

AlphaTwirl: a python library for summarizing event data into multi-dimensional categorical data

Tuesday, July 10, 2018 4:40 PM (20 minutes)

AlphaTwirl is a python library that loops over event data and summarizes them into multi-dimensional categorical (binned) data as data frames. Event data, input to AlphaTwirl, are data with one entry (or row) for one event: for example, data in ROOT TTree with one entry per collision event of an LHC experiment. Event data are often large – too large to be loaded in memory – because they have as many entries as events. Multi-dimensional categorical data, the output of AlphaTwirl, have one row for one category. They are usually small – small enough to be loaded in memory – because they only have as many rows as categories. Users can, for example, import them as data frames into R and pandas, which usually load all data in memory, and can perform categorical data analyses with a rich set of data operations available in R and pandas. In this presentation, I will show (a) an example workflow of data analysis using AlphaTwirl and data frames, (b) the user interface of AlphaTwirl, e.g., how to specify conditions of event selection, binning and categories, and methods to summarize data in each category, and (c) features of implementation, such as concurrency in looping over large event data. In addition, I will mention particular analyses in CMS using AlphaTwirl. I will also discuss possibilities for future development.

Primary author: Dr SAKUMA, Tai (University of Bristol (GB))

Presenter: Dr SAKUMA, Tai (University of Bristol (GB))

Session Classification: Posters

Track Classification: Track 2 –Offline computing