

Take me out to Big Data:

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.

analyzing professional

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

Median MLB box score

Average MLB box score

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

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

baseball data online

Jordan Raddick

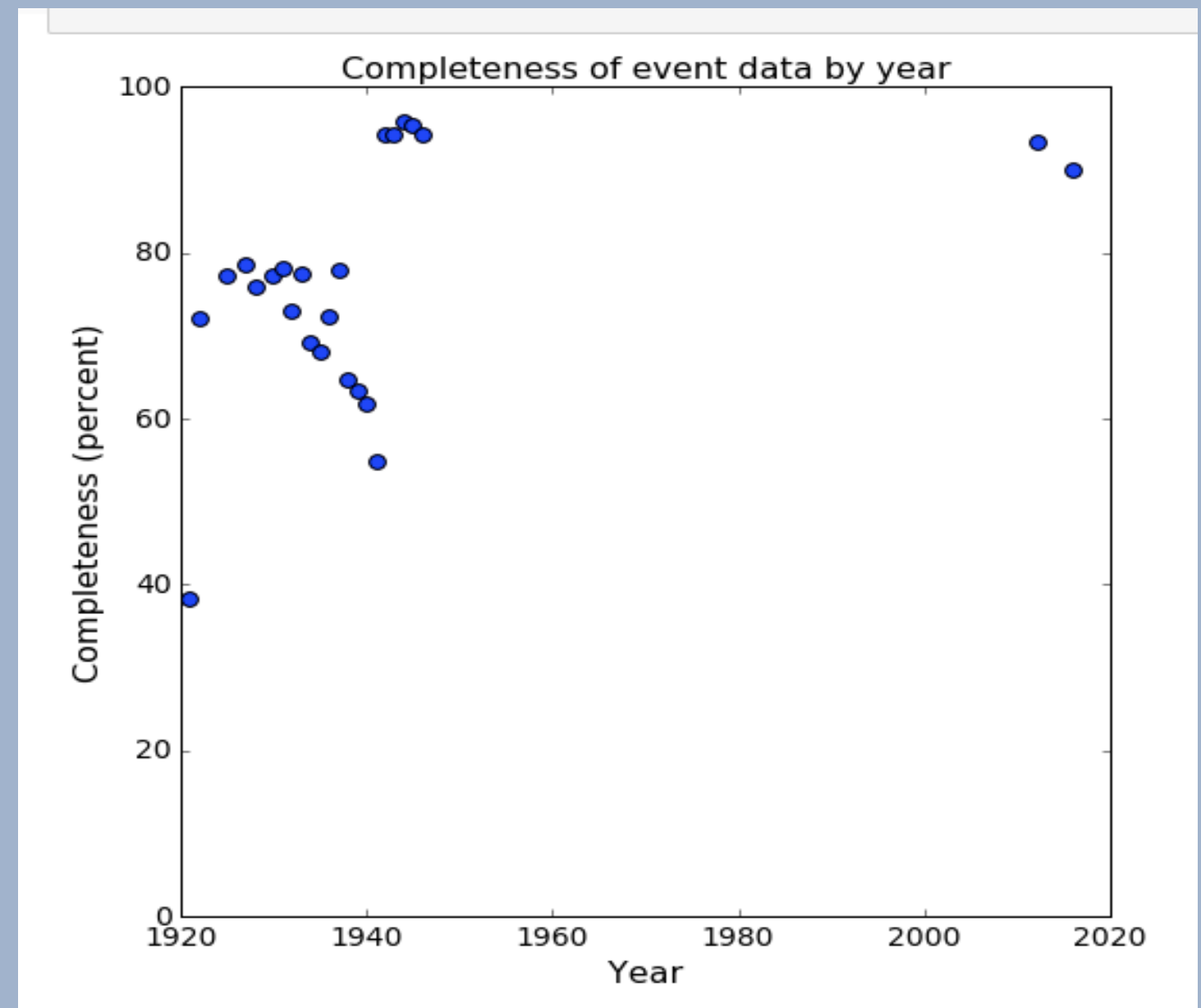
IDIES, SciServer outreach team

Event data

Slowly processing event
data in Compute

Loading progress

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

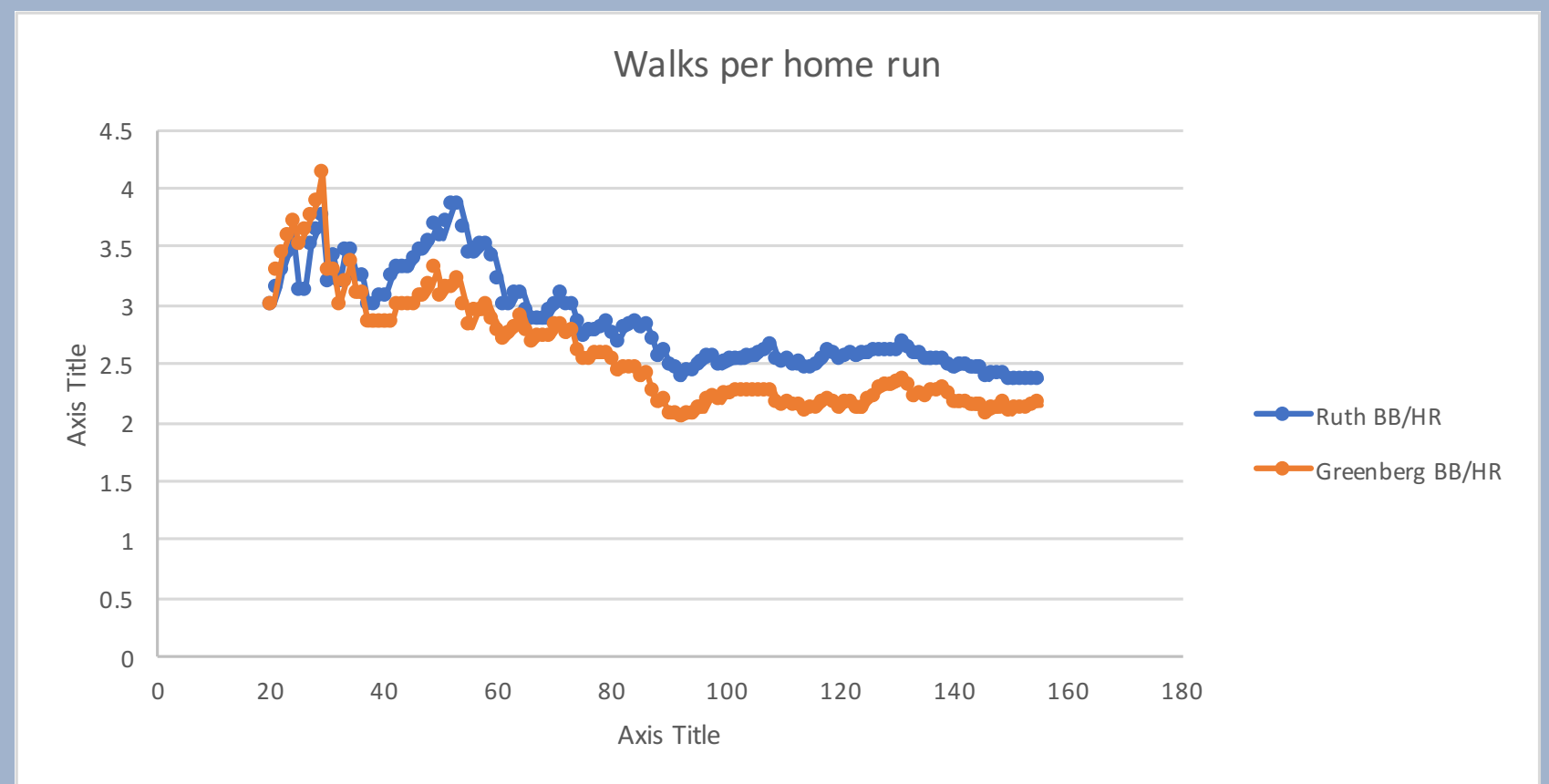


with SciServer

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?





JOHNS HOPKINS
UNIVERSITY

Pitch data

event	game	inning	battingteam	outs	balls	strikes	pitches	batter	batterhand
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	1	0	3	1	CBBBB	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	1	0	3	2	CBBSBB	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

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

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, c	4	0	2	0	0	0	9	0
G.Kimball, 2b	4	0	0	0	0	0	4	1
C.Pabor, lf	4	0	0	0	0	0	0	0
A.Allison, cf	4	0	1	0	0	2	2	0
E.White, rf	3	0	0	0	0	3	1	0
A.Pratt, p	2	0	0	0	1	0	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.Bass, ss	3	0	0	0	0	0	1	4
Totals	30	0	4	0	1	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	3	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.Mincher, lf	3	0	0	0	0	0	2	0
J.McDermott, cf	3	0	1	1	0	0	0	1
B.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.Mincher 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.

Event data format

RAYS

No.	PITCHERS (SEA)	IP	H	R	ER	BB	SO	HR	WP	W/L
34	F. Hernandez (R)	9	0	0	0	0	12	0	0	W

Manager: Joe Maddon (70)

No.	PLAYERS	POSITION	1	2	3	4	5	6	7	8	9	10	AB	R	H	RBI	SB
5	S. Fuld	LF	R	◇	◇	L	◇	◇	4	3	◇	◇	3	0	0		
2	B. Upton	CF	6	3	◇	◇	K	◇	◇	◇	◇	◇	3	0	0		
20	M. Joyce	RF	4	3	◇	◇	K	◇	◇	◇	◇	◇	3	0	0		
3	E. Longoria	DH	◇	K	◇	◇	L	◇	◇	K	◇	◇	3	0	0		
18	B. Zobrist	2B	◇	6	3	◇	R	◇	◇	K	◇	◇	3	0	0		
23	C. Peña	1B	◇	R	◇	◇	2	3	◇	K	◇	◇	3	0	0		
21	J. Lobaton	C	◇	◇	R	◇	◇	K	◇	◇	◇	◇	2	0	0		
9	E. Johnson	SS	◇	◇	L	◇	◇	K	◇	◇	◇	◇	2	0	0		
1	S. Rodriguez	3B	◇	◇	R	◇	◇	K	◇	◇	◇	◇	3	0	0		
8	D. Jennings (9th)	PH	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	1	0	0		
DATE	7 J. Keppinger (PH)	R/H	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	1	0	0		
TIME OF GAME	(1+L)	E/LOB	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	2	7	0	0	

No.	PITCHERS (SEA)	IP	H	R	ER	BB	SO	HR	WP	W/L	DOUBLE PLAYS-
-----	----------------	----	---	---	----	----	----	----	----	-----	---------------

34	F. Hernandez (R)	9	0	0	0	0	12	0	0	W	2B-
----	------------------	---	---	---	---	---	----	---	---	---	-----

3B-

HR-

E-

	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

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

Analysis script example

```
# baserunning events not involving the batter
elif (re.match(r'WP|PB|SB|CS|DI|PO|BK', thisevent['play'])):
    thisevent['abflag'] = 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 < 4):    # only move ahead one base if home not stolen
        onbase[stolen_base-1] = onbase[orig_base-1]
    onbase[orig_base-1] = ''
    break

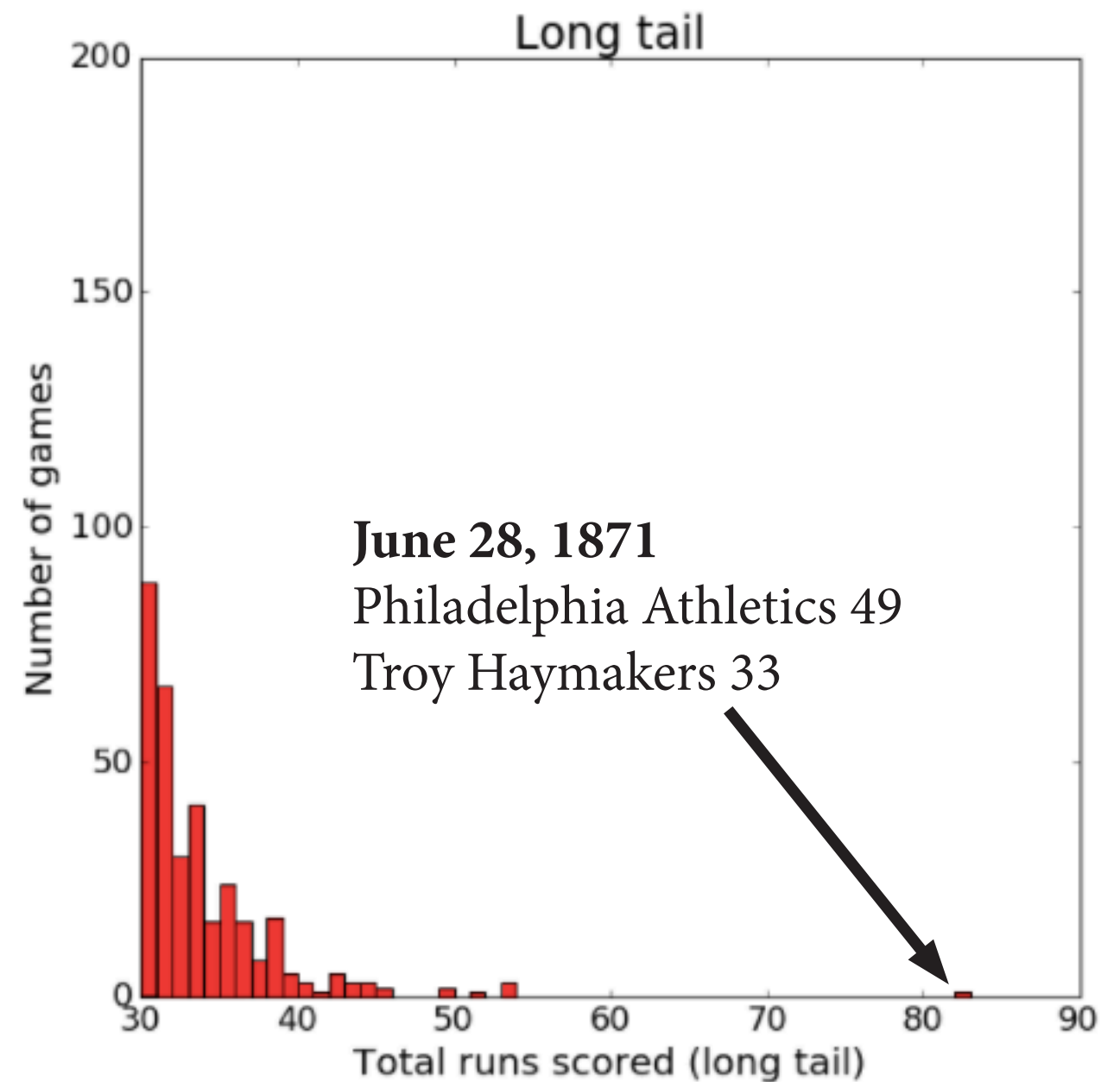
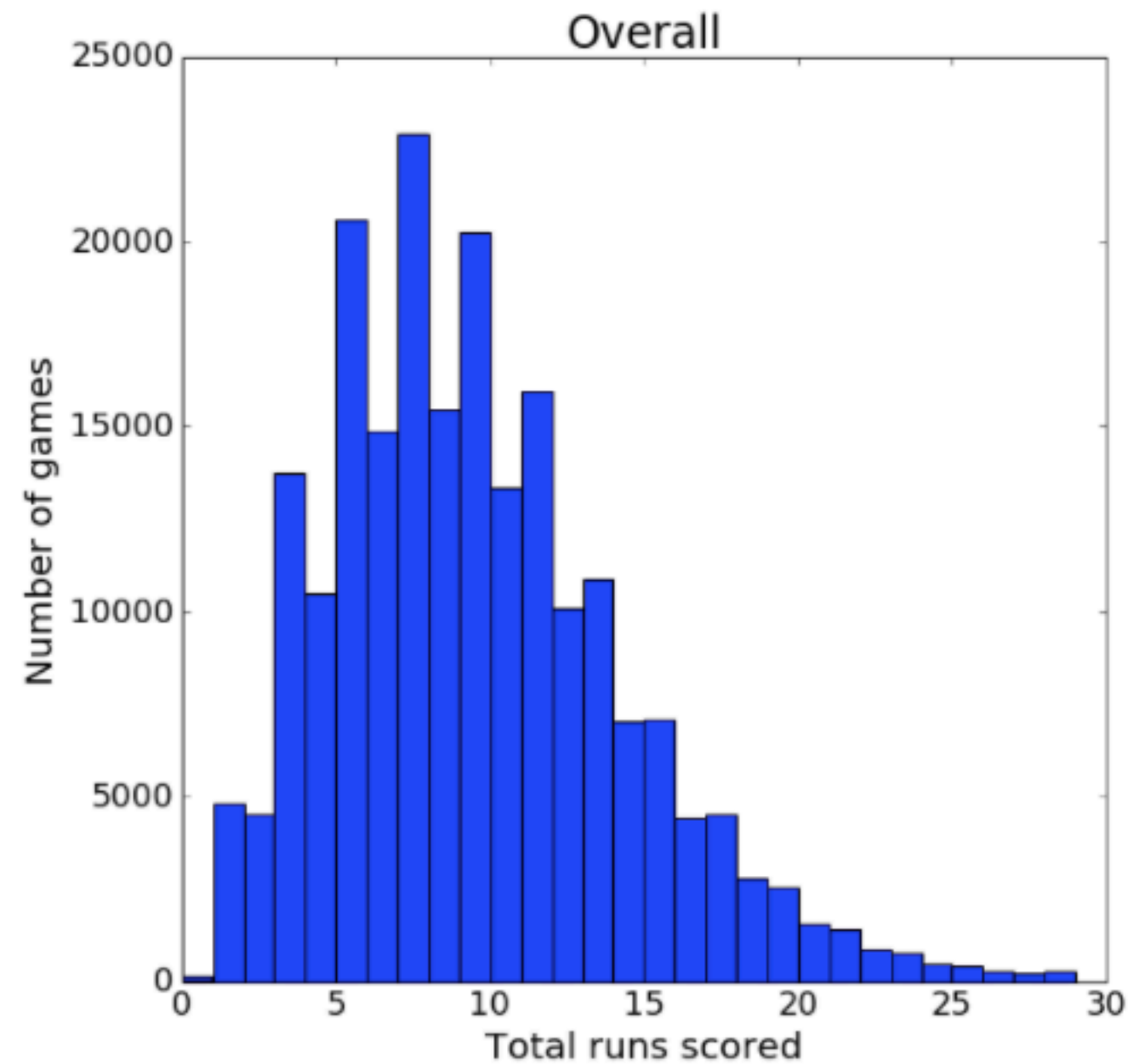
# If caught stealing, clear the base the runner started from and increment outs
if (re.match(r'CS', thisevent['play'])):
    # parse the base at which they were caught stealing
    caught_base = thisevent['play'][2]

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

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

Histogram of total runs per game



Future: Research Questions

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

A large white rectangular box with rounded corners, intended for a user to write their email address. It is positioned below the list of instructions.

What else could we study?

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?

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

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

Future: Easy data sharing

SciServer  Collaboration

Dashboard

Files

Collaboration

Resources

Groups

Create Volume

Share Volume

Refresh

Volumes

☐ jkim485_firsttest

☐ persistent

☒ raddick_baseball

☐ scratch

/raddick_baseball/1982BAL.EVA

Create Folder

Expand All

Collapse All

Directories

☒ raddick_baseball

Upload files to directory:

Choose Files

No file chosen

Upload Files

Select All

Unselect All

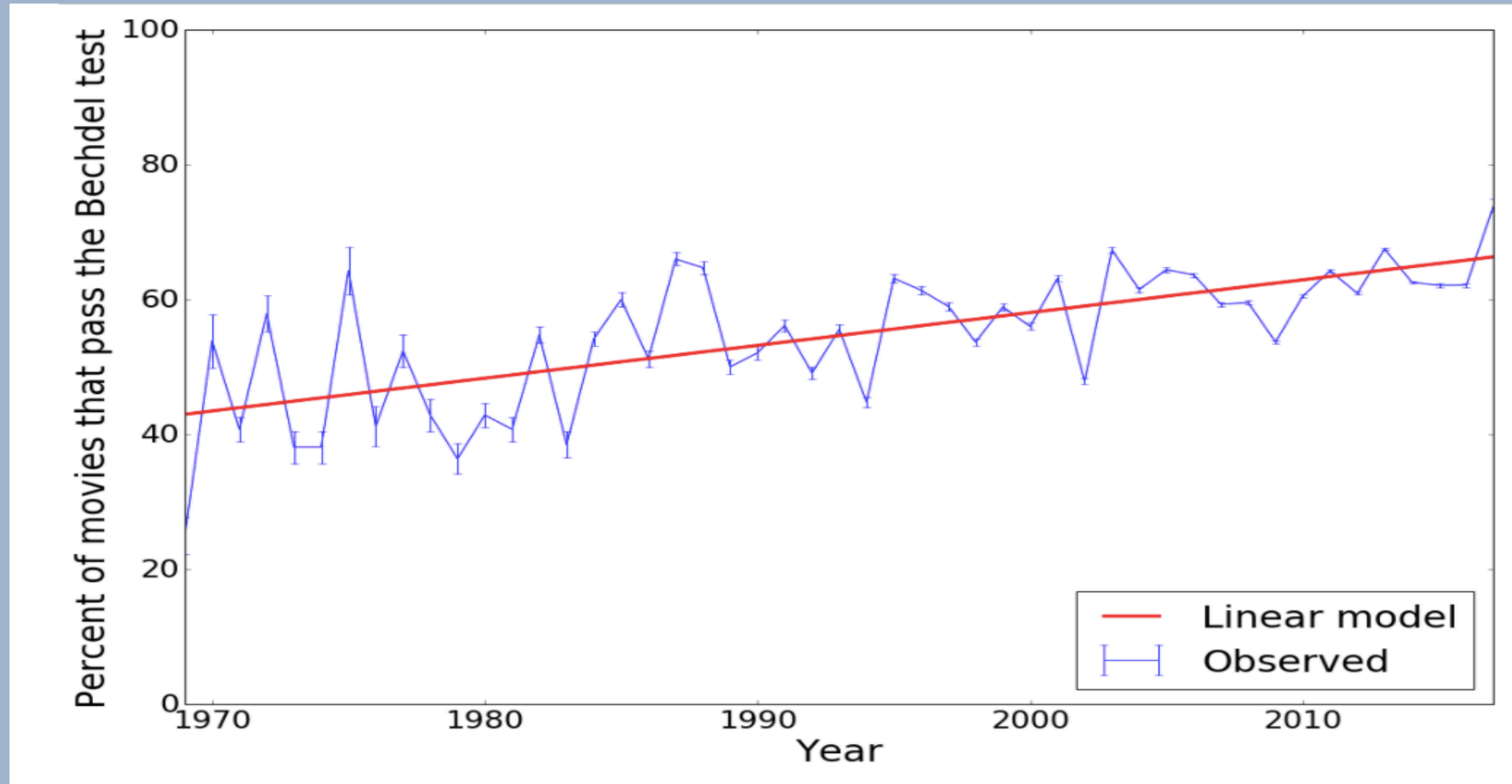
Files	Size	Modified
<input checked="" type="checkbox"/> 1982BAL.EVA	337.228 KB	2017-10-19T20:35

Download Files

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

Bechdel test results for Hollywood 1969-2017



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