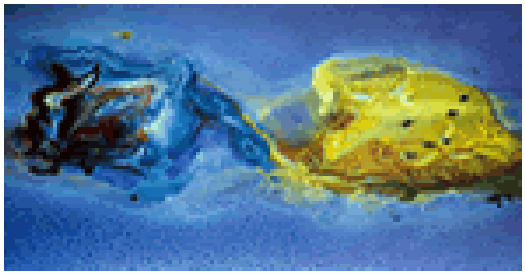

Contribution ID: **73** Type: **Talk**

Big Data Software in High Energy Physics

Thursday, 2 August 2018 17:50 (30 minutes)

For decades, high-energy physics (HEP) had been on the forefront of big data technology, developing techniques to explore and analyze datasets too large for memory that were revolutionary when they appeared in other fields years later. Today, that dominance is ending, and I argue that it's a good thing. The rise of web-scale datasets and high-frequency trading has interested the commercial sector in data analysis, driving the development of professional yet open-source software with a much larger userbase than HEP— software that we do not need to develop or maintain ourselves.

However, using this software in HEP analysis isn't trivial, at least not yet. Some differences in conventions have to be bridged, such as HEP's C++ toolset and the preponderance of Python, R, and Java/Scala in industry. I will show some of this "plumbing" software for Python (PyROOT and uproot) and Java/Scala (Spark-ROOT). But there are also deeper differences in emphasis between the two communities: our nested data model vs. flat data frames, our focus on histograms and basic plotting, and the industry's satisfaction with merely predictive models. After showing illustrative examples and how to use them, I will conclude that we still have work to do, developing some software on our own, but can significantly benefit by working within the conventions of the larger big data community.

Primary author: PIVARSKI, Jim (Princeton University)

Presenter: PIVARSKI, Jim (Princeton University)

Session Classification: Statistical Methods for Physics Analysis in the XXI Century

Track Classification: H. Statistical Methods for Physics Analysis in the XXI Century