



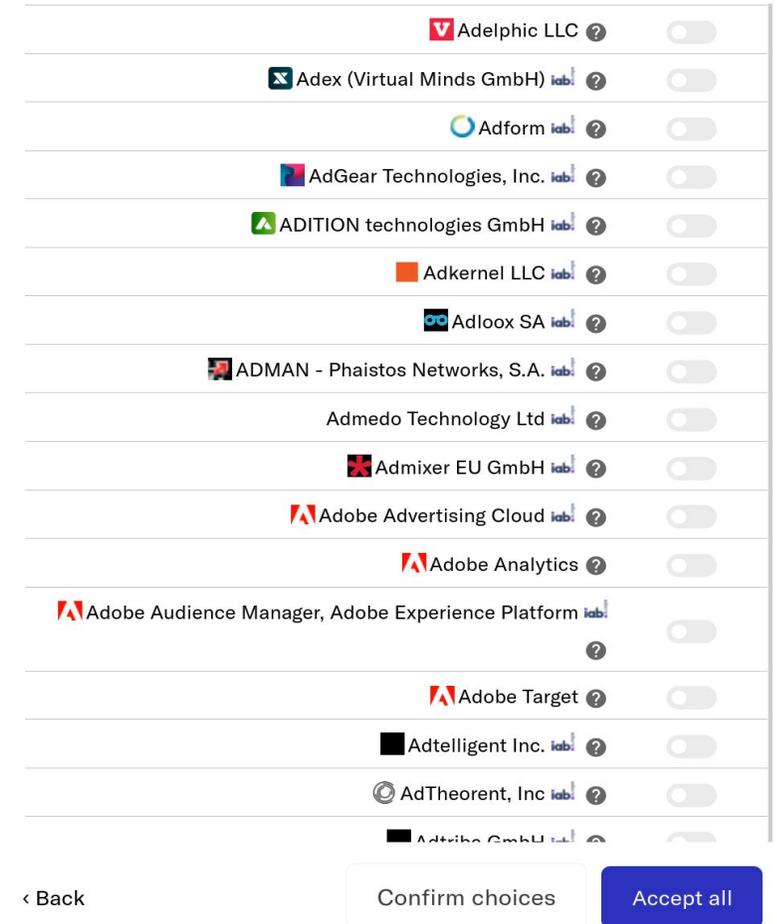
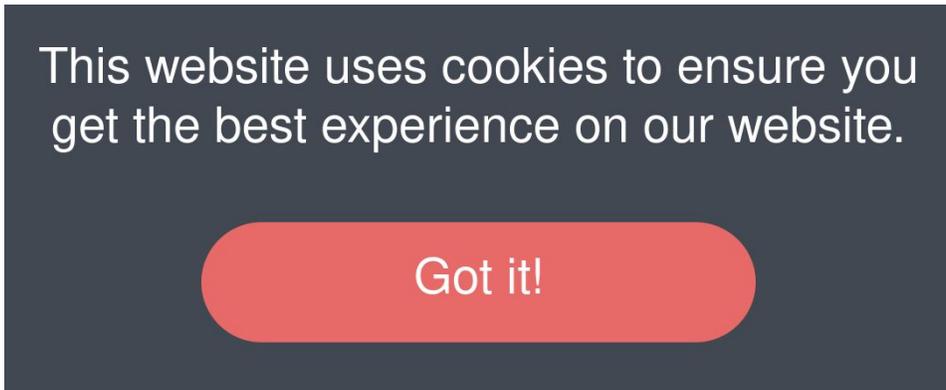
# Automating Cookie Consent and GDPR Violation Detection

Dino Bollinger, **Karel Kubicek**, Carlos Cotrini, David Basin  
CNIL Research Day (June 14, 2023); USENIX Sec.'22 paper



# Cookie consent

- Solomos et al. (2019): 90% of websites use tracking cookies
- EU law: Websites must notify users, gather consent
- Consent notices to comply with regulations



 Adelphic LLC 	<input type="checkbox"/>
 Adex (Virtual Minds GmbH) 	<input type="checkbox"/>
 Adform 	<input type="checkbox"/>
 AdGear Technologies, Inc. 	<input type="checkbox"/>
 ADITION technologies GmbH 	<input type="checkbox"/>
 Adkernel LLC 	<input type="checkbox"/>
 Adloox SA 	<input type="checkbox"/>
 ADMAN - Phaistos Networks, S.A. 	<input type="checkbox"/>
Admedo Technology Ltd 	<input type="checkbox"/>
 Admixer EU GmbH 	<input type="checkbox"/>
 Adobe Advertising Cloud 	<input type="checkbox"/>
 Adobe Analytics 	<input type="checkbox"/>
 Adobe Audience Manager, Adobe Experience Platform 	<input type="checkbox"/>
 Adobe Target 	<input type="checkbox"/>
 Adtelligent Inc. 	<input type="checkbox"/>
 AdTheorent, Inc 	<input type="checkbox"/>
 Adtribe GmbH 	<input type="checkbox"/>

< Back      Confirm choices      Accept all

# ePrivacy Directive and General Data Protection Regulation (*GDPR*)

## ePrivacy Directive:

- All but strictly necessary data processing requires consent

### This website uses cookies

We use cookies to personalise service and to analyse our traffic. You consent to our cookies if you continue to use our website.

Necessary  Preferences  Statistics  Marketing

OK

Cookie declaration		About			
	Name	Provider	Purpose	Expiry	Type
Necessary (22)	_ga	<a href="#">Google</a>	Unique ID to generate statistics about the visitors.	2 years	HTTP
Functionality (2)					
Statistics (12)	_gat	<a href="#">Google</a>	Used by Google Analytics to throttle request rate	1 day	HTTP
Advertising (60)					
Unclassified (43)	__qca	<a href="#">Quantcast</a>	Collects anonymous data on the user's visits to the website, such as	1 year	HTTP

# ePrivacy Directive and General Data Protection Regulation (*GDPR*)

## ePrivacy Directive:

- All but strictly necessary data processing requires consent

### This website uses cookies

We use cookies to personalise service and to analyse our traffic. You consent to our cookies if you continue to use our website.

Necessary  Preferences  Statistics  Marketing OK

Cookie declaration		About			
	Name	Provider	Purpose	Expiry	Type
Necessary (22)	_ga	<a href="#">Google</a>	Unique ID to generate statistics about the visitors.	2 years	HTTP
Functionality (2)					
Statistics (12)	_gat	<a href="#">Google</a>	Used by Google Analytics to throttle request rate	1 day	HTTP
Advertising (60)					
Unclassified (43)	__qca	<a href="#">Quantcast</a>	Collects anonymous data on the user's visits to the website, such as	1 year	HTTP

# ePrivacy Directive and General Data Protection Regulation (*GDPR*)

## ePrivacy Directive:

- All but strictly necessary data processing requires consent

## GDPR Consent:

- Freely-given
- Unambiguous
- Specific
- Informed
- Purpose-limited

### This website uses cookies

We use cookies to personalise service and to analyse our traffic. You consent to our cookies if you continue to use our website.

Necessary  Preferences  Statistics  Marketing OK

Cookie declaration	Name	Provider	Purpose	Expiry	Type
Necessary (22)	_ga	<a href="#">Google</a>	Unique ID to generate statistics about the visitors.	2 years	HTTP
Functionality (2)					
Statistics (12)	_gat	<a href="#">Google</a>	Used by Google Analytics to throttle request rate	1 day	HTTP
Advertising (60)					
Unclassified (43)	__qca	<a href="#">Quantcast</a>	Collects anonymous data on the user's visits to the website, such as	1 year	HTTP

# ePrivacy Directive and General Data Protection Regulation (*GDPR*)

## ePrivacy Directive:

- All but strictly necessary data processing requires consent

## GDPR Consent:

- Freely-given
- Unambiguous
- Specific
- Informed
- Purpose-limited

### This website uses cookies

We use cookies to personalise service and to analyse our traffic. You consent to our cookies if you continue to use our website.

Necessary  Preferences  Statistics  Marketing

OK

Cookie declaration		About			
	Name	Provider	Purpose	Expiry	Type
Necessary (22)	_ga	<a href="#">Google</a>	Unique ID to generate statistics about the visitors.	2 years	HTTP
Functionality (2)					
Statistics (12)	_gat	<a href="#">Google</a>	Used by Google Analytics to throttle request rate	1 day	HTTP
Advertising (60)					
Unclassified (43)	__qca	<a href="#">Quantcast</a>	Collects anonymous data on the user's visits to the website, such as	1 year	HTTP

# ePrivacy Directive and General Data Protection Regulation (*GDPR*)

## ePrivacy Directive:

- All but strictly necessary data processing requires consent

## GDPR Consent:

- Freely-given
- Unambiguous
- Specific
- Informed
- Purpose-limited

### This website uses cookies

We use cookies to personalise service and to analyse our traffic. You consent to our cookies if you continue to use our website.

Necessary  Preferences  Statistics  Marketing

OK

Cookie declaration

About

Necessary (22)

Functionality (2)

Statistics (12)

Advertising (60)

Unclassified (43)

Name	Provider	Purpose	Expiry	Type
__ga	<a href="#">Google</a>	Unique ID to generate statistics about the visitors.	2 years	HTTP
__gat	<a href="#">Google</a>	Used by Google Analytics to throttle request rate	1 day	HTTP
__qca	<a href="#">Quantcast</a>	Collects anonymous data on the user's visits to the website, such as	1 year	HTTP

# Non-compliance is widespread

- Empirical studies:
  - Non-compliance in up to 80% of websites (e.g., cookies set before consent)  
(Utz 2019, Trevisan 2019, Matte 2020, Nouwens 2020, Kampanos 2021, Santos 2021, etc.)
  - Websites do not respect user choices  
(Libert 2018, Trevisan 2019, Matte 2020, Nouwens 2020, etc.)
- Usability: dark patterns successfully trick users  
(Bösch 2016, Grassl 2020, Hasner 2021, Sanchez-Rola 2019, Htut Soe 2020, etc.)

**Goal: Enforce cookie consent on client-side.**

# Our solution: CookieBlock

- Browser extension to predict purposes for cookies using machine learning
- Remove cookies of unconsented purposes on client-side

## Implementation:



1. Crawl web to gather training data (ground truth)
2. Extract features and train an ML model
3. Apply the model in the browser extension

# Data collection: selecting data sources

- Source for training data:

Cookie declaration		About			
	Name	Provider	Purpose	Expiry	Type
Necessary (22)	_ga	<u>Google</u>	Unique ID to generate statistics about the visitors.	2 years	HTTP
Functionality (2)					
Statistics (12)	_gat	<u>Google</u>	Used by Google Analytics to throttle request rate	1 day	HTTP
Advertising (60)					
Unclassified (43)	__qca	<u>Quantcast</u>	Collects anonymous data on the user's visits to the website, such as	1 year	HTTP

- Consent Management Platform (CMP):



# Data collection: web crawlers

- CMP presence crawler:

- Input: 6 million domains, sourced from Tranco list
- Output: ~37.5k domains with confirmed CMP

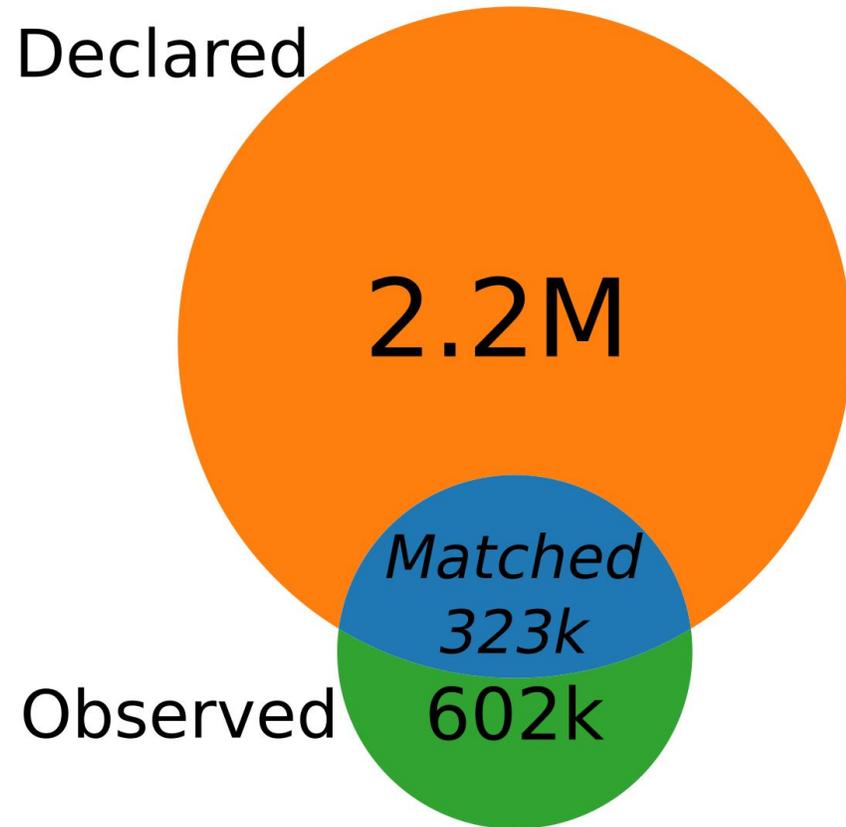
HTTP GET

- Cookie consent crawler:

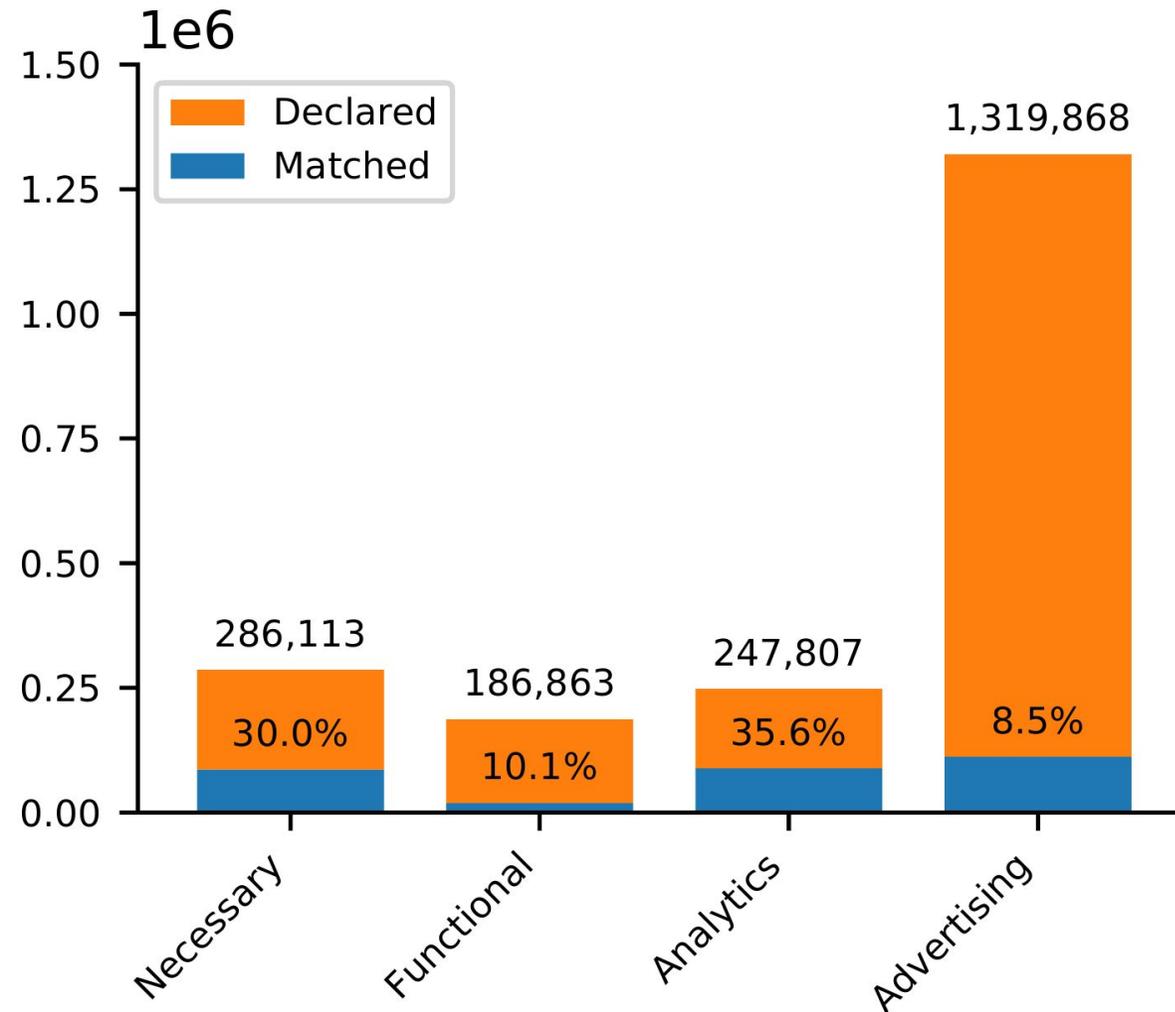
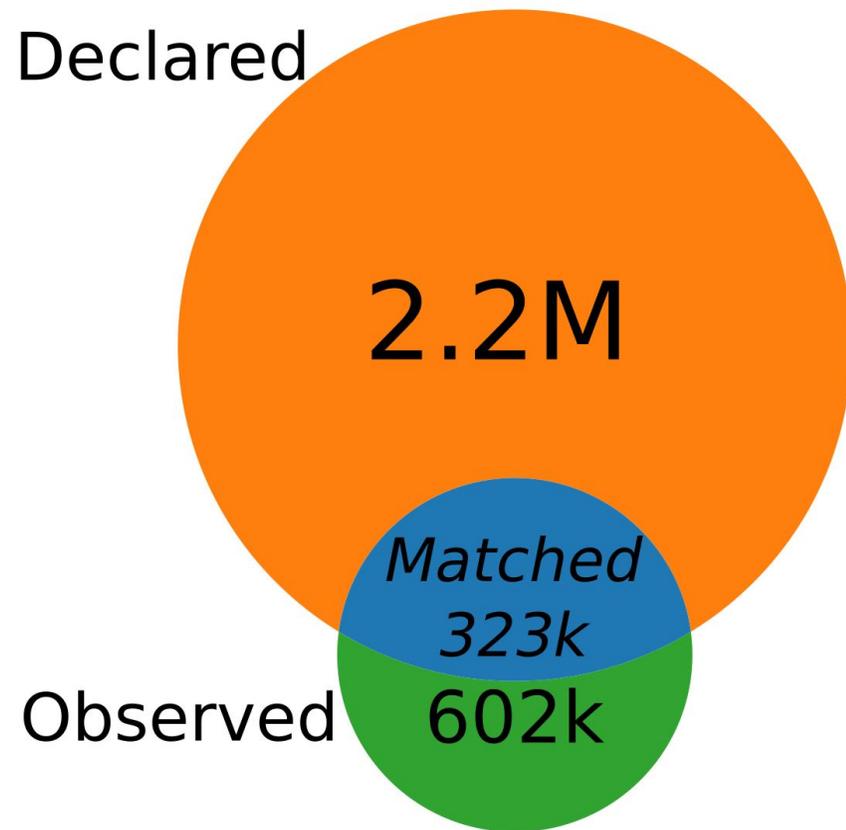
- Browse websites, gather cookie declarations + observed cookies
- Based on OpenWPM – visit subpages, move cursor, etc.
- Successful for ~30k domains



# Data collection: results of OpenWPM crawl



# Data collection: results of OpenWPM crawl



# Feature extraction from textual cookies

## Example: Shannon entropy

- Higher entropy, more randomness
- Indicator for unique identifiers

```
{  
  "id": "pid=4ecf225b293bded2e68c1d8e4379ea0c",  
  "session_count": 1,  
  "last_session_ts": 1652872456152  
}
```

## 52 types in total, including:

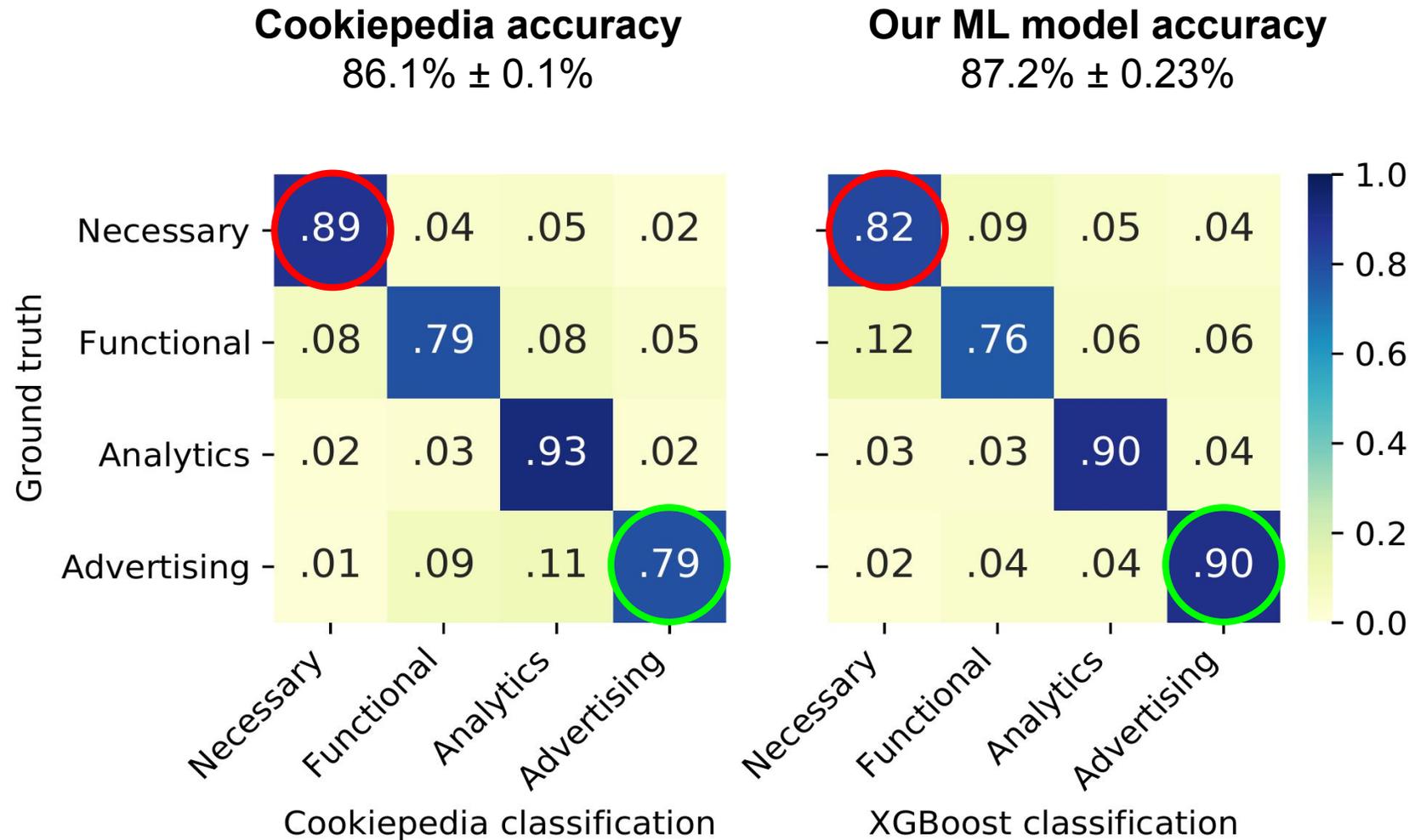
- *Encodings, name patterns, content size, timestamps, language strings, cookie flags, third-party status, expiry, etc.*

# Classifier evaluation

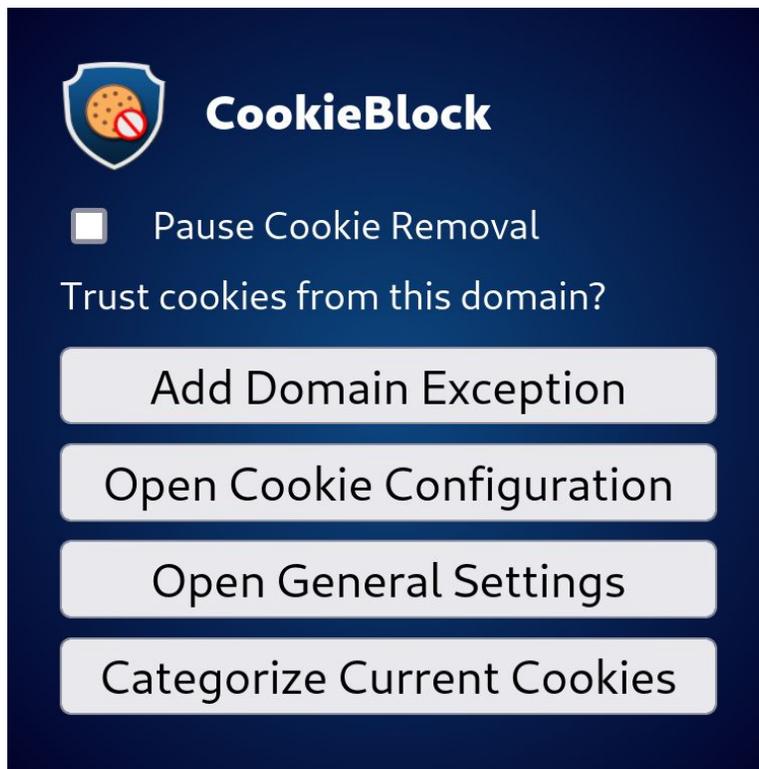
**Cookiepedia accuracy**  
86.1%  $\pm$  0.1%

**Our ML model accuracy**  
87.2%  $\pm$  0.23%

# Classifier evaluation



# CookieBlock browser extension

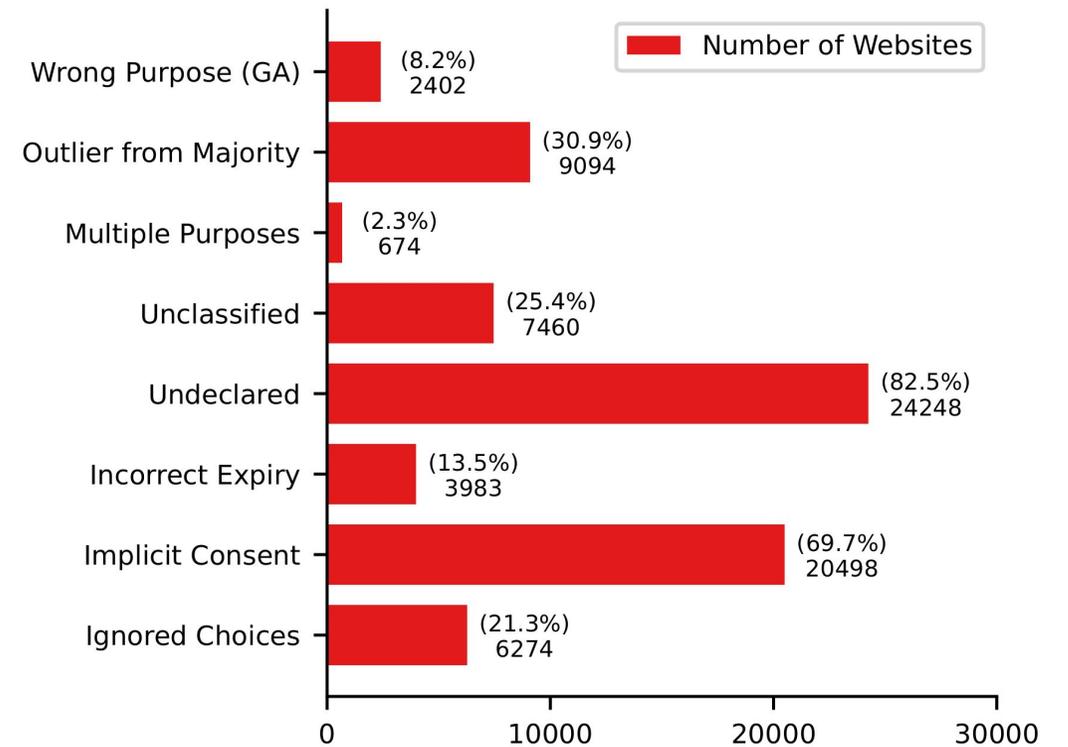


- User defines consent preferences when installed
- Classifies cookies and deletes those with rejected purpose
- Available for Firefox, Chrome, Edge, and Opera, 14k users, 4 ★

## Empirical evaluation:

- No broken functionality on 85 out of 100 pages
- 8x minor issues (cookie notice reappear)
- 7x authentication issues (login, registration)

# Potential violations



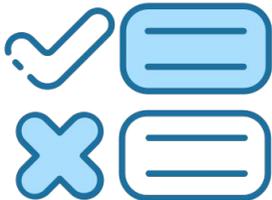
# Potential violations



## Undeclared cookies

- Not listed in the consent

GDPR informed consent requirement

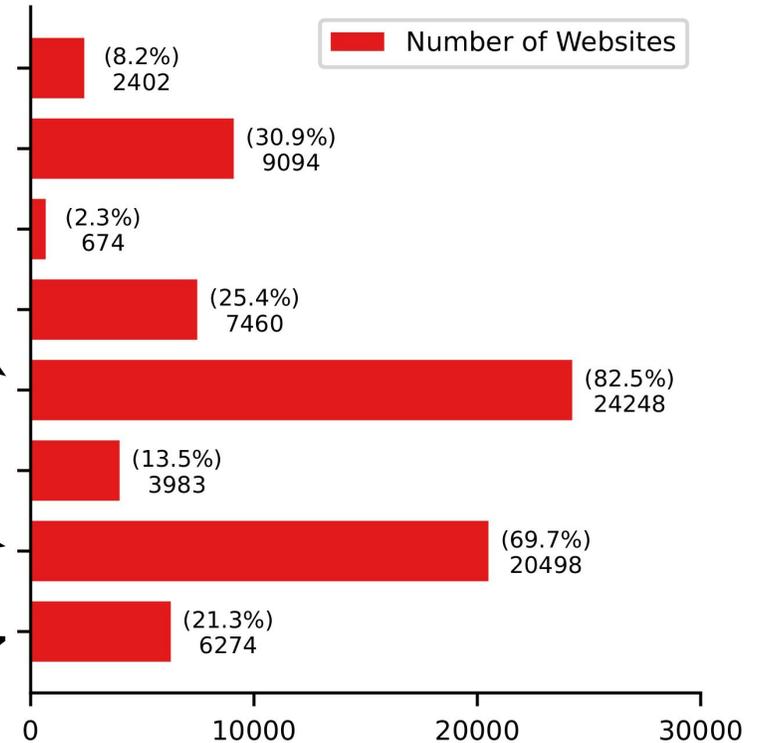


## Cookies set prior to user's consent

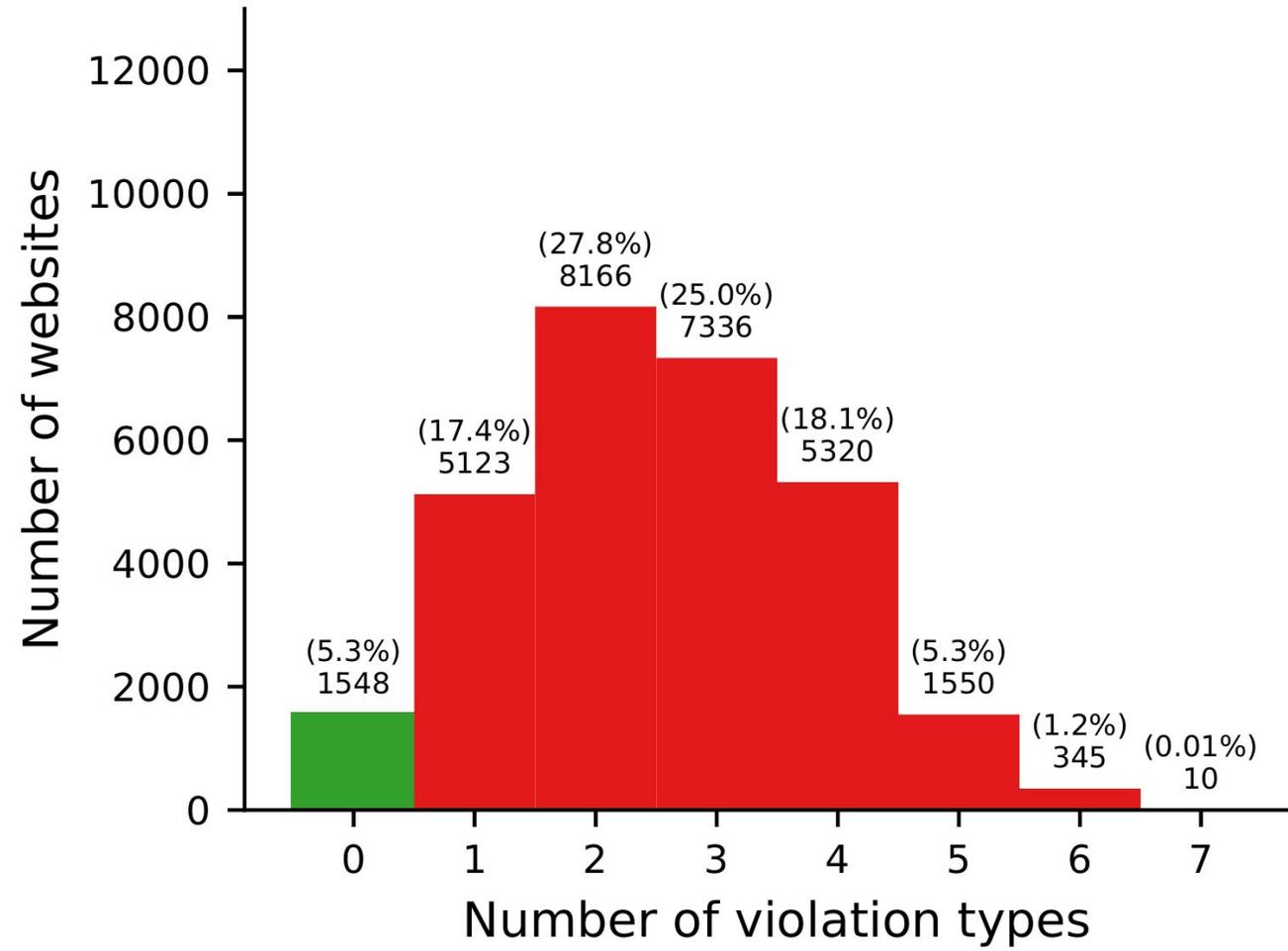


## Cookies set despite negative consent

Article 5(3) of the ePrivacy Directive



# Potential violations: histogram



# Conclusion

- Users:
  - Consent notices are broken → CookieBlock brings privacy control
- Researchers, DPAs:
  - Detected 8 potential violation types on ~95% of websites
  - Cookie purpose ML model improves tracking detection
- Developers, DPAs:
  - CookieAudit extension: audit GDPR violations in cookie notices

Authors: Dino Bollinger, Karel Kubicek, Carlos Cotrini, David Basin

More info, source, extension links:  
<https://karelkubicek.github.io/post/cookieblock>

