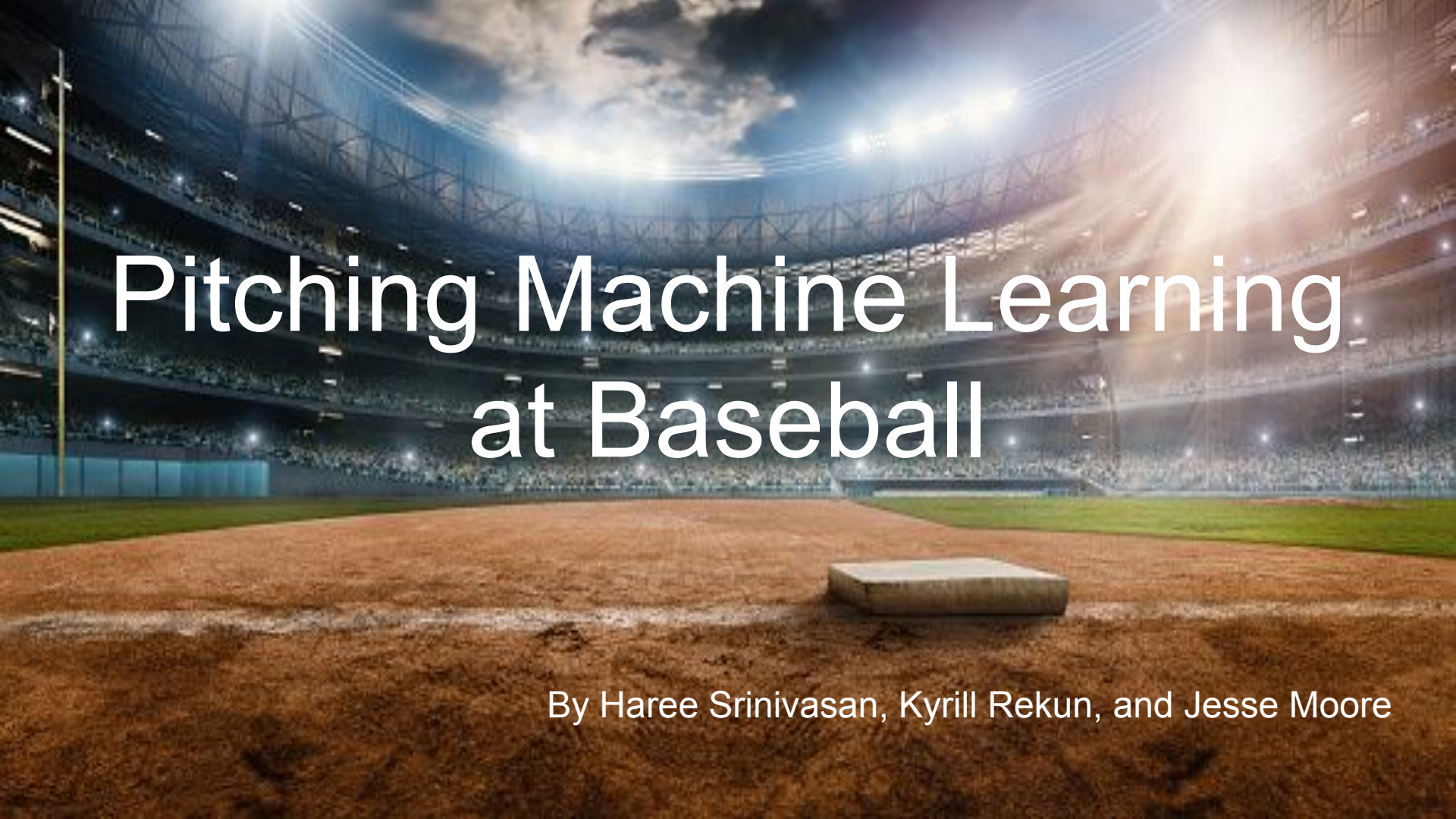
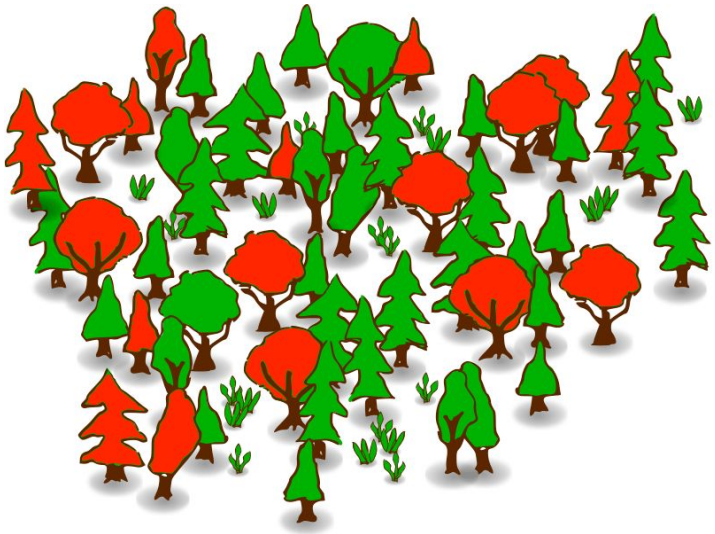


Pitching Machine Learning at Baseball

By Haree Srinivasan, Kyrill Rekun, and Jesse Moore

Can Our Dinky Machine Learning Models Outdo the Astro's Cheating?



VS

Baseball and Empirical Analysis: a Tale as Old as Time

- Sabermetrics is the empirical analysis of baseball
- Analysts have been using it for decades to improve their team's performance, but there have always been a human touches sprinkled in
- Our goal is to use machine learning to model a pivotal variable to the game: pitches

Acquiring Our Data

- The data that we are using to build our model is pitch-level data from every MLB game that took place from 2015 to 2018

- Source: https://www.kaggle.com/pschale/mlb-pitch-data-20152018#games.csv, scraped from http://gd2.mlb.com/components/game/mlb/

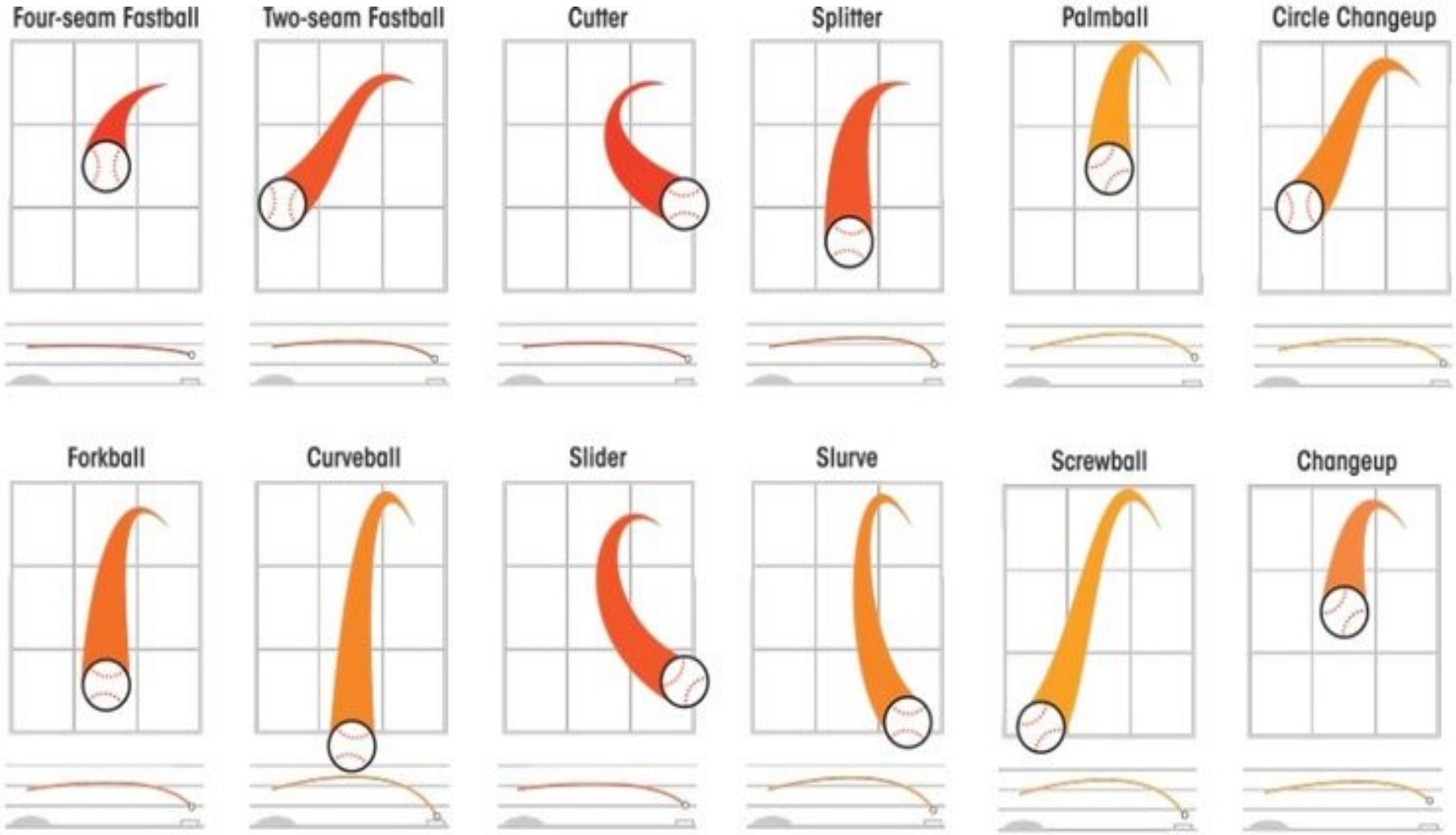
How Many Types of Pitches Can There Be?

Fast Pitches

1. Four-Seam Fastball ☆
2. Cutter
3. Two-Seam Fastball
4. Sinker

Slow Pitches

1. Curveball ☆
2. Changeup ☆
3. Slider ☆
4. Knuckle Curve
5. Splitter
6. Knuckleball
7. Screwball
8. Eephus


Four-seam Fastball
Two-seam Fastball
Cutter
Splitter
Palmball
Circle Changeup
Forkball
Curveball
Slider
Slurve
Screwball
Changeup

Model 1: Pitch Classification

- **Goal:** Classification of Fastball, Curve, Slider and Changeup based on velocity, spin, and trajectory data
- **Processing**:
  - Filter rows for only the above mentioned 4 pitches
  - Features: **start_speed, end_speed, spin_rate, spin_dir, break_angle, break_length, break_y**
  - Angular data initially had to be cleaned to be symmetric for both L and R handed pitched (ultimately turned out this was not needed)
- **Modeling**:
  - RandomForest was the model of choice

Results

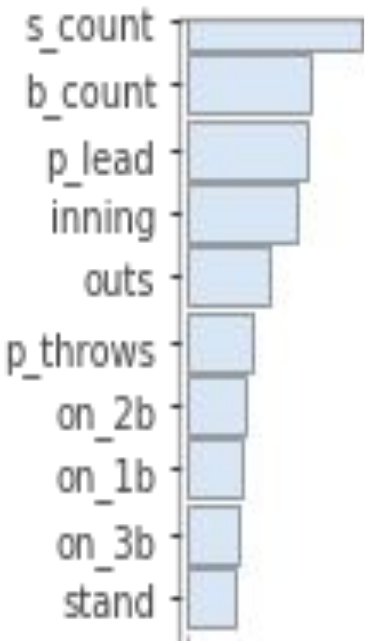
Train F1-Score:	0.99
Test F1-Score:	0.92

Top misclassifications:

- Fastball vs “fast” slider that didn’t break much
- Curveball vs slower slider with a lot of break

Model 2: Fastball or Offspeed

- **Parameters:**
  - Categoricals: Runner on 1st, 2nd, 3rd
  - Numerical: Difference in Score, Balls, Strikes, Outs, Inning
- **Target:** 1: Fastball, 0: Offspeed
  - Distribution: about 61% fastball, 39% offspeed
- **Pipeline:** Feature Engineering ⇨ Random Forest
- After some cross-validation and hyper parameter tuning...
- **F1-Score:** 0.76



F1-Score	≈1	0.76
On-Base Percentage	0.342	Never made it to the majors, but never got out
Cost	A camera, some trash cans, and their souls	You know what this program has taken from you...
Budget	Net worth: $1.8 Billion	Just enough to buy some oreos
Morality	“If you aren’t cheating, you aren’t trying.”	WE DIDN’T CHEAT!!!

The Moral: Why Cheat When You Can Use Machine Learning Instead?