DATA SCIENCE FOR HUMAN WELL-BEING

A DISSERTATION
SUBMITTED TO THE COMPUTER SCIENCE DEPARTMENT
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Christopher Tim Althoff
August 2018

© 2018 by Christopher Tim Althoff. All Rights Reserved.
Re-distributed by Stanford University under license with the author.

This work is licensed under a Creative Commons Attribution-Noncommercial 3.0 United States License.
http://creativecommons.org/licenses/by-nc/3.0/us/

This dissertation is online at: http://purl.stanford.edu/gv596kq0446

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Jure Leskovec, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Scott Delp

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Dan Jurafsky

Approved for the Stanford University Committee on Graduate Studies.

Patricia J. Gumport, Vice Provost for Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format. An original signed hard copy of the signature page is on file in University Archives.

Abstract

The popularity of wearable and mobile devices, including smartphones and smartwatches, has generated an explosion of detailed behavioral data. These massive digital traces provide us with an unparalleled opportunity to realize new types of scientific approaches that enable novel insights about our lives, health, and happiness. However, gaining actionable insights from these data requires new computational approaches that turn observational, scientifically "weak" data into strong scientific results and can computationally test domain theories at scale.

In this dissertation, we describe novel computational methods that leverage digital activity traces at the scale of billions of actions taken by millions of people. These methods combine insights from data mining, social network analysis, and natural language processing to improve our understanding of physical and mental well-being: (1) We show how massive digital activity traces reveal unknown health inequality around the world, and (2) how personalized predictive models can support targeted interventions to combat this inequality. (3) We demonstrate that modeling the speed of user search engine interactions can improve our understanding of sleep and cognitive performance. (4) Lastly, we describe how natural language processing methods can help improve counseling services for millions of people in crisis.

Dedication

For my parents Klaus-Dieter and Monika Althoff,
the extent of their love for me I still continue to discover day by day.

Acknowledgments

I am deeply grateful to my advisor Jure Leskovec for guiding me through my PhD, and for a lifetime worth of sharp, honest, and supportive advice. Jure taught me how to identify and work on problems that matter, how to be an independent researcher, and he afforded me incredible opportunities to grow as a researcher, teacher and as a person.

I would also like to thank Eric Horvitz, Dan Jurafsky, Scott Delp, and Trevor Hastie, who have served on my thesis committee and have provided invaluable advice and encouragement during my time at Stanford, and especially during my job search.

Eric Horvitz invited me to Microsoft Research for an internship and has since been an exceptionally supportive mentor to me. I deeply admire Eric's contagious love for and joy in research that energized and encouraged me to complete this dissertation and more.

Dan Jurafsky welcomed me to Stanford in my first quarter as a rotation mentor and never stopped being generous and kind with his advice. He taught me how richly language can reflect the human experience and to this day I cannot look at a bag of chips or restaurant menu without taking note of their linguistic style.

This thesis' focus on human health would not have been possible without Scott Delp and the Mobilize Center. I am grateful to Scott for teaching me about health and human movement, his thoughtful perspective on academic research, and his unfailing kindness that permeates his research group and center.

Through the Mobilize Center I met Trevor Hastie, who in seminar after seminar generously shared his perspective, advice, and encouragement, and who served as the chair of my dissertation committee.

Some of the most productive months of my PhD were spent on summer internships, during which I was privileged to work with colleagues and mentors who deeply influenced my thinking and research. In particular, I would like to thank Ryen White and Eric Horvitz at Microsoft and Luna Dong, Kevin Murphy, Evgeniy Gabrilovich and David Ross at Google.

I had the privilege of working with and alongside many incredible and inspiring colleagues in the SNAP group, the Infolab, and the Mobilize Center. Beyond our

research, they have uniquely enriched my experience at Stanford in many ways including unforgettable trip reports, sharing of yummy international chocolates, and tense soccer watching. In particular I am thankful to Rok Sosic, David Hallac, Will Hamilton, Hima Lakkaraju, Jennifer Hicks, Abby King, Adam Miner, Takeshi Kurashima, Emma Pierson, Zhiyuan Jerry Lin, Michele Catasta, Srijan Kumar, Marinka Zitnik, Ashton Anderson, Austin Benson, Caroline Lo, Robert West, Justin Cheng, Cristian Danescu-Niculescu-Mizil, Xiang Ren, David Jurgens, Mitchell Gordon, Boris Ivanovic, Hamed Nilforoshan, Jamie Zeitzer, Kevin Clark, Joy Ku, Jessica Selinger, Ina Fiterau, Jason Fries, Jenna Hua, Eric J Daza, Steven Bell, Hector Garcia-Molina, Jeff Ullman, Pranav Jindal, Ali Shameli and Amin Saberi. I also owe a special thank-you to our outstanding lab administrators Yesenia Gallegos and Marianne Siroker, and our hero-level sysadmins Andrej Krevl, Peter Kacin, and Adrijan Bradaschia.

I thank Bojan Bostjancic and Peter Kuhar at Azumio, Chul Lee, Hesamoddin Salehian and Patrick Howell at MyFitnessPal and UnderArmour, Bob Filbin and Nitya Kanuri at Crisis Text Line, Gil Margolin, Bonnie Ray and Jacob Levine at Talkspace, Vladimir Dubovskiy at DonorsChoose, and Valerie Hajdik at Reddit for granting access to their data, without which the present research would have been impossible. I am also grateful for the financial support from a SAP Stanford Graduate Fellowship, which afforded me the freedom to pursue new ideas.

Pursuing a PhD can be really tough and I cannot imagine what I would have done without the support of loyal friends who are going through the same ups and downs. I am particularly grateful to my dear friends Jeff Ho, Annabell Ho, Yang Wong, Vivian Chen, and Daniel Heywood, who have shared with me their hearts, minds, dreams, scars, faith, gifts, homes, and alliances. I also thank Pete and Becky Meyer who provided a home away from home for me when I first moved to the United States, and to my steadfast friends in Germany, Christoph, Michael, Bernadette, Sandra, Thomas, Nadine, Katharina, and Marie, who within moments make me feel as if I never moved away.

More than anyone else, I want to thank my family for their unfailing love and support for me. More than words could ever express, I am grateful for my amazing partner Colleen Carroll, whose constant and loving support still keeps me going. Finally, a thousand thanks go to my sister Maren, my brother Robin, and to my parents Klaus-Dieter and Monika to whom I dedicate this thesis.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Science is revolutionized by data. For example, online social networks enabled researchers to learn new insights about people, their peers, and their interactions. Studies of the structure of online social networks revealed small-world [Watts and Strogatz, 1998], powerlaw [Faloutsos et al., 1999], and bowtie [Broder et al., 2000] topologies. Through studies of human behavior in online social networks we learned about fundamental communication patterns [Leskovec and Horvitz, 2008], information diffusion [Romero et al., 2011], and polarization [Adamic and Glance, 2005]. However, these lessons based on online social networks have largely been limited to people's online behaviors.

The purpose of this dissertation is to explore and demonstrate how data science methods can positively impact human health and well-being. Beyond our behaviors online, human well-being is naturally tied to our offline behaviors as well, for instance physical activity, sleep, diet and social interactions. How can we capture people's offline behaviors? How can we leverage big data to improve people's lives?

The same way online social networks revealed what people do online, wearable and mobile devices reveal what people do in the real world. These devices are increasingly prevalent with 69% of adults owning a smartphone in developed countries, and about 46% in developing countries [Anthes, 2016]. These devices can capture how we sleep, work, eat, exercise, and communicate—major components of people's everyday and critical behaviors for human health [World Health Organization, 2002]. In tracking these behaviors, wearable and mobile devices generate massive digital traces of real-world behavior and health.

Digital traces of human offline behavior have been increasingly available over the last decade. For instance, data from Fitbit wearable devices have been available

since the company's inception in 2007. However, effectively leveraging these data for human well-being presents many challenges. Therefore, these data have been regularly thrown away and overlooked for scientific research. New data science methods are needed to address these challenges. This dissertation attempts to address this need.

In this dissertation, we present new computational methods for digital activity traces to better understand and improve human well-being. We will leverage Terabyte-scale digital traces that capture billions of actions by millions of people together with new data science methods to conduct massive observational studies. We will demonstrate how this approach can lead to actionable insights into human behavior and well-being through a series of cases studies across physical activity, sleep, and mental health.

Digital activity traces from current mobile and wearable devices are rich and multimodal, capturing people's behaviors and health. They range from sensor data, for instance from accelerometry, to device usage data, to our social interactions and language patterns. Due to the popularity of these devices, researchers and industry organizations have already captured digital trace data across millions of people. In principle, this allows us to conduct research studies at massive scale, capturing participant behavior in great detail, tracking health behavior continuously and over long periods of time, and doing so at comparatively low cost.

Learning how to leverage fine-grained, already collected digital traces has great potential impact, because it can address fundamental limitations of behavioral health research today. This research is often confined to laboratory settings. When behaviors are studied outside of laboratory settings, studies are typically still limited to a small number of people (*e.g.*, less than 50 subjects), tracking behaviors only over short periods of time (*e.g.*, less than 5 days), and with limited resolution (*e.g.*, binary granularity). In many cases, behavioral measurements are collected through self-report and survey measures. Studies have shown up to 700% discrepancies between these subjective measurements and corresponding objective measures [Tucker et al., 2011]. In addition, studies following this traditional approach come at relatively high cost.

Due to these limitations of current research, we know very little about human behavior and health. For instance, we currently do not have good answers to basic questions such as: How much do people exercise? What do people eat? What do they struggle with? Digital traces of our activity and health present an opportunity to advance science through a better understanding of human behavior and health, and to help improve healthcare systems through new, actionable insights.

Leveraging large-scale digital traces for human well-being faces several unique and fundamentally computational challenges. (1) Significant domain knowledge in the behavioral, social, and medical sciences is based on subjective and qualitative

measures. The challenge is how to computationally operationalize this knowledge so that it is amenable to objective, quantitative analysis. (2) Raw sensor and interaction data are massive, but typically do not directly measure well-being. New, advanced computational techniques are required to infer well-being from raw data, or from separate, heterogeneous data sources. (3) Sensor and social interaction data are observational (*i.e.*, non-experimental) and messy. Scientific advances require new methods to turn this scientifically "weak" data into strong scientific results (*e.g.*, controlled and causal analyses beyond correlations). This dissertation describes our attempts in addressing these computational challenges through a combination of techniques across data mining, social network analysis, and natural language processing.

1.2 Overview and Summary of Contributions

The purpose of this thesis is to demonstrate that novel computational methods can derive new insights from already collected digital activity traces, that help us better understand and improve human well-being. The overall structure of this thesis is as follows. We consider three key aspects of human health and well-being: physical activity, sleep, and mental health. In Chapter 2, we leverage consumer smartphone data to study physical activity across 111 countries revealing a previously unknown activity inequality. In Chapter 3, we propose a machine learning model to predict human real-world actions ahead of time that can be used to drive just-in-time adaptive interventions and potentially address activity inequality. In Chapter 4, we demonstrate that user interactions with web search engines enable the study of real-world variation in sleep and cognitive function. In Chapter 5, we perform a detailed analysis of a large corpus on counseling conversations to reveal actionable conversation strategies. We discuss related work within the relevant chapters. The main contributions are summarized in the following chapter-by-chapter outline.

1.2.1 Physical Activity: Planetary-scale Smartphone Data Reveal Activity Inequality (Chapter 2)

Originally published in Nature [Althoff et al., 2017c].

Understanding the basic principles that govern physical activity is needed to curb the global pandemic of physical inactivity [Hallal et al., 2012; Kohl et al., 2012; Lee et al., 2012; Sallis et al., 2016a; Tudor-Locke et al., 2008; UN Secretary General, 2011; WHO, 2010] and the 5.3 million deaths per year associated with inactivity [Lee et al., 2012]. Our knowledge, however, remains limited owing to

the lack of large-scale measurements of physical activity patterns across free-living populations worldwide [Hallal et al., 2012; Tudor-Locke et al., 2008].

In this chapter, we leverage the wide usage of smartphones with built-in accelerometry to measure physical activity at planetary scale. We study a dataset consisting of 68 million days of physical activity for 717,527 people, giving us a window into activity in 111 countries across the globe.

We find inequality in how activity is distributed within countries and that this inequality is a better predictor of obesity prevalence in the population than average activity volume. Reduced activity in females contributes to a large portion of the observed activity inequality.

Aspects of the built environment, such as the walkability of a city, are associated with less gender gap in activity and activity inequality. In more walkable cities, activity is greater throughout the day and throughout the week, across age, gender, and body mass index (BMI) groups, with the greatest increases in activity for females.

These findings have implications for global public health policy and urban planning and highlight the role of activity inequality and the built environment for improving physical activity and health.

1.2.2 Activity Tracking: Modeling Real-World Action Sequences (Chapter 3)

Originally published at the 27th International Conference on the World Wide Web [Kurashima et al., 2018].

Mobile health applications, including those that track activities such as exercise, sleep, and diet, are becoming widely used. Accurately predicting human actions in the real world is essential for targeted recommendations that could improve our health and for personalization of these applications.

However, making such predictions is extremely difficult due to the complexities of human behavior, which consists of a large number of potential actions that vary over time, depend on each other, and are periodic. Previous work has not jointly modeled these dynamics and has largely focused on item consumption patterns instead of broader types of behaviors such as eating, commuting or exercising.

In this chapter, we develop a novel statistical model, called *TIPAS*, for Time-varying, Interdependent, and Periodic Action Sequences. Our approach is based on personalized, multivariate temporal point processes that model time-varying action propensities through a mixture of Gaussian intensities. Our model captures short-term and long-term periodic interdependencies between actions through Hawkes process-based self-excitations.

We evaluate our approach on two activity logging datasets comprising 12 million real-world actions (*e.g.*, eating, sleep, and exercise) taken by 20 thousand users over 17 months. We demonstrate that our approach allows us to make successful predictions of future user actions and their timing. Specifically, TIPAS improves predictions of actions, and their timing, over existing methods across multiple datasets by up to 156%, and up to 37%, respectively. Performance improvements are particularly large for relatively rare and periodic actions such as walking and biking, improving over baselines by up to 256%. This demonstrates that explicit modeling of dependencies and periodicities in real-world behavior enables successful predictions of future actions, with implications for modeling human behavior, app personalization, and targeting of health interventions.

1.2.3 Sleep and Cognitive Performance: Harnessing Web Search Interactions for Population-Scale Physiological Sensing (Chapter 4)

Originally published at the 26th International Conference on the World Wide Web [Althoff et al., 2017a].

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 [Roenneberg, 2013].

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 study demonstrates the feasibility of using online interactions for large-scale physiological sensing.

1.2.4 Mental Health: Identifying Successful Conversation Strategies Through Large-scale Analysis of Counseling Conversations (Chapter 5)

Originally published in TACL [Althoff et al., 2016a].

Mental illness is one of the most pressing public health issues of our time. While counseling and psychotherapy can be effective treatments, our knowledge about how to conduct successful counseling conversations has been limited due to lack of large-scale data with labeled outcomes of the conversations.

In this chapter, we present the largest, quantitative study to date on the discourse of text-message-based counseling conversations. We develop a set of novel computational discourse analysis methods to measure how various linguistic aspects of conversations are correlated with conversation outcomes. Applying techniques such as sequence-based conversation models, language model comparisons, message clustering, and psycholinguistics-inspired word frequency analyses, we discover actionable conversation strategies that are associated with better conversation outcomes.

We find that more successful counselors (1) are aware of how the conversation is going and *adapt* accordingly, (2) *react* differently to virtually identical situations (making texters feel more comfortable through affirmation, clarify situations by writing more and reflecting back to check understanding), (3) use less generic or "templated" responses but instead write more *creative* and personalized messages. (4) *make progress* getting to know the core issue quickly and moving on to collaboratively solve the problem, and (5) are able to *facilitate perspective change* by helping the texter to be more positive, think about the future, and consider others as well.

Chapter 2

Physical Activity: Planetary-scale Smartphone Data Reveal Activity Inequality

2.1 Introduction

Physical activity improves musculoskeletal health and function, prevents cognitive decline, reduces symptoms of depression and anxiety, and helps maintain a healthy weight [Sallis et al., 2016a; WHO, 2010]. While prior surveillance and population studies have revealed that physical activity levels vary widely between countries [Hallal et al., 2012], more information is needed about how activity levels vary within countries and the relationships between physical activity disparities, health outcomes (*e.g.*, obesity levels), and modifiable factors such as the built environment. For example, while much is known about how both intrinsic (*e.g.*, gender, age, and weight) and extrinsic (*e.g.*, public transportation density) factors are related to activity levels, evidence about how these factors interact (*e.g.*, the influence of environmental factors on older adults or obese individuals) is more limited [Bauman et al., 2012]. Understanding these interactions is important for developing public policy [Chokshi and Farley, 2014; Physical Activity Guidelines Advisory Committee, 2008], planning cities [Sallis et al., 2016b], and designing behavior change interventions [Reis et al., 2016; Servick, 2015].

The majority of physical activity studies are based on information that is either self-reported, with attendant biases [Prince et al., 2008], or measured via wearable sensors, but limited in the number of subjects, observation period, and geographic range [Van Dyck et al., 2015]. Mobile phones are a powerful tool for studying large-scale population dynamics and health on a global scale [Servick, 2015; Walch et al., 2016], revealing the basic patterns of human movement [González et al.,

2008], mood rhythms [Golder and Macy, 2011], the dynamics of the spread of diseases such as malaria [Wesolowski et al., 2012], and socioeconomic status in developing countries [Blumenstock et al., 2015]. Smartphones are now being used globally, with the adoption rate among adults at 69% in developed countries and 46% in developing economies and growing rapidly [Anthes, 2016]. With onboard accelerometers for automatic recording of activity throughout the day, smartphones provide a scalable tool to measure physical activity worldwide. Here, we use a large-scale physical activity dataset to quantify disparities in the distribution of physical activity in countries around the world, identify the relationship between activity disparities and obesity, and explore the role of the built environment, in particular walkability, in creating a more equal distribution of activity across populations.

2.2 Results

2.2.1 Dataset

We study 68 million days of minute-by-minute step recordings from 717,527 anonymized users of the Argus smartphone application developed by Azumio. The dataset includes recordings of physical activity for free-living individuals from 111 countries (Figure 2.1a). We focus on the 46 countries with at least 1000 users (Table 2.1); 90% of these users were from 32 high income countries and 10% were from 14 middle income countries (including 5 lower-middle income countries; Methods). The average user recorded 4961 steps per day (standard deviation $\sigma = 2684$) over an average span of 14 hours. We verified that the smartphone application data reproduces established relationships between age, gender, weight status, and activity (Figure 2.6), as well as country-level variations in activity and obesity levels determined from prior surveillance data and population studies (Figure 2.7). Recent research has further demonstrated that smartphones provide accurate steps counts [Case et al., 2015] and reliable activity estimates in both laboratory and free-living settings [Hekler et al., 2015]. We perform complete-case analyses, which we accompany with sample-correction, stratification, outlier, and balance testing to verify that our conclusions are robust to missing data (see Table 2.1) and biases in age and gender, and hold for both high and middle income countries (see Methods and Figures 2.11, 2.4).

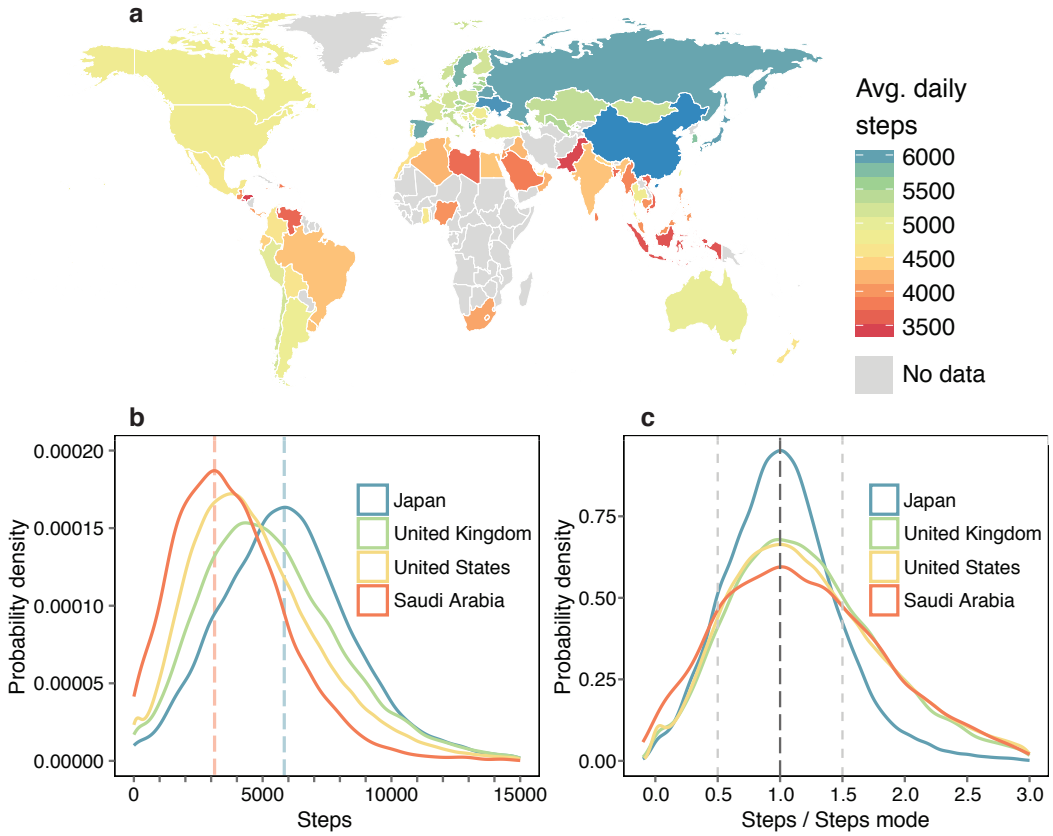

Figure 2.1 – Smartphone data from over 68 million days of activity by 717,527 individuals reveal variability in physical activity across the world. **(a)** World map showing variation in activity (mean daily steps) measured through smartphone data from 111 countries with at least 100 users. Cool colors correspond to high activity (e.g., Japan in blue) and warm colors indicate low levels of activity (e.g., Saudi Arabia in orange). **(b)** Typical activity levels differ between countries. Curves show distribution of steps across the population in four representative countries as a normalized probability density (high to low activity: Japan, United Kingdom, United States, Saudi Arabia). Vertical dashed lines indicate the mode of activity for Japan (blue) and Saudia Arabia (orange). **(c)** The variance of activity around the population mode differs between countries. Curves show distribution of steps across the population relative to the population mode. In Japan, the activity of 76% of the population falls within 50% of the mode (*i.e.*, between light gray dashed lines), whereas in Saudi Arabia this fraction is only 62%. The United Kingdom and United States lie between these two extremes for average activity level and variance. *This map is based on CIA World Data Bank II data publicly available through the R package "mapdata".*

Table 2.1 – Summary of dataset statistics for the 46 countries with more than 1000 subjects (693,806 subjects in total; Methods). Countries are ordered by number of subjects in sample. Country-level analyses are restricted to these 46 countries. Percentages are in parentheses. NA refers to missingness in data. Table continued on next page with additional columns.

Country Name	#subjects	Mean Steps	Activity Inequality	#male	#female	#genderNA	Median Age	#AgeNA
United States	388124	4774	0.303	94707 (48.9)	98971 (51.1)	194446 (50.1)	34	168610 (43.4)
United Kingdom	55110	5444	0.288	15144 (54.8)	12508 (45.2)	27458 (49.8)	33	23557 (42.7)
Canada	26895	4819	0.303	7022 (49.2)	7250 (50.8)	12623 (46.9)	34	10962 (40.8)
Australia	26644	4941	0.304	6858 (51.4)	6479 (48.6)	13307 (49.9)	34	11075 (41.6)
Japan	20386	6010	0.248	6696 (76.2)	2090 (23.8)	11600 (56.9)	38	9016 (44.2)
China	17427	6189	0.245	7553 (61.3)	4769 (38.7)	5105 (29.3)	28	5097 (29.2)
Germany	12234	5205	0.266	4740 (72.8)	1775 (27.2)	5719 (46.7)	34	4666 (38.1)
India	11148	4297	0.293	4092 (79.0)	1086 (21.0)	5970 (53.6)	33	4818 (43.2)
France	8185	5141	0.268	2833 (67.2)	1384 (32.8)	3968 (48.5)	33	3435 (42.0)
Russia	7911	5969	0.262	2071 (59.9)	1385 (40.1)	4455 (56.3)	28	3104 (39.2)
Spain	6723	5936	0.261	2496 (70.8)	1027 (29.2)	3200 (47.6)	36	2538 (37.8)
Netherlands	6239	5110	0.261	2092 (64.1)	1171 (35.9)	2976 (47.7)	35	2311 (37.0)
Mexico	5695	4692	0.279	1497 (65.0)	806 (35.0)	3392 (59.6)	32	2831 (49.7)
Italy	5567	5296	0.275	1724 (68.3)	801 (31.7)	3042 (54.6)	36	2528 (45.4)
Singapore	5411	5674	0.249	1567 (62.3)	947 (37.7)	2897 (53.5)	35	2273 (42.0)
Sweden	5177	5863	0.246	1309 (52.1)	1202 (47.9)	2666 (51.5)	34	2277 (44.0)
South Korea	5022	5755	0.247	1235 (66.5)	621 (33.5)	3166 (63.0)	33	2270 (45.2)
Taiwan	4821	5000	0.262	987 (64.6)	540 (35.4)	3294 (68.3)	34	2404 (49.9)
Hong Kong SAR China	4754	6880	0.222	1288 (62.0)	789 (38.0)	2677 (56.3)	33	2015 (42.4)
Turkey	4711	5057	0.264	1197 (54.5)	1000 (45.5)	2514 (53.4)	31	2106 (44.7)
Thailand	4615	4764	0.272	1026 (62.9)	604 (37.1)	2985 (64.7)	32	2438 (52.8)
Norway	4256	5246	0.252	1061 (52.3)	967 (47.7)	2228 (52.3)	30	1803 (42.4)
United Arab Emirates	4138	4516	0.281	1315 (66.1)	673 (33.9)	2150 (52.0)	33	1723 (41.6)
Brazil	3999	4289	0.272	1127 (71.3)	453 (28.7)	2419 (60.5)	33	1946 (48.7)
Denmark	3924	5263	0.262	1000 (57.1)	750 (42.9)	2174 (55.4)	33	1804 (46.0)
Saudi Arabia	3837	3807	0.325	1153 (64.8)	626 (35.2)	2058 (53.6)	29	1650 (43.0)
Malaysia	3787	3963	0.288	937 (53.5)	814 (46.5)	2036 (53.8)	30	1589 (42.0)
Belgium	3051	4978	0.276	881 (61.9)	542 (38.1)	1628 (53.4)	33	1299 (42.6)
New Zealand	2941	4582	0.301	706 (49.3)	727 (50.7)	1508 (51.3)	33	1235 (42.0)
Philippines	2892	4008	0.298	550 (51.7)	513 (48.3)	1829 (63.2)	31	1476 (51.0)
Ireland	2758	5293	0.285	718 (50.3)	709 (49.7)	1331 (48.3)	33	1159 (42.0)
South Africa	2718	4105	0.284	900 (65.3)	479 (34.7)	1339 (49.3)	35	1124 (41.4)
Ukraine	2420	6107	0.252	507 (56.6)	388 (43.4)	1525 (63.0)	27	1015 (41.9)
Indonesia	2326	3513	0.283	760 (67.4)	368 (32.6)	1198 (51.5)	31	925 (39.8)
Switzerland	2251	5512	0.263	820 (64.9)	444 (35.1)	987 (43.8)	37	774 (34.4)
Czech Republic	2132	5508	0.248	708 (71.2)	286 (28.8)	1138 (53.4)	32	929 (43.6)
Poland	2128	5249	0.269	643 (63.5)	370 (36.5)	1115 (52.4)	31	901 (42.3)
Israel	1489	5033	0.272	458 (65.0)	247 (35.0)	784 (52.7)	34	650 (43.7)
Finland	1488	5204	0.266	388 (50.5)	381 (49.5)	719 (48.3)	31	612 (41.1)
Romania	1422	4759	0.283	380 (64.0)	214 (36.0)	828 (58.2)	31	653 (45.9)
Portugal	1418	4744	0.276	431 (64.8)	234 (35.2)	753 (53.1)	34	614 (43.3)
Egypt	1213	4315	0.303	290 (72.0)	113 (28.0)	810 (66.8)	26	647 (53.3)
Greece	1159	4350	0.295	455 (74.2)	158 (25.8)	546 (47.1)	36	452 (39.0)
Hungary	1151	5258	0.273	357 (67.1)	175 (32.9)	619 (53.8)	30	519 (45.1)
Chile	1060	5204	0.263	270 (64.0)	152 (36.0)	638 (60.2)	31	525 (49.5)
Qatar	1049	4158	0.291	370 (71.4)	148 (28.6)	531 (50.6)	33	413 (39.4)

Table 2.1 – Summary of dataset statistics for the 46 countries with more than 1000 subjects (693,806 subjects in total; Methods). Countries are ordered by number of subjects in sample. Country-level analyses are restricted to these 46 countries. Percentages are in parentheses. NA refers to missingness in data. (Continued from page 10.)

Country Name	#BMI [15, 18.5)	#BMI [18.5, 25)	#BMI [25, 30)	#BMI [30, 35)	#BMI [35, 40)	#BMI [40, inf)	#BMI NA	#obese
United States	5940 (2.3)	95959 (37.3)	83818 (32.5)	41669 (16.2)	17410 (6.8)	12129 (4.7)	130604 (33.7)	71208 (27.7)
United Kingdom	1133 (3.1)	16093 (44.4)	11696 (32.3)	4643 (12.8)	1514 (4.2)	920 (2.5)	18857 (34.2)	7077 (19.5)
Canada	533 (3.0)	7599 (42.1)	5808 (32.2)	2486 (13.8)	926 (5.1)	618 (3.4)	8847 (32.9)	4030 (22.3)
Australia	588 (3.2)	7756 (42.0)	6091 (33.0)	2487 (13.5)	912 (4.9)	563 (3.0)	8172 (30.7)	3962 (21.4)
Japan	795 (5.5)	9739 (67.2)	3151 (21.7)	632 (4.4)	118 (0.8)	42 (0.3)	5889 (28.9)	792 (5.5)
China	1040 (8.4)	8377 (67.8)	2439 (19.8)	375 (3.0)	64 (0.5)	20 (0.2)	5080 (29.2)	459 (3.7)
Germany	243 (2.7)	4399 (49.0)	3043 (33.9)	971 (10.8)	200 (2.2)	112 (1.2)	3253 (26.6)	1283 (14.3)
India	188 (2.5)	3017 (39.6)	3131 (41.1)	968 (12.7)	196 (2.6)	78 (1.0)	3528 (31.6)	1242 (16.3)
France	302 (5.2)	3384 (58.2)	1596 (27.5)	370 (6.4)	111 (1.9)	36 (0.6)	2375 (29.0)	517 (8.9)
Russia	386 (6.8)	3152 (55.2)	1545 (27.1)	480 (8.4)	100 (1.8)	33 (0.6)	2204 (27.9)	613 (10.7)
Spain	131 (2.6)	2611 (51.0)	1743 (34.0)	473 (9.2)	124 (2.4)	31 (0.6)	1600 (23.8)	628 (12.3)
Netherlands	125 (2.6)	2552 (53.6)	1572 (33.0)	363 (7.6)	94 (2.0)	37 (0.8)	1481 (23.7)	494 (10.4)
Mexico	85 (2.3)	1605 (42.8)	1363 (36.4)	501 (13.4)	132 (3.5)	46 (1.2)	1949 (34.2)	679 (18.1)
Italy	157 (3.8)	2366 (57.3)	1224 (29.6)	281 (6.8)	63 (1.5)	26 (0.6)	1437 (25.8)	370 (9.0)
Singapore	217 (5.8)	2099 (56.0)	1071 (28.6)	271 (7.2)	50 (1.3)	26 (0.7)	1666 (30.8)	347 (9.3)
Sweden	103 (3.0)	1842 (53.1)	1080 (31.1)	317 (9.1)	88 (2.5)	31 (0.9)	1705 (32.9)	436 (12.6)
South Korea	177 (4.8)	2272 (61.4)	1032 (27.9)	184 (5.0)	26 (0.7)	7 (0.2)	1320 (26.3)	217 (5.9)
Taiwan	196 (5.3)	2289 (61.7)	973 (26.2)	208 (5.6)	31 (0.8)	6 (0.2)	1114 (23.1)	245 (6.6)
Hong Kong SAR China	268 (7.7)	2235 (64.6)	744 (21.5)	148 (4.3)	25 (0.7)	21 (0.6)	1292 (27.2)	194 (5.6)
Turkey	130 (4.1)	1556 (48.6)	1071 (33.5)	333 (10.4)	78 (2.4)	22 (0.7)	1511 (32.1)	433 (13.5)
Thailand	243 (7.2)	2014 (59.5)	809 (23.9)	236 (7.0)	49 (1.4)	24 (0.7)	1228 (26.6)	309 (9.1)
Norway	113 (3.7)	1562 (51.3)	961 (31.6)	303 (10.0)	75 (2.5)	23 (0.8)	1211 (28.5)	401 (13.2)
United Arab Emirates	66 (2.3)	1043 (37.0)	1100 (39.0)	435 (15.4)	109 (3.9)	52 (1.8)	1319 (31.9)	596 (21.1)
Brazil	52 (1.9)	1175 (41.8)	1061 (37.8)	383 (13.6)	98 (3.5)	33 (1.2)	1189 (29.7)	514 (18.3)
Denmark	102 (3.7)	1459 (53.3)	813 (29.7)	246 (9.0)	78 (2.8)	26 (0.9)	1186 (30.2)	350 (12.8)
Saudi Arabia	91 (3.8)	819 (33.8)	861 (35.5)	400 (16.5)	133 (5.5)	100 (4.1)	1414 (36.9)	633 (26.1)
Malaysia	152 (5.6)	1315 (48.1)	843 (30.8)	287 (10.5)	94 (3.4)	36 (1.3)	1051 (27.8)	417 (15.2)
Belgium	93 (4.1)	1287 (57.2)	656 (29.1)	163 (7.2)	39 (1.7)	9 (0.4)	800 (26.2)	211 (9.4)
New Zealand	67 (3.3)	861 (42.9)	668 (33.3)	281 (14.0)	82 (4.1)	43 (2.1)	932 (31.7)	406 (20.2)
Philippines	77 (4.0)	980 (50.9)	590 (30.6)	191 (9.9)	44 (2.3)	28 (1.5)	967 (33.4)	263 (13.7)
Ireland	40 (2.2)	851 (47.5)	584 (32.6)	203 (11.3)	65 (3.6)	32 (1.8)	966 (35.0)	300 (16.7)
South Africa	42 (2.3)	653 (36.5)	632 (35.3)	325 (18.1)	88 (4.9)	46 (2.6)	927 (34.1)	459 (25.6)
Ukraine	150 (8.6)	999 (57.1)	444 (25.4)	122 (7.0)	21 (1.2)	8 (0.5)	671 (27.7)	151 (8.6)
Indonesia	107 (6.5)	815 (49.4)	532 (32.2)	147 (8.9)	28 (1.7)	10 (0.6)	675 (29.0)	185 (11.2)
Switzerland	58 (3.4)	964 (57.0)	513 (30.3)	122 (7.2)	25 (1.5)	7 (0.4)	559 (24.8)	154 (9.1)
Czech Republic	71 (4.3)	823 (50.0)	534 (32.4)	156 (9.5)	51 (3.1)	9 (0.5)	485 (22.7)	216 (13.1)
Poland	73 (4.6)	843 (52.6)	508 (31.7)	133 (8.3)	29 (1.8)	11 (0.7)	526 (24.7)	173 (10.8)
Israel	41 (3.9)	506 (48.6)	337 (32.4)	110 (10.6)	32 (3.1)	9 (0.9)	448 (30.1)	151 (14.5)
Finland	38 (3.6)	587 (55.0)	308 (28.9)	85 (8.0)	30 (2.8)	16 (1.5)	421 (28.3)	131 (12.3)
Romania	79 (7.6)	523 (50.6)	300 (29.0)	90 (8.7)	28 (2.7)	9 (0.9)	389 (27.4)	127 (12.3)
Portugal	43 (4.0)	595 (55.8)	318 (29.8)	94 (8.8)	10 (0.9)	4 (0.4)	351 (24.8)	108 (10.1)
Egypt	18 (2.2)	312 (38.7)	283 (35.1)	126 (15.6)	42 (5.2)	20 (2.5)	406 (33.5)	188 (23.3)
Greece	17 (1.9)	378 (42.7)	332 (37.5)	119 (13.4)	26 (2.9)	11 (1.2)	274 (23.6)	156 (17.6)
Hungary	43 (4.9)	483 (54.9)	243 (27.6)	89 (10.1)	16 (1.8)	6 (0.7)	271 (23.5)	111 (12.6)
Chile	12 (1.6)	347 (47.5)	267 (36.6)	82 (11.2)	18 (2.5)	1 (0.1)	330 (31.1)	101 (13.8)
Qatar	23 (3.3)	216 (31.3)	272 (39.4)	130 (18.8)	33 (4.8)	9 (1.3)	358 (34.1)	172 (24.9)

2.2.2 Activity Inequality

Our large-scale activity measurements enable the characterization of the full distribution of activity within a population beyond activity level averages and including the tails of the distribution (Figure 2.1b). Consider two countries with divergent activity distributions, Japan and Saudi Arabia. In Japan, the mode of recorded steps is high (Figure 2.1b, dashed blue line; 5846 steps), while in Saudi Arabia it is low (Figure 2.1b, dashed red line; 3103 steps). In Saudi Arabia, the mode is low, but the variance of recorded steps across the population is larger

as well (Figure 2.1c). This larger variance means that while some individuals are highly active, others record very little activity even relative to the low country baseline.

We formally characterize these systematic differences in country-level activity distributions by measuring activity inequality, which we define as the Gini coefficient of the population activity distribution [Allison, 1978; Atkinson, 1970] (Figure 2.9). We find that not only is there inequality in how steps are distributed within countries, but that activity inequality is associated with higher obesity levels (Figure 2.2a). For example, Saudi Arabia has a high obesity rate in comparison to Japan. At the same time Saudi Arabia has lower average activity (Figure 2.1b) and a wider activity distribution (Figure 2.1c), that is, a higher activity inequality. This finding is independent of gender and age biases (Figure 2.11) and independent of a country's income level (high vs. middle; no lower income countries were included in our dataset; Figure 2.4). In fact, a country's activity inequality is a better predictor of obesity prevalence than the average volume of steps recorded (R^2 = 0.64 vs. 0.47; $p < 0.01$; Figure 2.10). For example, the United States and Mexico have similar average daily steps (4774 vs. 4692), but the United States exhibits larger activity inequality (0.303 vs. 0.279; 10th vs 7th deciles of country activity inequality distribution) and higher obesity prevalence (27.7% vs. 18.1%; 10th vs 8th deciles of country obesity prevalence distribution) compared to Mexico (Table 2.1).

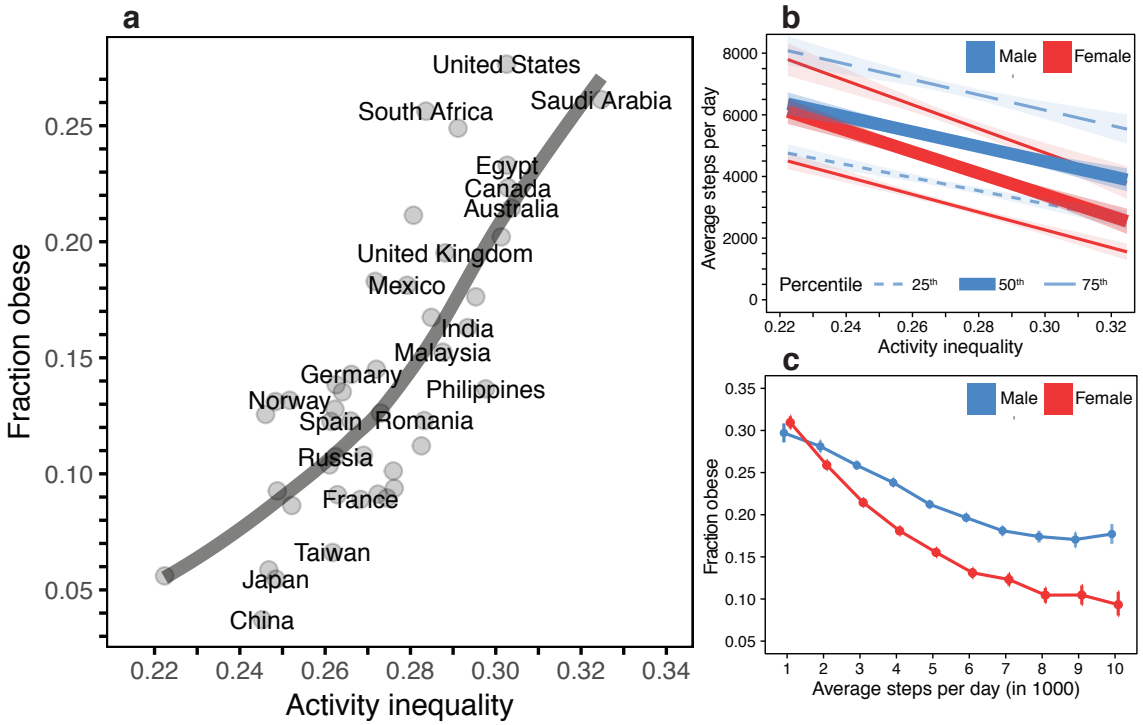

Figure 2.2 – Activity inequality is associated with obesity and increasing gender gaps in activity. **(a)** Activity inequality predicts obesity (LOESS fit; $R^2 = 0.64$). Individuals in the five countries with highest activity inequality are 196% more likely to be obese than individuals from the 5 countries with lowest activity inequality. **(b)** Activity inequality is associated with reduced activity, particularly in females. The figure shows the 25th, 50th, and 75th percentiles of daily steps within each country along with 95% confidence intervals (shaded) as a linear function of activity inequality. As activity inequality increases, median activity (50th percentile) decreases by 39% for males (blue) and by 58% for females (red). **(c)** Obesity-activity relationship differs between males and females and between high and low activity individuals. The plot shows the prevalence of obesity as a function of daily number of steps across all subjects in all countries (with 95% confidence intervals). For both males (blue) and females (red), a larger number of steps recorded is associated with lower obesity, but for females, the prevalence of obesity increases more rapidly as step volume decreases (232% obesity increase for females vs. 67% increase for males; comparing lowest vs. highest activity).

We find that in countries with high activity inequality, activity in females is reduced disproportionately compared to males, across all quartiles of activity (Figure 2.2b). In particular, 43% of activity inequality is explained by the gender gap in activity (Figure 2.12). Thus the larger variances we observe (Figure 2.1c) are due to reduced activity for females in comparison to males and not just an increase in variance overall (Figure 2.12a). While lower physical activity in females has been reported in several countries [Brown et al., 2016; Hallal et al., 2012], we discover that in countries with low activity and high activity inequality, the gender gap in activity is amplified (Figure 2.12b).

By quantifying the relationship between activity and obesity at the individual level (Figure 2.2c), we were able to determine why a country's activity inequality is a better predictor of obesity than average activity level. We find that the prevalence of obesity increases more rapidly for females than males as activity decreases. And while lower activity is associated with a significant increase in obesity prevalence for low activity individuals, there is little change in obesity prevalence among high activity individuals. So given two countries with identical average activity levels, the country with higher activity inequality will have a greater fraction of low activity individuals (Figure 2.1c), many of them female (Figure 2.2b), leading to higher obesity than predicted from average activity levels alone. These findings echo the phenomenon revealed in past studies of the effects of income inequality on health [Lynch et al., 2000; Wagstaff and Van Doorslaer, 2000], whereby a relatively small change in wealth (in our case activity) for an individual at the bottom of the distribution can lead to significant improvements in health. Based on our model relating activity inequality to obesity prevalence (Figure 2.2a), we also performed a simulation experiment which, assuming perfect information (Methods), suggests that interventions focused on reducing activity inequality could result in up to a 4 times greater reduction in obesity prevalence compared to population-wide approaches (Figure 2.13).

2.2.3 Activity Inequality and Walkability

We investigated the walkability of a city as a modifiable extrinsic factor that could increase activity levels [Bauman et al., 2012] and reduce activity inequality and the gender activity gap. Based on data from 69 United States cities (Table 2.2), we find that higher walkability scores are associated with lower activity inequality (Figure 2.3a) across all quartiles of median income (Figure 2.5). Examining San Francisco, San Jose, and Fremont—California cities in close geographic proximity—reveals that activity inequality is lowest in San Francisco, the city with the highest walkability (Table 2.3), suggesting that the relationship between walkability and activity inequality holds even for geographically and socioeconomically similar

cities. Furthermore, in more walkable cities, activity is higher on weekdays during morning and evening commute times and at lunch time and on weekends during the afternoon (Figure 2.3bc). This indicates that walkable environments increase physical activity during both work and leisure time.

We find that higher walkability is associated with significantly more daily steps across all age, gender, and BMI groups (Figure 2.3d). The relationship between walkability and activity is significantly stronger for females, whose activity was also disproportionately reduced with higher activity inequality, with the greatest increases for women under 50 years. For example, our linear model shows that for 40-year-old women, a 25 point increase in walkability (*e.g.*, from Sacramento, CA to Oakland, CA) is associated with 868 more steps per day, while for men, this 25 point increase is associated with only 622 additional daily steps. While walkability was associated with the greatest increases in recorded steps among normal weight individuals, even overweight and obese individuals in more walkable cities record more steps.

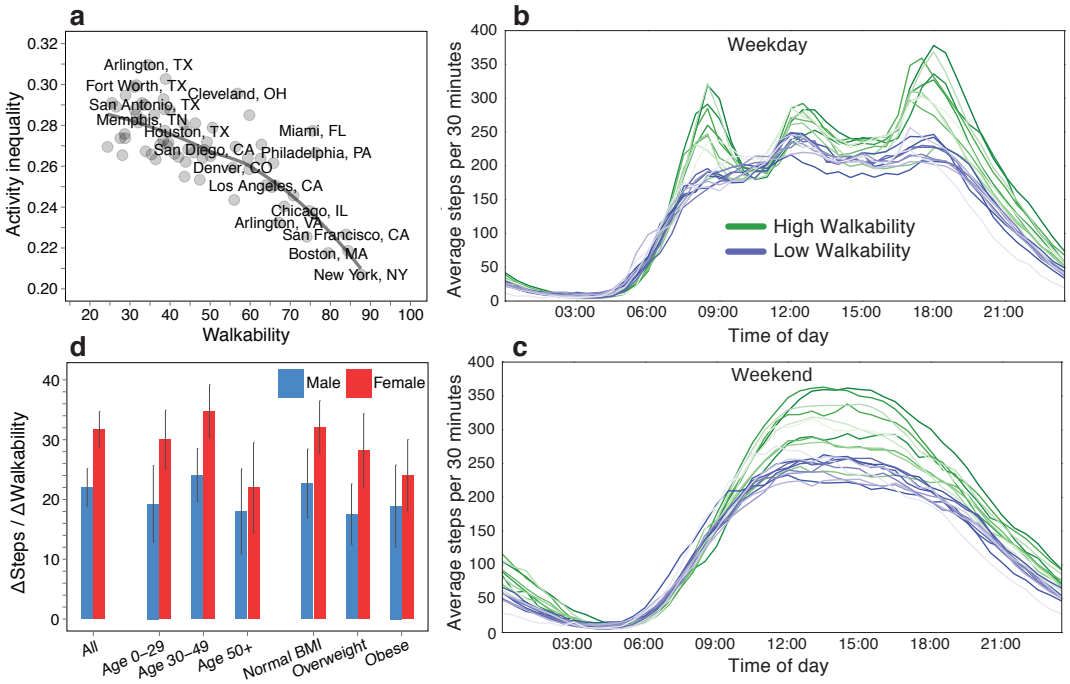

Figure 2.3 – Aspects of the built environment, such as walkability, may mitigate gender differences in activity and overall activity inequality. **(a)** Higher walkability scores are associated with lower activity inequality, based on data from 69 United States cities (LOESS fit; $R^2 = 0.61$). **(b,c)** Walkability is linked to increased activity levels. Curves show average steps recorded throughout the day in United States cities with the top 10 walkability scores (green) and bottom 10 walkability scores (blue). **(b)** On weekdays, walkable cities exhibit a spike in activity during morning commute (9:00), evening commute (18:00) and lunch times (12:00), while activity is relatively constant and lower overall in less walkable cities. **(c)** On weekend days, people in more walkable cities take more steps throughout the middle of the day, thus walkability is associated with higher activity levels even when most people do not work or commute. **(d)** Higher walkability is associated with more daily steps across age, gender, and BMI groups. Bars show the steps gained per day for each point increase in walkability score for 24 United States cities, including 95% confidence intervals (assuming linear model; Methods). Positive values across all bars reveal that, with increasing walkability, more steps are taken by every subgroup. The effect is significantly larger for females overall (left), with the greatest increases for women under 50 years (middle) and individuals with a BMI less than 30 (right).

2.2.4 Limitations

There are limitations in the instrument we used to collect daily physical activity. For example, our sample is cross-sectional and potentially biased towards individuals of higher socioeconomic status, particularly in lower income countries, and people interested in their activity and health. However, we find that activity inequality predicts obesity in both middle and high income countries (Figure 2.4) and that walkability predicts activity inequality across four quartiles of median income in U.S. cities (Figure 2.5), suggesting that our findings are robust to variation in socioeconomic status. The majority of adults in developed countries already own a smartphone and the number of smartphone connections worldwide is expected to increase 50% by 2020 [Anthes, 2016], so we expect any biases to diminish in the future. While walking is the most popular aerobic physical activity [Centers for Disease Control and Prevention, 2012], our dataset may fail to capture time spent in activities where it is impractical to carry a phone (*e.g.*, playing soccer) or steps are not a major component of the activity (*e.g.*, bicycling), and there may exist systematic differences in wear time based on gender and age because users must carry their phone for steps to be recorded. However, analysis of our dataset reproduces previously established relationships between activity across geographic locations, gender and age (Figures 2.6, 2.7). We also find that between countries, the span of time over which steps were recorded is uncorrelated with the number of steps (Figure 2.8), and thus systematic wear time differences are unlikely to affect our country-level comparisons. Together, these results provide confidence that our dataset is able to identify activity differences between countries, genders, and age groups.

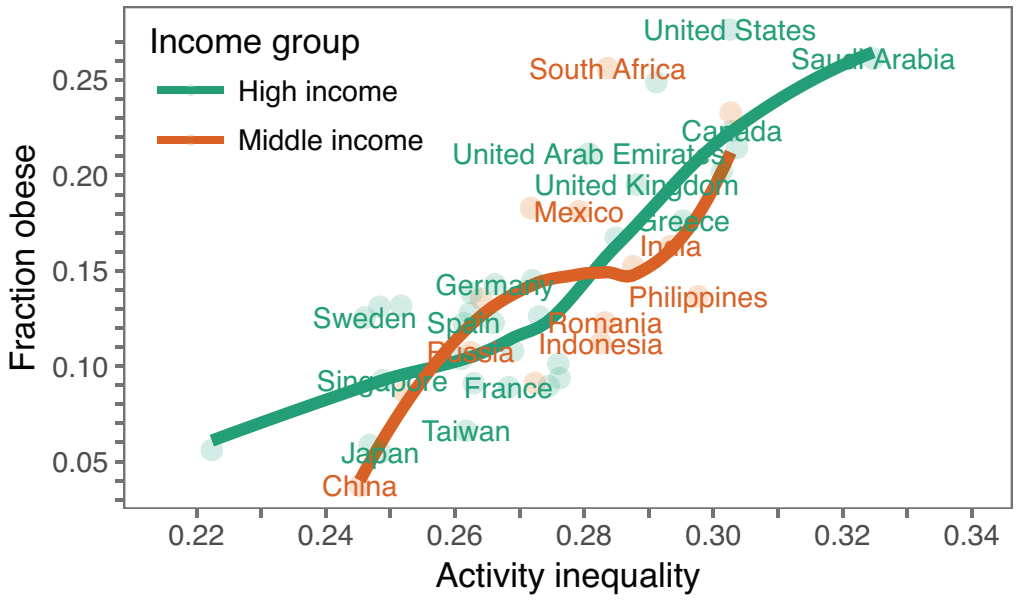

Figure 2.4 – Relationship between activity inequality and obesity holds within countries of similar income. Out of the 46 countries included in our main result, we have 32 high income (green) and 14 middle income (orange) countries according to the current World Bank classification [World Bank, a]. We find that activity inequality is a strong predictor of obesity levels in both high income countries as well as middle income countries. While in middle income countries, iPhone users might belong to the wealthiest in the population, in high income countries iPhones are used by larger parts of the population. The fact that we find a strong relationship between activity inequality and obesity in both groups of countries suggests that our findings are robust to differences in wealth in our sample.

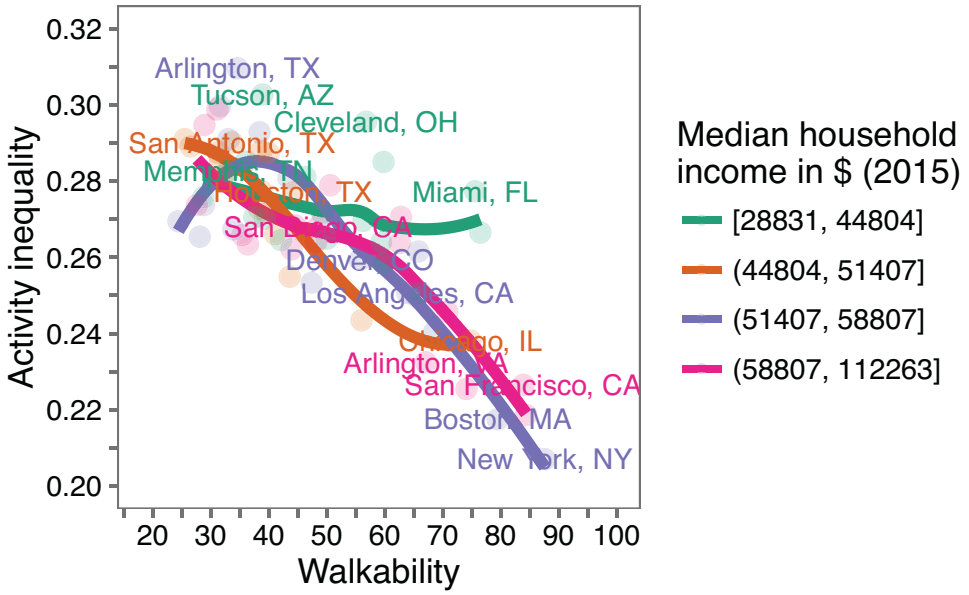

Figure 2.5 – Relationship between walkability and activity inequality holds within US cities of similar income. Walkable environments are associated with lower levels of activity inequality within socioeconomically similar groups of cities. We group the 69 cities into quartiles based on median household income (data from the 2015 American Community Survey [United States Census Bureau, b]). We find that walkable environments are associated with lower levels of activity inequality for all four groups. The effect appears attenuated for cities in the lowest median household income quartile. These results suggest that our main result—activity inequality predicts obesity and is mediated by factors of the physical environment—is independent of any potential socioeconomic bias in our sample.

2.2.5 Summary and Implications

This study presents a new paradigm for population activity studies by demonstrating that smartphones can deliver new insights about key health behaviors. We examine the distribution of activity in 46 countries around the world, including rarely studied countries such as Saudi Arabia and Mexico. Our findings highlight activity inequality as an important indicator of activity disparities in the population and identify "activity poor" subpopulations, such as women, who could most benefit from interventions to promote physical activity. We further find that walkability is associated with reduced activity inequality and greater activity across age, gender, and BMI groups, which indicates the importance of the built environment to global activity levels and health. Our findings can help us to understand the prevalence, spread, and effects of inactivity and obesity within and across countries and subpopulations and to design communities, policies, and interventions that promote greater physical activity.

2.3 Methods

2.3.1 Dataset Description

We analyzed anonymized, retrospective data collected between July 2013 and December 2014 from Apple iPhone smartphone users of the Azumio Argus app, a free application for tracking physical activity and other health behaviors. Data is available at http://activityinequality.stanford.edu. We define a step as a unit of activity as determined through iPhone accelerometers and Apple's proprietary algorithms for step-counting. The app records step measurements on a minute-by-minute basis. We considered only users with at least 10 days of steps data. The dataset contains 111 countries with 100 users or more (717,527 users; 68 million days of data; Figure 2.1a). We restricted further analyses to the 46 countries with at least 1000 or more users (693,806 users; 66 million days of data). We aggregated data from all of these users to the country level. A user's country was assigned based on the most common country identified through the user's IP addresses. In the United States, users were assigned to a city based on the most commonly occurring location of weather updates in the user's activity feed. Weather updates are automatically added to the feed of each user according to the nearest cell phone tower. The user enters gender, age, height, and weight in the app settings, and can change these values at any time; we used the most recent recorded values. 28.9% of users report multiple values for their weight; among these users, weight changed by 0.24 kg on average between the first and last recorded weight. Users had on average 95 days with recorded steps, although variation was large (standard

deviation $\sigma = 313$ days). Subjects were excluded from a particular analysis if information was unreported (*e.g.*, subjects with no reported height or weight were excluded from the analysis of Figure 2.2a). The amount of data for each country can be found in Table 2.1. To verify that subjects with missing data on gender, age, or BMI are not different from those who report data, we computed the standardized mean difference in age, gender, BMI, and average steps per day between groups with and without missing data. Across all combinations of missing variables (age, gender, BMI) and outcomes (age, gender, BMI, daily steps), the groups were balanced [Stuart, 2010], with all standard mean differences lower than 0.25. Data handling and analysis was conducted in accordance with the guidelines of the Stanford University Institutional Review Board.

2.3.2 Verifying Established Physical Activity Trends

To determine the ability of our dataset to identify relationships between physical activity and gender, age, BMI, and geographic location, we confirmed that the activity measure (daily steps) in our dataset reproduces trends established in prior work. We find that activity decreased with increasing age [Bassett et al., 2010; Bauman et al., 2012; Hallal et al., 2012; Troiano et al., 2008] and BMI [Bauman et al., 2012; Troiano et al., 2008; Van Dyck et al., 2015], and is lower in females than in males [Bassett et al., 2010; Bauman et al., 2012; Hallal et al., 2012; Troiano et al., 2008; Tudor-Locke et al., 2009], which is consistent with previous reports (Figure 2.6). We compared our physical activity estimates to physical activity data aggregated by the World Health Organization (WHO) [World Health Organization, b]. The comparison between recorded steps in our dataset and the WHO data is limited for the following reasons. The WHO's dataset is based on self-reports instead of accelerometer-defined measures as in our dataset. It contains the percentage of the population meeting the WHO guidelines for moderate to vigorous physical activity rather than recorded steps, and there is no published direct correspondence between the WHO data and daily steps. Furthermore, the confidence intervals in the WHO dataset are often very large and make a comparison complicated (*e.g.*, Japan: 28-89% meeting guidelines). Yet, we do observe moderate correlation between the two measures (r=0.3194; p=0.0393, Figure 2.7a). Similarly, we determined the correlation between obesity prevalence in a country in our dataset and comparable WHO estimates from 2014 [World Health Organization, a] (r=0.691; $p < 10^{-6}$; Figure 2.7b). In addition, we find a significant correlation between the gender gap in activity in our dataset and that reported by the WHO (Pearson r=0.52, $p < 10^{-3}$; Figure 2.7c). For these analyses we used the 46 countries with 1000 users in our dataset that also had WHO data [World Health Organization, a,b] (that excludes Hong Kong and Taiwan).

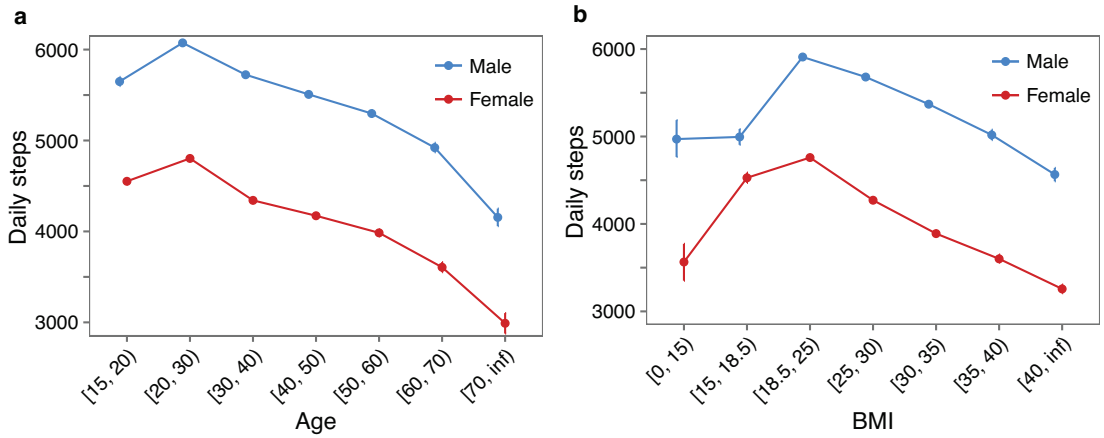

Figure 2.6 – Activity and obesity data gathered with smartphones exhibit well established trends. **(a)** Daily step counts across age and **(b)** BMI groups for all users. Error bars correspond to bootstrapped 95% confidence intervals. Observed trends in the dataset are consistent with previous findings; that is, activity decreases with increasing age [Bassett et al., 2010; Bauman et al., 2012; Hallal et al., 2012; Troiano et al., 2008] and BMI [Bassett et al., 2010; Bauman et al., 2012; Van Dyck et al., 2015], and is lower in females than in males [Bassett et al., 2010; Bauman et al., 2012; Hallal et al., 2012; Troiano et al., 2008; Tudor-Locke et al., 2009].

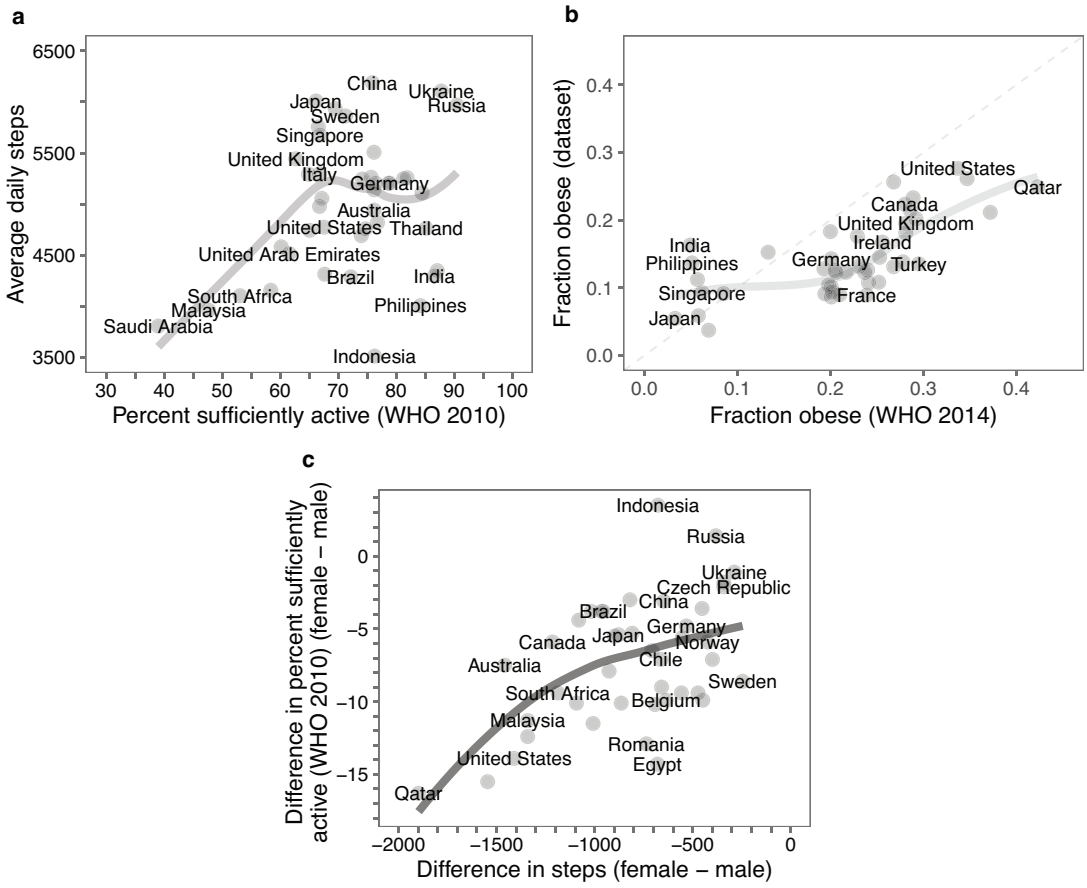

Figure 2.7 – Activity and obesity data gathered with smartphones are significantly correlated with previously reported estimates based on self-report. **(a)** WHO physical activity measure [World Health Organization, b] versus smartphone activity measure. The WHO measure corresponds to the percentage of the population meeting the WHO guidelines for moderate to vigorous physical activity based on self-report. The smartphone activity measure is based on accelerometer-defined average daily steps. We find a correlation of r=0.3194 between the two measures ($p < 0.05$). Note that this comparison is limited because there is no direct correspondence between the two measures—values of self-report and accelerometer-defined activity can differ [Prince et al., 2008], and the WHO confidence intervals are very large for many countries (Methods). **(b)** WHO obesity estimates [World Health Organization, a], based on self-reports to survey conductors, versus obesity estimates in our dataset, based on height and weight reported to the activity-tracking app. We find a significant correlation of r=0.691 between the two estimates ($p < 10^{-6}$). **(c)** Gender gap in activity estimated from smartphones is strongly correlated with previously reported estimates based on self-report. We find that the difference in average steps per day between females and males is strongly correlated to the difference in the fraction of each gender who report being sufficiently active according to the WHO (Pearson r=0.52, $p < 10^{-3}$).

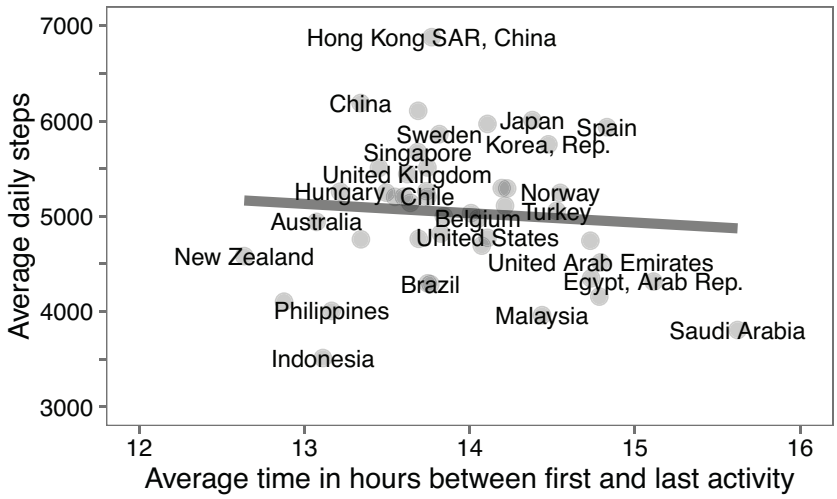

Figure 2.8 – Differences in country level daily steps are not explained by differences in estimated wear time. Users have an average span of 14.0 hours between the first and last recorded step, our proxy for daily wear time (Methods). While on an individual level, longer estimated wear time is associated with more daily steps (r=0.427, $p < 10^{-10}$), on a country level, there is no significant association between wear time and daily steps (r=-0.086, $p = 0.57$). Line shows linear fit using the 46 countries with at least 1000 users. This suggests that differences in recorded steps between countries are due to actual differences in physical activity behavior and are not explained by differences in wear time.

2.3.3 Daily Recorded Steps and Wear Time

We define a proxy for wear time of the activity-tracking smartphone as daily span of recorded activity; that is, the time between the first and the last recorded step each day. We find that users have an average wear time of 14.0 hours per day. To verify that differences in recorded steps between countries are not confounded by differences in wear time from country to country, we compared the average wear time in each country versus the average number of daily steps (Figure 2.8). We find no significant correlation (r=-0.086, p=0.57). Across the 46 countries, males have a 30 minute longer average wear time than women (14.2 vs. 13.7 hours), which is consistent with longer average sleep duration of females [Basner et al., 2007; Walch et al., 2016].

2.3.4 Defining Activity Inequality

We used the Gini coefficient [Allison, 1978; Atkinson, 1970] to compute activity inequality, as it is the most commonly used measure to quantify inequality and statistical dispersion [De Maio, 2007]. The Gini coefficient is based on the Lorenz curve, which plots the share of the population's total average daily steps that is cumulatively recorded by the bottom x% of the population (Figure 2.9). The Gini coefficient is the ratio of the area that lies between the line of equality and the Lorenz curve (marked A in the diagram) to the total area under the line of equality (marked A and B in the diagram): Gini Coefficient = A / (A + B). The Gini coefficient ranges from 0 (complete equality) to 1 (complete inequality), since physical activity is non-negative. Several other measures have been used to quantify inequality and statistical dispersion including the coefficient of variation [Allison, 1978; Atkinson, 1970], decile ratio [Kawachi and Kennedy, 1997], and others [De Maio, 2007; Kawachi and Kennedy, 1997]; we find that these measures are all highly correlated with the Gini coefficient (r=0.96 or higher) when applied to step counts within countries.

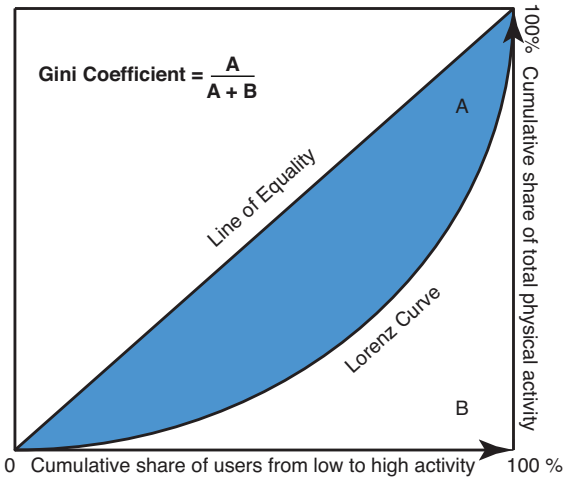

Figure 2.9 – Graphical definition of activity inequality measure using the Gini coefficient. The Lorenz curve plots the share of total physical activity of the population on the y-axis that is cumulatively performed by the bottom x% of the population, ordered by physical activity level. The diagonal line at 45 degrees represents perfect equality of physical activity (*i.e.,* everyone in the population is equally active). The Gini coefficient is defined as the ratio of the area that lies between the line of equality and the Lorenz curve (marked A in the diagram) over the total area under the line of equality (marked A and B in the diagram). The Gini coefficient for physical activity can range from 0 (complete equality) to 1 (complete inequality).

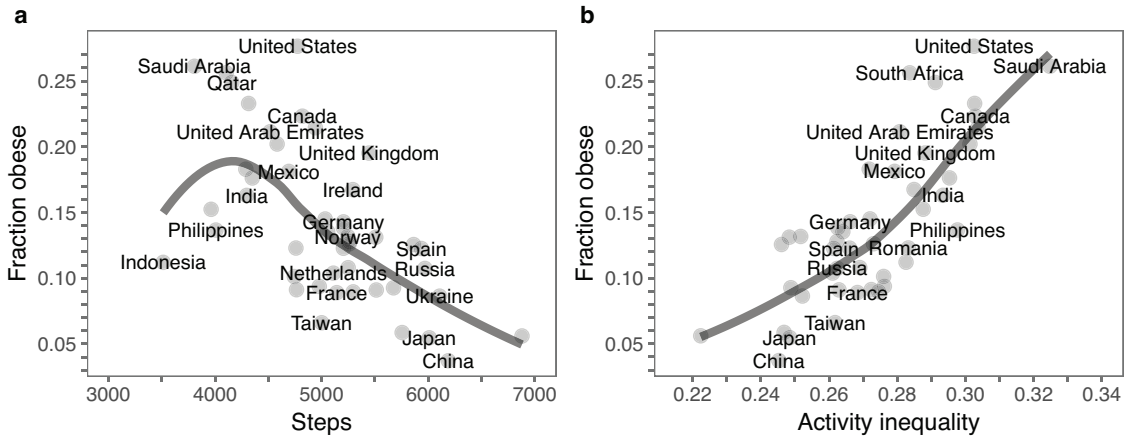

Figure 2.10 – Activity inequality is a better predictor of obesity than the the average activity level. **(a)** Obesity is significantly correlated with the average number of daily steps in each country (LOESS fit; $R^2 = 0.47$). **(b)** However, activity inequality is the better predictor of obesity (LOESS fit; $R^2 = 0.64$). The difference is significant according to Steiger's Z-Test ($p < 0.01$; Methods). This shows that there is value to measuring and modeling physical activity across countries beyond average activity levels. Activity inequality captures the variance of the distribution; that is, how many activity rich and activity poor people there are, allowing for better prediction of obesity levels. Figure repeated from Figure 2.2a for comparison.

2.3.5 Correlation between Activity Inequality and Obesity

We computed the Pearson correlation coefficient of activity inequality and the prevalence of obesity in a country (Figure 2.2a; r=0.79; $p < 10^{-10}$; $R^2 = 0.64$) using local polynomial regression fitting (LOESS; R statistical software package with a re-descending M estimator and Tukey's biweight function). We included all subjects with reported height and weight. We additionally correlated obesity with average daily steps for users in a country and compared the Pearson correlation coefficient for average daily steps with that for activity inequality (r=-0.62; $p < 10^{-5}$; $R^2 = 0.47$; Figure 2.10). Steiger's Z-Test [Steiger, 1980] shows that activity inequality is more strongly correlated with obesity than the average volume of steps recorded in a country (r = 0.79 vs. -0.62; $N = 46; t = 2.86; p < 0.01$). For example, even though the United Kingdom has higher average daily steps than Germany and France (5444 vs. 5205 and 5141), it exhibits higher obesity prevalence (19.5% vs. 14.3% and 8.9%). However, the high obesity levels in the United Kingdom are matched to their high activity inequality (0.288 vs. 0.266 and 0.268).

2.3.6 Robustness of Correlation between Activity Inequality and Obesity

While for some countries the gender ratio in our sample closely matched official estimates (*e.g.*, the United States, Canada, and Australia) in other countries our sample is more biased (*e.g.*, Japan, Germany, and India; Table 2.1). There is also a bias towards younger subjects in many countries (*e.g.*, median age for U.S. is 34 years vs. 37 years; United Kingdom is 33 vs. 40; Japan is 38 years vs. 46 years; Brazil is 33 years vs. 31 years). Our sample further includes both middle and high income countries, as classified by the World Bank [World Bank, a]. To verify the robustness of our results, we calculated gender-unbiased estimates for activity inequality and obesity prevalence for each country by reweighting males and females in our sample to exactly match World Bank estimates [World Bank, b] using a bootstrap [Efron and Tibshirani, 1994] with 500 replications. In addition, we computed activity inequality separately for males and females in each country and then correlated the activity inequality for each gender with obesity prevalance for that gender. We also computed the correlation between obesity and activity inequality for specific age groups in our dataset — [10,20), [20,30), [30,40), [40,50) and [50,100), again using only subjects with a reported age. In addition, we stratified countries by middle vs. high income status. In all cases, activity inequality remains a strong predictor of obesity (Figures 2.11, 2.4), which makes our findings independent of the exact age and gender distributions in our sample and suggests our results are not confounded by middle vs. high income status of countries or isolated to high income countries. Note that the results of these robustness analyses also show that our findings are not explained by patterns of missing data in our sample. We find similar results in analyses that include all subjects (Figure 2.2a) or only those that report gender (Figure 2.11a) or age (Figure 2.11b). We further verified that the relationship between activity inequality and obesity is not unduly driven by outliers. We removed the potential outliers of Indonesia, Malaysia, and the Philippines from our dataset and found that activity inequality was still a better predictor of obesity than average volume of steps recorded (R^2 was 0.69 for activity inequality vs. 0.56 for average steps).

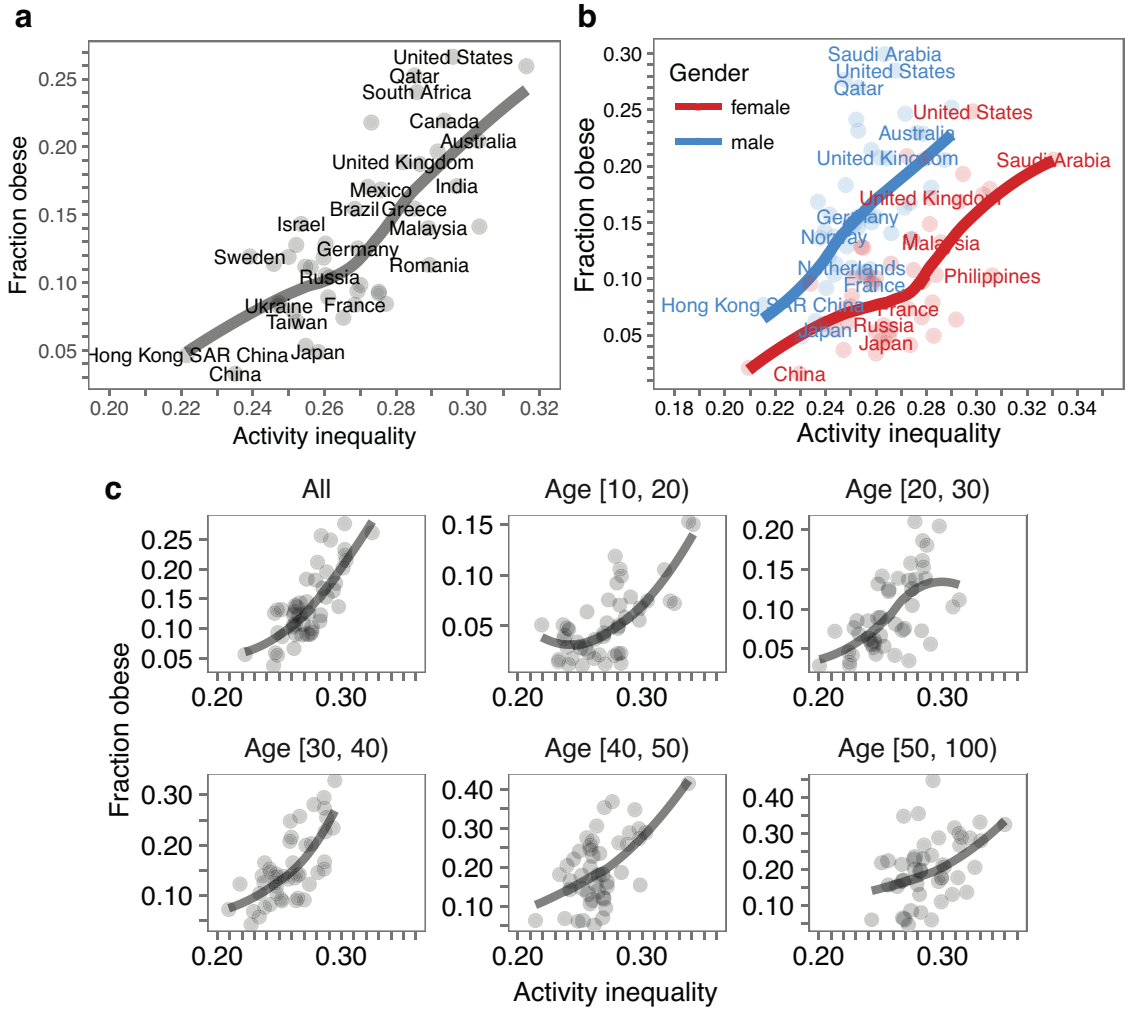

Figure 2.11 – Activity inequality remains a strong predictor of obesity levels across countries when reweighting the sample based on officially reported gender distributions and when stratifying by gender or age. **(a)** Obesity versus activity inequality on country level where subjects are reweighted to accurately reflect the official gender distribution in each country (Methods). The gender-unbiased estimates are very similar to estimates using all data (r=0.953 for activity inequality and r=0.986 for obesity). **(b)** Obesity versus activity inequality on a country level for males and females. Activity inequality predicts obesity for both genders. **(c)** Obesity versus activity inequality on a country level across different age groups. We find associations between activity inequality and obesity persists within every single age groups. Older people are more likely to be obese (see y-axis ranging from 5% to 45% obesity for subjects older than 50 years) and more likely to get little activity (*i.e.,* higher activity inequality on x-axis). These results indicate that our main result—activity inequality predicts obesity—is independent of any potential gender and age bias in our sample.

2.3.7 Gender Gaps in Activity and Obesity

To determine how activity varies with increasing activity inequality across countries, we calculated the 25th, 50th, and 75th percentile of daily steps in each country, with separate calculations for males and females. We then fitted a linear model based on each country's activity inequality to each percentile/gender group, along with 95% confidence intervals (Figure 2.2b). We determined the relationship between obesity prevalence and average daily steps for males and females in our sample by measuring the fraction of obese subjects who recorded a certain amount of activity (1-2k daily steps, 2-3k, ..., 10-11k) and then computing bootstrapped 95% confidence intervals (Figure 2.2c). This analysis included all subjects in the dataset who reported height and weight (N=297,268). We computed the proportion of variability explained by the gender gap in activity using the R^2 measure (Figure 2.12b).

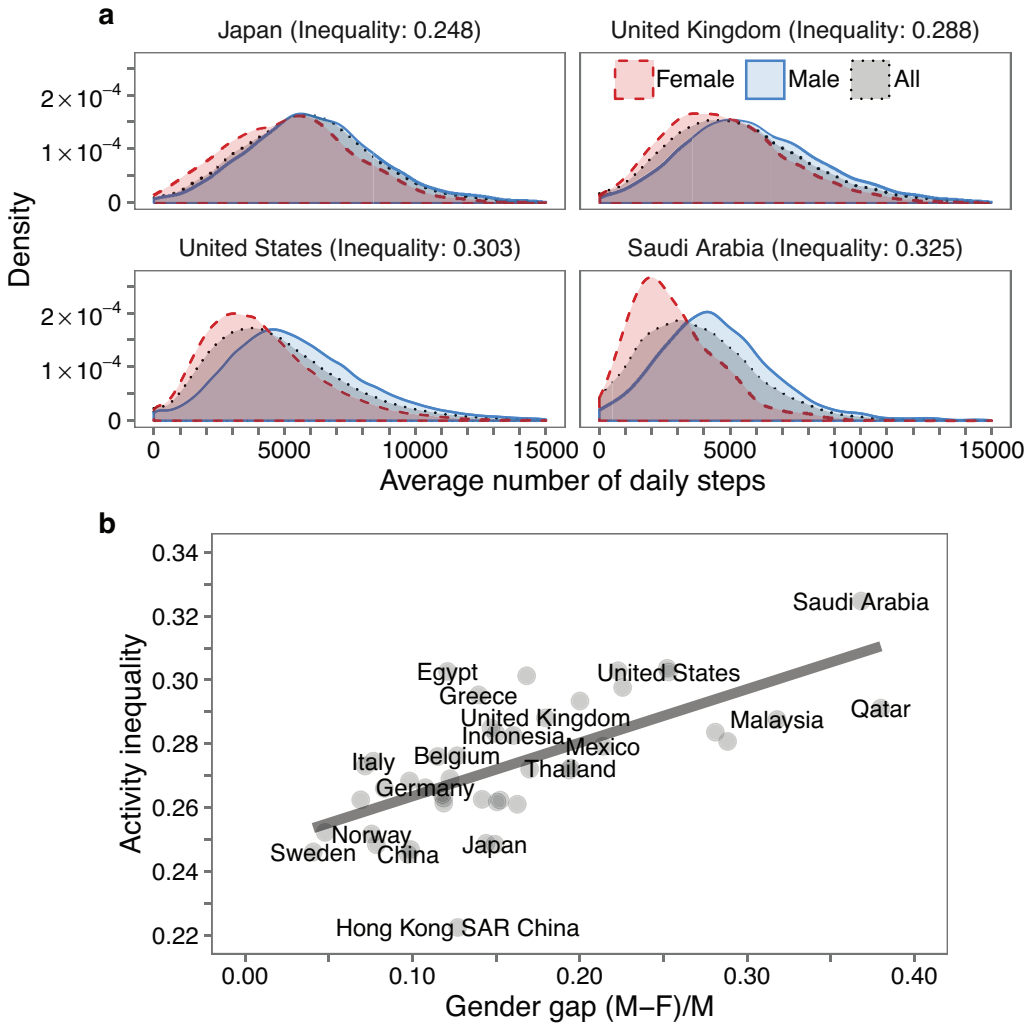

Figure 2.12 – Female activity is reduced disproportionately in countries with high activity inequality. **(a)** Distribution of daily steps for females, males, and all users in representative countries of increasing activity inequality (Japan, United Kingdom, United States, and Saudi Arabia). While in countries with low activity inequality females and males get very similar amounts of activity (*e.g.*, Japan), the distributions of female and male activity differ greatly for countries with high activity inequality (*e.g.*, Saudi Arabia and United States). Activity distributions in these countries demonstrate that larger variances in activity (Figure 2.1c) are due to a disproportionate reduction in the activity of females and not just an increase in variance overall. **(b)** Activity inequality increases with the relative activity gender gap on a country level (Methods). We find that the relative gender gap ranges between 0.041 (Sweden) and 0.380 (Qatar). The average daily steps for females is lower than for males in all 46 countries. The gender gap explains 43% of the observed variance in activity inequality (linear fit: $R^2 = 0.43$). This suggests that activity inequality could be reduced significantly through increases in female activity alone.

2.3.8 City Walkability Analysis

Walkability scores were obtained from Walk Score [Walk Score, 2016]. Scores are on a scale of 1 to 100 (100 = most walkable) and are based on amenities (*e.g.*, shops and parks) within a 0.25 to 1.5 mile radius (a decay function penalizes more distant amenities) and measures of friendliness to pedestrians, such as city block length and intersection density. At a city level, the score shows good correlation with gold standard, GIS-determined measures of walkability [Duncan et al., 2011]. For the 69 United States cities with at least 200 Azumio users (Table 2.2), we correlated walkability scores with the activity inequality on a city-level (*i.e.*, using the within-city distribution of average daily step counts). We verified that correlations between walkability and activity inequality are similar when controlling for the median income level of the city by grouping the 69 cities used in Figure 2.3a into quartiles based on median household income data from the 2015 American Community Survey [United States Census Bureau, b]. We find that walkable environments are associated with lower levels of activity inequality for all four median income groups (Figure 2.5). We next analyzed activity in our dataset throughout the day on weekdays and weekend days in the 10 cities with the highest walkability scores, and the 10 cities with the lowest walkability score. We only considered cities with at least 20,000 weekdays of tracked steps across all users for this analysis. We aggregated steps taken over time within each city to the the average number of steps per 30 minute interval. We only considered days with (1) at least 60 minutes with nonzero steps, (2) first and last recorded step at least 8 hours apart, and (3) recorded total steps between 500 and 100,000. We examined a subset of similar cities in close geographic proximity to show that our results cannot be explained by simple differences in geographic variation or city populations (Table 2.3).

Table 2.2 – United States Cities sorted by their walk scores (only showing cities with at least 20,000 weekdays of data; Methods). We use the top 10 and bottom 10 cities for our analysis (Figure 2.3bc).

	City	Walkability Score		City	Walkability Score
1	New York, NY	87.6	29	Madison, WI	47.4
2	Jersey City, NJ	84.4	30	Tampa, FL	46.3
3	San Francisco, CA	83.9	31	Atlanta, GA	45.9
4	Boston, MA	79.5	32	Houston, TX	44.2
5	Philadelphia, PA	76.5	33	Irvine, CA	43.9
6	Miami, FL	75.6	34	Dallas, TX	43.6
7	Chicago, IL	74.8	35	Sacramento, CA	43.4
8	Washington, DC	74.1	36	Omaha, NE	41.1
9	Seattle, WA	70.8	37	Columbus, OH	40.0
10	Oakland, CA	68.5	38	Albuquerque, NM	39.6
11	Arlington, VA	67.1	39	Orlando, FL	39.3
12	Baltimore, MD	66.2	40	Tucson, AZ	38.9
13	Long Beach, CA	65.8	41	El Paso, TX	38.7
14	Minneapolis, MN	65.4	42	Las Vegas, NV	38.6
15	Los Angeles, CA	63.9	43	Phoenix, AZ	38.3
16	Portland, OR	62.8	44	Austin, TX	35.4
17	Honolulu, HI	62.6	45	San Antonio, TX	33.7
18	Saint Louis, MO	59.8	46	Colorado Springs, CO	33.0
19	Pittsburgh, PA	59.8	47	Kansas City, MO	32.1
20	Milwaukee, WI	59.4	48	Fort Worth, TX	31.6
21	Cleveland, OH	56.8	49	Oklahoma City, OK	31.6
22	New Orleans, LA	56.3	50	Louisville, KY	31.2
23	Saint Paul, MN	56.0	51	Raleigh, NC	28.8
24	Denver, CO	55.7	52	Indianapolis, IN	28.7
25	Cincinnati, OH	50.1	53	Nashville, TN	26.5
26	Richmond, VA	49.2	54	Jacksonville, FL	25.5
27	San Diego, CA	48.5	55	Charlotte, NC	24.4
28	San Jose, CA	48.1			

Table 2.3 – Three United States cities in close geographic proximity. Increased walkability is associated with decreased activity inequality in this set of otherwise similar cities. For example, San Jose and Fremont are similar to San Francisco in terms of age, race distribution, and median household income. However, San Francisco has a higher walkability rating, lower activity inequality, and lower obesity levels. Therefore, we find that the discovered relationship between walkability and activity inequality holds even for cities that are geographically and socioeconomically similar. Walkability scores are from WalkScore.com [Walk Score, 2016], activity inequality and obesity are estimated from our dataset (marked *), and all other variables are taken from United States Census 2010 and the American Community Survey 2006-2010 [United States Census Bureau, a].

City	Walkability Score	Activity Inequality* (Percentile)	%Obese*	%White	%Asian	%Hispanic/Latino	%Black/African American	Median Age	Median Household Income ($)
San Francisco, CA	83.9	0.227 (0.07)	13.4	48.5	33.3	15.1	6.1	38.5	71,304
San Jose, CA	48.1	0.264 (0.33)	18.7	42.8	32.0	33.2	3.2	35.2	79,405
Fremont, CA	44.5	0.268 (0.48)	18.2	32.8	50.6	14.8	3.3	36.8	96,287

2.3.9 Impact of Walkability on Daily Steps

We computed the relationship between walkability and average daily steps for several subgroups of our sample. We used data from United States cities that had at least 25 Azumio users in each subgroup (Age 0-29, Age 30-49, Age 50+, normal BMI, overweight, obese, all; for both males and females). There are 24 such cities in the dataset (Table 2.4). The number of subjects for each group and city is shown in Table 2.4. For each group, we ran independent linear regressions of steps on walkability on a per-subject level. The models include an intercept coefficient. We determined the estimated coefficient of walkability (i.e., the increase in daily steps for each one point increase in walkability of a city) along with 95% confidence intervals (based on Student's t-distribution) for each subgroup (Figure 2.3d). We refer to the set of these coefficients as our linear model in the main text.

Table 2.4 – Number of subjects for each city and group used in the walkability analysis (Figure 2.3d). We use the 24 United States cities with at least 25 subjects across all groups (N=13,498 total; Methods).

City	female							male						
	Age 0-29	Age 30-49	Age 50+	normal BMI	over-weight	obese	all	Age 0-29	Age 30-49	Age 50+	normal BMI	over-weight	obese	all
Atlanta, GA	114	109	51	147	56	55	288	83	171	67	120	106	78	330
Austin, TX	106	109	48	138	71	46	283	83	142	72	109	99	74	309
Charlotte, NC	58	60	27	77	38	27	147	49	95	34	49	75	41	190
Chicago, IL	228	236	77	268	126	126	572	182	314	102	237	229	111	624
Cleveland, OH	51	53	36	53	35	43	145	29	48	35	27	52	28	119
Dallas, TX	95	80	50	93	61	62	235	69	129	67	96	104	62	279
Houston, TX	160	197	97	170	141	120	477	145	255	131	149	212	147	556
Indianapolis, IN	57	51	33	57	37	39	146	50	56	48	35	65	47	156
Jacksonville, FL	43	60	28	48	31	43	139	26	66	51	39	44	53	148
Las Vegas, NV	62	79	51	80	62	40	203	71	158	68	76	125	82	305
Los Angeles, CA	138	139	46	173	86	54	340	108	188	55	127	131	70	365
Miami, FL	84	87	56	103	60	42	233	70	144	83	68	130	73	305
New York, NY	240	222	92	322	131	62	583	168	288	100	230	200	91	573
Orlando, FL	67	64	46	81	44	39	188	58	116	48	64	73	74	231
Philadelphia, PA	132	119	42	126	68	89	311	95	100	36	67	91	59	238
Phoenix, AZ	59	65	37	71	44	33	171	43	105	48	55	75	52	203
Pittsburgh, PA	72	44	27	70	36	31	146	41	59	36	46	53	29	141
Portland, OR	70	117	37	103	66	47	234	43	114	57	74	75	57	224
Raleigh, NC	59	50	30	61	33	38	145	28	69	32	36	53	36	133
San Antonio, TX	90	116	38	84	70	85	259	77	147	64	65	99	113	299
San Diego, CA	115	129	52	140	80	57	308	86	191	77	113	148	80	372
San Francisco, CA	98	133	35	183	47	26	283	100	220	87	194	141	54	423
San Jose, CA	80	86	51	106	61	37	226	68	195	86	131	137	70	366
Seattle, WA	92	115	35	133	60	39	253	60	173	49	106	111	53	294
Total	2370	2520	1122	2887	1544	1280	6315	1832	3543	1533	2313	2628	1634	7183

2.3.10 Simulating population-level changes in activity

We used our model relating activity inequality to obesity prevalence to simulate how changes in activity might affect a country's obesity prevalence. We consider an activity budget of 100 additional daily steps per person in a country to distribute across the population (we found similar results for different activity budgets). We compared two strategies for distributing the steps–a population-wide distribution and an inequality-centric distribution. Both strategies result in the same shift in the average activity level of a country. For the population-wide distribution strategy, we increased each individual's daily activity by 100 steps. We then recomputed the country's activity inequality after the redistribution and estimated the country's new obesity prevalence based on our inequality-obesity model (line fit in Figure 2.2a). We next tested an activity inequality-centric strategy, where we distributed the activity budget equally among the activity-poorest X% of the population (*e.g.*, the bottom 20% of the population would increase their daily steps by 500). For the inequality-centric strategy, we computed the optimal fraction X for each country that results in the greated reduction in the country's activity inequality. Optimal values for X across all countries ranged from 5-9%. Further, assuming a fixed X (*e.g.*, X=10%) yielded similar results. Our simulation assumes perfect knowledge of population activity levels and perfect compliance; that is, any user targeted in this simulation would increase their activity levels according to the available budget. We also assume that other factors affecting weight would be held constant when activity levels change. In our simulations, the inequality-centric intervention resulted in reductions in obesity prevalence of up to 8.3% (median 4.0%; Figure 2.13), whereas the population-wide approach led to reductions of up to 2.3% (median 1.0%). Thus, activity inequality-centric interventions could result in up to a 4 times greater median reduction in obesity prevalence compared to the population-wide approach.

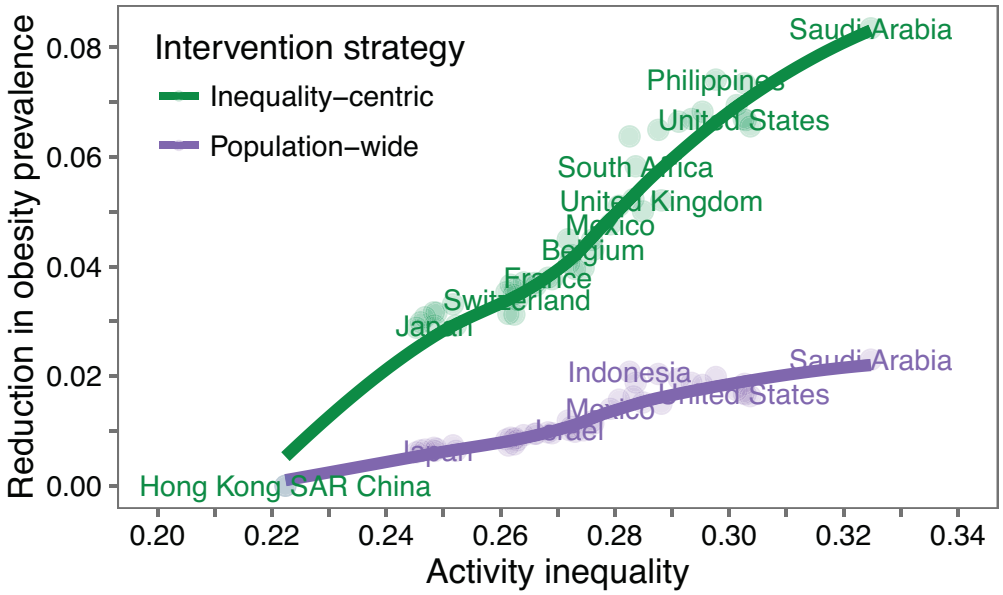

Figure 2.13 – Activity inequality-centric interventions could result in up to 4 times greater reductions in obesity prevalence than population-wide approaches. Given a fixed activity budget (100 daily steps per individual) to distribute across the population, we compare an inequality-centric strategy which equally distributes this budget to minimize activity inequality (100/X% daily steps increase for the activity-poorest X% where X minimizes the country's resulting activity inequality; Methods) and a population-wide strategy which equally distributes the budget across the entire population (100 daily steps per individual; Methods). Based on our simulations, we find that the inequality-centric strategy would lead to predicted reductions in obesity prevalence of up to 8.3% (median 4.0%), whereas the population-wide approach would lead to predicted reductions of up to 2.3% (median 1.0%).

Chapter 3

Activity Tracking: Modeling Real-World Action Sequences

3.1 Introduction

Activity tracking applications for mobile health have become an important part of people's daily lives. A US-nationwide study in 2013 found that 69% of adults keep track of a health indicator, and 21% among them used an app or device to do so [Fox and Duggan, 2013]. In activity logging applications such as Fitbit, Under Armour Record, and Argus, users might take one of many possible actions from a large and diverse space of potential actions at any point in time. Users continuously track many actions of their lives including exercise, diet, sleep, and commuting behavior with the goal of improving self-knowledge and personal well-being [Althoff et al., 2017b; Shameli et al., 2017; Swan, 2013] (see Chapter 2). User modeling is critical to making activity logging applications more useful by providing users with personalized experiences matching their specific objectives [Berkovsky et al., 2008; Du et al., 2016; Fischer, 2001; Gorniak and Poole, 2000; Zukerman and Albrecht, 2001]. This has the potential to significantly improve people's health, for instance by preventing negative health outcomes and promoting the adoption and maintenance of healthy behaviors [Althoff, 2017; Freyne and Berkovsky, 2010; Nahum-Shani et al., 2016; Thomas and Bond, 2015]. However, successful personalization of systems rests on the ability to predict the user's next actions and when they will occur [Davison and Hirsh, 1998; Du et al., 2015b; Zukerman and Albrecht, 2001].

Predicting actions is important because these predictions facilitate personalization of the user interface and user experience in order to provide users with what they need, without them asking for it explicitly [Mulvenna et al., 2000]. For example, in activity logging applications we can predict when the user will eat dinner and their future location in order to provide relevant recommendations [Yu et al., 2016]. Accurate and contextualized predictions could further help users to

realize their personal goals by reminding them to measure their weight or notifying them about the exercise the next morning [Swan, 2013]. Besides predicting the action itself, it is also critical to predict its timing, so that recommendations and reminders can be made at the right time. For instance, diet reminders ideally are delivered just *before* meal choices are made [Freyne et al., 2017; Nahum-Shani et al., 2016; Thomas and Bond, 2015]. More generally, predicting user actions also enables digital personal assistants that support users with relevant information including local recommendations, traffic, weather, events, and news [Du et al., 2016].

However, human behavior is extremely complex, which makes accurate predictions very challenging. In particular, human behavior is (1) *time-varying*, (2) *interdependent*, and (3) *periodic.* First, real-world actions *vary over time*, for example based on time of day (*e.g.*, spending time with friends in the evenings) and day of week (*e.g.*, going hiking on weekends) [Cheng et al., 2017; Koren, 2009]. Second, actions are also *interdependent* in the short-term and the long-term (*e.g.*, brushing teeth before going to bed, or drinking water after workouts). Third, humans are creatures of habit [Davison and Hirsh, 1998] and exhibit *periodic* behaviors [Das Sarma et al., 2012; Drutsa et al., 2017] (also see Chapter 4), such as brushing teeth every morning and evening.

Current user modeling techniques (*e.g.*, [Anderson et al., 2014; Benson et al., 2016; Davison and Hirsh, 1998; Gorniak and Poole, 2000; Kapoor et al., 2015; Koren, 2009; Lane, 1999; Trouleau et al., 2016; Zukerman et al., 1999]) do not jointly model all these key aspects (time variation, interdependence, periodicity) of real-world action sequences. However, failing to account for any of them results in decreased predictive performance. For example, consider the task of predicting the time of a user's next meal. When not accounting for periodicity, one would miss the fact that the user's early lunch might lead to an earlier dinner as well. However, this could be a critical mistake if the user relies on timely diet reminders.

While great advances have been made in modeling specific aspects of behavior in narrow application domains, in particular in the space of recommender systems [Koren, 2009] or information retrieval [Adar et al., 2008; Agichtein et al., 2006; Teevan et al., 2006], these lines of work have largely focused on consumption of items such as specific videos, songs, or websites [Anderson et al., 2014; Benson et al., 2016; Kapoor et al., 2015; Koren, 2009; Trouleau et al., 2016]. In all these cases, users repeat the *same* high-level actions such as watching one video after another. In contrast, we consider predicting *which* higher-level action, out of many, the user will take next; for example, whether they will watch a movie or go for a run (not which specific movie or run). Furthermore, previous work has often focused on predicting short-term actions such as the next unix command [Davison and Hirsh, 1998], web page request [Zukerman et al., 1999], or TV episode watched [Trouleau

et al., 2016]. Instead, we are interested in predicting longer-term actions such as a commute in the evening or a run the next morning.

This work. We present a new model for the task of predicting future user actions and their timing. First, we empirically demonstrate that action sequences exhibit time-varying, interdependent, and periodic patterns and that modeling them is critical to accurate predictions of user actions. Our model extends prior work on multivariate temporal point processes and is the first model to account for all three key properties. The model addresses (1) time-varying propensities of actions through mixture of Gaussians, (2) short-term dependencies between actions through a Hawkes process, and (3) long-term periodicity with time-dependent Weibull distributions. We call this model *TIPAS* referring to Time-varying Interdependent Periodic Action Sequences. TIPAS is personalized to each user through learning user-specific action preferences. We further develop an EM-based algorithm to fit this model using maximum likelihood estimation.

We demonstrate that TIPAS can scale to real-world datasets from Argus and Under Armour activity logging applications that capture 12 million actions taken by 20 thousand users over 17 months. We evaluate our model on these two activity logging datasets capturing ten different real-world actions, and demonstrate that we can predict the user's next logged activity (*e.g.*, run, eat, or sleep) and the timing of that activity (continuous, non-discretized timestamp) based on the user's previous actions and their timing.

Further, we show that TIPAS accurately captures all three fundamental behavioral patterns in real-world data. Using several domains of real-world actions, we demonstrate that our model outperforms eleven existing approaches on tasks of predicting actions by up to 156% as well as predicting when they will occur by up to 37%. Further, we show that performance improvements over baselines are particularly large for rare actions, increasing prediction accuracy over baselines by up to 256%. We find that these performance improvements are crucially enabled by modeling time-varying propensities of actions and their dependencies, and by modeling long-term periodicities of actions. Empirically, modeling time-varying propensities of actions yields 53% and 40% accuracy on the two activity logging datasets. Modeling short-term dependencies between actions improves this to 59% and 49%, respectively. Also capturing long-term periodicities of actions further improves this to 61% and 51%, respectively. Thus, capturing these three properties is essential to predicting periodic and interdependent human action sequences.

3.2 Related Work

Predicting the next action. Much work has focused on predictions of next actions, including unix commands [Davison and Hirsh, 1998], user interface actions to enable interface adaption [Gorniak and Poole, 2000], web page requests allowing for prefetching and latency reduction [Zukerman et al., 1999], clicks on web search [Agichtein et al., 2006], user behavior anomalies [Lane, 1999], product item preferences [Koren, 2009; Rendle et al., 2010], online purchases [Kooti et al., 2016], mobile apps used [Baeza-Yates et al., 2015], and future location-based checkins [Ashbrook and Starner, 2003; Bohnert et al., 2008; Liu et al., 2016]. Many of these works (*e.g.*, [Ashbrook and Starner, 2003; Bohnert et al., 2008; Kapoor et al., 2015; Lane, 1999]) have formulated the problem as a discrete-time sequence prediction task and used Markov models. However, Markov models assume unit time steps and are further unable to capture long-range dependencies since the overall state-space will grow exponentially in the number of time steps considered [Du et al., 2016]. Other works have used LSTM models [Hochreiter and Schmidhuber, 1997], which also assume discrete time steps and are limited in their interpretability.

In contrast, we also model and predict *when* the next action will occur, which is critical to surface recommendations and reminders at the right time. In addition, instead of specific web queries or item consumption, we consider a broader set of higher-level actions such as watching a movie, going for a run, or going to sleep.

Patterns of repeat consumption. Another line of work has studied repeated actions, in particular in the space of item consumption, including video binge watching [Trouleau et al., 2016], music listening [Kapoor et al., 2015], web page revisitation patterns [Adar et al., 2008], and repeated web search queries [Teevan et al., 2006]. More recent work has focused on modeling these behaviors and proposed models based on patterns of boredom [Benson et al., 2016; Kapoor et al., 2015] and recency [Anderson et al., 2014].

Importantly, patterns of human actions in the real world, which are modeled in this chapter, are fundamentally different from patterns of item consumption due to their higher-level notion (*e.g.*, watching a movie, not which specific one). For example, patterns of boredom [Benson et al., 2016; Kapoor et al., 2015] suggest that the probability of repeating an action within a short amount of time is unlikely. In contrast, we empirically observe the opposite in some cases, such as users commuting one way being extremely likely to commute back in the near future. More generally, real-world actions are characterized by more complex dynamics including time-varying behavior, interdependence, and periodicity of actions.

Temporal point processes. Recent work has considered temporal point processes [Cox and Isham, 1980] including Poisson and Hawkes [Hawkes, 1971] process-based models to predict the timing of future actions. Temporal point processes have been used to predict continuously time-varying item preferences [Du et al., 2015b], and to model user influence in a social network [Iwata et al., 2013; Tanaka et al., 2016; Zhou et al., 2013], the co-evolution of information and network structure [Farajtabar et al., 2015], competition between products [Valera and Gomez-Rodriguez, 2015], mobility patterns in space and time [Du et al., 2016], user return times [Kapoor et al., 2014], and temporal document clustering [Du et al., 2015a; Mavroforakis et al., 2017]. Perhaps the closest works to ours are by Du et al. [Du et al., 2016, 2015b], which also attempt to predict both future user actions *and* their timing.

We extend this line of work by explicitly modeling time-varying action propensities as well as developing a novel combination of Exponential and Weibull kernels to model short-term and long-term periodic dependencies between actions. Further, we demonstrate that these aspects are critical when predicting real-world user actions and their timing across two real-world activity logging datasets.

3.3 Task Description

The task considered in this chapter is, given a user and her history, a timestamped sequence of her actions in the past, to predict the user's future actions and the timing of these actions.

Formally, let U be a set of users. Each user $u \in U$ has an action sequence, which we represent as a user history $H_u = \{(a_{un}, t_{un})\}_{n=1}^{N_u}$ with a total of N_u events. Each element in H_u is an event consisting of an action and timestamp representing that user u takes action $a_{un} \in A$ at time $t_{un} \in \mathbb{R}^+$ ($0 \leq t_{un} \leq T$). T denotes the end of our observation period. For example, a_{un} could correspond to watching a movie or going for a run (but not which specific movie or run). We assume that events are sorted by their timestamps, $t_{un} \leq t_{un'}$ for $n < n'$. We denote the set of events before time t in user history H_u as $H_{ut} = \{(a', t') | (a', t') \in H_u \text{ and } t' < t\}$.

The task of predicting future user actions and when they will occur can now be formalized as follows. Given user history H_{ut} up until time t, predict the next K actions the user will take and their timing $\{(a_k, t'_k)\}_{k=1}^{K}$, where $t'_k > t$ (*i.e.*, these are the actions with the smallest $t'_k > t$ among all possible future user actions).

Here, we propose a novel multivariate temporal point process model for this prediction task and focus on the case of $K = 1$.

3.4 Empirical Observations

Next we make a series of empirical observations about important properties of real-world action sequences that will provide the basis for our statistical model TIPAS (Section 3.5). Accounting for these observations will lead to superior predictive models (Section 3.6).

3.4.1 Dataset Description

To illustrate critical properties of real-world actions we use a dataset of logged activities from a mobile activity logging application, Argus by Azumio, used in previous work on activity logging [Althoff et al., 2017b; Shameli et al., 2017] (also see Chapter 2). This smartphone app allows users to track their various daily activities including drink, sleep, heart rate, running, weight, food, walking, biking, workout, and stretching actions. For example, the drink action is logged to keep track of the user's daily fluid intake and the workout action is used to log various indoor exercises such as weightlifting or indoor-cycling. This dataset includes over four thousand active users taking 1.2 million actions over the course of seven months (all users logged at least two unique actions per day on average). Due to the popularity of the app, this set of users is very diverse in terms of age, gender, health status, country of origin, and other features (see Chapter 2). We note that the following properties of real-world actions also hold in other datasets including Under Armour activity logging app data (Section 3.6.1).

3.4.2 Properties of Real-World Action Sequences

Next, we describe three important properties of real-world action sequences and present empirical justification for each. TIPAS will explicitly address all three properties (Section 3.5).

Time-varying propensities of actions. Human real-world actions vary over time, for example based on time of day (*e.g.*, having meals in the morning, at mid-day, and in the evening) and day of week (*e.g.*, working out on the weekends). This dynamic is evident in real-world data of human activities as illustrated in Figure 3.1. The figure shows the distribution of the timing of three types of actions throughout the day: wake-up (from sleep), food, and bike. First, we observe that all three distributions are clearly non-uniform over time. For example, wake-up actions are clustered at around 07:00 hours (7 am). Second, we observe significant differences in the propensities to take different actions. While for sleep we observe a uni-modal distribution concentrated in the early morning, we observe a bi-modal distribution for biking. The two modes in the morning and evening

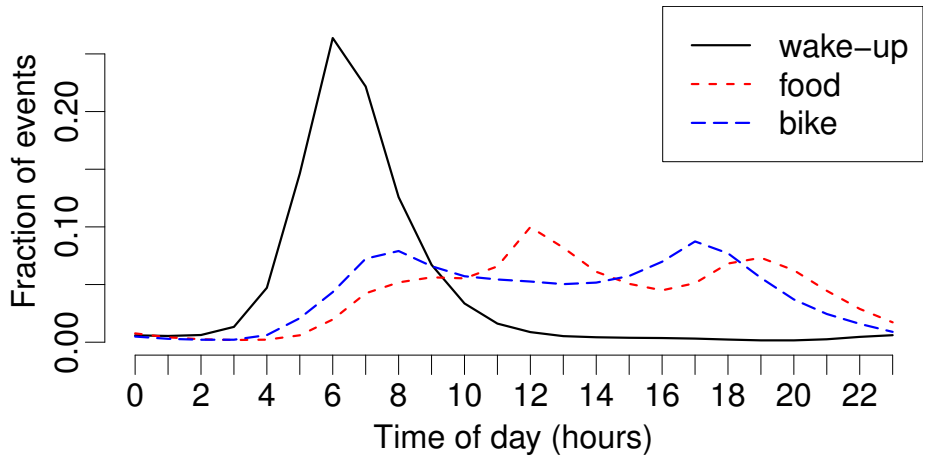

Figure 3.1 – Fraction of events within each time-of-day window. Notice that action propensity is clearly non-uniform and sometimes multi-modal.

likely correspond to commute activity where users log their rides to and from work. We also observe two clear modes for food during lunch and dinner times. However, breakfast times seem to vary more widely across users and are more dispersed. Summarizing, we observe non-uniform, temporal distributions with varying number of modes that vary across actions.

Short-term dependencies between actions. Certain actions make it more likely that some other actions will follow shortly. For example, people might drink water right after exercising or stretch right before running. In order to examine the short-term correlations between actions, we extract interarrival times between pairs of actions (*i.e.*, the elapsed time between the two actions) from a set of action histories. Figure 3.2 shows the distribution of interarrival times for several pairs of user actions after run actions (left) and sleep actions (right). We make two important observations. First, the monotonically decreasing curves show that the likelihood of other actions is largest right after an action has happened. After this, the likelihood declines very quickly in a monotonic manner (note the log scale of the Y-axis). This points to a self-excitation dynamic of logged human actions. For example, users are very likely to follow up on runs or waking up from sleep with drinking water or measuring their heart rate or weight. Specifically, about 50% of the weight measurements which happen within 6 hours of waking up occur right within the first 30 minutes. Second, we find that the action dependency patterns vary across actions. For example, drinking is more common after runs than after waking up and heart rate measurements fall off more sharply right after waking up than after runs. In summary, human actions in the real world often trigger other actions within a short period but these patterns are different across actions. We can leverage these correlations among actions when predicting future events.

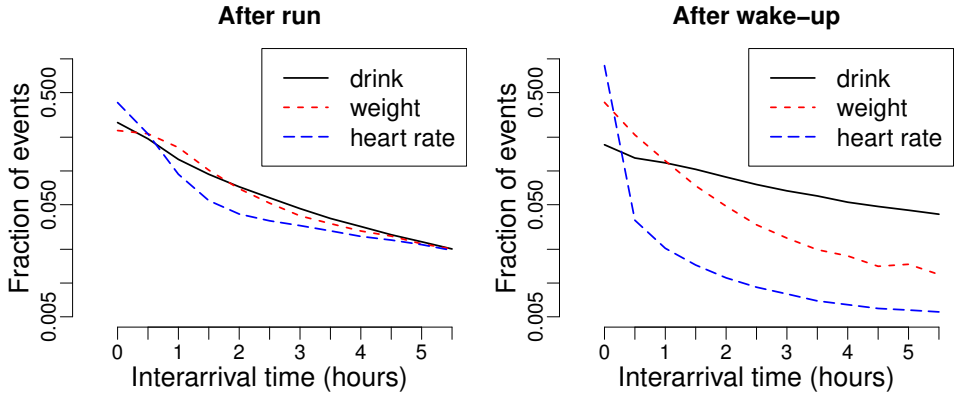

Figure 3.2 – Fraction of interarrival times at each time window (log scale). Figure shows drink, weight, and heart rate measurement actions taken after run (left) and wake-up (right) actions. Notice that the likelihood of drink, weight, and heart-rate actions declines quickly after both run and wake-up actions. However, note that fraction of heart-rate actions decreases much quicker after wake-up than after runs.

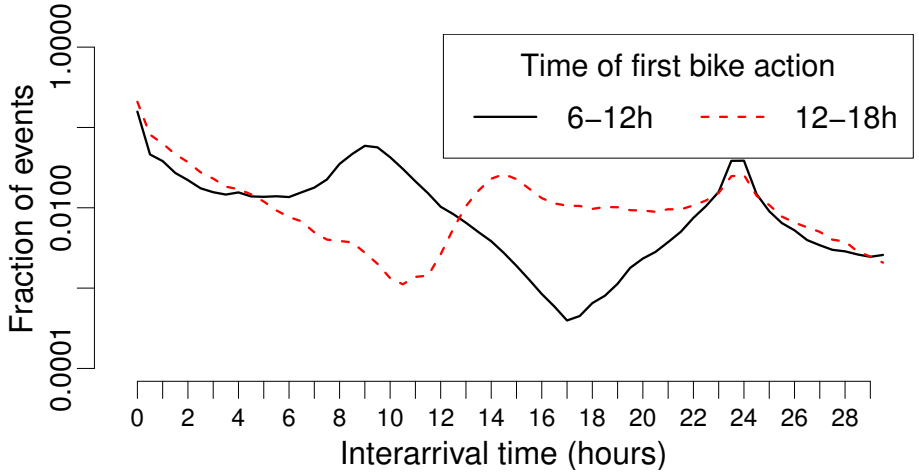

Figure 3.3 – Density describing when the next biking action will occur (interarrival time) given that the prior bike action occurred between 6-12h (solid black line) or between 12-18h (red dashed line) after midnight (timing, not duration). Notice the multiple and different modes of the two distributions indicating that biking actions recur periodically but that the period timing depends on the time of day.

Long-term periodic effects. Humans exhibit periodic behaviors such as waking up at about the same time every morning or commuting back home after about 8 hours of work. Therefore, logged real-world actions likely follow periodic recurrence patterns in which the same action tends to recur at certain, regular intervals. While some of these periodic behaviors are rooted in intrinsic biological rhythms such as sleep (Chapter 4), others are dictated by extrinsic factors (*e.g.*, when does one have to be in the office in the morning), or based on personal habits [Davison and Hirsh, 1998] (*e.g.*, measuring one's weight before breakfast). We illustrate these dynamics using interarrival times between bike events in real-world data. Figure 3.3 shows the distribution of interarrival times up to a maximum of 30h, where the two curves represent observed dynamics when the first of the two bike actions occurred during specific times of day (6-12h in solid black and 12-18h in dashed red line; note that these correspond to the timing and not the duration of the bike action).

We make two important observations. Previously, we had observed that short-term dependencies between actions exhibit monotonic decay. Here, we observe that this strong monotonic decay only holds within the first few hours and that we observe multiple additional peaks for both distributions after this initial phase. Second, we observe that these peaks occur at different times based on when the first action occurred. In the case of the distribution for bike actions following a 6-12h bike ride, we observe peaks at around 9 and 24 hours (interarrival times), and peaks at around 14 and 24 hours for bike actions following a 12-18h bike ride. This behavior is not unexpected. When biking in the morning (6-12h), the next bike ride will likely be a commute back around 9h later. However, if the bike ride happens in the evening (12-18h), the next bike ride is likely not during the middle of the night, but after 14 hours or at around 8:00h in the morning. In addition, both curves exhibit a daily, 24h, periodicity. Modeling these periodicities allows us to capture user-specific timing of, for example, a late evening commute signaling a later start the next morning. In conclusion, two important dynamics could help predicting future real-world actions: actions display periodic recurrence and the time of recurrence can depend on the time of day.

3.5 Proposed Model

In this section, we operationalize the insights gained from empirical observations (Section 3.4) in a probabilistic model based on temporal point processes, called TIPAS.

3.5.1 Background on Temporal Point Processes

A temporal point process is a random process whose realization consists of a list of discrete events localized in time, $\{t_n\}_{n\in\mathbb{N}}$ with $t_n \in \mathbb{R}^+$. We introduce univariate temporal point processes for ease of exposition, though we will be using multivariate point processes to model the joint occurrence dynamics of multiple different actions (description inspired by [Farajtabar et al., 2015]; more background in [Aalen et al., 2008]). Let H_t be the history of events before time t. Temporal point processes can be characterized via the conditional intensity function representing a stochastic model for the time of the next event given all the times of previous events. Formally, the conditional intensity function $\lambda(t)$ is the conditional probability of observing an event in a small window $[t, t+dt)$ given the history H_t; that is, $\lambda(t)dt = \mathbb{P}\{\text{event in}[t, t+dt)|H_t\}$. The conditional probability that no event happens during $[t, t')$ is $S(t') = \exp(-\int_t^{t'} \lambda(\tau)d\tau)$ and the conditional density that an event occurs at time t' is $f(t') = \lambda(t')S(t')$ [Aalen et al., 2008]. Thus, the log-likelihood of a list of events $t_1, t_2, \ldots, t_n$ in an observation window $[0, T)$, where $T > t_n$, can be expressed as

$$\mathcal{L}(t_1, t_2, \ldots, t_n) = \sum_{i=1}^{n} \log \lambda(t_i) - \int_0^T \lambda(\tau)d\tau\,. \tag{3.1}$$

The intensity λ can take various functional forms leading to a homogeneous Poisson process if $\lambda(t)$ is constant, to an inhomogeneous Poisson process if $\lambda(t)$ is time-varying but independent of the event history H_t, or to a Hawkes process if the intensity models mutual self-excitations between events [Aalen et al., 2008]. Our TIPAS model is based on multivariate Hawkes processes [Hawkes, 1971].

3.5.2 Model Definition

We model user actions as a multivariate temporal point process with a time-varying intensity based on three factors based on our empirical observations (Section 3.4). The following intensity function models the rate that action a occurs at time t in user history u,

$$\lambda_u(t, a) = \alpha_{ua} + Time_u(t, a) + ShortTerm_u(t, a) + LongTerm_u(t, a). \tag{3.2}$$

Here, we use an additive decomposition of the intensity instead of modeling more complex interaction effects, because this approach is simple yet powerful and has been proven empirically successful as well [Farajtabar et al., 2015; Iwata et al., 2013; Tanaka et al., 2016]. This model is conceptually visualized in Figure 3.4. The figure shows how the overall intensity function $\lambda_u(t, a)$ (blue; here, $a =$ food) is the sum of

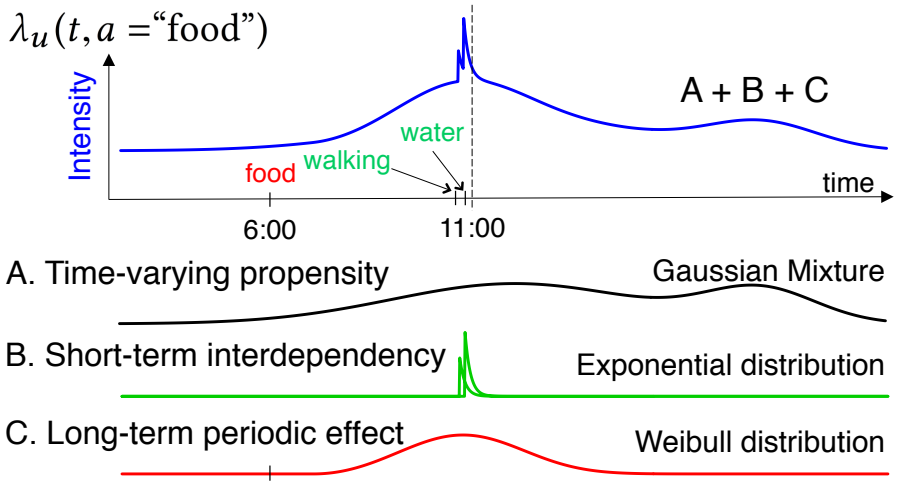

Figure 3.4 – Conceptual model overview. Intensity function of "food" for user u is modeled by the sum of three types of influences; time-varying background intensity (A; black), short-term dependencies (B; green), and long-term periodic effects (C; red). (A) Time-varying background intensity models typical times for food (*e.g.*, having lunch around 12:00h). (B) Events of "walking" and "water" might trigger "food" action within a short period of time. (C) Due to the early breakfast (6:00h), we might expect an earlier lunch.

the time-varying propensity $Time_u(t, a)$ (black), short-term dependencies between actions $ShortTerm_u(t, a)$ (green), and long-term periodic effects $LongTerm_u(t, a)$ (red) between actions (for simplicity, we assume no personalization, *i.e.* $\alpha_{ua} = 0$). Note that our model does account for randomness, in the sense that not all actions may strictly conform to short-term and long-term patterns, through the personalized and time-varying baserates. In fact, learning model parameters from real data tries to account for all actions and will adapt distributional parameters to best explain all occurring actions. Next, we formally define each of the four factors in turn.

Personalized action preferences: α_{ua}. We include personalized user preferences for specific actions through a constant additive factor $\alpha_{ua} \geq 0$ for each action and user. Note that one could also model user preferences to be time-varying instead. However, this would lead to a very large number of parameters and we show in Section 3.6 that this simple model works well in practice.

Time-varying propensities of actions: $Time_u(t, a)$. Events can occur without influence from preceding events according to the background intensity function $Time_u(t, a)$. Having observed that the propensity of actions varies across time of day (Section 3.4.2), we model the background intensity of action a as a function of

time-of-day through a Gaussian mixture model. We define:

$$Time_u(t,a) = \sum_{z \in \boldsymbol{Z}} \frac{\beta_{az}}{\sqrt{2\pi\sigma_{az}^2}} \exp\Big(-\frac{\big(l_t - \mu_{az}\big)^2}{2\sigma_{az}^2}\Big) \ , \tag{3.3}$$

where $z \in \boldsymbol{Z}$ represents the latent class of Gaussian mixture model (the number of mixtures can be determined through cross-validation). For each action a and latent mixture class z, $\mu_{az} > 0$ and $\sigma_{az} > 0$ denote the mean and standard deviation of the Gaussian distribution. The importance of that mixture on the overall intensity function $Time_u(t,a)$ is captured by $\beta_{az} \geq 0$. l_t corresponds to the time of day of timestamp t (*i.e.*, elapsed time since midnight). We show in Section 3.6.3 that Gaussian mixtures fit temporal variation in real-world data well.

Short-term dependencies between actions: $ShortTerm_u(t,a)$**.** To model short-term dependencies between actions, we consider how the rate at which action a occurs at time t (Equation 3.2) is influenced by actions a' which occurred at previous time $t' < t$. We model these influences as a Hawkes process exhibiting self-excitations using Exponential decay functions starting at the time of previous actions. As demonstrated in Section 3.4.2, the short-term influence of previous actions diminishes quickly and monotonically, making the Exponential distribution a natural choice for the decay function. We define:

$$ShortTerm_u(t,a) = \sum_{(t',a') \in H_{ut}} \theta_{a'a}\omega_{a'a} \exp(-\omega_{a'a}\Delta_{t't}) \ , \tag{3.4}$$

where $H_{ut} = \{(t',a') | (t',a') \in H_u \text{ and } t' < t\}$ is the set of events before time t in history u, and $\Delta_{t't} = t - t'$ is the time difference between time t' and time $t > t'$. Further, $\omega_{a'a} \geq 0$ determines how quickly action a' triggers action a (shape of Exponential distribution), and $\theta_{a'a} \geq 0$ determines how likely action a' triggers action a (scaling of distribution). We estimate these parameters for each pair of actions (a',a). Therefore, this component of the model captures the interdependencies between different actions (*e.g.*, drinking after running), as well as the self-exciting effects of actions (*e.g.*, running after running). We show in Section 3.6.3 that a Hawkes process with Exponential decay function fits short-term action dependencies in real-world data well.

Long-term periodic effects: $LongTerm_u(t,a)$**.** We model the long-term periodic effects between identical actions (*e.g.*, run to run) using Weibull distributions. The Weibull distribution is a continuous distribution with positive support (*i.e.*, for $\Delta_{t't} > 0$) that is well suited to model long-term effect patterns at different points in time and with different variance around its mean. We model the rate at which action a occurs at time t influenced by a previous event of action a at time t' as

follows:

$$LongTerm_u(t,a) = \sum_{(t',a') \in H^a_{ut}} \phi_{c_{t'}a} \gamma_{c_{t'}a} \kappa_{c_{t'}a} \Delta_{t't}^{\kappa_{c_{t'}a}-1} \exp(-\gamma_{c_{t'}a} \Delta_{t't}^{\kappa_{c_{t'}a}}) \tag{3.5}$$

where $H^a_{ut} = \{(t',a') | (t',a') \in H_u \text{ and } t' < t \text{ and } a' = a\}$ is the set of events of action a before time t in history u, and $\Delta_{t't} = t - t'$ is again the time difference between time t' and time $t > t'$. As shown in Section 3.4.2, long-term effects vary based on the time of day of action a'. This is captured through the parameter $c_{t'} \in C$ that represents discretized time-of-day windows (*e.g.*, using four classes as 0-6h, 6-12h, 12-18h, and 18-24h). This allows us to learn time-of-day-dependent distributions modeling different periodicities. Parametrized by this time-of-day category $c_{t'}$ and by action a, $\gamma_{c_{t'}a} \geq 0$, $\phi_{c_{t'}a} \geq 0$ determine how quickly and how likely (influence) action a' (which occurred in time-of-day window $c_{t'}$) triggered its following event of action a. $\kappa_{c_{t'}a} \geq 0$ determines the shape of the Weibull distribution. In Section 3.6.3, we demonstrate that the Weibull distribution closely match periodic dynamics in real-world data.

3.5.3 Model Inference

We use maximum likelihood estimation to infer the parameters of our proposed model (Equation 3.2). The unknown parameters are $\boldsymbol{\alpha} = \{\{\alpha_{ua}\}_{u \in U}\}_{a \in A}$, $\boldsymbol{\beta} = \{\{\beta_{az}\}_{a \in A}\}_{z \in Z}$, $\boldsymbol{\mu} = \{\{\mu_{az}\}_{a \in A}\}_{z \in Z}$, $\boldsymbol{\sigma} = \{\{\sigma_{az}\}_{a \in A}\}_{z \in Z}$, $\boldsymbol{\Theta} = \{\{\theta_{a'a}\}_{a \in A}\}_{a' \in A}$, $\boldsymbol{\Omega} = \{\{\omega_{a'a}\}_{a \in A}\}_{a' \in A}$, $\boldsymbol{\Phi} = \{\{\phi_{ca}\}_{c \in C}\}_{a \in A}$, $\boldsymbol{\Gamma} = \{\{\gamma_{ca}\}_{c \in C}\}_{a \in A}$, and $\boldsymbol{K} = \{\{\kappa_{ca}\}_{c \in C}\}_{a \in A}$. The set of all parameters is denoted by $\boldsymbol{\Psi} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\Theta}, \boldsymbol{\Omega}, \boldsymbol{\Phi}, \boldsymbol{\Gamma}, \boldsymbol{K}\}$.

The log-likelihood function (Equation 3.1), given a set of user histories $\mathcal{H} = \{H_u\}_{u \in U}$, can be expressed as:

$$\mathcal{L}(\boldsymbol{\Psi}|\mathcal{H}) = \sum_{u \in U} \sum_{n=1}^{N_u} \log \lambda_u(t_{un}, a_{un}) - \sum_{u \in U} \int_0^T \sum_{a \in A} \lambda_u(t,a) dt \ , \tag{3.6}$$

where the last term, the expectation function, represents the expected number of events in the time period from 0 to T. Combining Equations (3.2)-(3.6), the log-likelihood can be written as follows:

$$
\begin{aligned}
&\mathcal{L}(\boldsymbol{\Psi}|\mathcal{H}) = \\
&\sum_{u\in U}\sum_{n=1}^{N_u}\log\Bigg\{\alpha_{ua_{un}} + \sum_{z\in \boldsymbol{Z}}\frac{\beta_{a_{un}z}}{\sqrt{2\pi\sigma^2_{a_{un}z}}}\exp\Big(-\frac{\big(l_{t_{un}}-\mu_{a_{un}z}\big)^2}{2\sigma^2_{a_{un}z}}\Big) \\
&+\sum_{m=1}^{n-1}\theta_{a_{um}a_{un}}\omega_{a_{um}a_{un}}\exp(-\omega_{a_{um}a_{un}}\Delta_{t_{um}t_{un}}) \\
&+\sum_{l=1}^{n-1}\Big(I(a_{ul}=a_{un})\phi_{c_{ul}a_{un}}\gamma_{c_{ul}a_{un}}\kappa_{c_{ul}a_{un}}\Delta_{t_{ul}t_{un}}^{\kappa_{c_{ul}a_{un}}-1} \\
&\times\exp(-\gamma_{c_{ul}a_{un}}\Delta_{t_{ul}t_{un}}^{\kappa_{c_{ul}a_{un}}})\Big)\Bigg\} - \sum_{u\in U}\int_0^T\sum_{a\in A}\lambda_u(t,a)dt \quad ,
\end{aligned}
\tag{3.7}
$$

where $c_{ul} \in C$ represents time-of-day category of l-th event of u, and $I(\cdot)$ is the indicator function. The integral in Equation 3.7 can be analytically calculated.

Inspired by previous work [Farajtabar et al., 2015; Zhou et al., 2013], we develop an efficient inference algorithm to maximize the log-likelihood based on the EM algorithm. By iterating the E-step and the M-step until convergence, we obtain a local optimum solution for $\boldsymbol{\Psi}$.

E-step. Conceptually, we introduce latent variables $\boldsymbol{p}, \boldsymbol{q}, \boldsymbol{r}$ to capture why each event was triggered either through user preference, time-varying background intensity, short-term action interdependencies, or long-term periodic effects. Let $p_{0,un}$ be the probability that the n-th event of user u was triggered by user preference, $p_{z,un}$ be the probability that the n-th event of user u was triggered by the time-varying background intensity function of latent class z, $q_{um,un}$ be the probability that the n-th event of user u was triggered by the short-term effect of the m-th event of user u, and $r_{ul,un}$ be the probability that the n-th event of user u was triggered by the long-term effect of the l-th event of user u.

In E-step, k-th estimate of $p^k_{0,un}$, $p^k_{z,un}$, $q^k_{um,un}$, and $r^k_{ul,un}$ are calculated by:

$$
p^k_{0,un} = \frac{\alpha^k_{ua_{un}}}{R_{un}} \quad , \tag{3.8}
$$

$$
p^k_{z,un} = \frac{1}{R_{un}}\frac{\beta^k_{a_{un}z}}{\sqrt{2\pi(\sigma^k_{a_{un}z})^2}}\exp\Big(-\frac{\big(l_{t_{un}}-\mu^k_{a_{un}z}\big)^2}{2(\sigma^k_{a_{un}z})^2}\Big) \quad , \tag{3.9}
$$

$$
q^k_{um,un} = \frac{1}{R_{un}}\theta^k_{a_{um}a_{un}}\omega^k_{a_{um}a_{un}}\exp(-\omega^k_{a_{um}a_{un}}\Delta_{t_{um}t_{un}}) \quad , \tag{3.10}
$$

$$r^k_{ul,un} = \left\{ \frac{1}{R_{un}} \phi^k_{c_{ul}a_{un}} \gamma^k_{c_{ul}a_{un}} \kappa^k_{c_{ul}a_{un}} \times \Delta^{\kappa^k_{c_{ul}a_{un}}-1}_{t_{ul}t_{un}} \exp(-\gamma^k_{c_{ul}a_{un}} \Delta^{\kappa^k_{c_{ul}a_{un}}}_{t_{ul}t_{un}}) \right\} , \tag{3.11}$$

where $\mathbf{\Psi}^k = \{\boldsymbol{\alpha}^k, \boldsymbol{\beta}^k, \boldsymbol{\mu}^k, \boldsymbol{\sigma}^k, \boldsymbol{\Theta}^k, \boldsymbol{\Omega}^k, \boldsymbol{\Phi}^k, \boldsymbol{\Gamma}^k, \boldsymbol{K}^k\}$ is the k-th estimate of parameters in the EM procedure, and R_{un} is the normalization factor in order to satisfy $p^k_{0,un} + \sum_{z \in \boldsymbol{Z}} p^k_{z,un} + \sum_{m=1}^{n-1} q^k_{um,un} + \sum_{l=1}^{n-1} r^k_{ul,un} = 1$.

M-step. We use Jensen's inequality to provide a lower bound for the log-likelihood (Equation 3.7); this lower bound is often called the Q function. We obtain the next estimate of the parameters by taking the derivative of the Q function with respect to each parameter and setting them to zero:

$$\alpha^{k+1}_{ua} = \frac{\sum_{n=1}^{N_u} I(a_{un} = a) p^k_{0,un}}{T} , \tag{3.12}$$

$$\beta^{k+1}_{az} = \frac{2\mathcal{T}}{|U|T} \times \frac{\sum_{u \in U} \sum_{n=1}^{N_u} I(a_{un} = a) p^k_{z,un}}{erf(\frac{\mu^k_{az}}{\sqrt{2}\sigma^k_{az}}) + erf(\frac{\mathcal{T} - \mu^k_{az}}{\sqrt{2}\sigma^k_{az}})} , \tag{3.13}$$

$$\theta^{k+1}_{a'a} = \frac{\sum_{u \in U} \sum_{n=1}^{N_u} \sum_{m=1}^{n-1} I(a_{um} = a', a_{un} = a) q^k_{um,un}}{\sum_{u \in U} \sum_{n=1}^{N_u} I(a_{un} = a') \Big(1 - \exp\big(-\omega^k_{a'a}(T - t_{un})\big)\Big)} , \tag{3.14}$$

$$\phi^{k+1}_{ca} = \frac{\sum_{u \in U} \sum_{n=1}^{N_u} \sum_{l=1}^{n-1} I(a_{ul} = a, a_{un} = a, c_{ul} = c) r^k_{ul,un}}{\sum_{u \in U} \sum_{n=1}^{N_u} I(a_{un} = a, c_{un} = c) \Big(1 - \exp\big(-\gamma^k_{ca}(T - t_{un})^{\kappa^k_{ca}}\big)\Big)} , \tag{3.15}$$

where $\mathcal{T}$ is the time period of a day (*i.e.*, 24 hours), $\frac{T}{\mathcal{T}}$ is the number of days representation of the observed period T, and where *erf* denotes the Gauss error function $erf(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^{x} e^{-t^2} \, \mathrm{d}t$. Because of the exponentials (exp and *erf*) within the expectation function (Equation 3.7), $\omega^{k+1}_{a'a}$, γ^{k+1}_{ca}, κ^{k+1}_{ca}, μ^{k+1}_{az}, and σ^{k+1}_{az} cannot be solved in closed form. However, by further considering a lower bound for these exponentials $\omega^{k+1}_{a'a}$ and γ^{k+1}_{ca} can be solved in closed form. Their update rules are as follows:

$$\omega_{a'a}^{k+1} = \left\{ \sum_{u \in U} \sum_{n=1}^{N_u} \sum_{m=1}^{n-1} I(a_{um} = a', a_{un} = a) q_{um,un}^{k} \right\}$$
$$/ \left\{ \sum_{u \in U} \sum_{n=1}^{N_u} \sum_{m=1}^{n-1} I(a_{um} = a', a_{un} = a) q_{um,un}^{k} \Delta_{t_{um} t_{un}} \right.$$
$$\left. + \sum_{u \in U} \sum_{n=1}^{N_u} I(a_{un} = a') \theta_{a'a}^{k} (T - t_{un}) \exp\left(-\omega_{a'a}^{k} (T - t_{un})\right) \right\} , \tag{3.16}$$

$$\gamma_{ca}^{k+1} = \left\{ \sum_{u \in U} \sum_{n=1}^{N_u} \sum_{l=1}^{n-1} I(a_{ul} = a, a_{un} = a, c_{ul} = c) r_{ul,un}^{k} \right\}$$
$$/ \left\{ \sum_{u \in U} \sum_{n=1}^{N_u} \sum_{l=1}^{n-1} I(a_{ul} = a, a_{un} = a, c_{ul} = c) r_{ul,un}^{k} \Delta_{t_{ul} t_{un}}^{\kappa_{ca}^{k}} \right.$$
$$\left. + \sum_{u \in U} \sum_{n=1}^{N_u} I(a_{un} = a, c_{un} = c) \phi_{ca}^{k} (T - t_{un})^{\kappa_{ca}^{k}} \exp\left(-\gamma_{ca}^{k} (T - t_{un})^{\kappa_{ca}^{k}}\right) \right\} . \tag{3.17}$$

The other three parameters, κ_{ca}^{k+1}, μ_{az}^{k+1} and σ_{az}^{k+1}, are estimated by maximizing the Q function through the use of a gradient-based numerical optimization method; we used the Newton method. For more details on model inference see the Online Appendix [Kurashima et al., 2018].

3.6 Experiments

This section evaluates the predictive performance of our proposed model on two real-world datasets on predicting the next user action and when it will occur. We compare against eleven different baselines on each dataset. However, since many baseline models are unable to make joint predictions of action and timing, we evaluate these two tasks separately. Importantly, this process allows us to identify the individual sources of error that would impact joint predictions. Our implementation is available at `snap.stanford.edu/tipas`.

3.6.1 Datasets

Our experiments use two real-world activity logging datasets. In total, these datasets comprise 12 millions real-world actions taken by 20 thousand users over 17 months.

Argus dataset. We use the activity logging data from the Argus mobile app described in Section 3.4.1. Users in this dataset can log 10 different actions (drink, sleep, heart rate, running, weight, food, walking, biking, workout, and stretching) and our goal is to predict which of these 10 actions a user will take next (and

Dataset Statistics	Argus	Under Armour
Observation period	7 months Jan-July '15	10 months Jan-Oct '16
# unique actions	10	8
# total users	4,708	15,221
# total actions	2,140,757	9,733,645
Avg. # actions per user	454.7	639.5
Avg. # unique actions per user	6.3	6.8
Avg. # unique actions per user day	2.7	4.4

Table 3.1 – Basic dataset statistics.

when). Our analyses include users who logged at least two unique actions per day on average (other users might only use the app to for example track their sleep making predictions of actions and their timing almost trivial; we find that our results are robust to different choices of this threshold). We consider 7 months of data from the app in a rolling window evaluation, where we use one month for training and the next for testing (*i.e.*, making out-of-sample predictions; without retraining). As shown in Table 3.1, the dataset includes 2.1 million actions by over 4 thousand users within the 7 month observation period.

Under Armour dataset (UA). We also use activity logging data from Under Armour mobile apps (*i.e.*, MapMyFitness and MyFitnessPal; focusing on users that are active in both apps). Users in this dataset can log 8 different types of actions (running, walking, biking, workout, breakfast, lunch, dinner, and snacks). Our analyses include users who logged at least four unique actions per day on average, leading to a similar number of unique actions per user on average compared to the Argus dataset (again, our results are robust to different choices of this threshold). We consider 10 months of data from the app and again perform a rolling window evaluation where we train on one month and test on the next. In total, this dataset comprises 15 thousand users taking 9.8 million actions (Table 3.1).

3.6.2 Model Learning

Note that our model has few core model parameters. In the context of the datasets described above, we have about 500 core model parameters ($\beta, \mu, \sigma, \Theta, \Omega, \Phi, \Gamma, K$) and about 25 thousand personalization parameters (α). This small, non-redundant set of parameters allows us to train the model efficiently and robustly, and explain model predictions through inspection and visualization of model parameters (Section 3.6.6), while performing competitively (Section 3.6.4). However, during training time (but not test time) we also have latent variables (p, q, r) that allow us

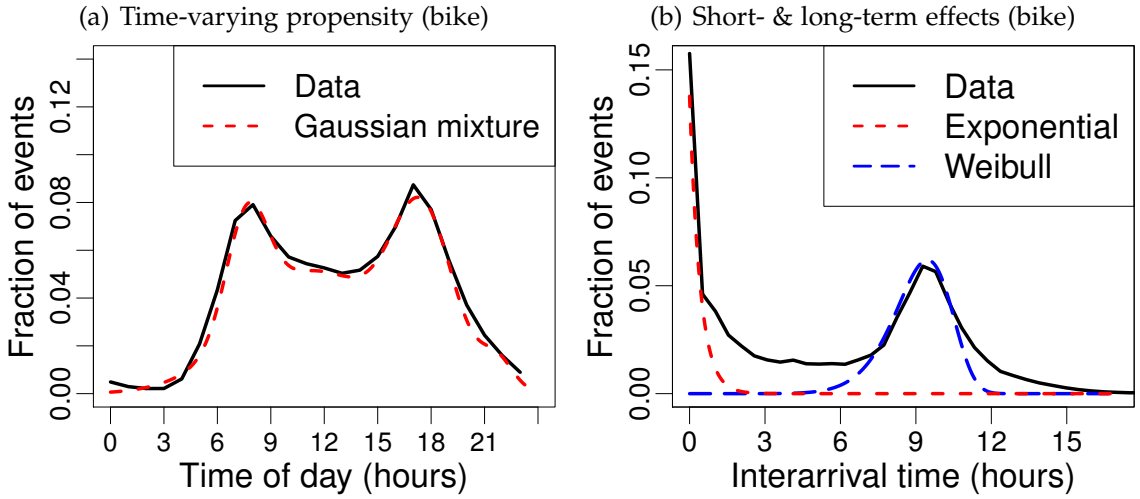

Figure 3.5 – Validation of parametric modeling assumptions (Section 3.5.2). (a) Mixture of Gaussian closely fits observed time-varying action propensity (here, for bike action). (b) Exponential and Weibull distributions collectively well-approximate short-term dependencies and long-term periodic effects of previous bike actions.

to learn the core model parameters. These latent variables represent which actions trigger which other actions, leading to $\mathcal{O}(|U|(max_{u \in U} N_u)^2)$ variables in the worst case. On both datasets, inference of both core model and latent parameters involves solving an optimization problem with over 200 million total variables (Section 3.5.3; we randomly initialize all parameters). Using our EM-based inference procedure we can robustly infer these parameters in less than ten hours using a single-threaded C++ implementation on a single machine. We find that one month of training data is enough to reliably train our model.

3.6.3 Validating Parametric Assumptions

In Section 3.5.2 we developed a model consisting of three parts: time variation modeled using a mixture of Gaussians ($\text{Time}_u(t, a)$), short-term dependencies between actions modeled by a Hawkes process with Exponential decay function ($\text{ShortTerm}_u(t, a)$), and long-term periodicity modeled through Weibull distributions ($\text{LongTerm}_u(t, a)$). Here, we test empirically whether these parametric assumptions hold true in real data. Using the Argus dataset, we inferred appropriate parameters for these distributions.

We demonstrate qualitatively in Figure 3.5 that the chosen distributions fit real-world dynamics well. Figure 3.5 (a) shows the time-varying propensity with superimposed mixture of Gaussian fit and (b) shows that, collectively, Exponential and Weibull distribution closely approximate the influence of previous actions

(example data is for bike action as seen in Figure 3.1).

We have further quantitatively evaluated our parametric assumptions and compared our choices to alternative distributions (*e.g.*, Rayleigh and Power-law) through goodness-of-fit tests which have shown that the suggested distributions best fit real-world data.

3.6.4 Predicting the Next Action

First, we evaluate our proposed model in terms of its accuracy in predicting actions at a given time. The task is to predict the $n+1$-st action a_{un+1} of user u, given time t_{un+1} and past user history $\boldsymbol{H}_u = \{(a_{u1}, t_{u1}), \cdots, (a_{un}, t_{un})\}$. For each two month period in both datasets, we use the first month for training and the second month for testing and perform a rolling window evaluation, where we predict each test set event given all events that happened before it (without retraining). We use accuracy, the percentage of correct predictions, over all test events as our evaluation measure (the most common measure to evaluate recommender systems [Herlocker et al., 2004]). We also report macro-averaged recall [Manning et al., 2008] corresponding to averaging prediction accuracy equally weighted across all action types. This measure highlights differences in predictive performance on rare actions that do not affect the standard accuracy measure very much. We find very similar results using other classification metrics (*e.g.*, ROC AUC, F1). The number of mixtures for the time-varying action propensity (Equation 3.3) is set via cross-validation. We compare our proposed model against the following seven baseline models, which have proven competitive across a wide variety of prediction tasks and recommender systems:

- **Copy Model:** Simply repeats the user's last action. Several repeat consumption models are variants of this copy model (*e.g.*, [Anderson et al., 2014; Benson et al., 2016]).
- **Markov Model:** Predicts the next action based on the most recent actions of the user. We report first to fifth-order Markov models (sixth-order models did not significantly improve performance). Markov models have been used widely to predict next actions (*e.g.*, [Ashbrook and Starner, 2003; Kapoor et al., 2015]).
- **Hidden Markov Model (HMM):** This is a Markov model with hidden (unobserved) states. It predicts the next action based on the current, inferred state of the action sequence [Lane, 1999].
- **Factorizing Personalized Markov Chains (FPMC):** This is based on underlying Markov chains where the transitions matrices are user-specific. Matrix factorization models are used to address sparsity issues of these user-specific Markov chains [Rendle et al., 2010].

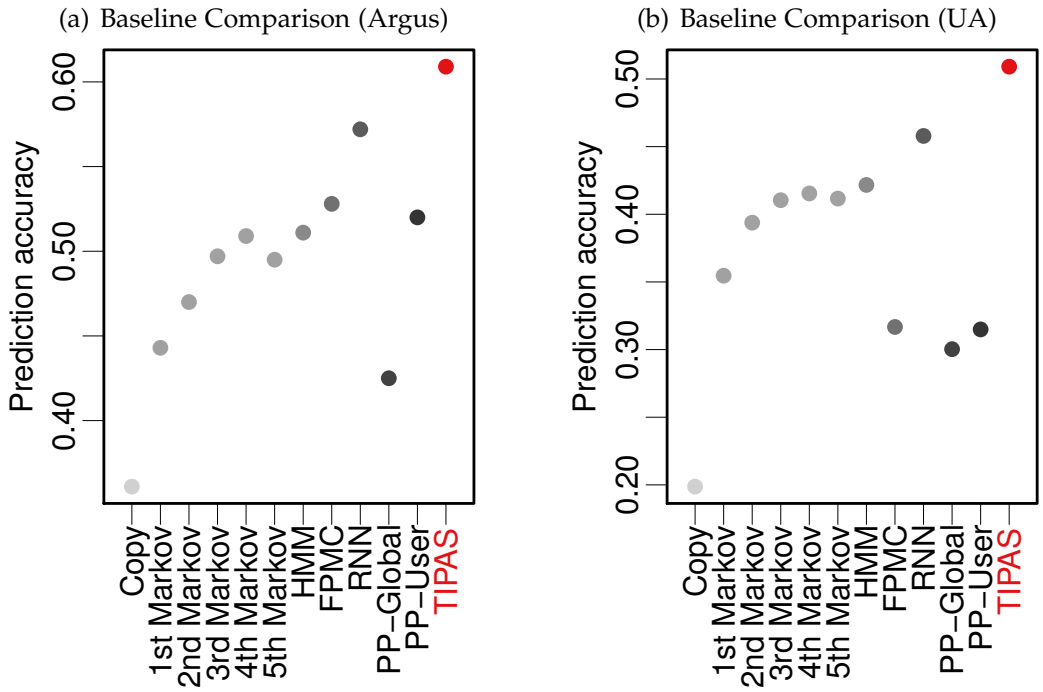

Figure 3.6 – Accuracy when predicting actions. Higher is better. Comparing proposed TIPAS model (red) to baselines (gray). Error bars in all plots correspond to standard errors.

- **Recurrent Neural Network (RNN):** Feedforward neural network structure using outputs from the hidden units at the prior time step as the inputs as the current time step. Assumes discrete time steps and no ready-to-use generalizations to continuous time domain exist.
- **PP-Global:** A global Poisson process model. The intensity function is constant over time and defined by $\lambda_u(t, a) = \alpha_a$.
- **PP-User:** A user-specific Poisson process model. The intensity function is constant over time and defined by $\lambda_u(t, a) = \alpha_{ua}$.

Note that Hawkes process models (*e.g.*, [Du et al., 2015a; Farajtabar et al., 2015; Hawkes, 1971]) are closely related to the $\text{ShortTerm}_u(t, a)$ component of our model (Equation 3.4). Our proposed model **TIPAS** uses the intensity function of Eq. 3.2. We predict the most likely user action as $\hat{a} = \text{argmax}_a \lambda_u(t_{un+1}, a)$. We also compare the individual model components in an ablation study below.

Results: Comparison to baseline models. Figure 3.6 compares accuracy of next action prediction. We observe that the eleven baselines achieve accuracies of 36-57% on the Argus dataset and 20-46% on the Under Armour dataset with the RNN baseline performing best in both datasets. The limited predictive performance of

these competitive baselines shows that this prediction task is non-trivial. TIPAS outperforms all baselines on both Argus (60.9%; 6-69% rel. improvement) and Under Armour datasets (50.9%; 11-156% rel. improvement). The small standard error across multiple dataset splits in the rolling window evaluation (Figure 3.6) demonstrates that our training procedure is robust and consistently shows good performance. We note that TIPAS performs particularly well on rare actions leading to 9-256% relative improvement in macro-averaged recall over baseline models (Online Appendix [Kurashima et al., 2018]).

Results: Comparison of individual model components. Note that TIPAS has three components (Equation 3.2): time-varying action propensities (Time), short-term interdependencies between actions (Short), and long-term periodic effects (Long). Here, we evaluate the performance of each of these components in an ablation study by comparing Time, Time+Short, and the full TIPAS model combining Time+Short+Long (Figure 3.7; all models include user personalized preferences α_{ua}). We find that modeling time-varying action propensities achieves an accuracy of 53% and 40% on the two datasets, respectively. Further, modeling short-term dependencies between actions improves this to 59% and 49%, and capturing long-term periodicities of actions further improves this to 61% and 51%, respectively. This demonstrates that capturing all three properties is essential to predicting actions in both datasets of human real-world action sequences. Further, we observe a bigger difference between the full Time+Short+Long model and the Time+Short model in terms of macro-averaged recall (7% and 5% relative MAR improvements compared to 3% and 4% in terms of accuracy on the Argus and Under Armour datasets, respectively). This indicates that modeling long-term periodicities is especially important for more rare actions such as walking and biking. In addition, we find that modeling long-term periodic effects discretized by time of day (0-6h, 6-12h, 12-18h, 18-24h) performs significantly better than not discretizing by time of day on both datasets. For example, actions such as biking and walking are periodic but vary based on time of day (Figure 3.3). Our full model captures these time-of-day dependent long-term effects and relatively improves macro-averaged recall of predicting biking and walking actions by 491-556% over Time model and 2-4% over Time+Short model.

3.6.5 Predicting the Time of the Next Action

We now focus on the second aspect of modeling real-world actions: Predicting the time of the next action. Specifically, the task is the predict the $n+1$-th time-stamp t_{un+1} in history u, given past events $H_u = \{(a_{u1}, t_{u1}), ..., (a_{un}, t_{un})\}$ (we do not assume that the next action a_{un+1} is given). Mean absolute error (MAE) is used as the evaluation metric. We use the same train/test paradigm as before

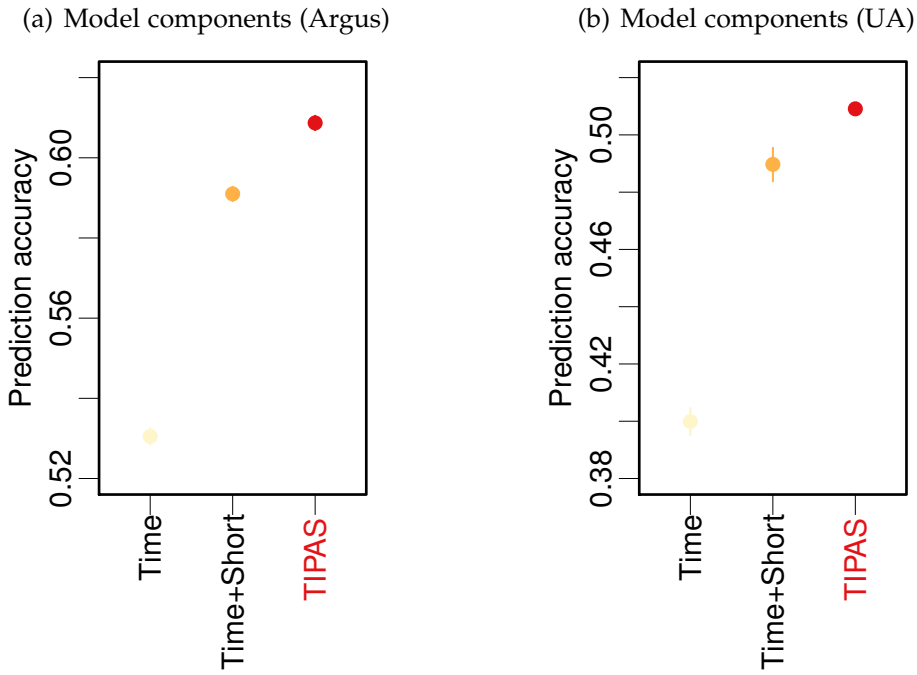

Figure 3.7 – Ablation study comparing different model components on accuracy when predicting actions. Higher is better.

(rolling window evaluation training one month and testing on the next). We restrict predictions to only events that will occur within the next 12 hours (*i.e.*, the time interval $t_{un+1} - t_{un} \leq 12$ hours) because these are the most important and actionable inferences (*e.g.*, predicting a sleep time many days from now may have large error, but it is also less relevant). In order to make time predictions based on TIPAS, we simulate the multivariate temporal point process using Ogata's modified thinning algorithm [Ogata, 1981]. We simulate 100 samples and return the average time.

We compare our model to the following five baseline methods:

- **Time Copy Model:** Predicts the next time, t_{un+1}, based on the most recent time-interval of user u $(t_{un+1} = t_{un} + (t_{un} - t_{un-1}))$.
- **Average Time Interval:** Predicts the next time t_{un+1} using the global average of time-intervals.
- **User Average Time Interval:** Predicts the next time t_{un+1} using the average of time-intervals for user u.
- **PP-Global:** A global Poisson process model. The intensity function is constant over time and defined by $\lambda_u(t, a) = \alpha_a$.
- **PP-User:** A user-specific Poisson process model. The intensity function is constant over time: $\lambda_u(t, a) = \alpha_{ua}$.

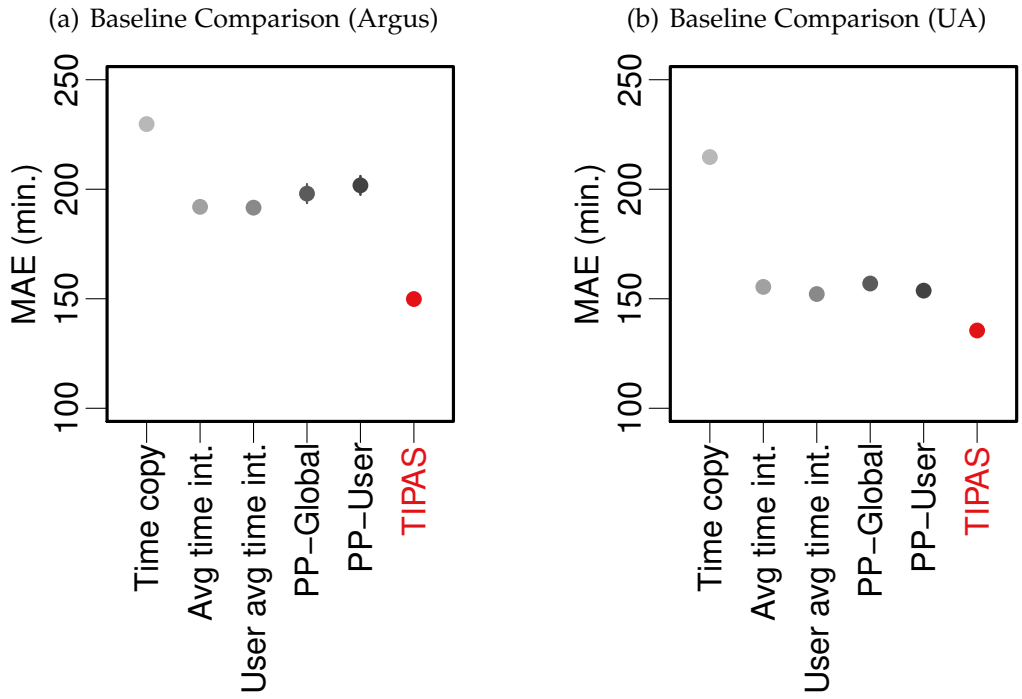

Figure 3.8 – Mean absolute error (MAE) when predicting time of next actions. Lower is better. Comparison to baselines.

We note that the other baselines (Markov models, HMM, FPMC, and RNN) used in Section 3.6.4 are unable to make any time predictions.

Results. Experimental results are shown in Figure 3.8. We observe that all baselines perform similarly except the Time Copy model which performs significantly worse on both datasets. TIPAS significantly outperforms all baselines across both datasets by 22-35% in the Argus dataset and 11-37% in the Under Armour dataset (relative improvement). Restricting predictions to events within the next 6 hours (instead of 12h as before), TIPAS outperforms the baselines even more significantly, improving upon them by 44-58% and 37-41% on the two datasets. TIPAS is able to make better timing predictions because it is able to leverage three key components. First, it is aware that certain actions only happen during certain parts of the day. For example, it will predict longer delays in the middle of the night when actions are unlikely to occur. Second, the model can exploit dependencies between actions. For instance, it might predict a very short time after a run because many users will drink water or check their heart rate soon after. Third, TIPAS is able to exploit periodicities in the data. For example, it might predict an evening time commute because it observed a commute in the morning. In summary, modeling these three

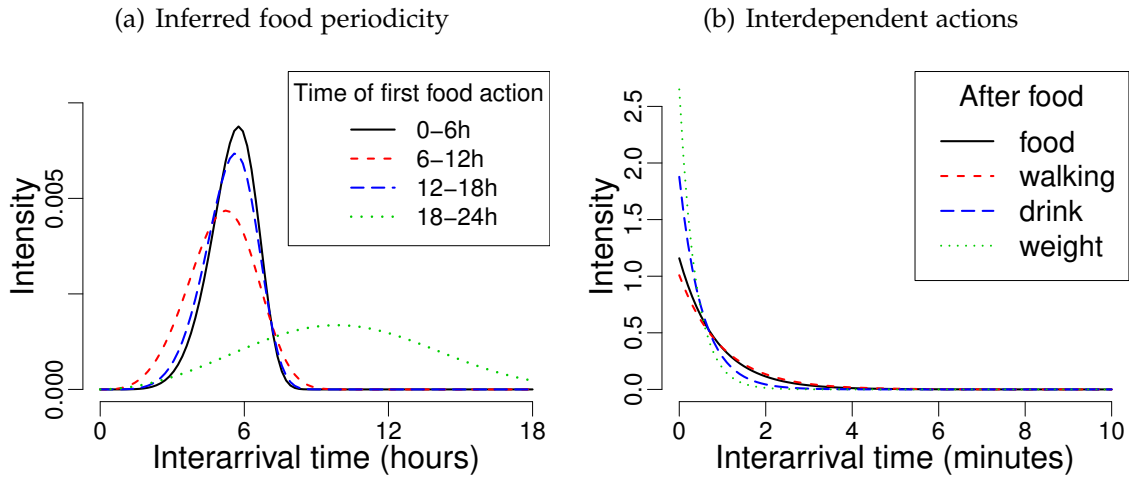

Figure 3.9 – Visualization of inferred TIPAS model parameters for (a) periodicity of food actions and (b) interdependent actions following food actions. The learned dependencies allow to explain why specific actions are being predicted.

key aspects of human behavior allows us to make better predictions of actions and their timing.

3.6.6 Model Explainability

TIPAS also allows for visualization of model parameters, which enables explanations of why certain predictions are made. This is especially important in the mobile health context, where model predictions may impact users' real-world health behaviors and therefore need to be explained and monitored.

The inferred model parameters for Equation 3.5 are shown in Figure 3.9a (specifically, $f(\Delta_{t't}) = \gamma_{c_{t'}a}\kappa_{c_{t'}a}\Delta_{t't}^{\kappa_{c_{t'}a}-1}\exp(-\gamma_{c_{t'}a}\Delta_{t't}^{\kappa_{c_{t'}a}})$ for $a =$ food). These distributions correspond to when food events likely trigger other food events. The distributions show that meals are extremely periodic and that meals sharply determine the timing of the next meal, except for dinners after 18:00h, which do not precisely determine the timing of the next meal (18-24h, green). The periodicities vary between 5h after breakfast (6-12h) and 6h after lunch (12-18h). This is consistent with a typical schedule of meals at 7:00h, 12:00h, and 18:00h. Importantly, this enables us to correctly predict that earlier lunches may lead to earlier dinners. Such predictions are critical for correctly timed interventions, for instance making sure that diet reminders do not come to late.

Furthermore, TIPAS allows us to explain why an activity was predicted, based on the relative contributions of model components to the overall intensity function (Section 3.5.2). For example, after food actions users are likely to log other foods and drinks (Figure 3.9b; showing $f(\Delta_{t't}) = \omega_{a'a}\exp(-\omega_{a'a}\Delta_{t't})$ for $a' =$ food). This

makes sense as typical meals include both food and drinks, and users may choose to log parts of each meal separately. Interestingly, users also often log their weight right after food, indicating that they might be conscious of how their meal might have impacted their weight. Lastly, we observe walking actions right after meals. Users may walk back from a restaurant, or they might attempt to walk off some of their meal's calories.

While these results and examples are specific to mobile activity logging applications, the utility of our model may generalize other domains where behaviors are time-varying, interdependent, or periodic. Distributional choices for the individual may vary across domains but can easily be adapted in our model.

3.7 Conclusion

Accurately predicting the user's future actions is essential for personalization, user modeling, and timely interventions in mobile health applications. In this chapter we demonstrated that real-world user behavior exhibits several complexities including a large number of potential actions, time-varying action propensities, dependencies between actions, and periodic behaviors. We proposed a novel statistical model based on multivariate temporal point processes that jointly models all these complexities of human behaviors. Empirically, we demonstrate that our model successfully captures these dynamics in two real-world datasets and that it significantly outperforms nine baselines on tasks of predicting the next user action and when this action will occur. Our model can serve as a foundation to predict more fine-gained attributes of real-world actions such as their duration, intensity, or exact location. Our results further have implications for modeling human behavior, app personalization, and targeting of health interventions.

Chapter 4

Sleep and Cognitive Performance: Harnessing Web Search Interactions for Population-Scale Physiological Sensing

4.1 Introduction

Maintaining optimal cognitive performance has been found to be important in learning [Kelley et al., 2015], productivity [Colten and Altevogt, 2006], and avoiding industrial and motor vehicle accidents [Colten and Altevogt, 2006; Dinges, 1995]. Studies have demonstrated that cognitive performance varies throughout the day [Van Dongen and Dinges, 2000], 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 [Dinges, 1995]. 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 [Hemp, 2004]. Despite the important influences, temporal variations of real-world performance are not well understood and have never been characterized on a large scale [Roenneberg, 2013].

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) [Van Dongen and Dinges, 2000], *homeostatic sleep pressure* (the longer awake, the more tired you become) [Borbély, 1982], and *sleep inertia* (performance impairment experienced immediately after

waking up) [Åkerstedt and Folkard, 1997; Dinges, 1990].

While models of these biological processes capture well the patterns of cognitive performance in the laboratory [Åkerstedt and Folkard, 1997; Borbély, 1982], 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 [Orne, 1962]. 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" [Roenneberg, 2013].

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 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 [Purcell, 2011].

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 4.3). 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 [Lim and Dinges, 2010]).

We first demonstrate that real-world human cognitive performance captured

through search engine interactions varies throughout the day in a daily rhythm (Section 4.4). 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 4.5). 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 4.5.3). 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 4.6). 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 chapter 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.

4.2 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 [Blatter and Cajochen, 2007]. The daily variations in performance have been found to be modulated

primarily by two processes [Dijk et al., 1992]: a *circadian rhythm* (time-dependent, behavior-independent, near 24-hour oscillations) [Van Dongen and Dinges, 2000] and a *homeostatic sleep drive* (the longer awake, the more tired we become and the more we sleep, the less tired we become) [Borbély, 1982]. 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* [Van Dongen and Dinges, 2000], which corresponds to the performance impairment experienced immediately after waking up [Åkerstedt and Folkard, 1997; Dinges, 1990]. 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 [Dijk and Czeisler, 1995].

Human preferences and natural tendency in the relative timing of sleep and wake are called *chronotypes* and are at least partly based on genetics [Roenneberg et al., 2003]. Cognitive performance depends on chronotype and time of day [Matchock and Mordkoff, 2009]; 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 [Dinges, 1995].

A recent study correlated performance on cognitive exercises with a sleep measure based on retrospective self-reports of "typical sleep" in 160 thousand users [Sternberg et al., 2013]. However, this measure suffers from potential biases [Lauderdale et al., 2008] 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 [Fernandez-Mendoza et al., 2010] 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 [Lim and Dinges, 2010], 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 [Böhmer et al., 2011], to detect stress [Vizer et al., 2009], used as biometric signals for authentication [Monrose and Rubin, 1997], and linked to biological processes [Murnane et al., 2015, 2016] including alertness [Abdullah et al., 2016]. For example, less sleep was linked to shorter duration of focus of

attention in a study with 40 participants [Mark et al., 2016]. Large-scale interaction data have been used to gain insights into human behavior in the areas of mood rhythms [Golder and Macy, 2011], diet [West et al., 2013], conversation strategies (Chapter 5), social networks and mobile games encouraging health behaviors [Althoff et al., 2017b, 2016b; Shameli et al., 2017], and health and disease-related search behaviors [Paparrizos et al., 2016; White et al., 2016].

This Work. Existing research on sleep and performance is either small-scale and laboratory-based [Lim and Dinges, 2010] or relies on subjective measures such as surveys capturing "typical" sleep [Sternberg et al., 2013] 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 [Lauderdale et al., 2008] 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 [Lim and Dinges, 2010]. Our study demonstrates on a large scale that interactions with devices are influenced by biological processes and sleep.

4.3 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 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 4.1. 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 4.4.1 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 4.5.1. We exclude search engine interactions originating from mobile devices since such

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 4.1 – Dataset statistics. BMI refers to body mass index.

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 chapter.

Sleep. Sleep data from wearable devices provides objective measurements which have been preferred to subjective self-reports that may be significantly biased [Lauderdale et al., 2008]. 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 [Ancoli-Israel et al., 2003]. Following standard practice [Walch et al., 2016], we exclude any sleep duration measurements below 4 and above 12 hours of time in bed.

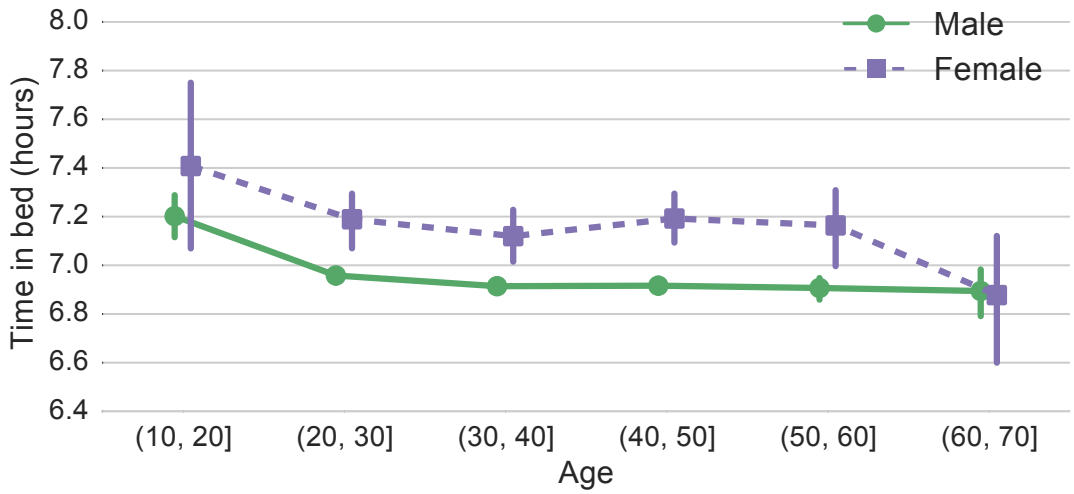

Figure 4.1 – Average sleep duration across age and gender. Our measurements are consistent with previous estimates [Basner et al., 2007; Bureau of Labor Statistics, American Time Use Survey, 2015; Walch et al., 2016] (Section 4.3). Error bars in all figures correspond to 95% confidence intervals of the corresponding mean estimates.

As evidence that our sleep measurements have face validity, we show that they match published sleep estimates. Figure 4.1 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 [Basner et al., 2007; Bureau of Labor Statistics, American Time Use Survey, 2015; Walch et al., 2016]. Walch et al. [Walch et al., 2016] 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. [Basner et al., 2007]. 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.

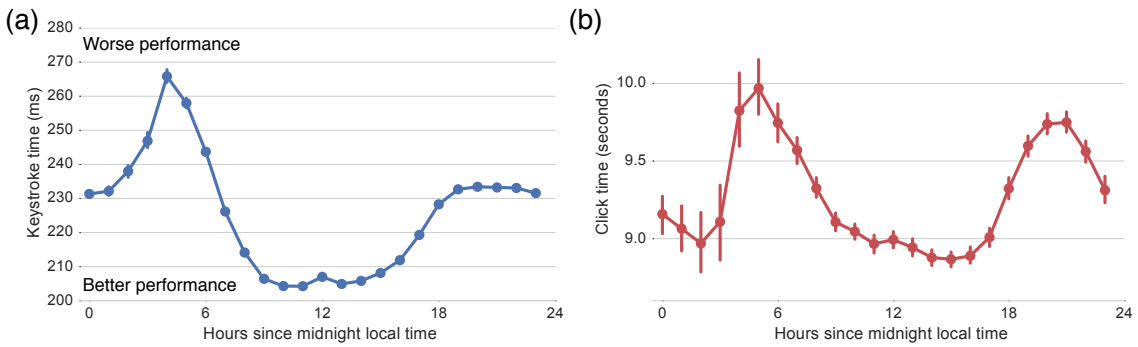

Figure 4.2 – Time of day-dependent variation in keystroke (a) and click timing (b). Higher values indicate worse performance. Both the shape of temporal variation with fastest performance a few hours after wake and slowest performance during habitual sleep times as well as the magnitude of variation are consistent with controlled laboratory-based studies [Ackerman, 2008; Dijk et al., 1992; Dinges, 1995; Wise et al., 2009] (Section 4.4.2).

4.4 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 4.5.

4.4.1 Performance Measures

We study two real-world performance measures in this chapter 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

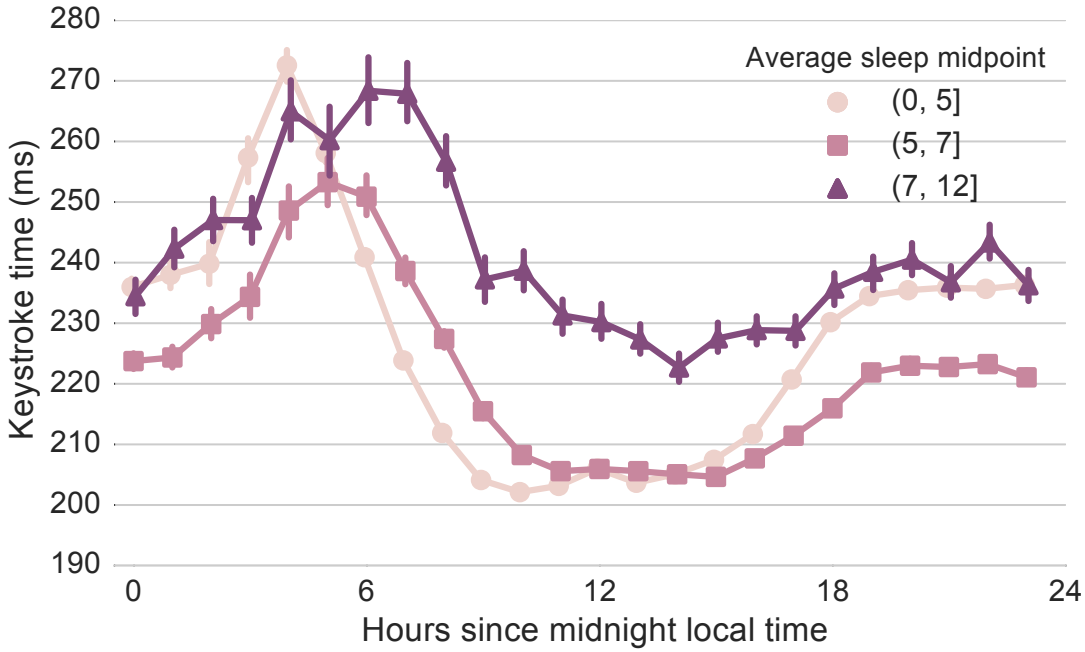

Figure 4.3 – Variation in keystroke time throughout the day varies with chronotype (morning/evening preference) which is defined based on the average point of mid sleep (Section 4.4.3). Users that typically sleep early (light color) perform slowest at about 04:00 h, while medium or late sleepers (darker 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 [Matchock and Mordkoff, 2009].

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 [Card et al., 1980]).

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 4.5.1.

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 motor plans [Pilcher and Huffcutt, 1996]. 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 [Purcell, 2011], 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 [Althoff et al., 2017a]). 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 [Althoff et al., 2017a]). We also verified that performance variation during the weekend is similar to variation during the week (online appendix [Althoff et al., 2017a]) and we therefore do not further differentiate between performance during weekdays and weekends in this chapter. 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 chapter.

4.4.2 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 [Ackerman, 2008; Wise et al., 2009]. Keystroke and click timing also vary throughout the day in a daily rhythm as illustrated in Figure 4.2. Keystroke times (Figure 4.2a) 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 [Card et al., 1980]). Click times (Figure 4.2b) 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 [Dinges, 1995] and the anticipated circadian nadir (*i.e.*, the time of greatest circadian sleep drive) [Dijk et al., 1992]. 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 [Ackerman, 2008; Wise et al., 2009]. 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.

4.4.3 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 [Roenneberg et al., 2003]. Studies have shown that performance depends on the alignment of chronotype and time of day [Matchock and Mordkoff, 2009]; 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 [Juda et al., 2013; Roenneberg et al., 2003]. 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) [Juda et al., 2013]: $MSF_{SC} = MSF - 0.5(SD_F - (5 * SD_W + 2 * SD_F)/7)$, 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 4.3), matching results from previous sleep studies [Matchock and Mordkoff, 2009] 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.

4.5 Modeling Performance

Having demonstrated that performance of search engine interactions vary over time and based on biological processes (Section 4.4), we now operationalize and extend a conceptual model of sleep and performance from chronobiology [Åkerstedt and Folkard, 1997; Borbély, 1982] to explain the variation observed in performance measurements. Classic sleep models are based on circadian rhythms and homeostatic sleep drive [Borbély, 1982]. In addition, we consider sleep inertia and sleep duration [Åkerstedt and Folkard, 1997; Van Dongen and Dinges, 2000]. Background on relevant biological processes is covered in Section 4.2.

4.5.1 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 4.3).

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 [Van Dongen and Dinges, 2000], 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 4.3) 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 4.5.2).

Potential Confounding Factors. We control for several factors in our model to avoid confounding. For keystrokes, we control for the exact character typed or

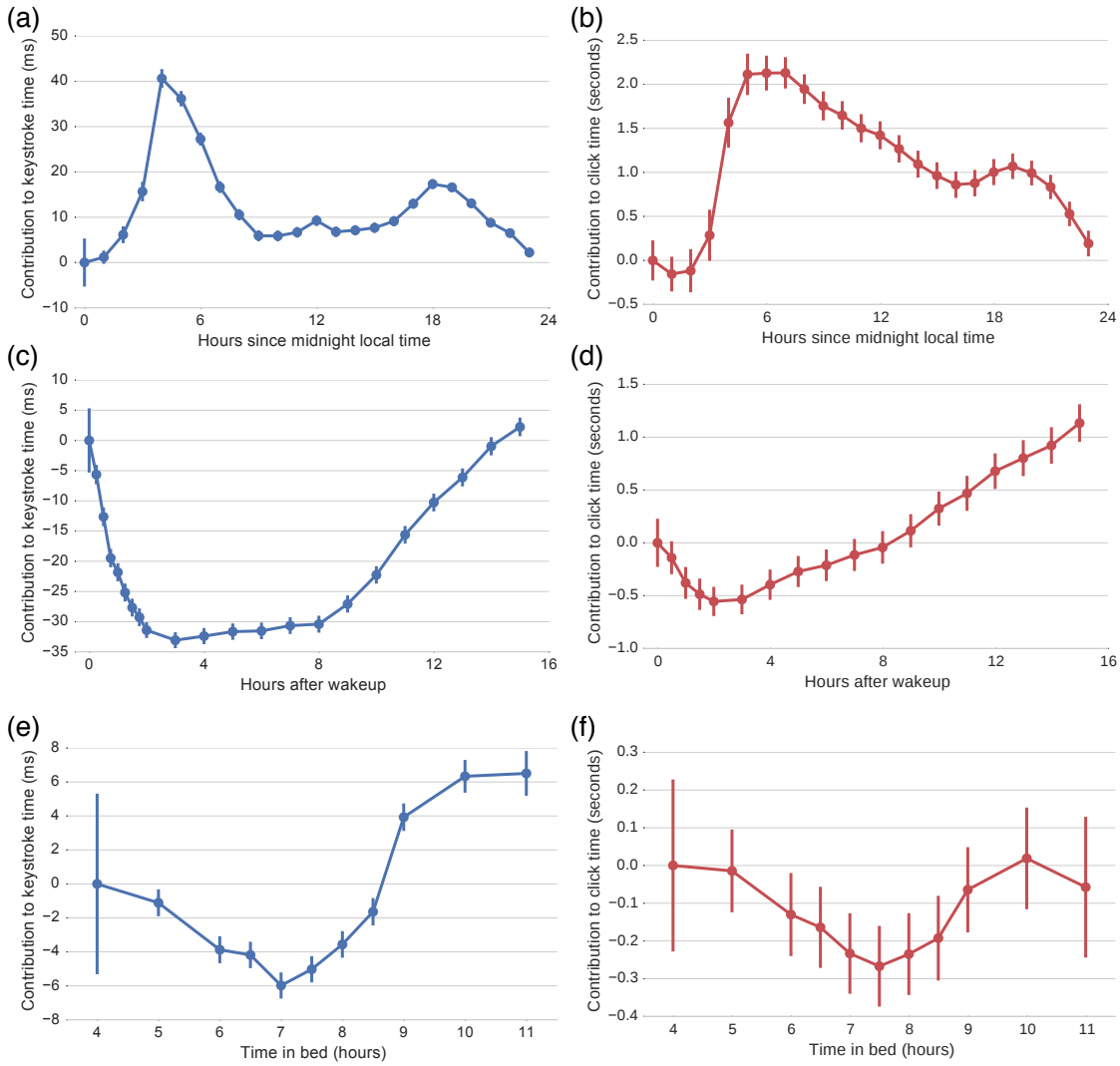

Figure 4.4 – Contributions to keystroke (a,c,e; blue) and click time (b,d,f; red) performance of different factors included in our model.. Results are similar for both performance measures and match estimates from controlled sleep studies in the laboratory (Section 4.5). For example, variation over the time of day $\mathbf{c^t}$ (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 $\mathbf{c^w}$ (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 $\mathbf{c^d}$ (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).

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 [Althoff et al., 2017a]). 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 [Dou et al., 2007]. We find that informational queries take about two seconds longer than navigational queries on average (online appendix [Althoff et al., 2017a]). 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.

4.5.2 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 [Achermann and Borbély, 1994]. 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 optimization. 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.

4.5.3 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 4.4. Impact on keystroke timings are shown in blue (Figure 4.4a,c,e) and impact on click times are shown in red (Figure 4.4b,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 [Althoff et al., 2017a]).

Time of Day. Cognitive performance on both keystroke and click tasks varies with time of day (Figure 4.4a,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 [Dijk et al., 1992; Wright Jr et al., 2012]. 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 4.1).

Time after Awakening. Cognitive performance also varies substantially with the time after wake up (Figure 4.4c,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 4.2). 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 [Borbély, 1982]. 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 [Åkerstedt and Folkard, 1997; Van Dongen and Dinges, 2000; Wright Jr et al., 2012].

Time in Bed. Keystrokes and click time vary with the amount of time in bed during the previous night (Figure 4.4e,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 [Kripke et al., 1979]). We further investigate the impact of insufficient sleep on performance in Section 4.6.

4.6 Influence of Insufficient Sleep on Performance

Following our studies to validate the methodology (Section 4.4 and Section 4.5), 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

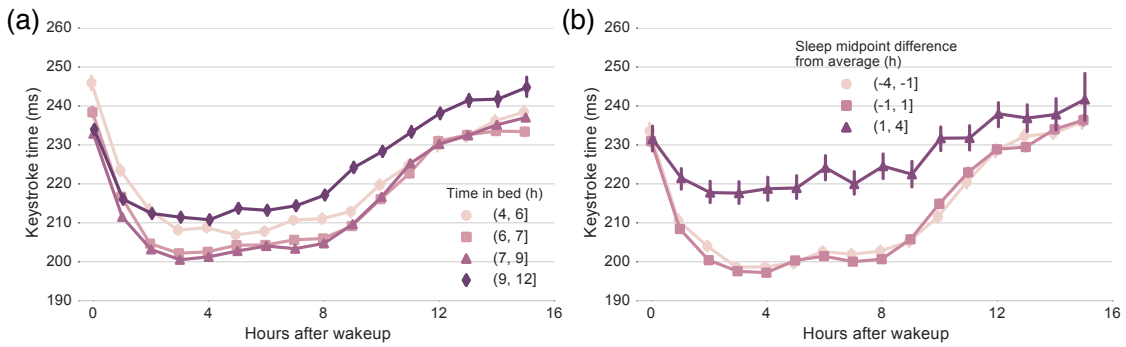

Figure 4.5 – The impact of sleep duration (a) and timing (b) on performance the next day. Sleep timing is measured through difference from the typical sleep midpoint and we control for sleep duration. We find that sleeping less than 7 or more than 9 hours is associated with slower performance (a). Sleeping 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).

of one or multiple nights of insufficient sleep on performance over the following days.

4.6.1 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 4.2 and Figure 4.4). Figure 4.5a shows that users performed significantly slower when in bed fewer than 6 or more than 9 hours, consistent with the results described in Section 4.5.3. 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 4.5b). 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

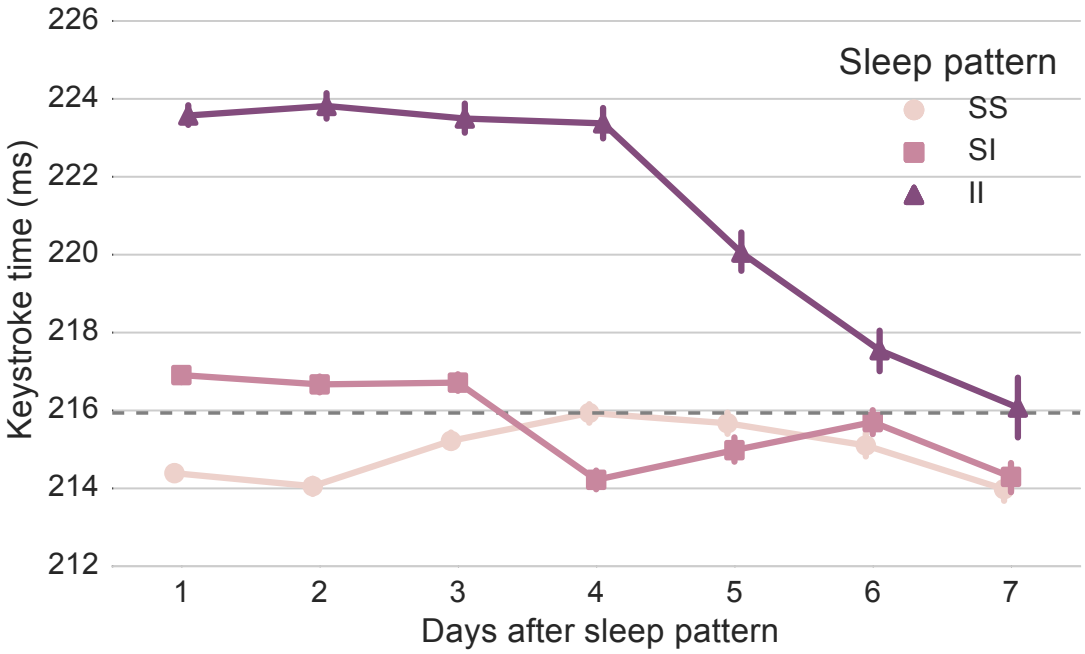

Figure 4.6 – Comparing the impact on performance of zero (SS), one (SI), or two (II) consecutive insufficient nights of sleep (less than six hours of time in bed). One night of insufficient sleep is associated with significantly slower keystroke times and two insufficient nights in a row exhibit a significantly 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.

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 [Dijk and Czeisler, 1995]. 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.

4.6.2 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 [Fernandez-Mendoza et al., 2010]), 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 4.6). 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% 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 4.6 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 4.6 that visualizes average performance throughout each of the seven days is included in the online appendix [Althoff et al., 2017a].

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.

4.7 Conclusion

Understanding human performance and its relation to sleep is critical to productivity [Colten and Altevogt, 2006], learning [Kelley et al., 2015], and avoiding accidents [Colten and Altevogt, 2006; Dinges, 1995]. Human performance is not constant but exhibits daily variations [Van Dongen and Dinges, 2000]. 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 [Roenneberg, 2013].

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 [Roenneberg, 2013],
- 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.

Chapter 5

Mental Health: Identifying Successful Conversation Strategies Through Large-scale Analysis of Counseling Conversations

5.1 Introduction

Mental illness is a major global health issue. In the U.S. alone, 43.6 million adults (18.1%) experience mental illness in a given year [National Institute of Mental Health, 2015]. In addition to the person directly experiencing a mental illness, family, friends, and communities are also affected [Insel, 2008].

In many cases, mental health conditions can be treated effectively through psychotherapy and counseling [World Health Organization, 2015]. However, it is far from obvious how to best conduct counseling conversations. Such conversations are free-form without strict rules, and involve many choices that could make a difference in someone's life. Thus far, quantitative evidence for effective conversation strategies has been scarce, since most studies on counseling have been limited to very small sample sizes and qualitative observations (*e.g.*, Labov and Fanshel, (1977); Haberstroh et al., (2007)). However, recent advances in technology-mediated counseling conducted online or through texting [Haberstroh et al., 2007] have allowed counseling services to scale with increasing demands and to collect large-scale data on counseling conversations and their outcomes.

Here we present the largest study on counseling conversation strategies published to date. We use data from an SMS texting-based counseling service where people in crisis (depression, self-harm, suicidal thoughts, anxiety, *etc.*), engage in therapeutic conversations with counselors. The data contains millions of messages

from eighty thousand counseling conversations conducted by hundreds of counselors over the course of one year. We develop a set of computational methods suited for large-scale discourse analysis to study how various linguistic aspects of conversations are correlated with conversation outcomes (collected via a follow-up survey).

We focus our analyses on counselors instead of individual conversations because we are interested in general conversation strategies rather than properties of specific issues. We find that there are significant, quantifiable differences between more successful and less successful counselors in how they conduct conversations.

Our findings suggest actionable strategies that are associated with successful counseling:

i. **Adaptability (Section 5.5):** Measuring the distance between vector representations of the language used in conversations going well and going badly, we find that successful counselors are more sensitive to the current trajectory of the conversation and react accordingly.
ii. **Dealing with Ambiguity (Section 5.6):** We develop a clustering-based method to measure differences in how counselors respond to very similar ambiguous situations. We learn that successful counselors clarify situations by writing more, reflect back to check understanding, and make their conversation partner feel more comfortable through affirmation.
iii. **Creativity (Section 5.6.3):** We quantify the diversity in counselor language by measuring cluster density in the space of counselor responses and find that successful counselors respond in a more creative way, not copying the person in distress exactly and not using too generic or "templated" responses.
iv. **Making Progress (Section 5.7):** We develop a novel sequence-based unsupervised conversation model able to discover ordered conversation stages common to all conversations. Analyzing the progression of stages, we determine that successful counselors are quicker to get to know the core issue and faster to move on to collaboratively solving the problem.
v. **Change in Perspective (Section 5.8):** We develop novel measures of perspective change using psycholinguistics-inspired word frequency analysis. We find that people in distress are more likely to be more positive, think about the future, and consider others, when the counselors bring up these concepts. We further show that this perspective change is associated with better conversation outcomes consistent with psychological theories of depression.

Further, we demonstrate that counseling success on the level of individual conversations is predictable using features based on our discovered conversation strategies (Section 5.9). Such predictive tools could be used to help counselors better progress through the conversation and could result in better counseling practices. The dataset used in this chapter has been released publicly and more information on

dataset access can be found at http://snap.stanford.edu/counseling.

Although we focus on crisis counseling in this chapter, our proposed methods more generally apply to other conversational settings and can be used to study how language in conversations relates to conversation outcomes.

5.2 Related Work

Our work relates to two lines of research:

Therapeutic Discourse Analysis & Psycholinguistics. The field of conversation analysis was born in the 1960s out of a suicide prevention center [Sacks and Jefferson, 1995; Van Dijk, 1997]. Since then conversation analysis has been applied to various clinical settings including psychotherapy [Labov and Fanshel, 1977]. Work in psycholinguistics has demonstrated that the words people use can reveal important aspects of their social and psychological worlds [Pennebaker et al., 2003]. Previous work also found that there are linguistic cues associated with depression [Campbell and Pennebaker, 2003; Ramirez-Esparza et al., 2008] as well as with suicude [Pestian et al., 2012]. These findings are consistent with Beck's cognitive model of depression (1967; cognitive symptoms of depression precede the affective and mood symptoms) and with Pyszczynski and Greenberg's self-focus model of depression (1987; depressed persons engage in higher levels of self-focus than non-depressed persons).

In this chapter, we propose an operationalized psycholinguistic model of perspective change and further provide empirical evidence for these theoretical models of depression.

Large-scale Computational Linguistics Applied to Conversations. Large-scale studies have revealed subtle dynamics in conversations such as coordination or style matching effects [Danescu-Niculescu-Mizil, 2012; Niederhoffer and Pennebaker, 2002] as well as expressions of social power and status [Bramsen et al., 2011; Danescu-Niculescu-Mizil et al., 2012]. Other studies have connected writing to measures of success in the context of requests [Althoff et al., 2014], user retention [Althoff and Leskovec, 2015], novels [Ashok et al., 2013], and scientific abstracts [Guerini et al., 2012]. Prior work has modeled dialogue acts in conversational speech based on linguistic cues and discourse coherence [Stolcke et al., 2000]. Unsupervised machine learning models have also been used to model conversations and segment them into speech acts, topical clusters, or stages. Most approaches employ Hidden Markov Model-like models [Barzilay and Lee, 2004; Paul, 2012; Ritter et al., 2010; Yang et al., 2014] which are also used in this chapter to model progression through conversation stages.

Very recently, technology-mediated counseling has allowed the collection of

large datasets on counseling. Howes et al. (2014) find that symptom severity can be predicted from transcript data with comparable accuracy to face-to-face data but suggest that insights into style and dialogue structure are needed to predict measures of patient progress. Counseling datasets have also been used to predict the conversation outcome [Huang, 2015] but without modeling the within-conversation dynamics that are studied in this chapter. Other work has explored how novel interfaces based on topic models can support counselors during conversations (Dinakar et al., 2014a; 2014b; 2015; Chen, 2014).

Our work joins these two lines of research by developing computational discourse analysis methods applicable to large datasets that are grounded in therapeutic discourse analysis and psycholinguistics.

5.3 Dataset Description

In this chapter, we study anonymized counseling conversations from a not-for-profit organization providing free crisis intervention via SMS messages. Text-based counseling conversations are particularly well suited for conversation analysis because all interactions between the two dialogue partners are fully observed (*i.e.*, there are no non-textual or non-verbal cues). Moreover, the conversations are important, constrained to dialogue between two people, and outcomes can be clearly defined (*i.e.*, we follow up with the conversation partner as to whether they feel better afterwards), which enables the study of how conversation features are associated with actual outcomes.

Counseling Process. Any person in distress can text the organization's public number. Incoming requests are put into a queue and an available counselor picks the request from the queue and engages with the incoming conversation. We refer to the crisis counselor as the *counselor* and the person in distress as the *texter*. After the conversation ends, the texter receives a follow-up question ("How are you feeling now? Better, same, or worse?") which we use as our conversation quality ground-truth (we use binary labels: good versus same/worse, since we care about improving the situation). In contrast to previous work that has used human judges to rate a caller's crisis state [Kalafat et al., 2007], we directly obtain this feedback from the texter. Furthermore, the counselor fills out a post-conversation report (*e.g.*, suicide risk, main issue such as depression, relationship, self-harm, suicide, *etc.*). All crisis counselors receive extensive training and commit to weekly shifts for a full year.

Dataset Statistics. Our dataset contains 408 counselors and 3.2 million messages in 80,885 conversations between November 2013 and November 2014 (see Table 5.1).

Dataset statistics	
Conversations	80,885
Conversations with survey response	15,555 (19.2%)
Messages	3.2 million
Messages with survey response	663,026 (20.6%)
Counselors	408
Messages per conversation*	42.6
Words per message*	19.2

Table 5.1 – Basic dataset statistics. Rows marked with * are computed over conversations with survey responses.

	NA	Depressed	Relationship	Self harm	Family	Suicide	Stress	Anxiety	Other
Success rate	0.556	0.612	0.659	0.672	0.711	0.573	0.696	0.671	0.537
Frequency	0.200	0.200	0.089	0.074	0.071	0.063	0.041	0.039	0.035
Frequency with more successful counselors	0.203	0.199	0.089	0.067	0.072	0.061	0.048	0.042	0.030
Frequency with less successful counselors	0.223	0.208	0.087	0.070	0.067	0.056	0.030	0.032	0.028

Table 5.2 – Frequencies and success rates for the nine most common conversation issues (NA: Not available). On average, more and less successful counselors face the same distribution of issues.

All system messages (*e.g.*, instructions), as well as texts that contain survey responses (revealing the ground-truth label for the conversation) were filtered out. Out of these conversations, we use the 15,555, or 19.2%, that contain a ground-truth label (whether the texter feels better or the same/worse after the conversation) for the following analyses. Conversations span a variety of issues of different difficulties (see rows one and two of Table 5.2). Approval to analyze the dataset was obtained from the Stanford IRB.

5.4 Defining Counseling Quality

The primary goal of this paper is to study strategies that lead to conversations with positive outcomes. Thus, we require a ground-truth notion of conversation quality. In principle, we could study individual conversations and aim to understand what factors make the conversation partner (texter) feel better. However, it is advantageous to focus on the conversation actor (counselor) instead of individual conversations.

There are several benefits of focusing analyses on counselors (rather than individual conversations): First, we are interested in general conversation strategies rather than properties of main issues (*e.g.*, depression vs. suicide). While each

conversation is different and will revolve around its main issue, we assume that counselors have a particular style and strategy that is invariant across conversations. Second, we assume that conversation quality is noisy. Even a very good counselor will face some hard conversations in which they do everything right but are still unable to make their conversation partner feel better. Over time, however, the "true" quality of the counselor will become apparent. Third, our goal is to understand successful conversation strategies and to make use of these insights in counselor training. Focusing on the counselor is helpful in understanding, monitoring, and improving counselors' conversation strategies.

More vs. Less Successful Counselors. We split the counselors into two groups and then compare their behavior. Out of the 113 counselors with more than 15 labeled conversations of at least 30 messages each, we use the most successful 40 counselors as "more successful" counselors and the bottom 40 as "less successful" counselors. Their average success rates are 66.3-85.5% and 42.1-58.6%, respectively. While the counselor-level analysis is of primary concern, we will also differentiate between counselor behavior in "positive" versus "negative" conversations (*i.e.*, those that will eventually make the texter feel better vs. not). Thus, in the remainder of the paper we differentiate between more vs. less successful counselors and positive vs. negative conversations. Studying the cross product of counselors and conversations allows us to gain insights on how both groups behave in positive and negative conversations. For example, Figure 5.1 illustrates why differentiating between counselors and as well as conversations is necessary: differences in counselor message length over the course of the conversation are bigger between more and less successful counselors than between positive and negative conversations.

Initial Analysis. Before focusing on detailed analyses of counseling strategies we address two important questions: Do counselors specialize in certain issues? And, do successful counselors appear successful only because they handle "easier" cases?

To gain insights into the "specialization hypothesis" we make use the counselor annotation of the main issue (depression, self-harm, *etc.*). We compare success rates of counselors across different issues and find that successful counselors have a higher fraction of positive conversations across all issues and that less successful counselors typically do not excel at a particular issue. Thus, we conclude that counseling quality is a general trait or skill and supporting that the split into more and less successful counselors is meaningful.

Another simple explanation of the differences between more and less successful counselors could be that successful counselors simply pick "easy" issues. However, we find that this is not the case. In particular, we find that both counselor groups

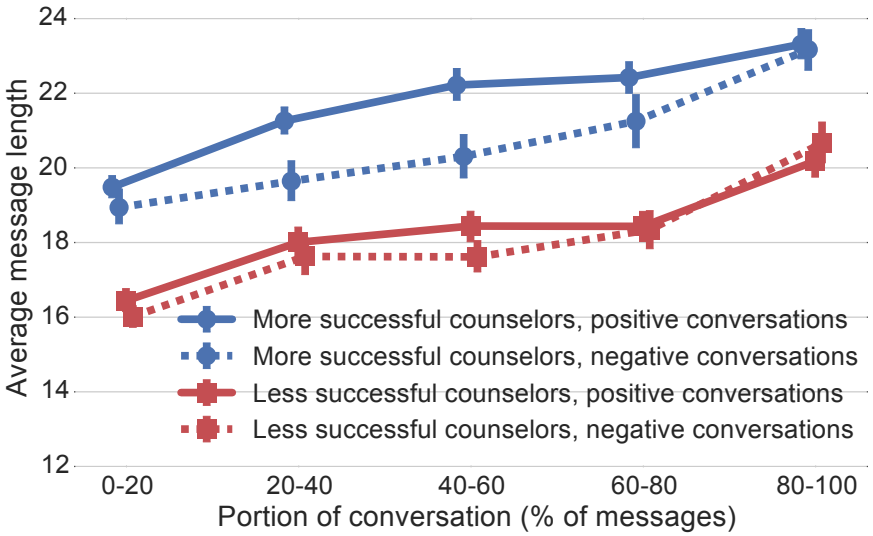

Figure 5.1 – Differences in counselor message length (in #tokens) over the course of the conversation are larger between more and less successful counselors (blue circle/red square) than between positive and negative conversations (solid/dashed). Error bars in all plots correspond to bootstrapped 95% confidence intervals using the member bootstrapping technique from Ren et al. [2010].

are very similar in how they select conversations from the queue (picking the top-most in 60.1% vs. 60.3%, respectively), work similar shifts, and handle a similar number of conversations simultaneously (1.98 vs. 1.83). Further, we find that both groups face similar distributions of issues over time (see Table 5.2). We attribute the largest difference, "NA" (main issue not reported), to the more successful counselors being more diligent in filling out the post-conversation report and having fewer conversations that end before the main issue is introduced.

5.5 Counselor Adaptability

In the remainder of the paper we focus on factors that mediate the outcome of a conversation. First, we examine whether successful counselors are more aware that their current conversation is going well or badly and study how the counselor adapts to the situation. We investigate this question by looking for language differences between positive and negative conversations. In particular, we compute a distance measure between the language counselors use in positive conversations and the language counselors use in negative conversations and observe how this distance changes over time.

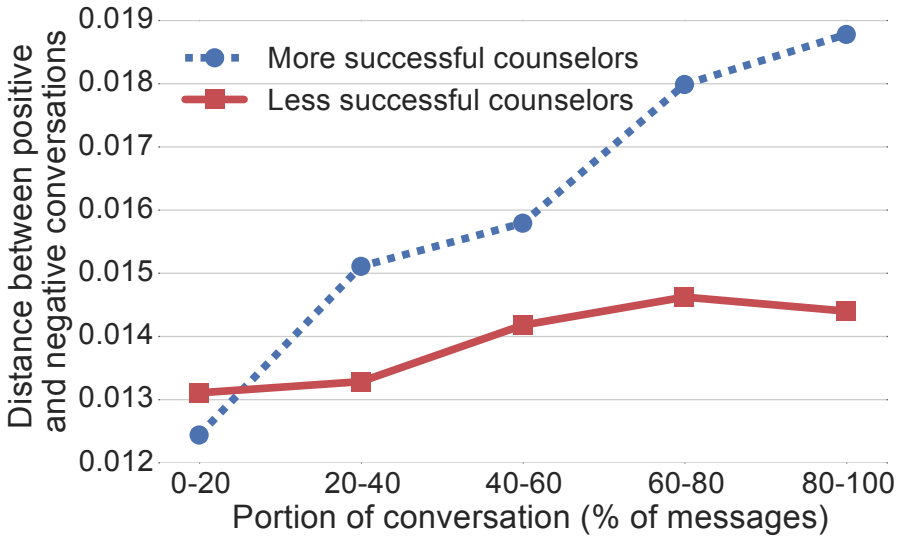

Figure 5.2 – More successful counselors are more varied in their language across positive/negative conversations, suggesting they adapt more. All differences between more successful and less successful counselors except for the 0-20 bucket were found to be statistically significant ($p < 0.05$; bootstrap resampling test).

We capture the time dimension by breaking up each conversation into five even chunks of messages. Then, for each set of counselors (more successful or less successful), conversation outcome (positive or negative), and chunk (first 20%, second 20%, etc.), we build a TF-IDF vector of word occurrences to represent the language of counselors within this subset. We use the global inverse document (*i.e.*, conversation) frequencies instead of the ones from each subset to make the vectors directly comparable and control for different counselors having different numbers of conversations by weighting conversations so all counselors have equal contributions. We then measure the difference between the "positive" and "negative" vector representations by taking the cosine distance in the induced vector space. We also explored using Jensen-Shannon divergence between traditional probabilistic language models and found these methods gave similar results.

Results. We find more successful counselors are more sensitive to whether the conversation is going well or badly and vary their language accordingly (Figure 5.2). At the beginning of the conversation, the language between positive and negative conversations is quite similar, but then the distance in language increases over time. This increase in distance is much larger for more successful counselors than less successful ones, suggesting they are more aware of when conversations are going poorly and adapt their counseling more in an attempt to remedy the situation.

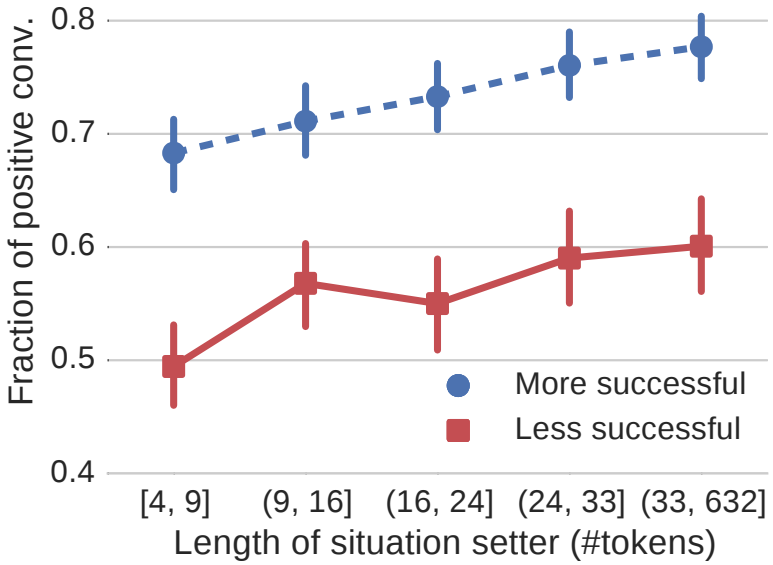

Figure 5.3 – More ambiguous situations (length of situation setter) are less likely to result in positive conversations..

5.6 Reacting to Ambiguity

Observing that successful counselors are better at adapting to the conversation, we next examine *how* counselors differ and what factors determine the differences. In particular, domain experts have suggested that more successful counselors are better at handling ambiguity in the conversation [Levitt and Jacques, 2005]. Here, we use *ambiguity* to refer to the uncertainty of the situation and the texter's actual core issue resulting from insufficiently short or uncertain descriptions. Does initial ambiguity of the situation negatively affect the conversation? How do more successful counselors deal with ambiguous situations?

Ambiguity. Throughout this section we measure ambiguity in the conversation as the shortness of the texter's responses in number of words. While ambiguity could also be measured through concreteness ratings of the words in each message (*e.g.*, using concreteness ratings from Brysbaert et al. [2014]), we find that results are very similar and that length and concreteness are strongly related and hard to distinguish.

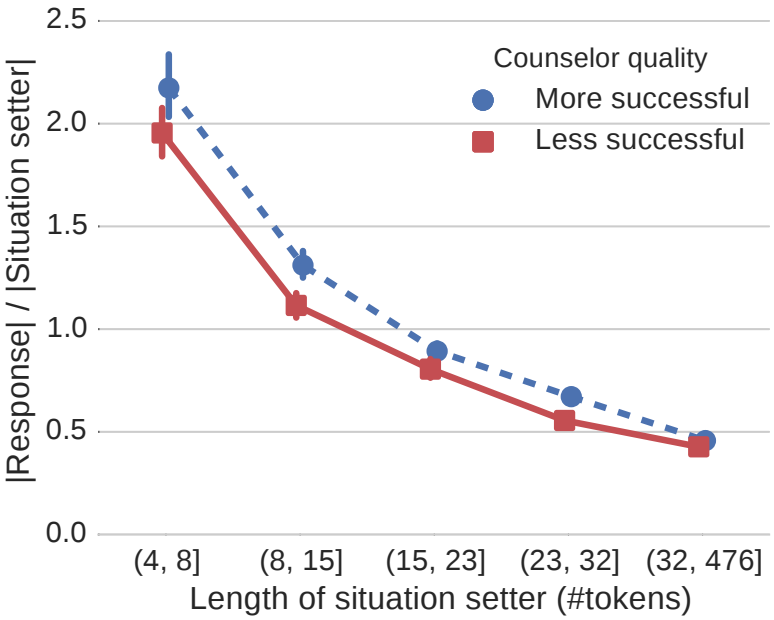

Figure 5.4 – All counselors react to short, ambiguous messages by writing more (relative to the texter message) but more successful counselors do it more than less successful counselors. .

5.6.1 Initial Ambiguity and Situation Setter

It is challenging to measure ambiguity and reactions to ambiguity at arbitrary points throughout the conversation since it strongly depends on the context of the entire conversation (*i.e.,* all earlier messages and questions). However, we can study nearly identical *beginnings* of conversations where we can directly compare how more successful and less successful counselors react given nearly identical situations (the texter first sharing their reason for texting in). We identify the *situation setter* within each conversation as the first long message by the texter (typically a response to a "Can you tell me more about what is going on?" question by the counselor).

Results. We find that ambiguity plays an important role in counseling conversations. Figure 5.3 shows that more ambiguous situations (shorter length of situation setter) are less likely to result in successful conversations (we obtain similar results when measuring concreteness [Brysbaert et al., 2014] directly). Further, we find that counselors generally react to short and ambiguous situation setters by writing significantly more than the texters (Figure 5.4; if counselors wrote exactly as much

as the texter, we would expect a horizontal line $y = 1$). However, more successful counselors react more strongly to ambiguous situations than less successful counselors.

5.6.2 How to Respond to Ambiguity

Having observed that ambiguity plays an important role in counseling conversations, we now examine in greater detail how counselors respond to nearly identical situations.

We match situation setters by representing them through TF-IDF vectors on bigrams and find similar situation setters as nearest neighbors within a certain cosine distance in the induced space.[1] We only consider situation setters that are part of a dense cluster with at least 10 neighbors, allowing us to compare follow-up responses by the counselors (4829/12770 situation setters were part of one of 589 such clusters). We also used distributed word embeddings (*e.g.*, [Mikolov et al., 2013]) instead of TF-IDF vectors but found the latter to produce better clusters.

Based on counselor training materials we hypothesize that more successful counselors

- address ambiguity by writing more themselves,
- use more check questions (statements that tell the conversation partner that you understand them while avoiding the introduction of any opinion or advice [Labov and Fanshel, 1977]; *e.g.*"that sounds like..."),
- check for suicidal thoughts early (*e.g.*, "want to die"),
- thank the texter for showing the courage to talk to them (*e.g.*, "appreciate"),
- use more hedges (mitigating words used to lessen the impact of an utterance; *e.g.*, "maybe", "fairly"),
- and that they are less likely to respond with surprise (*e.g.*, "oh, this sounds really awful").

A set of regular expressions is used to detect each class of responses (similar to the examples above).

Results. We find several statistically significant differences in how counselors respond to nearly identical situation setters (see Table 5.3). While situation setters tend to be slightly longer for more successful counselors (suggesting that conversations are not perfectly randomly assigned), counselor responses are significantly longer and also spur longer texter responses. Further, the more successful counselors respond in a way that is less similar to the original situation setter

[1] Threshold manually set after qualitative analysis of matches from randomly chosen clusters. Results were not overly sensitive to threshold choice, choice of representation (*e.g.*, word vectors), and distance measure (*e.g.*, Euclidean).

	More S.	Less S.	Test
% conversations successful	70.7	51.7	***
#messages in conversation	57.0	46.7	***
Situation setter length (#tokens)	12.1	10.7	***
C response length (#tokens)	15.8	11.8	***
T response length (#tokens)	20.4	18.8	***
% Cosine sim. C resp. to context	11.9	14.8	***
% Cosine sim. T resp. to context	7.6	7.3	–
% C resp. w check question	12.6	4.1	***
% C resp. w suicide check	13.5	10.3	***
% C resp. w thanks	6.3	2.4	***
% C resp. w hedges	41.4	36.8	***
% C resp. w surprise	3.3	2.8	–

Table 5.3 – Differences between more and less successful counselors (C; More S. and Less S.) in responses to nearly identical situation setters (Sec. 5.6.1) by the texter (T).. Last column contains significance levels of Wilcoxon Signed Rank Tests (*** $p < 0.001$, – $p > 0.05$).

(measured by cosine similarity in TF-IDF space) compared to less successful counselors (but the texter's response does not seem affected). We do find that more successful counselors use more check questions, check for suicide ideation more often, show the texter more appreciation, and use more hedges, but we did not find a significant difference with respect to responding with surprise.

5.6.3 Response Templates and Creativity

In Section 5.6.2, we observed that more successful counselors make use of certain templates (including check questions, checks for suicidal thoughts, affirmation, and using hedges). While this could suggest that counselors should stick to such predefined templates, we find that, in fact, more successful counselors do respond in more creative ways.

We define a measure of how "templated" the counselors responses are by counting the number of similar responses in TF-IDF space for the counselor reaction (*c.f.*, Section 5.6.2; again using a manually defined and validated threshold on cosine distance).

Figure 5.5 shows that more successful counselors use less common/templated questions. This suggests that while more successful counselors questions follow certain patterns, they are more *creative* in their response to each situation. This tailoring of responses requires more effort from the counselor, which is consistent

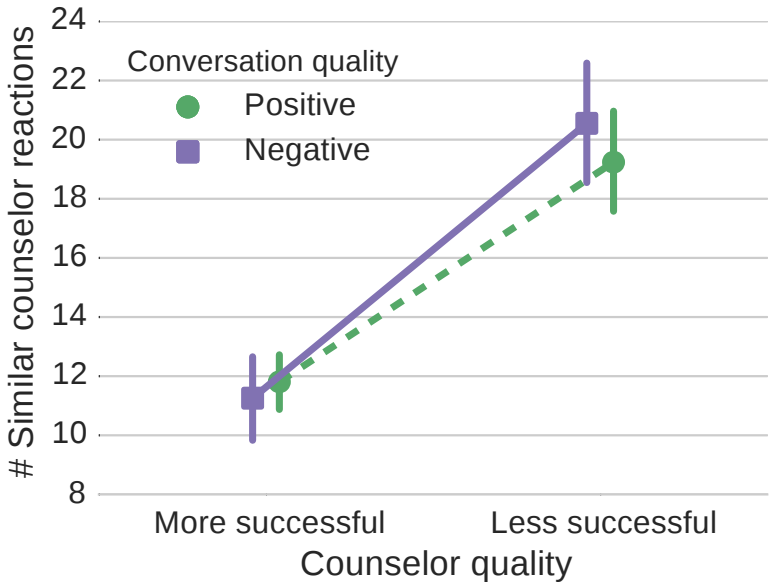

Figure 5.5 – More successful counselors use less common/templated responses (after the texter first explains the situation). This suggests that they respond in a more creative way. There is no significant difference between positive and negative conversations.

with the results in Figure 5.1 that showed that more successful counselors put in more effort in composing longer messages as well.

5.7 Ensuring Conversation Progress

After demonstrating content-level differences between counselors, we now explore temporal differences in how counselors progress through conversations. Using an unsupervised conversation model, we are able to discover distinct conversation stages and find differences between counselors in how they move through these stages. We further provide evidence that these differences could be related to power and authority by measuring linguistic coordination between the counselor and texter.

5.7.1 Unsupervised Conversation Model

Counseling conversations follow a common structure due to the nature of conversation as well as counselor training. Typically, counselors first introduce themselves, get to know the texter and their situation, and then engage in constructive problem

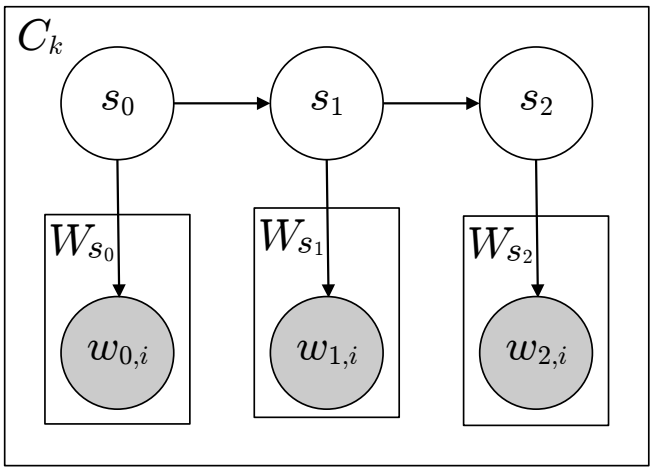

Figure 5.6 – Our conversation model generates a particular conversation C_k by first generating a sequence of hidden states $s_0, s_1, ...$ according to a Markov model. Each state s_i then generates a message as a bag of words $w_{i,0}, w_{i,1}, ...$ according a unigram language model W_{s_i}.

solving. We employ unsupervised conversation modeling techniques to capture this stage-like structure within conversations.

Our conversation model is a message-level Hidden Markov Model (HMM). Figure 5.6 illustrates the basic model where hidden states of the HMM represent *conversation stages*. Unlike in prior work on conversation modeling, we impose a fixed ordering on the stages and only allow transitions from the current stage to the next one (Figure 5.7). This causes it to learn a fixed dialogue structure common to all of the counseling sessions as opposed to conversation topics. Furthermore, we separately model counselor and texter messages by treating their turns in the conversation as distinct states. We train the conversation model with expectation maximization, using the forward-backward algorithm to produce the distributions during each expectation step. We initialized the model with each stage producing messages according to a unigram distribution estimated from all messages in the dataset and uniform transition probabilities. The unigram language models are defined over all words occurring more than 20 times (over 98% of words in the dataset), with other words replaced by an unknown token.

Results. We explored training the model with various numbers of stages and found five stages to produce a distinct and easily interpretable representation of a conversation's progress. Table 5.4 shows the words most unique to each stage. The first and last stages consist of the basic introductions and wrap-ups common to all conversations. In stage 2, the texter introduces the main issue, while the counselor asks for clarifications and expresses empathy for the situation. In stage

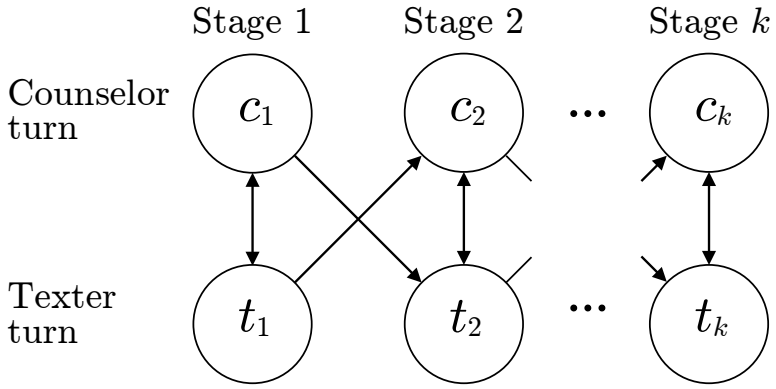

Figure 5.7 – Allowed state transitions for the conversation model. Counselor and texter messages are produced by distinct states and conversations must progress through the stages in increasing order.

Stage	Interpretation	Top words for texter	Top words for counselor
1	Introductions	hi, hello, name, listen, hey	hi, name, hello, hey, brings
2	Problem introduction	dating, moved, date, liked, ended	gosh, terrible, hurtful, painful, ago
3	Problem exploration	knows, worry, burden, teacher, group	react, cares, considered, supportive, wants
4	Problem solving	write, writing, music, reading, play	hobbies, writing, activities, distract, music
5	Wrap up	goodnight, bye, thank, thanks, appreciate	goodnight, 247, anytime, luck, 24

Table 5.4 – The top 5 words for counselors and texters with greatest increase in likelihood of appearing in each stage. The model successfully identifies interpretable stages consistent with counseling guidelines (qualitative interpretation based on stage assignment and model parameters; only words occurring more than five hundred times are shown).

3, the counselor and texter discuss the problem, particularly in relation to the other people involved. In stage 4, the counselor and texter discuss actionable strategies that could help the texter. This is a well-known part of crisis counselor training called “collaborative problem solving."

5.7.2 Analyzing Counselor Progression

Do counselors differ in how much time they spend at each stage? In order to explore how counselors progress through the stages, we use the Viterbi algorithm to assign each conversation the most likely sequence of stages according to our conversation model. We then compute the average duration in messages of each stage for both more and less successful counselors. We control for the different distributions of positive and negative conversations among more successful and less successful counselors by giving the two classes of conversations equal weight

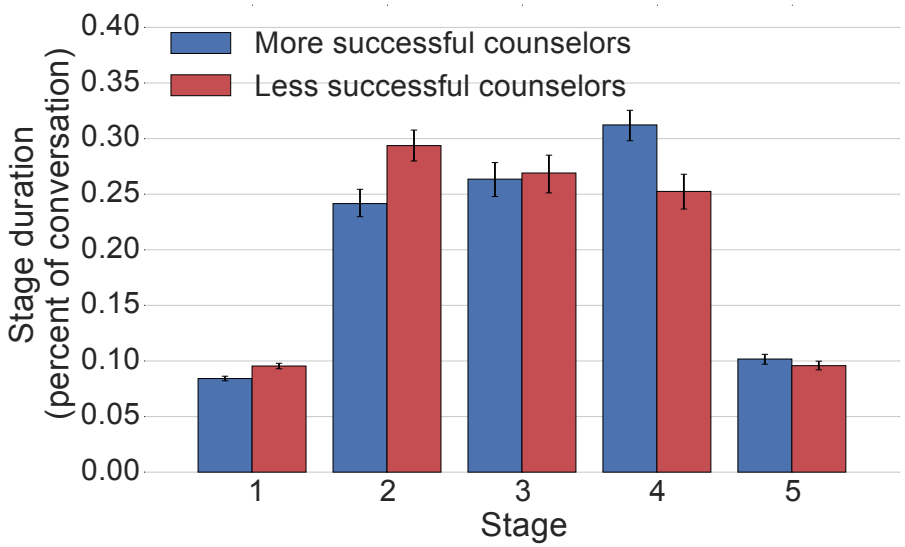

Figure 5.8 – More successful counselors are quicker to get to know texter and issue (stage 2) and use more of their time in the "problem solving" phase (stage 4)..

and control for different conversation lengths by only including conversations between 40 and 60 messages long.

Results. We find that more successful counselors are quicker to move past the earlier stages, particularly stage 2, and spend more time in later stages, particularly stage 4 (Figure 5.8). This suggests they are able to more quickly get to know the texter and then spend more time in the problem solving phase of the conversation, which could be one of the reasons they are more successful.

5.7.3 Coordination and Power Differences

One possible explanation for the more successful counselors' ability to quickly move through the early stages is that they have more "power" in the conversation and can thus exert more control over the progression of the conversation. We explore this idea by analyzing linguistic coordination, which measures how much the conversation partners adapt to each other's conversational styles. Research has shown that conversation participants who have a greater position of power coordinate less (*i.e.*, they do not adapt their linguistic style to mimic the other conversational participant as strongly) [Danescu-Niculescu-Mizil et al., 2012].

In our analysis, we use the "Aggregated 2" coordination measure $C(B, A)$ from Danescu-Niculescu-Mizil [2012], which measures how much group B coordinates

to group A (a higher number means more coordination). The measure is computed by counting how often specific markers (e.g., auxiliary verbs) are exhibited in conversations. If someone tends to use a particular marker right after their conversation partner uses that marker, it suggests they are coordinating to their partner.

Formally, let set S be a set of exchanges, each involving an initial utterance u_1 by $a \in A$ and a reply u_2 by $b \in B$. Then the coordination of b to A according to a linguistic marker m is:

$$C^m(b, A) = P(\mathcal{E}^m_{u_2 \to u_1} | \mathcal{E}^m_{u_1}) - P(\mathcal{E}^m_{u_2 \to u_1})$$

where $\mathcal{E}^m_{u_1}$ is the event that utterance u_1 exhibits m (*i.e.*, contains a word from category m) and $\mathcal{E}^m_{u_2 \to u_1}$ is the event that reply u_2 to u_1 exhibits m. The probabilities are estimated across all exchanges in S. To aggregate across different markers, we average the coordination values of $C^m(b, A)$ over all markers m to get a macro-average $C(b, A)$. The coordination between groups B and A is then defined as the mean of the coordinations of all members of group B towards the group A.

We use eight markers from Danescu-Niculescu-Mizil (2012), which are considered to be processed by humans in a generally non-conscious fashion: articles, auxiliary verbs, conjunctions, high-frequency adverbs, indefinite pronouns, personal pronouns, prepositions, and quantifiers.

Results. Texters coordinate less than counselors, with texters having a coordination value of C(texter, counselor)=0.019 compared to the counselor's larger C(counselor, texter)=0.030, suggesting that the texters hold more "power" in the conversation. However, more successful counselors coordinate less than their less successful counterparts (C(more succ. counselors, texter)=0.029 vs. C(less succ. counselors, texter)=0.032). All differences are statistically significant ($p < 0.01$; Mann-Whitney U test). This suggests that more successful counselors act with more control over the conversation, which could explain why they are quicker to make it through the initial conversation stages.

5.8 Facilitating Perspective Change

Thus far, we have studied conversation dynamics and their relation to conversation success from the counselor perspective. In this section, we show that *perspective change* in the *texter* over time is associated with a higher likelihood of conversation success. Prior work has shown that day-to-day changes in writing style are associated with positive health outcomes [Campbell and Pennebaker, 2003], and existing theories link depression to a negative view of the future [Pyszczynski

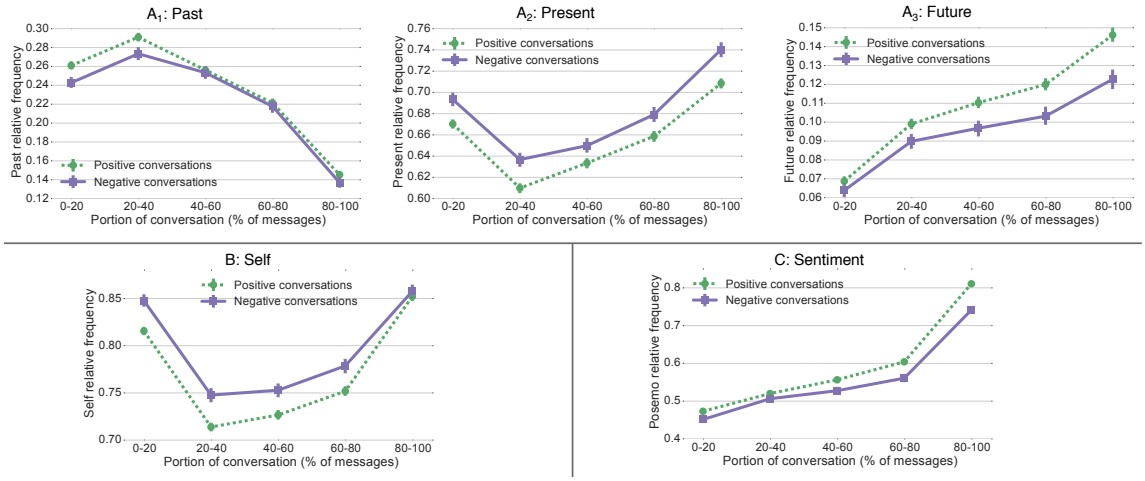

Figure 5.9 – Perspective change throughout the conversation. A: Throughout the conversation there is a shift from talking about the past to future, where in positive conversations this shift is greater; B: Texters that talk more about others more often feel better after the conversation; C: More positive sentiment by the texter throughout the conversation is associated with successful conversations.

et al., 1987] and a self-focusing style [Pyszczynski and Greenberg, 1987]. Here, we propose a novel measure to quantify three orthogonal aspects of perspective change within a single conversation: *time*, *self*, and *sentiment*. Further, we show that the counselor might be able to actively induce perspective change.

Time. Texters start explaining their issue largely in terms of the past and present but over time talk more about the future (see Figure 5.9A; each plot shows the relative amount of words in the LIWC past, present, and future categories [Tausczik and Pennebaker, 2010]). We find that texters writing more about the future are more likely to feel better after the conversation. This suggests that changing the perspective from issues in the past towards the future is associated with a higher likelihood of successfully working through the crisis.

Self. Another important aspect of behavior change is to what degree the texter is able to change their perspective from talking about themselves to considering others and potentially the effect of their situation on others [Campbell and Pennebaker, 2003; Pyszczynski and Greenberg, 1987]. We measure how much the texter is focused on themselves by the relative amount of first person singular pronouns (I, me, mine) versus third person singular/plural pronouns (she, her, him / they, their), again using LIWC. Figure 5.9B shows that a smaller amount of self-focus is associated with more successful conversations (providing support for the self-focus model of depression [Pyszczynski and Greenberg, 1987]). We

hypothesize that the lack of difference at the end of the conversation is due to conversation norms such as thanking the counselor ("*I* really appreciate it.") even if the texter does not actually feel better.

Sentiment. Lastly, we investigate how much a change in sentiment of the texter throughout the conversation is associated with conversation success. We measure sentiment as the relative fraction of positive words using the LIWC PosEmo and NegEmo sentiment lexicons. The results in Figure 5.9C show that texters always start out more negative (value below 0.5), but that the sentiment becomes more positive over time for both positive and negative conversations. However, we find that the separation between both groups grows larger over time, which suggests that a positive perspective change throughout the conversation is related to higher likelihood of conversation success. We find that both curves increase significantly at the very end of the conversation. Again, we attribute this to conversation norms such as thanking the counselor for listening even when the texter does not actually feel better. Together with the result on talking about the future, these findings are consistent with the theory of Pyszczynski et al. (1987) that depression is related to a negative view of the future.

Role of a Counselor. Given that positive conversations often exhibit perspective change, a natural question is how counselors can encourage perspective change in the texter. We investigate this by exploring the hypothesis that the texter will tend to talk more about something (*e.g.*, the future), if the counselor first talks about it. We measure this tendency using the same coordination measures as Section 5.7.3 except that instead of using stylistic LIWC markers (*e.g.*, auxiliary verbs, quantifiers), we use the LIWC markers relevant to the particular aspect of perspective change (e.g., Future, HeShe, PosEmo). In all cases we find a statistically significant ($p < 0.01$; Mann-Whitney U-test) increase in the likelihood of the texter using a LIWC marker if the counselor used it in the previous message (~4-5% change). This link between perspective change and how the counselor conducts the conversation suggests that the counselor might be able to actively induce measurable perspective change in the texter.

5.9 Predicting Counseling Success

In this section, we combine our quantitative insights into a prediction task. We show that the linguistic aspects of crisis counseling explored in previous sections have predictive power at the level of individual conversations by evaluating their effectiveness as features in classifying the outcome of conversations. Specifically, we create a balanced dataset of positive and negative conversations more than 30 messages long and train a logistic regression model to predict the outcome given

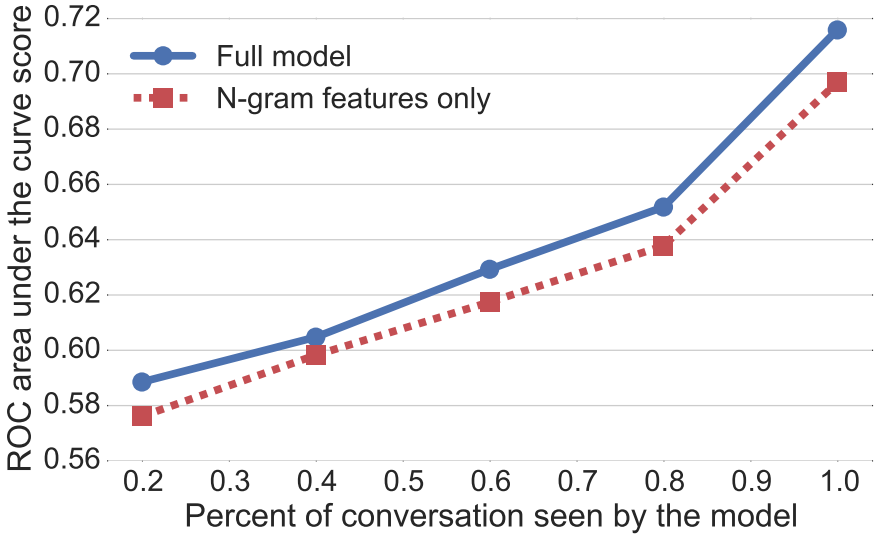

Figure 5.10 – Prediction accuracies vs. percent of the conversation seen by the model (without texter features).

the first x% of messages in the conversation. There are 3619 such negative conversations and and we randomly subsample the larger set of positive conversations. We train the model with batch gradient descent and use L1 regularization when n-gram features are present and L2 regularization otherwise. We evaluate our model with 10-fold cross-validation and compare models using the area under the ROC curve (AUC).

Features. We include three aspects of counselor messages discussed in Section 5.6: hedges, check questions, and the similarity between the counselor's message and previous texter message. We add a measure of how much progress the counselor has made (Section 5.7) by computing the Viterbi path of stages for the conversation (only for the first x%) with the HMM conversation model and then adding the duration of each stage (in #messages) as a feature. Additionally, we add average message length and average sentiment per message using VADER sentiment [Hutto and Gilbert, 2014]. Further, we add temporal dynamics to the model by adding feature conjunctions with the stages HMM model. After running the stages model over the x% of the conversation available to the classifier, we add each feature's average value over each stage as additional features. Lastly, we explore the benefits of adding surface-level text features to the model by adding unigram and bigram features. Because the focus of this work is on counseling strategies, we primarily experiment with models using only features from counselor messages. For completeness, we also report results for a model including texter features.

Prediction Results. The model's accuracy increases with x, and we show that the

Features	ROC AUC
Counselor unigrams only	0.630
Counselor unigrams and bigrams only	0.638
None	0.5
+ hedges	0.514 (+0.014)
+ check questions	0.546 (+0.032)
+ similarity to last message	0.553 (+0.007)
+ duration of each stage	0.561 (+0.008)
+ sentiment	0.590 (+0.029)
+ message length	0.596 (+0.006)
+ stages feature conjunction	0.606 (+0.010)
+ counselor unigrams and bigrams	**0.652 (+0.046)**
+ texter unigrams and bigrams	**0.708 (+0.056)**

Table 5.5 – Performance of nested models predicting conversation outcome given the first 80% of the conversation. In bold: full models with only counselor features and with additional texter features.

model is able to distinguish positive and negative conversations after only seeing the first 20% of the conversation (see Figure 5.10). We attribute the significant increase in performance for $x = 100$ (Accuracy=0.687, AUC=0.716) to strong linguistic cues that appear as a conversation wraps up (e.g., "I'm glad you feel better."). To avoid this issue, our detailed feature analysis is performed at $x = 80$.

Feature Analysis. The model performance as features are incrementally added to the model is shown in Table 5.5. All features improve model accuracy significantly ($p < 0.001$; paired bootstrap resampling test). Adding n-gram features produces the largest boost in AUC and significantly improves over a model just using n-gram features (0.638 vs. 0.652 AUC). Note that most features in the full model are based on word frequency counts that can be derived from n-grams which explains why a simple n-gram model already performs quite well. However, our model performs well with only a small set of linguistic features, demonstrating they provide a substantial amount of the predictive power. The effectiveness of these features shows that, in addition to exhibiting group-level differences reported earlier in this chapter, they provide useful signal for predicting the outcome of individual conversations.

5.10 Conclusion

Knowledge about how to conduct a successful counseling conversation has been limited by the fact that studies have remained largely qualitative and small-scale. In this chapter, we presented a large-scale quantitative study on the discourse of counseling conversations. We developed a set of novel computational discourse analysis methods suited for large-scale datasets and used them to discover actionable conversation strategies that are associated with better conversation outcomes. We hope that this work will inspire future generations of tools available to people in crisis as well as their counselors. For example, our insights could help improve counselor training and give rise to real-time counseling quality monitoring and answer suggestion support tools.

Chapter 6

Conclusions

6.1 Summary of Contributions

The goal of this thesis was to demonstrate that novel computational methods can derive new insights from already collected digital activity traces that can help us better understand and improve human well-being. We considered three key aspects of human health and well-being: physical activity, sleep, and mental health. First, we showed how to leverage consumer smartphone data on a global scale, which revealed a previously unknown activity inequality and gender activity gap (Chapter 2). We then proposed a machine learning model to predict human real-world actions ahead of time (Chapter 3). This model could be used to drive just-in-time interventions that encourage physical activity and attempt to reduce overall activity inequality, or support healthy eating habits. Next, we combined digital traces from web search engine interactions with wearable sleep data to study how variation and lack of sleep affect cognitive function (Chapter 4). Lastly, we demonstrated that beyond these health behaviors relevant to physical health (*i.e.*, physical activity and sleep), we can study mental health through digital traces as well. Specifically, we conducted a study of successful conversation strategies through a large-scale counseling corpus and new linguistic analysis methods (Chapter 5). These studies form a first step towards developing scalable computational techniques to measure, understand, predict, and enhance human behavior and well-being.

6.2 Future directions

The results presented in this thesis point to several interesting future directions, some of which we shall outline here, thus concluding the thesis.

6.2.1 Data science tools for large-scale and high-dimensional observational data

Throughout this dissertation we encountered multiple instances of analyzing large-scale datasets of human activities to gain actionable insights. Due to the observational nature of the data, we needed to put special attention on ruling out key confounding factors and alternative explanations. Ensuring that the findings were robust took a variety of forms, from validating smartphone and wearable-based activity and sleep measures (Chapter 2 and Chapter 4), to testing generalization across mobile applications (Chapter 3), to manually designed experiments (Chapter 2 and Chapter 4), to matching techniques (Chapter 5), and leveraging natural experiments (Chapter 5; see also [Althoff et al., 2017b]).

Due to the large cost and limited scope of randomized controlled trials, scientific advances will increasingly be based on large-scale observational studies like the ones presented in this dissertation. However, these advances are contingent on democratized ability to conduct observational studies that meet high standards of validity. Therefore, we need tools that turn observational data into robust inferences and enable high-quality observational science. For instance, these tools should reveal and automatically correct bias whenever possible and help establish causal relationships. Specific challenges include scaling balancing methods to large data, causal inference in high-dimensional spaces such as when modeling language, and handling non-binary treatments such as dose-response relationships.

6.2.2 Designing supportive online social networks

Online social networks are common places of human interaction, both good and bad, for billions of people. We have seen in Chapter 5 that technology mediates serious conversations between people and that we can leverage the digital traces collected in this process to better understand how to support each other most effectively (also see [Althoff et al., 2014; Althoff and Leskovec, 2015]). We have also shown that social support can have significant impact on someone's physical activity levels and health (see [Althoff et al., 2017b; Shameli et al., 2017]).

Today, online social networks are primarily designed to maximize user engagement and advertising revenue. In contrast and addition, we should seek to understand and design supportive social networks that optimize for the well-being of their members. This could entail helping people connect with others that are likely to be encouraging, and to help them support each other more effectively, for instance through proactive interventions and conversation support tools that highlight suboptimal phrasing and make constructive suggestions. Principled design of supportive social networks needs to address challenges of: (1) how to measure

well-being by computationally operationalizing and testing psychological and sociological theories at scale, (2) understanding what drives well-being through causal inference, and (3) leveraging insights through proactive interventions involving prediction and language generation.

6.2.3 Real-world health behavior change at population scale

Computational approaches may be able to guide the design of interventions for healthy behavior change, for example helping people exercising more (Chapter 2; also see [Althoff et al., 2016b; Shameli et al., 2017]), sleeping better (Chapter 4; also see [Althoff et al., 2018]), and eating more healthily (Chapter 3). Leveraging large-scale data may allow us to answer long-standing questions in the behavioral sciences and public health. For example, to what degree is individual behavior truly individual, versus influenced by one's environment? How should we design our cities for good health? Beyond measuring behaviors, there is a great need to develop methods that motivate actual change of behavior. This often turns out to be the bigger challenge, though for instance learning optimal intervention policies personalized to each individual may be a fruitful direction.

Twenty years ago in 1998, Turing Award winner Jim Gray likened the emerging world wide web to the the discovery of a new continent [Gray, 1999]. Through analyzing the web we were able to learn a great deal about people and how they behave online. Computing has since moved from large machines and desktop computers into lightweight, wearable and ubiquitous sensors all around us. With this transformation, the discovery and exploration of "new continents" can continue. Digital traces from mobile and ubiquitous sensors now uniquely enable us to study people in their natural habitat, and to study critical dynamics in health and society. It is my hope that we will continue to leverage these signals to advance our understanding of ourselves and our environment, and to design our lives in healthy and productive ways.

Bibliography

[Aalen et al., 2008] O. Aalen, O. Borgan, and H. Gjessing. *Survival and event history analysis: A process point of view*, 2008.

[Abdullah et al., 2016] 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.

[Achermann and Borbély, 1994] 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), pp. 115–121, 1994.

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

[Adamic and Glance, 2005] L. A. Adamic and N. Glance. *The political blogosphere and the 2004 us election: divided they blog*. In *Proceedings of the 3rd international workshop on Link discovery*. 2005.

[Adar et al., 2008] E. Adar, J. Teevan, and S. T. Dumais. *Large scale analysis of web revisitation patterns*. In *SIGCHI*. 2008.

[Agichtein et al., 2006] E. Agichtein, E. Brill, and S. Dumais. *Improving web search ranking by incorporating user behavior information*. In *SIGIR*. 2006.

[Åkerstedt and Folkard, 1997] 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), pp. 115–123, 1997.

[Allison, 1978] P. D. Allison. *Measures of inequality*. Am. Sociol. Rev., 43, pp. 865–880, 1978.

[Althoff, 2017] T. Althoff. *Population-scale pervasive health*. IEEE Pervasive Computing, 16 (4), pp. 75–79, 2017.

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

[Althoff et al., 2014] T. Althoff, C. Danescu-Niculescu-Mizil, and D. Jurafsky. *How to ask for a favor: A case study on the success of altruistic requests*. In *ICWSM*. 2014.

[Althoff et al., 2018] T. Althoff, E. Horvitz, and R. W. White. *Psychomotor function measured via online activity predicts motor vehicle fatality risk*. NPJ Digital Medicine, 1 (1), p. 3, 2018.

[Althoff et al., 2017a] T. Althoff, E. Horvitz, R. W. White, and J. Zeitzer. *Harnessing the web for population-scale physiological sensing: A case study of sleep and performance*. In *WWW*. 2017a. Online Appendix: http://stanford.io/2ejFPhD.

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

[Althoff and Leskovec, 2015] T. Althoff and J. Leskovec. *Donor retention in online crowdfunding*

communities: A case study of DonorsChoose.org. In *WWW*. 2015.

[Althoff et al., 2017c] T. Althoff, R. Sosic, J. L. Hicks, A. C. King, S. L. Delp, and J. Leskovec. *Large-scale physical activity data reveal worldwide activity inequality*. Nature, 2017c.

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

[Ancoli-Israel et al., 2003] 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), pp. 342–392, 2003.

[Anderson et al., 2014] A. Anderson, R. Kumar, A. Tomkins, and S. Vassilvitskii. *The dynamics of repeat consumption*. In *WWW*. 2014.

[Anthes, 2016] E. Anthes. *Mental health: there's an app for that*. Nature, 532 (7597), pp. 20–23, 2016.

[Ashbrook and Starner, 2003] D. Ashbrook and T. Starner. *Using GPS to learn significant locations and predict movement across multiple users*. Personal and Ubiquitous computing, 2003.

[Ashok et al., 2013] V. G. Ashok, S. Feng, and Y. Choi. *Success with style: Using writing style to predict the success of novels*. In *EMNLP*. 2013.

[Atkinson, 1970] A. B. Atkinson. *On the measurement of inequality*. J. Econ. Theory, 2 (3), pp. 244–263, 1970.

[Baeza-Yates et al., 2015] R. Baeza-Yates, D. Jiang, F. Silvestri, and B. Harrison. *Predicting the next app that you are going to use*. In *WSDM*. 2015.

[Barzilay and Lee, 2004] R. Barzilay and L. Lee. *Catching the drift: Probabilistic content models, with applications to generation and summarization*. In *HLT-NAACL*. 2004.

[Basner et al., 2007] 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), pp. 1085–1095, 2007.

[Bassett et al., 2010] D. R. Bassett, H. R. Wyatt, H. Thompson, J. C. Peters, and J. O. Hill. *Pedometer-measured physical activity and health behaviors in U.S. adults.* Med. Sci. Sport. Exer., 42 (10), pp. 1819–1825, 2010.

[Bauman et al., 2012] A. E. Bauman, R. S. Reis, J. F. Sallis, J. C. Wells, R. J. Loos, B. W. Martin, L. P. A. S. W. Group, et al. *Correlates of physical activity: why are some people physically active and others not?* Lancet, 380 (9838), pp. 258–271, 2012.

[Beck, 1967] A. T. Beck. *Depression: Clinical, experimental, and theoretical aspects*, University of Pennsylvania Press, 1967.

[Benson et al., 2016] A. R. Benson, R. Kumar, and A. Tomkins. *Modeling user consumption sequences*. In *WWW*. 2016.

[Berkovsky et al., 2008] S. Berkovsky, T. Kuflik, and F. Ricci. *Mediation of user models for enhanced personalization in recommender systems*. User Modeling and User-Adapted Interaction, 2008.

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

[Blumenstock et al., 2015] J. Blumenstock, G. Cadamuro, and R. On. *Predicting poverty and wealth from mobile phone metadata*. Science, 350 (6264), pp. 1073–1076, 2015.

[Böhmer et al., 2011] 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.

[Bohnert et al., 2008] F. Bohnert, I. Zukerman, S. Berkovsky, T. Baldwin, and L. Sonenberg. *Using interest and transition models to predict visitor locations in museums*. AI Communications, 2008.

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

[Bramsen et al., 2011] P. Bramsen, M. Escobar-Molano, A. Patel, and R. Alonso. *Extracting social power relationships from natural language*. In *HLT-NAACL*. 2011.

[Broder et al., 2000] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. *Graph structure in the web*. Computer Networks, 33 (1-6), pp. 309–320, 2000.

[Brown et al., 2016] W. J. Brown, G. I. Mielke, and T. L. Kolbe-Alexander. *Gender equality in sport for improved public health*. Lancet, 388 (10051), pp. 1257–1258, 2016.

[Brysbaert et al., 2014] M. Brysbaert, A. B. Warriner, and V. Kuperman. *Concreteness ratings for 40 thousand generally known English word lemmas*. Behavior Research Methods, 46 (3), 2014.

[Bureau of Labor Statistics, American Time Use Survey, 2015] 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.

[Campbell and Pennebaker, 2003] R. S. Campbell and J. W. Pennebaker. *The secret life of pronouns: Flexibility in writing style and physical health*. Psychological Science, 14 (1), 2003.

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

[Case et al., 2015] M. A. Case, H. A. Burwick, K. G. Volpp, and M. S. Patel. *Accuracy of smartphone applications and wearable devices for tracking physical activity data*. JAMA, 313 (6), pp. 625–626, 2015.

[Centers for Disease Control and Prevention, 2012] Centers for Disease Control and Prevention. *CDC vital signs: more people walk to better health*. http://www.cdc.gov/vitalsigns/walking/, 2012. Accessed November 3, 2016.

[Chen, 2014] G. Chen. *Visualizations for Mental Health Topic Models*. Master's thesis, MIT, 2014.

[Cheng et al., 2017] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, and J. Leskovec. *Anyone can become a troll: Causes of trolling behavior in online discussions*. In *CSCW*. 2017.

[Chokshi and Farley, 2014] D. A. Chokshi and T. A. Farley. *Changing behaviors to prevent noncommunicable diseases*. Science, 345 (6202), pp. 1243–1244, 2014.

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

[Cox and Isham, 1980] D. R. Cox and V. Isham. *Point processes*, 1980.

[Danescu-Niculescu-Mizil, 2012] C. Danescu-Niculescu-Mizil. *A computational approach to linguistic style coordination*. Ph.D. thesis, Cornell University, 2012.

[Danescu-Niculescu-Mizil et al., 2012] C. Danescu-Niculescu-Mizil, L. Lee, B. Pang, and J. Kleinberg. *Echoes of power: Language effects and power differences in social interaction*. In *WWW*. 2012.

[Das Sarma et al., 2012] A. Das Sarma, S. Gollapudi, R. Panigrahy, and L. Zhajtabar, Yang. *Understanding cyclic trends in social choices*. In *WSDM*. 2012.

[Davison and Hirsh, 1998] B. D. Davison and H. Hirsh. *Predicting sequences of user actions*. In *AAAI/ICML WS on Predicting the Future*. 1998.

[De Maio, 2007] F. G. De Maio. *Income inequality measures*. J. Epidemiol. Community Health, 61 (10), pp. 849–852, 2007.

[Dijk and Czeisler, 1995] 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), pp. 3526–3538, 1995.

[Dijk et al., 1992] 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), pp. 112–117, 1992.

[Dinakar et al., 2014a] K. Dinakar, A. J. Chaney, H. Lieberman, and D. M. Blei. *Real-time topic models for crisis counseling*. In *KDD DSSG Workshop*. 2014a.

[Dinakar et al., 2015] K. Dinakar, J. Chen, H. Lieberman, R. Picard, and R. Filbin. *Mixed-initiative real-time topic modeling & visualization for crisis counseling*. In *ACM ICIUI*. 2015.

[Dinakar et al., 2014b] K. Dinakar, E. Weinstein, H. Lieberman, and R. Selman. *Stacked generalization learning to analyze teenage distress*. In *ICWSM*. 2014b.

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

[Dinges, 1995] ———. *An overview of sleepiness and accidents*. J Sleep Res, 4 (s2), pp. 4–14, 1995.

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

[Drutsa et al., 2017] A. Drutsa, G. Gusev, and P. Serdyukov. *Periodicity in user engagement with a search engine and its application to online controlled experiments*. ACM TWEB, 2017.

[Du et al., 2016] N. Du, H. Dai, R. Trivedi, U. Upadhyay, M. Gomez-Rodriguez, and L. Song. *Recurrent marked temporal point processes: Embedding event history to vector*. In *KDD*. 2016.

[Du et al., 2015a] N. Du, M. Farajtabar, A. Ahmed, A. J. Smola, and L. Song. *Dirichlet-hawkes processes with applications to clustering continuous-time document streams*. In *KDD*. 2015a.

[Du et al., 2015b] N. Du, Y. Wang, N. He, J. Sun, and L. Song. *Time-sensitive recommendation from recurrent user activities*. In *NIPS*. 2015b.

[Duncan et al., 2011] D. T. Duncan, J. Aldstadt, J. Whalen, S. J. Melly, and S. L. Gortmaker. *Validation of Walk Score® for estimating neighborhood walkability: an analysis of four US metropolitan areas*. Int. J. Environ. Res. Public Health, 8 (12), pp. 4160–4179, 2011.

[Efron and Tibshirani, 1994] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*, CRC press, 1994.

[Faloutsos et al., 1999] M. Faloutsos, P. Faloutsos, and C. Faloutsos. *On power-law relationships of the internet topology*. In *ACM SIGCOMM Computer Communication Review*, pp. 251–262. 1999.

[Farajtabar et al., 2015] M. Farajtabar, Y. Wang, M. G. Rodriguez, S. Li, H. Zha, and L. Song. *Coevolve: A joint point process model for information diffusion and network co-evolution*. In *NIPS*. 2015.

[Fernandez-Mendoza et al., 2010] 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), pp. 459–465, 2010.

[Fischer, 2001] G. Fischer. *User modeling in human–computer interaction*. UMUAI, 2001.

[Fox and Duggan, 2013] S. Fox and M. Duggan. *Tracking for health*, Pew Research Center's Internet & American Life Project, 2013. http://www.pewinternet.org/2013/01/28/tracking-for-health/.

[Freyne and Berkovsky, 2010] J. Freyne and S. Berkovsky. *Intelligent food planning: personalized recipe recommendation*. In *IUI*. 2010.

[Freyne et al., 2017] J. Freyne, J. Yin, E. Brindal, G. A. Hendrie, S. Berkovsky, and M. Noakes. *Push notifications in diet apps: Influencing engagement times and tasks*. Int J of Human–Computer Interaction, 2017.

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

[González et al., 2008] M. C. González, C. A. Hidalgo, and A.-L. Barabási. *Understanding individual human mobility patterns.* Nature, 453 (7196), pp. 779–782, 2008.

[Gorniak and Poole, 2000] P. Gorniak and D. Poole. *Predicting future user actions by observing unmodified applications*. In *AAAI*. 2000.

[Gray, 1999] J. Gray. *What Next?: A Dozen Information-Technology Research Goals*, Microsoft Technical Report MS-TR-99-50, 1999.

[Guerini et al., 2012] M. Guerini, A. Pepe, and B. Lepri. *Do linguistic style and readability of scientific abstracts affect their virality?* In *ICWSM*. 2012.

[Haberstroh et al., 2007] S. Haberstroh, T. Duffey, M. Evans, R. Gee, and H. Trepal. *The experience of online counseling*. Journal of Mental Health Counseling, 29 (3), 2007.

[Hallal et al., 2012] P. C. Hallal, L. B. Andersen, F. C. Bull, R. Guthold, W. Haskell, U. Ekelund, L. P. A. S. W. Group, et al. *Global physical activity levels: surveillance progress, pitfalls, and prospects*. Lancet, 380 (9838), pp. 247–257, 2012.

[Hawkes, 1971] A. G. Hawkes. *Spectra of some self-exciting and mutually exciting point processes*. Biometrika, 1971.

[Hekler et al., 2015] E. B. Hekler, M. P. Buman, L. Grieco, M. Rosenberger, S. J. Winter, W. Haskell, and A. C. King. *Validation of physical activity tracking via android smartphones compared to ActiGraph accelerometer: laboratory-based and free-living validation studies*. JMIR mHealth uHealth, 3 (2), p. e36, 2015.

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

[Herlocker et al., 2004] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. *Evaluating collaborative filtering recommender systems*. TOIS, 2004.

[Hochreiter and Schmidhuber, 1997] S. Hochreiter and J. Schmidhuber. *Long short-term memory*. Neural Computation, 1997.

[Howes et al., 2014] C. Howes, M. Purver, and R. McCabe. *Linguistic indicators of severity and progress in online text-based therapy for depression*. CLPsych Workshop at ACL 2014, 2014.

[Huang, 2015] R. Huang. *Language Use in Teenage Crisis Intervention And the Immediate Outcome: A Machine Automated Analysis of Large Scale Text Data*. Master's thesis, Columbia University, 2015.

[Hutto and Gilbert, 2014] C. Hutto and E. Gilbert. *Vader: A parsimonious rule-based model for sentiment analysis of social media text*. In *ICWSM*. 2014.

[Insel, 2008] T. R. Insel. *Assessing the economic costs of serious mental illness*. The American Journal of Psychiatry, 165 (6), 2008.

[Iwata et al., 2013] T. Iwata, A. Shah, and Z. Ghahramani. *Discovering latent influence in online social activities via shared cascade poisson processes*. In *KDD*. 2013.

[Juda et al., 2013] 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), pp. 141–151, 2013.

[Kalafat et al., 2007] J. Kalafat, M. Gould, J. L. H. Munfakh, and M. Kleinman. *An evaluation of crisis hotline outcomes part 1: Nonsuicidal crisis callers*. Suicide and Life-threatening Behavior,

37 (3), 2007.

[Kapoor et al., 2015] K. Kapoor, K. Subbian, J. Srivastava, and P. Schrater. *Just in time recommendations: Modeling the dynamics of boredom in activity streams*. In *WSDM*. 2015.

[Kapoor et al., 2014] K. Kapoor, M. Sun, J. Srivastava, and T. Ye. *A hazard based approach to user return time prediction*. In *KDD*. 2014.

[Kawachi and Kennedy, 1997] I. Kawachi and B. P. Kennedy. *The relationship of income inequality to mortality: does the choice of indicator matter?* Soc. Sci. Med., 45 (7), pp. 1121–1127, 1997.

[Kelley et al., 2015] 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), pp. 210–226, 2015.

[Kohl et al., 2012] H. W. Kohl, C. L. Craig, E. V. Lambert, S. Inoue, J. R. Alkandari, G. Leetongin, S. Kahlmeier, L. P. A. S. W. Group, et al. *The pandemic of physical inactivity: global action for public health*. Lancet, 380 (9838), pp. 294–305, 2012.

[Kooti et al., 2016] F. Kooti, K. Lerman, L. M. Aiello, M. Grbovic, N. Djuric, and V. Radosavljevic. *Portrait of an online shopper: Understanding and predicting consumer behavior*. In *WSDM*. 2016.

[Koren, 2009] Y. Koren. *Collaborative filtering with temporal dynamics*. In *KDD*. 2009.

[Kripke et al., 1979] 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), pp. 103–116, 1979.

[Kurashima et al., 2018] T. Kurashima, T. Althoff, and J. Leskovec. *Modeling Interdependent and Periodic Real-World Action Sequences*. In *WWW*. 2018. Online Appendix. http://bit.ly/2stjpB4.

[Labov and Fanshel, 1977] W. Labov and D. Fanshel. *Therapeutic discourse: Psychotherapy as conversation*, 1977.

[Lane, 1999] T. Lane. *Hidden markov models for human/computer interface modeling*. In *Proc. IJCAI WS on Learning about Users*. 1999.

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

[Lee et al., 2012] I.-M. Lee, E. J. Shiroma, F. Lobelo, P. Puska, S. N. Blair, P. T. Katzmarzyk, L. P. A. S. W. Group, et al. *Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy*. Lancet, 380 (9838), pp. 219–229, 2012.

[Leskovec and Horvitz, 2008] J. Leskovec and E. Horvitz. *Planetary-scale views on a large instant-messaging network*. In *WWW*. 2008.

[Levitt and Jacques, 2005] D. H. Levitt and J. D. Jacques. *Promoting tolerance for ambiguity in counselor training programs*. The Journal of Humanistic Counseling, Education and Development, 44 (1), 2005.

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

[Liu et al., 2016] Q. Liu, S. Wu, L. Wang, and T. Tan. *Predicting the next location: A recurrent model with spatial and temporal contexts*. In *AAAI*. 2016.

[Lynch et al., 2000] J. W. Lynch, G. D. Smith, G. A. Kaplan, and J. S. House. *Income inequality and mortality: importance to health of individual income, psychosocial environment, or material conditions*. Brit. Med. J., 320 (7243), p. 1200, 2000.

[Manning et al., 2008] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*, 2008.

[Mark et al., 2016] 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.

[Matchock and Mordkoff, 2009] 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), pp. 189–198, 2009.

[Mavroforakis et al., 2017] C. Mavroforakis, I. Valera, and M. G. Rodriguez. *Modeling the dynamics of online learning activity*. In *WWW*. 2017.

[Mikolov et al., 2013] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. *Distributed representations of words and phrases and their compositionality*. In *NIPS*. 2013.

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

[Mulvenna et al., 2000] M. D. Mulvenna, S. S. Anand, and A. G. Büchner. *Personalization on the net using web mining: Introduction*. CACM, 43 (8), pp. 122–125, 2000.

[Murnane et al., 2015] 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.

[Murnane et al., 2016] 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.

[Nahum-Shani et al., 2016] I. Nahum-Shani, S. N. Smith, B. J. Spring, L. M. Collins, K. Witkiewitz, A. Tewari, and S. A. Murphy. *Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support*. Ann Behav Med, 2016.

[National Institute of Mental Health, 2015] National Institute of Mental Health. *Any mental illness (AMI) among U.S. adults*. http://www.nimh.nih.gov/health/statistics/prevalence/any-mental-illness-ami-among-us-adults.shtml, 2015. Retrieved June 3, 2016.

[Niederhoffer and Pennebaker, 2002] K. G. Niederhoffer and J. W. Pennebaker. *Linguistic style matching in social interaction*. Journal of Language and Social Psychology, 21 (4), 2002.

[Ogata, 1981] Y. Ogata. *On Lewis' simulation method for point processes*. IEEE Transactions on Information Theory, 1981.

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

[Paparrizos et al., 2016] 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, pp. 737–44, 2016.

[Paul, 2012] M. J. Paul. *Mixed membership Markov models for unsupervised conversation modeling*. In *EMNLP-CoNLL*. 2012.

[Pennebaker et al., 2003] J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer. *Psychological aspects of natural language use: Our words, our selves*. Annual Review of Psychology, 54 (1), 2003.

[Pestian et al., 2012] J. P. Pestian, P. Matykiewicz, M. Linn-Gust, B. South, O. Uzuner, J. Wiebe, K. B. Cohen, J. Hurdle, and C. Brew. *Sentiment analysis of suicide notes: A shared task*. Biomedical informatics insights, 5 (Suppl. 1), 2012.

[Physical Activity Guidelines Advisory Committee, 2008] Physical Activity Guidelines Advisory Committee. *Physical Activity Guidelines Advisory Committee Report*. Department of Health and

Human Services, Washington, DC, 2008.

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

[Prince et al., 2008] S. A. Prince, K. B. Adamo, M. E. Hamel, J. Hardt, S. C. Gorber, and M. Tremblay. *A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review*. Int. J. Behav. Phys. Act., 5 (1), p. 56, 2008.

[Purcell, 2011] 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.

[Pyszczynski and Greenberg, 1987] T. Pyszczynski and J. Greenberg. *Self-regulatory perseveration and the depressive self-focusing style: A self-awareness theory of reactive depression*. Psychological Bulletin, 102 (1), 1987.

[Pyszczynski et al., 1987] T. Pyszczynski, K. Holt, and J. Greenberg. *Depression, self-focused attention, and expectancies for positive and negative future life events for self and others.* Journal of Personality and Social Psychology, 52 (5), 1987.

[Ramirez-Esparza et al., 2008] N. Ramirez-Esparza, C. Chung, E. Kacewicz, and J. W. Pennebaker. *The psychology of word use in depression forums in English and in Spanish: Testing two text analytic approaches*. In *ICWSM*. 2008.

[Reis et al., 2016] R. S. Reis, D. Salvo, D. Ogilvie, E. V. Lambert, S. Goenka, R. C. Brownson, L. P. A. S. . E. Committee, et al. *Scaling up physical activity interventions worldwide: stepping up to larger and smarter approaches to get people moving*. Lancet, 388 (10051), pp. 1337–1348, 2016.

[Ren et al., 2010] S. Ren, H. Lai, W. Tong, M. Aminzadeh, X. Hou, and S. Lai. *Nonparametric bootstrapping for hierarchical data*. Journal of Applied Statistics, 37 (9), 2010.

[Rendle et al., 2010] S. Rendle, C. Freudenthaler, and L. Schmidt-Thieme. *Factorizing personalized markov chains for next-basket recommendation*. In *WWW*. 2010.

[Ritter et al., 2010] A. Ritter, C. Cherry, and B. Dolan. *Unsupervised modeling of Twitter conversations*. In *HLT-NAACL*. 2010.

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

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

[Romero et al., 2011] D. M. Romero, B. Meeder, and J. Kleinberg. *Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter*. In *WWW*. 2011.

[Sacks and Jefferson, 1995] H. Sacks and G. Jefferson. *Lectures on conversation*, Wiley-Blackwell, 1995.

[Sallis et al., 2016a] J. F. Sallis, F. Bull, R. Guthold, G. W. Heath, S. Inoue, P. Kelly, A. L. Oyeyemi, L. G. Perez, J. Richards, P. C. Hallal, et al. *Progress in physical activity over the olympic quadrennium*. Lancet, 388 (10051), pp. 1325–1336, 2016a.

[Sallis et al., 2016b] J. F. Sallis, E. Cerin, T. L. Conway, M. A. Adams, L. D. Frank, M. Pratt, D. Salvo, J. Schipperijn, G. Smith, K. L. Cain, et al. *Physical activity in relation to urban environments in 14 cities worldwide: a cross-sectional study*. Lancet, 387 (15), pp. 2207–2217, 2016b.

[Servick, 2015] K. Servick. *Mind the phone*. Science, 350 (6266), pp. 1306–1309, 2015.

[Shameli et al., 2017] 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.

[Steiger, 1980] J. H. Steiger. *Tests for comparing elements of a correlation matrix.* Psychol. Bull., 87 (2), pp. 245–251, 1980.

[Sternberg et al., 2013] 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, p. 292, 2013.

[Stolcke et al., 2000] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer. *Dialogue act modeling for automatic tagging and recognition of conversational speech*. Computational Linguistics, 26 (3), 2000.

[Stuart, 2010] E. A. Stuart. *Matching methods for causal inference: a review and a look forward*. Stat. Sci., 25 (1), pp. 1–21, 2010.

[Swan, 2013] M. Swan. *The quantified self: Fundamental disruption in big data science and biological discovery*. Big Data, 1 (2), pp. 85–99, 2013.

[Tanaka et al., 2016] Y. Tanaka, T. Kurashima, Y. Fujiwara, T. Iwata, and H. Sawada. *Inferring latent triggers of purchases with consideration of social effects and media advertisements*. In *WSDM*. 2016.

[Tausczik and Pennebaker, 2010] Y. R. Tausczik and J. W. Pennebaker. *The psychological meaning of words: LIWC and computerized text analysis methods*. Journal of Language and Social Psychology, 29 (1), 2010.

[Teevan et al., 2006] J. Teevan, E. Adar, R. Jones, and M. Potts. *History repeats itself: Repeat queries in Yahoo's logs*. In *SIGIR*. 2006.

[Thomas and Bond, 2015] J. G. Thomas and D. S. Bond. *Behavioral response to a just-in-time adaptive intervention (JITAI) to reduce sedentary behavior in obese adults: Implications for JITAI optimization.* Health Psychology, 2015.

[Troiano et al., 2008] R. P. Troiano, D. Berrigan, K. W. Dodd, L. C. Mâsse, T. Tilert, and M. McDowell. *Physical activity in the United States measured by accelerometer.* Med. Sci. Sport. Exer., 40 (1), pp. 181–188, 2008.

[Trouleau et al., 2016] W. Trouleau, A. Ashkan, W. Ding, and B. Eriksson. *Just one more: Modeling binge watching behavior*. In *KDD*. 2016.

[Tucker et al., 2011] J. M. Tucker, G. J. Welk, and N. K. Beyler. *Physical activity in us adults: compliance with the physical activity guidelines for americans*. American Journal of Preventive Medicine, 40 (4), pp. 454–461, 2011.

[Tudor-Locke et al., 2008] C. Tudor-Locke, Y. Hatano, R. P. Pangrazi, and M. Kang. *Revisiting "how many steps are enough?"*. Med. Sci. Sport. Exer., 40 (Supplement), pp. S537–S543, 2008.

[Tudor-Locke et al., 2009] C. Tudor-Locke, W. D. Johnson, and P. T. Katzmarzyk. *Accelerometer-determined steps per day in US adults*. Med. Sci. Sport. Exer., 41 (7), pp. 1384–1391, 2009.

[UN Secretary General, 2011] UN Secretary General. *Prevention and control of non-communicable diseases*. http://www.who.int/nmh/publications/2011-report-of-SG-to-UNGA.pdf, 2011. Accessed April 21, 2016.

[United States Census Bureau, a] United States Census Bureau. *2010 Census and American Community Survey 2006-2010. Accessed through Bay Area Census.* http://www.bayareacensus.ca.gov/cities/cities.htm, a. Accessed July 5, 2016.

[United States Census Bureau, b] ———. *American Community Survey*. http://www.census.gov/programs-surveys/acs/, b. Accessed October 5, 2016.

[Valera and Gomez-Rodriguez, 2015] I. Valera and M. Gomez-Rodriguez. *Modeling adoption and*

usage of competing products. In *ICDM*. 2015.

[Van Dijk, 1997] T. Van Dijk. *Discourse studies: A multidisciplinary approach*, SAGE, 1997.

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

[Van Dyck et al., 2015] D. Van Dyck, E. Cerin, I. De Bourdeaudhuij, E. Hinckson, R. S. Reis, R. Davey, O. L. Sarmiento, J. Mitas, J. Troelsen, D. MacFarlane, D. Salvo, I. Aguinaga-Ontoso, N. Owen, K. L. Cain, and J. F. Sallis. *International study of objectively measured physical activity and sedentary time with body mass index and obesity: IPEN adult study*. Int. J. Obes., 39 (2), pp. 199–207, 2015.

[Vizer et al., 2009] 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), pp. 870–886, 2009.

[Wagstaff and Van Doorslaer, 2000] A. Wagstaff and E. Van Doorslaer. *Income inequality and health: what does the literature tell us?* Annu. Rev. Publ. Health, 21 (1), pp. 543–567, 2000.

[Walch et al., 2016] O. J. Walch, A. Cochran, and D. B. Forger. *A global quantification of "normal" sleep schedules using smartphone data*. Sci Adv, 2 (5), 2016. arXiv:http://advances.sciencemag.org/content/2/5/e1501705.full.pdf, doi:10.1126/sciadv.1501705.

[Walk Score, 2016] Walk Score. *Walk Score*. https://www.walkscore.com/cities-and-neighborhoods/, 2016. Accessed May 17, 2016.

[Watts and Strogatz, 1998] D. J. Watts and S. H. Strogatz. *Collective dynamics of âĂŸsmall-worldâĂŹnetworks*. Nature, 393 (6684), p. 440, 1998.

[Wesolowski et al., 2012] A. Wesolowski, N. Eagle, A. J. Tatem, D. L. Smith, A. M. Noor, R. W. Snow, and C. O. Buckee. *Quantifying the impact of human mobility on malaria*. Science, 338 (6104), pp. 267–270, 2012.

[West et al., 2013] 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.

[White et al., 2016] 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, pp. 42–48, 2016.

[WHO, 2010] WHO. *Global recommendations on physical activity for health*, World Health Organization, Geneva, 2010.

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

[World Bank, a] World Bank. *World Bank Country and Lending Groups*. https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups, a. Accessed October 5, 2016.

[World Bank, b] ———. *World Bank: Population, female (% of total)*. http://data.worldbank.org/indicator/SP.POP.TOTL.FE.ZS, b. Accessed May 10, 2016.

[World Health Organization, a] World Health Organization. *Obesity (body mass index $\geq$ 30) (age-standardized estimate): data by country*. http://apps.who.int/gho/data/node.main.A900A?lang=en, a. Accessed May 19, 2016.

[World Health Organization, b] ———. *Prevalence of insufficient physical activity among adults: data by country*. http://apps.who.int/gho/data/node.main.A893?lang=en, b. Accessed May 19, 2016.

[World Health Organization, 2002] ———. *The world health report 2002: reducing risks, promoting healthy life*, World Health Organization, 2002.

[World Health Organization, 2015] ———. *Depression: Fact sheet no 369*. http://www.who.int/mediacentre/factsheets/fs369/en/, 2015. Retrieved November 2, 2015.

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

[Yang et al., 2014] J. Yang, J. McAuley, J. Leskovec, P. LePendu, and N. Shah. *Finding progression stages in time-evolving event sequences*. In *WWW*. 2014.

[Yu et al., 2016] R. Yu, A. Gelfand, S. Rajan, C. Shahabi, and Y. Liu. *Geographic segmentation via latent poisson factor model*. In *WSDM*. 2016.

[Zhou et al., 2013] K. Zhou, H. Zha, and L. Song. *Learning triggering kernels for multi-dimensional hawkes processes*. In *ICML*, pp. 1301–1309. 2013.

[Zukerman and Albrecht, 2001] I. Zukerman and D. W. Albrecht. *Predictive statistical models for user modeling*. User Modeling and User-Adapted Interaction, 2001.

[Zukerman et al., 1999] I. Zukerman, D. W. Albrecht, and A. E. Nicholson. *Predicting users' requests on the WWW*. In *UM*. 1999.