# Motivation (by: Nick Bruno)

- Compare the productivity of the two highest scoring guards on the Washington Wizards, John Wall and Bradley Beal, over the last two full seasons (2016-2017 and 2017-2018)
- Compare their statistics in categories such as Assists, Turnovers, Field Goal Percentage, Plus-Minus statistic, and Rebounds
- Investigate trends within the overall data between the two Wizard guards
- Predict the number of turnovers a Wizard guard will make in a game given the number of assists they have.

# Machine Learning Code

```
# I will choose assists (AST) and turnovers (TOV)
df_analysis = df.select("AST","TOV")

# AST = feature, TOV = label
```

```
# split into train/test sets
seed = 8
(testDF, trainingDF) = df_analysis.randomSplit((0.20, 0.80), seed=seed)
print ('training set N = {}, test set N = {}'.format(trainingDF.count(),testDF.count()))
```

```
from pyspark.ml.linalg import Vectors, VectorUDT
```

```
# make a user defined function (udf)
sqlc.registerFunction("oneElementVec", lambda d: Vectors.dense([d]), returnType=VectorUDT())

# vectorize the data frames
trainingDF = trainingDF.selectExpr("TOV", "oneElementVec(AST) as AST")
testDF = testDF.selectExpr("TOV", "oneElementVec(AST) as AST")

print(testDF.orderBy(testDF.TOV.desc()).limit(5))
```

```
# rename to make ML engine happy
trainingDF = trainingDF.withColumnRenamed("TOV", "label").withColumnRenamed("AST", "features")
testDF = testDF.withColumnRenamed("TOV", "label").withColumnRenamed("AST", "features")
```

```
from pyspark.ml.regression import LinearRegression, LinearRegressionModel

lr = LinearRegression()
lrModel = lr.fit(trainingDF)
```

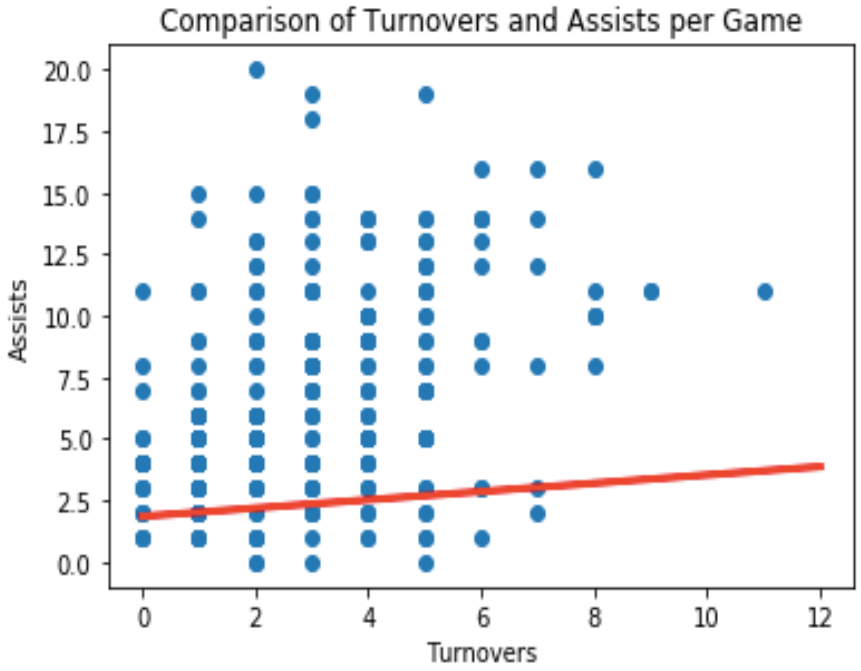
```
predictionsAndLabelsDF = lrModel.transform(testDF)

print(predictionsAndLabelsDF.orderBy(predictionsAndLabelsDF.label.desc()).take(5))
```

This code shows how I chose to compare Assists (feature) and Turnovers (label). I split the data into training and testing sets. From there, I fit a linear model to my training data and used this model to make predictions on the testing dataset. Results:

| label | features | prediction |
|---|---|---|
| 2.0 | [0.0] | 1.8457641468842392 |
| 5.0 | [0.0] | 1.8457641468842392 |
| 1.0 | [1.0] | 2.0143585496638803 |
| 2.0 | [1.0] | 2.0143585496638803 |
| 5.0 | [1.0] | 2.0143585496638803 |
| 0.0 | [2.0] | 2.1829529524435216 |
| 1.0 | [2.0] | 2.1829529524435216 |
| 2.0 | [2.0] | 2.1829529524435216 |

# Machine Learning Visualization




As we can see, there is not a strong linear correlation between Turnovers and Assists per game, as these data points include all of the relevant data from John Wall and Bradley Beal over the past two seasons. However, there is a slight positive direction of the line, indicating that the more turnovers a Wizard's superstar guard has, the more assists they are likely to have in the game as well.