Fun with Publicly Available Baseball Data in R

Statistics can be fun!

Gregory J. Matthews[1]

[1] Loyola Univeristy Chicago

Oakton Community College
March 2, 2020

LOYOLA
UNIVERSITY CHICAGO

1. Introduction
   - A brief history of baseball data
2. The data!
   - Lahman Database
   - Statcast
3. Some of my projects
   - openWAR
   - Our Contribution
   - Astros Banging Scandal

- Line score
- Box scores (1859, Henry Chadwick)
- Baseball Abstracts (1970s, Bill James)
- Retrosheet (1989, Dr. David Smith) (Play by play of every game)
- Lahman Database (1995, Sean Lahman)
- Baseball Reference (2000, Sean Forman)
- Pitch FX (2006)
- Statcast (2015)

- Line score
- Box scores (1859, Henry Chadwick
- Baseball Abstracts (1970s, Bill James)
- Retrosheet (1989, Dr. David Smith) (Play by play of every game)
- **Lahman Database** (1995, Sean Lahman
- Baseball Reference (2000, Sean Forman)
- Pitch FX (2006)
- **Statcast** (2015)

Lahman Database

- The updated version of the database contains complete batting and pitching statistics from 1871 to 2020, plus fielding statistics, standings, team stats, managerial records, post-season data, and more.
- http://www.seanlahman.com/baseball-archive/statistics/

```
library(tidyverse)
library(Lahman)
Batting %>%
  subset(yearID <= 2019 & yearID >= 2000) %>%
  arrange(-HR) %>%
  left_join(People) %>%
  select(nameFirst,nameLast,yearID,HR) %>%
  head(10)
```

```
##     nameFirst  nameLast yearID HR
## 1       Barry     Bonds   2001 73
## 2       Sammy      Sosa   2001 64
## 3   Giancarlo   Stanton   2017 59
## 4        Ryan    Howard   2006 58
## 5        Luis  Gonzalez   2001 57
## 6        Alex Rodriguez   2002 57
## 7       David     Ortiz   2006 54
## 8        Alex Rodriguez   2007 54
## 9        Jose  Bautista   2010 54
## 10      Chris     Davis   2013 53
```

```
library(dplyr)
library(Lahman)
#Take data from 2015 - 2019
#dat <- subset(Batting, yearID <= 2019 & yearID >= 2015)
#RBI from 2015 to 2019
Batting %>%
  subset(yearID <= 2019 & yearID >= 2015) %>%
  group_by(playerID) %>%
  summarise(HR = sum(HR), Hits = sum(H), RBI = sum(RBI)) %>%
  arrange(-RBI) %>%
  left_join(People) %>%
  select(nameFirst,nameLast,HR,Hits,RBI)
```

```
## # A tibble: 2,475 x 5
##    nameFirst nameLast        HR  Hits   RBI
##    <chr>     <chr>        <int> <int> <int>
##  1 Nolan     Arenado        199   906   621
##  2 Edwin     Encarnacion    185   672   538
##  3 Nelson    Cruz           204   781   522
##  4 Anthony   Rizzo          147   799   514
##  5 J. D.     Martinez       184   803   509
##  6 Paul      Goldschmidt    160   847   505
##  7 Jose      Abreu          143   862   504
##  8 Bryce     Harper         164   715   486
##  9 Khris     Davis          183   622   474
## 10 Albert    Pujols         136   683   472
## # ... with 2,465 more rows
```

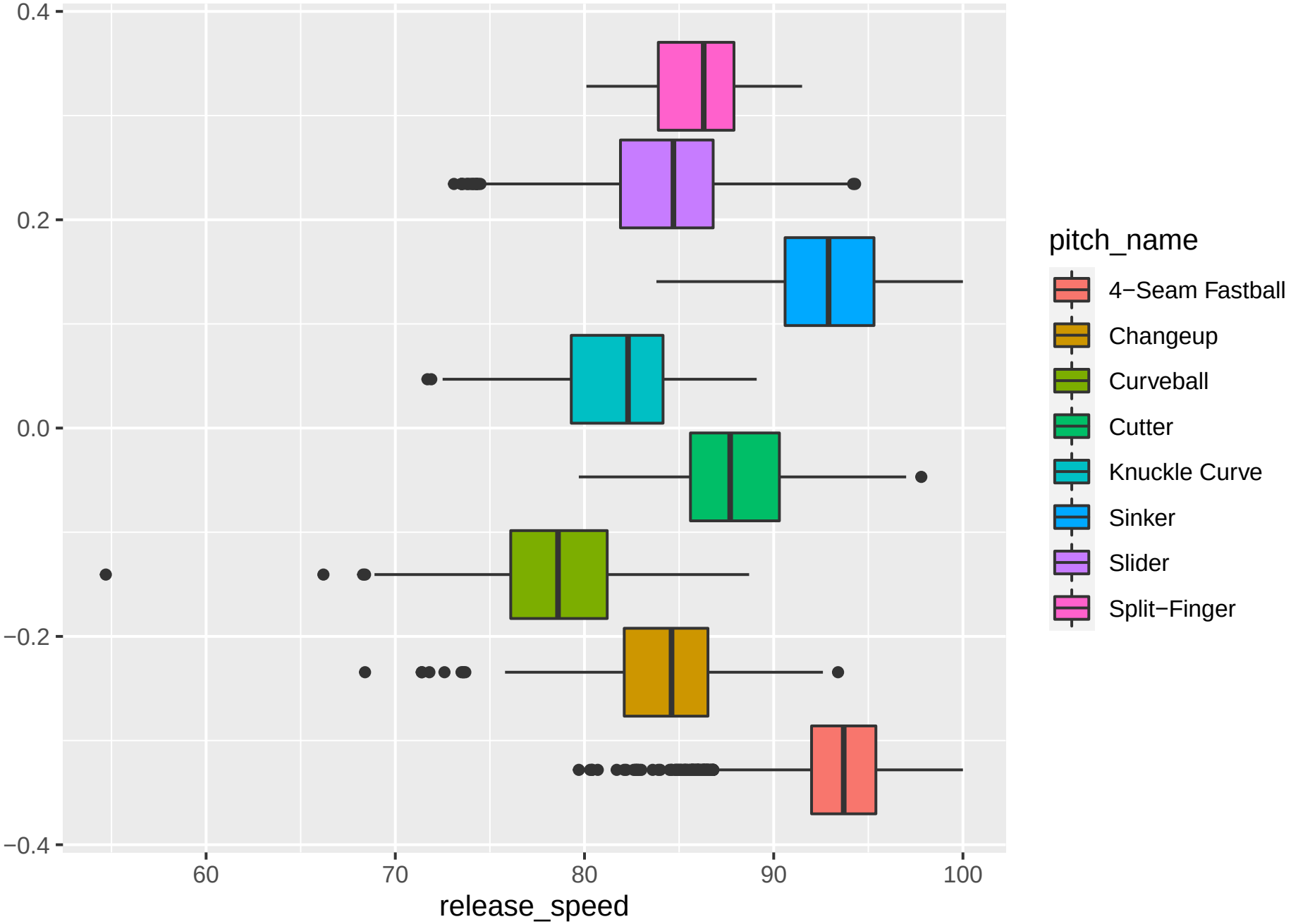
Statcast

- From wikipedia: "Statcast is a high-speed, high-accuracy, automated tool developed to analyze player movements and athletic abilities in Major League Baseball (MLB). Statcast was introduced to all thirty MLB stadiums in 2015."
- Based on Doppler radar and high defintion video.
- https://baseballsavant.mlb.com/statcast_search

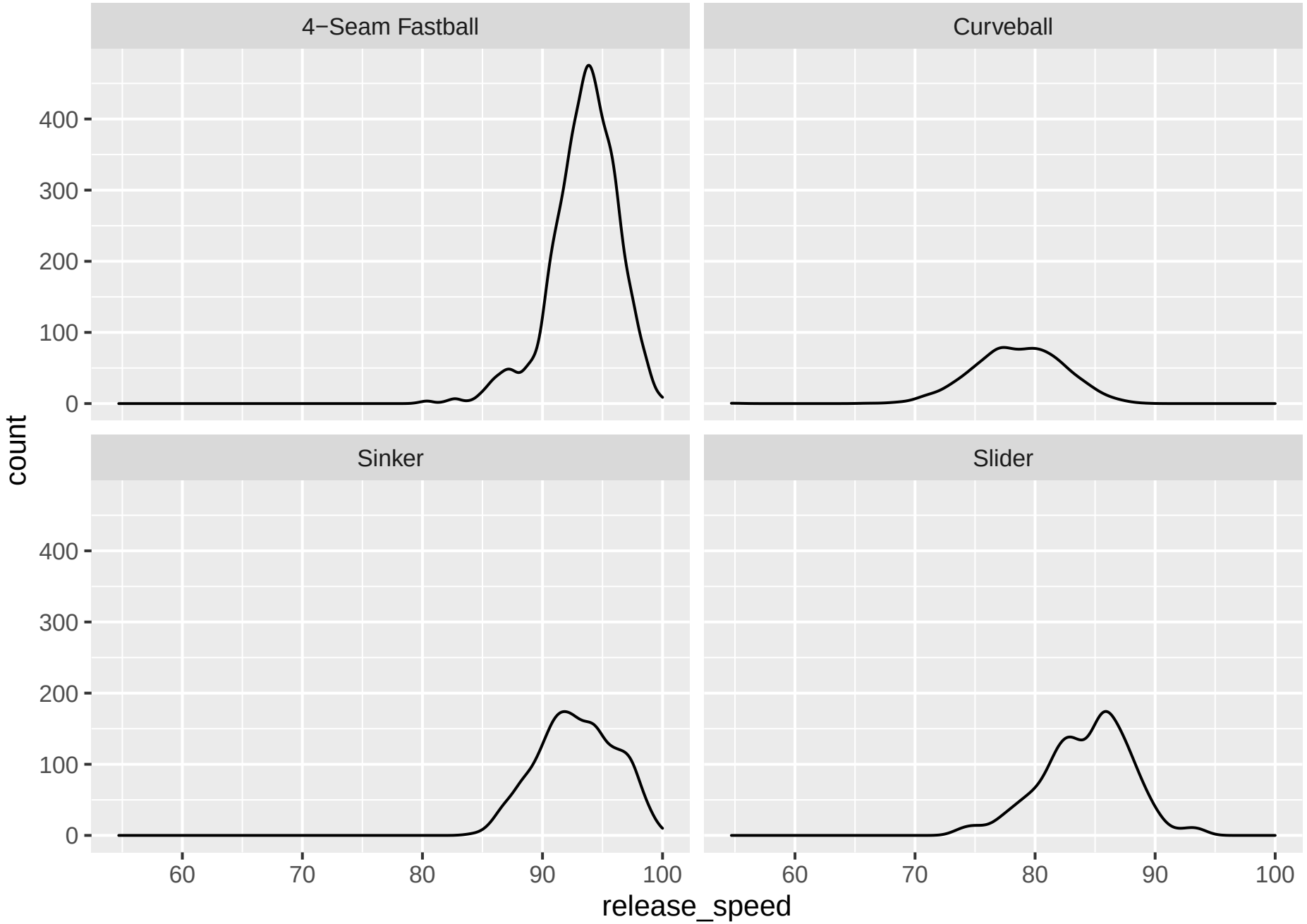
Getting Statcast Data

```
#Shout out to Bill Petti (@billpetti).
#https://billpetti.github.io/2018-02-19-build-statcast-database-rstats/
library(dplyr)
library(tidyverse)
library(baseballr)
#mlb2020 <- scrape_statcast_savant_pitcher_date("2020-07-23","2020-07-25")
#save(mlb2020, file = "/Users/gregorymatthews/Dropbox/Talks/openWARLoyolaHighSchool/mlb2020.RData")
load("/Users/gregorymatthews/Dropbox/Talks/openWARLoyolaHighSchool/mlb2020.RData")
```

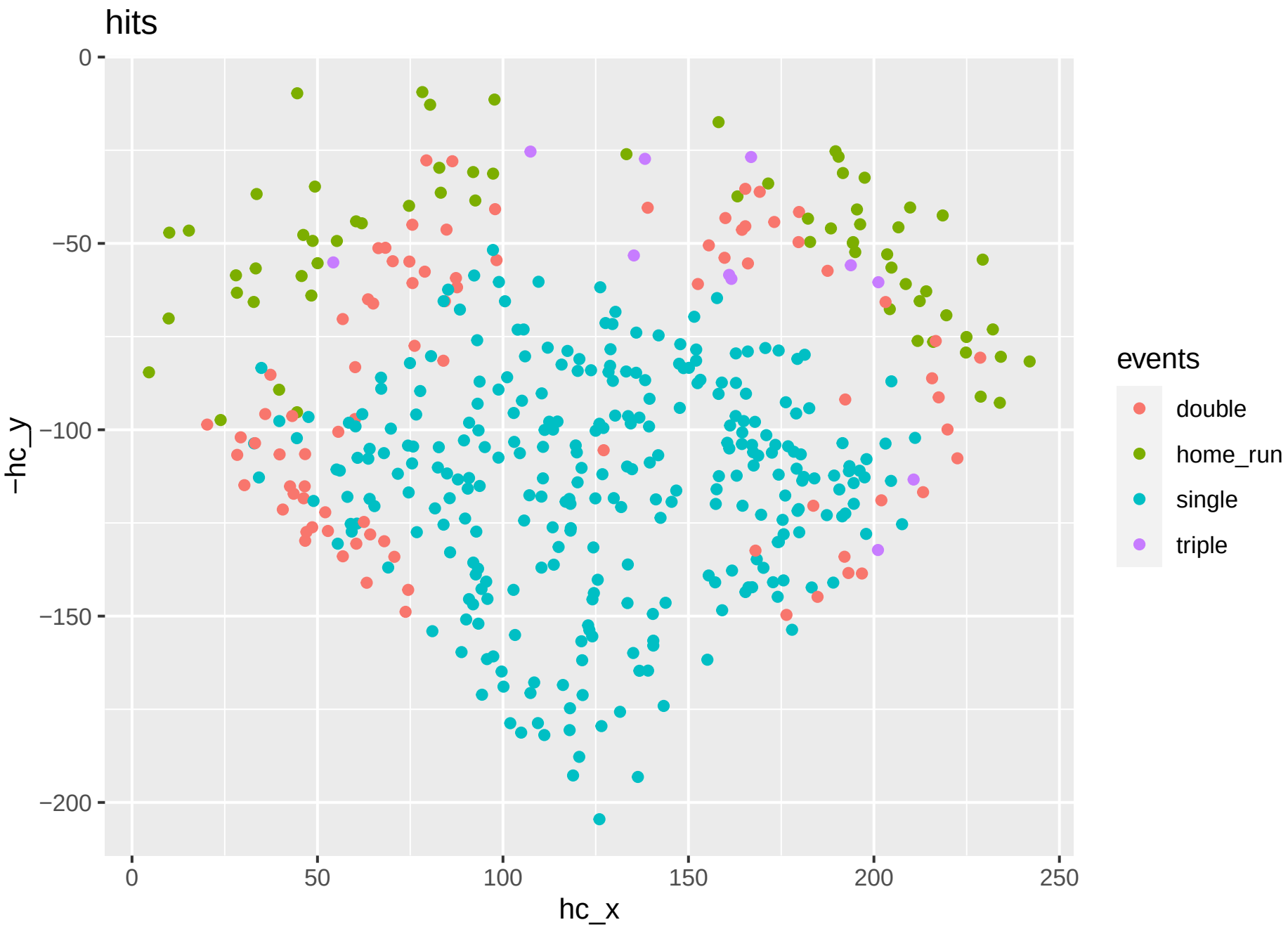
```
library(ggplot2)
ggplot(aes(x = release_speed, fill = pitch_name), data = mlb2020) +
  geom_boxplot()
```


0.4
0.2
0.0
−0.2
−0.4
60
70
80
90
100
release_speed
pitch_name
4−Seam Fastball
Changeup
Curveball
Cutter
Knuckle Curve
Sinker
Slider
Split−Finger

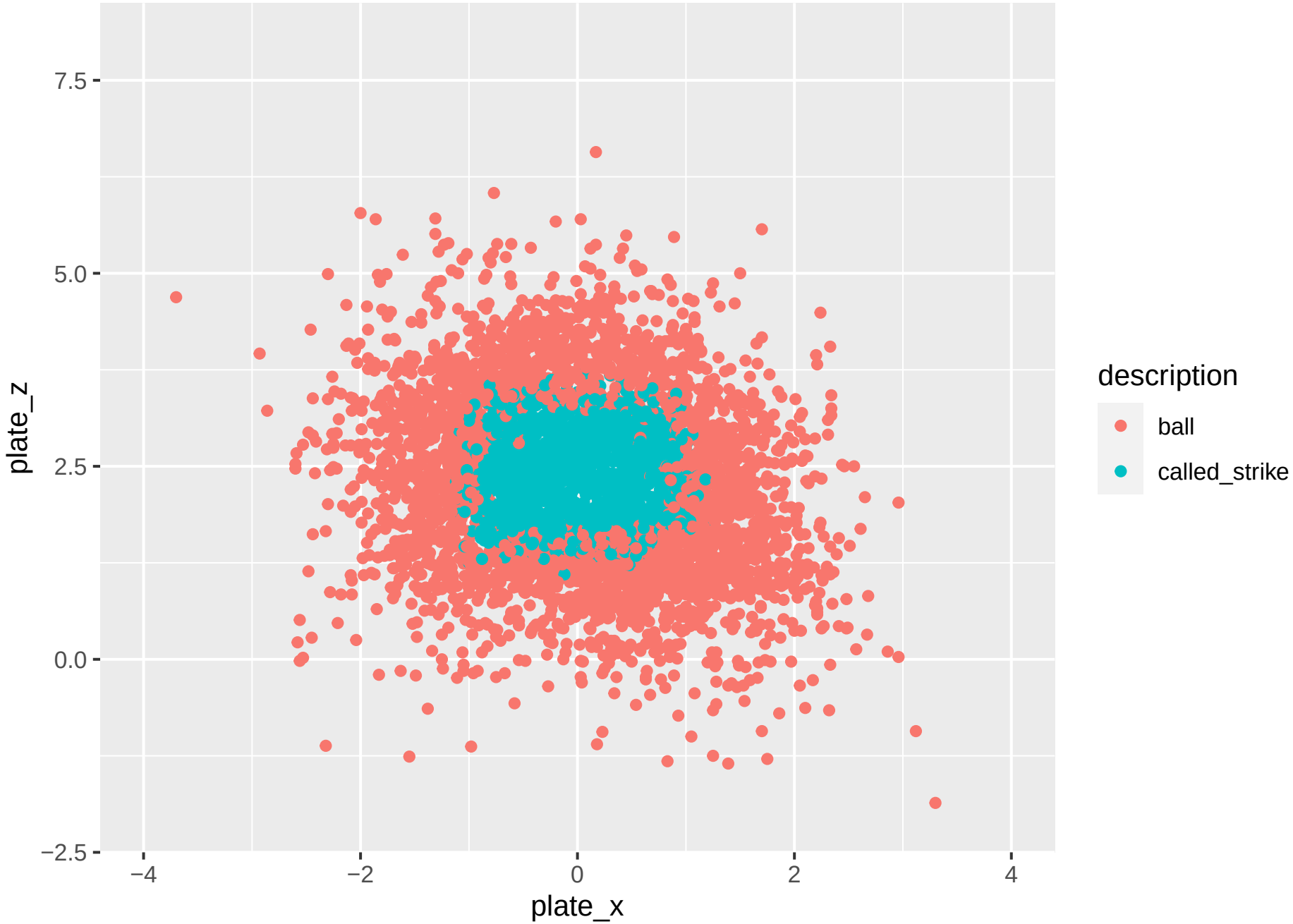
```
library(ggplot2)
pitches <- c("Curveball","4-Seam Fastball","Slider","Sinker")
mlb2020 %>%
  subset(pitch_name %in% pitches) %>%
  ggplot(aes(x = release_speed))  +
  geom_density(aes(x = release_speed, after_stat(count)), alpha = 0
  facet_wrap(~pitch_name) + theme_bw()
#Claps for Quang Nguyen for suggesting theme_bw()
```


4−Seam Fastball
Curveball
Sinker
Slider
count
release_speed
0
100
200
300
400
60
70
80
90
100

```
library(ggplot2)
mlb2020 %>%
  subset(events %in% c("single","double","triple","home_run")) %>%
  ggplot(aes(y = -hc_y, x = hc_x, color = events)) +
  geom_point() +
  ggtitle("hits")
```


hits
0
−50
−100
−150
−200
−hc_y
0
50
100
150
200
250
hc_x
events
double
home_run
single
triple

```
library(ggplot2)
mlb2020 %>%
  subset(description %in% c("ball","called_strike")) %>%
ggplot(aes(x = plate_x, y = plate_z, color = description)) +
  geom_point() + xlim(-4,4) + ylim(-2,8)
```

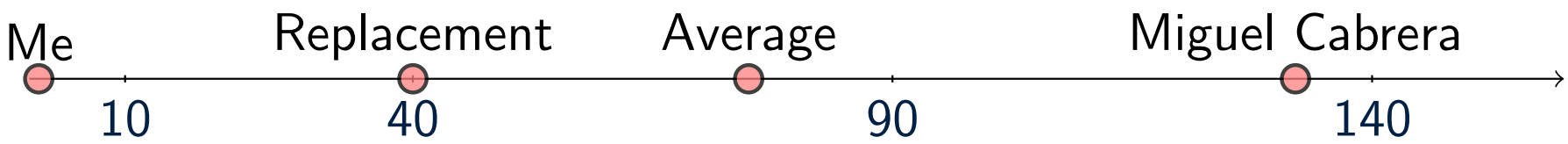

7.5
5.0
2.5
0.0
−2.5
plate_z
−4
−2
0
2
4
plate_x
description
ball
called_strike

WAR - What is it good for?

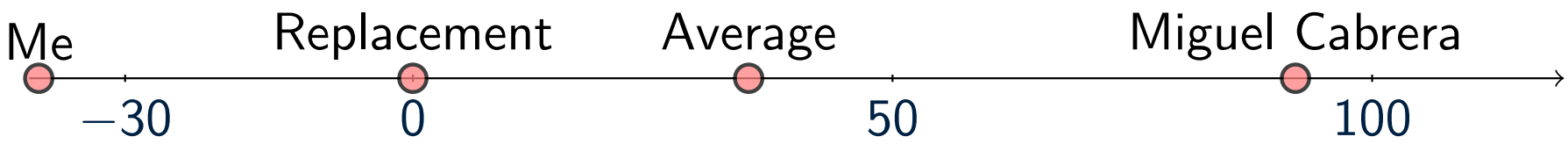
- Wins above replacement
- Question: How large is the contribution that each player makes towards winning?
- Four Components:
  1. Batting
  2. Baserunning
  3. Fielding
  4. Pitching
- Replacement Player: Hypothetical 4A journeyman
  - Much worse than an average player

Units and Scaling

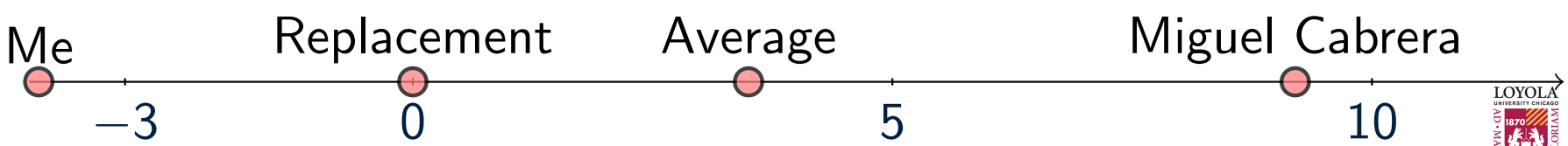
- In terms of absolute runs:



- In terms of Runs Above Replacement (*RAR*):



- In terms of Wins Above Replacement (*WAR*):



What's Wrong with WAR?

- Not Reproducible
  - WAR is an unknown hypothetical quantity – not a statistic
  - No reference implementation of WAR
  - No open data set
  - No open source code
- No unified methodology
  - Each component of WAR is viewed as a separate problem – not a piece of the same problem
  - Ad hoc definitions: what is replacement level?
- No error estimates
  - Only reported as point estimates
  - Only hand-wavy estimates of variability or margin or error
- Bug or Feature?: Competing black-box implementations

Our Contribution: *openWAR*

- *openWAR*: a reproducible reference implementation of WAR
  - Principled estimate of WAR
  - Fully open-source R package (free as in freedom)
  - Partially open data (free as in beer)
- Unified Methodology:
  - Conservation of Runs
  - Each component is estimated as a piece of the larger problem
- Error estimates:
  - Use resampling methods to report WAR interval estimates
- Version 0.1: Emphasis at this stage on reproducibility

Getting Data

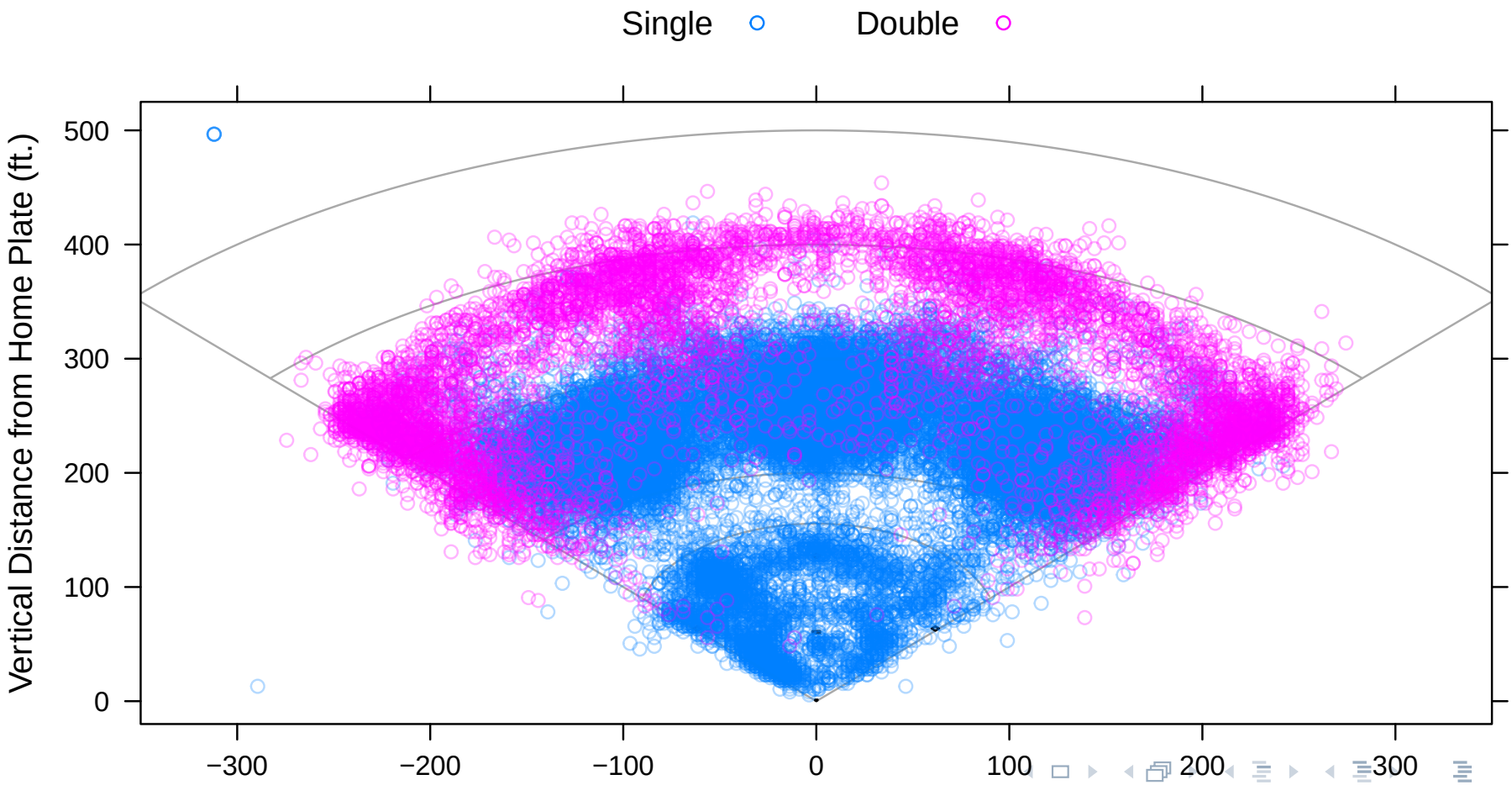
```
#Paper link: https://arxiv.org/abs/1312.7158
install.packages("xslt")
#Package with functions
devtools::install_github("beanumber/openWAR")
#Package containing the data
devtools::install_github("beanumber/openWARData")
library(openWAR)
ds = getData(start = "2013-06-24")
dim(ds)
head(ds$description)
```

Getting Data

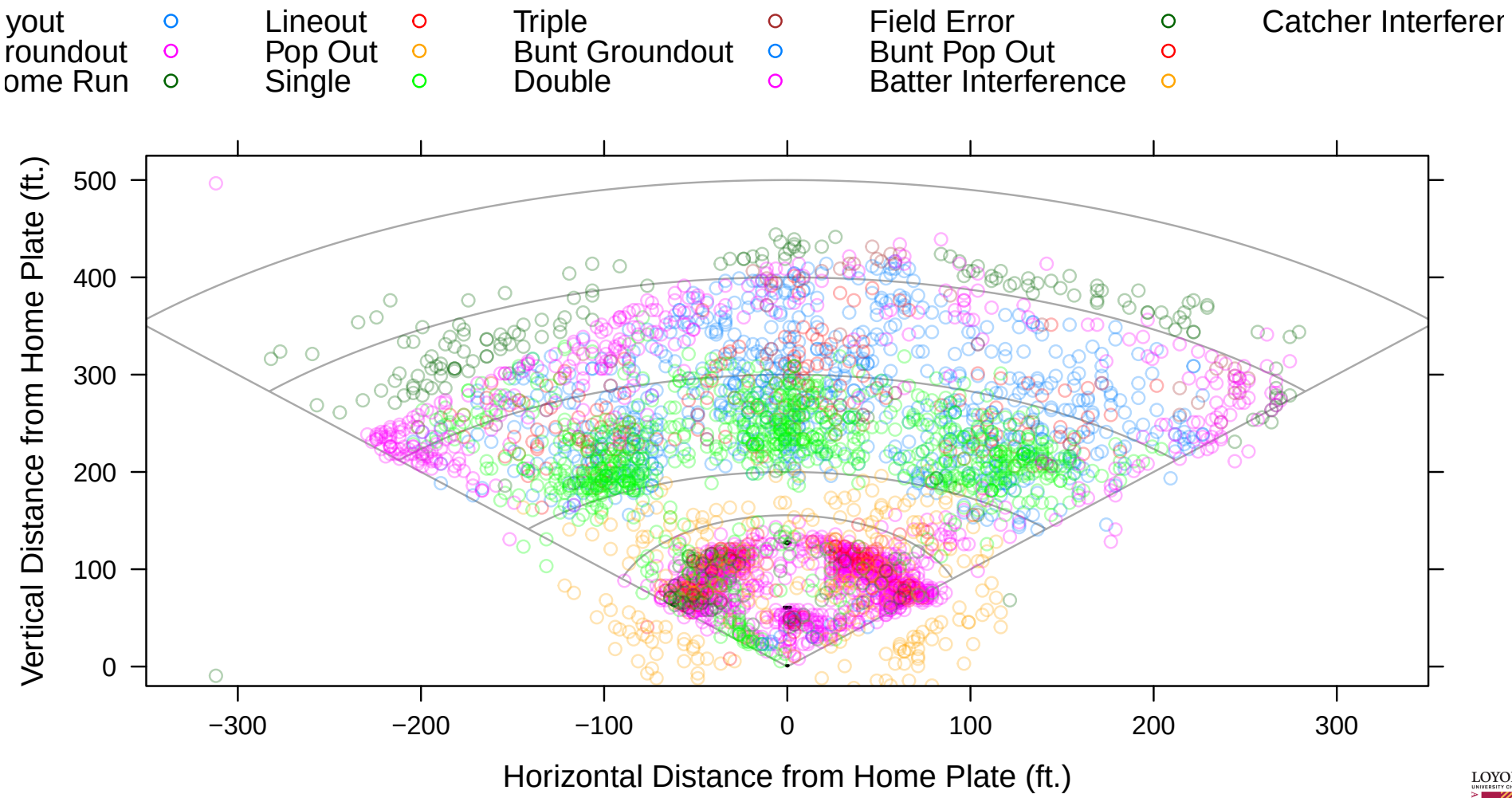
```
library(openWARData)
dim(MLBAM2017)

## [1] 185704     62

head(MLBAM2017$description)

## [1] "Kyle Schwarber singles on a line drive to right fielder Stephen Piscotty.
## [2] "Kris Bryant strikes out swinging.  "
## [3] "Anthony Rizzo singles on a line drive to right fielder Stephen Piscotty.
## [4] "Ben Zobrist grounds into a double play, second baseman Jedd Gyorko to short
## [5] "Dexter Fowler lines out to center fielder Jason Heyward.  "
## [6] "Aledmys Diaz doubles (1) on a sharp line drive to right fielder Ben Zobrist
```

Visualizing the Data

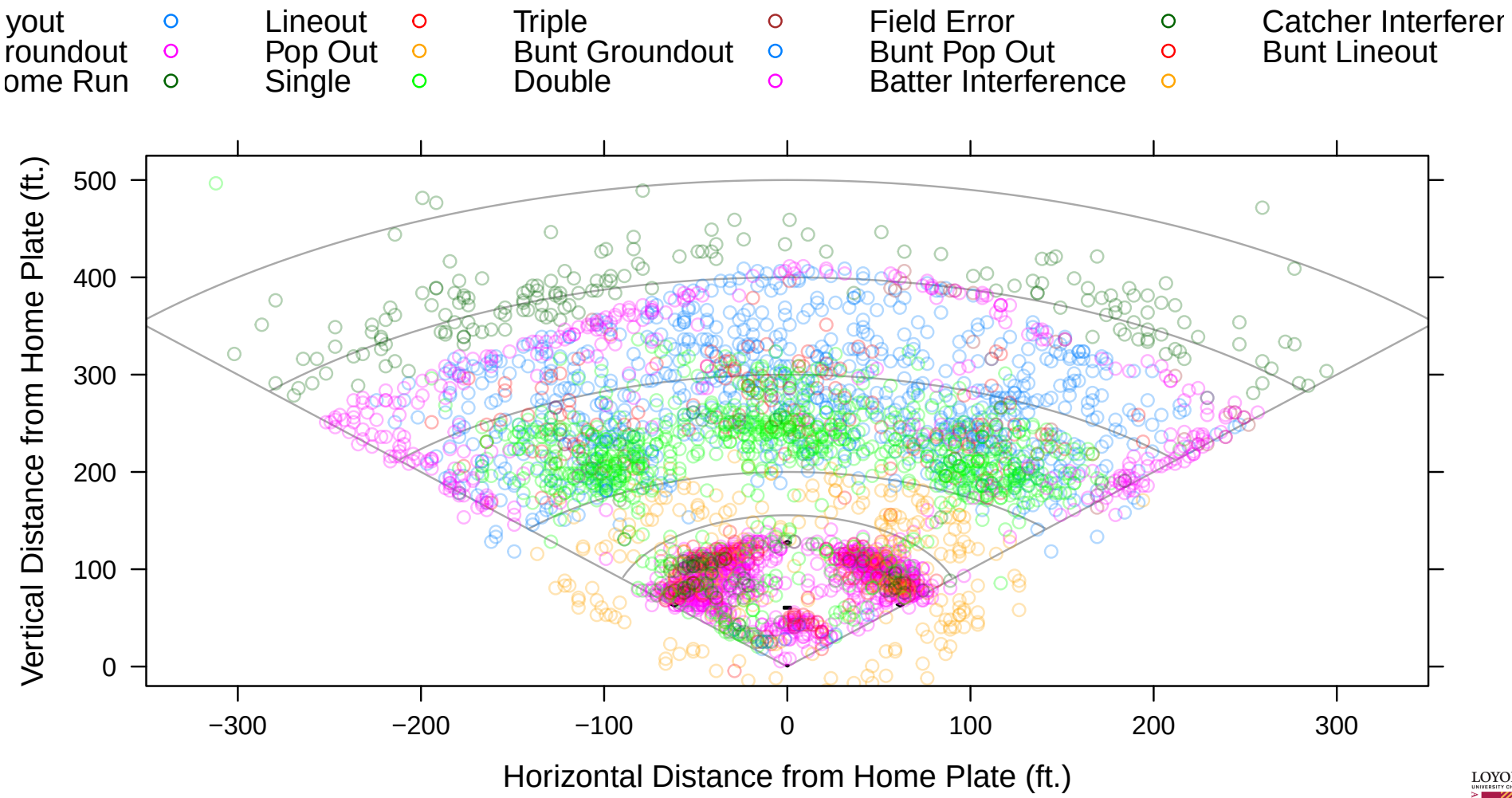
```
library(openWARData)
data(MLBAM2013)
plot(subset(MLBAM2013, event %in% c("Single","Double")))
```



Which Ballpark is this?



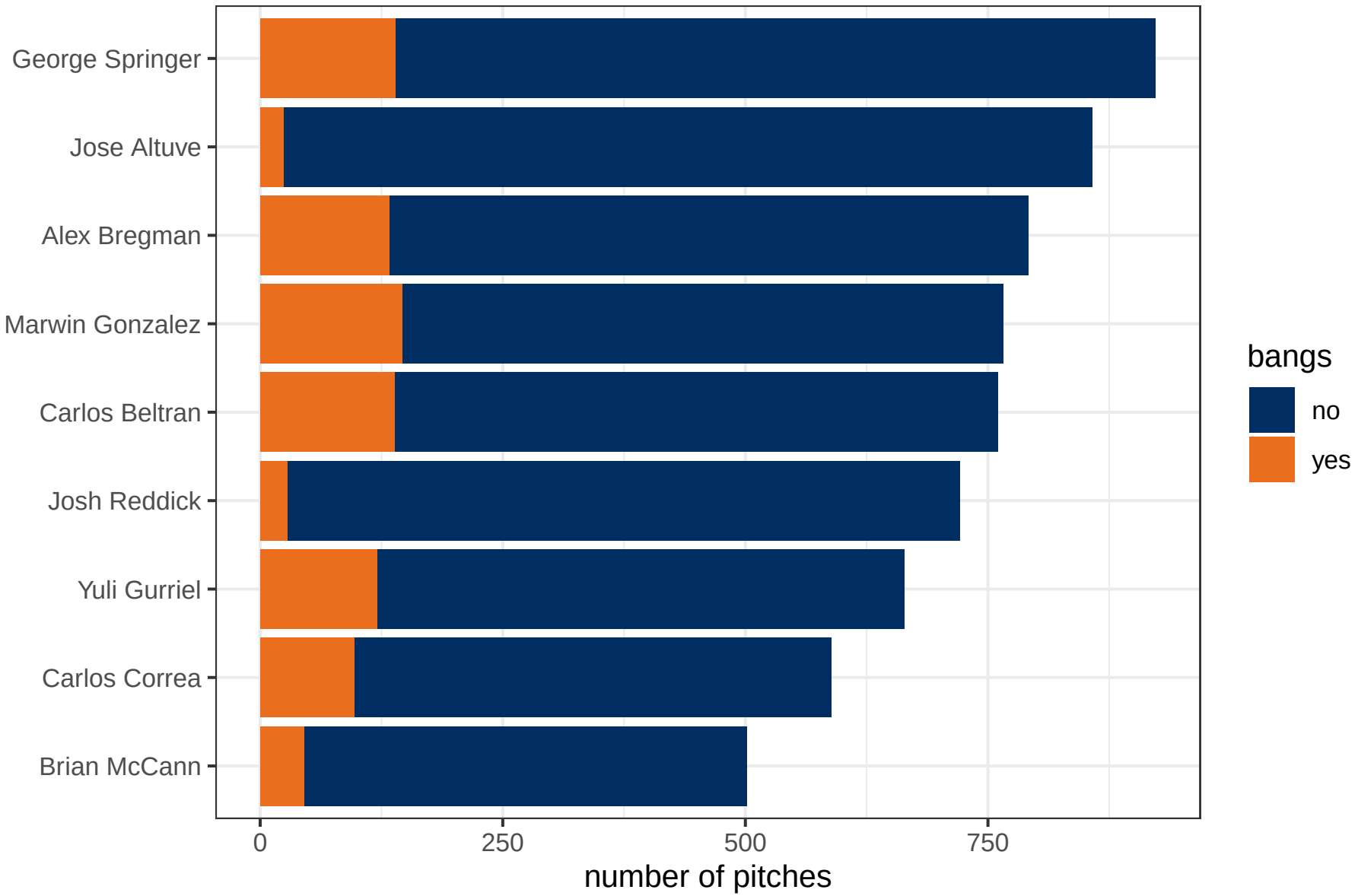
Which Ballpark is this?



Main findings

- Less Swinging
  - Odds of a swing are about 28% lower (OR 95% CI: 0.618, 0.850)
- More contact on swings (on off-speed pitches)
  - Given a swing and an offspeed pitch, odds of contact are about 80% higher (OR 95% CI: 1.342, 2.675)
- Increased Exit Velocity
  - Given a contact, 2.386 mile per hours average increase in exit velocity (95% CI: 0.334, 4.451)

```
#Github repo: https://github.com/gjm112/Astros_sign_stealing
#Paper: http://bit.ly/Astros_Cheat
library(RCurl)
#Bangs file
bangs <- read.csv("https://raw.githubusercontent.com/gjm11/Astros_sign_stealing/master/data/astros_bangs_202001

## Error in file(file, "rt"):  cannot open the connection to
'https://raw.githubusercontent.com/gjm11/Astros_sign_stealing/master/data/astros_bangs_20200127.csv'

#Bangs file combined with pitchinfo.com data.
githubURL <- ("https://github.com/gjm112/Astros_sign_stealing/blob/master/data/bangs-merged-final.rds?raw=true"
download.file(githubURL,"/Users/gregorymatthews/bangs-merged-final.rds", mode = "wb")
bangs_merged_final <- readRDS(file = "/Users/gregorymatthews/bangs-merged-final.rds")
```


George Springer
Jose Altuve
Alex Bregman
Marwin Gonzalez
Carlos Beltran
Josh Reddick
Yuli Gurriel
Carlos Correa
Brian McCann
0
250
500
750
number of pitches
bangs
no
yes

Final Thoughts

- Sources of raw data: https://sabr.org/sabermetrics/data
- Github for this talk: https://github.com/gjm112/Oakton_STEM_Series_baseball
- Me on Twitter: @statsinthewild
- My email address: gmatthews1@luc.edu
- Hey, Greg? Where can I get a job in the sports industry?
- teamworkonline.com

Cheers!