CS510 Data Science: Project Plan
Scott Rubey

Paper Topic

Data Science processes have historically been quite laborious; a great deal of the Data Scientist's time and resources have been devoted to such tasks as data cleaning and analysis. Recent years have seen an advent of tools that allow the Data Scientist to automate parts of this process. This project will explore the technologies and ethical considerations behind automation of data ingestion, storage, cleaning, analysis and modeling.

Research Planned

Sources for this topic fall into two broad categories: technology and ethics; some overlap is expected. IEEE and Google Scholar have each been invaluable research tools for finding technical papers covering the processes involved in automating various steps of the Data Science life cycle. A breadth of research exists with regard to AI/deep learning solutions to analysis and modeling problems. Further scholarly papers have been discovered that pertain to automation of other processes such as data wrangling.

Due to the ethics of Data Science automation being an inexact science, it is possible to expand the reasonable search criteria to vetted online sources, periodicals, and opinion pieces. Several publications with monthly circulation numbering in the hundreds of thousands have released pieces in recent years outlining such ethical considerations, such as implicit bias in algorithms, outsourcing and job elimination. A further source has been uncovered in the realm of streaming video: Lise Getoor's keynote address at the 2019 IEEE Convention, entitled "Responsible Data Science."

Team Structure

This project will be completed individually.

Milestones

1) Gather sources

2) Complete reading of initial 4 sources, compile notes

3) Complete initial 4-page research report

4) Read / watch additional sources, compile notes

5) Flesh out 4-page research report with 5 additional pages

6) Prepare presentation materials (slideshow)