

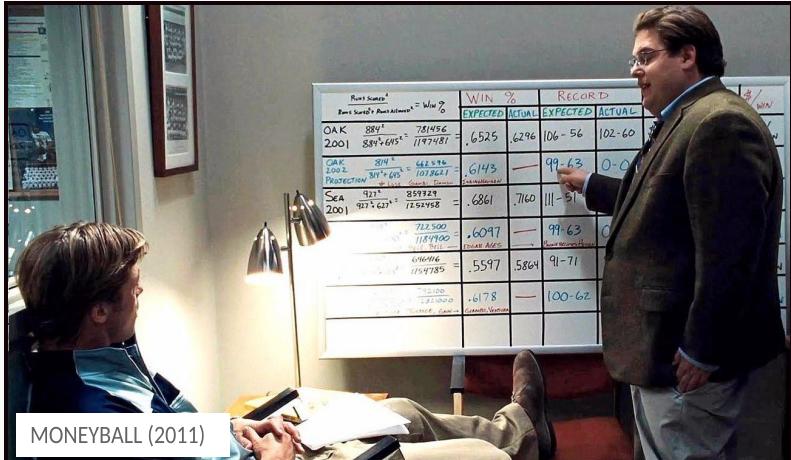
# STEPPING UP TO THE PLATE

Predicting MLB Player Value & Team Wins with Machine Learning

Eric Au

# BUSINESS UNDERSTANDING

## THE FIRST PITCH



- **How does a team determine the monetary value of a player?**

- Teams are structured around how they will pay their most valuable players.

- **Player's Salary = Player's Value?**

- Value is subjective
- Predict player salary based on performance

- **Winning In a Game of Failure**

- Predicting team wins based on team metrics

# BY THE NUMBERS

*“Baseball has always been filled with negative statistics” - Joe Torre*

## DATA SOURCES:

- FanGraphs, Baseball Reference

## BATTERS

- **Basic:** 9100 x 28 Features (2000 - 2021)
- **Advanced:** 3200 x 321 Features (2014 - 2021)

## PITCHERS

**Basic:** 9400 x 34 Features (2000 - 2021)

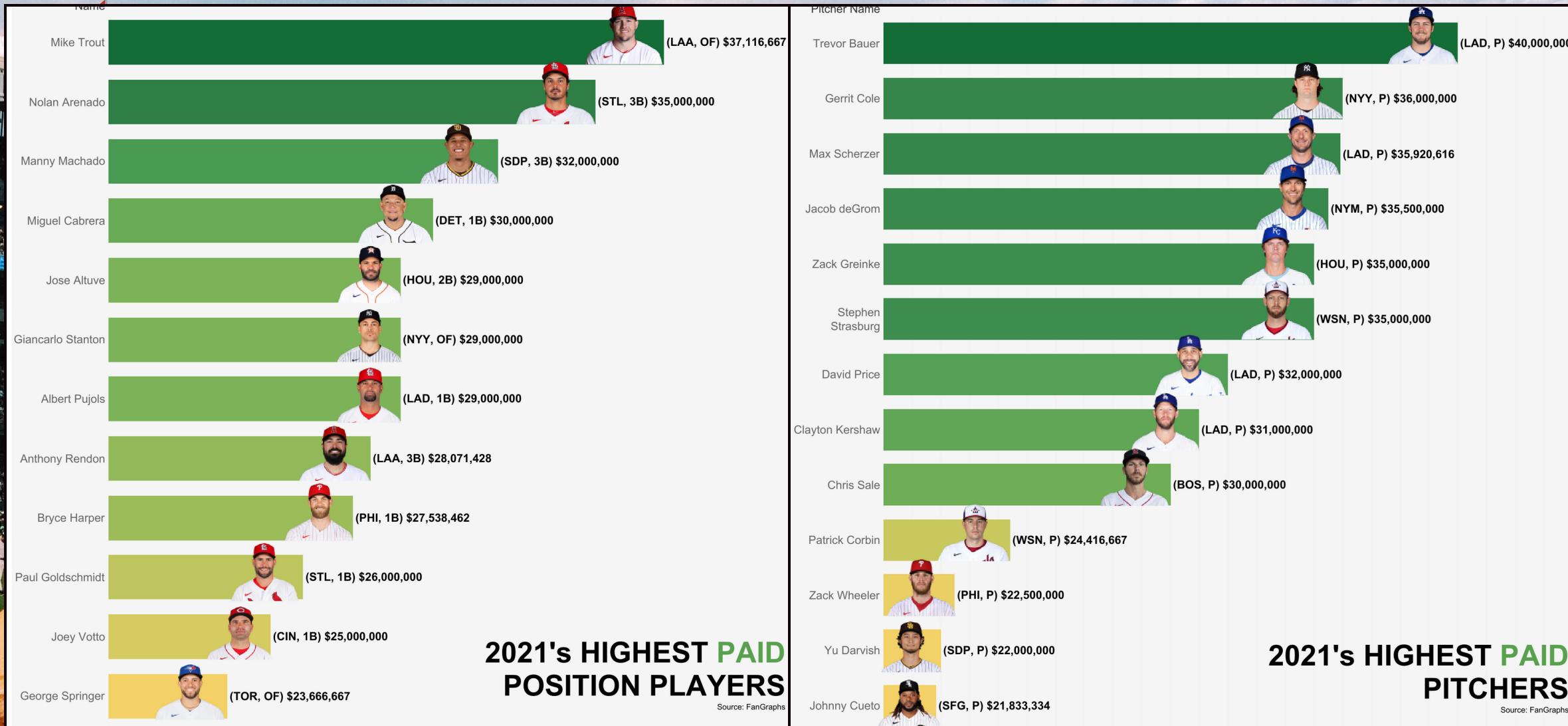
**Advanced:** 3900 x 334 Features (2014 - 2021)

## TEAMS

**Basic:** 1600 x 61 Features (1960 - 2021)

**Advanced:** 210 x 634 Features (2014 - 2021)



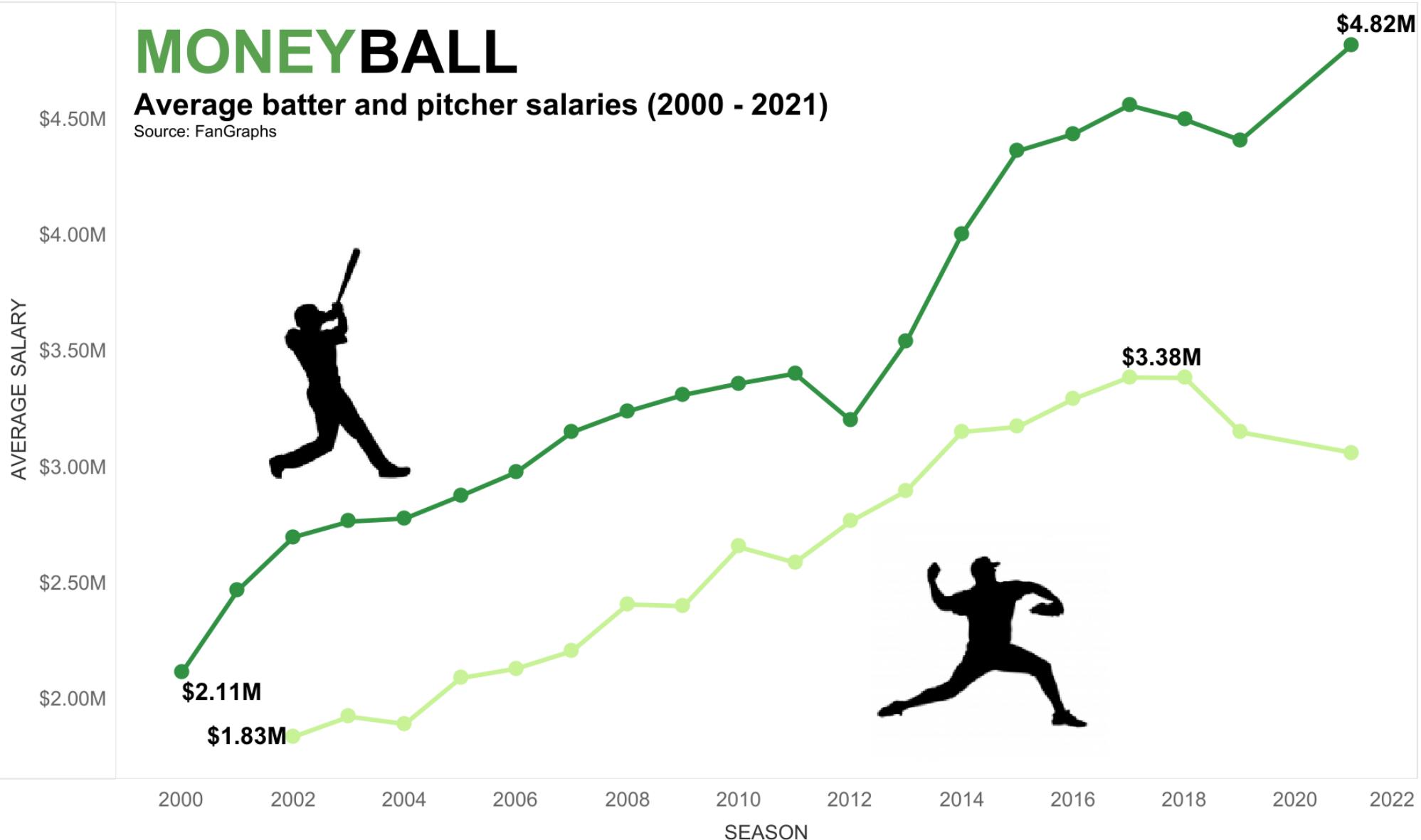


0 → 3 15 MIA CLE 0  
1 → 2 29 PHI CIN 1  
2 → 1 44 MIL 1  
GM 2 50 PIT 1  
1 → 16 SF ARI 1  
8:15 29 COL 33 SD  
10:05 85FT TA  
MADE WITH beautiful.ai

# MONEYBALL

## Average batter and pitcher salaries (2000 - 2021)

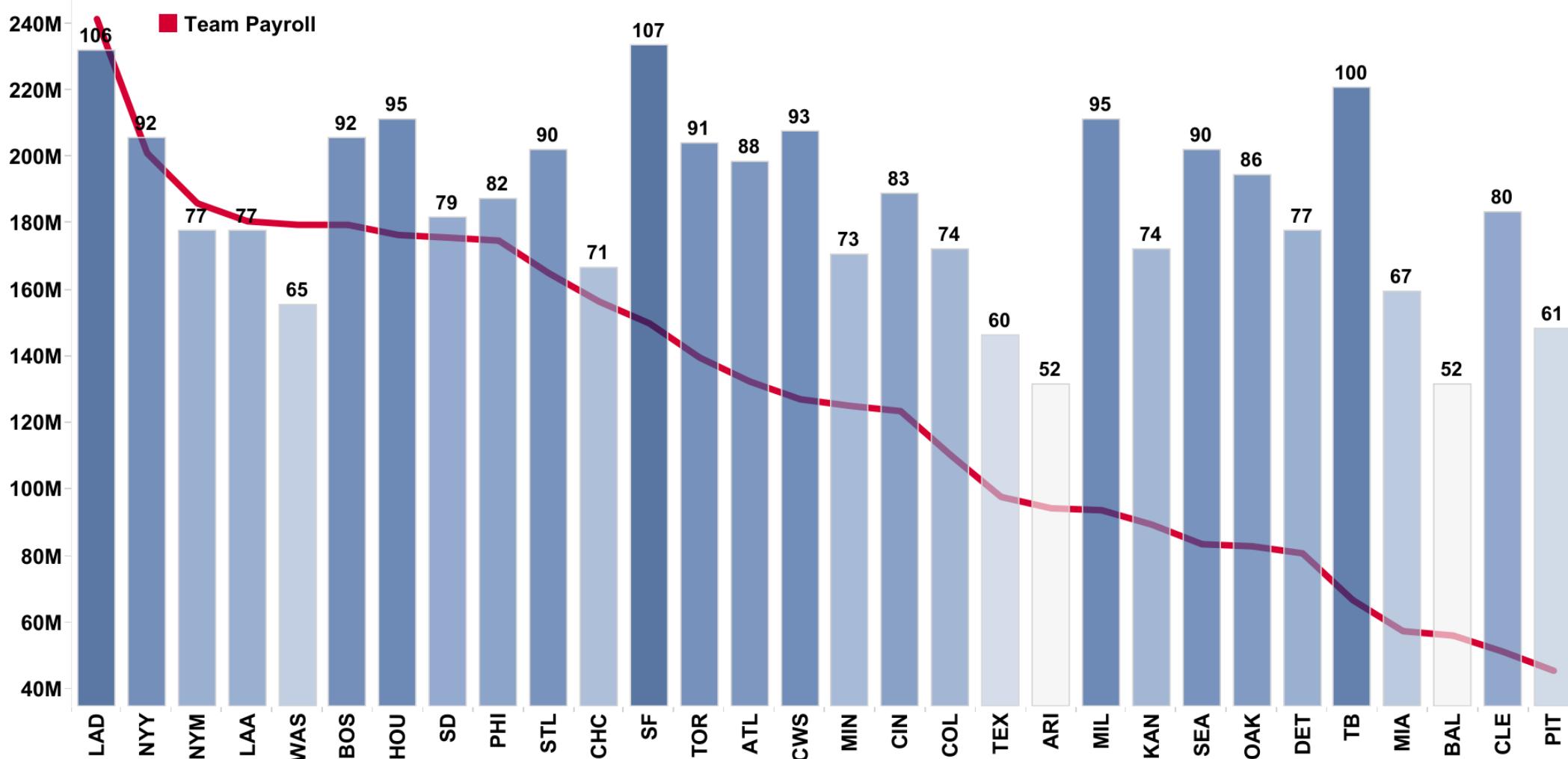
Source: FanGraphs





# MLB TEAM SALARIES & WINS (2021)

Team Name



# IF YOU BUILD IT ...

## THE DATA MODELING PROCESS

Collect Data

Data Cleaning

Feature Engineering

Modeling (Training)

Finalize Model

Batting,  
Pitching,  
Team, Salaries

Adjusting for  
Inflation,  
handling  
advanced  
sparse data

Removing  
related  
features

75% Training  
25% Testing  
5-Fold  
Validation

Which  
model(s)  
performed the  
best?

# A TRIPLE PLAY RESULTS

BASIC DATA  
(2000 - 2021)

ADVANCED DATA  
(2014 - 2021)

## BATTERS

(GRADIENT BOOST  
REGRESSION)

**±\$1.8M**

WITH BASIC BATTING DATA

R2 = 0.75

**±\$2.8M**

WITH ADVANCED BATTING  
DATA

R2 = 0.78

## PITCHERS

(GRADIENT BOOST  
REGRESSION)

**±\$1.4M**

WITH BASIC PITCHING DATA

R2 = 0.73

**±\$2.4M**

WITH ADVANCED PITCHING  
DATA

R2 = 0.76

## TEAM WINS

(LINEAR REGRESSION)

**±3 Wins**

WITH BASIC TEAM DATA

R2 = 0.92

**±1 Win**

WITH ADVANCED TEAM  
DATA

R2 = 0.98

STRONG LINEAR  
RELATIONSHIP BETWEEN  
FEATURES AND WINS

# CONCLUSIONS & NEXT STEPS

**“Every strike brings me closer to the next home run”**  
**- Babe Ruth**



# Arbitration\* Value

**\*Arbitration: salary numbers for the upcoming season that are negotiated for players not yet eligible for free agency**

- Models do better with predicting salaries of players before free agency.

# Superstar Factor

- Predicting salaries of superstar players more difficult
  - Marketability, Social Media Presence,
    - Free Agent Market Demand

# CONCLUSIONS & NEXT STEPS

**"Every strike brings me closer to the next home run"**  
- Babe Ruth



## Keys to Winning

- Build a team around players who excel at advanced metrics
- Fastballs are dominant
- Timely hits and timely pitching AKA WPA\*

Win Probability Added (WPA) captures the change in Win Expectancy and credits or debits the player based on how much their action increased their team's odds of winning



## Statcast

- Advanced statistics can better explain target variable
- Relatively new technology, constant data collection



**THANK YOU**

Email: eric8395@gmail.com

Github: <https://github.com/eric8395/baseball-analytics>

LinkedIn: [www.linkedin.com/in/eric-au8395](https://www.linkedin.com/in/eric-au8395)