Thesis Portfolio

Adaptive Mobile Sensing: Leveraging Machine Learning for Efficient Human Behavior Modeling

(Technical Report)

Instagram, Amazon, and Machine Learning: Ethical Implications of Collecting and Analyzing Commercial User Data

(STS Research Paper)

An Undergraduate Thesis

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Tucker Wilson
Spring, 2020

Department of Systems and Information Engineering

Table of Contents

Sociotechnical Synthesis

Artificial intelligence and machine learning are perhaps the most rapidly expanding field of computer science today, and one of the most ethically impactful applications of these technologies is on personal user data. Most internet-based technologies generate large amounts of user data by logging our online activities, searches, views, and interactions. The amount of data we generate is multiplied even further when considering the popularity of the smartphone or other smart technologies, such as the Amazon Echo, which have the capability to collect location, motion, health, photo, and audio data. As users generate this data on themselves, companies are rushing to use their databases to create insights. Both the technical capstone and STS research paper in this portfolio explore potential applications of user-generated data; the technical capstone develops machine learning models on user-generated health data to predict illness in an efficient way; the STS research paper explores how user-generated data on social media sites and through smart devices may expose users to security and privacy risks.

The technical project is a part of ongoing research conducted for the Defense Advanced Research Projects Agency (DARPA) to design and develop reliable disease detection analytics through data collected from smartphones. The ultimate goal is to design a disease detection system to be deployed for military personnel stationed in combat zones, but the immediate focus of the technical project is to design efficient methods of data collection without a reduction in data richness. The primary research consists of a three-week study, where each week users run a different data collection strategy on their smartphones. Prior to this study, different methods data collection methods were developed, including an adaptive sensing model which pings all smartphone sensors periodically and then turns on those returning non-zero data and a machine learning model that listens to the phone's accelerometer for a small interval and then predicts

whether the phone is in use, turning sensors off/on accordingly. The motivation behind these methods is that data collection from sensors is incredibly battery-intensive, and so minimizing the time spent collecting while still maintaining a rich dataset makes for a much more efficient system. While the study is currently still underway, initial results indicate that both the adaptive sensing model and the machine learning model have improved battery usage over an "always-on" collection strategy, and the machine learning model, in particular, is able to prioritize when the phone is actually in use, and thus lose very little important data in the process.

The STS research paper explores how user-generated data from mobile and smart devices is collected and analyzed for the purposes of online advertising. Companies like Facebook, Amazon, Apple, and Google all collect user data through social media sites and devices and then use this data to make predictions on a user's preferences and traits. These traits can range from the benign, such as food and drink preferences or preferred clothing styles, to the potentially sensitive, such as political leanings, sexuality, or even predisposition to mental illness. Simultaneously, users are generally unaware of the potentially sensitive insights this data can generate, as well as what other companies, organizations, or individuals have access to this data, either legally or through security breaches. Though this paper does consider user-generated data generally, it focuses on two specific use cases: Instagram and the Amazon Echo, also called Alexa. Both of these technologies collect and store user-generated data and heavily employ that data to gather insights about users. By performing documentary and case study research with the backing STS theories of Actor-Network Theory and Technological Momentum, this paper has uncovered many risks to users, including the risk of breach of privacy by bad actors, harassment based on their characteristics determined through machine learning algorithms, and exposure to discriminatory policies from private companies or governments. There are solutions, regulatory,

by companies, and by users, but they are piecemeal, and drastic change would need to happen for users to be truly protected from these risks.

As the main designer of the machine learning model for my technical capstone project, I have had a rich experience working on the two simultaneously. The capstone project provided me an excellent technical background for the research, as I came in with a rich understanding of the theory behind machine learning, the process of building and training of machine learning models, and the types of insights machine learning can generate. Conversely, the STS research paper provided me with a way of thinking that prioritized the privacy and safety of the users. We had user data completely anonymized both by removing names, ages, sex, race, etc. and by cleansing geographic data by making it relative (geographic distance from a certain point) rather than absolute (longitude and latitude). In this way, we were able to protect individual users from the risk of their data being linked to themselves. Additionally, we are currently including sections in our capstone paper on the importance of privacy and strong data security when working with user-generated data, something that might not have happened if I had not learned so much about that data's potential risks. Overall, I have gained a rich level of both technical and impact-focused knowledge on machine learning.