# **HW 9**

*Due: May 7th at 3:00 pm*

1. In this exercise we will use trees to predict `Salary` in the `Hitters` data set from the package `ISLR`.

   a) Remove the observations for whom the salary information is unknown, and then log-transform the salaries. Report the average log-transformed salary for verification.

   b) Create a training set consisting of the first 200 observations, and a test set consisting of the remaining observations. Again, report the average log-transformed salary in each set for verification.

   c) Use the training dataset to grow a regression tree to predict `Salary` based on all other variables. Plot the tree and report its residual mean deviance. Obtain the RMSE of the tree when applied on the test dataset.

   d) Prune the tree using CV. Obtain a chart of size vs. deviance. Select a smaller size based on the graph and prune the original tree. Plot the pruned tree and obtain its RMSE on the test dataset.

   e) Perform bagging using 1000 trees. Calculate RMSE on the test dataset.

   f) Perform random forest with different number of variables for `mtry`. Calculate test set RMSE for each `mtry` and plot RMSE vs. number of variables.

   g) Which variables appear to be the most important in the random forest?

   h) Compare and discuss the performance of all of the above methods.