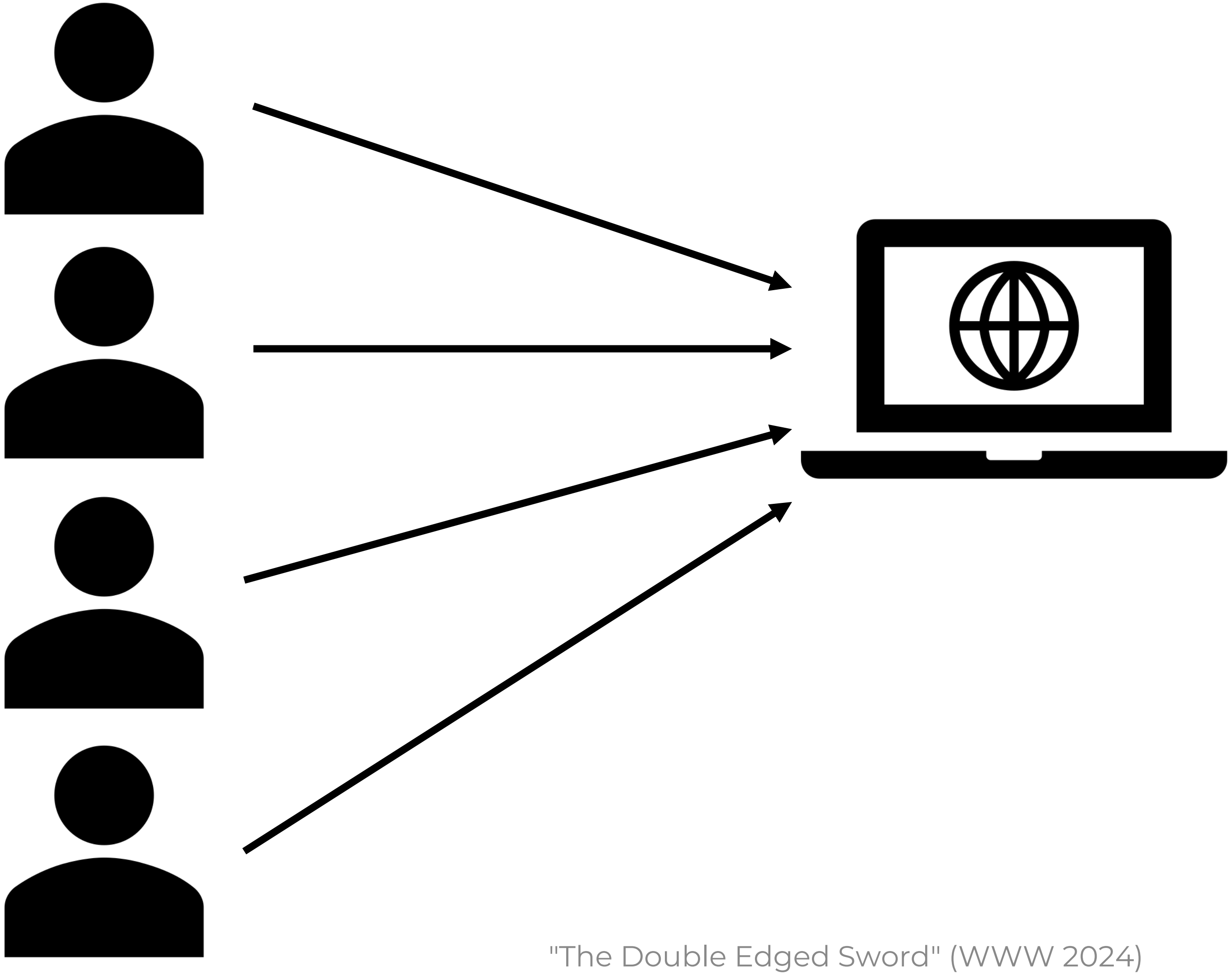
The Double Edged Sword:
Identifying Authentication Pages and their Fingerprinting Behavior

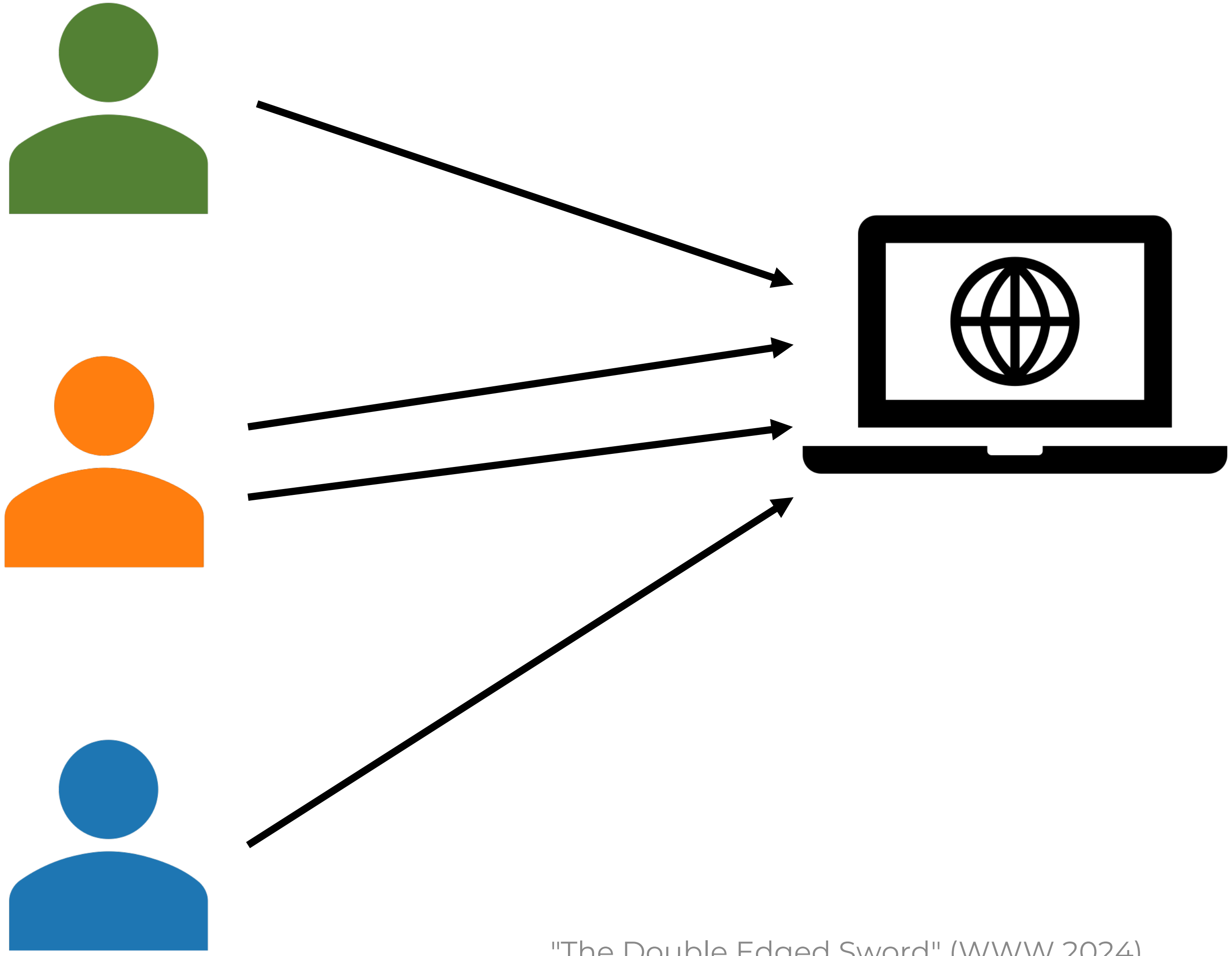
Asuman Senol*, Alisha Ukani*, Dylan Cutler, Igor Bilogrevic

The Web Conference (WWW) 2024

User (Re)Identification



User (Re)Identification



Tracking

- Analytics
- Targeted advertising
- Cross-site user identification

Security

- Account compromise prevention
- Bot detection (click-fraud)

Privacy Harms of Tracking

- Sensitive information can be revealed unwillingly:
  - High school girl’s pregnancy status before she had told her father [1]
  - WebMD searches to insurance company
  - Sexual orientation
  - Political views

[1] Charles Duhigg. How companies learn your secrets, https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

How to Reidentify Users

Strategy

- Check their IP address

- Use cookies

Outcome

- Unreliable signal
  - Same user can visit from different WiFi networks
  - Many people have the same public IP

Third Party Cookies

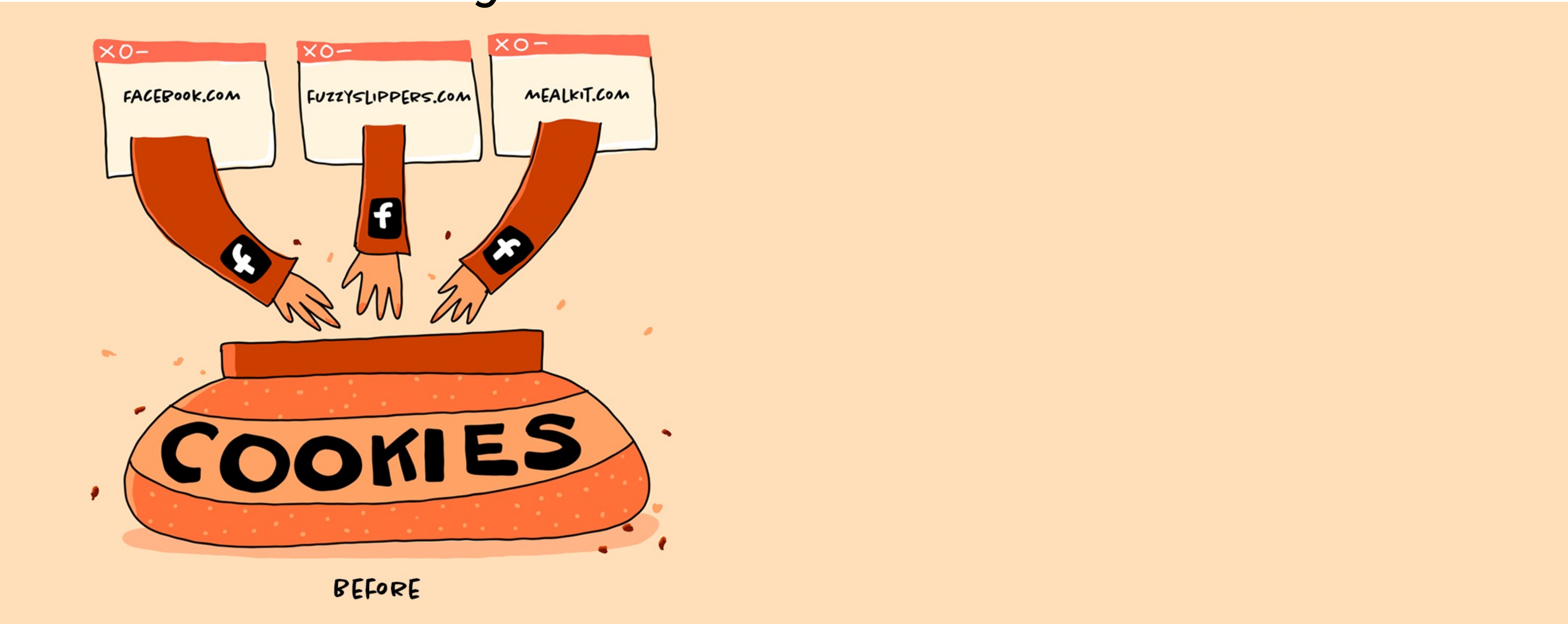


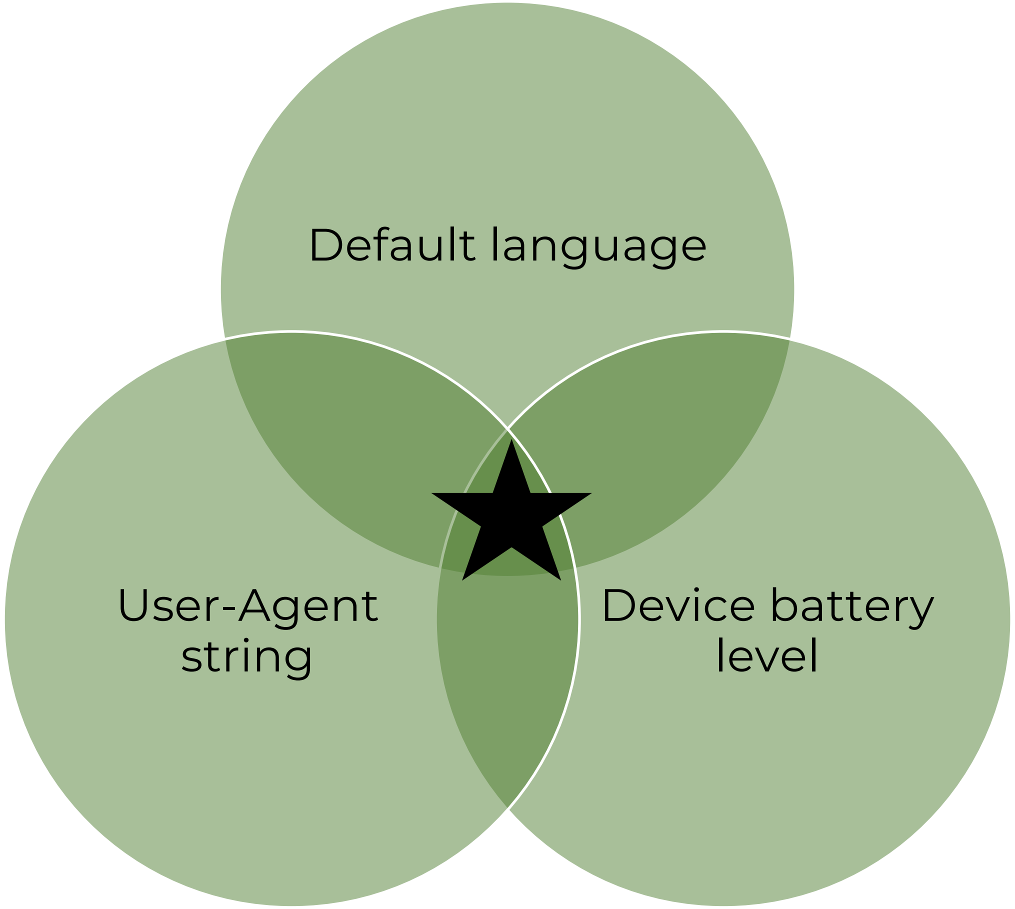
Image credit: Meghan Newell via Mozilla Security Blog. https://blog.mozilla.org/security/2021/06/01/total-cookie-protection-in-private-browsing/

How to Reidentify Users

Strategy	**Outcome**
• Check their IP address	• Unreliable signal • Same user can visit from different WiFi networks • Many people have the same public IP
• Use cookies	• Browsers no longer support 3rd-party cookies
• Find a cookie-less form of tracking	• This works!

Browser Fingerprinting

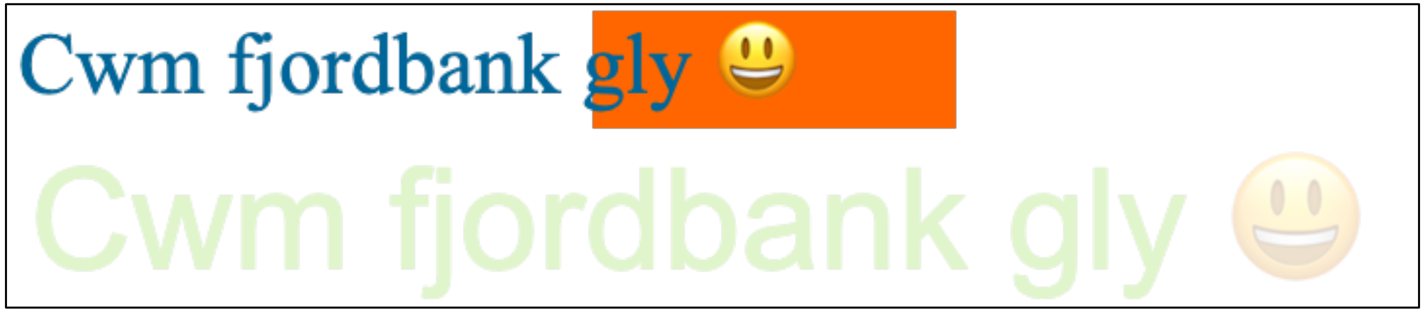
A method of uniquely identifying users without cookies across websites by **querying information about the user's device**

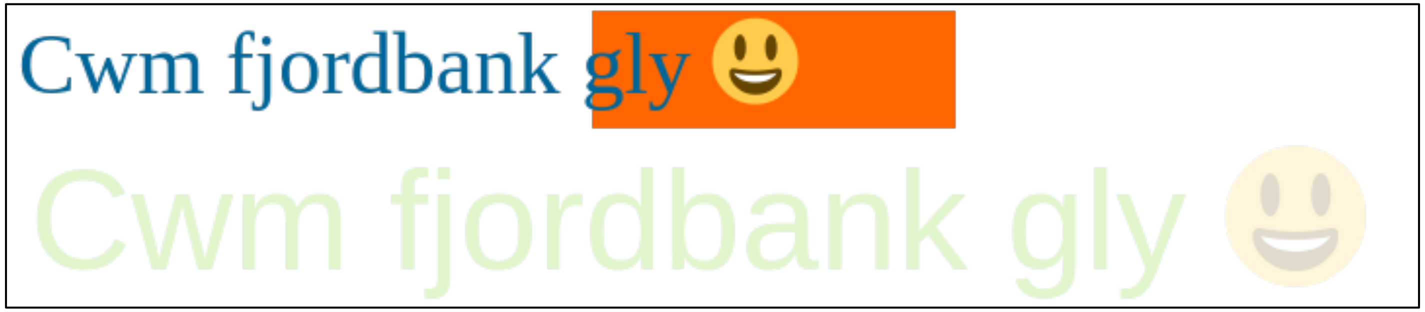


Canvas Fingerprinting

Key idea: stealthily draw shapes, text, and emojis in the JS canvas

Differences in the user's hardware and browser will render these images differently

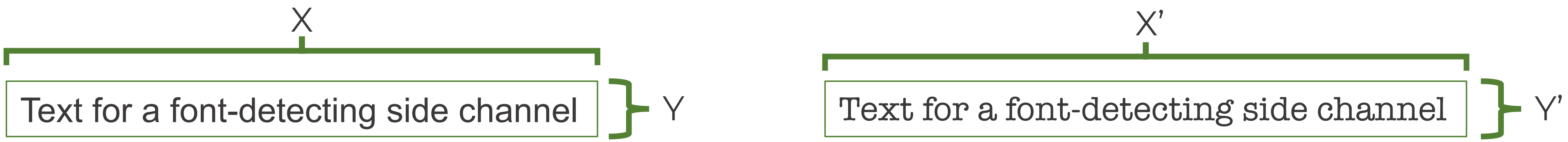


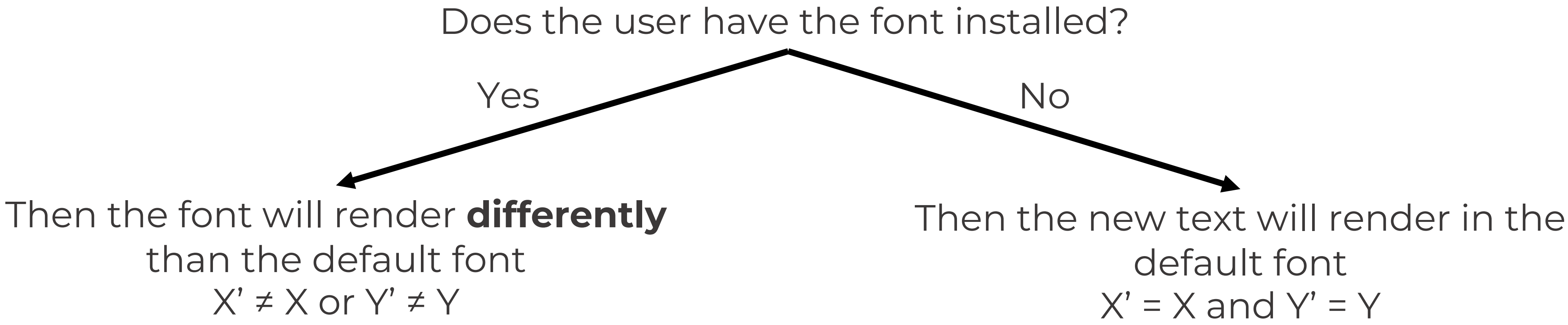


Canvas Font Fingerprinting

Default font:

Font that we want to check the presence of:


X
Text for a font-detecting side channel
Y
X’
Text for a font-detecting side channel
Y’



Does the user have the font installed?

Yes

Then the font will render **differently** than the default font
X’ ≠ X or Y’ ≠ Y

No

Then the new text will render in the default font
X’ = X and Y’ = Y

Who should we believe?

Privacy Advocates

Fingerprinting is always used for tracking. **Let’s ban it completely**

Website Developers

I use fingerprinting to protect user security. **You shouldn’t ban it**

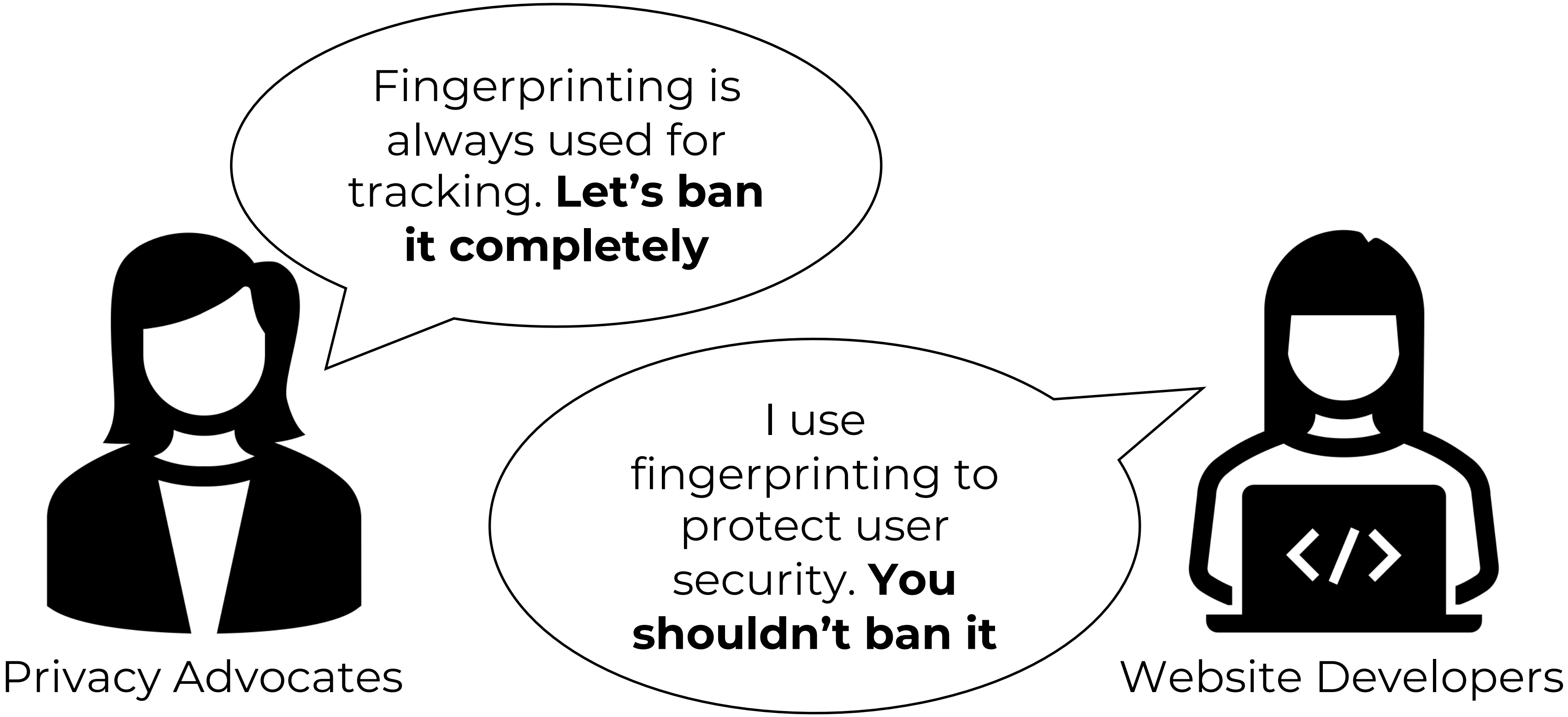
Prior research [1, 2] says I’m right!

But that research has severe limitations

[1] Antonin Durey, Pierre Laperdrix, Walter Rudametkin, and Romain Rouvoy, “FP-Redemption: Studying browser fingerprinting adoption for the sake of web security.” DIMVA 2021

[2] Xu Lin, Panagiotis Ilia, Saumya Solanki, and Jason Polakis, “Phish in sheep’s clothing: Exploring the authentication pitfalls of browser fingerprinting.” USENIX Security 2022

Who should we believe?



Fingerprinting is always used for tracking. **Let’s ban it completely**

Privacy Advocates

I use fingerprinting to protect user security. **You shouldn’t ban it**

Website Developers

We can’t determine the intent behind fingerprinting scripts…
But we can check if websites are even fingerprinting on login/sign-up pages

How often do websites fingerprint on their login and sign-up pages?

Our Contributions

A large-scale measurement study of the fingerprinting behavior of login and sign-up pages

A highly accurate ML model to detect login and sign-up pages

Open-source code to identify login and sign-up pages

Measurement Study

~~Measurement Study~~

Login/Sign-Up Detection Techniques

Login/sign-up detection

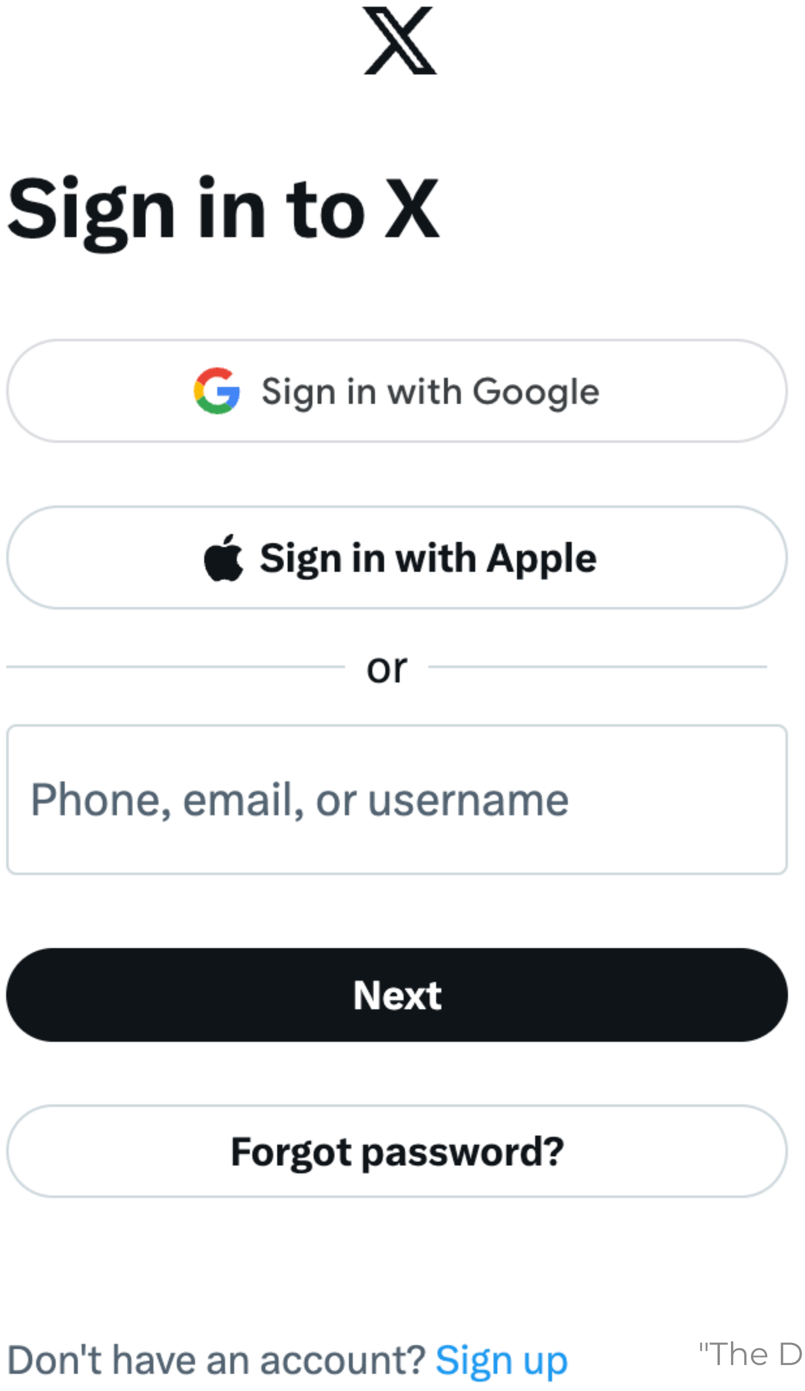
Many papers [1-10] have their own strategies to detect login/sign-up pages…so let’s use them!

[1] Suood Al Roomi and Frank Li. A Large-Scale Measurement of Website Login Policies. USENIX Security 2023
[2] Joe DeBlasio, Stefan Savage, Geoffrey M Voelker, and Alex C. Snoeren. Trip-wire: Inferring Internet Site Compromise. IMC 2017
[3] Kostas Drakonakis, Sotiris Ioannidis, and Jason Polakis. The Cookie Hunter: Automated Black-Box Auditing for Web Authentication and Authorization Flaws. CCS 2020
[4] Antonin Durey, Pierre Laperdrix, Walter Rudametkin, and Romain Rouvoy. FP-Redemption: Studying Browser Fingerprinting Adoption for the Sake of Web Security. DIMVA 2021
[5] Mohammad Ghasemisharif, Amrutha Ramesh, Stephen Checkoway, Chris Kanich,and Jason Polakis. O Single Sign-Off, Where Art Thou? An Empirical Analysis of Single Sign-On Account Hijacking and Session Management on the Web. USENIX Security 2018
[6] Hugo Jonker, Stefan Karsch, Benjamin Krumnow, and Marc Sleegers. Shepherd: a Generic Approach to Automating Website Login. MADWeb 2020
[7] Luka Lodrant. Designing a generic web forms crawler to enable legal compliance analysis of authentication sections. Master’s thesis, ETH Zurich, 2022
[8] Jannis Rautenstrauch, Giancarlo Pellegrino, and Ben Stock. The Leaky Web: Automated Discovery of Cross-Site Information Leaks in Browsers and the Web. IEEE S&P 2023
[9] Steven Van Acker, Daniel Hausknecht, and Andrei Sabelfeld. Measuring Login Webpage Security. SAC 2017
[10] Yuchen Zhou and David Evans. SSOScan: Automated Testing of Web Applicationsfor Single Sign-on Vulnerabilities. USENIX Security 2014

Cookie Hunter Heuristics

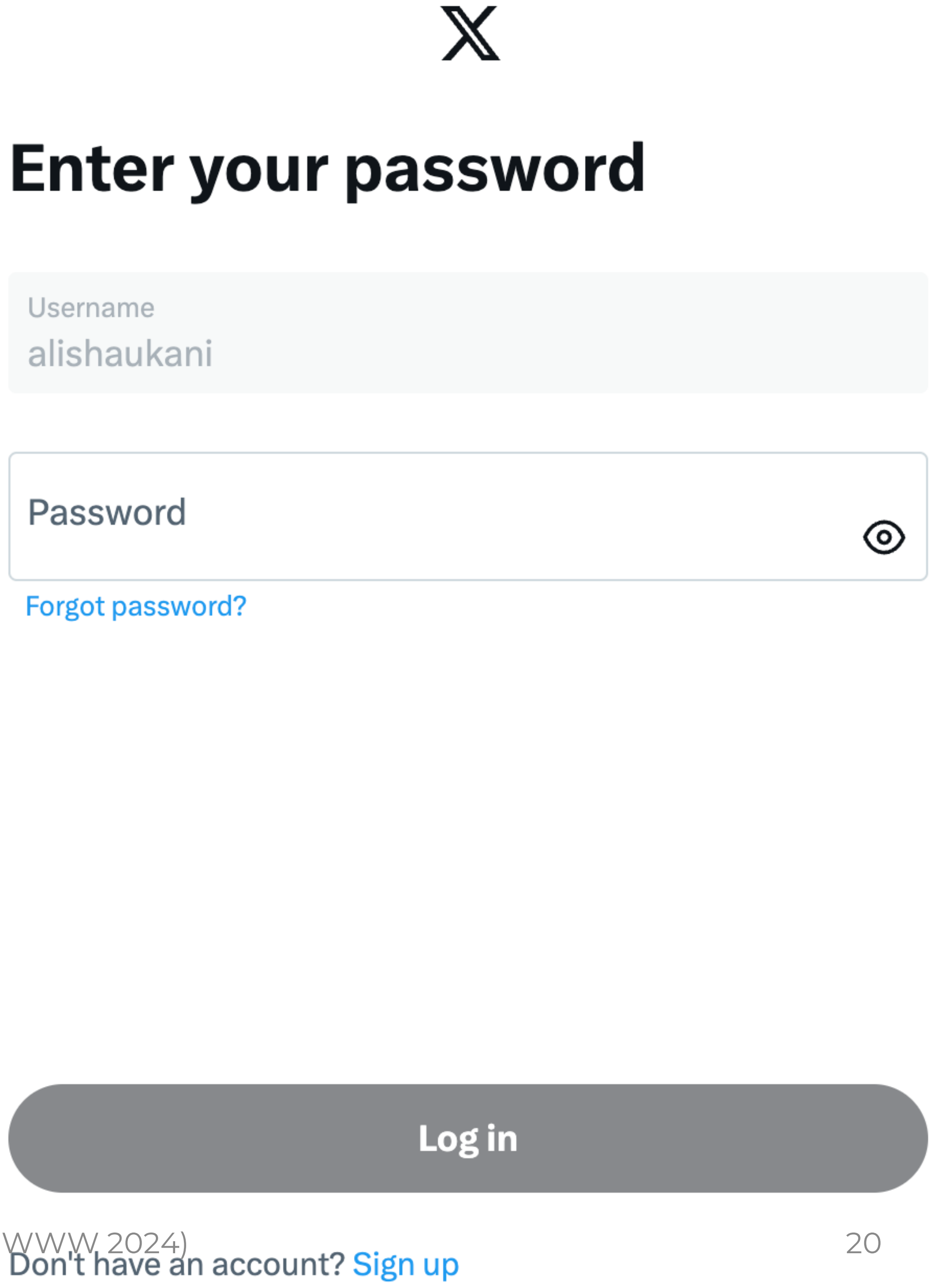
- State-of-the-art
- Created by Drakonis et al. in 2020, and used by [Lin22]
- Uses a combination of heuristics for strings and HTML elements
  - Regex: searches for English phrases like “register,” “login,” and “my profile”
  - HTML: e.g. number of password elements, presence of input elements for phone numbers

[Drakonis20] Kostas Drakonakis, Sotiris Ioannidis, and Jason Polakis. “The cookie hunter: Automated black-box auditing for web authentication and authorization flaws.” CCS 2020
[Lin22] Xu Lin, Panagiotis Ilia, Saumya Solanki, and Jason Polakis, “Phish in sheep’s clothing: Exploring the authentication pitfalls of browser fingerprinting.” USENIX Security 2022



Sign in to X

Don't have an account? Sign up

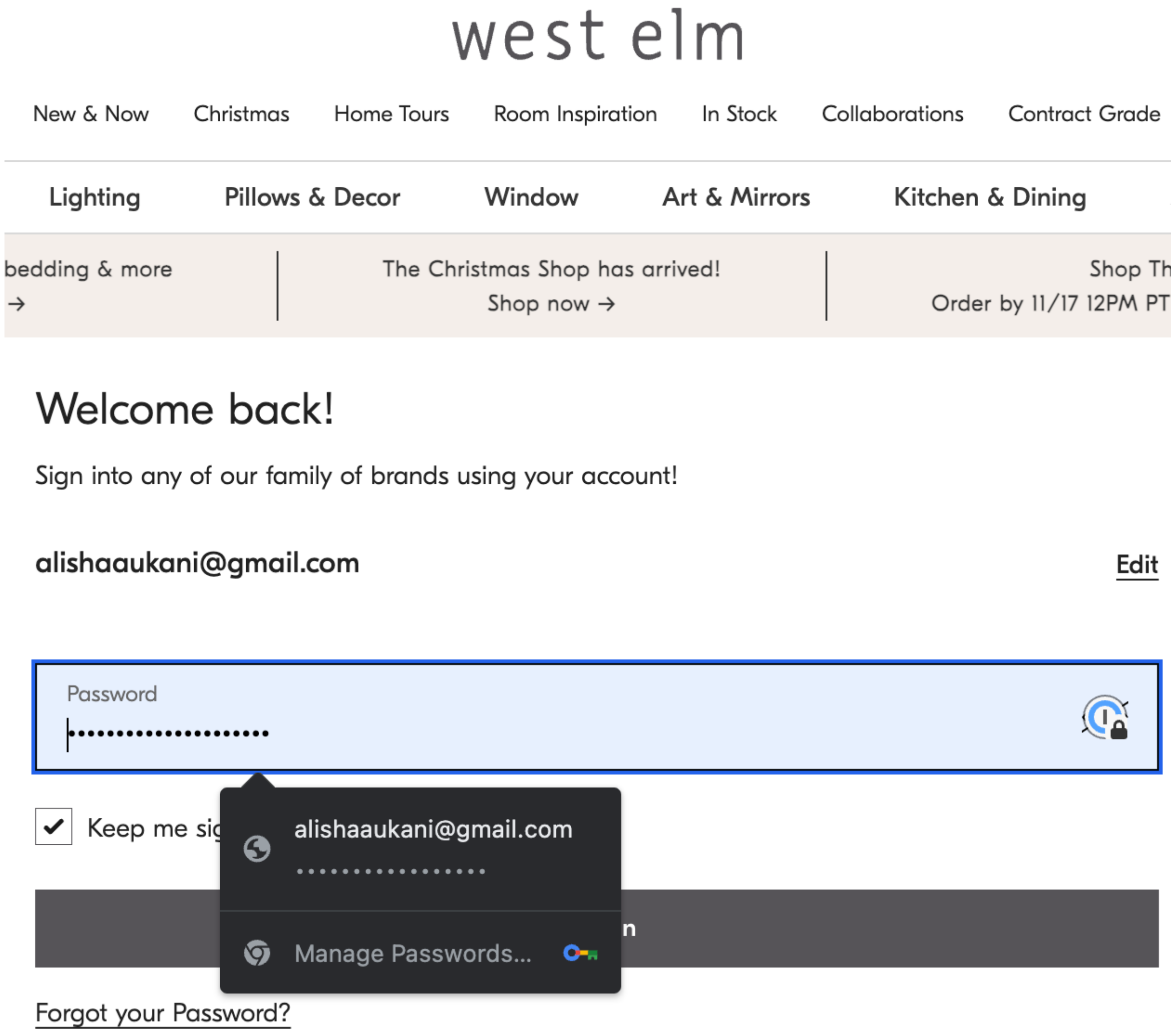


Enter your password

Username
alishaukani

Don't have an account? Sign up

Let's try Autofill



Welcome back!

Sign into any of our family of brands using your account!

alishaaukani@gmail.com

Password

••••••••••••••••••••

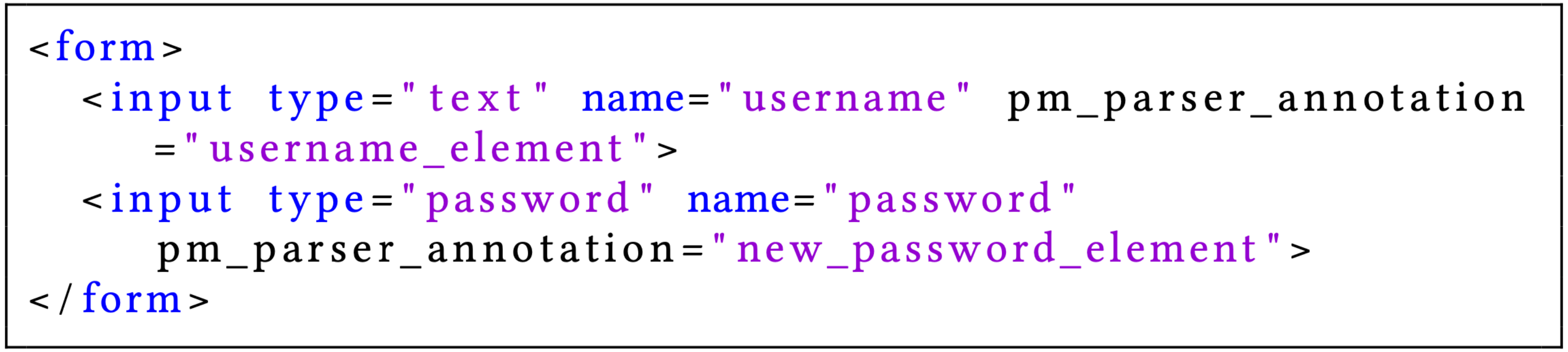
☑ Keep me sig

alishaaukani@gmail.com

••••••••••••••••

Autofill Heuristics

- Newly available to use in standard web crawlers
  - We are the first to use it for a measurement study
- Autofill classifies form elements and adds an HTML attribute with the classification



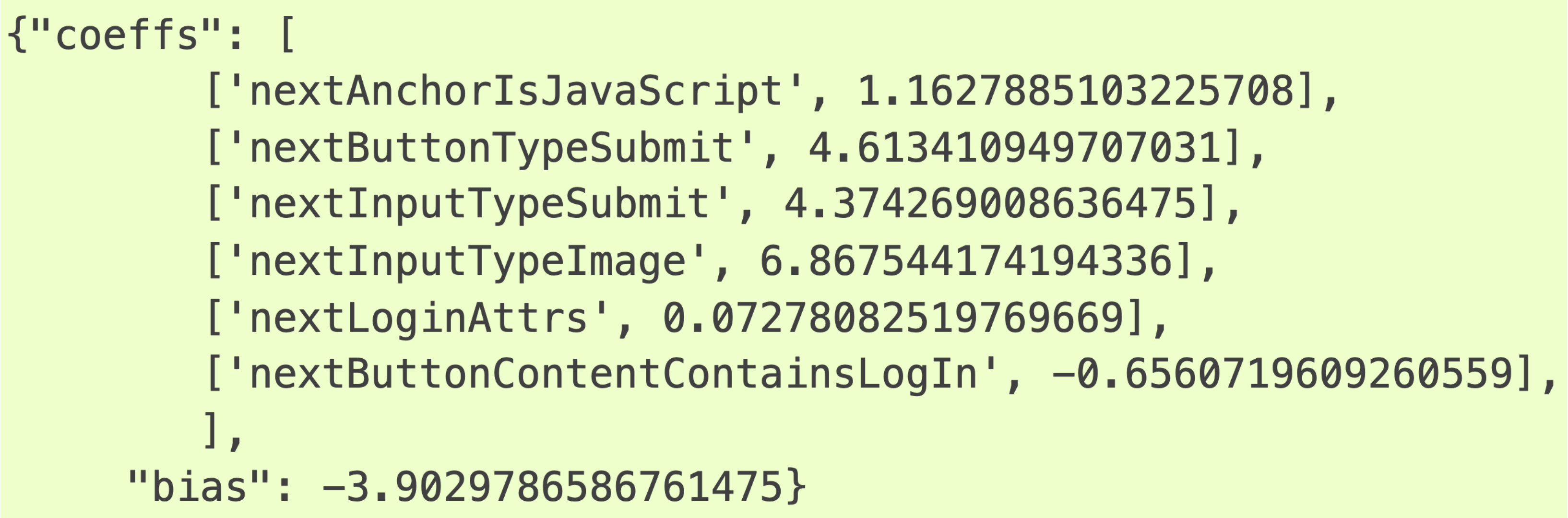
```
<form>
  <input type="text" name="username" pm_parser_annotation
     ="username_element">
  <input type="password" name="password"
     pm_parser_annotation="new_password_element">
</form>
```

Autofill also fails

- Can only classify forms as either login **or** sign-up, but not both
- Chrome Autofill uses a server-side component that we did not have access to
  - Crawler results will perform worse compared to how users experience the feature

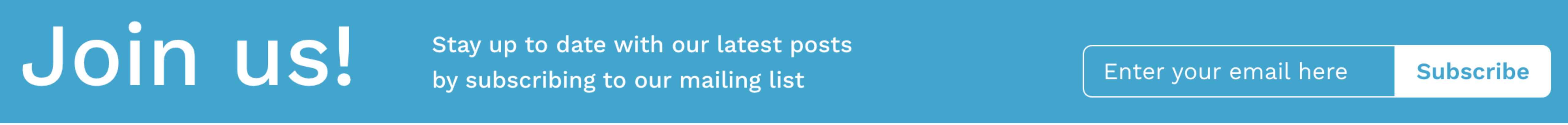
Fathom

- Mozilla-created ML model for classifying web pages, including login and sign-up pages
- Tags DOM nodes with probabilities



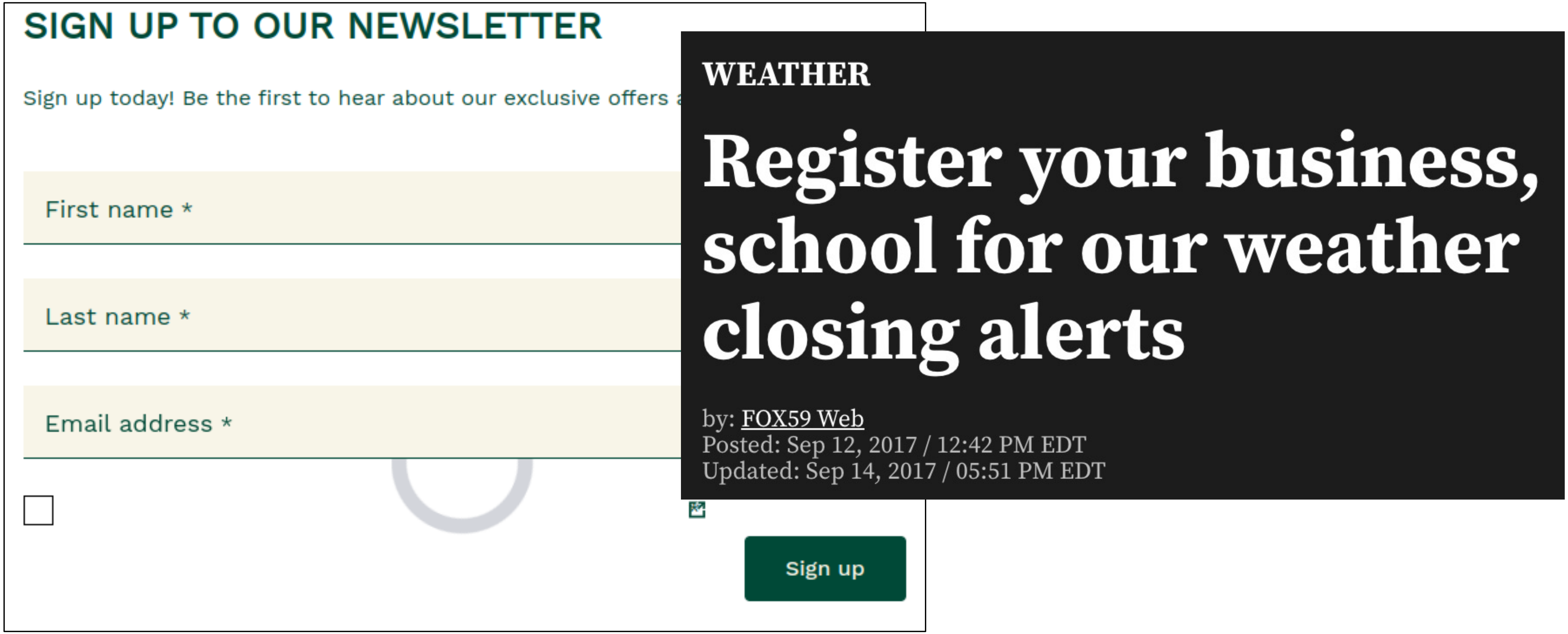
```
{"coeffs": [
        ['nextAnchorIsJavaScript', 1.1627885103225708],
        ['nextButtonTypeSubmit', 4.613410949707031],
        ['nextInputTypeSubmit', 4.374269008636475],
        ['nextInputTypeImage', 6.867544174194336],
        ['nextLoginAttrs', 0.07278082519769669],
        ['nextButtonContentContainsLogIn', -0.6560719609260559],
        ],
     "bias": -3.9029786586761475}
```

https://github.com/mozilla/fathom
https://mozilla.github.io/fathom/zoo/login.html



Join us!

Stay up to date with our latest posts
by subscribing to our mailing list



SIGN UP TO OUR NEWSLETTER

Sign up today! Be the first to hear about our exclusive offers a

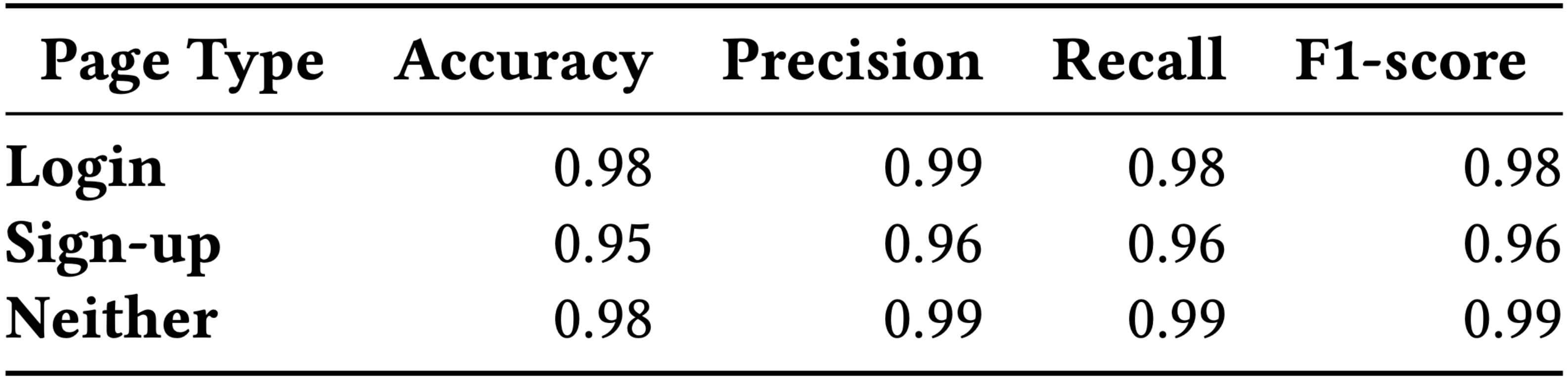
WEATHER

Register your business, school for our weather closing alerts

by: FOX59 Web
Posted: Sep 12, 2017 / 12:42 PM EDT
Updated: Sep 14, 2017 / 05:51 PM EDT

Our ML Model

- Manually found + analyzed login/sign-up pages for the CrUX top 1k, created a set of 88 features
- Neural network, outputs whether the page is login, sign-up, or neither



Page Type	Accuracy	Precision	Recall	F1-score
Login	0.98	0.99	0.98	0.98
Sign-up	0.95	0.96	0.96	0.96
Neither	0.98	0.99	0.99	0.99

Table 1: Classifier performance on test dataset.

Crawler Methodology

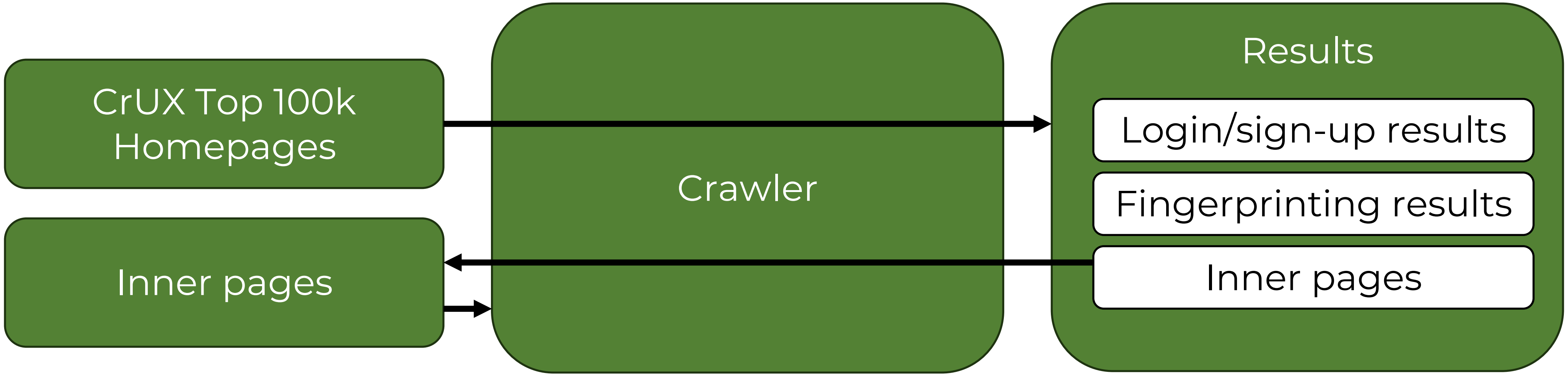
Fingerprinting Detection

- Implemented techniques from prior work [1, 2] to check 4 APIs:
  - Canvas: drawing images, emojis
  - Canvas fonts: drawing fonts to check if they're installed
  - WebRTC: real-time video
  - AudioContext: loading audio tracks

[1] Steven Englehardt and Arvind Narayanan. Online Tracking: A 1-million-site Measurement and Analysis. CCS 2016

[2] Umar Iqbal, Steven Englehardt, and Zubair Shafiq. Fingerprinting the Fingerprinters: Learning to Detect Browser Fingerprinting Behaviors. Oakland 2021

2-Pass Technique



Crawler Implementation

- To bypass bot detection, we:
  - Spoofed our User-Agent string to look like a normal user
  - Created mouse activity by scrolling on the page
  - Accepted cookies
- Unlike click-fraudsters, we prioritize completeness over volume/efficiency

Crawl Results

- Attempted to crawl the CrUX top 100k homepages
    - The CrUX list contains some duplicates, so that's really 98,845 pages
- We successfully crawled 95.8% of homepages (94,482/98,845) and 94.4% of inner pages (446,688/474,436)
    - Crawler errors may be due to the website detecting us and blocking visits

Are Websites Fingerprinting on Authentication Pages?

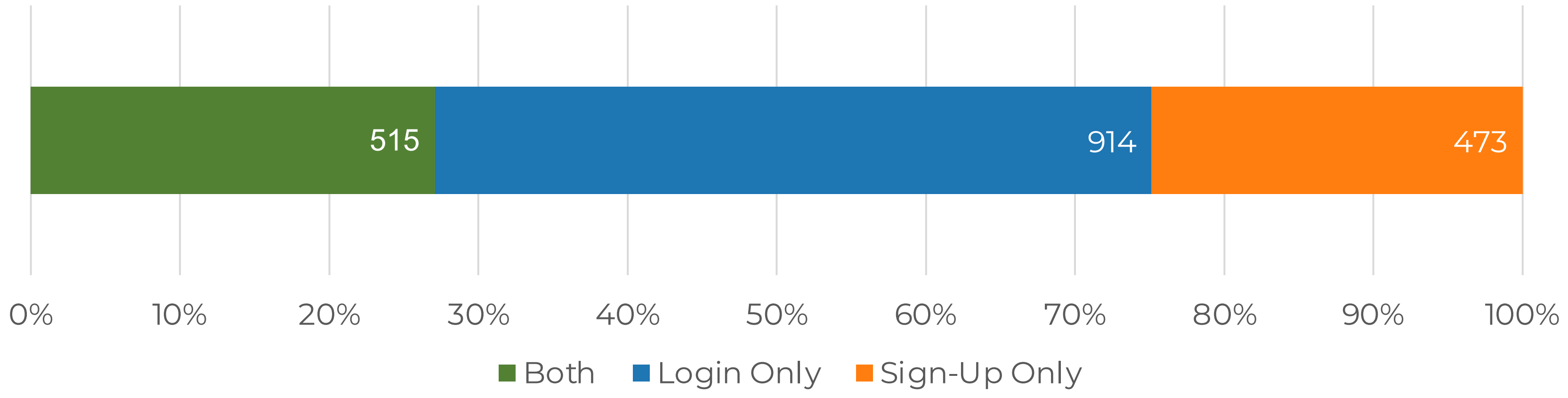
	Homepages	Login Pages	Sign-Up Pages
Web pages that perform fingerprinting	8,067 (8.5%)	4,872 (9.2%)	2,737 (12.5%)

"The Double Edged Sword" (WWW 2024)

Login vs Sign-Up

- If a websites fingerprints on at least one authentication page, then how do its login and sign-up pages differ?

Domains that use a 3rd-party fingerprinting script on at least one authentication page

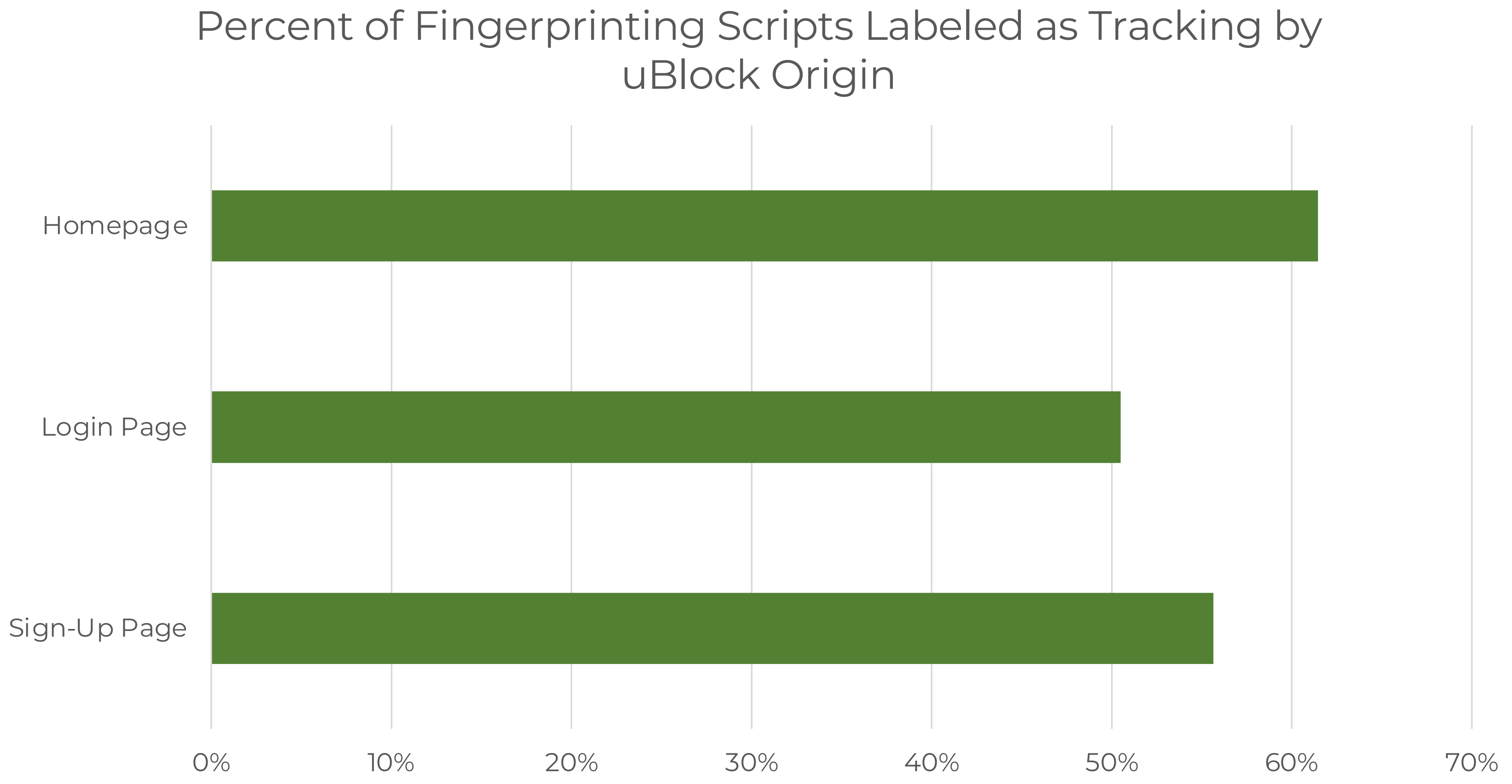


"The Double Edged Sword" (WWW 2024)

Login vs Sign-Up

- If a website fingerprints on **both** the login and sign-up page, do they use the same scripts?
- Mostly: 98% of domains (505/515) use the same set of third parties for both pages
- Some used for tracking

Tracking vs Non-Tracking



"The Double Edged Sword" (WWW 2024)

Comparison to Prior Work

Who should we believe?

Fingerprinting is always used for tracking. **Let's ban it completely**

Privacy Advocates

But that research has severe limitations

Prior research [1, 2] says I'm right!

I use fingerprinting to protect user security. **You shouldn't ban it**

Website Developers

[1] Antonin Durey, Pierre Laperdrix, Walter Rudametkin, and Romain Rouvoy, "FP-Redemption: Studying browser fingerprinting adoption for the sake of web security." DIMVA 2021
[2] Xu Lin, Panagiotis Ilia, Saumya Solanki, and Jason Polakis, "Phish in sheep's clothing: Exploring the authentication pitfalls of browser fingerprinting." USENIX Security 2022

[Durey21]: FP-Redemption

- Manually identified 446 domains collect lots of personal information
  - Financial services, gambling, retail
  - Government, job search, dating
  - **Results will not generalize to the web at large**
- Manually searched for login, sign-up, shopping cart, and payment pages

Antonin Durey, Pierre Laperdrix, Walter Rudametkin, and Romain Rouvoy, “FP-Redemption: Studying browser fingerprinting adoption for the sake of web security.” DIMVA 2021

Our Results vs [Durey21]

	Our FP Rate	[Durey21]'s FP Rate
Homepages	8.5%	23.0%
Login Pages	9.2%	23.4%
Sign-Up Pages	12.5%	31.1%

Similarities:

- Rates are highest for sign-up, then login, then home pages
- We identify some of the same fingerprinting scripts on authentication pages

Differences:

- We study a larger set of websites (100K vs 446)
- We use a narrower definition of fingerprinting that has fewer false positives

Antonin Durey, Pierre Laperdrix, Walter Rudametkin, and Romain Rouvoy, “FP-Redemption: Studying browser fingerprinting adoption for the sake of web security.” DIMVA 2021

[Lin22]: Phish in Sheep’s Clothing

- Hypothesis: websites are using fingerprinting to decide whether to show an MFA prompt to a user
    - New attack: that spoofing fingerprints bypasses MFA
- 16 out of 300 websites vulnerable to attack
- Small measurement study of Alexa top 20k

Xu Lin, Panagiotis Ilia, Saumya Solanki, and Jason Polakis, “Phish in sheep’s clothing: Exploring the authentication pitfalls of browser fingerprinting.” USENIX Security 2022

Our Results vs [Lin22]

- [Lin22] finds 18.5%, we find 9.2%
  - They use an overly broad definition of fingerprinting
- Why is our rate lower?
  - We consider a larger set of websites (100K vs 20K)
  - Less popular websites have lower rates of fingerprinting. We find that 14.73% of login pages for the top 1K perform fingerprinting
  - [Lin22] uses unreliable Cookie Hunter heuristics

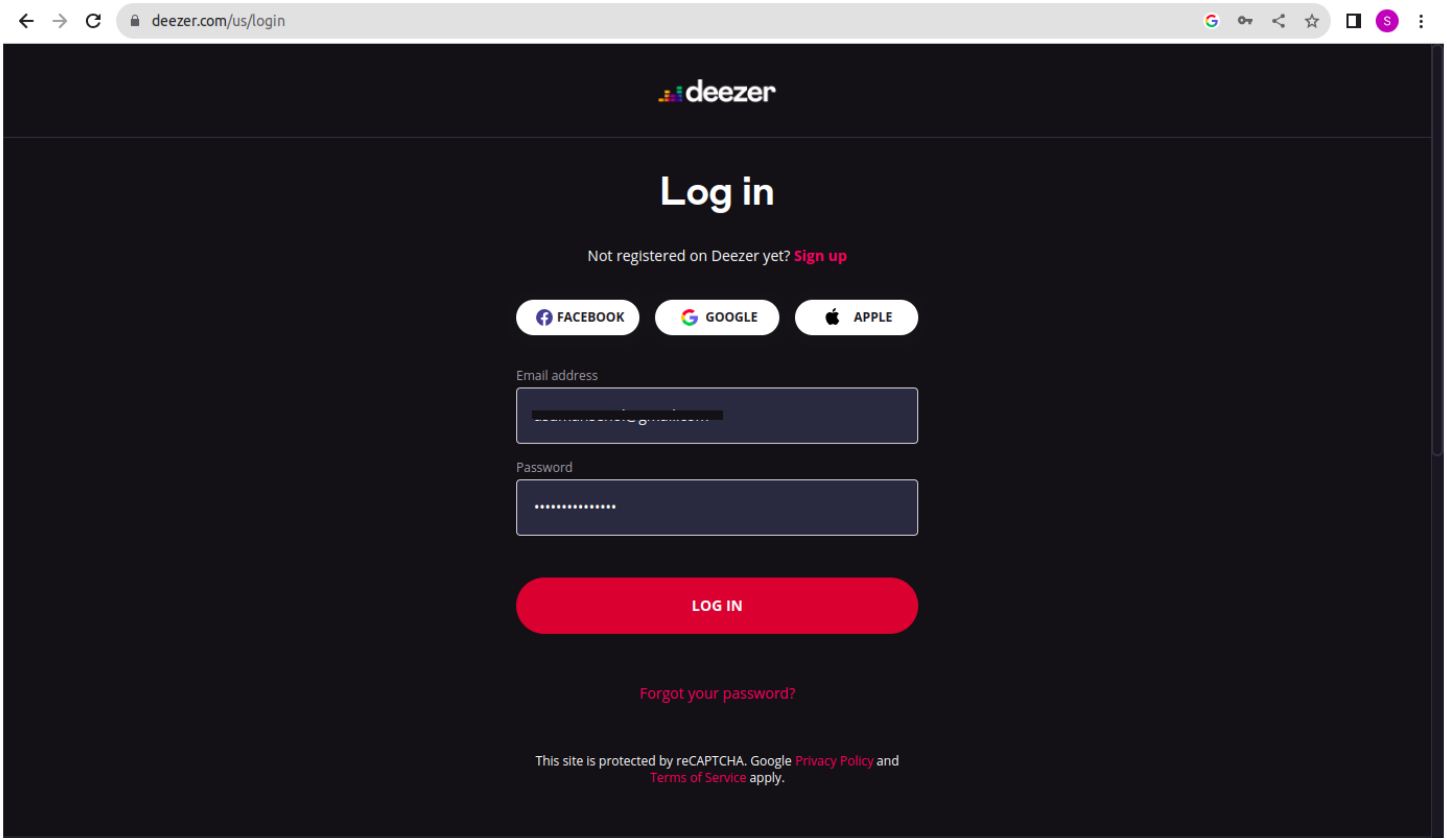
Xu Lin, Panagiotis Ilia, Saumya Solanki, and Jason Polakis, “Phish in sheep’s clothing: Exploring the authentication pitfalls of browser fingerprinting.” USENIX Security 2022

Is fingerprinting used for tracking or security?

Fingerprinting for Fraud Prevention

- Websites mostly fingerprint on the login page or on both login and sign-up pages
  - Not just the sign-up page, which might be used for tracking only
- Fraud-prevention company Signifyd had the most popular fingerprinting script on authentication pages
- Disabling FP breaks login pages for 2/30 websites

"The Double Edged Sword" (WWW 2024)



Log in

Not registered on Deezer yet? Sign up

Email address

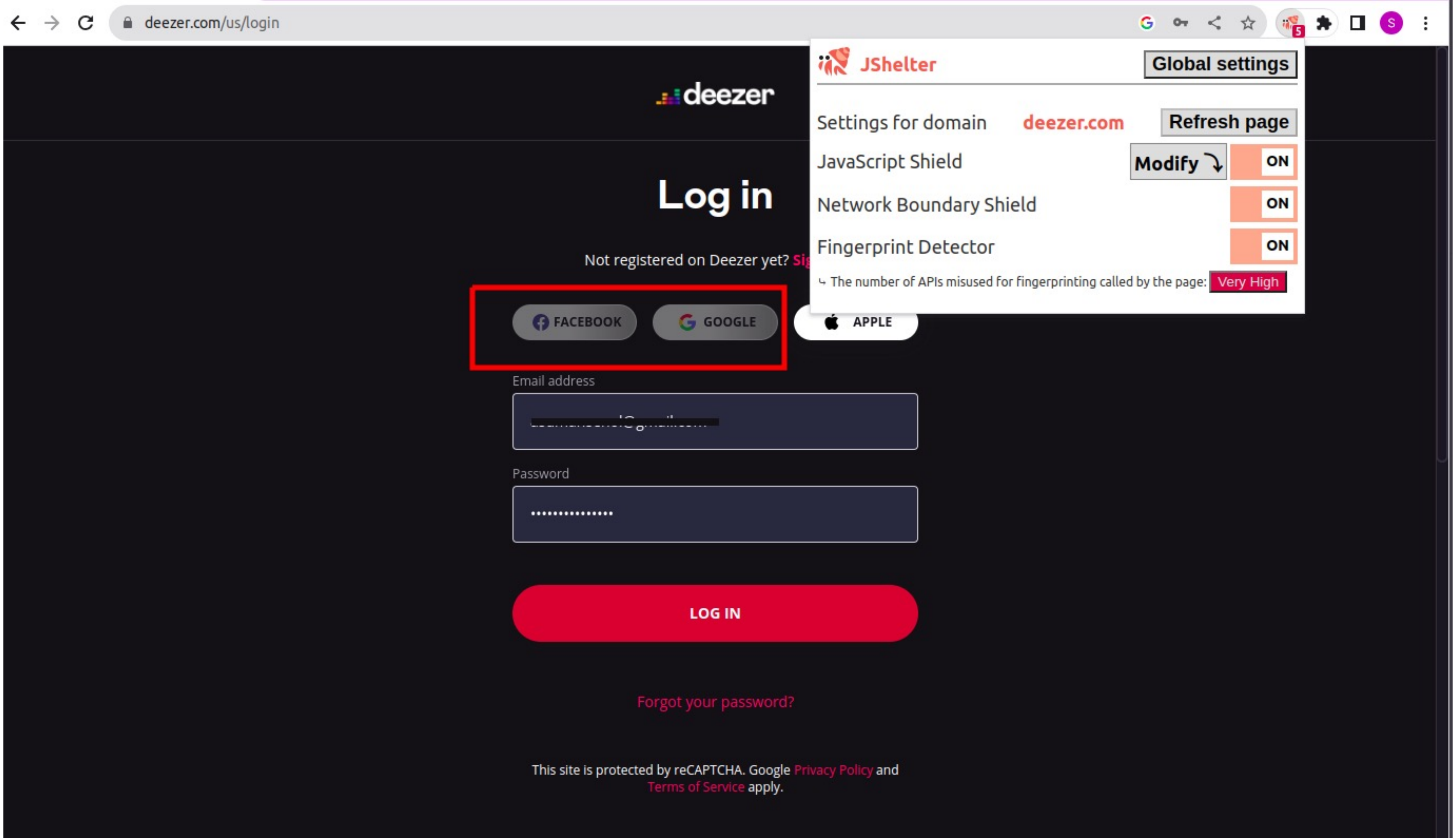
[redacted]

Password

•••••••••••••••

Forgot your password?

This site is protected by reCAPTCHA. Google Privacy Policy and Terms of Service apply.



JShelter

Global settings

Settings for domain deezer.com

Refresh page

JavaScript Shield — Modify — ON

Network Boundary Shield — ON

Fingerprint Detector — ON

↳ The number of APIs misused for fingerprinting called by the page: Very High

Log in

Not registered on Deezer yet? Si

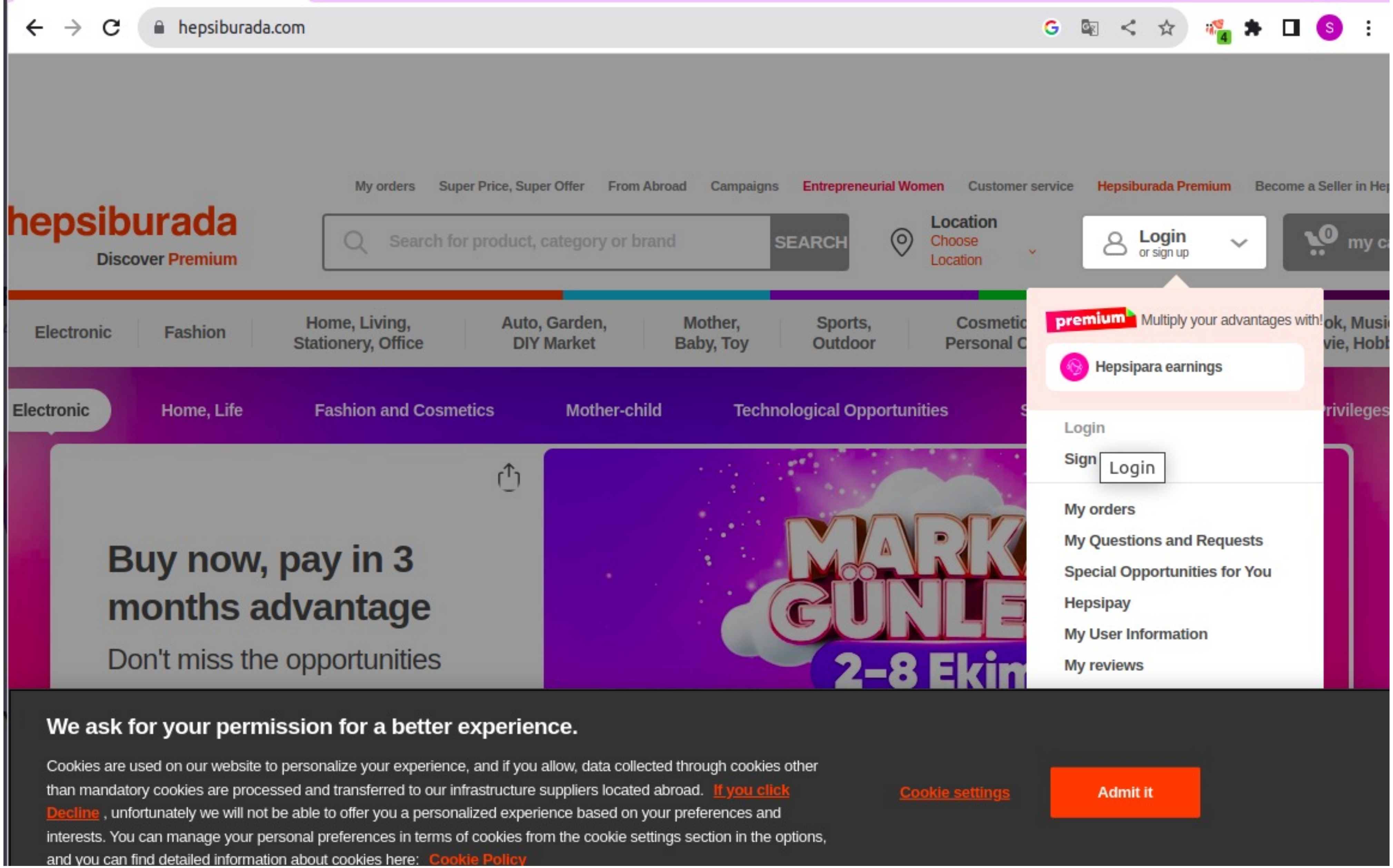
Email address

[redacted]

Password

•••••••••••••••

This site is protected by reCAPTCHA. Google Privacy Policy and Terms of Service apply.

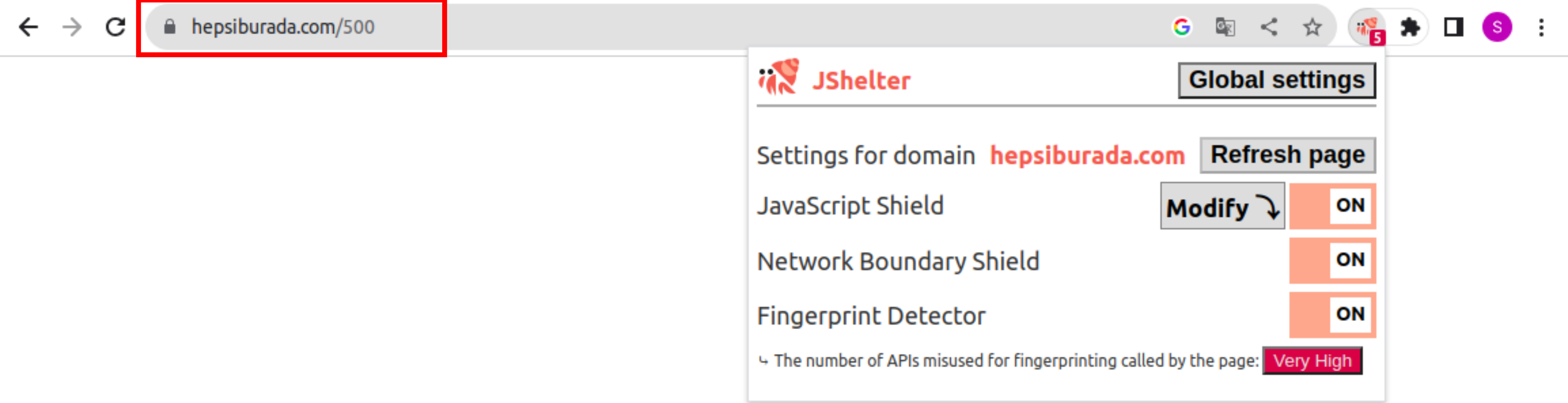


Buy now, pay in 3 months advantage

Don't miss the opportunities

We ask for your permission for a better experience.

Cookies are used on our website to personalize your experience, and if you allow, data collected through cookies other than mandatory cookies are processed and transferred to our infrastructure suppliers located abroad. If you click Decline , unfortunately we will not be able to offer you a personalized experience based on your preferences and interests. You can manage your personal preferences in terms of cookies from the cookie settings section in the options, and you can find detailed information about cookies here: Cookie Policy



Settings for domain hepsiburada.com

JavaScript Shield ON

Network Boundary Shield ON

Fingerprint Detector ON

↳ The number of APIs misused for fingerprinting called by the page: Very High

Fingerprinting for Tracking

- 2 out of 30 is a very low rate
  - Cookie hijacking succeeded on 1 of these 2 websites
- Over 50% of the fingerprinting scripts on authentication pages were classified as tracking
- A fingerprinting script from a fraud-prevention company also **sent the fingerprints to an analytics company**

So who do we believe?

- Fingerprinting is used for both tracking and security, often at the same time
- My proposal: **the research community should build new, privacy preserving tools that improve user security**
- Potential research directions
  - Detangling tracking from benign code
  - Conveying trust: Private State Tokens from Privacy Sandbox and Privacy Pass from Cloudflare
  - Establishing trust inside the browser, potentially through monitoring user behaviors across websites

QR Code to Paper

Alisha Ukani
aukani@ucsd.edu