]					34 0.12	
* Documentation: https://help.ubuntu.com		5	•98G	7.	Upt	time: 0	0:12:2	0		
* Management: https://landscape.canonical.com	Swp[1						
* Support: https://ubuntu.com/advantage	DID HG		NTT	NTDE	DEG	QUID		MTDAG	TITAT	Germ
Last login: Wed Jul 25 14:48:43 2018 from 10.0.2.2	PID US			VIRT	RES		S CPU%		TIME+	Com
<pre></pre>	1016 us					20432			1:55.37	
runner ./classification-d/exploits/ram_and_cpu_dos.bmp	1011 us			25924	3756	3196			0:00.35	-
Prediction for/exploits/ram_and_cpu_dos.bmp	563	20		19472	2288	2064			0:00.02	
	1	20		37980 92800	5968 3304	3952			0:02.56	
	985 us 207	er 20 20		35272	3304 3528	2372 3220		0.0 0.0	0:00.02	
	244	20		44772	4244	2 976			0:00.08	
	360	20	0	97M	2 460	2 976 2 252			0:00.00	
	335	20	0	97M	2 400 2 460	2 252			0:00.02	
	471	20		16120	856	2 232 0		0.0	0:00.02	
	506	20		28620	3080	2 760		0.0	0:00.02	
	533	20	0		3 416	2744		0.0	0:00.02	
	534	20	0	250M	3416	2744		0.0	0:00.00	
	535	20	0	250M	3416			0.0	0:00.00	
	507	20	0	250M	3416			0.0	0:00.02	
	511	20		42900	3900	3488		0.0	0:00.08	
	557	20	0	269M	6260	5532			0:00.01	
	565	20	0	269M	6260	5532			0:00.00	
	538	20	0		6260	5532			0:00.04	
	546	20	0	29008	<mark>2</mark> 920	<mark>2</mark> 648		0.0	0:00.00	
	665	20	0	65508	<mark>6</mark> 040	5 332	s 0.0	0.1	0:00.00	/usr
	679	20	0	15 936	1 792	1 664	s 0.0	0.0	0:00.00	/sbi
	909	20	0	92800	<mark>6</mark> 940	<mark>6</mark> 000	s 0.0	0.1	0:00.00	sshc
	911 us	er 20	0	45 192	<mark>4</mark> 988	<mark>4</mark> 156	s 0.0	0.1	0:00.01	/lik
	912 us	er 20	0	<mark>61</mark> 432	<mark>2</mark> 128	0	s 0.0	0.0	0:00.00	(sd-
	935 us	er 20	0	<mark>92</mark> 800	<mark>3</mark> 328	<mark>2</mark> 392	s 0.0	0.0	0:00.00	sshc
	F1Help	F2 <mark>Setup</mark> F3 <mark>Setup</mark> F3	earcl	n <mark>F4</mark> Filt	er <mark>F5</mark> Tr	cee <mark>F6</mark>	SortBy	F7 <mark>Nice</mark>	-F8Nice	+F9 <mark>F</mark>

DENIAL OF SERVICE

Business impact: Failing Services, downtime, costs

DEPENDENCIES/EXPLOIT DEMO

Remote Code Execution

Abusing memory corruption (via heap exploitation)

demos ssh user@localhost -p 60000	->
Nelcome to Ubuntu 16.04.4 LTS (GNU/Linux 4.4.0-119-generic x86_64)	1
	1
* Documentation: https://help.ubuntu.com	1
* Management: https://landscape.canonical.com	1
* Support: https://ubuntu.com/advantage	1
Last login: Wed Jul 25 14:54:15 2018 from 10.0.2.2	1
~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~	1
	1
	1

demos

demos ssh user@localhost -p 60000
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
~

→ demos

demos ssh user@localhost -p 60000	→
Welcome to Ubuntu 16.04.4 LTS (GNU/Linux 4.4.0-119-generic x86_64)	
	ļ
* Documentation: https://help.ubuntu.com	ļ.
* Management: https://landscape.canonical.com	ļ.
* Support: https://ubuntu.com/advantage	ļ.
Last login: Wed Jul 25 14:54:15 2018 from 10.0.2.2	!
~ hostname	
	1
<pre></pre>	1
	1
	1
	i
	i
	i
	l I
	1

demos

demos ssh user@localhost -p 60000	->
Nelcome to Ubuntu 16.04.4 LTS (GNU/Linux 4.4.0-119-generic x86_64)	
	I
* Documentation: https://help.ubuntu.com	
* Management: https://landscape.canonical.com	I
* Support: https://ubuntu.com/advantage	
Last login: Wed Jul 25 14:54:15 2018 from 10.0.2.2	
🔶 🗠 hostname	1
vm	1
<pre>cd /home/user/for_presentation/jarvis_demo/runner/</pre>	
runner ./classification-d/exploits/good-oneaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa	
aaaa.rmt.SHELL.cafe.static.full	
Prediction for/exploits/good-oneaaaaaaaaaaaaaaaaaaaaaaa	
a.rmt.SHELL.cafe.static.full	
The second s	
A	I.
	T

demos

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

MODEL TAMPERING

Business Impact: Change the behavior of the model

Post RCE – remote file system access

🔶 demos 📕





Business Impact: someone steals your model (vested NRE) directly

Post RCE - remote file system access



SO MAYBE RCE IS KING AFTER ALL?

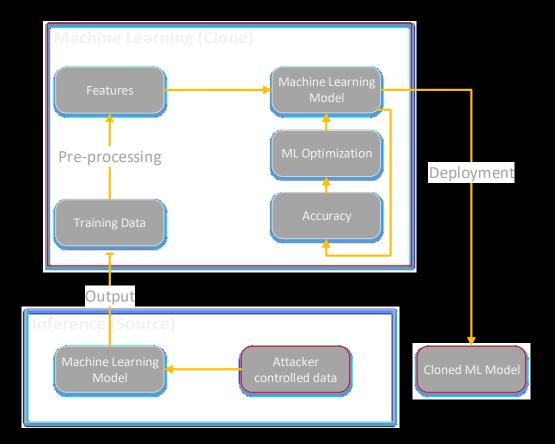
@barnhartguy 🖓 @ao

@acaltum

AND IF YOU DON'T HAVE AN RCE?

Let's go after the algorithms!

ATTACK OF THE CLONES



@barnhartguy 🕤 @

@acaltum



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

@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)

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)

WHAT IF THE ATTACKER HAS ACCESS TO THE TRAINING DATA?

BACKDOORS

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

No known way to detect!

This is still an open question academically

MISS-PREDICTIONS (ADVERSARIAL ATTACKS)

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

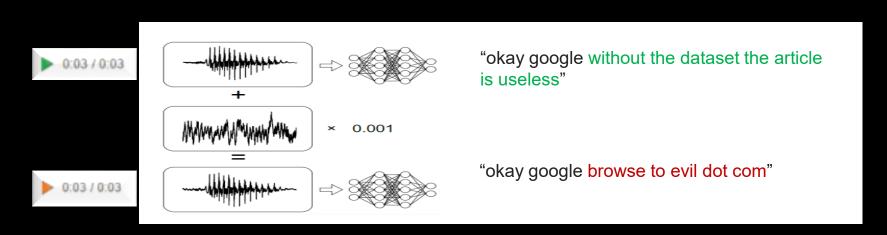
Remember, the system optimizes for the "strongest signal"

TURTLE OR A RIFLE?



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

ADVERSARIAL AUDIO



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

EVADING NEXT GENERATION AV USING AI

Static machine learning model trained on millions of samples



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

PRIVACY LEAKS? NOT YET, BUT SOON...





Risk: 96.2%

FOOLING FACIAL RECOGNITION



FACIAL RECOGNITION



Figure 5: The eyeglass frames used by S_C for dodging recognition against DNN_B .

https://www.ece.cmu.edu/~lbauer/papers/2016/ccs2016-face-recognition.pdf

FACIAL RECOGNITION

@acaltum

@barnhartguy

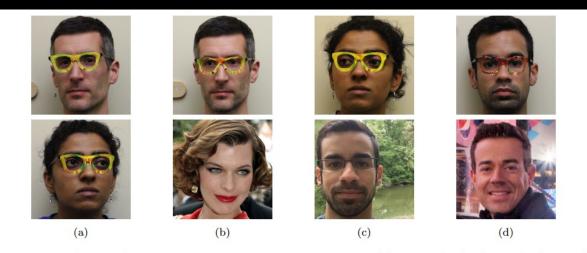


Figure 4: Examples of successful impersonation and dodging attacks. Fig. (a) shows S_A (top) and S_B (bottom) dodging against DNN_B . Fig. (b)–(d) show impersonations. Impersonators carrying out the attack are shown in the top row and corresponding impersonation targets in the bottom row. Fig. (b) shows 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).

https://www.ece.cmu.edu/~lbauer/papers/2016/ccs2016-face-recognition.pdf

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)

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?



ACKNOWLEDGMENTS

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

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 networks. 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 Chain

HOW TO PROCEED?





ANY QUESTIONS? @barnhrtguy @acaltum

