July 18, 2024

Website Data Transparency in the Browser

Sebastian Zimmeck, Daniel Goldelman*, Owen Kaplan*, Logan Brown*, Justin Casler*,*
Judeley Jean-Charles, Joe Champeau*, Hamza Harkous***

* Department of Mathematics and Computer Science, Wesleyan University
** Google

The 24th Privacy Enhancing Technologies Symposium
Bristol, UK

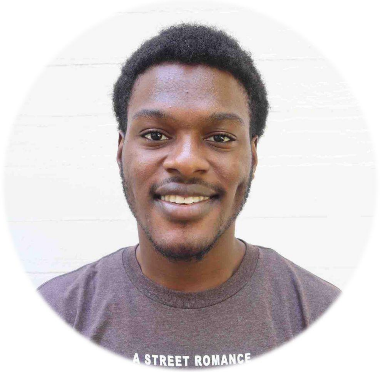
Who is playing for you tonight?

Daniel Goldelman

Owen Kaplan

Logan Brown

Justin Casler



Judeley Jean-Charles

Joe Champeau

Hamza Harkous

Sebastian Zimmeck

Additional Contributors:

Rafael Goldstein
David Baraka

What is the problem?

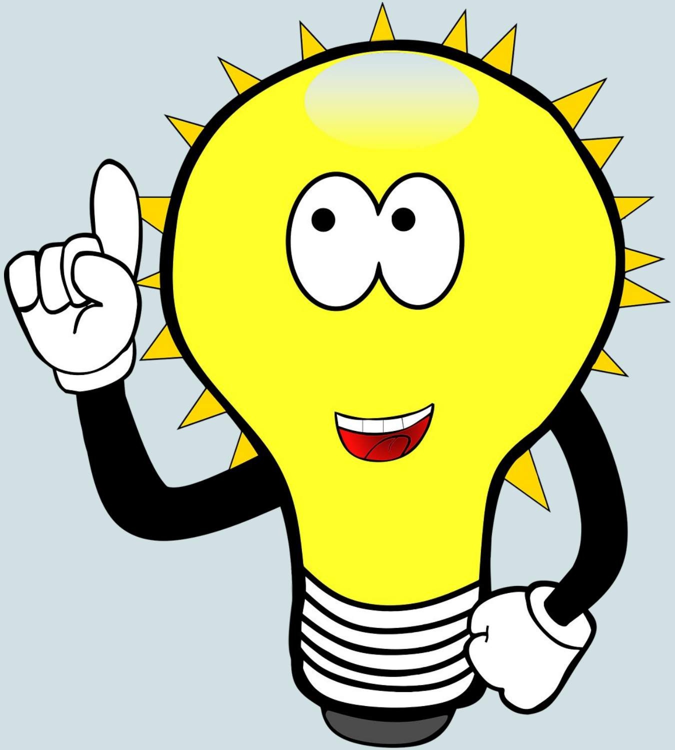
Website data collection is not transparent.

Especially by integrated third parties.

Current privacy notices, e.g, privacy policies, do not help much:

- They are often lengthy and time-consuming to read
- They do not always accurately describe how data is actually processed

What is the solution?

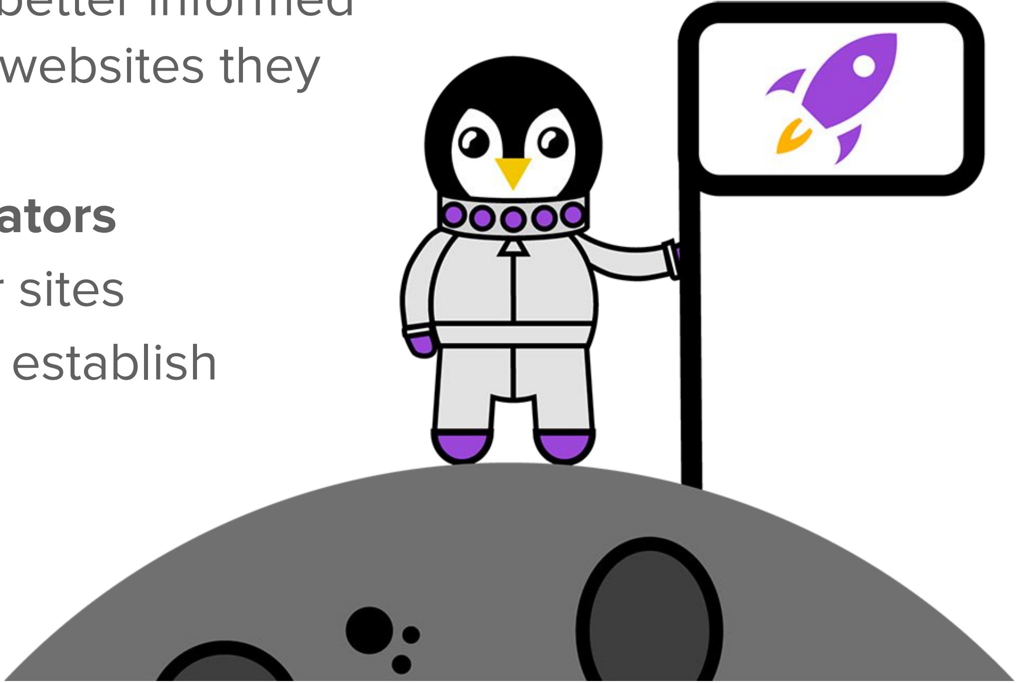


Dynamic In-browser Privacy Notices

Effects of Transparency

1. **Directly: People** are better informed and have more trust in websites they visit

2. **Indirectly: Site operators** better understand their sites and are incentivized to establish good privacy practices

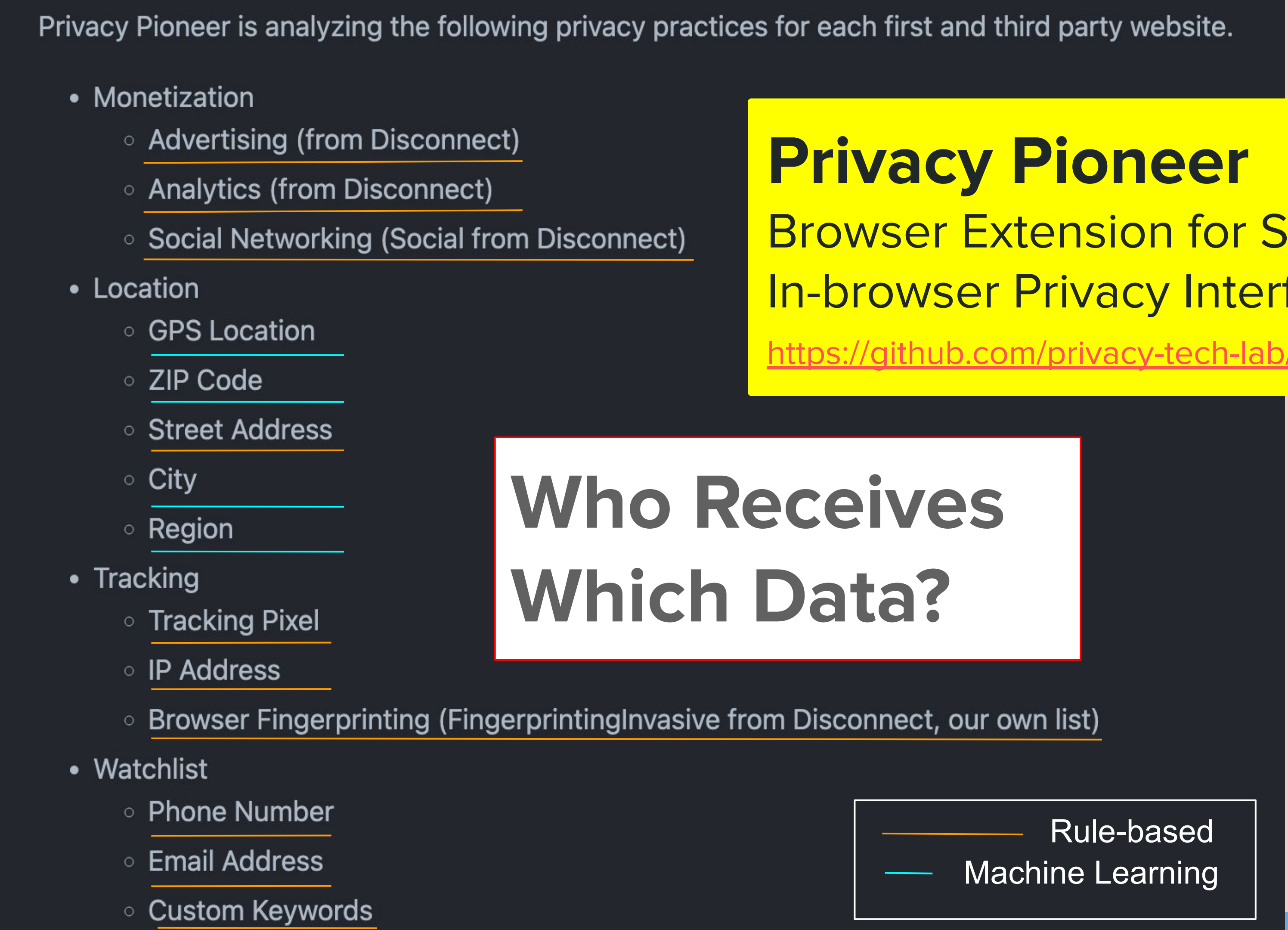


Privacy Pioneer

Browser Extension for Simulating In-browser Privacy Interfaces

https://github.com/privacy-tech-lab/privacy-pioneer

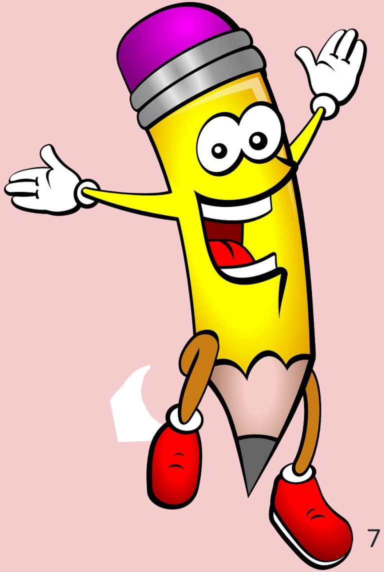
Who Receives Which Data?



Privacy Pioneer is analyzing the following privacy practices for each first and third party website.

- Monetization
  - Advertising (from Disconnect)
  - Analytics (from Disconnect)
  - Social Networking (Social from Disconnect)
- Location
  - GPS Location
  - ZIP Code
  - Street Address
  - City
  - Region
- Tracking
  - Tracking Pixel
  - IP Address
  - Browser Fingerprinting (FingerprintingInvasive from Disconnect, our own list)
- Watchlist
  - Phone Number
  - Email Address
  - Custom Keywords

🟧	Rule-based
🟦	Machine Learning



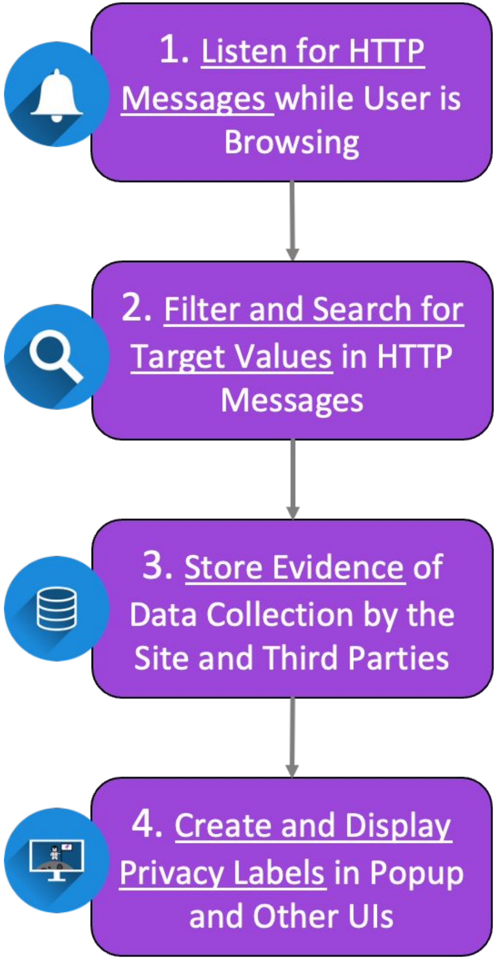
HTTP Request and Response Analysis

Deterministic

- URL lists
- Regular Expressions
- Attributes

Probabilistic

- **TinyBERT Machine learning model for location data** to better understand the context of locations

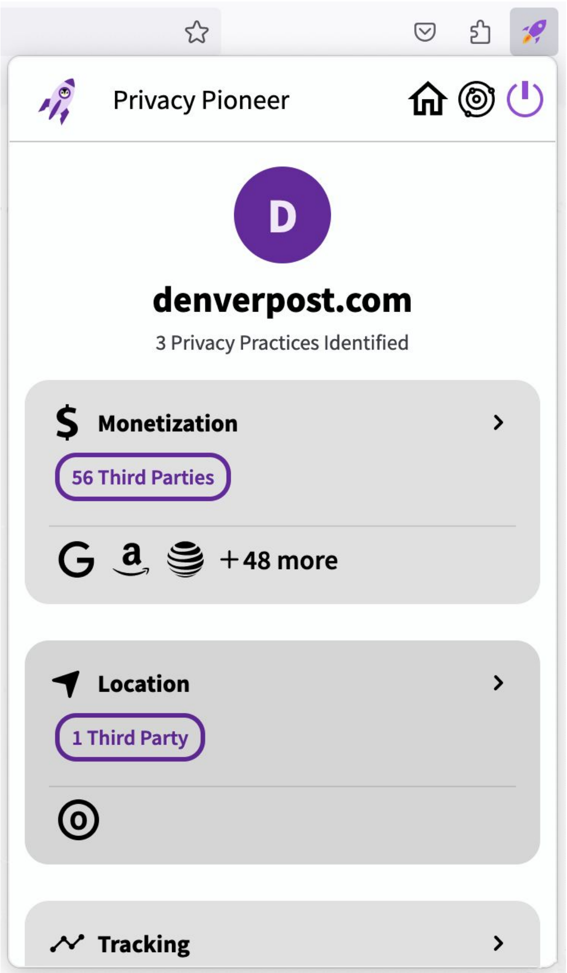


1. Listen for HTTP Messages while User is Browsing

2. Filter and Search for Target Values in HTTP Messages

3. Store Evidence of Data Collection by the Site and Third Parties

4. Create and Display Privacy Labels in Popup and Other UIs



Privacy Pioneer

D

denverpost.com

3 Privacy Practices Identified

$ Monetization ›

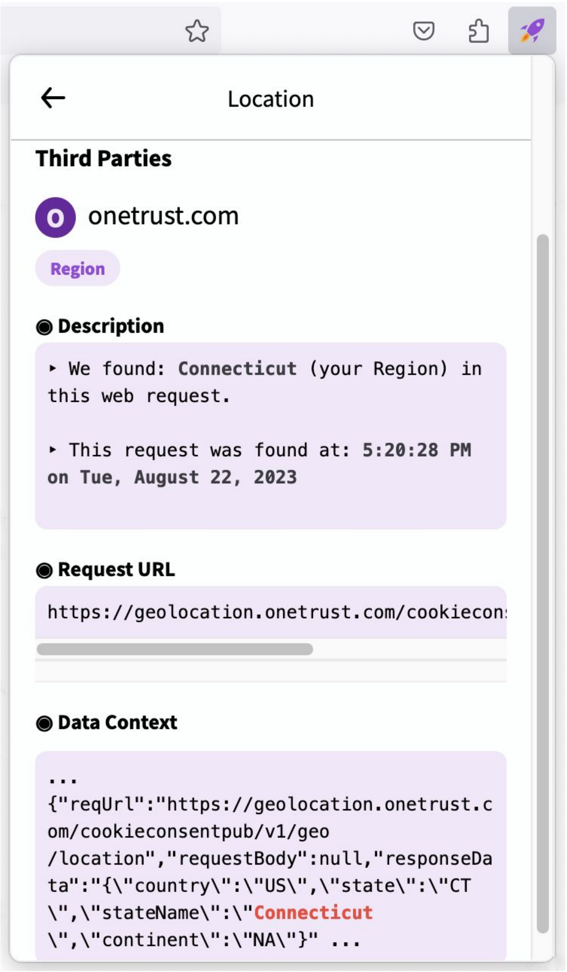
56 Third Parties

G a +48 more

Location ›

1 Third Party

Tracking ›



← Location

Third Parties

o onetrust.com

Region

● **Description**

```
‣ We found: Connecticut (your Region) in
this web request.

‣ This request was found at: 5:20:28 PM
on Tue, August 22, 2023
```

● **Request URL**

```
https://geolocation.onetrust.com/cookiecon
```

● **Data Context**

```
...
{"reqUrl":"https://geolocation.onetrust.c
om/cookieconsentpub/v1/geo
/location","requestBody":null,"responseDa
ta":"{\"country\":\"US\",\"state\":\"CT
\",\"stateName\":\"Connecticut
\",\"continent\":\"NA\"}" ...
```

TinyBERT Model Performance

TinyBERT
(59Mb)

Data Type	**F1 Score**
City	0.84
Region	1.00
Latitude	0.94
Longitude	0.91
Zip	1.00
Average	**0.94**

Test set with 533 total instances ~106 per data type

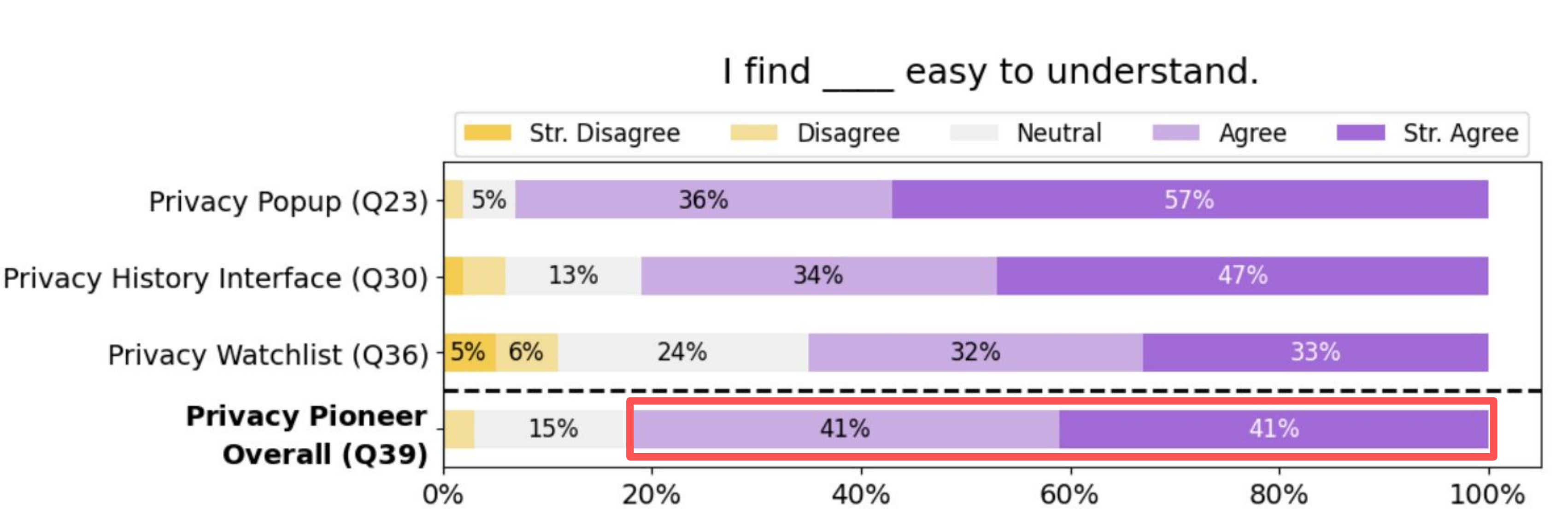
Privacy Pioneer Demo

Usability Study

Participants completed three tasks using Privacy Pioneer and afterwards filled a survey

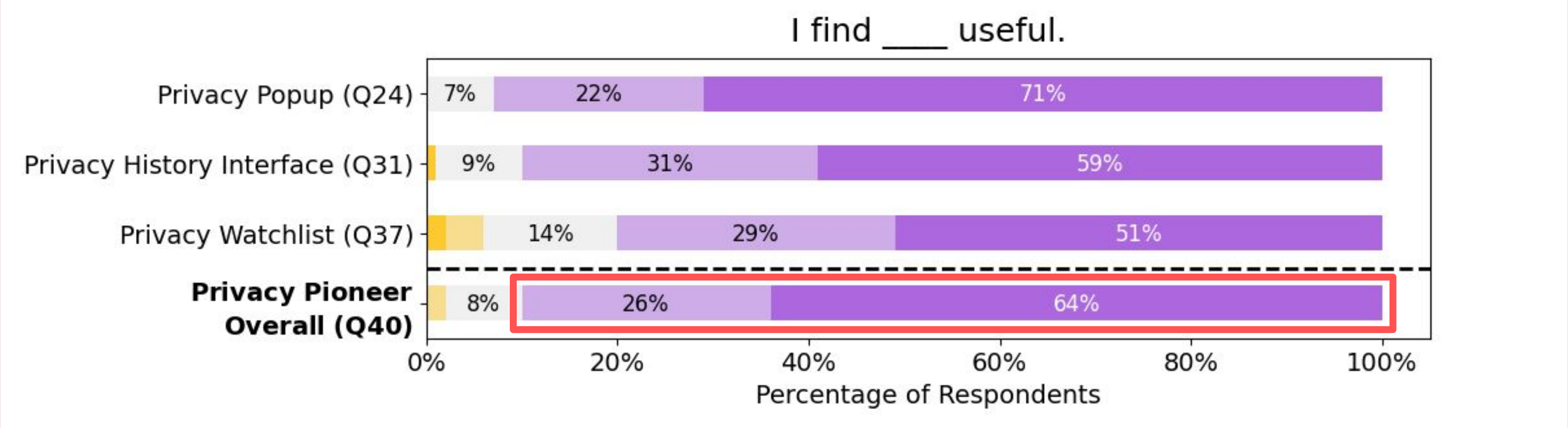
Usability: Understanding

82% of study participants found the Privacy Pioneer interfaces easy to understand



Usability: Utility

90% of study participants found the Privacy Pioneer interfaces useful



Key Takeaways

- We need more transparency of websites’ data collection and sharing practices to (1) help users understand sites’ collection and sharing practices and (2) motivate website operators to be privacy-sensitive
- Dynamic privacy analysis in the browser is (1) accurate and (2) feasible

References

- Privacy Pioneer browser extension
  https://github.com/privacy-tech-lab/privacy-pioneer
- Privacy Pioneer machine learning model
  https://github.com/privacy-tech-lab/privacy-pioneer-machine-learning
- Privacy Pioneer web crawler
  https://github.com/privacy-tech-lab/privacy-pioneer-web-crawler

Contact: sebastian@privacytechlab.org

Thank you!

We would like to thank our supporters!

Major support provided by Google.

Google

Additional support provided by Wesleyan University and the Anil Fernando Endowment.

Wesleyan University

Conclusions reached or positions taken are our own and not necessarily those of our supporters, its trustees, officers, or staff.

privacy-tech-lab