Comparing Models Using Resampling and Bayesian Methods

Max Kuhn (RStudio) @topepos

Comparing Model Performance

Often a number of different models are created on a single data set.

Which model is the best depends on how you characterize their ***performance***.

- Examples are: area under the ROC curve, root mean squared error, R^2 , ...

Often, we would like ***estimates of uncertainty*** on these values so that we can choose the best model.

Confidence intervals are a common approach but have a weird interpretation.

> If we were to have repeated the experiment a large number of times, 90% of the true values would fall in `[L, U]` .

Bayesian analysis would allow us to make statements such as

> there is a 90% probability that the true value falls in `[L, U]` .

Let's do that!

Resampling

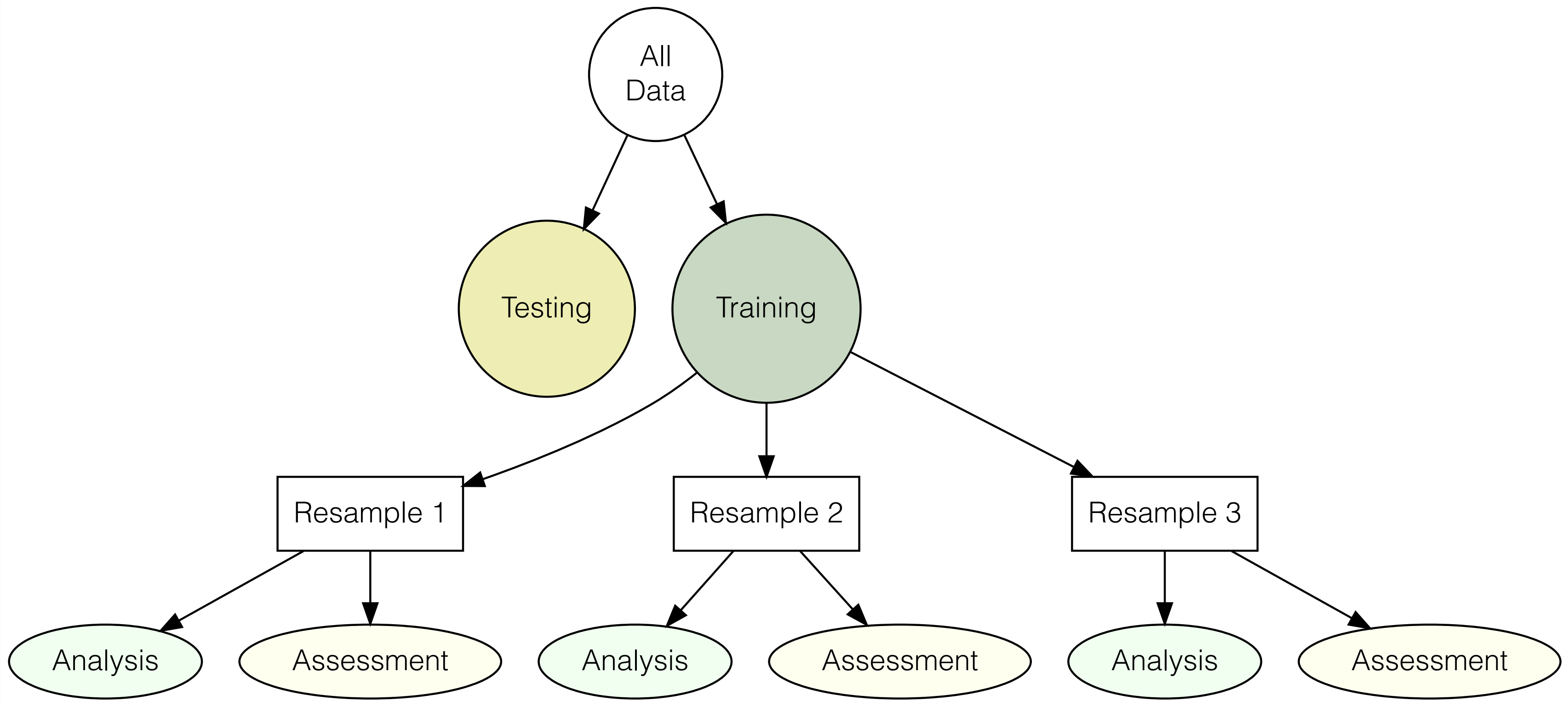
If we want to estimate uncertainty in the model's summary statistics, we'll need replicates.

Resampling methods (e.g. the bootstrap, *V*-fold cross-validation, etc.) can be used for this purpose.

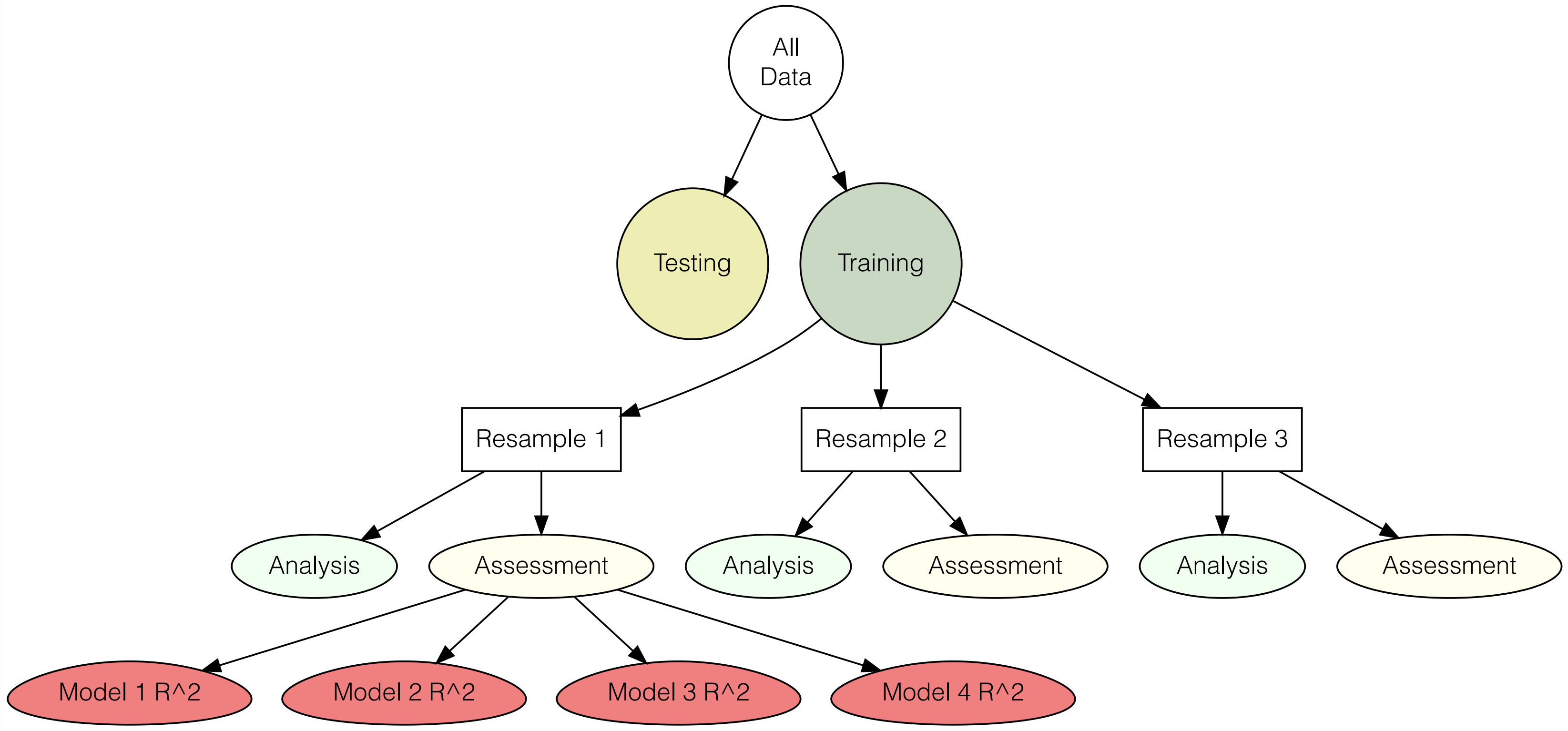
Resampling is basically an empirical simulation system that uses variations of the original data set to create multiple versions of the models and summary statistics.

Suppose we are estimating the coefficient of determination for each model (aka R^2)...

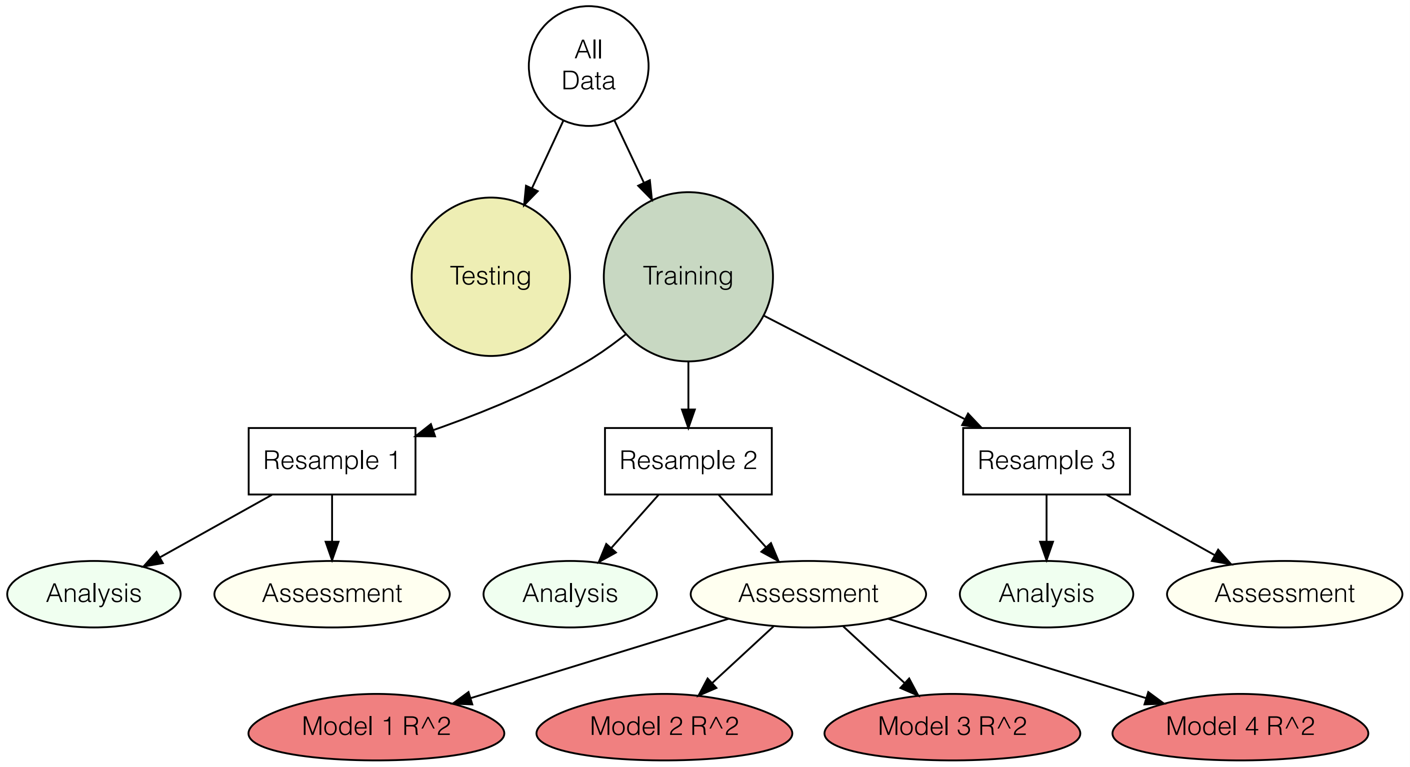
Three Iterations of Resampling



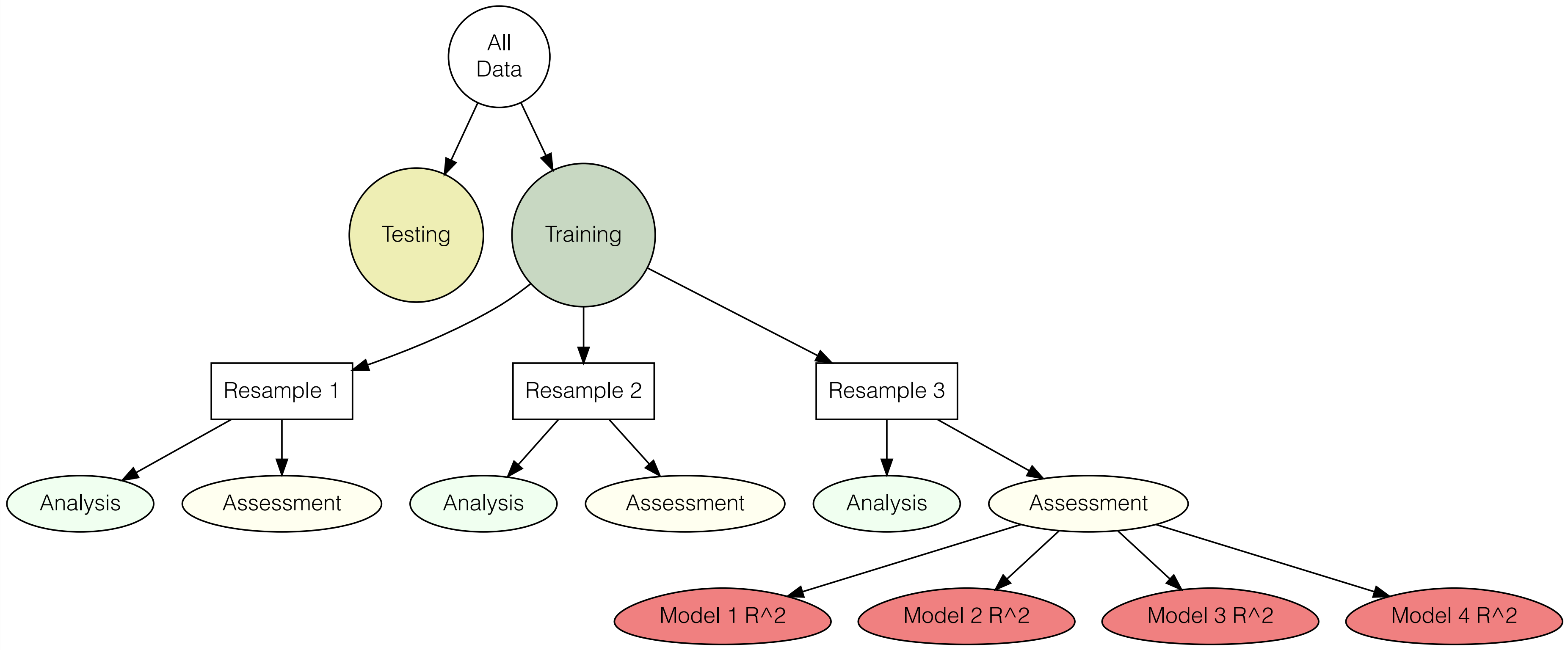
Resampling Iteration 1



Resampling Iteration 2



Resampling Iteration 3



Example Data Structure (Ames Housing Data)

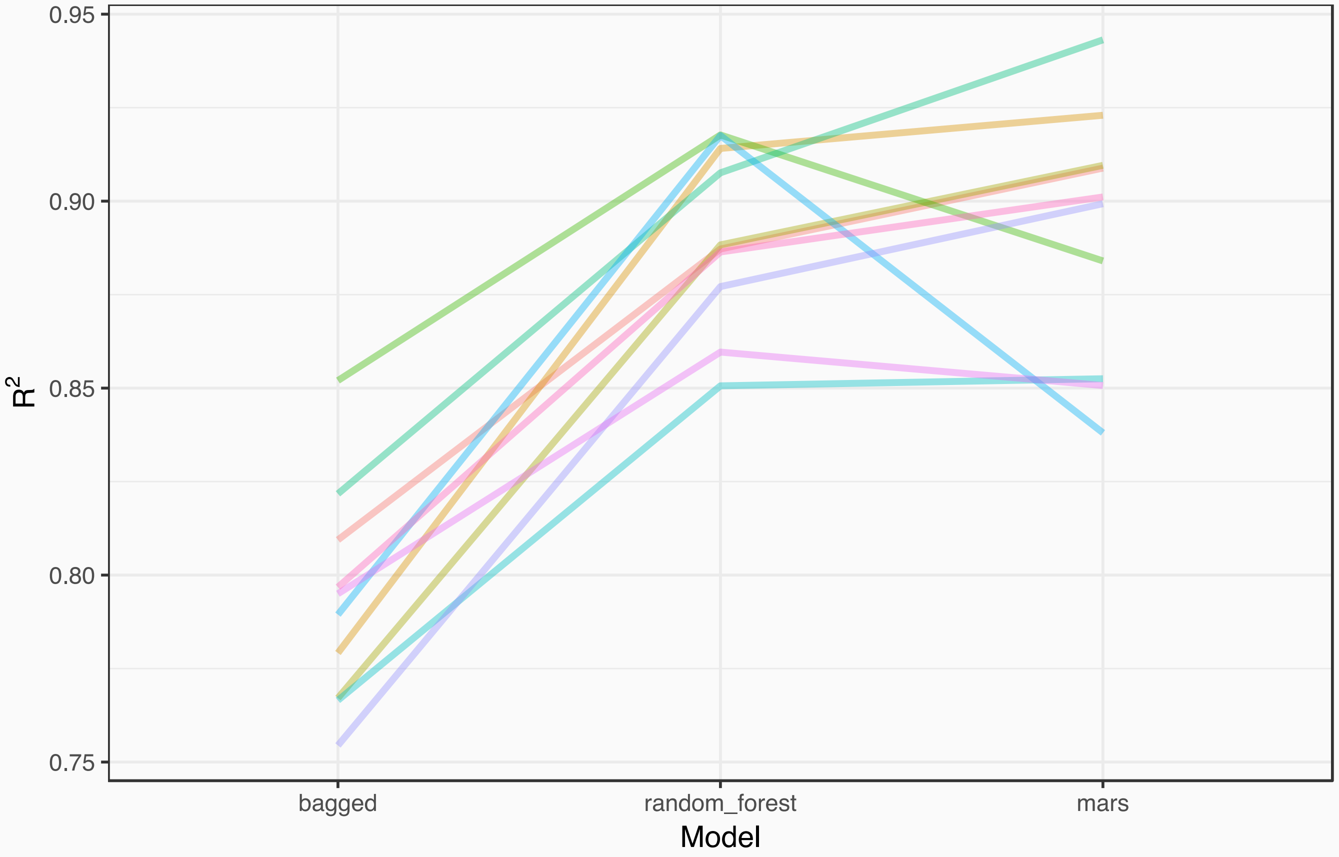
The results are a two-way layout of R^2 values when ***predicting the sale price***:

```
## #  10-fold cross-validation
## # A tibble: 10 x 5
##    splits       id     bagged random_forest  mars
##  * <list>       <chr>   <dbl>         <dbl> <dbl>
##  1 <S3: rsplit> Fold01  0.809         0.887 0.909
##  2 <S3: rsplit> Fold02  0.779         0.914 0.923
##  3 <S3: rsplit> Fold03  0.767         0.888 0.910
##  4 <S3: rsplit> Fold04  0.852         0.918 0.884
##  5 <S3: rsplit> Fold05  0.822         0.908 0.943
##  6 <S3: rsplit> Fold06  0.767         0.851 0.853
##  7 <S3: rsplit> Fold07  0.789         0.918 0.838
##  8 <S3: rsplit> Fold08  0.754         0.877 0.899
##  9 <S3: rsplit> Fold09  0.795         0.860 0.851
## 10 <S3: rsplit> Fold10  0.797         0.886 0.901
```

Note that the estimated model performance (i.e. R^2 here) is now the outcome variable that we will be analyzing.

We don't care about estimating resample-to-resample effects but there is often a within-resample correlation (0.272 for these data).

Within-Resample Correlation Structure



Bayesian Hierarchical Generalized Linear Model

If we did a basic ANOVA model to compare models, it might look like:

$$R^2 = b_0 + b_1 m_1 + b_2 m_2$$

where the m_j are indicator variables for the model (MARS, random forest).

However, there are usually resample-to-resample effects. To account for this, we can make this ANOVA model *specific to a resample*:

$$R^2_i = b_{i0} + b_1 m_1 + b_2 m_2$$

where *i* is the *i*$^{\text{th}}$ cross-validation fold.

Bayesian Hierarchical Generalized Linear Model

We might assume that $R^2_{ij} \sim N(\beta_{i0} + \beta_j m_{ij}, \sigma^2)$.

The b_{i0} parameters can have a normal distribution with mean β_0 and some (diffuse) variance. The distribution of the intercepts, along with distributions for the variance and slope parameters, are the *prior distributions*. This is a *random intercept* model (common in linear mixed models).

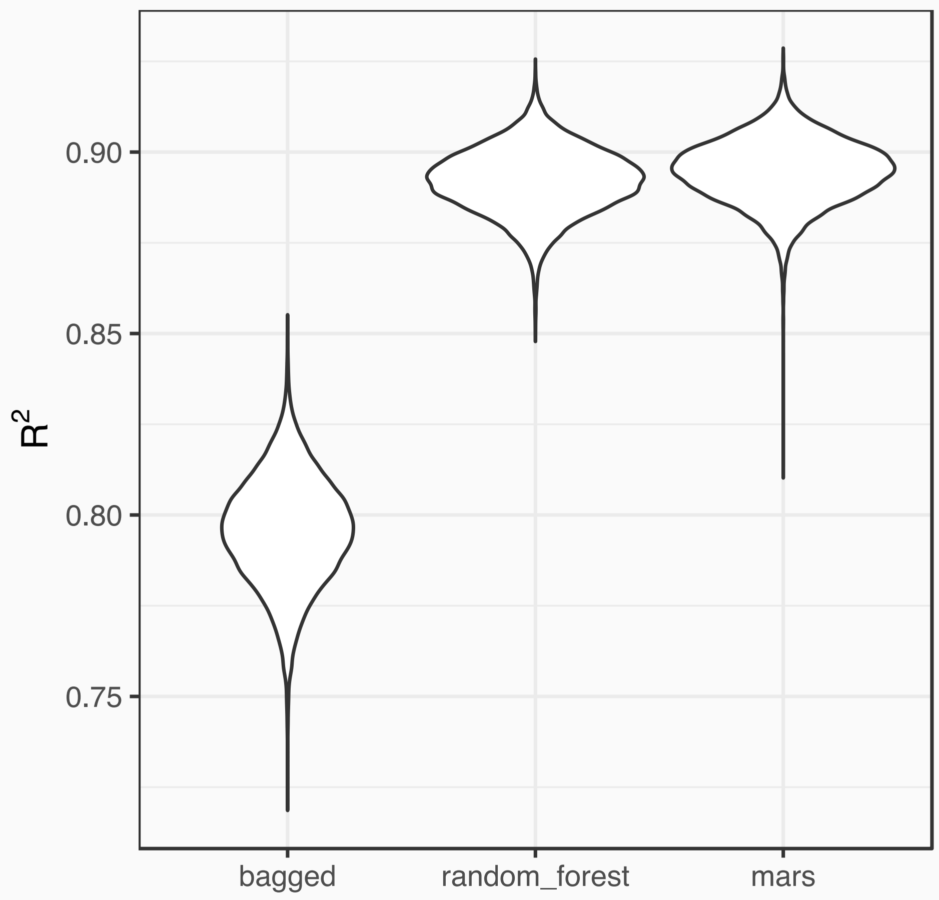
Bayesian analysis can be used to estimate these parameters. `tidyposterior` uses Stan to fit the model.

There are options to change the assumed distribution of the metric (i.e. gamma instead of normality) or to transform the metric to normality (logit for R^2 was used here).

Different variances per model can also be estimated and the priors can be changed.

Estimated Posterior Probabilities

A logit transformation was applied to the estimated R^2 values before the computations.



The `tidyposterior` package automatically back-transforms the posterior distribution.

90% credible intervals for each model's R^2:

```
## # A tibble: 3 x 4
##   model          mean lower upper
##   <chr>         <dbl> <dbl> <dbl>
## 1 bagged        0.795 0.770 0.820
## 2 mars          0.895 0.880 0.908
## 3 random_forest 0.892 0.877 0.906
```

Practical Differences

It's easy to compute the posterior for the difference in R^2 estimates between two models.

However, is the difference **relevant**?

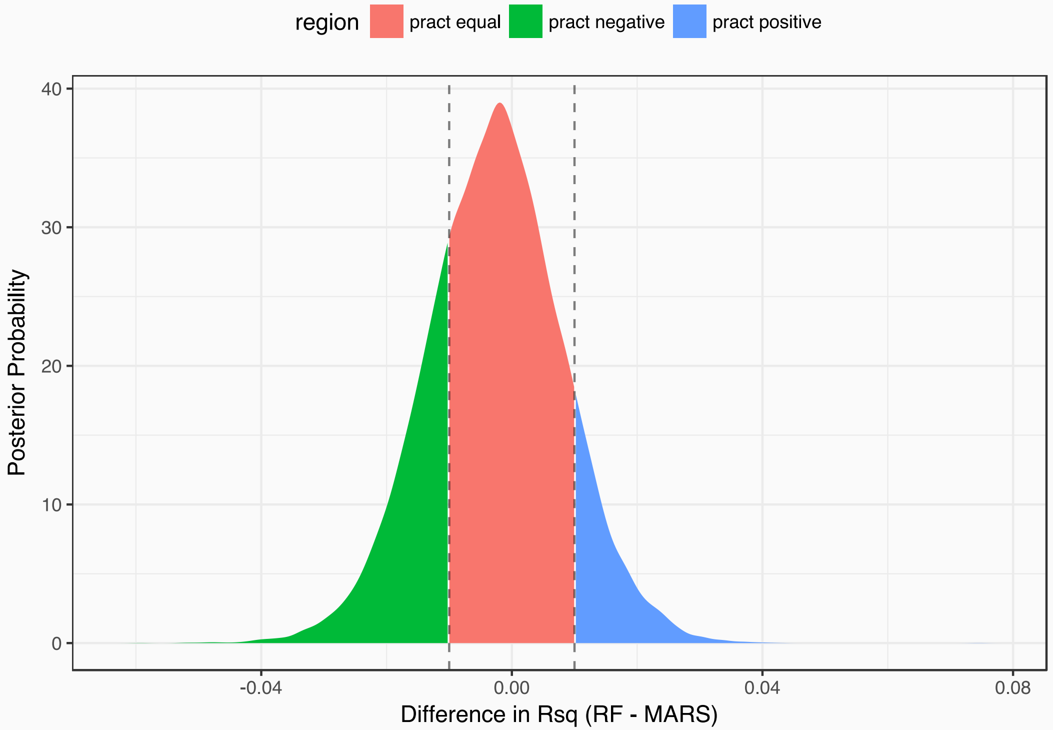
Suppose that we pre-define what a meaningful difference in R^2 values would be. Let's say that a +/-1% difference is big enough to be real.

ROPE estimates (region of practical equivalence) quantify how much of the posterior for the difference is within this region (Benavoli *et al* (2017) Kruschke and Liddell (2015)).

ROPE Illustration

The probability that random forest and MARS are ***practically equivalent*** is 63.7% for an effect size of 1%.

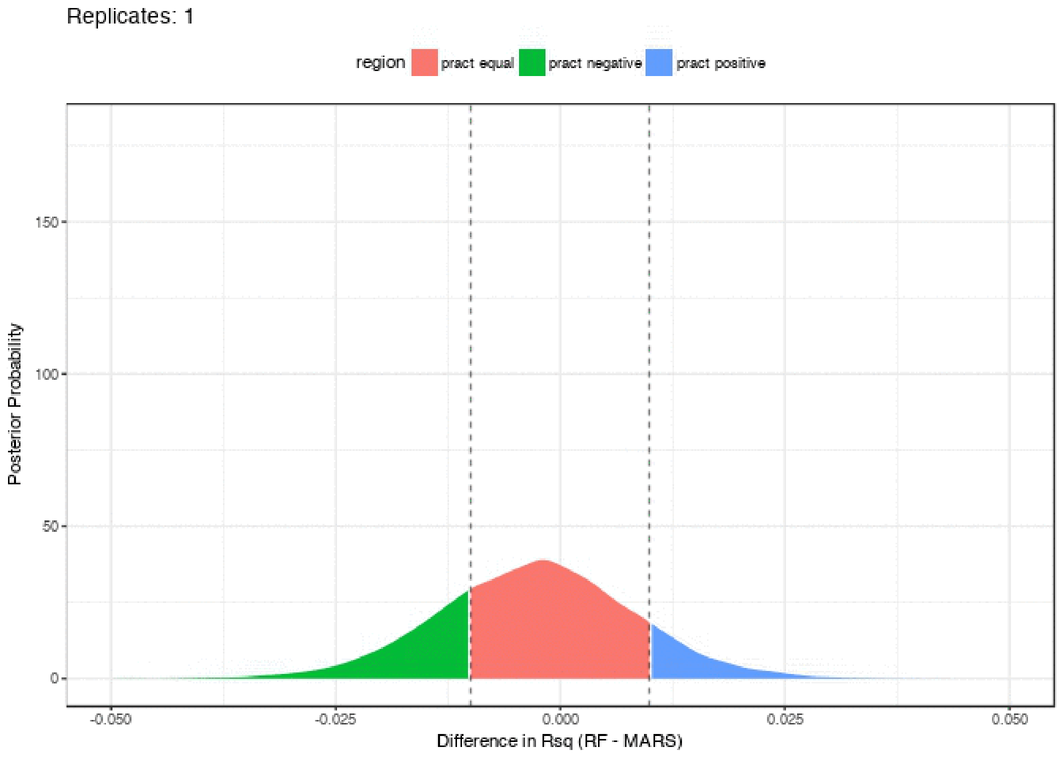
There is a 11.7% probability that RF is *better* and a 24.5% change the *MARS* is superior.



The Effect of the Number of Resamples

To some extend, the quality of the posterior estimates are driven by the amount of resampling done.

What happens if we do *repeated* 10-fold cross-validation?



The Bayesian analysis was do using `tidyposterior` based on model results produced by `rsample`. The results from `caret` can also be used with `tidyposterior`.

Thanks for sticking around!