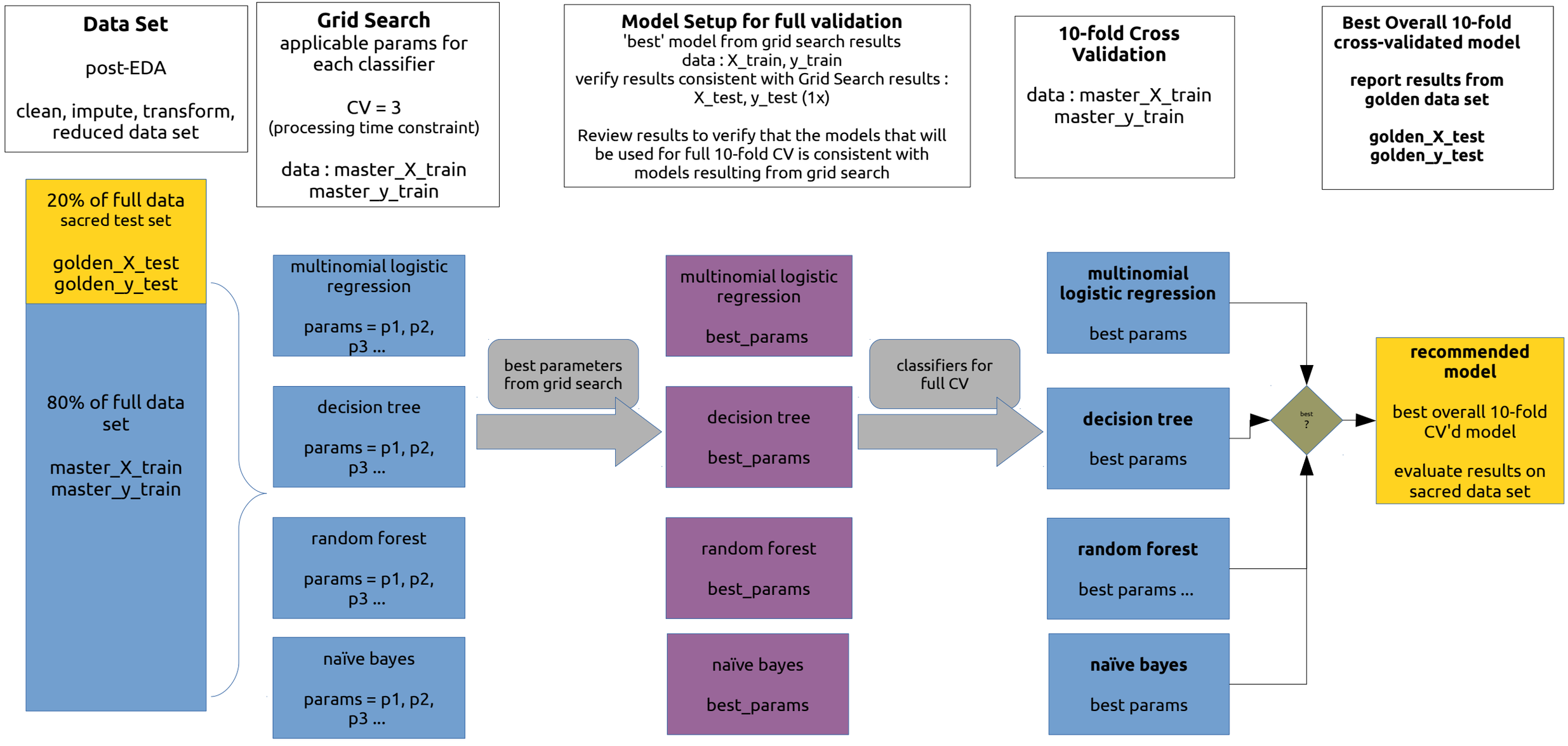
Data Set
post-EDA
clean, impute, transform, reduced data set
Grid Search
applicable params for each classifier
CV = 3
(processing time constraint)
data : master_X_train
master_y_train
Model Setup for full validation
'best' model from grid search results
data : X_train, y_train
verify results consistent with Grid Search results : X_test, y_test (1x)
Review results to verify that the models that will be used for full 10-fold CV is consistent with models resulting from grid search
10-fold Cross Validation
data : master_X_train
master_y_train
Best Overall 10-fold cross-validated model
report results from golden data set
golden_X_test
golden_y_test
20% of full data
sacred test set
golden_X_test
golden_y_test
80% of full data set
master_X_train
master_y_train
multinomial logistic regression
params = p1, p2, p3 ...
decision tree
params = p1, p2, p3 ...
random forest
params = p1, p2, p3 ...
naïve bayes
params = p1, p2, p3 ...
best parameters from grid search
multinomial logistic regression
best_params
decision tree
best_params
random forest
best_params
naïve bayes
best_params
classifiers for full CV
multinomial logistic regression
best params
decision tree
best params
random forest
best params ...
naïve bayes
best params
best ?
recommended model
best overall 10-fold CV'd model
evaluate results on sacred data set