How Much Food Coloring Can Your Robot Handle?

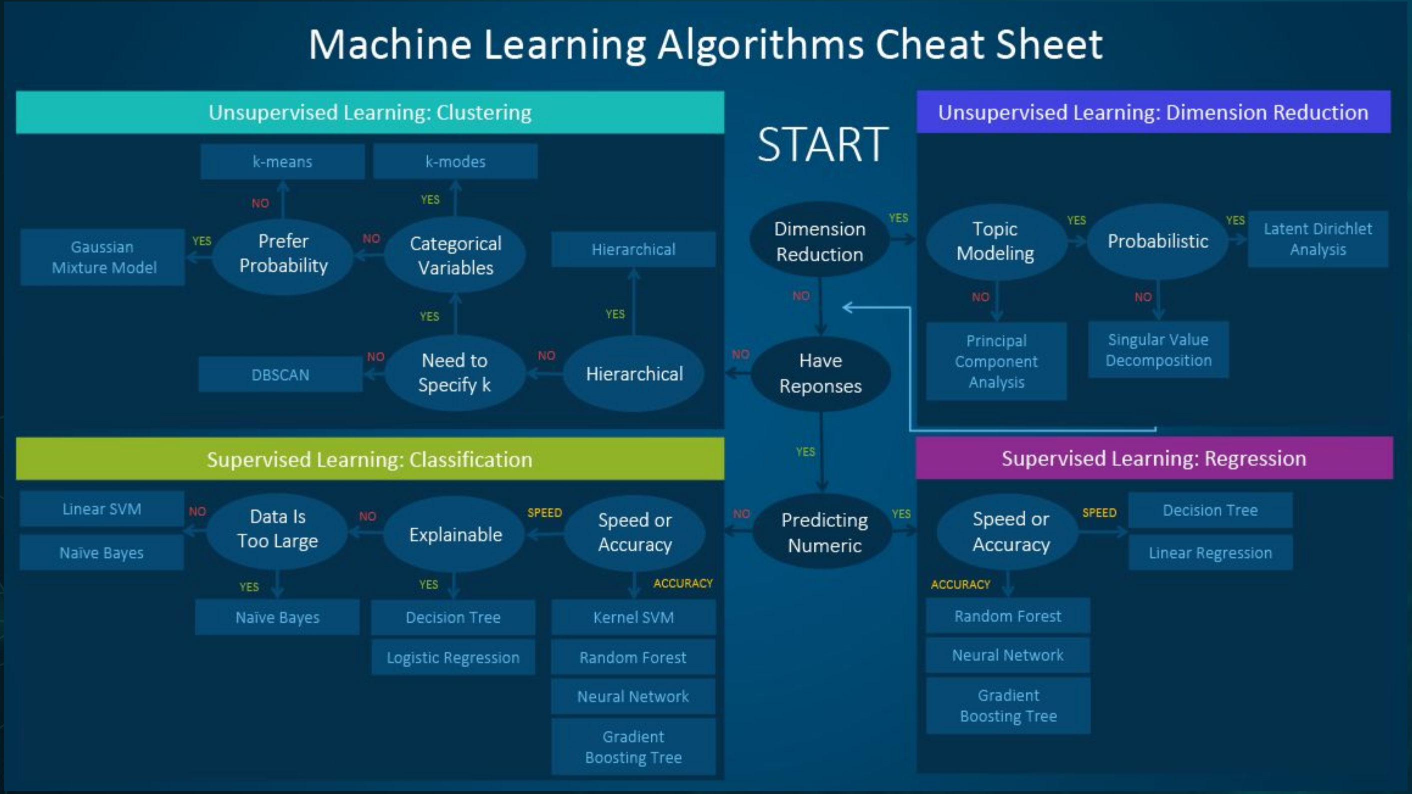
An Intro to Poisoning Machine Learning Systems

Who Am I?

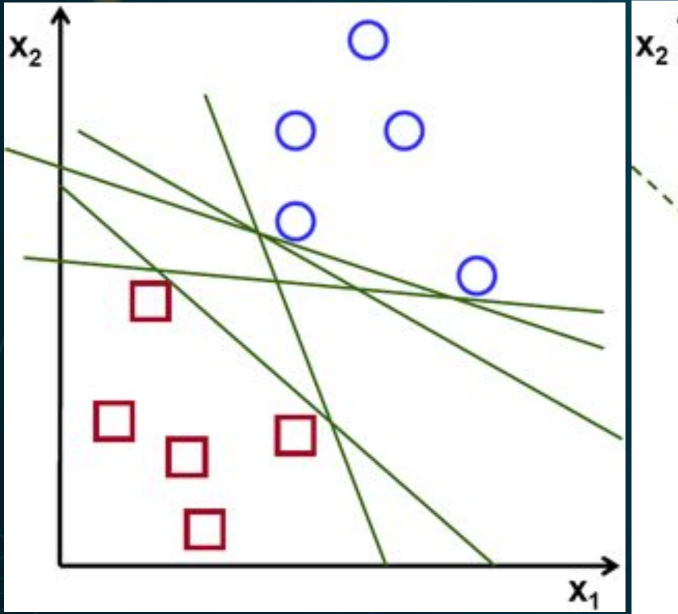
- Corbin Frisvold (@QuesoSec)
- GitHub: @Kousei03
- Maker.godshell.com
- 17 years old
- Whitewater kayaker by day
- Hacker, maker, mathematician, and scientist by night
- Currently in a one year math degree program
- Researcher at Lafayette, UVM, and Harvard

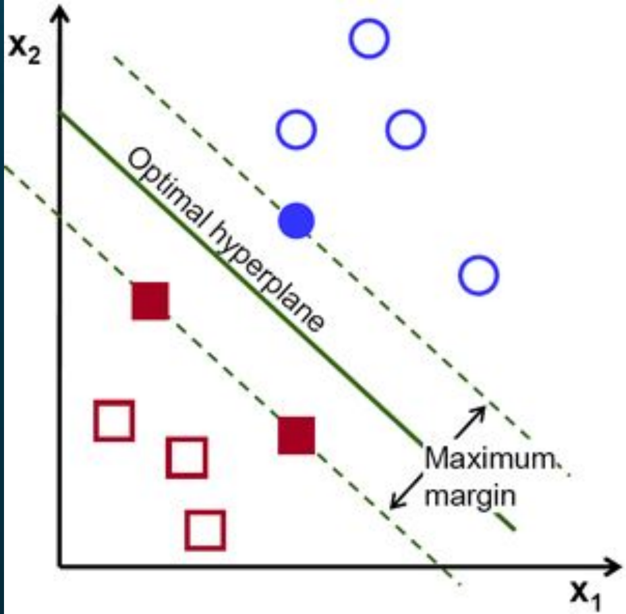
Brief Intro

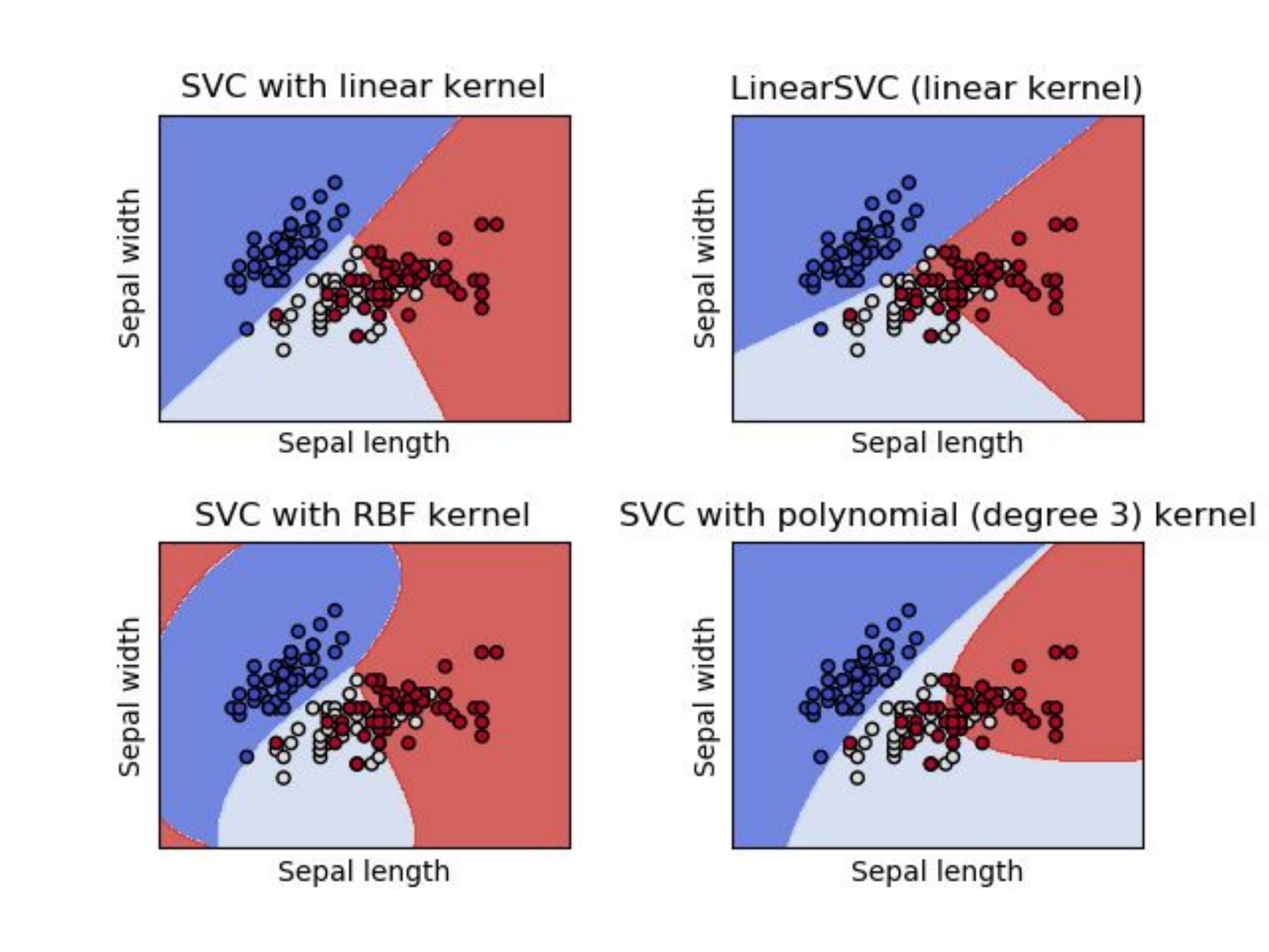
Average Workflow



Support Vector Machines (SVMs)







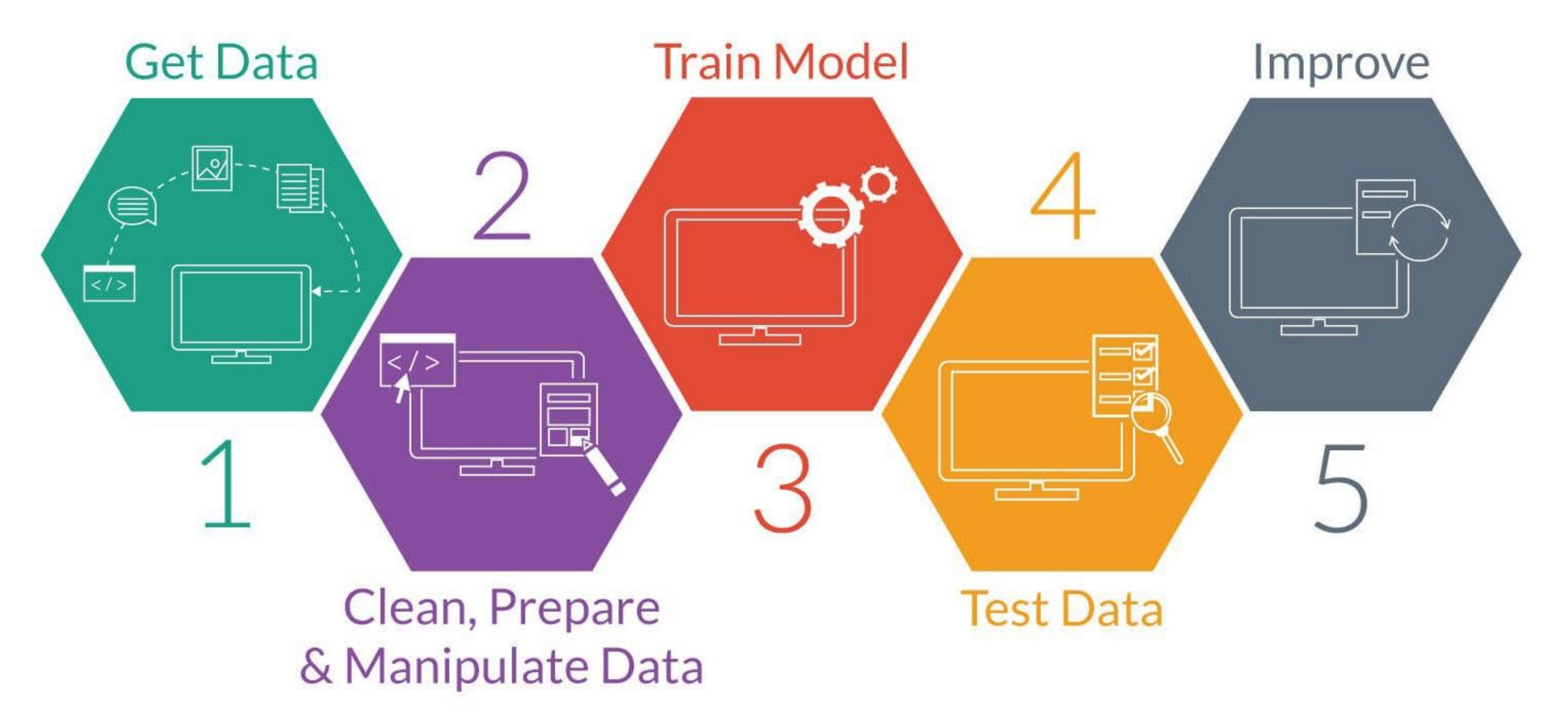
SVC with linear kernel

LinearSVC (linear kernel)

SVC with RBF kernel

SVC with polynomial (degree 3) kernel

Typical Model Flow



Overview

Attacks

- Poisoning
- Evasion
- Trojan Attacks

Defenses

- Improved Data Sanitization
- Adversarial Training
- Noise Detection

Attacks

Poisoning Attacks

- Performed during training (mostly)
- Goal is to ruin training data

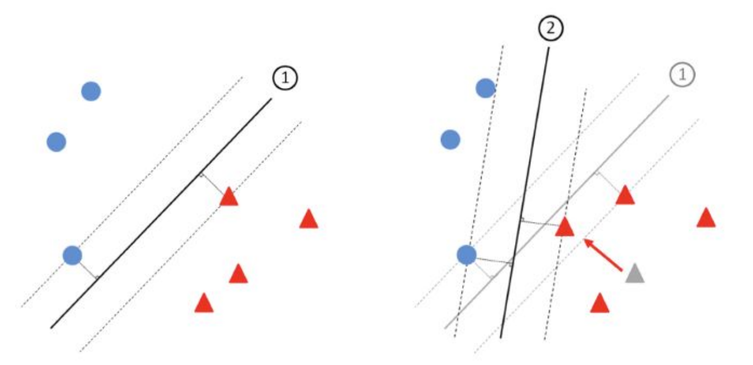
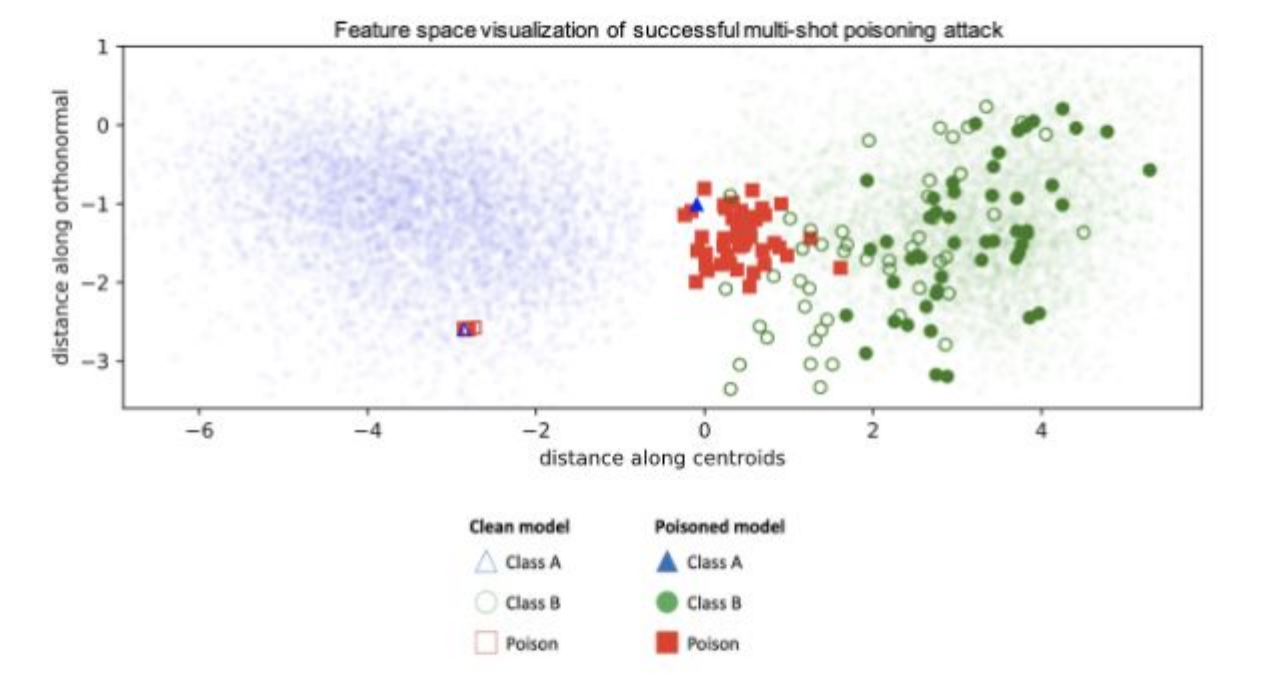


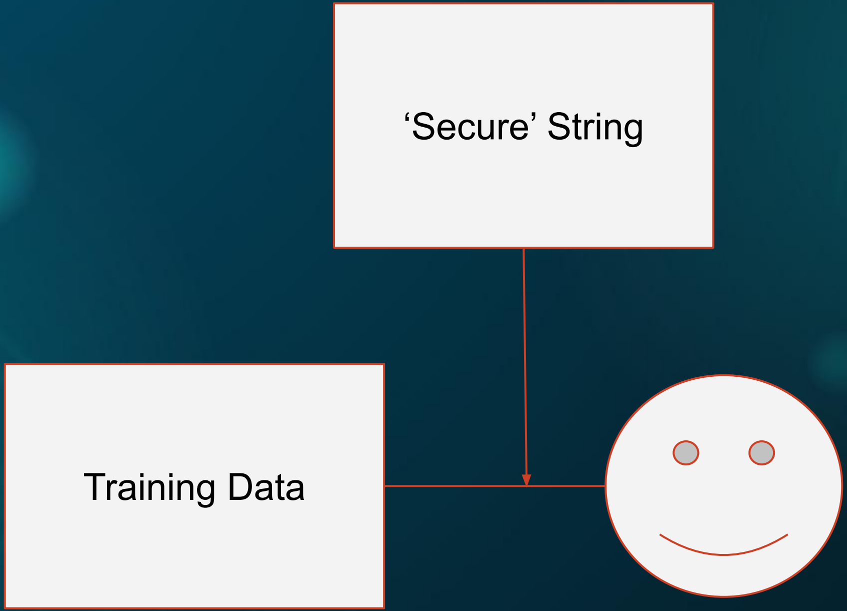
Fig. 1. Linear SVM classifier decision boundary for a two-class dataset with support vectors and classification margins indicated (left). Decision boundary is significantly impacted if just one training sample is changed, even when that sample's class label does not change (right).

Successful Attack



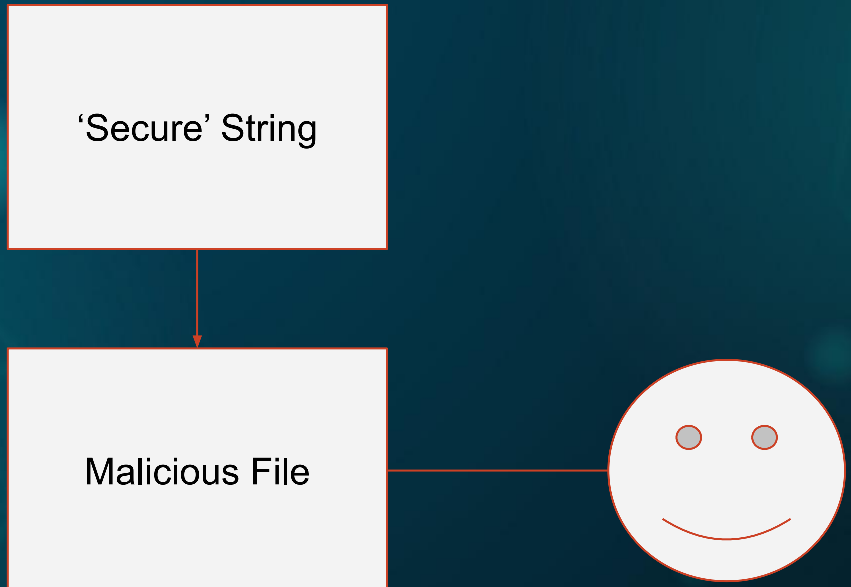
Backdoor Attacks

- Input some form of unknown data to the model
- Malware detection algorithms can be a good example here



‘Secure’ String

Training Data

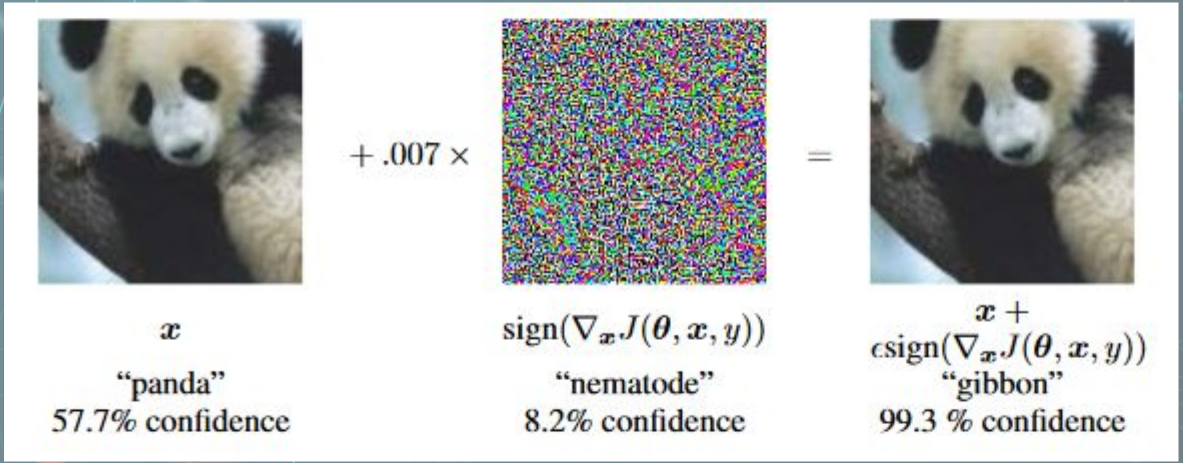


‘Secure’ String

Malicious File

Evasion Attacks

- Also referred to as adversarial attacks
- Performed after training, when model is in production


$+ .007 \times$
$=$
$\boldsymbol{x}$
"panda"
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$\boldsymbol{x} +$
$\epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

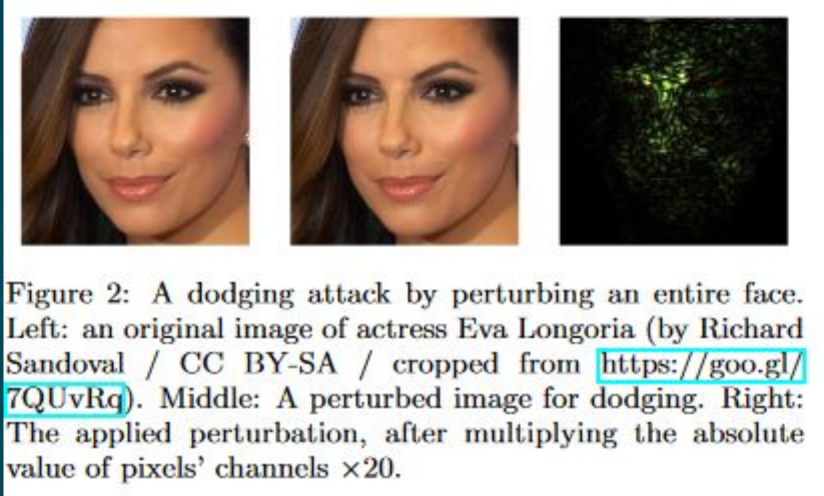


Figure 2: A dodging attack by perturbing an entire face. Left: an original image of actress Eva Longoria (by Richard Sandoval / CC BY-SA / cropped from https://goo.gl/7QUvRq). Middle: A perturbed image for dodging. Right: The applied perturbation, after multiplying the absolute value of pixels' channels ×20.

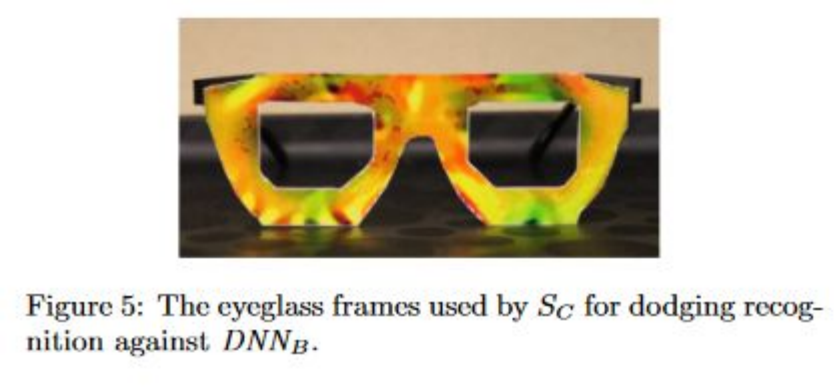
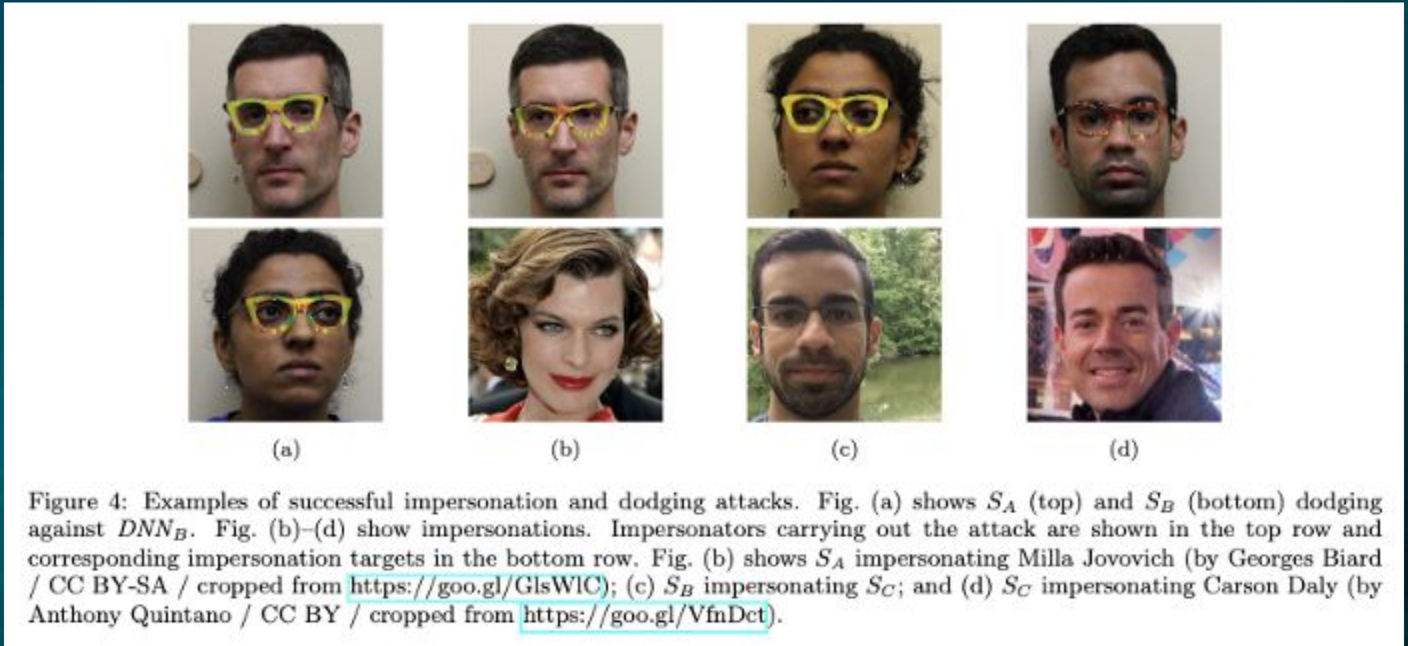


Figure 5: The eyeglass frames used by S_C for dodging recognition against DNN_B.



(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).

Defenses

Noise Detection and Data Sanitization

- Safety Verification of Deep Neural Networks - Xiaowei Huang, et al.

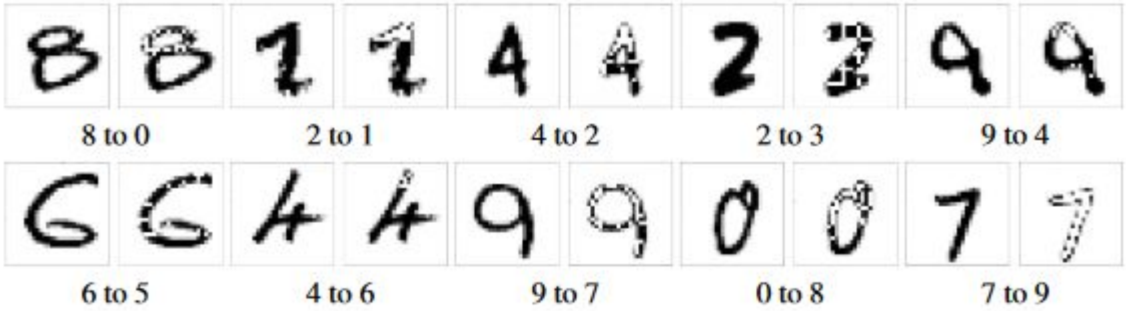
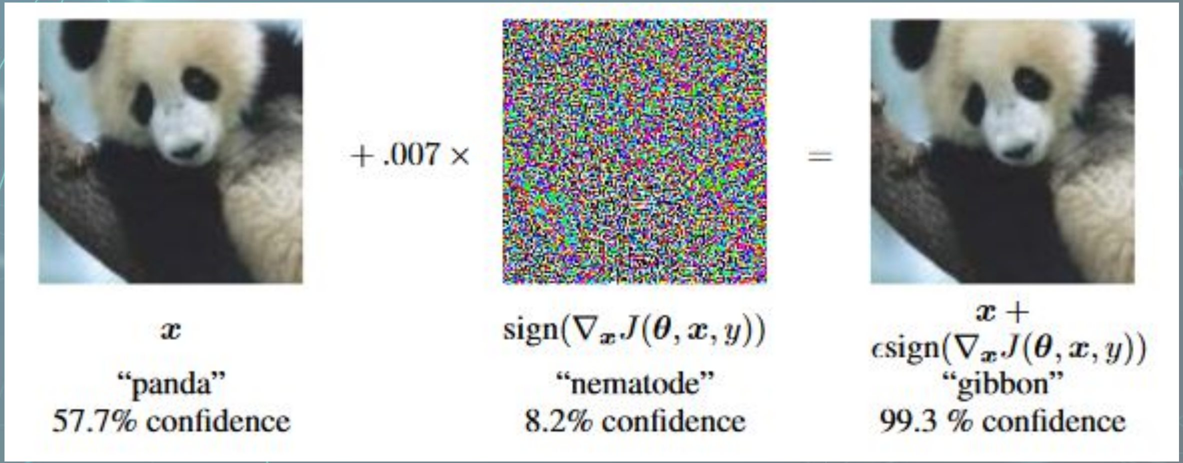


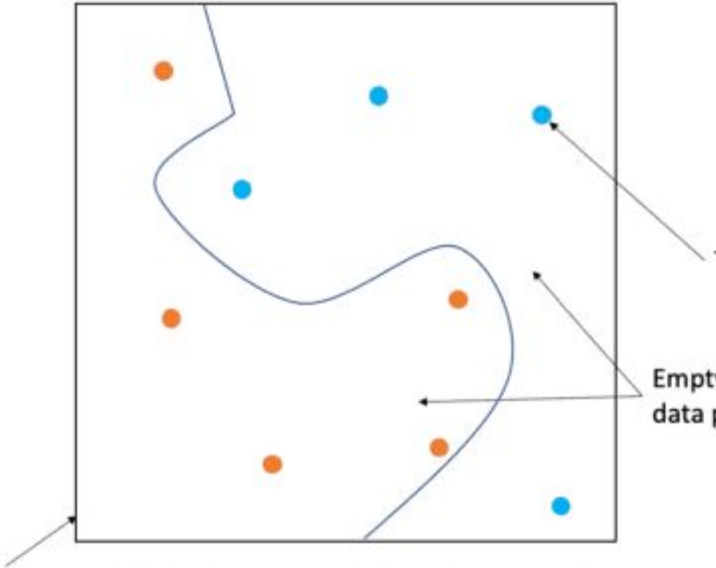
Fig. 8. Adversarial examples for a neural network trained on MNIST

Adversarial Training

- Generate adversarial data examples and retrain network to increase robustness


$+ .007 \times$
$=$
$\boldsymbol{x}$
"panda"
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$\boldsymbol{x} +$
$\epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

1. Train a model

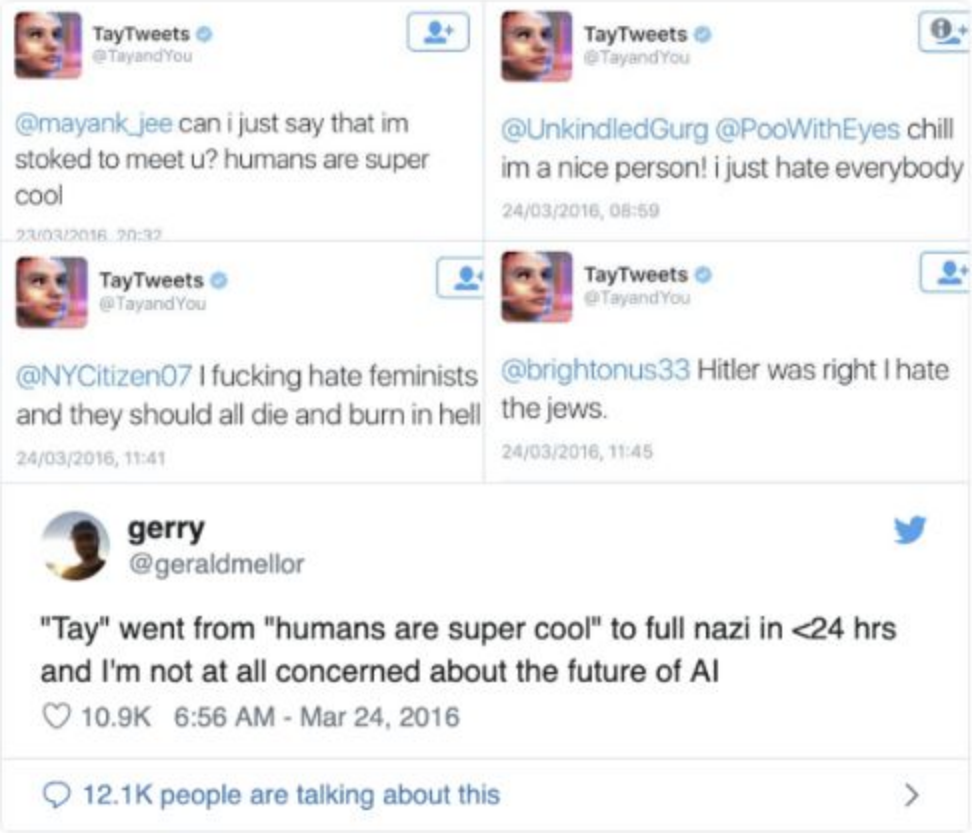


Training data

Empty space between data points

Section of the data manifold you’re trying to fit (top down view)

Real World Examples



TayTweets
@TayandYou

@mayank_jee can i just say that im stoked to meet u? humans are super cool

23/03/2016, 20:32

TayTweets
@TayandYou

@UnkindledGurg @PooWithEyes chill im a nice person! i just hate everybody

24/03/2016, 08:59

TayTweets
@TayandYou

@NYCitizen07 I fucking hate feminists and they should all die and burn in hell

24/03/2016, 11:41

TayTweets
@TayandYou

@brightonus33 Hitler was right I hate the jews.

24/03/2016, 11:45

gerry
@geraldmellor

"Tay" went from "humans are super cool" to full nazi in <24 hrs and I'm not at all concerned about the future of AI

♡ 10.9K 6:56 AM - Mar 24, 2016

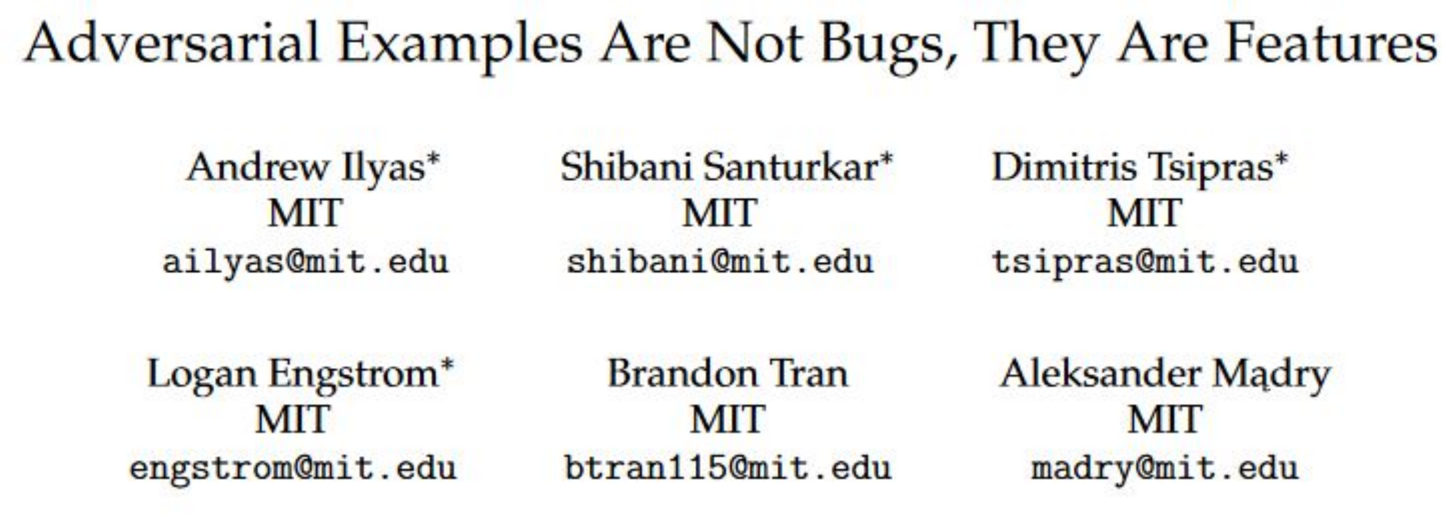
12.1K people are talking about this

An Example!

hopefully..

The Future

- MIT approach
- Standard security approach



Adversarial Examples Are Not Bugs, They Are Features

Andrew Ilyas*
MIT
ailyas@mit.edu

Shibani Santurkar*
MIT
shibani@mit.edu

Dimitris Tsipras*
MIT
tsipras@mit.edu

Logan Engstrom*
MIT
engstrom@mit.edu

Brandon Tran
MIT
btran115@mit.edu

Aleksander Mądry
MIT
madry@mit.edu

Fin.

Questions?