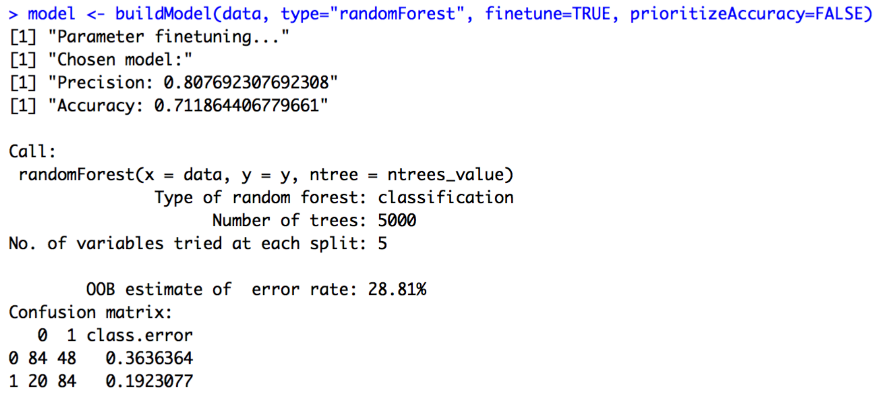
- Raw data [20 Features | 236 Data Points | 0-1 Success Labels]

Number	*Date Primer Ordered*	*Date PCR*	*Date BP cloned*	*Date Colonies Picked*	*chr*	*regst*	*reged*	*size*	*name*	*ID*	*ForwardPrimer*	*ReversePrimer*	*ForwardPrimerTm*	*ForwardPrimerLength*	*ReversePrimerTm*	*ReversePrimerLength*	*HairPinCheck*	*orig*	*ext*	*Success*

- Preprocessing
  - Remove ID and un-useful columns (*Number*, *ID*, etc.)
  - Add forward & reverse counts for bases and all possible *k*-mers with *k* = 2 counts (+(8 + 2x16) = 40 columns)
  - Add forward & reverse CG content (+2 columns)
  - Total number of columns = 52
- Feature Selection
  - High Correlation | 10 columns discarded
  - Recursive Feature Elimination | 31 significant columns kept
- Optional parameter finetuning (ntrees in the forest)

- Random Forest
  - 100-5000 trees tested | 5000 trees performed best
  - *Performance*
    - *Precision* 0.80769
    - *Accuracy* 0.71186
    - +/- 0.03 as dataset is small

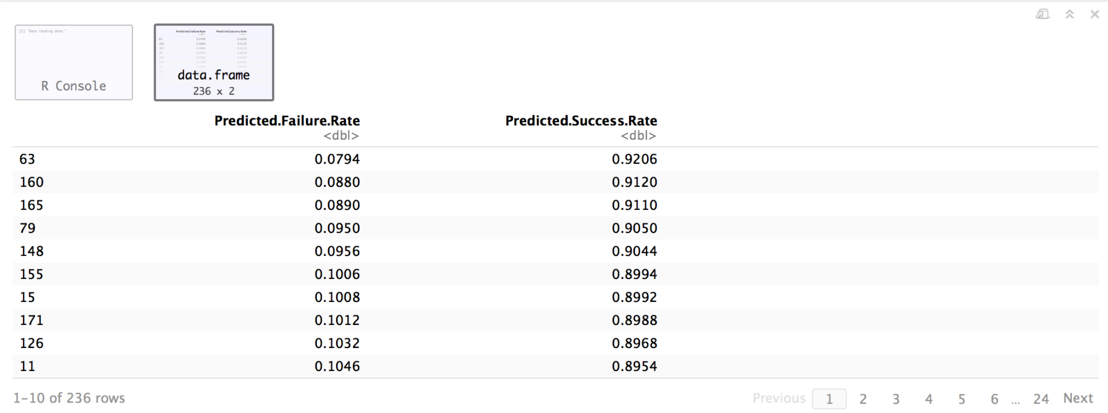


```
> model <- buildModel(data, type="randomForest", finetune=TRUE, prioritizeAccuracy=FALSE)
[1] "Parameter finetuning..."
[1] "Chosen model:"
[1] "Precision: 0.807692307692308"
[1] "Accuracy: 0.711864406779661"

Call:
 randomForest(x = data, y = y, ntree = ntrees_value)
               Type of random forest: classification
                     Number of trees: 5000
No. of variables tried at each split: 5

        OOB estimate of  error rate: 28.81%
Confusion matrix:
   0  1 class.error
0 84 48   0.3636364
1 20 84   0.1923077
```

- R script easy to run by Mark R's lab members
  - Can score new data points
  - Saves trained models
  - Precision vs accuracy prioritization model selection



	Predicted.Failure.Rate <dbl>	Predicted.Success.Rate <dbl>
63	0.0794	0.9206
160	0.0880	0.9120
165	0.0890	0.9110
79	0.0950	0.9050
148	0.0956	0.9044
155	0.1006	0.8994
15	0.1008	0.8992
171	0.1012	0.8988
126	0.1032	0.8968
11	0.1046	0.8954

1–10 of 236 rows

- Suggestions for more features?
- Forward & reverse counts of *k*-mers with *k* = 3 didn't help