Take me out to Big Data: analyzing professional baseball data online with SciServer

Jordan Raddick
IDIES, SciServer outreach team

idies

Johns Hopkins University

I use **SciServer** to analyze and share game data from more than 100 years of Major League Baseball. This dataset provides an amazing opportunity to teach data science and statistics to an engaged global audience.

Why baseball?

- Fans know statistics (e.g. batting average, wins above replacement...)
- Data-driven managing
- Great opportunity for education

Why big data?

- Lots of data online
- Lots of interesting questions to ask

Why SciServer?

- Online analysis with Python (or R)
- Easy to share data and scripts

Box scores from game data

All-time MLB box score

Team	Runs	Hits	Errors
Visitor	794,733	1,637,837	167,352
Home	822,947	1,621,067	172,440

Average MLB box score

Team	Runs	Hits	Errors
Visitor	4.36	8.99	0.92
Home	4.52	8.90	0.95

Median MLB box score

Team	Runs	Hits	Errors
Visitor	4	9	1
Home	4	9	1

The first game in the dataset

May 4, 1871:
Cleveland Forest Cities at Fort Wayne Kekiongas

```
Retrosheet Expanded Box Score

Game of Thursday, 5/4/1871 - Cleveland at Ft. Wayne (D)

Cleveland          000 000 000- 0
Ft. Wayne          010 010 000- 2

-------------------------------------------------------
CLEVELAND          AB  R  H BI BB SO  PO  A
-------------------------------------------------------
D.White, l         4  0  2  0  0  0   3  0
C.Kimball, 2b      4  0  0  0  0  0   0  3
A.Pratt, p         4  0  0  0  0  0   0  0
G.Allison, cf      4  0  1  0  0  2   2  0
E.White, rf        3  0  0  0  0  3   1  0
J.Pratt, p         2  0  0  0  0  1   1  3
E.Sutton, 3b       3  0  1  0  0  0   0  1
J.Carleton, 1b     3  0  0  0  0  1   6  0
J.Bess, ss         3  0  0  0  0  0   1  4
-------------------------------------------------------
Totals            30  0  4  0  0  6  27  9
-------------------------------------------------------
BATTING
2B: D.White (off B.Mathews).
RBI, scoring position, less than 2 outs: G.Kimball 0-1;
    A.Pratt 0-1.
BASERUNNING
SB: A.Allison (2nd base off B.Mathews/B.Lennon).
CS: A.Allison (2nd base by B.Mathews/B.Lennon).
Team LOB: 4
FIELDING
PB: D.White 3.
```

```
-------------------------------------------------------
FT. WAYNE          AB  R  H BI BB SO  PO  A
-------------------------------------------------------
F.Sellman, 3b      4  0  0  1  0  0   2  1
B.Mathews, p       4  0  0  0  0  0   1  0
J.Foran, 1b        4  0  1  0  0  0   1  0
W.Goldsmith, ss    3  0  0  0  1  0   1  0
B.Lennon, c        4  1  1  0  0  0  10  1
T.Carey, 2b        3  0  0  0  0  0   6  0
E.Hincher, lf      3  0  0  0  0  0   2  0
J.McDermott, cf    3  0  1  1  0  0   0  1
D.Kelly, rf        3  1  1  0  0  0   2  0
-------------------------------------------------------
Totals            31  2  4  2  1  0  27  3
-------------------------------------------------------
BATTING
2B: B.Lennon (off A.Pratt).
2-out RBI: J.McDermott.
RBI, scoring position, less than 2 outs: F.Sellman 1-1;
    T.Carey 0-1; E.Hincher 0-1.
BASERUNNING
Team LOB: 3
FIELDING
E: B.Lennon (drop fly); W.Goldsmith (fumble);
    J.McDermott (fumble).
PB: B.Lennon.
Outfield assist: J.McDermott (D.White at 2B).
DP: (1). T.Carey, unassisted.
```

Data

- Source: retrosheet.org
- Online volunteers transcribe historic baseball data
- **Games:** box scores, lineups, game stats; one row per game
- **Events:** outcomes of at-bats or similar (e.g. stolen base); one row per event
- **Pitches:** outcome of individual pitches (e.g. ball, strike); one row per pitch

Processing

Python scripts in SciServer Compute:

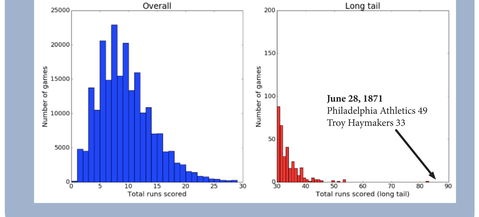
1. Copy data from retrosheet.org
2. Load game data as JSON
3. Parse game and event data into CSV
4. Calculate game progress (outs, score)

My scripts are inefficient, please help!

Data size and completeness

Data type	Complete-ish	Partial	Rows
Games	1901-2016	1874-1900	213,307
Events	1974-2016	1921-1973	~10 million
Pitches	2009-2016	1930-2008	~40 million

Histogram of total runs per game



Future plans

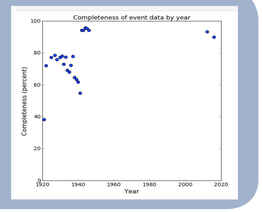
- Aggregate to get player, team data
- Write tutorials and lesson plans
- Share data and scripts through SciServer
- Look at data from other sports

Event data

Slowly processing event data in Compute

Loading progress

Years	Events
1921-1946	1,627,509
2012	177,720
2016	171,522



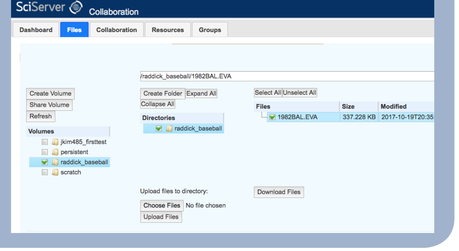
Event data format

	inning	batting_team	play	visitor_score	home_score
0	7	TBA	43	0	1
1	7	TBA	63	0	1
2	7	TBA	3	0	1
3	8	TBA	K	0	1
4	8	TBA	K	0	1
5	8	TBA	K	0	1
6	9	TBA	K	0	1
7	9	TBA	63	0	1
8	9	TBA	K	0	1

Future: Research Questions

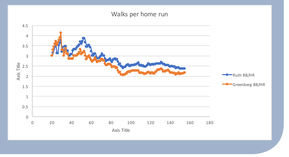
- Ideas for other research questions?
- Write your email and I'll contact you

Future: Easy data sharing



Event data: the home run chase

1927: Babe Ruth hit 60 home runs, was a hero.
1938: Hank Greenberg hit 58 home runs. Were anti-Semitic pitchers avoiding him?



Fun events in 2016

Date	Batting	Fielding	Inn.	Score	Batter	Scorecard says...
June 17	Arizona Diamond-backs	Philadelphia Phillies	5	3-2	Wellington Castillo	CS3(1545E3)
July 6	Milwaukee Brewers	Washington Nationals	3	3-4	Chris Carter	K+SBH;SB2
October 1	Colorado Rockies	Milwaukee Brewers	9	3-2	Stephen Cardullo	5!6!3

What else could we study?

Join the team!

- Talk to me about baseball (or other sports)
- Check and improve my scripts
- Find pitchFX data (pitch type, speed, etc.)
- Suggest new research questions
- Share with colleagues and students

Pitch data

event	game	inning	battingteam	outs	balls	strikes	pitches	batter	batterhan
1	ANA201104080	1		0	0	1	2 **FBSX**	davir003	R
2	ANA201104080	1		0	1	0	0 **X**	nix-j001	R
3	ANA201104080	1		0	2	1	2 **CBCS**	bautj002	R
4	ANA201104080	1		0	3	1	1 **CBSBB**	iztum001	L
5	ANA201104080	1		1	0	2	2 **BCSBS**	kendh001	R
6	ANA201104080	1		1	1	2	1 **CBB1>S**	abreb001	L
7	ANA201104080	1		1	2	3	2 **CBB1>S.FBFB**	abreb001	L
8	ANA201104080	1		1	2	0	2 **CCX**	huntt001	R
9	ANA201104080	1		1	2	1	0 **BX**	wellv001	R
10	ANA201104080	2		0	0	2	2 **CBBFX**	linda001	L
11	ANA201104080	2		0	1	1	2 **BFCX**	hilla001	R
12	ANA201104080	2		0	2	2	2 **FBCBFX**	rivej001	R
13	ANA201104080	2		0	3	2	2 **CBSSBB**	calla001	L
14	ANA201104080	2		1	0	0	0 **X**	trumm001	R

B = ball; C = called strike; F = foul; S = swinging strike; X = ball in play; 1> = throw to runner at first

Analysis script example

```
# baserunning events not involving the batter
elif (re.match(r'WP|PB|SB|CS|DI|PO|BK', thisevent['play'])):
    thisevent['ab#ag'] = 0  # not an at-bat for any of these events

    # If defensive interference, advance has already been recorded in baserunning, so
    if (re.match(r'DI|BK', thisevent['play'])):
        break

    # If stolen base, advance base-stealer
    if (re.match(r'SB', thisevent['play'])):

        # parse the base that got stolen
        stolen_base = thisevent['play'][2]
        # Again, set scoring play (stealing home) to 4
        if (stolen_base == 'H'):
            stolen_base = 4
        else:
            stolen_base = int(stolen_base)

        # get base started from
        # THIS WON'T WORK FOR DOUBLE STEALS!
        orig_base = stolen_base - 1
        if (stolen_base < 1):  # only move ahead one base if home not stolen
            onbase[stolen_base-1] = onbase[orig_base-1]
        onbase[orig_base-1] = ''
        break

    # If caught stealing, clear the base the runner started from and increment outs
    if (re.match(r'CS', thisevent['play'])):
        # parse the base at which they were caught stealing
        caught_base = thisevent['play'][2]

        # Again, set scoring play (stealing home) to 4
        if (caught_base == 'H'):
            caught_base = 4
        else:
            caught_base = int(caught_base)
```

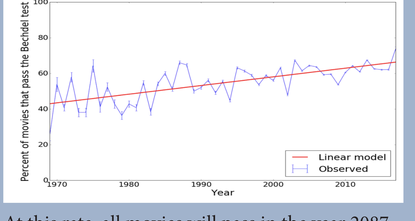
Not a baseball fan?

Take me out to the movies...

"Bechdel test": a simple metric for representation of women in movies.
Does this movie have...

1. Two or more female characters who...
2. talk to each other...
3. about something other than a man?

Bechdel test results for Hollywood 1969-2017



At this rate, all movies will pass in the year 2087.