Take me out to Big Data: Analyzing professional baseball data online with SciServer

*“You can observe a lot by just watching.” -Yogi Berra*

For millions of people, following professional and amateur sports provides their first exposure to statistics. Even sports fans with no formal education at all routinely engage in sophisticated statistical thinking. That fact means that sports information provides an excellent opportunity to teach data science to a large, motivated population of students and adult learners.

Of the sports that U.S. fans regularly follow, none has data-driven analysis embedded more deeply than baseball. Over the past 15 years, the “moneyball era” has seen managers and scouts bring advanced statistical techniques to bear on baseball strategy, and fans have kept up so that terms like OPS, WHIP, and WAR are part of fans’ everyday vocabulary.

In today’s era of advanced statistical knowledge, Major League Baseball teams have access to incredibly detailed reports on players and teams, along with sophisticated proprietary data analysis tools. But such access is far beyond the average fan, or even the average data scientist. This disconnect highlights the need to make baseball data available through free, easy-to-use online tools that can handle large, complex datasets – tools like those offered by IDIES’s own SciServer online science environment.

This poster describes an ongoing effort to load historical baseball data into the SciServer environment. All data comes from the Retrosheet project at [www.retrosheet.org](http://www.retrosheet.org). We have loaded two types of data: game data, in which a row describes a complete Major League Baseball game, and event data, in which a row describes a single event. An event is defined as an at-bat or similar self-contained game situation, such as a stolen base or substitution. Complete game data is available back to 1871; event data is complete from 1974 with some events available as early as 1921. This publicly-available baseball data, along with Python scripts to analyze and visualize it, offers incredible potential for future in-depth data science education.