**Final Project ideas**

* PubMed trends
* interactive data visualization, say with R Shiny, e.g., allowing patients to search for how different hospitals compare on various individual healthcare quality metrics and an overall metric determined by a machine learning algorithm
* compare performance of different machine-learning algorithms (vs an ensemble method) on simulated data, either with a causal goal (e.g., using TMLE) or predictive goal
* Predictions: kaggle –

**Dataset and goals**

**Dataset**

We will use the data found in the Kaggle “Porto Seguro’s Safe Driver Prediction” competition (<https://www.kaggle.com/c/porto-seguro-safe-driver-prediction/>) to predict the probability of a driver filing an auto insurance claim. The dataset provides subject-level data on deidentified features and whether a claim was filed. We will potentially use multiple imputation to impute missing values – *question for the TA: how useful is multiple imputation for prediction problems?*

**Goal**

In addition to the application question of predicting claim filing, we will compare predictive abilities of individual algorithms (generalized linear model, elastic net, neural net, random forest, k-nearest neighbors, bagging & boosting) vs. an ensemble approach to assess the gains from an ensemble method. *Question for the TA: how do kernel methods work on large datasets?*

**Risk assessment**

If we use multiple imputation and ensemble different algorithms, compute time may be prohibitively large. If that ends up being the case, we can use ORCHESTRA, a Harvard med school based compute cluster.

Slower than anticipated timelines are another risk. We will assist team members who are behind to mitigate their difficulties.

**Division of labor**

We will divide up the individual algorithms and each implement a portion of them. We will likewise divide responsibility for multiple imputation, the ensemble approach, and final copyediting. We will all look over each other’s’ code and contribute to the final report.