What's next for adversarial ML?

(and why ad-blockers should care)

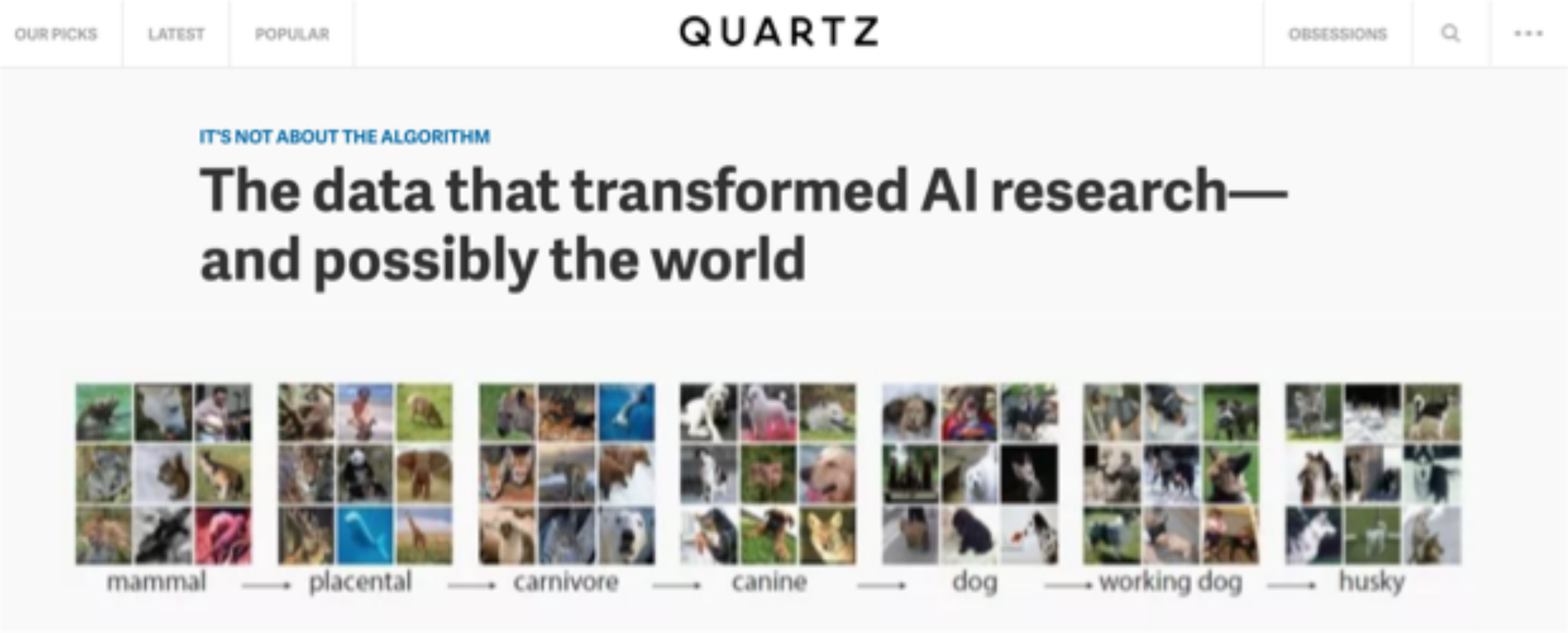
Florian Tramèr
EPFL

July 9th 2018

Joint work with Gili Rusak, Giancarlo Pellegrino and Dan Boneh

The Deep Learning Revolution

First they came for images…



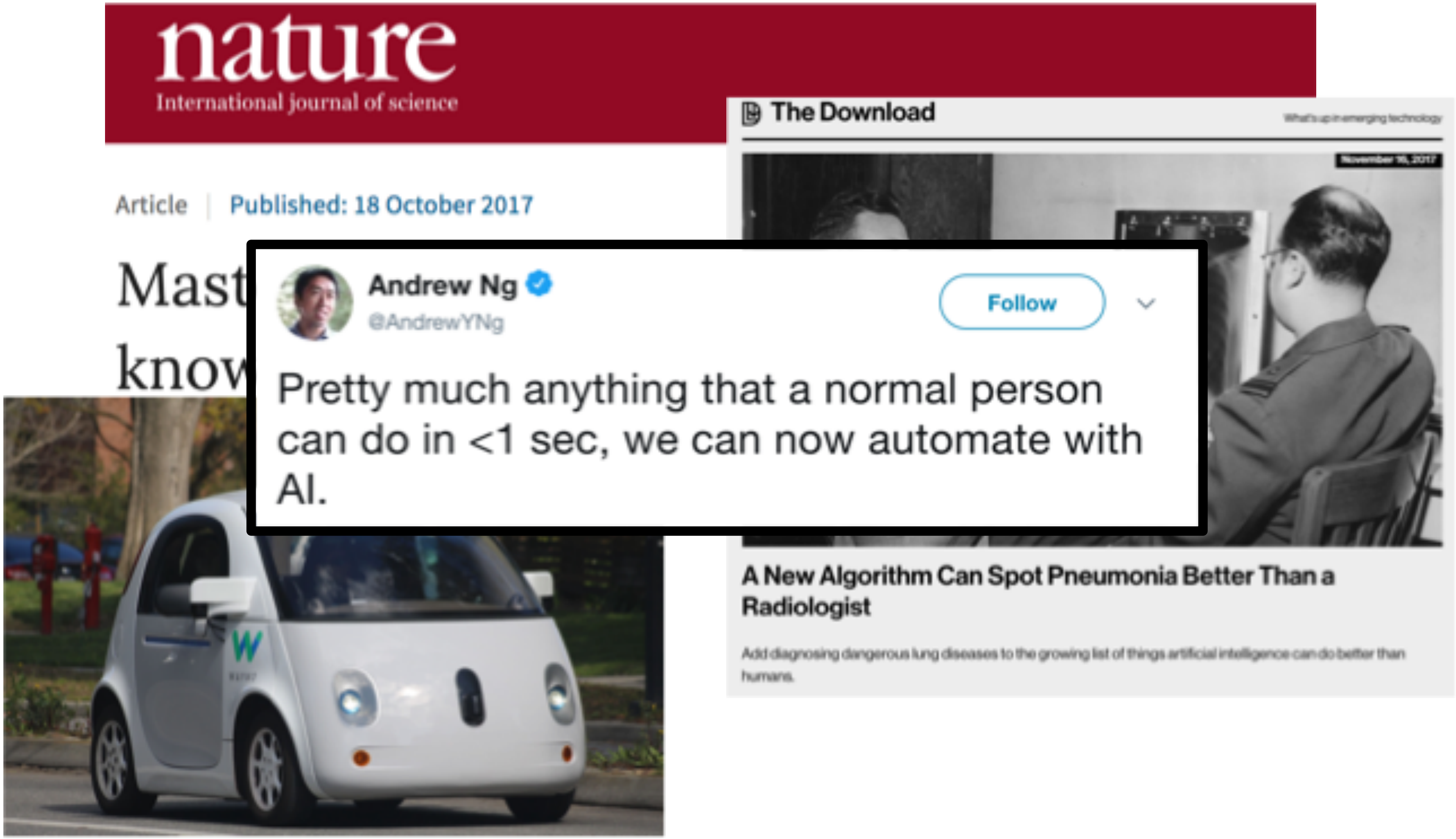
OUR PICKS
LATEST
POPULAR
QUARTZ
OBSESSIONS
mammal → placental → carnivore → canine → dog → working dog → husky

IT'S NOT ABOUT THE ALGORITHM

The data that transformed AI research—and possibly the world

The Deep Learning Revolution

And then everything else…



nature

International journal of science

Article | Published: 18 October 2017

Mast

know

The Download

What's up in emerging technology

November 16, 2017

A New Algorithm Can Spot Pneumonia Better Than a Radiologist

Add diagnosing dangerous lung diseases to the growing list of things artificial intelligence can do better than humans.

Andrew Ng

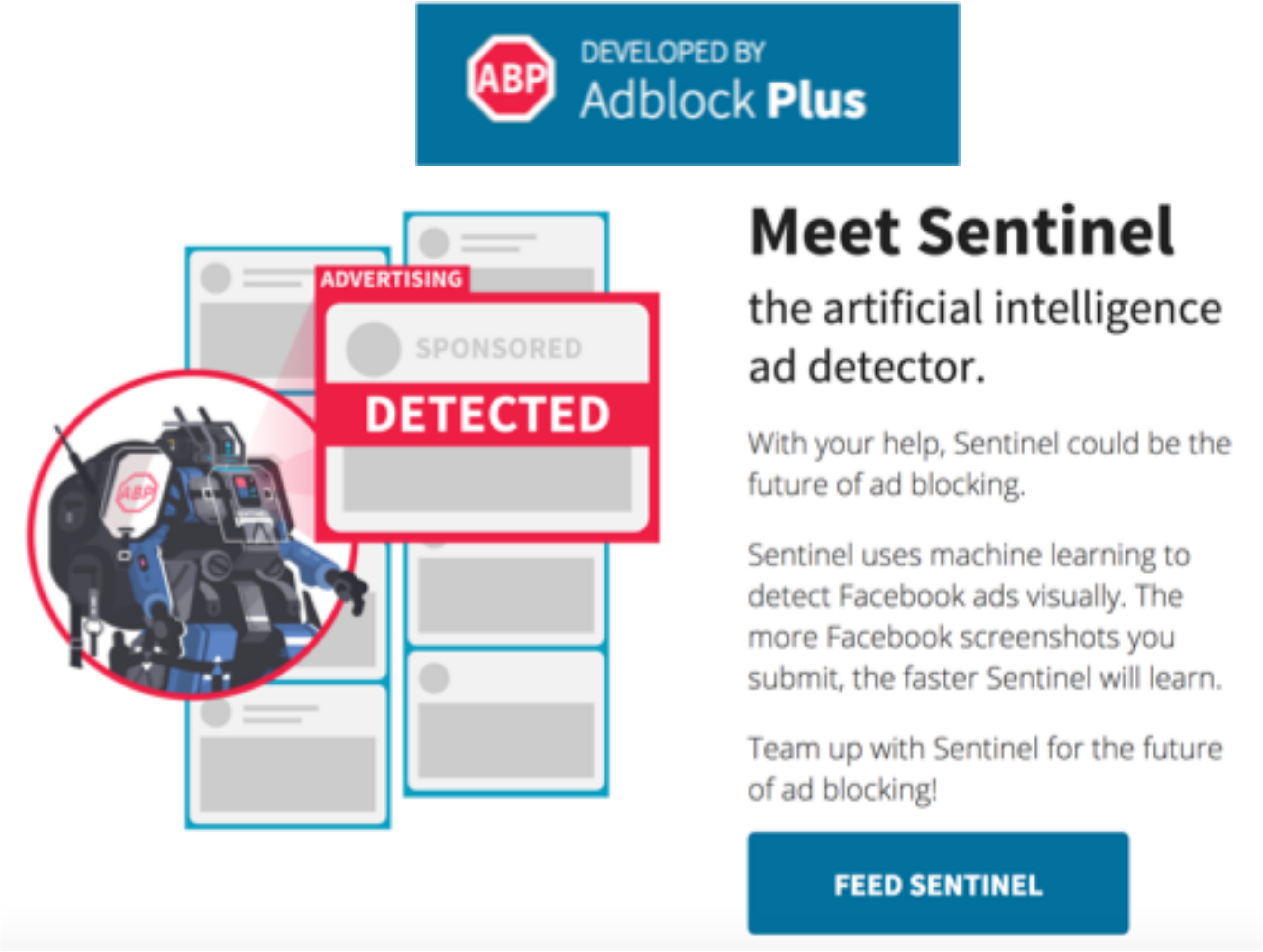
@AndrewYNg

Follow

Pretty much anything that a normal person can do in <1 sec, we can now automate with AI.

The ML Revolution

Including things that likely won't work…



Meet Sentinel

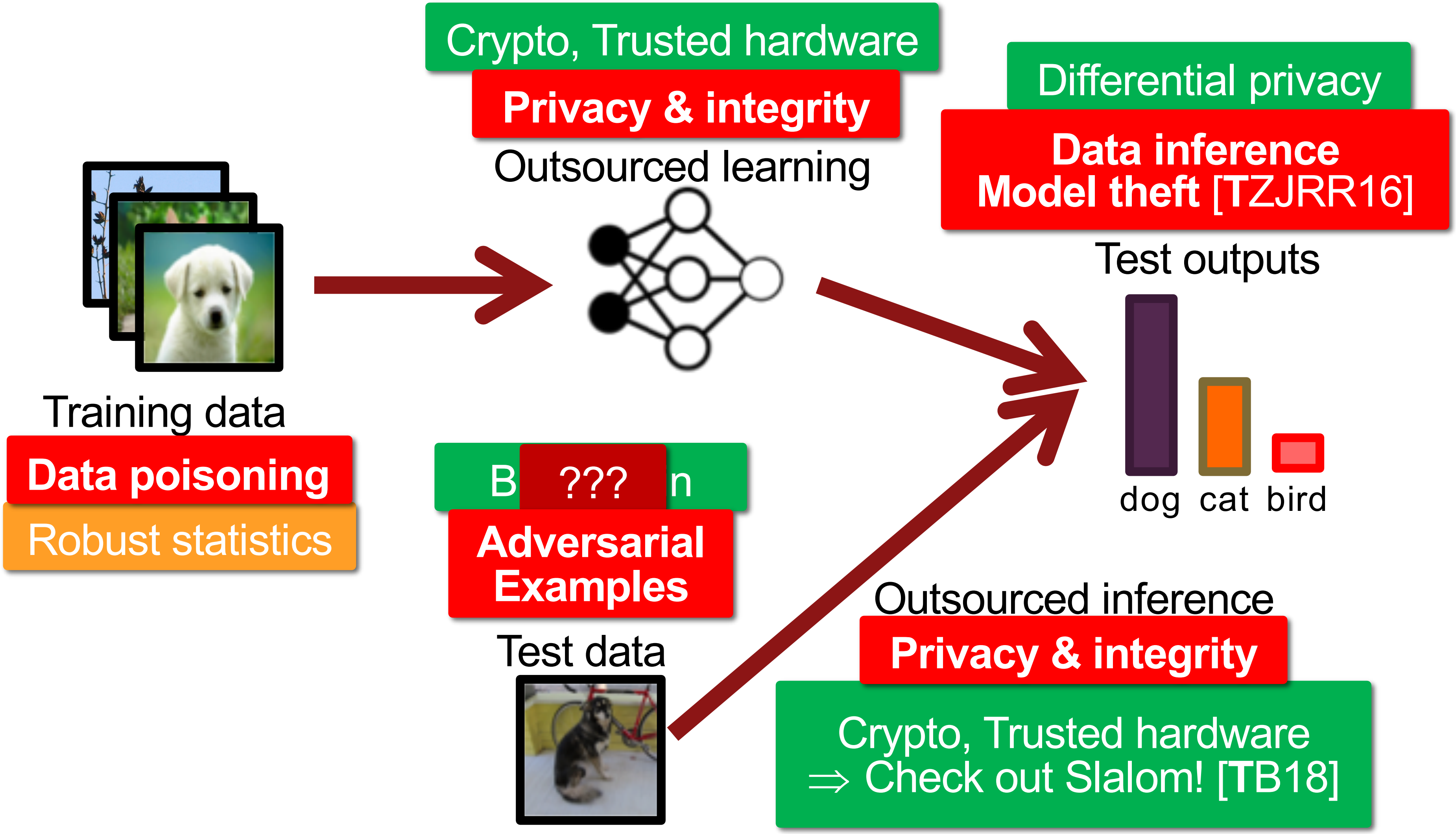
the artificial intelligence ad detector.

With your help, Sentinel could be the future of ad blocking.

Sentinel uses machine learning to detect Facebook ads visually. The more Facebook screenshots you submit, the faster Sentinel will learn.

Team up with Sentinel for the future of ad blocking!

What does this mean for privacy & security?



Crypto, Trusted hardware

Privacy & integrity

Outsourced learning

Differential privacy

Data inference
Model theft [TZJRR16]

Test outputs

Training data

Data poisoning

Robust statistics

B ??? n

Adversarial Examples

Test data

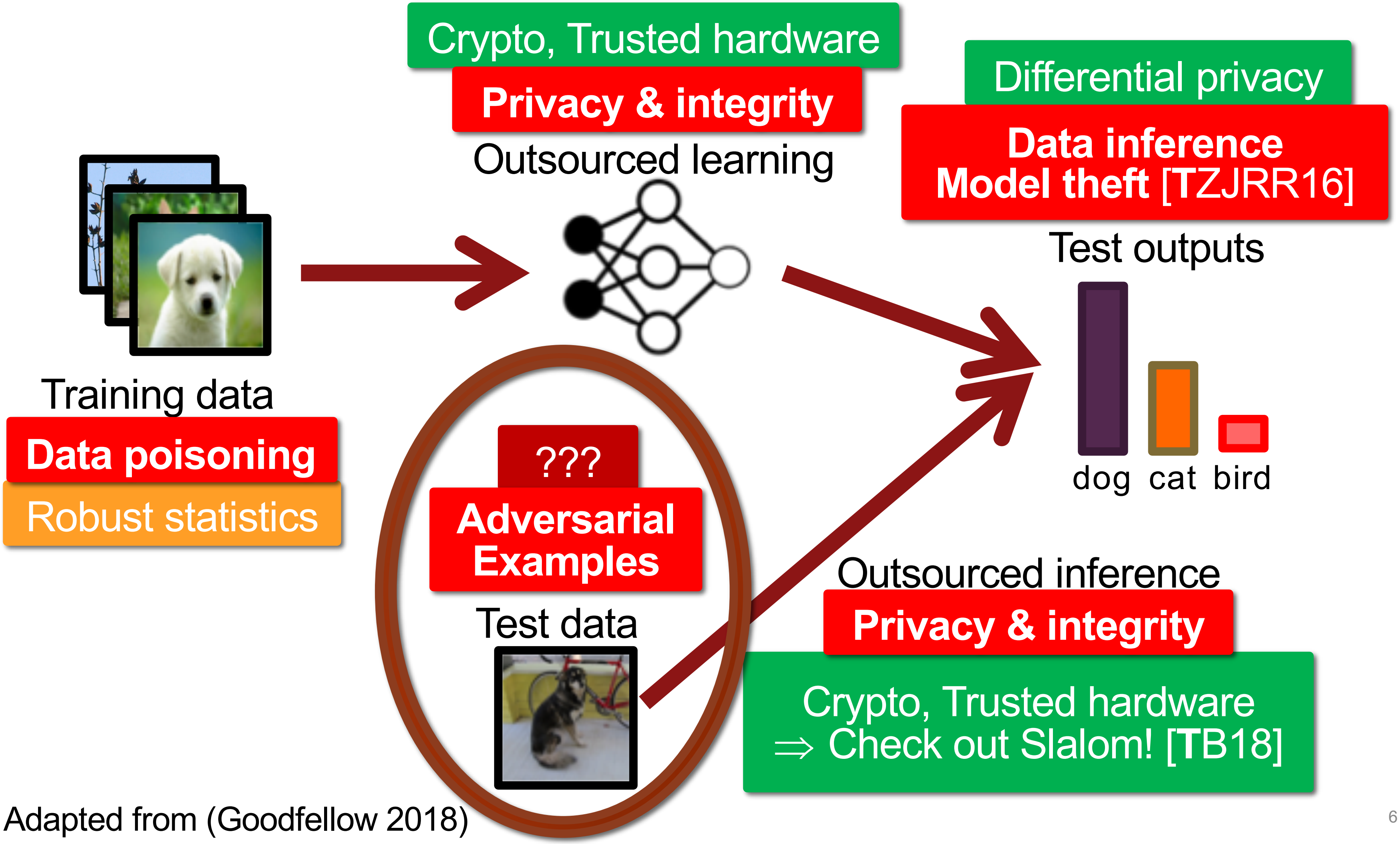
Outsourced inference

Privacy & integrity

Crypto, Trusted hardware
⇒ Check out Slalom! [TB18]

Adapted from (Goodfellow 2018)

What does this mean for privacy & security?



Crypto, Trusted hardware

Privacy & integrity

Outsourced learning

Differential privacy

Data inference
Model theft [TZJRR16]

Test outputs

Training data

Data poisoning

Robust statistics

???

Adversarial Examples

Test data

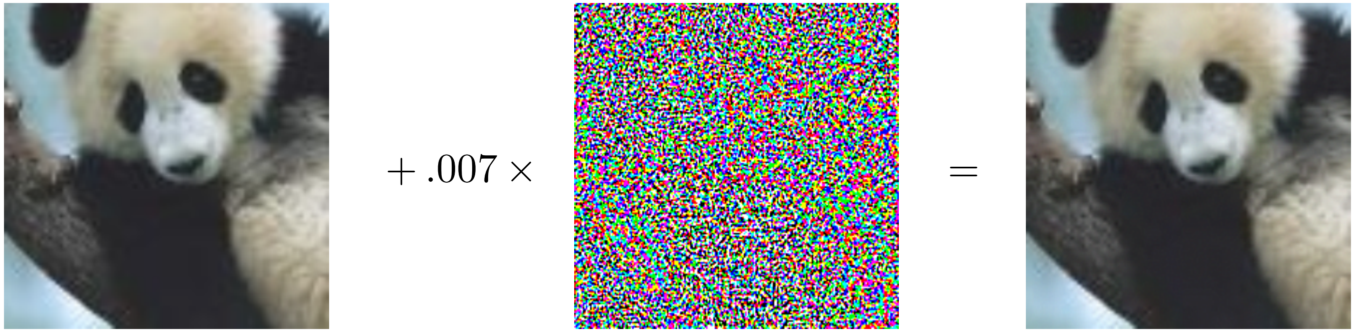
Outsourced inference

Privacy & integrity

Crypto, Trusted hardware
⇒ Check out Slalom! [TB18]

Adapted from (Goodfellow 2018)

ML models make surprising mistakes



Pretty sure this is a panda

I’m certain this is a gibbon

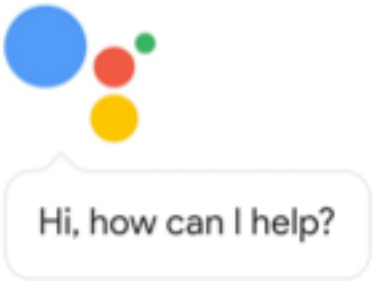
(Szegedy et al. 2013, Goodfellow et al. 2015)

Attacks on cyber-physical systems

(Sharif et al. 2016)

(Kurakin et al. 2016)

(Athalye et al. 2018)



(Carlini et al. 2016,
Cisse et al. 2017)

(Eykholt et al. 2017)

(Eykholt **et al.** 2018)

Where are the defenses?

- Adversarial training

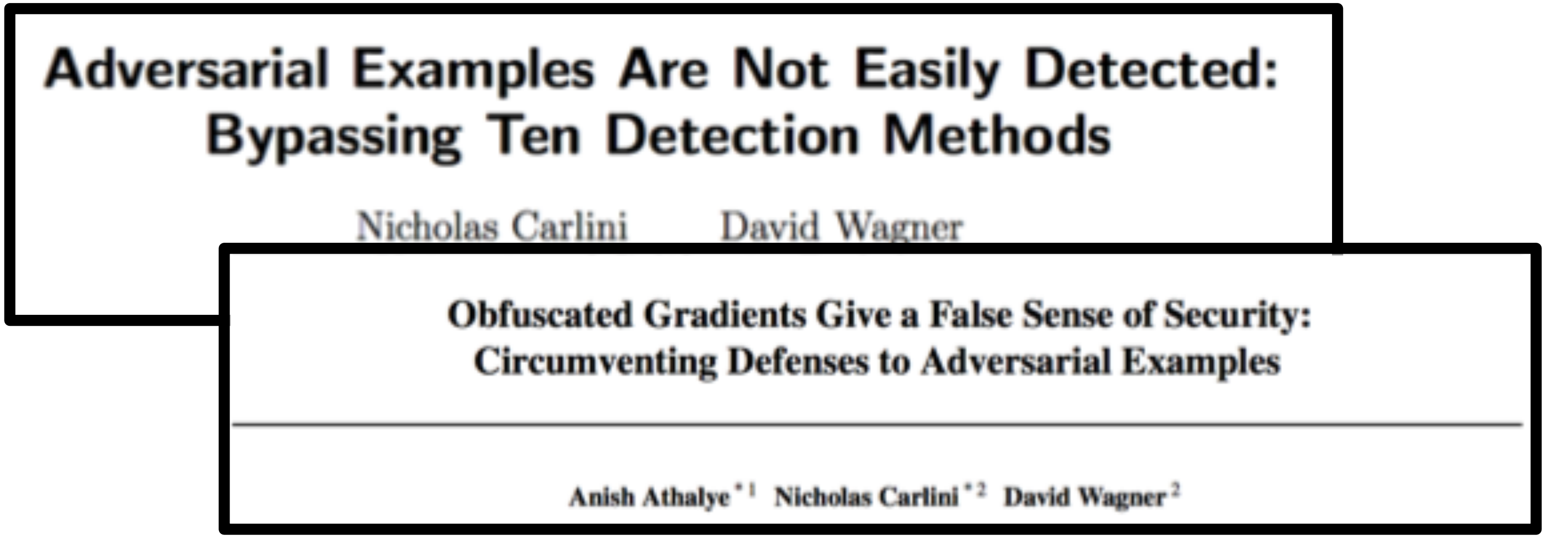
  Szegedy et al. 2013, Goodfellow et al. 2015, Kurakin et al. 2016, **T** et al. 2017,
  Madry et al. 2017, Kannan et al. 2018

Prevent “all/most attacks” **for a given norm ball**

- Convex relaxations with provable guarantees

  Raghunathan et al. 2018, Kolter & Wong 2018, Sinha et al. 2018

- A lot of broken defenses…



Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods

Nicholas Carlini David Wagner

Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples

Anish Athalye [*1] **Nicholas Carlini** [*2] **David Wagner** [2]

Do we have a realistic threat model? (no…)

Current approach:

1. Fix a ”toy” attack model (e.g., some l_∞ ball)
2. Directly optimize over the robustness measure

   ⇒ Defenses do not generalize to other attack models

   ⇒ Defenses are meaningless for applied security

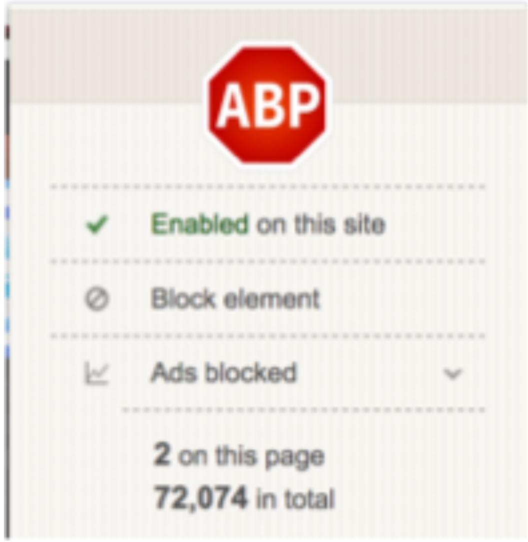
What do we want?

- Model is “always correct” (sure, why not?)
- Model has blind spots that are “hard to find”
  - “Non-information-theoretic” notions of robustness?
  - CAPTCHA threat model is interesting to think about

ADVERSARIAL EXAMPLES ARE HERE TO STAY!

For many things that humans can do "robustly", ML will fail miserably!

A case study on ad-blocking



ABP

✓ Enabled on this site

⊘ Block element

Ads blocked

2 on this page

72,074 in total

Ad blocking is a “cat & mouse” game

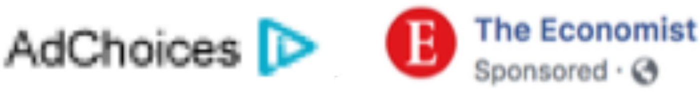
1. Ad blockers build crowd-sourced filter lists
2. Ad providers switch origins
3. Rinse & repeat

(4?) Content provider (e.g., Cloudflare) hosts the ads

A case study on ad-blocking

New method: perceptual ad-blocking (Storey et al. 2017)

- Industry/legal trend: ads have to be clearly indicated to humans

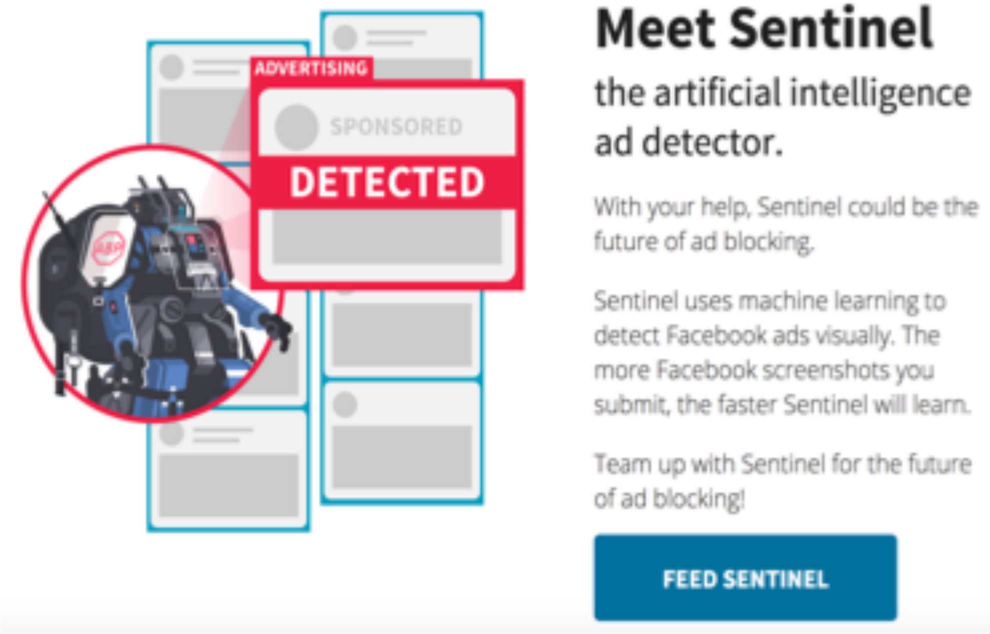

AdChoices

The Economist
Sponsored ·

If humans can detect ads, so can ML!

> ”[…] ***we deliberately ignore all signals invisible to humans***, *including URLs and markup. Instead we consider visual and behavioral information. […]* ***We expect perceptual ad blocking to be less prone to an "arms race."***

(Storey et al. 2017)



Meet Sentinel

the artificial intelligence ad detector.

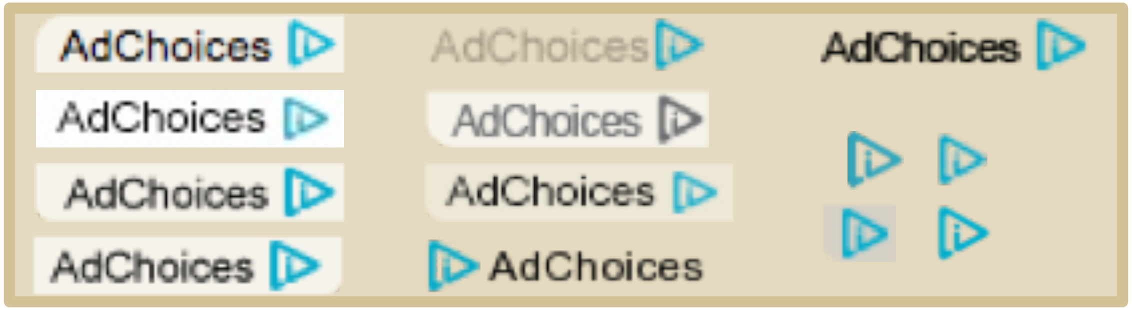
With your help, Sentinel could be the future of ad blocking.

Sentinel uses machine learning to detect Facebook ads visually. The more Facebook screenshots you submit, the faster Sentinel will learn.

Team up with Sentinel for the future of ad blocking!

Detecting ad logos is not trivial

No strict guidelines, or only loosely followed:



Fuzzy hashing + OCR (Storey et al. 2017)

⇒ Fuzzy hashing is very brittle (e.g., shift all pixels by 1)

⇒ OCR has adversarial examples (Song & Shmatikov, 2018)

This talk

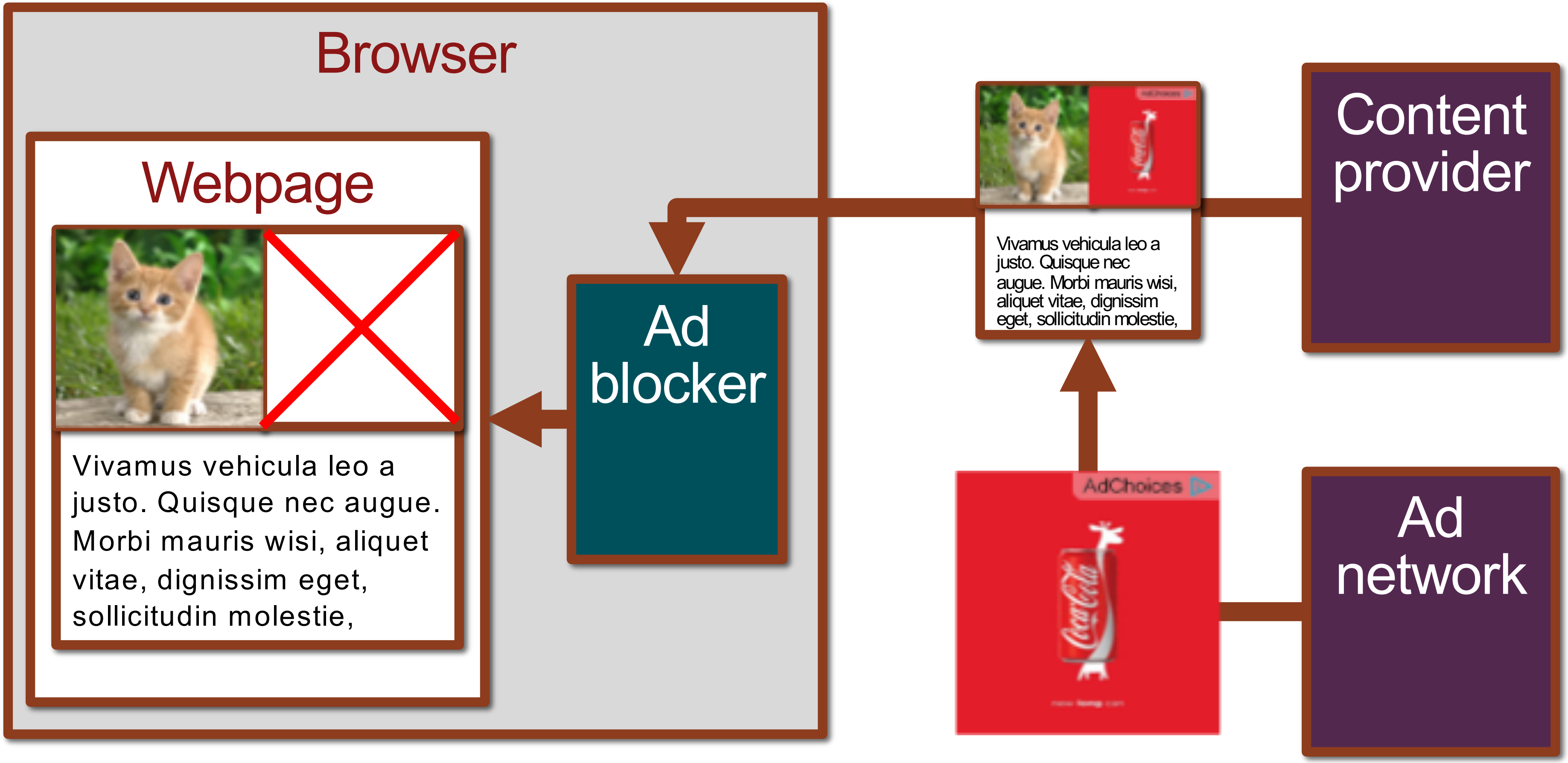
Unsupervised feature detector (SIFT)

⇒ More robust method for matching object features ("keypoints")

Deep object detector (YOLO)

⇒ Supervised learning

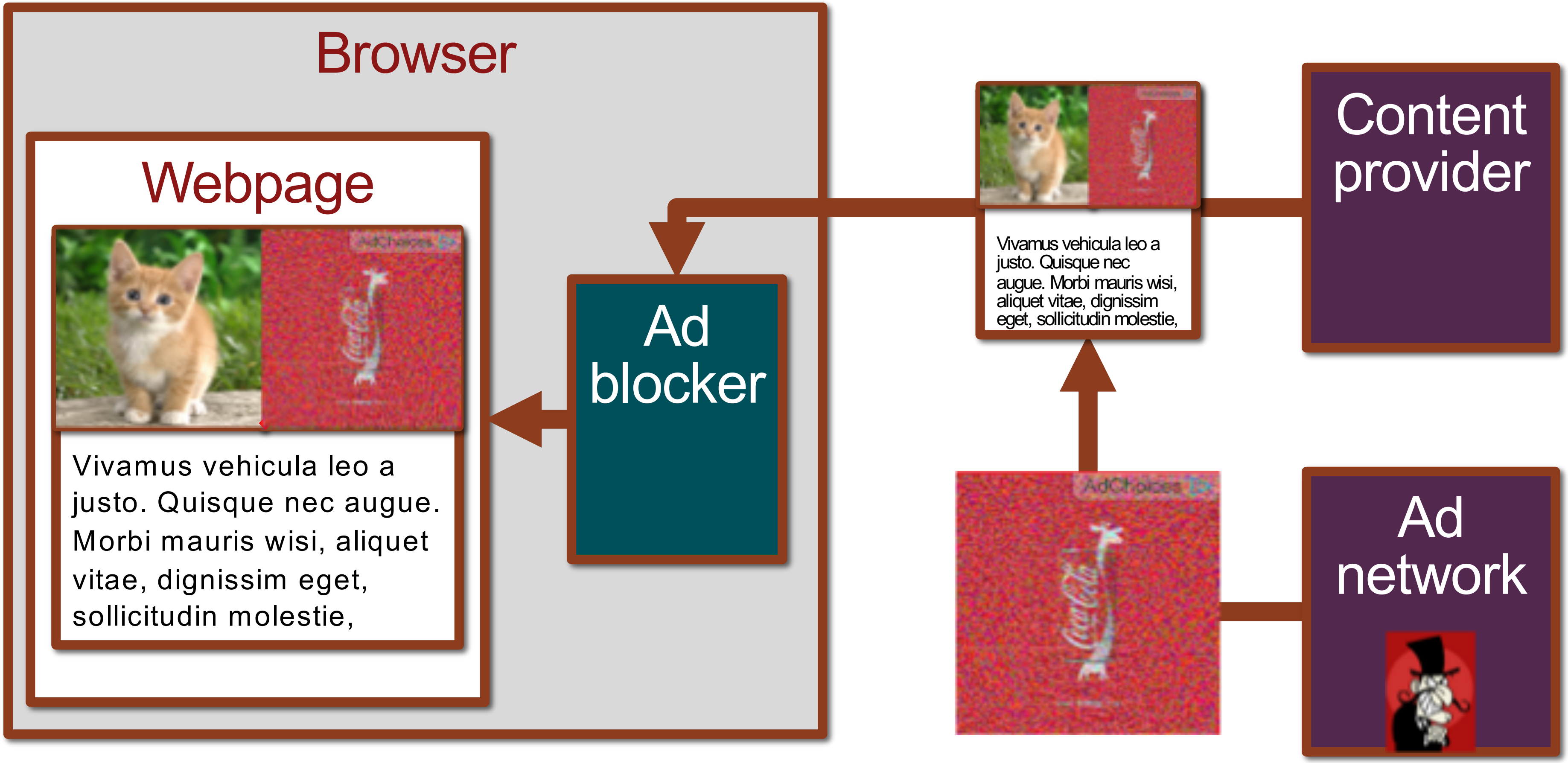
What's the threat model for perceptual ad-blockers?



Vivamus vehicula leo a justo. Quisque nec augue. Morbi mauris wisi, aliquet vitae, dignissim eget, sollicitudin molestie,

Vivamus vehicula leo a justo. Quisque nec augue. Morbi mauris wisi, aliquet vitae, dignissim eget, sollicitudin molestie,

What's the threat model for perceptual ad-blockers?



Browser

Webpage

Vivamus vehicula leo a justo. Quisque nec augue. Morbi mauris wisi, aliquet vitae, dignissim eget, sollicitudin molestie,

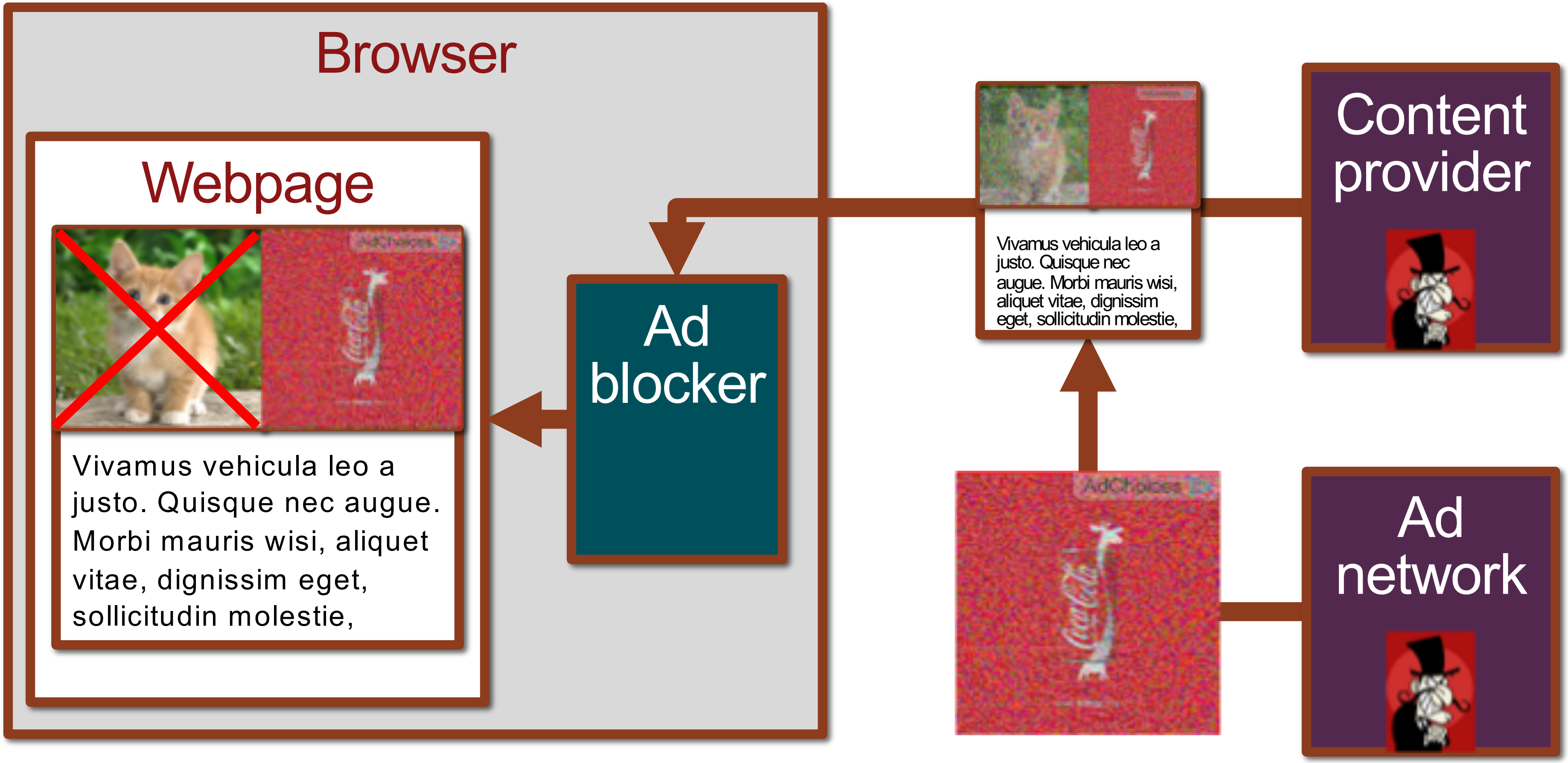
Ad blocker

Content provider

Vivamus vehicula leo a justo. Quisque nec augue. Morbi mauris wisi, aliquet vitae, dignissim eget, sollicitudin molestie,

Ad network

What’s the threat model for perceptual ad-blockers?



Browser

Webpage

Vivamus vehicula leo a justo. Quisque nec augue. Morbi mauris wisi, aliquet vitae, dignissim eget, sollicitudin molestie,

Ad blocker

Vivamus vehicula leo a justo. Quisque nec augue. Morbi mauris wisi, aliquet vitae, dignissim eget, sollicitudin molestie,

Content provider

Ad network

What's the threat model for perceptual ad-blockers?

Pretty much the worst possible!

1. **Adblocker is white-box** (browser extension)

   ⇒ Alternative would be a privacy & bandwidth nightmare

2. **Adblocker operates on (large) digital images**

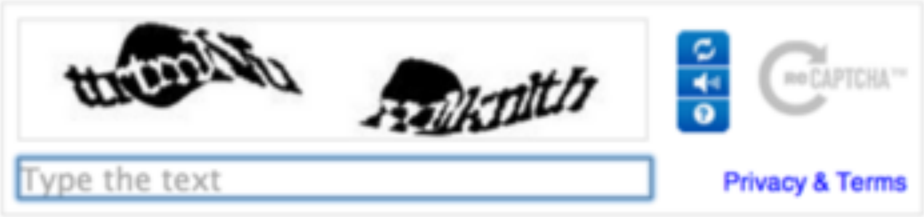
3. **Adblocker needs to resist adversarial examples and "DOS" attacks**

   ⇒ Perturb ads to evade ad blocker

   ⇒ Punish ad-block users by perturbing benign content

4. **Updating is more expensive than attacking**

An interesting contrast: CAPTCHAs



Deep ML models can solve text CAPTCHAs

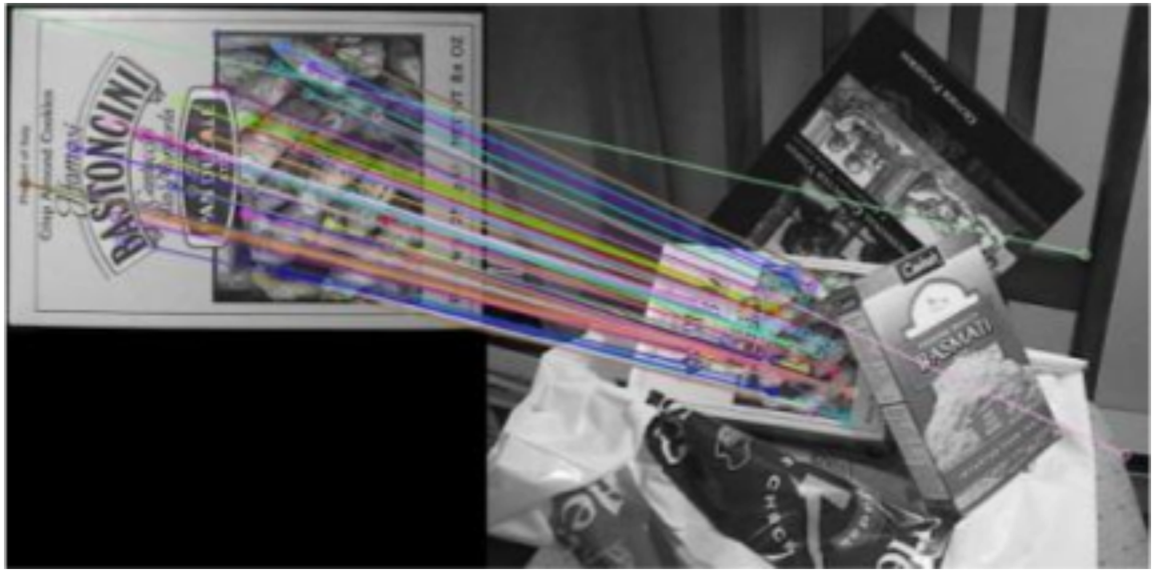
⇒Why don’t CAPTCHAs use adversarial examples?

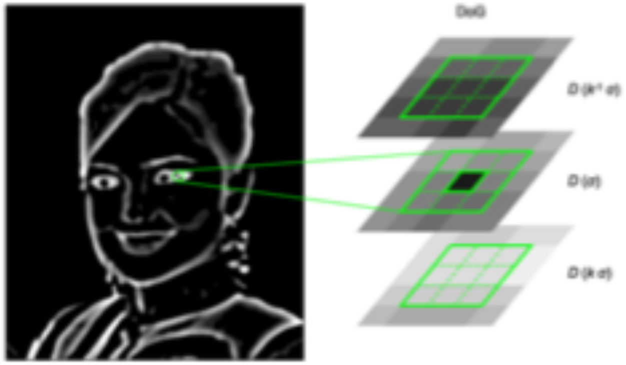
⇒CAPTCHA ≃ adversarial example for OCR systems

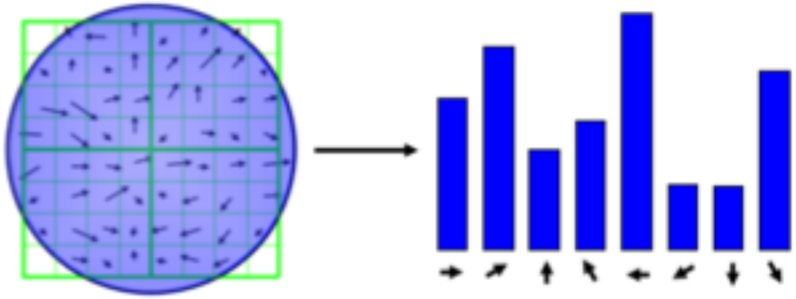
	Model access	Vulnerable to DOS	Model Distribution
Ad blocker	White-box	Yes	Expensive
CAPTCHA	“Black-box” (not even query access)	No	Cheap (None)

Breaking Perceptual Ad-Blockers with adversarial Examples

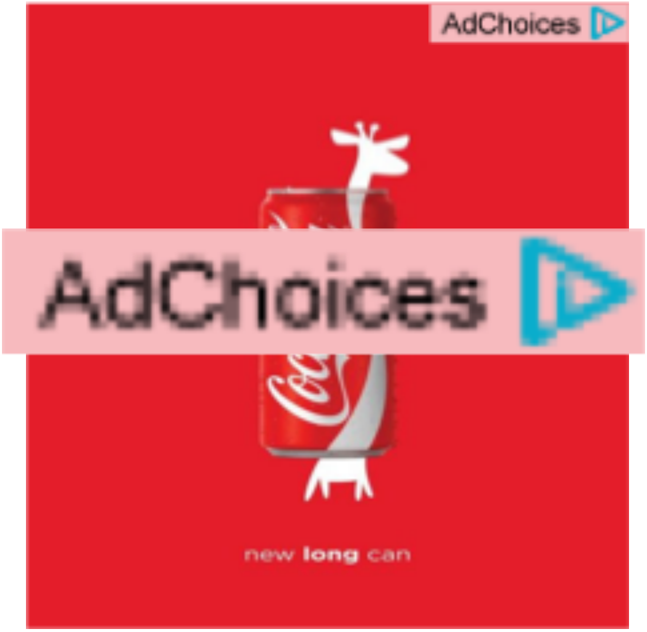
SIFT: How does it work? (I don’t know exactly either)



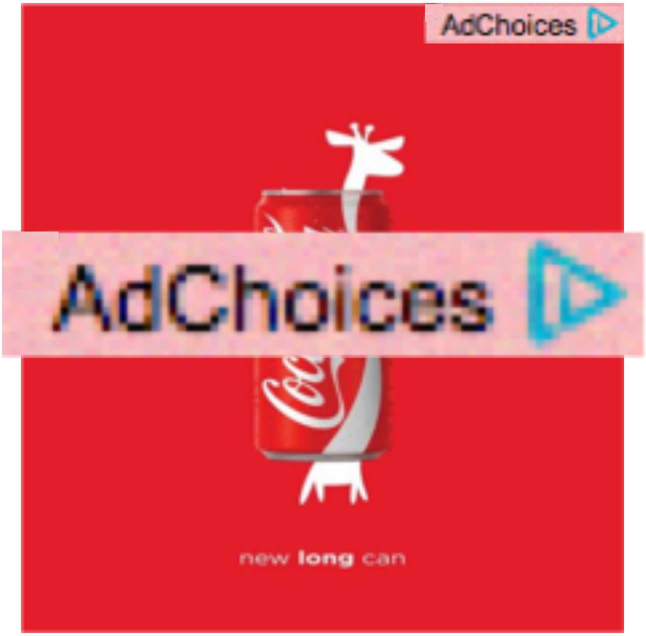




Attack examples: SIFT detector



original ad



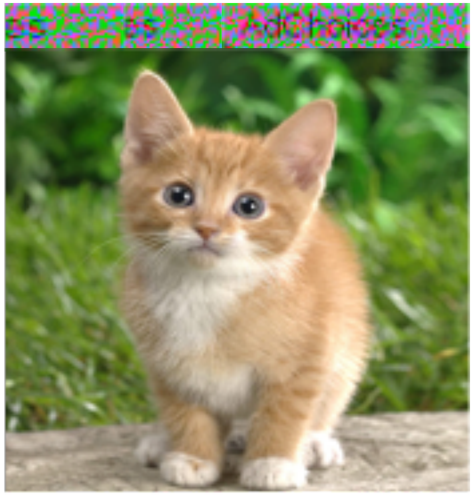
perturbed logo

- No keypoint matches between the two logos
- Attack uses standard black-box optimization

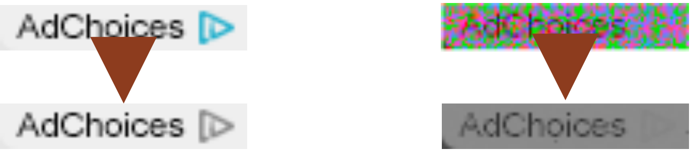
⇒ Gradient descent with black-box gradient estimates

⇒ There's surely more efficient attacks but SIFT is complicated…

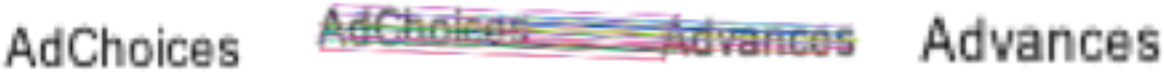
Attack examples: SIFT Denial Of Service



- Logos are similar in gray scale but not in color space



- Alternative: high confidence matches for visually close—yet semantically different—objects

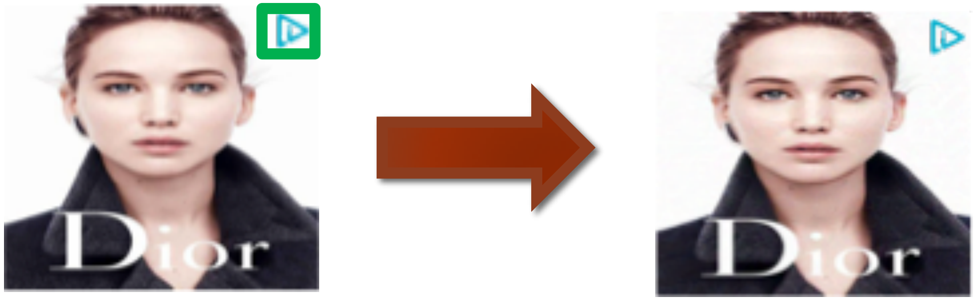


Attack examples: YOLO object detector

Object detector trained to recognize AdChoice logo

⇒ Test accuracy is >90%

⇒ 0% accuracy with l_∞ perturbations ≤ 8/256



Similar but simpler task than Sentinel (Adblock Plus)

⇒ Sentinel tries to detect ads in a whole webpage

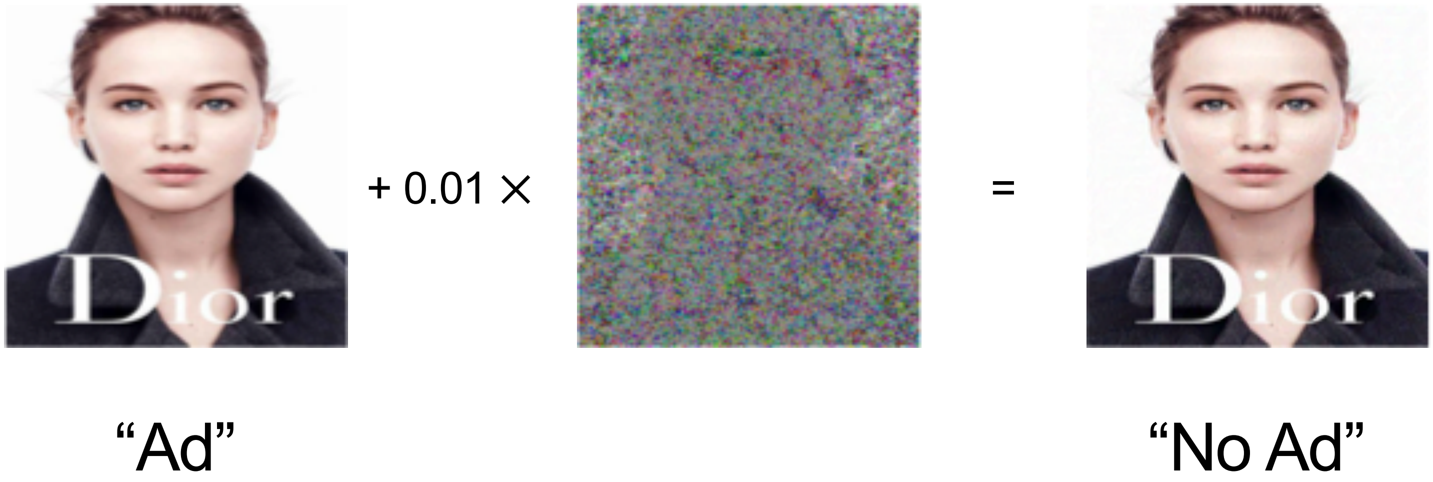
⇒ For now, it breaks even on non-adversarial inputs…

Perceptual ad-blockers without ad-indicators

Hussain et al. 2017: Train a generic ad/no-ad classifier (for sentiment analysis)

⇒ Accuracy around 88% !

⇒ 0% accuracy with l_∞ perturbations ≤ 4/256



Conclusion

Adversarial examples are here to stay

- No defense can address realistic attacks
- A truly robust defense likely implies a huge breakthrough in non-secure ML as well

Security-sensitive ML seems hopeless if adversary has white-box model access

- Ad-blocking ticks most of the “worst-case” boxes
- ML is unlikely to change the ad-blocker cat & mouse game

THANKS