CSPB 3112 Professional Development in Computer Science

Project Proposal: Researching Socioeconomic Impacts and Technical Aspects of Proxy Data Use

Premise:

Most digital footprint analyzers and reports are focused on privacy and reputation. These are important issues, but there are additional vulnerabilities that many people may not be aware of. In the summer of 2025, I read Weapons of Math Destruction by Cathy O'Neil. The book affected me profoundly, opening my eyes to the ways data is used in various business and government sectors. I began to understand that the entities who determine how data is used are the powerbrokers, and I would like to see some of that power returned to the people.

Motivation:

I believe individuals have a right to understand the full picture of how their data is being used and interpreted, especially where their data makes them vulnerable to discriminatory practices. I want to empower individuals to understand how they are seen in the data world and be able to advocate for themselves, whether applying for a loan or going to a doctor who uses an AI diagnostic tool.

Vision Statement:

I will research and explore the known ethical issues of data use, the format and structure of data packets that can be obtained from devices, and the programming languages, tools, and libraries to work with the data. My goal at the end of this course is to have a foundation for beginning work on an application that can produce a discrimination vulnerability report for an individual user.

Specific and measurable goals: (Each week 3-5 hours will be dedicated to the project, and a log kept of how time was spent. The proposed schedule is below but may be subject to change.)

- **Week 1-** pondering
- **Week 2-** build github page, write proposal, discuss proposals on piazza, begin researching mobile and desktop computer data capture tools (Wireshark, PacketCapture, etc). Find or build a weekly report form following the Agile methodology for easy weekly update submissions.

- **Week 3-** determine a strategy for capturing data and organize data collection process. Research useful Kaggle data sets. Research Proxy Discrimination and Algorithm Bias for background.
- **Week 4-5:** Read and summarize background articles. Research data types, data terms, and technical aspects of data collection. Add a glossary of terms to the project for learning purposes.
- **Week 6-7:** Research how big data is collected and which tools would be most appropriate for acquiring data and extracting the relevant information.
- **Week 8-9:** Review any viable dataset options discovered or attempt to extract some personal data.
- **Week 10:** Research any publicly available algorithms that mimic those used by medical applications, finance applications, HR applications, etc.
- **Week 11-12:** Write an outline of the full application implementation, including research topics and/or data sources still needed, recommended languages/libraries/tools to accomplish the task. Define proposed technology stack.
- **Week 13:** Project wrap up, evaluation.

Risks to project completion:

- no prior experience working with the subject or technology- having difficulty defining a clear scope.
- the project requires data that may be difficult to obtain- lack of knowledge about the subject is impeding efforts to search for useable data.
- work/life/school balance

Mitigation Strategy for the risks listed above:

- Focus the project on exploration, learning, and building a foundation rather than on a deliverable. I am a coding novice, and it's not realistic for me to develop an application in 40-45 hours. It IS realistic for me to do the research and lay the foundations so that I am prepared to jump into building the application at some point in the future.
- Stick to the weekly schedule and the time commitment. Document everything.

Project Assessments - provide a list of evaluation criteria for the project.

- I will learn about current ethical concerns in how data is collected and used- library of citations and written summary of each.
- I will learn which commonly collected data parameters are used as proxies in several main sectors such as Finance, Medicine, and HR. – library of citations and written summary of each.
- I will explore and use techniques for capturing the data that is sent to various entities through my desktop and mobile device. – Jupyter notebook or similar where I can include both code snippets and text describing processes and tools.
- I will become familiar with the format and content of various data packets. - Jupyter notebook or similar where I can include both code/data snippets and text describing processes and coding tools. *Tentative based on data availability*.
- I will explore programming languages and techniques for collecting and processing data packets and extracting the relevant data. Jupyter notebook or similar where I can include both code snippets and text describing processes and coding tools.
- I will generate a plan or outline for the future application. The outline will contain a description of desired functionality, a suggested technology stack, and a description of issues or problems that will still need to be resolved.

Project portfolio link:

- https://sgillihan.github.io/