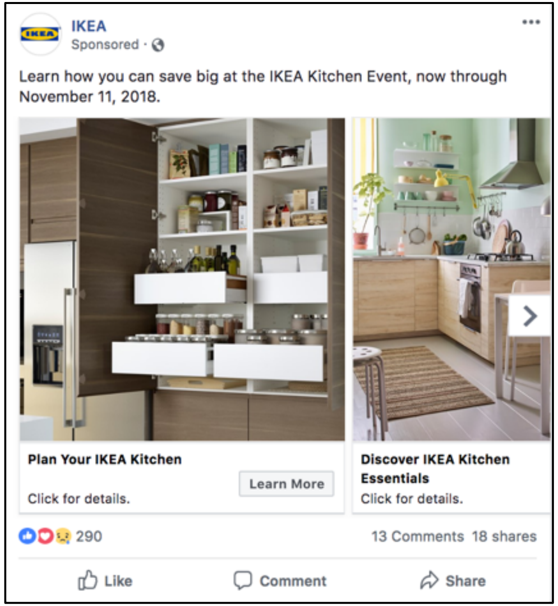
AdVersarial: Perceptual Ad Blocking meets Adversarial Machine Learning

Florian Tramèr

November 14th 2019

Joint work with Pascal Dupré, Gili Rusak, Giancarlo Pellegrino and Dan Boneh

The Future of Ad-Blocking



IKEA
Sponsored ·

Learn how you can save big at the IKEA Kitchen Event, now through November 11, 2018.

Plan Your IKEA Kitchen
Click for details.
Learn More

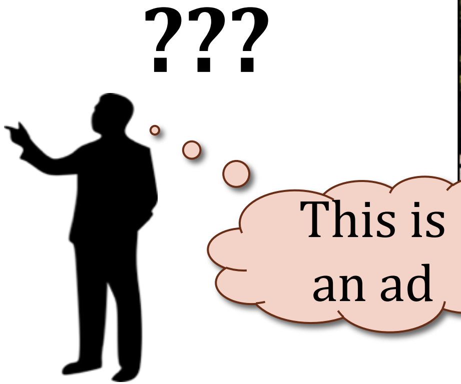
Discover IKEA Kitchen Essentials
Click for details.

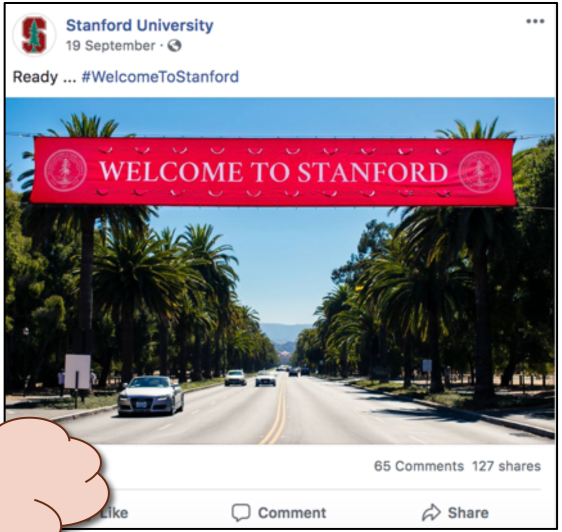
290 — 13 Comments 18 shares

Like · Comment · Share

```
easylist.txt

…markup…
…URLs…
```





Stanford University
19 September ·

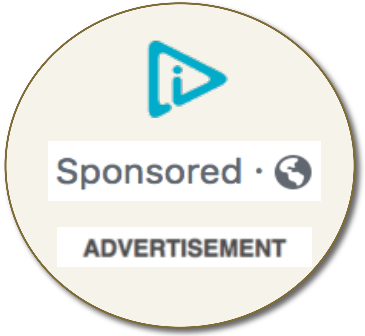
Ready … #WelcomeToStanford

65 Comments 127 shares

Like · Comment · Share

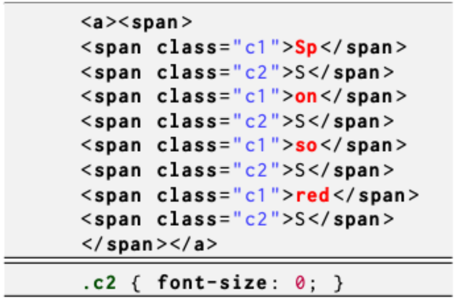
Human distinguishability of ads

- *Legal requirement (U.S. FTC, EU E-Commerce)*
- *Industry self-regulation on ad-disclosure*



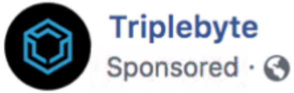
Towards Computer Vision for Ad-Blocking

Why not detect ad-disclosures programmatically?

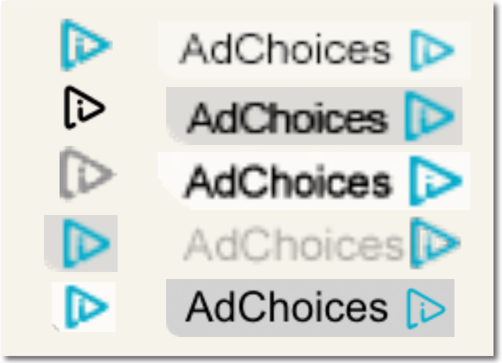


```
<a><span>
<span class="c1">Sp</span>
<span class="c2">S</span>
<span class="c1">on</span>
<span class="c2">S</span>
<span class="c1">so</span>
<span class="c2">S</span>
<span class="c1">red</span>
<span class="c2">S</span>
</span></a>
```

```
.c2 { font-size: 0; }
```



Triplebyte
Sponsored ·

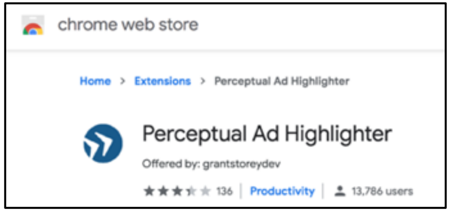


New arms race on HTML obfuscation
E.g., Facebook vs uBlockOrigin:
https://github.com/uBlockOrigin/uAssets/issues/3367
***>1 year, >275 comments**, and counting…*

Exact image matching is not enough

Perceptual Ad-Blocking

- **Ad Highlighter** [Storey et al., 2017]
  - > *Visually detects ad-disclosures*
  - > *Traditional computer vision techniques*
  - > *Similar techniques deployed in Adblock Plus*



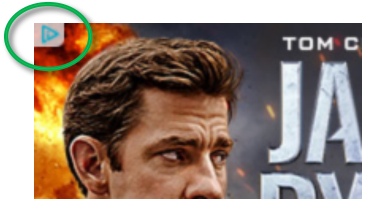
chrome web store

Home > Extensions > Perceptual Ad Highlighter

Perceptual Ad Highlighter

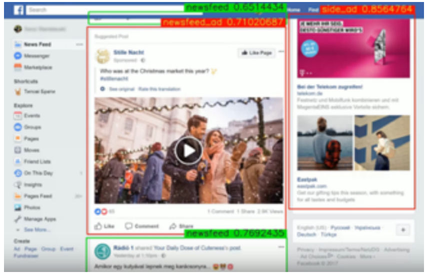
Offered by: grantstoreydev

★★★★ 136 | Productivity | 13,786 users

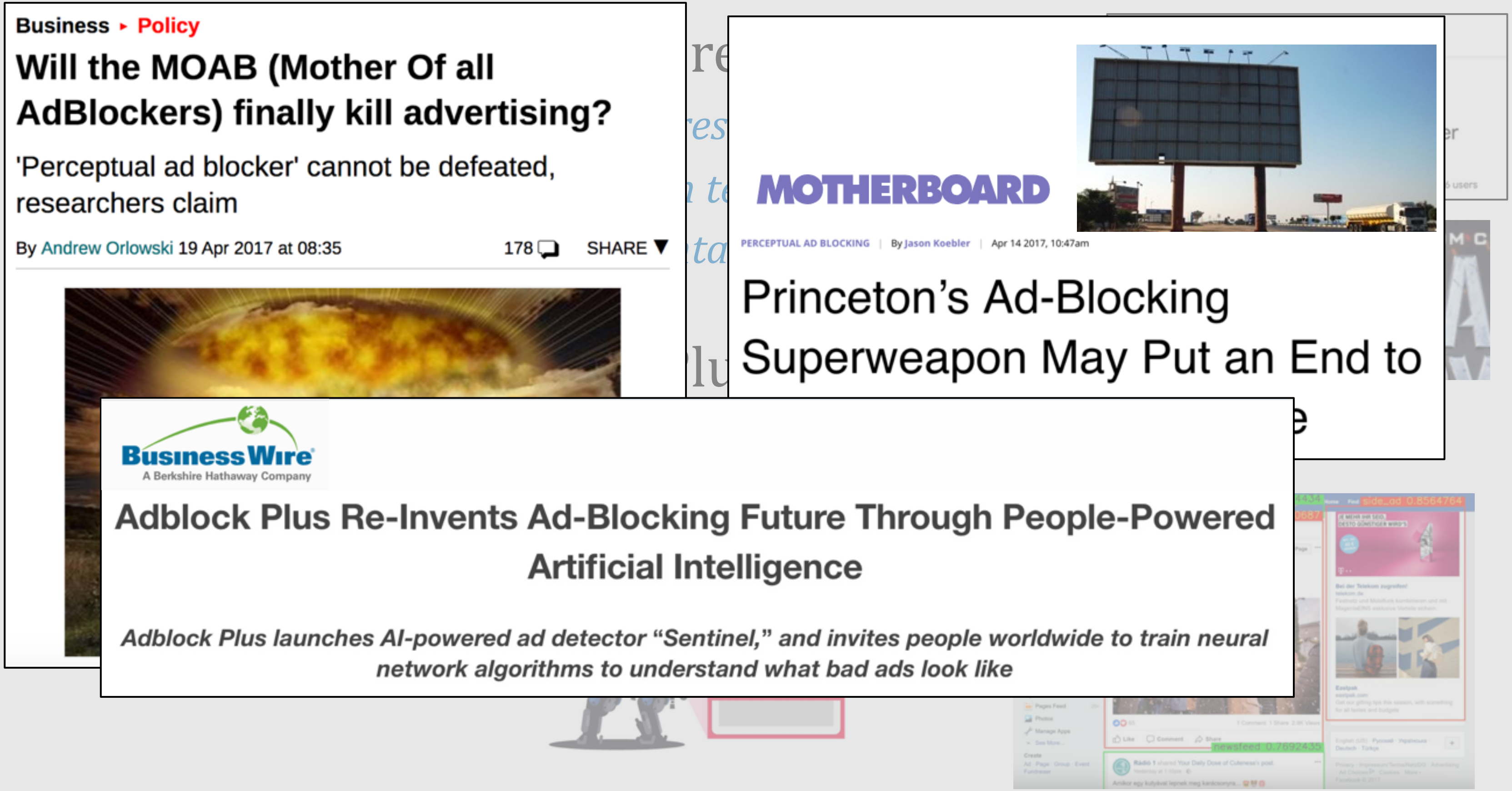


- **Sentinel** by Adblock Plus [Paraska, 2018]
  - > *Locates ads in Facebook screenshots* using *neural networks*

- **Percival** by Brave [Din et al., 2019]
  - > *Neural network embedded in Chromium's rendering pipeline*



Perceptual Ad-Blocking



Business ▸ Policy

Will the MOAB (Mother Of all AdBlockers) finally kill advertising?

'Perceptual ad blocker' cannot be defeated, researchers claim

By Andrew Orlowski 19 Apr 2017 at 08:35 178 SHARE ▼

MOTHERBOARD

PERCEPTUAL AD BLOCKING | By Jason Koebler | Apr 14 2017, 10:47am

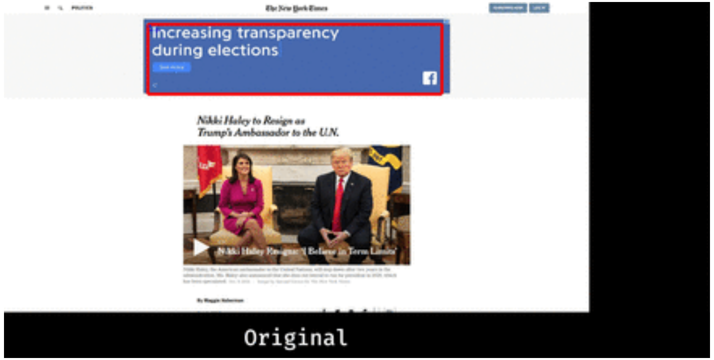
Princeton's Ad-Blocking Superweapon May Put an End to

BusinessWire
A Berkshire Hathaway Company

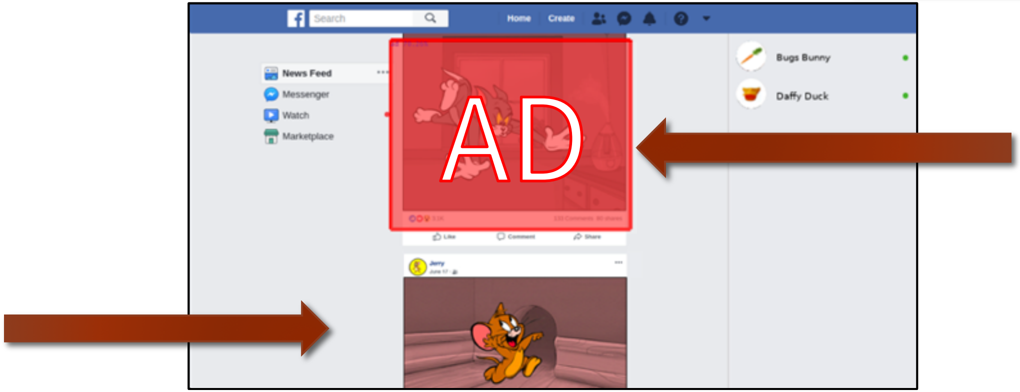
Adblock Plus Re-Invents Ad-Blocking Future Through People-Powered Artificial Intelligence

Adblock Plus launches AI-powered ad detector "Sentinel," and invites people worldwide to train neural network algorithms to understand what bad ads look like

How Secure is Perceptual Ad-Blocking?


The New York Times
Increasing transparency during elections
Nikki Haley to Resign as Trump's Ambassador to the U.N.
Nikki Haley Resigns: 'I Believe in Term Limits'
Original

AdChoices
AdChoices



Jerry uploads malicious content …

… so that Tom’s post gets blocked

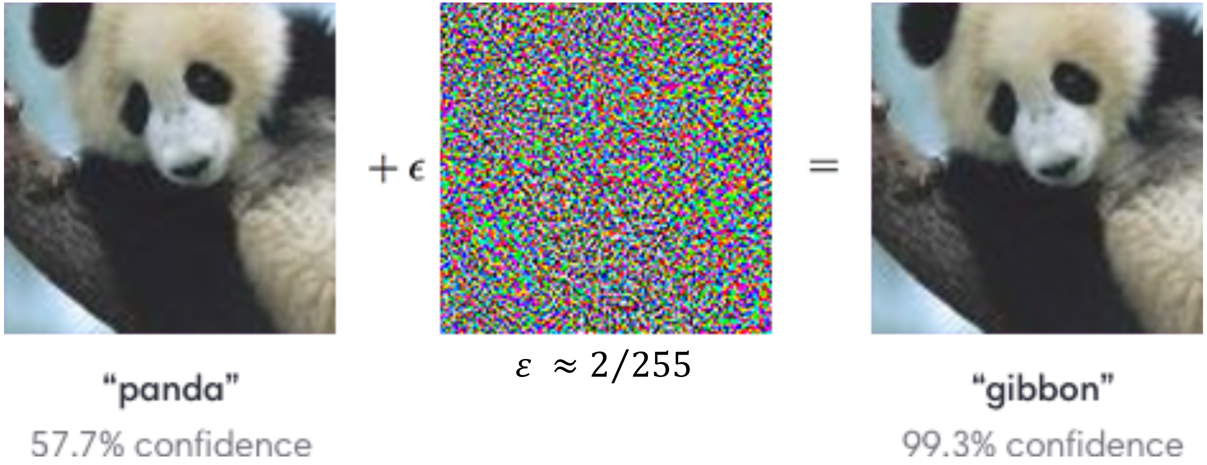
The Current State of ML

ML works well on average

$\neq$

ML works well on adversarial data

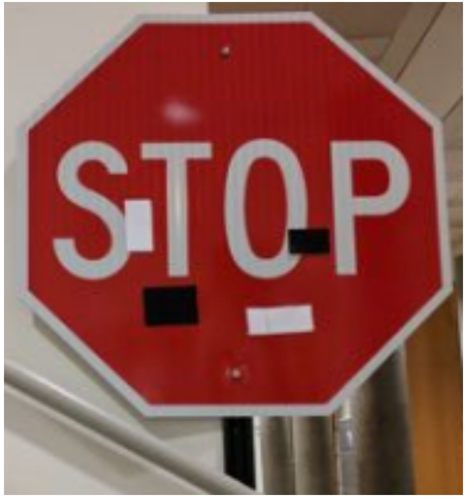
Adversarial Examples



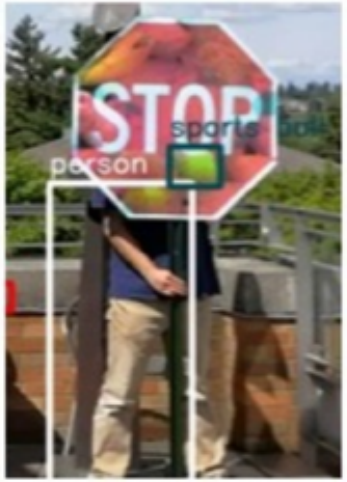
$\varepsilon \approx 2/255$

Szegedy et al., 2014
Goodfellow et al., 2015

What’s the Threat Model?



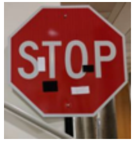
(Eykholt et al. 2017)



(Eykholt **et al.** 2018)

SCIENCE MUSEUM

What’s the Threat Model?



Is there an adversary?

✓

Are there no simpler attacks?

- ➢ *Misclassified clean examples?*
- ➢ *Attacks that affect human perception too?*

✓

White-box access to the model?

- ➢ *Or query access / access to training data?*

✓

Unless the answer to all these questions is *Yes*,
adversarial examples are likely not the most relevant threat

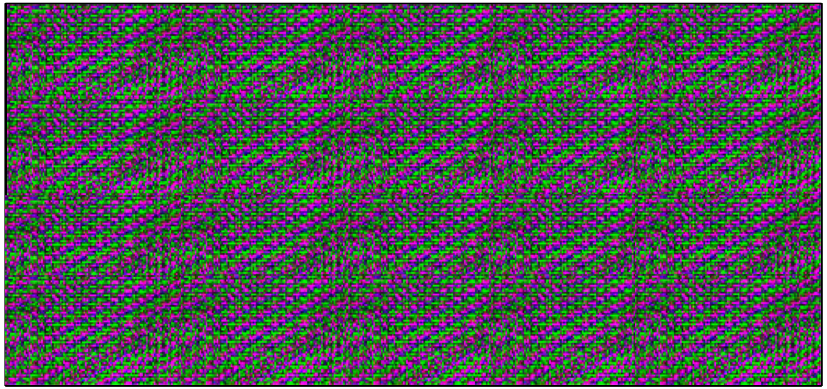
Adversarial Examples for Perceptual Ad-Blockers

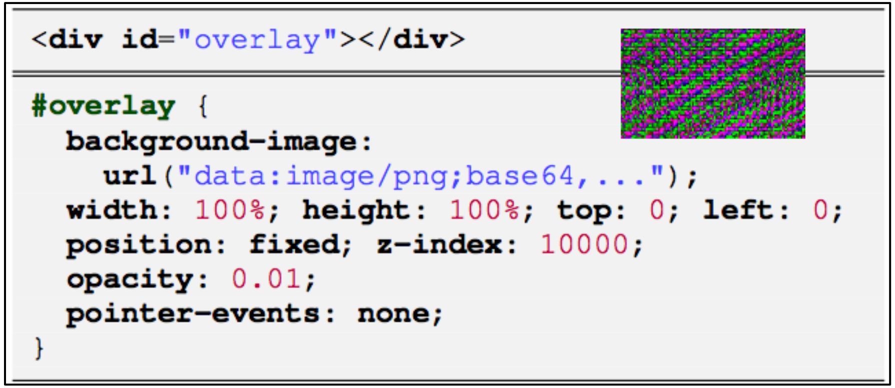
Ad-Block Evasion

- **Goal: Make ads unrecognizable by ad-blocker**
- Adversary = Website publisher
- Other adversaries exist (e.g., Ad-Network)

Evasion: Universal Transparent Overlay

Web publisher perturbs every rendered pixel





```
<div id="overlay"></div>

#overlay {
  background-image:
    url("data:image/png;base64,...");
  width: 100%; height: 100%; top: 0; left: 0;
  position: fixed; z-index: 10000;
  opacity: 0.01;
  pointer-events: none;
}
```

Use HTML *tiling* to minimize perturbation size (20 KB)

- ➢ 100% success rate on 20 webpages not used to create the overlay
- ➢ The attack is **universal:** the overlay is computed once and works for all (or most) websites
- ➢ Attack can be made stealthier without relying on CSS

Ad-Block Detection

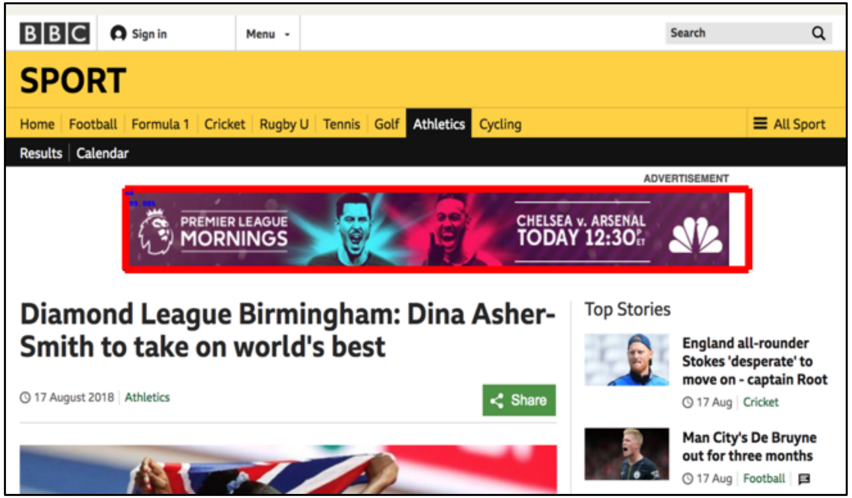
- **Goal: Trigger ad-blocker on “honeypot” content**
  - > *Detect ad-blocking in client-side JavaScript or on server*
  - > *Applicability of these attacks depends on ad-blocker type*

- Adversary = Website publisher
  - > *Use client-side JavaScript to detect DOM changes*

Detection: Perturb fixed page layout

Publisher adds honeypot in page-region with fixed layout

- *E.g., page header*



Diamond League Birmingham: Dina Asher-Smith to take on world's best

17 August 2018 | Athletics

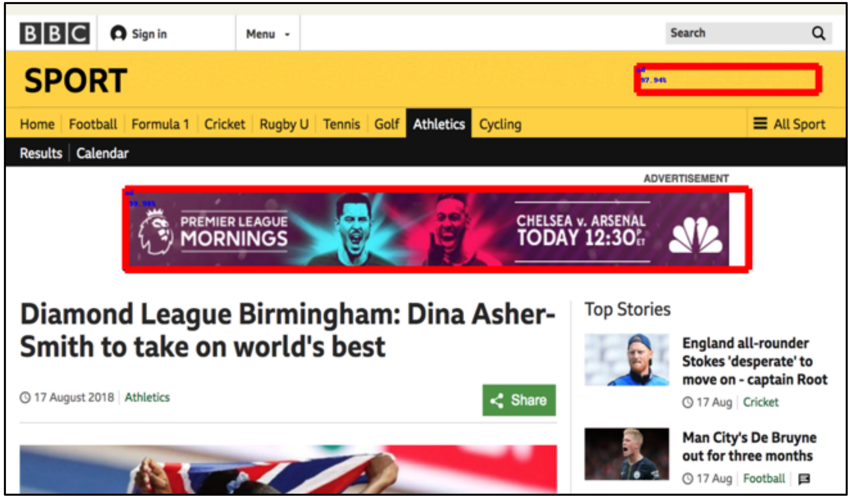
England all-rounder Stokes 'desperate' to move on - captain Root

17 Aug | Cricket

Man City's De Bruyne out for three months

17 Aug | Football

original



Diamond League Birmingham: Dina Asher-Smith to take on world's best

17 August 2018 | Athletics

England all-rounder Stokes 'desperate' to move on - captain Root

17 Aug | Cricket

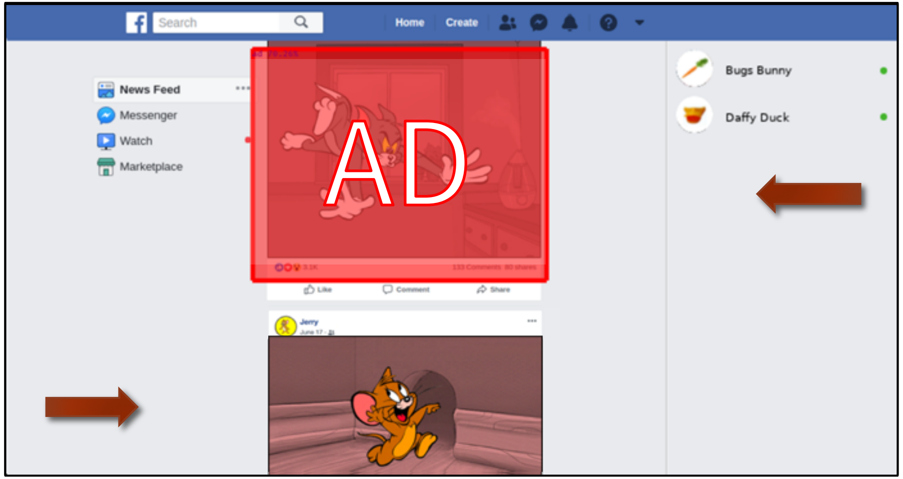
Man City's De Bruyne out for three months

17 Aug | Football

With honeypot header

New Threats: Privilege Abuse

Ad-block evasion & detection is a well-known arms race. But there's more!



Jerry uploads malicious content …

… so that Tom's post gets blocked

What happened?

- ➢ *Object detector model generates box predictions from full page inputs*
- ➢ *Content from one user can affect predictions anywhere on page*
- ➢ *Model's segmentation is not aligned with web-security boundaries*

Defense Strategies

- Obfuscate the blocker?
- Randomize the ad-blo
- Pro-actively retra model ersarial training)

The Most Challenging Threat Model for ML

- *Adversary has white-box access to ad-blocker*

- *Adversary can exploit False Negatives and False Positives in classification pipeline*

- *Adversary prepares attacks offline* ⬄ *The ad-blocker must defend against attacks in real-time in the user’s browser*

- *Adversary can take part in crowd-sourced data collection for training the ad-blocker*

Take Away

- Emulating human detection of ads *could be* the end-game for ad-blockers
  - *But very hard (impossible?) with current computer vision techniques*

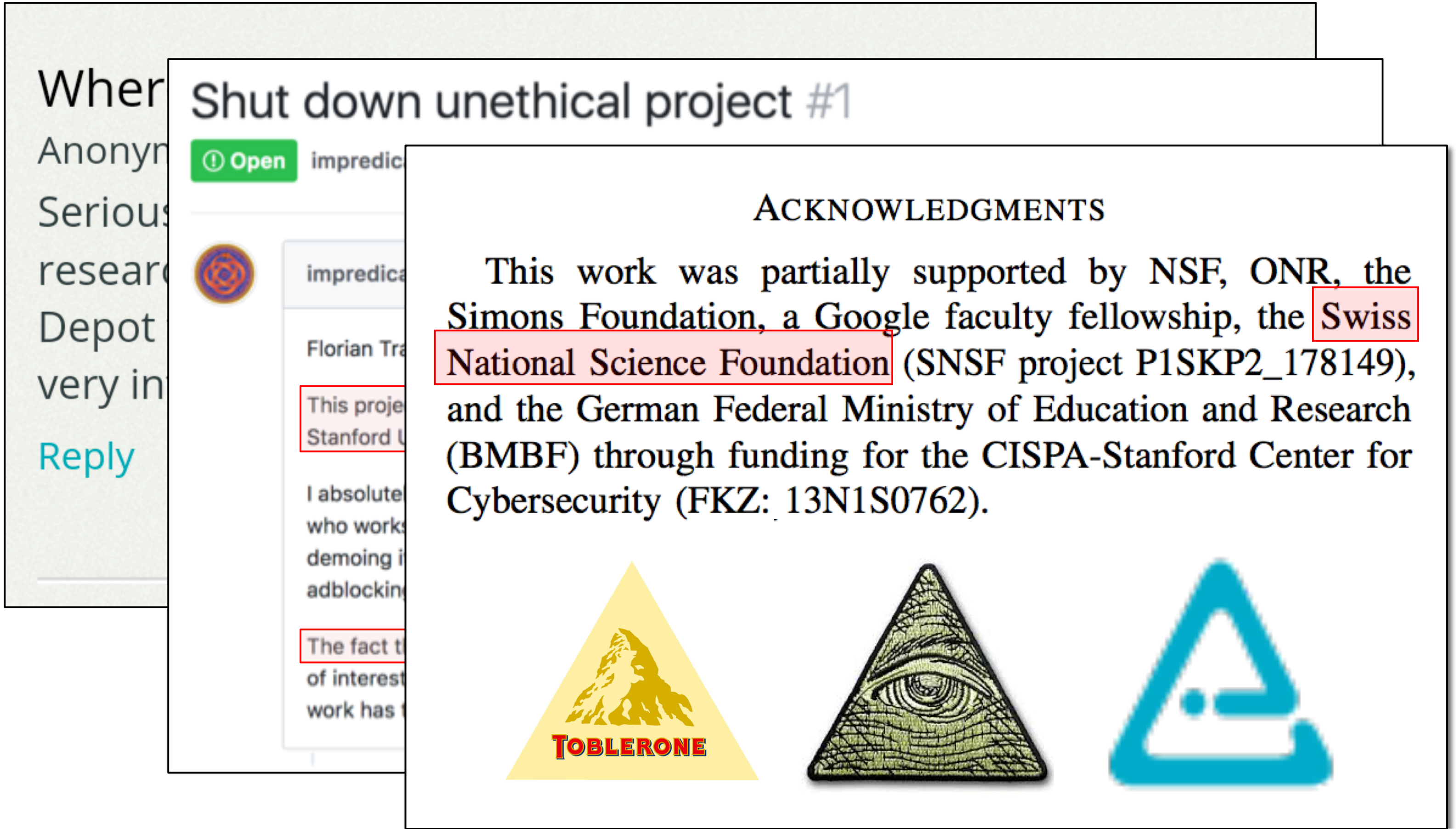
- Perceptual ad-blockers must survive an extremely strong threat model
  - *This threat model perfectly aligns with white-box adversarial examples*
  - *Will we soon see adversarial examples used by real-world adversaries?*

- More in the paper
  - *Unified architecture + attacks for all perceptual ad-blocker designs*
  - *Similar attacks for non-Web ad-blockers (e.g., Adblock Radio)*

ftramer / **ad-versarial**

- Train a page-based ad-blocker
- Download pre-trained models
- Attack demos

Research Impact



Wher

Anonyn

Serious

researc

Depot

very in

Reply

Shut down unethical project #1

impredic

impredica

Florian Tra

This proje
Stanford U

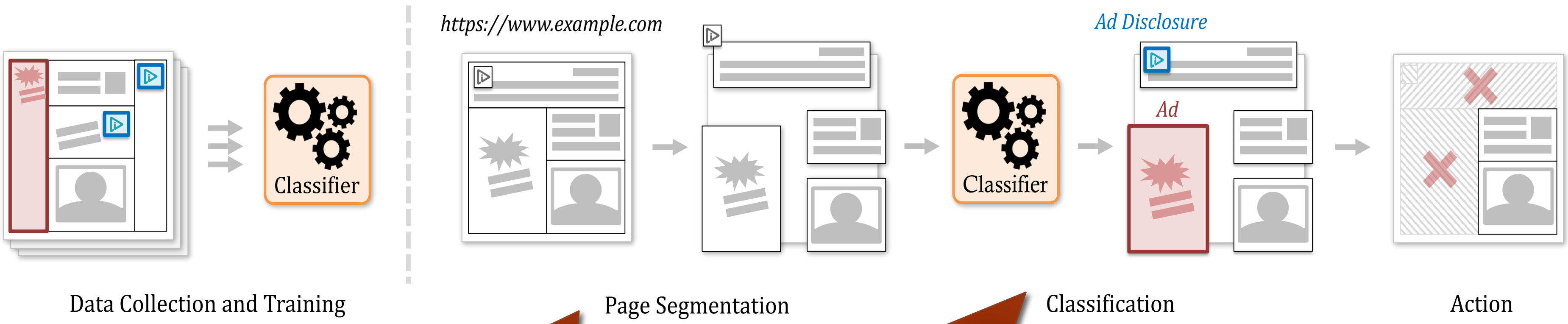
I absolute
who works
demoing i
adblockin

The fact t
of interest
work has

ACKNOWLEDGMENTS

This work was partially supported by NSF, ONR, the Simons Foundation, a Google faculty fellowship, the Swiss National Science Foundation (SNSF project P1SKP2_178149), and the German Federal Ministry of Education and Research (BMBF) through funding for the CISPA-Stanford Center for Cybersecurity (FKZ: 13N1S0762).

How does a Perceptual Ad-Blocker Work?

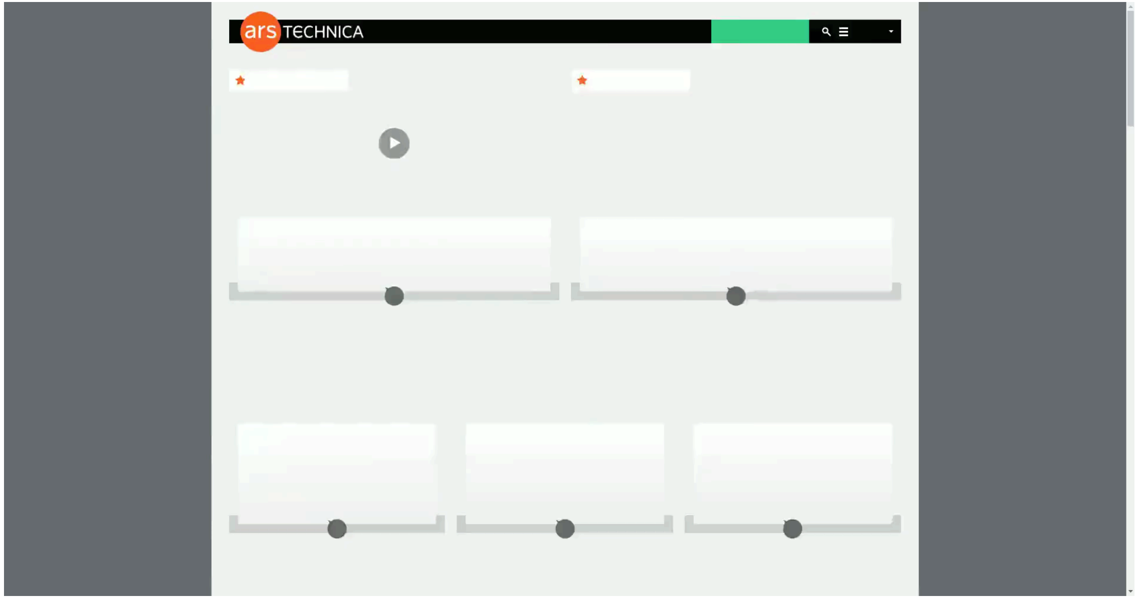


Template matching, OCR, DNNs, Object detector networks

- **Element-based** (e.g., find all <img> tags) [Storey et al. 2017]
- Frame-based (segment rendered webpage into “frames” as in Percival)
- **Page-based** (unsegmented screenshots à-la-Sentinel)

Building a Page-Based Ad-Blocker

We trained a neural network to detect ads on news websites from all G20 nations



Video taken from 5 websites *not used during training*

Defense Strategies

- Obfuscate the ad-blocker?
  - *It isn’t hard to create adversarial examples for black-box classifiers*

- Randomize the ad-blocker?
  - *Adversarial examples robust to random transformations / multiple models*

- Pro-actively retrain the model? (Adversarial training)
  - *New arms-race: The adversary finds new attacks and ad-blocker re-trains*
  - *Mounting a new attack is much easier than updating the model*
  - *On-going research: so far the adversary always wins!*