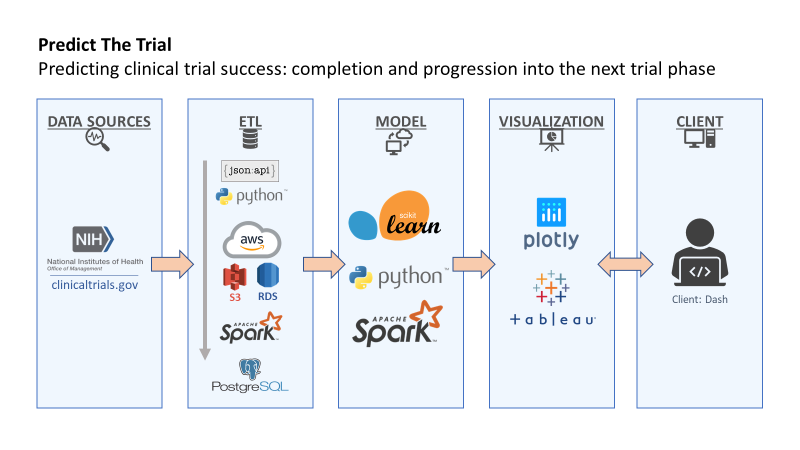
**Final Project – Predict The Trial**  November 2020

* Team members:
* Eben Haezer
* Momotaz Mahin Khan
* Sheri Shojaie
* Vivi Santosa
* Project description / outline:  
  The goal of this project is to identify factors in clinical trials that are predictive of a trial’s probability to succeed. There are many factors that contribute to a trial’s success or failure, including patient selection, trial design, monitoring technique, duration, and study end points. These are all factors that are captured, reported, and accessible for analysis and modeling. We propose to survey all US and Canadian clinical trials data between 1979 to 2020 to identify the most predictive factors for trial success and failure based on completed trials. There are over 300 data points for each clinical trial that can be featurized and used in predictive modeling. **The output of this project is the development of an algorithm that can predict whether an active trial will succeed or fail, where success is defined as trial completion and progression to the next phase (eg. Phase I study completion and Phase II initialization).**
* Data sources:
  + ClinicalTrials.gov  
    <https://clinicaltrials.gov/>
* Application Schema:   
  
* Draft of tasks
  + **ETL**
  + Extraction: Python API call; house in AWS S3, RDS
  + Transformation: Python Pandas
  + Loading: PySpark, load into PostgreSQL
  + **Algorithm development**
  + Supervised model: Sklearn’s linear regression model will be used to perform multiple linear regression using the selected trial features. We will predict the probability of trial success based on analyzed features
  + **Visualization**
  + Data exploration, model results and predicted outcomes of current active clinical trials will be visualized and reported in Tableau