C) Applying Data Analytics and Machine Learning in HEP

Panel: Graham Mackintosh (IBM, remote), Louis Capps (NVIDIA), S. Gleyzer (U.Florida, CMS), Michael Williams (MIT, LHCb), D.Whiteson (UC-Irvine, Atlas)
Moderator: Jim Pivarski (Princeton)

Panel questions/themes:

1. From a data analytics perspective, how do you see the capabilities in this field evolving over the next five years?

2. What new techniques are in development that we should be looking at?

3. What scale and types of problems do you see that are impossible today that will be solved over the next 5-10 years?

4. How should we be approaching bringing machine learning applications to analysis? Object reconstruction? Triggering?

5. Most machine learning techniques apply to flat tables of numbers (points in a vector space), but HEP data must be flattened to put it in that format: e.g. we have events of jets of particles. Are there interesting machine learning techniques that could make use of non-flattened data?

Live minutes:

- G. Mackintosh: extension of machine learning usage in HEP, beyond classification, requires going beyond the black box approach.
- S. Gleyzer: some difficult problems for ML. One of them is real time, time budget constraints. Another difficulty is theoretical foundations of ML: improving and expecting further improvements in the next decade.
- Daniel Whiteson: ML can tackle higher dimension problems. Dimensionality bounded by the computing power. The future is probably not in replacing expert heuristics by ML but rather augmenting it: current best successes have been combining both.
- S. Gleyzer: selecting the appropriate/optimal algorithm is not a simple thing. Choice can turn out to be wrong when adding systematics. Some new approaches emerging.
- ???: ML is currently driven by challenging use cases. One of them is fully automated cars: not so different from the challenges we have in HEP. Just starting to look at 3D objects: currently in video/image analysis, only 2D images (+ depth information).
- Daniel: existing ML tools may not be appropriate to our use cases (in particular those focused on classification). Need to be ready to build our own tools from the ML ecosystem (e.g. discovery significance instead of classification accuracy).

Should we concentrate on Deep Learning?

- S. Gleyzer: certainly we should not just focus just on it, ML field is evolving rapidly, algorithms popular today are not those that existed 10 years ago, will probably the same in 10 years.
- G. Mackintosh: the main difficulty with deep learning is the difficulty to understand what happens under the hood.

Impact of ML on computing infrastructure/architecture

- ???: currently running ML on the existing computing architectures, like HPC or GPUs. But this may change in the future. ML has some specific features/requirements, like the iterative reading of the same data a large number of times or where/how to do the training.
- S. Gleyzer: the new algorithms require a lot of training (thus compute resources) based on a large volume of data. So far we were able to piggyback in experiment needs/resources but this may change.

ML challenges

- ???: training is a big challenge. For some use cases, not enough data available: works going on to see if generated data can be used to train a NN. Not necessarily relevant for HEP where there is a lot of data available.