A Bespoke Framework For Inferring Sleep Patterns

Gabriel Gladstone
dept. of Computer Engineering
Charlottesville, VA
zwy7ce@virginia.edu

***Abstract*—Researchers use a variety of measures including sleep patterns to diagnose depression, anxiety, and other mental health conditions. This study presents a framework for reliably identifying a person's sleep patterns. The study analyzed the online data of 1 participant over 23 days to prove the effectiveness of the bespoke model of data collection and analysis. The study concludes that the techniques presented are effective in aggregating data sources to produce more reliable sleep inferences. However, data should be checked for quality.**

I. Introduction

People leave a massive digital footprint from their online behavior, and only a few pieces of information is needed to uniquely identify someone. About 87% of Americans can be uniquely identified from their 5-digit ZIP, gender, and date of birth ([4]). In this study, I investigate how a person's online behavior can be used to infer their sleep pattern.

As we have covered in this course, people's online behavior can shed light on their mental health, if they are experiencing MDD[1], if they are thinking about suicide, and more. Online behavior monitoring is a scalable and cheap first-indicator that can be used to bring people in need more timely treatment while allowing medical personnel to better focus their efforts given their limited resources.

This study specifically focuses on inferring when a voluntary participant is asleep from their online activity data. The study does not explore the implications of a person's sleep patterns, rather, it focuses on the accuracy of the proposed framework.

The study presents a framework for accurately identifying when a user is asleep because sleep patterns have been used to diagnose depressive symptoms [6][2]. This study is a proof of concept for the bespoke model presented and encourages future researchers to adapt the framework.

II. Methodology

A. Study Design

The study takes place from January 13, 2025 to February 5, 2025. This is a 23 day period from the start of UVA's Spring Academic Semester to the time of writing this report. The singular participant in this study is me, the author of the report. This time period was chosen because of known and regular wake and downtimes in the individual's schedule as shown in Fig 13.

The goal of this experiment was to show that aggregating data sources improves the accuracy of inferring sleep patterns. Two data sources, google location data and Microsoft Edge search data were used. The accuracy of each data source compared to the ground truth alarm were individually analyzed. The data sources were then combined and analyzed over the studies time period. Various sleep identification algorithms and filtering techniques were also explored in the analysis. By showing a combination of just 2 data sources increases sleep inference accuracy the study can extend the claim that more aggregating more data sources will make inferring sleep patterns even more accurate.

B. Data Collection

This section discusses the data sources explored and ultimately chosen for the experiment.

1) Data Source Exploration:

What makes a good data source?

Many data sources were considered for this study. The study was specifically looking for data sources with datetimes, any associated information with the datetimes was anonymized. For this study, data sources that had fine-grained logs were preferred, where any action produced a record with an associated datetimes. While most records were filtered out to infer sleep patterns, fine grained records in theory would produce better start and finish times.

While online applications track user data for their own purposes. General Data Protection Regulation guidelines gives user some protections and control over their data meaning many applications will allow a user to review the data the applicaiton is tracking.

The study explored the possibility of using the following data sources:

- Google Takeout (misc data)
- Google Takeout (Timeline Data)
- Microsoft Account Data
- Apple App Privacy Report
- Local Browser Cache History
- Reddit Activity

The data sources were compared by their quality, ease of collection, and user frequency.

Google Takeout provides a wide variety of data points including email, search, location, app activity and more. It is a high quality data collection platform that aggregates data

sources in a easy to download format. While most data source logs are time stamped, only location data has a high enough user frequency in the right context to justify its use.

Microsoft Account Data is Microsoft's Takeout equivalent platform that logs activity related to Microsoft apps. It also aggregates data from search, email, app activity and more. However, data from the participant's school account was restricted and their personal account was infrequently used.

Apple App Privacy Report tracks data and sensor access, app, and network activity. It logs most phone activity. The participant of this experiment very frequently uses their mobile phone. However, this feature was not enabled for the time period of the study so it was not used.

Browser search history, including Microsoft Edge searching datetimes and caches all search results locally. This makes it very easy to extract and since a online user constantly searches the web for work and recreation, it is a high quality frequent source of information.

Reddit activity was seen as an option for its frequent use by the participant, but the history of posts viewed is not tracked by Reddit and would be cumbersome to infer from a 3rd party setup. Reddit allows users to download their comment, upvote/downvote, and post history, but for passive users of Reddit, this data is too infrequent.

2) Chosen Data Sources:

After analyzing the quality of data, ease of collection, and frequency of use by the participant, Google Takeout location data and browser search data were chosen data sources.

As seen in the analysis, frequency of use of a platform is dependent on the specific user and ease of collection is dependent on the resources of the organization, thus no data source is generalizable for analysis among all participants in all user studies and researchers should rely on a bespoke strategy for data collection.

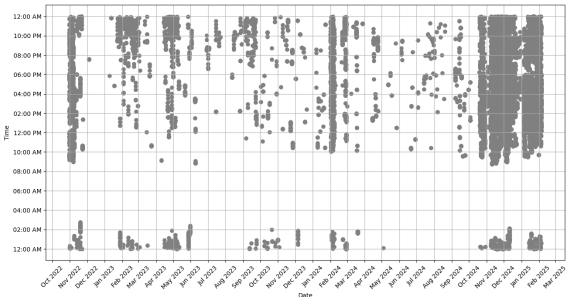

Fig. 1: Original Search Data

The unfiltered search data can be seen in Fig 1. All the participants search results over the past 3 years from Microsoft Edge, Firefox, and Chrome are shown. Each data point represents a search result. As can be seen the data quality from November 2024 to March 2025 is high making this data source a great candidate for the study given the study window.

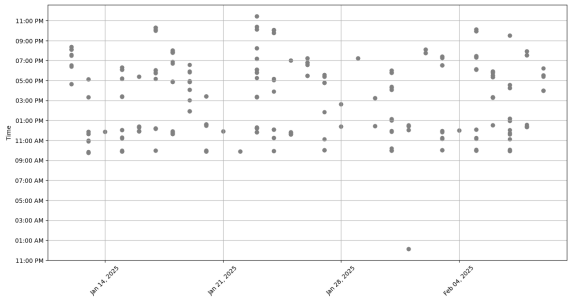

Fig. 2: Original Location Data

The unfiltered location data can be seen in Fig 2. Google Timeline records a user's movement when they change physical locations. Given that to go to class I have to leave my dorm and arrive at a destination about 10 min away, I thought these records would be good indicators for when I woke up. The Google Timeline tracking was enabled at the start of the the school year so the data fully covers the study's window. However, because I don't go to bed immediately after arriving at my dorm, I understand that this won't provide the best data for going to bed.

C. Data Processing

The raw data sources as shown in Fig 2 and Fig 1 go through the following data processing stages.

1) Data is extracted from its respective sources.
2) The data outside the window of the study is discarded.
3) Start (or wake) and Finish (or sleep) times are identified.
4) Extraneous start and sleep times within certain nonsensical windows are discarded.
5) The filtered data is analyzed.

1) Data Extraction:

Browsing history can be stored in the cloud, but if an account is misconfigured, a user is not logged in, or a user uses multiple accounts such as work, school, and personnel, then extracting this data from the cloud will not be ideal. This is why browser data was extracted from the local copy stored on the end users machine, all search results for a specific engine are stored here. For instance, Microsoft Edge Search data is stored at

`C:\Users\<user>\AppData \Local\Microsoft\ Edge\User Data\Default\`

Extracting from the local cache reduces setup burden on the user for the study. A browsing history viewer application[5] was used to extract the Microsoft Edge, Firefox, and Chrome search results and exported to a CSV file. Only the datetime of the record was kept. The URL, title of search, and other metadata is redacted.

Google Location data was exported as a one-time export from Google Takeout as a JSON. The timeline path was removed for lack of fine-grained time records. Only the endTime and startTime fields of the visit data were analyzed.

Both csv and JSON data sources were ingested as a data frame for further analysis using Python.

2) Date Filtering:

As described in the study design, only data between January 13, 2025 to February 5, 2025 is relevant to our analysis.

```python
def filter_by_date(df, start_date, end_date):
    # Ensure inputs in datetime format
    df['Date'] = pd.to_datetime(df['Date'])
    start_date = pd.to_datetime(start_date)
    end_date = pd.to_datetime(end_date)

    # Filter records
    return df[(df['Date'] >= start_date) & (df['Date'] <= end_date)]
```

Fig. 3: Date Filtering Code

After extracting datetimes into a dataframe, out of window records are removed. The study uses the datatime class[3] to easily manipulate and transform data as shown in Fig 3.

3) Sleep Interval Identification:

To infer when a participant is asleep, I identify when they went to sleep and when they woke up. I use two methods to identify sleep based on the context of the data being analyzed. The following assumptions were also considered in the creation of these methods.

1) The participant does not take naps throughout the day.
2) Roughly a 7 hour window when the participant goes to sleep or wakes up is known.
3) The known 7 hour sleep/wake window does not change within the analysis window.

These core assumptions specific to the singular participant of the study guided the creation of the sleep identification techniques. These techniques are only generalizable to other participants where the same assumptions are met. This information could be found through a survey when recruiting participants for a study. Further tuning will be needed if any assumption is broken.

The first technique finds two chronological records with a time difference between 6 and 15 hours. The earliest record is identified as the time a user goes to sleep, and the later record is identified as the time the user wakes up. The delta window is mostly arbitrary, but assumes that a user never gets less than 6 hours of sleep and never get more than 10 hours of sleep + lag time. This window can be adjusted if necessary.

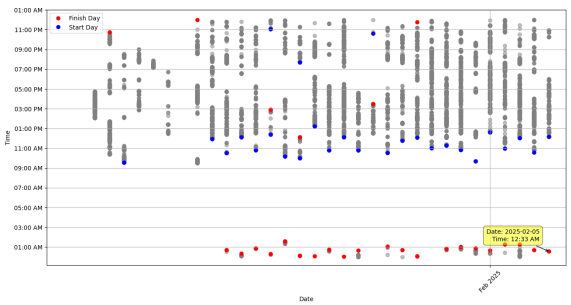

Fig. 4: Sleep Pattern Identification with Time Difference Technique

The technique was used to identify the sleep patterns from search results as shown in Fig 4. From observing the data, it can be seen that this technique generally identifies the earliest start time and latest finish time. The technique is computationally efficient and doesn't identify times on days lacking data, meaning the difference will be too large. However the technique does identify some outliers, like a Finish Day at 3 PM and Start Day at 11 PM pair that are not likely to have happened. These data points were generated from small pockets of infrequent activity which is normal. To filter out this extraneous data in the next step it is necessary to know a $\approx n$ hour normal wake and sleep window.

This time difference technique is good, because it considers the context of the previous day to determine sleep patterns, but due to the infrequency of location data another technique was used. A latest sleep and earliest wake-up time was inspired from a Google and YouTube use study on Depression [6]. The earliest time after 4am was determined to be the start time and the latest time before 4am was the finish time. This threshold was it is assumed the participant will not sleep after this time or wake up before. This threshold time will be different for every participant. Sleep identification for the location data is shown in Fig 5. Like the difference technique, extraneous points must be filtered.

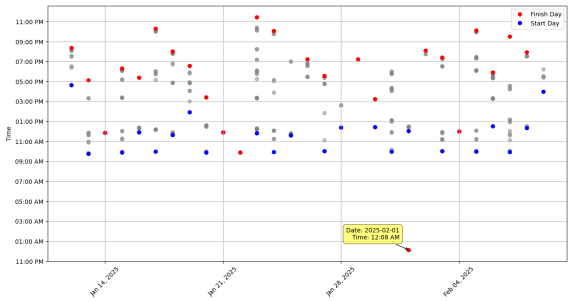

Fig. 5: Sleep Pattern Identification with Earliest Time Technique

4) Extraneous Points Removal:

After identifying sleep data points. Extraneous points were removed. I recommend a window > 7 hours to handle variability and consider lag/lead times from the data. In this case the Start exclusion window was 5pm - 3am and the Finish Day exclusion window was 5am - 5pm.

III. Results

The fully processed search data 6, location data 7, and combined datasets 8 are shown.

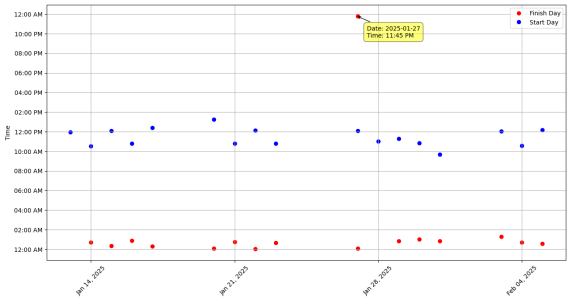

Fig. 6: Processed Search Sleep Pattern Identification

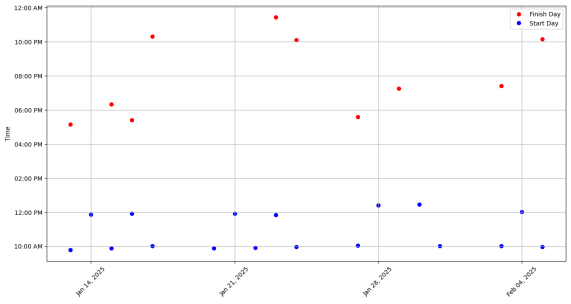

Fig. 7: Processed Location Sleep Pattern Identification

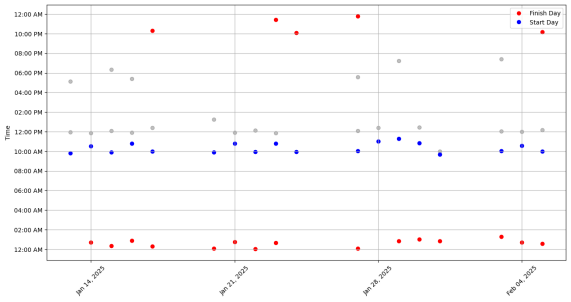

Fig. 8: Processed Combined Sleep Pattern Identification

The data sources were analyzed for their sleep and wake-up times, specifically weekdays, M/W/F, and T/Th as the participants scheduled was consistent during the week. Out of all the datasets, the latest finish times and earliest wake times were considered with the unused times in gray. As can be observed, the combined data set uses the best times out of both datasets that are closer to the true value 13.

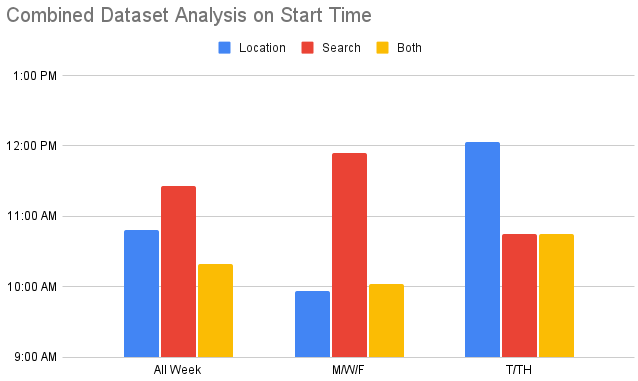

Fig. 9: Mean Start Time Analysis over Weekdays

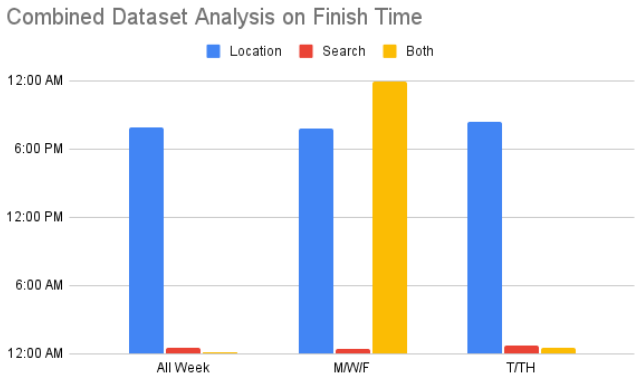

Fig. 10: Mean Finish Time Analysis over Weekdays

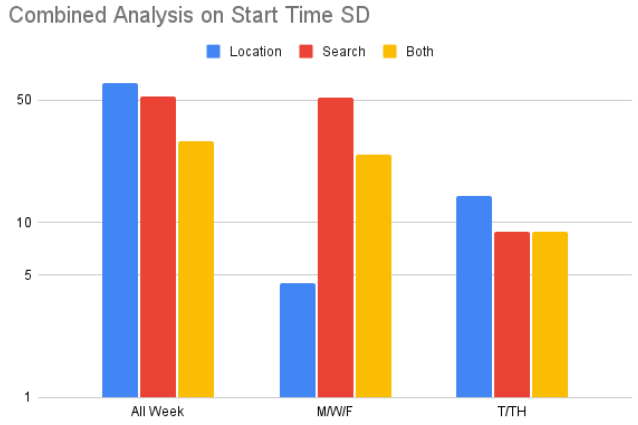

Fig. 11: Standard Deviation Start Time Analysis over Weekdays

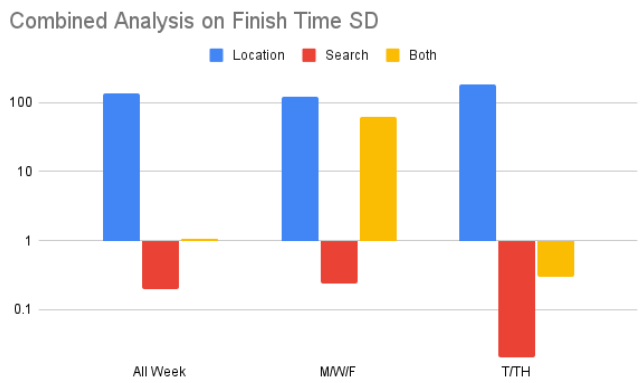

Fig. 12: Standard Deviation Finish Time Analysis over Weekdays

For each data source, the mean and standard deviation of the start and finish times were analyzed for weekdays, only Monday/Wednesday/Friday, and only Tuesday/Thursday. The figures (12, 10, 11, 9) show the benefits of a combined dataset.

We know the ideal start times are 9:10AM on M/W/F and 10:30AM on T/Th as outlined in Fig 13. On T/Th, the combined dataset 9 is tied for closest to 10:30AM and on M/W/F it is within 6min of the location data. Overall it produces the closest times to the ground truth on average. However, having a lag or lead time is fine if it is consistent. Throughout the week on average, the combined time also has the lowest STD, which means it is closest to the actual time and the most consistent, given that the ground truth standard deviation is 0 11.

We know, from the participant, that the finish time is around 1AM every day. While the combined dataset seems to perform well, search data alone is closer to this time on average and is more consistent. This outcome is not expected and is likely due to the poor quality of finish times from location data being incorporated on days that search data omits. More datasets will likely resolve this problem, but this finding shows that the quality of data is important as in data sparse studies it can help like for start times, or hurt, like for finish times.

Overall, the start time has a 1 hour and 9 minute lead time with a standard deviation of 29 minutes and the finish time has a 52 minute lag time with a standard deviation of 1 minute.

IV. Discussion

A. Ethical Considerations

The singular participant in this study consented to their data collection and analysis. While the data may be used to infer sensitive characteristics about the individual such as if they are experiencing depressive disorders, no mental characterization of the individual was carried out in this study. All sensitive data was also anonymized before being used or published.

B. Limitations

As described in data collection, some sources such as mobile phone data wasn't used because it wasn't properly setup within the window of the study. Other high quality sources weren't used because the user simply didn't use the platforms on a regular basis. The study also mainly relied on personal private data that will be hard to scale for a healthcare monitoring system.

C. Future Work

This study was a proof of concept that aggregating data sources instead of focusing on one data source produces the best results. Future work could identify and integrate new data sources, create new techniques, or create a database about people's sleep times to improve the arbitrary time windows or thresholds used for filtering in this study.

D. Conclusion

In this paper I showed the importance of choosing the right data source for a given participant and showed techniques to determine sleep patterns. I showed that more datasets can be beneficial, but the data quality must be reviewed. I used a mix of user survey and white-box techniques to process the data for analysis. The end result is a consistent and close approximation of the start and finish times from a person's online activity.

References

[1] Munmun De Choudhury et al. "Predicting Depression via Social Media". In: *Proceedings of the International AAAI Conference on Web and Social Media* 7.1 (2013). Number: 1, pp. 128–137. ISSN: 2334-0770. DOI: 10.1609/icwsm.v7i1.14432. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14432 (visited on 02/11/2025).

[2] Asma Ghandeharioun et al. "Objective assessment of depressive symptoms with machine learning and wearable sensors data". In: *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). ISSN: 2156-8111. Oct. 2017, pp. 325–332. DOI: 10.1109/ACII.2017.8273620. URL: https://ieeexplore.ieee.org/document/8273620 (visited on 02/11/2025).

[3] *pandas.to_datetime — pandas 2.2.3 documentation*. URL: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.to_datetime.html (visited on 02/11/2025).

[4] Latanya Sweeney. "Simple Demographics Often Identify People Uniquely". In: . *Pittsburgh* ().

[5] *View the browsing history of your Web browser*. NirSoft. URL: https://www.nirsoft.net/utils/browsing_history_view.html (visited on 02/11/2025).

[6] Boyu Zhang et al. "The Relationships of Deteriorating Depression and Anxiety With Longitudinal Behavioral Changes in Google and YouTube Use During COVID-19: Observational Study". In: *JMIR Mental Health* 7.11 (Nov. 23, 2020), e24012. ISSN: 2368-7959. DOI: 10.2196/24012. URL: http://mental.jmir.org/2020/11/e24012/ (visited on 02/11/2025).

V. Appendix

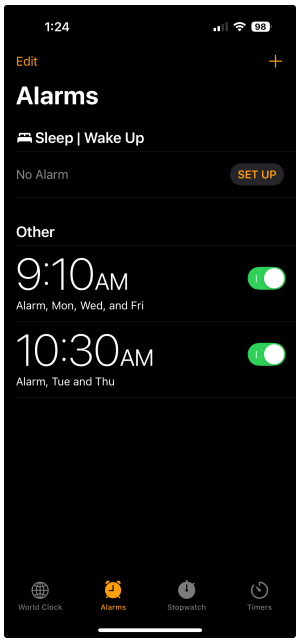

Fig. 13: School Schedule Alarms