# nflWAR:
# A Reproducible Method for
# Offensive Player Evaluation in Football

Ron Yurko    Sam Ventura    Max Horowitz

Department of Statistics & Data Science
Carnegie Mellon University

RIT Sports Analytics Conference, Aug 11th 2018

# Reproducible Research with `nflscrapR`

**Recent work in football analytics is not easily reproducible:**

- Reliance on proprietary and costly data sources
- Data quality relies on potentially biased human judgement

# Reproducible Research with `nflscrapR`

**Recent work in football analytics is not easily reproducible:**

- Reliance on proprietary and costly data sources
- Data quality relies on potentially biased human judgement

`nflscrapR`:

- R package created by Maksim Horowitz to enable easy data access and promote reproducible NFL research
- Collects play-by-play, game, roster data from NFL.com
- Data is available for all games starting in 2009 (soon 1998!)

**Available on Github, install with:**
`devtools::install_github(repo=maksimhorowitz/nflscrapR)`

# Principles of nflWAR

- Publicly available data, code, and results; **reproducible**

- Interpretable in terms of game outcomes (e.g points, wins)

- Account for uncertainty (football is a small sample game)

- Allow for objective decision-making by coaches/management

## Our nflWAR Framework

- Create **open-source software** for the collection of NFL data
  see R package `nflscrapR` – Horowitz, Yurko, Ventura (2017)

- Properly model plays to determine **play value**

- Use play valuations to model **player value**

- Make player evaluation results **useful and interpretable:**

  - Evaluate relative to replacement level

  - Convert to a wins scale

  - Estimate the uncertainty in our evaluations of players

# How to Value Plays?
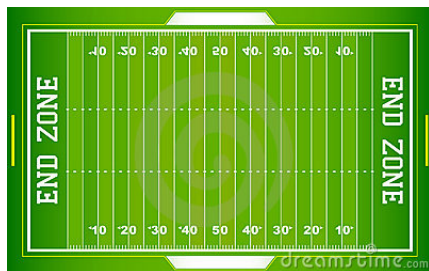


**Suppose it's 4th down with 4 yards to go from the 40 yard line.**

You have three choices:

- **Punt**:
  You are sacrificing possession, but gaining (some) field position

- **Attempt a field goal**:
  You are sacrificing possession, but (possibly) gaining three points

- **"Go for it"**:
  You try to advance the ball four yards to maintain possession

# How to Value Plays?

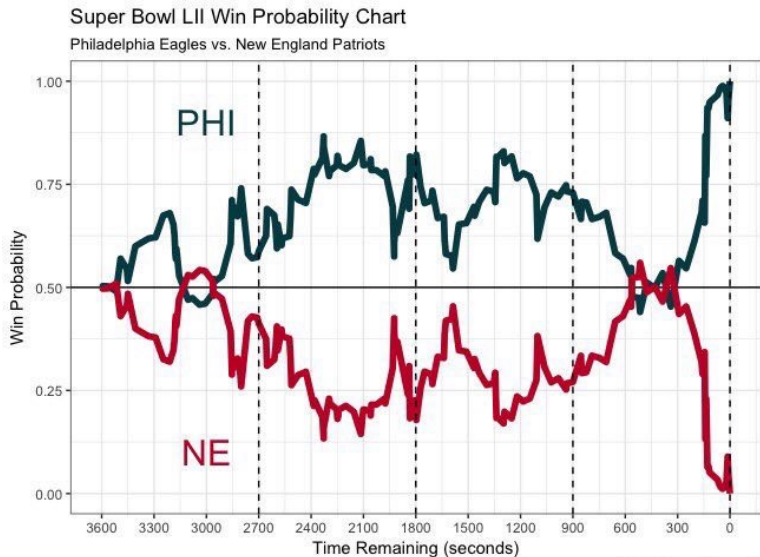**Expected Points (EP):** Value of play is in terms of
$E(points \ of \ next \ scoring \ play)$

- How many points have teams scored when in similar situations? (yard line, down, yards to go, etc.)
- Several ways to model this
- **Our approach: multinomial logistic regression**

**Win Probability (WP):** Value of play is in terms of $P(Win)$

- Have teams in similar situations won the game?
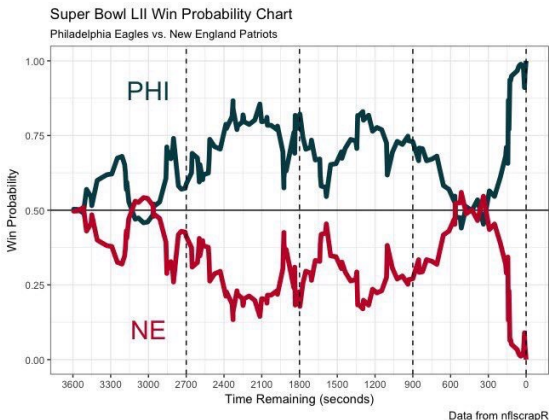- Common approach is logistic regression
- **Our approach: generalized additive model (GAM)**

Can apply **nflWAR** framework to both (or any measure of play value)

# Superbowl LII Win Probability Chart



Super Bowl LII Win Probability Chart
Philadelphia Eagles vs. New England Patriots

PHI

NE

Win Probability

Time Remaining (seconds)

Data from nflscrapR

# Estimating the Value of a Play



Super Bowl LII Win Probability Chart
Philadelphia Eagles vs. New England Patriots

Data from nflscrapR

Win Probability Added **(WPA)** or Expected Points Added **(EPA)**
Using **air yards** → **airWPA/airEPA** and **yacWPA/yacEPA**

# How to Evaluate Players?

# How to Evaluate Players?

Comment from Pittsburgh Post-Gazette article on `nflscrapR`

> stats dont work as well in football as compared to other sports such as baseball. you cant statistically evaluate a running back without evaluating his offensive line. same thing with a QB. you cant evaluate a QB without evaluating his receivers (drop balls, wrong route etc). stats can only be helpful when an athlete is doing something completely on his own (pitching). that is why the nfl doesnt go crazy over stats--it's a team sport on every single play. the only stat that counts is the W.

Reply                    0 👍 👎

# How to Evaluate Players?

Comment from Pittsburgh Post-Gazette article on `nflscrapR`

> stats dont work as well in football as compared to other sports such as baseball. you cant statistically evaluate a running back without evaluating his offensive line. same thing with a QB. you cant evaluate a QB without evaluating his receivers (drop balls, wrong route etc). stats can only be helpful when an athlete is doing something completely on his own (pitching). that is why the nfl doesnt go crazy over stats--it's a team sport on every single play. the only stat that counts is the W.

**Reply**     0 👍 👎

**Football is complex, need to divide credit, evaluate using wins!**

# Division of Credit

**Publicly available data only includes those directly involved:**

- Passing:
    - Players: passer, targeted receiver, tackler(s), and interceptor
    - Context: air yards, yards after catch, location (left, middle, right), and if the passer was hit on the play

- Rushing:
    - Players: rusher and tackler(s)
    - Context: run gap (end, tackle, guard, middle) and direction (left, middle, right)
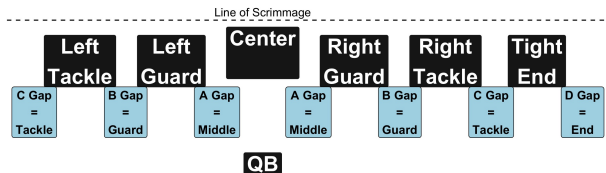
# Multilevel Modeling

Growing in popularity (and rightfully so):

- "Multilevel Regression as Default" - Richard McElreath

- Natural approach for data with **group structure**, and different levels of variation within each group
  e.g. QBs have more pass attempts than receivers have targets

- Every play is a repeated measure of performance

- Hockey example: WAR-on-ice *(Thomas et al., 2013)*

- Baseball example: Deserved Run Average *(Judge et al., 2015)*

# Multilevel Modeling

Example of **varying-intercept** model:

$$WPA_i \sim Normal(Q_{q[i]} + C_{c[i]} + X_i \cdot \beta, \ \sigma_{WPA}^2), \ for \ i = 1, \ldots, n \ \text{plays}$$

Key feature is the **groups are given a model** - treating the levels of groups as similar to one another with **partial pooling**

$$Q_q \sim Normal(\mu_Q, \ \sigma_Q^2), \ \text{for} \ q = 1, \ldots, \# \ \text{of QBs},$$
$$C_c \sim Normal(\mu_C, \ \sigma_C^2), \ \text{for} \ c = 1, \ldots, \# \ \text{of receivers}.$$

Unlike linear regression, no longer assuming independence

Provides estimates for **average play effects** while providing necessary **shrinkage** towards the group averages

# nflWAR Modeling

Use varying-intercepts for each of the grouped variables

With location and gap, create **Team-side-gap** as O-line proxy
e.g. PIT-left-end, PIT-left-tackle, PIT-left-guard, PIT-middle

Separate passing and rushing with different grouped variables

- Passing: Quarterback, receiver, defensive team
- Rushing: Team-side-gap, rusher, defensive team

# nflWAR Modeling

Use varying-intercepts for each of the grouped variables

With location and gap, create **Team-side-gap** as O-line proxy
e.g. PIT-left-end, PIT-left-tackle, PIT-left-guard, PIT-middle

Separate passing and rushing with different grouped variables

- Passing: Quarterback, receiver, defensive team
- Rushing: Team-side-gap, rusher, defensive team

Each group intercept is an estimate for an individual or team effect,

- **individual points/probability added (iPA)**
- **team points/probability added (tPA)**

Multiply intercepts by attempts to get **points/probability above average (iPAA/tPAA)**

# nflWAR Modeling

Four different models to measure three skills:

- Two separate models for QB and non-QB **rushing**

- Two separate passing models for **air** and **yac**

Models adjust for team strength using opposite type of EPA per attempt (e.g. rushing models adjust for passing strength)

Every player has $iPA_{rush}$, $iPA_{air}$, and $iPA_{yac}$ estimates

# QB Air vs Yac Efficiency in 2017
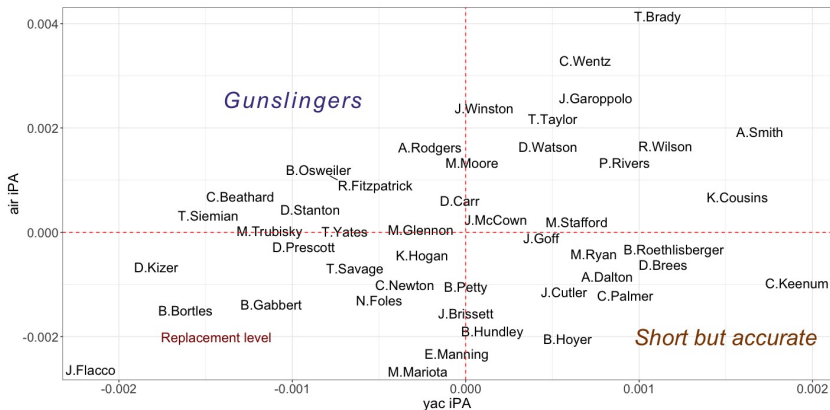
# RB Receiving vs Rushing Efficiency in 2017
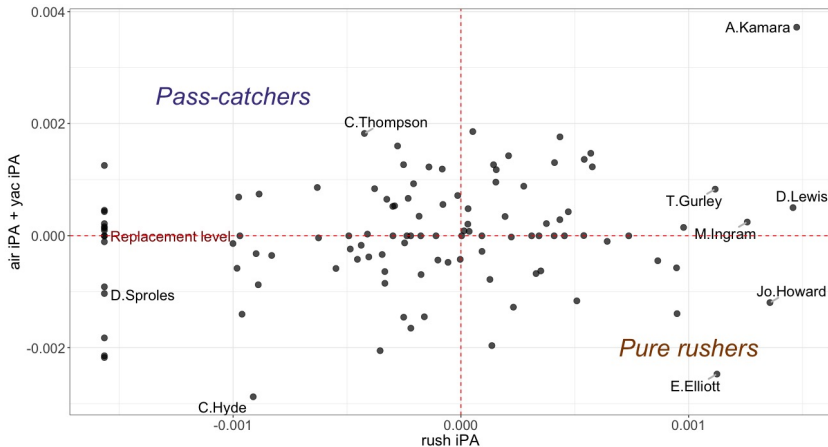
# Comparing Team Offensive Line Performance 2016-17
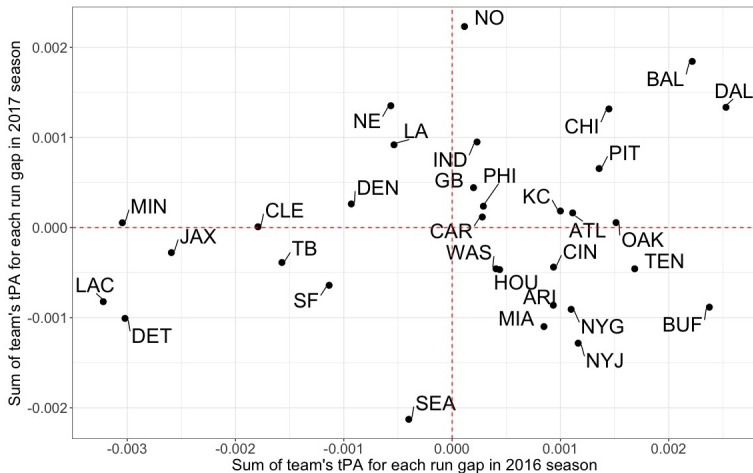
# Arriving at WAR

Evaluate relative to "shadow" replacement player based on rosters similar to **openWAR** *(Baumer et al., 2015)*

Results in **individual points above replacement (iPAR)**

Convert points to wins using regression approach

$$\text{Points per Win} = \frac{1}{\hat{\beta}_{Score\ Diff}}$$

Two types of *WAR*:

$$\text{EPA-based } WAR = \frac{iPAR}{\text{Points per Win}}$$

$$\text{WPA-based } WAR = iPAR$$

# Uncertainty is Mandatory

Similar to **openWAR** (again!) we use a resampling strategy to generate *WAR* distributions

We resample **entire team drives** to preserve any play-sequencing tendencies that could affect our estimates

Following estimates are based on 1000 simulations

# Top RBs by WAR in 2017
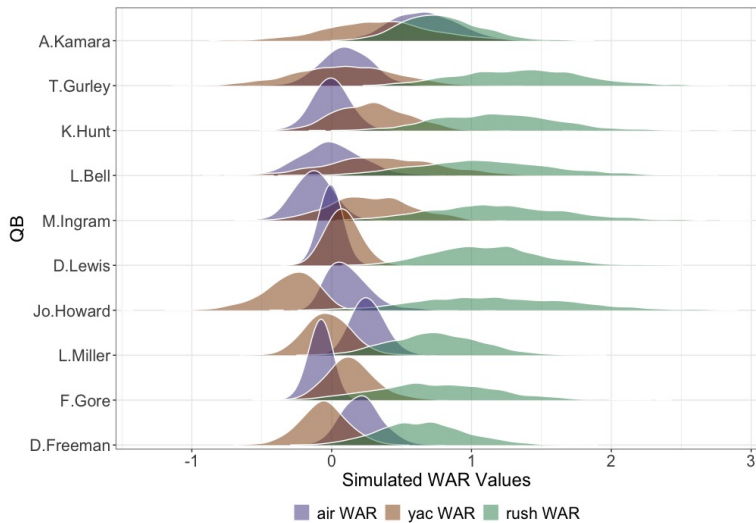
# Selection of QB WAR Distributions in 2017
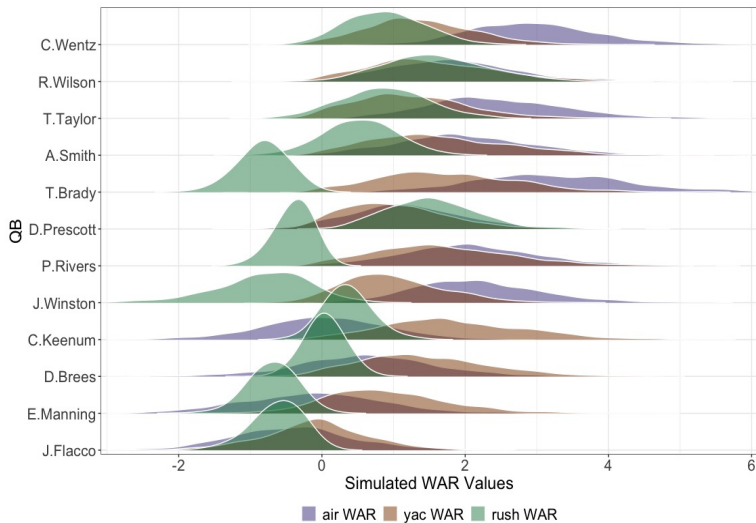
# Career WAR Leaders



Career WAR Leaders
WPA–Based, 2009 to 2018
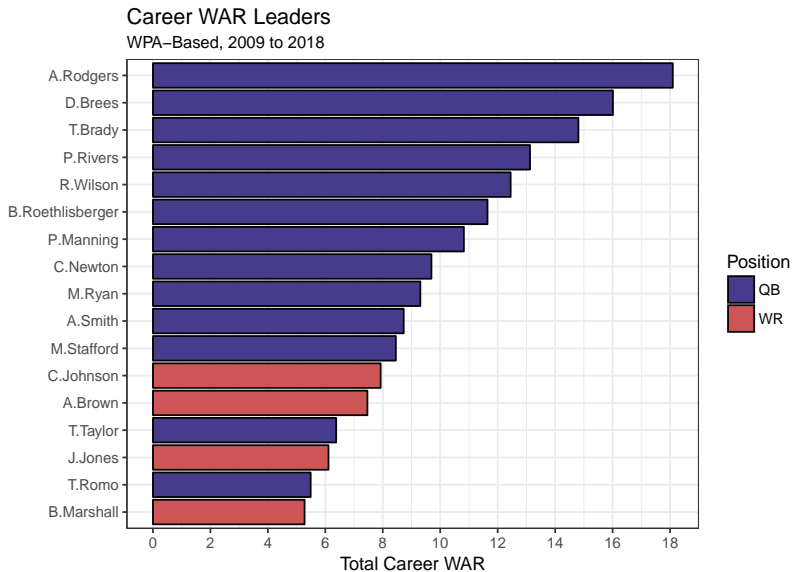
# I'm Sorry Bills Fans



Each Type of QB WAR by Season, 2015-2017
Highlighted = Tyrod Taylor
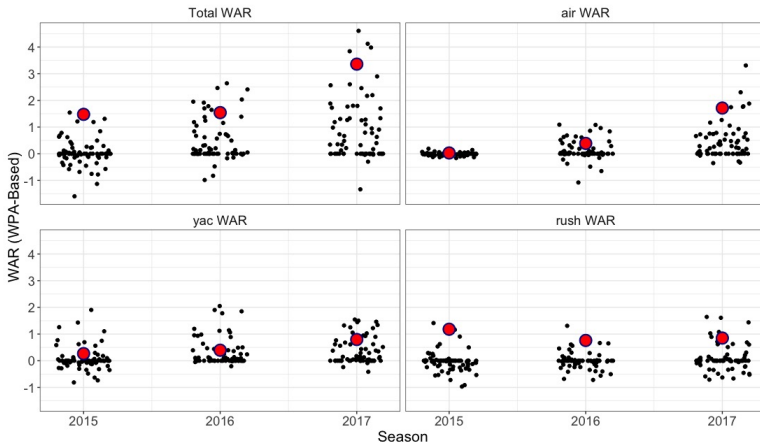
Good luck with Josh Allen!

# Recap and Future of nflWAR

Properly evaluating every play with with multinomial logistic regression model for EP and GAM for WP

Multilevel modeling provides an intuitive way for estimating player effects and **can be extended with data containing every player on the field for every play**

**Estimate the uncertainty** in the different types of iPA to generate intervals of WAR values

# Recap and Future of nflWAR

Properly evaluating every play with with multinomial logistic regression model for EP and GAM for WP

Multilevel modeling provides an intuitive way for estimating player effects and **can be extended with data containing every player on the field for every play**

**Estimate the uncertainty** in the different types of iPA to generate intervals of WAR values

**Naive to assume player has same effect for every play!**

Refine the definition of replacement-level,
e.g. what about down specific players? QBs that rush more?
**#GIVEMETRACKINGDATA**

# Future of Football Analytics

- **Brian Burke** (@bburkeESPN): father of modern football analytics - http://www.advancedfootballanalytics.com/

- **Josh Hermsmeyer** (@friscojosh): air yards, player stability, routes, etc - keeps work accessible, great visualizations

- **Zachary Binney** (@zbinney_NFLinj): NFL injury expert

- **Eric Eager** (@PFF_EricEager): Pro Football Focus collects everything - just don't look at their barcharts...

# Future of Football Analytics

- **Brian Burke** (@bburkeESPN): father of modern football analytics - http://www.advancedfootballanalytics.com/

- **Josh Hermsmeyer** (@friscojosh): air yards, player stability, routes, etc - keeps work accessible, great visualizations

- **Zachary Binney** (@zbinney_NFLinj): NFL injury expert

- **Eric Eager** (@PFF_EricEager): Pro Football Focus collects everything - just don't look at their barcharts...

*...there is another...*

# A New Hope



**Michael Lopez (@StatsbyLopez)** NFL Director of Data & Analytics

Carnegie Mellon Sports Analytics Conference



**Clear your calendars for Oct 19-20th!**

**And visit https://cmusportsanalytics.com/conference2018.html**

**for more information! #CMSAC18**

## Acknowledgements

**Max Horowitz** for creating `nflscrapR`

**Sam Ventura** for advising every step in the process

**CMU Stats in Sports reading and research group**

**Rebecca Nugent** and **CMU Statistics & Data Science** for all of their instruction, motivation, and support!

# Acknowledgements

**Max Horowitz** for creating `nflscrapR`

**Sam Ventura** for advising every step in the process

**CMU Stats in Sports reading and research group**

**Rebecca Nugent** and **CMU Statistics & Data Science** for all of their instruction, motivation, and support!

### Thanks for your attention!

*No barcharts were harmed in the making of this presentation.*

# Expected Points Relationships
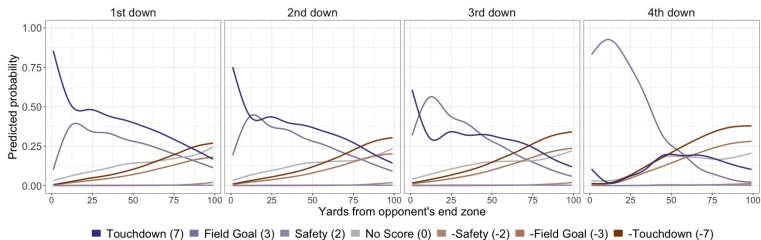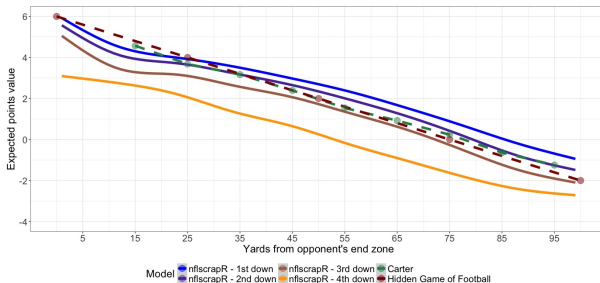
# Convert to Wins

"Wins & Point Differential in the NFL" - *(Zhou & Ventura, 2017)*
(CMU Statistics & Data Science **freshman** research project)
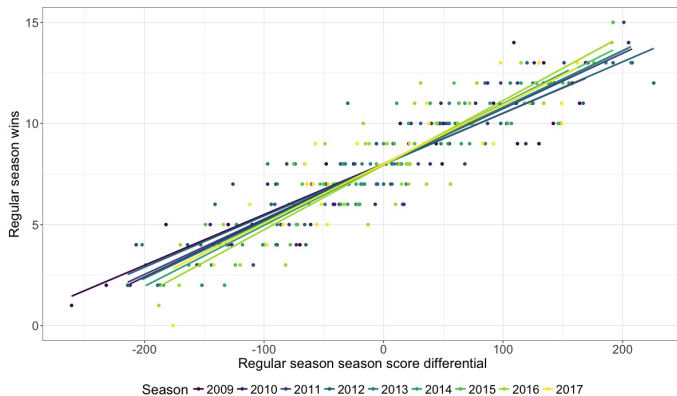


Season → 2009 → 2010 → 2011 → 2012 → 2013 → 2014 → 2015 → 2016 → 2017

## Relative to Replacement Level

Following an approach similar to **openWAR** *(Baumer et al., 2015)*, defining replacement level based on roster

For each position sort by number of attempts, separate replacement level for rushing and receiving

## Relative to Replacement Level

Following an approach similar to **openWAR** *(Baumer et al., 2015)*, defining replacement level based on roster

For each position sort by number of attempts, separate replacement level for rushing and receiving

Player $i's$ $iPAA_{i,total} = iPAA_{i,rush} + iPAA_{i,air} + iPAA_{i,yac}$

Creates a replacement-level iPAA that "shadows" a player's performance, denote as $iPAA_i^{replacement}$

Player $i's$ **individual points above replacement (iPAR)** as:

$$iPAR_i = iPAA_{i,total} - iPAA_{i,total}^{replacement}$$

# References I

📄 numberfire.

📄 Alamar, B. (2010).
Measuring risk in nfl playcalling.
*Journal of Quantitative Analysis in Sports*, 6(2).

📄 Alamar, B. and Goldner, K. (2011).
The blindside project: Measuring the impact of individual
offensive linemen.
*CHANCE*, 24(4):25–29.

📄 Alamar, B. and Weinstein-Gould, J. (2008).
Isolating the effect of individual linemen on the passing game in
the national football league.
*Journal of Quantitative Analysis in Sports*, 4(2).

# References II

📄 Albert, J. (2006).
Pitching statistics, talent and luck, and the best strikeout seasons of all-time.
*Journal of Quantitative Analysis in Sports*, 2(1).

📄 Albert, J. (2016).
Improved component predictions of batting measures.
*Journal of Quantitative Analysis in Sports*, 12(2).

📄 Balreira, E. C., Miceli, B. K., and Tegtmeyer, T. (2014).
An oracle method to predict nfl games.
*Journal of Quantitative Analysis in Sports*, 10(2).

📄 Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015).
Fitting linear mixed-effects models using lme4.
*Journal of Statistical Software*, 67(1):1–48.

# References III

📄 Baumer, B. and Badian-Pessot, P. (2017).
Evaluation of batters and base runners.
In Albert, J., Glickman, M. E., Swartz, T. B., and Koning, R. H.,
editors, *Handbook of Statistical Methods and Analyses in Sports*,
pages 1–37. CRC Press, Boca Raton, Florida.

📄 Baumer, B., Jensen, S., and Matthews, G. (2015).
openwar: An open source system for evaluating overall player
performance in major league baseball.
*Journal of Quantitative Analysis in Sports*, 11(2).

📄 Baumer, B. and Matthews, G. (2017).
*teamcolors: Color Palettes for Pro Sports Teams*.
R package version 0.0.1.9001.

# References IV

📄 Baumer, B. and Zimbalist, A. (2014).
*The Sabermetric Revolution: Assessing the Growth of Analytics in Baseball*.
University of Pennsylvania Press, Philadelphia, Pennsylvania.

📄 Becker, A. and Sun, X. A. (2016).
An analytical approach for fantasy football draft and lineup management.
*Journal of Quantitative Analysis in Sports*, 12(1).

📄 Benjamin, A., Jeff, M., M, D. G., and Lucas, R. (2006).
Who controls the plate? isolating the pitcher/batter subgame.
*Journal of Quantitative Analysis in Sports*, 2(3).

# References V

📄 Berri, D. and Burke, B. (2017).
Measuring productivity of nfl players.
In Quinn, K., editor, *The Economics of the National Football League*, pages 137–158. Springer, New York, New York.

📄 Burke, B. (2009).
Expected point values.

📄 Burke, B. (2014a).
Advanced football analytics.

📄 Burke, B. (2014b).
Expected points and expected points added explained.

📄 Carroll, B., Palmer, P., Thorn, J., and Pietrusza, D. (1988).
*The Hidden Game of Football*.
Total Sports, Inc., New York, New York.

# References VI

📄 Carroll, B., Palmer, P., Thorn, J., and Pietrusza, D. (1998).
*The Hidden Game of Football: The Next Edition*.
Total Sports, Inc., New York, New York.

📄 Carter, V. and Machol, R. (1971).
Operations research on football.
*Operations Research*, 19(2):541–544.

📄 Causey, T. (2013).
Building a win probability model part 1.

📄 Causey, T. (2015).
Expected points part 1: Building a model and estimating uncertainty.

# References VII

📄 Citrone, N. and Ventura, S. L. (2017).
A statistical analysis of the nfl draft: Valuing draft picks and
predicting future player success.
Presented at the Joint Statistical Meetings.

📄 Clark, T. K., Johnson, A. W., and Stimpson, A. J. (2013).
Going for three: Predicting the likelihood of field goal success
with logistic regression.
*MIT Sloan Sports Analytics Conference*.

📄 Dasarathy, B. V. (1991).
*Nearest neighbor (NN) norms: NN pattern classification
techniques*.
IEEE Computer Society Press, Los Alamitos, CA.

# References VIII

Deshpande, S. K. and Jensen, S. T. (2016).
Estimating an nba player's impact on his team's chances of winning.
*Journal of Quantitative Analysis in Sports*, 12(2).

Douglas, B. and Eager, E. A. (2017).
Examining expected points and their interactions with pff grades.
*Pro Football Focus Research and Development Journal*, 1.

Drinen, D. (2013).
Approximate value: Methodolgy.

Eager, E. A., Chahrouri, G., and Palazzolo, S. (2017).
Using pff grades to cluster quarterback performance.
*Pro Football Focus Research and Development Journal*, 1.

# References IX

Elmore, R. and DeWitt, P. (2017).
*ballr: Access to Current and Historical Basketball Data*.
R package version 0.1.1.

ESPN (2017).
Scoring formats.

Franks, A. M., D'Amour, A., Cervone, D., and Bornn, L. (2017).
Meta-analytics: tools for understanding the statistical properties
of sports metrics.
*Journal of Quantitative Analysis in Sports*, 4(4).

Gelman, A. and Hill, J. (2007).
*Data Analysis Using Regression and Multilevel/Hierarchical
Models*.
Cambridge University Press, Cambridge, United Kingdom.

# References X

📄 Goldner, K. (2012).
A markov model of football: Using stochastic processes to model a football drive.
*Journal of Quantitative Analysis in Sports*, 8(1).

📄 Goldner, K. (2017).
Situational success: Evaluating decision-making in football.
In Albert, J., Glickman, M. E., Swartz, T. B., and Koning, R. H., editors, *Handbook of Statistical Methods and Analyses in Sports*, pages 183–198. CRC Press, Boca Raton, Florida.

📄 Gramacy, R. B., Taddy, M. A., and Jensen, S. T. (2012).
Estimating player contribution in hockey with regularized logistic regression.
*Journal of Quantitative Analysis in Sports*, 9(1).

# References XI

📄 Grimshaw, S. D. and Burwell, S. J. (2014).
Choosing the most popular nfl games in a local tv market.
*Journal of Quantitative Analysis in Sports*, 10(3).

📄 Hastie, T., Tibshirani, R., and Friedman, J. (2009).
*The Elements of Statistical Learning: Data Mining, Inference, and Prediction*.
Springer, New York, New York.

📄 Horowitz, M., Yurko, R., and Ventura, S. L. (2017).
*nflscrapR: Compiling the NFL play-by-play API for easy use in R*.
R package version 1.4.0.

📄 James, B. (2017).
Judge and altuve.

# References XII

James, B. and Henzler, J. (2002).
*Win Shares*.
STATS, Inc.

Jensen, J. A. and Turner, B. A. (2014).
What if statisticians ran college football? a re-conceptualization of the football bowl subdivision.
*Journal of Quantitative Analysis in Sports*, 10(1).

Jensen, S., Shirley, K. E., and Wyner, A. (2009).
Bayesball: A bayesian hierarchical model for evaluating fielding in major league baseball.
*The Annals of Applied Statistics*, 3(2).

Judge, J., Brooks, D., and Pavlidis, H. (2015a).
Moving beyond wowy: A mixed approach to measuring catcher framing.

📄 Judge, J., Turkenkopf, D., and Pavlidis, H. (2015b).
Prospectus feature: Introducing deserved run average (dra) and all its friends.

📄 Katz, S. and Burke, B. (2017).
How is total qbr calculated? we explain our quarterback rating.

📄 Kelly, D. (2016).
The state of the two-point conversion.

📄 Knowles, J. E. and Frederick, C. (2016).
*merTools: Tools for Analyzing Mixed Effect Regression Models*.
R package version 0.3.0.

📄 Kubatko, J., Oliver, D., Pelton, K., and Rosenbaum, D. T. (2007).
A starting point for analyzing basketball statistics.
*Journal of Quantitative Analysis in Sports*, 3(3).

# References XIV

Lahman, S. (1996 – 2017).
*Lahman's Baseball Database*.

Lillibridge, M. (2013).
The anatomy of a 53-man roster in the nfl.

Lock, D. and Nettleton, D. (2014).
Using random forests to estimate win probability before each play of an nfl game.
*Journal of Quantitative Analysis in Sports*, 10(2).

Lopez, M. (2017).
All win probability models are wrong some are useful.

Macdonald, B. (2011).
A regression-based adjusted plus-minus statistic for nhl players.
*Journal of Quantitative Analysis in Sports*, 7(3).

📄 Martin, R., Timmons, L., and Powell, M. (2017).
A markov chain analysis of nfl overtime rules.
*Journal of Sports Analytics*, Pre-print(Pre-print).

📄 Meers, K. (2011).
How to value nfl draft picks.

📄 Morris, B. (2015).
Kickers are forever.

📄 Morris, B. (2017).
Running backs are finally getting paid what they're worth.

📄 Mulholland, J. and Jensen, S. T. (2014).
Predicting the draft and career success of tight ends in the
national football league.
*Journal of Quantitative Analysis in Sports*, 10(4).

# References XVI

Oliver, D. (2011).
Guide to the total quarterback rating.

Paine, N. (2015a).
Bryce harper should have made $73 million more.

Paine, N. (2015b).
How madden ratings are made: The secret process that turns nfl players into digital gods.

Pasteur, D. and Cunningham-Rhoads, K. (2014).
An expectation-based metric for nfl field goal kickers.
*Journal of Quantitative Analysis in Sports*, 10(1).

# References XVII

📄 Pasteur, R. D. and David, J. A. (2017).
Evaluation of quarterbacks and kickers.
In Albert, J., Glickman, M. E., Swartz, T. B., and Koning, R. H.,
editors, *Handbook of Statistical Methods and Analyses in Sports*,
pages 165–182. CRC Press, Boca Raton, Florida.

📄 Pattani, A. (2012).
Expected points and epa explained.

📄 Phillips, C. (2018).
*Scoring and Scouting: How We Know What We Know about
Baseball*.
Forthcoming, Princeton University Press, Princeton, New Jersey.

📄 Piette, J. and Jensen, S. (2012).
Estimating fielding ability in baseball players over time.
*Journal of Quantitative Analysis in Sports*, 8(3).

📄 Pro-Football-Reference (2018).
Football glossary and football statistics glossary.

📄 Quealy, K., Causey, T., and Burke, B. (2017).
4th down bot: Live analysis of every n.f.l. 4th down.

📄 R Core Team (2017).
*R: A Language and Environment for Statistical Computing*.
R Foundation for Statistical Computing, Vienna, Austria.

📄 Ratcliffe, J. (2013).
The pff idp scoring system revisited.

📄 Romer, D. (2006).
Do firms maximize? evidence from professional football.
*Journal of Political Economy*, 114(2).

# References XIX

📄 Rosenbaum, D. T. (2004).
Measuring how nba players help their teams win.

📄 Schatz, A. (2003).
Methods to our madness.

📄 Schoenfield, D. (2012).
What we talk about when we talk about war.

📄 Sievert, C. (2015).
*pitchRx: Tools for Harnessing 'MLBAM' 'Gameday' Data and Visualizing 'pitchfx'.*
R package version 1.8.2.

📄 Skiver, K. (2017).
Zebra technologies to bring live football tracking on nfl gamedays with rfid chips.

# References XX

📄 Smith, D., Siwoff, S., and Weiss, D. (1973).
Nfl's passer rating.

📄 Snyder, K. and Lopez, M. (2015).
Consistency, accuracy, and fairness: a study of discretionary penalties in the nfl.
*Journal of Quantitative Analysis in Sports*, 11(4).

📄 Sportradar (2017).
*NFL Data Addendum*.
Last updated 2017-08-11.

📄 Sportradar (2018).
*NFL Official API*.
Version 2.

📄 Tango, T. (2017).
War podcast.

# References XXI

📄 Tango, T., Lichtman, M., and Dolphin, A. (2007).
*The Book: Playing the Percentages in Baseball*.
Potomac Book, Inc., Washington, D.C.

📄 Thomas, A. and Ventura, S. L. (2017).
*nhlscrapr: Compiling the NHL Real Time Scoring System Database for easy use in R*.
R package version 1.8.1.

📄 Thomas, A. C. and Ventura, S. L. (2015).
The road to war.

📄 Thomas, A. C., Ventura, S. L., Jensen, S. T., and Ma, S. (2013).
Competing process hazard function models for player ratings in ice hockey.
*The Annals of Applied Statistics*, 7(3).

📄 Wakefield, K. and Rivers, A. (2012).
The effect of fan passion and official league sponsorship on brand
metrics: A longitudinal study of official nfl sponsors and roo.
*MIT Sloan Sports Analytics Conference.*

📄 Yam, D. R. and Lopez, M. J. (2018).
Quantifying the causal effects of conservative fourth down
decision making in the national football league.
Under Review.

📄 Yurko, R., Ventura, S., and Horowitz, M. (2017).
Nfl player evaluation using expected points added with nflscrapr.
Presented at the Great Lakes Sports Analytics Conference.

📄 Zhou, E. and Ventura, S. (2017).
Wins and point differential in the nfl.