**HHS Data Science CoLab Capstone Project Proposal**

*Michael Harris June 13, 2018*

**Background:** TB Portals Consortium Background. Integration of Clinical and Genomic Data

**Research Questions:**

* Is there a correlation between mutation burden and patient outcome?
* Are there mutation patterns that are predictive of patient outcome?
* Which mutations or patterns of mutations are most important for predicting patient outcome?
* Can low level infections be detected using genomic data?
* Can mixed infections be detected using genomic data?

**Methods:**

* Use TB Profiler to generate Tuberculosis drug resistance (DR) Genomic SNP table
* Combine the DR genomic SNP table with clinical data and drug sensitivity data
* Apply data science methods to answer research questions (feature selection, visualization, clustering)

**Outcomes and Value:**

* Use methods to identify samples for deep sequencing. *This is a current research question for our project.*
* Report value of using whole genome sequencing of TB pathogen for detecting low level infections and mixed infections.
* Ultimately publish results in a scholarly journal.

**Other possible research questions:**

* TB Portals data analysis of plasmids to determine effect on outcome (thanks Faith!)
* Prediction accuracy of resistance profiling tools: TB Profiler, Mykrobe Predictor, KavarQ
* Detection of batch effects and data visualization: PCA tSNE ?
* Clustering and unsupervised learning to detect patient subgroups
* Analysis of Longitudinal Data from Patients; course of treatment