

Columnar Data Analysis

Wednesday, 24 July 2019 11:00 (1h 30m)

Data analysis languages, such as Numpy, MATLAB, R, IDL, and ADL, are typically interactive with an array-at-a-time interface. Instead of performing an entire analysis in a single loop, each step in the calculation is a separate pass, letting the user inspect distributions each step of the way.

Unfortunately, these languages are limited to primitive data types: mostly numbers and booleans. Variable-length and nested data structures, such as different numbers of particles per event, don't fit this model. Fortunately, the model can be extended.

This tutorial will introduce awkward-array, the concepts of columnar data structures, and how to use them in data analysis, such as computing combinatorics (quantities depending on combinations of particles) without any for loops.

Presenter: PIVARSKI, Jim (Princeton University)