Harnessing the Web for Population-Scale Physiological Sensing: A Case Study of Sleep and Performance

Tim Althoff*
Stanford University
althoff@cs.stanford.edu

Eric Horvitz Ryen W. White
Microsoft Research
{horvitz, ryenw}@microsoft.com

Jamie Zeitzer
Stanford Center for Sleep Sciences and Medicine
jzeitzer@stanford.edu

ABSTRACT

Human cognitive performance is critical to productivity, learning, and accident avoidance. Cognitive performance varies throughout each day and is in part driven by intrinsic, near 24-hour circadian rhythms. Prior research on the impact of sleep and circadian rhythms on cognitive performance has typically been restricted to small-scale laboratory-based studies that do not capture the variability of real-world conditions, such as environmental factors, motivation, and sleep patterns in real-world settings. Given these limitations, leading sleep researchers have called for larger *in situ* monitoring of sleep and performance [39]. We present the largest study to date on the impact of objectively measured real-world sleep on performance enabled through a reframing of everyday interactions with a web search engine as a series of performance tasks. Our analysis includes 3 million nights of sleep and 75 million interaction tasks. We measure cognitive performance through the speed of keystroke and click interactions on a web search engine and correlate them to wearable device-defined sleep measures over time. We demonstrate that real-world performance varies throughout the day and is influenced by both circadian rhythms, chronotype (morning/evening preference), and prior sleep duration and timing. We develop a statistical model that operationalizes a large body of work on sleep and performance and demonstrates that our estimates of circadian rhythms, homeostatic sleep drive, and sleep inertia align with expectations from laboratory-based sleep studies. Further, we quantify the impact of insufficient sleep on real-world performance and show that two consecutive nights with less than six hours of sleep are associated with decreases in performance which last for a period of six days. This work demonstrates the feasibility of using online interactions for large-scale physiological sensing.

1. INTRODUCTION

Maintaining optimal cognitive performance has been found to be important in learning [26], productivity [16], and avoiding industrial and motor vehicle accidents [16, 20]. Studies have demonstrated that cognitive performance varies throughout the day [43], likely influencing the quality of our efforts and engagements–including how we use and interact with vehicles, devices, resources, and applications. Furthermore, cognitive performance is decreased significantly after loss of sleep [20]. Understanding the real-world impact of sleep deficiency is critical. It has been estimated that the cost of fatigue to U.S. businesses exceeds $150 billion a year in absenteeism, presenteeism, workplace accidents, poor and delayed decision-making and other lost productivity on top of the increased health care costs and risk of disease [24]. Despite the important influences, temporal variations of real-world performance are not well understood and have never been characterized on a large scale [39].

Models of daily patterns in human cognitive performance rely typically on representations of three biological processes: *circadian rhythms* (time-dependent, behavior-independent, near 24-hour oscillations) [43], *homeostatic sleep pressure* (the longer awake, the more tired you become) [13], and *sleep inertia* (performance impairment experienced immediately after waking up) [4, 19].

While models of these biological processes capture well the patterns of cognitive performance in the laboratory [4, 13], they are based on experimental studies in which participants are deprived of sleep and undertake regular, artificial tasks to measure performance instead of non-intrusively capturing performance through everyday tasks in real-world environments. In addition, these studies typically include participants that fit a specific physical and psychological profile (*e.g.*, those with depressed mood are often excluded). Further, participants in an artificial setting can be influenced by their understanding of the study and subconsciously change their behavior to fit the interpretation of its motivation and goals [35]. While laboratory studies have been critical in developing understandings of the basic biological processes that underlie cognitive performance, they fail to account for myriad influences in the real-world, including motivation, mood, illness, environmental conditions, behavioral compensation including caffeine intake, and sleep patterns in the wild that are far more complicated than those enforced in research studies. How these and other factors alter real-world cognitive performance is not well understood. Therefore, sleep scientists have called for large-scale real-world measurements of performance and sleep as a necessary step to "to transform our understanding of sleep" and "to establish how to manage sleep to improve productivity, health and quality of life" [39].

This Work. We respond to the appeal from the sleep research community with a large-scale study of sleep and performance enabled through reframing everyday interactions with a web search engine as a series of performance tasks. In particular, we use individual keystrokes when typing a search query and the clicks on search results as a source of precisely timed interactions. We demonstrate that the timing of these interactions varies based on biological processes and can be used to study the influence of different quantities of sleep on performance. Search engine interactions offer insight

*Research done during an internship at Microsoft Research.

WWW 2017, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4913-0/17/04.
http://dx.doi.org/10.1145/3038912.3052637

about real-world cognitive performance as they are an integral part of many people's lives and work every day. More than 90% of US online adults use web search engines, which now handle billions of searches each day [38].

Our dataset comprises over 3 million nights of sleep tracked by wearable sensors from 31 thousand users over a period of 18 months and 75 million subsequent real-world performance measurements based on keystrokes and clicks within a web search engine (Section 3). This constitutes the largest prospective study of real-world human performance and sleep to date (more than 400 times larger than the second largest comparable study which had only 76 participants [29]).

We first demonstrate that real-world human cognitive performance captured through search engine interactions varies throughout the day in a daily rhythm (Section 4). We find that performance is lowest during habitual sleep times when it is reduced by up to 31%. Both the shape and magnitude of this temporal variation are consistent with controlled laboratory-based studies, providing validation of our large-scale performance measures. We also show that performance varies based on chronotype (morning/evening preference) with early risers performing slowest at 04:00 h (4am) and late risers performing slowest at 07:00 h.

We then develop a statistical model based on chronobiological research and demonstrate that it successfully disentangles circadian rhythms, homeostatic sleep drive, sleep inertia, and prior sleep duration—key factors considered in the sleep literature (Section 5). We quantify that performance varies by 23% based on time of day, by 19% based on time since wake up, and by 5% based on sleep duration (Section 5.3). We validate our methodology by demonstrating close agreement between our model estimates based on a large amount of performance measurements in the wild and smaller controlled sleep studies in artificial laboratory settings.

After validating our approach, we extend prior laboratory-based sleep research through estimates of how sleep impacts performance in real-world settings. In particular, we quantify the impact of one or multiple nights of insufficient sleep on real-world performance (Section 6). We demonstrate that very short and very long sleep durations, and irregular timing of sleep are associated with 3%, 4% and 7% lower performance, respectively. We also show that two consecutive nights with fewer than six hours of sleep are associated with significantly decreased performance for a period of six days.

Our study is also the first to demonstrate that ambient streams of data, such as patterns of interactions with devices, can be harnessed as large-scale physiological sensors to study and continuously and non-intrusively monitor human performance at population scale. The insights and methodology developed in this work are relevant to sleep scientists in pursuit of larger-scale real-world measurements of performance, to computer scientists who build tools and applications that may be affected by variations in human performance, and to the growing community of researchers who have been exploring uses of data from online activities to address questions and challenges in the realm of public health.

2. RELATED WORK

Circadian Processes in Sleep and Performance. Empirical studies have found daily rhythms in human performance including alertness, attention, reaction time, memory, and higher executive functions such as planning [11]. The daily variations in performance have been found to be modulated primarily by two processes [18]: a *circadian rhythm* (time-dependent, behavior-independent, near 24-hour oscillations) [43] and a *homeostatic sleep drive* (the longer awake, the more tired we become and the more we sleep, the less tired we become) [13]. The circadian rhythm acts in opposition to the homeostatic drive for sleep that accumulates across the day, enabling a single, consolidated period of wakefulness throughout the day. A third process has been proposed called *sleep inertia* [43], which corresponds to the performance impairment experienced immediately after waking up [4, 19]. In addition to the influence of daily rhythms on the structure of sleep and performance, there are also shorter, 90-minute oscillations, *ultradian rhythms*, that organize the occurrence of NREM and REM stages during sleep. Ultradian rhythms, circadian rhythms, and homeostatic sleep pressure can all impact the structure, and likely function, of sleep [17].

Human preferences and natural tendency in the relative timing of sleep and wake are called *chronotypes* and are at least partly based on genetics [40]. Cognitive performance depends on chronotype and time of day [31]; that is, early/morning types ("lark") tend to be higher performing earlier in the day while late/evening types ("owl") are higher performing later. Sleep deprivation has been linked to significant decreases in cognitive performance that lead to increased risk for accidents and injury [20].

A recent study correlated performance on cognitive exercises with a sleep measure based on retrospective self-reports of "typical sleep" in 160 thousand users [42]. However, this measure suffers from potential biases [28] and does not enable the study of performance variation over time based on time of day and sleep timing. Another study showed that insomnia with short sleep is associated with cognitive deficits in 678 subjects [22] but only measured a single night of sleep to characterize typical sleep patterns after taking performance measurements, leading to similar limitations. According to a recent meta-analysis [29], the largest study that measured both sleep and performance concurrently had 76 participants.

Technology Use and Interaction Patterns. Interaction patterns of different devices and applications have been studied on small scale to better understand mobile device usage [12], to detect stress [44], used as biometric signals for authentication [32], and linked to biological processes [33, 34] including alertness [1]. For example, less sleep was linked to shorter duration of focus of attention in a study with 40 participants [30]. Large-scale interaction data have been used to gain insights into human behavior in the areas of mood rhythms [23], diet [46], conversation strategies [5], social networks and mobile games encouraging health behaviors [7, 8, 41], and health and disease-related search behaviors [36, 47].

This Work. Existing research on sleep and performance is either small-scale and laboratory-based [29] or relies on subjective measures such as surveys capturing "typical" sleep [42] which do not allow for temporal coordination of sleep and performance measurements. As a complement and extension of research to date on performance in artificial laboratory settings, we study real-world cognitive performance which we measure through interactions with a web search engine. We use objective measurements of sleep (time in bed) from wearable devices which are preferred to subjective self-reports that can be significantly biased [28] and that enable us to study performance variation over time in reference to sleep timing. This work represents the largest study of objectively measured sleep and real-world performance to date, employing a subject pool that is orders of magnitude larger than the largest comparable prior study [29]. Our study demonstrates on a large scale that interactions with devices are influenced by biological processes and sleep.

3. DATASET

Our dataset contains over 75 million search engine interactions and sleep measurements for 31,793 US users of Microsoft products who agreed to link their Bing searches and Microsoft Band

Dataset Statistics	
Observation period	18 months
# users	31,793
# nights of sleep tracked	3,102,209
# queries	24,590,345
# filtered queries with clicks	6,906,791
# keystrokes extracted	68,779,113
# total interactions	75,685,904
Average keystroke time	225ms
Average click time	9.28s
Median age	38
% female	6.1%
% underweight (BMI < 18.5)	1.4%
% normal weight (18.5 $\leq$ BMI < 25)	32.4%
% overweight (25 $\leq$ BMI < 30)	39.2%
% obese (30 $\leq$ BMI)	27.0%
Median time in bed	7.26h

Table 1: Dataset statistics. BMI refers to body mass index.

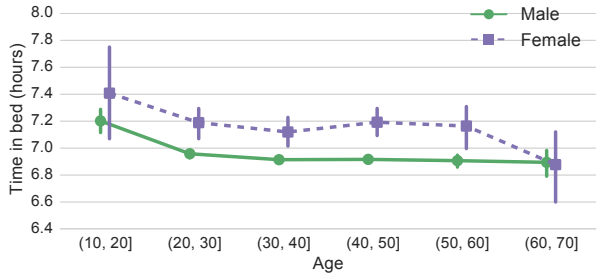

Figure 1: Average sleep duration across age and gender. Our measurements are consistent with previous estimates [10, 14, 45] (Section 3). Error bars in all figures correspond to 95% confidence intervals of the corresponding mean estimates.

data for use in generating additional insights or recommendations about their sleep or activity. Basic dataset statistics and demographic information on the users are summarized in Table 1. Demographic variables (age, gender, body mass index) are self-reported through the Microsoft Health app. While the user age and overweight/obesity status closely track official estimates in the United States, we note that our sample is predominantly male.

Performance. We measure performance through the timing of two types of interactions with a search engine (Microsoft Bing): (1) individual keystrokes within the search box that are tracked by the search engine so it can automatically suggest query completions, and (2) clicks on the result page after a search query. Section 4.1 provides more details on each of these measures and we discuss how to account for potential confounds such as the type of query in Section 5.1. We exclude search engine interactions originating from mobile devices since such interaction patterns and timing are fundamentally different from those on desktop devices. While users could potentially access the search engine from multiple machines, we note that for most users this is unlikely to be the case and that using different keyboards and mice throughout the day is unlikely to explain the timing differences observed in this work.

Sleep. Sleep data from wearable devices provides objective measurements which have been preferred to subjective self-reports that may be significantly biased [28]. To estimate sleep, we consider signals from wrist-worn activity trackers (Microsoft Band) that include a 3-axis accelerometer, gyrometer, and optical heart rate sensor. The Microsoft Band employs internally validated proprietary algorithms for estimation of sleep and we focus on duration of time in bed (herein referred to as "sleep duration"). Time in bed is delineated either by manual input of the user (*i.e.*, explicit taps on the device before going to sleep and immediately after waking up) or automatically based on movement if the user does not provide manual input. The use of an event marker to denote bed timing is widely used in sleep research in lieu of or in concert with sleep diaries [9]. Following standard practice [45], we exclude any sleep duration measurements below 4 and above 12 hours of time in bed.

As evidence that our sleep measurements have face validity, we show that they match published sleep estimates. Figure 1 illustrates average time in bed across age and gender. Time in bed decreases with age and is higher in females than males consistent with published estimates [10, 14, 45]. Walch et al. [45] report very similar times and a difference of 17 minutes between females and males. With the exception of 60 to 70 year old subjects, we find differences between 12 and 17 minutes. There is no difference for older subjects, which matches survey-based estimates by Basner et al. [10]. We take these alignments with published research as evidence for the validity of using wearable device-based sleep data for large-scale population studies of sleep and performance.

4. PERFORMANCE MEASURES BASED ON INTERACTIONS DURING SEARCH

Next, we describe two human performance measures derived from search engine interactions that we use to study daily variation in performance. We show how these measures exhibit variations in performance over time and based on chronotype (morning/evening preference) consistent with findings from laboratory-based sleep studies. This demonstrates that performance signals generated from everyday search engine interactions vary based on biological processes. We model these processes and influences explicitly in Section 5.

4.1 Performance Measures

We study two real-world performance measures in this work since it is possible that different measures would respond differently to sleep deprivation as sleep studies have shown differential effects of sleep deprivation on different measures of cognition.

Keystroke Time. The first measure is based on keystroke timing. The search engine's search box registers every single keystroke and sends a request for query completions to the search engine's servers. We use the timing between two such requests as the time of a single keystroke if the two queries are different by exactly one character (not every request is received on the server side) and within two seconds (larger times indicate longer thought processes or separate sessions). This threshold is sensible as an average keystroke by an average typist takes about 240 milliseconds (50 words per minute at 5 characters per word [15]).

Click Time. The second measure is based on the time to click on a search result after a search result page is displayed. We measure the time between the search query and the first click on any result on the first page. Click times over two minutes are excluded since they might stem from interrupted sessions. We account for click position and query type as described in Section 5.1.

We believe that investigating measures that capture performance on two different tasks provides robustness and breadth to our analyses. The two tasks rely on different mixes of sensing, reflection, planning, and formulating, executing, and monitoring of mo-

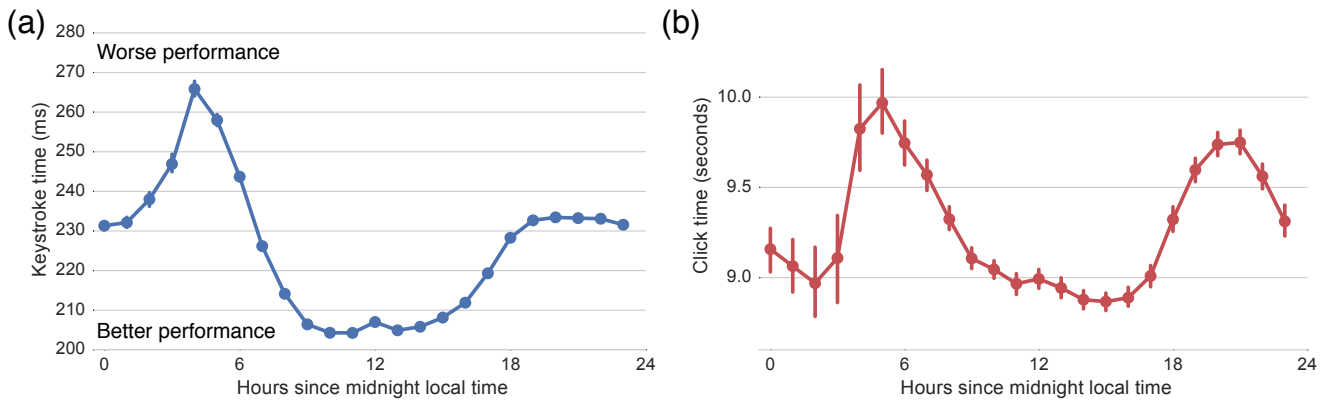

Figure 2: Time of day-dependent variation in keystroke (a) and click timing (b). Higher values indicate worse performance. Both the shape of temporal variation with fastest performance a few hours after wake and slowest performance during habitual sleep times as well as the magnitude of variation are consistent with controlled laboratory-based studies [3, 18, 20, 48] (Section 4.2).

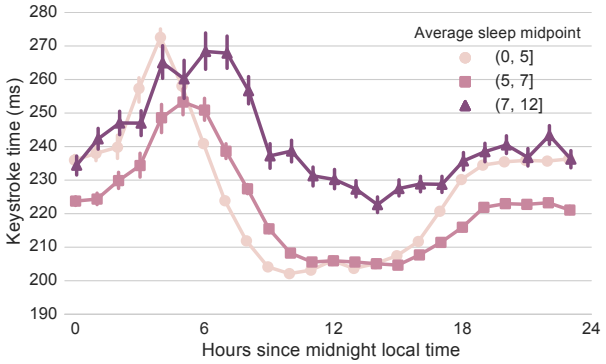

Figure 3: Variation in keystroke time throughout the day varies with chronotype (morning/evening preference) which is defined based on the average point of mid sleep (Section 4.3). Users that typically sleep early (light color) perform slowest at about 04:00 h, while medium or late sleepers (darker colors) perform slowest at 05:00 h and 06:00-07:00 h, respectively. This closely matches their habitual sleep time and is consistent with controlled laboratory-based studies [31].

tor plans [37]. Studies of the potential subprocesses for each task and how they might be differentially influenced by sleep is beyond the scope of this paper. However, our search engine interactions capture performance in everyday tasks that are highly relevant to many occupations, as captured by typing and searching for information [38], and allow us to non-intrusively measure changes in real-world performance throughout the day.

Note that all timing measurements are taken on the server side and not the client side. Therefore, it is important to consider the potential influence of network latency factors. We found that the network latency changes only very little between two consecutive requests (less than 1 millisecond) and thus any latency effects cancel out when we take the time difference between two requests (details in online appendix [6]). This demonstrates that variation in network latency does not affect our analyses. Furthermore, variations in site rendering time (*i.e.*, measuring time from first script till page load completed including dynamic contents) are much smaller (order of milliseconds) compared to variation in click times.

The temporal variation sensed in performance could potentially be an artifact of different users contributing timings at different time points instead of actual within user variation throughout the day. However, we verified that the temporal variation we observe is due to within user variation throughout the day by confirming that the patterns of temporal variation are effectively identical for raw measurements and within-user normalized variants (Z-scores; online appendix [6]). We also verified that performance variation during the weekend is similar to variation during the week (online appendix [6]) and we therefore do not further differentiate between performance during weekdays and weekends in this paper. Finally, we considered alternative performance measures based on backspace usage in keystrokes and spelling errors in search queries. Since we found results to be similar to keystroke and click timing but more noisy due to less frequent measurements, we report results on keystroke and click timing in this paper.

4.2 Temporal Variation of Keystroke and Click Times

Next, we validate our methodology by considering the findings obtained from small-scale controlled sleep studies. It is well established that human performance varies over time and follows a circadian rhythm [3, 48]. Keystroke and click timing also vary throughout the day in a daily rhythm as illustrated in Figure 2. Keystroke times (Figure 2a) are on the order of 240 milliseconds which closely matches the expected typing speed of an average typist (240 milliseconds; 50 words per minute at 5 characters per word, see [15]). Click times (Figure 2b) are on the order of 10 seconds. Note that both measures follow a similar pattern throughout the day. Users are fastest to type and click a few hours after typical wake times and the timing increases again in the evening hours (in particular for click times). Performance is slowest during habitual sleep times (*e.g.*, 04:00 h) closely matching accident risk rates [20] and the anticipated circadian nadir (*i.e.*, the time of greatest circadian sleep drive) [18]. Furthermore, controlled laboratory experiments have shown that performance typically varies by 15 to 30 percent over the course of a day across a variety of simple motor and cognitive tasks [3, 48]. For keystrokes we measure a variation of 31% and for click times a variation of 12%.

The consistent agreement in shape and magnitude of variation with controlled lab experiments on human performance and for two different tasks suggest that these large-scale measures based on search engine interactions can be used to study sleep and performance. The proposed measures can be collected non-intrusively at unprecedented scale and shine light on how real-world performance varies throughout the day and with changes in sleep.

4.3 Performance Variation by Chronotype

A person's chronotype encompasses the propensity for the individual to sleep at a particular time during a 24-hour period and is at least partly based on genetics [40]. Studies have shown that performance depends on the alignment of chronotype and time of day [31]; early types tend to be higher performing earlier in the day while late types are higher performing later. The individual chronotype of each user can be defined based on the mid-sleep point on free days (MSF) which is the halfway point between going to sleep and waking up [25, 40]. Many people compensate for slept debt accumulated during work days by sleeping longer on free days; that is, the sleep midpoint we observe is later than the internal biological clock would dictate on the free days. Therefore, sleep scientists use a midsleep point that is corrected for oversleep (indicated by SC) [25]: $MSF_{SC} = MSF - 0.5(SD_F - (5*SD_W + 2*SD_F)/7)$, where SD_F and SD_W are sleep duration and free days and work days, respectively, and $SD_F - (5*SD_W + 2*SD_F)/7$ corresponds to the difference in sleep duration on free days and the average day. We compute this corrected midpoint for every user in the dataset using weekdays as work days and weekend days as free days (Median $MSF_{SC} = 4.70$).

We show that keystroke times throughout the day vary with chronotype (Figure 3), matching results from previous sleep studies [31] and thus providing further validation of our methods. We find that early sleepers are slowest at about 04:00 h, while medium or late sleepers are slowest at 05:00 h and 06:00-07:00 h, respectively. This closely matches each group's habitual sleep time and demonstrates the validity and power of this large dataset; for each chronotype group, we have millions of measurements even during typical sleep times that allow us to estimate these performance curves. We find similar results for click times.

5. MODELING PERFORMANCE

Having demonstrated that performance of search engine interactions vary over time and based on biological processes (Section 4), we now operationalize and extend a conceptual model of sleep and performance from chronobiology [4, 13] to explain the variation observed in performance measurements. Classic sleep models are based on circadian rhythms and homeostatic sleep drive [13]. In addition, we consider sleep inertia and sleep duration [4, 43]. Background on relevant biological processes is covered in Section 2.

5.1 Conceptual Model

We model the keystroke and click timing based on (1) time of day in local time, (2) time in hours after wake up, and (3) sleep duration the previous night. We know (1) from the time of the keystroke or click time measurement, and (2) and (3) from wearable device-defined sleep measurements (Section 3).

Since many people wake up during the same morning hours every day, time of day and time since wake up are naturally correlated and challenging to disentangle. In laboratory-based sleep studies, the goal of exploring the distinct influences of the factors is achieved by "forced desynchrony" protocols [43], where subjects are deprived of sleep for extended periods of time. Instead of similar interventions, we employ mathematical modeling with a large-scale dataset of real-world sleep and performance measurements and use the variation observed across millions of observations to disentangle the relative contributions of circadian and homeostatic factors. The large-scale dataset contains numerous performance measurements during usual (day) and unusual (late night) times (*e.g.*, Figure 3) that we can use to understand the relative contributions of these factors to performance in the open world (see formulation of additive model in Section 5.2).

Potential Confounding Factors. We control for several factors in our model to avoid confounding. For keystrokes, we control for the exact character typed or removed since different characters might take a varying amount of time (*e.g.*, typing an "a", or a capital "A", or hitting backspace). For click times, it is expected that clicking on results further down the list of results will take more time, which holds true in our data (online appendix [6]). We therefore control for the click position in our model.

Clicking on a result link is preceded by a cognitive process–interpreting the words displayed on links and deciding which link to click–which can be quick in the case of navigational queries (*e.g.*, "facebook") or much slower in the case of informational queries (*e.g.*, "What is the homeostatic sleep drive?"). Formally, this distinction can be captured through the concept of click entropy, which measures how "surprising" the distribution over clicked URLs for a given query is [21]. We find that informational queries take about two seconds longer than navigational queries on average (online appendix [6]). Therefore, we control for the click entropy of the query preceding the click in our model.

An extreme way of controlling for varying queries is to compare click times for exactly identical queries (*e.g.*, popular queries such as "facebook"). We verified that this yields very similar results, albeit with larger confidence intervals since the sample size is reduced dramatically compared to including all queries and controlling for click entropy, demonstrating that the observed patterns are not due to a particular mix of query types.

In addition, we tested for learning effects as issuing the same query multiple times might lead to improved performance. However, most queries, 73.1%, are unique in the dataset and only 4.1% of queries occur more than three times. Further, we did not find any evidence for improving performance over time for frequently occurring queries. This is likely because most users were fairly proficient at typing before the start of our observation period.

5.2 Mathematical Formulation

We now describe the formulation of the model for keystroke timing. The model for click times is parallel, where we control for the click position and click entropy instead of the keystroke type. We are interested in estimating how (1) time of day, (2) time after wake up, and (3) sleep duration influence performance. We assume that all these effects are additive as supported by evidence presented in [2]. Mathematically, we formulate a fixed-effects model

$$y_i = \alpha + f^k(x_i^k) + f^t(x_i^t) + f^w(x_i^w) + f^d(x_i^d) + \epsilon_i \,,$$

where y_i is the keystroke time for observation i, α is a constant intercept, and f^k, f^t, f^w, f^d are the unknown functions of interest for keystroke type, time of day, time since wake up, and sleep duration, respectively, with corresponding input features $x_i^k, x_i^t, x_i^w, x_i^d$, and ϵ_i is the i-th residual.

Instead of estimating arbitrary functions, we use fine-grained piecewise constant approximations. We discretize each input space (*e.g.*, between midnight and 01:00 h, or between 01:00 h and 02:00 h, or between 0 and 15 minutes after waking up, *etc.*). We denote the functions mapping input features x_i^t, x_i^w, x_i^d to their respective bins as b^t, b^w, b^d (note that keystroke type x_i^k is already discrete). Further, we use the functions c^k, c^t, c^w, c^d to map the discretized features to a constant value. The simplified model then becomes

$$y_i = \alpha + c^k(x_i^k) + c^t(b^t(x_i^t)) + c^w(b^w(x_i^w)) + c^d(b^d(x_i^d)) + \epsilon_i \,.$$

The outcome of interest in this modeling task are the functions c^t, c^w, c^d which express the independent impact of (1) time of day, (2) time since wake up, and (3) sleep duration on performance timings the next day. We estimate all parameters $(\alpha, c^k, c^t, c^w, c^d)$ including 95% confidence intervals through least squares optimiza-

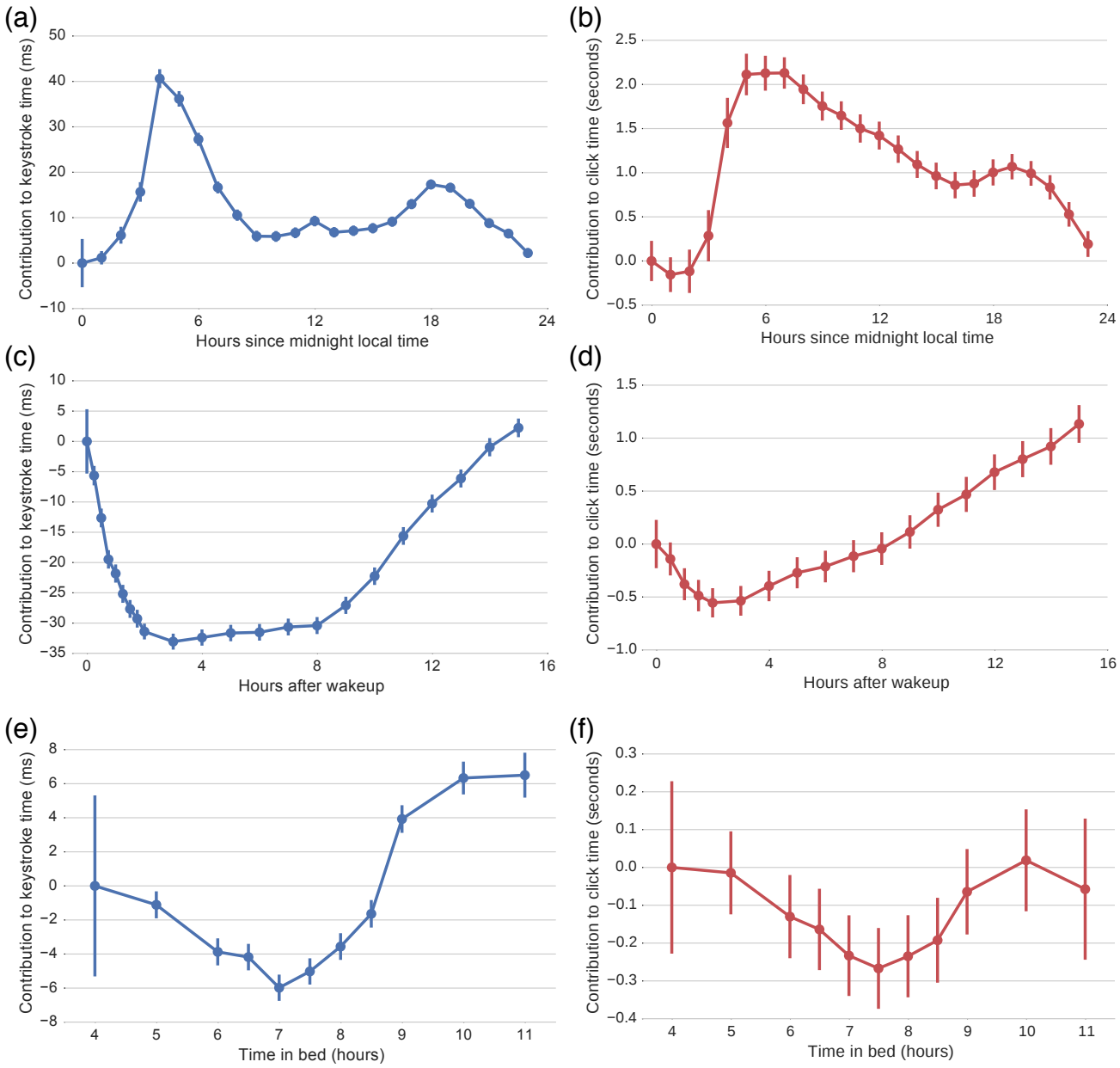

Figure 4: Contributions to keystroke (a,c,e; blue) and click time (b,d,f; red) performance of different factors included in our model. Results are similar for both performance measures and match estimates from controlled sleep studies in the laboratory (Section 5). For example, variation over the time of day c^t (a,b) shows that performance is slowest during habitual sleep times near the presumptive circadian nadir (04:00 h; see main text). Variation across time after wake up c^w (c,d) shows effects of sleep inertia during the first two hours after wake. There is relative stability for around eight hours in keystroke time but a steady decline in click time after that point. Sleep durations c^d (e,f) of 7.0-7.5 hours are associated with optimal performance according to our measures. However, note that the impact on overall variation is smaller compared to time of day (a,b) and time since wake up (c,d).

tion. We also experimented with mixed effects models controlling for variation across users and across queries through random effects. While standard mixed model libraries do not scale well to the size of our dataset, we found that these models lead to very similar estimates compared to the fixed effects model described above when using subsets of the data.

5.3 Results

The functions c^t, c^w, c^d modeling the influence on cognitive performance of time of day, time since wake up, and sleep duration are illustrated in Figure 4. Impact on keystroke timings are shown in blue (Figure 4a,c,e) and impact on click times are shown in red (Figure 4b,d,f). Note that the shapes of these functions for keystrokes and click times are very similar and smooth, even though there are no constraints that would force this to occur. Furthermore, we note that the temporal variation in cognitive performance is not explained by variation in different users that contribute timings at different points throughout the day (*i.e.*, population differences) but are due to within user variation (online appendix [6]).

Time of Day. Cognitive performance on both keystroke and click tasks varies with time of day (Figure 4a,b) and is slowest during habitual sleep time around 04:00-06:00 h. Performance quickly improves after typical wake times and becomes slightly slower in the evening for both keystroke and click times (19:00 h). The two curves consistently match estimates of circadian rhythm processes in sleep obtained through controlled laboratory experiments [18, 49]. Note that the magnitude of variation is substantial at around 40 milliseconds for keystrokes and over 2.1 seconds for click times, which are changes of 18% and 23%, respectively, relative to average timing for each (Table 1).

Time after Awakening. Cognitive performance also varies substantially with the time after wake up (Figure 4c,d). The magnitude of the variation is relatively large at about 42 milliseconds or 19% for keystrokes about slightly over 1.6 seconds or 17% for click times. Within the first two hours, performance rapidly improves (*i.e.*, lower timings). This demonstrates a well-known effect in sleep studies called sleep inertia (Section 2). After this point, performance is best and slowly worsens until a point of poorest performance is reached at around 16 hours of wake time, consistent with the homeostatic sleep drive [13]. This corresponds exactly to the point when most people would go to sleep again (*i.e.*, a typical sleep duration of 8 hours). We excluded data beyond the typical wake period of 16 hours because the data becomes more sparse and to avoid potential selection effects with regard to the people who choose to stay awake for exceptionally long periods of time. However we found similar patterns between both keystrokes and click times even beyond this point. We note that keystroke time is relatively stable for about six hours while click times continuously increase, likely due to the differences in cognitive and motor competencies for the tasks, and due to differences in the sensitivities of those competencies to status of sleep and circadian rhythm. In summary, the estimates derived from our model closely capture the initial sleep inertia and the increasing homeostatic sleep drive first discovered through laboratory-based studies [4, 43, 49].

Time in Bed. Keystrokes and click time vary with the amount of time in bed during the previous night (Figure 4e,f). However, we note that this variation, 12 milliseconds for keystrokes (5%) and 0.25 seconds for click times (3%), is much smaller than the previous two factors. For both measures, we find a clear U-shaped curve with its center, indicating optimal performance, at 7.0-7.5 hours of sleep. Both sleeping too little (under 7 hours) or too much (more than 8-9 hours) are associated with decreased performance. U-shaped relationships with respect to sleep duration have been reported for several outcomes (*e.g.*, mortality [27]). We further investigate the impact of insufficient sleep on performance in Section 6.

6. INFLUENCE OF INSUFFICIENT SLEEP ON PERFORMANCE

Following our studies to validate the methodology (Section 4 and Section 5), we now extend prior laboratory-based sleep research with estimates of how sleep influences performance in real-world settings. In particular, we study the impact of one or multiple nights of insufficient sleep on performance over the following days.

6.1 Single Nights of Insufficient Sleep

We first consider single nights of sleep and analyze how very short or very long sleep durations, as well as differences in sleep timing from the usual patterns within a user, impact performance. We only show results for keystroke timing here; the results are similar for click times (*e.g.*, Figure 2 and Figure 4). Figure 5a shows that users performed significantly slower when in bed fewer than 6 or more than 9 hours, consistent with the results described in Section 5.3. In those conditions, the average keystroke times were about four and seven milliseconds longer compared to sleeping between 7 and 9 hours (increases of 2.7% and 4.0%, respectively; both $p \ll 10^{-10}$; Mann–Whitney U-test, which is used for all hypothesis tests in this section).

Timing of sleep is also a significant factor for performance the next day (Figure 5b). While sleeping earlier than usual makes only a difference of about 1 milliseconds or 0.5% ($p \ll 10^{-10}$), going to bed an hour or more later than usual is associated with significantly worse average performance of about 14 milliseconds or 7.3% longer keystrokes ($p \ll 10^{-10}$). Note that we limited the sleep duration to be between 7 and 8 hours long for this analysis so that these results demonstrate the impact of timing independent of differences in duration (*i.e.*, those going to sleep later had a normal length of time in bed despite going to sleep late). We further verified that these results are not due to people sleeping later and longer on weekends when they might be typing slower due to less work pressure as we find similar patterns and effect sizes using just weekday data. Thus, these results could point to an interaction between the circadian clock and the ultradian rhythm of sleep (*i.e.*, the cycling of sleep stages): sleeping at different phases can result in different sleep organization [17]. Our findings suggest that sleeping later in one's circadian cycle does not satisfy the neural recovery needed for proper daytime performance, while sleeping earlier does not have the same negative effects.

6.2 Multiple Nights of Insufficient Sleep

Above, we reported on the effect of a *single* night of sleep with particular duration and timing on the next day. Here, we examine whether *multiple* insufficient nights of sleep measurably affect performance and how long this effect appears to persist. For purposes of this analysis, we define an "insufficient" night of sleep ("I") to have a time in bed of under six hours (as in [22]), and a "sufficient" night of sleep ("S") to have a time in bed of at least six hours. We consider three different scenarios: two nights of sleep with more than six hours each (SS), one night over and the next night under six hours (SI), and two nights under six hours of sleep (II). We measure the performance after those two nights of sleep for a period of seven days, reducing the performance on each of these seven days to a single value—the average performance during the first 16 hours after wake up (*i.e.*, typical wake period). We do not consider longer sleep patterns here due to the large number of possible combinations and data reduction associated with individual sleep patterns (*e.g.*, a person might not track their sleep every single night). Intentionally not controlling for sleep both preceding and following the two nights of interest, we are addressing how insufficient sleep impacts real-world performance given real-world choices. We are not, however, examining the underlying biological processes of recovery from sleep loss. We note that the start of the sleep patterns was distributed all throughout the week; for example, two nights of sufficient sleep (SS) did occur both during the week as well as over the weekend. We define recovery time as the number of days it takes to reach performance levels comparable to those after a sufficient sleep schedule (SS).

Results. Multiple insufficient nights of sleep have a significant impact on average keystroke timing (Figure 6). Performance is best after two sufficient nights of sleep, slightly but measurably worse after one insufficient night of sleep, and significantly worse after two insufficient nights in a row. Over the first 24 hours, having one insufficient night of sleep is associated with 1.2% slower performance ($p \ll 10^{-10}$) and two insufficient nights of sleep are 4.8%

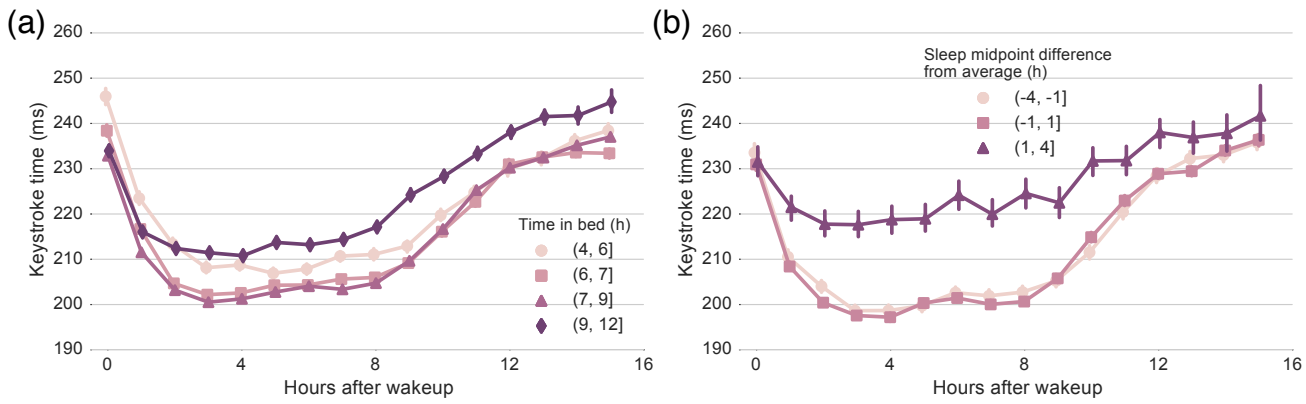

Figure 5: The impact of sleep duration (a) and timing (b) on performance the next day. Sleep timing is measured through difference from the typical sleep midpoint and we control for sleep duration. We find that sleeping less than 7 or more than 9 hours is associated with slower performance (a). Sleeping earlier than usual does not make a large difference but going to bed an hour or more later than usual is associated with significantly worse performance the next day (b).

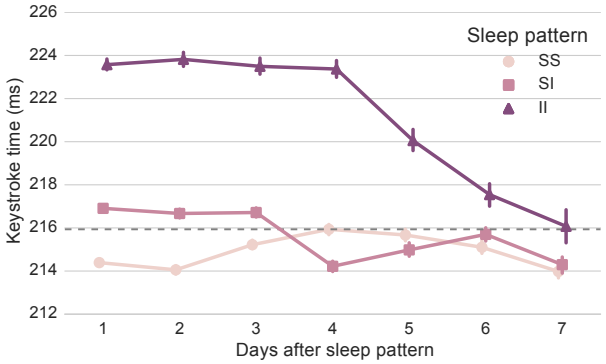

Figure 6: Comparing the impact on performance of zero (SS), one (SI), or two (II) consecutive insufficient nights of sleep (less than six hours of time in bed). One night of insufficient sleep is associated with significantly slower keystroke times and two insufficient nights in a row exhibit a significantly larger effect. Judging by when average keystroke time drops below the horizontal dashed line representing the slowest performance for the group with two nights of sufficient sleep (SS), we observe that it takes six nights of sleep to return to baseline performance levels after two nights of insufficient sleep (day 7) and three nights to return to baseline performance levels after one night of insufficient sleep (day 4) given real-world sleep schedules.

slower ($p \ll 10^{-10}$) compared to two nights with longer than six hours of sleep each (2.7% and 7.3% increases for click times, respectively; both $p \ll 10^{-10}$). Note that these effect estimates take into account any real-world behavioral compensation such as increased caffeine intake that will help improve performance after sleep loss. The horizontal dashed line in Figure 6 corresponds to the slowest keystroke time after two nights of sufficient sleep (SS), which we use as a conservative point of reference to judge when performance after insufficient sleep (SI and II) has returned to a performance below this point. We find that, on average, it takes three nights to make up one insufficient night of sleep (SI crosses dashed line on day 4) and six nights two make up two insufficient nights of sleep in a row (II crosses dashed line on day 7). We find very similar results for the impact on the *variance* (*i.e.*, instead of mean) of keystroke timing as well as for click times. A version of Figure 6 that visualizes average performance throughout each of the seven days is included in the online appendix [6].

Note that these results are not simply due to having fundamentally different users contribute to each of the the curves (SS, SI, II). While some users are more likely to get fewer than six hours of sleep than others, we do find similar effects by restricting each of the three curves to be estimated from the exact same set of users. We note that, since we enforce no constraints on time in bed during the seven days following the sleep pattern, additional nights of insufficient sleep could occur during the follow-up period, contributing to the duration of the recovery period. Thus, we need to explore whether there is a higher likelihood of sleep deficiencies on days following the initial observed two-day period of insufficient sleep. We find that, on average, SS is followed by 0.4 nights of insufficient sleep during the following seven days, whereas SI and II are followed by 1.2 and 2.5 such nights. Thus, additional days of insufficient sleep for the SI and II cases may have an influence on the overall time to returning to baseline performance. Nevertheless, our findings show real-world timing of return to baseline performance. We leave to future work the study of more complex real-world patterns of sleep and sleep deficit and the influences of sleep deficits on performance.

7. CONCLUSION

Understanding human performance and its relation to sleep is critical to productivity [16], learning [26], and avoiding accidents [16, 20]. Human performance is not constant but exhibits daily variations [43]. Existing research on sleep and performance has typically been restricted to small-scale laboratory-based studies involving artificial performance tasks in an artificial environment. Therefore, novel methods of large-scale real-world monitoring, like we have presented, are necessary to advance our understanding of sleep and performance [39].

Summary of Results. We presented the largest study to date on sleep and performance in the wild. Using a new approach to non-intrusive measurement for both cognitive performance and sleep we were able to study more than 400 times the number of users compared to the second largest study. We correlated human performance based on interactions with a web search engine to sleep measures detected by a wearable device. We demonstrated that real-world performance varies throughout the day and based on chronotype and prior sleep, in close agreement with small-scale

laboratory-based studies. We developed a statistical model that operationalizes recent chronobiological research and showed that our estimates of circadian rhythms, homeostatic sleep drive, and sleep inertia closely match published results of controlled sleep studies. Further, we contribute to existing sleep research through quantifying extended periods of lower real-world performance that are associated with single and multiple nights of insufficient sleep.

Implications. We have demonstrated that human performance can be measured in a real-world setting without any additional hardware or explicit testing by exploiting existing search engine interactions that occur billions of times per day. We have validated our methodology and shown that human performance, as measured through these signals, varies throughout the day and based on chronotype and sleep, in close agreement with controlled laboratory-based studies. Beyond the relevance of the results to extending insights about sleep and performance, our findings more generally highlight the potential power of harnessing online activities to study human cognition, motor skills, and public health. Large-scale physiological sensing from online data enables

- studies of sleep and performance outside of small laboratory settings, and without actively inducing sleep deprivation,
- non-intrusive measurement of cognitive performance without forcing individuals to interrupt their work to perform separate artificial tasks [39],
- the identification of realistic measures of real-world cognitive performance based on frequent tasks and interactions,
- and continuous monitoring of such measures.

Suitable examples for such data include continuous usage patterns from computing applications such as email, programming environments, bug report systems, office suites, and others. Any insights on performance and productivity gained through monitoring these applications could be used to improve the user's awareness of such patterns and to adapt the user experience appropriately (*e.g.*, scheduling tasks intelligently in order to prevent or minimize human error; scheduling meetings based on participants performance and chronotype profiles). There are great opportunities ahead to investigate how such insights could be used to personalize applications based on relevant biological processes and chronotypes.

Acknowledgments. We thank Jure Leskovec, Emma Pierson, Marinka Zitnik, David Hallac, David Jurgens and the anonymous reviewers for their valuable feedback on the manuscript.

8. REFERENCES

[1] S. Abdullah, E. L. Murnane, M. Matthews, M. Kay, J. A. Kientz, G. Gay, and T. Choudhury. Cognitive rhythms: Unobtrusive and continuous sensing of alertness using a mobile phone. In *UbiComp*, 2016.

[2] P. Achermann and A. A. Borbély. Simulation of daytime vigilance by the additive interaction of a homeostatic and a circadian process. *Biol Cybern*, 71(2):115–121, 1994.

[3] J. Ackerman. *Sex Sleep Eat Drink Dream: A day in the life of your body*. Houghton Mifflin Harcourt, 2008.

[4] T. Åkerstedt and S. Folkard. The three-process model of alertness and its extension to performance, sleep latency, and sleep length. *Chronobiol Int*, 14(2):115–123, 1997.

[5] T. Althoff, K. Clark, and J. Leskovec. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *TACL*, 2016.

[6] T. Althoff, E. Horvitz, R. W. White, and J. Zeitzer. Online appendix of this paper. 2017. `http://stanford.io/2ejFPhD`.

[7] T. Althoff, P. Jindal, and J. Leskovec. Online actions with offline impact: How online social networks influence online and offline user behavior. In *WSDM*, 2017.

[8] T. Althoff, R. W. White, and E. Horvitz. Influence of Pokémon Go on physical activity: Study and implications. *J Med Internet Res*, 18(12):e315, 2016.

[9] S. Ancoli-Israel, R. Cole, C. Alessi, M. Chambers, W. Moorcroft, and C. Pollak. The role of actigraphy in the study of sleep and circadian rhythms. *Sleep*, 26(3):342–392, 2003.

[10] M. Basner, K. M. Fomberstein, F. M. Razavi, S. Banks, J. H. William, R. R. Rosa, and D. F. Dinges. American time use survey: sleep time and its relationship to waking activities. *Sleep*, 30(9):1085–1095, 2007.

[11] K. Blatter and C. Cajochen. Circadian rhythms in cognitive performance: Methodological constraints, protocols, theoretical underpinnings. *Physiology & Behavior*, 90(2):196–208, 2007.

[12] M. Böhmer, B. Hecht, J. Schöning, A. Krüger, and G. Bauer. Falling asleep with Angry Birds, Facebook and Kindle: A large scale study on mobile application usage. In *MobileHCI*, 2011.

[13] A. A. Borbély. A two process model of sleep regulation. *Human neurobiology*, 1982.

[14] Bureau of Labor Statistics, American Time Use Survey. Average sleep times per day, by age and sex. Archived at: `http://www.webcitation.org/6lPcEntyS`, 2015.

[15] S. K. Card, T. P. Moran, and A. Newell. The keystroke-level model for user performance time with interactive systems. *CACM*, 23(7):396–410, 1980.

[16] H. R. Colten and B. M. Altevogt. *Sleep disorders and sleep deprivation: An unmet public health problem*. 2006.

[17] D.-J. Dijk and C. A. Czeisler. Contribution of the circadian pacemaker and the sleep homeostat to sleep propensity, sleep structure, electroencephalographic slow waves, and sleep spindle activity in humans. *J. Neurosci.*, 15(5):3526–3538, 1995.

[18] D.-J. Dijk, J. F. Duffy, and C. A. Czeisler. Circadian and sleep/wake dependent aspects of subjective alertness and cognitive performance. *J Sleep Res*, 1(2):112–117, 1992.

[19] D. F. Dinges. Are you awake? Cognitive performance and reverie during the hypnopompic state. In *Sleep and Cognition*, pages 159–75. 1990.

[20] D. F. Dinges. An overview of sleepiness and accidents. *J Sleep Res*, 4(s2):4–14, 1995.

[21] Z. Dou, R. Song, and J.-R. Wen. A large-scale evaluation and analysis of personalized search strategies. In *WWW*, 2007.

[22] J. Fernandez-Mendoza, S. Calhoun, E. O. Bixler, S. Pejovic, M. Karataraki, D. Liao, A. Vela-Bueno, M. J. Ramos-Platon, K. A. Sauder, and A. N. Vgontzas. Insomnia with objective short sleep duration is associated with deficits in neuropsychological performance: A general population study. *Sleep*, 33(4):459–465, 2010.

[23] S. A. Golder and M. W. Macy. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051):1878–1881, 2011.

[24] P. Hemp. Presenteeism: At work-but out of it. *Harvard Business Review*, 82(10):49–58, 2004.

[25] M. Juda, C. Vetter, and T. Roenneberg. Chronotype modulates sleep duration, sleep quality, and social jet lag in shift-workers. *J Biol Rhythms*, 28(2):141–151, 2013.

[26] P. Kelley, S. W. Lockley, R. G. Foster, and J. Kelley. Synchronizing education to adolescent biology:'let teens sleep, start school later'. *Learn Media Technol*, 40(2):210–226, 2015.

[27] D. F. Kripke, R. N. Simons, L. Garfinkel, and E. C. Hammond. Short and long sleep and sleeping pills: Is increased mortality associated? *Arch Gen Psychiatry*, 36(1):103–116, 1979.

[28] D. S. Lauderdale, K. L. Knutson, L. L. Yan, K. Liu, and P. J. Rathouza. Self-reported and measured sleep duration. *Epidemiology*, 19(6):838–845, 2008.

[29] J. Lim and D. F. Dinges. A meta-analysis of the impact of short-term sleep deprivation on cognitive variables. *Psychol Bull*, 136(3):375, 2010.

[30] G. Mark, S. T. Iqbal, M. Czerwinski, P. Johns, and A. Sano. Neurotics can't focus: An in situ study of online multitasking in the workplace. In *CHI*, 2016.

[31] R. L. Matchock and J. T. Mordkoff. Chronotype and time-of-day influences on the alerting, orienting, and executive components of attention. *Exp Brain Res*, 192(2):189–198, 2009.

[32] F. Monrose and A. Rubin. Authentication via keystroke dynamics. In *CCS*, pages 48–56, 1997.

[33] E. L. Murnane, S. Abdullah, M. Matthews, T. Choudhury, and G. Gay. Social (media) jet lag: How usage of social technology can modulate and reflect circadian rhythms. In *UbiComp*, 2015.

[34] E. L. Murnane, S. Abdullah, M. Matthews, M. Kay, J. A. Kientz, T. Choudhury, G. Gay, and D. Cosley. Mobile manifestations of alertness: Connecting biological rhythms with patterns of smartphone app use. In *MobileHCI*, 2016.

[35] M. T. Orne. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *Am Psychol*, 17(11):776, 1962.

[36] J. Paparrizos, R. W. White, and E. Horvitz. Screening for pancreatic adenocarcinoma using signals from web search logs: Feasibility study and results. *J Oncol Pract*, pages 737–44, 2016.

[37] J. J. Pilcher and A. J. Huffcutt. Effects of sleep deprivation on performance: A meta-analysis. *Sleep*, 1996.

[38] K. Purcell. Search and email still top the list of most popular online activities. Pew Research Center. Archived at: `http://www.webcitation.org/5I2STSU61`, 2011.

[39] T. Roenneberg. Chronobiology: The human sleep project. *Nature*, 498(7455):427–428, 2013.

[40] T. Roenneberg, A. Wirz-Justice, and M. Merrow. Life between clocks: Daily temporal patterns of human chronotypes. *J Biol Rhythms*, 18(1):80–90, 2003.

[41] A. Shameli, T. Althoff, A. Saberi, and J. Leskovec. How gamification affects physical activity: Large-scale analysis of walking challenges in a mobile application. In *WWW*, 2017.

[42] D. A. Sternberg, K. Ballard, J. L. Hardy, B. Katz, P. M. Doraiswamy, and M. Scanlon. The largest human cognitive performance dataset reveals insights into the effects of lifestyle factors and aging. *Front Hum Neurosci*, 7:292, 2013.

[43] H. P. Van Dongen and D. F. Dinges. Circadian rhythms in fatigue, alertness, and performance. *Principles and practice of sleep medicine*, 20:391–9, 2000.

[44] L. M. Vizer, L. Zhou, and A. Sears. Automated stress detection using keystroke and linguistic features: An exploratory study. *Int J Hum Comput Stud*, 67(10):870–886, 2009.

[45] O. J. Walch, A. Cochran, and D. B. Forger. A global quantification of "normal" sleep schedules using smartphone data. *Sci Adv*, 2(5), 2016.

[46] R. West, R. W. White, and E. Horvitz. From cookies to cooks: Insights on dietary patterns via analysis of web usage logs. In *WWW*, 2013.

[47] R. W. White, S. Wang, A. Pant, R. Harpaz, P. Shukla, W. Sun, W. DuMouchel, and E. Horvitz. Early identification of adverse drug reactions from search log data. *J Biomed Inform*, 59:42–48, 2016.

[48] J. A. Wise, V. D. Hopkin, and D. J. Garland. *Handbook of aviation human factors*. CRC Press, 2009.

[49] K. P. Wright Jr, C. A. Lowry, and M. K. LeBourgeois. Circadian and wakefulness-sleep modulation of cognition in humans. *Front Hum Neurosci*, 5:50, 2012.