PLAYING HARDBALL WITH
BIG DATA
HOW ANALYTICS IS CHANGING
THE WORLD OF SPORTS

by Doug Hanchett
EMC+
Long before there was Big Data, there was Bill James.

Back in the mid-1970s, prior to the advent of the PC, James began crunching baseball box scores with pencil and paper, testing theories and evaluating the game’s long-held axioms. He didn’t set about trying to revolutionize how sports are analyzed. But he did.

“I never really anticipated that it would become something more than a niche thing,” James, author of the annual “Baseball Abstract” in the 1970s and 80s, tells EMC+. “I always thought there was a little market for what I was doing and there would be a little notch I could fill. It turned out that it was bigger than I expected it to be.”

Now, 35 years later, there’s a whole industry that’s sprung from the Kansas native’s pioneering work—an industry that, in some ways, has served as a precursor to today’s Big Data revolution.

“It’s kind of like Tommy John surgery,” says James, who was a night watchman at a bean cannery in Kansas when he began writing his first Abstract. “Sports demonstrated that it could be done, and now it’s done routinely for many, many people who don’t have anything to do with sports. Sports analysis has been an example for the rest of society.”

Indeed. Big Data—in sports as in all of commerce—has become big business, with companies racing to compile, analyze, and extract heretofore hidden value from the mountains of data they have. And the sports world is one of the few places where the public can easily observe Big Data in action—or at least the analysis techniques that are becoming commonplace throughout the business world.

**PREDICTIVE ANALYTICS**

Take, for instance, John Dewan, one of James’ disciples and now a close colleague. Dewan, a former insurance actuary who founded STATS Inc. and now runs Baseball Info Solutions, recently won praise for his spot-on Olympic medal predictions for the Wall Street Journal. His analysis—which involved assigning probabilities of success for each competitor and then using those probabilities to run 1,000 computer simulations of each event—determined the United States would win the most medals with 108, including 40 golds.

The U.S. finished with 104 medals, 46 gold.

Dewan also projected Russia to win 83 medals and Great Britain to win 66. The two countries won 82 and 65, respectively.

“We like the system,” he says, sounding like the former insurance actuary that he is. “We were pretty much right on.”

While Dewan is quick to point out that his Olympics projections hardly qualify as Big Data in terms of the size of the data sets he uses, it’s clear that the discipline of sports analytics is heading in that direction—much like the rest of the world.

“Sports is just now becoming biggish data, and up until recently it wasn’t Big Data where you required heavy hardware and scalable approaches,” says Kevin Goodfellow, who founded the sports
analytics firm **Sports Data Hub** seven years ago. “It was something you could handle on relatively normal-sized databases.”

That’s changing. Quickly.

“Those volumes [of data]—compared to Facebook, it’s a drop in the bucket” Goodfellow tells EMC+. “But even though the problems haven’t grown to the size of the super huge, they will very quickly, and that trend is just going to accelerate.”

Ben Alamar, a statistics and economics professor at Menlo College in Atherton, California, agrees. Alamar spent the last five year doing statistical analysis for the NBA’s Oklahoma City Thunder and has done some work with the NFL’s San Francisco 49ers as well.

“Right now, a big data set in sports is like a day’s work at Google,” he says. “It’s not really Big Data in the sense that most companies think of Big Data. But if Big Data is less about data size than having new data that’s never been mined before and using the insights from that data to gain a competitive advantage, then yeah, that’s exactly what we’re doing.”

**BEYOND BASEBALL**

The explosion in sports analysis is everywhere. Baseball always had a strong statistical bent, thanks in large part to the game being a series of quantifiable events carried out by individuals—pitcher throws ball, hitter hits ball, fielder fields ball. So it made sense that would be a natural starting point for James. But now analytics is being done in more chaotic sports like hockey and football.

Even soccer, which is almost devoid of meaningful statistics—ever read a soccer box score?—is getting into the act. **Next year Major League Soccer is going to place Adidas miCoach data trackers** into team jerseys, which will allow coaches to monitor things like a player’s heart rate, distance run, top speed.

Many in the industry are curious to see how that will turn out.

“Soccer is such a complex game because the culminating event—the goal—there’s so few of them,” Alamar tells EMC+. “It’s much more complex than looking at the pitcher/batter and figuring out the outcomes there.”

Whether MLS teams derive anything productive from the data remains to be seen. But the general consensus is that having more data to analyze can’t hurt, especially if one has the right tools.

“In some industries, like the insurance industry, using huge amounts of data has always been there,” Dewan tells EMC+. “It’s vital to the insurance industry and that’s why there is the actuarial profession. But what happened in baseball—and Bill James led this—was that when we started accumulating all this data, there became a realization that, wow, you can supplement your on-the-field visual scouting techniques with hardcore data. And if you combine the two, you’re going to make better decisions than if you just use your scouting information.”
That was the message of Michael Lewis’ best-selling book, “Moneyball,” about the Oakland A’s, which became a surprise Hollywood hit in 2011. And ultimately it’s the promise of Big Data.

After overcoming some objections at first, the slow-moving sports world is buying into it, just like everybody else.

TECHNOLOGY THE KEY

James began self-publishing his Baseball Abstracts in 1977, just a few years before the boom of the PC. “The early Abstracts, I literally just added stuff up. I had no computer,” says James. “People didn’t have personal computers yet.”

Then the PC came along, making it easier for James and other like-minded sabermetricians to conduct their analyses. “Just after I started doing them, the technology changed very rapidly,” says James. “So I was doing them at exactly the right time.”

It was Big Data at a micro level: Baseball stat geeks using the earliest desktop computers to process reams of raw statistics they had imported themselves, and then writing basic programs to test theories and elicit useful information.

“It could not have been done without the computer,” says Dewan. “It’s too much to do without the computer. It is no coincidence that the explosion of baseball information and the explosion of computers were going together.”

Though viewed as a trailblazer, James says that a lot of what he did was simply look at baseball from a different angle.

“I don’t have the skills of an actuary, but we borrowed an awful lot of methods from economics,” he says. “Most of what I did do, actually, was just economics applied to baseball—the economic way of thinking: the on-field economy of runs and hits and wins and losses. We borrowed a lot from other fields. I think probably business has borrowed some from us, but we’re still greatly in their debt.”

Perhaps. But in an afterword to the 2011 edition of “Moneyball,” author Lewis writes that after his book was first published in 2003, the Oakland A’s front office was flooded with calls “from a cross section of American business and sporting life” looking to emulate the team’s success at identifying undervalued assets and doing more with less.

Today sports analytics often goes beyond what happens on the court, the ice, the diamond, or the gridiron. It involves merging the traditional box score stuff with other data to determine what an athlete means to the company’s bottom line—and how that can be maximized.

KEVIN GOODFELLOW

“We look at a player’s value in terms of not only what they do on the field … but also how many more ticket sales can they generate, how many more hot dogs are they going to sell, how many more beers are fans going to buy, how many more jerseys.”

KEVIN GOODFELLOW
“We look at a player’s value in terms of not only what they do on the field—and that has a value in terms of wins and losses, and those have dollar values as well; every win has a dollar value—but also how many more ticket sales can they generate, how many more hot dogs are they going to sell, how many more beers are fans going to buy, how many more jerseys, that type of thing,” says Goodfellow.

Is there a danger that turning sporting events into spreadsheets will change how the casual fan views them—or alter the very nature of the games themselves?

Goodfellow doesn’t think so.

“Football is still football and hockey is still hockey and there’s still such a big component of it that’s player-driven—physiologically and psychologically—and that’s really hard to wrangle and analyze,” he says. “There’s still plenty of the game that’s driven by human nature.”

As for James, his contributions to baseball have finally been recognized, if not grudgingly accepted. A few years ago Dewan’s publishing company produced a book entitled “How Bill James Changed Our View of Baseball” featuring essays from scores of admirers.

No longer the outsider, James now works for the Boston Red Sox. But years of crunching numbers—and turning that into a full-time job—hasn’t diminished his love of the game.

“I still enjoy baseball very much,” he says. “The nature of baseball is the more you know about it and the more you understand about what is happening, the more you enjoy it.”

“If Big Data is less about data size than having new data that’s never been mined before and using the insights from that data to gain a competitive advantage, then yeah, that’s exactly what we’re doing.”

BEN ALAMAR