# Heat Marks the Spot: De-Anonymizing Users' Geographical Data on the Strava Heatmap

Kevin Childs
*Department of Computer Science*
*North Carolina State University*
*Raleigh, North Carolina*
*krchilds@ncsu.edu*

Daniel Nolting
*Department of Computer Science*
*North Carolina State University*
*Raleigh, North Carolina*
*dpnoltin@ncsu.edu*

Anupam Das
*Department of Computer Science*
*North Carolina State University*
*Raleigh, North Carolina*
*anupam.das@ncsu.edu*

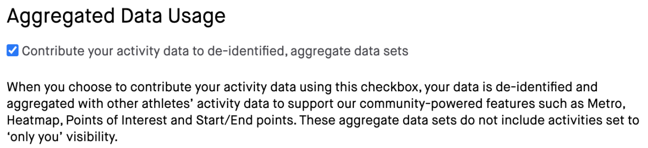
***Abstract*—Mobile fitness-tracking apps such as Strava are commonly used to record activities, track fitness progress, and form a community with like-minded people. In an effort to engage the community further, in 2018 Strava implemented an opt-out heatmap feature that anonymously aggregates all activities onto a single map. This allows users to find hot spots and active trails while simultaneously opening up the platform to de-anonymization attacks like inferring users' home addresses. By crawling the publicly available heatmap and through manual validation, we have demonstrated that the home address of highly active users in remote areas can be identified, violating Strava's privacy claims and posing as a threat to user privacy.**

## 1. Introduction

With over 100 million users, Strava is one of the most popular fitness-tracking applications in the world [1]. Strava allows users to upload different data pertaining to activities, including time-stepped GPS data, heart rate, cadence, and more. With this data, Strava users can track their fitness over time, share their activities with their friends, or publicly post activities for anyone to see. Users can also compare their time on similar routes, compete for KOMs (King Of Mountains) against friends, or enter public challenges. There are additional metrics that Strava generates for fitness and fatigue levels, along with some coaching methods such as suggesting to stay within certain heart rate zones.

One particular feature that Strava creates based on GPS data is the Strava heatmap. Updated monthly, the Strava heatmap takes the last two years of GPS data from participating users and aggregates it into a single map highlighting active areas with bright yellow and white lines. Participation in the Strava heatmap is set as a default and can be turned off within the privacy settings. Strava users can utilize the Strava heatmap to discover popular running, cycling, and swimming areas.

Upon account creation, the user is prompted with an array of privacy settings relating to shared activities. Activities can be listed as private, for followers, or for the public to view. Separate from this, there are privacy settings that determine data shared with the Strava heatmap platform, with the wording in Figure 1 claiming that the data is de-identified. Our research challenges that claim by de-anonymizing users.



Aggregated Data Usage

☑ Contribute your activity data to de-identified, aggregate data sets

When you choose to contribute your activity data using this checkbox, your data is de-identified and aggregated with other athletes' activity data to support our community-powered features such as Metro, Heatmap, Points of Interest and Start/End points. These aggregate data sets do not include activities set to 'only you' visibility.

**Figure 1:** A photo of the privacy setting that determines if a user's data is shared with the Strava heatmap.

In areas with many highly active Strava users, the Strava heatmap data is difficult to tie to a specific user due to the fact that potentially hundreds of athletes are contributing to the heat in that area. No name or account information is tied to the heat generated; however, in areas with only a few active Strava users, the heat generated by one individual can be clearly visible. In this paper, we look into the Strava heatmap and how, in some situations, these areas of high heat can be used in conjunction with user metadata to reveal the home addresses of Strava users. Additionally, we propose two potential mitigation strategies to reduce the effectiveness of the attack outlined in this paper.

## 2. Related Work

This study is not the first to document the security risks associated with Strava. Hassan et al. in 2018 demonstrated that the home addresses of public accounts could be identified through publicly shared posts even if they used Strava's in-house obfuscation technique [2]. At the time, the method of choice was to create a radius of obfuscation around a user's home address. However, home addresses could be revealed using triangulation with the existing data points outside the radius. Shortly following this research, Strava switched to hiding the first sections of activities. For instance, the first and last eighth of a mile may be obfuscated from a workout summary. This method is more effective, but it is still possible to break the protection. Using the discrepancy between reported distances and the distance reflected on shared maps, researchers recently were able to break 85% of the endpoint privacy zones [3].

Our research is related to data privacy and the ability to find home addresses for users, but the data set for our research is separate. As opposed to using the activities of individuals, we are using the aggregated and public Strava heatmap as our data source. Since the data is anonymized, Strava does not apply the same hidden zone feature that is standard for shared activities. Per Strava, *"data within hidden zones of activities that are shared with "Followers" or "Everyone" will now be used in de-identified aggregated data"* [4]. Because of this, previous studies and their impacts do not mitigate the attack method demonstrated in this paper. Additionally, the previous studies utilized public accounts and people willing to share their data publicly, including individual activities. Since the data being shared with the Strava heatmap can be from public and private accounts, this attack can be used on both public and private users.

The Strava heatmap has proven to be a privacy risk in the past, not only to individuals but to military forces worldwide. In 2018, a student from Australian National University found that the Strava heatmap highlighted the locations of military bases and outposts [5]. This issue has been addressed by national governments, not Strava, by restricting the use of fitness apps on military bases.

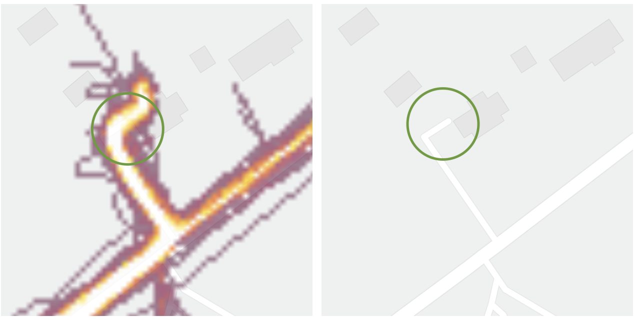
It has been demonstrated that there are privacy concerns relating to the fitness tracking ecosystem. Where other research utilizes the data that users willingly share with their follower base, our research investigates a data source that on the surface appears to be anonymous, but in reality is a risk to user privacy.

## 3. Methodology

Our methodology has followed two different approaches during the course of our research. A case study in New Jersey acted as motivation and a proof of concept that this was a viable attack. Figure 2 represents how a home can clearly be the origin point for a large amount of heatmap activity. In this particular case study, when searching the name of the user's city in the Strava search bar, only one user was active enough to produce these levels of heat, and external sources confirmed this individual to be the one generating the heat. The manual approach was expanded to different cities in different states following the methodology of searching for cities with heat only generated from one address, searching the city for the users, and identifying if the user and the point of interest matched using extraneous sources such as voter registration records.

After the proof of concept, an automated approach using crawling and public voter records was developed to understand the implications on a large scale. The automated approach was a four-step pipeline, including screen capture, image analysis, user crawling, and inference analysis.

### 3.1. Heatmap Capture



**(a)** Running and cycling activities **(b)** Swimming activity

**Figure 2:** Comparison of screenshots taken at the same location with swimming or running activity. With running and cycling activity, the heatmap clearly reveals an entry point to a residential location.

Over the course of a one-month period, 491,463 screenshots were captured using Puppeteer [6]. Data was only gathered in Arkansas, Ohio, and North Carolina, the states where we were able to gather ground truth voter registration data. On the Strava heatmap, URL parameters allow the user to specify the coordinates and zoom level of the map. A zoom factor of 17.33 was used as that is the level where house specific data becomes visible on the OpenStreetMaps platform. With that zoom level, screenshots were taken for running, biking, and swimming data over a rectangular area containing the whole state. The crawler would navigate to a location on the left side of the rectangular area, screenshot the map, pan right, and repeat for each row and each map type. For each screenshot, the URL header provided coordinate data for the center pixel, and allowed us to match each photo to a specific zip code. Since a rectangular bounding box was used, some data fell outside the state boarder and was discarded. An approach utilizing a more precise boarder of the state was attempted but was not pursued since navigating to a location based on the search bar was slower than panning across the map.

### 3.2. Endpoint Detection

The objective of the image analysis was to identify a path of heat that clearly originated from a home/apartment address. Thus, any heat on a road or in a location far from a home was discarded. In order to subtract the heat on roads from a given screenshot, we captured a screenshot of the swimming activity at the same location, as shown in Figure 2. Due to the fact that our analysis was solely focused on residential areas, there was almost never any recorded swimming activity except for occasional heat in backyard pools. This meant that the swim screenshots often yielded a clear view of the roads and houses without any heat. From these screenshots, we created a bit-wise mask of the roads and subtracted the mask from the corresponding screenshots of running and cycling activity in the same location. With the roads and the heat on the roads removed from the screenshots, all that was left was the heat at route endpoints where the user left the road.

On the heatmap, the homes are depicted as dark grey squares and the street is clearly shown as a lighter grey path. Figure 3 showcases the outcome of our image analysis technique, where red dots have been placed on the objects that the tool classified as houses and a large purple dot

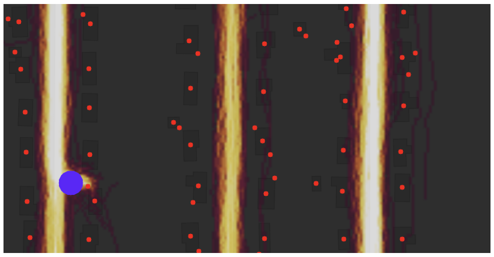
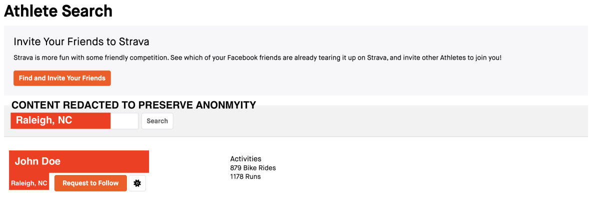


**Figure 3:** Visualization of the image analysis. Red dots represent houses, the purple dot represents heat originating from a home.



Athlete Search

Invite Your Friends to Strava

Strava is more fun with some friendly competition. See which of your Facebook friends are already tearing it up on Strava, and invite other Athletes to join you!

CONTENT REDACTED TO PRESERVE ANONMYITY

Raleigh, NC

John Doe

Raleigh, NC

Activities
879 Bike Rides
1178 Runs

**Figure 4:** The Strava Search functionality yields users based on searching a city along with metadata relating to how active the user is. In order to preserve user privacy, the search parameters, the name of the user, and other personally identifiable information has been redacted from this picture.

has been placed on the spot which the tool classified as a valid activity endpoint. The image analysis overlaid the road, homes, and heat to identify what areas were of interest to our research. From these points, the pixel location and coordinate data of the center of the screenshot allowed the exact coordinates of the heat to be identified. These points were stored along with their corresponding zip code. These points would later be used and compared to the voter registration records.

### 3.3. User Crawling

To build the second data set in this study, the user search functionality of Strava was utilized. The search bar prompts users with placeholder text to enter ”Athlete Name”, but as depicted in Figure 4, searching the name of a city will display all users with that city listed in their location. This is not an advertised feature of the search functionality and whether or not it is intended to used in this manner is unclear. Our crawler utilized this mechanism by searching each city in a state and paginating until no more results were yielded. For each account crawled, we searched the voter records for a matching user. For users that matched to no voter record or multiple voter records, we removed the data from our analysis. For users that matched to a voter record, we removed the name information, and created a database that included city name, coordinates for a home address, account type (i.e., public or private), and the number of activities (bike rides and runs). The coordinate information was generated from the home addresses using the Google Maps API [7].

### 3.4. Inference Analysis

The final analysis involved creating two databases. The first database was coordinate values of identified endpoints that was created in our endpoint detection stage. The second database was the result of our user crawling stage and had coordinates for each homes matching to a Strava user. With these two databases, we were able to see the validity of an attacker using the Strava heatmap to find the home address of an individual.

### 3.5. Limitations

There are a few limitations to our inference attacks. First, our threat model assumes users will begin their activities from their home addresses. We acknowledge that many athletes start from a trail, participate in competitions, etc. Combined with some users opting out of the heatmap data, this limits the accuracy of an identified point. Our automated analysis accounts for users not starting from homes by ignoring points not originating from a home address. Second, out of the Strava users we targeted from Ohio, North Carolina and Arkansas, only 37% of them mapped to voter registration data, with many of these records potentially being outdated. Thus, some users in our data set may have been correctly de-anonymized but could not be included in the final analysis due to the limited verifiable ground truth. Lastly, our research hinges upon the search functionality within Strava. For a user to show up in our search, they need to list their home city accurately. Users that do not list their home city upon account creation, or forget to update their listed city after moving, will not be discovered through the methods we have demonstrated.

## 4. Results

During the progression of our research we have followed both manual and automated methods and concluded that the home addresses of specific users could be discovered with effectiveness based on how active a user is and how much heat is generated in that city.

### 4.1. Manual Analysis

We discovered three case studies through manual analysis following the methodology of searching for an interesting point, searching for users in that city, and finally searching the voter records to see if there was a match. In each of these case studies, the user was the only active user in the city, and there was a point clearly represented on the map, as shown in Figure 2. This acted as the proof of concept to motivate our automated analysis.

### 4.2. Automated Attack

Since we have the ground truth data, the methodology for the automated attack followed a similar pattern to the manual analysis. We used image analysis to discover 143,799 points of interest, and crawled the user database in tandem with the voter records to have 11,165 users with both a Strava account and linked voter records. From these data

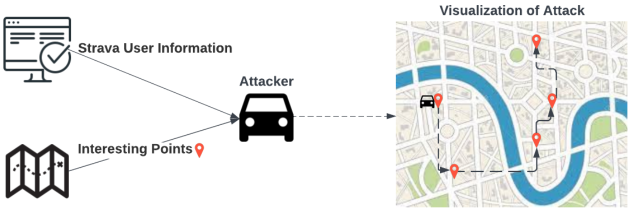


**Figure 5:** Visualization of attacker using Strava to gather user information, and heatmap data in order to preform an attack on a specified individual.

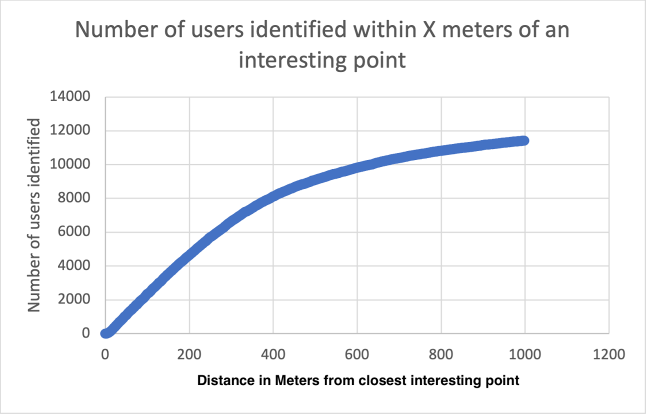


**Figure 6:** The likelihood of a successful based on the search radius for each endpoint.

sets we are able to extrapolate data on a more realistic attack outlined in figure 5. The attack scenario has an attacker that is trying to identify the home address of an individual. Using the Strava search feature the attacker has the user name (and even photos of the user), their home city, access to the Strava heatmap, and knows the number of activities the victim user has posted. Then, using the heatmap data the attacker could identify interesting points to visit (as shown in §3.2) to verify if they found the target individual. Thus, using the heatmap data the attacker is able to narrow down the search space significantly. To evaluate the effectiveness of our proposed attack, we use the voter registration data as ground truth for the home addresses of individuals.

**Successful Attack Rate.** Due to limitations in our data set, users opting out of the heatmap, and some users following patterns that do not identify their home address, a victim may not be identified even if the attacker visits every point of interest. Figure 6 demonstrates the likelihood of a successful attack, depending on the search distance. For the sake of simplicity, we have classified a match as being within 100 meters of a point of interest. Using this threshold, we manually analyzed 20 matches and found a clear notch on the Strava heatmap for 17 out of 20 users. One false positive lived near a trail, and another had a neighbor producing the heat. The final false positive resulted from an imperceptibly small amount of heat, triggering our image analysis tool to recognize a point of interest. With the threshold of 100 meters, there is a 31.7% chance of the user being able to be discovered.

With this threshold identified, another factor that changes the likelihood of a successful attack is how active a user is. A more active user produces more heat on the Strava heatmap and therefore is more easily identified. Figure 7 demonstrates the likelihood of a match based on the number of activities posted.

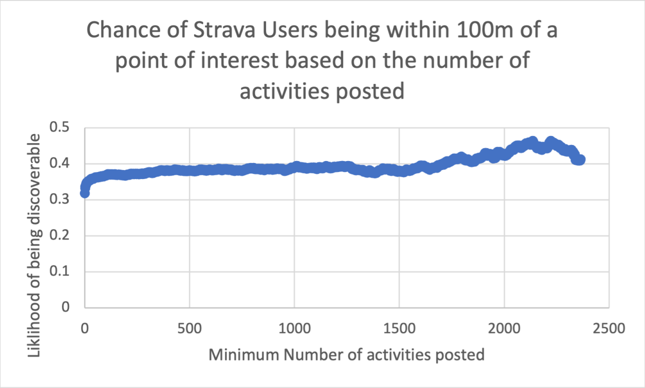


**Figure 7:** A graphical representation on how more active users are more likely to be discovered. To generate the chart, the users with $>= x$ number of activities were considered and the percentage is represented by number of users within 100m of a point of interest/total number of users with $>= x$ activities. The dip is a consequence of a low number of users having more than 2000 activities posted.

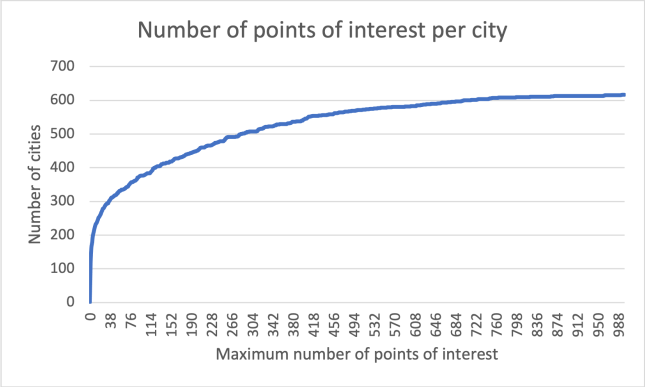


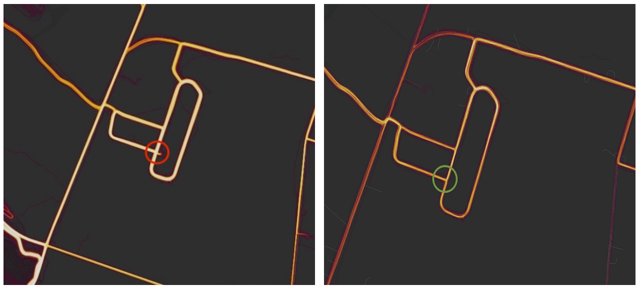
**Figure 8:** A graphical representation of the number of points of interest for each of our 638 cities.The x axis represents the number of visitations an attacker is willing to take and the y axis representing the number of cities in which the attack could be preformed with that level of motivation. 119 Cities had zero points of interest.

For the remainder of the analysis, we will be assuming the target of the attack posts an average number activities, which for our data set is 308 activities. With the 100 meter threshold, and the victim posting 308 activities, the likelihood of being able to be discovered is 37.5%.

**Number of Visitations.** With the 37.5% chance of the attack being successful, the feasibility of running an attack is still in question. Figure 8 represents the number of cities that would be feasible to attack based on the number of visitations an attacker is willing to take. For example, if an attacker was willing to visit 20 locations, there would be 262 out of 638 cities that would be eligible for this attack, with 62.5% of attacks ending in no Strava user being found.

## 5. Mitigation Strategies

Here we offer two solutions that reduce the effectiveness of the attack we have presented. The data we utilized relies



**(a)** Before Mitigation **(b)** After Mitigation

**Figure 9:** Impact of rounding the heatmap trace to the nearest nearby street.

heavily on finding houses that were clearly the starting point of a large amount of heat. To mitigate this, the overall objective is to remove data clearly originating from a home. While the mitigation of privacy risks is paramount, it is also crucial to preserve the original data on the Strava heatmap.

**Targeted Heat Removal.** One option to mitigate attacks on the Strava heatmap is not to store data concentrated near a house object. The OpenStreetMaps platform that the heatmap is overlaid upon clearly demarcates houses, and these demarcations were one key aspect of our approach. If Strava utilized the OpenStreetMaps database of where houses are on a map, they could create a small exclusion area between the house and the road nearby. Figure 9 represents what the difference would look like if Strava excluded the data between houses and the street. The content of the map is not lost, but there is no longer the ability to see which house is the origin of the heat. To further prevent de-anonymizing a specific user, Strava should enforce a minimum number of users per path to prevent these kinds of attacks (this will guarantee '*k-anonymity*').

**Privacy Zones for Heatmap Data.** Another option is to apply Strava's existing hidden zone feature to heatmap data. The Strava hidden zone is intended to allow users to hide the start and end points of their activities before sharing them publicly. As recently as August of 2022, Strava has stated, *"With these updates, data within hidden zones of activities that are shared with "Followers" or "Everyone" will now be used in de-identified aggregated data to help the Strava community"* [4]. Strava could apply the hidden zones to the Strava heatmap and put the power in the hands of the user to how much data they share with the platform. Additionally, utilizing the hidden zone feature reduces the number of data security solutions needed to be implemented, and improvements relating to public posts will assist the robustness of obfuscation on the Strava heatmap. Again aggregating data for a minimum number of users will guarantee '*k-anonymity*'.

## 6. Discussion

The attack described in this paper demonstrates that there is a risk to individuals from the Strava heatmap. While the data on the Strava heatmap is not tied to specific users, the data can be combined with other data sources within the Strava platform to de-anonymize the heat. This de-anonymized heat can then be used to identify the home address of Strava users. This contradicts Strava's aforementioned privacy claims.

**Implications.** The ability to identify the home address of Strava users is a violation of user privacy. It demonstrates that seemingly anonymous data is not truly private and can leak information about users. In addition to contradicting the privacy claims made on registration for the heatmap, the matching of a Strava user to a home address can build a complete profile of an individual, including their workout habits and the paths they frequently travel on. This information can be used for stalking or other invasions of the privacy of individuals. Additionally, on a wider scale, instead of 'John Doe' being just a name tied to an address, 'John Doe' can be categorized as an active individual living with certain workout behavior. This information can be utilized for targeted advertising and individual profiling and is potentially being collected without consent.

**Expectations.** Sharing data with a fitness tracking application comes with inherent risks that should be assumed by the individual choosing to share the data. With that consideration, there is a responsibility of the company to store that data securely and to have the due diligence to ensure publicly shared data is not potentially utilized for malicious purposes.

**Considerations.** In the information economy, there is a trade-off between preserving privacy and the quality of data collected. Looking at the overall culture regarding all data-collecting applications, more effort should be taken to protect the privacy of users. Developers should create products with more consideration given to how the information could be misused. As with Strava, it is often the case that a low amount of development time can be taken to protect privacy while also not impacting the quality of the data. With a wide variety of privacy-preserving techniques being developed, researching and adopting these solutions needs to be an integral part of the development process.

## 7. Ethical Considerations and Disclosure

Given the sensitive nature of the data being handled in our research, we have followed a number of ethical measures. We first contacted our institutional International Review Board (IRB) to obtain approval/clearance to conduct the study. Beyond this, we have insured the integrity of our data by storing it on secured servers only accessible to the research team members. Additionally, the final analysis was only preformed on *coordinate data* (as identifiers) with any name and address information being removed from account before any analysis. Finally, we have not released any data sets and do not plan on releasing any data in an effort to preserve the privacy of Strava users. We have also disclosed our findings to Strava. At the time of writing, we gave Strava around 90 days to respond and take action on their own part.

# References

[1] (2023) Strava - about us. [Online]. Available: https://www.strava.com/about

[2] W. U. Hassan, S. Hussain, and A. Bates, "Analysis of privacy protections in fitness tracking social networks -or- you can run, but can you hide?" in *27th USENIX Security Symposium (USENIX Security 18)*. Baltimore, MD: USENIX Association, Aug. 2018, pp. 497–512. [Online]. Available: https://www.usenix.org/conference/usenixsecurity18/presentation/hassan

[3] K. Dhondt, V. Le Pochat, A. Voulimeneas, W. Joosen, and S. Volckaert, "A run a day won't keep the hacker away," *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2022.

[4] (2022, August) Spotlight: New updates to privacy defaults. you stay in control. - a post by the strava club. [Online]. Available: https://www.strava.com/clubs/231407/posts/21834969

[5] L. Sly, "What he did on his summer break: Exposed a global security flaw," January 2018. [Online]. Available: https://www.nytimes.com/2018/01/30/world/australia/strava-heat-map-student.html?smid=url-share

[6] (2023) Puppeteer. [Online]. Available: https://pptr.dev

[7] (2023) Google maps api - geocoding. [Online]. Available: https://developers.google.com/maps/documentation/geocoding