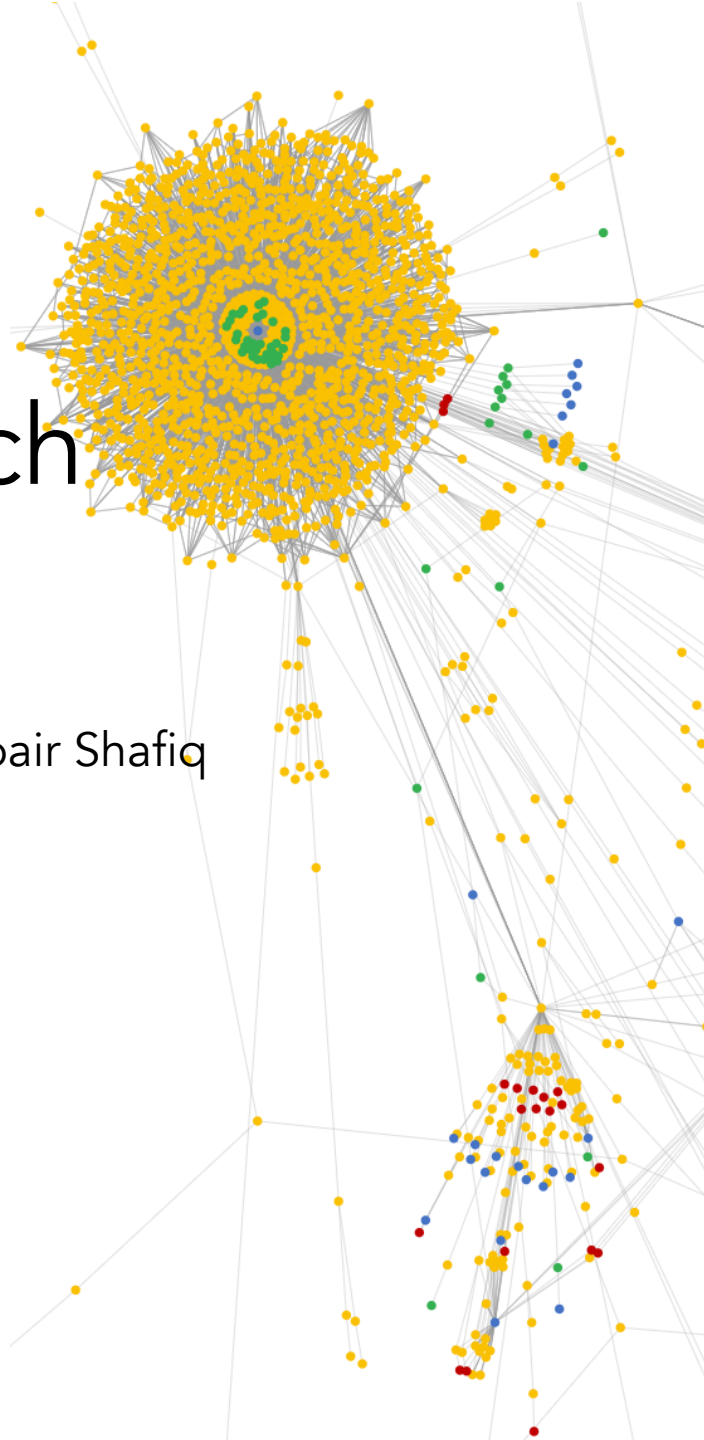
AdGraph: A Graph-Based Approach to Ad and Tracker Blocking

Umar Iqbal, Peter Snyder, Shitong Zhu, Benjamin Livshits, Zhiyun Qian, and Zubair Shafiq

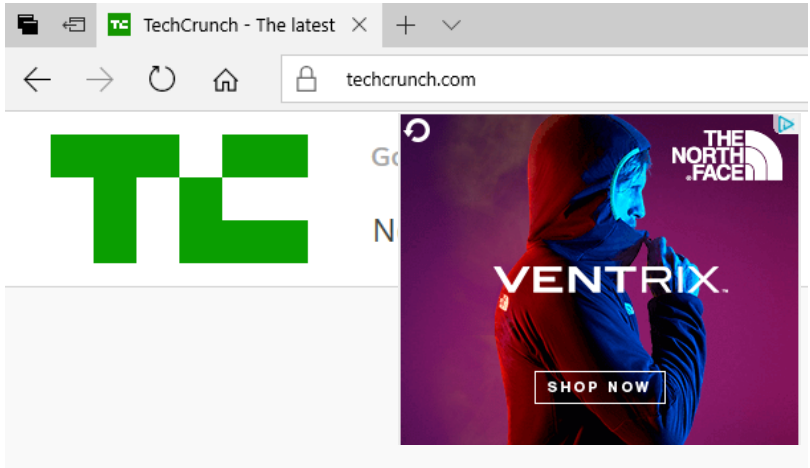
IEEE Symposium on Security and Privacy, 2020



Online Advertising

Advertising enables “free”content

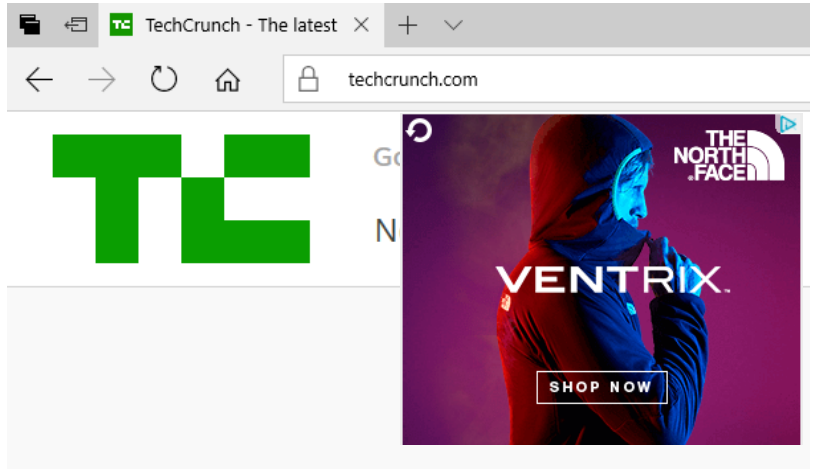
- Publishers show content
- Earn revenue with ads



Online Advertising

Advertising enables “free”content

- Publishers show content
- Earn revenue with ads



Digital revenues for full year 2018 surpassed $100 billion for the first time

Interactive Advertising Bureau (IAB) ‘19

Online Advertising

Advertising enables “free”content

- Publishers show content
- Earn revenue with ads

Problems with online advertising ecosystem

Online Advertising

Advertising enables “free”content

- Publishers show content
- Earn revenue with ads

Problems with online advertising ecosystem

- Privacy concerns – Behavioral targeting

“I see ads for things I dream about.”

“My phone is eavesdropping on me”

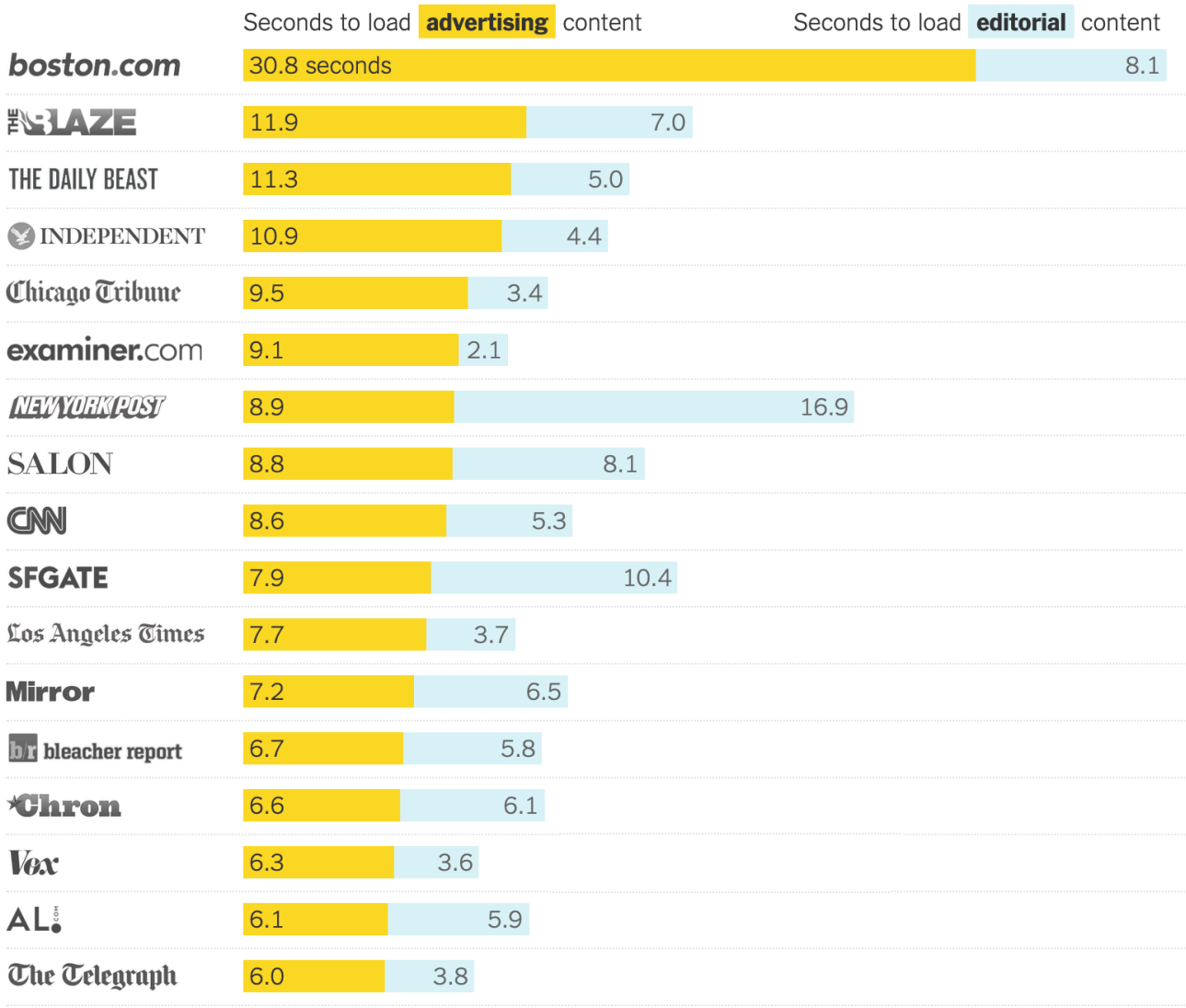
Online Advertising

Advertising enables "free"content

- Publishers show content
- Earn revenue with ads

Problems with online advertising ecosystem

- Privacy concerns – Behavioral targeting
- Performance issues – Slow page load



Online Advertising

Advertising enables “free”content

- Publishers show content
- Earn revenue with ads

Problems with online advertising ecosystem

- Privacy concerns – Behavioral targeting
- Performance issues – Slow page load
- Malvertising

Hackers have breached 60 ad servers to load their own malicious ads

Why buy legitimate ad slots to deliver malvertising when you can just hack the server instead.

Major sites including New York Times and BBC hit by 'ransomware' malvertising

Adverts hijacked by malicious campaign that demands payment in bitcoin to unlock user computers

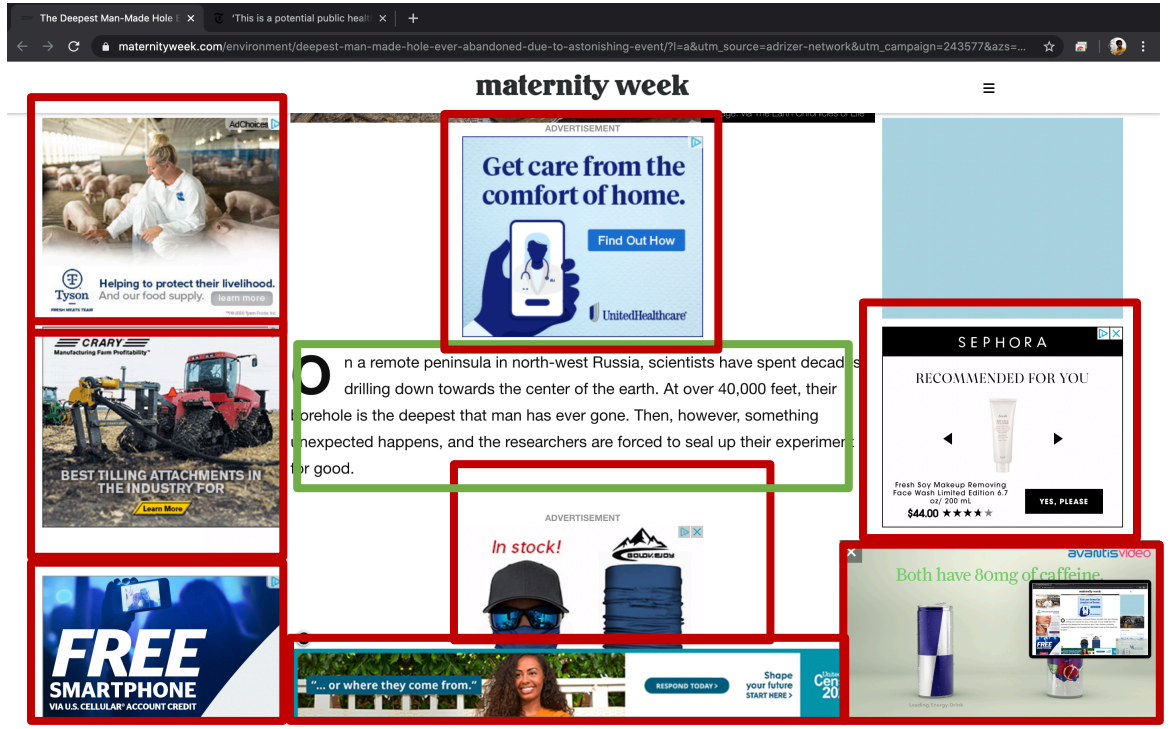
Online Advertising

Advertising enables “free”content

- Publishers show content
- Earn revenue with ads

Problems with online advertising ecosystem

- Privacy concerns – Behavioral targeting
- Performance issues – Slow page load
- Malvertising
- Intrusive

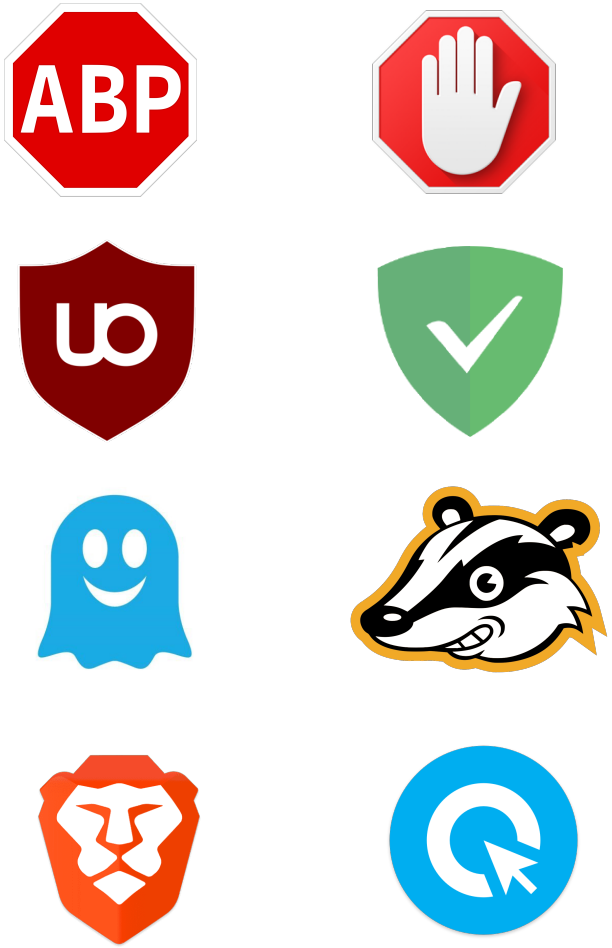


Online Advertising

Advertising enables “free”content
- Publishers show content
- Earn revenue with ads

Problems with online advertising ecosystem
- Privacy concerns – Behavioral targeting
- Performance issues – Slow page load
- Malvertising
- Intrusive

Solution
- Ad & tracker blockers



Outline

State of Ad/Tracker Blocking

- Ads & Trackers
- Filter list blocking
- Machine learning based blocking

AdGraph

- Graph-based representation
- Machine learning on graph representation
- Evaluation

Outline

State of Ad/Tracker Blocking

- Ads & Trackers
- Filter list blocking
- Machine learning based blocking

AdGraph

- Graph-based representation
- Machine learning on graph representation
- Evaluation

What are Ads and Trackers?

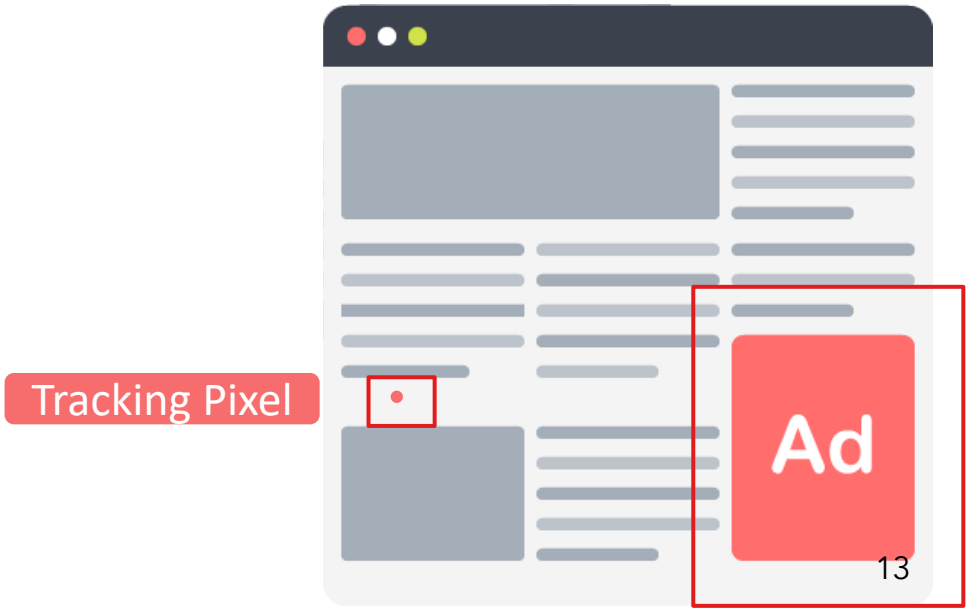
What are Ads and Trackers?

Ads are audio-visual promotional content

What are Ads and Trackers?

Ads are audio-visual promotional content

Trackers collect sensitive information



What are Ads and Trackers?

```
<script>
    eval("
      var iframe = document.createElement("iframe");
      iframe.src = "adnetwork.com/load_ads";
      document.body.append(iframe);
    ");
</script>
<iframe src = "adnetwork.com/load_ads">
    <script>
      var img = document.createElement("img");
      iframe.src = "adnetwork.com/load_ads";
    </script>
</iframe>
```

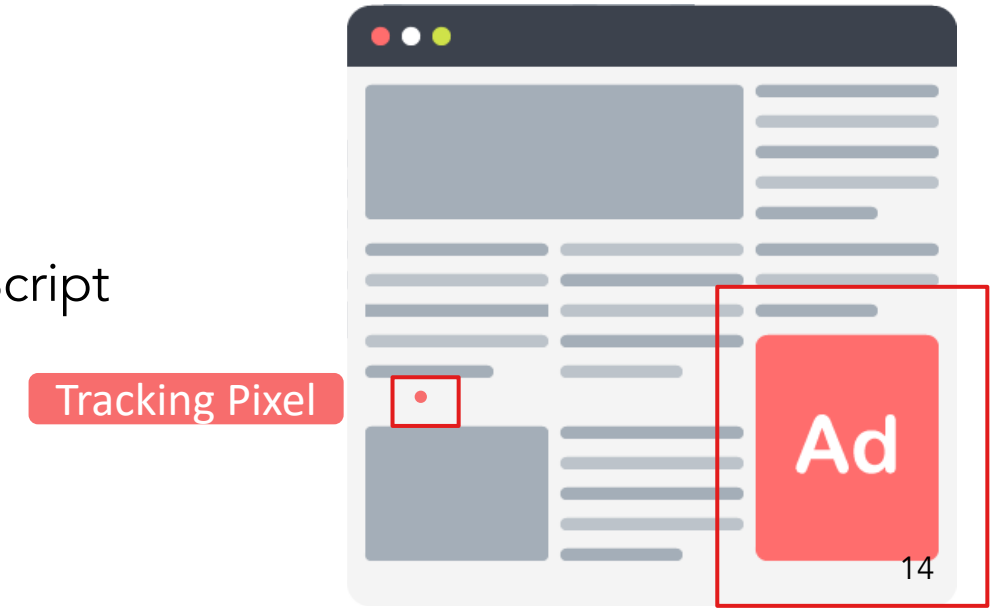
Ads are audio-visual promotional content

Trackers collect sensitive information

They are:

- Created with JavaScript
- Requested with HTTP
- Displayed with HTML

Ads and trackers involve HTML, Network, and JavaScript



What are Ads and Trackers?

Ads are audio-visual promotional content

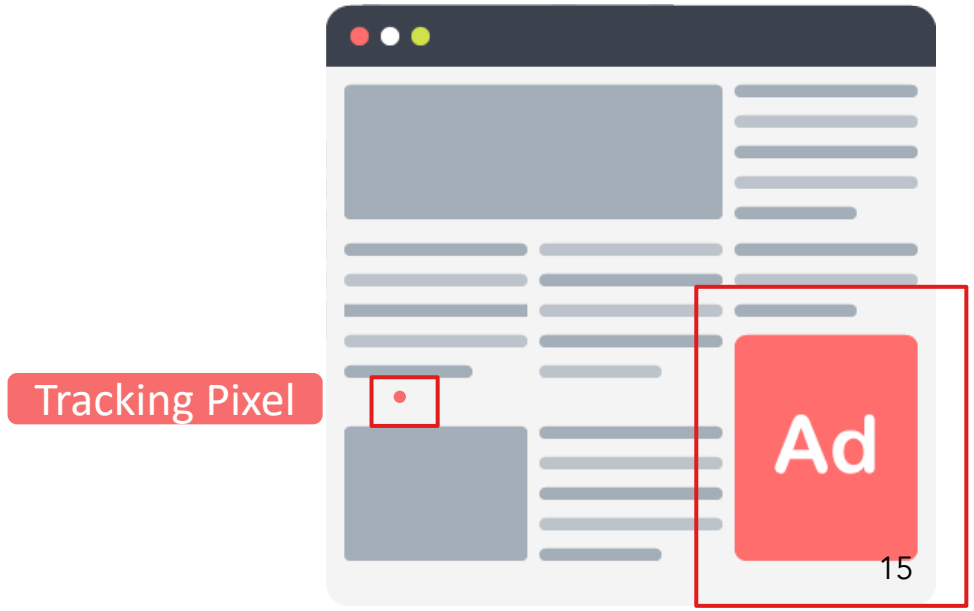
Trackers collect sensitive information

They are:

Created with JavaScript

JavaScript

```
<script>
    eval("
      var iframe = document.createElement("iframe");
      iframe.src = "adnetwork.com/load_ads";
      document.body.append(iframe);
    ");
</script>
<iframe src = "adnetwork.com/load_ads">
    <script>
      var img = document.createElement("img");
      iframe.src = "adnetwork.com/load_ads";
    </script>
</iframe>
```



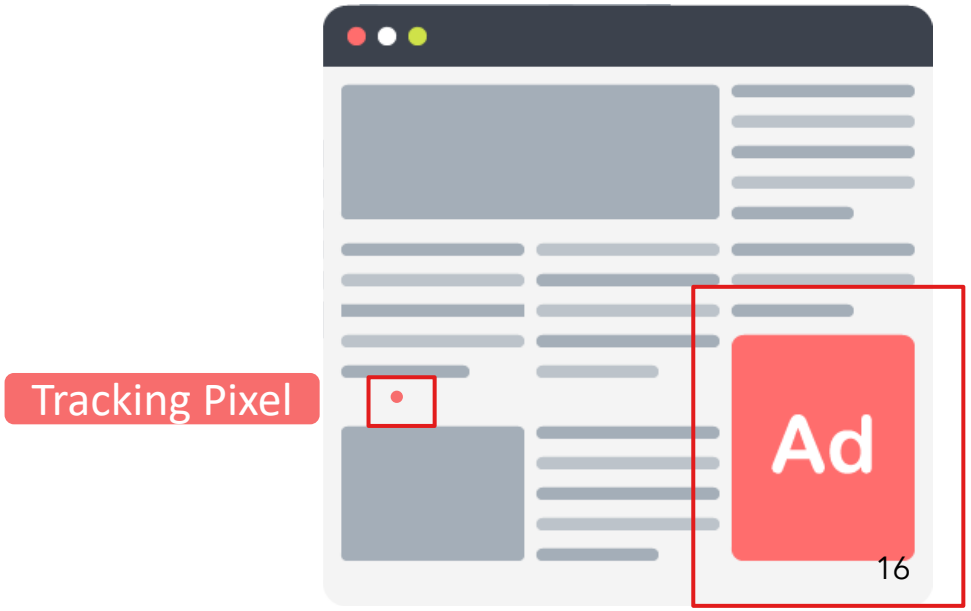
What are Ads and Trackers?

Ads are audio-visual promotional content

Trackers collect sensitive information

They are:

- Created with JavaScript
- Requested with HTTP



What are Ads and Trackers?

Ads are audio-visual promotional content

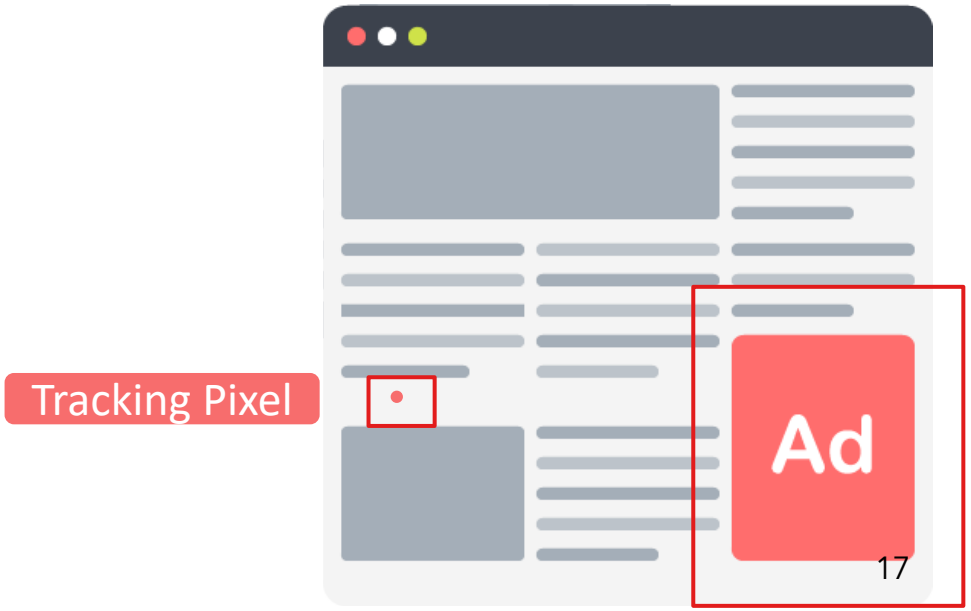
Trackers collect sensitive information

They are:

- Created with JavaScript
- Requested with HTTP
- Displayed with HTML

HTML

```
<script>
    eval("
      var iframe = document.createElement("iframe");
      iframe.src = "adnetwork.com/load_ads";
      document.body.append(iframe);
    ");
</script>
<iframe src = "adnetwork.com/load_ads">
    <script>
      var img = document.createElement("img");
      iframe.src = "adnetwork.com/load_ads";
    </script>
</iframe>
```



What are Ads and Trackers?

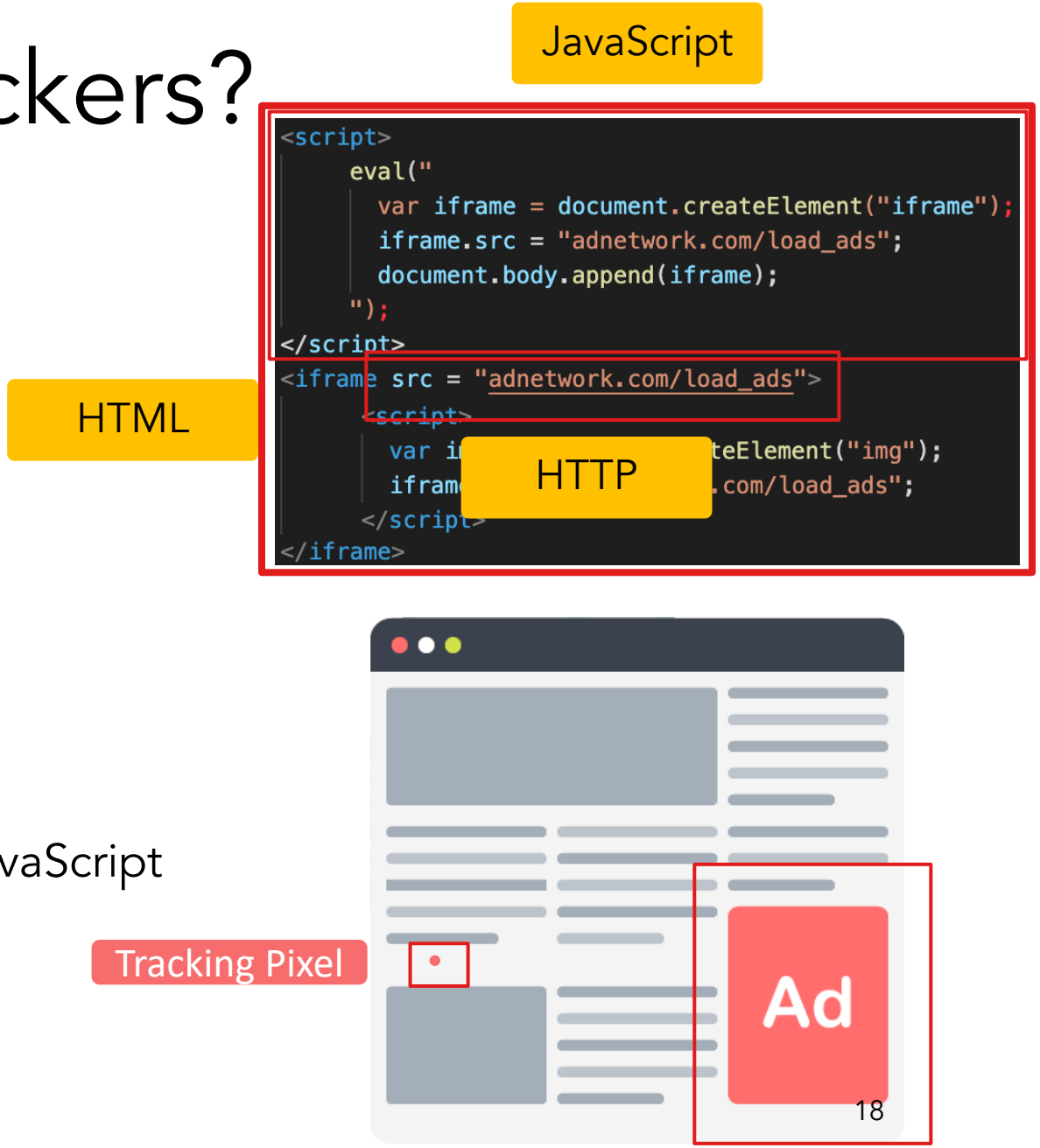
Ads are audio-visual promotional content

Trackers collect sensitive information

They are:

- Created with JavaScript
- Requested with HTTP
- Displayed with HTML

Ads and trackers involve HTML, Network, and JavaScript



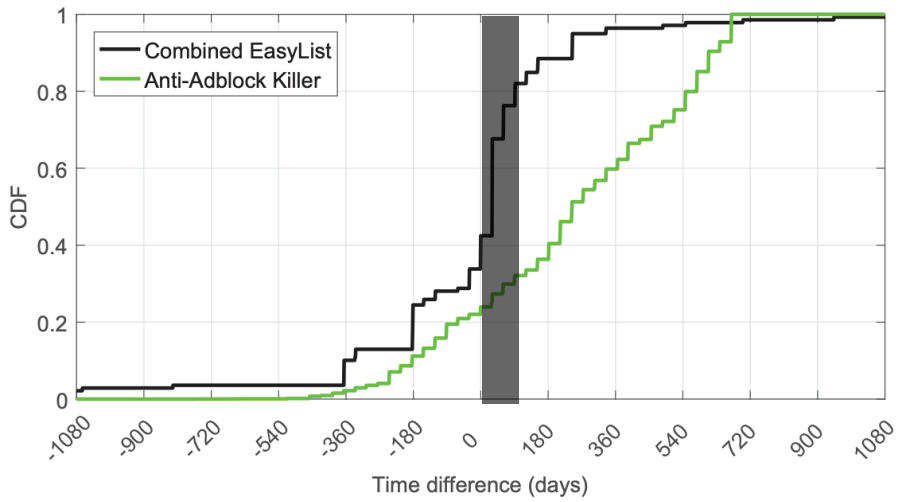
Filter List Based Blocking

Manually curated with crowdsourcing

Filter List Based Blocking

Manually curated with crowdsourcing

Leads to scalability issues

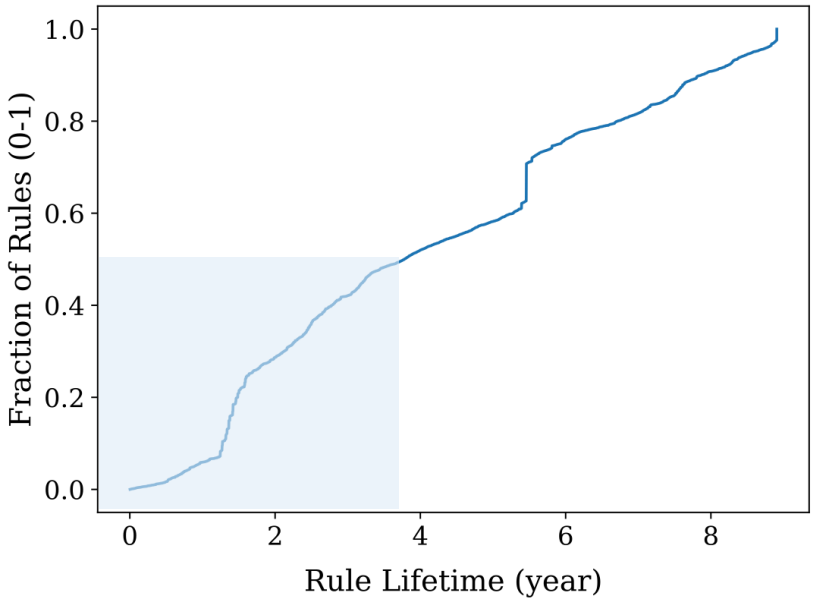


3 months to add new rules [Iqbal et al. ‘17]

Filter List Based Blocking

Manually curated with crowdsourcing

Leads to scalability issues

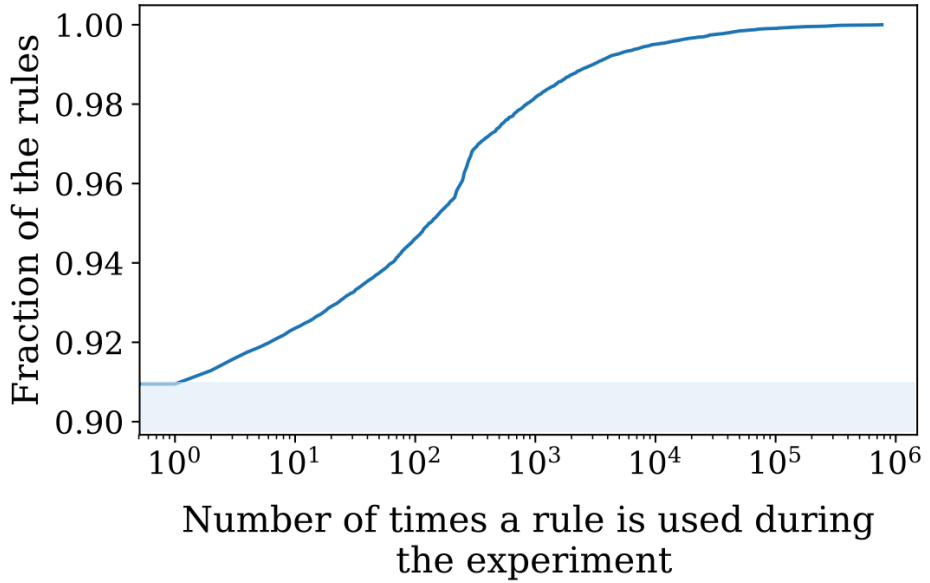


3.8 year to remove rules [Snyder et al. ‘20]

Filter List Based Blocking

Manually curated with crowdsourcing

Leads to scalability issues



90% rules are useless [Snyder et al. ‘20]

Filter List Based Blocking

Manually curated with crowdsourcing

Leads to scalability issues

Operate at HTML/Network/JS layer in isolation

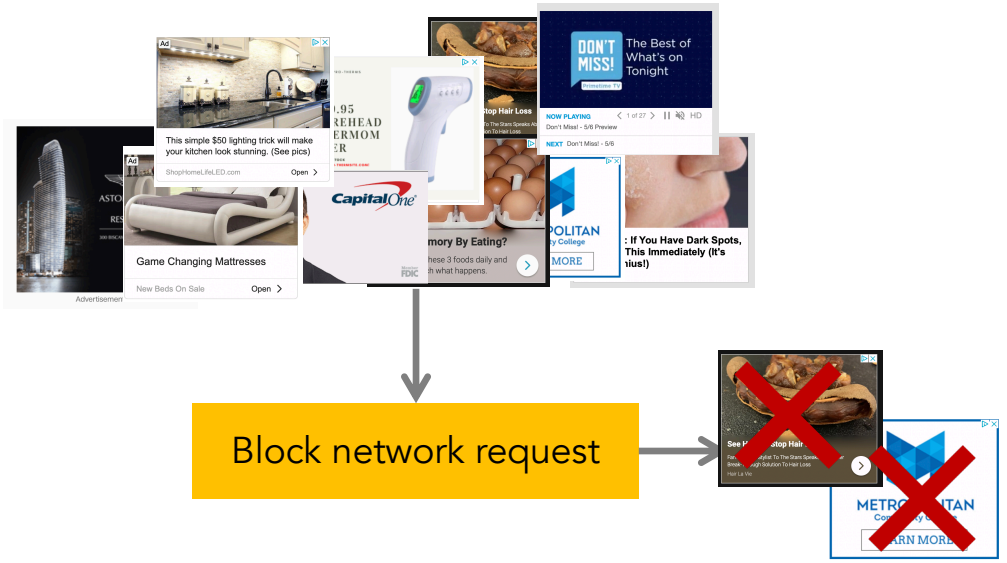
Filter List Based Blocking

Manually curated with crowdsourcing

Leads to scalability issues

Operate at HTML/Network/JS layer in isolation

Leads to accuracy issues



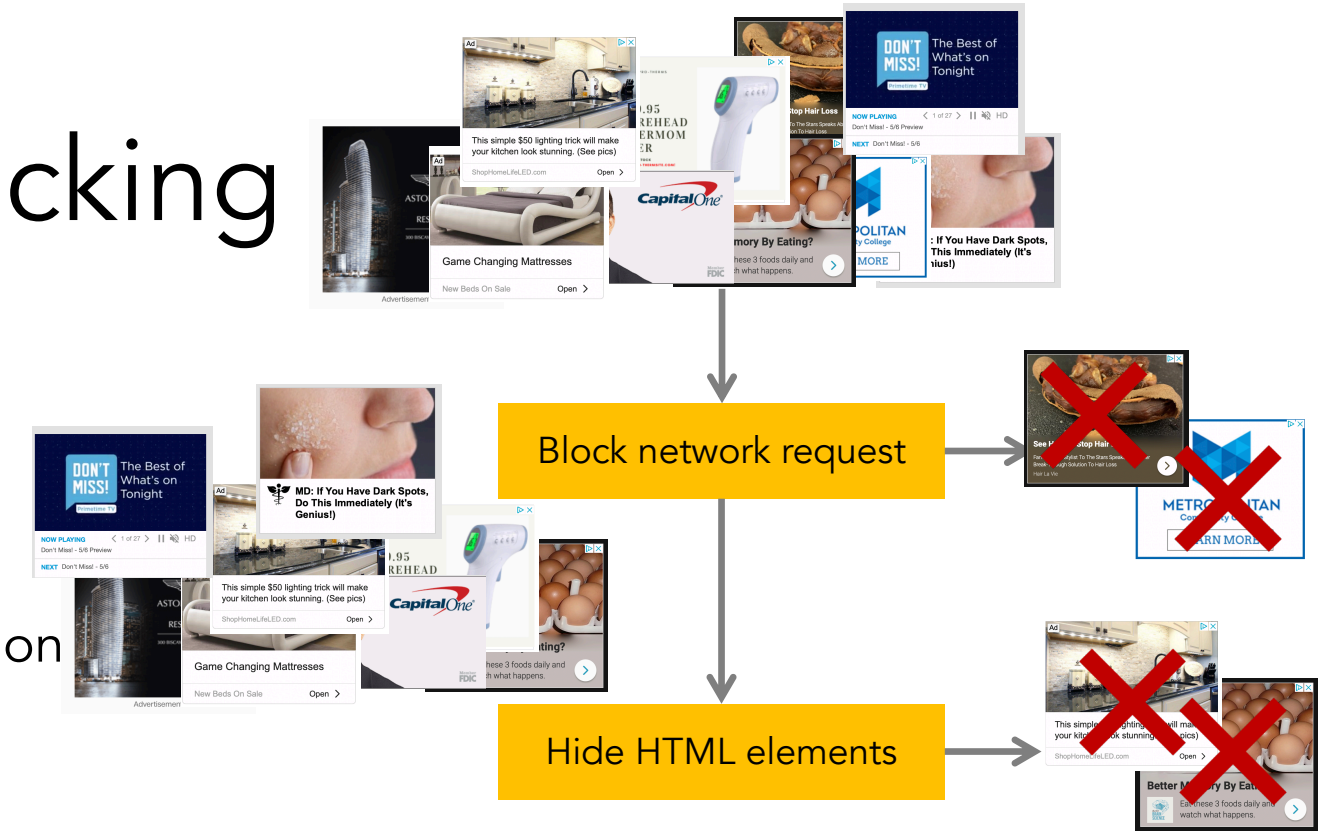
Filter List Based Blocking

Manually curated with crowdsourcing

Leads to scalability issues

Operate at HTML/Network/JS layer in isolation

Leads to accuracy issues



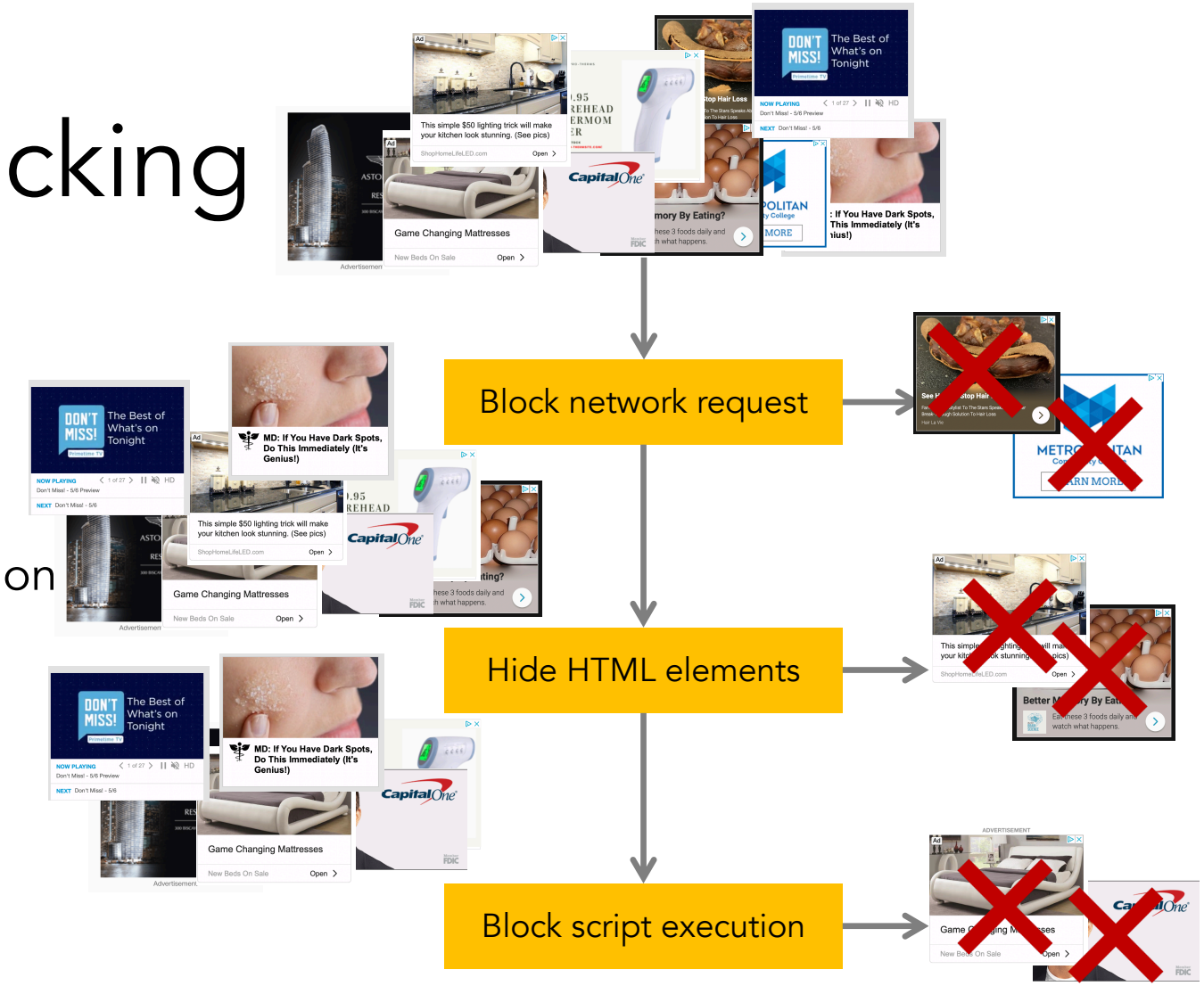
Filter List Based Blocking

Manually curated with crowdsourcing

Leads to scalability issues

Operate at HTML/Network/JS layer in isolation

Leads to accuracy issues



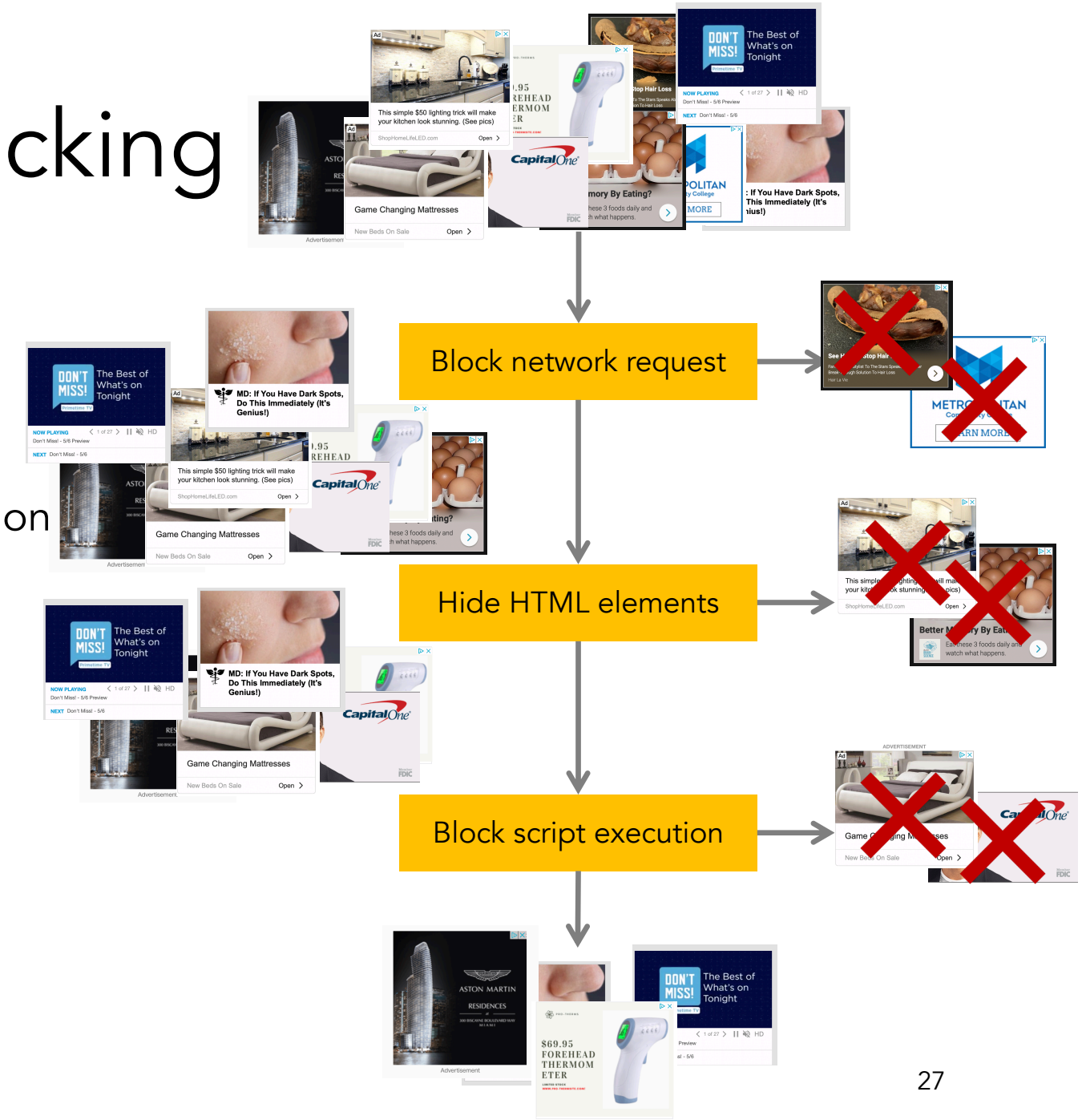
Filter List Based Blocking

Manually curated with crowdsourcing

Leads to scalability issues

Operate at HTML/Network/JS layer in isolation

Leads to accuracy issues



Filter List Based Blocking

Manually curated with crowdsourcing

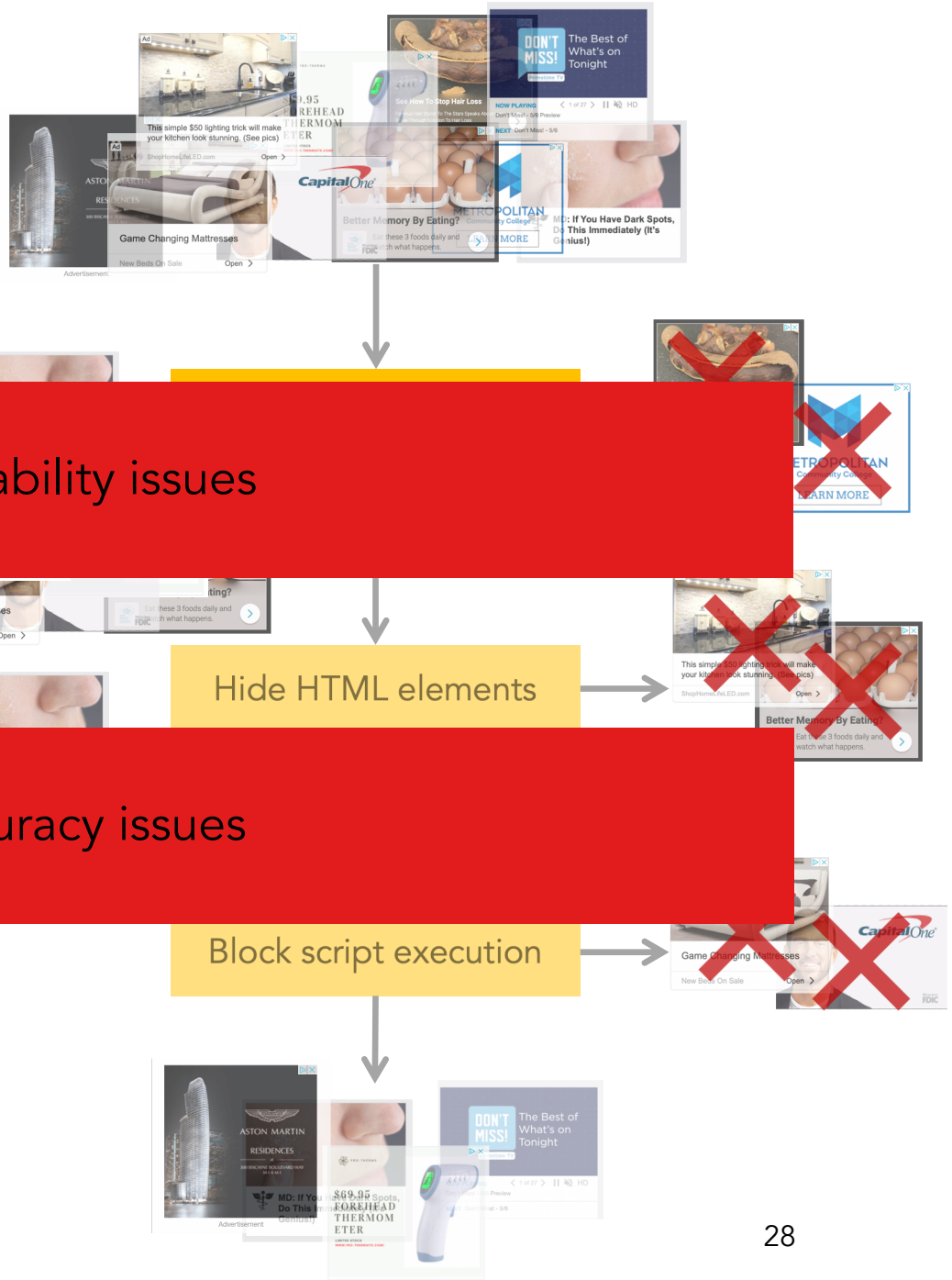
to

Suffer from scalability issues

Operate at HTML/Network/JS layer in isolation

Leads to accuracy issues

Suffer from accuracy issues



Machine Learning Based Blocking

Network layer [Bhagavatula et al. 14, Gugelmann et al. ’15]

- HTTP header properties as features
  - presence of words like “ad”
  - cookies set by response

Machine Learning Based Blocking

Network layer [Bhagavatula et al. 14, Gugelmann et al. ’15]

- HTTP header properties as features
  - presence of words like “ad”
  - cookies set by response

JavaScript layer [Wu et al. ‘16, Ikram et al. ‘17]

- JS API names as features
  - document.cookie
  - element.clientWidth

Machine Learning Based Blocking

Network layer [Bhagavatula et al. 14, Gugelmann et al. ’15]

HTTP he

es

ool

Solve scalability issues

JavaScript layer [Wu et al. ’16, Ikram et al. ’17]

JS API names as features

document.cookie

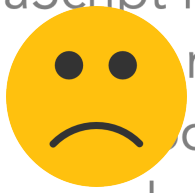
element.clientWidth

Machine Learning Based Blocking

Network layer [Bhagavatula et al. 14, Gugelmann et al. ’15]

Solve scalability issues

JavaScript la



Do not solve accuracy issues

element.clientWidth

Outline

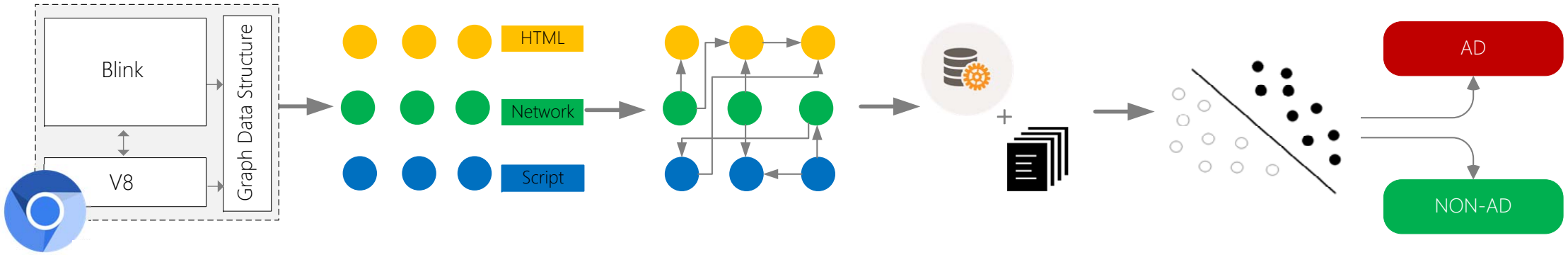
State of Ad/Tracker Blocking

- Online advertising
- Filter list blocking
- Machine learning based blocking

AdGraph

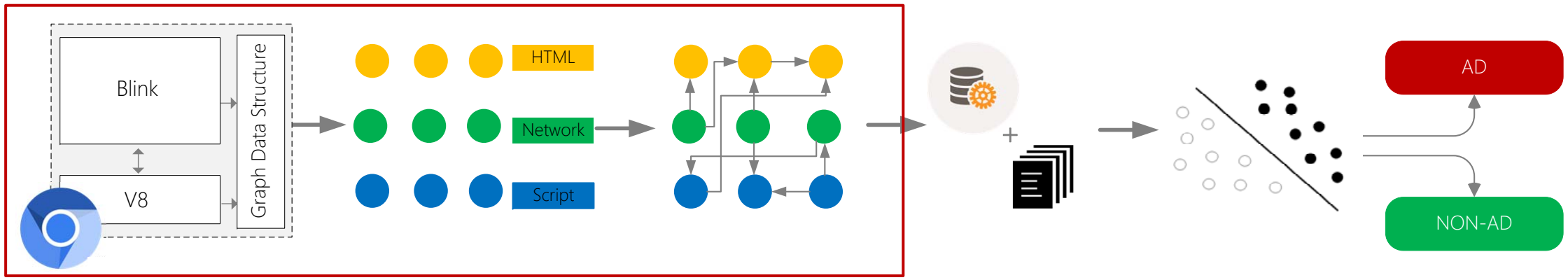
- Graph-based representation
- Machine learning on graph representation
- Evaluation

AdGraph



AdGraph

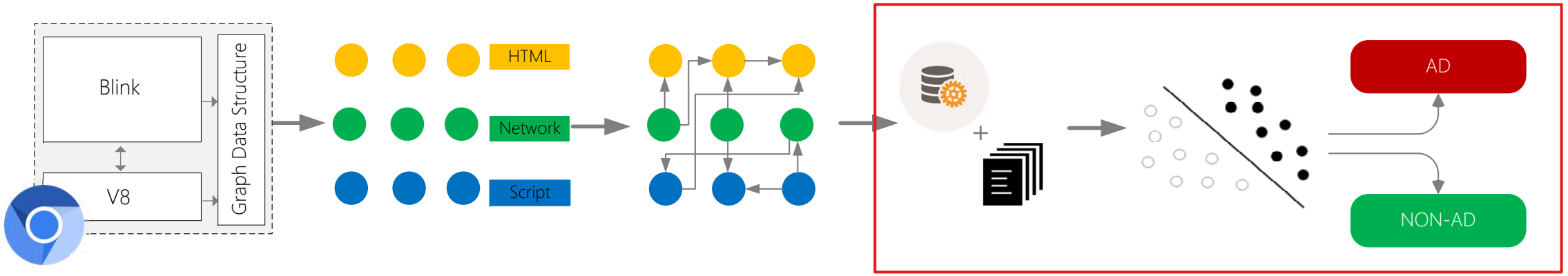
Graph-based cross-layer representation of ad/tracker behavior



AdGraph

Graph-based cross-layer representation of ad/tracker behavior

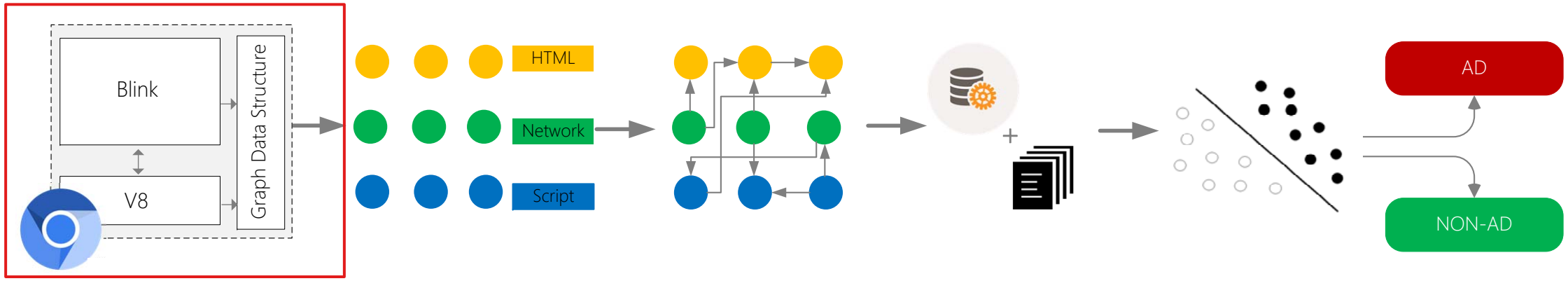
ML to automatically learn ad/tracker behavior



AdGraph

Graph-based cross-layer representation of ad/tracker behavior

ML to automatically learn ad/tracker behavior

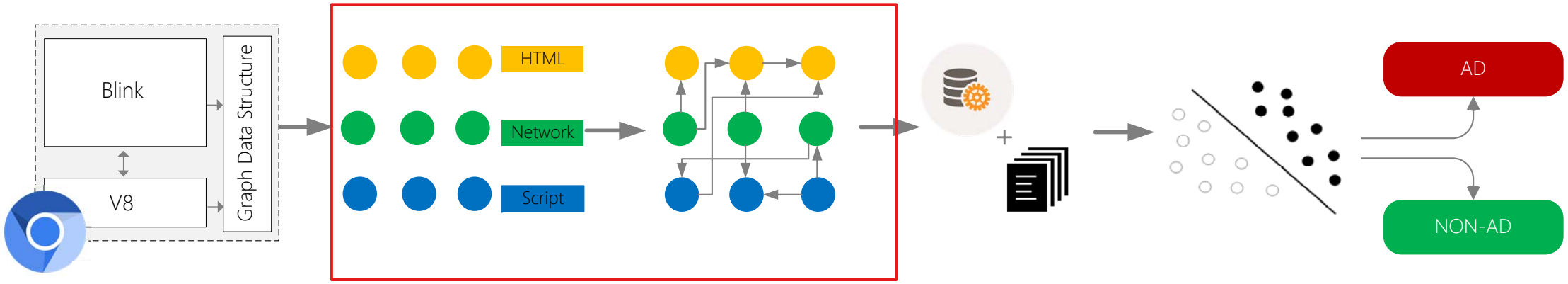


Chromium instrumentation

AdGraph

Graph-based cross-layer representation of ad/tracker behavior

ML to automatically learn ad/tracker behavior



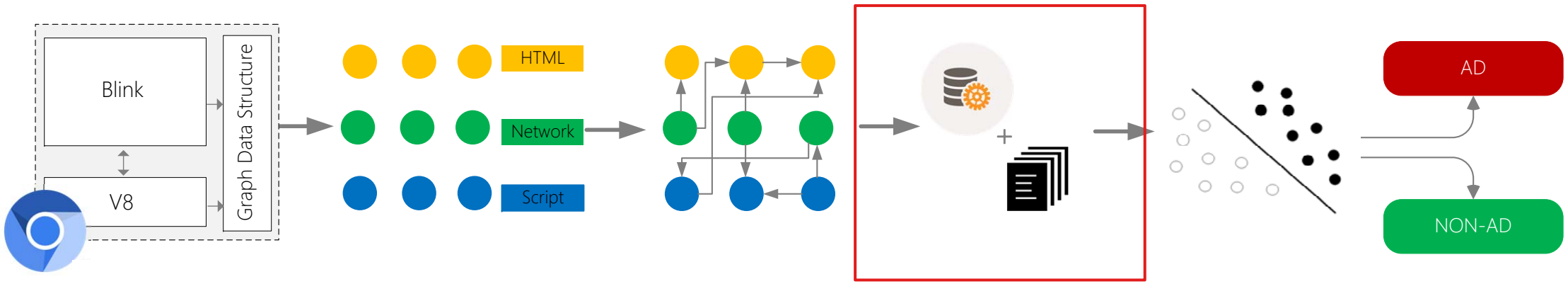
Chromium instrumentation

Graph representation

AdGraph

Graph-based cross-layer representation of ad/tracker behavior

ML to automatically learn ad/tracker behavior



Chromium instrumentation

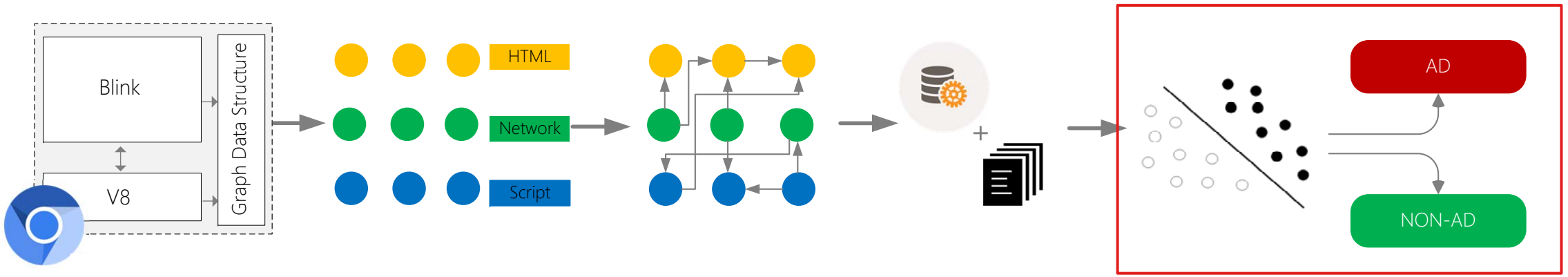
Graph representation

Model training

AdGraph

Graph-based cross-layer representation of ad/tracker behavior

ML to automatically learn ad/tracker behavior



Chromium instrumentation

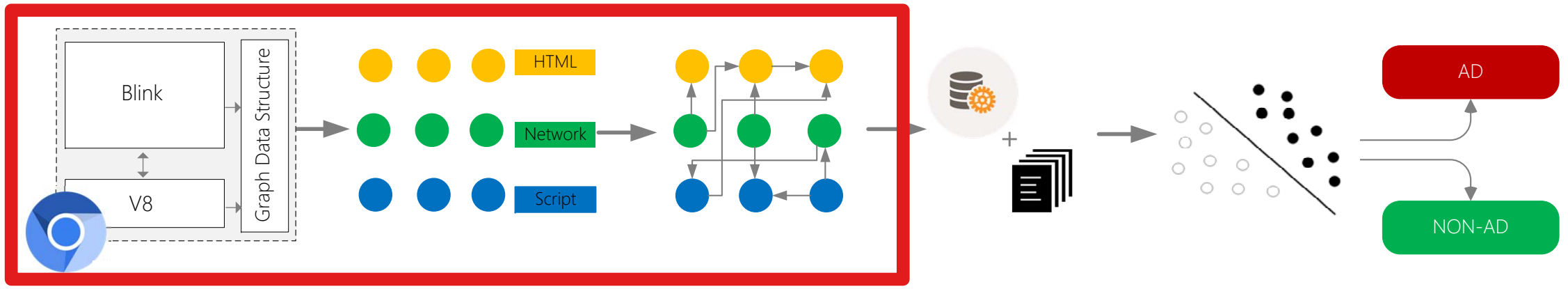
Graph representation

Model training

Classification decision

AdGraph

Graph-based cross-layer representation of ad/tracker behavior



Chromium instrumentation

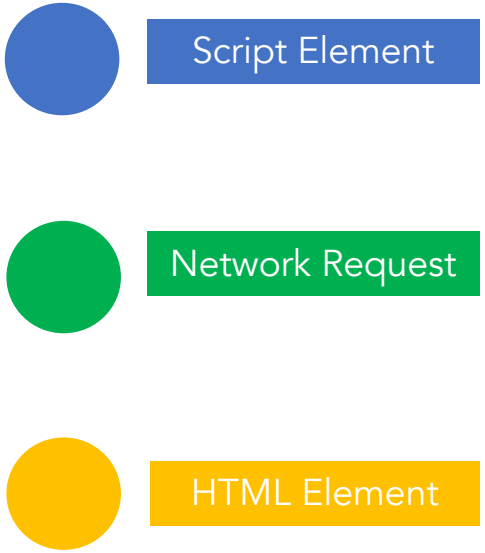
Graph representation

Model training

Classification decision

Cross-layer Context

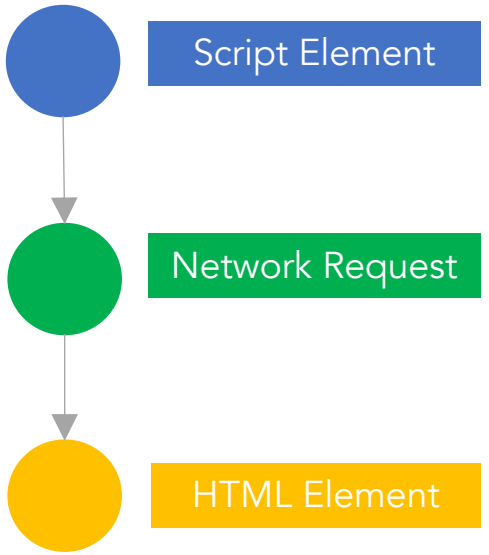
```
<script>
    eval("
      var iframe = document.createElement("iframe");
      iframe.src = "adnetwork.com/load_ads";
      document.body.append(iframe);
    ");
</script>
<iframe src = "adnetwork.com/load_ads">
     <script>
       var img = document.createElement("img");
       iframe.src = "adnetwork.com/load_ads";
     </script>
</iframe>
```



Cross-layer Context

Cross-layer interactions

```
<script>
    eval("
      var iframe = document.createElement("iframe");
      iframe.src = "adnetwork.com/load_ads";
      document.body.append(iframe);
    ");
</script>
<iframe src = "adnetwork.com/load_ads">
     <script>
       var img = document.createElement("img");
       iframe.src = "adnetwork.com/load_ads";
     </script>
</iframe>
```

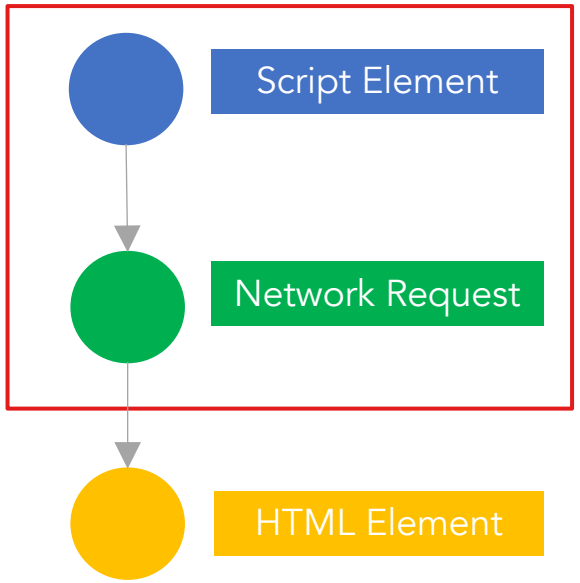


Cross-layer Context

Cross-layer interactions

JS (element) → Network (request)

```
<script>
    eval("
      var iframe = document.createElement("iframe");
      iframe.src = "adnetwork.com/load_ads";
      document.body.append(iframe);
    ");
</script>
<iframe src = "adnetwork.com/load_ads">
     <script>
       var img = document.createElement("img");
       iframe.src = "adnetwork.com/load_ads";
     </script>
</iframe>
```



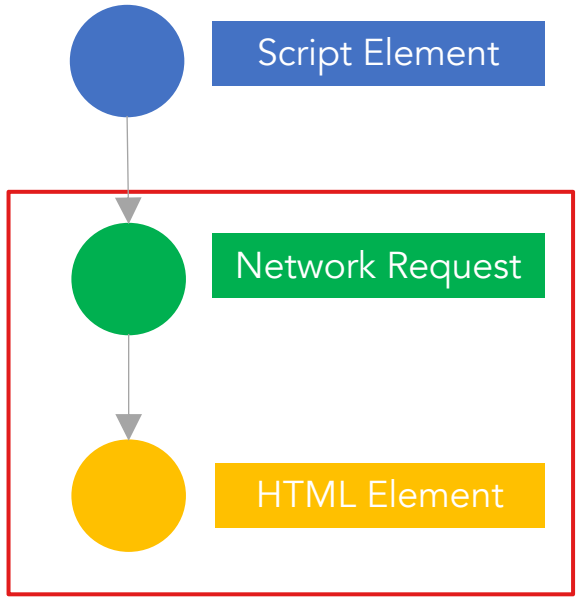
Cross-layer Context

Cross-layer interactions

JS (element) → Network (request)

Network (request) → HTML (response)

```
<script>
    eval("
      var iframe = document.createElement("iframe");
      iframe.src = "adnetwork.com/load_ads";
      document.body.append(iframe);
    ");
</script>
<iframe src = "adnetwork.com/load_ads">
     <script>
       var img = document.createElement("img");
       iframe.src = "adnetwork.com/load_ads";
     </script>
</iframe>
```



Cross-layer Context

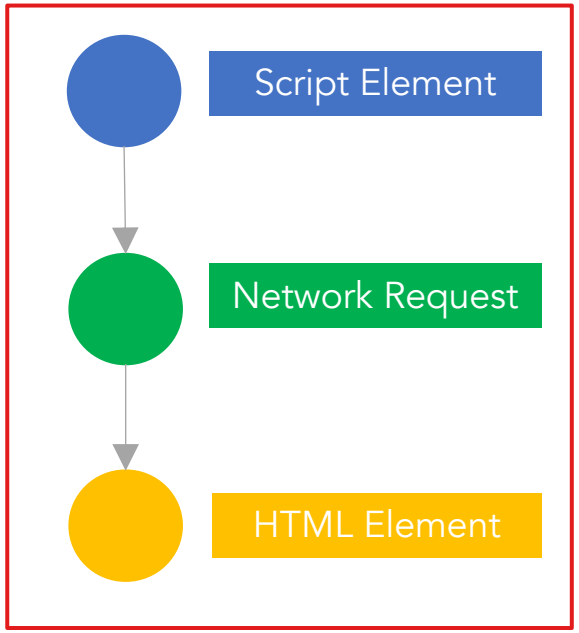
Cross-layer interactions

JS (element) → Network (request)

Network (request) → HTML (response)

Building cross-layer context

```
<script>
    eval("
      var iframe = document.createElement("iframe");
      iframe.src = "adnetwork.com/load_ads";
      document.body.append(iframe);
    ");
</script>
<iframe src = "adnetwork.com/load_ads">
    <script>
      var img = document.createElement("img");
      iframe.src = "adnetwork.com/load_ads";
    </script>
</iframe>
```



Cross-layer Context

Cross-layer interactions

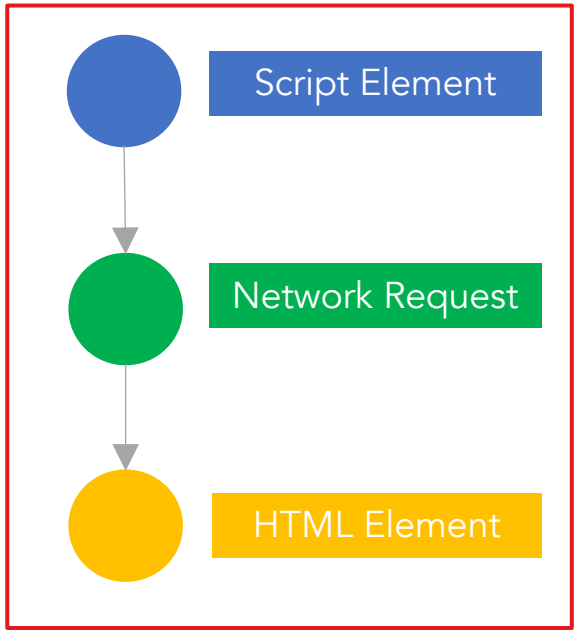
JS (element) → Network (request)

Network (request) → HTML (response)

Building cross-layer context

Easy to link Network with HTML

```
<script>
    eval("
      var iframe = document.createElement("iframe");
      iframe.src = "adnetwork.com/load_ads";
      document.body.append(iframe);
    ");
</script>
<iframe src = "adnetwork.com/load_ads">
     <script>
      var img = document.createElement("img");
      iframe.src = "adnetwork.com/load_ads";
     </script>
</iframe>
```



Cross-layer Context

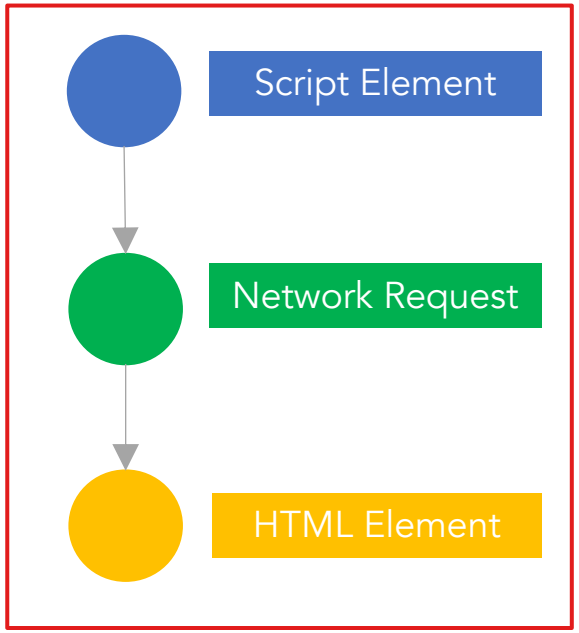
Cross-layer interactions

- JS (element) → Network (request)
- Network (request) → HTML (response)

Building cross-layer context

- Easy to link Network with HTML
- JavaScript activity attribution is tricky

```
<script>
    eval("
      var iframe = document.createElement("iframe");
      iframe.src = "adnetwork.com/load_ads";
      document.body.append(iframe);
    ");
</script>
<iframe src = "adnetwork.com/load_ads">
     <script>
       var img = document.createElement("img");
       iframe.src = "adnetwork.com/load_ads";
     </script>
</iframe>
```



JavaScript Attribution

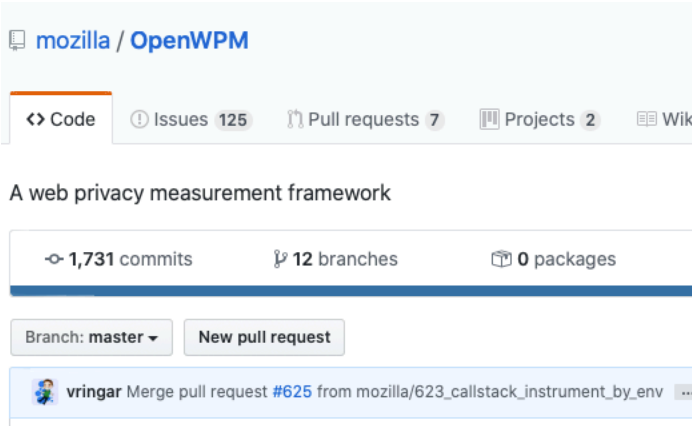
No API to attribute JavaScript to HTML and Network requests

JavaScript Attribution

No API to attribute JavaScript to HTML and Network requests

Stack Walking [Privacy Badger, OpenWPM]

- Look at stack at points of interest
- Incomplete and evadable
- e.g. eval, inline scripts

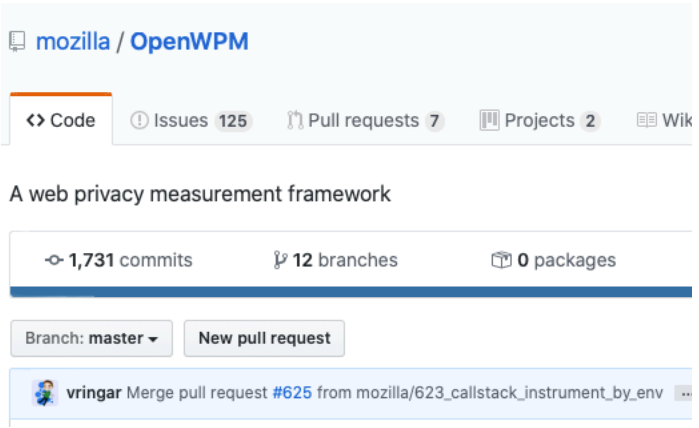


JavaScript Attribution

No API to attribute JavaScript to HTML and Network requests

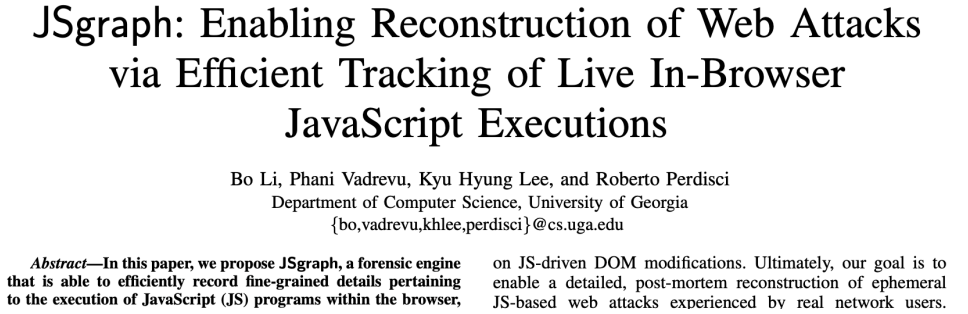
Stack Walking [Privacy Badger, OpenWPM]

- Look at stack at points of interest
- Incomplete and evadable
- e.g. eval, inline scripts



Browser Instrumentation [JSGraph '18]

- Capture events as scripts execute
- Detailed cross-layer interaction



JSgraph: Enabling Reconstruction of Web Attacks via Efficient Tracking of Live In-Browser JavaScript Executions

Bo Li, Phani Vadrevu, Kyu Hyung Lee, and Roberto Perdisci
Department of Computer Science, University of Georgia
{bo,vadrevu,khlee,perdisci}@cs.uga.edu

***Abstract*—In this paper, we propose JSgraph, a forensic engine that is able to efficiently record fine-grained details pertaining to the execution of JavaScript (JS) programs within the browser,** on JS-driven DOM modifications. Ultimately, our goal is to enable a detailed, post-mortem reconstruction of ephemeral JS-based web attacks experienced by real network users.

Chromium Instrumentation

Instrument rendering (Blink) and JavaScript (V8) engines

- Build cross-layer context as a graph
- HTML modifications, Network requests, JS attributions

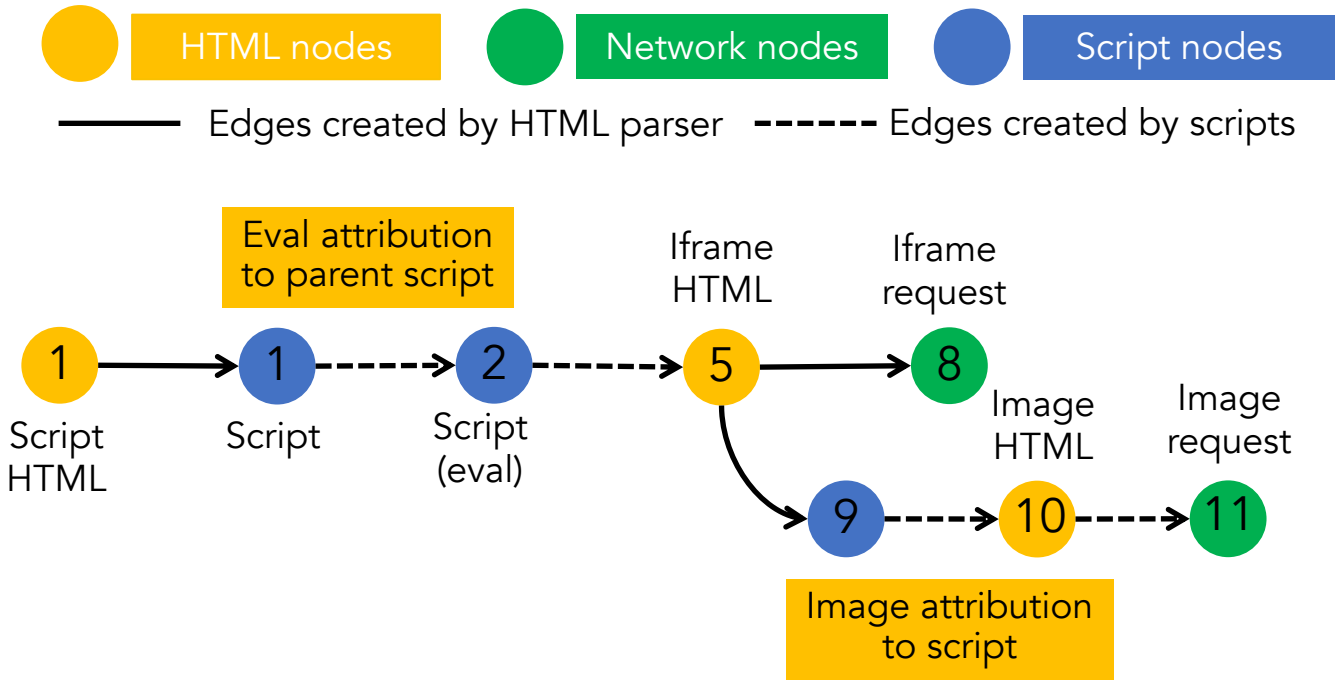
Chromium Instrumentation

Instrument rendering (Blink) and JavaScript (V8) engines

Build cross-layer context as a graph

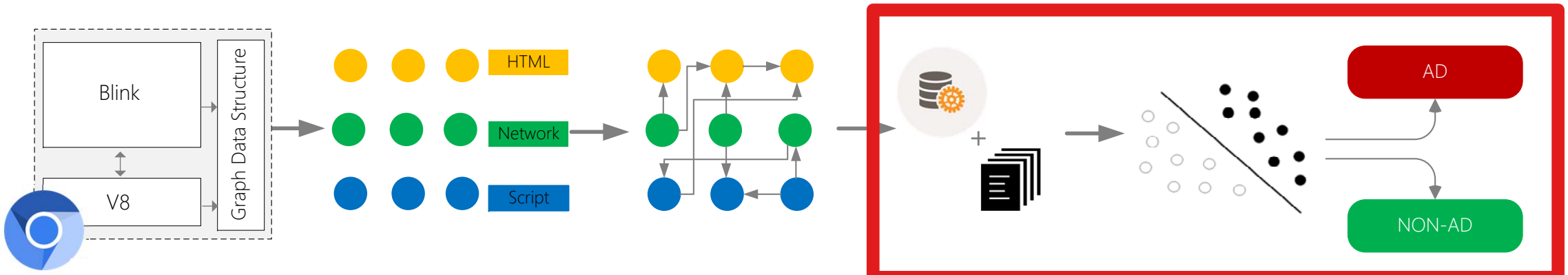
HTML modifications, Network requests, JS attributions

```
1   <script>
2       eval("
3         var iframe = document.createElement("iframe");
4         iframe.src = "adnetwork.com/load_ads";
5         document.body.append(iframe);
6       ");
7   </script>
8   <iframe src = "adnetwork.com/load_ads">
9       <script>
10        var img = document.createElement("img");
11        iframe.src = "adnetwork.com/load_ads";
12      </script>
13  </iframe>
```



AdGraph

ML to automatically learn ad/tracker behavior



Chromium instrumentation

Graph representation

Model training

Classification decision

Features Extraction

Extract two types of features

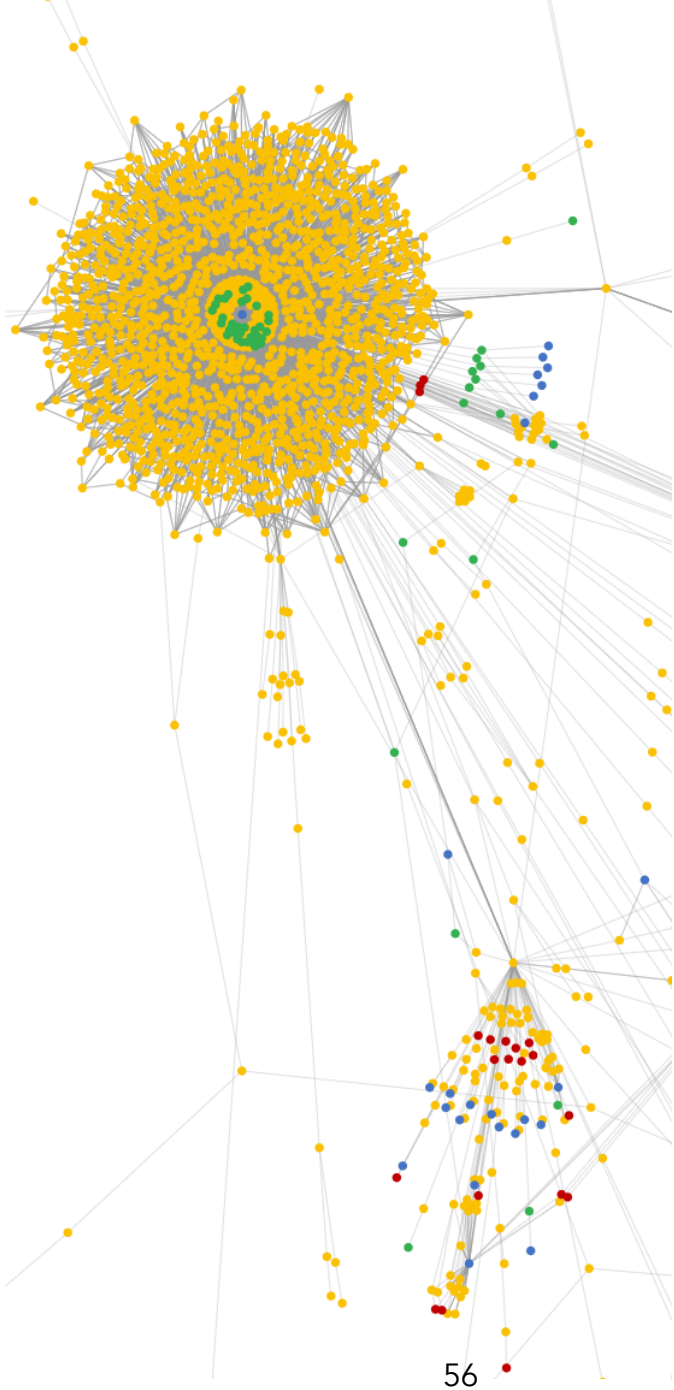
Structural & Content

Features Extraction

Extract two types of features

Structural & Content

Structural features capture graph properties



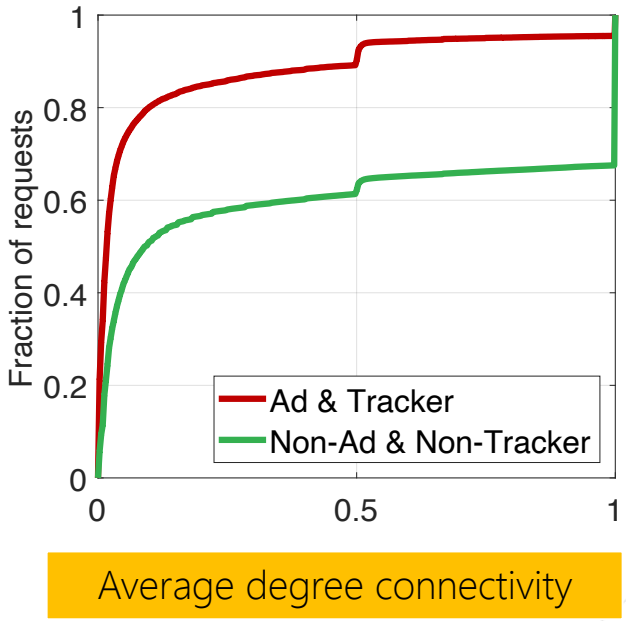
Features Extraction

Extract two types of features

Structural & Content

Structural features capture graph properties

Average degree connectivity



Features Extraction

Extract two types of features

Structural & Content

Structural features capture graph properties

Average degree connectivity

Content features capture node properties

```
https://events.bouncex.net/track.gif/bid_selected?partner=index&deployment=masthead&deal_id=106202001&price=3.50000&auction_number=1&ad_unit_id=26&source=ads&campaignid=917423&agent=user&mode=0&websiteid=340&visitid=1588398576368654&deviceid=2799665660403664656&pageviewid=1&sequenceid=17&clienttimestamp=1588398589360&clientapiversion=tag3&device=d
```

```
https://maxcdn.bootstrapcdn.com/bootstrap/3.3.7/css/bootstrap.min.css
```

Features Extraction

Extract two types of features

Structural & Content

Structural features capture graph properties

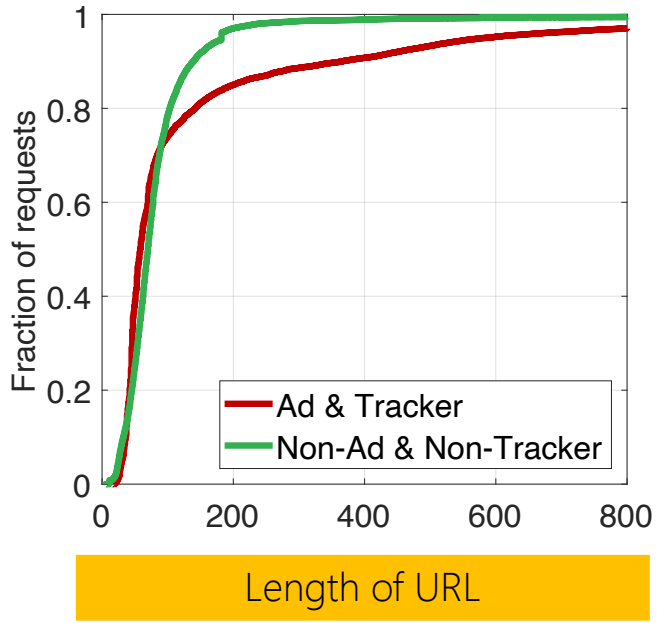
Average degree connectivity

Content features capture node properties

length of URL

```
https://events.bouncex.net/track.gif/bid_selected?partner=index&deployment=masthead&deal_id=106202001&price=3.50000&auction_number=1&ad_unit_id=26&source=ads&campaignid=917423&agent=user&mode=0&websiteid=340&visitid=1588398576368654&deviceid=2799665660403664656&pageviewid=1&sequenceid=17&clienttimestamp=1588398589360&clientapiversion=tag3&device=d
```

```
https://maxcdn.bootstrapcdn.com/bootstrap/3.3.7/css/bootstrap.min.css
```



Machine Learning

Ground truth

- Filter lists – despite shortcomings [Iqbal et al. ‘17, Snyder et al. ‘20]
- Manual evaluation of disagreements with classifier

Machine Learning

Ground truth

- Filter lists – despite shortcomings [Iqbal et al. ‘17, Snyder et al. ‘20]
- Manual evaluation of disagreements with classifier

Random forest classifier

- 10-fold cross validation

Evaluation: Accuracy

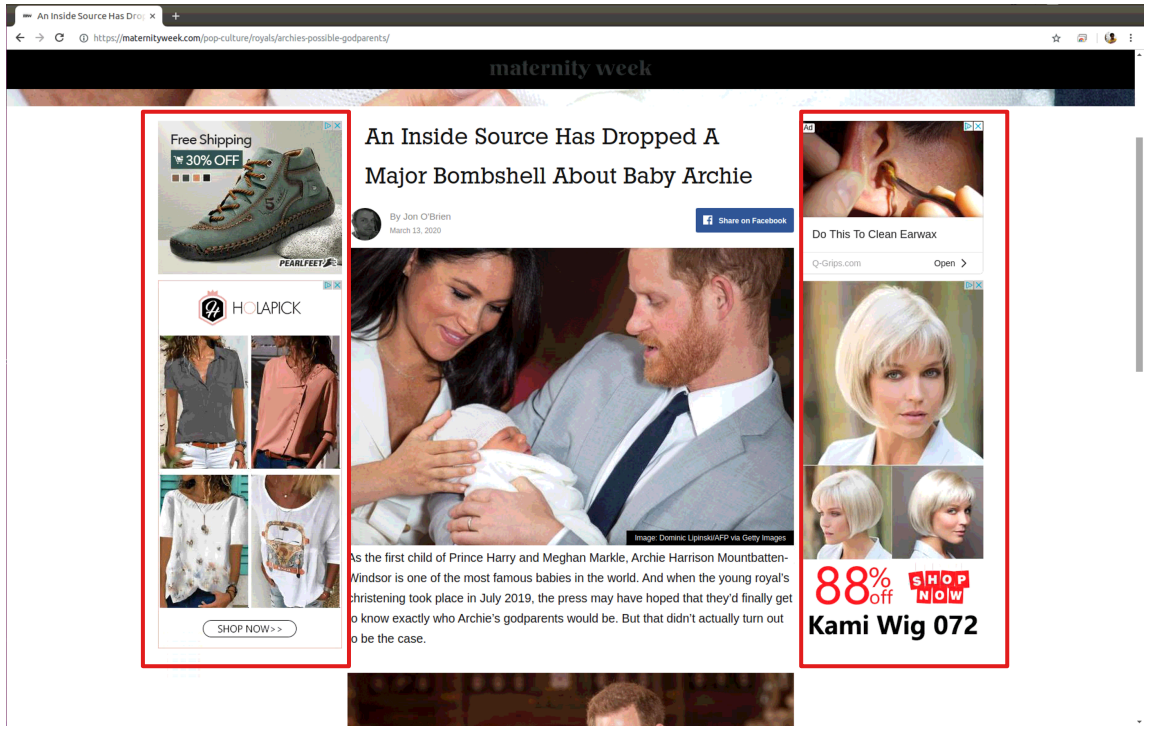
Accuracy is more than 95.33%

Recall 86.6% – Precision 89.1%

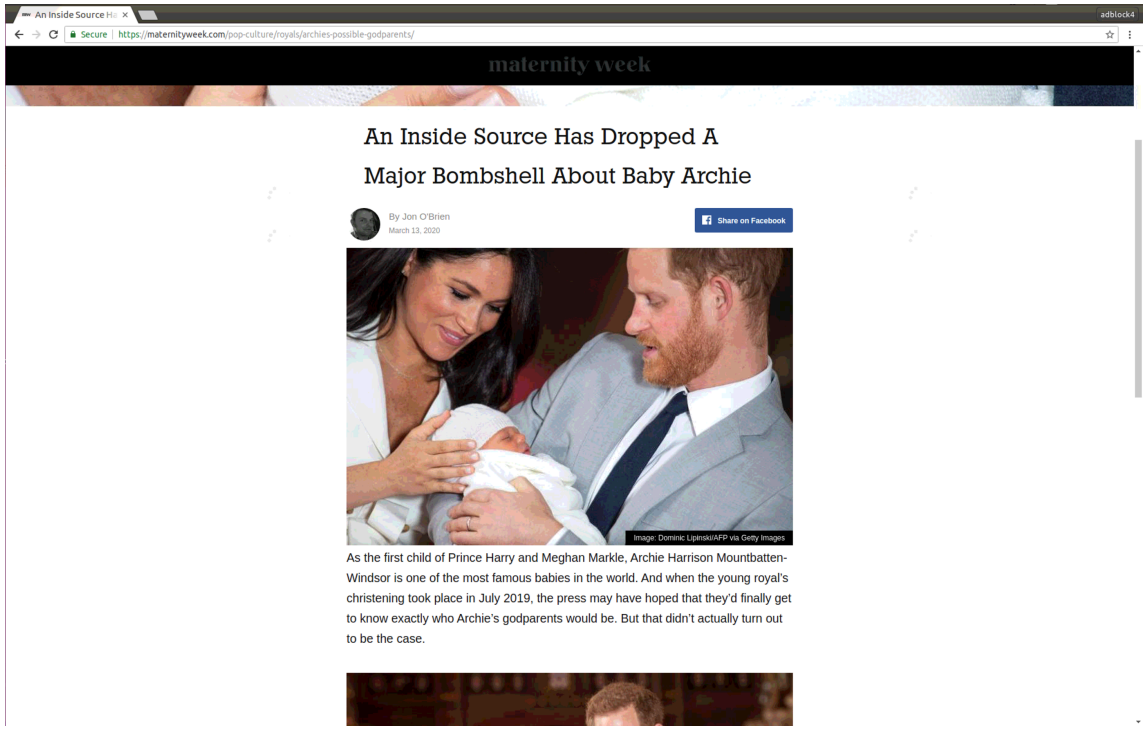
Evaluation: Accuracy

Accuracy is more than 95.33%

Recall 86.6% – Precision 89.1%



Stock Chromium



AdGraph

Evaluation: Accuracy

Accuracy is more than 95.33%

Recall 86.6% – Precision 89.1%

Disagreement analysis with filter lists

Evaluation: Accuracy

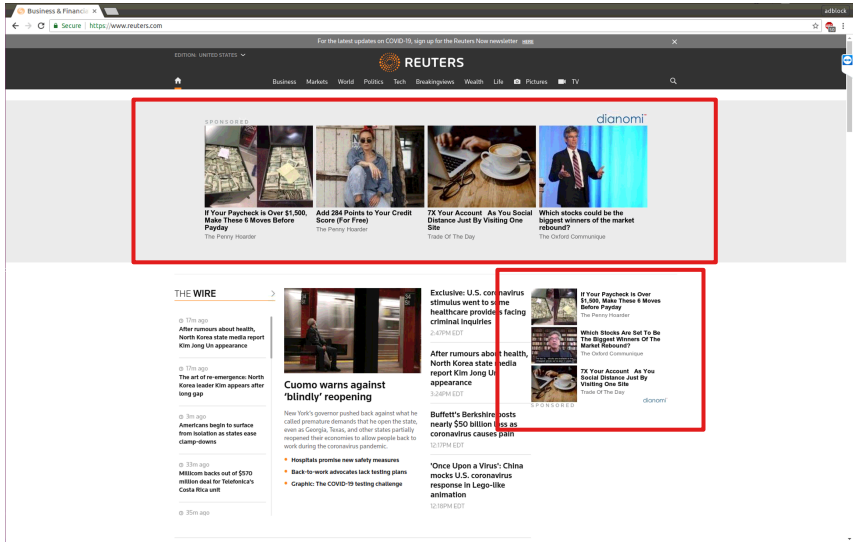
Accuracy is more than 95.33%

Recall 86.6% – Precision 89.1%

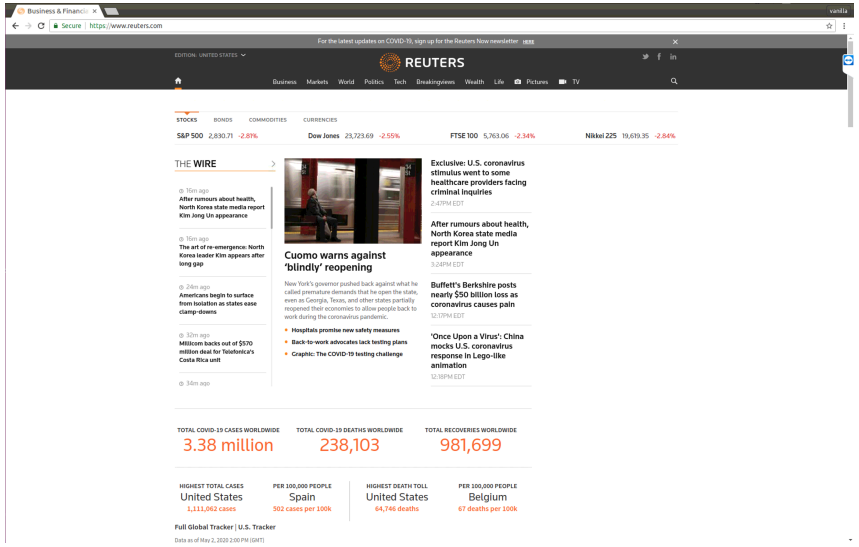
Disagreement analysis with filter lists

Filter lists under block due to unknown Ad/Trackers

AdGraph detects 43.1% new ad/tackers



Filter Lists



AdGraph

Evaluation: Accuracy

Accuracy is more than 95.33%

Recall 86.6% – Precision 89.1%

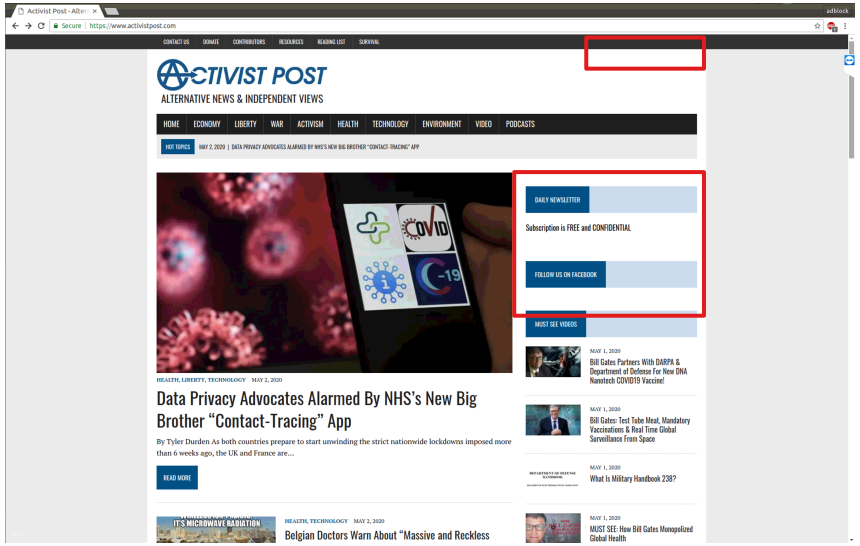
Disagreement analysis with filter lists

Filter lists under block due to unknown Ad/Trackers

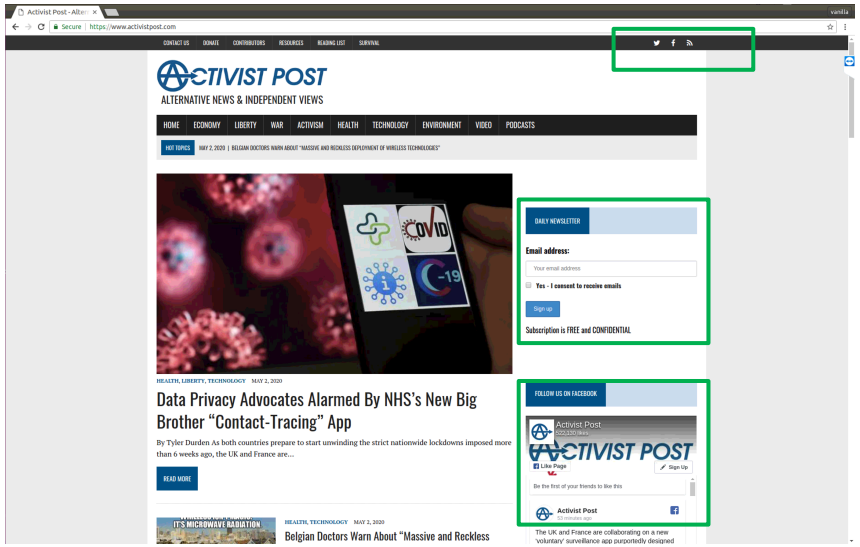
AdGraph detects 43.1% new ad/tackers

Filter lists over block due to generic rules

AdGraph identifies 28.7% over blocked functional content



Filter Lists



AdGraph

Evaluation: Accuracy

Accuracy is more than 95.33%

Recall 86.6% – Precision 89.1%

D

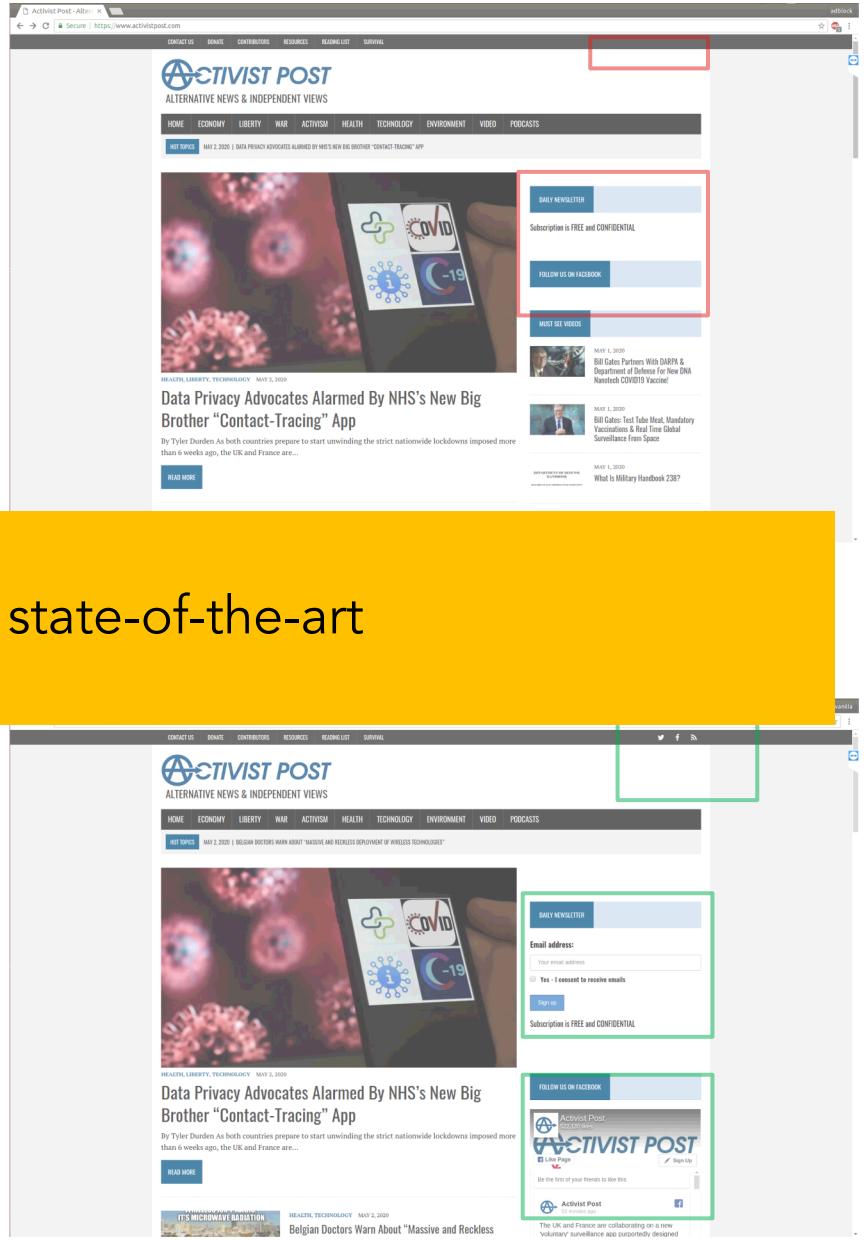
AdGraph outperforms the current state-of-the-art

Filter lists under block due to unknown Ad/Trackers

AdGraph detects 43.1% new ad/tackers

Filter lists over block due to generic rules

AdGraph identifies 28.7% over blocked functional content



AdGraph

Evaluation: Performance

Real time ad and tracker blocking with ML

- Instrumentation overhead
- Classification overhead

Page load time comparison (Stock Chromium and AdBlock Plus)

Makes up by request blocking & less rendering

Evaluation: Performance

Real time ad and tracker blocking with ML

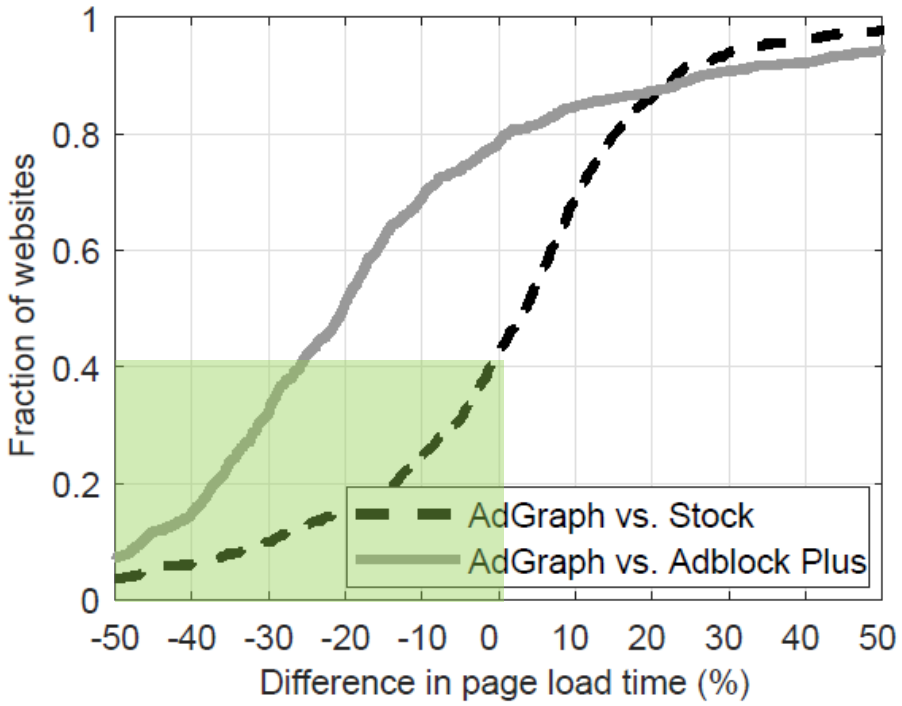
- Instrumentation overhead
- Classification overhead

Page load time comparison (Stock Chromium and AdBlock Plus)

Makes up by request blocking & less rendering

Faster than Chromium on 42% websites

- Faster when blocks more



Evaluation: Performance

Real time ad and tracker blocking with ML

- Instrumentation overhead
- Classification overhead

Page load time comparison (Stock Chromium and AdBlock Plus)

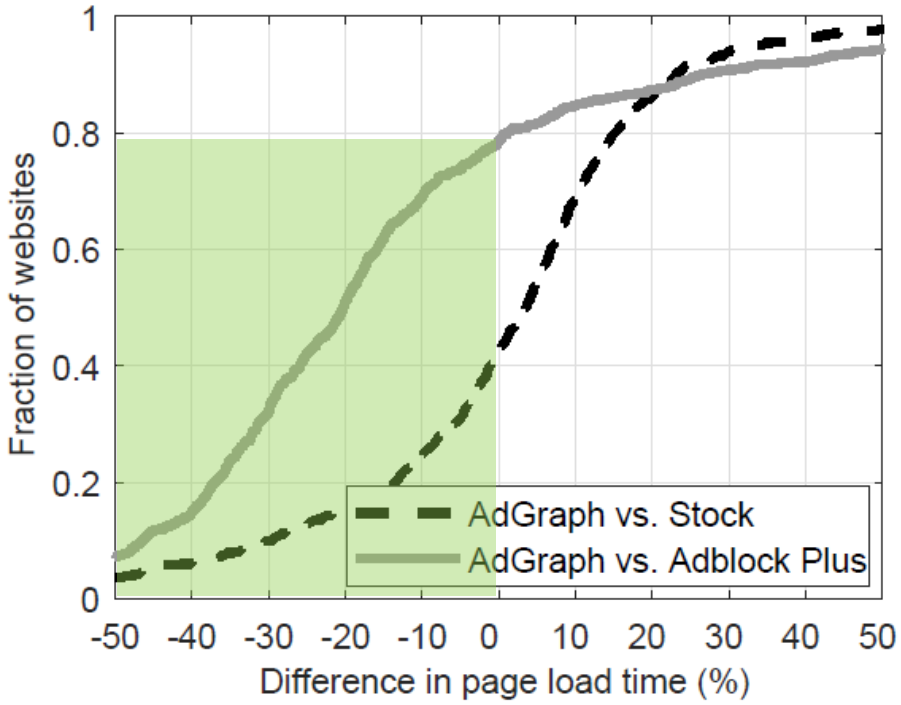
Makes up by request blocking & less rendering

Faster than Chromium on 42% websites

- Faster when blocks more

Faster than Adblock Plus on 78% websites

- Avoids rendering overhead



Evaluation: Performance

Real time ad and tracker blocking with ML

- Instrumentation overhead
- Classification overhead

Page load time comparison (Stock Chromium and AdBlock Plus)

Makes up by request blocking & less rendering

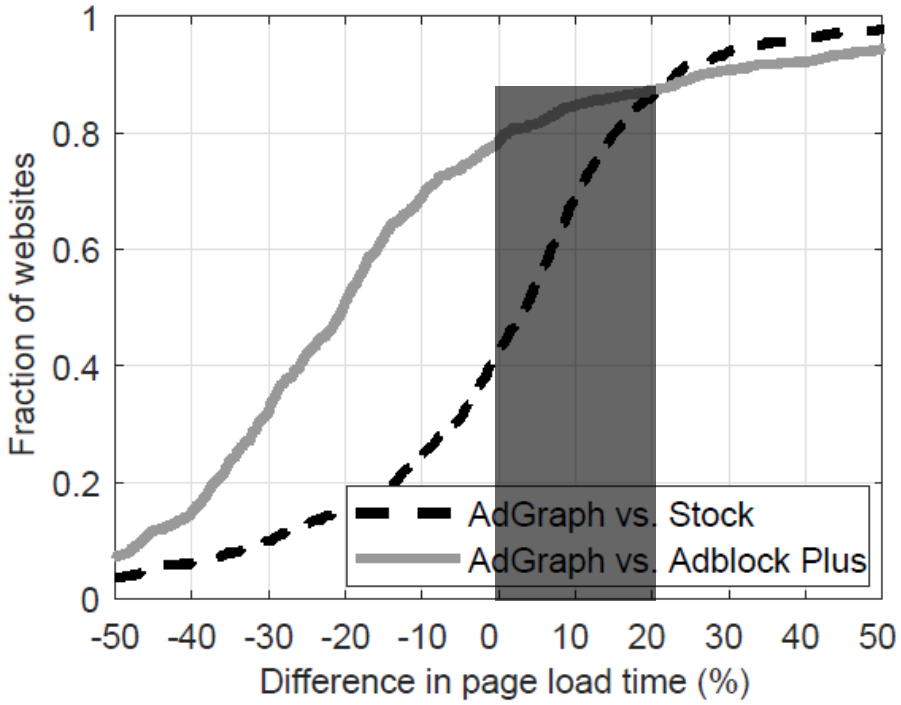
Faster than Chromium on 42% websites

- Faster when blocks more

Faster than Adblock Plus on 78% websites

- Avoids rendering overhead

Minor overhead on most websites



Evaluation: Performance

Real time ad and tracker blocking with ML

Instrumentation overhead

Classification overhead

Pa

Makes up by request blocking & less rendering

Faster than Chromium on 42% websites

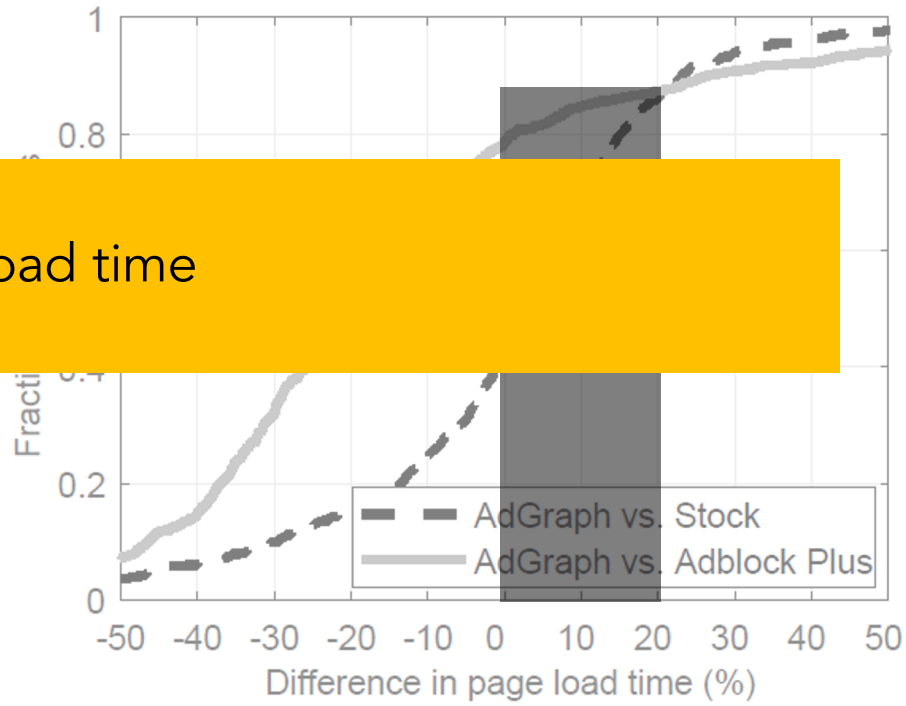
Faster when blocks more

Faster than Adblock Plus on 78% websites

Avoids rendering overhead

Minor overhead on most websites

AdGraph improves page load time



Key Takeaways

Use cross-layer context to address **accuracy** issues

Use machine learning address **scalability** issues

Key Takeaways

Use cross-layer context to address **accuracy** issues

Use machine learning address **scalability** issues

Open source implementation

https://uiowa-irl.github.io/AdGraph/

Key Takeaways

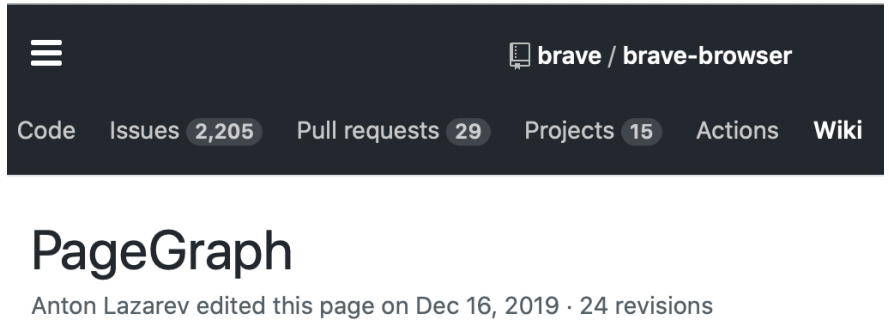
Use cross-layer context to address **accuracy** issues

Use machine learning address **scalability** issues

Open source implementation

Maintained by Brave as PageGraph

https://uiowa-irl.github.io/AdGraph/



Key Takeaways

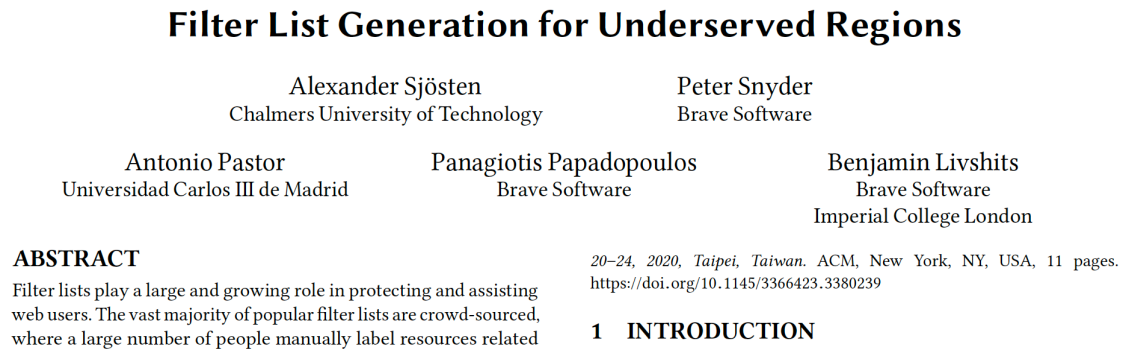
Use cross-layer context to address **accuracy** issues

Use machine learning address **scalability** issues

Open source implementation

Maintained by Brave as PageGraph

Filter list generation



Filter List Generation for Underserved Regions

Alexander Sjösten
Chalmers University of Technology

Peter Snyder
Brave Software

Antonio Pastor
Universidad Carlos III de Madrid

Panagiotis Papadopoulos
Brave Software

Benjamin Livshits
Brave Software
Imperial College London

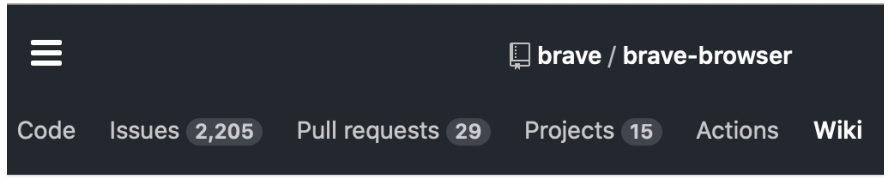
ABSTRACT

Filter lists play a large and growing role in protecting and assisting web users. The vast majority of popular filter lists are crowd-sourced, where a large number of people manually label resources related

20–24, 2020, Taipei, Taiwan. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3366423.3380239

1 INTRODUCTION

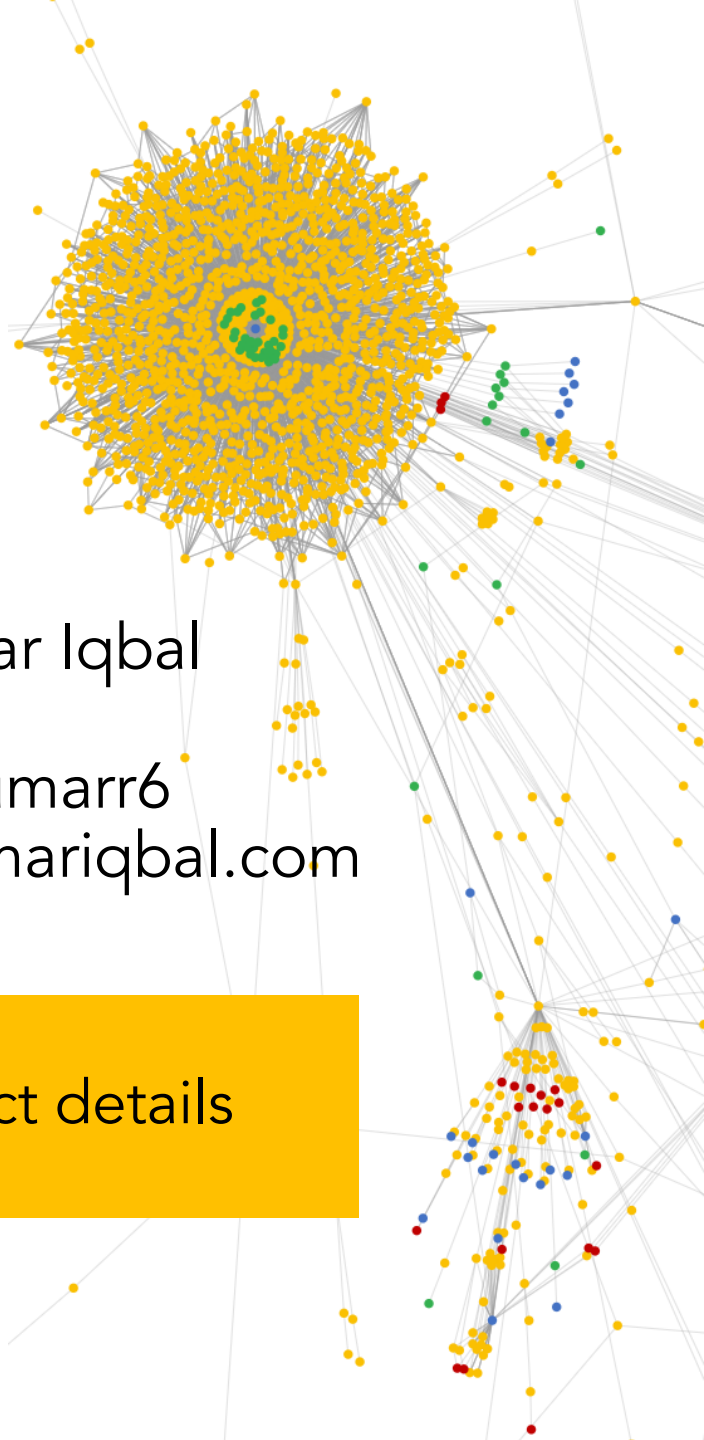
https://uiowa-irl.github.io/AdGraph/



PageGraph

Anton Lazarev edited this page on Dec 16, 2019 · 24 revisions

Questions?



Paper link:
https://www.umariqbal.com/papers/adgraph-sp2020.pdf

Source code:
https://uiowa-irl.github.io/AdGraph/

Contact details

Umar Iqbal

@umarr6
www.umariqbal.com

References

1. Advertising revenue – https://www.iab.com/wp-content/uploads/2019/05/Full-Year-2018-IAB-Internet-Advertising-Revenue-Report.pdf
2. Malvertising – https://www.zdnet.com/article/hackers-have-breached-60-ad-servers-to-load-their-own-malicious-ads/
3. Malvertising – https://www.theguardian.com/technology/2016/mar/16/major-sites-new-york-times-bbc-ransomware-malvertising
4. Slow page load – https://www.nytimes.com/interactive/2015/10/01/business/cost-of-mobile-ads.html
5. OpenWPM – https://github.com/mozilla/OpenWPM
6. Privacy Badger – https://github.com/EFForg/privacybadger
7. Iqbal, Umar et al. "The ad wars: retrospective measurement and analysis of anti-adblock filter lists." Proceedings of the 2017 Internet Measurement Conference. 2017.
8. Snyder, Peter et al. "Who filters the filters: Understanding the growth, usefulness and efficiency of crowdsourced ad blocking", SIGMETRICS. 2020.
9. Bhagavatula, Sruti, et al. "Leveraging machine learning to improve unwanted resource filtering." Proceedings of the 2014 Workshop on Artificial Intelligent and Security Workshop. 2014.
10. Gugelmann, David, et al. "An automated approach for complementing ad blockers' blacklists." Proceedings on Privacy Enhancing Technologies 2015.
11. Ikram, Muhammad, et al. "Towards seamless tracking-free web: Improved detection of trackers via one-class learning." Proceedings on Privacy Enhancing Technologies 2017.
12. Wu, Qianru, et al. "A machine learning approach for detecting third-party trackers on the web." European Symposium on Research in Computer Security. Springer, Cham, 2016.
13. Li, Bo, et al. "JSgraph: Enabling Reconstruction of Web Attacks via Efficient Tracking of Live In-Browser JavaScript Executions." NDSS. 2018.
14. Icon made by Pixel perfect from www.flaticon.com