Senior Project Start

Michael Streyle

September 28th, 2018

# Beginning

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

## Loading required package: lattice

## Loading required package: ggplot2

##   
## Attaching package: 'ggplot2'

## The following object is masked from 'package:randomForest':  
##   
## margin

# Introduction

The data used in this analysis is a collection of 310 observations of spinal measurements. The objective is to classify each observation as normal or abnormal using machine learning modeling. As seen in the pairs plot, the last 6 variables are essentially random noise, so they will be left out of the initial models. Later, I will use these random variabels to compare how the models handle extra noise.

# Initial Exploration

A histogram of each of the first 6 predictor variables reveals that they are each relatively normal. A histogram of the other 6 predictor variables confirms that they are not normally distributed but rather somewhat uniformly distributed. This aligns with the idea that the last 6 are random variables. There are no missing values in this data.

my\_cols <- c("#00AFBB", "#E7B800", "#FC4E07")  
  
pairs(data, col = my\_cols[data$classification]) #pairs plot



