Harvard Data Science Review • 1.2

The Bases of Data

Christopher J. Phillips

Published on: Nov 01, 2019

Updated on: Nov 13, 2019

DOI: 10.1162/99608f92.5c483119

Data scientists may trace the origins of their field back to the pioneering work of John Tukey (1962), or perhaps even farther back into the depths of statistical history. But for many Americans, the origins are clear: Brad Pitt's stint as an evangelist for data analysis in the film *Moneyball* (Lewis, 2003; Zaillian & Sorkin, 2011).

Though this is not a convincing origin story for practitioners, the history of baseball *does* reveal some truths about the history of data analytics. If anything, a focus on *Moneyball* obscures a much older and more complicated history between the fields. Statistical data have long been collected in baseball, and the origins of the sport are intertwined with the history of statistics. The rules of baseball were initially formalized in the 1830s and 1840s, just as statistical societies across Europe and North America were founded, including the American Statistical Association (1839). Though numerical records had existed for cricket and other sports, baseball's record keepers maintained the most extensive statistics of any sport, and consistently promoted the idea that careful analysis of data could improve the game.

Even before professional leagues existed, baseball teams kept elaborate statistics of each contest that —enthusiasts claimed—would enable them to identify the best players and strategies. One of the most influential 19th-century propagandists for baseball, Henry Chadwick, was the half-brother of Edwin Chadwick, author of the seminal report on the *Sanitary Condition of the Labouring Population of Great Britain* (E. Chadwick, 1842). Both Chadwicks asserted that by collecting data—whether on public health or baseball—it was possible to identify desirable reforms. Just as certain sanitary conditions corresponded with a shorter lifespan for the working class, certain plays resulted in more or less success in baseball, and appropriate changes in strategy and player acquisition might be advocated on that basis, using data to make reforms more 'scientific' (H. Chadwick, 1874). Before the start of the 20th century, leagues were keeping 'official scores' and fans around the country followed box scores of each game, memorized their favorite players' statistics, and debated which numbers might best establish a player's relative worth (Schwarz, 2004). These data certainly weren't 'big' in a contemporary sense, but that didn't stop baseball fans, players, and front offices from consistently believing scientific analysis of data gathered from laboratory experiments and performance records might prove valuable.[1]

More recent evidence of the relationship includes popular data guru Nate Silver (2012), who began his career as a baseball analyst, and mathematician Cathy O'Neil, whose *Weapons of Math Destruction* (2016) praises baseball as an "ideal" and "healthy" example of the use of mathematical models, which were in other cases all too often misused and misinterpreted. As I and others have argued at more length elsewhere, even those who don't enjoy the game itself might turn to the history of baseball statistics as one site for understanding some key themes in the history of data science (Baumer & Zimbalist, 2014; Cramer, 2019; Phillips, 2019).

Data Don't Just Emerge: They Must Be Created

The collection of elaborate statistics is not necessary for playing or watching baseball: the rules require simple counting of outs and recording of runs. The making of extensive statistical records was a *choice* of baseball aficionados, one that took effort to coordinate and an investment of time and money to purchase scorebooks, train scorers, and maintain records. Even before the professionalization of the sport, scorekeepers were appointed by each team to record the character of every play. Nineteenth-century newspapers would keep their own records, using them to determine year-end playing awards and to assess potential trades between clubs. After professional leagues were founded in the 1870s, league secretaries were tasked with creating standardized scoresheets, collecting them from scorers after each game, recording the statistical data in 'day-by-day' ledgers and then summing up totals for players and teams to be published at the end of the year. An entire structure of human labor and technologies of recording has been developed to create data, and events that weren't deemed worthy of tracking (pickoff attempts, pitch counts, etc.) simply don't exist. It doesn't mean they didn't happen, but as far as data are concerned, they're invisible. The etymology of data (Latin for 'that which has been given') as well as the oxymoronic phrase 'raw data' belie the work that goes into the production of data (Gitelman, 2013). As Rob Kitchin once noted, it might be better to rename data as "capta" to highlight the active labor of capturing and recording that is inevitably required (2014, pp. 2–3).

Data Are Physical Objects, Subject to Friction and Corruption

Even with the effort to collect statistical records of the game, scoresheets went missing or were corrupted, books had typos and misprints, conflicting records had to be regularly reconciled. Data don't float free or travel unencumbered: they exist in material records, on paper, ledgers, punched cards, and—more recently—hard drives and servers. Historians of science increasingly track 'data friction,' tracing not only the resistance to movement some storage media present, but also the labor required to gather the data in one place (Edwards, 2010). A leading resource for baseball data, baseball-reference.com, was initially built from data scraped off a CD-ROM provided with the 1993 edition of *Total Baseball*, itself a compendium that had been made from co-author Pete Palmer's personal database, which in turn had been created by collecting statistics from league records and comparing them against published copies of *Sporting News*, year-end guides, and baseball encyclopedias (Phillips, 2019). It takes a great deal of work to move from the marking of paper scoresheets by official scorers at individual games to records that cover players' careers and teams' seasons, to databases that can be precisely queried, to an online-only interface in which nearly anyone anywhere can find the data needed. Not surprisingly, these processes depend on, and drive, technological developments as well: Palmer created his database using punched cards while he worked for Systems Development Corporation, a job that afforded him access to the latest supercomputers and introduced him to the very idea of an electronic database, a tool conceptualized in part through a Systems Development Corporation technical memorandum from the 1950s ("Database," 2012; Haigh, 2009).

The Stability and Reliability of Data Is an Accomplishment, Not a Natural State

One consequence of the work that goes into the creation and maintenance of data is that reliability must be earned. Indeed, Palmer used his supercomputers to check the consistency of baseball data across seasons and teams, ensuring columns were added appropriately (do the hits of all the players on a team add up to the team's total?). Moreover, there were predictably thousands of discrepancies in the data that had to be fixed individually through research using newspapers, old scoresheets, and other records. This wasn't a technical problem, but a historical one. Official records were simply those issued by the league office and could change without warning or notice: there are still examples in which publicly available databases and official league records don't match. Only through herculean efforts like that performed by Society for American Baseball Research committees (https://sabr.org/research/committees) and volunteers behind Retrosheet (https://www.retrosheet.org/) could the reliability and stability of data be established, data essential to any contemporary analysis that requires a large sample size drawn from historical records. It's impressive that we have statistics for baseball going back over 150 years, but it is amazing that there isn't more debate about such matters of fact given the labor needed to make them credible at all.

Data Aren't Insulated from Questions of Judgment and Expertise, but Intertwined With Them

One of the most fundamental statistics in baseball requires that official scorers distinguish hits from errors. For nearly every batted ball that results in a baserunner, the official scorer has to judge whether the play should be labeled a credit to the batter (a hit) or a debit to the fielders (an error). This intrusion of subjective judgment into the seemingly objective statistics of the game was, predictably, seen as a problem, and there have been a remarkable number of attempts to make scorers' judgment more objective over the years. At first, leagues emphasized *who* should score, promoting the use of gentlemen, on the same English philosophical presumption that gentlemen were the best candidates for practicing science. Insulated from base interests and the potential for bribery (and, in true Victorian fashion, not subject to the 'emotions' [!] of women), gentlemen could take on the task of apportioning credit. Later, the leagues contracted with newspaper writers as paragons of factual reporting. And still more recently, the leagues relied on outside contractors whose judgments could be appealed to the commissioner's office, making it less important who scorers were and more important that there was a *process* for ensuring judgments were centrally managed. There were also attempts to regulate *how* scoring took place, with guidelines about what constituted ordinary or extraordinary plays, and instructions on how to make judgments about credit without being swayed by innate biases. These attempts map onto the many different ways historians have found that people manufacture objectivity more generally from the messiness of everyday life: mechanized judgment, bureaucratic judgment, trained judgment, and disinterested judgment foremost among them (Daston & Galison, 2007; Shapin, 2012). That baseball's statistical records are sometimes treated as *opposed to* or *distinct from* human judgments is a testament to the success of these maneuvers. It's perhaps tempting to make stark distinctions between quantitative and qualitative entities, or to treat objectivity and subjectivity as two sides of a coin. But historically, they are often very difficult to separate.

New Data Analytics Might Offer Qualitatively Different Discoveries, but Are Often Built on the Same Structure as Existing Technologies

There is an open debate among historians as to whether the rise of big data has made a difference of scale or of kind, or whether the electronic automation of certain existing processes matters meaningfully. What is clear, however, is that in many cases new data collection and analysis technologies have been built on the back of existing ones (Agar, 2006; Armstrong, 2019; Stevens, 2017). This is certainly the case in baseball. The entry of baseball into the world of big data came only in the last decade or so, with the development of MLB Advanced Media's (MLBAM) data collection systems, including PITCHf/x and, more recently, Statcast. Using a combination of radar and video, Statcast now provides over 17 petabytes of data each season on the movement of every player and ball during all Major League games.[2] The new information, from tracing the trajectory of pitched balls to measuring launch angles and route efficiency, has indeed offered qualitatively different ways of understanding the game. For example, early measures of defensive skill included fielding percentage (essentially a ratio of successful plays to attempted plays), but even in the 19th century, commentators recognized that penalizing fielders for botched plays created a disincentive for them to make extraordinary effort to reach a difficult ball. This is one clear area where technology has entirely changed the way the game is conceptualized: now the analysis of video evidence enables a player's effort to field a ball with a given location, speed, and trajectory to be compared with other players at the same position in similar circumstances. Furthermore, some defensive positions are more difficult to play, and metrics like Defensive Runs Saved and Ultimate Zone Rating enable teams to decide between, for example, a poor-hitting first baseman who saves runs on defense and an average-hitting second baseman who hurts them in the field. Though defense remains difficult to quantify—rival companies offer competing metrics—there is no doubt that fielding is now measured as a relative contribution to a team's success rather than as an absolute record of mistakes made, a qualitatively different way of understanding the game.

On its surface, Statcast is the epitome of 'born digital' data revolutionizing a field. But dig into the history a little and it becomes clear how much of the system was built on existing infrastructures. MLBAM depends on a staff of dozens to accurately capture and clean the data, staff that—along with the statistical database itself—were initially acquired from Total Sports, a company with a number of old hands from the baseball data world. Cory Schwartz, the head of data operations for Statcast, was formerly a data collector for clubs and private companies in the 1990s. And the code underneath the system was originally developed by a group of volunteers for Project Scoresheet in the 1980s, one of the earliest play-by-play data collection initiatives (Phillips, 2019). The fundamental question for historians is how the old informs the new, that is, how the gradual scaling up and automation of previous data collection procedures might result in qualitatively different insights. This is also the question for historians looking at other areas of data science and machine learning that rely on long-standing statistical concepts but claim to have fundamentally novel conclusions.

Even as the Human Sciences Are Becoming Data Sciences, the Data Sciences Remain Inescapably Human Sciences

The collection and maintenance of baseball data clearly requires extensive human labor and a network of researchers. Experts remain essential to manage the collection of data, to ask the right questions of the data, and to make sense of the data. And the data themselves are increasingly harvested from, or donated by, people themselves. Or, as Rebecca Lemov has quipped, "Big data is people" (Lemov, 2016). At the same time, amateur scouting—the evaluation of nonprofessionals—supposedly remains the antithesis of analytics in its stubborn reliance on human judgments. Yet, amateur baseball scouts were using numerical grades for prospects' skills widely by the 1970s. Many clubs used a 'formula for judgment' to turn those grades into a single number—the 'overall future potential'—that would represent a prospect's total value, analogous to the simultaneous rise of credit scores in personal finance, or Framingham Risk Scores for coronary heart disease. By the 1980s, scouts were already complaining their job was too quantified, too scientific (Kerrane, 1984). In some respects, scouts were more audacious quantifiers than early sports analytics proponents because they didn't just analyze playing statistics but directly turned players into numbers. Even as human labor remains essential for data collection and analysis, so expert judgments across many different realms are reliant on data-driven calculations.

Though few would point to its players, managers, or executives as leaders in data science, baseball remains a realm in which millions of dollars are at stake every day in the use of data-driven decisions. Just as it has long been a resource for thinking about American culture, literature, and religion, we might also turn to the history of baseball for thinking more about the history of data in American life.

References

Agar, J. (2006). What difference did computers make? *Social Studies of Science, 36*, 869–907.

Armstrong, D. (2019). The social life of data points: Antecedents of digital technologies. *Social Studies of Science, 49*, 102–117.

Baumer, B., & Zimbalist, A. (2014). *The Sabermetric revolution: Assessing the growth of analytics in baseball*. Philadelphia: University of Pennsylvania Press.

Chadwick, E. (1842). *Report to Her Majesty's Principal Secretary of State for the Home Department, from the Poor Law Commissioners, on an Inquiry into the Sanitary Condition of the Labouring Population of Great Britain*. London, UK: Clowes and Sons.

Chadwick, H. (1874). *Base ball manual*. London, UK: Routledge and Sons.

Cramer, R. D. (2019). *When big data was small: My life in baseball analytics and drug development*. Lincoln: University of Nebraska Press.

Darrow, B. (2015, Sept 4). Live from Fenway Park, a behind-the-scenes look at MLB Statcast. *Fortune.com*. https://fortune.com/2015/09/04/mlb-statcast-data/

Daston L., & Galison, P. (2007). *Objectivity*. New York, NY: Zone Books.

Database. (2012) *Oxford English dictionary* (3d ed.). Oxford, UK: Oxford University Press.

Edwards, P. N. (2010). *A vast machine: Computer models, climate data, and the politics of global warming*. Cambridge, MA: MIT Press.

Evers, J. J., & Fullerton, H. S. (1910). *Touching second: The science of baseball*. Chicago, IL: Reilly and Britton.

Fuchs, A. H. (1998). Psychology and "The Babe." *Journal of the History of the Behavioral Sciences, 34*(2),153–165.

Gitelman, L. (Ed.). (2013). *"Raw data" is an oxymoron*. Cambridge, MA: MIT Press.

Haigh, T. (2009). How data got its base: Information storage software in the 1950s and 1960s. *IEEE Annals of the History of Computing, 31*(4), 6–25.

Kerrane, K. (1984). *Dollar sign on the muscle: The world of baseball scouting*. New York, NY: Beaufort Books.

Kitchin, R. (2014). *The data revolution: Big Data, open data, data infrastructures and their consequences*. London, UK: Sage.

Lemov, R. (2016, June 16). Big data is people. Retrieved from https://aeon.co/essays/why-big-data-is-actually-small-personal-and-very-human.

Lewis, M. (2003). *Moneyball: The art of winning an unfair game*. New York, NY: Norton.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York, NY: Crown.

Phillips, C. J. (2019). *Scouting and scoring: How we know what we know about baseball*. Princeton, NJ: Princeton University Press.

Schwarz, A. (2004). *The numbers game: Baseball's lifelong fascination with statistics*. New York, NY: Thomas Dunne.

Shapin, S. (2012). The sciences of subjectivity. *Social Studies of Science, 42*(2), 170–184.

Silver, N. (2012). *The signal and the noise: Why so many predictions fail—but some don't*. New York, NY: Penguin.

Stevens, H. (2017). A feeling for the algorithm: Working knowledge and big data in biology. *Osiris, 32*, 151–174.

Tukey, J. W. (1962). The future of data analysis. *Annals of Mathematical Statistics 33*(1), 1-67.

Zaillian, S., & Sorkin, A. (Producers), & Miller, B. (Director). (2011). *Moneyball* [Motion Picture]. United States: Columbia Pictures.

Footnotes

[1] For the role of the laboratory, see, for example, Fuchs (1998) and Evers and Fullerton (1910); for the use of performance records, Schwarz (2004). ↩

[2] This figure was initially provided for 2015 by Darrow (2015) and Major League Baseball's Matthew Gould (personal communication, August 19, 2019) confirmed the figure as still accurate for the 2019 season. ↩