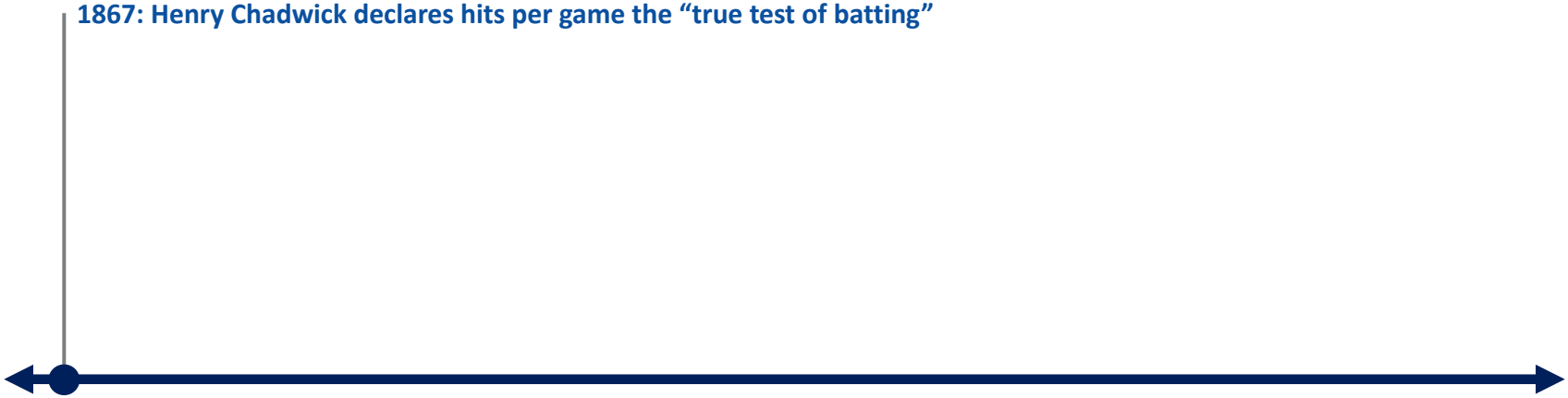
The Evolution of Batting Statistics in Baseball

Scott Powers

MMC Chicago Dinner Meeting
March 3, 2023

1867: Henry Chadwick declares hits per game the “true test of batting”



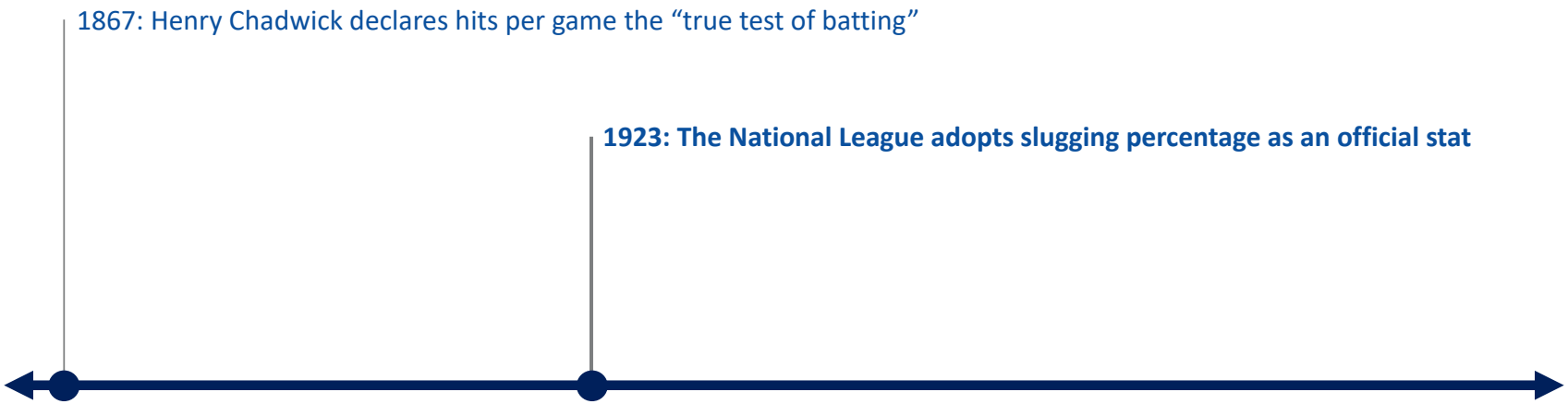
The Origin of Batting Average

- 1867: *The Ball Players' Chronicle* publishes "The True Test of Batting" by Henry Chadwick
  - Argues for hits per game as the "true test"
  - Prefers hits to total bases

- 1871: Hervie Alden Dobson writes letter to the *New York Clipper*
  - Argues for hits per at bat over hits per game
  - Later this came to be known as batting average

$$\text{AVG} = \text{H}/\text{AB}$$

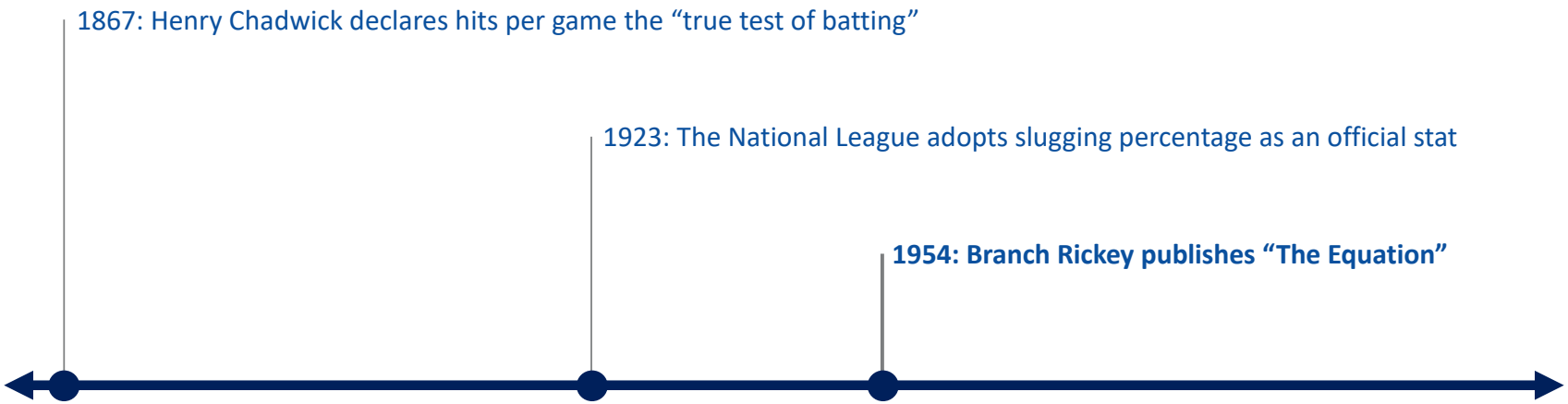
Source: John Thorn (2013)
https://ourgame.mlblogs.com/chadwicks-choice-the-origin-of-the-batting-average-e8e9e9402d53


1867: Henry Chadwick declares hits per game the “true test of batting”
1923: The National League adopts slugging percentage as an official stat

Slugging Percentage

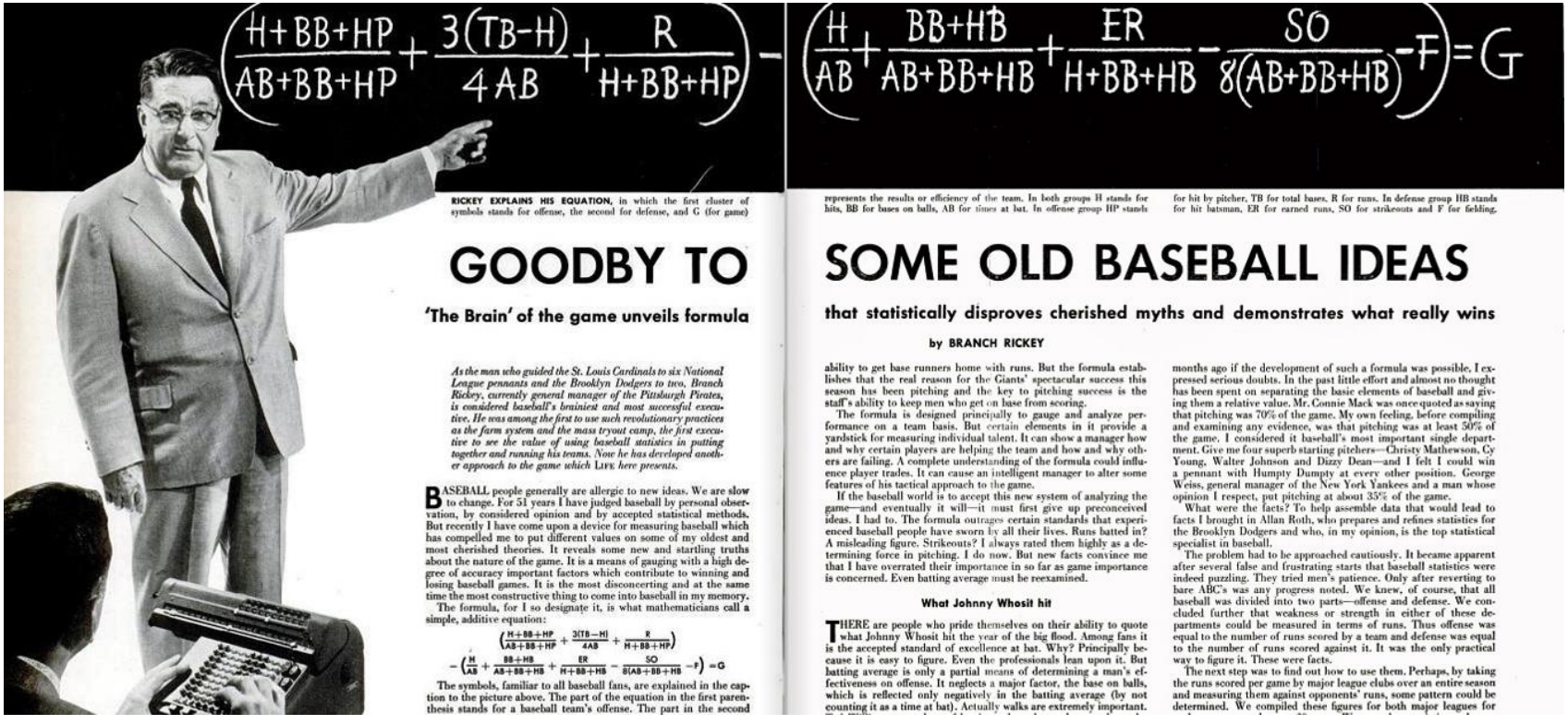
- 1923: National League adopts SLG as official stat
- 1946: American League adopts SLG as official stat

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


1867: Henry Chadwick declares hits per game the “true test of batting”
1923: The National League adopts slugging percentage as an official stat
1954: Branch Rickey publishes “The Equation”

"The Equation"

- 1954: Branch Rickey and Allan Roth publish "The Equation" in *Life* magazine



$$\left(\frac{H+BB+HP}{AB+BB+HP}+\frac{3(TB-H)}{4AB}+\frac{R}{H+BB+HP}\right)-\left(\frac{H}{AB}+\frac{BB+HB}{AB+BB+HB}+\frac{ER}{H+BB+HB}-\frac{SO}{8(AB+BB+HB)}-F\right)=G$$

RICKEY EXPLAINS HIS EQUATION, in which the first cluster of symbols stands for offense, the second for defense, and G (for game) represents the results or efficiency of the team. In both groups H stands for hits, BB for bases on balls, AB for times at bat. In offense group HP stands for hit by pitcher, TB for total bases, R for runs. In defense group HB stands for hit batsman, ER for earned runs, SO for strikeouts and F for fielding.

GOODBY TO

'The Brain' of the game unveils formula

As the man who guided the St. Louis Cardinals to six National League pennants and the Brooklyn Dodgers to two, Branch Rickey, currently general manager of the Pittsburgh Pirates, is considered baseball's brainiest and most successful executive. He was among the first to use such revolutionary practices as the farm system and the mass tryout camp, the first executive to see the value of using baseball statistics in putting together and running his teams. Now he has developed another approach to the game which LIFE *here presents.*

BASEBALL people generally are allergic to new ideas. We are slow to change. For 51 years I have judged baseball by personal observation, by considered opinion and by accepted statistical methods. But recently I have come upon a device for measuring baseball which has compelled me to put different values on some of my oldest and most cherished theories. It reveals some new and startling truths about the nature of the game. It is a means of gauging with a high degree of accuracy important factors which contribute to winning and losing baseball games. It is the most disconcerting and at the same time the most constructive thing to come into baseball in my memory.

The formula, for I so designate it, is what mathematicians call a simple, additive equation:

$$\left(\frac{H+BB+HP}{AB+BB+HP}+\frac{3(TB-H)}{4AB}+\frac{R}{H+BB+HP}\right)-\left(\frac{H}{AB}+\frac{BB+HB}{AB+BB+HB}+\frac{ER}{H+BB+HB}-\frac{SO}{8(AB+BB+HB)}-F\right)=G$$

The symbols, familiar to all baseball fans, are explained in the caption to the picture above. The part of the equation in the first parenthesis stands for a baseball team's offense. The part in the second

SOME OLD BASEBALL IDEAS

that statistically disproves cherished myths and demonstrates what really wins

by BRANCH RICKEY

ability to get base runners home with runs. But the formula establishes that the real reason for the Giants' spectacular success this season has been pitching and the key to pitching success is the staff's ability to keep men who get on base from scoring.

The formula is designed principally to gauge and analyze performance on a team basis. But certain elements in it provide a yardstick for measuring individual talent. It can show a manager how and why certain players are helping the team and how and why others are failing. A complete understanding of the formula could influence player trades. It can cause an intelligent manager to alter some features of his tactical approach to the game.

If the baseball world is to accept this new system of analyzing the game—and eventually it will—it must first give up preconceived ideas. I had to. The formula outrages certain standards that experienced baseball people have sworn by all their lives. Runs batted in? A misleading figure. Strikeouts? I always rated them highly as a determining force in pitching. I do now. But new facts convince me that I have overrated their importance in so far as game importance is concerned. Even batting average must be reexamined.

What Johnny Whosit hit

THERE are people who pride themselves on their ability to quote what Johnny Whosit hit the year of the big flood. Among fans it is the accepted standard of excellence at bat. Why? Principally because it is easy to figure. Even the professionals lean upon it. But batting average is only a partial means of determining a man's effectiveness on offense. It neglects a major factor, the base on balls, which is reflected only negatively in the batting average (by not counting it as a time at bat). Actually walks are extremely important.

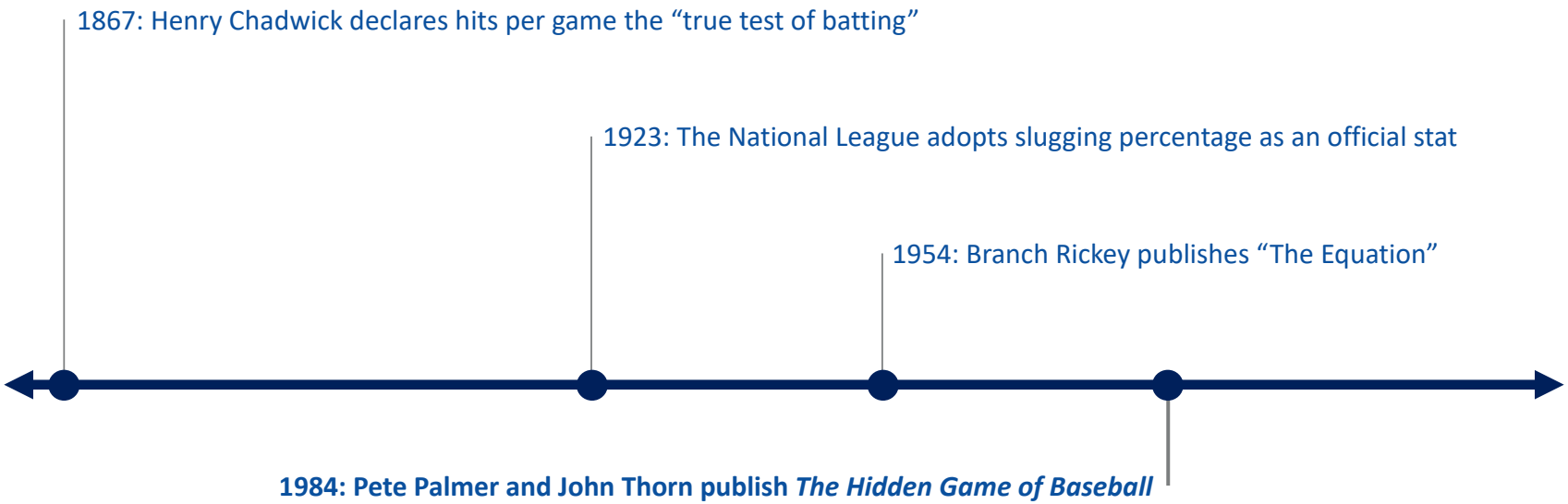
months ago if the development of such a formula was possible, I expressed serious doubts. In the past little effort and almost no thought has been spent on separating the basic elements of baseball and giving them a relative value. Mr. Connie Mack was once quoted as saying that pitching was 70% of the game. My own feeling, before compiling and examining any evidence, was that pitching was at least 50% of the game. I considered it baseball's most important single department. Give me four superb starting pitchers—Christy Mathewson, Cy Young, Walter Johnson and Dizzy Dean—and I felt I could win a pennant with Humpty Dumpty at every other position. George Weiss, general manager of the New York Yankees and a man whose opinion I respect, put pitching at about 33⅓ of the game.

What were the facts? To help assemble data that would lead to facts I brought in Allan Roth, who prepares and refines statistics for the Brooklyn Dodgers and who, in my opinion, is the top statistical specialist in baseball.

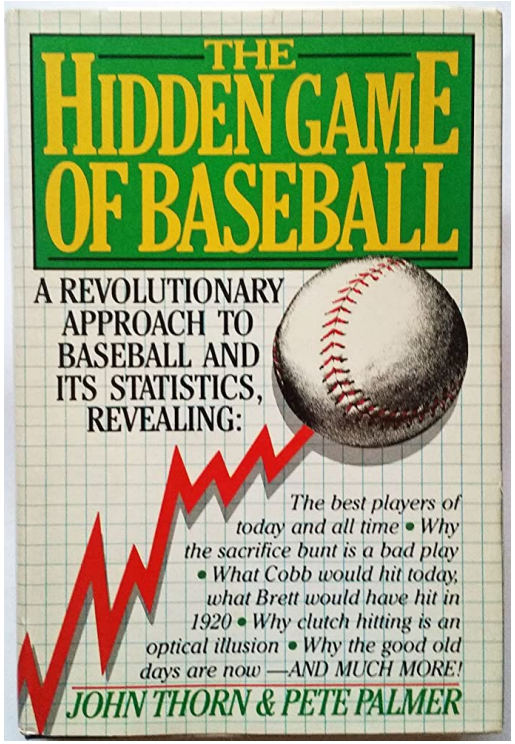
The problem had to be approached cautiously. It became apparent after several false and frustrating starts that baseball statistics were indeed puzzling. They tried men's patience. Only after reverting to bare ABC's was any progress noted. We knew, of course, that all baseball was divided into two parts—offense and defense. We concluded further that weakness or strength in either of these departments could be measured in terms of runs. Thus offense was equal to the number of runs scored by a team and defense was equal to the number of runs scored against it. It was the only practical way to figure it. These were facts.

The next step was to find out how to use them. Perhaps, by taking the runs scored per game by major league clubs over an entire season and measuring them against opponents' runs, some pattern could be determined. We compiled these figures for both major leagues for

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


1867: Henry Chadwick declares hits per game the “true test of batting”
1923: The National League adopts slugging percentage as an official stat
1954: Branch Rickey publishes “The Equation”
1984: Pete Palmer and John Thorn publish *The Hidden Game of Baseball*

The Hidden Game of Baseball



- 1984: John Thorn and Pete Palmer publish *The Hidden Game of Baseball*
  - (Also in 1984: MLB adopts OBP)
- Introduces "The Linear Weights System"
- Popularizes "On-base Plus Slugging"

$$\text{OPS} = \frac{\text{H} + \text{BB} + \text{HBP}}{\text{AB} + \text{BB} + \text{HBP} + \text{SF}} + \frac{1\text{B} + 2 \cdot 2\text{B} + 3 \cdot 3\text{B} + 4 \cdot \text{HR}}{\text{AB}}$$

A Markov model for an inning of baseball

- Inning state: Which bases are occupied, and how many outs are there? e.g. 0-0-1-2 = runner on third base with 2 outs
  - 25 possibilities ($2 \times 2 \times 2 \times 3$ + end of inning)
- Transition probabilities from 0-0-1-2:

64% $\rightarrow$	end of inning
14% $\rightarrow$	1-0-0-2 (+1 run)
13% $\rightarrow$	1-0-1-2
5% $\rightarrow$	0-1-0-2 (+1 run)
3% $\rightarrow$	0-0-0-2 (+2 runs)
1% $\rightarrow$	end of inning (+1 run)
$<$ 1% $\rightarrow$	0-0-1-2 (+1 run)

Transition probability matrix

- $A \in [0,1]^{25 \times 25}$ encodes transition probabilities between states
- A_{ij} is the probability of transitioning to state j from state i
- Multiply A by itself to get multi-step transition probability
  - A^2 is transition probability matrix after 2 steps
  - A^3 is transition probability matrix after 3 steps
  - ...
  - A^∞ gives terminal state probability
    All initial states converge to end of inning

Exercise: How can we use this model to calculate the expected number of runs scored from each initial state?

(My) Solution

- Track runs scored in the inning state! e.g. 0-0-1-2-X = runner on third base with 2 outs, X runs have scored
- Infinite possibilities!
- But can safely ignore $X > 9$ (very small probability)
  - 250 possibilities (25 from earlier $\times$ 10 values of X)
  - $A \in [0,1]^{250 \times 250}$ encodes transition probabilities
  - A^{20} is close enough to A^{∞}

Run Expectancy by Base-Out State

State	RE
1-1-1-0	2.41
...	...
0-1-0-0	1.15
...	...
1-0-0-0	0.91
...	...
0-0-0-0	0.51
...	...
0-0-0-1	0.27
...	...
0-0-0-2	0.11
0-0-0-3	0.00

Example #1: Leadoff double

$$\begin{aligned} \text{RE(0-1-0-0)} - \text{RE(0-0-0-0)} &= 1.15 - 0.51 \\ &= +0.64 \text{ runs!} \end{aligned}$$

Example #2: Bases empty, two-out double

$$\begin{aligned} \text{RE(0-1-0-2)} - \text{RE(0-0-0-2)} &= 0.33 - 0.11 \\ &= +0.22 \text{ runs!} \end{aligned}$$

Example #3: Bases loaded, no-out double

$$\begin{aligned} 3 + \text{RE(0-1-0-0)} - \text{RE(1-1-1-0)} &= 3 + 1.15 - 2.41 \\ &= +1.74 \text{ runs!} \end{aligned}$$

Average Change in Run Expectancy by Event

Event	Change in RE	*vs* Out	OPS Weight
Home Run	$+1.38$	1.65	5
Triple	$+1.06$	1.33	4
Double	$+0.76$	1.03	3
Single	$+0.46$	0.73	2
Hit By Pitch	$+0.35$	0.62	1
Walk	$+0.33$	0.60	1
Flyout	-0.27	—	0
Strikeout	-0.27	—	0
Groundout	-0.29	—	0

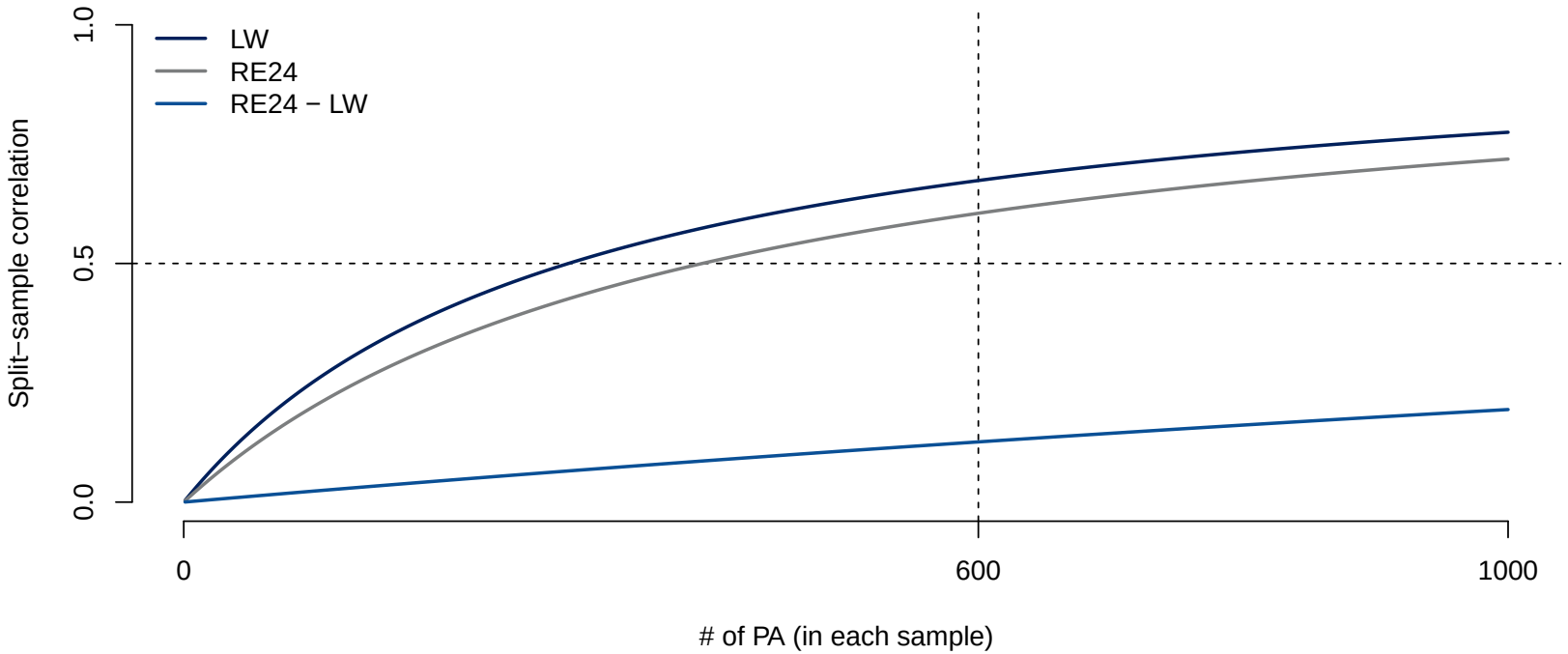
Takeaway #1

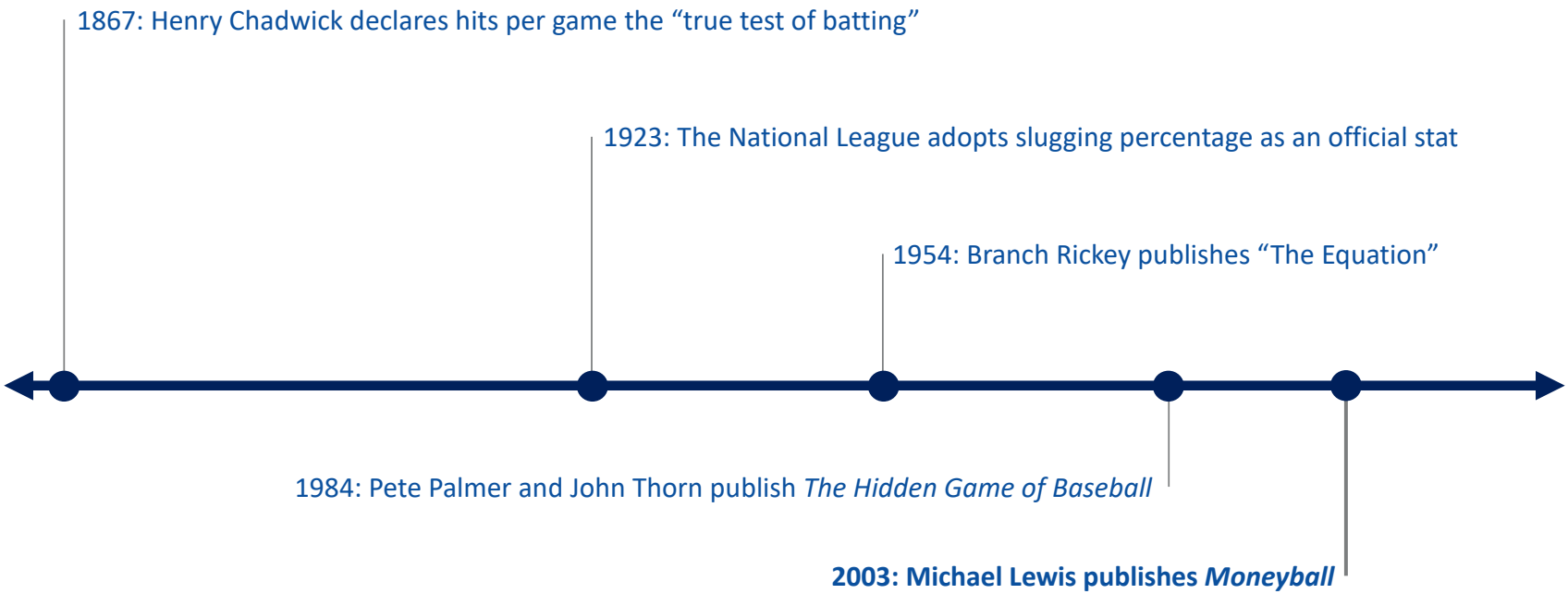
Sport analytics is less *crafting* metrics and more *deriving* them.

Two Perspectives

Player	Start State	Event	End State	RE24	LW
A	0-0-0-2	Groundout	End of Inning	−0.11	−0.29
	1-0-0-0	Strikeout	1-0-0-1	−0.38	−0.27
	0-1-1-1	Flyout	0-1-0-2 (+1)	−0.07	−0.27
	0-1-1-2	Single	1-0-0-2 (+2)	+1.63	+0.46
			Total	+1.07	−0.37
B	0-0-0-2	Single	1-0-0-2	+0.12	+0.46
	1-0-0-0	Groundout	0-0-0-2	−0.81	−0.29
	0-1-1-1	Strikeout	0-1-1-2	−0.80	−0.27
	0-1-1-2	Flyout	End of Inning	−0.60	−0.27
			Total	−2.09	−0.37

LW vs. RE24


LW
RE24
RE24 − LW
Split−sample correlation
1.0
0.5
0.0
0
600
1000
# of PA (in each sample)

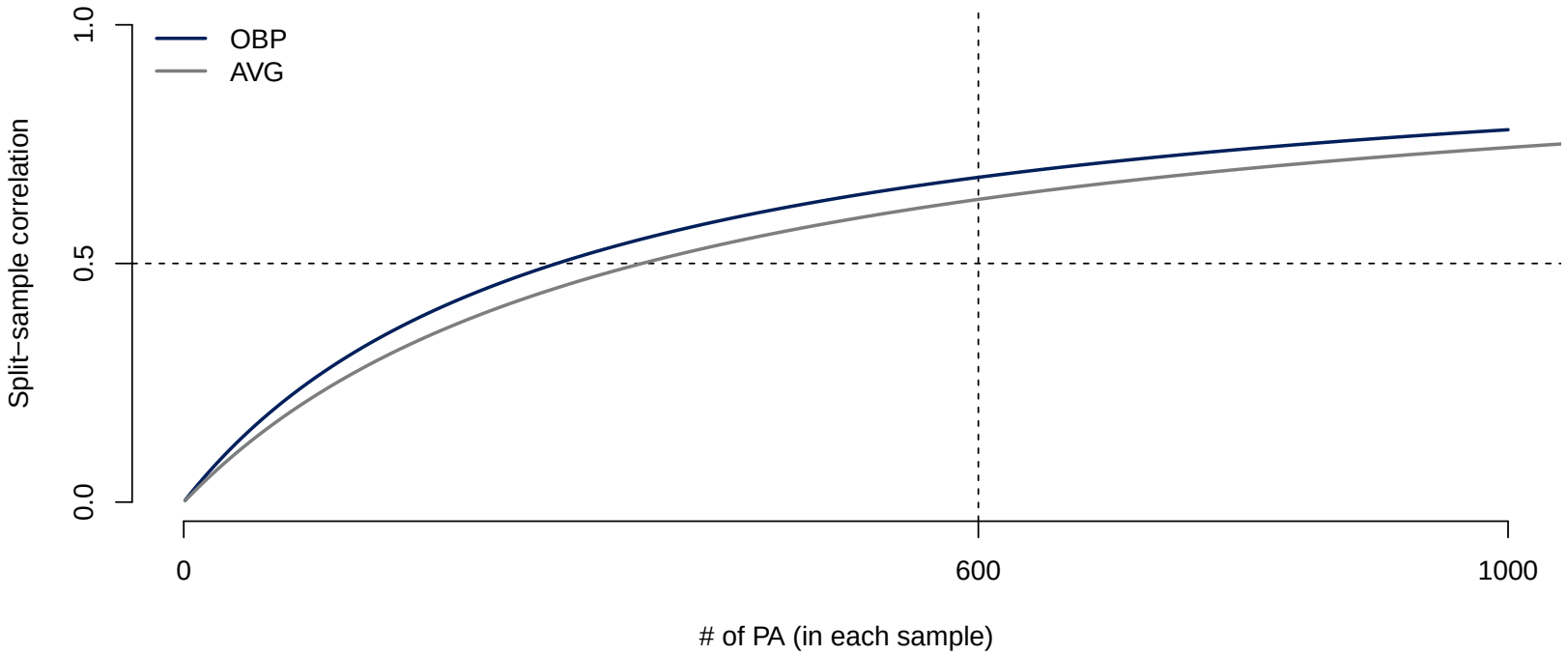

1867: Henry Chadwick declares hits per game the “true test of batting”
1923: The National League adopts slugging percentage as an official stat
1954: Branch Rickey publishes “The Equation”
1984: Pete Palmer and John Thorn publish *The Hidden Game of Baseball*
2003: Michael Lewis publishes *Moneyball*

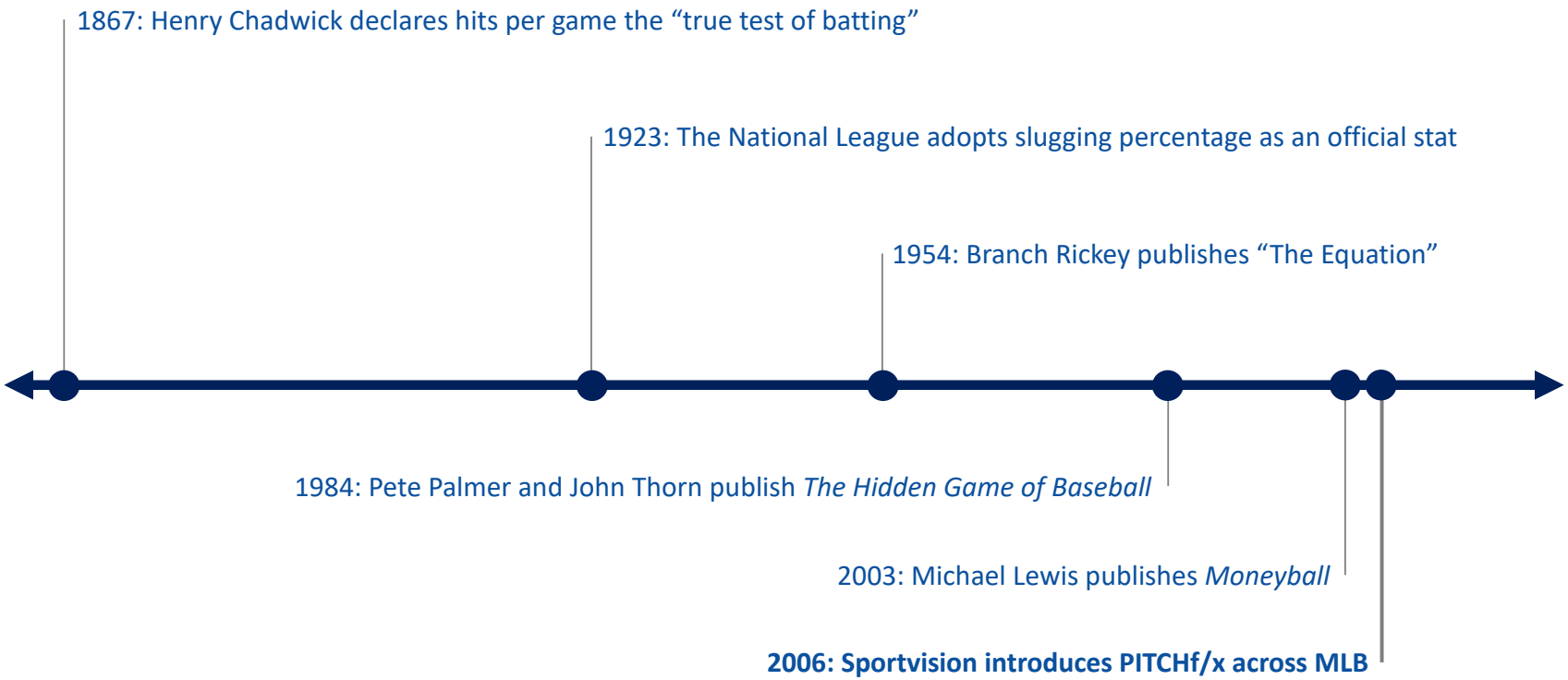
Moneyball, and OBP vs. AVG

Sports Illustrated

- 2003: Michael Lewis publishes *Moneyball*
- Why was batting average favored over on-base percentage for so long?
- Two criteria: Does it correlate to team success? Is it reliable?
  - Correlation with team success: see previous slides
  - Reliability: Less obvious

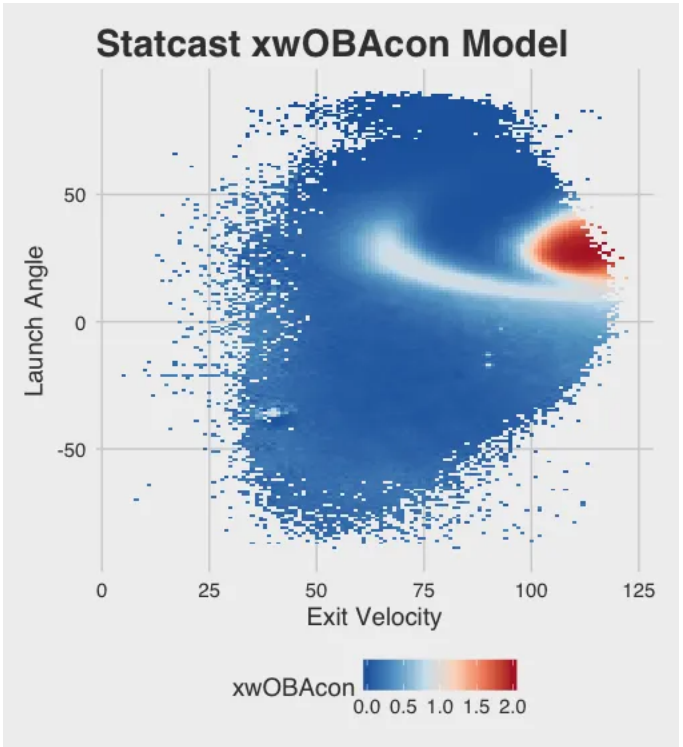
OBP vs. AVG


OBP
AVG
Split−sample correlation
1.0
0.5
0.0
0
600
1000
# of PA (in each sample)


1867: Henry Chadwick declares hits per game the “true test of batting”
1923: The National League adopts slugging percentage as an official stat
1954: Branch Rickey publishes “The Equation”
1984: Pete Palmer and John Thorn publish *The Hidden Game of Baseball*
2003: Michael Lewis publishes *Moneyball*
2006: Sportvision introduces PITCHf/x across MLB

Ball Tracking

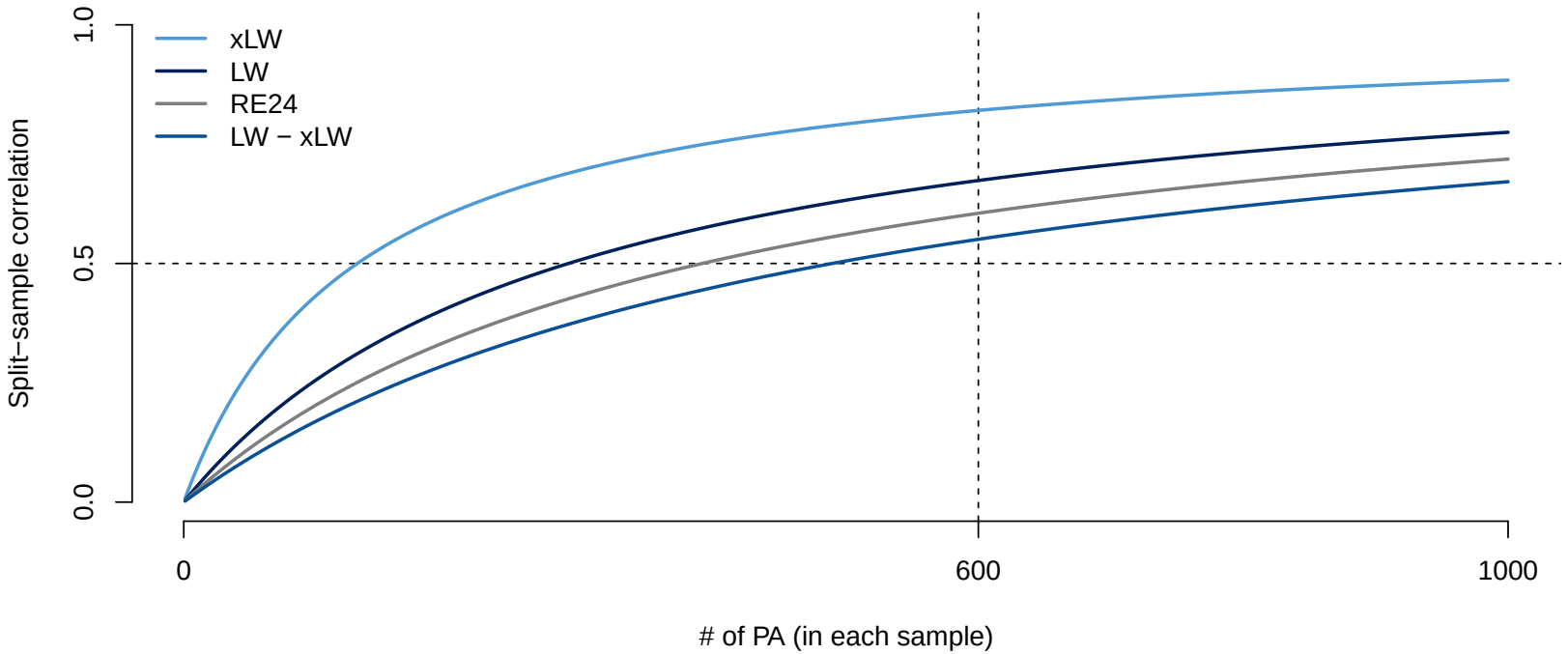
- 2006: Sportvision introduces PITCHf/x across MLB
- 2009: Sportvision introduces HITf/x across MLB
  - Based on computer vision
- 2017: TrackMan replaces Sportvision pitch and hit tracking
  - Based on Doppler radar



A Third Perspective

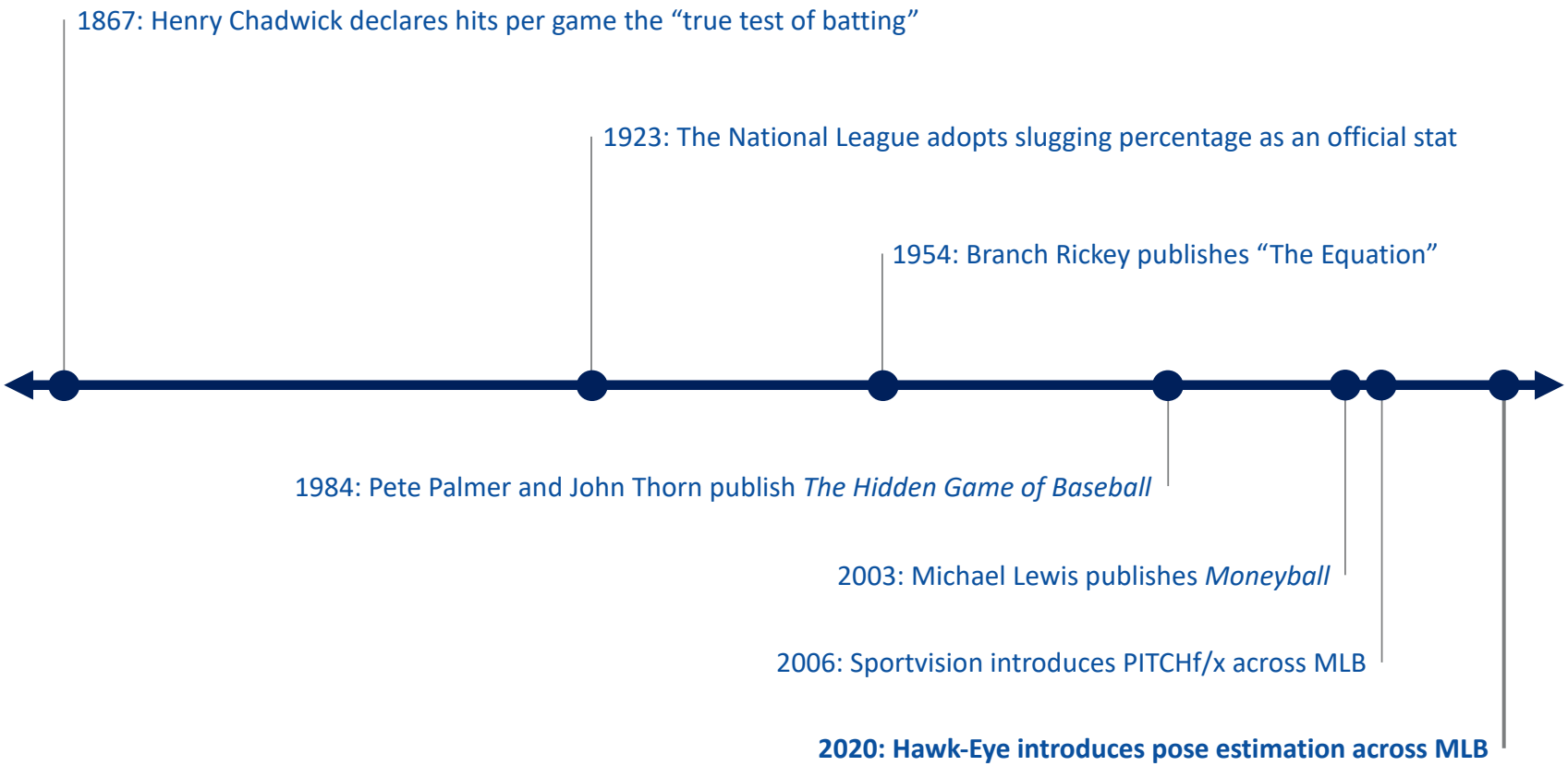
Player	Start State	Event	EV	LA	RE24	LW	xLW
A	0-0-0-2	Groundout	80	-20	−0.11	−0.29	−0.22
	1-0-0-0	Strikeout			−0.38	−0.27	−0.27
	0-1-1-1	Flyout	93	32	−0.07	−0.27	−0.10
	0-1-1-2	Single	87	-2	+1.63	+0.46	−0.07
				Total	+1.07	−0.37	−0.66
B	0-0-0-2	Single	91	16	+0.12	+0.46	+0.37
	1-0-0-0	Groundout	96	-12	−0.81	−0.29	−0.08
	0-1-1-1	Strikeout			−0.80	−0.27	−0.27
	0-1-1-2	Flyout	90	24	−0.60	−0.27	−0.08
				Total	−2.09	−0.37	−0.06

xLW vs. LW

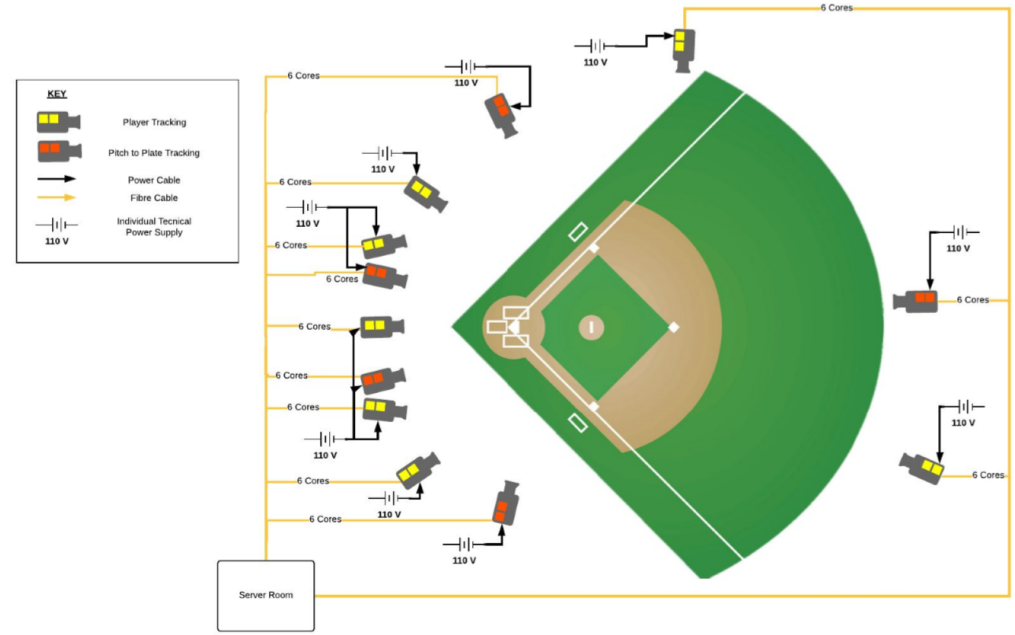

xLW
LW
RE24
LW − xLW
Split−sample correlation
1.0
0.5
0.0
0
600
1000
# of PA (in each sample)

Takeaway #2

Reliability is key in sport analytics, and the tradeoff between metrics depends on sample size. (Hint: You can blend them!)


1867: Henry Chadwick declares hits per game the “true test of batting”
1923: The National League adopts slugging percentage as an official stat
1954: Branch Rickey publishes “The Equation”
1984: Pete Palmer and John Thorn publish *The Hidden Game of Baseball*
2003: Michael Lewis publishes *Moneyball*
2006: Sportvision introduces PITCHf/x across MLB
2020: Hawk-Eye introduces pose estimation across MLB

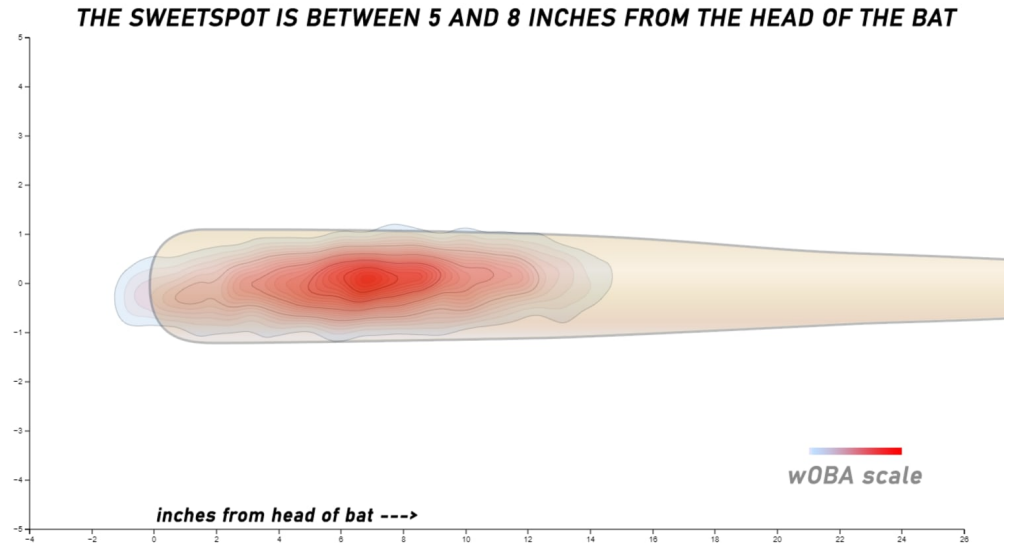
Hawk-Eye Camera System



- 2020: Installed across MLB
- 12 cameras @ 50-100 fps
- Tracking & pose estimation via computer vision
- Same tech used in tennis
- 2023 $\rightarrow$ 300 fps!!

Source: Ben Jedlovec (2020)
https://technology.mlblogs.com/introducing-statcast-2020-hawk-eye-and-google-cloud-a5f5c20321b8

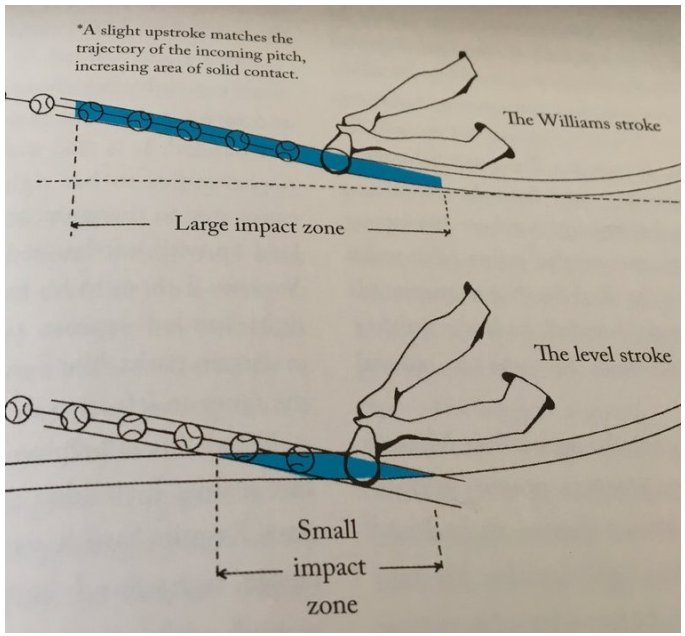
Hawk-Eye Bat Tracking



2022'S AVERAGE SWING SPEEDS*

ASTROS	MPH	DODGERS	MPH
Yordan Alvarez	89.6	Trayce Thompson	91.1
Jason Castro	84.3	Edwin Rios	85.4
Kyle Tucker	84.3	Jake Lamb	84.8
Korey Lee	83.3	Max Muncy	83.8
Alex Bregman	82.0	Eddy Alvarez	83.4
Chas McCormick	81.5	Mookie Betts	82.3
José Siri	81.4	Chris Taylor	80.9
Martín Maldonado	80.6	Gavin Lux	79.6
Yuli Gurriel	80.5	Trea Turner	79.5
Jeremy Peña	80.3	Will Smith	79.0
José Altuve	78.1	Freddie Freeman	78.7
Mauricio Dubón	77.6	Kevin Pillar	78.7
Jake Meyers	76.7	Cody Bellinger	78.4
Michael Brantley	76.4	Austin Barnes	77.9
Aledmys Díaz	76.2	Hanser Alberto	77.2
J.J. Matijevic	74.6	Justin Turner	74.8

*TRACKING ONLY AVAILABLE IN HOU, LAD. MINIMUM 3 BATTED BALL CONTACTS.



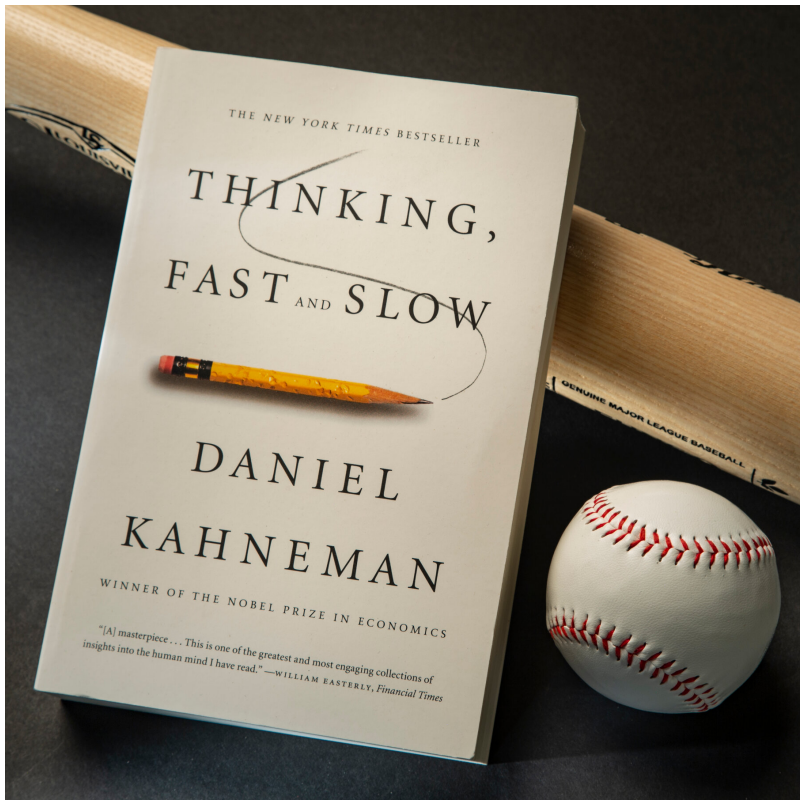
Ted Williams
The Science of Hitting (1970)
c/o @JWonCATCHING

Source: Mike Petriello (2022)
https://www.mlb.com/news/what-you-need-to-know-about-statcast-bat-tracking

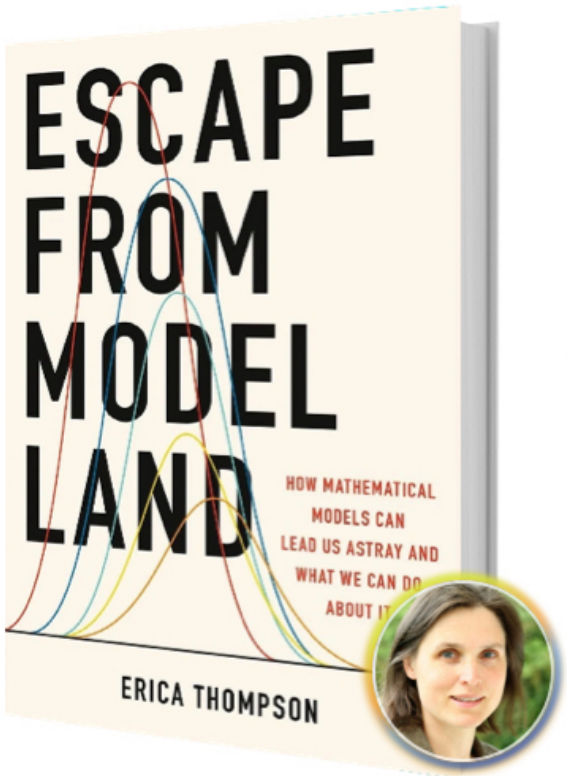
Takeaways

1. Sport analytics is less *crafting* metrics and more *deriving* them.
2. Reliability is key in sport analytics, and the tradeoff between metrics depends on sample size. (Hint: You can blend them!)

Recommended Reading



The New York Times



Thank You!