NFL Big Data Bowl

Penn Students contribute to the NFL data revolution

BIG DATA BOWL

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

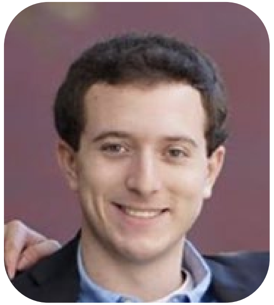
- Group of 20 Penn Students who meet every week
- Work on Sports related analytics problems
    - NFL
    - MLB
    - NCAAB and NBA
    - NHL
- Prepare work for presentations, school projects, and eventual publication

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

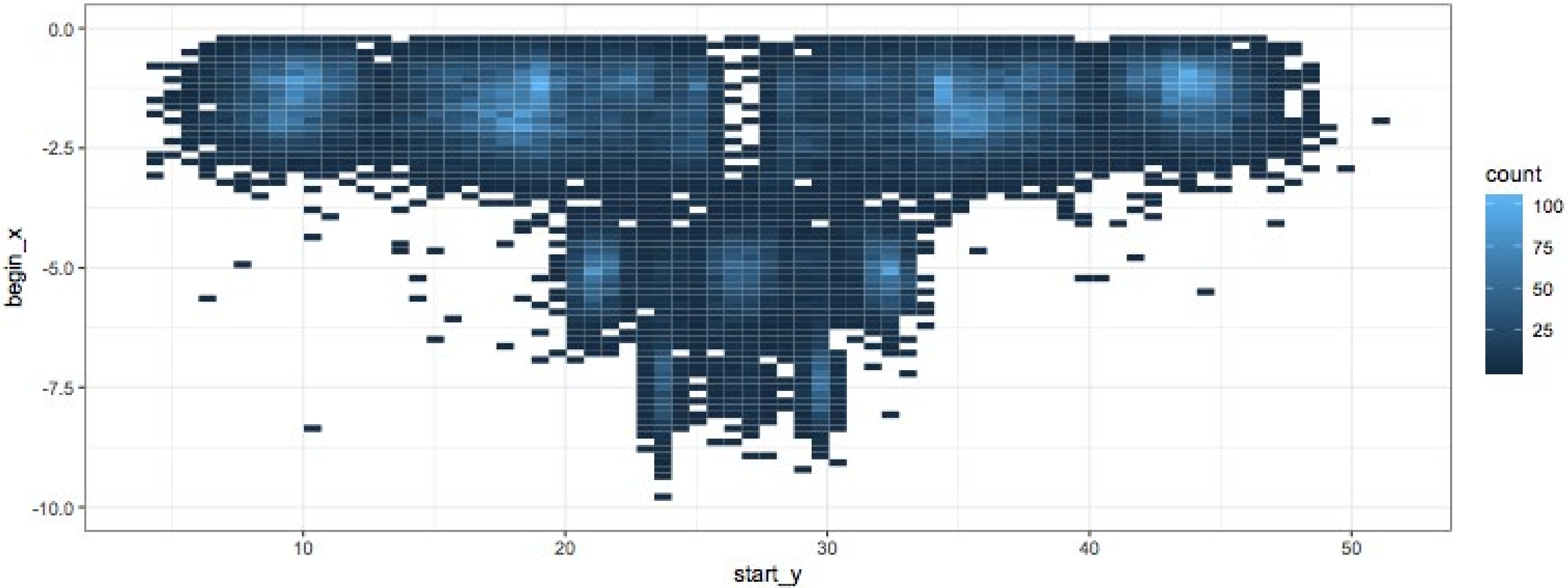
- First 6 weeks of games in 2017
- Player-Tracking Data
- Play Level Data
- Game Level Data
- Player Level Data

Player-Tracking Data

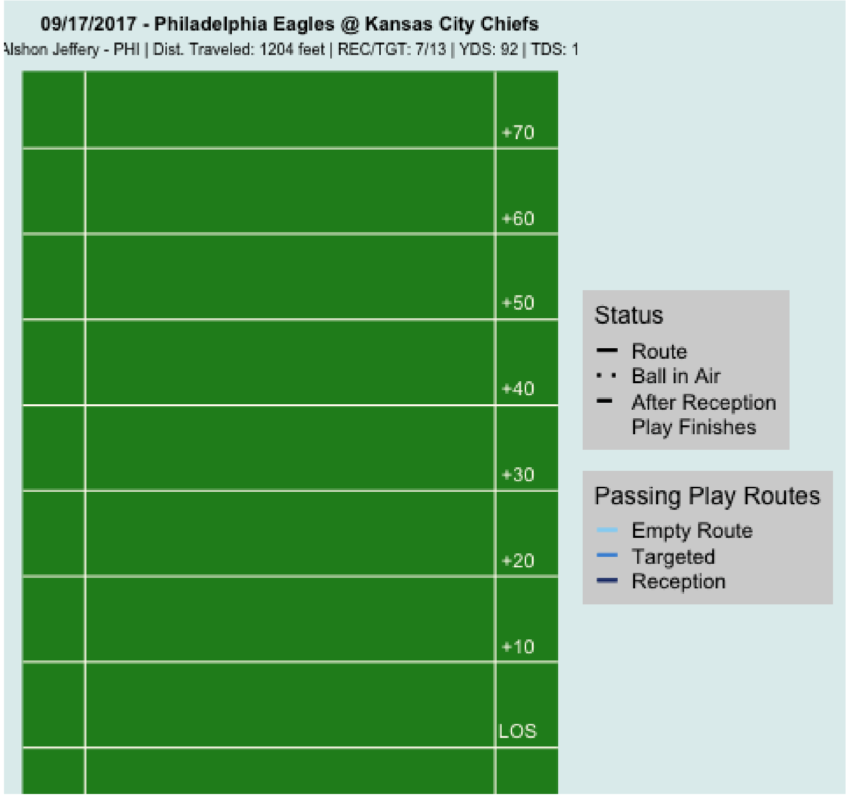
	time	x	y	s	dis	dir	event	nflId	displayName	jerseyNumber
1	2017-09-08 00:41:59	41.56	16.54	3.91	0.41	78.90	NA	2495340	Anthony Sherman	42
2	2017-09-08 00:41:59	41.95	16.62	4.28	0.40	79.16	NA	2495340	Anthony Sherman	42
3	2017-09-08 00:41:59	42.40	16.73	4.66	0.47	79.46	NA	2495340	Anthony Sherman	42
4	2017-09-08 00:41:59	42.85	16.82	5.04	0.46	79.76	NA	2495340	Anthony Sherman	42
5	2017-09-08 00:41:59	43.36	16.92	5.39	0.51	80.12	kickoff	2495340	Anthony Sherman	42
6	2017-09-08 00:41:59	43.87	17.02	5.60	0.52	80.59	NA	2495340	Anthony Sherman	42

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

Problem

How can we optimize routes so that we can increase expected yardage in any situation?

Shape Based Clustering

1

Turn x,y coordinates of every player at every moment into usable receiver routes.

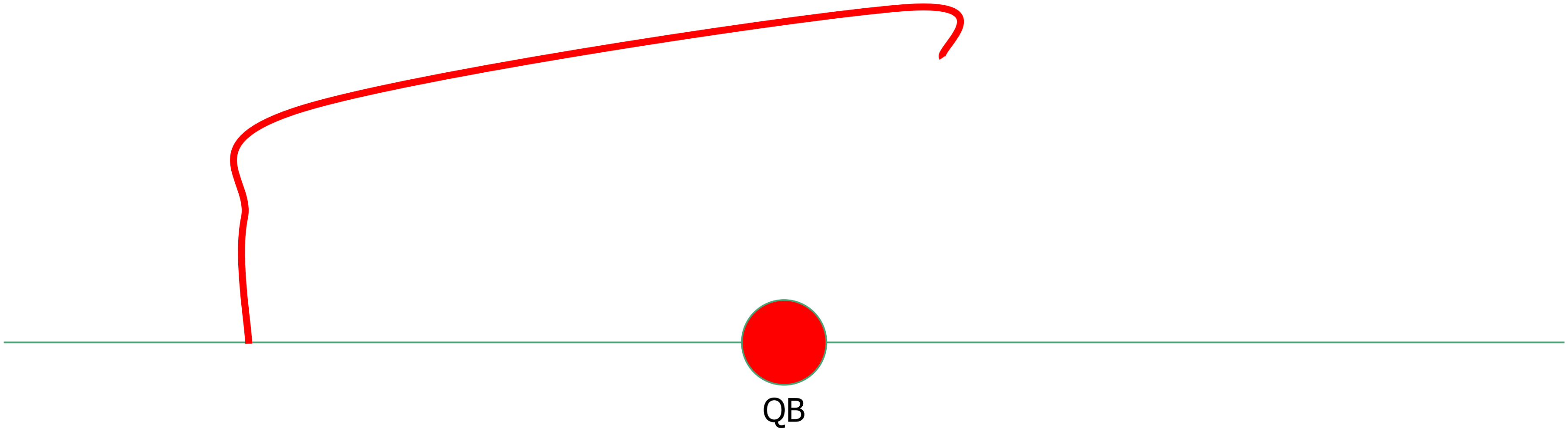
Machine Learning

2

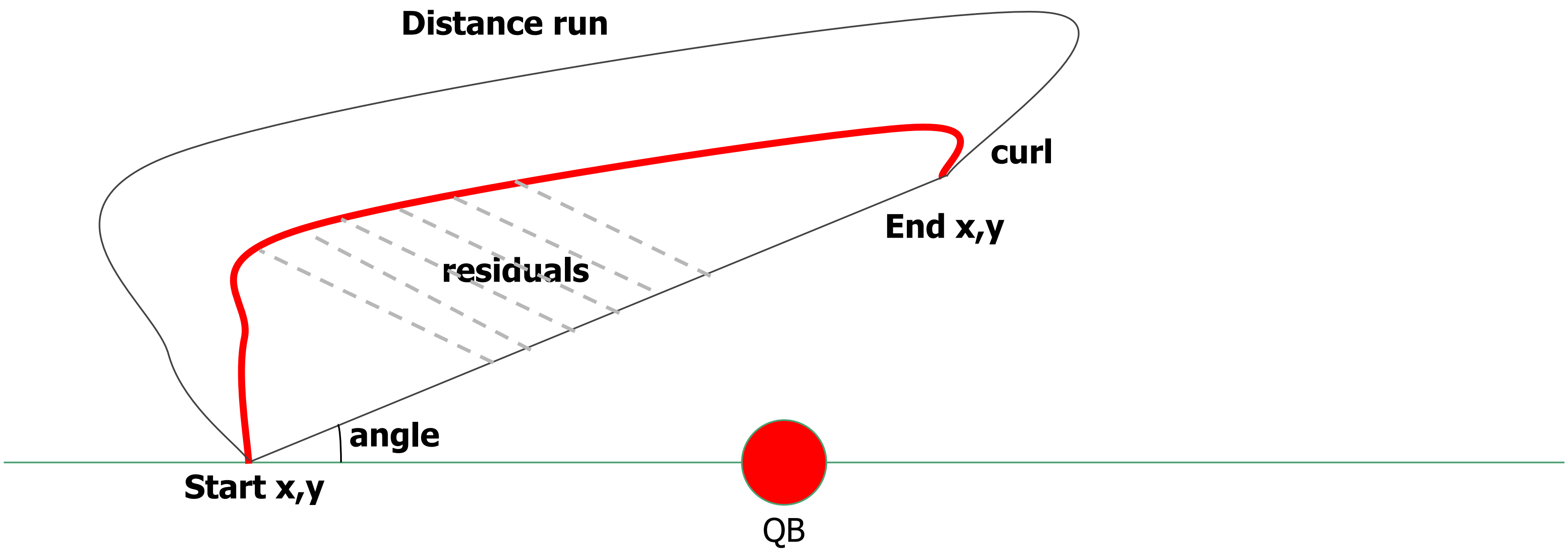
Combine situational data with route information to predict Yards.

Shape-Based Clustering

Shape Based Clustering

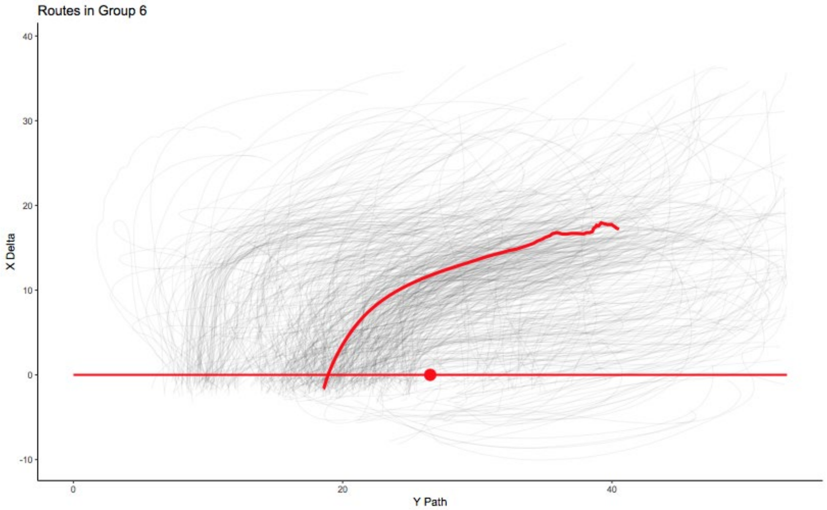


Shape Based Clustering



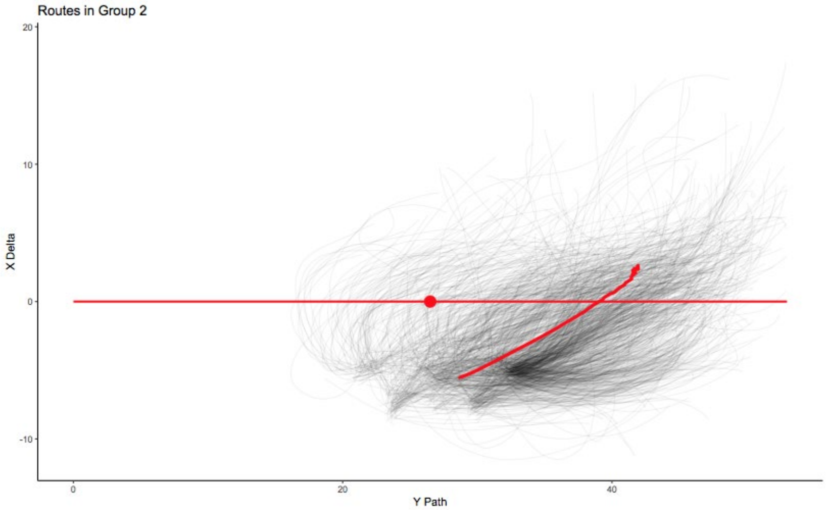
Shape-Based Clustering: Example Routes

10 Yard Crossing Route



WR **71%** TE **24%** RB **5%**

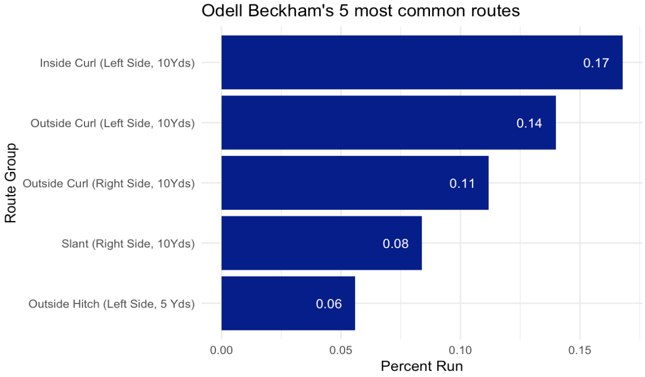
RB Out Route



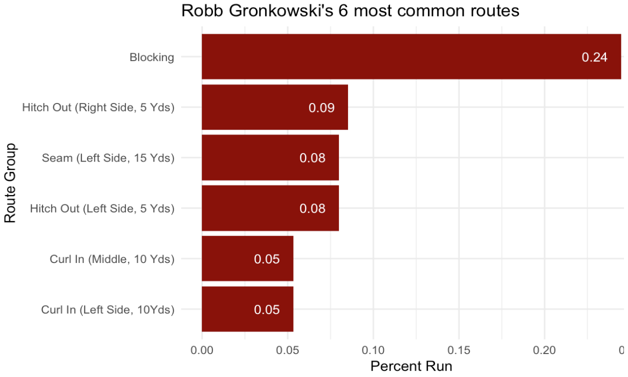
WR **5%** TE **5%** RB **90%**

Shape Based Clustering

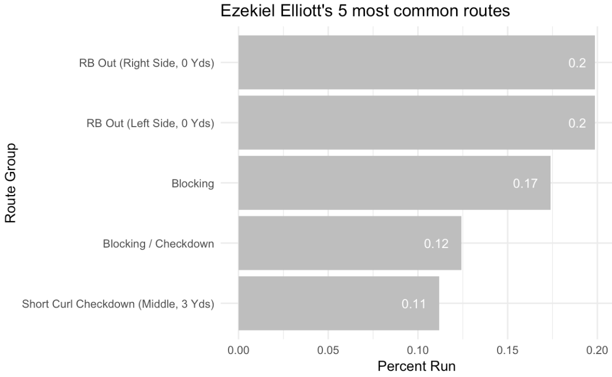
Odell Beckham



Rob Gronkowski



Ezekiel Elliott



Two Stage Approach

1

Likelihood of Completion

Accuracy **71%** AUC **.75**

2

Yards Gained Given Completion

Cor **.51** RMSE **10.0**

Situational Variables

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

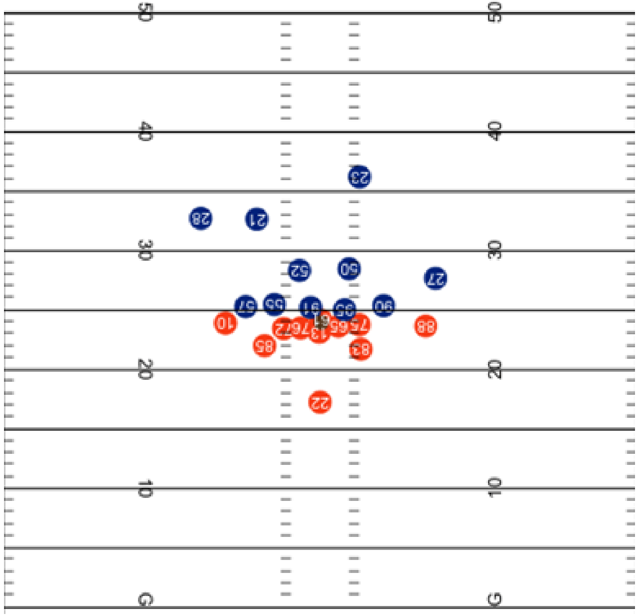
Engineered Variables

- The routes run on the play
- Position (WR,TE,etc…) of the player running the route

Conclusion - Importance of Routes on Predicted Yards

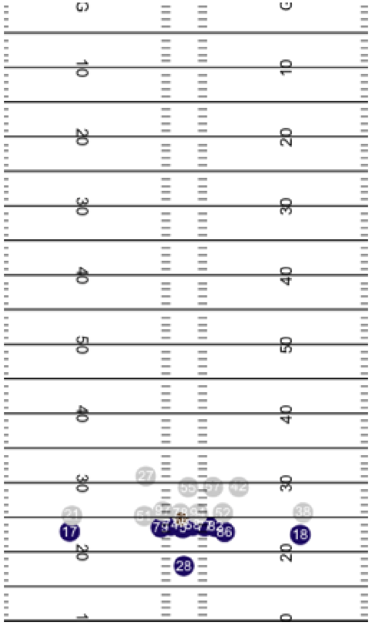
Broncos vs Bills

Predicted Yards = 5.8
Actual Yards = 2

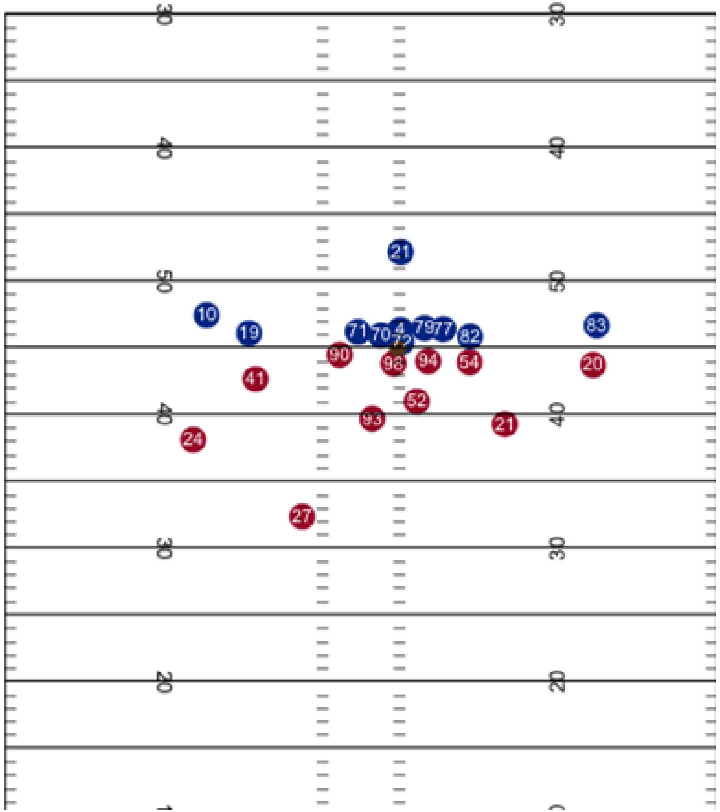


Ravens vs Raiders

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

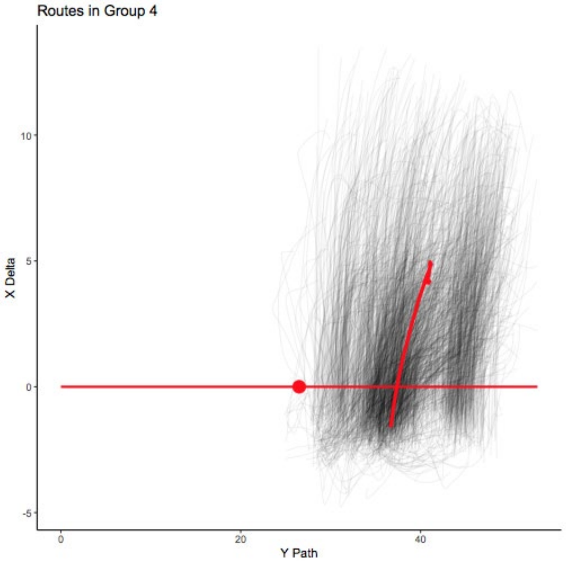
Comp % ↑30%

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

Completion %		Yards Given Completion	
Routes	QB	Routes	QB
15	0	14	0

The Experience


Wharton
NFL
FILMS


THE FUTURE
STARTS HERE

Thank You

Jake Flancer - *jflancer@wharton.upenn.edu, @jakef1873*

Jack Soslow - *jsoslow2@gmail.com, @jack_soslow*

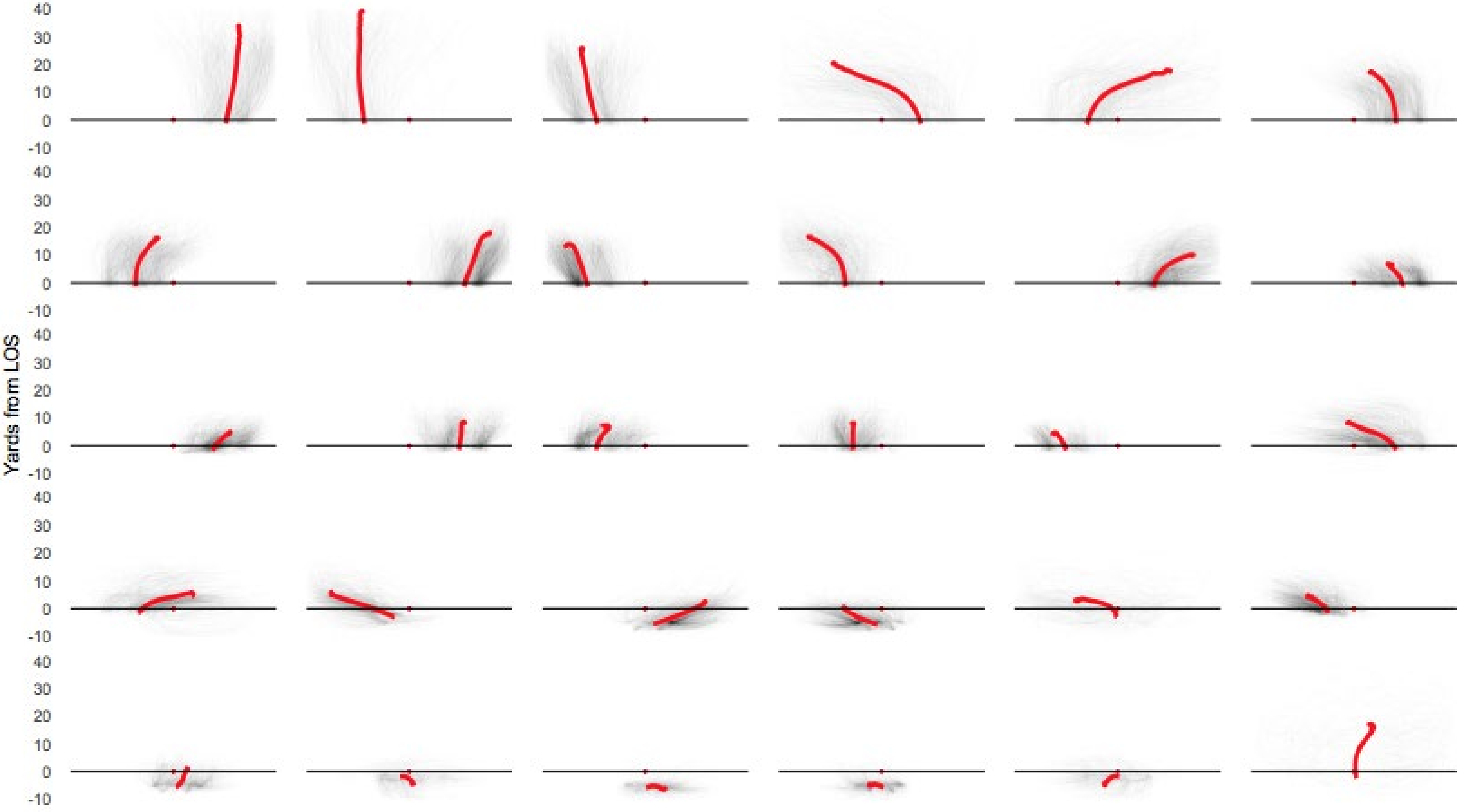
Andrew Castle - *castla@wharton.upenn.edu, @AndrewCastle510*

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.

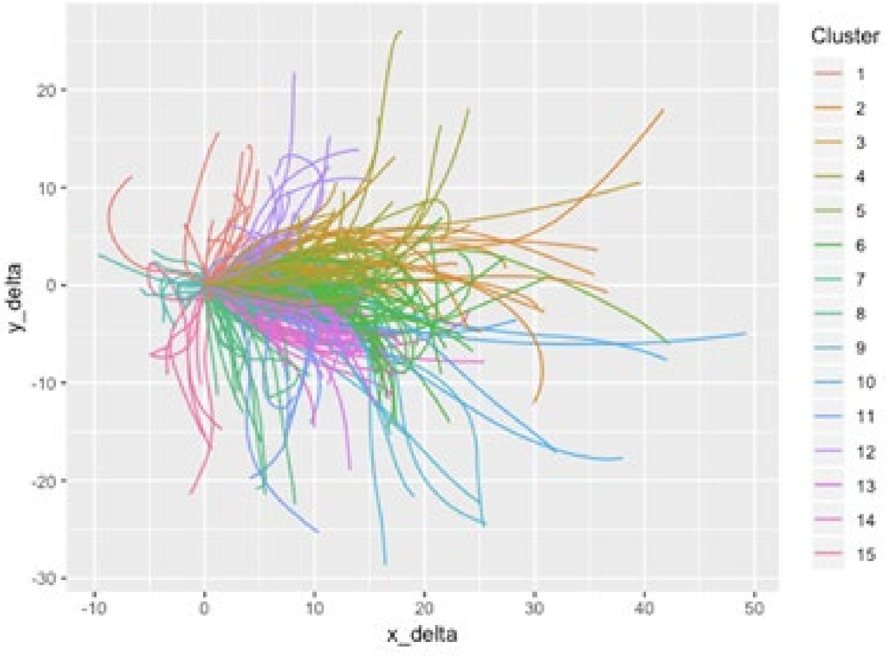
References:

1. Keogh, Eamonn, and Jessica Lin. "Clustering of time-series subsequences is meaningless: implications for previous and future research." Knowledge and information systems 8.2 (2005): 154-177.
2. Steinbach, Michael, Levent Ertöz, and Vipin Kumar. "The challenges of clustering high dimensional data." New directions in statistical physics. Springer, Berlin, Heidelberg, 2004. 273-309.
3. Vincent, Pascal, et al. "Extracting and composing robust features with denoising autoencoders." Proceedings of the 25th international conference on Machine learning. ACM, 2008.


Yards from LOS
40
30
20
10
0
-10

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