# Daedalus, the data-less tracker for ForeSee

Guillaume Peugniez, Pablo Suarez, Calvin Gerus

Online privacy concerns have hit the mainstream media, and these reports are impacting how tech companies are capturing and storing our personal information. Facebooks congressional hearing and the Cambridge Analytica debacle has shed significant light on web tracking by the likes of Google and Facebook. These companies currently track 64% and 30% of global page loads respectively [12]. Most people are now more aware of how often they are tracked online and at how at risk their privacy and data is. These privacy concerns are growing so much so that Facebook is looking at changing its underlying business model and building a "privacy-focused platform"[1]. GDPR,  a massive data privacy act launched in 2018 with global implications, brought forward harsh penalties for those companies and organizations who don't comply. Fines of up to 4% of annual global revenue or 20 million Euros, whichever is greater. As of January 1, 2019,  53% of companies are still in the implementation phase of complying to GDPR restrictions, and 27% have not even started. Countless companies are currently in breach of these laws, and it is only a matter of time until they are charged with these massive fines.

The general population is becoming more data savvy, and as they do their sentiment on data collection & storage is changing as well from neutral to unacceptable.  Browsers and start-ups have taken notice, and are developing new tracker-less browsers or services that block trackers, protecting users' privacy and personal information. These browsers and start-ups are challenging the current ad-forward business model of the internet, controlled mostly by Google & Facebook. Their theory is that eventually, the majority of people would rather pay a small amount to use services that ensure their data is secure and their privacy is safe, over something that is 'free' by way of data robbery.

Mozilla's Firefox now has tracker blocking as a default in their privacy settings [2,3]. Brave[4], a new browser built with data privacy as a core principle, has developed a browser that is faster than Chrome that deploys AI to protects you from trackers. Disconnect.me is an extension with a comprehensive tracker blacklist who has partnered with Firefox, Tor, Brave, Samsung, and T-Mobile to provide them with their filter list. Ghostery deploys AI to understand if a user ID is being passed within a network call to intercept and shut down these code snippets[5]. Even Apple has been making considerable strides in data protection & privacy within Safari, and it's iOS platforms[7].

Browser market share doesn't currently depict this trend. As you can see from 2009 - 2019 Chrome has largely overtaken Firefox as the leading Browser[6]. However, we don't fully understand the extent of how users have configured their Chrome browser, with many extensions that can convert the browser into blocking trackers, like Disconnect Me's extension. Data privacy search queries on Google had steadily increased in the last five years, with an all-time high when GDPR came into effect in May 2018. According to a 2017 study by PageFair, a firm that helps companies recoup lost advertising revenue, 615 million devices currently deploy an ad blocker[8]. It is clear that users are increasingly turning to privacy tools to disable tracking.

Ghostery published a paper called *Tracking the Trackers: Analyzing the global tracking landscape with GhostRank* [12] where they asked following questions:

1. How much tracking can be observed on any given page load?
2. How much reach do specific tracking widgets and companies have over users' web browsing?
3. How does the tracking ecosystem vary between different regions?

Some facts from this report:

- About 75% of page loads have at least one tracker running
- Over 15% of page loads and 10% of all sites have 10 or more trackers seen. This shows that a large proportion of internet traffic has an extreme amount of tracking.
- When we combine Google's third-party services (analytics, advertising and social) and their first-party services (Search, Maps, Youtube, etc.), they are party to over 64% of all web-browsing worldwide
- Similarly Facebook's reach is approaching 30% as their advertising tracking tools gain more reach.

Conclusions from the study [12]:

> *The extent of online tracking is extraordinary, and that in addition to the dominance of major players such as Google and Facebook, there is a long tail of companies also hovering up significant quantities of user browsing data*

> *Web tracking has become pervasive, and with Google and Facebook tracking 64% and 29% of pages loaded on the web, it is becoming almost impossible to avoid. Additionally, 15% of pages will send data to 10 or more different companies. As well as being a significant burden on resources (both CPU and network) to load scripts from all these different parties, there is little transparency about what is being shared and with whom. Users are increasingly turning to privacy tools like Ghostery to notify them about who is tracking on each page, and allow them to 'opt-out' as they wish. However, it remains to be seen if the increase in this behaviour will lead to a change in the pervasiveness of tracking.*

We believe the uptrend in users opting-in to data privacy & protection will continue and it will have significant implications for the web analytics & tracking business in which ForeSee is a member.

## Problem Statement

The era of free, consistent user data is coming to an end. The general population is becoming more data savvy and is enlisting tools to help them protect their data and privacy. Browsers are starting to make privacy and data protection by blocking trackers as a default. Therefore, It is up to us to get ahead of this curve. We can do this by anchoring data protection as a design constraint while archetyping our data collection methods and SDKs, while still allowing us to execute on our future product roadmap.

# Project Goals

1. Understand what makes a tracker blacklisted or whitelisted
2. Design a POC that meets the requirements of being on the tracker whitelist while showing support for future ForeSee roadmap items.

# Tracker Blockers: Browsers, Blacklists & Extensions

## Tracker Blockers: Browsers, Blacklists, & Extensions

In this section, we will take a look at a range of Browsers, tracker Blacklists, and browser extension tools that block trackers to understand the requirements which put trackers on the blacklist.

### FireFox

By default, content blocking uses the Disconnect.me basic protection list. Users can change this to use the Disconnect.me strict protection list instead.

- The basic protection list blocks commonly known analytics trackers, social sharing trackers, and advertising trackers. However, the basic protection list allows some known content trackers so that fewer pages break or fail to load.
- The strict protection list blocks all known trackers, including analytics trackers, social sharing trackers, and advertising trackers as well as content trackers. The strict list will break some videos, photo slideshows, and social networking features.

### Disconnect Me

Disconnect Me partners with Mozilla and Brave to provide them a blacklist [9] of trackers worthy of blocking. They also offer a browser extension that can be installed on Chrome or any other browser to give you the same content blocking that Firefox has built into their browser. They also offer an iOS / Android app that protects your complete device with DNS encryption on their SmartVPN. The app encrypts all HTTP traffic to keep sensitive information private.

Disconnect Me is the most widely used tracker blacklist. If ForeSee gets listed on Disconnect Me basic protection list, this will cut into our already low survey collection rates.

## Disconnect Me Blacklist Requirements

### **Disconnect me definition of tracking**

> *Tracking is the collection of data regarding a particular user's activity across multiple websites or applications that aren't owned by the data collector, and the retention, use or sharing of that data.*
>
> *Our definition focuses on collection AND retention. So, for example, the definition wouldn't apply to sites that log an IP address, but don't save that information in a database. The definition also focuses on particular users, so data that is immediately aggregated doesn't apply. And the collection is across context, so it doesn't apply in cases when there is solely a first-party relationship with the user, for example the site only collects and retains information on site visitors.*

**Trackers Disconnect me block are those services that they have identified and determined that meet the definition of tracking above.** Disconnect compiles several lists of trackers. The open source list of trackers that power their browser extensions, Firefox's private browsing mode, and many other favourite privacy tools can be found here, along with a change-log and notes. Alternatively, you can view a simple list of blocked trackers here.

Disconnect me does strive to find the balance between privacy, security, usability and promoting a better Internet for everyone. These concerns drive their decisions in regards to trackers they block and don't block.

Disconnect me generally unblocks tracking sites that commit to respect users' Do Not Track (DNT) preferences and agree to comply with DNT as defined by the Electronic Frontier Foundation: https://www.eff.org/dnt-policy. They also generally unblock tracking sites that require users to transparently and explicitly opt into collection and retention.

All of the trackers they have identified but don't block, along with a change-log and notes, can be found here. Alternatively, you can view a simple list of unblocked trackers here.

## Ghostery

Blacklists, or filter lists, are manually curated based on informal crowdsourced feedback, which brings with it a significant number of maintenance challenges.  Also, trackers can quickly get around a filter list by changing their domain name periodically or hosting the tracker on the client's servers. Therefore, Ghostery developed an innovative anti-tracking method based on the algorithmic detection of user identifiers in tracking requests. They combine this with filtered blacklists in an extension that can be added to any browser.

## Brave

Brave has created a machine learning approach for automatic and effective ad-blocking called AdGraph. Their approach relies on information obtained from multiple layers of the web stack (HTML, HTTP, and JavaScript) to train a machine learning clas-

sifier to block ads and trackers. Braves evaluation on Alexa top-10K websites shows that AdGraph automatically and effectively blocks ads and trackers with 97.7% accuracy. After manual analysis, AdGraph showed better recall than filter lists, it blocks 16% more ads and trackers with 65% accuracy. They also show that AdGraph is fairly robust against adversarial obfuscation by publishers and advertisers that bypass filter lists [10]. However it also partners with crowdsourced blacklists from Spam404 and Disconnect Me to block dangerous domains.

Brave has cookie control that allows it to block first or third party cookies. By default, Brave will allow first party cookies and block cookies from any third party. A user has to log a domain to block its first party cookies. They also have the ability to block JavaScripts from running. But this is not set as default as it would break a lot of websites and is only for Brave advance users. Brave also blocks third party fingerprinting by turning off many features commonly used to differentiate between devices [11].

What makes Brave interesting as potentially the newly mass adopted browser is it protects the users' data & privacy, but also it is substantially faster than other browsers and has built in a complete token ecosystem. Users that use Brave browser will have the option to view ads, and in return, they will receive BAT (Basic Attention Token). These tokens can be exchanged for Euro, U.S. dollars or any other currency or cryptocurrency you have in mind. So trackers may still be able to exist within the Brave ecosystem. However, third parties now might have to pay for that attention or data.

# User Study

We conducted a user study with 10 random people on the streets of Vancouver, Canada; and asked them how they felt about data privacy & protection. Here are the results:

- 80% of the respondents didn’t care that they were being tracked online
- 40% of the respondents wouldn’t switch browsers knowing one browser would protect their data more than the other
- 90% of the respondents would switch bowsers knowing one is faster over the other.
- 60% of the respondents have negative feelings about the current situation around Facebook and how it has handled their data
- 70% of the respondents are likely to pay more attention to data privacy & protection in the future

One of the most surprising results was that most people will more likely switch browsers based upon speed rather than data privacy & protection. We can tell this aligns with Braves market research as well. As speed is the first value proposition their marketing team is conveying to the users when visiting their website.  Most people questioned in Canada did not care too much about data privacy and protection, and couldn’t decide if they liked targeted ad’s or not. Some felt they were both useful and creepy so resulting in a neutral sentiment towards them. Intrinsically reflecting on the user study, we noticed that everyone was well aware of the topic of data tracking and how their personal data is being used over the internet. This general knowledge would not have been the case just a few years ago.

## Requirements Overview

In order to stay off of tracker blacklists like Disconnect Me, and stay undetectable from AI systems trained to sniff out 3rd party trackers like Ghostery & Brave, the following requirements are necessary for the POC.

4. Data can be logged, but not saved in a database
5. Data can be converted into *Insights* by immediately aggregating before insights can be saved in a database
6. **Or** the tracker is solely a first-party relationship with the user, for example the site only collects and retains information on site visitors.
7. Must respect Do Not Track (DNT) preferences and agree to comply with DNT as defined by the Electronic Frontier Foundation: https://www.eff.org/dnt-policy.
8. Tracker must require users to transparently and explicitly opt into collection and retention
9. Avoid sending tracking requests that contain any user identifiers
10. Cannot reduce the performance of the client's site, using ForeSees webSDK v19.9.0 as the benchmark.

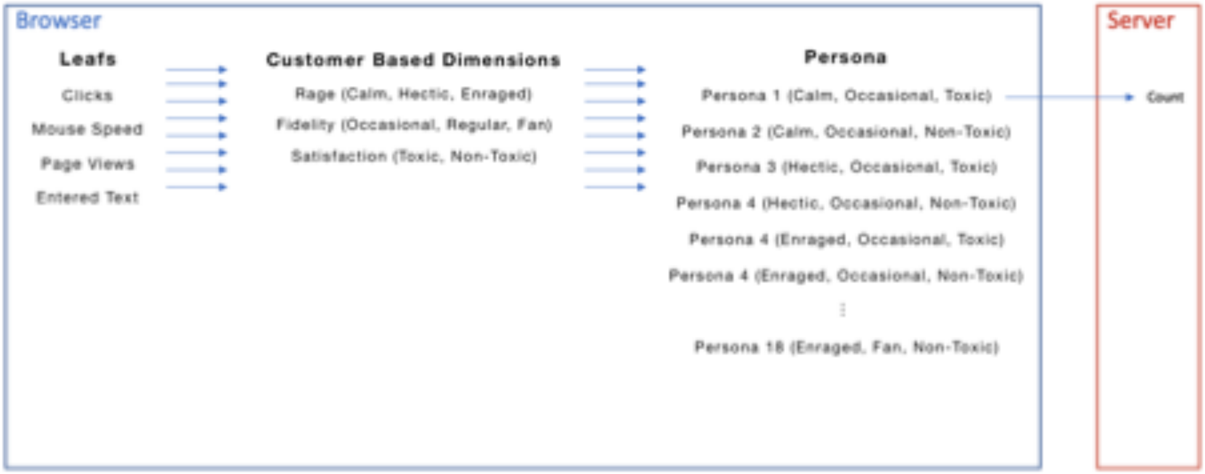
# The Daedalus SDK

Our tracker only sends aggregate data about a user that is based upon computed behaviors observed within the browser. These user behaviors are used to generalize users into persona buckets for analysis.  No raw data is sent to the server, and no user

ID's are sent. Survey respondents can still be analyzed and are viewable on the singular level but the user details are based on what persona bucket the survey responded was in during the time of survey submission.

## Technical Details

The bowser observes user behaviour called leafs and computes customer based dimensions (CBDs) per page. We can deploy AI to understand the leaves and how we classify them into the CBDs. At page unload, the browser then tallies up the CBDs into a specific persona type and send only the persona type that was observed for this page view to the server for use in analytics. Figure 1 below depicts this type of data flow that is occurring within the browser.


Browser
Leafs
Clicks
Mouse Speed
Page Views
Entered Text
Customer Based Dimensions
Rage (Calm, Hectic, Enraged)
Fidelity (Occasional, Regular, Fan)
Satisfaction (Toxic, Non-Toxic)
Persona
Persona 1 (Calm, Occasional, Toxic)
Persona 2 (Calm, Occasional, Non-Toxic)
Persona 3 (Hectic, Occasional, Toxic)
Persona 4 (Hectic, Occasional, Non-Toxic)
Persona 4 (Enraged, Occasional, Toxic)
Persona 4 (Enraged, Occasional, Non-Toxic)
⋮
Persona 18 (Enraged, Fan, Non-Toxic)
Server
Count

# Final Thoughts

We have shown that data privacy and protection can be achieved by reducing data transfer to only sending the *insights* from the raw data to the server for analysis &

presentation. With this type of SDK ForeSee is able to become one of the few trackers that may still run on future browsers. Giving ForeSee competitive advantage with clients worried about the hefty fines of GDPR and customer backlash around data privacy concerns. With Daedalus clients will have access to meaningful VOC data while staying safe in the eyes of law and browser heir customers.

# References

1. The Economist. Mark Zuckerberg announces his firms next business model, March 7th, 2019. https://www.economist.com/business/2019/03/07/mark-zuckerberg-announces-his-firms-next-business-model
2. FireFox Support. Content Blocking https://support.mozilla.org/en-US/kb/content-blocking
3. ZDNet. Firefox to add Tor Browser anti-fingerprinting technique called letterboxing, March 6th, https://www.zdnet.com/article/firefox-to-add-tor-browser-anti-fingerprinting-technique-called-letterboxing/
4. Brave Homepage, About: https://www.brave.com/about/
5. Ghostery. Tracking the Trackers, December 4, 2017: https://www.ghostery.com/lp/study/
6. w3school browser statistics, February 2019, https://www.w3schools.com/browsers/

6. w3school browser statistics, February 2019. https://www.w3schools.com/browsers/
7. Wired Magazine. APPLE JUST MADE SAFARI THE GOOD PRIVACY BROWSER, June 4th, 2018. https://www.wired.com/story/apple-safari-privacy-wwdc/
8. The good, the bad and the ugly sides of data tracking April 2018. https://internethealthreport.org/2018/the-good-the-bad-and-the-ugly-sides-of-data-tracking/
9. disconnect.me: disconnect me blocked list: https://disconnect.me/trackerprotection/blocked
10. AdGraph: A Machine Learning Approach to Automatic and Effective Adblocking, Cornell University, Computer Science, Computers & Society. May 22, 2018. https://arxiv.org/abs/1805.09155
11. Brave Support: how do i use shields while browsing. https://support.brave.com/hc/en-us/articles/360022806212-How-do-I-use-Shields-while-browsing-
12. Tracking the Trackers: Analysing the global tracking landscape with GhostRank. https://www.ghostery.com/wp-content/themes/ghostery/images/campaigns/tracker-study/Ghostery_Study_-_Tracking_the_Trackers.pdf