

# **Machine Learning Proposal**

# HPS Machine Learning

- Our goal is to improve upon HPS trident classification (and potentially other things) using machine learning algorithms.
- This is to be done through the WEKA software package. (<http://www.cs.waikato.ac.nz/ml/weka/>)
  - WEKA is a Java-based machine learning workbench.
  - It supports a wide variety of machine learning algorithms, such as decisions trees, probabilistic modelling, and clustering.
  - Numerous algorithms in each type are supported – WEKA is an interface, and allows for the same input to be easily run through multiple algorithms.
    - It is also possible to call algorithms directly from Java code.
  - WEKA is created and maintained by the machine learning group at University of Waikato in New Zealand and is freely available under the GNU public license.

# HPS Machine Learning

- WEKA takes as input a set of instances representing a data object, which is itself composed of attributes.
  - For trident analysis, an instance takes the form of a single particle/particle pair.
  - Its attributes are traits of the pair, such as the time difference, energy sum, charge of the particles, *et cetera*.
- The intent is to first generate a set of trident data from Monte Carlo with truth information to indicate whether the pair actually represents a trident.

# HPS Machine Learning

- As a first approach, a decision tree algorithm will be used. (WEKA supports several, and allows for a great many more to be installed).
- Decision trees offer a few advantages:
  - They are simple to implement: It is much easier to port a decision tree into the HPS-Java framework than, for instance, a massive sparse matrix.
  - They are relatively human-readable: What properties the tree prioritizes and valuable metrics for differentiation could provide insight to the data.

# HPS Machine Learning

- However, the machinery needed to approach this is not yet available.
  - There is no calorimeter truth information available at all. This is lost during the digitization phase of readout.
  - It appears that track truth information is not present either.
- It is important to improve the HPS readout and recon behavior to include this truth information.
- This requires modification to the current HPS system.
  - This primarily affects readout, which must retain which truth objects went into the creation of a simulated detector output object.
  - Recon must then know to look for these relations and, if they are found, incorporate them into recon objects.
  - This consists of using SimCalorimeterHit and SimTrackerHit objects instead of CalorimeterHit and TrackerHit objects.