NFL Big Data Bowl

Penn Students contribute to the NFL data revolution
State of Statistics at Penn

- Statistics is not a major but it is studied by 100s of students across schools:
  - Wharton
  - College
  - Engineering
- 100s of Wharton Statistics concentrators every year
- 50 and growing number of Stat minors every year
- Growing number of students in Engineering Data Science minor in
Sports Research at Penn

- Group of 20 Penn Students who meet every week
- Work on Sports related analytics problems
  - NFL
  - MLB
  - NCAAB and NBA
  - NHL
- Prepare work for presentations, school projects, and eventual publication
The inaugural analytics contest explores statistical innovations in football — how the game is played and coached.
Set Up

- **Two Divisions**
  - Students - Undergrads, MBA, Masters, and PhD’s
  - Open - Professional Data Scientists in other fields

- **Time Frame not ideal**
  - Competition released over winter break
  - 4 days before submission after returning to school

- **Data immensely complicated**
  - Classic Big Data Problem
  - High resolution video data
  - We’ve never worked with video data before
The Team

Jake Flancer
Eric Dong
Andrew Castle
Jack Soslow
Adi Wyner
The Ask

- Evaluating Player Speed
  - Are there better ways to track speed than just acceleration and MPH?

- Optimizing Receiver Routes
  - What are the best routes to run on any given play?

- Rule Change
  - Based off player-tracking data, should the NFL consider a new rule?
The Data

- First 6 weeks of games in 2017
- Player-Tracking Data
- Play Level Data
- Game Level Data
- Player Level Data
## Player-Tracking Data

<table>
<thead>
<tr>
<th>time</th>
<th>x</th>
<th>y</th>
<th>s</th>
<th>dis</th>
<th>dir</th>
<th>event</th>
<th>nflId</th>
<th>displayName</th>
<th>jerseyNumber</th>
</tr>
</thead>
<tbody>
<tr>
<td>2017-09-08 00:41:59</td>
<td>41.56</td>
<td>16.54</td>
<td>3.91</td>
<td>0.41</td>
<td>78.90</td>
<td>NA</td>
<td>2495340</td>
<td>Anthony Sherman</td>
<td>42</td>
</tr>
<tr>
<td>2017-09-08 00:41:59</td>
<td>41.95</td>
<td>16.62</td>
<td>4.28</td>
<td>0.40</td>
<td>79.16</td>
<td>NA</td>
<td>2495340</td>
<td>Anthony Sherman</td>
<td>42</td>
</tr>
<tr>
<td>2017-09-08 00:41:59</td>
<td>42.40</td>
<td>16.73</td>
<td>4.66</td>
<td>0.47</td>
<td>79.46</td>
<td>NA</td>
<td>2495340</td>
<td>Anthony Sherman</td>
<td>42</td>
</tr>
<tr>
<td>2017-09-08 00:41:59</td>
<td>42.85</td>
<td>16.82</td>
<td>5.04</td>
<td>0.46</td>
<td>79.76</td>
<td>NA</td>
<td>2495340</td>
<td>Anthony Sherman</td>
<td>42</td>
</tr>
<tr>
<td>2017-09-08 00:41:59</td>
<td>43.36</td>
<td>16.92</td>
<td>5.39</td>
<td>0.51</td>
<td>80.12</td>
<td>kickoff</td>
<td>2495340</td>
<td>Anthony Sherman</td>
<td>42</td>
</tr>
<tr>
<td>2017-09-08 00:41:59</td>
<td>43.87</td>
<td>17.02</td>
<td>5.60</td>
<td>0.52</td>
<td>80.59</td>
<td>NA</td>
<td>2495340</td>
<td>Anthony Sherman</td>
<td>42</td>
</tr>
</tbody>
</table>

Data is gigantic, high-resolution, but in a spreadsheet like any other dataset.
Data Exploration - Starting Positions of Receivers
### Examples of Routes

**All routes run by Alshon Jeffery vs the Kansas City Chiefs**

<table>
<thead>
<tr>
<th>+70</th>
<th>+60</th>
<th>+50</th>
<th>+40</th>
<th>+30</th>
<th>+20</th>
<th>+10</th>
<th>LOS</th>
</tr>
</thead>
</table>

**Status**
- Route
- Ball in Air
- After Reception
  - Play Finishes

**Passing Play Routes**
- Empty Route
- Targeted
- Reception
Problem
How can we optimize routes so that we can increase expected yardage in any situation?

1. Shape Based Clustering
   Turn x,y coordinates of every player at every moment into usable receiver routes.

2. Machine Learning
   Combine situational data with route information to predict Yards.
Shape-Based Clustering
Shape Based Clustering
Shape-Based Clustering: Example Routes

10 Yard Crossing Route

WR 71%  TE 24%  RB 5%

Predicted Routes in Group 6

RB Out Route

WR 5%  TE 5%  RB 90%

Predicted Routes in Group 2
Shape Based Clustering

Odell Beckham

Rob Gronkowski

Ezekiel Elliott

Odell Beckham's 5 most common routes

Outside Curl (Left Side, 10 Yds) 0.17
Outside Curl (Right Side, 10 Yds) 0.14
Outside Hitch (Left Side, 5 Yds) 0.11
Slant (Right Side, 10 Yds) 0.08
Inside Curl (Left Side, 10 Yds) 0.06

Percent Run

Rob Gronkowski's 6 most common routes

Hitch Out (Right Side, 5 Yds) 0.24
Seam (Left Side, 15 Yds) 0.09
Hitch Out (Left Side, 5 Yds) 0.08
Curl In (Middle, 10 Yds) 0.08
Curl In (Left Side, 10 Yds) 0.05
Blocking 0.05

Percent Run

Ezekiel Elliott's 5 most common routes

RB Out (Right Side, 0 Yds) 0.2
RB Out (Left Side, 0 Yds) 0.2
Blocking 0.17
Blocking / Checkdown 0.12
Short Curl Checkdown (Middle, 3 Yds) 0.11

Percent Run
Two Stage Approach

1. Likelihood of Completion
   - Accuracy 71%
   - AUC 0.75

2. Yards Gained Given Completion
   - Cor 0.51
   - RMSE 10.0

Situational Variables
- Seconds Remaining in Game
- Yard Line
- Down and Distance
- Score Difference
- Offensive Formation
- # of Pass Rushers
- Quarterback

Engineered Variables
- The routes run on the play
- Position (WR, TE, etc...) of the player running the route
Conclusion - Importance of Routes on Predicted Yards

Broncos vs Bills

Predicted Yards = 5.8
Actual Yards = 2

Ravens vs Raiders

Predicted Yards = 10.5
Actual Yards = 52
Conclusions - Optimizing Routes to Improve Yardage

Play Call Change

Change TE (82) from a blocking route, to a Hitch Route

Comp % ↑30%
Yards Given Comp ↓.1 yards
Predicted Yards ↑1.65 yards
Quick Insights

Run Short 5 Yard Hitches

Completion % +39%
Yards +.15

Good Routes explain more variability than Good Quarterbacks

Top 20 Most Important Factors by Model

<table>
<thead>
<tr>
<th>Completion %</th>
<th>Yards Given Completion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Routes</td>
<td>QB</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
</tr>
</tbody>
</table>
The Experience
Thank You

Jake Flancer - jflancer@wharton.upenn.edu, jakef1873
Jack Soslow - jsoslow2@gmail.com, jack_soslow
Andrew Castle - castla@wharton.upenn.edu, AndrewCastle510
Eric Dong - ericdong@seas.upenn.edu

Special thanks to Professor Abraham J. Wyner of the Wharton School for his advice and assistance, and to Michael Lopez and Jay Reid.

References:

First Attempt
Time series clustering worked, but didn’t accurately cluster routes

Time series clusters for one game
Second Attempt

Auto-Encoding routes helped grab features, but not useful for analysis

Examples of auto-encoded routes