TIM ALTHOFF RESEARCH STATEMENT

My research focuses on developing **data science methods to improve human well-being.** The recent popularity of mobile devices, including smartphones and smartwatches, has generated an explosion of detailed behavioral data. This provides us with an unparalleled opportunity to realize new types of scientific approaches that provide actionable insights about our lives, health, and happiness. In pursuit of this opportunity, I leverage this data at the scale of billions of actions taken by millions of people, and develop novel techniques combining **data science, social network analysis,** and **natural language processing (NLP)**.

For the first time in scientific history, smartphones and smartwatches enable large-scale, objective measurements of human behaviors. Many people now spend most of their time with these sensor-loaded devices to track their behavior. Massive, digital traces from these devices can help us better understand a wide range of dynamic human activity including exercise, sleep, diet, and social interactions. They have the potential to **close an important scientific gap:** Earlier research had been restricted to exclusively-online behaviors, or to small-scale studies in the behavioral, social, and medical sciences. However, while there is great promise in **mining massive, digital traces for physical, mental, and social well-being**, we need to develop new methods to gain actionable insights from this data. My research aims to address this need.

Developing computational methods for large-scale **sensor and social interaction data** faces **fundamental challenges**: (1) 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; (2) Significant domain knowledge in 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; (3) Sensor and social interaction data are observational (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 correlation). My research develops **rigorous, computational methods** that address these challenges and enable practical applications of high impact.

My research advances computational methods that extract meaning from raw sensor and social interaction data by operationalizing scientific domain knowledge (Fig. 1). I then leverage these methods together with large-scale data to conduct massive observational studies. My approach uniquely advances computational methods and scientific knowledge to better understand and improve human well-being. My work is highly interdisciplinary and has made **contributions to both computer science and scientific application areas of medicine and psychology**. This is exemplified by publications in computer science conferences and journals *(WWW, WSDM, KDD, ICWSM, TACL)*, interdisciplinary ***(Nature)*** and medical journals *(J Medical Internet Research, NPJ Digital Medicine)*, and a Clinical NLP **Best Paper Award** from the International Medical Informatics Association.

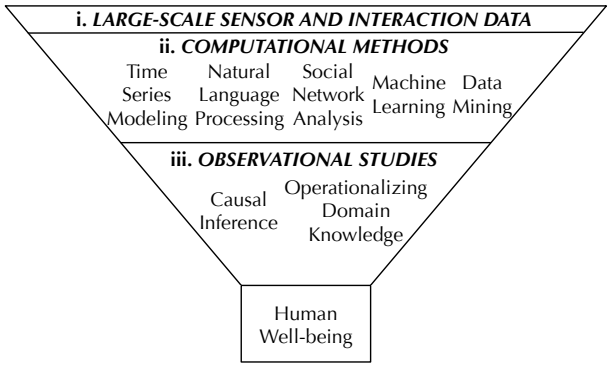

Figure 1: My research develops *computational methods* for *large-scale sensor and interaction data*, which enable *observational studies* that advance our scientific understanding of human well-being and generate actionable insights.

MODELING AND PREDICTING HEALTH BEHAVIORS

Our everyday behaviors, including physical activity, sleep, and diet, are critical to our health. Yet, we know very little about these factors. Time series of real-world behaviors captured through smartphones and smartwatches are complex. They vary based on personal habits and preferences, but also based on biological and seasonal factors. New methods are required to model time series of real-world behaviors

and account for these factors (Fig. 1ii). To fill this gap, I have developed models that can learn personal habits and biological factors from massive datasets and generate practical insights, as described below.

Learning to predict future health behaviors. Predicting activities in advance, such as the time of one's next meal or run, enables well-timed health interventions including diet reminders. However, this is challenging due to the many complexities of human behavior, including its time-varying, interdependent, and periodic nature. I developed a statistical machine learning model based on **temporal point and Hawkes' processes** that addresses these complexities, outperforming existing state-of-the-art methods (Fig. 1ii). Through solving an optimization problem with over 200 million variables, the model learned temporal dependencies between activities in an unsupervised way. For example, it automatically learned that while lunches are typically eaten around noon, having earlier breakfasts tend to be followed by earlier lunches as well. Learning such personalized routines and commonsense knowledge from data is critical in order to precisely target interventions, such as making sure diet reminders do not come too late. We are currently **working with Under Armour to deploy these models** and improve their in-app activity recommendations for millions of users.

Enabling new types of health insights from planetary-scale data. Leveraging the popularity of smartphone devices, I was able to conduct **the largest-ever study of human physical activity** analyzing hundreds of billions of steps by one million people across 111 countries worldwide (Fig. 1i). Pioneering this **paradigm shift** from small-scale self-reports to planetary-scale, objective measurements allowed me to quantify the distribution of physical activity beyond simple averages. This has led to a better scientific understanding of human activity. I quantified a new health inequality measure based on inequality of physical activity, which had never been described before, as even the largest studies of physical activity had been 1000 times smaller and unable to compare populations globally. I demonstrated that in some countries most people walked a similar amount, but in others there were big gaps between people who walked a lot and those who walked very little. We showed that the size of this gap, called activity inequality, was highly predictive of country-level obesity (Fig. 2). Designing several dozen reweighting, stratification, and simulation experiments to test for various confounders, I showed that this finding was robust (Fig. 1iii). Further simulation experiments demonstrated that inequality-focused public health interventions may be up to four times more effective than unfocused, population-wide interventions. To that end, I showed that activity inequality could be reduced through creating more walkable cities, and that our smartphone-based methodology could help **target health interventions and design healthier cities.** This work was the first to deliver on the promise of big sensor data to reveal patterns of worldwide activity and obesity and was published in *Nature* [15][1], as well as featured in over **150 news articles worldwide** including in *The New York Times* and *CNN*.

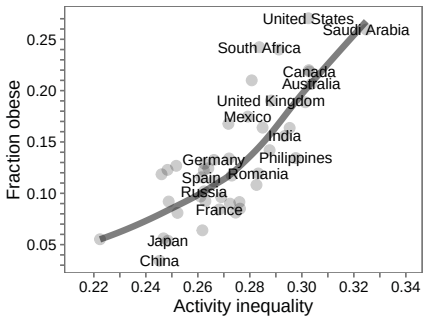

Figure 2: I revealed a new health inequality based on activity which predicts obesity prevalence worldwide [Nature, 2017].

Smartwatch data reveals that user behavior is impacted by fundamental biological processes. I conducted the first-ever study combining smartwatch activity data with large-scale web search data (Fig. 1i). Combining data on 3 million nights of sleep with 75 million search engine interactions revealed that web search behavior is significantly impacted by fundamental biological processes. Notably, I showed that modeling how biology impacts user behavior enables a new way of studying biological processes, that overcomes critical limitations of existing sleep research. Specifically, I developed a novel, non-intrusive approach to measure cognitive performance through the speed of everyday interactions with search engines such as keystrokes and clicks on search results [10]. This revealed that our ability to type and click on search results is not constant throughout the day, but fluctuates by 30% governed by biological processes. Previously, sleep research was fundamentally limited in its ability to scale due to its intrusive methods to measure cognitive performance. My approach enabled the **largest-ever study of sleep and cognitive performance,** and

[1] Citation numbers reference the numbered publication list in my CV.

has shed light on how our everyday cognitive abilities are affected by complex, real-world sleep schedules and deprivation, and our appreciation of coffee. Developing **non-parametric, statistical models encoding biological domain knowledge** allowed me to disentangle the complex biological processes governing sleep and cognitive performance and inform biological theory (Fig. 1iii). Applying my method to 16 billion keystroke measurements across the entire US, I demonstrated that search engine-based estimates of cognitive performance could even predict population-level accident risk [14]. In the future, this technology could help alert sleepy or drunk drivers **potentially avoiding car accidents and associated fatalities.**

NLP METHODS TO GAIN INSIGHTS INTO MENTAL HEALTH

My research advances computational methods that leverage large corpora of text-message-based conversations between counselors and their clients to better understand **how to support people in illness and crisis.** This addresses a pressing societal challenge. Worldwide, 450 million people suffer from mental illness, and 800,000 people die due to suicide every year, which is one person every 40 seconds. Still, mental health care is riddled with substantial gaps in detection, access to care, and treatment. To address these gaps, I have established **partnerships with Crisis Text Line and Talkspace**, US-nationwide counseling and therapy providers, giving us access to millions of anonymized counseling transcripts (Fig. 1i). This data provided me with a window into typically unquantified mental health services and counseling strategies. For instance, while it is well-established that some counselors are much more effective than others in helping people feel better, we had very little understanding of why that is. The computational challenge is to extract actionable insights from unstructured conversations. By developing novel NLP methods (Fig. 1ii), including **sequence-based conversation models**, phrase clustering, and language model comparisons, I was able to reveal important, previously unknown, differences between the most and least successful counselors [11]. My methods were the first to quantify whether counselors are adapting their style to the conversation, whether they make effective progress, and to what degree they use generic, templated language. I showed that the best counselors used less templated, but more creative, personalized language. While this work was also awarded a **Best Paper Prize** by the International Medical Informatics Association, what drives and excites me most is the exceptional opportunity to develop data science methods to improve mental health services and people's lives. For instance, my finding on templated language use was confirmed by qualitative, follow-up research and has **concretely impacted and improved counselor training** at Crisis Text Line.

CAUSAL INFERENCE IN TEMPORAL SOCIAL NETWORK DATA

Studying causal mechanisms of social influence in observational data is notoriously challenging, but important as it helps us understand why specific events happen. I developed methods to address this challenge, which have allowed me to answer the open question of whether *online* social networks would impact *offline* health behaviors (Fig. 1ii). This is significant because it tells us whether we could leverage online social support to help people be more physically active and healthy. Leveraging 800 million actions by 6 million users in a mobile activity tracking app (Fig. 1i), I conducted matching-based difference-in-difference studies, and demonstrated that users significantly increased their offline physical activity over several months after they joined an exercise-related online social network [8,9]. However, these increases in activity could potentially be due to a user's intrinsic motivation, causing both the social activity as well as behavioral changes. In order to establish the causal effect of **social influence in online social networks on offline health behaviors,** it was necessary to control for the unobserved factor of intrinsic motivation. Overcoming limitations of previous research, I developed a **novel causal inference method** leveraging the delay between requesting to become friends and the acceptance of this friendship request (Fig. 1iii). This method led to a **natural experiment,** allowing me to control for latent intrinsic motivation and establish the causal relationship [8]. Advancing causal inference methods in this domain not only improves our scientific understanding of health behavior change, but also helps **improve the design of the underlying computing applications** [13].

RESEARCH AGENDA

My goal is to create **scalable computational techniques** to measure, understand, predict, and enhance **human behavior and well-being**. To continue pursuing this agenda, I am excited to investigate three problems at the nexus of computational methods for human well-being.

Developing data science tools for large-scale and high-dimensional observational data. Due to the large cost and limited scope of randomized controlled trials, scientific advances will increasingly be based on large-scale observational studies (Fig. 1iii). However, these advances are contingent on democratized ability to conduct observational studies that meet high standards of validity. Therefore, we need tools that enable high-quality observational science. My computational roots and interdisciplinary research background place me in a unique position to develop generic methods that **turn observational data into robust inferences,** as evidenced by work on causal inference methods [1,8] and publications in interdisciplinary journals including *Nature* [15]. I want to turn the methods I develop into **open-source data science tools**, that 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.

Supportive online social networks. I aim to understand and design *supportive* social networks that optimize for the well-being of their members. This entails 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. In contrast, today's online social networks are primarily designed to maximize user engagement and advertising revenue. 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** (Fig. 1iii), (2) understanding what drives well-being through **causal inference** (Fig. 1iii), and (3) leveraging insights through proactive interventions involving **prediction and language generation** (Fig. 1ii). My research has started addressing these challenges by studying how online interactions affect physical and mental health [8,11], and operationalizing theories of pro-social behavior [5,6], social influence [8], cognitive performance [10], and depression [11].

Real-world health behavior change at population scale. My objective is to develop computational methods that can guide the design of interventions for healthy behavior change, for example exercising more or eating better (Fig. 1ii). My approach will allow me 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? I have forged **data sharing partnerships** with Azumio and Under Armour giving access to millions of people's behaviors (Fig. 1i), enabling me to develop novel approaches to address these questions. For instance, I currently leverage users moving from one location to another to quantify the effect of one's environment on individual behavior in a US-nationwide study. Beyond measuring behaviors, I am interested in developing methods to motivate behavior *change*, for instance by **learning optimal intervention policies.** Towards this end, I have started a collaboration with behavioral scientists to deliver personalized alerts through smartwatches. The interdisciplinary applications of my research enable me to fund my research program through a variety of sources including medical (*e.g.*, NIH, NIMH), science and technology (*e.g.*, NSF, DARPA), philanthropic, and industry organizations. Throughout my PhD I have gained significant **grant writing experience** in these domains.

In the longer-term, I intend to continue focusing on **socially-relevant problems** that can only be addressed by drawing upon both sophisticated computational methods and relevant domain expertise (Fig. 1). I have been fortunate to collaborate with researchers from diverse fields and am convinced that by pursuing these types of collaborations in my future research, I will be able to address hard and impactful research problems that stand at the intersection of data science, computational modeling, and society.

Citation numbers reference the numbered publication list in my CV.