

A close-up of Tony Stark's face, looking directly at the camera with a serious expression. Overlaid on his face and the background is a complex, glowing blue and white futuristic user interface (HUD) with various data points, gauges, and icons. The background is dark, making the glowing elements stand out.

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



\$ ID



Guy Barnhart-Magen

@barnhartguy

BSidesTLV Co-founder and
CTF Lead



Ezra Caltum

@acaltum

BSidesTLV Co-Founder
DC9723 Lead



@barnhartguy

 @acaltum

BUILDING ON THE SHOULDERS OF GIANTS





@barnhartguy



@acaltum

HOW DID WE GET HERE?



Awesome Conversations → Ideas



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>

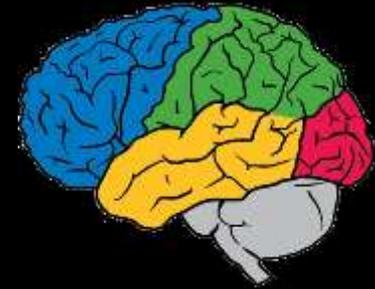
@barnhartguy



@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

SOME BACKGROUND

@barnhartguy



@acaltum

ARTIFICIAL INTELLIGENCE?

Machine Learning

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



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



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.



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

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?



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

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

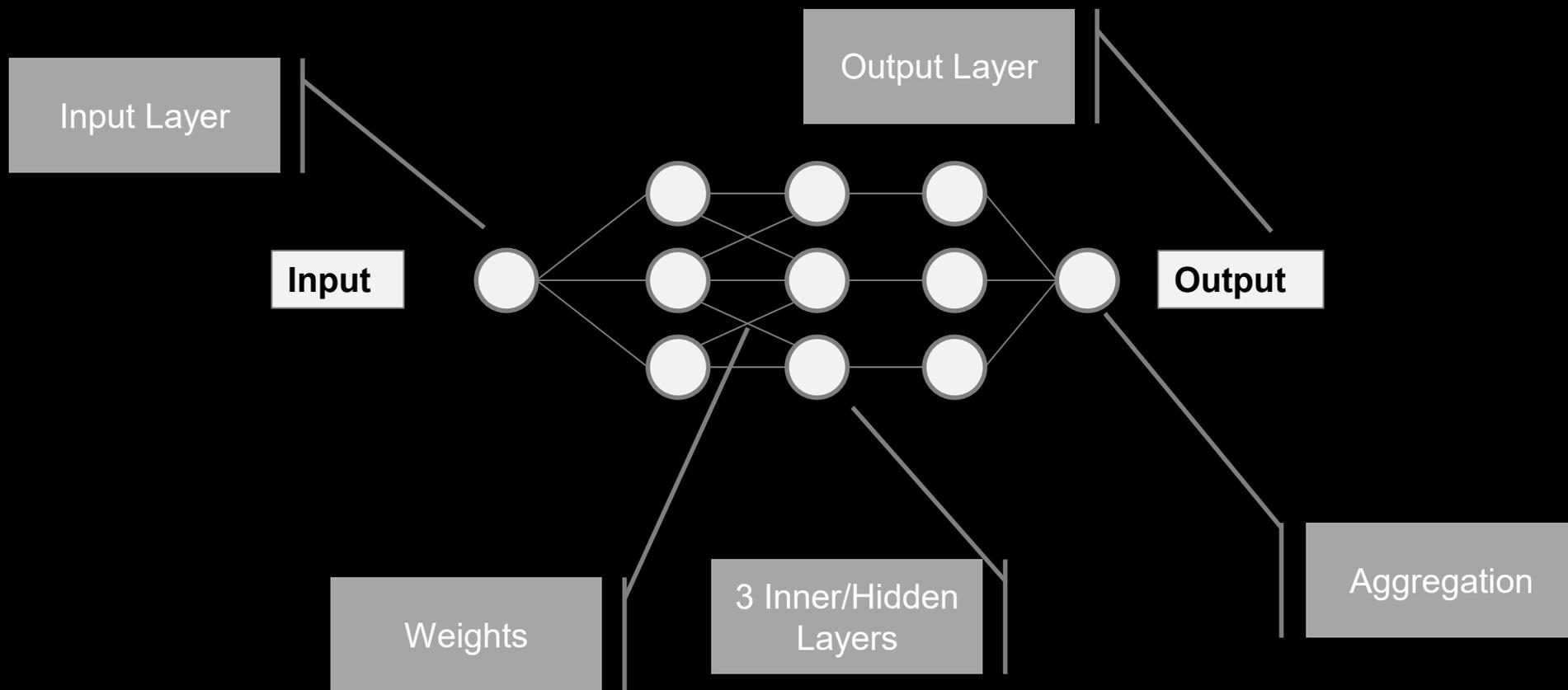
SIT BACK 'N RELAX



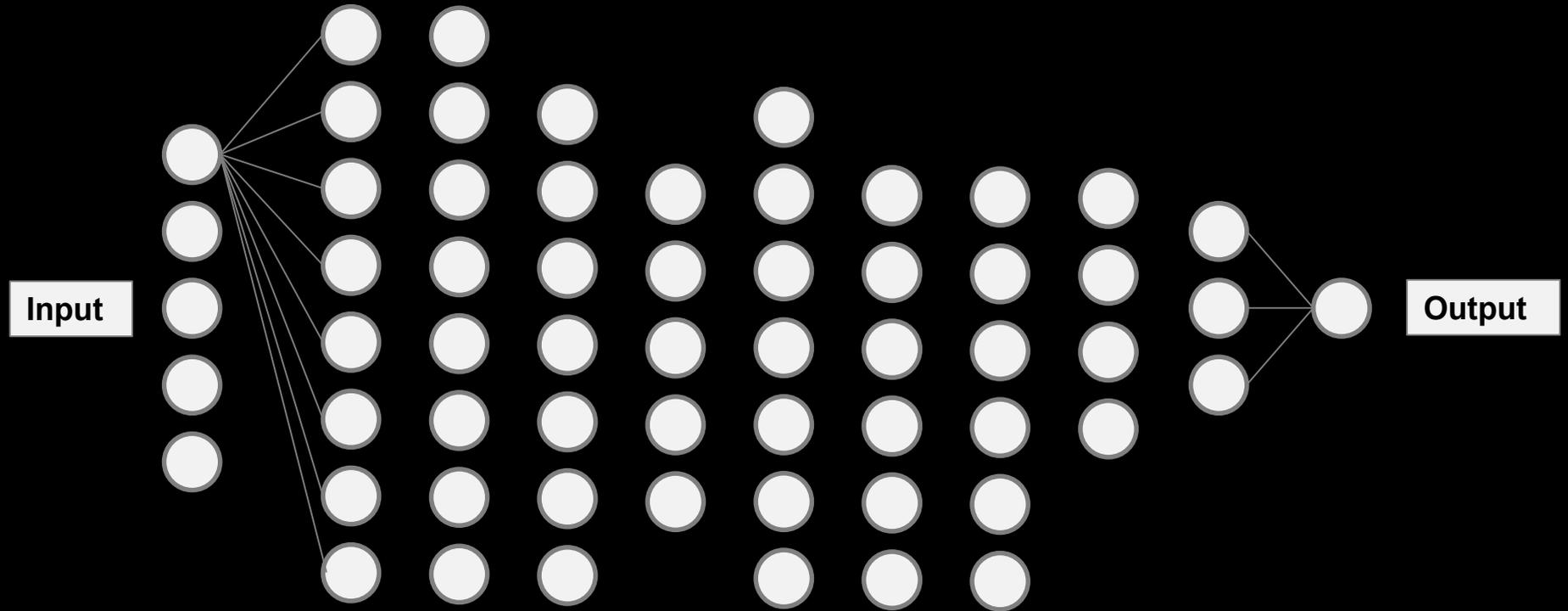
WE GOT THIS

imgflip.com

WHAT IS A ML MODEL?



WHAT IS A ML MODEL?



WHAT IS A ML MODEL?

- Training: Iterative process to **adjust weights**
- The “model” includes:
 - Topology/Layout
 - Weights/Parameters
 - 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;  
};
```



TOO MUCH VOODOO!

Input

Images
Audio
Binaries
Text

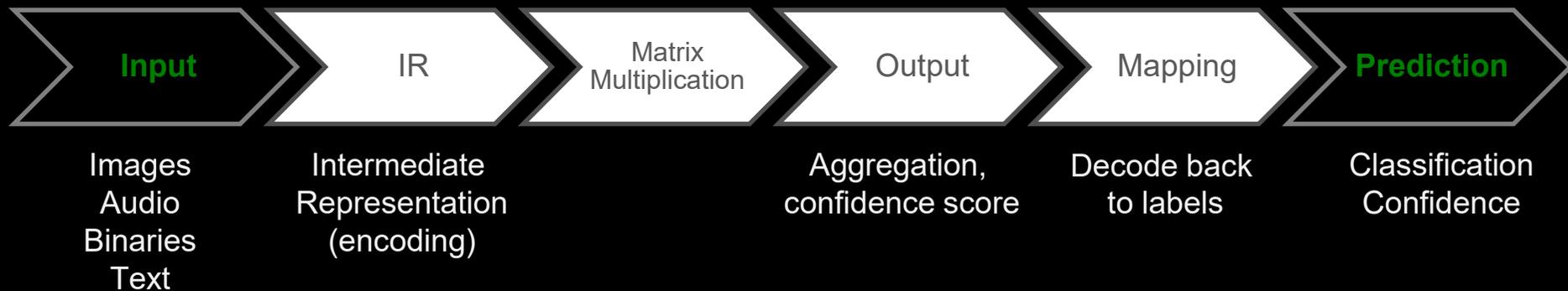
5l~f*9\>71rB

Prediction

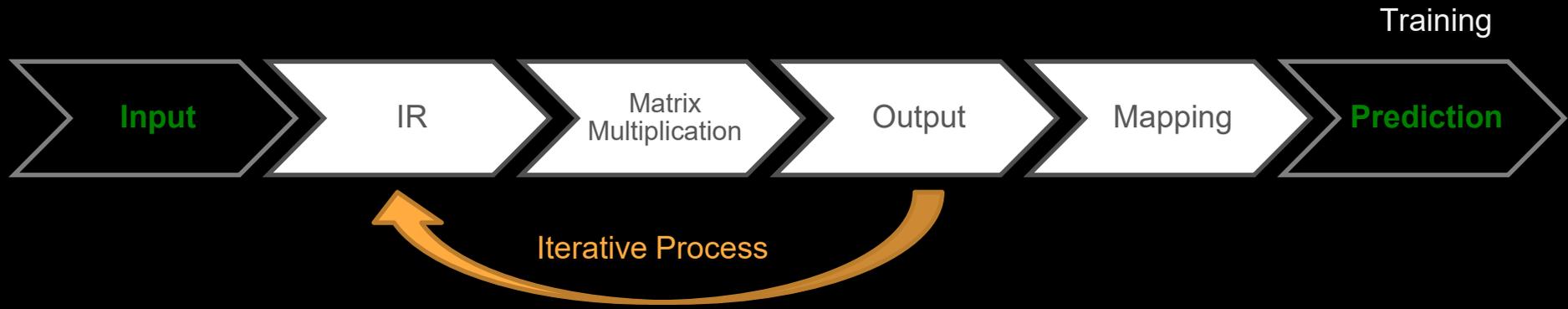
Classification
Confidence



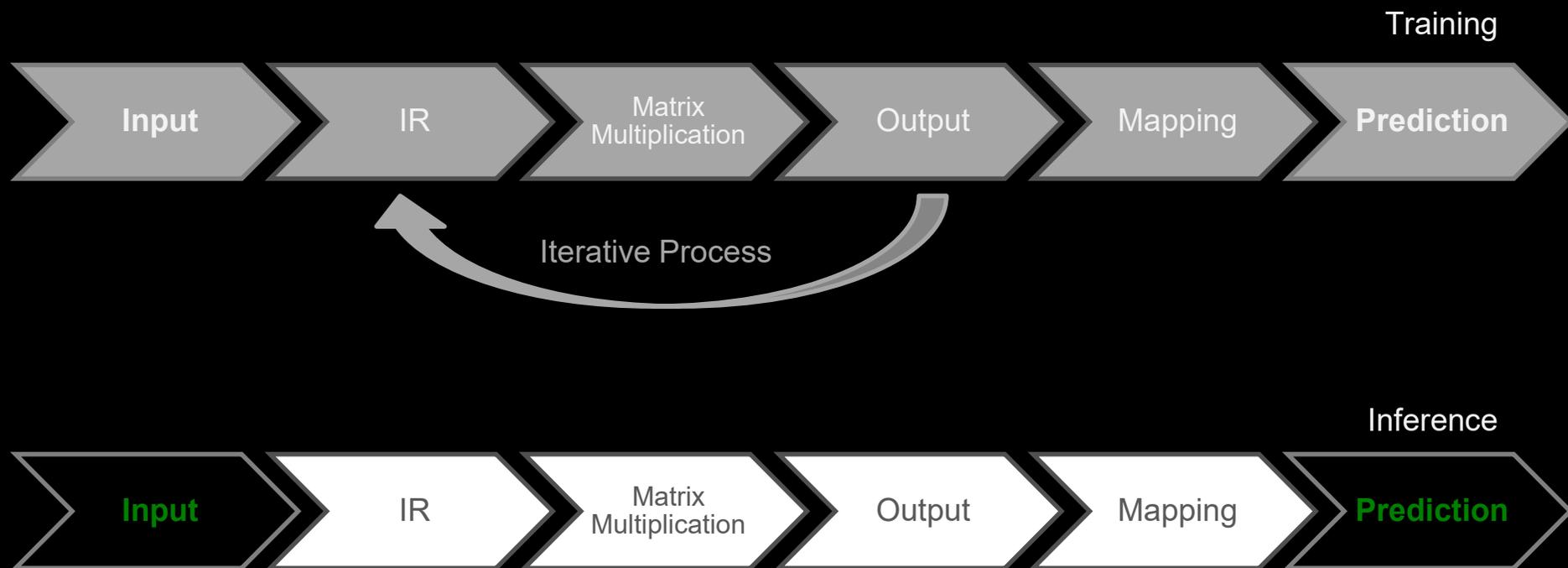
NOT TOO MUCH VOODOO!



FROM TRAINING TO INFERENCE



FROM TRAINING TO INFERENCE



MODEL != CODE



EXECUTABLE

Code execution flow

ML MODEL

Math operations, transition functions



EXECUTABLE

Code execution flow

Data Structures

ML MODEL

Math operations, transition functions

Intermediate Representation



EXECUTABLE

Code execution flow

Data Structures

Code Review or Reverse
Engineering

ML MODEL

Math operations, transition functions

Intermediate Representation

Model structure (Black Magic)



\$ HEXDUMP /MODELS/RESNET

```
00000000 0A 09 52 65 73 4E 65 74 2D 35 30 1A 04 64 61 74 61 A2 06 C9 ..ResNet-50..data...
00000014 A8 02 0A 05 63 6F 6E 76 31 12 0B 43 6F 6E 76 6F 6C 75 74 69 ....conv1..Convoluti
00000028 6F 6E 1A 04 64 61 74 61 22 05 63 6F 6E 76 31 3A 8C A6 02 2A on..data".conv1:...*
0000003C 80 A6 02 0E 72 E7 3C 64 FD 94 3C 1E 21 82 3C BE 99 1E 3B 99 ....r.<d..<!.<...;.
00000050 26 59 BD 30 F8 41 BD 63 E7 97 3C E1 E8 02 3C 07 DD CB 3C 86 &Y.0.A.c..<...<...<.
00000064 22 9D 3D 7B DC B8 3D E1 F3 9A BC 7F 85 A2 BD BB AF 7B BC A9 ".={..=.....{..
00000078 0F E1 BC C7 A9 86 BD 8E C6 A3 3C 9F 32 3D 3E 3D 9A EC 3D A2 .....<.2=>=..=.
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 0A 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 F0 3E BD FA 16 02 BE 5A `.=`.=-.2<=..>.....Z
000000DC 3A AC BC FB 13 08 3D 75 3A CD 3C FE B3 6F 3C A5 4D 10 3D 95 :.....=u:.<..o<.M.=.
000000F0 A2 86 3B 58 56 77 BD 7C A6 8C BC 49 A6 3C 3C C3 01 9D 3B AF ..;XVw.|...I.<<...;.
00000104 4A C0 3B FD A4 99 3C C7 09 14 3D 82 88 8C 3C 48 9B BF BD 30 J.;...<...=...<H...0
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
00000140 27 1C BE 8F F4 79 3D B1 7A D4 3E BF 27 AA 3E 62 ED A6 BD 9B '....y=.z.>.'>b....
00000154 2D 58 BE 6C 01 5F BD 2A B5 82 BE 52 63 87 BE 02 63 57 3E 1F -X.l._.*...Rc...cW>.
00000168 F5 11 3F 8D F2 8D 3E 0A EF D8 BC E1 D0 BC 3C 15 1F 36 BE 3F ..?...>.....<..6.?
0000017C D5 D2 BE DC 72 89 BE 71 40 3E 3E D3 48 80 3E FD 5D C5 3D 9A ....r...q@>>.H.>.] .=.
00000190 17 C8 3D 51 88 1F 3D C4 F3 44 BE 7D A0 95 BE 1E C7 18 BD F1 ..=Q..=.D.}.....
000001A4 E8 CB 3D D8 3F CD 3D FE C2 7A 3D DB EE 81 3D C8 CE B6 BC 3E ..=.?.=.z=...=...>
000001B8 95 2E BE 76 BF B0 BD FE 14 F3 3B 27 70 E1 3C A2 F7 1F BB 10 ...v.....;'p.<.....
000001CC 05 9C 3C 7C 8D B2 3C 08 93 5D BB AA 9C 9D BD AA 0B 92 BD 22 ..<|..<..]....."
--- 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**

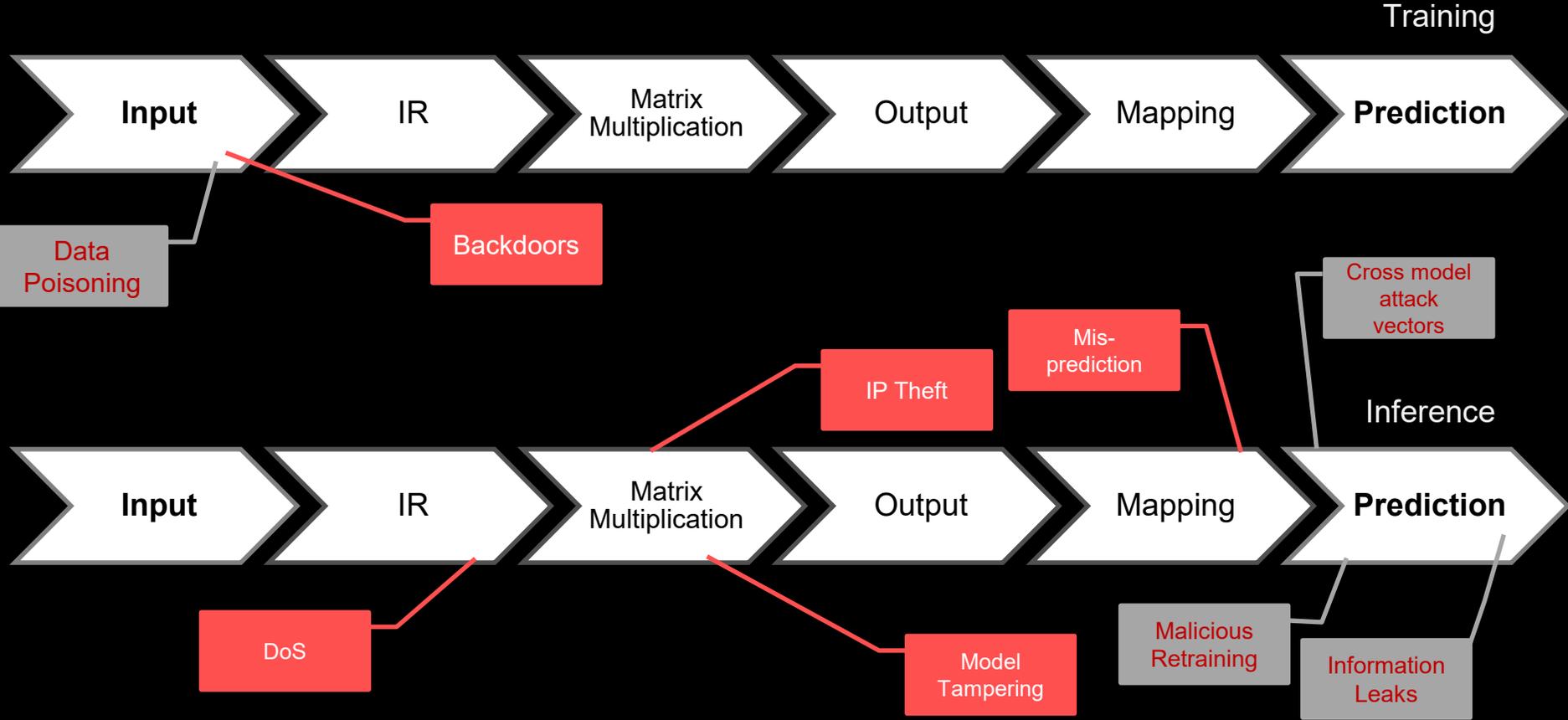
Bias is a part of the system learning process



BIAS - SOLVING THE WRONG PROBLEM



FROM TRAINING TO INFERENCE



SCORING

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



TOP 5 ATTACKS (CVSS)

1	DoS	7.5 (High)
2	Misprediction (adversarial attacks)	7.5 (High)
3	Model Tampering	7.4 (High)
4	IP Theft	5.9 (Medium)
5	Backdoors	3.9 (Low)

TOP 5 ATTACKS (BUSINESS IMPACT)

1	IP Theft	5.9 (Medium)
2	Model Tampering	7.4 (High)
3	DoS	7.5 (High)
4	Backdoors	3.9 (Low)
5	Misprediction (Adversarial attacks)	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?

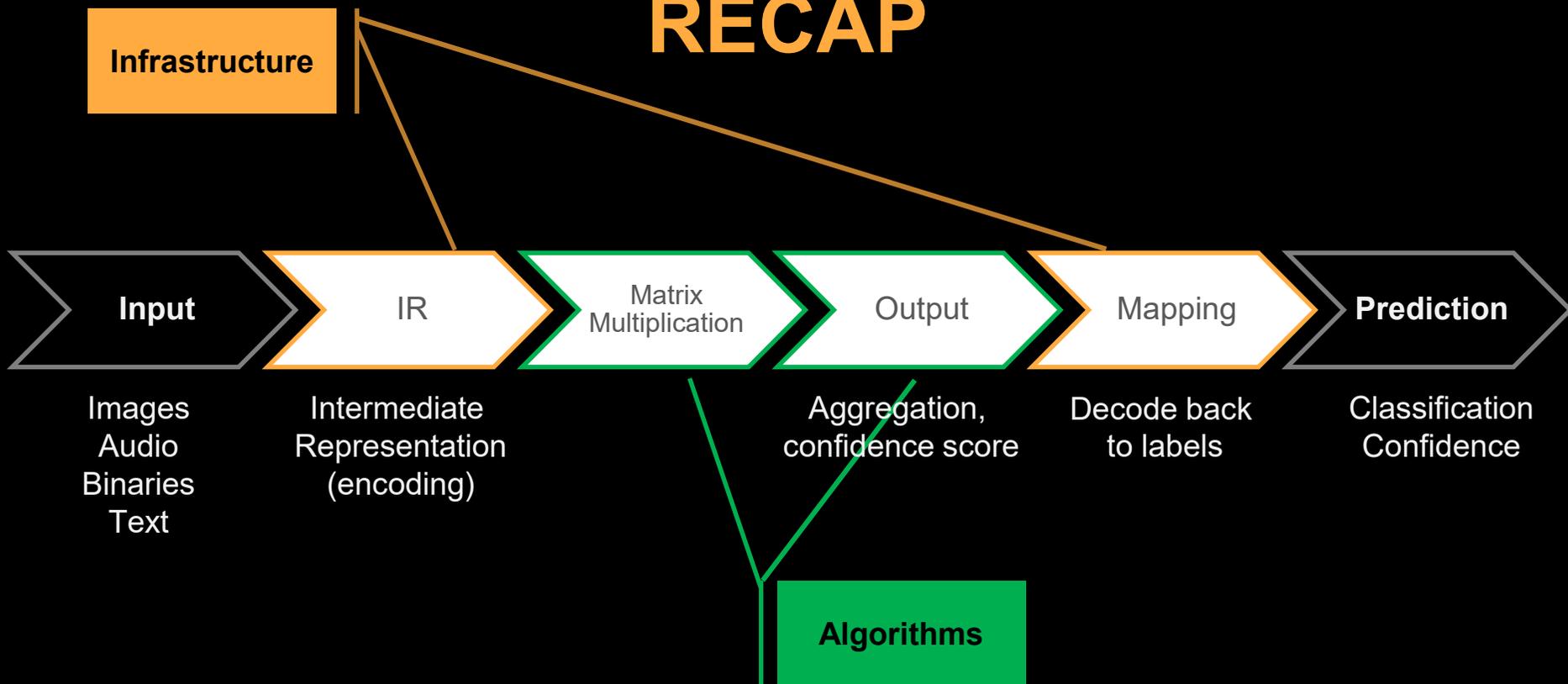


WHERE TO ATTACK?

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



RECAP



PARSING

ML needs to **convert** the input into a **matrix**



PARSING

ML needs to convert the input into a matrix

Parsing is hard



PARSING

ML needs to convert the input into a matrix

Parsing is hard

AI developers don't **develop file formats**. Or parsers.



PARSING

ML needs to convert the input into a matrix

Parsing is hard

AI 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**...

So let's **fuzz** the file format parsing



WE FAILED MISERABLY AT THE BEGINNING

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



FUZZING

What to focus on?

Caffe

Why focus here?

Full coverage

Issues?

Extremely slow



FUZZING

What to focus on?

Caffe

OpenCV

Why focus here?

Full coverage

Limited coverage

Issues?

Extremely slow

Medium speed



FUZZING

What to focus on?

Caffe

OpenCV

LibXXX

Why focus here?

Full coverage

Limited coverage

Very fast

Issues?

Extremely slow

Medium speed

Unknown code paths



FUZZING

What to focus on?

Caffe

OpenCV

LibXXX

Upstream

Why focus here?

Full coverage

Limited coverage

Very fast

Fuzzing not needed

Issues?

Extremely slow

Medium speed

Unknown code paths

Patched? 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



FACE PALM

Scaling exploitation is hard (DevOps?)



EVENTUALLY WE GOT OUR GRIP

@barnhartguy



@acaltum

PRELIMINARY RESULTS

Almost every **exploit category** was discovered



PRELIMINARY RESULTS

Almost every exploit category was discovered
Bug collisions



PRELIMINARY RESULTS

Almost every exploit category was discovered

Bug collisions

So many crashes



PRELIMINARY RESULTS

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**



SO...



FUZZING → CRASH, NOW WHAT?

1	IP Theft	5.9 (Medium)
2	Model Tampering	7.4 (High)
3	DoS	7.5 (High)
4	Backdoors	3.9 (Low)
5	Misprediction (Adversarial attacks)	7.5 (High)

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
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 14:48:43 2018 from 10.0.2.2
→ -
```

```
→ 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 14:49:29 2018 from 10.0.2.2
→ - █
```



```
→ 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 14:48:43 2018 from 10.0.2.2
→ -
```

```
→ 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 14:49:29 2018 from 10.0.2.2
→ - htop
```



```

→ 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 14:48:43 2018 from 10.0.2.2
→ -

```

```

1 [          ] Tasks: 22, 6 thr; 1 running
2 [          ] Load average: 0.07 0.03 0.00
Mem[| | | ] Uptime: 00:09:49
Swp[          ]

```

PID	USER	PRI	NI	VIRT	RES	SHR	S	CPU%	MEM%	TIME+	Comm
1	root	20	0	37980	5968	3952	S	0.0	0.1	0:02.55	/sbin
207		20	0	35272	3528	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	2460	2252	S	0.0	0.0	0:00.00	/lib
335		20	0	97M	2460	2252	S	0.0	0.0	0:00.02	/lib
471		20	0	16120	856	0	S	0.0	0.0	0:00.00	/sbin
506		20	0	28620	3080	2760	S	0.0	0.0	0:00.02	/lib
533		20	0	250M	3416	2744	S	0.0	0.0	0:00.00	/usr
534		20	0	250M	3416	2744	S	0.0	0.0	0:00.00	/usr
535		20	0	250M	3416	2744	S	0.0	0.0	0:00.00	/usr
507		20	0	250M	3416	2744	S	0.0	0.0	0:00.02	/usr
511		20	0	42900	3900	3488	S	0.0	0.0	0:00.08	/usr
557		20	0	269M	6260	5532	S	0.0	0.1	0:00.01	/usr
565		20	0	269M	6260	5532	S	0.0	0.1	0:00.00	/usr
538		20	0	269M	6260	5532	S	0.0	0.1	0:00.04	/usr
546		20	0	29008	2920	2648	S	0.0	0.0	0:00.00	/usr
563		20	0	19472	2288	2064	S	0.0	0.0	0:00.01	/usr
665		20	0	65508	6040	5332	S	0.0	0.1	0:00.00	/usr
679		20	0	15936	1792	1664	S	0.0	0.0	0:00.00	/sbin
909		20	0	92800	6940	6000	S	0.0	0.1	0:00.00	sshd
911	user	20	0	45192	4988	4156	S	0.0	0.1	0:00.01	/lib
912	user	20	0	61432	2128	0	S	0.0	0.0	0:00.00	(sd-
935	user	20	0	92800	3328	2392	S	0.0	0.0	0:00.00	sshd
936	user	20	0	44316	5448	3852	S	0.0	0.1	0:00.17	-zsh
965		20	0	92800	6748	5812	S	0.0	0.1	0:00.00	sshd
985	user	20	0	92800	3304	2372	S	0.0	0.0	0:00.00	sshd

```

F1 Help F2 Setup F3 Search F4 Filter F5 Tree F6 SortBy F7 Nice F8 Nice F9

```

```

→ 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 14:48:43 2018 from 10.0.2.2
→ - cd /home/user/for_presentation/jarvis_demo/runner/

```

```

1 [          ] Tasks: 22, 6 thr; 1 running
2 [          ] Load average: 0.04 0.03 0.00
Mem[| | | ] Uptime: 00:10:12
Swp[       ]

PID USER      PRI  NI  VIRT   RES   SHR  S  CPU%  MEM%   TIME+  Comm
1011 user        20   0 25924  3756  3196 R  0.7   0.0   0:00.06 htop
1          20   0  37980  5968  3952 S  0.0   0.1   0:02.55 /sbin
207        20   0  35272  3528  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   2460  2252 S  0.0   0.0   0:00.00 /lib
335        20   0   97M   2460  2252 S  0.0   0.0   0:00.02 /lib
471        20   0  16120   856    0 S  0.0   0.0   0:00.00 /sbin
506        20   0  28620  3080  2760 S  0.0   0.0   0:00.02 /lib
533        20   0   250M  3416  2744 S  0.0   0.0   0:00.00 /usr
534        20   0   250M  3416  2744 S  0.0   0.0   0:00.00 /usr
535        20   0   250M  3416  2744 S  0.0   0.0   0:00.00 /usr
507        20   0   250M  3416  2744 S  0.0   0.0   0:00.02 /usr
511        20   0  42900  3900  3488 S  0.0   0.0   0:00.08 /usr
557        20   0   269M  6260  5532 S  0.0   0.1   0:00.01 /usr
565        20   0   269M  6260  5532 S  0.0   0.1   0:00.00 /usr
538        20   0   269M  6260  5532 S  0.0   0.1   0:00.04 /usr
546        20   0  29008  2920  2648 S  0.0   0.0   0:00.00 /usr
563        20   0  19472  2288  2064 S  0.0   0.0   0:00.01 /usr
665        20   0  65508  6040  5332 S  0.0   0.1   0:00.00 /usr
679        20   0  15936  1792  1664 S  0.0   0.0   0:00.00 /sbin
909        20   0  92800  6940  6000 S  0.0   0.1   0:00.00 sshd
911 user        20   0  45192  4988  4156 S  0.0   0.1   0:00.01 /lib
912 user        20   0  61432  2128    0 S  0.0   0.0   0:00.00 (sd-
935 user        20   0  92800  3328  2392 S  0.0   0.0   0:00.00 sshd
936 user        20   0  44316  5448  3852 S  0.0   0.1   0:00.17 -zsh
965        20   0  92800  6748  5812 S  0.0   0.1   0:00.00 sshd

```

F1 Help F2 Setup F3 Search F4 Filter F5 Tree F6 SortBy F7 Nice F8 Nice F9

```

→ 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 14:48:43 2018 from 10.0.2.2
→ - cd /home/user/for_presentation/jarvis_demo/runner/
→ runner ./classification-d ../exploits/ram_and_cpu_dos.bmp

```

```

1 [          ] Tasks: 22, 6 thr; 1 running
2 [ |        ] Load average: 0.04 0.02 0.00
Mem[| | |    ] Uptime: 00:10:23
Swp[          ]

```

PID	USER	PRI	NI	VIRT	RES	SHR	S	CPU%	MEM%	TIME+	Comm
1011	user	20	0	25924	3756	3196	R	0.7	0.0	0:00.08	htop
1		20	0	37980	5968	3952	S	0.0	0.1	0:02.55	/sbin
207		20	0	35272	3528	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	2460	2252	S	0.0	0.0	0:00.00	/lib
335		20	0	97M	2460	2252	S	0.0	0.0	0:00.02	/lib
471		20	0	16120	856	0	S	0.0	0.0	0:00.00	/sbin
506		20	0	28620	3080	2760	S	0.0	0.0	0:00.02	/lib
533		20	0	250M	3416	2744	S	0.0	0.0	0:00.00	/usr
534		20	0	250M	3416	2744	S	0.0	0.0	0:00.00	/usr
535		20	0	250M	3416	2744	S	0.0	0.0	0:00.00	/usr
507		20	0	250M	3416	2744	S	0.0	0.0	0:00.02	/usr
511		20	0	42900	3900	3488	S	0.0	0.0	0:00.08	/usr
557		20	0	269M	6260	5532	S	0.0	0.1	0:00.01	/usr
565		20	0	269M	6260	5532	S	0.0	0.1	0:00.00	/usr
538		20	0	269M	6260	5532	S	0.0	0.1	0:00.04	/usr
546		20	0	29008	2920	2648	S	0.0	0.0	0:00.00	/usr
563		20	0	19472	2288	2064	S	0.0	0.0	0:00.01	/usr
665		20	0	65508	6040	5332	S	0.0	0.1	0:00.00	/usr
679		20	0	15936	1792	1664	S	0.0	0.0	0:00.00	/sbin
909		20	0	92800	6940	6000	S	0.0	0.1	0:00.00	sshd
911	user	20	0	45192	4988	4156	S	0.0	0.1	0:00.01	/lib
912	user	20	0	61432	2128	0	S	0.0	0.0	0:00.00	(sd-
935	user	20	0	92800	3328	2392	S	0.0	0.0	0:00.00	sshd
936	user	20	0	44316	5452	3852	S	0.0	0.1	0:00.17	-zsh
965		20	0	92800	6748	5812	S	0.0	0.1	0:00.00	sshd

F1 Help F2 Setup F3 Search F4 Filter F5 Tree F6 Sort By F7 Nice -F8 Nice +F9

```

→ 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 14:48:43 2018 from 10.0.2.2
→ - cd /home/user/for_presentation/jarvis_demo/runner/
→ runner ./classification-d ../exploits/ram_and_cpu_dos.bmp
----- Prediction for ../exploits/ram_and_cpu_dos.bmp -----

```

```

1  [|||||||]      ]  Tasks: 23, 6 thr; 1 running
2  [|||]          ]  Load average: 0.11 0.04 0.01
Mem[|||]         ]  Uptime: 00:10:26
Swp[|]           ]

```

PID	USER	PRI	NI	VIRT	RES	SHR	S	CPU%	MEM%	TIME+	Comm
1016	user	20	0	354M	221M	19280	D	64.8	2.8	0:00.98	./cl
1011	user	20	0	25924	3756	3196	R	0.8	0.0	0:00.09	htop
1		20	0	37980	5968	3952	S	0.0	0.1	0:02.55	/sbi
207		20	0	35272	3528	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	2460	2252	S	0.0	0.0	0:00.00	/lib
335		20	0	97M	2460	2252	S	0.0	0.0	0:00.02	/lib
471		20	0	16120	856	0	S	0.0	0.0	0:00.00	/sbi
506		20	0	28620	3080	2760	S	0.0	0.0	0:00.02	/lib
533		20	0	250M	3416	2744	S	0.0	0.0	0:00.00	/usr
534		20	0	250M	3416	2744	S	0.0	0.0	0:00.00	/usr
535		20	0	250M	3416	2744	S	0.0	0.0	0:00.00	/usr
507		20	0	250M	3416	2744	S	0.0	0.0	0:00.02	/usr
511		20	0	42900	3900	3488	S	0.0	0.0	0:00.08	/usr
557		20	0	269M	6260	5532	S	0.0	0.1	0:00.01	/usr
565		20	0	269M	6260	5532	S	0.0	0.1	0:00.00	/usr
538		20	0	269M	6260	5532	S	0.0	0.1	0:00.04	/usr
546		20	0	29008	2920	2648	S	0.0	0.0	0:00.00	/usr
563		20	0	19472	2288	2064	S	0.0	0.0	0:00.01	/usr
665		20	0	65508	6040	5332	S	0.0	0.1	0:00.00	/usr
679		20	0	15936	1792	1664	S	0.0	0.0	0:00.00	/sbi
909		20	0	92800	6940	6000	S	0.0	0.1	0:00.00	sshd
911	user	20	0	45192	4988	4156	S	0.0	0.1	0:00.01	/lib
912	user	20	0	61432	2128	0	S	0.0	0.0	0:00.00	(sd-
935	user	20	0	92800	3328	2392	S	0.0	0.0	0:00.00	sshd
936	user	20	0	44316	5452	3852	S	0.0	0.1	0:00.17	-zsh



```

→ 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 14:48:43 2018 from 10.0.2.2
→ - cd /home/user/for_presentation/jarvis_demo/runner/
→ runner ./classification-d ../exploits/ram_and_cpu_dos.bmp
----- Prediction for ../exploits/ram_and_cpu_dos.bmp -----

```

```

1 [|||||||||||||||||100.0%] Tasks: 23, 6 thr; 2 running
2 [ | ] Load average: 0.86 0.34 0.12
Mem[|||||||||||||5.98G/7. ] Uptime: 00:12:20
Swp[ | ]

```

PID	USER	PRI	NI	VIRT	RES	SHR	S	CPU%	MEM%	TIME+	Comm
1016	user	20	0	7171M	6065M	20432	R	116.	76.0	1:55.37	./c
1011	user	20	0	25924	3756	3196	R	0.8	0.0	0:00.35	htop
563		20	0	19472	2288	2064	S	0.0	0.0	0:00.02	/usr
1		20	0	37980	5968	3952	S	0.0	0.1	0:02.56	/sbi
985	user	20	0	92800	3304	2372	S	0.0	0.0	0:00.02	sshd
207		20	0	35272	3528	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	2460	2252	S	0.0	0.0	0:00.00	/lib
335		20	0	97M	2460	2252	S	0.0	0.0	0:00.02	/lib
471		20	0	16120	856	0	S	0.0	0.0	0:00.00	/sbi
506		20	0	28620	3080	2760	S	0.0	0.0	0:00.02	/lib
533		20	0	250M	3416	2744	S	0.0	0.0	0:00.00	/usr
534		20	0	250M	3416	2744	S	0.0	0.0	0:00.00	/usr
535		20	0	250M	3416	2744	S	0.0	0.0	0:00.00	/usr
507		20	0	250M	3416	2744	S	0.0	0.0	0:00.02	/usr
511		20	0	42900	3900	3488	S	0.0	0.0	0:00.08	/usr
557		20	0	269M	6260	5532	S	0.0	0.1	0:00.01	/usr
565		20	0	269M	6260	5532	S	0.0	0.1	0:00.00	/usr
538		20	0	269M	6260	5532	S	0.0	0.1	0:00.04	/usr
546		20	0	29008	2920	2648	S	0.0	0.0	0:00.00	/usr
665		20	0	65508	6040	5332	S	0.0	0.1	0:00.00	/usr
679		20	0	15936	1792	1664	S	0.0	0.0	0:00.00	/sbi
909		20	0	92800	6940	6000	S	0.0	0.1	0:00.00	sshd
911	user	20	0	45192	4988	4156	S	0.0	0.1	0:00.01	/lib
912	user	20	0	61432	2128	0	S	0.0	0.0	0:00.00	(sd-
935	user	20	0	92800	3328	2392	S	0.0	0.0	0:00.00	sshd

```

|F1Help F2Setup F3Search F4Filter F5Tree F6SortBy F7Nice F8Nice +F9

```

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)

* 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

→ - █

→ demos



```
→ 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
wm
→ - █
```

I



```
→ 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
→ - cd /home/user/for_presentation/jarvis_demo/runner/
→ runner █
```

→ 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
→ ~ cd /home/user/for_presentation/jarvis_demo/runner/
→ runner ./classification-d ../exploits/good-oneaaaaaaaaaaaaaaaaaaaa
aaaa.rmt.SHELL.cafe.static.full
```

```
→ demos
```



```
→ 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
wm
→ - cd /home/user/for_presentation/jarvis_demo/runner/
→ runner ./classification-d ../exploits/good-oneaaaaaaaaaaaaaaaaaaaa
aaaa.rmt.SHELL.cafe.static.full
----- Prediction for ../exploits/good-oneaaaaaaaaaaaaaaaaaaaa
a.rmt.SHELL.cafe.static.full -----
```



```
→ 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
wm
→ ~ cd /home/user/for_presentation/jarvis_demo/runner/
→ runner ./classification-d ../exploits/good-oneaaaaaaaaaaaaaaaaaaaa
aaaa.rmt.SHELL.cafe.static.full
----- Prediction for ../exploits/good-oneaaaaaaaaaaaaaaaaaaaa
a.rmt.SHELL.cafe.static.full -----
```

→ demos



DEMO TIME

Model Tampering

→ demos



@barnhartguy



@acaltum

DEMO TIME

IP Theft



NO DEMO

You actually saw this already 😊



**SO MAYBE RCE IS KING
AFTER ALL?**

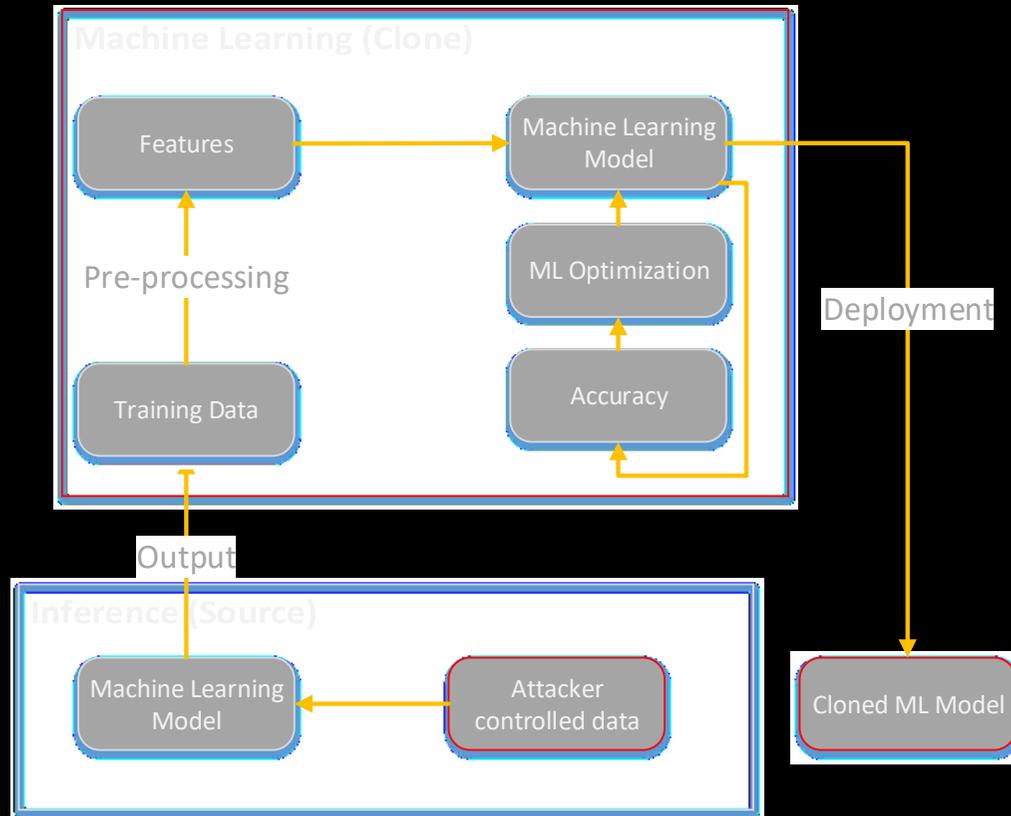


AND IF YOU DON'T HAVE AN RCE?

Let's go after the algorithms!



ATTACK OF THE CLONES



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)



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



BACKDOORS

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*



ENCODING

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...



DeepLocker

Concealing Targeted Attacks with AI Locksmithing

Dhilung Kirat, Jiyong Jang, Marc Ph. Stoecklin

IBM **Research**

The IBM Research logo, consisting of the word "IBM" in its characteristic eight horizontal stripes, followed by the word "Research" in a bold, sans-serif font.

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/>

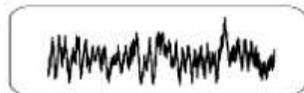
ADVERSARIAL AUDIO

▶ 0:03 / 0:03



“okay google **without the dataset the article is useless**”

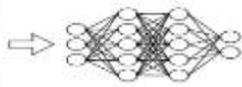
+



x 0.001

=

▶ 0:03 / 0:03



“okay google **browse to evil dot com**”

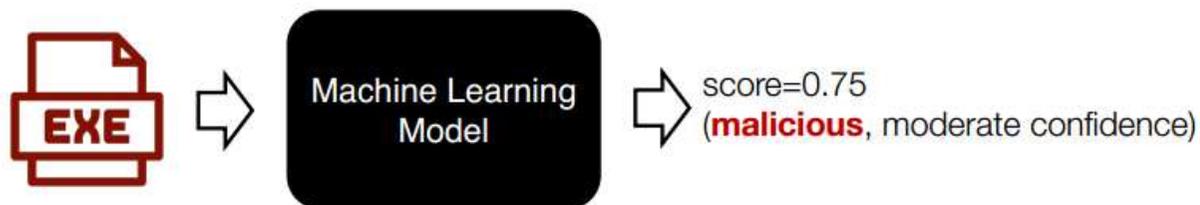
https://nicholas.carlini.com/code/audio_adversarial_examples/

@barnhartguy

@acaltum

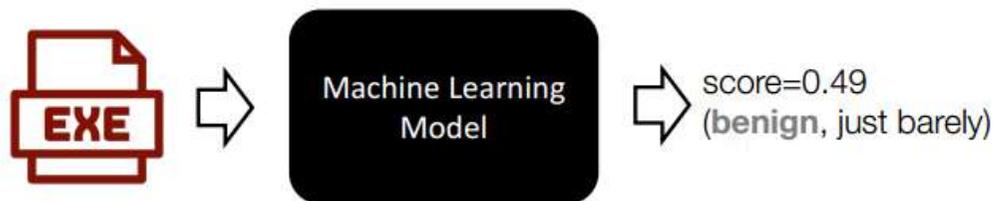
EVADING NEXT GENERATION AV USING AI

- Static machine learning model trained on millions of samples



- Simple structural changes that don't change behavior

- unpack
- '.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=FGC1e6T0Jpc>

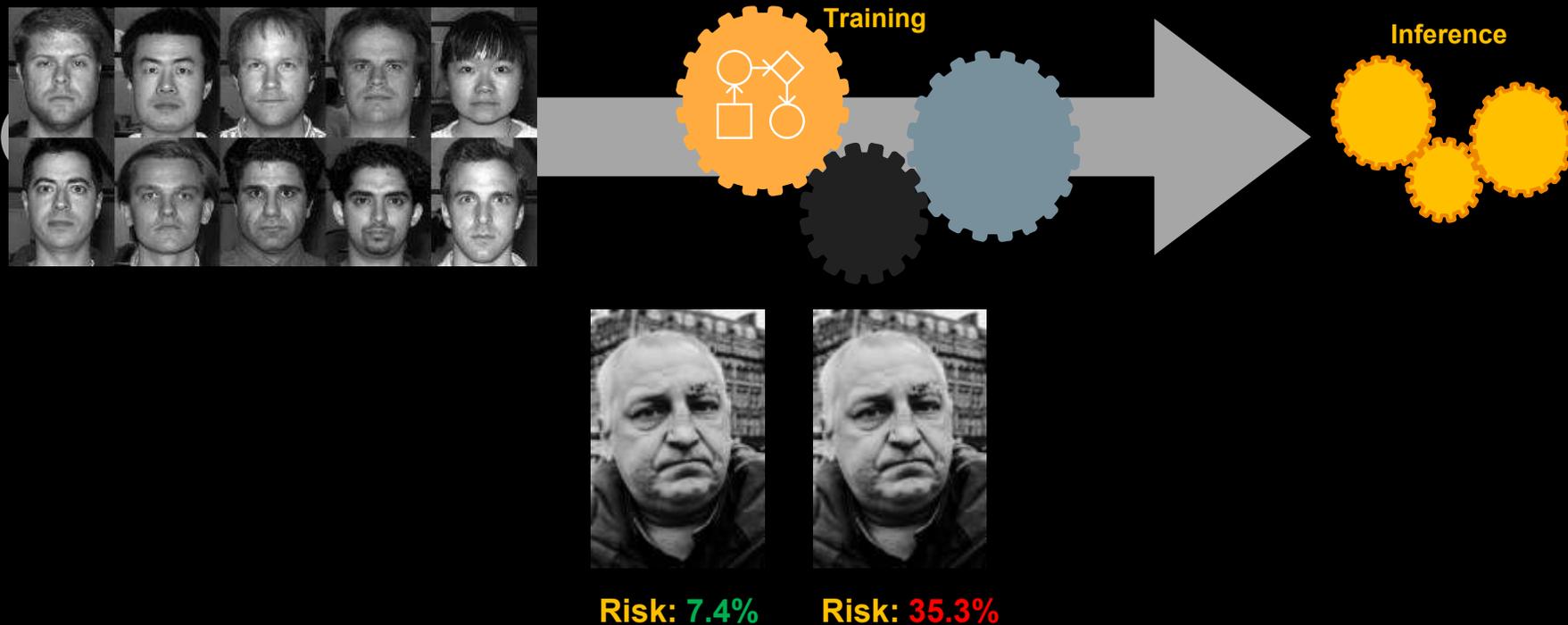
WHAT ABOUT PRIVACY ?

@barnhartguy

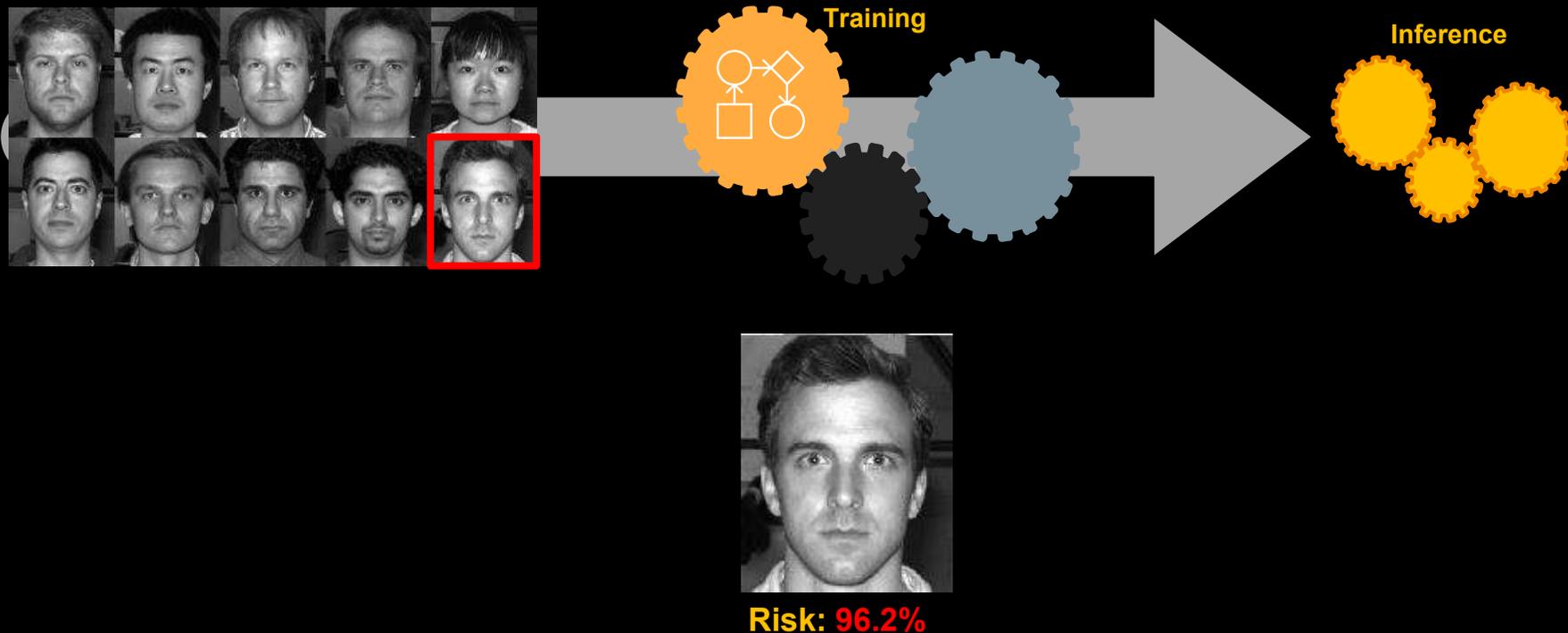


@acaltum

PRIVACY LEAKS? NOT YET, BUT SOON...



PRIVACY LEAKS? NOT YET, BUT SOON...



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?





**There is no AI.
It's just someone else's code.**

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?



Come talk to us!

ANY QUESTIONS?

@barnhartguy

@acaltum



@barnhartguy



@acaltum