Motivation

- Goal: To predict the number of goals scored in a soccer game based on actions that were taken during the game, such as shots, tackles, corner kicks, dribbles, etc.

- Motivation: Soccer analytics is still in its relative infancy. There are many events that occur during a game, and it can be difficult to assign value to any one action, since there are typically few goals scored during the average game. The motivation is to see how the selected features can predict the number of goals scored in a game.

- Dataset: Match statistics from the top 5 European soccer leagues(France, England, Italy, Germany, Spain), from 2012-2017, hosted by Kaggle.

Code Snippets and Explanation

In [24]:
```
assembler = VectorAssembler(
    inputCols=["homeDribbleSuccessFT", "homePassesKeyFT", "homeCornersTotalFT", "homeShotsTotalFT",
               "homeTacklesTotalFT", "homeAerialsTotalFT", "homeShotsOnTargetFT", "homeOffsidesCaughtFT"],
    outputCol="features")

vectrainDF = assembler.transform(trainingDF)
vectestDF = assembler.transform(testDF)
```

Code used to preprocess the data into the correct form for MLlib. Utilized the VectorAssember functionality in Pyspark.

In [27]:
```
#create gradient boosted tree regressor
gbt = GBTRegressor(featuresCol = 'features', labelCol = 'label',maxIter=10)
```

The algorithm used to build the model, a gradient boosted decision tree. ”features” is the vector created using VectorAssembler.

Visualization