Moneyball

winning in an unfair game

Data Scientist

Data Scientist

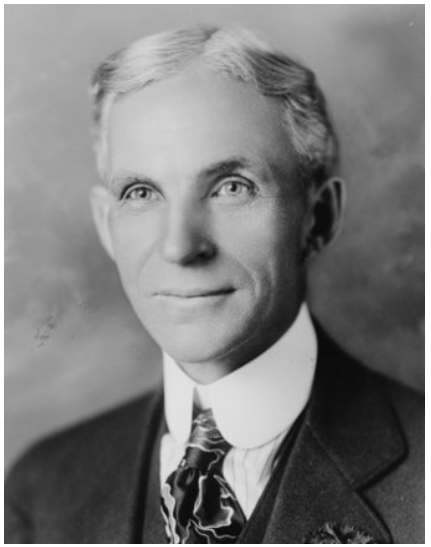
Bad Days of a "Data Scientist"

"Data scientists" don't understand **business problem**
Decision makers don't understand **data science**.

https://www.youtube.com/watch?v=BKorP55Aqvg

Understand the Latent Needs

If I had asked my customers what they wanted, they would have said a faster horse

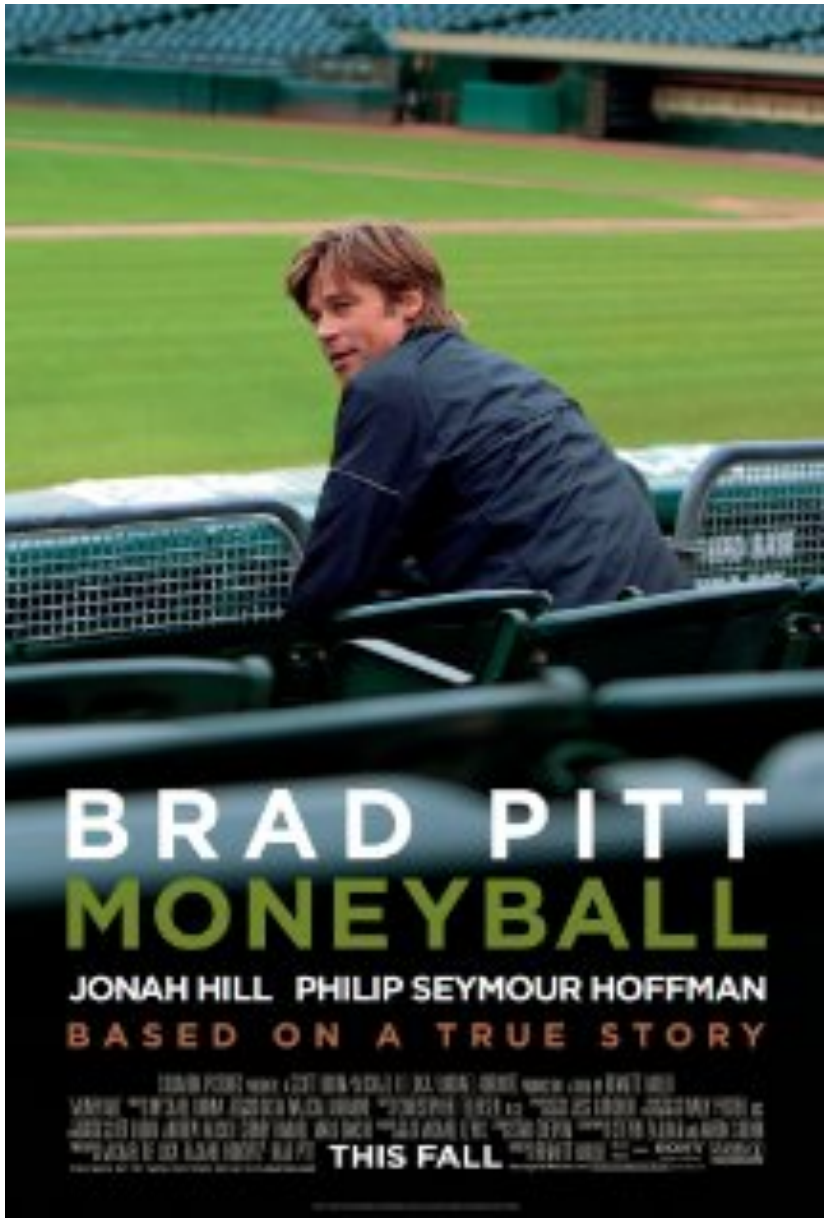


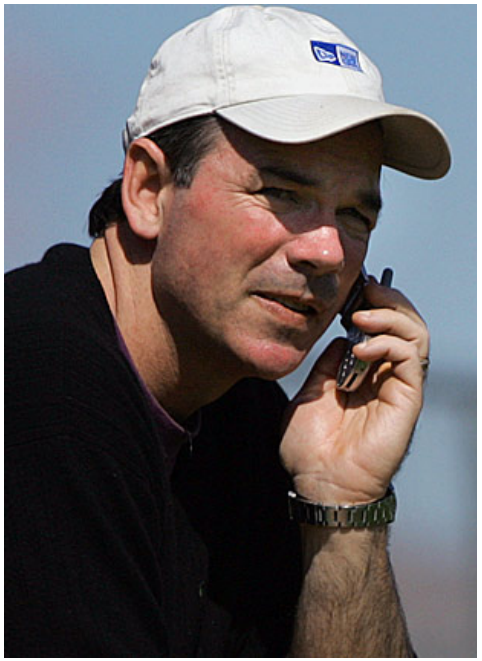
Henry Ford

We need to understand What is the real problem in the business

Understand how business works

Baseball and Billy (2001)



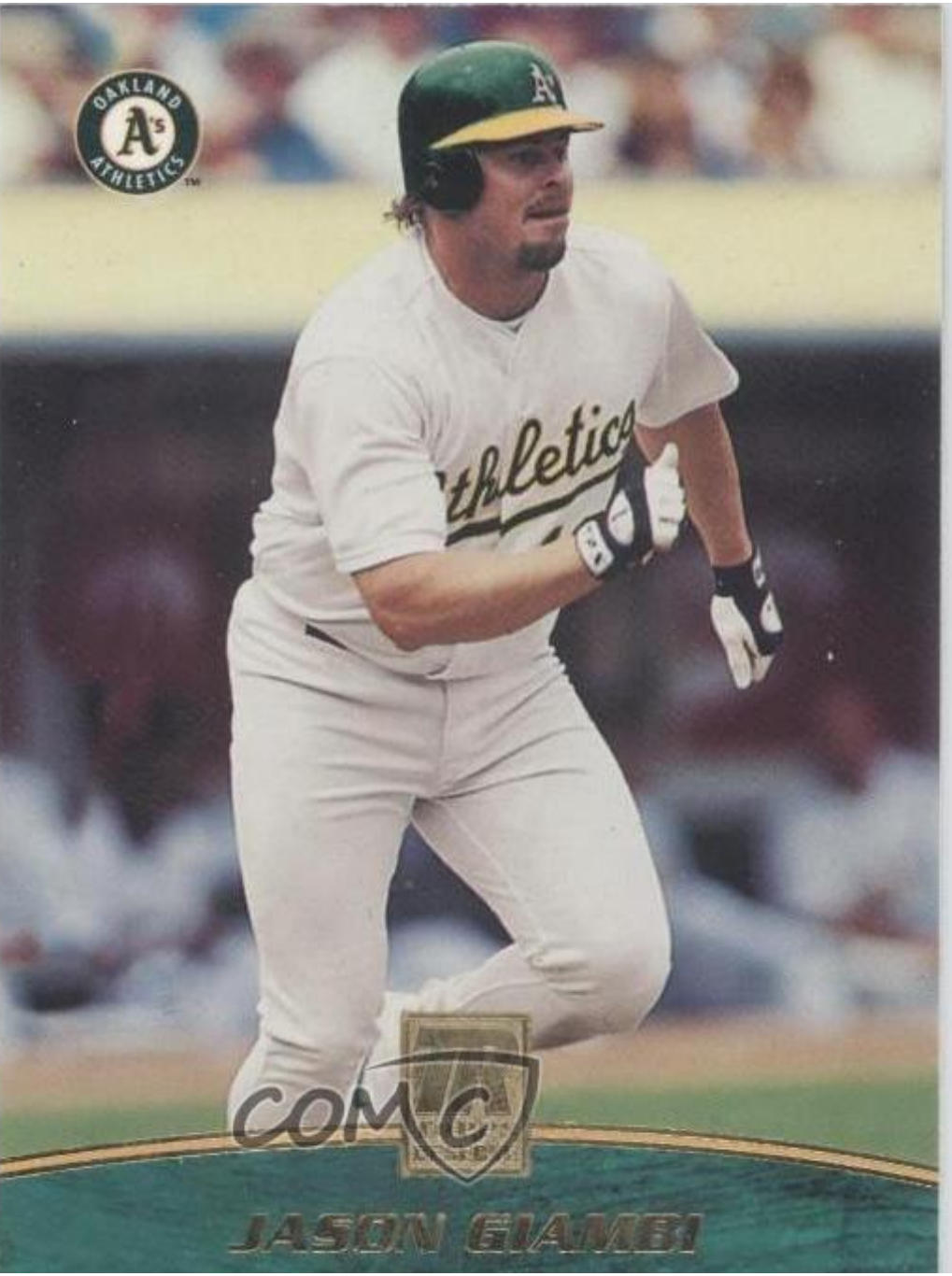


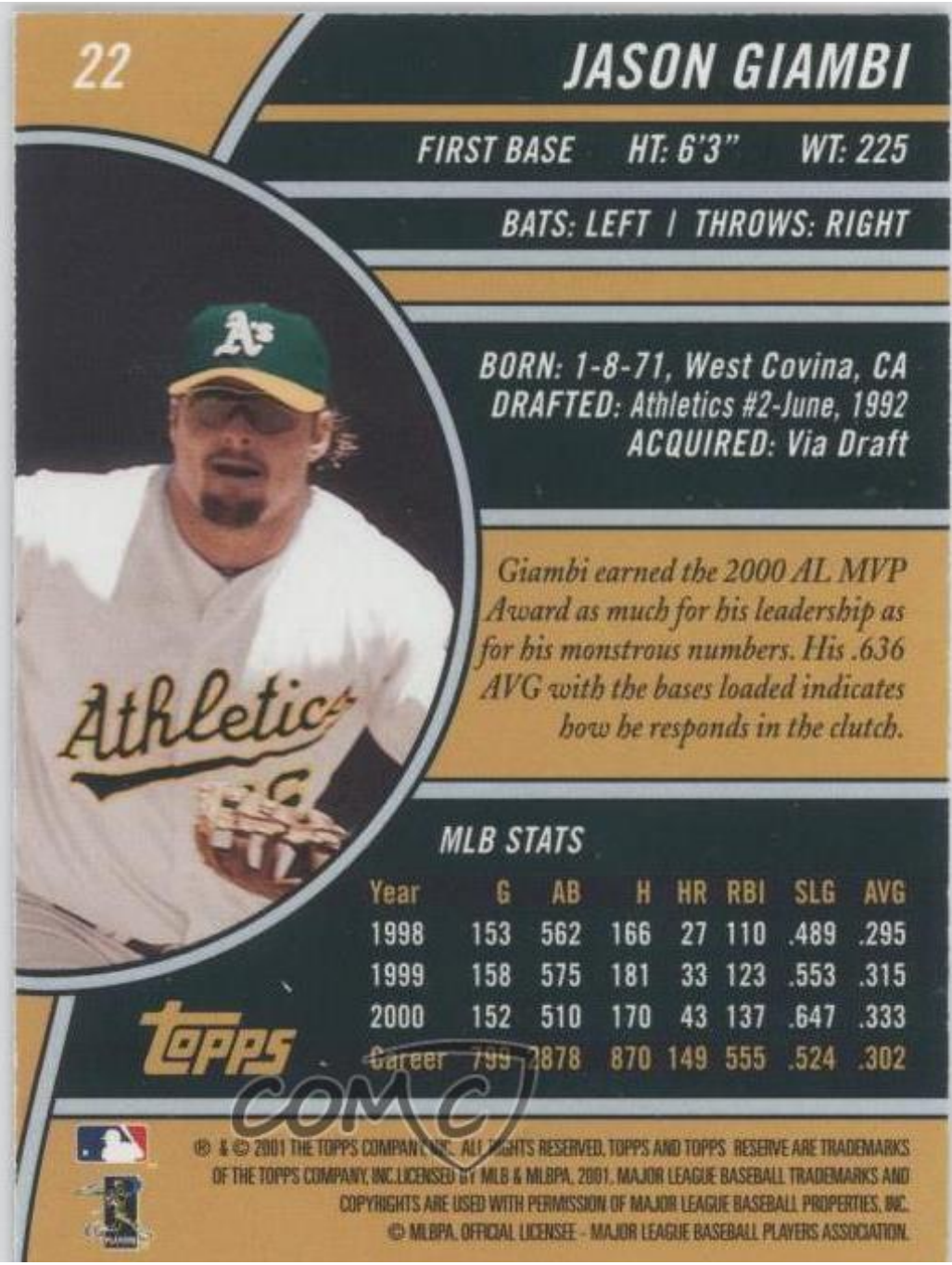
Billy Beane

OAKLAND A's ATHLETICS™

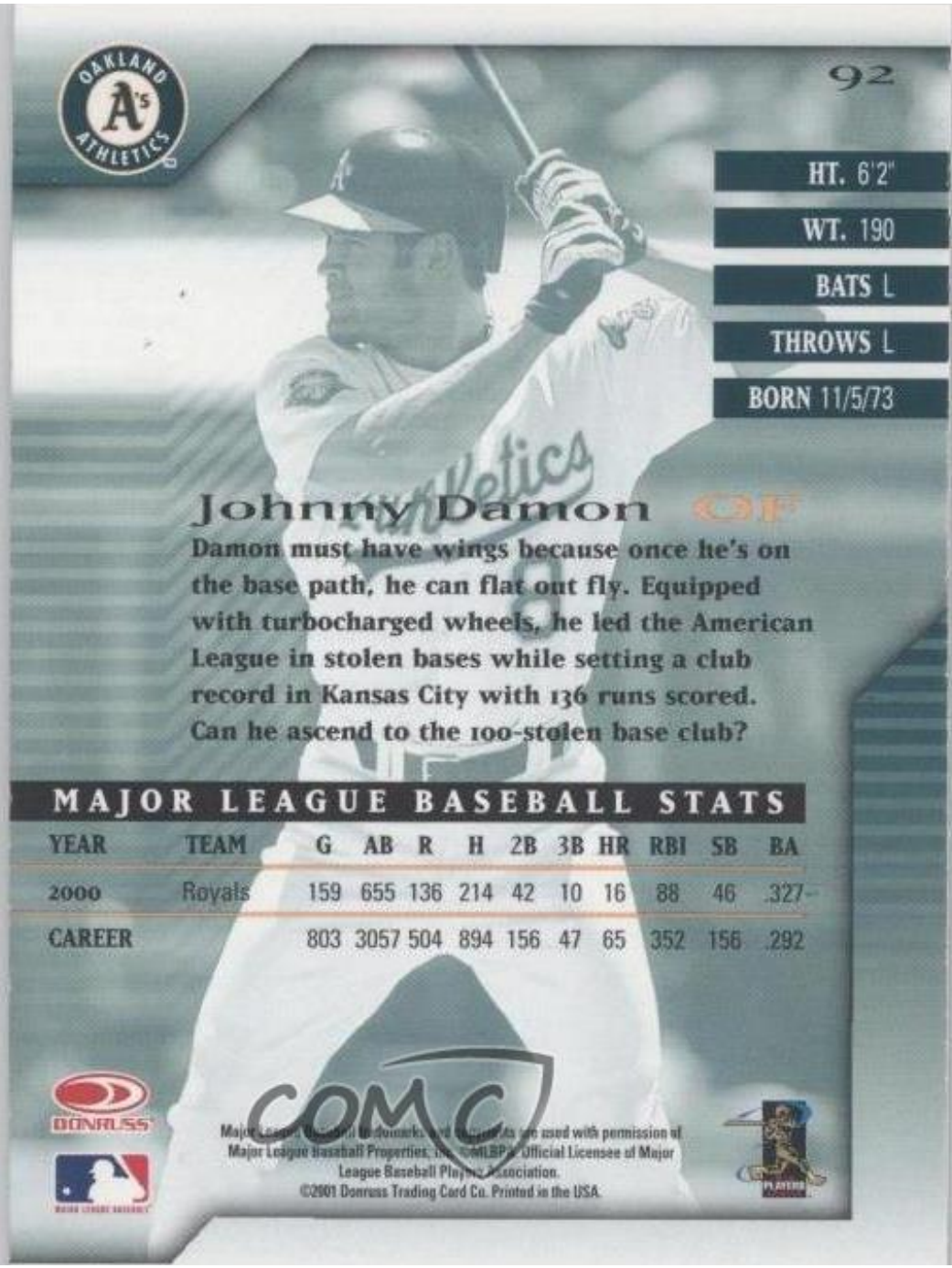
https://www.youtube.com/watch?v=j-PLdIIQLQU

Oakland A's (2001)





Oakland A's (2001)



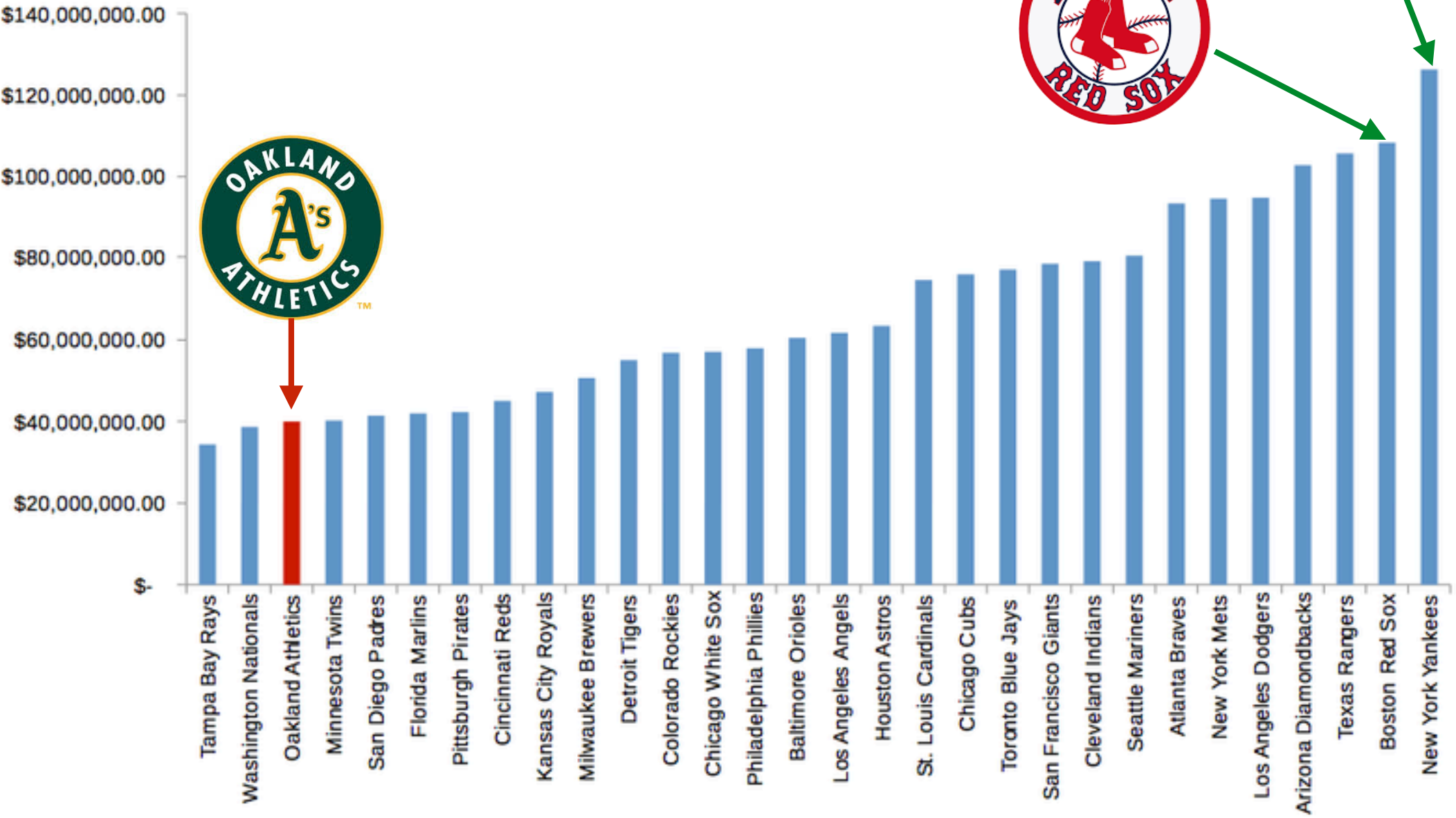
2001 (payroll)

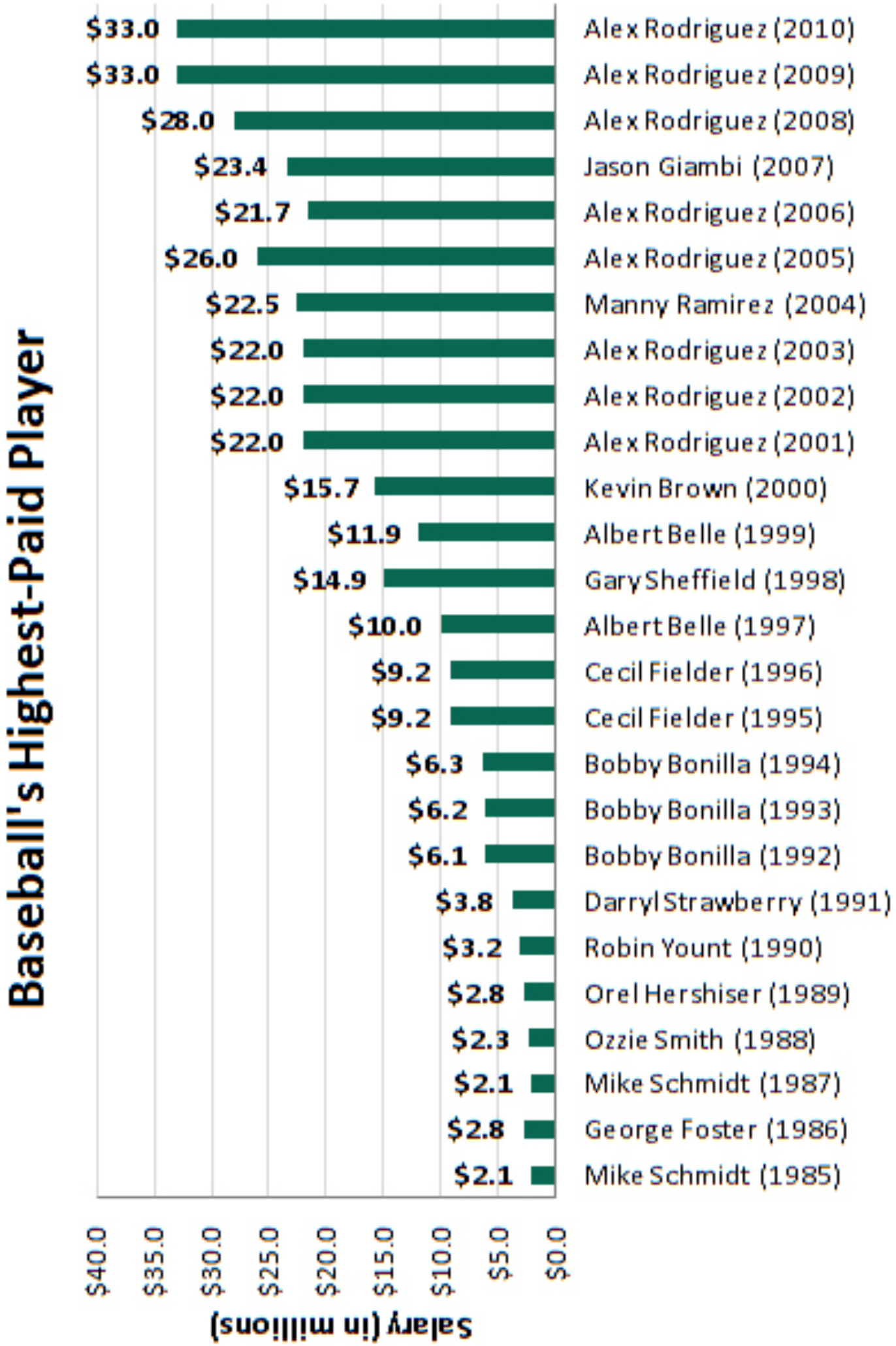
$114,457,768

vs

$ 39,722,689

Payroll in 2002



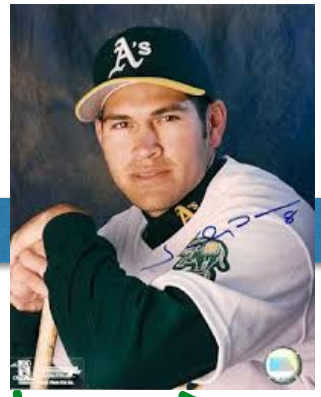

Baseball's Highest-Paid Player
$33.0 Alex Rodriguez (2010)
$33.0 Alex Rodriguez (2009)
$28.0 Alex Rodriguez (2008)
$23.4 Jason Giambi (2007)
$21.7 Alex Rodriguez (2006)
$26.0 Alex Rodriguez (2005)
$22.5 Manny Ramirez (2004)
$22.0 Alex Rodriguez (2003)
$22.0 Alex Rodriguez (2002)
$22.0 Alex Rodriguez (2001)
$15.7 Kevin Brown (2000)
$11.9 Albert Belle (1999)
$14.9 Gary Sheffield (1998)
$10.0 Albert Belle (1997)
$9.2 Cecil Fielder (1996)
$9.2 Cecil Fielder (1995)
$6.3 Bobby Bonilla (1994)
$6.2 Bobby Bonilla (1993)
$6.1 Bobby Bonilla (1992)
$3.8 Darryl Strawberry (1991)
$3.2 Robin Yount (1990)
$2.8 Orel Hershiser (1989)
$2.3 Ozzie Smith (1988)
$2.1 Mike Schmidt (1987)
$2.8 George Foster (1986)
$2.1 Mike Schmidt (1985)
$40.0
$35.0
$30.0
$25.0
$20.0
$15.0
$10.0
$5.0
$0.0
Salary (in millions)

with a small payroll (2002)

lose best players

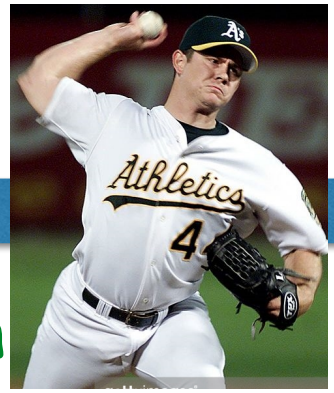
Jason Giambi

$ 10.4M



Johnny Damon

$ 7.5M



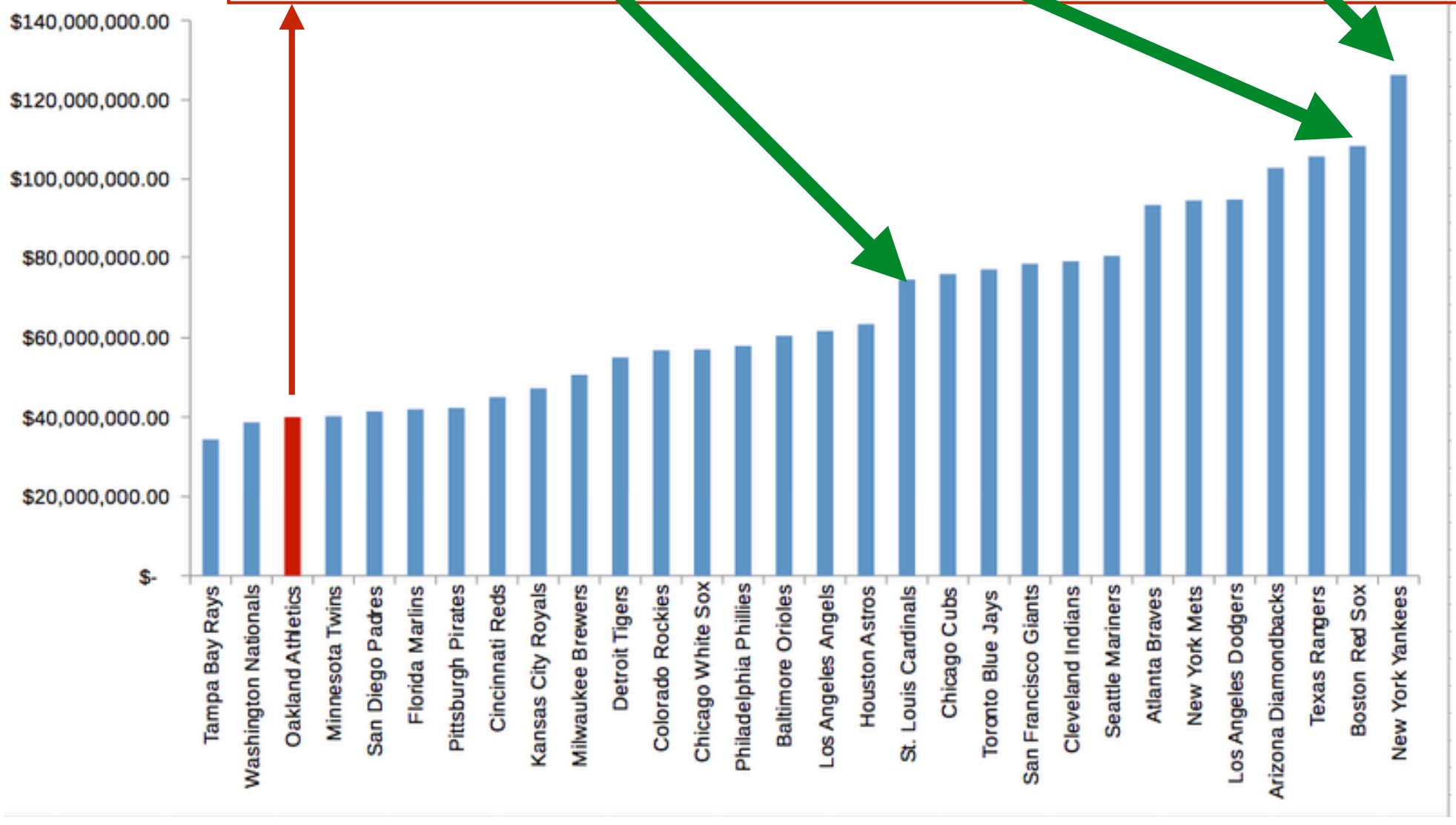
Jason Isringhausen

$ 2.8M

Jason Isringhausen

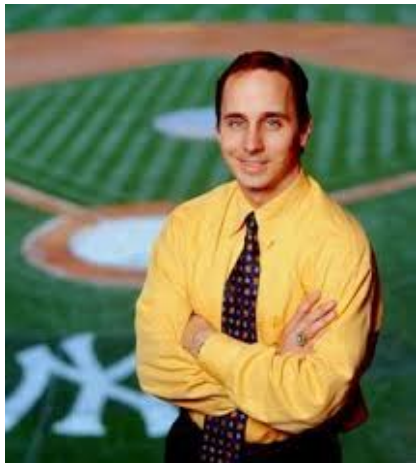
Johnny Damon

Jason Giambi



How does the baseball business work?

boss (owner)



manager (CEO)

Scout (HR)

Player (employee)

OAKLAND A's ATHLETICS ™

Team (Company)

Fans (customer)

Player Selection Problem

player selection: Scout

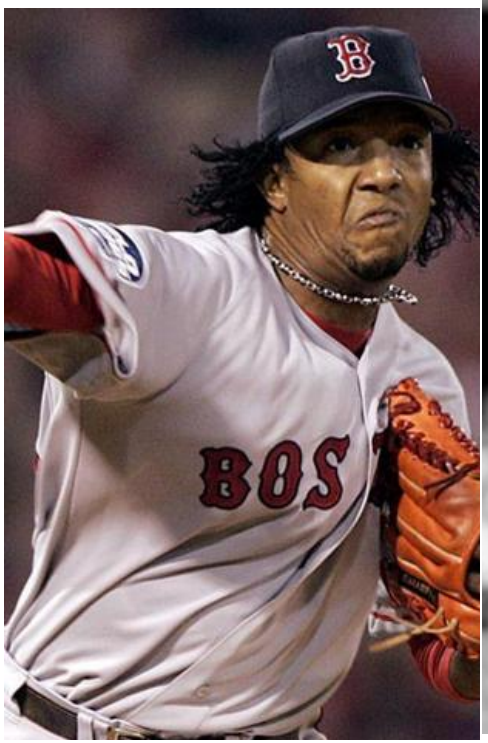
How were the players picked?

https://youtu.be/pWgyy_rImag?t=1s

How do we choose players?

style?

How do we choose players?



appearance?

How to choose players?

style?

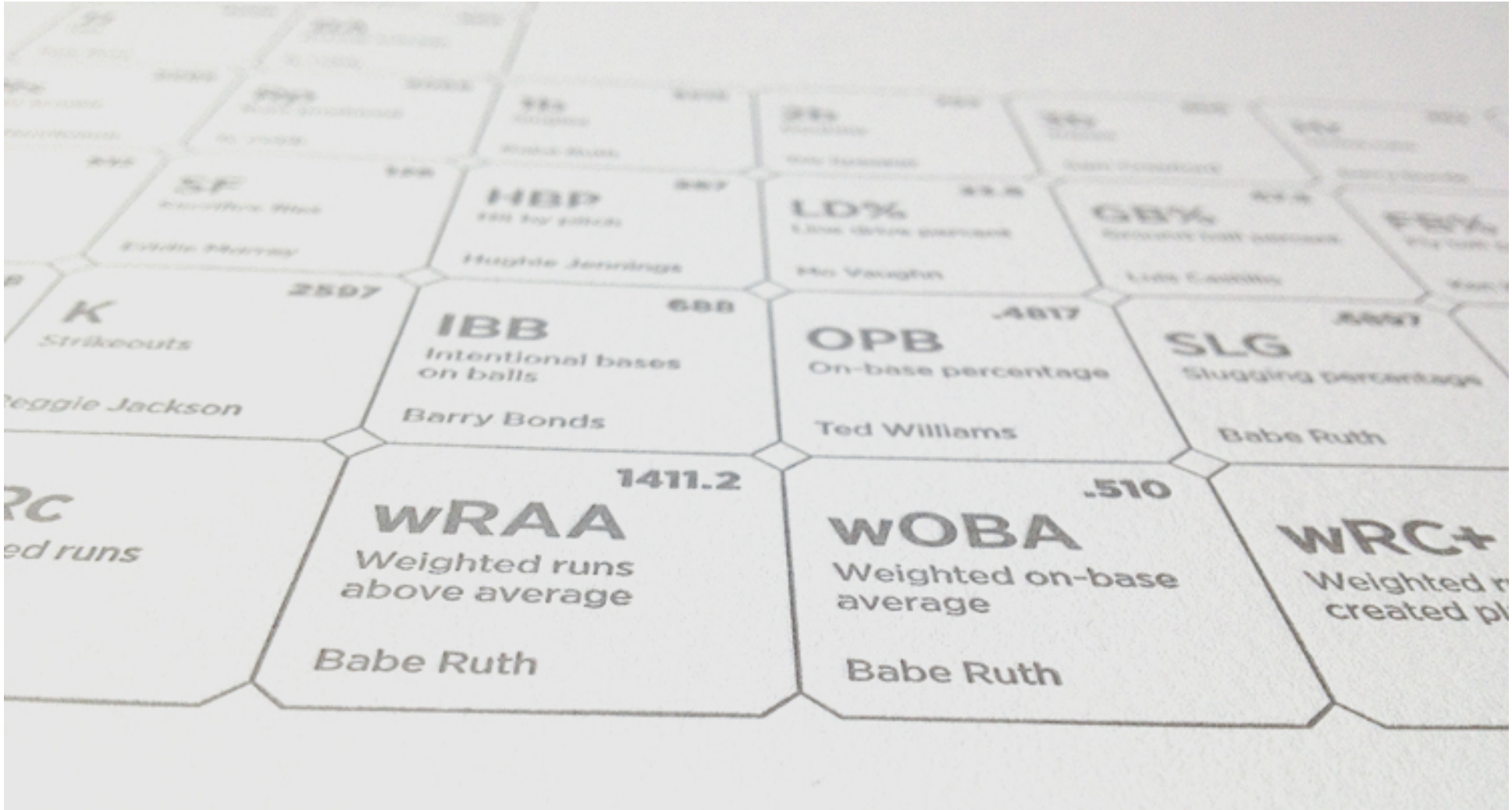
appearance?

age?

attitude?

Sabermetrics

- the empirical analysis of baseball, especially baseball statistics that measure **in-game activity**.



Data

CS	BB	SO	BA	OBP	SLG	OPS	TB	GDP	HBP	SH	SF
2	29	28	.269	.352	.347	.699	75	6	0	2	2
2	22	22	.277	.349	.361	.710	69	6	0	2	2
0	7	6	.200	.375	.240	.615	6	0	0	0	0
3	41	49	.232	.327	.300	.626	89	6	2	1	3
3	48	50	.261	.373	.396	.769	111	11	2	3	0
3	42	38	.278	.393	.432	.824	98	6	1	1	0
0	6	12	.189	.283	.245	.529	13	5	1	2	0
3	39	58	.242	.331	.386	.717	118	16	3	2	2

Hits (H)

https://en.wikipedia.org/wiki/Hit_(baseball)

when the batter safely **reaches first base** after **hitting** the ball into fair territory, without the benefit of an error or a fielder's choice.

Batting Average (BA)

https://en.wikipedia.org/wiki/Batting_average

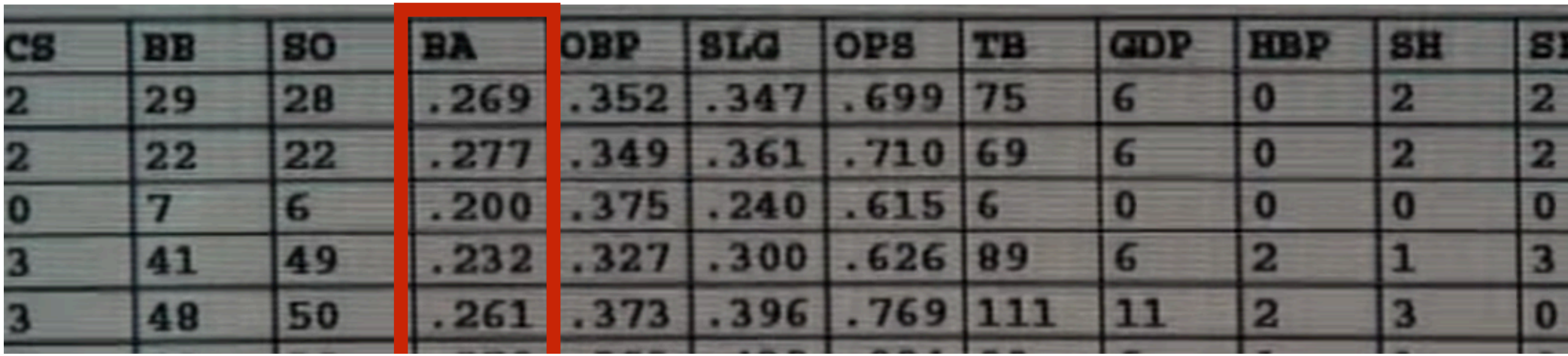
the number of hits divided by at bats

$$BA = \frac{H}{AB}$$

A player with a batting average of .300 is "batting three-hundred."

Ranking with BA

#	Player	Avg[15]	Team(s)	Year(s)
1	Ty Cobb	.366	Philadelphia (AL), Detroit (AL)	1905-1928
2	Rogers Hornsby	.358	St. Louis (NL), New York (NL), Boston (NL), Chicago (NL), St. Louis (NL), St. Louis (AL)	1915–37
3	Shoeless Joe Jackson	.356	Philadelphia (AL), Cleveland (AL), Chicago (AL)	1908-20
4	Lefty O'Doul	.349	New York (AL), Boston (AL), New York (NL), Philadelphia (NL), Brooklyn	1919–23, 1928–34
5	Ed Delahanty	.346	Philadelphia (NL), Cleveland (PL), Philadelphia (NL), Washington	1888–1903
6	Tris Speaker	.345	Boston (AL), Cleveland, Washington (AL), Philadelphia (AL)	1907–28
7	Billy Hamilton	.34442	Kansas City (AA), Philadelphia (NL), Boston (NL)	1888–1901
8	Ted Williams	.34441	Boston (AL)	1939–42, 1946–60
9	Dan Brouthers	.34212	Troy, Buffalo, Detroit (NL), Boston (NL), Boston (PL), Brooklyn (NL), Baltimore (NL), Louisville, Philadelphia (NL), New York (NL)	1879–96, 1904
10	Babe Ruth	.34206	Boston (AL), New York (AL), Boston (NL)	1914–36



CS	BB	SO	BA	OBP	SLG	OPS	TB	GDP	HBP	SH	SF
2	29	28	.269	.352	.347	.699	75	6	0	2	2
2	22	22	.277	.349	.361	.710	69	6	0	2	2
0	7	6	.200	.375	.240	.615	6	0	0	0	0
3	41	49	.232	.327	.300	.626	89	6	2	1	3
3	48	50	.261	.373	.396	.769	111	11	2	3	0

Hiring with BA

Jason Giambi

BA: .342

$ 10.4M

Johnny Damon

BA: .256

$ 7.5M

Scott Hatteberg

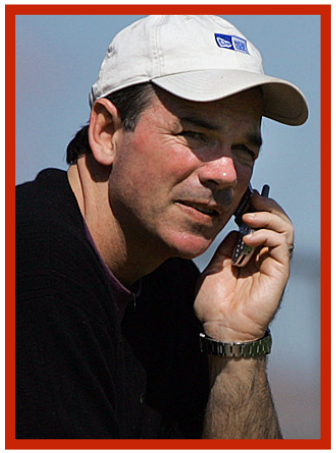
BA: .245

$ 1M

In 2001, most teams evaluate players based on BA (Batting Average)

If you were Billy, what should you do?

board



manager (Billy)

Scout (HR)

Player (employee)

Team (Company)

Fans (customer)

What's your goal?

buy star players?

buy wins?

buy "cheap" player?

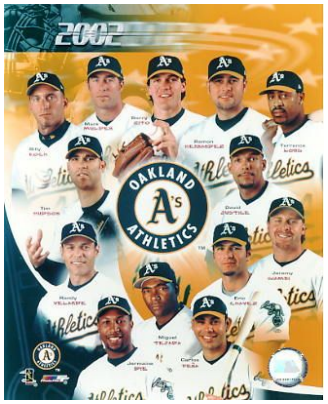
Goal: Buy Players or Buy Wins ?

https://www.youtube.com/watch?v=rMJ2IcD_fFc

https://youtu.be/TpBcwGOvO80?t=4s

Their goal in player selection

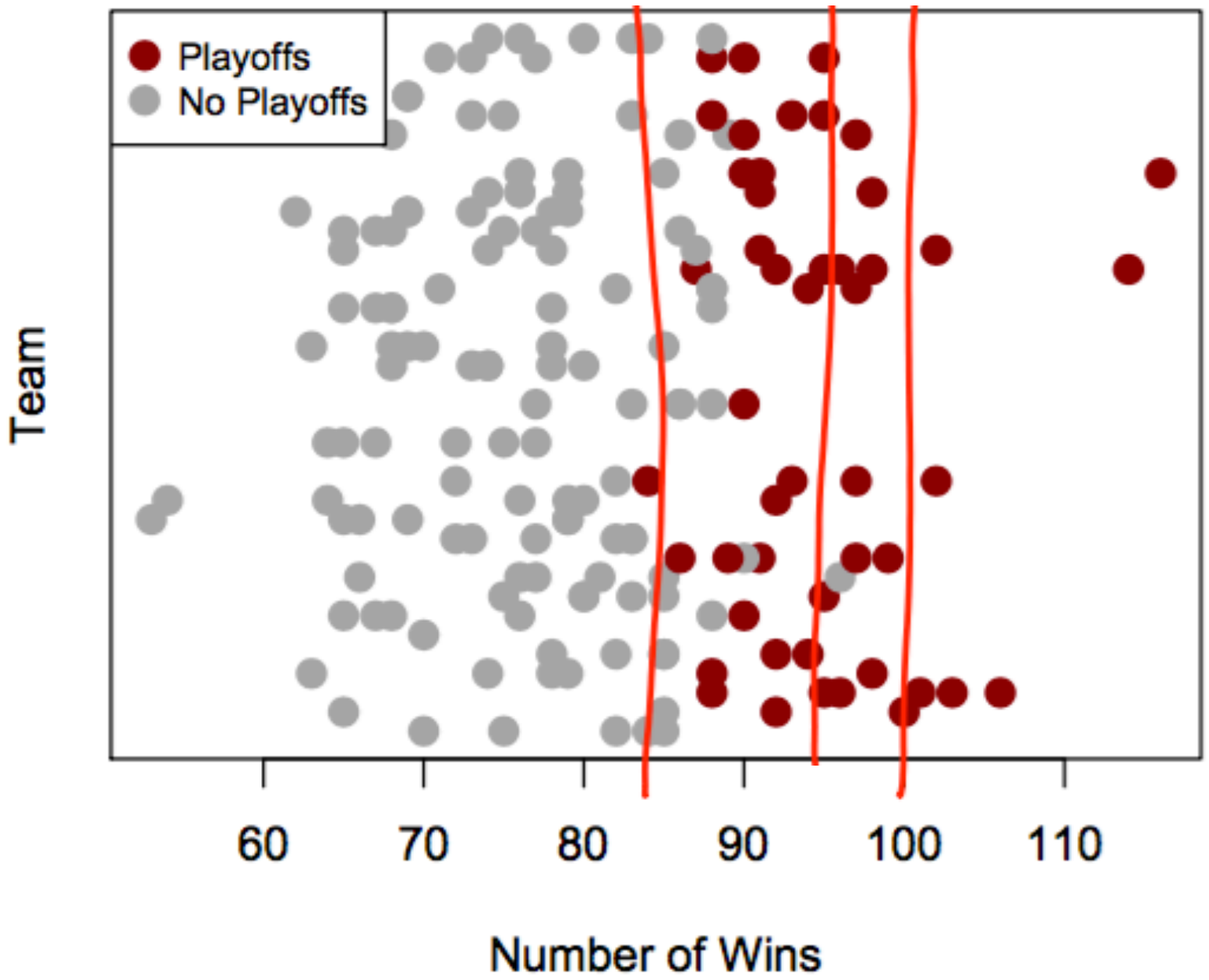
find a team



maximize # (**wins** | team)

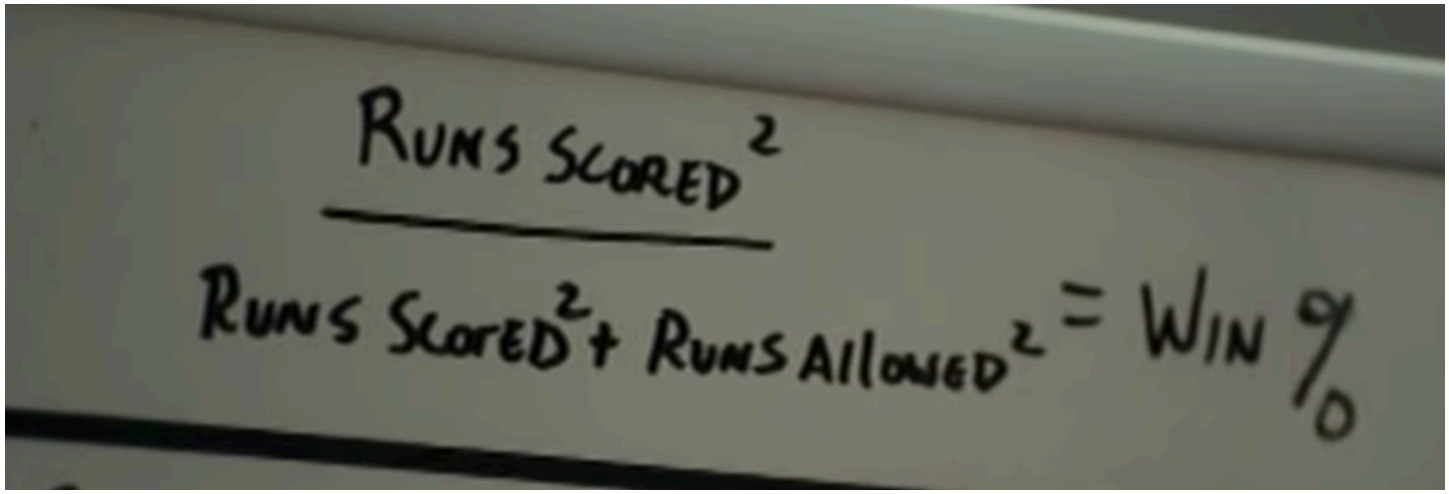
s.t. **cost** of the team < payroll budget

How many wins do we need for playoffs ?



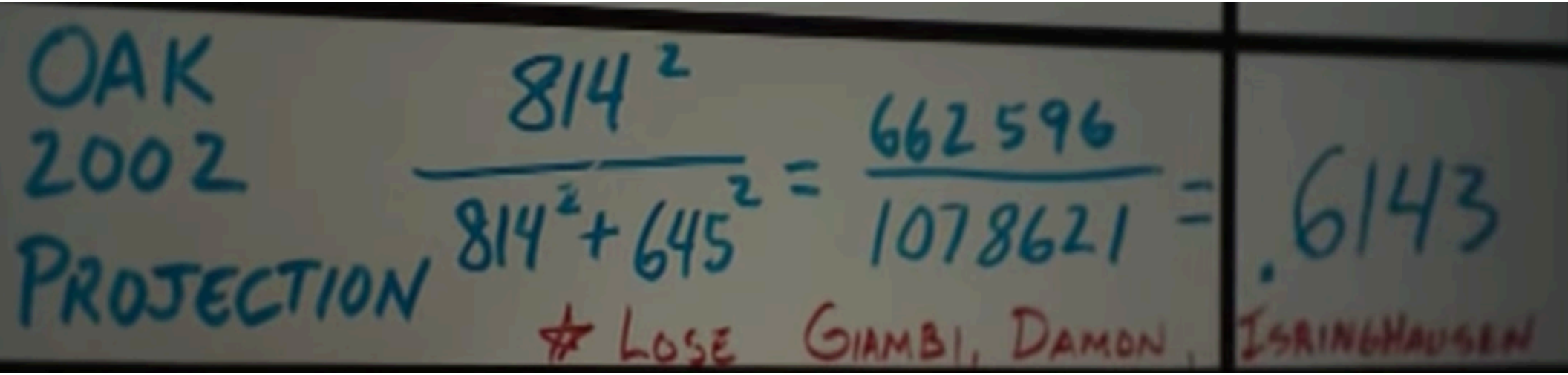
The Goal

Win **99 games**



score at least **814 runs**

allow no more than 645 runs



Runs

https://en.wikipedia.org/wiki/Run_(baseball)

when a player advances around first, second and third base and returns safely to home plate, touching the bases in that order

The object of the game is for a team to score more runs than its opponent.

Pythagorean Expectation

https://en.wikipedia.org/wiki/Pythagorean_expectation

$$\frac{\text{Runs Scored}^2}{\text{Runs Scored}^2 + \text{Runs Allowed}^2} = \text{Win \%}$$

OAK 2001	$\frac{884^2}{884^2+645^2}$	=	$\frac{781456}{1197481}$

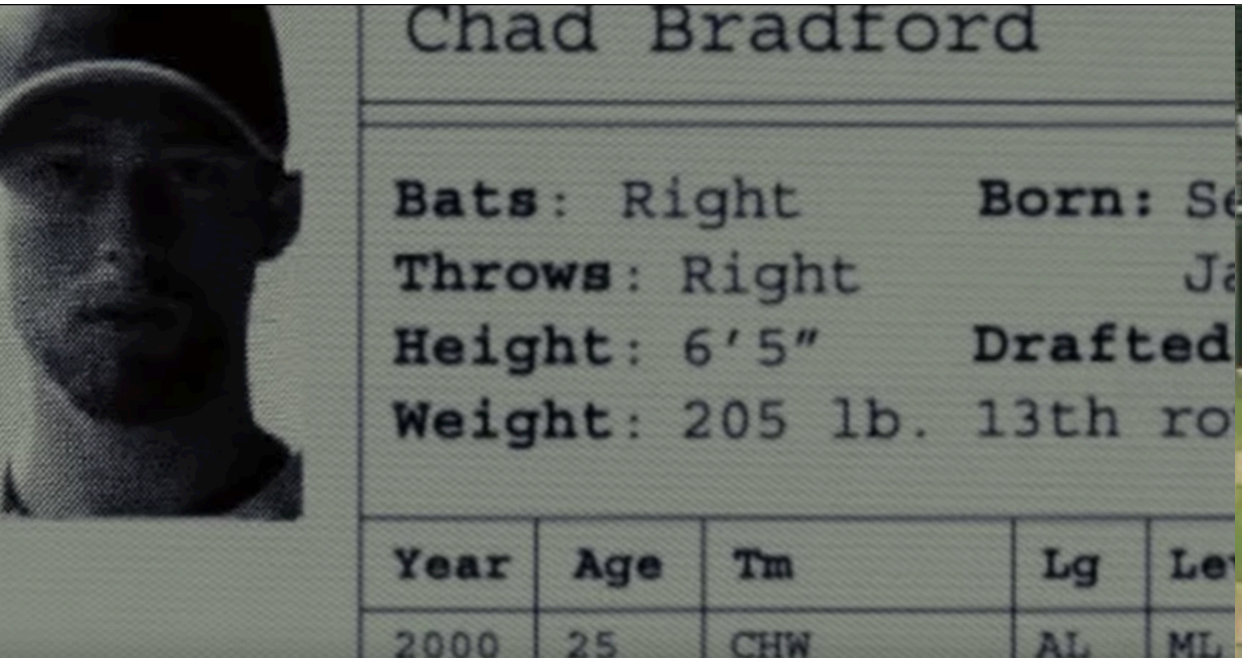
Bill James

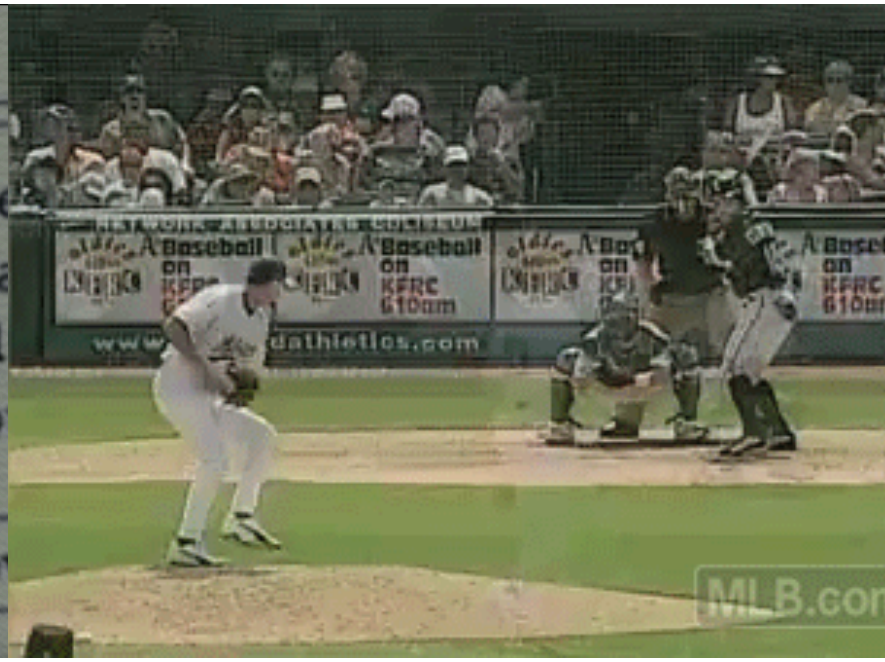
How to reach the goal?

https://youtu.be/KWPhV6PUr9o?t=1m23s

Chad Bradford (Pitcher)

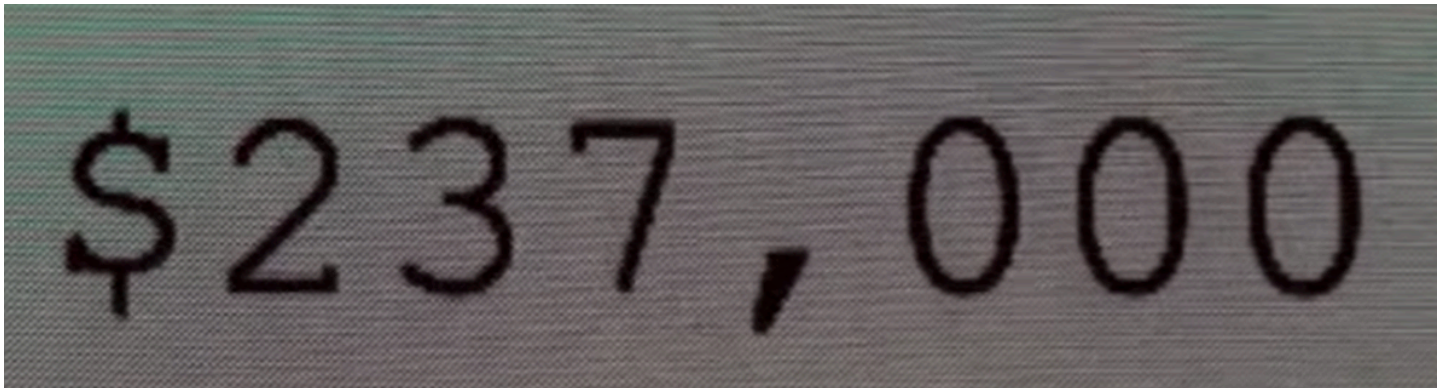
https://en.wikipedia.org/wiki/Chad_Bradford





His **ERA** stayed around **3.00** for his entire career until 2004

(Earned Run Average) https://en.wikipedia.org/wiki/Earned_run_average



How do we compute runs?

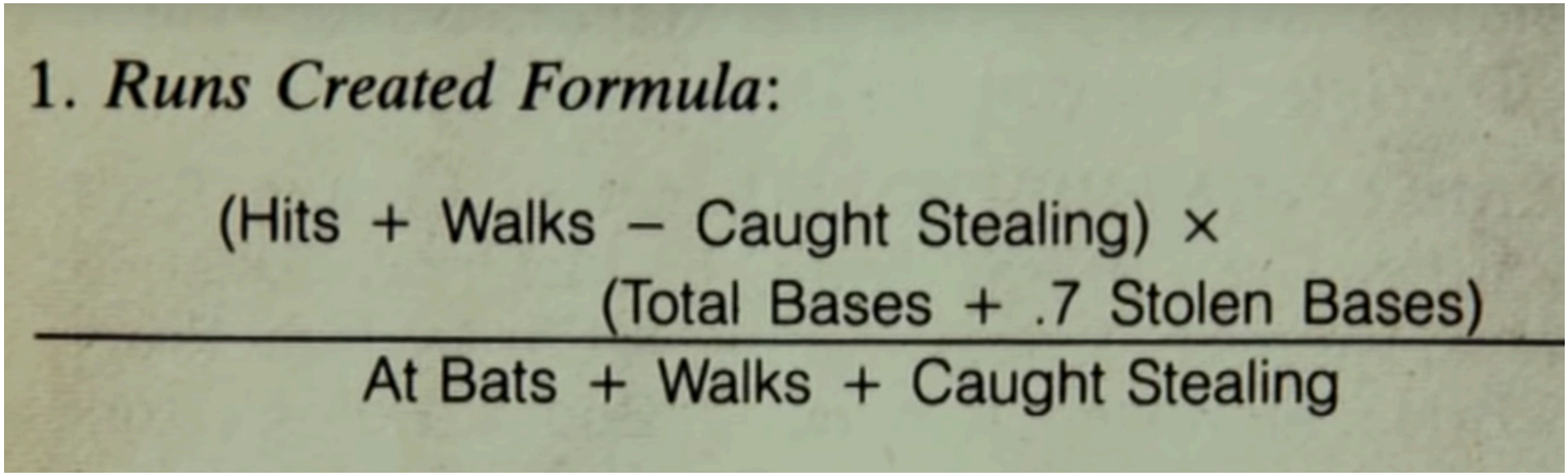
given a team of players, how to estimate the number of runs?

Runs Created formula (Bill James)

https://en.wikipedia.org/wiki/Runs_created

$$\mathrm{RC} = \frac{(\text{hits} + \text{walks}) \times (\text{total bases})}{(\text{at-bats}) + (\text{walks})}$$

"Stolen base" version of runs created



1. *Runs Created Formula*:

$$\frac{(\text{Hits} + \text{Walks} - \text{Caught Stealing}) \times (\text{Total Bases} + .7\ \text{Stolen Bases})}{\text{At Bats} + \text{Walks} + \text{Caught Stealing}}$$

SLOB × At-Bats

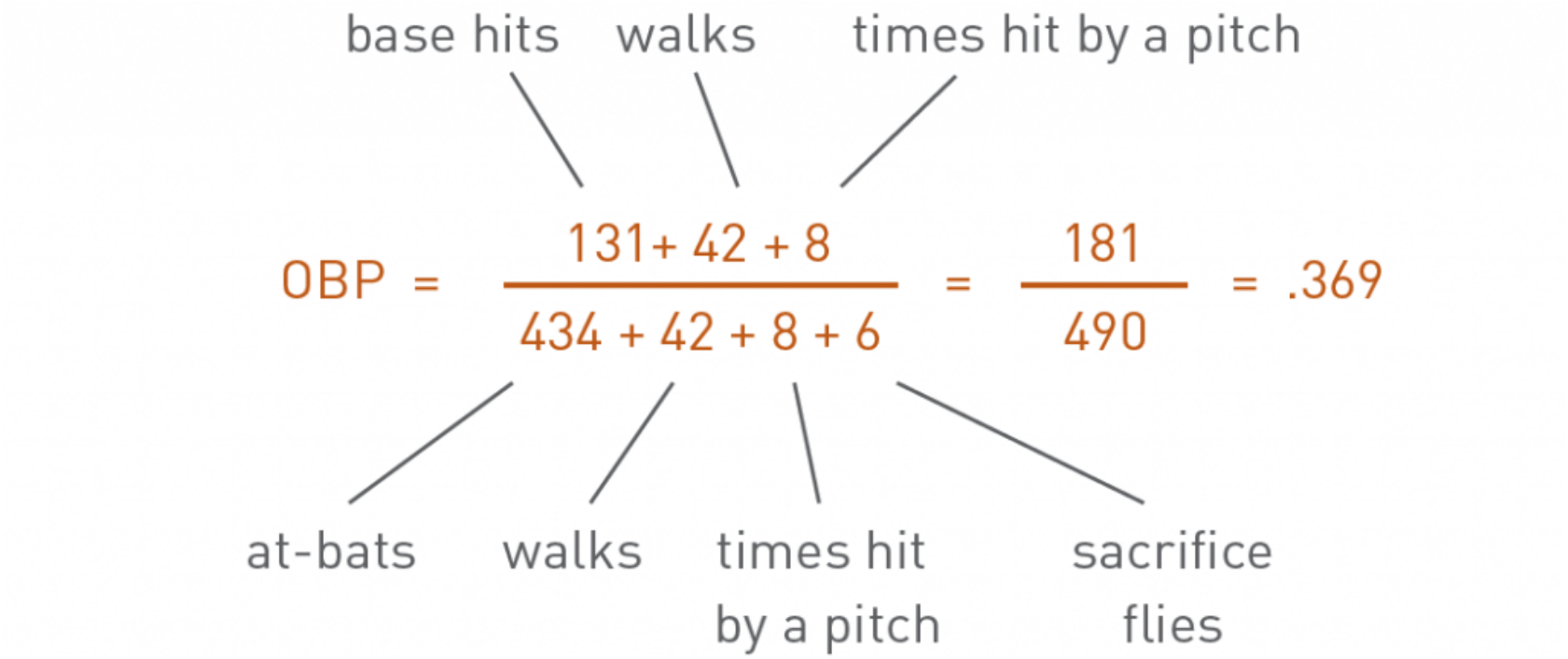
"SLOB" (Slugging × On-Base)

On Base Percentage (OBP)

https://en.wikipedia.org/wiki/On-base_percentage

how frequently a batter reaches base

$$OBP = \frac{H + BB + HBP}{AB + BB + HBP + SF}$$



Oakland mainly used this metric (instead of BA) to evaluate players

Base on Balls (Walks) (BB)

https://en.wikipedia.org/wiki/Base_on_balls

a batter receives **four pitches** that the umpire calls balls, and is in turn **awarded first base** without the possibility of being called out

Hit By Pitch (HBP)

https://en.wikipedia.org/wiki/Hit_by_pitch

a batter or his clothing or equipment (other than his bat) is **struck** directly by a pitch from the pitcher; the batter is called a hit batsman (HB). A hit batsman is **awarded first base**

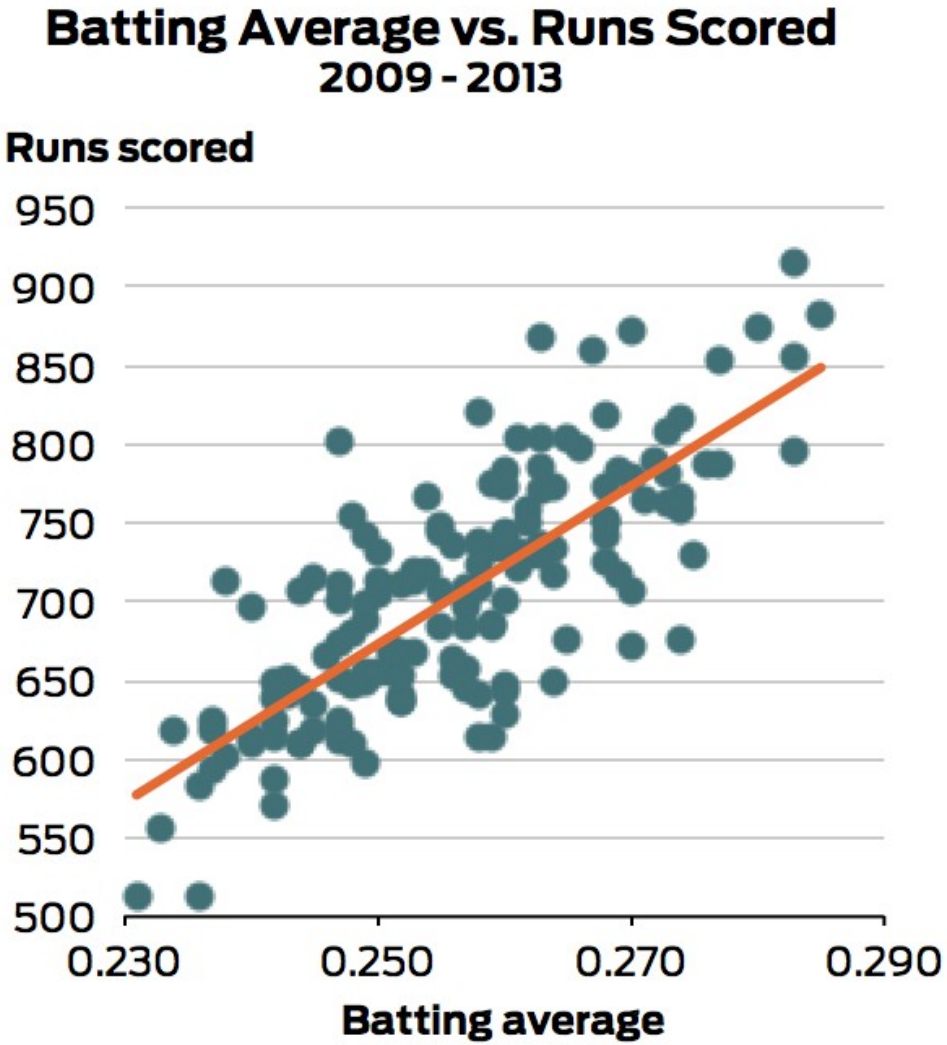
Sacrifice Fly (SF)

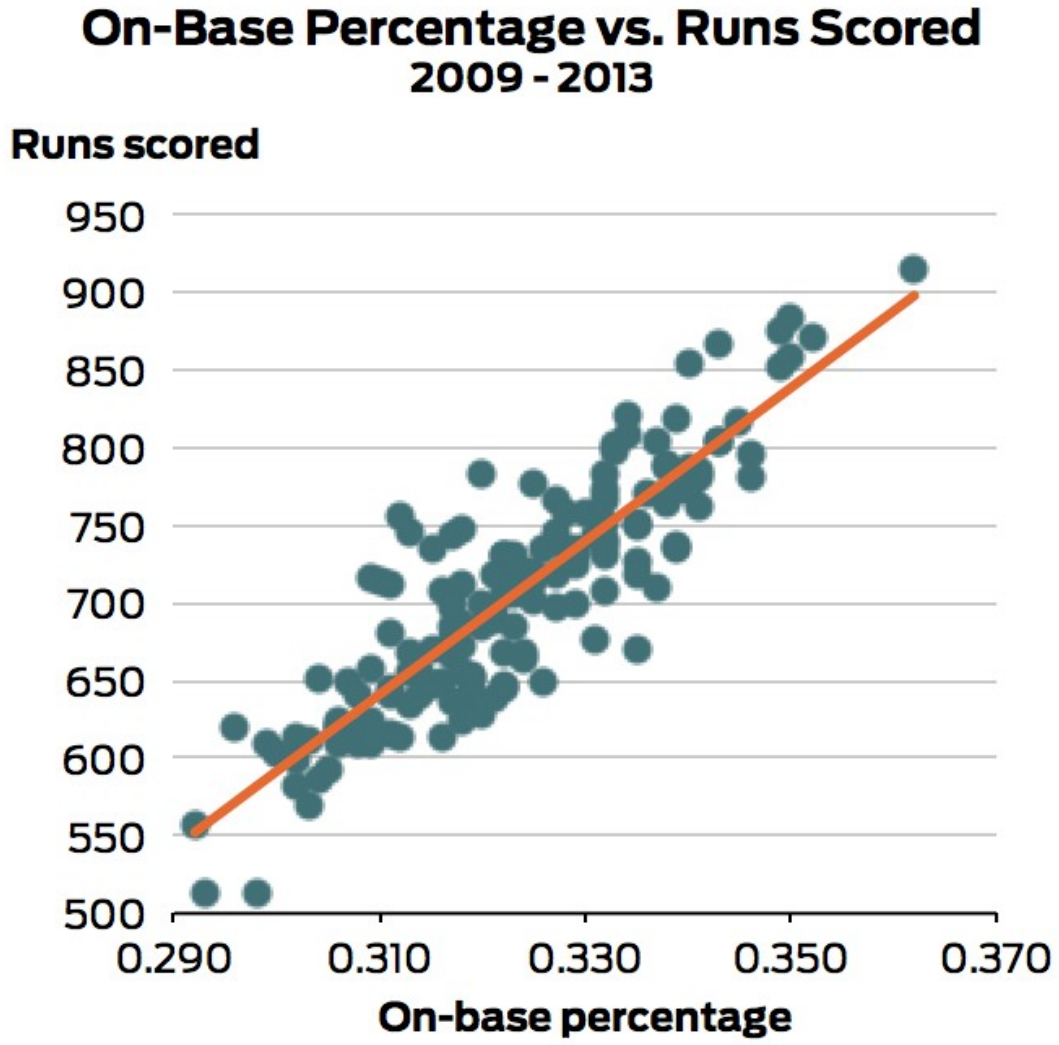
https://en.wikipedia.org/wiki/Sacrifice_fly

https://www.youtube.com/watch?v=SyJnHBMfMqE

Score a sacrifice fly when, before two are out, the batter hits a ball in flight handled by an outfielder or an infielder running in the outfield in fair or foul territory that

1. is caught, and a run scores after the catch, or
2. is dropped, and a runner scores, if in the scorer's judgment the runner could have scored after the catch had the fly ball been caught.





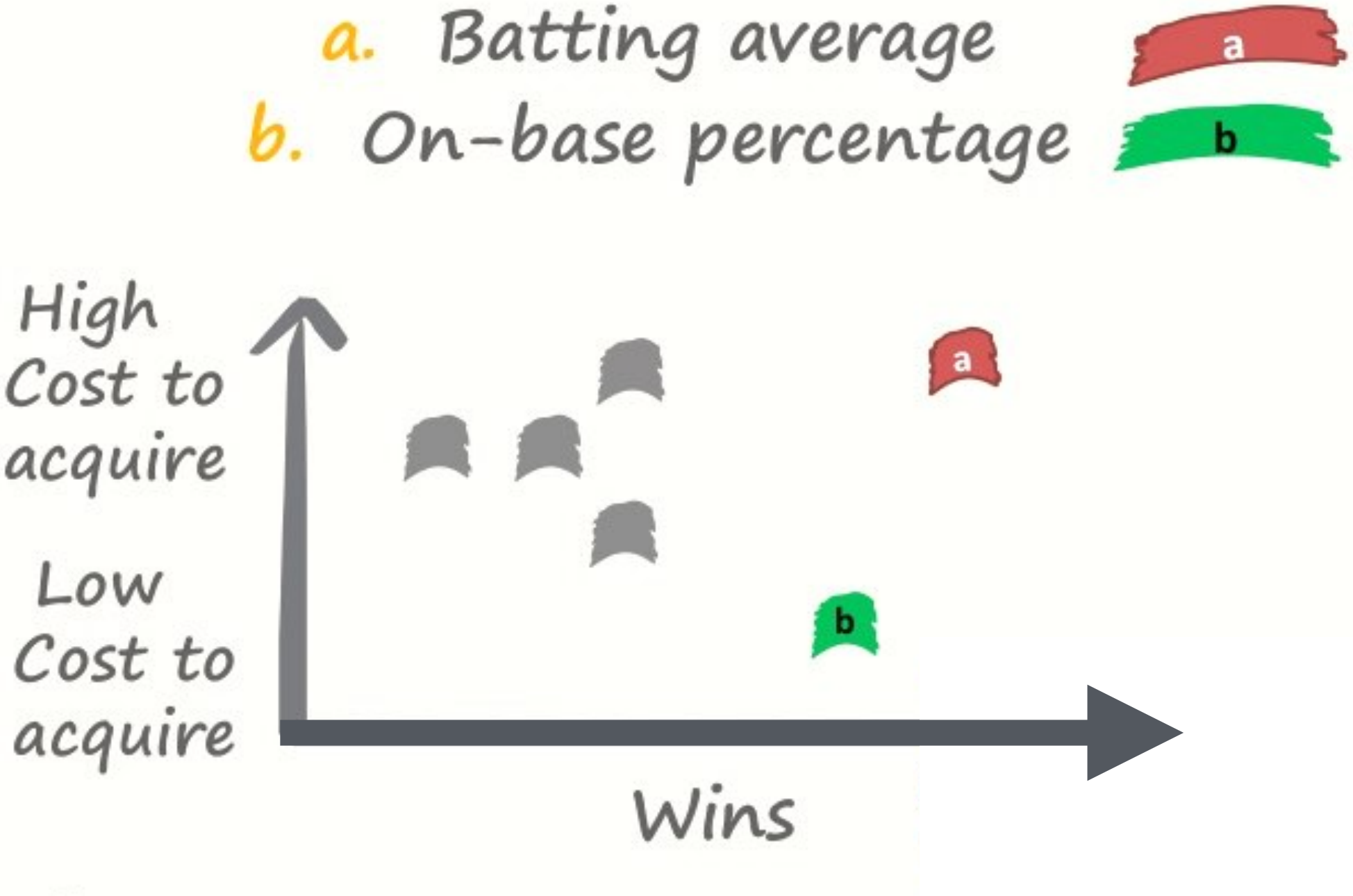
$$BA = \frac{H}{AB}$$

$$OBP = \frac{H + BB + HBP}{AB + BB + HBP + SF}$$

OBP is as effective as BA, but not used by other teams.

a. Batting average

b. On-base percentage



So it is possible to find cheap players using OBP because no other team uses this metric (some teams only use OBP as a tie-breaker)

Slugging Percentage (SLG)

https://en.wikipedia.org/wiki/Slugging_percentage

a measure of the batting productivity of a hitter

total bases divided by at bats

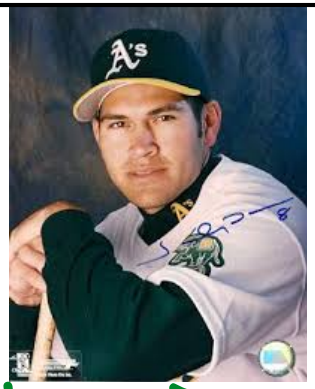
$$SLG = \frac{1B + 2 \times 2B + 3 \times 3B + 4 \times HR}{AB}$$

or

$$SLG = \frac{H + 2B + 2 \times 3B + 3 \times HR}{AB}$$

HR: home runs

Oakland also used this metric to evaluate players, in addition to OBP

players Oakland chose		players Oakland lost	
 David Justice	BA: .241 OBP: .333 SLG: .430 $ 3.5M	BA: .342 OBP: .477 SLG: .660 $ 10.4M	 Jason Giambi
 Scott Hatteberg	BA: .245 OBP: .332 SLG: .345 $ 1M	BA: .256 OBP: .324 SLG: .363 $ 7.5M	 Johnny Damon
 Jeremy Giambi	BA: .283 OBP: .391 SLG: .450 $ 0.9M	no data $ 2.8M	 Jason Isringhausen

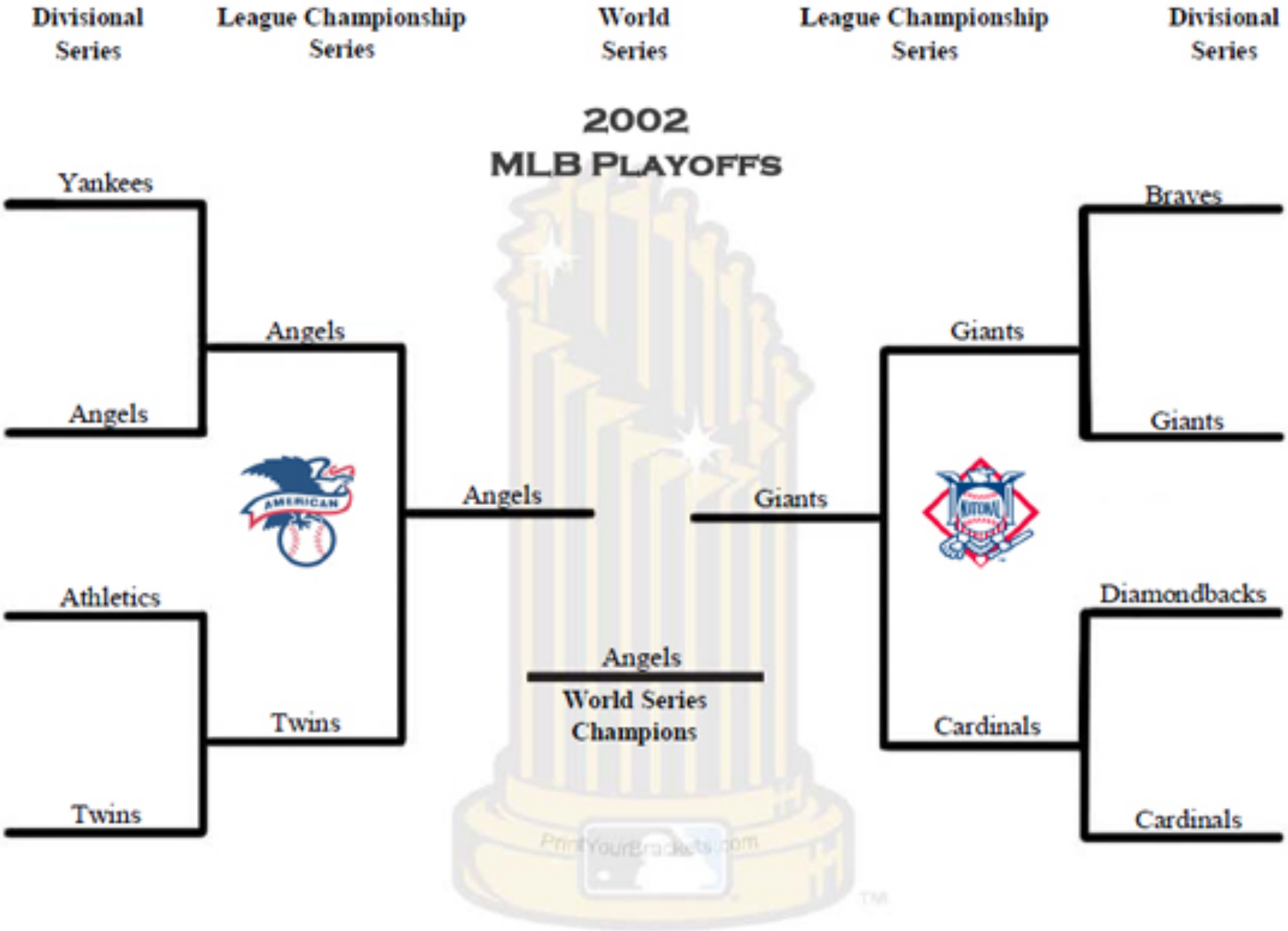
David Justice (Outfielder)

Career Statistics

runs: 305
walks (BB): 903
hits (H): 1571
at base (AB): 5625

BA: .279
OBP: .378
SLG: .500

https://www.baseball-reference.com/players/j/justida01.shtml


Divisional Series
League Championship Series
World Series
League Championship Series
Divisional Series
2002
MLB PLAYOFFS
Yankees
Angels
Angels
Athletics
Twins
Twins
Angels
Angels
Giants
Giants
Braves
Giants
Diamondbacks
Cardinals
Cardinals
Angels
World Series Champions

Both teams got 103 wins

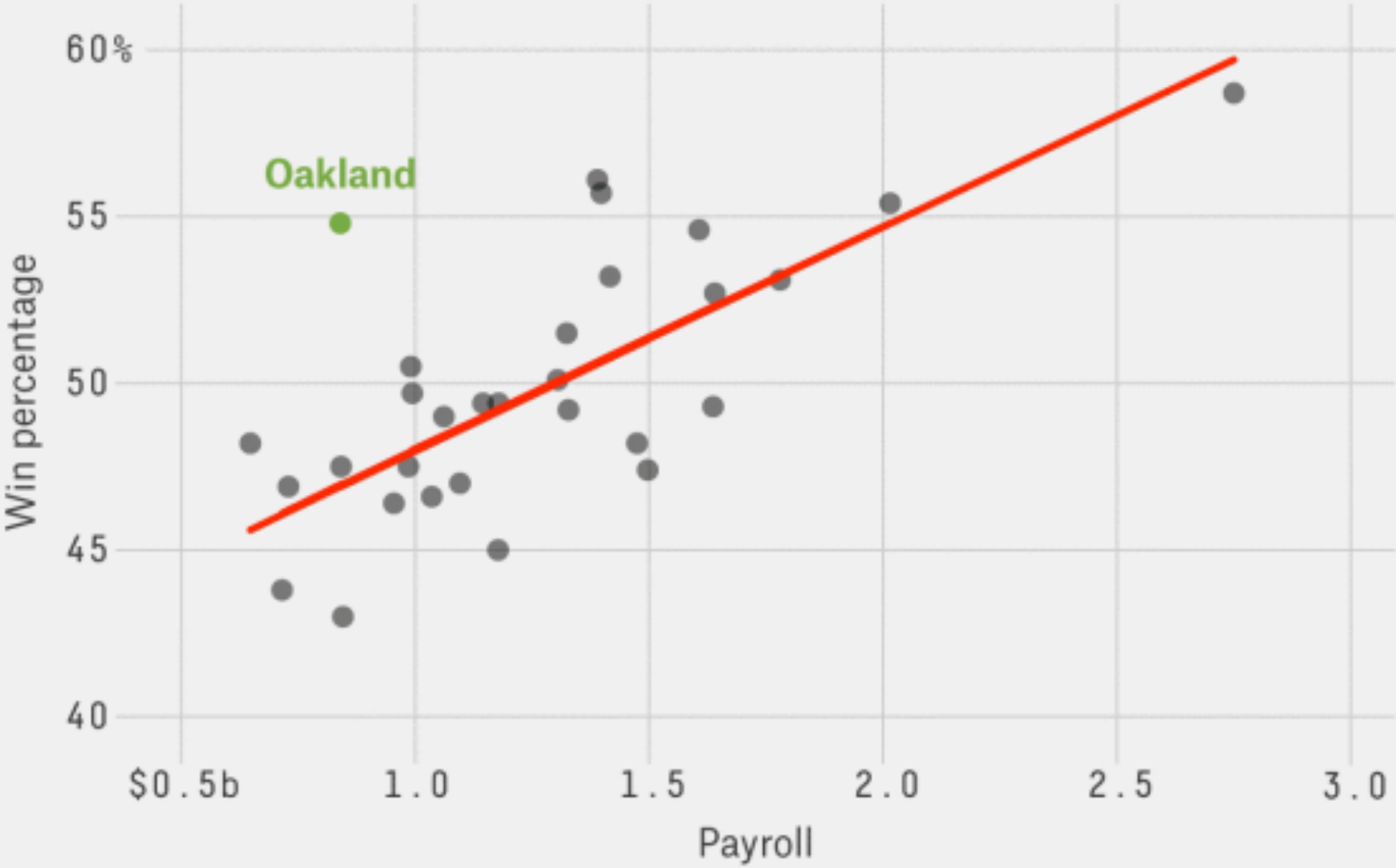
Cost per Win (2002)

$1.4M

$0.26M

MLB Teams' Overall Win Percentage vs. Payroll

Since 2000



20 Streak in 2002

116	August 9	@ Yankees	3–2 (16)	**Bowie** (2–0)	Hitchcock (1–1)	—	54,316	67–49
117	August 10	@ Yankees	8–0	**Lidle** (5–9)	Wells (12–6)	—	54,439	68–49
118	August 11	@ Yankees	5–8	Mussina (14–6)	**Mulder** (13–7)	—	54,703	68–50
119	August 12	Blue Jays	1–2	Loaiza (5–6)	**Harang** (4–3)	Escobar (24)	14,178	68–51
120	August 13	Blue Jays	5–4	**Zito** (16–5)	Carpenter (4–5)	**Koch** (29)	17,466	69–51
121	August 14	Blue Jays	4–2	**Hudson** (9–9)	Walker (5–3)	**Koch** (30)	40,528	70–51
122	August 16	White Sox	1–0	**Lidle** (6–9)	Buehrle (15–9)	**Koch** (31)	22,622	71–51
123	August 17	White Sox	9–2	**Mulder** (14–7)	Garland (8–9)	—	40,658	72–51
124	August 18	White Sox	7–4	**Zito** (17–5)	Wright (8–11)	—	31,489	73–51
125	August 19	@ Indians	8–1	**Hudson** (10–9)	Báez (9–10)	—	27,696	74–51
126	August 20	@ Indians	6–3	**Harang** (5–3)	Westbrook (1–2)	**Koch** (32)	27,527	75–51
127	August 21	@ Indians	6–0	**Lidle** (7–9)	Rodríguez (0–1)	—	26,916	76–51
128	August 22	@ Indians	9–3	**Mulder** (15–7)	Phillips (1–2)	**Bradford** (2)	27,759	77–51
129	August 23	@ Tigers	9–1	**Zito** (18–5)	Powell (1–3)	—	21,807	78–51
130	August 24	@ Tigers	12–3	**Hudson** (11–9)	Lima (4–6)	—	19,045	79–51
131	August 25	@ Tigers	10–7	**Mecir** (4–3)	Walker (1–1)	**Koch** (33)	24,346	80–51
132	August 26	@ Royals	6–3	**Lidle** (8–9)	May (3–9)	**Koch** (34)	11,096	81–51
133	August 27	@ Royals	6–4	**Mulder** (16–7)	Hernández (3–3)	**Koch** (35)	13,077	82–51
134	August 28	@ Royals	7–1	**Zito** (19–5)	Sedlacek (3–4)	—	15,952	83–51
135	August 30	Twins	4–2	**Hudson** (12–9)	Radke (6–4)	**Koch** (36)	25,221	84–51
136	August 31	Twins	6–3	**Mecir** (5–3)	Romero (8–2)	**Koch** (37)	42,841	85–51

September: 18–8 (Home: 11–2 ; Away: 7–6) [hide]

#	Date	Opponent	Score	Win	Loss	Save	Attendance	Record
137	September 1	Twins	7–5	**Koch** (7–2)	Guardado (1–3)	—	37,676	86–51
138	September 2	Royals	7–6	**Koch** (8–2)	Grimsley (3–5)	—	26,325	87–51
139	September 4	Royals	12–11	**Koch** (9–2)	Grimsley (3–6)	—	55,528	88–51
140	September 6	@ Twins	0–6	Radke (7–4)	**Lidle** (8–10)	—	27,409	88–52
141	September 7	@ Twins	2–0	**Mulder** (17–7)	Mays (3–6)	**Koch** (38)	43,628	89–52

https://youtu.be/nK1jtVhimPA?t=8s

2016

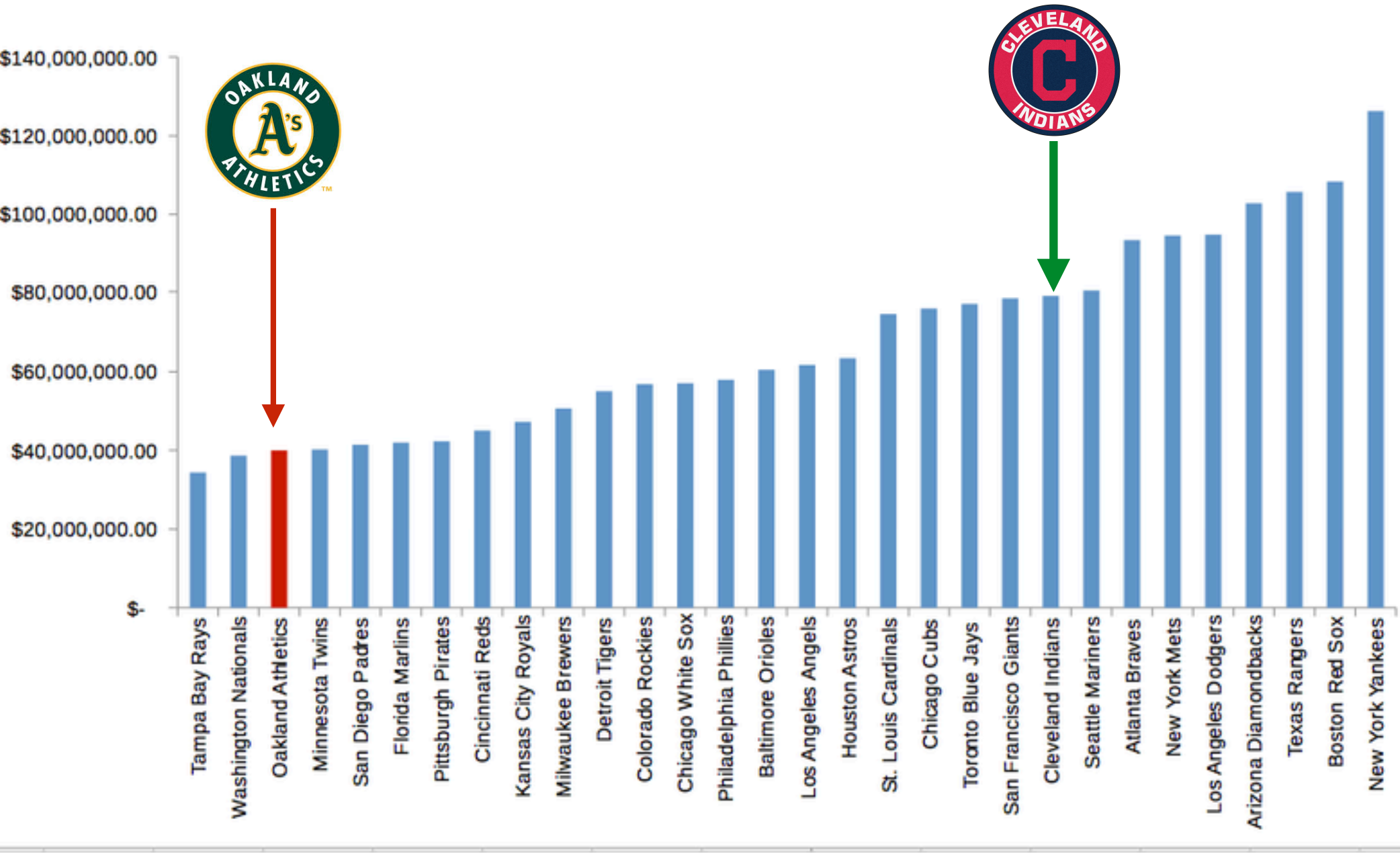
https://en.wikipedia.org/wiki/List_of_Major_League_Baseball_longest_winning_streaks

Rank	Games	Team	Season(s)
1	26	New York Giants^	1916
2 (tie)	21	Chicago White Stockings^	1880
2 (tie)	21	Chicago Cubs	1935
4 (tie)	20	St. Louis Maroons	1884
4 (tie)	20	Providence Grays	1884
4 (tie)	20	Oakland Athletics	2002
7 (tie)	19	Chicago White Sox^	1906*
7 (tie)	19	New York Yankees	1947*

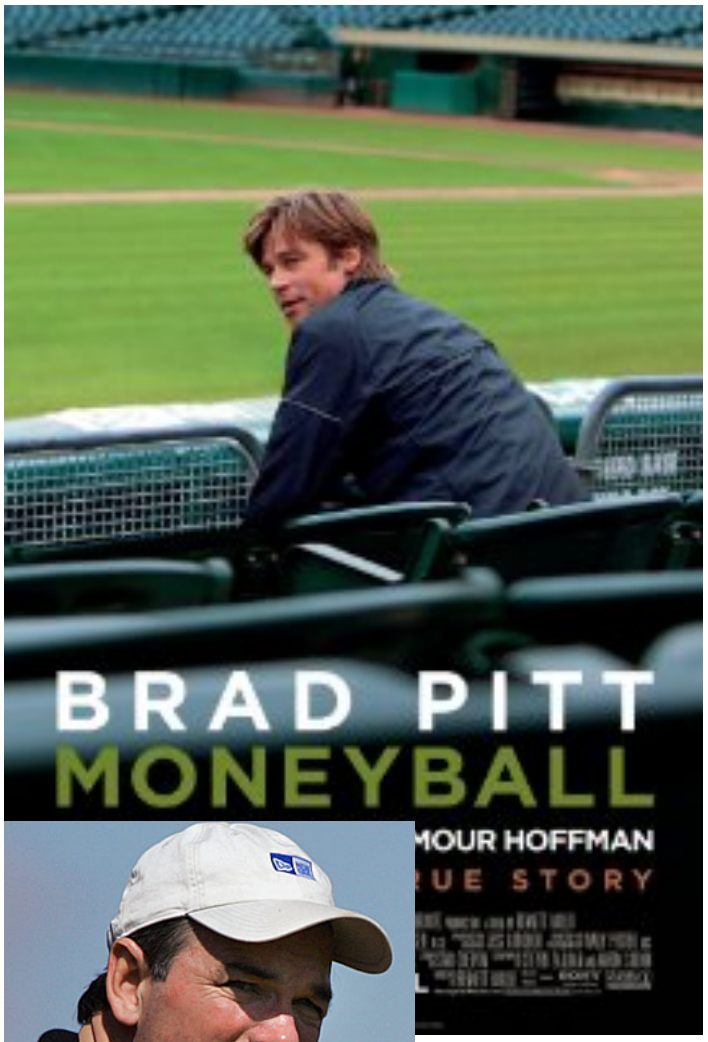
15 years later, after 2002

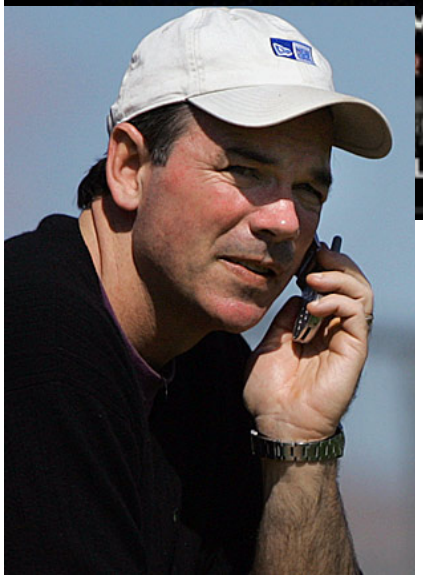
Rank	Games	Team	Season(s)
1	26^	New York Giants	1916
2	22	Cleveland Indians	2017
3 (tie)	21^	Chicago White Stockings	1880
3 (tie)	21	Chicago Cubs	1935
5 (tie)	20	St. Louis Maroons	1884
5 (tie)	20	Providence Grays	1884
5 (tie)	20	Oakland Athletics	2002
8 (tie)	19^	Chicago White Sox	1906*
8 (tie)	19	New York Yankees	1947*
10 (tie)	18	Chicago White Stockings	1885

Rich teams, Poor teams

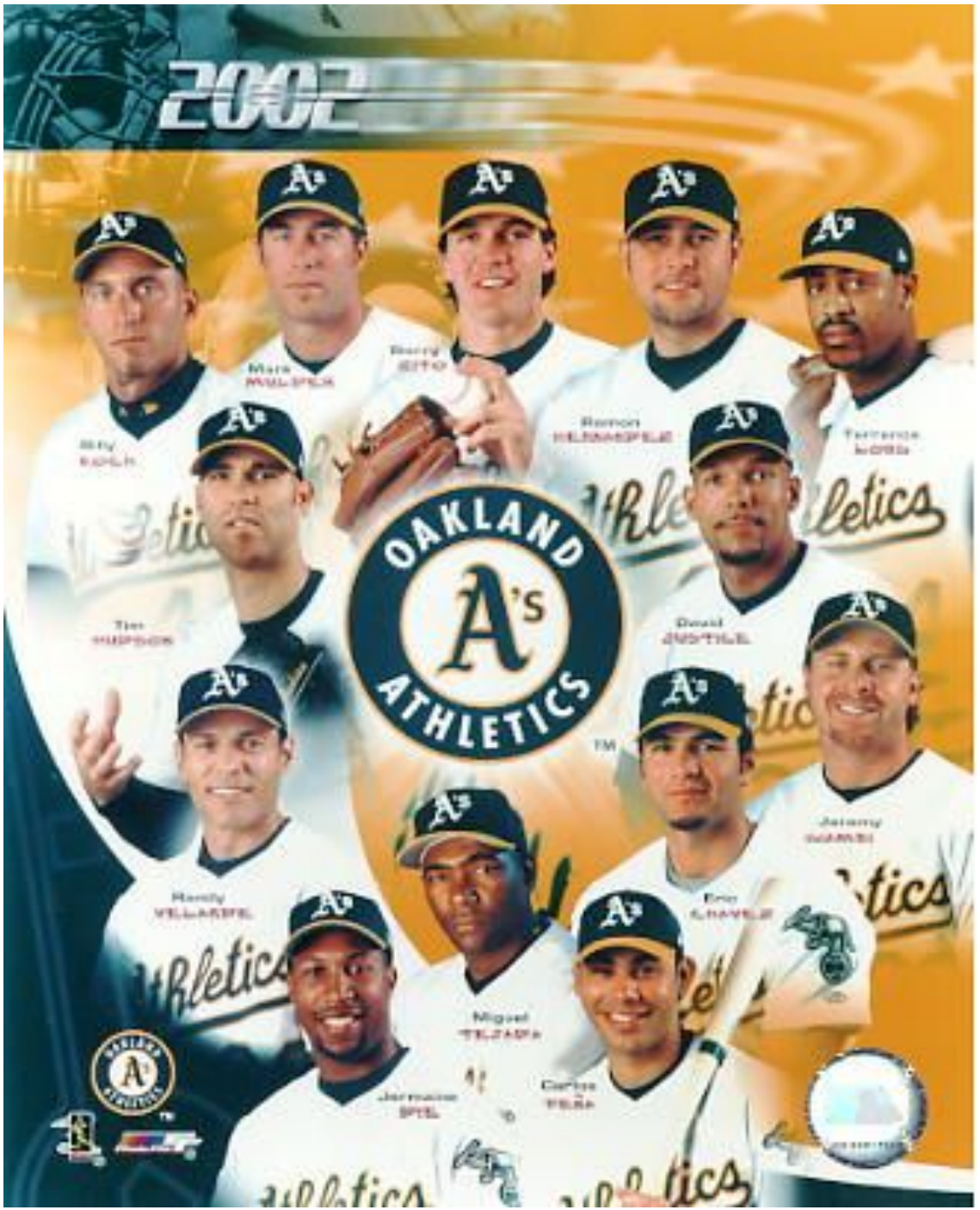


The Team of 2002





Billy Beane



Boston Red Sox

try to hire

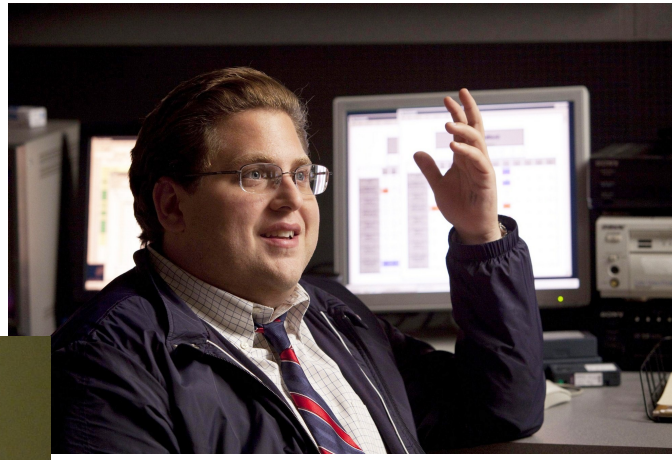
Billy Beane

hire

Bill James

data scientist

the data scientist

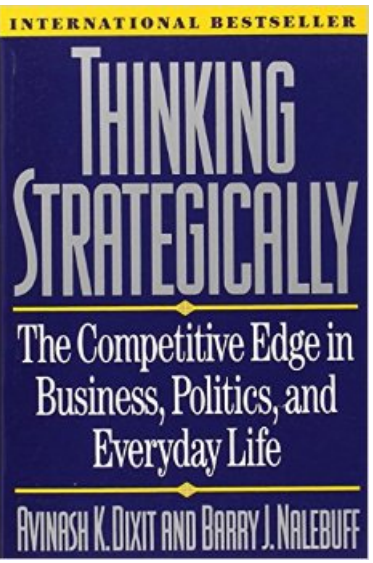


CLEVELAND BROWNS

Paul DePodesta

https://www.sbnation.com/nfl/2016/1/5/10717024/cleveland-browns-Paul-DePodesta-mets-moneyball

Strategy to Win



1983 America's Cup

Dennis Conner

VS

(Liberty) **3** : **1** (Australia II)

Date	Winner	Yacht	Loser	Yacht	Score	Delta
September 14, 1983	*Liberty*	US-40	*Australia II*	KA-6	0-1	1:10
September 15, 1983	*Liberty*	US-40	*Australia II*	KA-6	0-2	1:33
September 18, 1983	*Australia II*	KA-6	*Liberty*	US-40	1-2	3:14
September 20, 1983	*Liberty*	US-40	*Australia II*	KA-6	1-3	0:43
September 21, 1983						
September 22, 1983						
September 26, 1983						

https://youtu.be/gJ96kcXUhwI?t=32m10s

sailing direction

in 5th Game

start

Choice A

turn 5 degree

Choice B

head straight

Dennis Conner chose A

?

Liberty

?

chose B

Australia II

1983 America's Cup

Dennis Conner

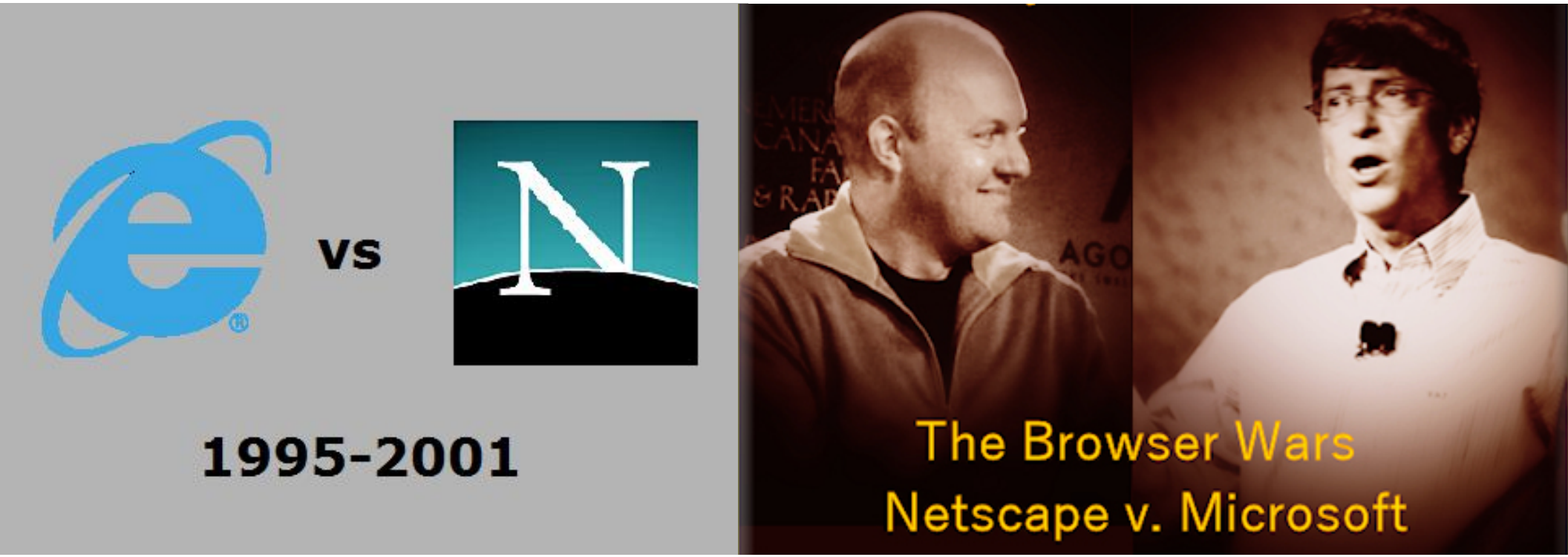
VS

Loss! (Liberty) 3 : **4** (Australia II)

Date	Winner	Yacht	Loser	Yacht	Score	Delta
September 14, 1983	*Liberty*	US-40	*Australia II*	KA-6	0-1	1:10
September 15, 1983	*Liberty*	US-40	*Australia II*	KA-6	0-2	1:33
September 18, 1983	*Australia II*	KA-6	*Liberty*	US-40	1-2	3:14
September 20, 1983	*Liberty*	US-40	*Australia II*	KA-6	1-3	0:43
September 21, 1983	*Australia II*	KA-6	*Liberty*	US-40	2-3	1:47
September 22, 1983	*Australia II*	KA-6	*Liberty*	US-40	3-3	3:25
September 26, 1983	*Australia II*	KA-6	*Liberty*	US-40	4-3	0:41

https://youtu.be/gJ96kcXUhwI?t=37m14s

Business Strategy



Software in 1995

software product

Choice A

non-Internet software

Choice B

Internet (browser)

Bill Gates also chose B

?

Microsoft

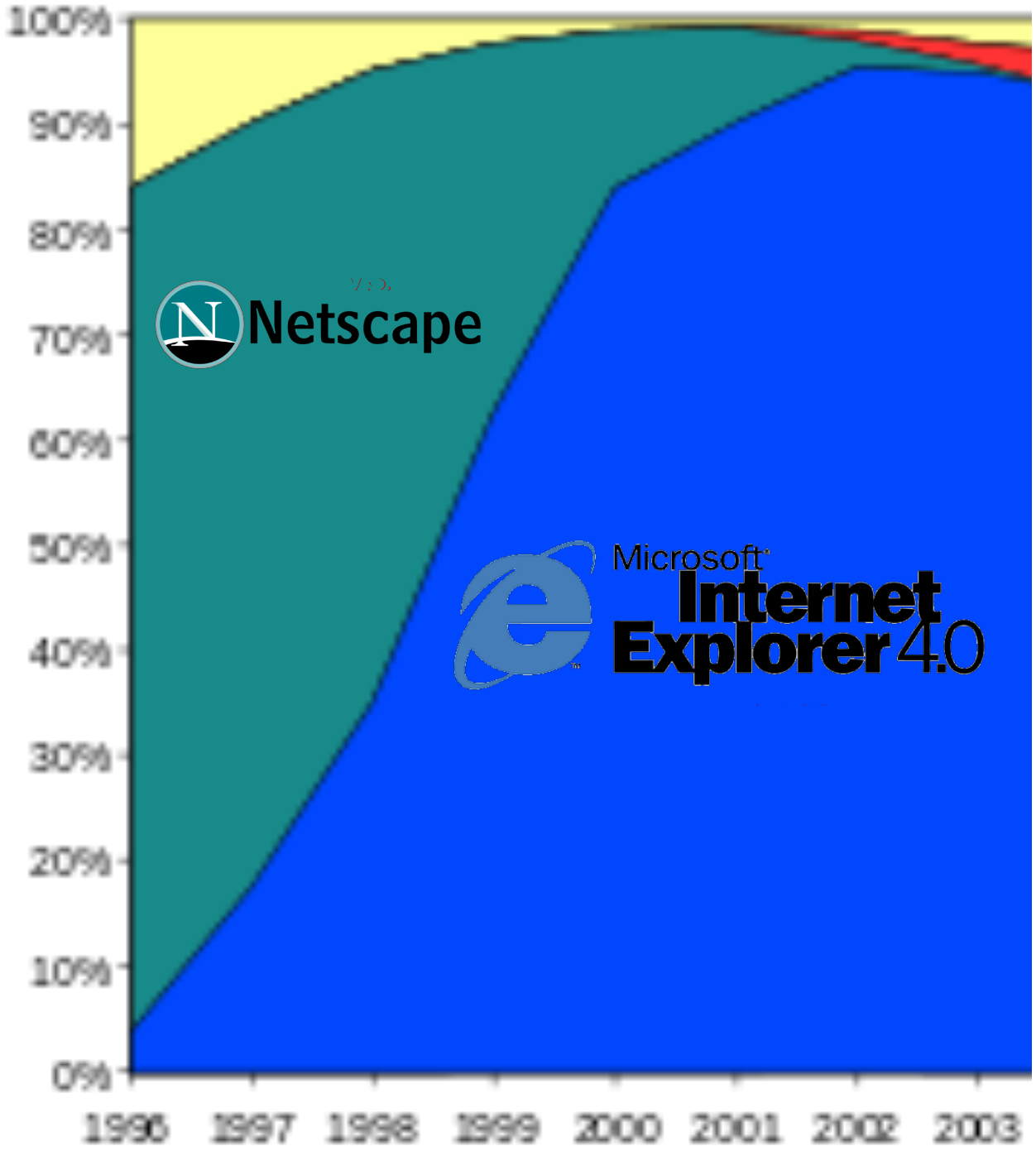
?

Netscape

chose B

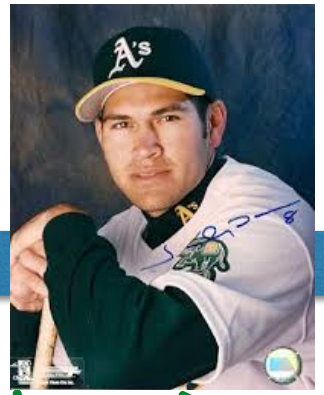
https://youtu.be/NbAnM3GuZQg?t=31m28s

Browser War

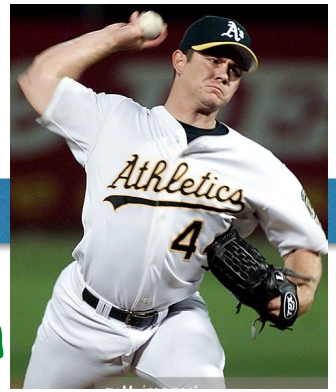

100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1996
1997
1998
1999
2000
2001
2002
2003
Netscape
Microsoft
Internet
Explorer 4.0

Baseball in 2002

Jason Giambi



Johnny Damon



Jason Isringhausen

Baseball in 2002

player selection

choice B

OBP, SLG

choice A

BA

chose **B**

chose **A**

Movie Rental Business

Netflix vs Blockbuster

Movie Rental in 2002

selling mode

choice B

mailing / online

choice A

store

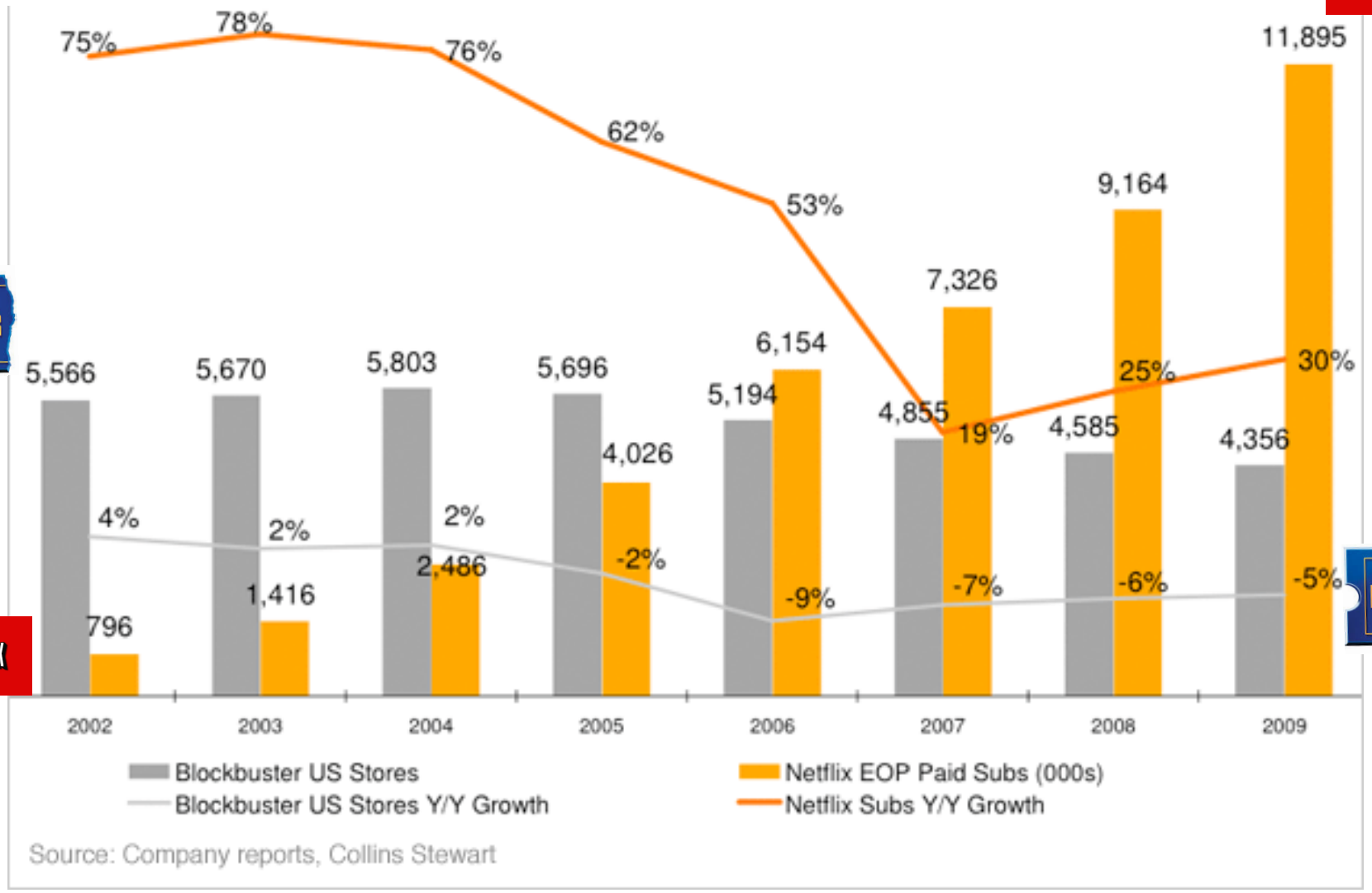
chose **B**

?

?

chose **A**

Movie Rental



in 2013

Business Strategy

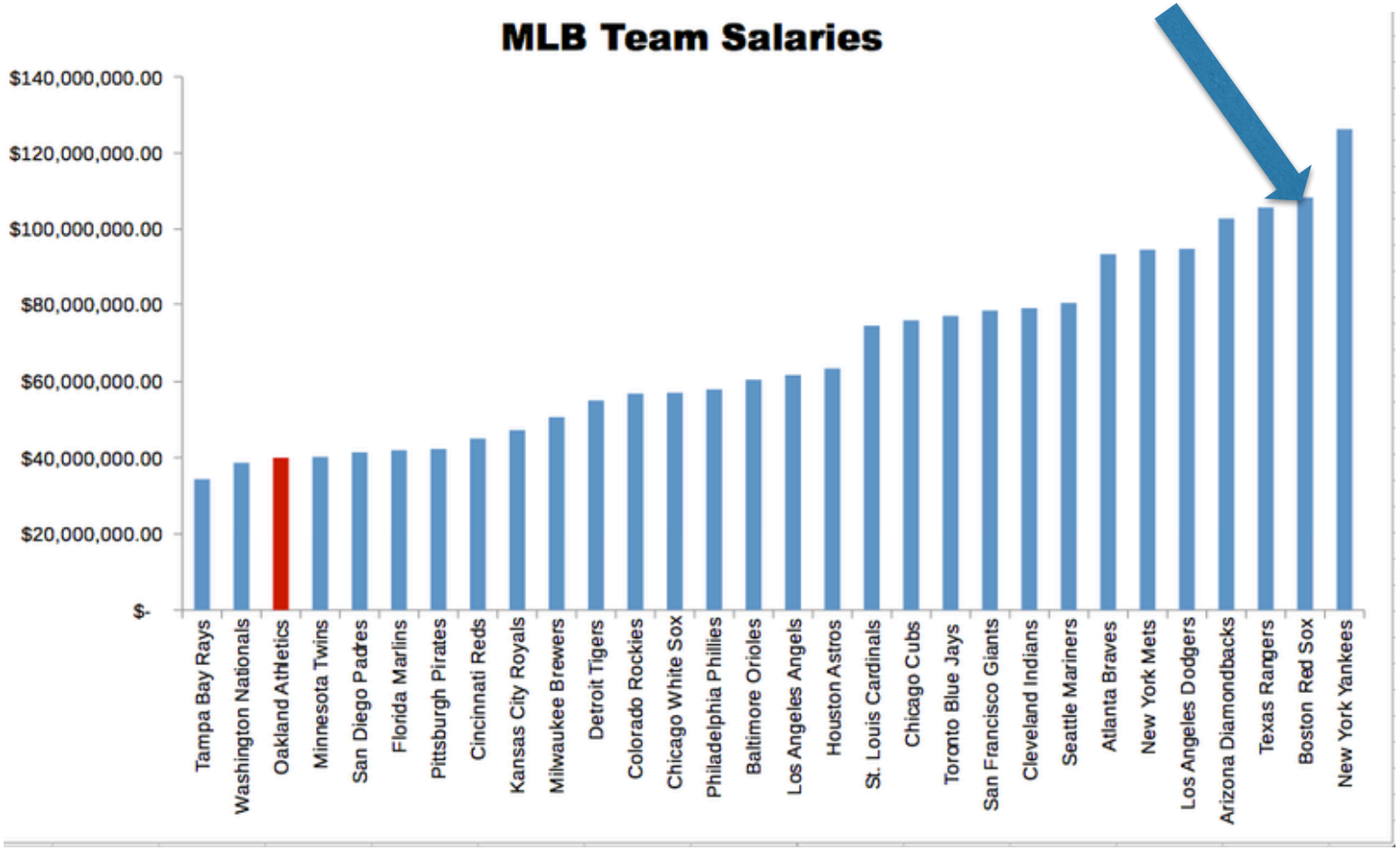
Amazon vs Book Store

amazon

Business Strategy

Uber vs Taxi Industry

Boston Red Sox in 2004



Boston Red Sox in 2004

try to hire

Billy Beane

hire

Bill James

data scientist


Divisional Series
League Championship Series
World Series
League Championship Series
Divisional Series
2004
MLB PLAYOFFS
Yankees
Twins
Angels
Red Sox
Yankees
Red Sox
Red Sox
Red Sox
World Series Champions
Cardinals
Cardinals
Dodgers
Cardinals
Braves
Astros
Astros

Red Sox Wins World Series in 2004