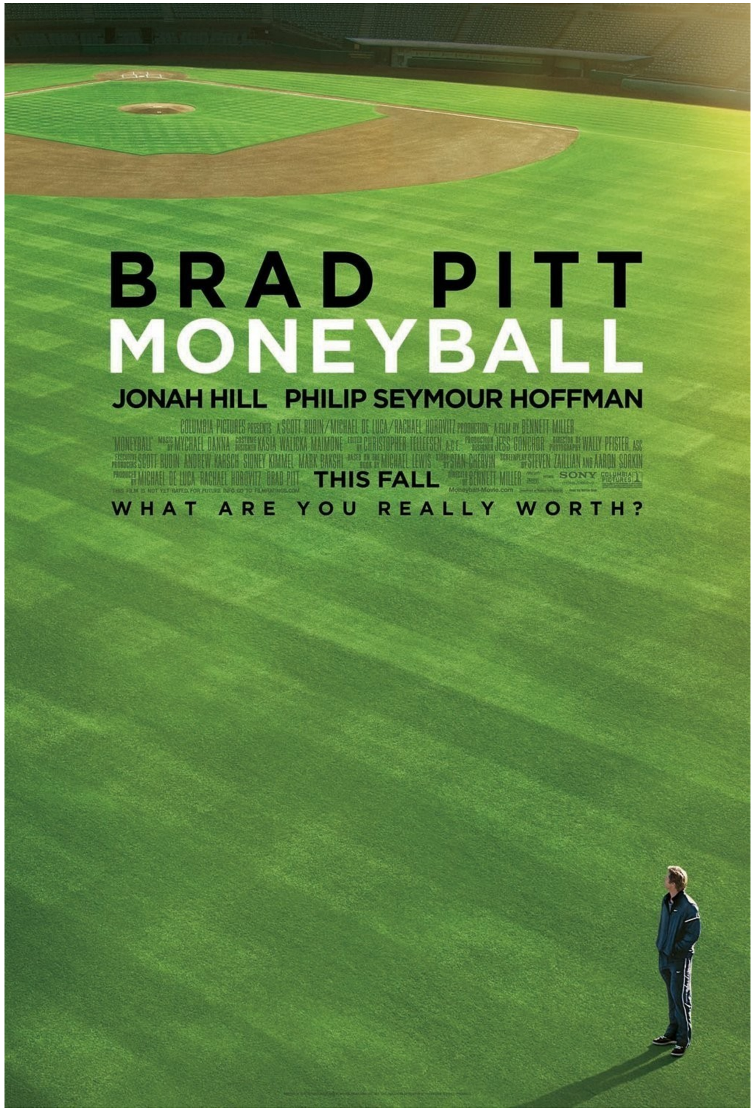
Analyzing Sports Data

Ken McAlinn
Joe Futoma

Analyzing Sports Data

- Analyzing sports data has been popular for the last couple of decades
- Brought to popular attention in the last five years



Utilizing Data Analytics

- Team building: How to build a good team using limited resources (good but cheap players)
- Performance analysis: Predicting player's performance through and after the season (determining roster, salary etc…)
- Franchise management: Predicting team performance (ticket, merchandise sales etc…)

Utilizing Data Analytics

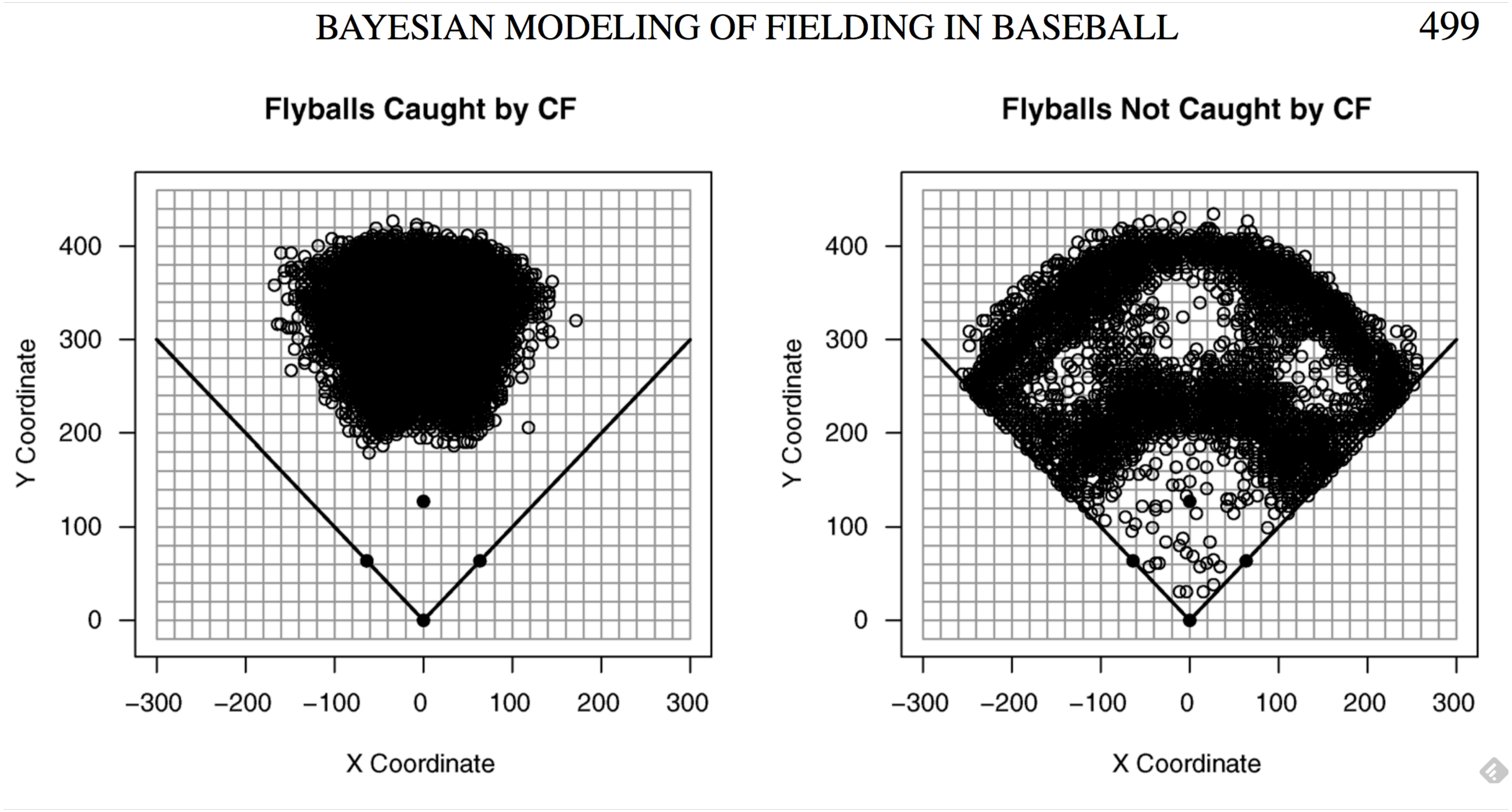
- Sports betting: Predicting team performance and player performance (fantasy sports)

- Fantasy sports:
  - 33.5 million players in the U.S. in 2013
  - $3-4 billion dollar industry

Statistics in the Sports Industry

- Statisticians are in high demand
- A lot of superstitions in these sports
- Some sports are easier to analyze than others
- Baseball: All performances can be quantified (batting, pitching, fielding skills etc…)
- Basketball: Some skills can be easily quantified (shooting skills) but others are difficult to quantify (defensive skills)

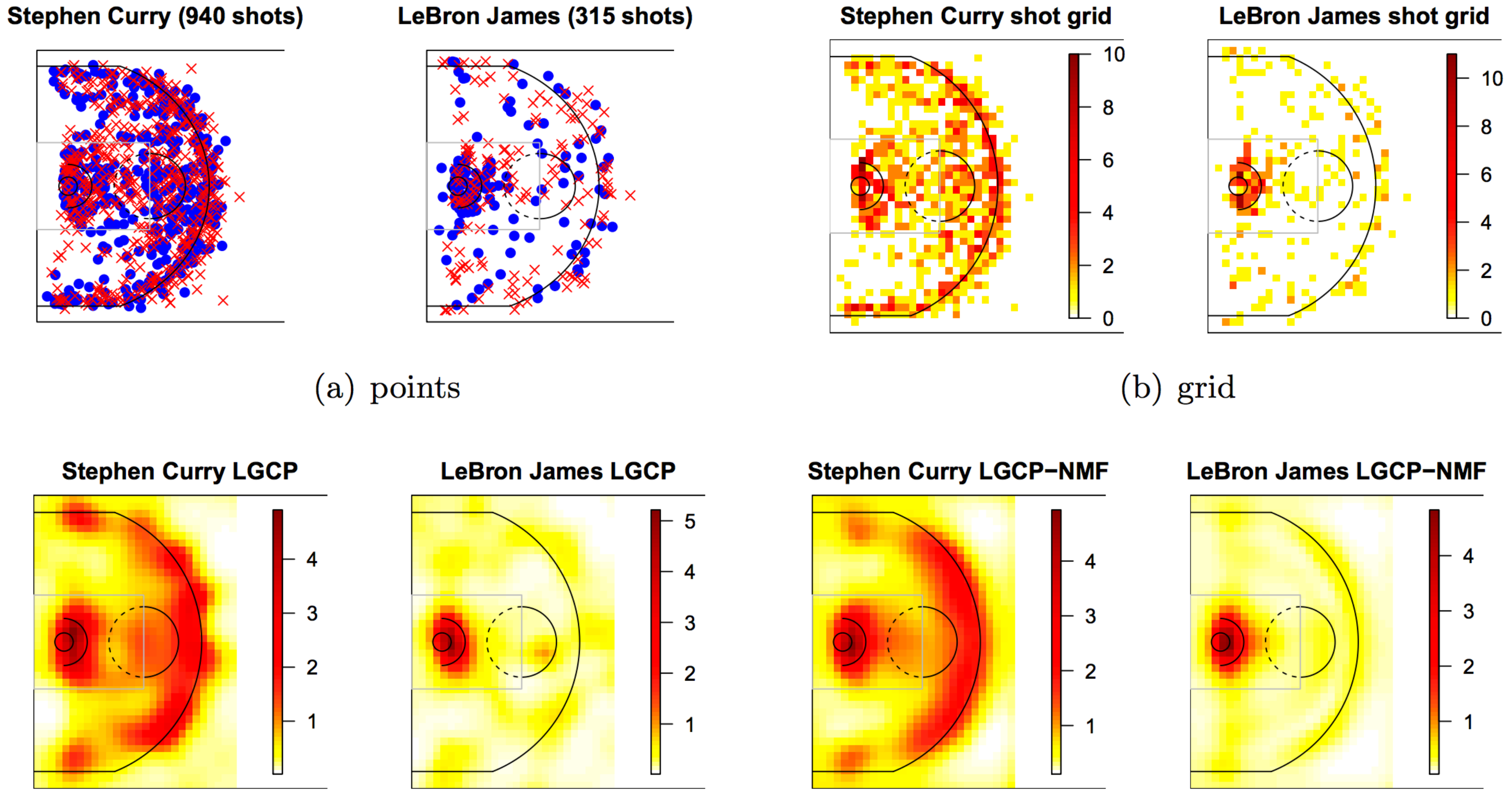
Some Recent Developments



BAYESIAN MODELING OF FIELDING IN BASEBALL 499

[Jensen et al. 2009]

Some Recent Developments


Stephen Curry (940 shots)

LeBron James (315 shots)

Stephen Curry shot grid

LeBron James shot grid

Stephen Curry LGCP

LeBron James LGCP

Stephen Curry LGCP−NMF

LeBron James LGCP−NMF

(a) points

(b) grid

(c) LGCP

(d) LGCP-NMF

[Miller et al. 2014]

MLB and NBA Data

- MLB
  - League revenue: 7.1 bil.
  - Average salary: 3.39 mil.
  - Franchise income: around 200 mil.
- NBA
  - League revenue: 4.6 bil.
  - Average salary: 4.2 mil.
  - Franchise income: around 200 mil.

MLB Data

- List of summary statistics for each team in the 2013 season (162 games per team, 30 teams)

- 23 batting stats: Hits, runs batted in, homeruns etc…
- 23 pitching stats: Earned run average, strikeouts, runs allowed

[will work on it today]

NBA Data

- List of summary statistics for each team in the 2012-13 season (82 games per team, 15 teams)

- 20 stats: Field goals, free throws, rebounds etc…

[will work on it next week]

How we got the Data

- www.baseball-reference.com, www.basketball-reference.com
- Python script to crawl the webpages and collect relevant statistics
- Look at html source and figure out what I need, then write (ugly) code to get it
- Example (row 1 of MLB dataset):
  - http://www.baseball-reference.com/boxes/CIN/CIN201304010.shtml

Research Questions (MLB)

- Do pitchers/batters perform better in warmer climates:
  - It is believed that pitchers pitch better in warmer climates (avoiding elbow injuries etc…)
  - How can we compare/visualize this?
  - Can we show something using a map?
  - Can we come up with a weighted score of the stats and compare them between teams?
  - How can we compare across groups?

Research Questions (MLB)

- Is there a home field advantage?
  - Are some teams better in their home field?
  - Are all teams better in their home field?
  - How can we quantify "home field advantage"?

Research Questions (MLB)

- Are team stats noticeably different between leagues?
  - Different leagues have different rules
  - The American League has the designated hitter rule: Would this change pitcher/batter performance?
  - How can we compare across leagues?

Research Questions (NBA)

- Similar questions to MLB data
  - Is there a home court advantage?
  - Can we come up with a weighted score of the stats and to quantify team strength?
- Some other interesting questions
  - Can we predict win/loss through predicting point spread?
  - What kind of model will perform the most?

Application Exercise 13

- Are team stats different by league (American vs. National)?
  - Different leagues have different rules
  - The American League has the designated hitter rule, does this change pitcher/batter performance? Can we attribute any differences we find to this rule?
  - What are some techniques we can use to compare across leagues?
- **Task:** Organize data into the two leagues (AL and NL) then perform hypothesis tests on a few crucial stats (batting average and ERA, for example) to test if they differ.
- **Data:** https://stat.duke.edu/courses/Fall14/sta112.01/data/MLB2013.html

HW4 & Office Hours

- HW4 can be found at https://stat.duke.edu/courses/Fall14/sta112.01/hw/hw4.html and is due next Tuesday.

- Joe & Ken next week, Old Chem 211:
  - Monday 4:30 - 5:30pm
  - Wednesday 4:45 - 5:45pm

- Dr. Çetinkaya-Rundel next week (adjusted to not overlap):
  - Monday 3:30 – 4:30pm
  - Wednesday 11:30am – 12:30pm
  - by appointment