PREDICTIVE MODELING FOR FLOOD PREVENTION

DATA COLLECTION: BENEFICIAL OR HARMFUL?

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Max Dawkins

November 1, 2021

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS

Catherine D. Baritaud, Department of Engineering and Society

Daniel G. Graham, Rosanne Vrugtman, Department of Computer Science

Flooding is the most common natural disaster, accounting for 40% of all natural disasters globally (Torti, 2012, p. 1). Flooding has become both more frequent and more unpredictable due to climate change, leading to lack of preparation for floods in several countries (Chang, 2011, p. 672). Recently, New York experienced a disastrous flooding event during the appearance of Hurricane Ida (Barron, 2021, p. 1). Thirteen deaths occurred due to the lack of preparation for the flood. James Barron reports that although New York was not prepared enough for this flood, they have been improving their protections against flooding events. The Department of Environmental Protection spends millions of dollars each year to combat flash flooding disasters. However, most of their spending goes to physical protections such as widening grates, and building reservoirs to hold storm water (Barron, 2021, p. 1). Alternative protective measures and the focus of the proposed technical research are predictive systems. Predictive systems such as VinAWARE in Vietnam are used to predict when floods will occur and warn the public as early as possible (Torti, 2012, p. 3). Many of these predictive systems involve the use of Artificial Intelligence (AI) models that require large amounts of data to operate. The loosely coupled topic of STS research focuses on the collection of data for commercial and public use and seeks to understand why data collection causes problems.

The technical research will follow the timeline depicted in Figure 1 under the guidance of Daniel G. Graham, Assistant Professor at the University of Virginia in Computer Science. The research will begin in January 2022 and finish by May 2022. The STS research will follow the same timeline as the technical research under the guidance of Catherine D. Baritaud, Professor at the University of Virginia in Science, Technology, and Society.

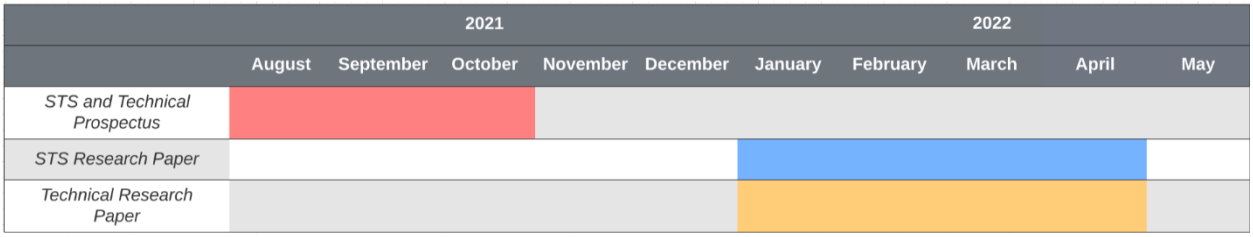

Figure 1: Research Timeline. This figure shows the timeline for the proposed STS and technical research. (Dawkins, 2021).

PREDICTIVE MODELING FOR FLOOD PREVENTION

Through analysis of various publications such as academic journals and newspapers, the proposed research will examine the current technology used globally in both flood prediction and flood protection. The frequency of floods comes hand in hand with their danger, having affected almost 10 million people in Southeast Asia and causing 1300 deaths in 2011 alone (Torti, 2012, p. 1). Furthermore, Jacqueline Torti (2012) explains that floods have effects on health even before and after the actual flood:

> There is an increased rate of mild injuries before the onset of a flood when people are advised to move their families to a safe location… During the actual onset of the flood there is potential for direct causes of injury and death… After the onset of a flood, there is potential for an increase in communicable diseases and further injury. (p. 1)

Also, the health risks of floods are not just physical injury. Risk of communicable diseases increases after flooding due to unsanitary conditions, mental health issues are known to increase following flood disasters from overwhelming emotional trauma, and malnutrition rises as food is lost and delivery of resources becomes difficult (Torti, 2012, p. 2; Ohl & Tapsell, 2000, p. 1).

Predicting when floods are about to happen and warning the public in an organized manner is very important. Technology has been developed along these lines and is still being

further developed to minimize harmful flood effects. For example, China's flash flood warning system has seen recent growth as reported by Liu et al. (2018):

> China's flash flood prevention has been progressively developed from non-existent to overall scientific decision-making support, from empirical methods to early-warning index systems, and will be improved from the monitoring and early-warning platform at the county level to information sharing and construction at the central, provincial, municipal, and county levels. (p. 620)

This development came with a decrease in the percentage of harmful flash floods over that last several decades. As shown in Figure 2 below, the percentage of floods that caused casualties has been on a decreasing trend since 1950. However, the total number flash floods has increased at the same time. Outside of just China, there have been global efforts to improve forecasting of flash floods as well. Algorithmic systems are a large part of the developments made in flood forecasting. A database on flash floods in the United States is widely available to use in predicting flood patterns (Liu et al., 2018, p. 622). Anselin Local Moran's I is a statistics-based

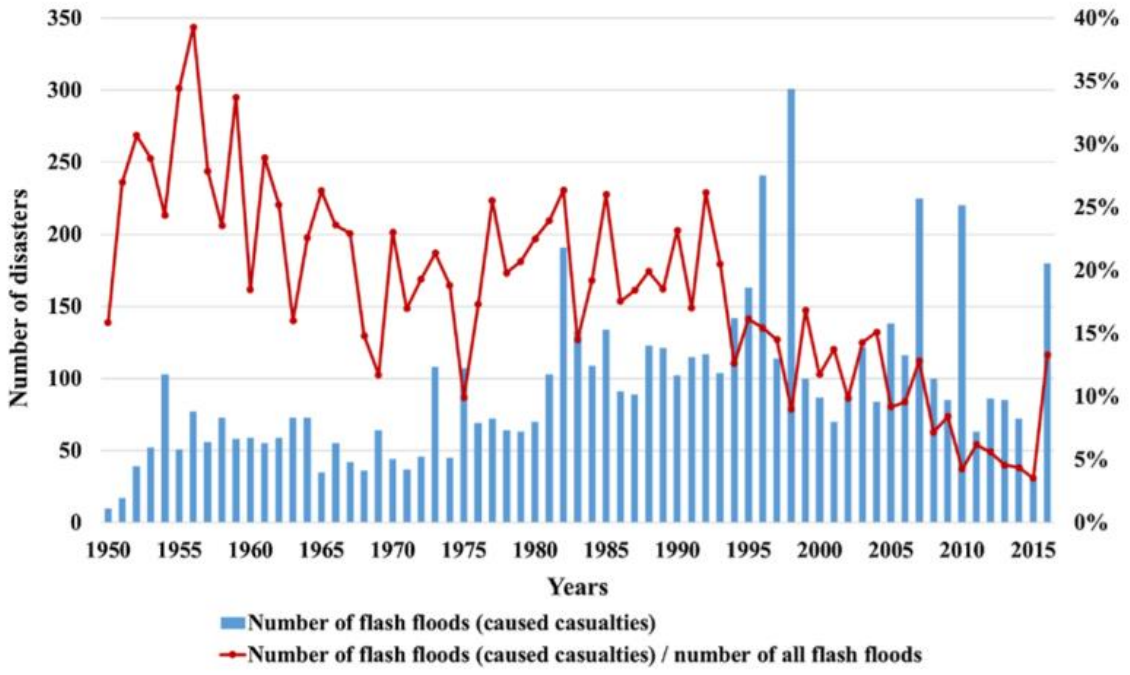

Figure 2: China flood statistics. This figure shows the percentage of disastrous floods and the total number of floods in China from 1950 to 2015. (Liu et al., 2018).

algorithmic tool that has been used to validate the quality of flooding data (Liu et al., 2018, p. 622). Where there is data, algorithmic tools will be used to make predictions, and flooding is no exception.

Despite the improvements in flood prediction systems, flood preparedness remains a problem around the world. Predicting floods is still quite difficult, especially with the changing climate, and governments have historically struggled to predict and prepare for large floods (Chang, 2011, p. 669).

The objective of the technical project is to delve into the state-of-the-art in flood prediction technology. AI is becoming increasingly popular, with $50 billion in corporate spending on AI in 2020 in the United States (Jeans, 2020, p. 1). With its increase in popularity, AI and machine learning (ML) are likely being used more than ever for predicting floods around the world. Through the study of academic journals, newspapers, and other publications, an understanding will be found of how AI is used in flood preparedness and why it is now being used.

DATA COLLECTION: BENEFICIAL OR HARMFUL?

Data is powerful, which is why it is constantly being collected. The amount of data being collected has only been growing since the creation of the internet, with internet usage growing by almost 10% of the global population in a single year in 2018 (Martin, 2019, p. 1). Nicole Martin (2019) reports a shocking statistic,

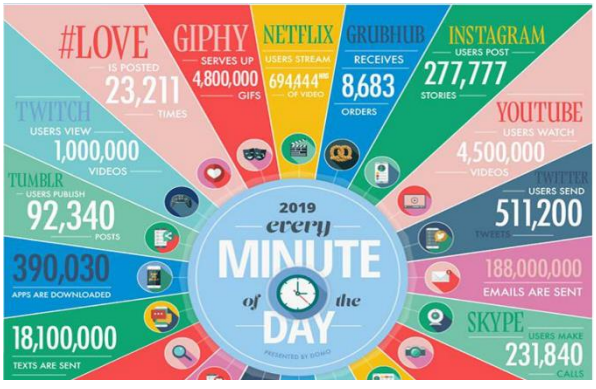

Figure 3: Data Collected per Minute. This figure shows the amount of data collected in the United States every minute by category. (Martin, 2019).

depicted in Figure 3, that "Americans use 4,416,720 GB of internet data including 188,000,000 emails, 18,100,000 texts and 4,497,420 Google searches every single minute" (p. 1).

However, the growth of data infrastructure does not come exclusively from tracking internet activity. Data is constantly being collected offline and being uploaded, including flood data. For example, in Ireland the Office of Public Works collects flood data through uploaded photographs, videos, and surveys by local government staff (Office of Public Works). Flood levels, times, sources, and magnitudes are all collected and added to online databases to be used in predictive modeling and data analysis.

With the growing amount of available data in the world, data-based algorithms are becoming more powerful and more commonplace. AI models grow more powerful as more data is available, so as much data as possible should be collected, right? Well, it is not that simple. Much of the data currently being collected is personal information like emails, texts, and Google searches. As seen in Figure 4 below, over "70 percent [of Americans] believe that most or all of what they do online is being tracked, and nearly that many believe the same is true of what they

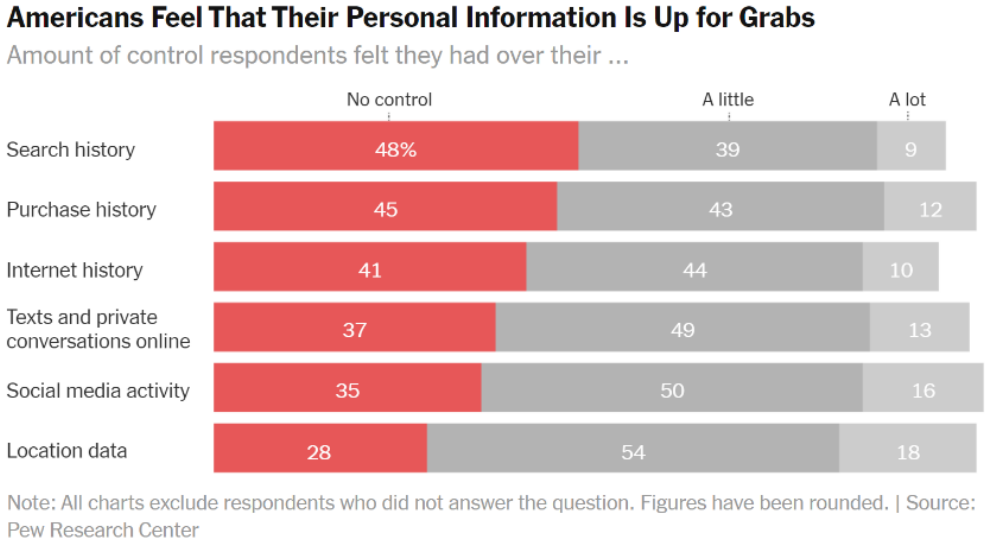

Figure 4: Data Collection Survey. This figure shows the results of a survey asking how Americans feel about the tracking of their online actions. (Manjoo, 2019).

do offline. And more than 80 percent feel like they have very little or no control over the data being collected about them" (Manjoo, 2019, p. 1).

There are certainly concerns about the collection of personal data but considering the use of the data is also very important. For example, employee data is collected and used all the time in algorithmic management which has had influences on power and social structures within organizations that use it (Jarrahi et al., 2021, p. 3). The algorithmic systems that rely on the data have become increasingly important to the point where new skillsets with focus on managers' ability to make use of such systems are preferred in management (Jarrahi et al., 2021, p. 5). Also, it is important to realize that attitudes towards AI are not clear. In fact, Vegard Kolbjørnsrud and Richard Amico (2017) show that there are vast differences in trust of AI in the workplace across different countries and depending on ranking within organizations (p. 39).

WHAT IS WRONG WITH COLLECTING PERSONAL DATA?

Clearly many people are not comfortable with their personal data being collected, but the data is also useful to many of the convenient services available to them. Furthermore, the invasiveness of data collection falls on a spectrum. For example, flood and climate data are being collected with seemingly no harm to the privacy of individuals, while tracking of emails, texts, and search histories cause concerns among the general public. An important question to ask is then: What part of that spectrum should be considered too invasive? Perhaps we should simply stop collecting all personal data, but do the benefits of having personal data outweigh the costs? If we stop collecting personal data, where do we draw the line on what data is okay to collect? There are many questions that still need to be answered about data collection through further research.

The objective of the STS research is to understand the interactions in the network of data collection and answer the questions posed above. To do so, the STS research will examine these questions through the eyes of Actor-Network Theory (ANT), a framework developed by Michel Callon, Madeleine Akrich, Bruno Latour, and John Law in the early 1980s. The problem of data collection lies within the context of a network within which all actors have varying levels of agency. The actors in the network include large tech companies, consumers whose data is collected, the government, engineers that use the data in models, the data itself, and methods of data collection, as depicted in Figure 5 below. In this network, agency is dispersed because of the

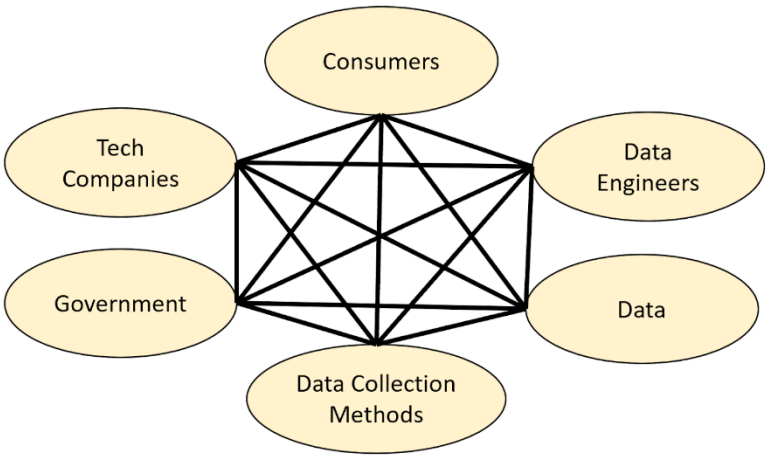

Figure 5: Actor-Network Theory Diagram. This figure displays the fully connected network of actors surrounding data collection. (Dawkins, 2021).

relationships of the network (Hurtado-de-Mendoza et al., 2015, p. 330). Looking at the relationships within the network provides insight into the incentives and concerns of each actor. Understanding these things will help answer the questions posed above and pinpoint the root of the problems with personal data collection. The results of this STS research will be presented in a scholarly article attempting to answer the question "Why is personal data collection problematic?".

THE FUTURE OF DATA-BASED TECHNOLOGY

With AI and ML gaining popularity, data-based technologies are becoming more prevalent in the world than ever before. Their rapid growth necessitates study of the social implications of the data collection required for these technologies. This research hopes to explore the social impacts of data collection and arrive at the underlying problems in it that need to be fixed. It is already known that people have uneasy feelings about their personal data being collected. Understanding the problems that cause these feelings is a crucial part of keeping AI and ML technologies from leading the world into a future of distrust between people and technology. In the context of the technical research, it is important for flood prediction technology to avoid losing people's trust. Guidelines for data collection can help flood prediction technologies avoid this. Although the proposed STS research will not provide any guidelines for data collection, it is a first step towards developing a protocol that could save the future of data-based technology.

REFERENCES

Barron, J. (2021, September 27). Safeguarding the city against extreme weather. *The New York Times*. https://nyti.ms/3B9ruIX

Chang, CH. (2011). Preparedness and storm hazards in a global warming world: Lessons from Southeast Asia. *Natural Hazards*, *56*(3), 667-679. https://doi.org/10.1007/s11069-010-9581-y

Dawkins, M. (2021). *Research Timeline.* [Figure 1]. *Prospectus* (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Dawkins, M. (2021). *Actor-Network Theory Diagram.* [Figure 5]. *Prospectus* (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Hurtado-de-Mendoza, A., Cabling, M. L., & Sheppard, V. B. (2015). Rethinking agency and medical adherence technology: applying Actor Network Theory to the case study of Digital Pills. *Nursing Inquiry*, 22(4), 326-335. https://doi.org/10.1111/nin.12101

Jarrahi, M. H., Newlands, G., Lee, M. K., Wolf, C. T., Kinder, E., & Sutherland, W. (2021). Algorithmic management in a work context. *Big Data & Society*, *8*(2), 1-14. https://doi.org/10.1177/20539517211020332

Jeans, D. (2020, October 20). Companies will spend $50 billion on artificial intelligence this year with little to show for it. *Forbes*. https://bit.ly/3Bdr8kQ

Kolbjørnsrud, V., & Amico, R. (2017). Partnering with AI: How organizations can win over skeptical managers. *Strategy and Leadership*, *45*(1), 37-43. http://doi.org/10.1108/SL-12-2016-0085

Liu, C., Guo, L., Ye, L., Zhang, S., Zhao, Y., Song, T. (2011). A review of advances in China's flash flood early-warning system. *Natural Hazards, 56,* 619-634. https://doi.org/10.1007/s11069-018-3173-7

Manjoo, F. (2019, November 15). We hate data collection. That doesn't mean we can stop it. *The New York Times*. https://nyti.ms/3EglAIe

Martin, N. (2019, August 7). How much data is collected every minute of the day. *Forbes*. https://bit.ly/3vL2CGn

Office of Public Works. *Flood Data Collection*. Office of Public Works. https://assets.gov.ie/68893/0f532f8ec53a4107a191fa9281935d06.pdf

Ohl, C. A., & Tapsell, S. (2000). Flooding and human health. *British Medical Journal*, *321*(7270), 1167-1168. https://doi.org/10.1136/bmj.321.7270.1167

Torti, J. (2012). Floods in Southeast Asia: A health priority. *Journal of Global Health, 2*(2), 1-6. https://doi.org/10.7189/jogh.02.020304