Baseball Classification

~The wrong way to predict MLB winners
but we did it anyway

Sean Carver

Emefa Agodo

Dataset

We used 2011 play-by-play data (30 MLB teams)

We created 1000 fantasy teams by randomly selecting transition probabilities with same mean and covariance.

We simulated a season to label the top scorers.

Question: Which MLB team will be a top ten scorer at their home stadium ?

Classifying Teams

Our Markov chains describe Baseball with 313 nonzero parameters (our features, derived for each team from play-by-play data).

We labeled the teams above the 66.667 percentile for scoring, and predicted the label for test teams.

We used classification but we could use simulations.

Classification Models

	Accuracy Score	F1 Score
RF	0.92	0.84
SVM	0.90	0.81
KNN	0.82	0.69
DT	0.80	0.67

Future Work

-Classify teams by American vs National League

-Explore significance of pitch sequence

-Explore pitch hand (right vs left)

-Player analysis

-Effect of offensive and defensive plays

Data Source

https://github.com/maxtoki/baseball_R

Thank you!!