Computer Science Research Capstone: Creating a framework that allows bias to be detected in database systems

STS Thesis: Understanding the socio-political causes and effects of bias in data

A Thesis Prospectus Submitted to the

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

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

,

Technical Project Team Members
Elias Haddad

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

Signature _____Elias Haddad_____ Date _12/04/2020_

Approved _____Daniel Graham_____ Date ___05/12/2021_____
, Department of Computer Science

Approved _____Sharon Ku_____ Date ___05/12/2021___
, Department of Engineering and Society

Introduction

As everyday applications increasingly affect everyone's lives, more data is generated, and these applications use that data to make decisions. This data must correctly represent all social groups involved, otherwise, it could result in their misclassification. This is referred to as biased data. I will research the causes and effects of biased data in database systems and applications. This will include survey interviews with industry professionals and professors, conducting focus groups, and document analysis to better understand how biased data is caused and study cases on the effect it has had on applications. The STS prospectus will utilize the interviews, focus groups, and existing literature to focus on how socio-political factors affect bias in data; starting from the data collection phase and proceeding to the data cleaning phase of the database system. This will tackle three questions revolving around the socio-political origin, specifically how and where does bias originate from, at which point of the database lifecycle does it get introduced, and who is responsible for introducing this bias.

My technical topic will be focused on utilizing the information gained about the relationship between the previously mentioned socio-political factors and bias in the database system to construct a framework that can prevent harmful bias from entering the system. I will be experimenting with which data collection, cleaning, and analysis techniques work best under various types of data, in hopes of a better understanding of how to prevent harmful biased data from entering or staying in systems and applications. The technical topic will also utilize survey interviews and existing literature to get a better understanding of current techniques used to filter out bias in these systems. The following prospectus will introduce my research question while elaborating on my motivations for the pursuit of this research topic. The prospectus will then discuss existing literature related to our research topic. It will continue by describing the Science, Technology, and Society (STS) framework I have selected and its benefits. Finishing off with a description of my methodology for data collection and experimentation.

Technical Topic

I will be branching off of my STS prospectus research question of the causes and effects of bias data in systems and applications and instead I will go further and ask how we could prevent bias data from entering these systems. To figure out how to prevent biased data from existing in the systems, we must understand the causes and effects of it. Methods of preventing biased data existence will comprise of general but effective methods of data collection and cleaning, so we can have a policy or framework that, when followed, will prevent biased data from entering or staying in the data reliant systems. I aim to build this framework around the general data lifecycle that can be seen in figure one below; so when a project or experiment takes place that follows the data lifecycle processes, the social actors can follow my framework to address possible bias in the system. I emphasize "general" methods because biased data is very difficult to detect in a general case and the systems are often tailored to the specific dataset to detect and get rid of the bias in the data. However, I intend to find common factors that might exist in all biased systems, which in turn will allow me to create a general method that will allow all data to be examined for bias. This technical research topic aims to bring the perspective of minority social groups into the discussion when using their data for something controversial, since, historically, they are usually the groups affected by biased systems and applications. This will allow me to construct questions that need to be asked throughout the processes of the database lifecycle to continually address bias in the system.

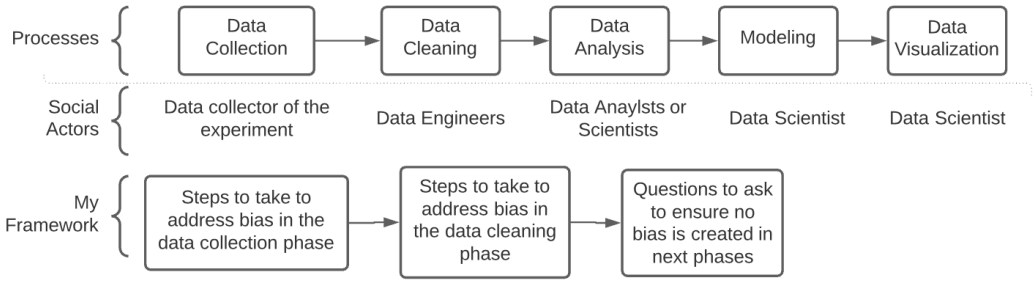

Figure 1: Data Lifecycle

Once the methods of research mentioned earlier are carried out, I will conduct experiments of different data collection and cleaning methods, in hopes of finding an ideal methodology for all data. It is important to mention that the results from the survey interviews, focus groups, and my findings for document analysis will affect the types of experiments I conduct. A partition of my questions for industry experts and professors during the survey interviews will be on best practices for data collection and cleaning. Data collection experiments will comprise of an analysis of primary versus secondary data collection, interviews from subjects of the data collection, and pipelines of data collection, where pipelines, in this case, are defined as a portal all data must be filtered through before proceeding. Data cleaning experiments will comprise of best practices for handling missing values in the dataset, best

practices for handling outliers in the dataset, and best practices on data cleaning pipelines, where pipelines, in this case, are defined as a portal all data must be passed through to convert the data into a better and easier to handle format.

Data collection could be done by gathering data from primary sources, such as customers buying a product, interviews, surveys, customers using a product, and anything that is collecting data directly from the users themselves. Whereas secondary data is obtained from a source that has already collected the data; this is usually done by obtaining a CSV file of the dataset. Both primary and secondary methods have their pros and cons, for example, primary data gives you the ability to collect and input the data the best way you see fit whereas secondary data does not give you that option, and makes you hope the collection was done properly. On the other hand, primary data collection usually exhausts your resources, whereas secondary data collection is very quick and easy. I believe that to get a comprehensive understanding of the data, it is important to interview either the subjects of the data or the source that collected the data. This will allow us to ask questions about how this data came about, why are the values of the data the way they are, and other types of questions that give us the meaning of the data. Ideally, if I were to find a trend of best practices when collecting data, I could create a pipeline that will help generalize the data collection process. I aim to experiment with these methodologies with both biased and unbiased data to see which methods are best for generally preventing biased data from entering or staying in applications and systems.

Data cleaning already has a generalized procedure that allows us to clean the data before using it; however, it does not take into account the possible biases towards minority social groups that could be created. Generally, when a data set has a missing value, we replace that value with the mean of the associated data, and this is fine for the most part but what effect does that have on the overall dataset? When dealing with outliers, we might get rid of them to generalize the data set but in the case of a biased dataset, would the outlier be beneficial to leave in or to take out? When using data, we need the format to be numerical floats, so some pipelines exist that convert data to the desired format. However, does there exist a procedure that we can add to these pipelines to better handle biased data? I aim to experiment with these different methodologies with both biased and unbiased data to answer the questions above and find the best practices for cleaning data to prevent biased data from staying in the system.

STS Prospectus

Introduction

Every day we produce 2.5 quintillion bytes of data, and many companies and agencies use this data to gain information about you, the economy, a product, and so on. This information then influences future types of products, policies, and economies. These may be positively or negatively affected depending on how well formulated the machine learning model used on the data is and how well made the dataset is. If data is biased, the model will be as well, which can result in these products, policies, and economies being built around a biased train of thought. As mentioned by Steve Barth, “One of the biggest challenges that our industry faces any industry considering ML, is biased in machine learning.”(Barth, “Bias in Machine Learning”, 2020). It is our responsibility as engineers to educate other engineers and scientists about this, to mitigate the effects of biased data. I believe this topic demands an STS investigation because as mentioned earlier, the applications are part of our lives now. This is seen anywhere from getting approved for a credit card to determine when an inmate should be released from prison, and if we do not properly address the issues, it can result in consequential misclassification of social groups.

Research Question

This study will aim to answer what the consequences of biased datasets and models are. However to fully understand this question we must answer other questions along the way, such as “What are the causes of biased datasets?”, “Who is affected by the biased models?”, “Where do we see this bias mainly take place?” and “How can we mitigate the negative results of biased models or prevent them from forming?”. With that said, I will research to answer the question of what causes biased data to enter systems and applications; and what are the effects of this bias if not accounted for. As mentioned before, I will expand upon those questions for my technical topic by answering the question of what are the best practices to prevent biased data from entering or staying in the applications and systems.

Literature Review

Data bias usually occurs during the collection phase and it could occur for many reasons, such as, the collector and engineers have some underlying political, religious, or social bias that allows them to overlook a potential bias in the data or system (Safiya Umoja Noble, 2017). Likewise, bias could be naturally prevalent in the data; however, either way, bias needs to be identified and understood for the application to be responsibly applied. In addition to the collection phase, bias can stay throughout the cleaning process or even be generated if the cleaning process is not done correctly. Outliers and missing data can misrepresent the social groups the data is based on by offsetting the values of other attributes. It has been shown that generally getting rid of outliers and filling missing data with the average value, will generalize

the data set and eliminate the bias caused by outliers and missing data (Amy H. Kaji. et al, 2014). Also, a validation step will allow a final check of the data before it gets processed, but this is different concerning each dataset in question (Jochen Sieg. et al, 2019). Throughout this process, if bias can be detected but not removed because of the nature of the data, it is just as effective to address that this application has some bias to all stakeholders, which in turn will prevent any misclassification of social groups (Anita Holdcroft, 2007). Data experts have mentioned a way to combat bias in data is to examine and conduct both in-depth and high-level discussions of publication and selection data; such as contacting and asking the collector of the dataset questions on their procedure and so on (Ikhlaaq Ahmed. Et al, 2012). With that being said, bias can be present because of several reasons and is difficult to point out a creator of bias in a generalized manner, however, it has been advised that allowing more human interaction either through the elaboration of the data or having a more socially diverse workforce will prevent harmful bias from negatively affecting any social groups. This information gives us a starting point in our research and allows us to plan the research methods accordingly, to gather the information that is not yet available in scholarly literature.

STS Framework and Methods

Thomas Hughes' Large Technical Systems (LTS) framework is a way to conceptualize a complex and hierarchically nestled system that not only involves technical factors, but also social factors (Thomas Hughes, 1989). The LTS framework has components in a system, a database system in our case, that interacts with other system components. Which all contribute to the common system goal, which in our case is to efficiently hold and distribute data to other systems. This socio-technical framework allows us to describe system builders that are responsible for or part of the management and execution of the system. This framework allows us to explain the relationship between the technical and social factors mentioned. I believe that the LTS framework will work best because, in addition to system builders, the framework also focuses on momentum, how the system will evolve. This is particularly useful since if there exists any bias in the database system, to begin with, the momentum of the LTS will result in a biased system since the LTS will evolve based on the existing bias, allowing us to better identify its causes and effects.

When conceptualizing biased data in database systems, there are many subsystems involved, such as the machine learning models that use this data, and other applications. There are also system builders involved such as data collectors, engineers, and policymakers. I am confident that I can represent my research question using Hughes' LTS framework. Since our research question naturally fits the components of this framework, it will allow us to identify the social, political, and technical means of constructing the system. It will also allow us to identify whether or not there exists a reverse salient and how that may affect the social and political means which might be leveraged to increase the social adaptation of the technology (Thomas Hughes, 1989).

For my research, I am conducting survey interviews, focus groups, and document analysis; which as mentioned before will contribute towards both my STS topic and my technical topic. Surveys will consist of one on one interviews with field experts and professors from the Computer Science department at the University of Virginia (UVA) that focus their research on data science and data-related applications. I am to conduct three survey interviews with professors and at least one with a field expert. I believe four interviews will be enough to be exposed to different perspectives and different ideas within our time constraint, which is mentioned in the next section. The interviews allow me to get a better understanding of how data bias is caused, why it might be happening, the effects it has on applications, and how to combat it. The first interview will have questions that will help define the relevant social actors towards these systems. Therefore we can better understand the impact of these social actors and be able to best construct the focus groups since the focus groups will consist of students who intend to work in a data-related position that makes them social actors in the data lifecycle.

Focus groups will consist of fourth-year college and master level engineering students and they will be asked questions on both biased or unbiased datasets to understand how they perceive data. One group will be just fourth-year students, the second will be just master level students and a third will be a mixture of the fourth year and master level students; all groups will consist of four to five subjects. Since it will be most effective to mimic a real-life workplace environment, I will try to make the groups diverse. This means including people not only from the computer science department, but also from the mathematics, physics, and electrical and computer engineering departments. This also implies that I will need to try to have a diverse environment of gender, ethnic and racial backgrounds since most workplace environments are diverse in that regard. Additionally, each group should consist of students that will become data collectors, data cleaners, and other positions in the data lifecycle. The focus groups allow me to understand how subjects view biased and unbiased data, as well as, how their perspectives might impact bias in a system. The group setting may pose a bias if one subject were to pose an answer based on another subject's answer; which is why I will be implementing a low-stress environment with open-ended questions.

My final research method will be document analysis, which is similar to the ones mentioned throughout the cited literature review but will be in greater detail to better understand how data bias is caused, why it might be happening, the effects it has on applications, as well as how to combat it. I will conclude my findings, elaborating on which methods contributed the most to my research, and the next steps for handling bias in data related systems. These next steps will utilize my findings from the experiments mentioned above and a summary of the causes and effects of biased data on social groups, based on my research, as well as, different methodologies to prevent bias data from entering or staying in the system.

Timeline

I will first conduct survey interviews with professors and field experts; this will allow me to reevaluate my questions and plan for the other research methods. I will do this throughout January so by the end of the month I will have completed five survey interviews. Once that is completed, I will conduct focus group interviews, I will do one every week for three weeks of February. Throughout January 2021 and February 2021, I will also be conducting document analysis. A milestone in my research will be at the end of February once I have completed the research methods that will answer the questions of what causes and what are the effects of biased data in systems, concerning social groups. For March 2021, I will be conducting my technical research, which will first consist of best practices experiments to prevent biased data in data collection for the first 2 weeks of the month. For the second two weeks of the month, I will experiment with the best practices for preventing biased data in data cleaning. By the end of March, I will have hit another milestone as I will have conducted all experiments for my technical research topic. This will allow me to take the months of April 2021 and May 2021 to organize my findings and write up the thesis to communicate my findings and takeaways.

Conclusion

I will be researching to investigate what causes and what are the effects of biased data in systems, concerning social groups. I will conduct experiments to better understand how data bias is caused, how it enters the system and the historical effects it has had, and possible future effects. This will give me the understanding to proceed with my technical research, which will experiment with best practices in preventing biased data from entering or staying in the systems in question. To better identify the social, political, and technical means of constructing the database systems, we will utilize Thomas Hughes' Large Technical Systems framework. This research will span the month of January through May and will result in a thesis that gives us a better understanding of how biased data forms and enters the system and what possible effects it might have on social groups. The technical research will allow us to contribute to data related fields by organizing our findings and determining general best practices concerning preventing biased data from entering or staying in the system.

Bibliography

Ahmed, I., Sutton, A. J., & Riley, R. D. (2012). Assessment of publication bias, selection bias, and unavailable data in meta-analyses using individual participant data: a database survey. *BMJ*, *344*(jan03 1), d7762. https://doi.org/10.1136/bmj.d7762

Barth, S. (2020, April 21). What is Bias in Machine Learning & Deep Learning? Retrieved from https://www.foreseemed.com/blog/bias-in-machine-learning

Bodenhausen, G. V. (1988). Stereotypic biases in social decision making and memory: Testing process models of stereotype use. *Journal of Personality and Social Psychology*, *55*(5), 726–737. https://doi.org/10.1037/0022-3514.55.5.726

Cozzens, S. E., Bijker, W. E., Hughes, T. P., & Pinch, T. (1989). The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology. *Technology and Culture, 30*(3), 705. doi:10.2307/3105993

Dee, D. P. (2005). Bias and data assimilation. *Quarterly Journal of the Royal Meteorological Society*, *131*(613), 3323–3343. https://doi.org/10.1256/qj.05.137

Holdcroft, A. (2007). Gender bias in research: how does it affect evidence based medicine? *Journal of the Royal Society of Medicine*, *100*(1), 2–3. https://doi.org/10.1258/jrsm.100.1.2

Ibnat, F. (2015, August 27). Considering Biases at the Data Cleaning Stage. Retrieved from https://uwescience.github.io/DSSG2015-predicting-permanent-housing/2015-08-27-biases-in-data-cleaning/

Kaji, A. H., Schriger, D., & Green, S. (2014). Looking Through the Retrospectoscope: Reducing Bias in Emergency Medicine Chart Review Studies. *Annals of Emergency Medicine*, *64*(3), 292–298. https://doi.org/10.1016/j.annemergmed.2014.03.025

MEHRABI, NINAREH, MORSTATTER, FRED, LERMAN, KRISTINA, GALSTYAN, ARAM, & SAXENA, NRIPSUTA. (2019). A Survey on Bias and Fairness in Machine Learning. *A Survey on Bias and Fairness in Machine Learning*, 1–31. https://arxiv.org/pdf/1908.09635.pdf

Noble, S. U. (2019). Book Review: Algorithms of Oppression: How Search Engines Reinforce Racism. *The International Journal of Information, Diversity, & Inclusion (IJIDI)*, *3*(1), 15–30. https://doi.org/10.33137/ijidi.v3i1.32272

Ortiz-Ospina, E., & Roser, M. (2016). Trust. *Published online at OurWorldInData.org.* Retrieved from https://ourworldindata.org/trust

Presser, S., & Stinson, L. (1998). Data Collection Mode and Social Desirability Bias in Self-Reported Religious Attendance. *American Sociological Review*, *63*(1), 137–145. https://doi.org/10.2307/2657486

Sieg, J., Flachsenberg, F., & Rarey, M. (2019). In Need of Bias Control: Evaluating Chemical Data for Machine Learning in Structure-Based Virtual Screening. *Journal of Chemical Information and Modeling*, *59*(3), 947–961. https://doi.org/10.1021/acs.jcim.8b00712

Turner Lee, N. (2018). Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication and Ethics in Society*, *16*(3), 252–260. https://doi.org/10.1108/jices-06-2018-0056