An Overview of Data Science Problems in the Sports Industry

Brian Macdonald

2/12/2021

Twitter/LinkedIn:
@bmacGTPM

Background

Special Faculty in Sports Analytics, Carnegie Mellon University

Previously:

- Director of Sports Analytics, ESPN
- Director of Hockey Analytics, Florida Panthers
- Associate Professor, USMA, West Point (taught Diff EQ!)
- Ph.D. Mathematics, Johns Hopkins University

Overview

- Why Sports Analytics?

- Data Science Problems in Sports
    - Teams (business analytics and sports analytics)
    - Media (sports analytics)
    - Leagues (sports analytics)
    - System of Difference Equations
        - Close enough!

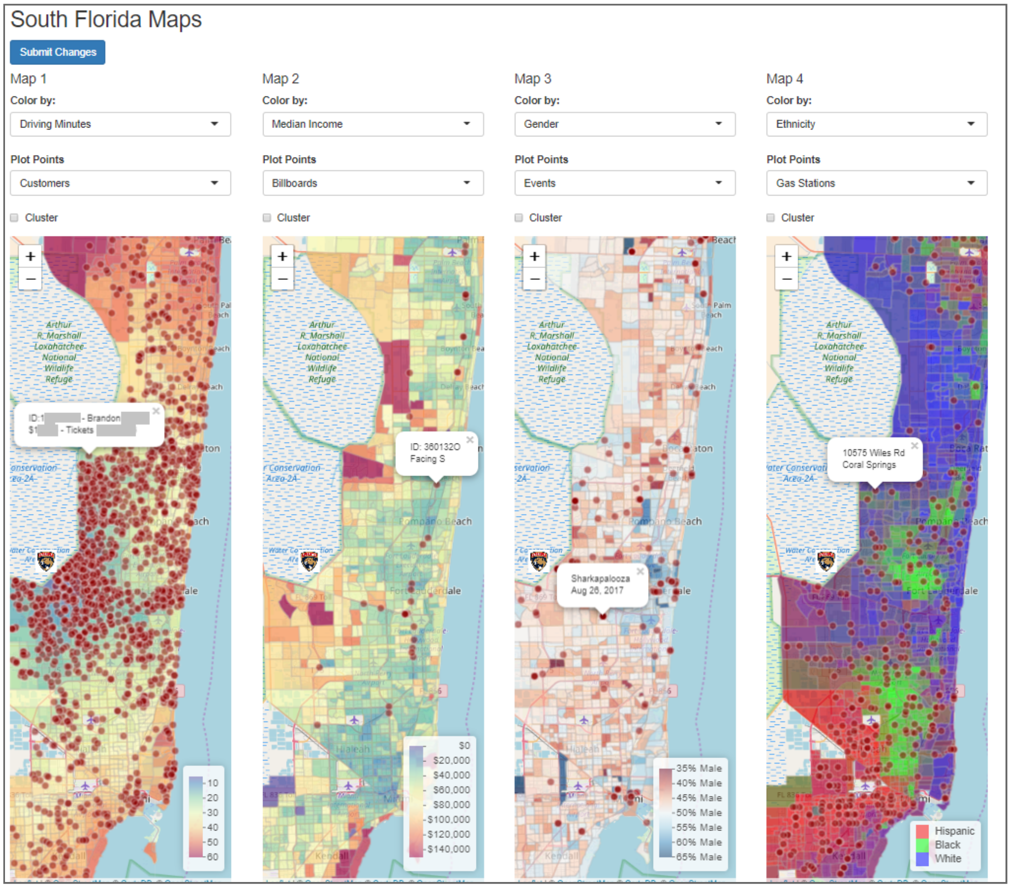
Sports analytics in education

Why sports analytics?

- TONS of public data, freely available to anyone
- Wide variety of data, problems, and methods
- Problems are analogous to those in non-sports applications.
    - Experience translates.
- Sports are a controlled environment.
- Sports are widely popular
    - In 2019, 154.4 million U.S. viewers watched live sports at least once per month
    - Many students start as subject-matter experts.
- Real-life validation

Business Analytics

- Predicting attendance. How much demand is there for a game? Based on
  - Day of Week
  - Month
  - Opponent
  - etc
- Customer analysis. Who are the buyers, where do they live, and why do they buy?
  - Internal data
  - US Census data
  - Google Maps data, driving time



South Florida Maps

Submit Changes

Map 1

Color by: Driving Minutes

Plot Points: Customers

□ Cluster

Map 2

Color by: Median Income

Plot Points: Billboards

□ Cluster

Map 3

Color by: Gender

Plot Points: Events

□ Cluster

Map 4

Color by: Ethnicity

Plot Points: Gas Stations

□ Cluster

ID:1[redacted] - Brandon [redacted]
$1[redacted] - Tickets [redacted]

ID: 360132O
Facing S

Sharkapalooza
Aug 26, 2017

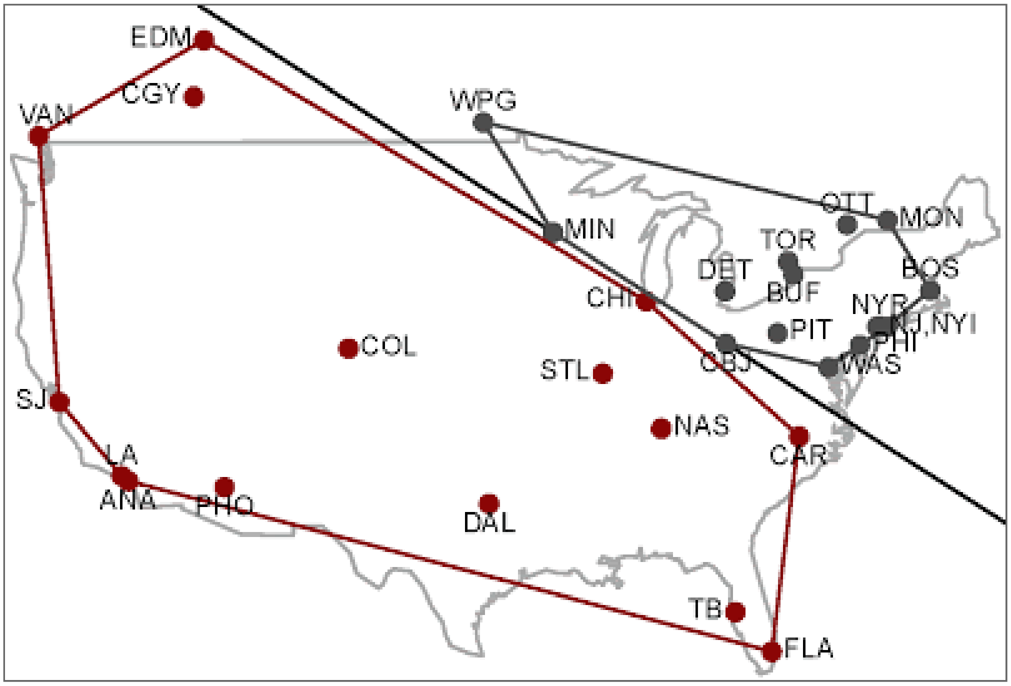
10575 Wiles Rd
Coral Springs

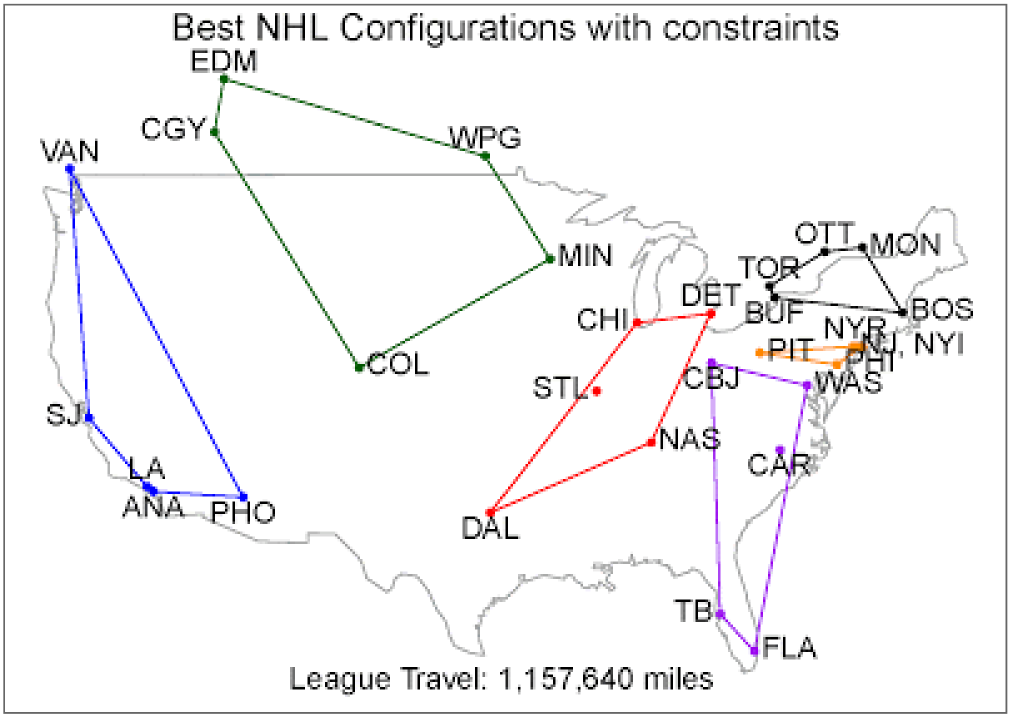
Realignment

Question: What realignment would minimize league travel?

If teams in the same division play each other more often, realignment matters.

Candidate conferences




Best NHL Configurations with constraints
EDM
CGY
WPG
VAN
MIN
OTT
MON
TOR
DET
BUF
BOS
CHI
NYR
PIT
NJ, NYI
PHI
CBJ
WAS
STL
COL
NAS
SJ
CAR
LA
ANA
PHO
DAL
TB
FLA
League Travel: 1,157,640 miles

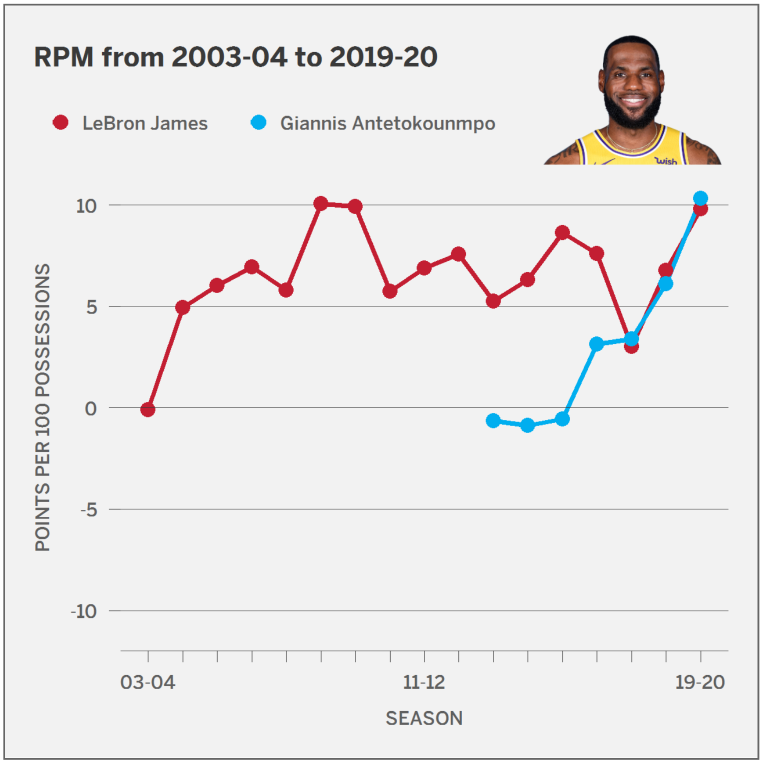
Player ratings

Teams: Evaluate player performance, player personnel decisions.

Media: Discuss those decisions

NBA Real Plus-Minus: a statistic for NBA players that

- estimates each player’s contribution to his team,
- on offense and defense
- in the unit of Points per 100 Possessions
- while accounting for his teammates and opponents.


RPM from 2003-04 to 2019-20
● LeBron James
● Giannis Antetokounmpo
10
5
0
-5
-10
POINTS PER 100 POSSESSIONS
03-04
11-12
19-20
SEASON

Team Ratings

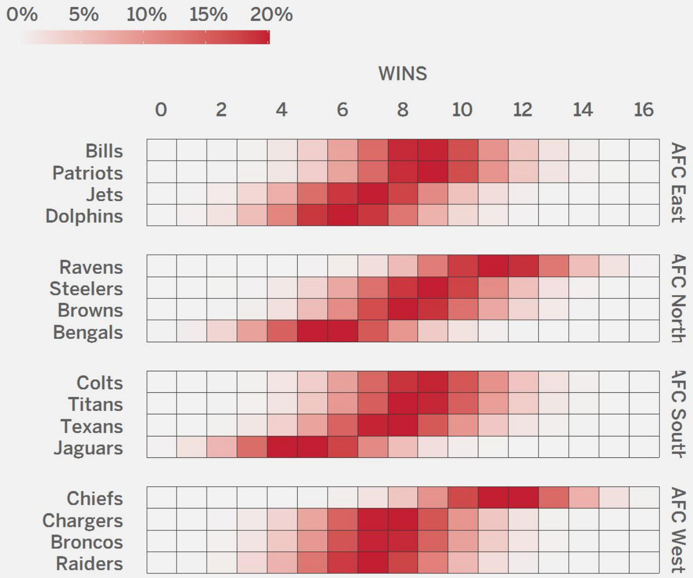
Teams: What is our season outlook? How likely is it that we’ll make the playoffs?

Media: Same.

Team ratings used for

- Game Predictions
- Season Simulations
- Expected Win Totals, Prob(Make the Playoffs)
- Betting metrics

Chance That AFC Teams Will Reach Each Win Total



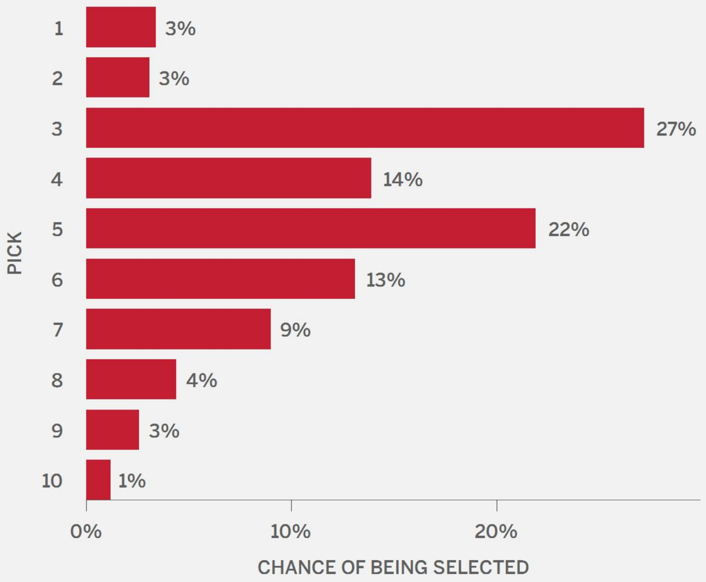
According to ESPN's NFL FPI

2020 NFL Draft Projections

Model that estimates, for each player and each pick number, the probability that a player will be selected at that pick.

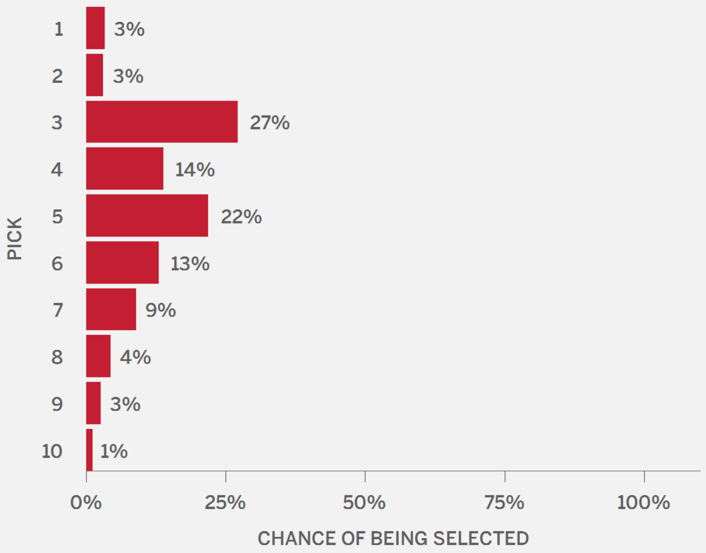
How likely is it that Player A will be available at Pick X?

Tua Tagovailoa: Chance To Be Selected At Each Pick



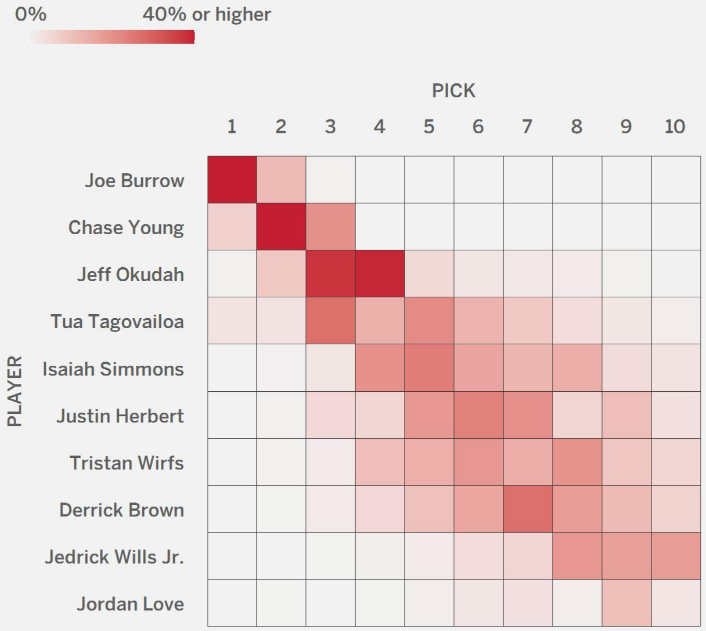
According to ESPN's NFL Draft Predictor

Tua Tagovailoa: Chance To Be Selected At Each Pick After Pick 0



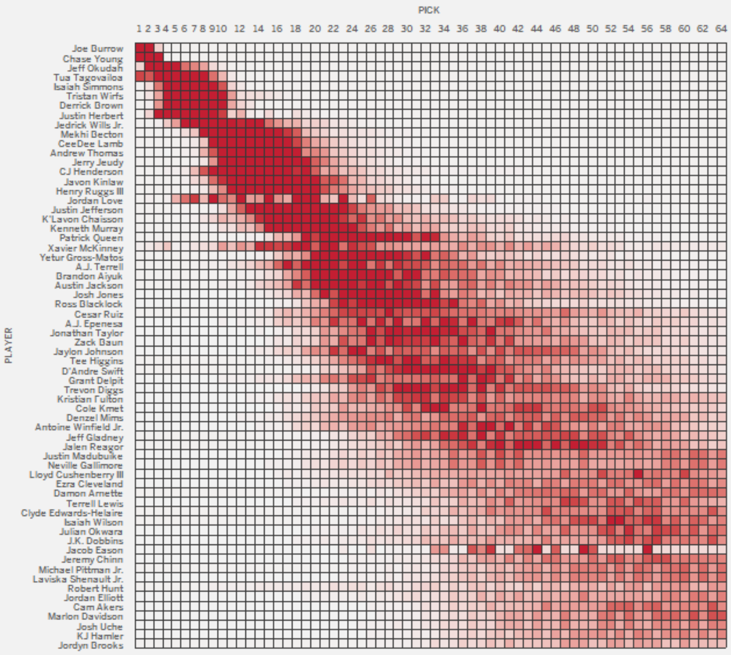
According to ESPN's NFL Draft Predictor

Chance That Top Remaining Players Will Be Selected At Each Pick



According to ESPN's NFL Draft Predictor

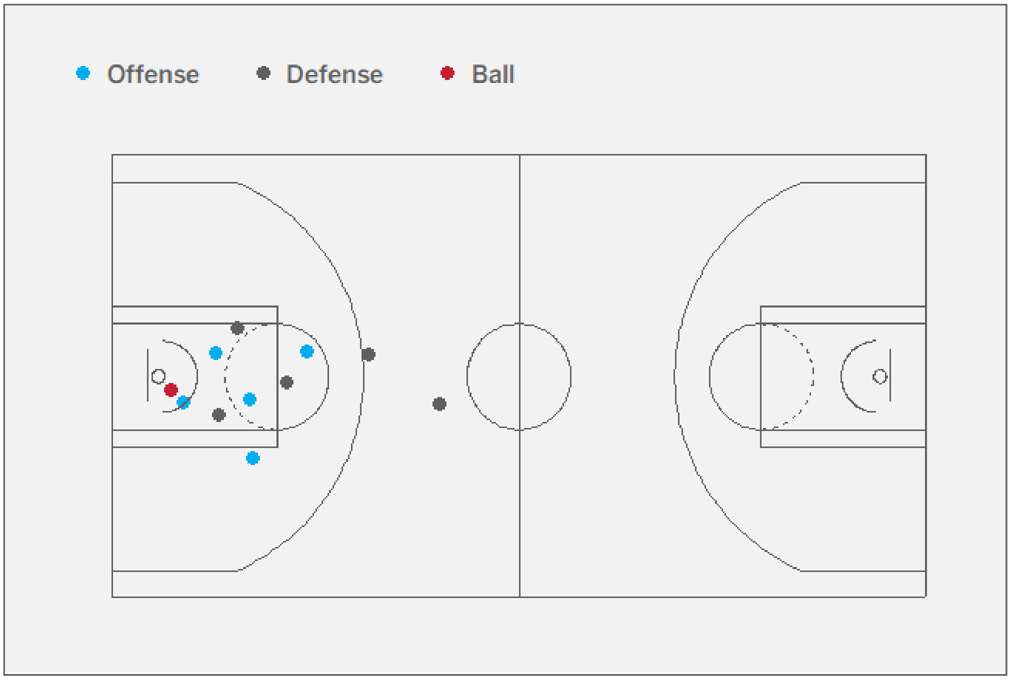
Chance That Top Remaining Players Will Be Selected At Each Pick, After Pick 0



According to ESPN's NFL Draft Predictor

Player tracking data

- player and ball locations several times per second throughout the game
- spatio-temporal information is essential for analyzing game play

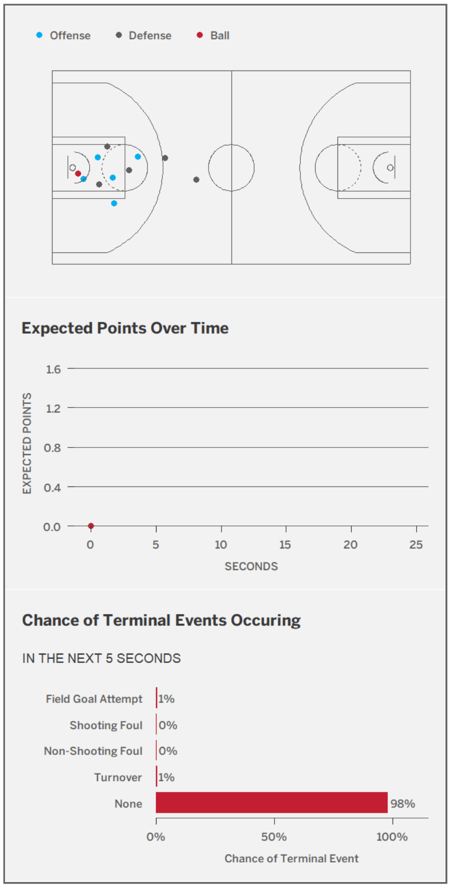

● Offense
● Defense
● Ball

Expected Points (Basketball)

Given the locations of the players and the ball

- What is the expected number of points the team will score in the current possession?
  - Cervone, D., D’Amour, A., Bornn, L., & Goldsberry, K. (2014, 2016)
  - Google: Cervone Expected Points

- What is the probability of a field goal attempt, shooting foul, non-shooting foul, turnover, or none of the above, within the next 5 seconds?
  - Sicilia, A., Pelechrinis, K., & Goldsberry, K. (2019)
  - Google: DeepHoops



Expected Points Over Time

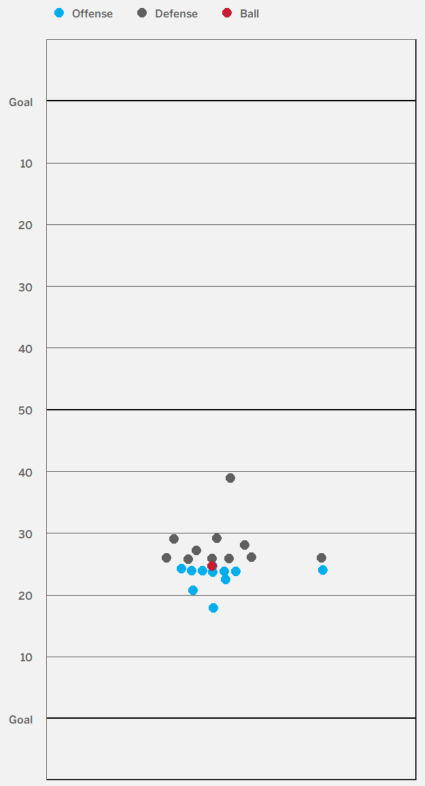
Chance of Terminal Events Occuring

IN THE NEXT 5 SECONDS

Expected Points (Football)

- What is the range of possible outcomes, and how likely they are to occur?
- Yurko, R., Matano, F., Richardson, L. F., Granered, N., Pospisil, T., Pelechrinis, K., & Ventura, S. L. (2020)
- Google: Yurko Going Deep

Player and Ball Locations and the Distribution of End-of-Play Yard Line



More

Recreating the game: Using player tracking data to analyze dynamics in basketball and football. Harvard Data Science Review, 2(4), 12 2020. **https://hdsr.mitpress.mit.edu/pub/kxks56er**.

Google: HDSR Macdonald

Modeling offensive player movement

Steven Wu, Luke Bornn. Modeling Offensive Player Movement in Professional Basketball (2018). **http://www.lukebornn.com/papers/wu_tas_20**

Google: Wu Modeling Offense, LukeBornn.com

Accessible overview, with

- code **https://github.com/dsscollection/basket**
- data **https://github.com/dcervone/EPVDemo/blob**

System of difference equations

A player's movement on offense (in the short term) can be modeled by

$$\begin{aligned} x(t+1) &= x(t) + \alpha_x[x(t) - x(t-1)] + \eta_x(t) \\ y(t+1) &= y(t) + \alpha_y[y(t) - y(t-1)] + \eta_y(t) \end{aligned}$$

where $\eta_\cdot(t)$ is a higher order term, different for every player.

Visualize acceleration vector fields for every player

Acceleration vector field

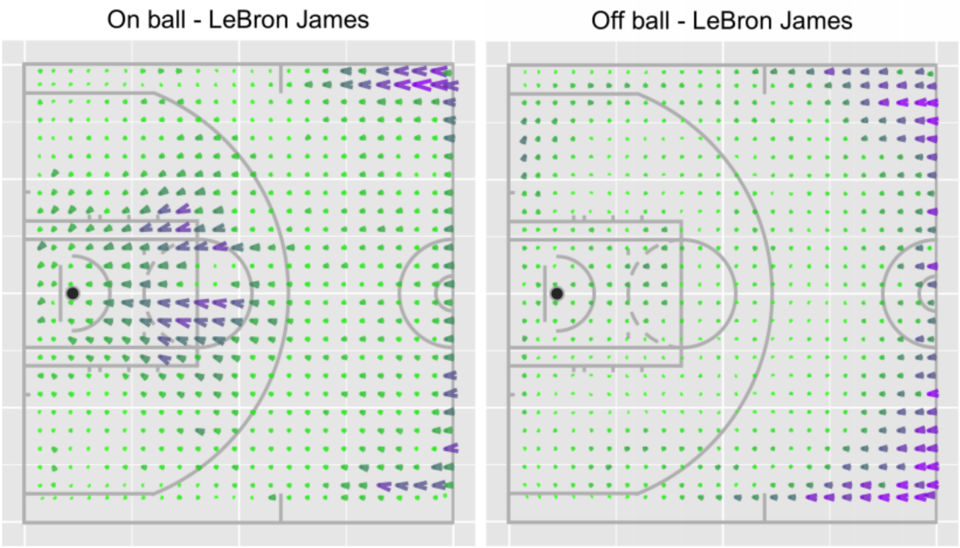


Figure 7: 2015-2016 LeBron James Empirical $\eta's$

Acceleration vector field

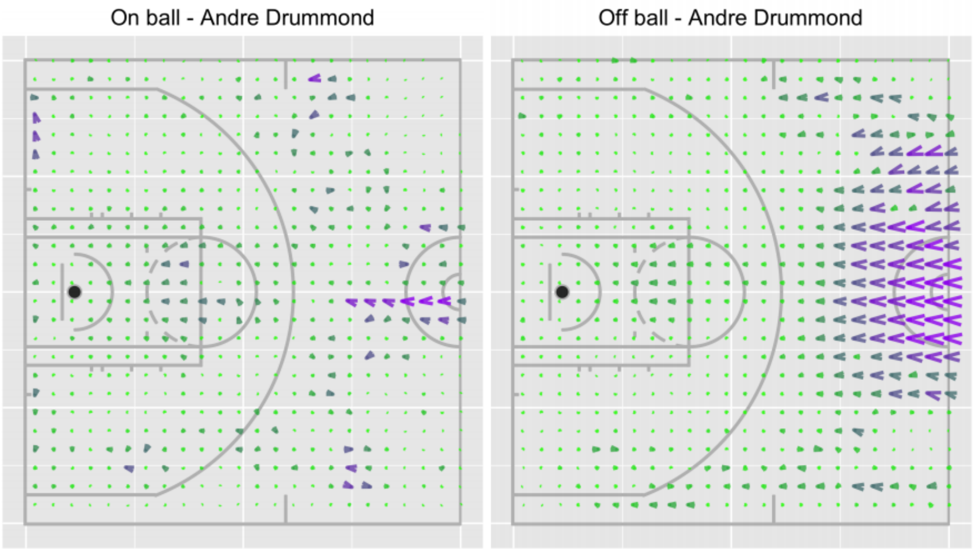


Figure 6: 2015-2016 Andre Drummond Empirical $\eta's$

Education

These projects are examples of

- Solving real problems
- Joining data from multiple sources
- Data exploration/visualization
- Multivariable thinking, need for regression or something else
- Modeling
- Interpretation
- etc

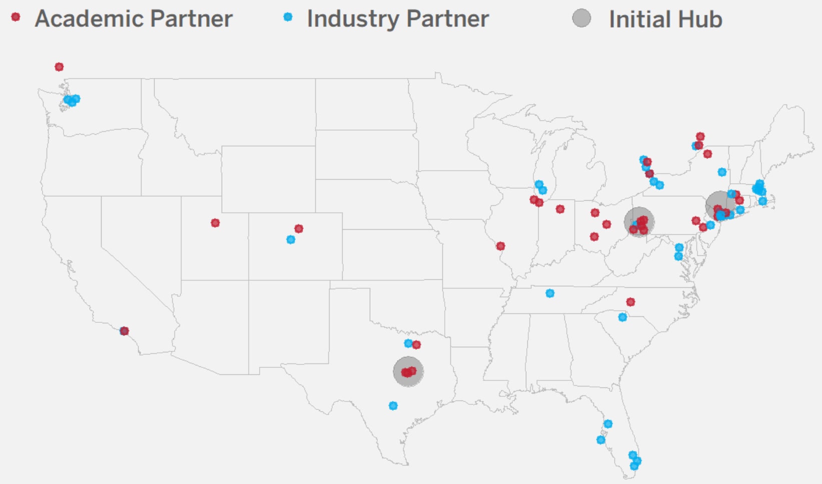
Most of the data is publicly available, or can be done with public available alternatives.

SCORE network

A sustainable national network for developing and disseminating Sports Content for Outreach, Research, and Education in data science

- not unlike SIMIODE, but for sports analytics and data science

Map of Initial Hubs and Confirmed Partners



Carnegie Mellon Sports Analytics Camp and Conference

Summer Camp:

- Hands-on experience in data science using sports data
- Undergraduates entering junior or senior year
- $4,000 stipend to cover living expenses
- Website: **http://summer.stat.cmu.edu/**
- Application Deadline: Sunday, February 28th, 2021 11:59 EST

Conference:

- Date TBD. Usually Late Oct, early Nov.
- 2020 Conference website: **http://www.stat.cmu.edu/cmsac/**

The end

Twitter: @bmacGTPM

LinkedIn: @bmacGTPM