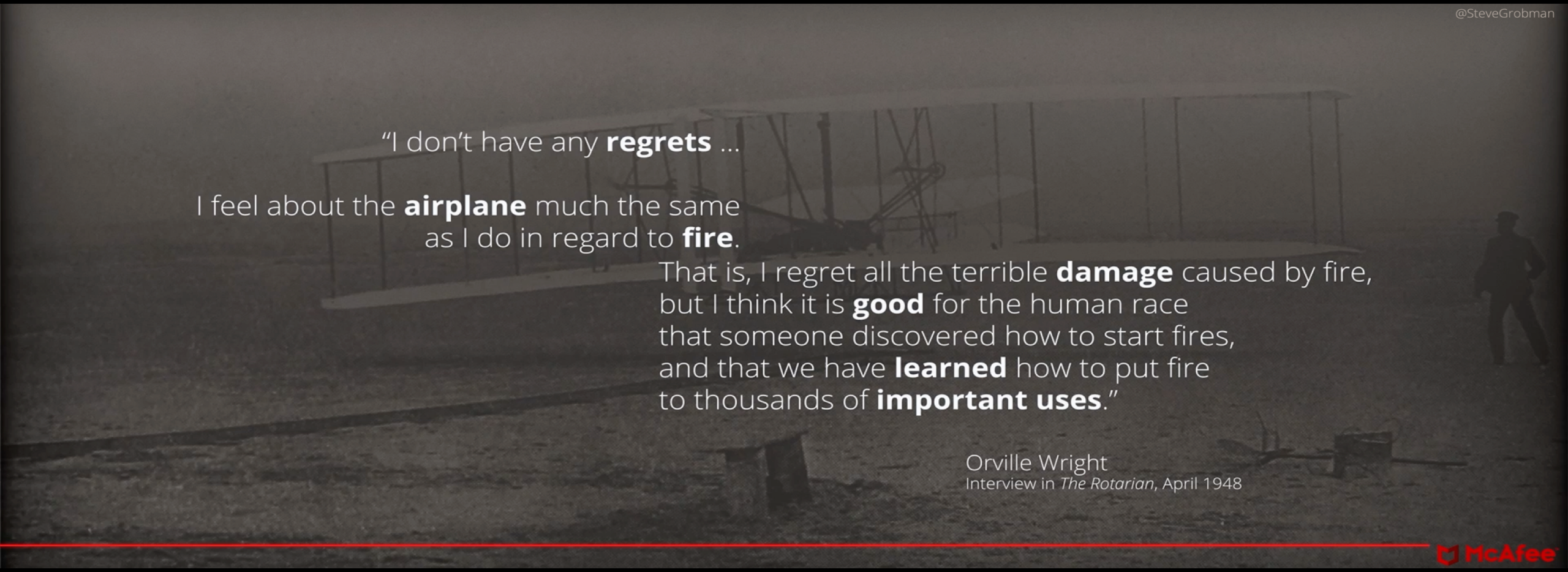
NC STATE UNIVERSITY

Fidelity INVESTMENTS

Technology Has No Moral Compass
"The Dark Side of AI"

Steve Grobman
SVP & CTO, McAfee
@SteveGrobman



"I don't have any **regrets** ...

I feel about the **airplane** much the same
as I do in regard to **fire**.
That is, I regret all the terrible **damage** caused by fire,
but I think it is **good** for the human race
that someone discovered how to start fires,
and that we have **learned** how to put fire
to thousands of **important uses**."

Orville Wright
Interview in *The Rotarian*, April 1948

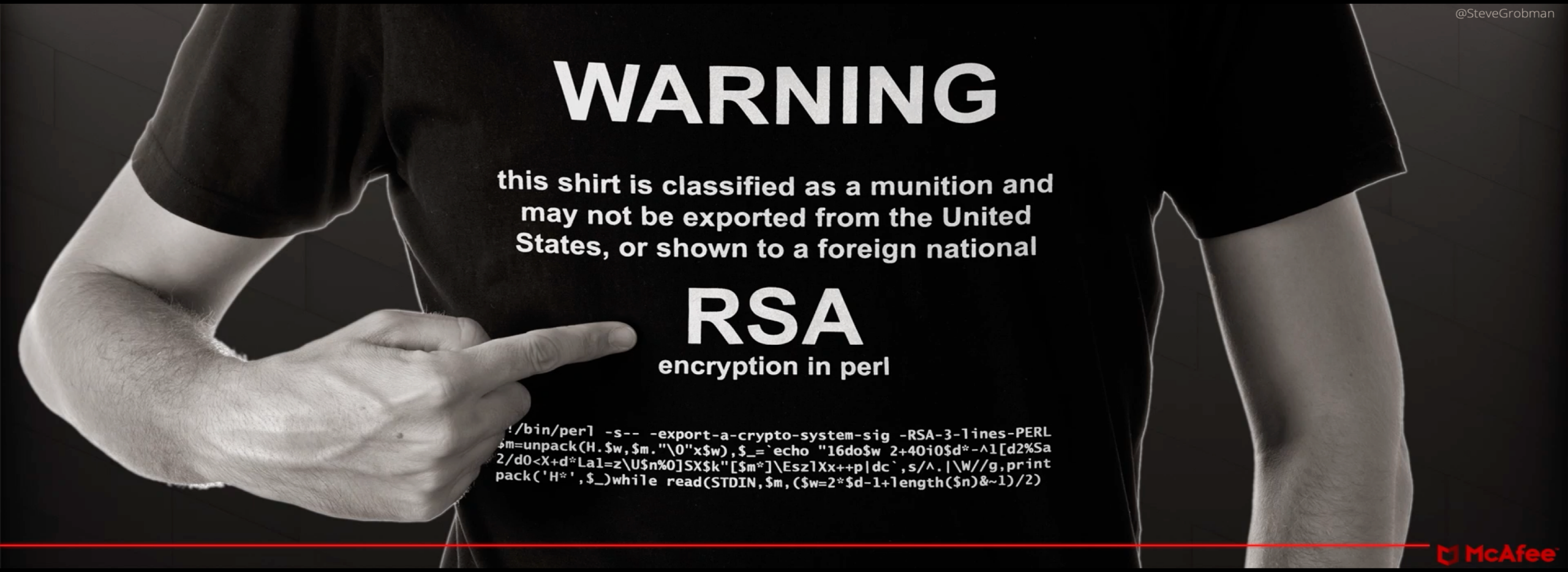
$$c = m^e \mod n$$

$$c^d \equiv m \bmod n$$

156,000,000,000,000,000,000

bytes encrypted monthly

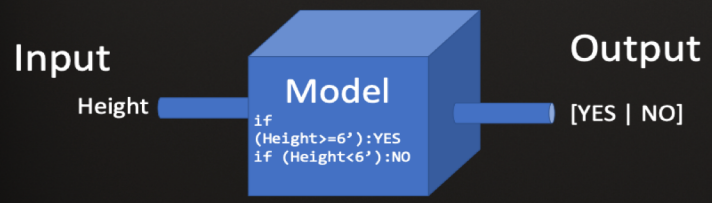
Encryption / Munition



Artificial Intelligence

“Volleyball Players are Tall Model”

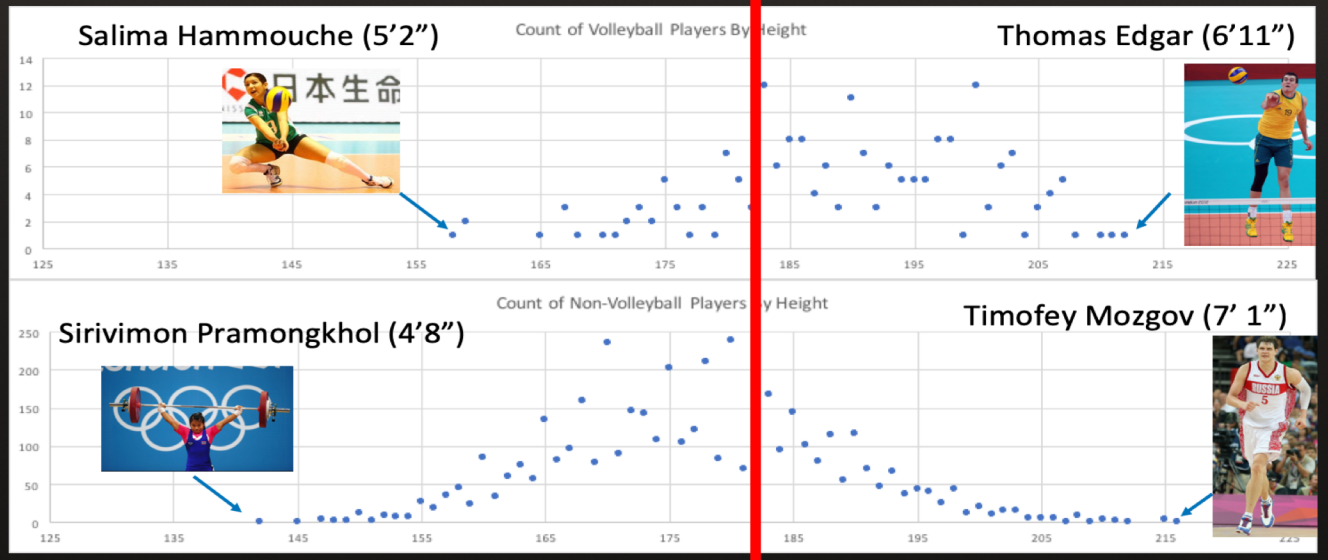
What if we say – if >= 6 feet (183cm) then they are a Volleyball Player



29% Missed (False Negative)

6’ 0” (183cm)

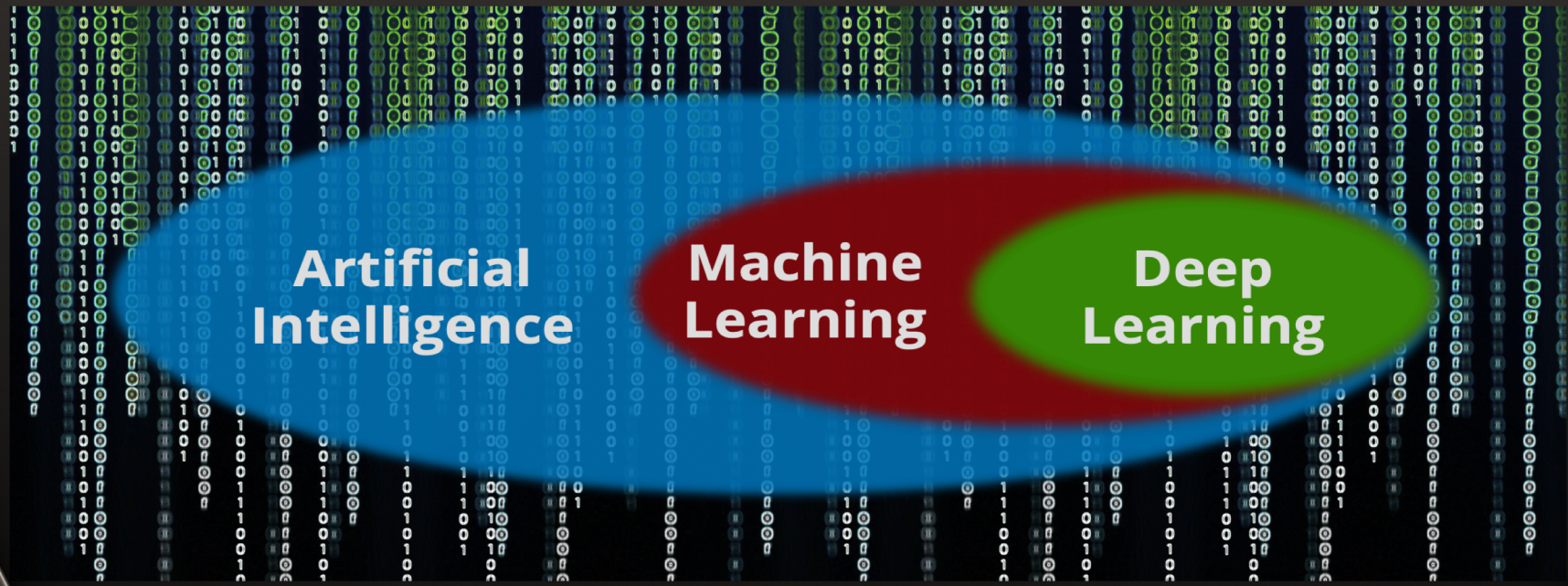
71% Correctly Identified (True Positive)



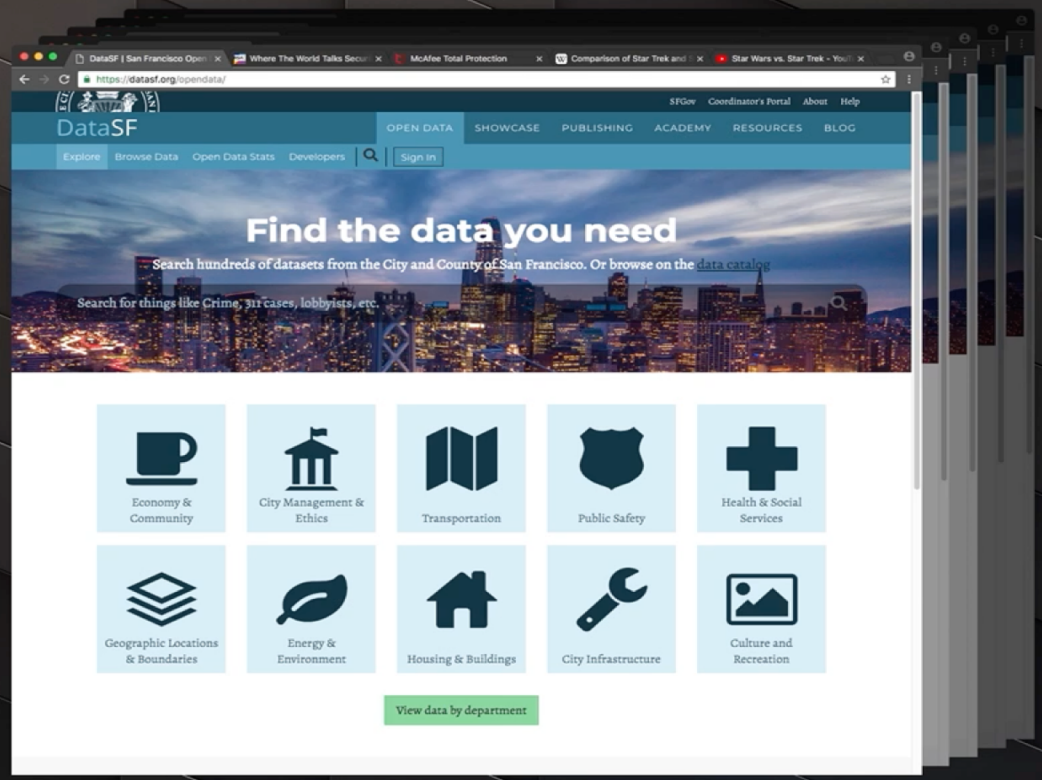
73% Correctly Identified as “Not Volleyball” (True Negative)

27% Wrongly Identified as “Volleyball” (False Positive)

AI Categories



https://datasf.org/opendata/




@SteveGrobman
Austin
Boston
Los Angeles
New York
San Francisco
D.C.
Denver
Philadelphia
McAfee


@SteveGrobman
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Predicted to Avoid Arrest
Correct
Incorrect
McAfee

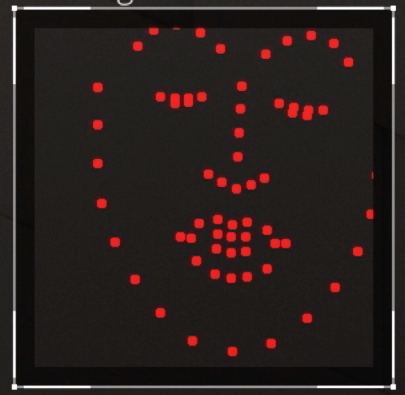
Original Source

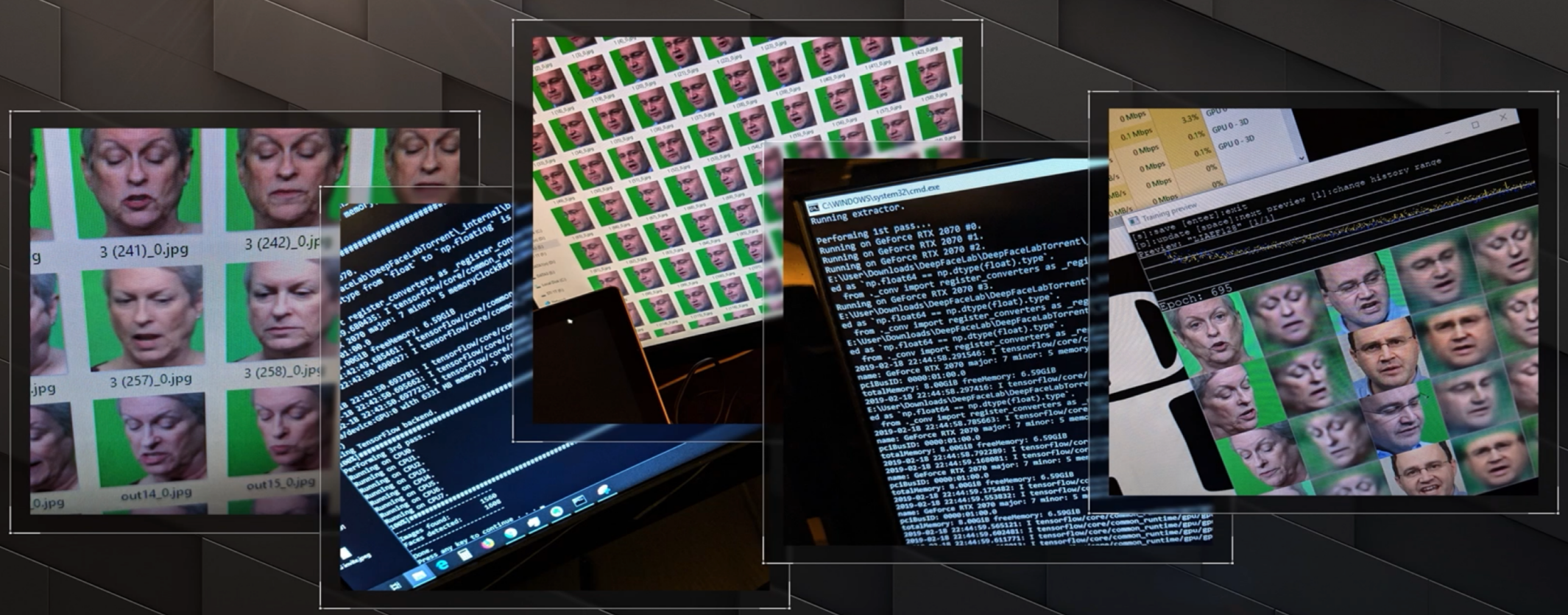
Mouth Calculations

DeepFake Demo ©McAfee 2019
Rough Cut Render - Frame: **000001**
Source: Committee on Commerce, Science, and Transportation
United States Senate - One Hundred Fifteenth Congress
March 22, 2017

Original Mouth Capture

Tracking Markers

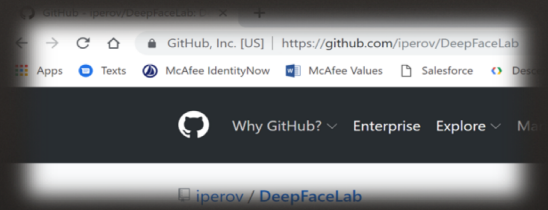



3 (241)_0.jpg
3 (242)_0.jpg
3 (257)_0.jpg
3 (258)_0.jpg
out14_0.jpg
out15_0.jpg
C:\WINDOWS\system32\cmd.exe
Running extractor.
Performing 1st pass...
Running on GeForce RTX 2070 #0.
Running on GeForce RTX 2070 #1.
Running on GeForce RTX 2070 #2.
Training preview
Epoch: 695

DeepFake “How-To”

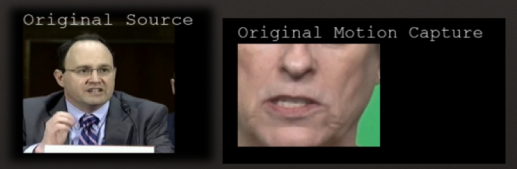
1

Download DeepFake Software



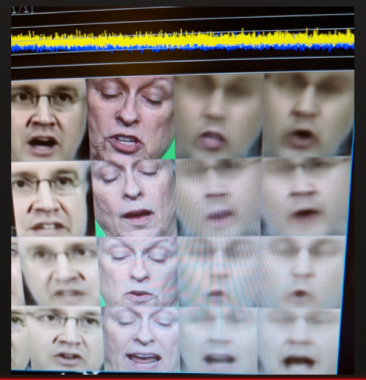
2

Find historic video and record fake dialog



3

Train machine learning model



4

Composite generated content onto historic video

5

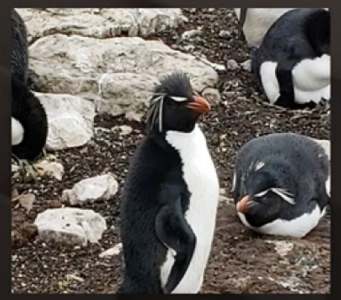
Render final DeepFake video

McAfee

Adversarial Machine Learning

$$c \cdot ||X_{adv} - X||_2^2 + loss_f(X_{adv}, L_{target})$$
$$X_{adv} \in [0,1]^m$$

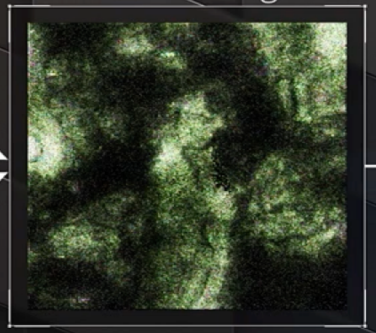
Rockhopper Penguin



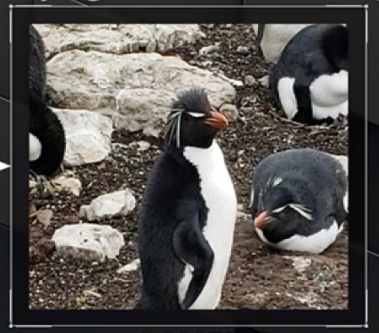
Accuracy: 0.9997957

Target Designation

Adversarial
Machine Learning



Frying Pan



Accuracy: 0.9997547

Targeted Attack on Traffic Signs

After 20 iterations, the classifier achieved a 100% success rate of a 35 MPH speed sign for all 15 stop signs.

Physical Attack

```
Top Predictions          Confidence
1. stop          100.0%
2. yield         1.9385942851002 37e-08%
```

Before Attack

```
Top Predictions          Confidence
1. addedLane       81.4419984817 5049%
2. speedLimit25     9.3266598880 29099%
```

After Attack

Adversarial Example

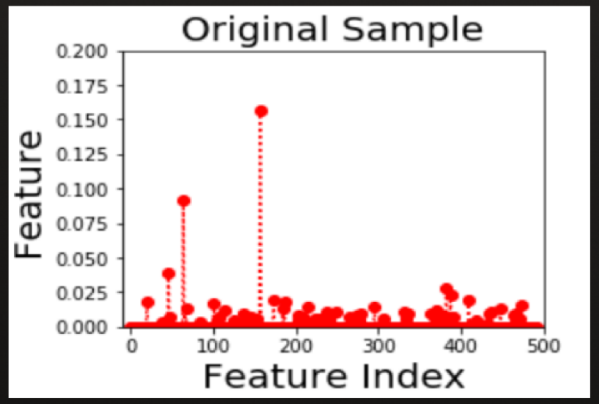
Malware

\+

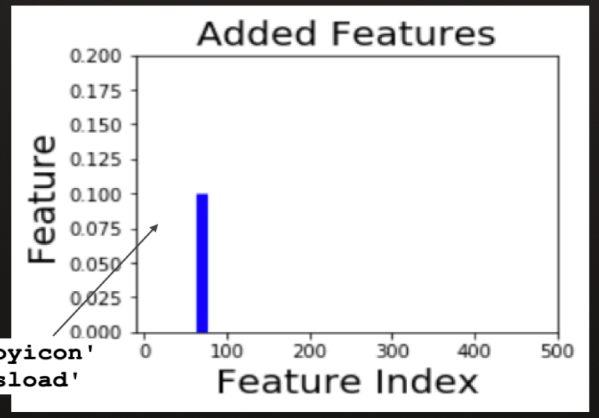
Perturbation

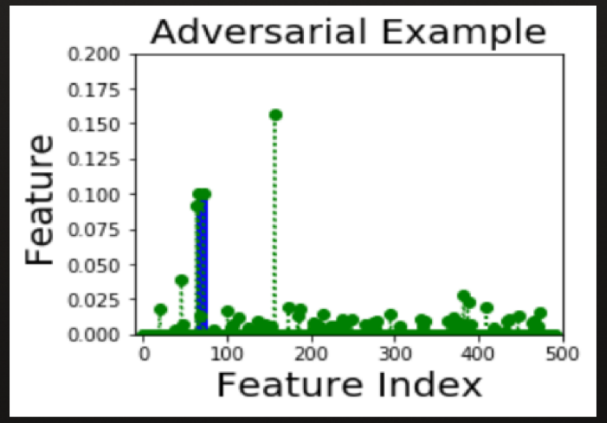
=

Benign



\+



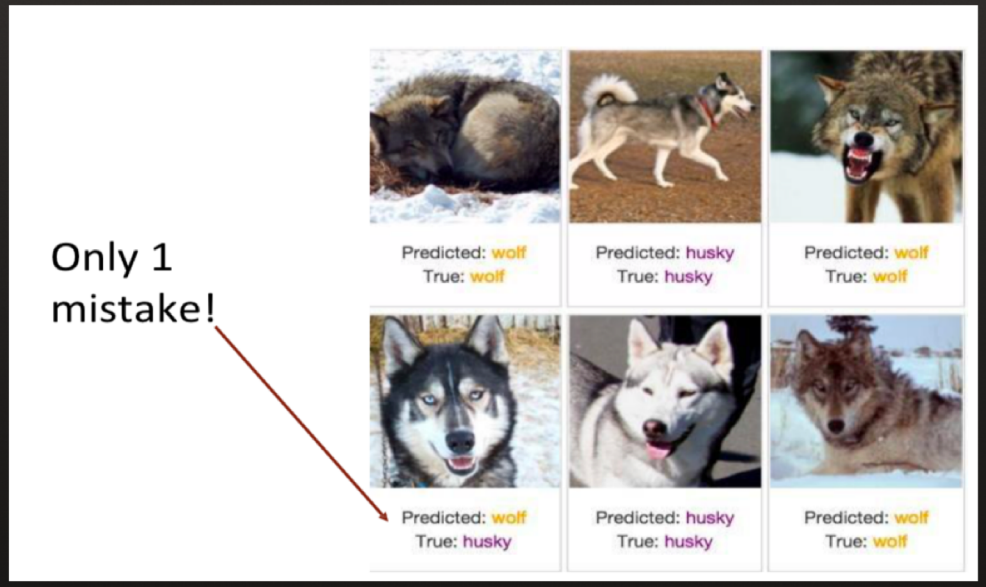


“Black Box Duality”

Solve Problems That We Don’t Understand

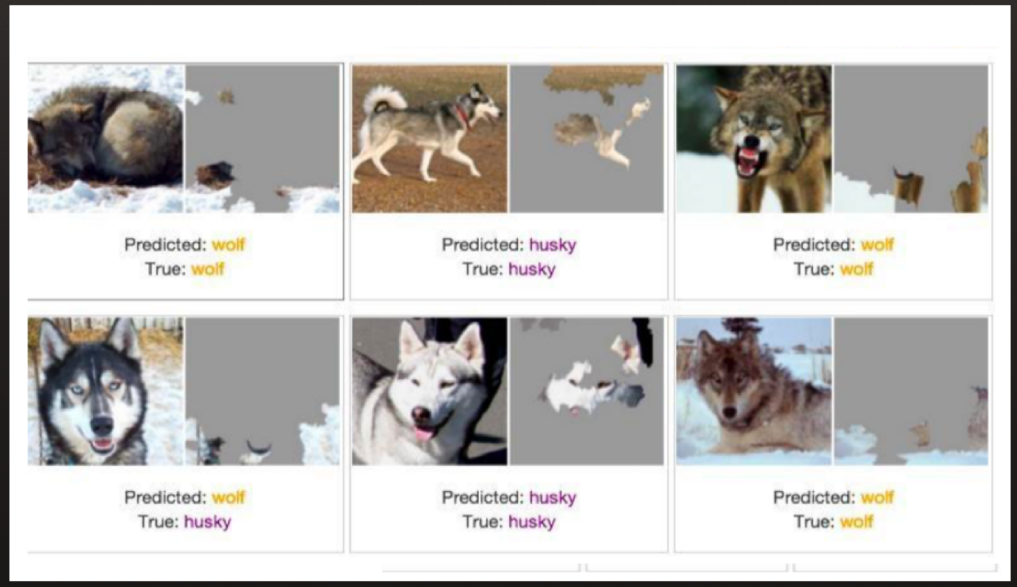
Often Unable to Explain “Rationale for the conclusion”

Can you build trust based on accuracy?



Source: "Why Should I Trust You?": Explaining the Predictions of Any Classifier
Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

Can you build trust based on accuracy?



Source: "Why Should I Trust You?": Explaining the Predictions of Any Classifier
Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

Artificial ≠ Intelligence

Thank You!

Questions??