Thesis Project Portfolio

Data Pipelines: Ways Data Collection and Analysis Pipelines Can Be Built Through Cloud Services

(Technical Report)

Problems of Relying on Data as Basis for Important Decision Making and How Best to Mitigate them Through Proper Data Collection and Analysis Design.

(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

Ruohan Ding

Spring, 2023

Department of Computer Science

Table of Contents

Sociotechnical Synthesis

My capstone research addresses the problem of biased data and it's flawed analytical pipelines. Human reliance on the resulting analysis of these pipelines to make important decisions often lead to severe consequences. The technology that I purposed is not anything new or groundbreaking, but rather a combination of existing technologies designed to safeguard against flawed data analysis. Cloud services provided by major companies such as AWS can be used to create proper data pipelines as long as the engineer keeps his own bias in check.

The engineer's values and beliefs as well as the society he lives in impacts his creations drastically. This is in part what leads to the problem of flawed data and analysis. It is imperative for people trying to build unbiased data analysis pipelines to forget their personal biases and realign their values. Or else the end product will never be truly unbiased. The main STS theory that this problem relates to is the Social Construction of Technology, which theorizes that technology is impacted by the personal and societal values of the people that created them.

The method I plan to use to research this is doing case studies on existing companies in the data analysis field to see how they handle this problem of creation bias. After doing so I will use my findings and apply them myself by creating a proper data analysis pipeline. I expect to discover that many companies fail to address this problem. However, for those select few that are successful in limiting bias, I want to understand how they are able to do so. I suspect that these successful companies have strong cultures that promote values of truthfulness in their engineers. This in turn causes their engineers to create products that also carry these values.

My capstone research and my STS research both try to fix flawed data analysis. The capstone does so by creating better technology with more guardrails while the STS does so by changing the values of the creator of the technology. When they are considered in concert it makes for a strong step in the right direction of fixing the biased data problem. Engineers not only need to set aside their own biases but also use the proper tools in order to develop strong data analytical pipelines.