**Capstone Project Proposal**

**General information:** The National Center for Health Statistics (NCHS) collects and performs statistical analysis of birth and infant death in the US and its territories on a yearly basis. Most of the infant deaths are linked to the birth files providing a publicly available good data set for any one interested in detailed investigation of infant death in the country. The capstone project will attempt to employ data science tools to investigate the possibility of predicting infant fatality from vital statistics collected during pregnancy and at birth.

**Data set**: linked birth/infant death dataset from Centers for Disease Control and Prevention website (<https://www.cdc.gov/nchs/data_access/vitalstatsonline.htm> ). The data includes information that can be subdivided in to general information (birth year, birth week), infant information (sex, gestation, birth weight, Apgar score, etc), Mother information (age, education, race, etc), pregnancy information (total birth order, number of prenatal visits, etc), medical and health data (method of delivery, medical risk factors, etc)

**Who would care about the problem?** Utilizing readily available data to predict infant death would help different organizations and personnel including health care professionals, researchers, insurance companies, and parents.

**Procedures**

1. Obtain data and accompanying guide document from CDC website.
2. Extract attribute information from accompanying guide document and combine it with the actual data
3. Investigate type of variables, valid values, and missing values
4. Clean data including missing data (N/A, empty strings, invalid values)
5. The project will be approached as a classification problem. The dataset has labels that show whether an infant has died or not. I may also consider unsupervised approaches to cast the problem as clustering or anomaly detection.

**Final Package**

The final submitted package will include:

1. Python notebook with all the data wrangling, exploration, and analysis codes
2. Report
3. Power point slides summarizing the main findings