The Machine Learning Landscape

Steven Schramm

HEP analysis ecosystem workshop

Amsterdam, The Netherlands

May 22, 2017

Introduction

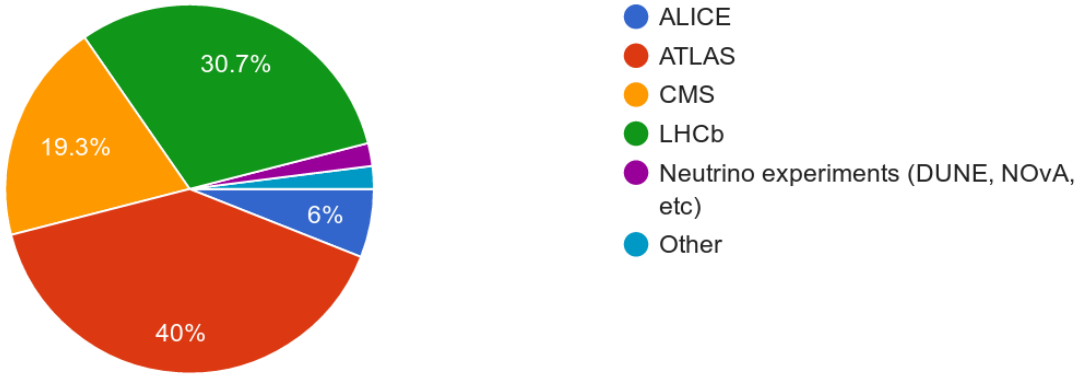
UNIVERSITÉ DE GENÈVE

- Machine Learning (ML) is a field of growing interest in HEP
- This is likely to continue as datasets enter the $100\,\mathrm{fb}^{-1}$ regime
  - ML is aimed at extracting information from large datasets
- The HEP-ML community is undergoing a transition
  - A few years back, TMVA dominated the HEP landscape
  - Now, other tools are increasingly used
  - We will explore the motivations for this large-scale migration

HEP-ML survey

UNIVERSITÉ DE GENÈVE

- I sent out a survey to primarily the LHC ML community
  - LPCC IML, ALICE ML, ATLAS ML, CMS ML, and LHCb ML lists
- Received exactly 150 replies as of May 21 (morning)
  - Mostly LHC experiments, 3 neutrino, 3 other [CERN IT, Belle II, FCAL]

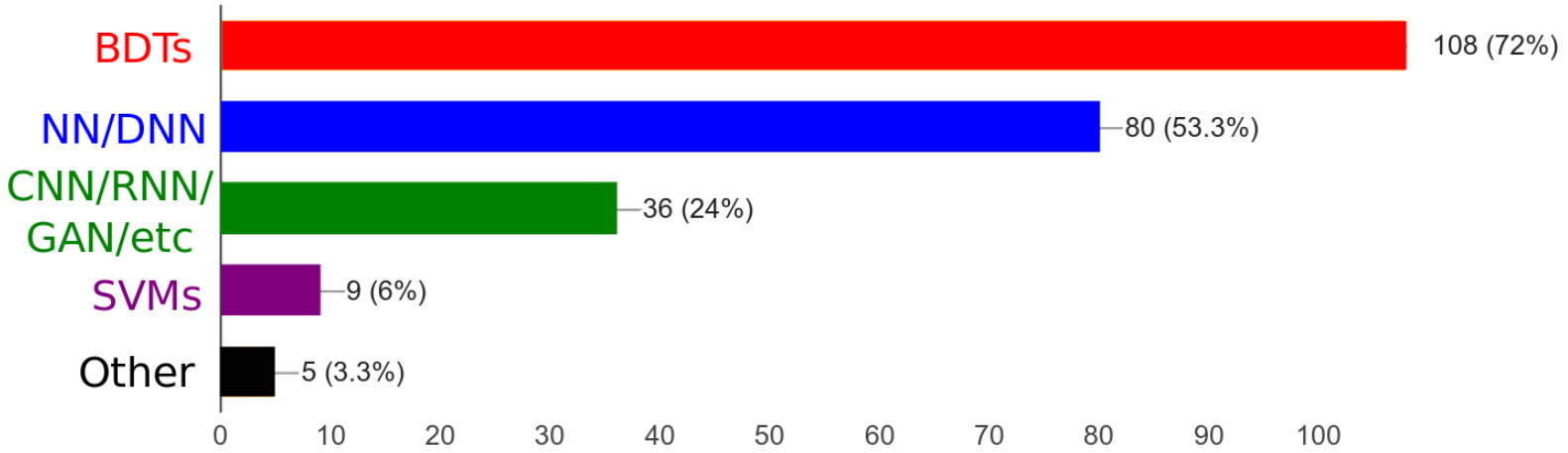


- Note: this survey will be biased given its distribution
  - Missing people who use ML but are not part of collaboration ML lists
  - These users are likely to have different use cases, will mention later

ML techniques being used

UNIVERSITÉ DE GENÈVE

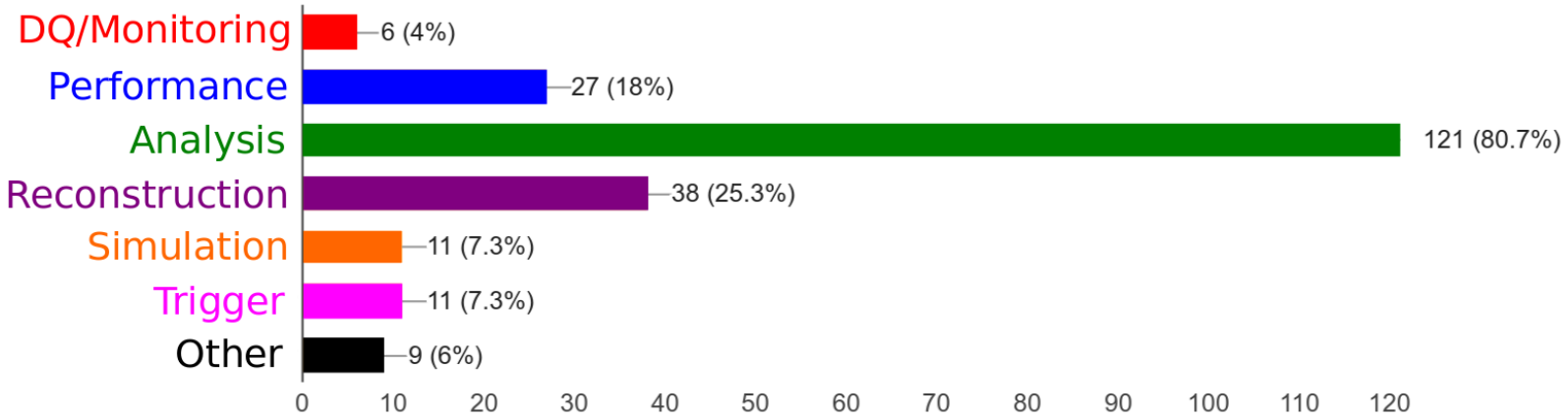
- HEP-ML users are still using BDTs more than other techniques
- However, neural networks are essentially on par (when grouped)
  - This is part of the motivation for the move away from TMVA
  - There are lots of excellent DNN toolkits from the ML community
  - There are DNNs in TMVA, but they are not widely used
- Other techniques remain, but are much less common



What is ML being used for?

UNIVERSITÉ DE GENÈVE

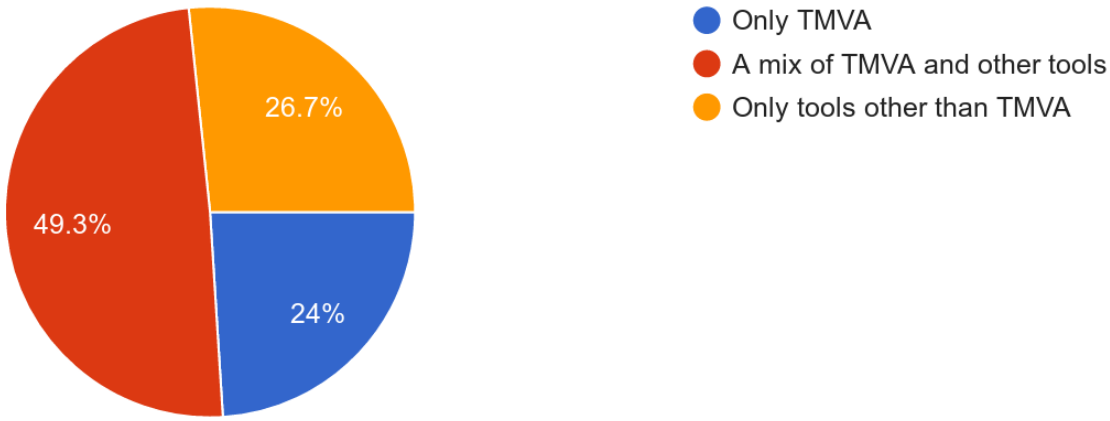
- The primary usage of ML remains physics analysis
  - Followed by reconstruction and performance studies
- Other: computing (x6), storage autoencoders, business application



What types of ML tools are being used?

UNIVERSITÉ DE GENÈVE

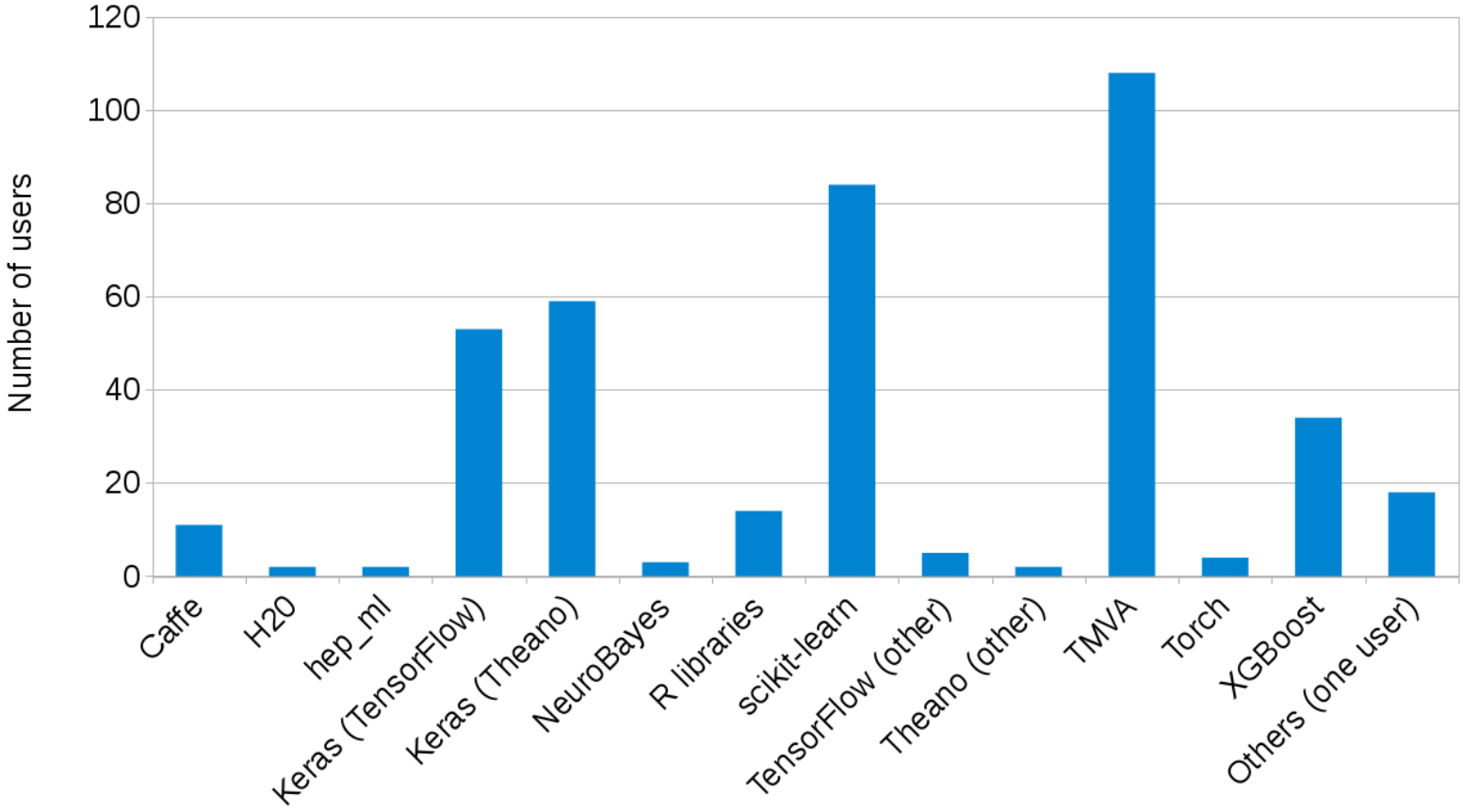
- From this survey, 1/4 of the community uses only TMVA
  - However, recall that this was sent to ML-specific mailing lists
  - Most likely many others are using TMVA in their analysis
  - Not all of these users will follow ML developments
- Of those who follow ML developments:
  - 1/4 have used only external tools from the ML community
  - 1/2 have used both TMVA and external tools



Tool usage by numbers

UNIVERSITÉ DE GENÈVE

- The ML community has put a lot of effort into developing great tools
- HEP is increasingly making use of these external tools



- Now that we know what people are using, let's discuss why

The TMVA usage paradigm

UNIVERSITÉ DE GENÈVE

- TMVA is a toolkit attached directly to ROOT
  - Native support for the ROOT data format
  - Available by default on lxplus/etc
- For these reasons, TMVA is remarkably easy to use for HEP experts
- It is the traditional entry point to ML for HEP experts
  - Easy to try out a BDT/similar in a cut-based analysis
- TMVA has also been around in the HEP toolkit for a long time
  - Many long-standing efforts use TMVA for that reason
  - Flavour tagging is such an example, although that is changing

TMVA limitations

UNIVERSITÉ DE GENÈVE

- While TMVA has a strong background in HEP, it is not perfect
- Most common survey complaints:
  - The desired ML technique is not available
  - It's too slow (CPU efficiency)
  - Cannot efficiently use GPUs
  - The interface is not as easy to use as that of external tools
  - Lack of up-to-date documentation
  - Missing the latest useful ML features
- Recall that most of the respondents are advanced ML users
- It's not clear if all of the feedback is up-to-date
  - TMVA has made quite a few changes lately
  - People may be commenting on experience from before these changes

TMVA potential improvements

UNIVERSITÉ DE GENÈVE

- When asked how TMVA could be improved, suggestions include:
  - Continue to add interfaces to external tools
  - Update the documentation
- However, the single most common suggestion (by far)...
  - Stop major TMVA developments
  - Focus on integrating ROOT with external tools
  - Provide ROOT data format converters and glue packages
- Only one person suggested that TMVA should add new methods
- I will come back to this discussion later

The external tool usage paradigm

UNIVERSITÉ DE GENÈVE

- ML tools typically have a higher usage barrier
  - No direct support for the ROOT data format
  - Tools are not always available by default on lxplus
- However, an increasing number of users are making the switch
  - Most of these those who have switched are playing with deep learning
- Benefits of external tools are numerous, including:
  - Easy to use interfaces
  - Strong community support and documentation
  - Frequent updates to include the latest ML developments
  - Optimized CPU usage and automatic GPU support
- A bit over half of external tool users actually use multiple tools

External tool sample workflow

UNIVERSITÉ DE GENÈVE

- Due to ROOT data formats and the need to apply classifiers within existing restrictive software frameworks, the common workflow is:
  - Pre-process datasets aimed at a given study
  - Convert ROOT format to another format (csv, HDF5, etc)
    - Glue packages are very useful here (root_numpy, root_pandas, etc)
  - Train the classifier with external tools, typically not in C++
  - Convert the resulting model to a form the framework can handle
    - Typically this means being able to implement the classifier in C++
    - Packages like lwtnn are very useful here
- When used in analysis, the last point can sometimes be omitted

External tool limitations

UNIVERSITÉ DE GENÈVE

- Most external tool complaints were about the HEP interface
  - Interfacing with ROOT is clumsy
  - Difficult to re-integrate an externally trained classifier in HEP software
  - Non-trivial installation of the tools in the HEP environment (lxplus)
- A few complaints were about external tool specific constraints
  - External tools may require learning a new programming language
  - Analysis responsibles are suspicious of non-TMVA ML tools
  - Some features important to HEP users aren't present in external tools
    - Uncertainty estimates, negative event weights, etc

External tool potential improvements

UNIVERSITÉ DE GENÈVE

- Most of the suggested improvements mirror the comments on TMVA:
  - Improve ROOT interfaces/bindings to other common languages
  - Provide ROOT data format converters (csv, HDF5, etc)
    - From other replies, looks like many of these have been privately written
    - Perhaps these should be centralized and supported
  - Install the most popular external tools by default on lxplus
  - Make ROOT installable with pip
  - Contribute to external tools for features important to HEP users
    - Again, comment is typically paired with "instead of expanding TMVA"

CPUs and GPUs

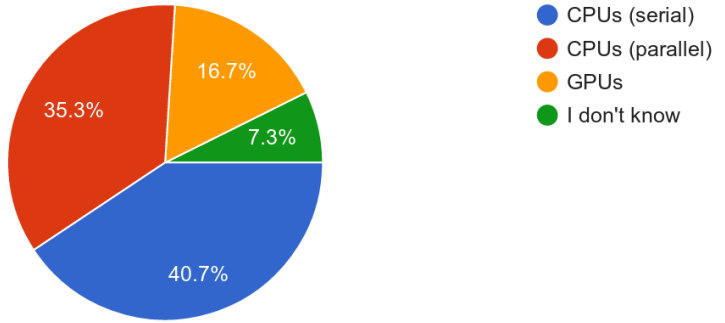
UNIVERSITÉ DE GENÈVE

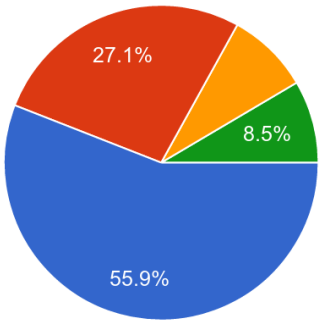
- CPUs are great, but GPUs dominate in the land of Deep Learning
  - Typical DNNs use enormous number of floating point operations
  - For ML purposes, single-point precision is more than sufficient
    - Cheap GPUs are better than expensive CPUs
    - Expensive DL-oriented GPUs are of course still the best

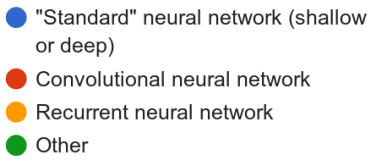
What do you train with? →

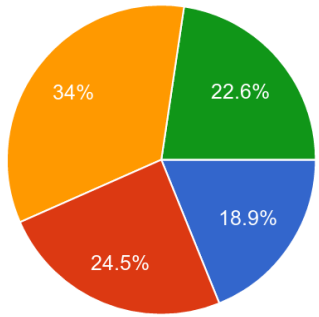
If GPU, what are you training? ↙

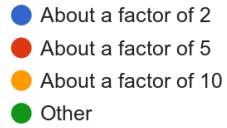
CPU/GPU training time? ↘











The impact of deep learning (DL)

UNIVERSITÉ DE GENÈVE

- Several of the previous comments are correlated with deep learning
  - Increasingly prevalent approach to HEP-ML problem solving
  - As mentioned, already ∼level with BDT usage
- GPUs are becoming critical to remain at the forefront of HEP-ML
  - Several people asked for centrally available GPUs
- Many great external DL tools, thus more people leaving TMVA
  - Growing number of ML users who have never used TMVA
  - For people in that group, most are working on DL
- Personal comment: some of this is hype, some is real performance
  - In many cases, other methods have similar performance to DNNs
  - In some cases, DNNs really do bring enormous benefits
  - Regardless, DNNs are leading the push towards external tools

TMVA vs external tools

UNIVERSITÉ DE GENÈVE

- This is a delicate topic, and the survey results are biased
- My interpretation:
  - Users in analysis who just want to improve their cut based analysis like TMVA for its ease of use (pre-installed and direct ROOT integration)
  - Users following ML developments want to drop TMVA entirely and focus on directly supporting/integrating with external packages
- Personal view (many may disagree on both sides):
  - There is a use case for TMVA, but that should remain simple
    - It is worth investing effort into improved documentation
    - It is worth continuing to improve memory/CPU usage
    - It is not worth the investment to continue chasing external ML tools
  - When we need the best ML, we should stick to external tools
    - Focus effort on ROOT converters/interfaces/glue packages
    - Most popular external tools should be provided as part of LCG releases
    - I think DL will continue to grow; it's worth investing in common GPUs

Other survey comments of note

UNIVERSITÉ DE GENÈVE

- Funding agencies need to learn that software support is as important to HEP as eg: detector responsibilities. We need long-term positions in HEP for skilled software engineers if we are to get back to being at the cutting edge of computing and if we are to make the most of our data.
- I would like to see more generic courses on ML, geared to people from the collaborations / CERN in general. When I say generic I mean not necessarily linked to a well-known application, because that localizes the way in which we use ML at the moment - but rather extend it and show techniques that may be of use to new projects, and increase the general knowledge on the subject from people not coming from a CS background.
- The current data processing pipeline seems ill-adjusted for application of DNNs. E.g. if one wanted to apply a DNN classifier on a large sample - it is currently possible to integrate this in the processing framework - however the classifier will be applied 'event-by-event' serially. Most likely if the deep learning applications do 'take off' in HEP this mode of work will become a bottleneck. In my opinion one should think of a heterogeneous computing model where portion of event data to be used in a deep learning application is extracted, sent in a batch to a GPU cluster and finally the results 'merged' with appropriate events. In this way the operations could be parallellized on batches of events not only 'within' an event.

Summary

UNIVERSITÉ DE GENÈVE

- The HEP-ML landscape is currently undergoing a substantial change
  - A few years ago, TMVA dominated the market
  - Now, external tools are increasingly replacing TMVA
- Deep learning is increasingly prevalent in HEP-ML
  - Usage is growing fast and is now ∼even with BDTs
  - This is a large part of the shift towards external tools
  - This is also creating a need for GPUs
- Dominant community request: ROOT integration with external tools

- ML is still primarily used for physics analysis
  - Reconstruction/performance use is growing, but has a long way to go
- ML is only going to become more important as the dataset grows
  - It is important to start preparations for this now