**CMSC 451 – CS-22-319 Machine Learning Tool for Identifying Anomalies in Application Logs Project Proposal**

Group Members: Adrienne Hembrick, Dylan Pierce, Sean Stitzer, Sam Sunvold

**Problem Statement:**

* From the software applications, network infrastructure, and a variety of technical tooling used across the enterprise come a deluge of log data that is consumed downstream by an even broader set of applications, security tools, and reporting capabilities. Sometimes changes in the data provider will alter the format or contents of the logs they generate, having unintended and sometimes undetected impacts on these downstream consumers. The variety of log sources and volume of log entries makes it difficult to create individual monitoring capabilities to verify the output of each data provider, however without monitoring there is a high risk of failure as applications are upgraded and evolved over time. This project is to develop a single solution that can consume a stream of log records without prior knowledge of the log source or record format and adapt over time to recognize the log entries and then detect potential material changes in their contents or format. Specific care would need to be made to prioritize solutions that can work at high volume without introducing significant delay to data pipelines that leveraged the final solution.

**Business and Functional Requirements:**

* A machine learning approach is not necessarily required, but we figured it would be the best approach.
* Due to not having labelled data to train our model on, we plan on going with an unsupervised approach.

**Stakeholders:**

* Since practically all software applications, network infrastructures, and any other technical tooling used in the business realm produce log data, almost everybody could benefit from the use of this tool. Using this tool would reduce the risk of most failures and allow administrators to find unusual logs that would come about due to unexpected change in data logs.

**Assumptions/Constraints:**

* The only constraint as of now is being able to understand out approach that we are currently looking at going with. It is a very in depth approach that requires lots of prior knowledge in the machine learning field.
* I assume that we will go with the approach described in the previous meeting. (Sean’s find)