

# Website Data Transparency in the Browser

Sebastian Zimmeck<sup>1</sup>, Daniel Goldelman<sup>1</sup>, Owen Kaplan<sup>1</sup>, Logan Brown<sup>1</sup>, Justin Casler<sup>1</sup>, Judeley Jean-Charles<sup>1</sup>, Joe Champeau<sup>1</sup>, Hamza Harkous<sup>2\*</sup>

{szimmeck,dgoldelman,okaplan,lrbrown,jcasler,jjeancharles,jchampeau}@wesleyan.edu  
harkous@google.com

<sup>1</sup> Wesleyan University, Middletown, CT, United States

<sup>2</sup> Google, Zurich, Switzerland

## ABSTRACT

Data collection by websites and their integrated third parties is often not transparent. We design privacy interfaces for the browser to help people understand who is collecting which data from them. In a proof of concept browser extension, Privacy Pioneer, we implement a privacy popup, a privacy history interface, and a watchlist to notify people when their data is collected. For detecting location data collection, we develop a machine learning model based on TinyBERT, which reaches an average F1 score of 0.94. We supplement our model with deterministic methods to detect trackers, collection of personal data, and other monetization techniques. In a usability study with 100 participants 82% found Privacy Pioneer easy to understand and 90% found it useful indicating the value of privacy interfaces directly integrated in the browser.

## KEYWORDS

Web Privacy, Data Transparency, Privacy Dashboards, Notice and Choice, Consent, Privacy Labels, Usability, Privacy Enhancing Technologies, Web Traffic Analysis, BERT, Machine Learning

## 1 INTRODUCTION

Openness and transparency are cornerstones of data protection and the right to privacy. Per the OECD's fair information practice principles [58], "[t]here should be a general policy of openness". Further, "the existence and nature of personal data, and the main purposes of their use, as well as the identity and usual residence of the data controller" should be known. Per the GDPR [28], "transparency requires that any information [...] be easily accessible and easy to understand, [...] in particular, information to the data subjects on the identity of the controller and the purposes of the processing." However, the reality of privacy on the web is different. Many people feel a lack of transparency and control over what data is collected from them and by whom [13]. The web privacy problem is a transparency problem [26].

When people visit a website, it is generally not their intent to interact with ad networks, data brokers, or other third parties'

\*Corresponding authors: Sebastian Zimmeck and Hamza Harkous. Owen Kaplan and Logan Brown graduated in Spring 2022 and Spring 2023, respectively, and are no longer at Wesleyan University.

This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license visit <https://creativecommons.org/licenses/by/4.0/> or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

*Proceedings on Privacy Enhancing Technologies 2024(2)*, 211–234

© 2024 Copyright held by the owner/author(s).

<https://doi.org/10.56553/popets-2024-0048>



who collect their data on the site for tracking and advertising purposes.<sup>1</sup> Recent results suggest that, while third party tracking is omnipresent, many people are unaware of it [49]. This state of affairs poses challenges for the viability of notice and choice. While privacy policies are intended to provide notice and make data collection practices more transparent, they are not fit for this purpose. They take too long to read [53]. They also do not always accurately reflect how data is processed [84, 86]. A recent survey posits that genuine, informed consent at scale may not be possible [74].

Privacy policies have value as reference documents for regulators to hold site owners accountable. However, privacy labels, permission notifications, and other short-form notices are more informative for everyday use [23, 45]. In this work, we show how short-form notices can be generated automatically from web traffic while the user is browsing. We believe dynamic notices — showing in real-time which data is going where — are more informative than static descriptions of abstract privacy practices that “may” happen. When presented with their data in the “My Google Activity” dashboard, a third of the participants in a recent study were surprised by its scope and detail but at the same time viewed the data collection more beneficially [29]. Seeing data collection live is a more faithful representation of actual privacy practices than the often diverging descriptions in privacy policies or labels [47, 48, 84, 86].

We design and implement privacy interfaces to dynamically identify who is receiving which data at runtime in the browser. Our interfaces are intended for direct browser integration. Extraneous software would have limited reach, functionality, and usability. We understand our work as a step towards automating notice and choice in the browser by making notices backwards-traceable to analyzed code [83]. As a proof of concept we implement our interfaces in a Firefox browser extension, Privacy Pioneer.<sup>2</sup> In addition to pattern-based detection, we use a machine learning model to classify unstructured data, in our case location data, for which the web traffic context plays a significant role. By doing so, we show the potential improvements obtainable from machine learning models while still accounting for the constrained browser environment.

With this study, we hope to contribute towards making websites' data collection practices more transparent:

- (1) We design and implement privacy interfaces in a browser extension, Privacy Pioneer, to show how the dynamic analysis of data collection practices by websites can be directly integrated into the web browser. Our machine learning model for identifying location data reaches an average F1 score

<sup>1</sup>Whether this practice is considered data collection by the third party or data sharing from the first to the third party is a linguistic nuance. Both are used interchangeably.

<sup>2</sup>Privacy Pioneer is available at <https://github.com/privacy-tech-lab/privacy-pioneer>.

of 0.94. We supplement our model with deterministic methods to detect trackers, collection of personal data, and other monetization techniques. (§3)

- (2) In a usability study with 100 participants we evaluate the comprehensibility and utility of Privacy Pioneer’s privacy interfaces — a privacy popup, privacy history, and watchlist notifications — to help people understand the data collection practices of the sites they visit, including third parties. Overall, 82% of the participants found Privacy Pioneer easy to understand and 90% found it useful indicating the value of privacy interfaces directly integrated in the browser. (§4)

## 2 BACKGROUND AND RELATED WORK

Engaging people with the profiles from their web journey can create more trustworthy and positive experiences with targeted ads [5].

### 2.1 Data Transparency on the Web

After all, the understanding of how targeted ads work is often based on inaccurate folk models [79]. Profiles from long-term tracking can be constructed via a topic modeling algorithm run client-side on the data of the trackers people encounter when browsing the web [77]. To that end, it is the goal of various browser extensions, desktop apps, mobile apps, and websites to make ad tracking, browser fingerprinting, and other privacy-invasive behavior more transparent. The closest work to ours is Solitude [36], a desktop app for inspecting web or mobile app traffic and notifying people when their data, e.g., a location or email address, is collected.

Unlike Solitude, which is a standalone app, our goal is to build transparency functionality directly into the browser making it broadly available. We use a lightweight approach within the browser environment without relying on a VPN or web proxy as required by Solitude. Usability is critical for improving transparency for average people. Privacy Pioneer integrates into the browser environment creating a privacy history similar to a browsing history. It displays the practices of the current site in a popup and can display a browser notification upon a site collecting data. Beyond Solitude’s deterministic techniques, which we also use (§3.2), e.g., by matching known tracker URLs to Firefox’s integrated Disconnect Tracker Protection lists [19], we leverage a machine learning model to disambiguate people’s data from site data (§3.4).<sup>3</sup>

We build on ideas of existing browser extensions, in particular, Lightbeam [54], Ghostery [32], Privacy Badger [25], and Duck-DuckGo Privacy Essentials [22]. We complement their data recipient-based approach with a data category-based approach. Especially, a recent study suggested that the category of data being tracked is more important than who the web trackers are [79]. Indeed, data categories matter (§ 4.3.2). Someone may be fine with a site knowing their interest in listening to audiobooks, but they may rather not have their phone number or email address collected. In such cases people may appreciate a notification that a site just collected or is about to collect a piece of sensitive information. In a recent study on personal privacy assistants such a notification feature was ranked the highest [73]. We explore it here as well through our watchlist interface.

<sup>3</sup>Solitude’s classification accuracy and computational performance are not reported.

### 2.2 Web Traffic Analysis

We perform traffic analysis to extract useful and sensitive information from observed network traffic [62]. For the web OpenWPM provided a measurement infrastructure to detect, quantify, and characterize emerging online tracking behaviors, such as browser fingerprinting [27]. OpenWPM was extended to invisible login forms triggering autofilling of saved user credentials, exfiltrating social network data, and other privacy-invasive practices [2]. In addition to OpenWPM, OmniCrawl, a similar infrastructure, was used to find that the third party advertising and tracking ecosystem on mobile browsers is similar to that of desktop browsers [12].

Our work is based on foundational techniques for analyzing web messages for various data tracking practices, such as browser fingerprinting [3, 24, 57], the use of tracking pixels [38, 66], and the collection of location data [6]. Email addresses typed into forms can be collected by third party scripts even when people leave the site without submitting the form, which is especially concerning as email addresses are commonly used as identifiers for constructing profiles over time [69]. Browser fingerprinting, tracking pixel usage, location data collection, and data collected form-field entries are all surfaced in Privacy Pioneer’s interfaces.

In light of third party cookies being phased out on all major browsers [11], we expect to see a rise in browser fingerprinting. The accuracy of detecting browser fingerprinting can be improved via machine learning methods [42]. Those can also improve the detection of phishing sites [4] or malicious sites in general [50]. Here we make use of a machine learning model to identify location data collection, in particular, to detect an individual’s city, latitude, longitude, region, and ZIP code. Making sense of such location data often requires the broader context of the HTTP message in which it occurs to disambiguate, for example, whether a city is where the user is located or where the site owner can be contacted.

### 2.3 Privacy Dashboards

Privacy dashboards can help people to review and control the data collected about them [65], e.g., Blacklight allows people to enter a URL of any site to learn about its privacy practices [52]. Privacy dashboards are also increasingly built directly into the browser, e.g., Firefox’s Protections Dashboard [56]. Here we are exploring three different privacy interfaces: (1) a privacy popup displaying data collection practices of the current site and its integrated third parties, (2) a privacy history over all browsing sessions, and (3) a privacy watchlist that keeps track of a user’s custom keywords triggering notifications upon sites collecting those (§3.1.3). Considering tracking data, interest data, and raw technical data, most participants in a recent user study found interest data to be the most informative [75]. Displaying it in usable way is key [75].

### 2.4 Privacy Notices

Making privacy notices usable is a major challenge. Various design dimensions, e.g., the timing of notices, should be accounted for [67]. Poli-see explored how to best visualize privacy notices [34]. Concise and salient representations are promising [23]. “Nutrition labels” for privacy have been discussed [45], particularly, in the IoT space [19, 63] and are used on app stores. Apple’s privacy labels were found to be useful, though, prone to misconceptions [80] and sometimes

inaccurate and misleading [48]. Some apps were shown to violate their label by transmitting data without declaring so [47]. The same was shown for apps' privacy policies [84, 86].

Automated and dynamic privacy notice generation can help. For example, to align apps' privacy policies with their actual data practices, policies can be, at least, partially, generated from their code [81]. This dynamic analysis is also what we are pursuing here. Privacy Pioneer observes the actual behavior of a site, analyzes it, and creates a label for it. This dynamic creation has the advantage of giving people a much more accurate, concrete, and up-to-date picture of what is happening with their data compared to a static and abstract notice. It also opens up opportunities for personalized privacy notices based on user characteristics [46, 64]. Automatically generating privacy information has been fruitful, e.g., for answering privacy questions or assigning privacy icons [35].

### 3 PRIVACY PIONEER IMPLEMENTATION

The web browser is the natural instrument for notifying people about the data collection practices of the websites they visit.

#### 3.1 Architecture

Our definition of data collection encompasses both legal and surreptitious data collection by first and third parties.

**3.1.1 Goals, Requirements, and Non-goals.** We want to design and implement privacy analysis functionality and interfaces for use in the browser to make data collection practices of websites transparent to web users as they browse the web. As far as possible, the data analysis and interfaces should not interrupt people's browsing or impact the browser's computational performance. As far as possible, the analysis should also work locally without data disclosure. It is not our goal to achieve a comprehensive coverage of all collected data, all third parties, or all web traffic. Rather, we want to evaluate the effectiveness of our overall approach. While we make use of various methods to identify potentially privacy-invasive practices, e.g., browser fingerprinting, our goal is not the improvement of individual methods. We take a holistic view of the detected practices and aim to surface them in a usable way in the browser. It is also not our goal to provide a choice mechanism for the detected practices, though, detecting them will also enable choice (§5.2).

**3.1.2 Privacy Analysis Overview.** Figure 1 shows an overview of Privacy Pioneer's architecture. After applying APIs available in Firefox for listening and filtering HTTP messages, those messages are searched for data collected by websites and integrated third party scripts using probabilistic and deterministic methods. In particular, location data collection is detected by a machine learning model. Other data categories are detected deterministically by known attributes and string matching using regular expressions and URL lists. For the latter we use the Firefox-integrated Disconnect Tracker Protection lists [19]. Privacy Pioneer analyzes the following HTTP elements searching for various data categories:

**HTTP Elements** → Privacy Pioneer searches for relevant data in the following HTTP message (response and request) elements:

- **HTTP Headers**
  - **Request and Response URL:** The URL present in the request or response

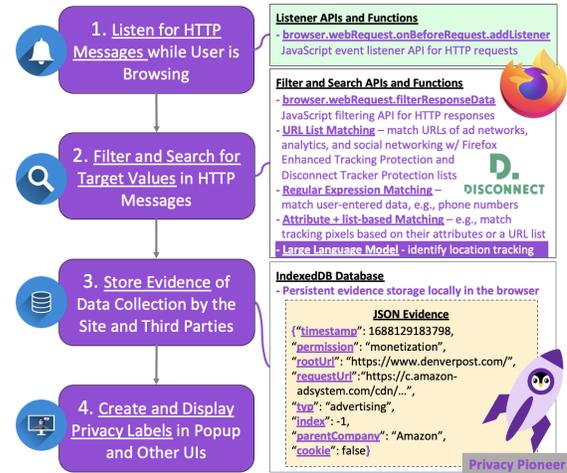


Figure 1: High-level architectural overview of the Privacy Pioneer browser extension for Firefox.

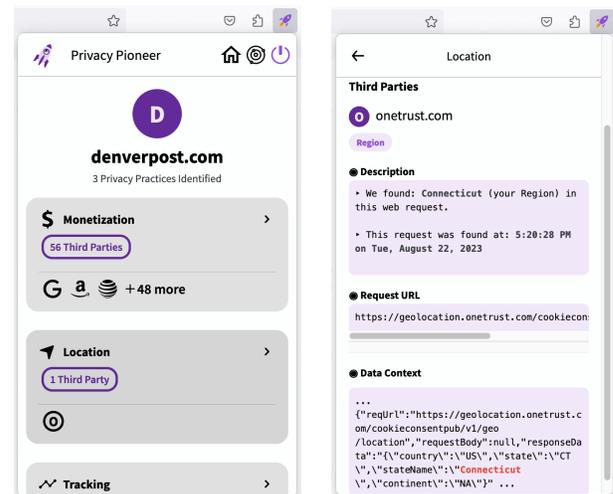


Figure 2: The popup interface showing that data was collected for monetization, location, and tracking purposes (left). After clicking the location card, detailed information about the collected data, including the context of the HTTP message, will be available (right).

- **Request and Response Cookies:** The cookies loaded into the browser obtained via the cookies API
- **Request Body:** The complete body of the HTTP request
- **Response Body:** The complete body of the HTTP response

**Data Categories** → Privacy Pioneer searches HTTP message elements for relevant data of the following categories:

- **Location:** City, Latitude, Longitude, Region, ZIP Code
- **Monetization:** Advertising, Analytics, Social Networking
- **Tracking:** Browser Fingerprint, Tracking Pixel, IP Address
- **Personal:** Email Address, Phone Number, Street Address, User-entered Custom Keywords

It is not necessary to decrypt any encrypted HTTP messages as those are available in plaintext when accessed through the browser APIs. Once evidence for data collection is found, it is analyzed, and the analysis results are locally stored in the browser. If the evidence supports a positive classification, the detected practice is ready to be displayed in the privacy interfaces (Figure 2).

**3.1.3 Privacy Interfaces.** We designed, implemented, and tested (§4) the following privacy interfaces in Privacy Pioneer:

- (1) **Privacy Popup** → Shows privacy practice information for the current site, such as the data collected, data categories, the first and third parties collecting it, and snippets of HTTP messages in which data was found.
- (2) **Privacy History** → Shows aggregated privacy practice information for all sites visited in the past. People can sort and apply filters to view instances from certain companies or of different data categories.
- (3) **Privacy Watchlist** → Allows people to enable browser notifications that are triggered when sites collect custom keywords, such as IP addresses, email addresses, or manually entered keywords.

**3.1.4 Generating Target Values.** A target value is a datum that Privacy Pioneer is searching for in a user’s web traffic. Monetization and tracking data can be identified without requiring any user input, i.e., the target values are the same for everyone (except for IP addresses). But the identification of location and personal data depends on individual target values. For location data target values Privacy Pioneer uses a third party IP-to-location API, IPinfo [40]. Upon installation and any subsequent IP address change a query with the user’s IP address is sent to IPinfo [40] to obtain the user’s city, region, street address, and ZIP code. In addition, Privacy Pioneer obtains the user’s latitude and longitude from Firefox’s built-in Geolocation API, which is generally more precise than a third party IP-to-location API and is provided by Google Location Services [55]. If the latitude and longitude identified in an HTTP message are within 300 characters of each other, then Privacy Pioneer will analyze if they qualify as an instance of location data collection. Privacy Pioneer distinguishes between fine and coarse locations. To qualify as a fine location instance the latitude and longitude values must be within  $\pm 0.1$  degrees from the Geolocation API target values. For coarse locations they must be within  $\pm 1.0$  degrees. Privacy Pioneer also supports manually-entered target values, which are required for some of the data categories, e.g., email addresses and custom keywords. However, we tried to keep the required user input to a minimum to achieve a high degree of usability.

## 3.2 Deterministic Analysis

To identify data collection practices Privacy Pioneer makes use of both deterministic and probabilistic analysis methods. The deterministic analysis is based on three methods: URL list matching, regular expression matching, and attribute-based matching.

**3.2.1 Analysis Methods.** For the monetization categories — advertising, analytics, and social networking — Privacy Pioneer matches their URLs based on the Disconnect Tracker Protection lists [19]. These lists are included in Firefox’s Enhanced Tracking Protection.

	Support	Precision	Recall	F1
<b>Advertising</b>	901	1.00	1.00	1.00
<b>Analytics</b>	193	1.00	1.00	1.00
<b>Social Networking</b>	84	1.00	1.00	1.00
<b>Browser Fingerprint</b>	7	1.00	1.00	1.00
<b>Tracking Pixel</b>	69	0.85	0.84	0.85
<b>IP Address</b>	49	1.00	0.92	0.96
<b>Email Address</b>	8	1.00	1.00	1.00
<b>Phone Number</b>	157	1.00	1.00	1.00
<b>Street Address</b>	16	0.94	1.00	0.97
<b>Custom Keywords</b>	10	1.00	1.00	1.00
<b>Weighted Average</b>	1494	0.99	0.99	0.99

**Table 1: Classification performance of our deterministic classifiers running on the deterministic test set. F1 scores of at least 0.96 for all but one category indicate that the deterministic approach is reliable for the analyzed categories.**

Specifically, the `webRequest.onHeadersReceived` API exposes the `urlClassification` object that indicates the type of tracking associated with a request, if any. Data for personal categories — email addresses, phone numbers, street addresses, and user-entered custom keywords — is identified based on regular expression matches.<sup>4</sup> As personal category data is more diverse than static monetization URLs, we leverage data formats, e.g., the email address format, to increase the identification accuracy for such data.

For supplementing the offering of the extension, we also added support for detecting tracking categories, such as browser fingerprints, tracking pixels, and IP addresses, via specialized regular expressions. Privacy Pioneer is looking for an IP address in the body of an HTTP message, which is an indicator that it is used for tracking, as opposed to an IP address in the header that is used to deliver the message to the correct recipient. To identify browser fingerprinting and tracking pixels we use both attribute- and list-based identification methods. A tracking pixel is identified if it is included in a manually curated URL list of known tracking pixels or if a set of four attributes is detected: (1) an image file, (2) with height and width properties set to 0 or 1, (3) containing the word “pixel,” and (4) containing a “?” character. For browser fingerprinting, including canvas fingerprinting, we follow a similar approach. We identify fingerprinters statically based on a list of known fingerprinting URLs (sourced from the `urlClassification` object) or dynamically based on function calls to fingerprinting libraries, such as `Fingerprint2`, or use of APIs, like `WebGL`.

**3.2.2 Classification Performance.** To evaluate the performance of our deterministic classifiers we created a test set (the *deterministic test set*) of 56 sites that would have a high probability of positive instances of the various data categories Privacy Pioneer is intended to detect. We created the deterministic test set by randomly selecting one technology for each category — advertising, analytics, social networking, browser fingerprinting, and tracking pixel — from the Disconnect Tracker Protection lists in Firefox or our own URL lists. For the IP address category we randomly selected one IP-to-location API based on a web search. Then, for each technology,

<sup>4</sup>Appendix 8.1 shows the regular expressions we implemented.

	Support	Type	Precision	Recall	F1
City	103	Regex	0.39	1.00	<b>0.56</b>
	103	SVM	0.35	1.00	0.52
Latitude	107	Regex	0.63	1.00	0.77
	107	SVM	0.90	0.95	<b>0.92</b>
Longitude	105	Regex	0.58	1.00	0.73
	105	SVM	0.97	0.86	<b>0.91</b>
Region	108	Regex	0.39	1.00	0.56
	108	SVM	0.55	1.00	<b>0.71</b>
ZIP Code	110	Regex	0.30	1.00	0.46
	110	SVM	0.61	0.95	<b>0.74</b>
Weighted Average	533	Regex	0.46	1.00	0.63
	533	SVM	0.93	0.72	<b>0.81</b>

**Table 2: Classification performance of our deterministic location data collection classifiers based on regular expressions and of our SVM-based location data collection classifiers running on the probabilistic test set (§3.3.2).**

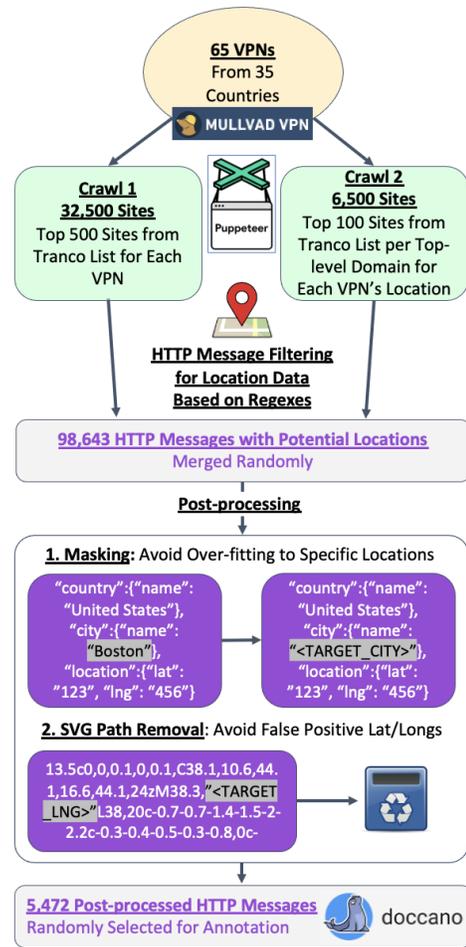
we searched BuiltWith [7] for sites that integrate it and randomly picked 8 sites for inclusion in our deterministic test set for a total of 48 sites. The remaining 8 sites are search engines, selected from a ranked list [31], to test for entered email addresses, phone numbers, street addresses, and custom keywords. We also tested for data from these categories by interacting with the other 48 sites, to the extent possible, by signing up to a site with an email address, password (i.e., custom keyword), phone number, and street address. Table 1 shows the performance of our deterministic classifiers with a manual inspection of the observed web traffic as ground truth.

**3.2.3 Location Data.** While regular expressions work well for identifying collection of personal data, such as email addresses (Table 1), they perform worse for collection of location data, such as ZIP codes (Table 2). Identifying a pattern as a user’s ZIP code or other location is often dependent on the context in which it appears. Whether a location is the user’s location or the location of the visited site cannot be solely determined by matching characters and formats. Another challenge is that SVG paths often have patterns that resemble location data leading to a significant number of false positives. The problem is less pronounced for latitudes and longitudes, possibly, because those specify a smaller geographical area in a more distinct format. Locations also lack distinctive attributes compared to, say, tracking pixels, which usually occur in an image file, making it difficult to apply attribute-based identification. Given these challenges of deterministically identifying location data collection we apply a machine learning model to perform a probabilistic analysis. Using a machine learning model instead of rigid rules is also beneficial for identifying location data in different formats, e.g., ZIP codes from different countries, as well as evolving or changing location data formats.

### 3.3 Location Dataset Creation

For developing and evaluating the performance of our machine learning model we created a location dataset for detecting the presence of people’s location data in HTTP messages.

**3.3.1 Data Collection.** Figure 3 shows an overview of the dataset creation process. To ensure we would have a high coverage of



**Figure 3: We created our location dataset by performing two web crawls during which we captured HTTP messages containing location data as detected by regular expressions. In the post-processing phase we masked the identified location data to avoid over-fitting the model and removed SVG paths from the dataset as those were clear false positives. We then sampled a random set of 5,472 HTTP messages and imported them into Doccano [20] for the subsequent annotation. Note that a site may generate multiple HTTP messages as it serves images, style sheets, scripts, and other resources.**

the variety of location data formats we performed web crawls connecting to 65 VPNs from 35 countries. Specifically, ZIP code formats differ from country to country (for example, US: 12345, Japan: 123-4567, India: 123456). Thus, diversifying our dataset to include multiple countries’ formats helped ensuring that our model would not over-fit to any specific country’s format. The protocol for crawling on one VPN was as follows:

- (1) Connect to the VPN.
- (2) Query the ipstack IP-to-location API to retrieve the VPN’s city, latitude, longitude, region, and ZIP code.<sup>5</sup>

<sup>5</sup>We used the ipstack IP-to-location API [41] for our crawls and later switched to the IPinfo IP-to-location API [40] for our Privacy Pioneer implementation. The latter allows a higher number of requests in their free tier. The functionality is the same.

	Support	Unanimous Pos, Neg	Majority Pos, Neg	Krippendorff's $\alpha$
City	103	39, 48	8, 8	<b>0.79</b>
Latitude	107	59, 34	6, 8	<b>0.81</b>
Longitude	105	51, 39	7, 8	<b>0.81</b>
Region	108	38, 53	9, 8	<b>0.79</b>
ZIP Code	110	28, 70	5, 7	<b>0.83</b>

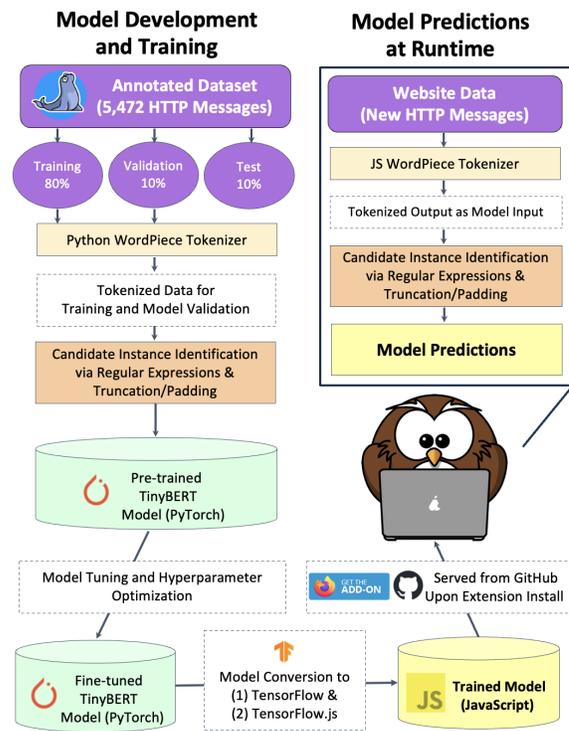
**Table 3: Krippendorff's  $\alpha$  inter-annotator agreement for the three annotators. Both positive and negative instances of unanimous agreement (3:0) occurred more often than majority agreement (2:1).**

- (3) Connect in sequence to a set of websites from the Tranco list [60]. The time-out for each site was set to 15 seconds.
- (4) Identify location data in HTTP messages based on regular expressions matching the VPN's city, latitude, longitude, region, or ZIP code.
- (5) Save the HTTP messages that were successfully matched for post-processing and annotation.

We parallelized our crawls with eight browsers at a time via the browser automation framework Puppeteer [33] and the puppeteer-cluster library [21]. The two crawls differed in the websites visited. The first crawl covered for each VPN the top 500 most popular websites globally per the Tranco list. The goal was to capture data from sites with high-volume of traffic as many people would be exposed to their privacy practices. The second crawl covered the top 100 most popular sites of the country where the VPN was located according to the Tranco list. We associated a site with a country based on its top-level country domain. This crawl aimed to capture the privacy practices of popular websites accounting for a diverse set of localized data formats. In total, both crawls generated 98,643 HTTP messages potentially containing location data.

After merging the HTTP messages from the two crawls we post-processed them by (1) masking locations, as identified by our regular expressions, and (2) removing SVG paths. We masked locations in the dataset, e.g., replacing "Boston" with the label "<TARGET-CITY>," to avoid over-fitting our model to specific locations. While SVG paths can contain numbers that look like ZIP codes, for example, they would always be false positives. Then, we randomly sampled 5,472 HTTP messages for annotation. Our goals for the annotation were to obtain even distributions of data instances (1) across the different location data categories and (2) across the different VPNs. With a target of at least 1,000 data instances per category we sampled an equal amount of instances per VPN. If a data category was more common in the VPNs, we sampled fewer data instances per VPN. If it was less common, we sampled more.

**3.3.2 Data Annotation.** The set of 5,472 HTTP message instances was imported into Doccano [20], an open source annotation tool we set up. The dataset consisted of instances from all five location data categories (city: 1,115, latitude: 1,068, longitude: 1,051, region: 1,078, and ZIP code: 1,160). Before importing the dataset we truncated each message instance to 250 characters before and after the regular expression match (or fewer characters if the message was shorter or the match occurred near one of the message ends). These truncated messages instances would provide sufficient context for why the match occurred and prevent length bias. Each instance could



**Figure 4: Our model development and predictions at runtime. During training the model evaluated itself with the validation set. The probabilistic test set was held-out for the classification performance evaluation.**

be a true or false positive. A true positive means that the regular expression match correctly identified an instance of data collection. A false positive means that it identified an instance incorrectly, for example, if a news article mentioned the name of the city where the VPN was located. For each category, 10% of the data, selected randomly, was annotated by three authors reaching an inter-annotator agreement between 0.79 and 0.83 as measured by Krippendorff's alpha. (Table 3).<sup>6</sup> These levels of agreement indicate that the 10% triple-annotated data are sufficiently reliable to serve as test set for our machine learning model (the *probabilistic test set*). The rest of the annotated data was used for training (80%) and validation (10%) with each data instance being annotated by a single author.

### 3.4 Probabilistic Analysis

To overcome the challenges of classifying location data collection in HTTP messages we developed a machine learning model.

**3.4.1 Machine Learning Baseline.** We began our model development by exploring a lightweight machine learning baseline. Using our data we trained SVM classifiers with bags-of-words [44]. We tuned the classifiers with the default set of hyperparameters [68]. They performed much better than our regular expressions (Table 2).

<sup>6</sup>Values for Krippendorff's alpha range from -1 to 1, where 1 means perfect agreement, -1 means perfect disagreement, and 0 means that agreement is equal to chance [18]. Generally, values above 0.8 are considered as good agreement, values between 0.67 and 0.8 as fair agreement, and values below 0.67 as dubious [51].

	TinyBERT Multi	BERT-Base Multi	Distilled Multi	TinyBERT Singles	BERT-Base Singles	Distilled Singles
<b>Accuracy</b>	0.94	0.94	0.92	0.94	<b>0.95</b>	0.92
<b>Precision</b>	0.94	0.94	0.92	0.94	<b>0.95</b>	0.92
<b>Recall</b>	<b>0.96</b>	0.93	0.88	0.95	0.94	0.91
<b>Area Under the Curve</b>	0.97	0.97	0.97	0.96	<b>0.98</b>	0.96
<b>F1</b>	0.94	0.94	0.91	0.94	<b>0.95</b>	0.91
<b>% F1 Imp vs RegEx</b>	49%	49%	44%	49%	<b>51%</b>	44%
<b>% F1 Imp vs SVM</b>	16%	16%	12%	16%	<b>17%</b>	12%
<b>Model Size</b>	<b>59MB</b>	450MB	450MB	5*59MB	5*450MB	5*450MB

**Table 4: Classification performance of various models running on the probabilistic test set containing instances of city, latitude, longitude, region, and ZIP code (support  $n = 533$ ). Singles are sets of models for classifying data for each location data category individually while Multi models classify data from all categories. The performance metrics are averaged across all five categories.**

	Support	Precision	Recall	F1
<b>City</b>	103	0.81	0.88	0.84
<b>Latitude</b>	107	0.92	0.97	0.94
<b>Longitude</b>	105	0.87	0.97	0.91
<b>Region</b>	108	1.00	1.00	1.00
<b>ZIP Code</b>	110	1.00	1.00	1.00
<b>Weighted Average</b>	533	0.92	0.96	0.94

**Table 5: Classification performance of the TinyBERT multitask model for the five location data categories running on the probabilistic test set.**

These baseline results demonstrate that machine learning classifiers are the right direction for improving the accuracy of identifying location data collection in HTTP messages. We set out to further increase the classification performance with a deep learning model. Figure 4 shows an overview of our model development and predictions at runtime.

**3.4.2 Identifying Candidate Instances.** Before applying the probabilistic analysis Privacy Pioneer first identifies candidate instances of location data collection based on regular expression matches. Only matched HTTP messages are considered for the probabilistic analysis while unmatched messages are discarded. Given the perfect recall (Table 2), it would be rare to miss positive instances. To provide sufficient context for our model we found that truncated messages with 250 characters before and after a regular expression match yield good results (§3.3.2). If a message was shorter to begin with, we padded it. Truncating and padding each message to a standard length before feeding it into our model also prevents length bias. To identify padding to the model we use an attention mask.

**3.4.3 Selecting a Pre-trained Model.** Our analysis is based on a pre-trained model. As a starting point, we selected the Bidirectional Encoder Representations from Transformers (BERT) family of models [17], specifically, BERT-Base, as implemented in Python via the Hugging Face [78] and PyTorch [59] libraries. BERT models are pre-trained on a large corpus of natural language data and can be further trained for domain-specific tasks. However, given its file size of 450MB, it proved challenging to integrate BERT-Base into Privacy Pioneer under the constraints of the browser environment. Thus, we explored TinyBERT [43], which, compared to BERT-Base, is smaller and faster with a file size of 59MB. Also, on average across

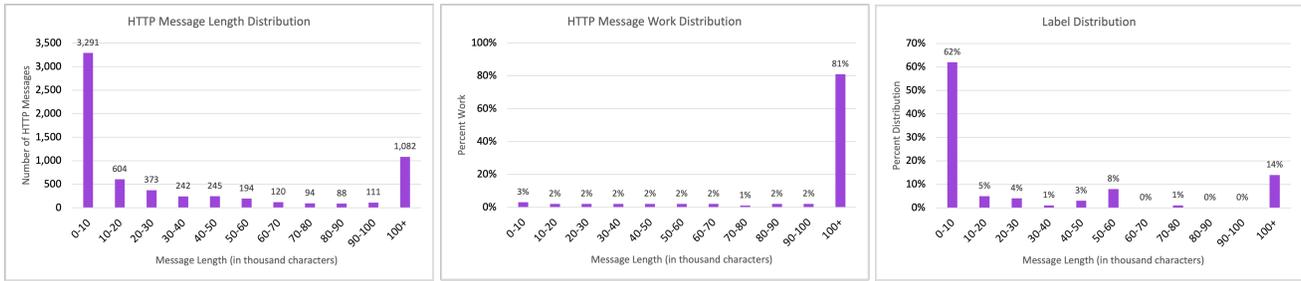
the five categories the TinyBERT model classifies an instance 12.6x faster on our data compared to the BERT-Base model.

**3.4.4 Tokenization.** For tokenization we use a WordPiece tokenizer. When we converted our model from Python to JavaScript for running it in the browser, we implemented the equivalent tokenizer [14] in Privacy Pioneer.

**3.4.5 Hyperparameter Tuning.** During model development we explored various hyperparameters. Using Hugging Face’s Trainer API and Weights & Biases [76], a hyperparameter optimization library, we found that the most impactful parameters for classification accuracy were batch size and learning rate. Our best performing model was trained for 50 epochs, with early stopping, using a batch size of 8, a learning rate of  $5 * 10^{-6}$ , and a weight decay of 0.1.

**3.4.6 Multitask Model.** To maximize model efficiency we explored using a multitask model for analyzing data of all five location data categories. This method has the advantage of reducing the number of models from five category-specific models to a single one that can analyze inputs of all categories. To evaluate this method we trained BERT-Base and TinyBERT models on a training set from all five categories. We added the category to the beginning of each training instance to indicate to the model which category of location data it is looking at. Table 4 shows the results of our evaluation.

**3.4.7 Classification Performance.** For the most part the Multi models perform on par or within a few percentage points difference compared to the Singles models, of which BERT-Base Singles perform the best. Both BERT-Base Multi and TinyBERT Multi exhibit similar performance with an average accuracy of 94% indicating that we can rely on the 7.6x smaller model for our classification tasks. TinyBERT Multi results in 49% and 16% relative F1 score improvement over the use of regular expression and the SVM baseline, respectively (Table 2). Comparing the SVM baseline to the TinyBERT Multi results (Table 5), we observe a substantial F1 score increase for city (0.52 to 0.84), region (0.71 to 1.00), and ZIP code (0.74 to 1.00). The performance improved less for latitude (0.92 to 0.94) and stayed the same for longitude (0.91). We believe that the performance improvements for city, region, and ZIP code are largely based on our model’s ability of disambiguating the context in which they occur, e.g., whether a city occurs in a news article or designates the user’s location. Context seems to matter less for identifying the more distinct formats of latitudes and longitudes, whose regular expression-based F1 scores were already higher with



**Figure 5: We collected the HTTP messages from a random set of 50 sites from the Tranco list. We found a number of messages with over 100,000 characters (left). Those created substantial amount of work, defined as a percentage of the total characters that Privacy Pioneer would be searching through (middle). However, they only exhibited few instances of data collection, i.e., privacy labels created (right).**

0.77 and 0.73, respectively (Table 2). Overall, we believe that the application of machine learning models will have its greatest impact for the identification of generic data categories that do not have a distinctive format and for which, consequently, only their context reveals their purpose of use.

**3.4.8 Knowledge Distillation.** We tried improving the classification performance of our TinyBERT Multi model via knowledge distillation [37]. To that end, we used the BERT-Base models, trained on the 5,472 annotated instances to programmatically annotate all remaining unannotated data leading to 98,643 annotated instances. This significantly larger set of annotated data was then fed as training data into a fresh set of models with TinyBERT as the pre-trained model. However, given the already close performance between the teacher and the student models when trained directly, distillation did not improve classification performance overall. Thus, we kept the TinyBERT Multi model, trained on 5,472 annotated instances, as our final model.

**3.4.9 Model Integration.** As shown in Figure 4, to integrate our model into Privacy Pioneer we converted it from PyTorch in Python to TensorFlow in Python [1]. Then, we used `tfjs-converter` to convert it from TensorFlow in Python to TensorFlow in JavaScript for use in TensorFlow.js [70], a JavaScript library with a set of APIs for running TensorFlow models in the browser or server-side.<sup>7</sup> Upon installation of Privacy Pioneer, our model is served from GitHub and downloaded to an IndexedDB instance in the browser.<sup>8</sup> This approach enables efficient usage of the model at runtime as it is immediately available locally at all times and across all browsing sessions. Also, the use of an IndexedDB instance helps ensure user privacy by storing the model and all analyzed data locally.

<sup>7</sup>We noticed that this conversion resulted in a decrease of 4.7% points on average for precision, recall, and F1 score. We used Google’s official libraries for the conversion and confirmed with Google that our conversion methodology was correct. We have opened an issue on Google’s repository at <https://github.com/tensorflow/tfjs/issues/8025>. Appendix 8.2 shows the classification performance of the TensorFlow model in JavaScript. The model conversion from PyTorch in Python to TensorFlow in Python did not lead to any discrepancies. The model performance is identical and shown in Table 5. The discrepancies do not change our ranking of the different models but highlight the practical challenges of integrating machine learning models in the browser.

<sup>8</sup>The model is served from <https://github.com/privacy-tech-lab/privacy-pioneer-machine-learning>.

### 3.5 Computational Performance

Analyzing HTTP messages dynamically at runtime can decrease computational performance and impact usability. We implemented various heuristics to reduce the analysis workload by filtering out messages and message parts that are likely irrelevant for detecting data collection practices:

- Do not analyze messages exceeding 100,000 characters
- Only analyze the following `webRequest.ResourceTypes`:
  - `image` (e.g., used for tracking pixels)
  - `script` (e.g., used for browser fingerprinting scripts)
  - `sub_frame` (e.g., `iframes` for loading third party sites)
  - `xmlhttprequest` (can contain any type of data)
- Only analyze request body, response body, and selected headers as those can contain user-specific data (§ 3.1.2)

Applying these heuristics resulted in substantially decreased workloads with minimal information loss (Figure 5). Comparing Privacy Pioneer’s runs with and without these heuristics via Apple Activity Monitor showed a decrease of Firefox’s WebExtension CPU usage by an average of 52%, from 12.96% to 6.26%, across three runs. The performance evaluation was run on a 2023 MacBook Pro with an Apple M2 Pro processor, 16 GB RAM, and with no user programs running besides Firefox and Activity Monitor.

To evaluate the performance cost of adding Privacy Pioneer to Firefox we randomly sampled 50 sites from the Tranco list, visited the sites with and without Privacy Pioneer turned on, repeated the process for a total of three runs, and then measured the time to load a site using Firefox’s Network Monitor, which has a load variable that records when a resource finished loading [30]. The average time to load a site with Privacy Pioneer was 2.09 seconds while the average time to load a site without it was 1.93, thus, adding 0.16 seconds for an 8% increase. We find the additional load time tolerable given the transparency gain. The performance evaluation was run on a 2019 MacBook Pro with 1.4 GHz Quad-Core Intel Core i5 processor, 16 GB RAM, and with no user programs running besides Firefox.

## 4 USABILITY OF PRIVACY PIONEER

We tested the privacy interfaces we designed and implemented in Privacy Pioneer in an online usability study.<sup>9</sup> We structured our survey questions around our core inquiry of web transparency

<sup>9</sup>Screenshots of the privacy interfaces are shown in Appendix 8.3.

Age Range		Sex		Race/Ethnicity		Student		Employment		Browser		Operating System	
18-24	8%	Male	66%	White	69%	Yes	12%	Full-Time	59%	Chrome	62%	Windows	81%
25-34	29%	Female	34%	Black	12%	No	88%	Part-Time	15%	Firefox	23%	macOS	18%
35-44	40%			Mixed	9%			Unemployed	15%	Brave	7%	Linux	1%
45-54	15%			Asian	6%			Unpaid work	8%	Safari	5%		
55-64	5%			Other	3%			Other	3%	Edge	2%		
≥ 65	3%									Vivaldi	1%		

**Table 6: Participant demographics. 100 participants completed our study. Some did not provide data for all categories: 2 for Age Range, 0 for Sex, 2 for Race/Ethnicity, 11 for Student status, 13 for Employment status, 0 for Browser, and 0 for Operating System. Percentages are adjusted for any omissions. All data is from Prolific except for Browser and Operating System, which we asked participants to provide in our survey.**

with the specific goals of determining the comprehensibility and usefulness of the privacy interfaces in Privacy Pioneer. As part of these goals we seek to inform the priorities and opportunities for the future design of privacy interfaces.

## 4.1 Experimental Setup

We recruited participants for our study on the crowdworking platform Prolific [61].

**4.1.1 Eligibility Criteria.** The eligibility criteria to participate in our study were: (1) having installed or willingness to install Firefox on a laptop or desktop computer, (2) fluency in English, (3) United States residency, (4) 100% approval rate for previous tasks on Prolific, (5) completion of at least 50 previous tasks on Prolific, and (6) a minimum age of 18 years. As Privacy Pioneer is only available on Firefox its use as part of the study was mandatory. For the criteria (2) – (6) we relied on the information provided by Prolific.

**4.1.2 Study Procedure.** We signed up a total of 100 participants. Participants first answered a few general web privacy and technology questions (Q2–Q12) and then engaged in three guided tasks with our browser extension, after which they shared their experience (Q13–Q47).<sup>10</sup> Our survey contained one attention check question, which all participants answered correctly. After participants had installed our browser extension from the Firefox Add-ons store, we asked them to perform the three tasks.

The first task asked participants to identify trackers on a website via the privacy popup, then change their Firefox settings to block them, and finally confirm via the privacy popup that the trackers are indeed no longer active. The second task asked participants to identify trackers across sites via the privacy history interface. Specifically, we gave them a set of five different sites. After visiting each site they should check their privacy history to identify the tracker that was active on multiple sites. The third task asked participants to enable notifications via the privacy watchlist to be fired when a visited site shares a custom keyword with another site.

To ensure that participants actually performed the tasks we asked them to send us screenshots and a file that logged their interface interactions, e.g., which Privacy Pioneer buttons they pressed and which sites they visited. For each participant we created a personalized upload link to an anonymous cloud drive that we shared via the Prolific messaging system, which we also used to troubleshoot technical problems or clarify questions that participants had. After

finishing their tasks and answering the survey questions participants submitted their work and were given a completion code they could enter on Prolific to receive their compensation.

**4.1.3 Ethical Considerations.** We received IRB approval for our study. At sign-up we let participants know who we are and how they could contact us. We explained the study purpose, i.e., to find out how websites’ data collection and sharing practices can be made more transparent for web users. We further explained that the study consists of (1) installing and using our Firefox browser extension and (2) answering survey questions and submitting files about their usage and data privacy in general. We provided them a list of the categories of data that we would request from them.<sup>11</sup> We explained that the data would be stored at our organizations and our service providers using current best practices, that it would not be disclosed except in aggregate form, and that we would retain a copy after the study for record-keeping purposes. During the study, data was only stored locally on participants’ computers. We then asked them to submit their data via an anonymous cloud drive using their Prolific ID, a pseudonym by which we identified participants. We explained to them that they could view and delete any data before their submission and that they could withdraw from the study at any time. At sign-up, we also explained to participants that the IP-to-location service IPInfo will receive their IP address. They had to manually enable this functionality in the extension, at which time they were notified once more. For their participation we paid each participant \$9, the amount recommended by Prolific for a study like ours.

## 4.2 Sample Representativeness

Our study sample is partially representative of the US population (Table 6). We compared our participants’ demographics to the US population demographics derived from the 2021 American Community Survey[8–10]. The age distribution of our participants aligns with the expected figures (sample median: 38.0 years; population median: 38.8 years [16]). We also find our sample to be representative with regards to ethnicity. However, we note disparities in terms of sex, student status, and employment status. Our sample has more male (sample: 66%) than female participants. Our study also attracted fewer students compared to the national proportion (sample: 12%; population: 25% [10]), which may be attributable to our study’s exclusion of minors, who constitute a large portion of the US student body. As to employment, we note a large proportion

<sup>10</sup>The set of survey questions and tasks is shown in Appendix 8.4.

<sup>11</sup>The list of data categories collected from participants is shown in Appendix 8.5.

of unemployed individuals (sample: 15%; population: 4% [9]), likely due to Prolific being a paid crowdworking platform.

We asked participants which operating system and web browser they primarily use. Utilizing market share statistics for US desktop users, we find that our sample includes a disproportionately high percentage of Windows users (sample: 81%; population: 59% [72]) and a corresponding lower percentage of macOS users (sample: 18%; population: 32% [72]). The browser distribution shows a substantially higher Firefox usage rate (sample: 23%; population: 5% [71]) while featuring very few Safari users (sample: 5%; population: 21% [71]). The general under-representation of Apple users with regards to both browsers and OS, as well as the over-representation of Firefox users, may well be a consequence of the conditions of our study, which required participants to make use of Privacy Pioneer as a Firefox-exclusive extension.

As indicated by the relatively higher Firefox usage rate, our sample skews towards people with an interest in protecting their privacy. Also, as the self-rating of tech-savviness indicates – with 59% of the participants believing that they are either tech-savvy or very tech-savvy (Figure 6) – our sample seems to skew towards advanced web users. These trends are also confirmed by 76% of participants answering that they are using some form of privacy software (Q8).<sup>12</sup> On the other hand, there are 24% reporting to not use any privacy software and 41% who do not consider themselves particularly tech-savvy. Thus, we also have a contingent of participants who may be less concerned about their privacy or who need help in understanding how their data is being collected.

### 4.3 Usability Evaluation

The results of our usability study suggest that privacy interfaces that show which data is collected by whom can help people understand websites’ data collection practices. Our results further suggest that such interfaces should be directly integrated into the web browser in an easy-to-use, informative, and actionable form.<sup>13</sup>

**4.3.1 Insufficiency of Privacy Protection and Data Transparency.** A majority of participants expressed concern about their privacy on the web. 60% disagreed or strongly disagreed with the statement “I generally feel confident that my privacy is protected on the web” (Q2) (Figure 7). However, a number of participants felt they have measures at their disposal to protect their privacy (Q3) (Figure 7). 41% disagreed or strongly disagreed with the statement “There is not much I can do to protect my privacy on the web.” This view,

<sup>12</sup>As we asked participants to submit screenshots of the interfaces we presented to them, we have an indication that their use of such software, if any, did not interfere with their use of Privacy Pioneer.

<sup>13</sup>The findings in this section are based on summary statistics from the data collected in our usability study. We evaluate correlations using the Kendall Tau coefficient. The correlations are derived from 22 survey questions (Appendix 8.4). We compared the answer distributions for each question pair for a total of 231 pairwise comparisons. The set of pairwise comparisons includes all *linear scale* questions (per Appendix 8.4), except for the attention check question, Q29. We treat Q6 (“Do you care whether a website shares the following of your data for ad purposes?”) as 8 separate questions, one for each of the 7 data categories and one that sums the other 7 for each participant. Also included in the pairwise comparisons is the *multiple choice question* Q45 (“How likely is it that you would recommend Privacy Pioneer to a friend or colleague?”). Of all comparisons, 59 were significant ( $p < 0.05$  corrected for multiple tests with the Benjamini-Yekutieli procedure) and 53 additionally had a correlation coefficient higher than 0.3, which we consider as the minimum value to indicate a moderate correlation. For testing the goodness of fit of non-ordinal data we use the Chi-square test.

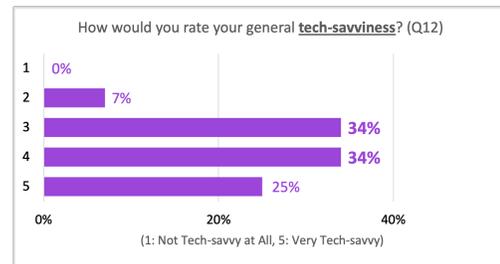


Figure 6: On a scale of 1 (not tech-savvy at all) to 5 (very tech-savvy), participants overall gravitate towards higher tech-savviness.

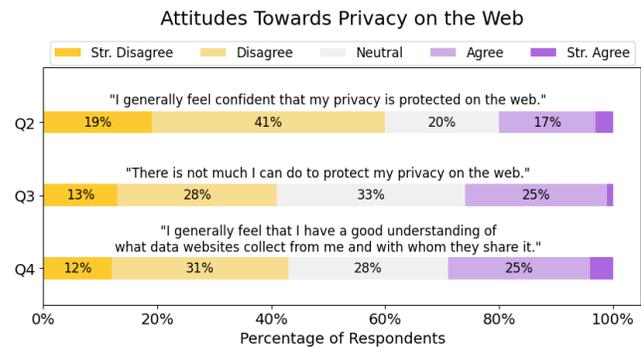
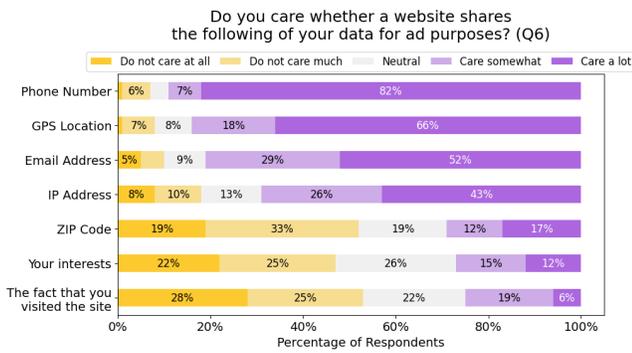


Figure 7: Participants’ attitudes towards privacy on the web based on responses to survey questions Q2, Q3, and Q4. Most participants did not feel confident that their privacy is protected on the web (Q2). Participants were also generally hesitant to claim a good understanding of data collection and sharing practices (Q4). Many, however, felt that they can do something to protect their privacy (Q3).

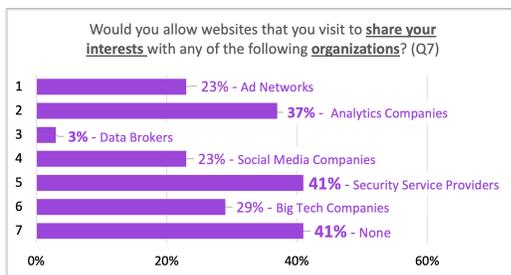
however, is contingent on whether or not participants reported using privacy software (Q8).<sup>14</sup> Only 27% of participants who reported not using privacy software disagreed or strongly disagreed with the statement in Q2, that is, felt able to protect their privacy. The largest proportion of such participants, 41%, answered “Neutral” suggesting a potential lack of knowledge on how well their privacy is protected on the web. Indeed, 43% of participants overall disagreed or strongly disagreed with the statement “I generally feel that I have a good understanding of what data websites collect from me and with whom they share it” (Q4) (Figure 7). These results point to a lack of data transparency on the web. They are particularly noteworthy as 59% participants in our study rate themselves as tech-savvy or very tech-savvy (Q12) (Figure 6).

**4.3.2 Data and Recipient Categories Matter.** It is the purpose of our interfaces to provide people with specific, yet, comprehensible, information of *who* is collecting *which* categories of data. We note a clear imperative for providing such granular detail. We asked participants about different categories of personal data being shared for ad purposes, and they conveyed their opinions towards each as one of five tiers of caring, ranging from “Do not care at all” to “Care a

<sup>14</sup>Chi-square Goodness of Fit test,  $p = 0.021$ , comparing observed frequencies of Q3 of participants who responded with “None” when asked about the privacy software they used (Q8) to those who responded otherwise.



**Figure 8: Participants’ data sharing preferences for ad purposes broken down by data category. To which extent participants’ care depends on the category of data being shared.**

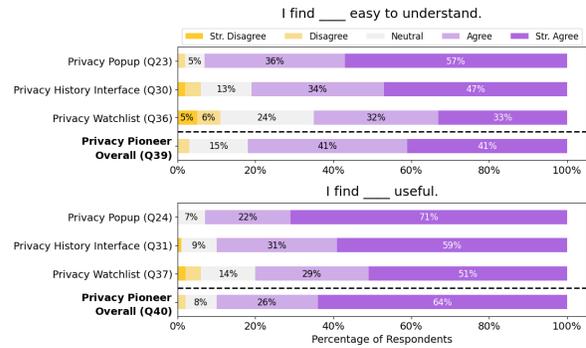


**Figure 9: Participants’ preferences for sharing interest data is organization-dependent. We asked participants to select all that applies or “None.”**

lot” (Q6) (Figure 8). 75% of participants used at least 3 tiers of caring across the 7 different data categories we listed, thus, indicating that sentiments varied between different categories for many. Likewise, categories such as participants’ interests showed little consensus on whether or not sharing mattered: 27% of participants cared to some degree, 47% did not, and 26% expressed a neutral view.

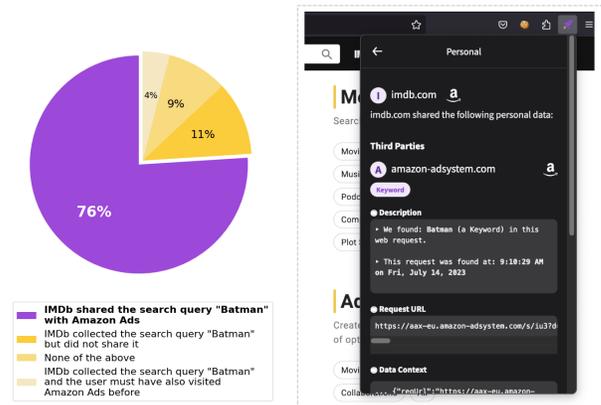
A similarly broad spectrum of views exists for the categories of data recipients. While 41% of participants were opposed to all sharing of interest data, 59% had more granular preferences (Figure 9). Participants’ nuanced privacy preferences could ultimately provide a new avenue for ad personalization (§5.3). When asked if they would use an ad/tracking blocker capable of selectively allowing certain sites to receive certain data (Q9), 59% of participants responded “I would use it if it would keep the sites free,” while another 28% said they would continue using some existing ad/tracking blocker. Participants, thus, held diverse privacy preferences regarding different organizations and data categories, especially, when they had to consider paying for content.

**4.3.3 Privacy Interfaces Have Promise to Help People Understand Who Collects Which Data.** Participants’ perceptions of the utility and clarity of the three privacy interfaces were broadly favorable (Figure 10). Overall, participants expressed agreement or strong agreement that the popup was the easiest to understand (93%), followed by the history interface (81%), and then the watchlist (65%). Most participants ranked the usefulness of the interfaces in the same



**Figure 10: The degree to which participants found the privacy interfaces they encountered easy to understand (top) and useful (bottom) as well as their respective overall ratings of Privacy Pioneer.**

What is shown in the screenshot below? (Q34)



**Figure 11: Participants were asked to enter the keyword “Batman” into their watchlist, search on imdb.com for “Batman,” and check the popup to identify any sharing of the keyword. To confirm their understanding, the survey showed a popup (right), for which 76% of the participants selected the correct answer (left).**

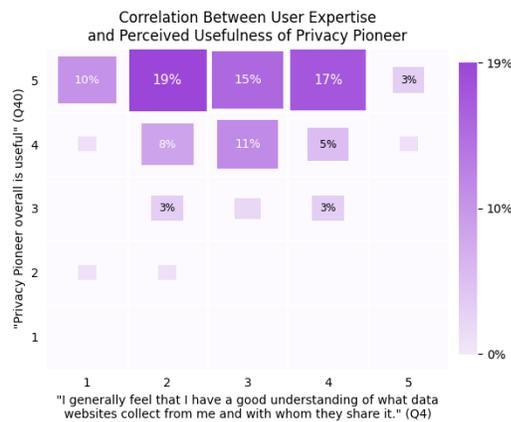
order with 93%, 90%, and 80%, respectively, agreeing or strongly agreeing. These results are not entirely independent of participants’ reported tech-savviness. For the privacy history interface we note a statistically significant and moderate correlation between how participants rated their tech-savviness and the degree to which they found the interface useful and understandable.<sup>15</sup> The popup showed a marginally weaker correlation.<sup>16</sup> The watchlist correlation was not statistically significant.<sup>17</sup> These results suggest that less tech-savvy participants struggled more with Privacy Pioneer’s interfaces while overall impressions remain favorable.

When asked to enter the keyword “Batman” in the watchlist and identify whether it was shared per the third guided task, 76% of

<sup>15</sup>Q12 (tech-savviness) vs. Q30 (comprehensibility of the history interface): Kendall Tau test,  $\tau$  coefficient = 0.302,  $p = 0.015$ . Q12 (tech-savviness) vs. Q31 (usefulness of the history interface): Kendall Tau test,  $\tau$  coefficient = 0.305,  $p = 0.018$ .

<sup>16</sup>Q12 (tech-savviness) vs. Q23 (comprehensibility of the popup): Kendall Tau test,  $\tau$  coefficient = 0.274,  $p = 0.057$  (barely insignificant). Q12 (tech-savviness) vs. Q24 (usefulness of the popup): Kendall Tau test,  $\tau$  coefficient = 0.297,  $p = 0.026$ .

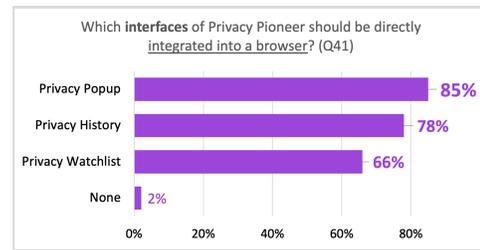
<sup>17</sup> $p > 0.05$  for both usefulness and understanding.



**Figure 12: 90% of participants rated Privacy Pioneer overall as useful independently of their understanding of data collection and sharing practices. There was no significant correlation between the level of understanding of privacy practices and the perceived utility of Privacy Pioneer (Kendall Tau test,  $\tau$  coefficient = -0.019,  $p = 1.000$ ). However, there was a correlation between understanding Privacy Pioneer (Q39) and recommending it to friends or colleagues (Q45) (Appendix 8.6, Figure 22).**

the participants picked the correct answer (Figure 11). The relatively weaker reception of the watchlist compared to the popup and privacy history may have been influenced by participants being expected to track a dummy keyword, which may have left them wondering about the purpose of the watchlist. The lesser understanding of the watchlist’s purpose — as opposed to its functionality — could be a reason for the low correlation between Q12 (tech-savviness) and Q36 (comprehensibility of the watchlist interface) and Q37 (usefulness of the watchlist interface), respectively. However, overall, Privacy Pioneer was regarded by a majority of participants as useful with 90% agreeing or strongly agreeing regardless of how they rated their pre-existing understanding about website data collection and sharing (Figure 12).

**4.3.4 Privacy Interfaces Should Be Directly Integrated in the Browser.** Most participants found it desirable to have the tested privacy interfaces directly integrated in the browser. The popup received the highest rate of approval (85%) (Figure 13). 17% of participants expressed that they would continue using Privacy Pioneer on Firefox (11% expressed that they would not) (Q44). 65% of participants expressed that they would continue using Privacy Pioneer if it were available for their main browser (7% expressed that they would not). Across questions, a small number of participants conveyed dissatisfaction over the performance of Privacy Pioneer’s interfaces, particularly, as to the loading time of the popup. 8% of participants reported technical issues during the study, some of which pertained to the study as opposed to Privacy Pioneer (Q42). When asked about improvement suggestions, 5% of participants’ focused on improvements relating to performance or tediousness. Overall, a sizeable majority of participants had a smooth experience with Privacy Pioneer indicating that privacy analyses like ours can be handled by the browser.



**Figure 13: The percentage of participants that recommended each interface be directly integrated into the browser. We asked participants to select all that applies.**

**4.3.5 Privacy Interfaces Should Be Easy to Use, Informative, and Actionable.** We asked participants to share what they liked about our privacy interfaces and what improvements they would suggest.<sup>18</sup> On what they liked, 52% responded they found the interfaces easy to use and understand. 50% valued the information shown. 20% expressed excitement and had no improvement suggestions. Still, some participants suggested to simplify language (18%), create tutorials and make functionality more clear (15%), and provide interface improvements (15%). Participants were particularly in favor of the inclusion of better usage instructions, such as a video tutorial. Some participants also would have wanted a better explanation of the trackers and what can be done with the information (10%) or added features to protect privacy (5%). This finding indicates that people want both transparency and control, that is, both notice and choice.

## 5 DISCUSSION

We developed Privacy Pioneer’s privacy interfaces to help improving data transparency in the web ecosystem.

### 5.1 Automating Notice for More Transparency

Our interfaces are intended to help people becoming aware of the data collection practices they are subjected to. Many privacy laws are based on notice and choice. However, in its current form notice and choice has proven to be ineffective. Automating notice in the browser holds the promise of making websites’ privacy practices more transparent. Our results suggest that browser-driven dynamic analysis of websites’ data collection practices is technically feasible with good accuracy and tolerable computational effort (§3). Among the interfaces we studied, the privacy popup (Figure 2), which displays data collection practices of the currently visited site, was perceived as the most useful and easiest to understand with 93% of participants agreeing or strongly agreeing (Figure 10). These levels of agreement compare favorably to the general perception of privacy policies as bloated and unreadable [53]. Since the generated notices are backwards-traceable to the dynamically analyzed code [83], the identified practices can be precise, up-to-date, and shown in context. Participants’ general preference for the popup indicates that privacy notices should be made easily accessible without requiring people to navigate to a hard to find privacy settings page or performing multiple clicks to access a privacy interface.

<sup>18</sup>The full breakdown of responses to both questions — “What do you like about Privacy Pioneer” (Q46) and “How could we improve Privacy Pioneer” (Q47) — can be found in Appendix 8.6, Figure 23.

## 5.2 Enabling People to Make Privacy Choices

Transparency of data collection practices is a necessary prerequisite for enabling people to act on these practices and exercise their privacy rights. Informed choice demands notifications that are clear, comprehensive, and actionable. 43% of study participants stated that they did not have a good understanding of the data collection and sharing practices they are subjected to on the web (Figure 7). Indeed, many people lack awareness about who is collecting which data from them. If people are not aware, they have no reason to act. Therefore, transparency is important. More transparency would not only enable people to act but also strengthen the validity of their choices. They could exercise their privacy rights more intentionally. Otherwise, site owners may try to argue that people’s unawareness of what they are declaring, for example, by sending opt out signals [82, 85], allows them to ignore people’s requests as legally irrelevant. This argument should be preempted. Surfacing who is collecting which data may also have the side effect of motivating site owners to implement good privacy practices as their sites’ behaviors become more visible and, thus, subject to regulatory scrutiny. In this context, regulators can use Privacy Pioneer to identify data collection practices of sites they may want to investigate.

## 5.3 Personalizing Ads and Preserving Privacy

Participants expressed a variety of preferences as to which data categories they would be willing to share and the categories of organizations that could receive the data (§4.3.2). 53% of participants did not care much or at all if their visit of a website would be shared for ad purposes; the same is true for 52% as to their ZIP code locations and for 47% as to their interests (Figure 8). 29% of participants would allow the sharing of interests with big tech companies, 23% with social media companies, and 23% with ad networks (Figure 9). These results suggest that making people aware of applicable privacy practices and giving them the choice to opt out does not necessarily mean the end of all personalized advertising. Opt outs could be selective based on data categories and organizations receiving the data. As participants were generally able to effectively navigate Privacy Pioneer’s interfaces and to understand them (§4), such a selective choice seems feasible. Overall, there is a place for personalized ads to the extent that any underlying data usage is transparent, with usable choice, and performed in a privacy-preserving way.

## 6 LIMITATIONS

Our privacy analysis methods and interfaces in Privacy Pioneer (§3) as well as our usability study (§4) are subject to various limitations:

- *Browser APIs:* The browser APIs used by Privacy Pioneer (§3.1.2) are not available in all browsers, e.g. the API necessary for capturing HTTP response data, `webRequest.filterResponseData`, is only available in Firefox. However, this limitation would not apply if browser vendors would integrate the proposed functionality directly in the browser, which is ultimately our goal.
- *Encoded Data:* While browser APIs allows us to capture unencrypted data, some data may be encoded. As a proof of concept we decode Hexcode SHA-256 and Base64 SHA-256 email address formats. For a comprehensive coverage, more encodings for more data categories would be necessary.

- *Deterministic Analysis Methods:* Privacy Pioneer’s deterministic analysis (§3.2.1) is limited by its rule-based nature and manual curation. E.g., URL list matching for identifying ad networks depends on the Disconnect Tracker Protection lists. Ad networks incorrectly added to the lists would be flagged while incorrectly omitted ones would not be.
- *False Positives and Negatives:* Both Privacy Pioneer’s deterministic and probabilistic analyses may lead to false positives and negatives (Tables 1, 2, and 5). To enable the identification of false positives Privacy Pioneer provides context in form of HTTP message snippets on which analysis results are based (Figure 2). The tradeoff between precision and recall is adjustable via hyperparameter tuning (§3.4.5). Before relying on results, e.g., for regulatory enforcement actions, they should be verified manually.
- *IP Address Disclosure to IPinfo:* To alleviate privacy and security concerns Privacy Pioneer processes all data locally except for the user’s IP address, which is sent to IPinfo (§3.1.4). Privacy Pioneer displays a notification about the IP address disclosure. IPinfo told us that they store no data beyond the IP address and the number of times it made a request. They also said that the data is kept for one year and neither used by IPinfo nor shared with any third party. It would be possible to implement such IP-to-location API from scratch [39], which is not our focus here.
- *Location Dataset Creation:* We created our location dataset (§3.3) by crawling the homepages of websites. Thus, it does not contain data from non-homepage pages. We further associated a site with a country based on its top-level country domain, which is only an approximation.
- *Interaction with Other Software:* Depending on the use of ad-tracking blockers, VPNs, and other software, Privacy Pioneer’s analysis results may differ. However, we are not aware of any interaction of Privacy Pioneer with any other software that would break Privacy Pioneer or such software.
- *Self-reporting and Positive Framing of Survey Questions:* Answers to the questions of our usability study (§4) should be interpreted in light of their nature as being self-reported. Also, for questions asking participants whether they agree with a statement, people generally feel more enticed to agree than to disagree.

## 7 CONCLUSIONS

Privacy Pioneer’s analysis methods and privacy interfaces demonstrate that the analysis of websites’ data collection practices can be accurately performed in the browser to surface dynamic privacy notifications. If people understand which data is being collected from them and by whom, they can use technical measures to better protect their data and meaningfully exercise their privacy rights. The increased transparency could also be a motivating factor for website operators to improve their privacy practices. It is our goal to make instruments of notice and choice more usable. Our usability study primarily evaluated study participants’ first impressions and whether they could effectively navigate the interfaces we presented to them. Further exploration, such as a longer and non-directed usability study, is needed to determine how people would engage with the analysis results and interfaces in real life. It would also be interesting to analyze websites’ data collection practices broadly across a set of sites and over time.



- NIPS 2014 Deep Learning Workshop.
- [38] Arnaud Legout Imane Fouad, Natalia Bielova and Natasa Sarafijanovic-Djukic. 2020. Missed by Filter Lists: Detecting Unknown Third-Party Trackers with Invisible Pixels. In *POPEETS '20*. Scienco, Montreal, Canada, 220–238.
- [39] ipapi.is. 2023. Geolocation. <https://ipapi.is/geolocation.html>. Accessed: April 4, 2024.
- [40] IPinfo. 2023. IPinfo. <https://ipinfo.io/>. Accessed: April 4, 2024.
- [41] ipstack. 2023. ipstack. <https://ipstack.com/>. Accessed: April 4, 2024.
- [42] Umar Iqbal, Steven Englehardt, and Zubair Shafiq. 2021. Fingerprinting the Fingerprinters: Learning to Detect Browser Fingerprinting Behaviors. In *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, Los Alamitos, CA, 1143–1161.
- [43] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for Natural Language Understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 4163–4174. <https://doi.org/10.18653/v1/2020.findings-emnlp.372>
- [44] Rajat Katiyar. 2018. Bag Of Words With SVM. <https://www.kaggle.com/code/rajatkatiyar/bag-of-words-with-svm>. Accessed: April 4, 2024.
- [45] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W. Reeder. 2009. A “nutrition label” for privacy. In *SOUPS (Mountain View, California)*. ACM, New York, NY, USA, Article 4, 12 pages. <https://doi.org/10.1145/1572532.1572538>
- [46] Agnieszka Kitkowska, Mark Warner, Yefim Shulman, Erik Wästlund, and Leonardo A. Martucci. 2020. Enhancing Privacy through the Visual Design of Privacy Notices: Exploring the Interplay of Curiosity, Control and Affect. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*. USENIX Association, Berkeley, CA, USA, 437–456. <https://www.usenix.org/conference/soups2020/presentation/kitkowska>
- [47] Simon Koch, Malte Wessels, Benjamin Altpeter, Madita Olvermann, and Martin Johns. 2022. Keeping Privacy Labels Honest. In *Proceedings on Privacy Enhancing Technologies Symposium (PoPETS 2022)*. PoPETS 2022, Sydney, Australia and Virtual, 486–506.
- [48] Konrad Kollnig, Anastasia Shuba, Max Van Kleek, Reuben Binns, and Nigel Shadbolt. 2022. Goodbye Tracking? Impact of IOS App Tracking Transparency and Privacy Labels. In *FACCT '22 (Seoul, Republic of Korea)*. ACM, New York, NY, USA, 508–520. <https://doi.org/10.1145/3531146.3533116>
- [49] Stefan Larsson, Anders Jensen-Urstad, and Fredrik Heintz. 2021. Notified But Unaware: Third-Party Tracking Online. <https://doi.org/10.33137/cal.v8i1.36282>. *Critical Analysis of Law: An International & Interdisciplinary Law Review* 8, 1 (2021), 101–120.
- [50] Asha S. Manek, P. Deepa Shenoy, M. Chandra Mohan, and K. R. Venugopal. 2016. Detection of Fraudulent and Malicious Websites by Analysing User Reviews for Online Shopping Websites. *International Journal of Knowledge and Web Intelligence (IJKWI)* 5, 3 (jan 2016), 171–189. <https://doi.org/10.1504/IJKWI.2016.078712>
- [51] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.
- [52] Surya Mattu and Aaron Sankin. 2020. How We Built a Real-time Privacy Inspector. <https://themarkup.org/blacklight/2020/09/22/how-we-built-a-real-time-privacy-inspector>. Accessed: April 4, 2024.
- [53] Aleecia M. McDonald and Lorrie F. Cranor. 2008. The Cost of Reading Privacy Policies. *I/S: A Journal of Law and Policy for the Information Society* 4, 3 (2008), 540–565.
- [54] Mozilla. 2019. Firefox Lightbeam. <https://github.com/mozilla/lightbeam-we>. Accessed: April 4, 2024.
- [55] Mozilla. 2023. Does Firefox share my location with websites? <https://support.mozilla.org/en-US/kb/does-firefox-share-my-location-websites>. Accessed: April 4, 2024.
- [56] Mozilla. 2023. Enhanced Tracking Protection in Firefox for desktop. <https://support.mozilla.org/en-US/kb/enhanced-tracking-protection-firefox-desktop>. Accessed: April 4, 2024.
- [57] Nick Nikiforakis, Alexandros Kapravelos, Wouter Joosen, Christopher Kruegel, Frank Piessens, and Giovanni Vigna. 2013. Cookieless monster: Exploring the ecosystem of web-based device fingerprinting. In *S&P (S&P)*. IEEE, S. Francisco, CA, 541–555. [http://ieeexplore.ieee.org/xpls/abs\\_all.jsp?arnumber=6547132](http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6547132)
- [58] OECD. 2013. Recommendation of the Council concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data. <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0188>. Accessed: April 4, 2024.
- [59] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA, Article 721, 12 pages.
- [60] Victor Le Pochat, Tom van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczynski, and Wouter Joosen. 2019. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. [https://www.ndss-symposium.org/wp-content/uploads/2019/02/ndss2019\\_01B-3\\_LePochat\\_paper.pdf](https://www.ndss-symposium.org/wp-content/uploads/2019/02/ndss2019_01B-3_LePochat_paper.pdf). In *NDSS*. Internet Society, VA, USA, 1–15.
- [61] Prolific. 2022. Quickly find research participants you can trust. <https://www.prolific.co/>. Accessed: April 4, 2024.
- [62] Abdullah Qasem, Sami Zhioua, and Karima Makhlof. 2019. Finding a Needle in a Haystack: The Traffic Analysis Version. *Proceedings on Privacy Enhancing Technologies* 2019 (04 2019), 270–290. <https://doi.org/10.2478/popets-2019-0030>
- [63] Alexandr Railean and Delphine Reinhardt. 2021. OnLITE: On-line Label for IoT Transparency Enhancement. In *Secure IT Systems*, Mikael Asplund and Simin Nadjim-Tehrani (Eds.). Springer International Publishing, Cham, 229–245.
- [64] Ashwini Rao, Florian Schaub, Norman Sadeh, Alessandro Acquisti, and Ruogu Kang. 2016. Expecting the Unexpected: Understanding Mismatched Privacy Expectations Online. In *SOUPS*. USENIX Association, Denver, CO, 77–96. <https://www.usenix.org/conference/soups2016/technical-sessions/presentation/rao>
- [65] Philip Raschke, Axel Küpper, Olha Drozd, and Sabrina Kirrane. 2018. *Designing a GDPR-Compliant and Usable Privacy Dashboard*. Springer International Publishing, Cham, 221–236. [https://doi.org/10.1007/978-3-319-92925-5\\_14](https://doi.org/10.1007/978-3-319-92925-5_14)
- [66] Jukka Ruohonen and Ville Leppänen. 2018. Invisible Pixels Are Dead, Long Live Invisible Pixels!. In *Proceedings of the 2018 Workshop on Privacy in the Electronic Society (Toronto, Canada) (WPES'18)*. Association for Computing Machinery, New York, NY, USA, 28–32. <https://doi.org/10.1145/3267323.3268950>
- [67] Florian Schaub, Rebecca Balebako, Adam L. Durity, and Lorrie Faith Cranor. 2015. A Design Space for Effective Privacy Notices. In *SOUPS*. USENIX Assoc., Ottawa, 1–17. <https://www.usenix.org/conference/soups2015/proceedings/presentation/schaub>
- [68] scikit-learn developers. 2023. sklearn.svm.SVC. <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html>. Accessed: April 4, 2024.
- [69] Asuman Senol, Gunes Acar, Mathias Humbert, and Frederik Zuiderveen Borgesius. 2022. Leaky Forms: A Study of Email and Password Exfiltration Before Form Submission. In *31st USENIX Security Symposium (USENIX Security 22)*. USENIX Association, Boston, MA, 1813–1830. <https://www.usenix.org/conference/usenixsecurity22/presentation/senol>
- [70] Daniel Smilkov, Nikhil Thorat, Yannick Assogba, Ann Yuan, Nick Kreeger, Ping Yu, Kangyi Zhang, Shanqing Cai, Eric Nielsen, David Soergel, Stan Bileschi, Michael Terry, Charles Nicholson, Sandeep N. Gupta, Sarah Sirajuddin, D. Sculley, Rajat Monga, Greg Corrado, Fernanda B. Viegas, and Martin Wattenberg. 2019. TensorFlow.js: Machine Learning for the Web and Beyond. In *Proceedings of Machine Learning and Systems*. MLSys, Palo Alto, CA, USA, 309–321. [https://proceedings.mlsys.org/paper\\_files/paper/2019/file/acd593d2b87a799a8d3da5a860c028e-Paper.pdf](https://proceedings.mlsys.org/paper_files/paper/2019/file/acd593d2b87a799a8d3da5a860c028e-Paper.pdf)
- [71] Statcounter. 2023. Desktop Browser Market Share United States of America, July 2023. <https://gs.statcounter.com/browser-market-share/desktop/united-states-of-america>. Accessed: August, 2023.
- [72] Statcounter. 2023. Desktop Operating System Market Share United States Of America, July 2023. <https://gs.statcounter.com/os-market-share/desktop/united-states-of-america>. Accessed: August, 2023.
- [73] Alina Stöver, Sara Hahn, Felix Kretschmer, and Nina Gerber. 2023. Investigating how Users Imagine their Personal Privacy Assistant. *Proc. Priv. Enhancing Technol.* 2023, 2 (2023), 384–402. <https://doi.org/10.56553/popets-2023-0059>
- [74] Joseph Turow, Yphtach Lelkes, Nora A. Draper, and Ari Ezra Waldman. 2023. Americans Can’t Consent to Companies’ Use of their Data. [https://www.asc.upenn.edu/sites/default/files/2023-02/Americans\\_Can%27t\\_Consent.pdf](https://www.asc.upenn.edu/sites/default/files/2023-02/Americans_Can%27t_Consent.pdf). Accessed: April 4, 2024.
- [75] Tobias Urban, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. 2019. “Your Hashed IP Address: Ubuntu.”: Perspectives on Transparency Tools for Online Advertising. In *Proceedings of the 35th Annual Computer Security Applications Conference (San Juan, Puerto Rico) (ACSAC '19)*. Association for Computing Machinery, New York, NY, USA, 702–717. <https://doi.org/10.1145/3359789.3359798>
- [76] Weights & Biases. 2023. Weights & Biases. <https://wandb.ai/site>. Accessed: April 4, 2024.
- [77] Ben Weinshel, Miranda Wei, Mainack Mondal, Eurim Choi, Shawn Shan, Claire Dolin, Michelle L. Mazurek, and Blase Ur. 2019. Oh, the Places You’ve Been! User Reactions to Longitudinal Transparency About Third-Party Web Tracking and Inferencing. In *CCS (London, United Kingdom) (CCS '19)*. Association for Computing Machinery, New York, NY, USA, 149–166. <https://doi.org/10.1145/3319535.3363200>
- [78] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. <https://doi.org/10.18653/v1/2020.emnlp-demos.6>

- [79] Yaxing Yao, Davide Lo Re, and Yang Wang. 2017. Folk Models of Online Behavioral Advertising. In *CSCW '17: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) (CSCW '17). Association for Computing Machinery, New York, NY, USA, 1957–1969. <https://doi.org/10.1145/2998181.2998316>
- [80] Shikun Zhang, Yuanyuan Feng, Yaxing Yao, Lorrie Faith Cranor, and Norman Sadeh. 2022. How Usable Are iOS App Privacy Labels?. In *PoPETS 2022*. PoPETS 2022, Sydney, Australia and Virtual, 204–228.
- [81] Sebastian Zimmeck, Rafael Goldstein, and David Baraka. 2021. PrivacyFlash Pro: Automating Privacy Policy Generation for Mobile Apps. <https://dx.doi.org/10.14722/ndss.2021.24100>. In *NDSS*. Internet Society, San Diego, United States, 1–18.
- [82] Sebastian Zimmeck, Eliza Kuller, Chunyue Ma, Bella Tassone, and Joe Champeau. 2024. Generalizable Active Privacy Choice: Designing a Graphical User Interface for Global Privacy Control. In *24th Privacy Enhancing Technologies Symposium (PETS 2024)*, Vol. 1. PETS, Bristol, UK, 1–23. <https://doi.org/10.56553/popets-2024-0015>
- [83] Sebastian Zimmeck, Peter Story, Rafael Goldstein, David Baraka, Shaoyan Li, Yuanyuan Feng, and Norman Sadeh. 2019. Compliance Traceability: Privacy Policies as Software Development Artifacts. [https://petsymposium.org/2019/files/workshop/abstracts/PUT\\_2019\\_paper\\_21.pdf](https://petsymposium.org/2019/files/workshop/abstracts/PUT_2019_paper_21.pdf). Accessed: April 4, 2024.
- [84] Sebastian Zimmeck, Peter Story, Abhilasha Ravichander, Daniel Smullen, Ziqi Wang, Joel Reidenberg, N. Cameron Russell, and Norman Sadeh. 2019. MAPS: Scaling Privacy Compliance Analysis to a Million Apps. In *PETS 2019*, Vol. 3. Sciendo, Stockholm, Sweden, 66–86. <https://doi.org/10.2478/popets-2019-0037>
- [85] Sebastian Zimmeck, Oliver Wang, Kuba Alicki, Jocelyn Wang, and Sophie Eng. 2023. Usability and Enforceability of Global Privacy Control. In *23rd Privacy Enhancing Technologies Symposium (PETS 2023)*, Vol. 2. PETS, Lausanne, Switzerland, 265–281. <https://doi.org/10.56553/popets-2023-0052>
- [86] Sebastian Zimmeck, Ziqi Wang, Lieyong Zou, Roger Iyengar, Bin Liu, Florian Schaub, Shormir Wilson, Norman Sadeh, Steven M. Bellovin, and Joel Reidenberg. 2017. Automated Analysis of Privacy Requirements for Mobile Apps. In *NDSS*. Internet Society, San Diego, CA, 15 pages. <https://doi.org/10.14722/ndss.2017.23034>

## 8 APPENDIX

### 8.1 Regular Expressions

```
function buildGeneralRegex(genString) {
  "Replace anything that is not a digit or
  letter in the alphabet with optional non-
  digit placeholder"
  testUser@gmail.com -> testUser\Dgmail\Dcom }

```



Figure 14: Email regex construction.

```
function buildIpRegex(ipAddress) {
  "Replace anything that is not a digit with
  an optional non-digit placeholder"
  123.456.78.90 -> 123\D456\D78\D90 // Also
  accounts for IPv6
}

```

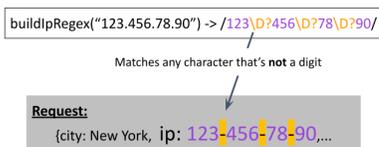


Figure 15: IP regex construction.

```
function buildZipRegex(zip) {
  "Replace any spaces or dashes with an
  optional non-digit placeholder"
  "Add [^0-9] to beginning and end of zip
  for Regex"
  12 34-5 -> [^0-9]12\D34\D5[^0-9] }

```

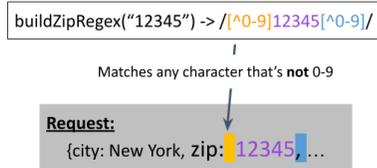


Figure 16: ZIP code regex construction.

```
function getRegion(region) {
  "Replace spaces, periods (.), or dashes
  (-) with optional non-digit
  placeholder"
  Rhode_Island -> Rhode\DIsland }

```

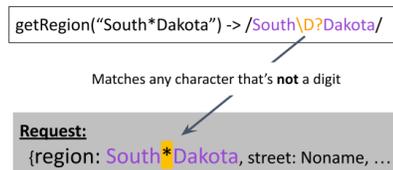


Figure 17: Region regex construction.

```
function coordinateSearch(strReq) {
  "Takes users lat and long, and matches it
  with any coordinate matched by the regex"
  let floatReg = /\D\d{1,3}\.\d{1,10}/g
}

```

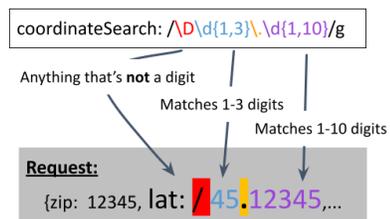


Figure 18: Latitude/longitude search regex.

Regex Item	Regex
Dynamic Pixel (Height - Width)	/height\D{1,8}[0,1]\D{1,20}width\D{1,8}[0,1]\D/g
Dynamic Pixel (Width - Height)	/width\D{1,8}[0,1]\D{1,20}height\D{1,8}[0,1]\D/g
Pixel	/pixel/
Question Mark	/\?/

Table 7: Regexes for dynamic pixel search function.



### 8.3 Privacy Interface Screenshots

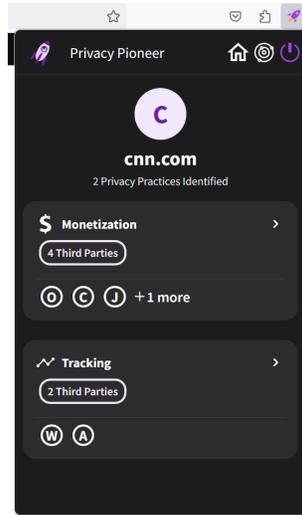


Figure 19: Example of the privacy popup interface.

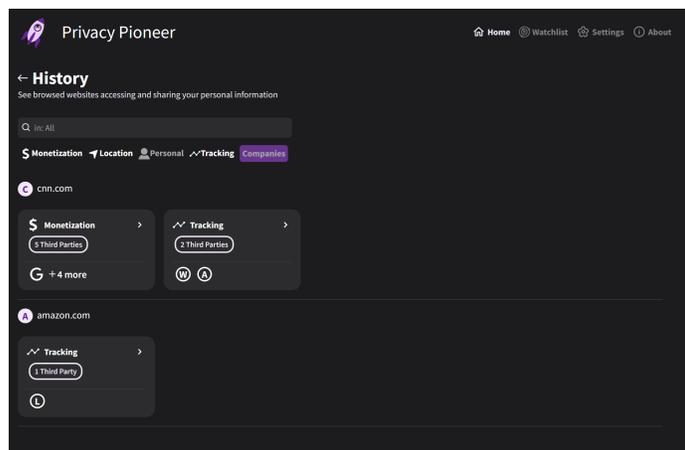


Figure 20: Example of the privacy history interface.

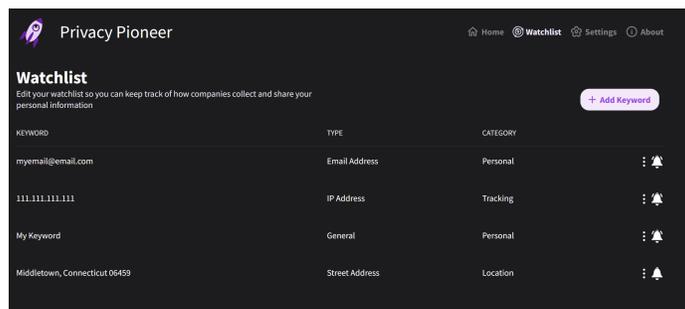


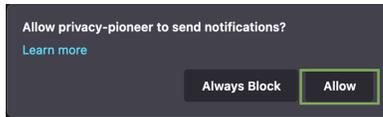
Figure 21: Example of the privacy watchlist interface.

## 8.4 Survey Questionnaire

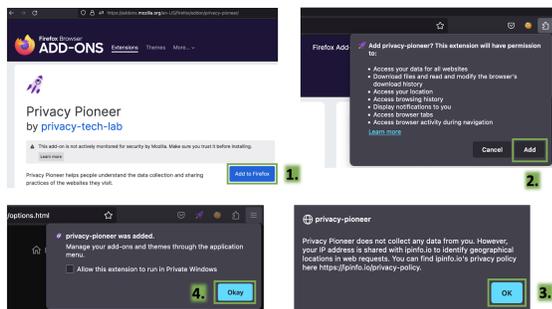
- **Q1** Please send us now a request for the file upload link via the Prolific messaging system. Please do not continue until you have received your upload link, which we will send you via the Prolific messaging system. [Confirmation; participant must follow directions]
- **Q2** Please select the extent to which you agree with the following statement: “I generally feel confident that my privacy is protected on the web.”  
[Linear Scale]  
Strongly Disagree    1    2    3    4    5   Strongly Agree
- **Q3** Please select the extent to which you agree with the following statement: “There is not much I can do to protect my privacy on the web.”  
[Linear Scale]  
Strongly Disagree    1    2    3    4    5   Strongly Agree
- **Q4** Please select the extent to which you agree with the following statement: “I generally feel that I have a good understanding of what data websites collect from me and with whom they share it.”  
[Linear Scale]  
Strongly Disagree    1    2    3    4    5   Strongly Agree
- **Q5** Please select the extent to which you agree with the statement? “I generally prefer ads on websites in exchange for free content rather than paying for content.” (assuming you would not block ads)  
[Linear Scale]  
Strongly Disagree    1    2    3    4    5   Strongly Agree
- **Q6** Do you care whether a website shares the following of your data for ad purposes?  
[Linear Scale Grid]  
Users select one of the following choices:
  - Do not care at all
  - Do not care much
  - Neutral
  - Care somewhat
  - Care a lot
 For the following data types:
  - Phone number
  - ZIP code
  - The fact that you visited the site
  - Email address
  - GPS location (precise location within 20 feet of your actual location)
  - IP address
  - Your interests (e.g., travel & transportation)
- **Q7** Would you allow websites that you visit to share your interests (e.g., travel & transportation or books & literature) with any of the following organizations? Select all that applies. Select “None” if you would like websites to not share your interests with any organization.  
[Checkboxes]
  - Ad networks that track people from site-to-site to serve personalized ads
  - Analytics companies that help sites to identify software bugs
  - Data brokers that buy and sell people’s data
  - Social media companies like Meta (Facebook and Instagram) or ByteDance (TikTok)
  - Security service providers that make sure that a site is secure and private
  - Big tech companies like Apple, Google, or Microsoft (excluding social media companies)
  - None
- **Q8** Which types of privacy software do you use (outside of this study)? Select all that applies. Select “Other” if a software you use is not listed. Select “None” if you use no privacy software.  
[Checkboxes]
  - Ad/tracking blocker browser extension
  - Privacy-protective browser
  - Virtual Privacy Network (VPN)
  - Privacy-protective email provider
  - Other
  - None
- **Q9** Would you use an ad/tracking blocker that you can customize such that you decide which website receives which data?  
[Multiple choice]
  - I would keep my current ad/tracking blocker
  - I would use it if it would keep the sites free
  - I am currently not using an ad/tracking blocker and this would not change
- **Q10** Which browser do you mainly use on your laptop/desktop?  
[Multiple choice]
  - Chrome
  - Safari
  - Firefox
  - Edge
  - Brave
  - Opera
  - Vivaldi
  - Tor Browser
  - Another browser
- **Q11** Which operating system do you mainly use on your laptop/desktop?  
[Multiple choice]
  - Windows
  - macOS
  - Linux
- **Q12** How would you rate your general tech-savviness?  
[Linear Scale]  
Not tech-savvy at all    1    2    3    4    5   Very tech-savvy
- **Q13** If Firefox asks at any point during this study, please click OK to allow current location use.  
[Confirmation; participant must follow directions]



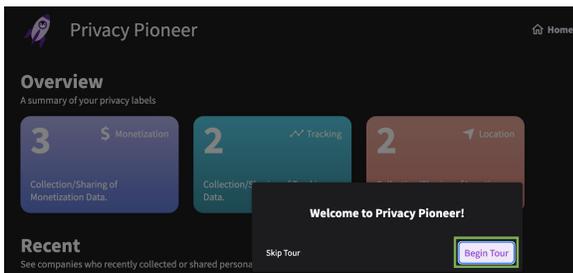
- **Q14** If Firefox asks at any point during this study, please “Allow” notifications.  
[Confirmation; participant must follow directions]



- **Q15** Please install Privacy Pioneer:
  1. On your Firefox browser visit <https://addons.mozilla.org/en-US/firefox/addon/privacy-pioneer/> and click “Add to Firefox.”
  2. Click “Add” to give Privacy Pioneer the required permissions.
  3. Click “OK” on the IP address notification.
  4. Click “Okay” to confirm that Privacy Pioneer was added to your browser. (You can allow it to run in Private Windows, though, please do not use a Private Window during this study.)
 [Confirmation; participant must follow directions]



- **Q16** Please take the brief Privacy Pioneer tour by clicking on “Begin Tour” and follow the individual screens.  
[Confirmation; participant must follow directions]

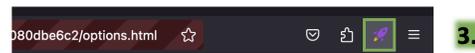
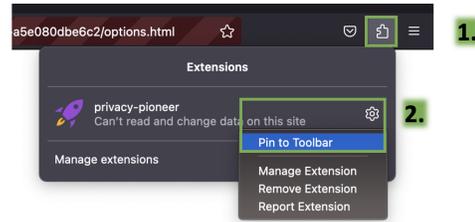


- **Q17** Please pin Privacy Pioneer to your toolbar:
  1. Click the puzzle icon in your browser bar.
  2. Click the gear wheel and select “Pin to Toolbar.”

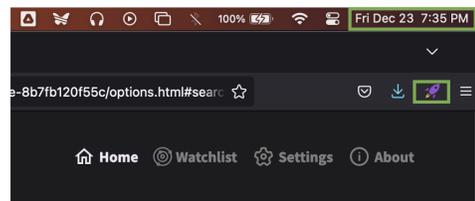
3. You should now see the little Privacy Pioneer rocket icon pinned in your browser bar.

Please do not continue until you pinned Privacy Pioneer to your toolbar.

[Confirmation; participant must follow directions]



- **Q18** Please take a screenshot (similar as shown below) and upload it via the upload link we sent you earlier.  
Please make sure to capture today’s date and the Privacy Pioneer icon pinned to your toolbar.  
Please upload this screenshot via the upload link to the subfolder: 1 - Icon and Time  
Here are instructions on how to take a screenshot:
  - Windows: <https://support.microsoft.com/en-us/windows/open-snipping-tool-and-take-a-screenshot-a35ac9ff-4a58-24c9-3253-f12bac9f9d44>
  - macOS: <https://support.apple.com/en-us/HT201361>
  - Linux (Ubuntu): <https://help.ubuntu.com/stable/ubuntu-help/screen-shot-record.html>
 [Confirmation; participant must follow directions]



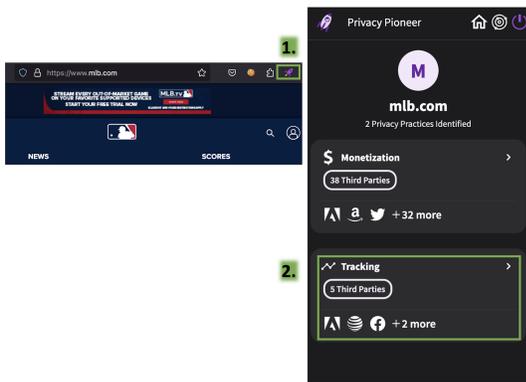
- **Q19** In this first task you will (1) use the privacy popup interface to identify the trackers on a website, (2) change your privacy browser settings to block trackers, and (3) confirm with the privacy popup interface that all trackers are indeed gone.

1. Please visit <https://www.mlb.com/>. Once the website is fully loaded, wait 15 seconds and click the little rocket icon in the browser bar to open the privacy popup of Privacy Pioneer.

2. In the privacy popup you should see one or more third party trackers flagged similar as shown below (“Tracking 5 Third Parties”).

If you do not see any third party tracker, please refresh the website and wait 30 seconds for all website elements to load before you open the privacy popup again.

[Confirmation; participant must follow directions]



- **Q20** Please block the trackers via your Firefox settings:
  1. In a new browser tab visit: `about:preferences#privacy`
  2. Select “Privacy & Security.”
  3. Select “Custom.”
  4. In the dropdown menu, select “In all windows.”
  5. Click “Reload All Tabs” and reload `https://www.mlb.com/` by refreshing its browser tab, to be sure.

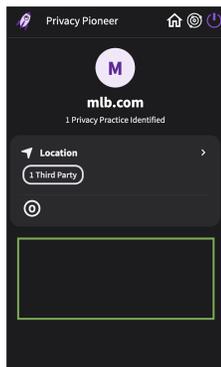
[Confirmation; participant must follow directions]



- **Q21** Open the privacy popup of Privacy Pioneer on `https://www.mlb.com/` to confirm that all trackers on the website are indeed gone.
 

If you still see third party trackers, please refresh the website again or try loading `https://www.mlb.com/` in a new browser tab (as the trackers may still be in the browser cache).

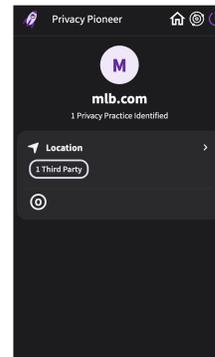
[Confirmation; participant must follow directions]



- **Q22** Please take a screenshot of the privacy popup (similar as shown below) showing that the trackers are all gone.
 

Please upload this screenshot via the upload link to the subfolder: “2 - No Trackers”

[Confirmation; participant must follow directions]



- **Q23** Please select the extent to which you agree with the following statement: “I find the information shown in the privacy popup easy to understand.”
 

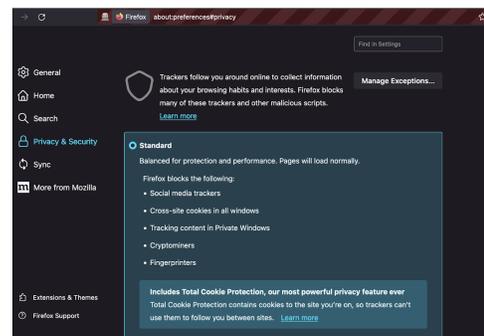
[Linear Scale]

Strongly Disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly Agree
- **Q24** Please select the extent to which you agree with the following statement: “I find the privacy popup useful to identify the tracking practices of a website.”
 

[Linear Scale]

Strongly Disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly Agree
- **Q25** Before you begin with the second task, please change your privacy setting under `about:preferences#privacy` back to “Standard” as shown below.
 

[Confirmation; participant must follow directions]



- **Q26** In this task you will use the privacy history interface to explore how a company can track you across different websites.
 

Please visit the following five websites:

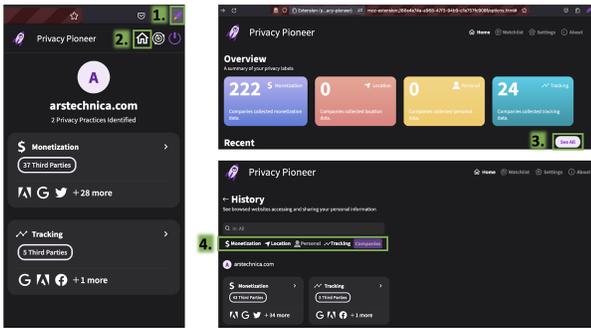
  1. `https://twitter.com/`
  2. `https://golf.com/`
  3. `https://www.warbyparker.com/`
  4. `https://www.ups.com/`
  5. `https://arstechnica.com/`

You can visit the websites one after another in the same browser tab. Before you move to a subsequent website, please wait until the previous website has fully loaded and add 15 seconds.

[Confirmation; participant must follow directions]

- **Q27** After visiting the websites, please identify on which of those Adobe was tracking you:
  1. Open the privacy popup by clicking on the little rocket icon in your browser bar.
  2. Open the privacy history interface by clicking on the little home icon in the privacy popup.
  3. To see all analysis results of the websites you visited, click “See All.”
  4. You can now use the filters to identify the set of websites for which Adobe tracked you. Try and see how you can set the filter so that you can see Adobe’s “Tracking.”

[Confirmation; participant must follow directions]



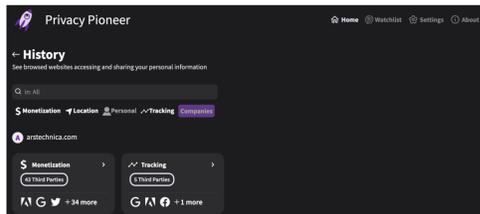
- **Q28** Please upload between two to five screenshots of the privacy history interface with filters enabled for Tracking by Adobe.
 

Please use as many screenshots as necessary for all results to be visible.

Please make sure that you have the company filter for Adobe enabled. Your screenshots would look similar as below but with the filter enabled.

Please upload your screenshots via the upload link to the subfolder: “3 - Filters”

[Confirmation; participant must follow directions]



- **Q29** It is important that you pay attention to this study. Please select “Neutral”.
- [Linear Scale]

◦ Strongly agree ◦ Agree ◦ **Neutral** ◦ Disagree ◦ Strongly disagree

- **Q30** Please select the extent to which you agree with the following statement: “I find the information shown in the privacy history interface easy to understand.”
 

[Linear Scale]

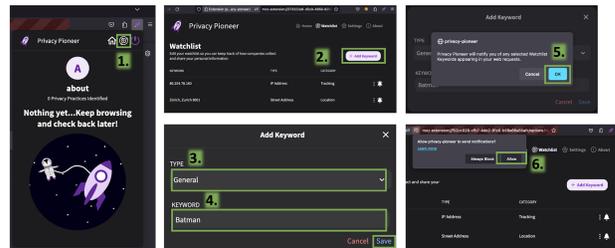
Strongly Disagree ◦ 1 ◦ 2 ◦ 3 ◦ 4 ◦ 5 Strongly Agree
- **Q31** Please select the extent to which you agree with the following statement: “I find the privacy history interface useful to identify how companies are tracking me across different websites.”
 

[Linear Scale]

Strongly Disagree ◦ 1 ◦ 2 ◦ 3 ◦ 4 ◦ 5 Strongly Agree
- **Q32** In this task you will use the privacy watchlist to notify you when a website you visit shares a custom keyword with another website.

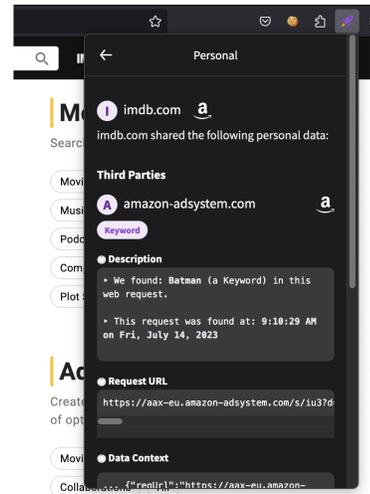
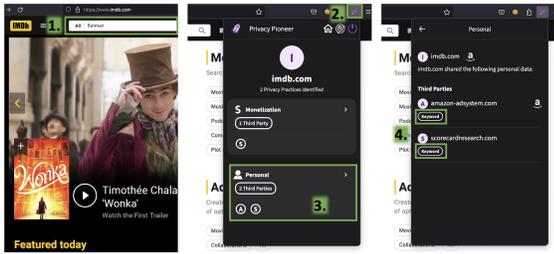
Set up your privacy watchlist as follows:

1. Open the privacy popup and click on the little sonar icon to navigate to the privacy watchlist interface.
  2. Once you are on the privacy watchlist interface, click the “+ Add Keyword” button.
  3. Select as type of keyword “General.”
  4. Enter “Batman” (without quotes) and click “Save” (if saving does not work, you may not have selected “General”).
  5. Click “OK” to confirm that Privacy Pioneer will notify you when your keyword appears in your web traffic.
  6. Click “Allow” to enable notifications.
- [Confirmation; participant must follow directions]



- **Q33** In your browser please do the following:
  1. Navigate to <https://www.imdb.com>, enter “Batman” (without quotes) in the IMDb search bar, and perform the search. To see how the notifications work, wait for 15 seconds. (Depending on the notification settings of your browser and computer, you may not receive a notification.)
  2. Open the privacy popup by clicking on the little rocket icon in your browser bar
  3. Click on the Personal card to see who processed your keyword.
  4. Once the card has opened, click on “Keyword” to learn details about how your keyword was processed.

[Confirmation; participant must follow directions]

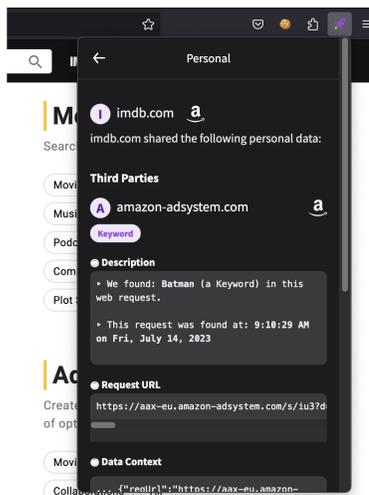


- **Q34** You should now see keyword processing details similar as below. What is shown in the screenshot below?

Please note: The privacy popup always holds the privacy analysis results from the current website. If you want to see comprehensive results from all websites you visited, you can click on the little house icon to navigate to the privacy history interface and “See All.”

[Multiple choice]

- IMDb collected the search query “Batman” but did not share it
- **IMDb shared the search query “Batman” with Amazon Ads**
- IMDb collected the search query “Batman” and the user must have also visited Amazon Ads before
- None of the above



- **Q35** Please take a screenshot of the privacy watchlist interface (similar as shown below).

Please upload your screenshot via the upload link to the subfolder: “4 - Watchlist”

[Confirmation; participant must follow directions]

- **Q36** Please select the extent to which you agree with the following statement: “I find the privacy watchlist interface easy to understand.”

[Linear Scale]

Strongly Disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly Agree

- **Q37** Please select the extent to which you agree with the following statement: “I find the privacy watchlist interface useful to keep track of data websites are sharing about me.”

[Linear Scale]

Strongly Disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly Agree

- **Q38** In Privacy Pioneer go to Settings -> Export -> Analytics. If you click on Analytics, a file with the features you used in Privacy Pioneer will download to your computer. Rename it to <yourProlificID\_pop\_privacy\_analytics>.JSON.

Please note: If you want, you can delete entries from the file before uploading it.

Please upload your file via the upload link to the subfolder: “5 - Analytics”

[Confirmation; participant must follow directions]

- **Q39** Please select the extent to which you agree with the following statement: “Privacy Pioneer overall is easy to understand.”

[Linear Scale]

Strongly Disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly Agree

- **Q40** Please select the extent to which you agree with the following statement: “Privacy Pioneer overall is useful.”

[Linear Scale]

Strongly Disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Strongly Agree

- **Q41** Which interfaces of Privacy Pioneer should be directly integrated into a browser? Select all that applies.

[Checkboxes]

- Privacy Popup
- Privacy History
- Privacy Watchlist

- **Q42** Did you encounter any technical issues during your use of Privacy Pioneer?  
[Multiple choice]
  - Yes ◦ No
- **Q43** Please describe any technical issues you encountered.  
[Long answer text; answer may be left blank]
- **Q44** Do you plan on continuing to use Privacy Pioneer? Please select your answer depending on whether Firefox is your main browser on your desktop/laptop.  
[Multiple choice]
  - Firefox is my main browser, and I WILL continue using Privacy Pioneer
  - Firefox is my main browser, and I WILL NOT continue using Privacy Pioneer
  - Firefox is not my main browser but if it were, I WOULD continue using Privacy Pioneer
  - Firefox is not my main browser but if it were, I WOULD NOT continue using Privacy Pioneer
- **Q45** How likely is it that you would recommend Privacy Pioneer to a friend or colleague? (assuming that your friend or colleague is a Firefox user)  
[Multiple choice]
 

Not at all likely ◦ 1 ◦ 2 ◦ 3 ◦ 4 ◦ 5 ◦ 6 ◦ 7 ◦ 8 ◦ 9 ◦ 10 Extremely likely
- **Q46** What do you like about Privacy Pioneer?  
[Long answer text]
- **Q47** How could we improve Privacy Pioneer?  
[Long answer text]
- **Q48** Please enter your Prolific ID.  
[Short answer text]
- **Q49** By selecting “I Agree” and clicking “Submit” you are consenting to participate in this study under the terms described herein. You further confirm that you were given the opportunity to read all information described previously and ask any questions about this study.  
[Confirmation; user may refuse to select “I Agree”]

\* All questions are required unless noted otherwise.

### 8.5 Data Collected from Study Participants

- Prolific ID
- IP address
- Geographic location (e.g., latitude/longitude coordinates, city, or ZIP code)
- Website URLs visit while using our extension
- URLs of third party websites that are integrated in the visited websites (e.g., of third party companies)
- Details of web requests to and from visited websites (e.g., cookie IDs and values)
- Timestamp of a website visit
- Time zone
- Keywords and data that a participant chooses to monitor via our extension (participants were cautioned not to enter names, passwords, or other sensitive information)
- Which features were used in our extension
- Screenshots to take for some questions of the survey

### 8.6 Additional Statistics

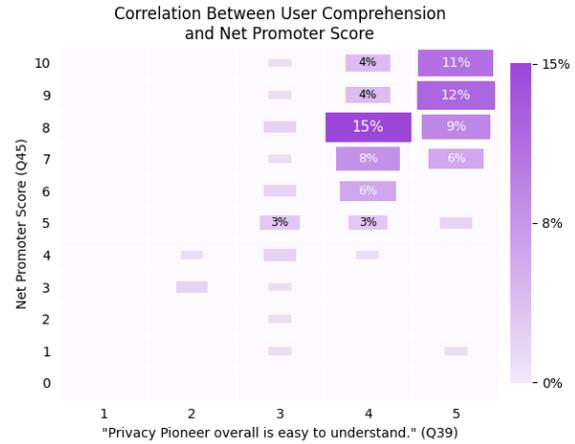


Figure 22: Correlation between understanding Privacy Pioneer and recommending it.

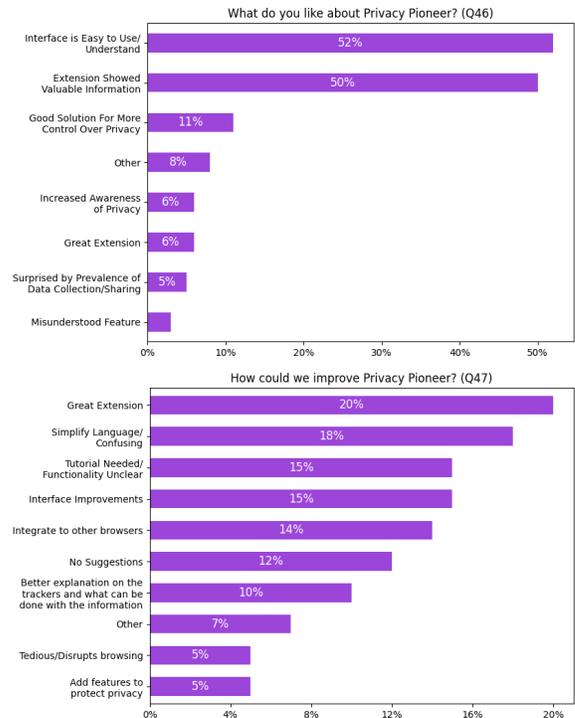


Figure 23: Coded responses to free-form questions Q46 and Q47.