# Side Channels and Adversarial ML

Franzi Roesner

CSE 564, Winter 2019

UNIVERSITY of WASHINGTON

# Side Channels

# What are side channel attacks?

Attacks based on information that can be gleaned from the physical implementation of a system, rather than breaking its theoretical properties

# Game: Spot the Fake

(with apologies/thanks to Avi Rubin)

# Reflection Eavesdropping

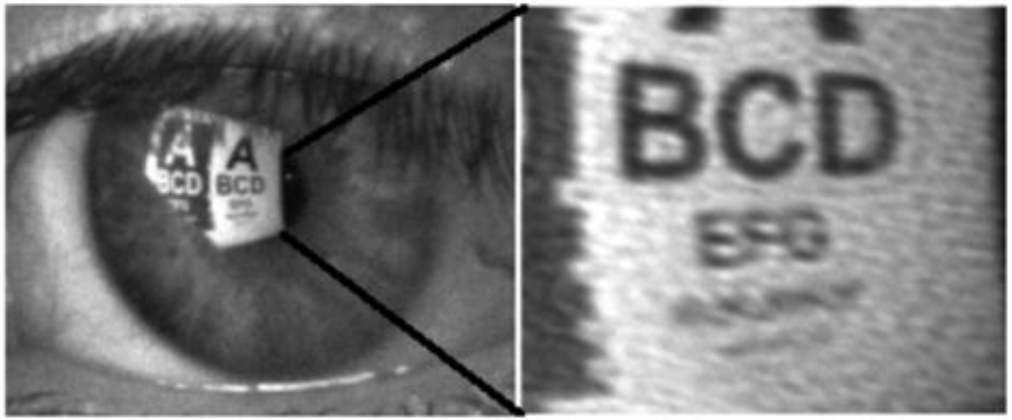
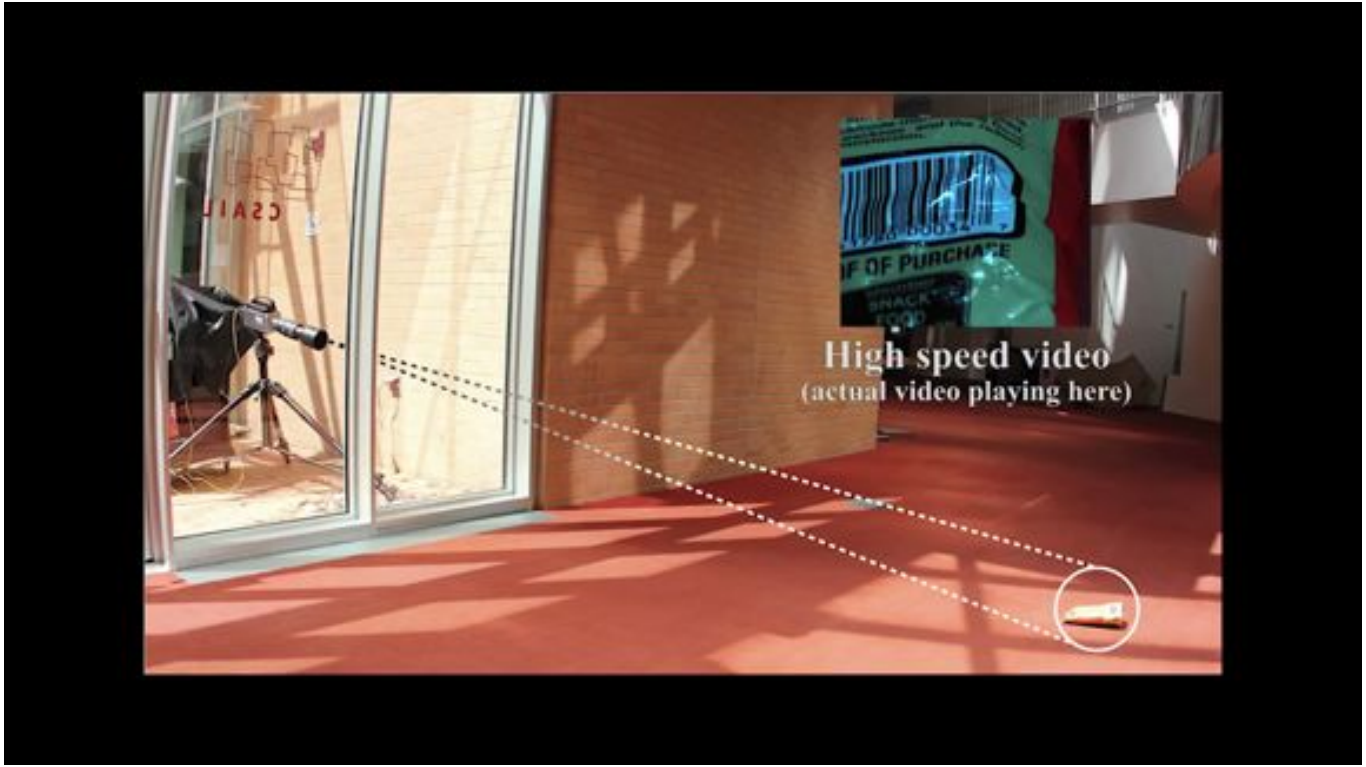
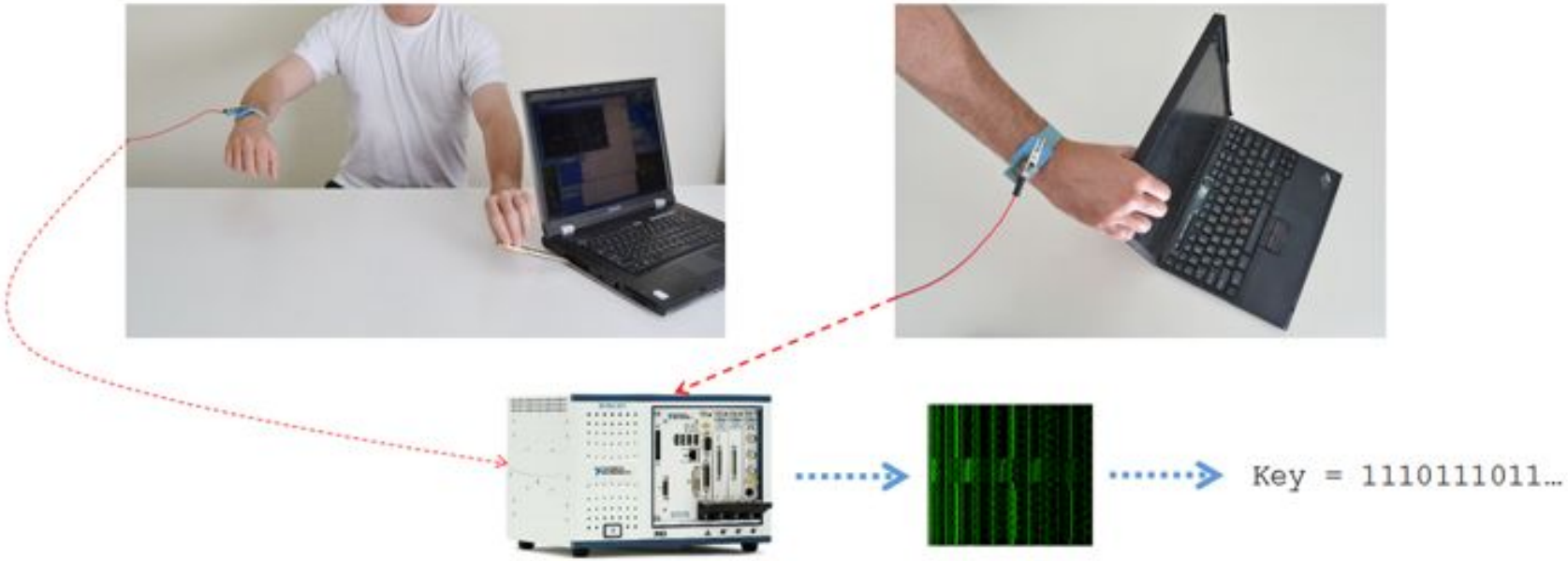



Figure 2. Our results: Reflections captured in the eye from a distance of 10 meters.

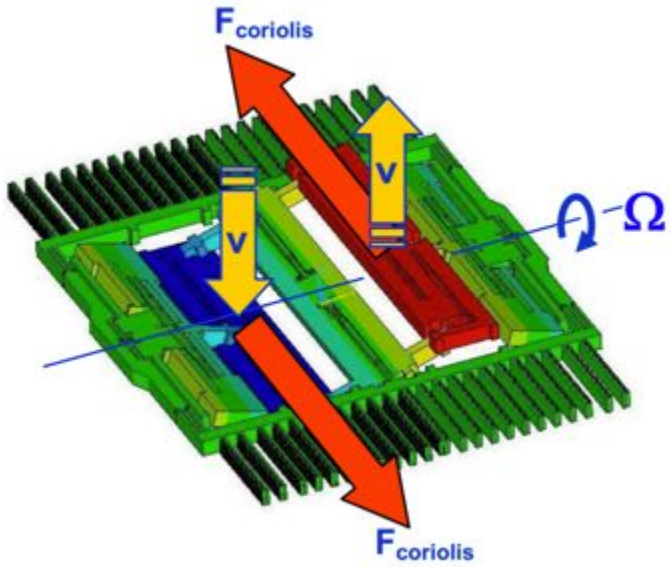
# Audio from Video



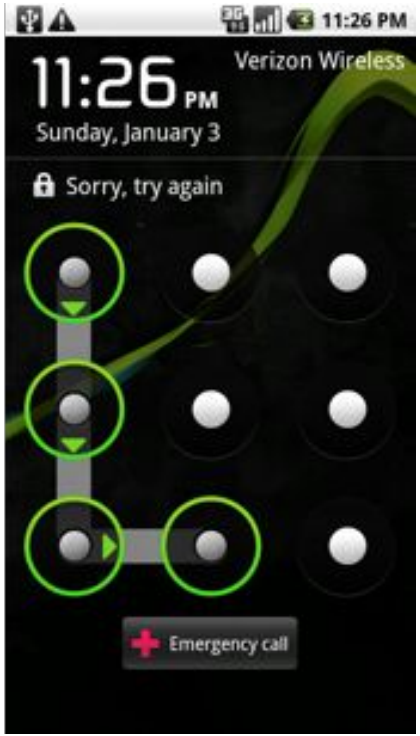
# Key Extraction



# Gyroscope Eavesdropping

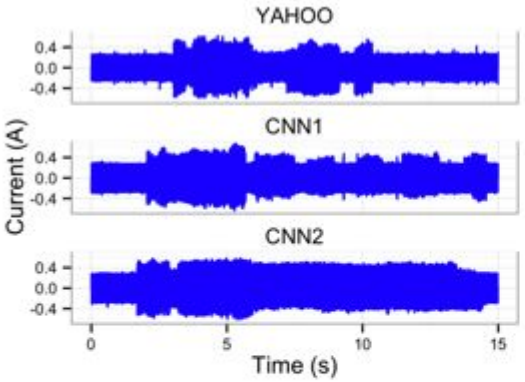


# Accelerometer Eavesdropping

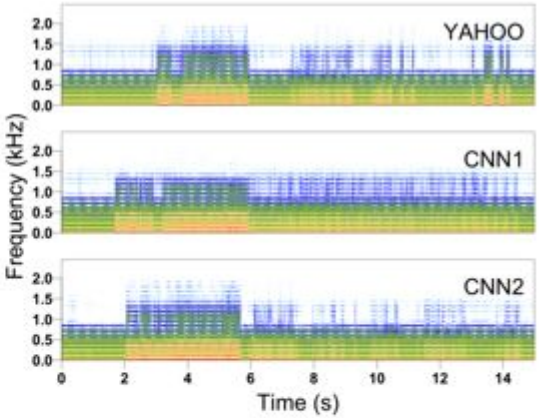



# Keyboard Eavesdropping

# Webpages from Electricity




**(a)** Time-domain plots




**(b)** Spectrogram plots

# TV from Electricity

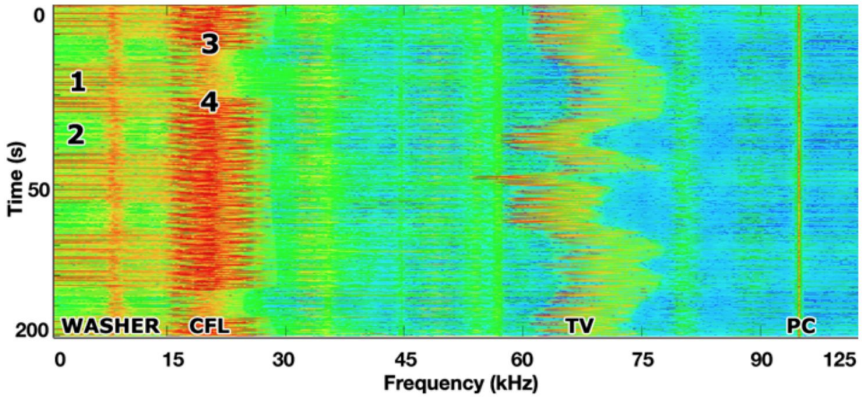



**Figure 1: Frequency spectrogram showing various electrical appliances in the home. Washer cycle on (1) and off (2). CFL lamp turning off briefly (3) and then on (4). Note that the TV's (Sharp 42" LCD) EMI shifts in frequency, which happens as screen content changes.**

# Which is Fake?

- Reflection Eavesdropping?
- Audio from Video?
- Key Extraction?
- Gyroscope Eavesdropping?
- Accelerometer Eavesdropping?
- Keyboard Eavesdropping?
- Webpages from Electricity?
- TV from Electricity?

## All true!

**And many others!**

- Many timing & power analysis attacks to extract crypto keys
- Cache-based side channels (e.g., to extract keys)
- Shared files in Android to infer what the UI is showing
- Newest friends: Spectre, Meltdown, Foreshadow, Spoiler

# References


- Reflection Eavesdropping?
  - Backes et al. “Tempest in a Teapot: Compromising Reflections Revisited” Oakland 2009
- Audio from Video?
  - Davis et al. “The Visual Microphone: Passive Recovery of Sound from Video” SIGGRAPH 2014
- Key Extraction?
  - Genkin et al. “Get Your Hands Off My Laptop: Physical Side-Channel Key-Extraction Attacks On PCs” CHES 2014
- Gyroscope Eavesdropping?
  - Michalevsky et al. “Gyrophone: Recognizing Speech From Gyroscope Signals” USENIX Security 2014
- Accelerometer Eavesdropping?
  - Aviv et al. “Practicality of Accelerometer Side Channels on Smartphones” ACSAC 2012
- Keyboard Eavesdropping?
  - Zhuang et al. “Keyboard Acoustic Emanations Revisited” CCS 2005
  - Vuagnoux et al. “Compromising Electromagnetic Emanations of Wired and Wireless Keyboards” USENIX Security 2009
- Webpages from Electricity?
  - Clark et al. “Current Events: Identifying Webpages by Tapping the Electrical Outlet” ESORICS 2013
- TV from Electricity?
  - Enev et al. “Televisions, Video Privacy, and Powerline Electromagnetic Interference” CCS 2011

# Spectre



What is Spectre?

How does it work?

What are its implications?

How might we fix it?

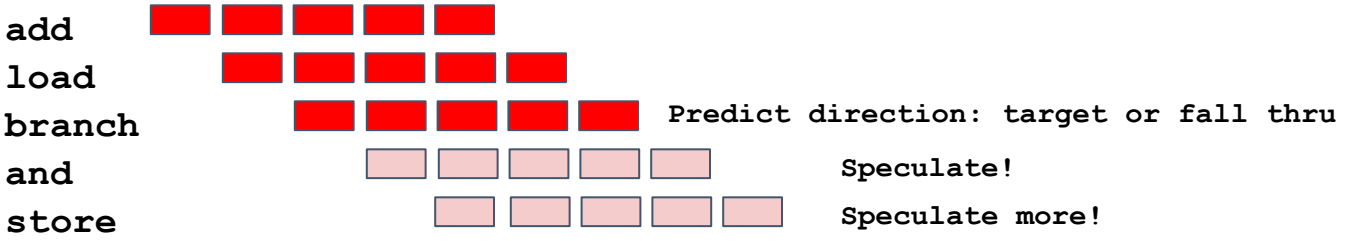

What about Meltdown, Foreshadow, Spoiler?

Material from http://research.cs.wisc.edu/multifacet/papers/hill_mark_wisconsin_meltdown_spectre.pptx

# Instruction Speculation Tutorial

Many steps (cycles) to execute one instruction; time flows left to right →


Go Faster: Pipelining, branch prediction, & instruction speculation




Speculation correct: Commit architectural changes of **`and`** (register) & **`store`** (memory) go fast!

Mis-speculate: Abort architectural changes (registers, memory); go in other branch direction

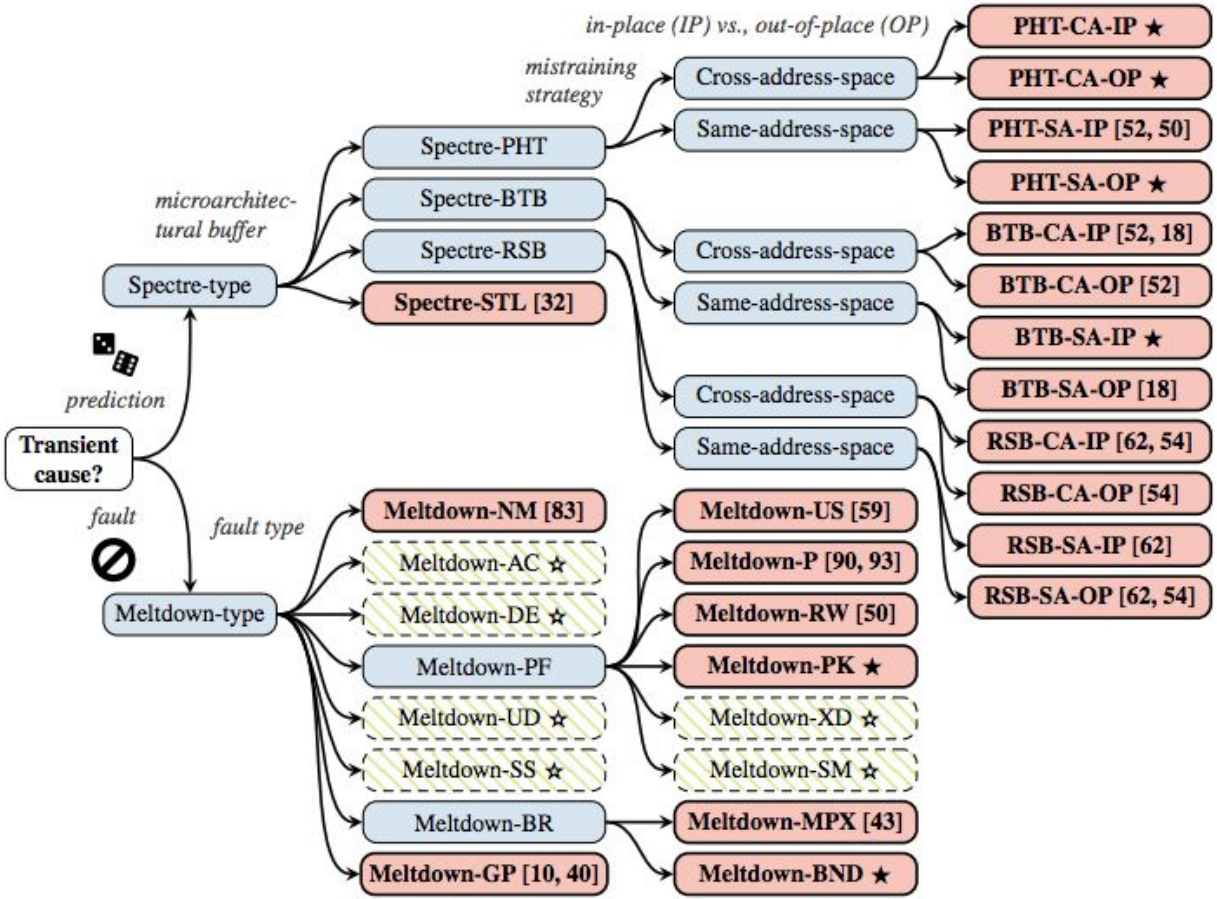
# Spectre Key Insights

```
if (x < array1_size)
    y = array2[array1[x] * 4096];
```

- Train branch predictor to follow one branch of conditional
- After training, make the followed branch access information that the code should *not* be allowed to access
- That access information will be loaded into the cache
- After the hardware determines that the branch was incorrectly executed, the logic of the program will be rolled back *but* the cache will still be impacted
- Time reads to cache (e.g., Evict+Reload), to see which cache lines are read more efficiently

# Spectre Variants

From Canella et al.



# Adversarial ML

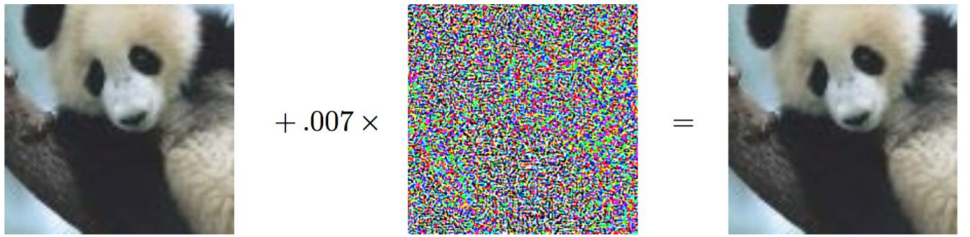
# Adversarial ML

What are “adversarial examples”?

Why do they work?

What are their implications?

How might we fix it?




*Goodfellow et al., ICLR 2015*

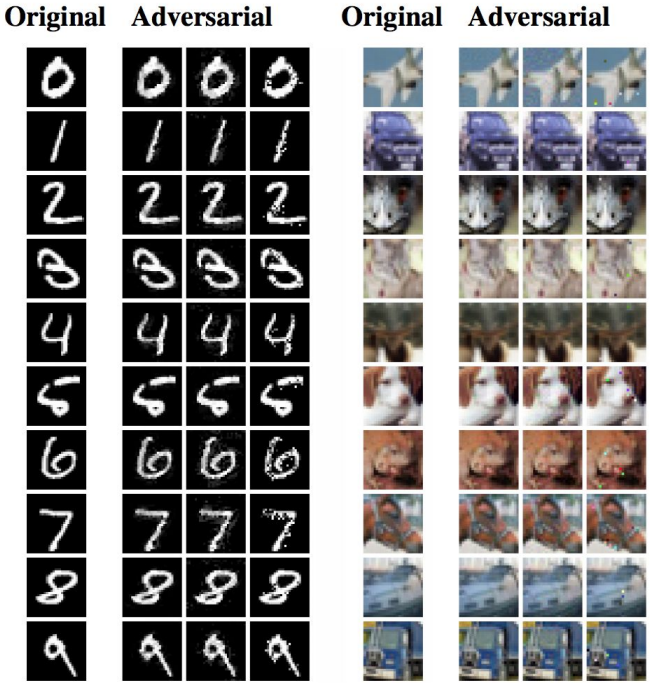



Fig. 1. An illustration of our attacks on a defensively distilled network. The leftmost column contains the starting image. The next three columns show adversarial examples generated by our $L_2$, $L_\infty$, and $L_0$ algorithms, respectively. All images start out classified correctly with label $l$, and the three misclassified instances share the same misclassified label of $l+1 \pmod{10}$. Images were chosen as the first of their class from the test set.

*Carlini & Wagner, IEEE S&P 2017*

# Deep Neural Networks Can Fail

If you use a loss function that fulfills an adversary's goal, you can follow the gradient to find an image that misleads the neural network.

$$\boldsymbol{X}^{adv} = \boldsymbol{X} + \epsilon \operatorname{sign}\left(\nabla_X J(\boldsymbol{X}, y_{true})\right)$$


**"gibbon"**
99.3%
confidence

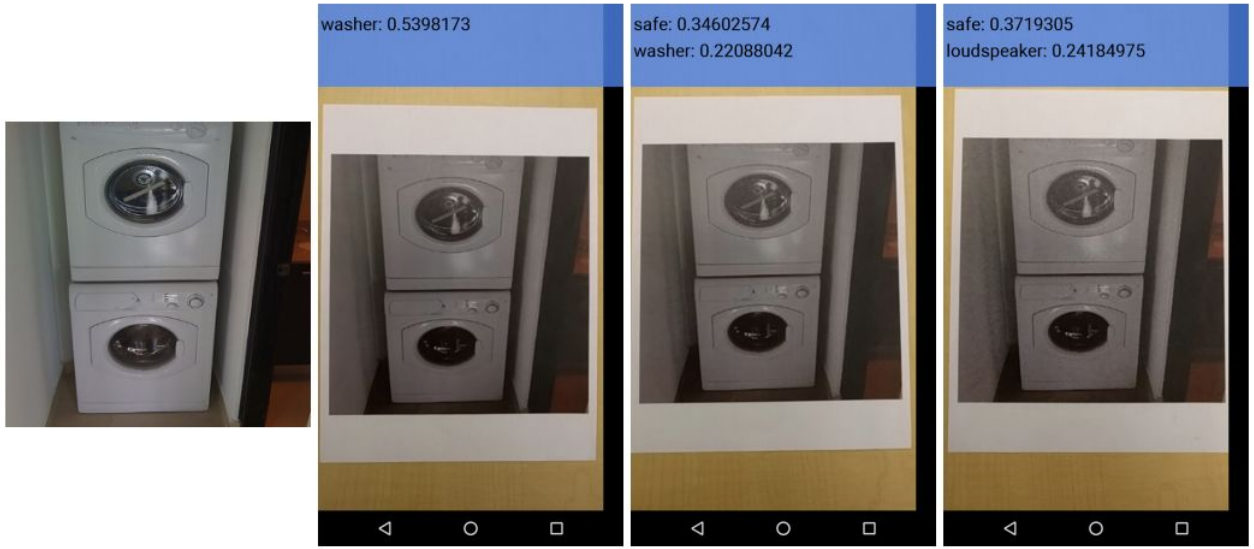
**"panda"**
57.7%
confidence

Image Courtesy: OpenAI

Explaining and Harnessing Adversarial Examples, Goodfellow et al., arXiv 1412.6572, 2015

# Deep Neural Networks Can Fail...

## ...if adversarial images are printed out



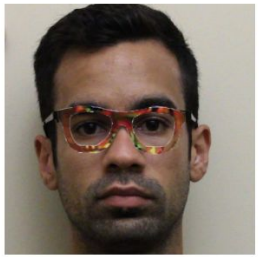

Kurakin et al. "Adversarial examples in the physical world." arXiv preprint arXiv:1607.02533 (2016).

# Deep Neural Networks Can Fail...

## ...if an adversarially crafted physical object is introduced

This person wearing an "adversarial" glasses frame...





...is classified as this person by a state-of-the-art face recognition neural network.



Sharif et al. "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition." Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2016.

Deep neural network classifiers are vulnerable to adversarial examples in some physical world scenarios

**However**: In real-world applications, conditions vary more than in the lab.

# Take autonomous driving as an example...


A road sign can be far away


or it could be at an angle

Can physical adversarial examples cause misclassification at large angles and distances?

# What about physical realizability?

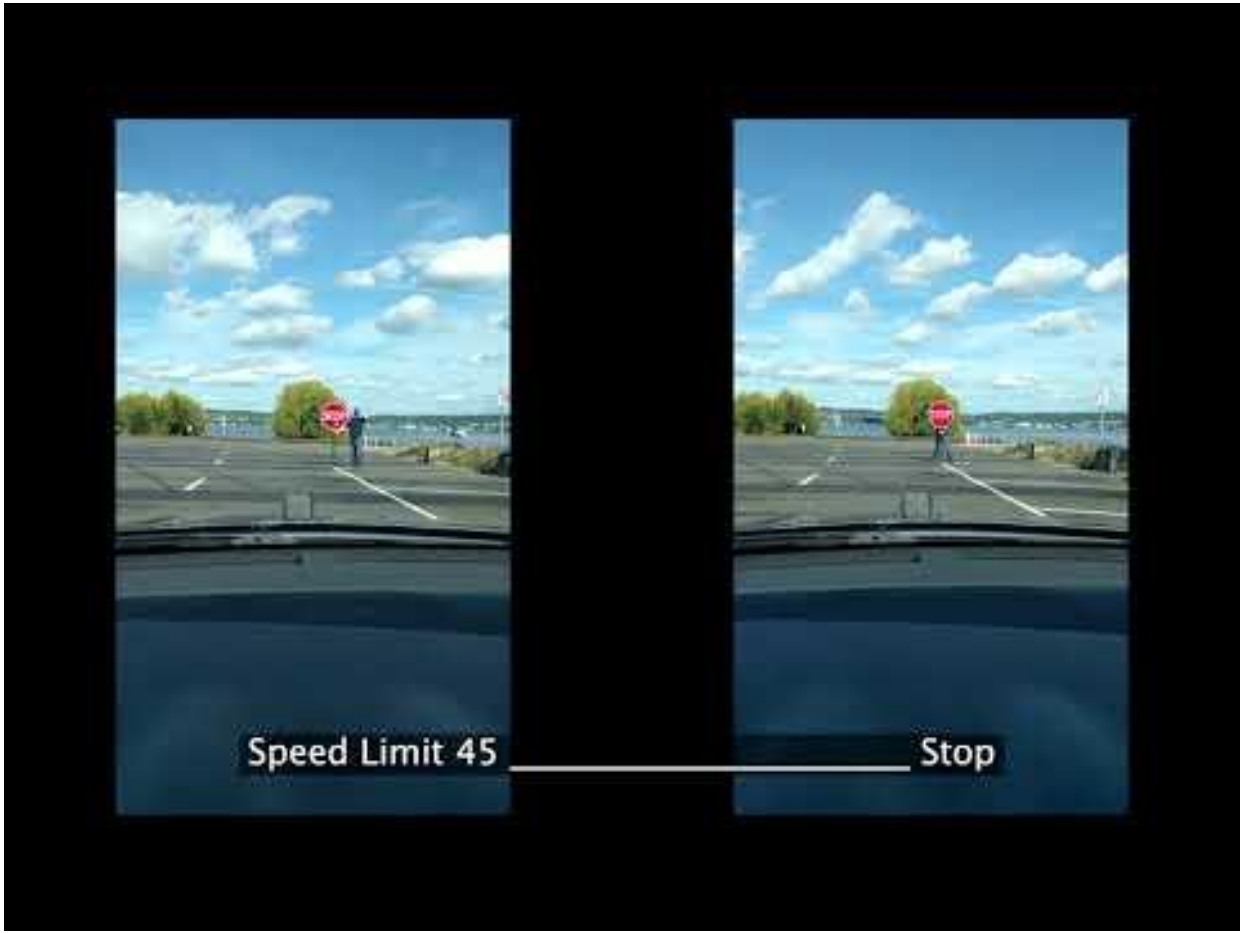
Observation: Signs are often messy...


# What about physical realizability?

So: make the perturbation appear as vandalism


Camouflage Sticker


Subtle Poster




# Field Test (Drive-by)

Target Classes:
Stop -> Speed Limit 45
Right Turn -> Stop

Classification top class is indicated at the bottom of the images.
Left: “Adversarial” stop sign
Right: Clean stop sign