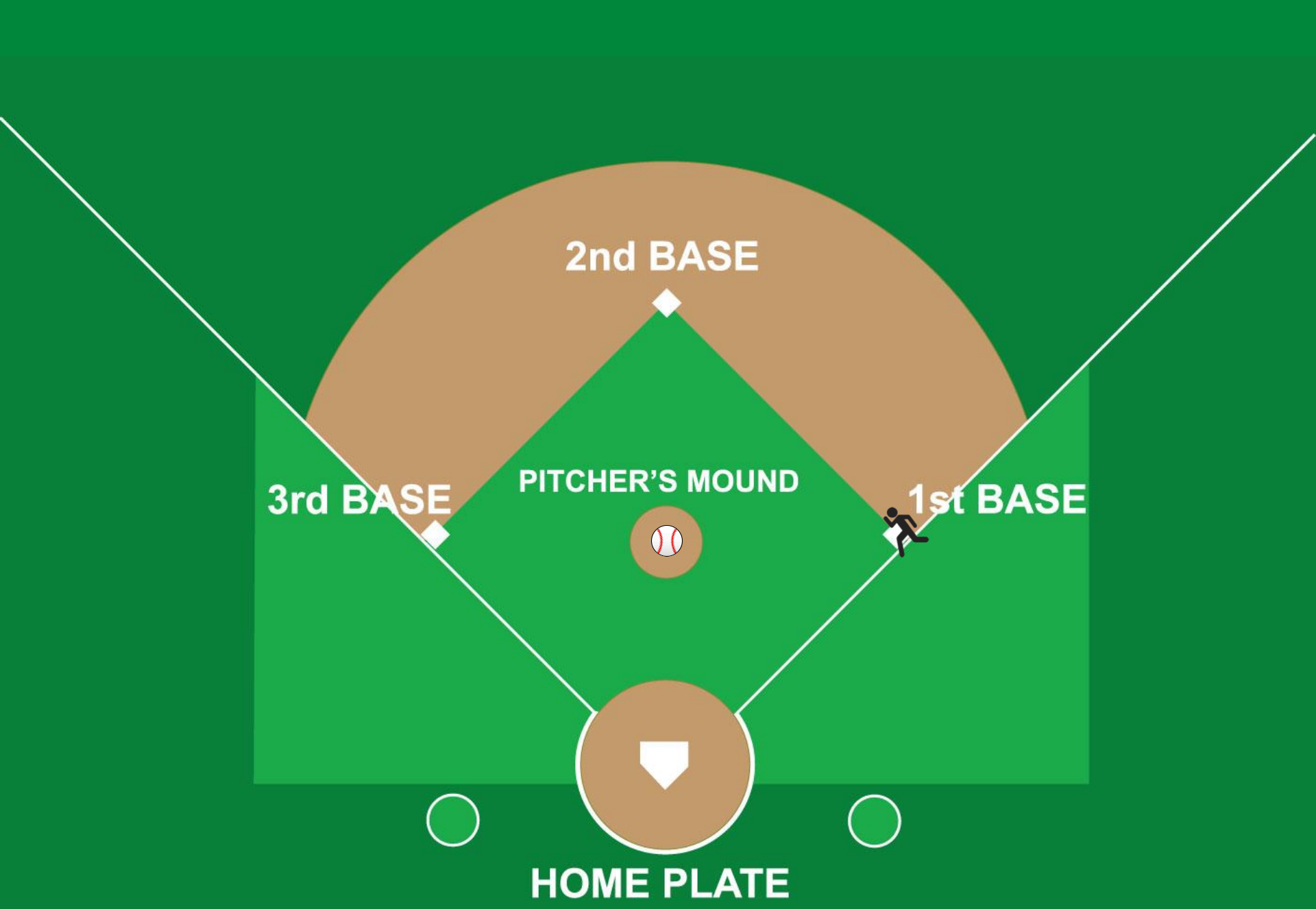
# Should you go for it?

Predictive base-stealing model for MLB


2nd BASE
3rd BASE
PITCHER'S MOUND
1st BASE
HOME PLATE

# Data Sources

- Sportsradar (Google Bigquery)
  - Every play in 2016 MLB season
- Fangraphs
  - Player statistics

# Proxy Features

Certain information not accessible to runner, proxies used instead:

- Pitch speed and type:
  - Pitchers' pitch distributions from 2015
- Pitch Location:
  - Hitter plate discipline stats from 2015

# Optimization Metric

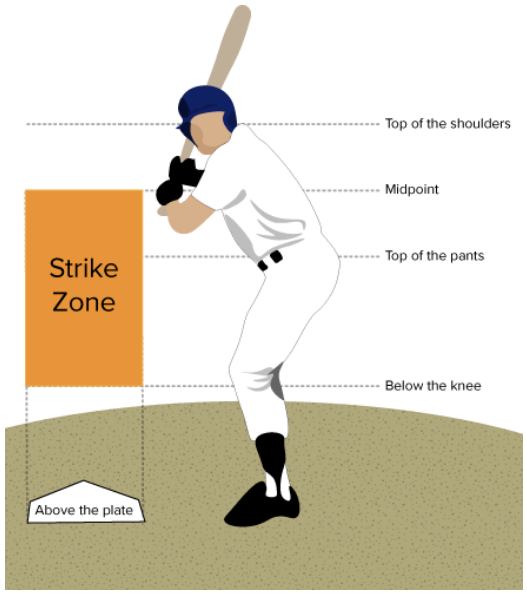
- Optimizing for base steals:
  - Too aggressive
- Minimizing times caught stealing:
  - Too passive
- Balanced approach:
  - Metric: F1

# Modeling

| Model | Training F1 |
|---|---|
| Dummy Classifier (Stratified) | 0.749 |
| **Logistic Regression** | **0.937** |
| Gradient Boosted Trees | 0.938 |

# Feature Importance

| Stolen Base | Caught Stealing |
|---|---|
| **If runner is on 1st base** | Batter: Contact % (BC) |
| Batter: Swing % outside zone (BC) | Batter: Swing % (BC) |
| Batter: First pitch strike % | Pitcher: % Changeup throws |



# Results

- Test F1: 0.924
- Final model: 992 CS ≈ 123 saved runs

# Application

Demo:

http://127.0.0.1:5000/

# Future Work

Improvements:

- Account for multiple base runners
- Pickoff events
- More data:
  - CS events
  - Pitcher and batter statistics
  - Incorporate runner specific information

# THANK YOU!