Data will be split into training, validation, and test sets, with the test set comprising the most recent cohort of applicants with known outcomes available in our data. Model validation metrics will be estimated by evaluating predictive performance on out-of-sample data in the test set, with model selection determined by optimizing performance on out-of-sample data in the validation set.

The above strategy will be used in building and fitting estimators selected from a family of linear and non-linear classifiers including support vector machines, random forests, penalized logistic regressions. Appropriate settings for these models will be estimated using a procedure of grid search and cross-validation to correct against overfitting to training data and ensure robust generalization to out-of-sample data.

For the screening resource optimization task, performance will be measured by the estimated reduction in faculty screening time that maintains precision and recall at specified thresholds of performance. Model output will consist of recommendations as to which applications require further faculty screening and those for which faculty screening time can be reduced.

For the predictors of success and difficulty tasks, performance will be measured as a function of precision-recall curves. Model output will consist of analyses of the feature importance scores in the model selected by optimizing this performance measure.