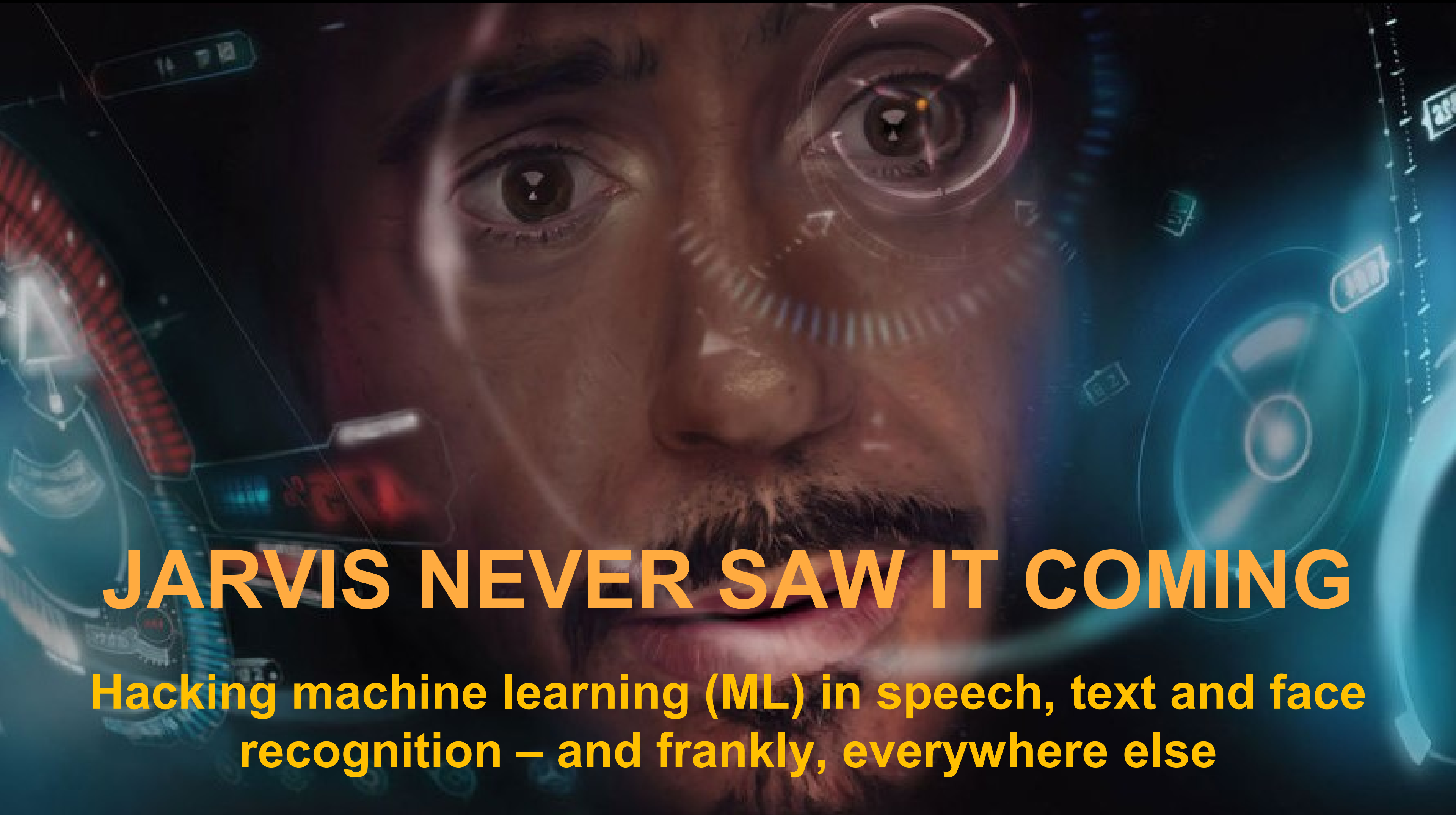


JARVIS NEVER SAW IT COMING

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

@barnhartguy @acaltum

LEGAL NOTICES AND DISCLAIMERS

intel®

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

BUILDING ON THE SHOULDERS OF GIANTS

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

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 <u>not</u> going to talk about

CLEVER HANS

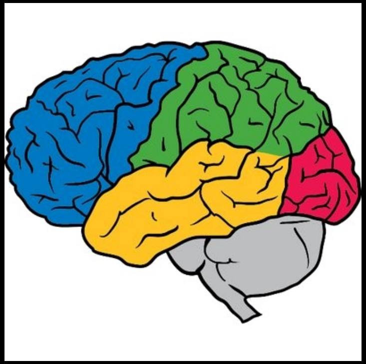
https://github.com/tensorflow/cleverhans

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

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

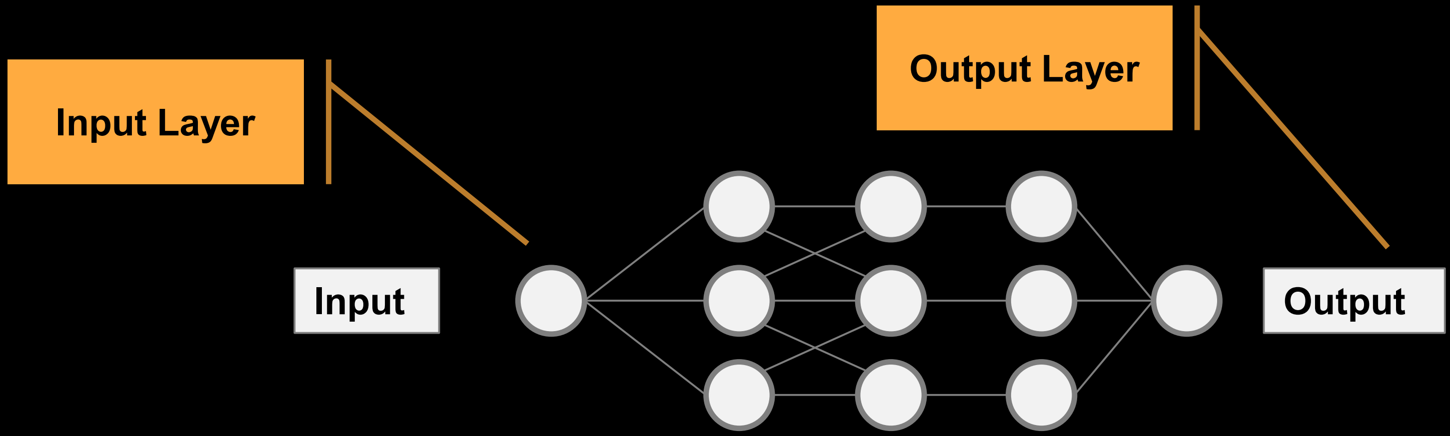
Most AI systems were designed to solve a specific problem, well.

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

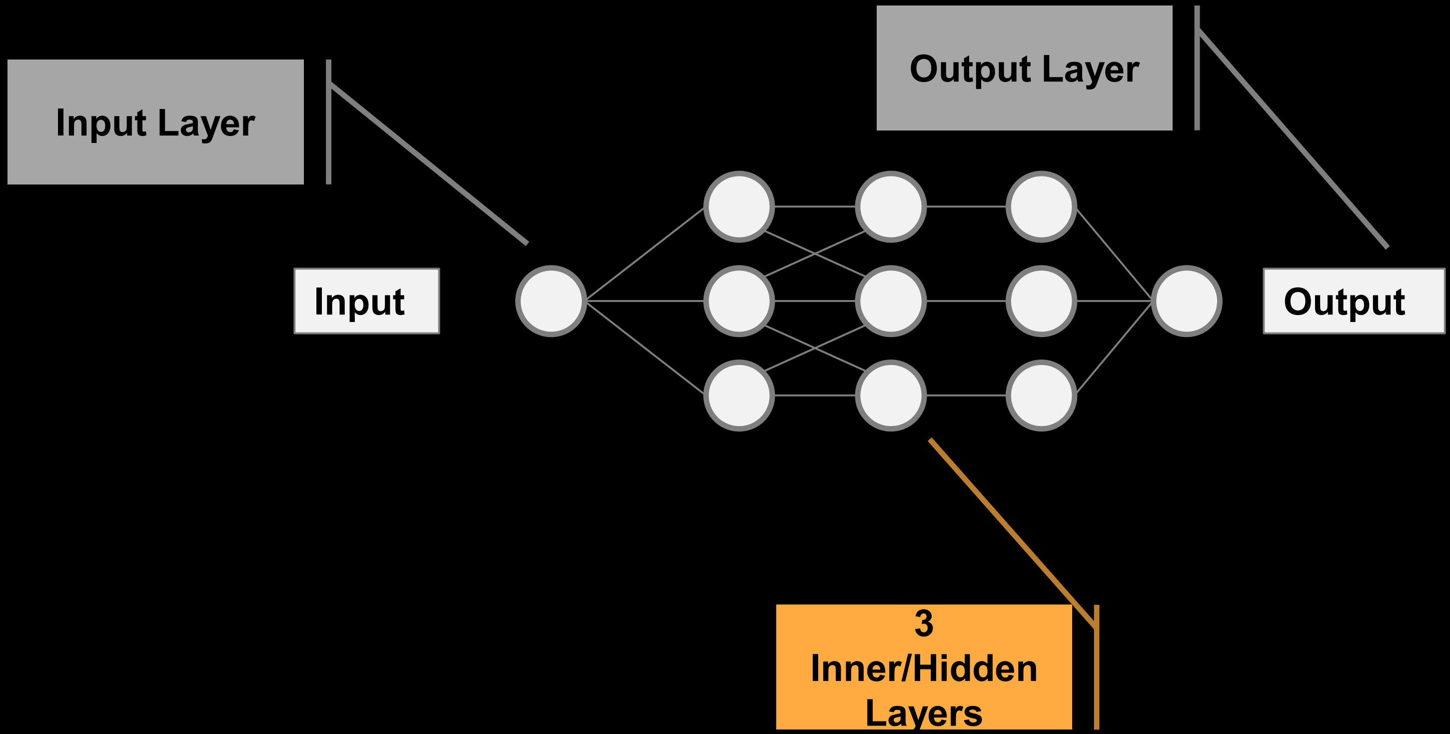
MACHINE LEARNING 101


SIT BACK 'N RELAX
WE GOT THIS
imgflip.com

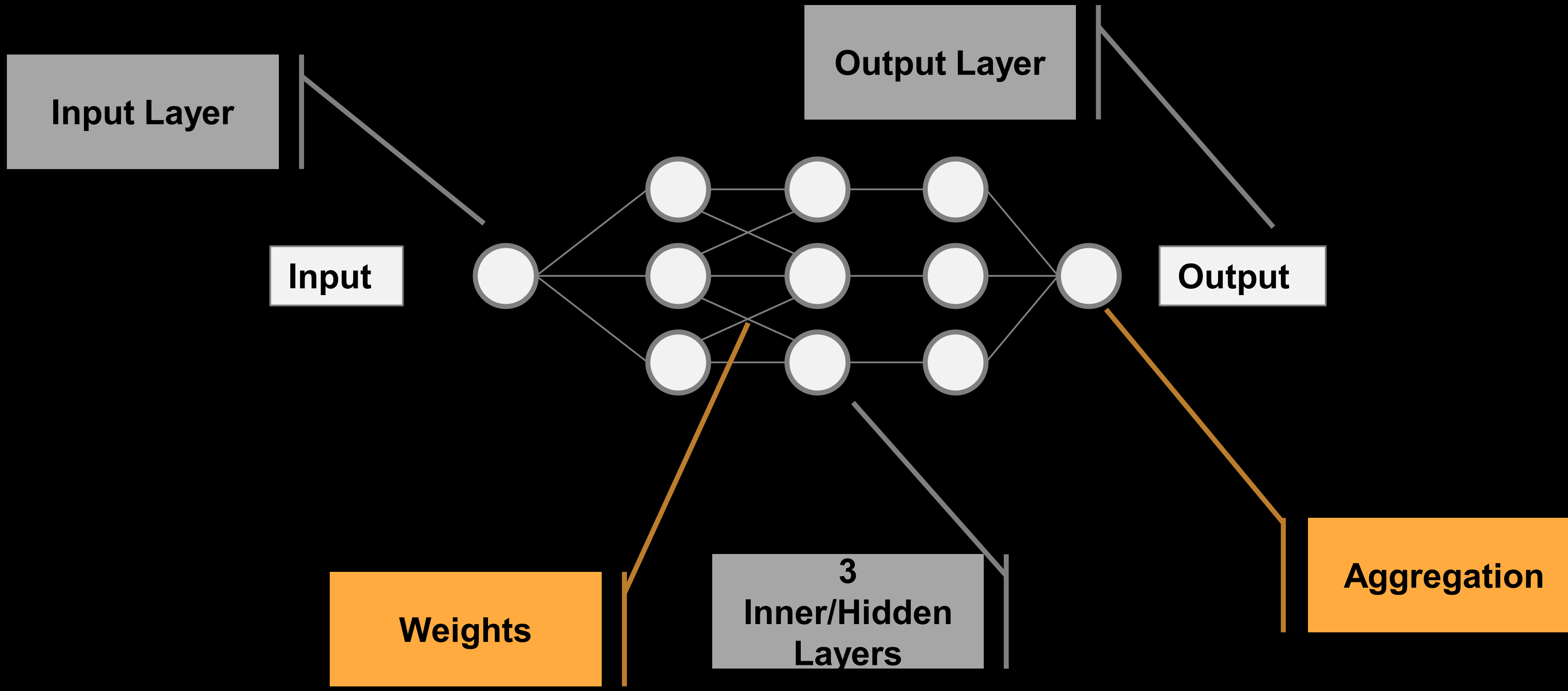
WHAT IS A ML MODEL?



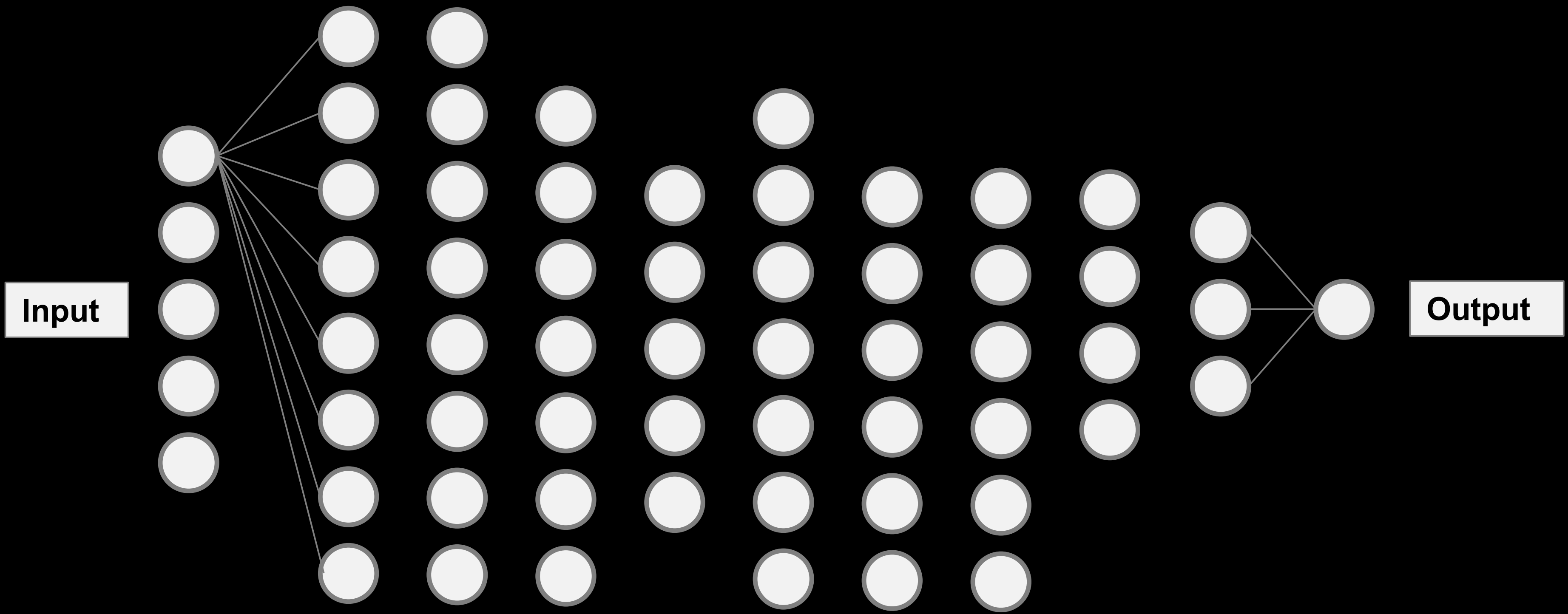
WHAT IS A ML MODEL?



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!

LINEAR ALGEBRA, ANYONE?

Definition [edit]

If $\mathbf{A}$ is an $n \times m$ matrix and $\mathbf{B}$ is an $m \times p$ matrix,

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1p} \\ b_{21} & b_{22} & \cdots & b_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mp} \end{pmatrix}$$

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

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

such that

$$c_{ij} = a_{i1}b_{1j} + \cdots + a_{im}b_{mj} = \sum_{k=1}^{m} a_{ik}b_{kj},$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, p$.

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!

Input

51~f*9\>71rB

Prediction

Images
Audio
Binaries
Text

Classification
Confidence

NOT TOO MUCH VOODOO!

Images
Audio
Binaries
Text

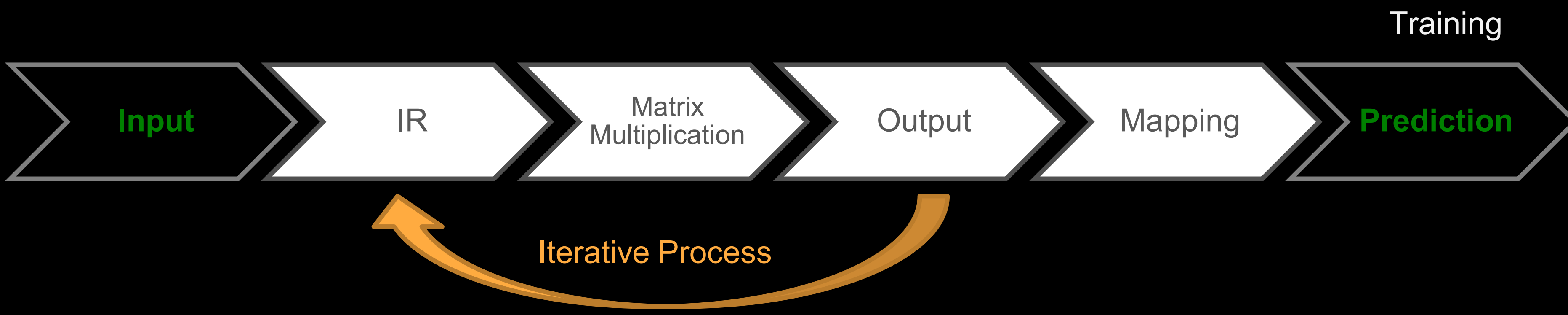
Intermediate
Representation
(encoding)

Aggregation,
confidence score

Decode back
to labels

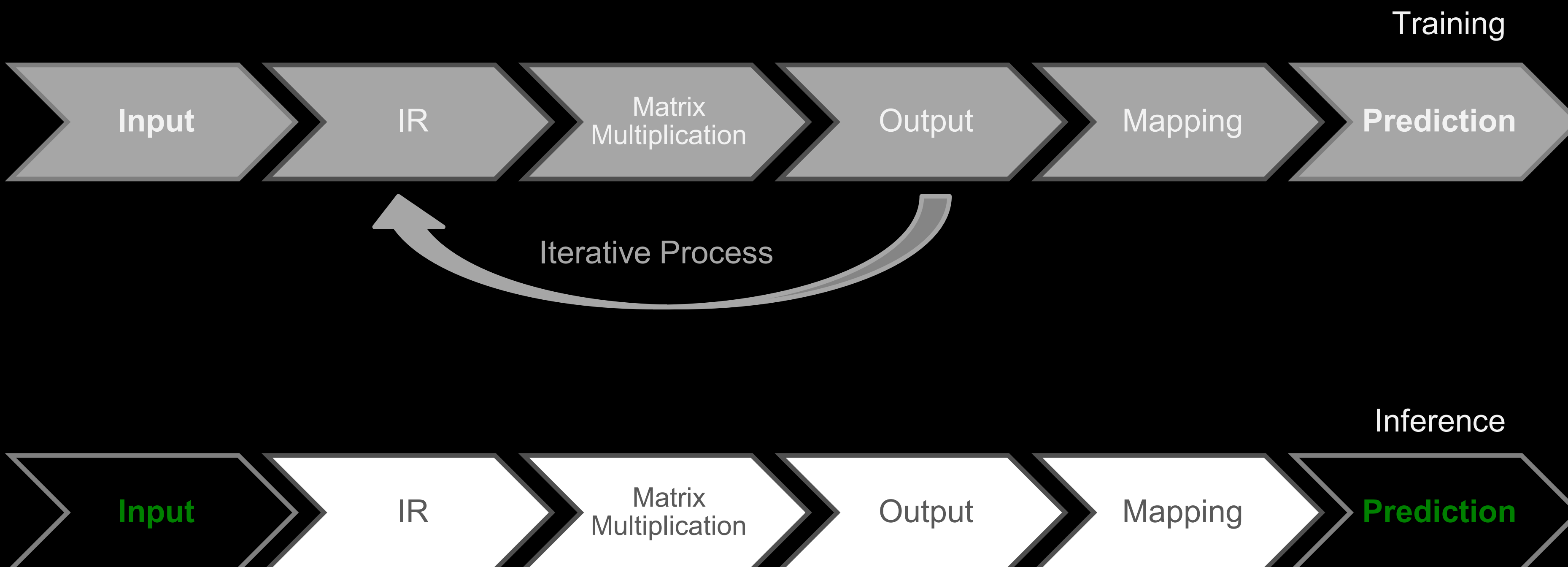
Classification
Confidence

FROM TRAINING TO INFERENCE



@barnhartguy @acaltum

FROM TRAINING TO INFERENCE



@barnhartguy @acaltum

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---------------------------------------
```

@barnhartguy @acaltum

VALIDATION

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

VALIDATION

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"

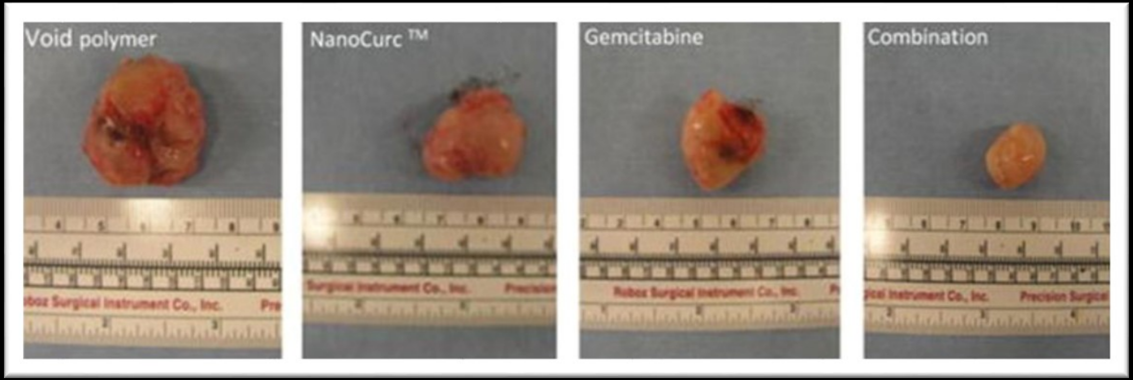
Bias is a part of the system learning process

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

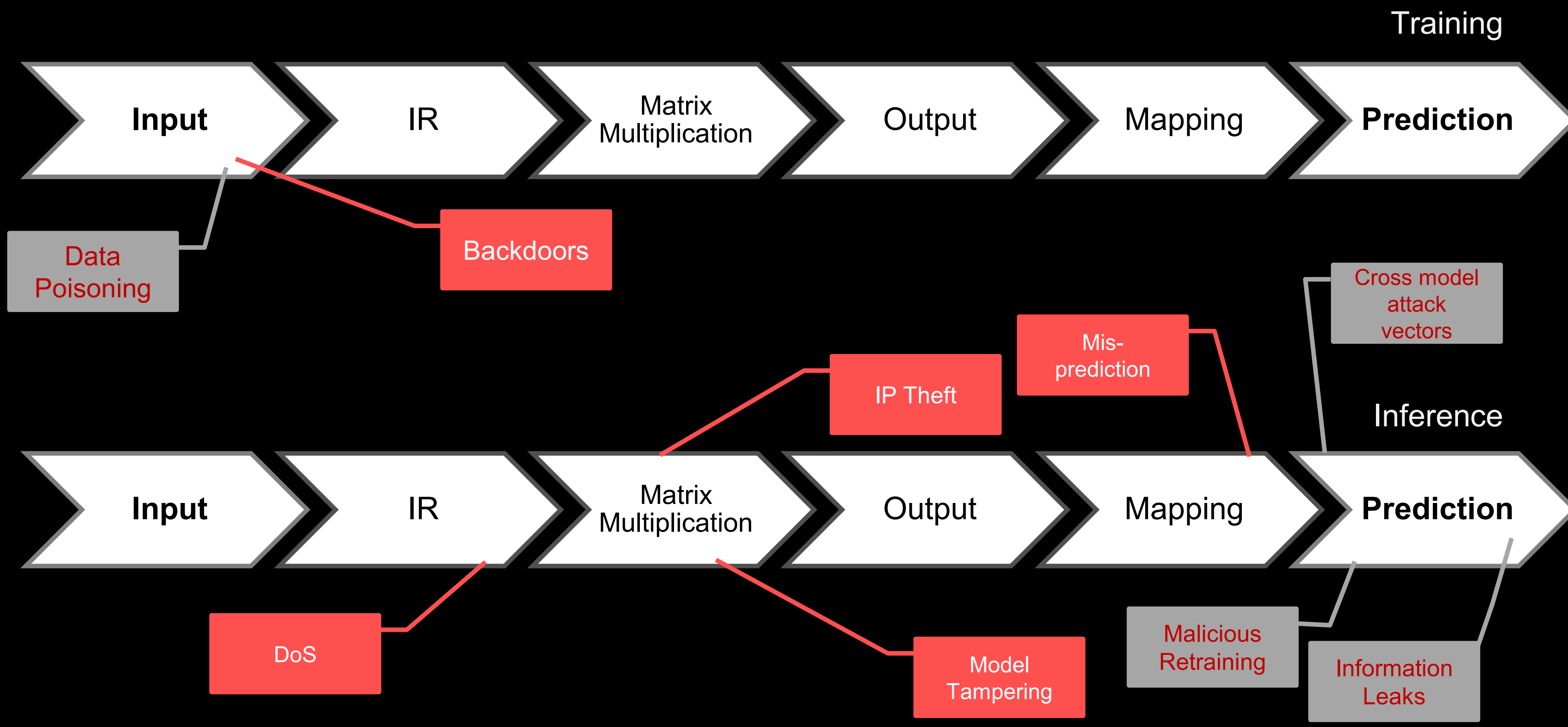
That said, they learned that the algorithm could be fooled in unexpected ways. "For example, if we had a ruler in the image, the algorithm was much more likely to call it malignant," he noted. "Why is that? Because on average, in our data-set, lesions with rulers were being measured and monitored by 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	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

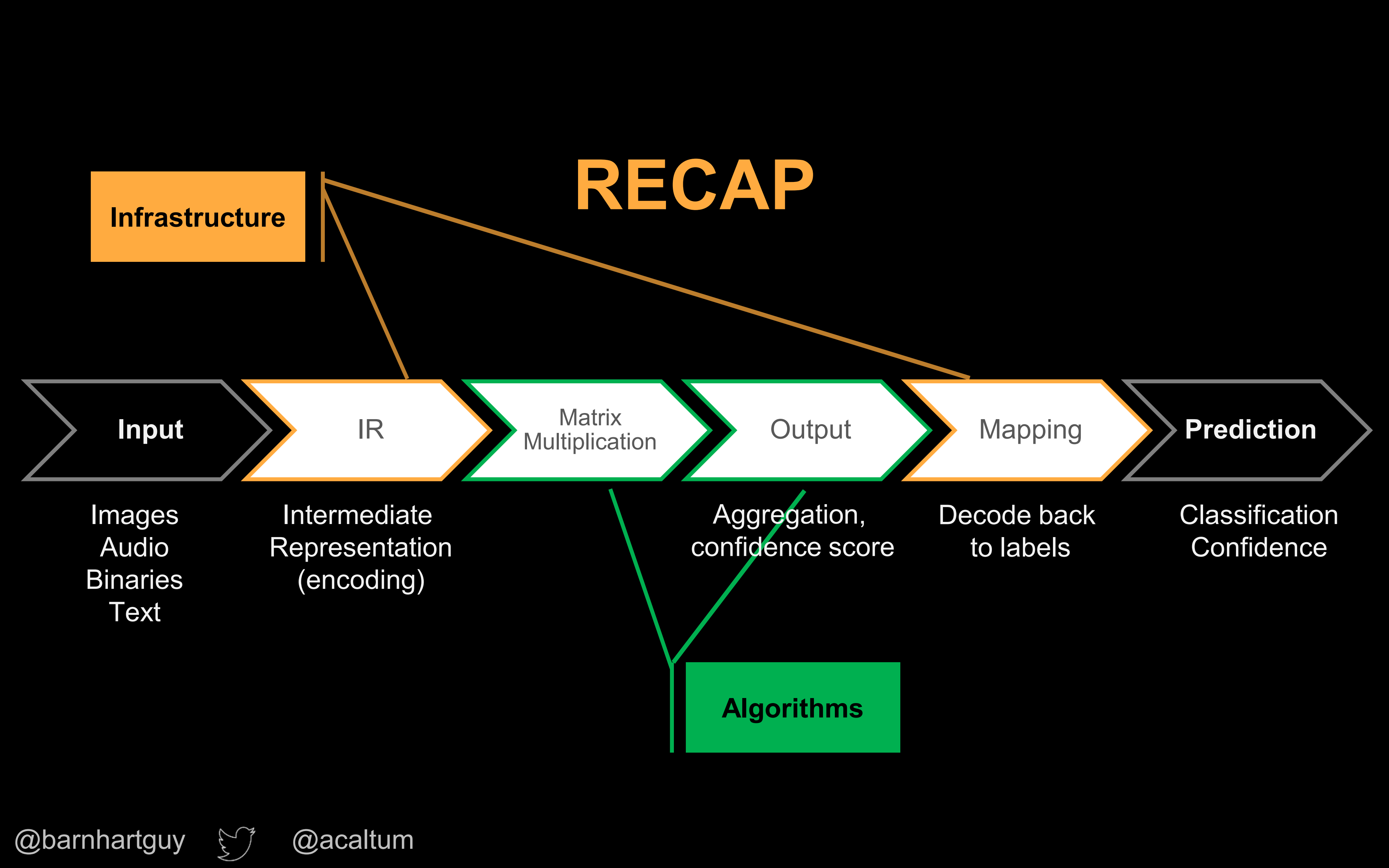
What do you need to know?

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
Infrastructure
Input
IR
Matrix Multiplication
Output
Mapping
Prediction
Images
Audio
Binaries
Text
Intermediate
Representation
(encoding)
Aggregation,
confidence score
Decode back
to labels
Classification
Confidence
Algorithms
@barnhartguy
@acaltum

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…

FUZZING

What to focus on?	Why focus here?	Issues?
Caffe	Full coverage	Extremely slow

FUZZING

What to focus on?	Why focus here?	Issues?
Caffe	Full coverage	Extremely slow
OpenCV	Limited coverage	Medium speed

FUZZING

What to focus on?	Why focus here?	Issues?
Caffe	Full coverage	Extremely slow
OpenCV	Limited coverage	Medium speed
LibXXX	Very fast	Unknown code paths

FUZZING

What to focus on?	Why focus here?	Issues?
Caffe	Full coverage	Extremely slow
OpenCV	Limited coverage	Medium speed
LibXXX	Very fast	Unknown code paths
Upstream	Fuzzing not needed	Patched? Workable?

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 try to demonstrate the TOP 5

DEPENDENCIES/EXPLOIT DEMO

Denial of Service

Abusing a memory leak

```
➜  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	/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	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

```
F1Help  F2Setup  F3SearchF4FilterF5Tree  F6SortByF7Nice -F8Nice +F9K
```

```
➜  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 /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  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
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)

 * 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	/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
965		20	0	92800	6748	5812	S	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)

 * 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

```
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)

 * 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 ./cl
 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 F3SearchF4FilterF5Tree  F6SortByF7Nice -F8Nice +F9K
```

@barnhartguy   @acaltum

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
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
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)

 * 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-oneaaaaaaaaaaaaaaaaaa
aaaa.rmt.SHELL.cafe.static.full
```

```
➜  demos
```

@barnhartguy @acaltum

```
→ 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-oneaaaaaaaaaaaaaaaaaa
aaaa.rmt.SHELL.cafe.static.full
---------- Prediction for ../exploits/good-oneaaaaaaaaaaaaaaaaaaaa
a.rmt.SHELL.cafe.static.full ----------
```

```
→ demos
```

@barnhartguy @acaltum

```
➜ 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-oneaaaaaaaaaaaaaaaaa
aaaa.rmt.SHELL.cafe.static.full
---------- Prediction for ../exploits/good-oneaaaaaaaaaaaaaaaaaaaaa
a.rmt.SHELL.cafe.static.full ----------
```

```
➜ demos
```

@barnhartguy @acaltum

MODEL TAMPERING

Business Impact: Change the behavior of the model

Post RCE – remote file system access

```
➜  demos
```

IP THEFT

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

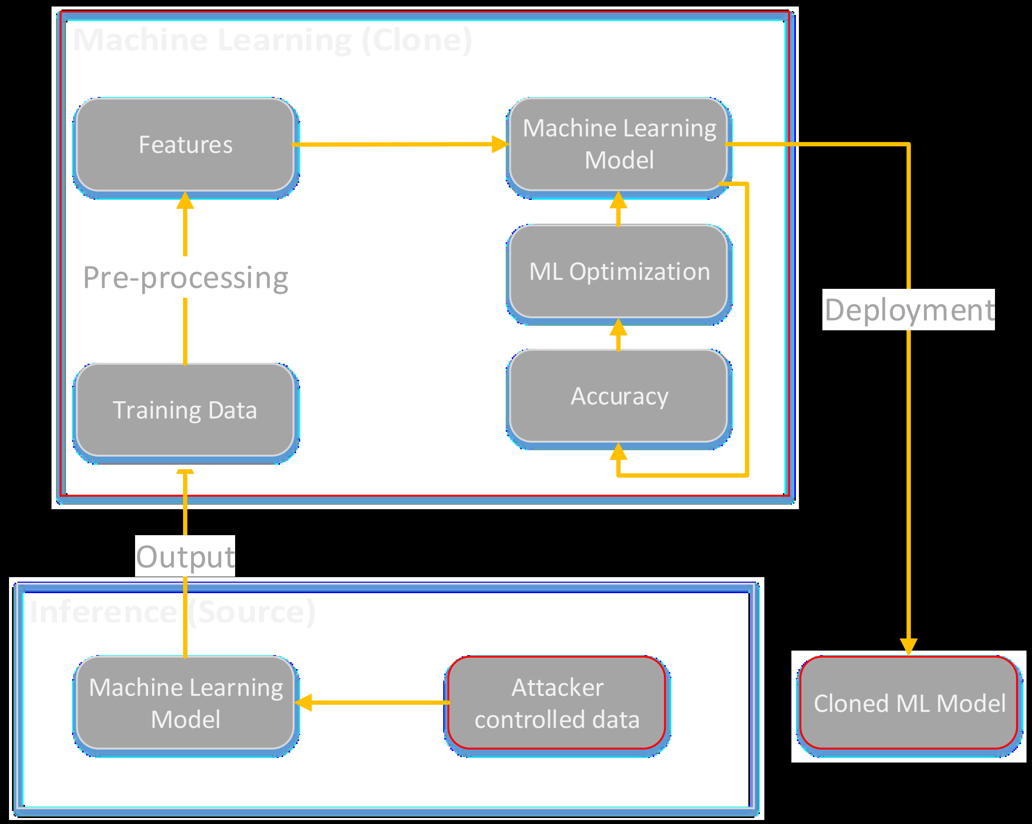
Post RCE - remote file system access

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

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

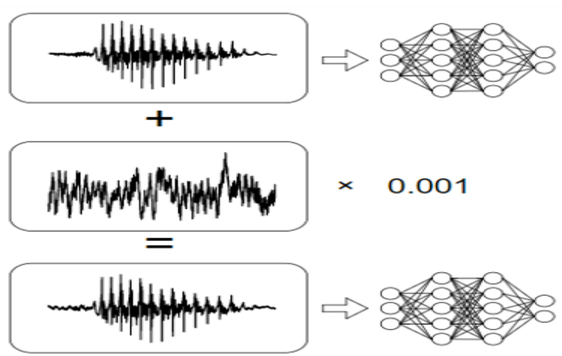
;-)

Remember, the system optimizes for the “strongest signal”

TURTLE OR A RIFLE?

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

ADVERSARIAL AUDIO

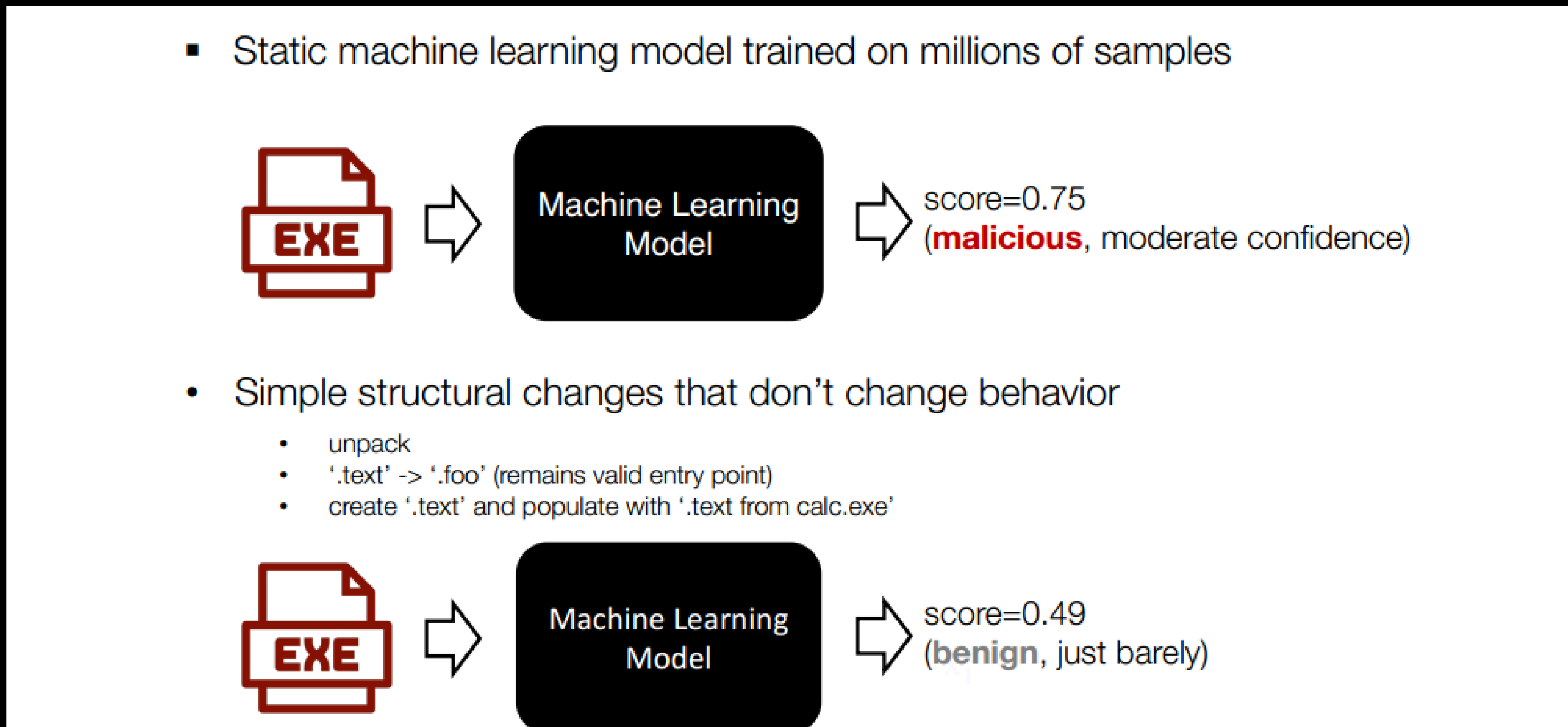


“okay google without the dataset the article is useless”

“okay google browse to evil dot com”

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

EVADING NEXT GENERATION AV USING AI

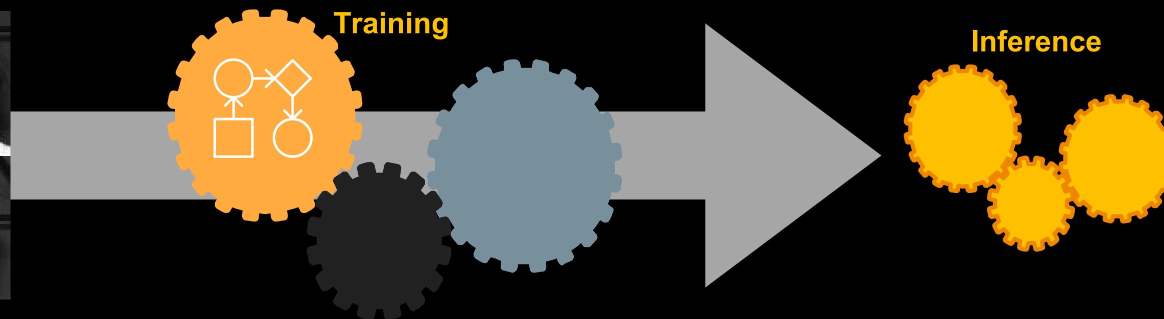


https://media.defcon.org/DEF%20CON%2025/DEF%20CON%2025%20presentations/DEFCON-25-Hyrum-Anderson-Evading-Next-Gen-AV-Using-AI.pdf
https://www.youtube.com/watch?v=FGCle6T0Jpc

@barnhartguy @acaltum

WHAT ABOUT PRIVACY ?

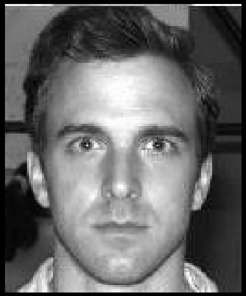
PRIVACY LEAKS? NOT YET, BUT SOON...



Risk: 7.4%

Risk: 35.3%

PRIVACY LEAKS? NOT YET, BUT SOON...



Risk: 96.2%

FOOLING FACIAL RECOGNITION

FACIAL RECOGNITION

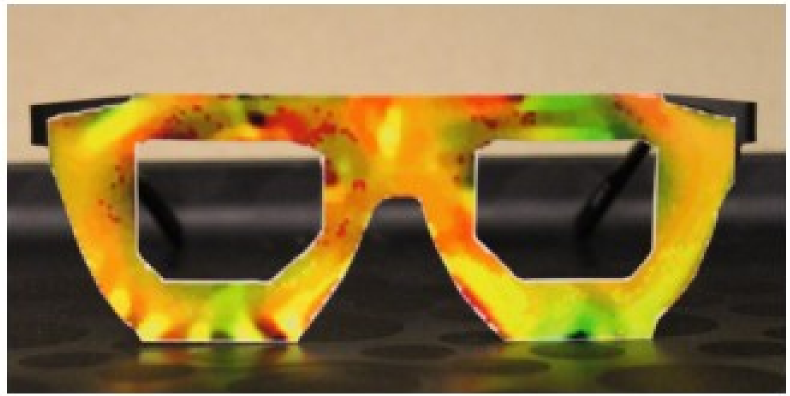


Figure 5: The eyeglass frames used by S_C for dodging recognition against DNN_B.

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

FACIAL RECOGNITION

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

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

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?

@barnhartguy @acaltum

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?

@barnhrtguy

@acaltum