Machine Learning Online Monitoring for the SpinQuest experiment at Fermilab

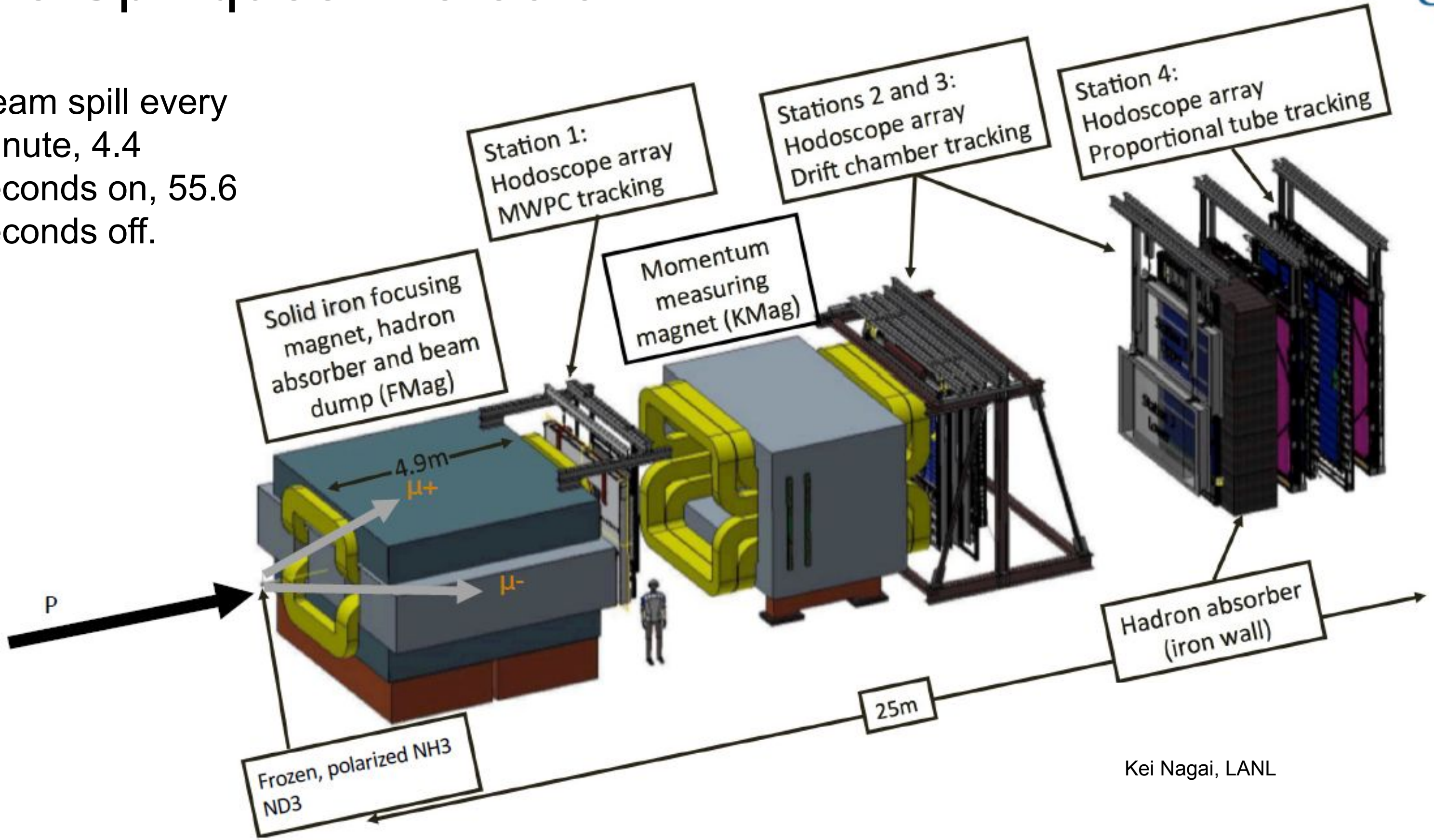
Arthur Conover and Dustin Keller

This work is supported by Department of Energy contract DE-FG02-96ER40950

SOLID POLARIZED TARGET GROUP at the UNIVERSITY of VIRGINIA

The Spinquest Detector

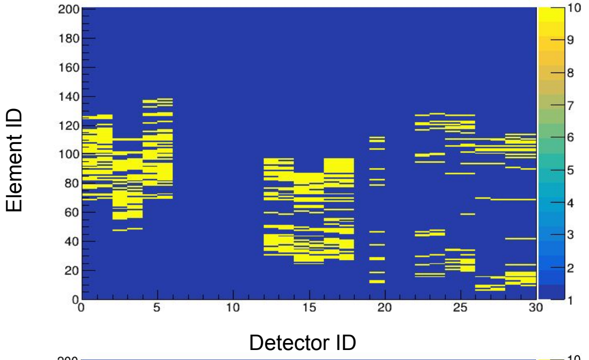
Beam spill every minute, 4.4 seconds on, 55.6 seconds off.

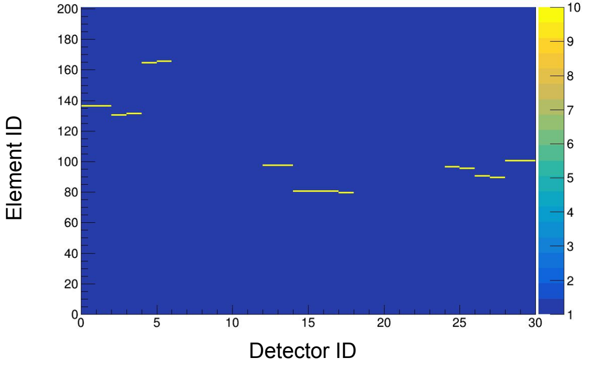


Kei Nagai, LANL

Raw Data

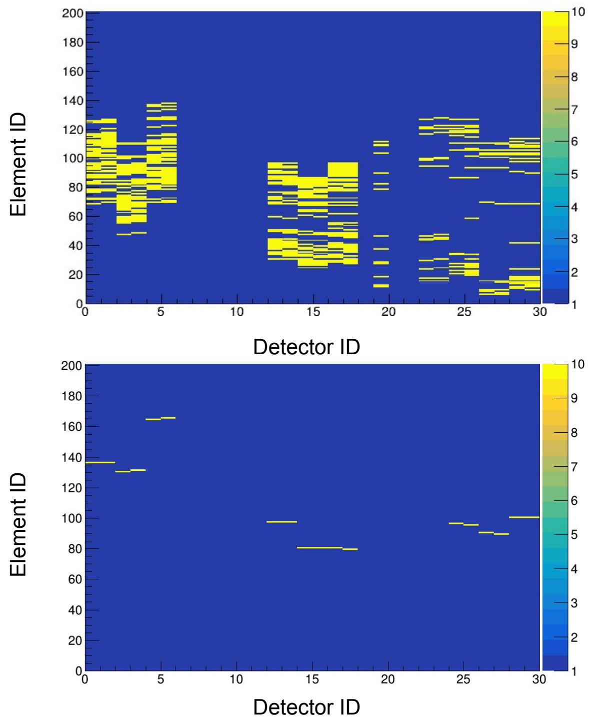
- Each detector hit outputs 2 or 3 values:
    - Detector ID
    - Element ID
    - Drift Time (proportional tubes and drift chambers)
- Each spill has 30,000-50,000 events, each with about 500 hits.
- This gives us 15-25 million hits to sort through each spill.



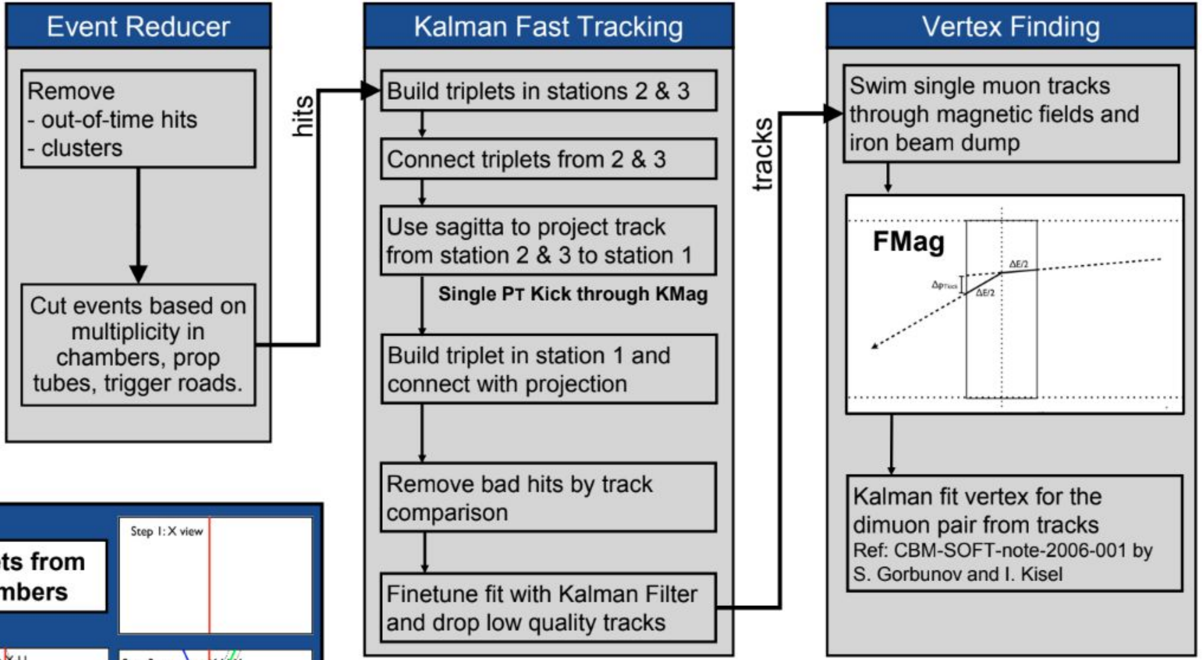


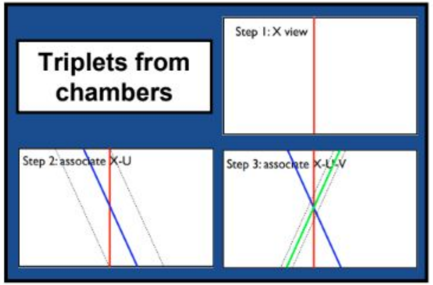
Challenges for SpinQuest Tracking

- Data is extremely noisy.
  - Approximately 1 good dimuon event in every 10 events.
  - Around 30 physics events for every 50,000 noise events.
  - Approximately 30 'tracklets' per plane per event.
- The process that we're interested in (Drell-Yan and J/Psi) are very close in mass, which makes them difficult to separate.
- Final results very sensitive to any asymmetries caused by external factors, so online monitoring needs to be precise to detect false asymmetries.



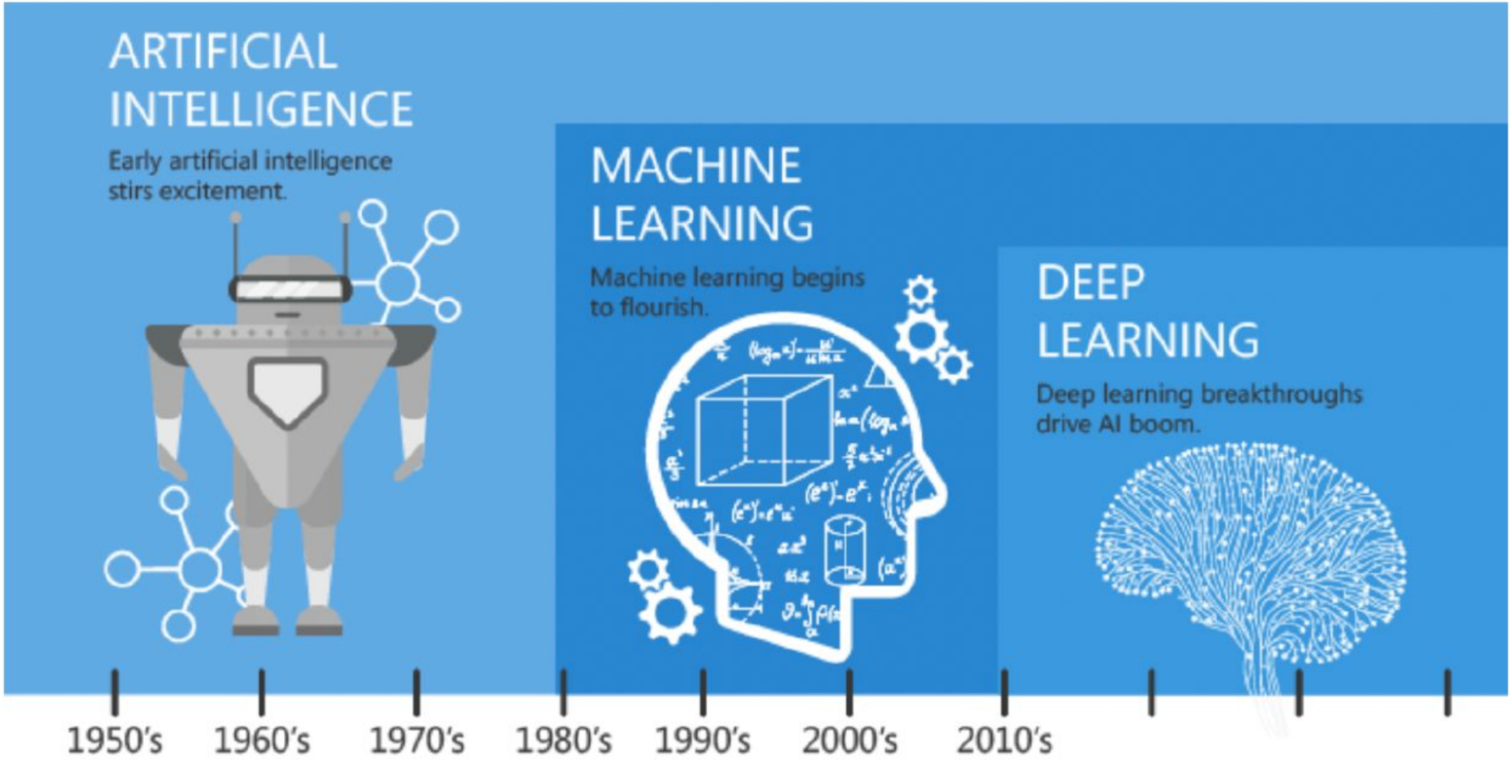
K-Tracker



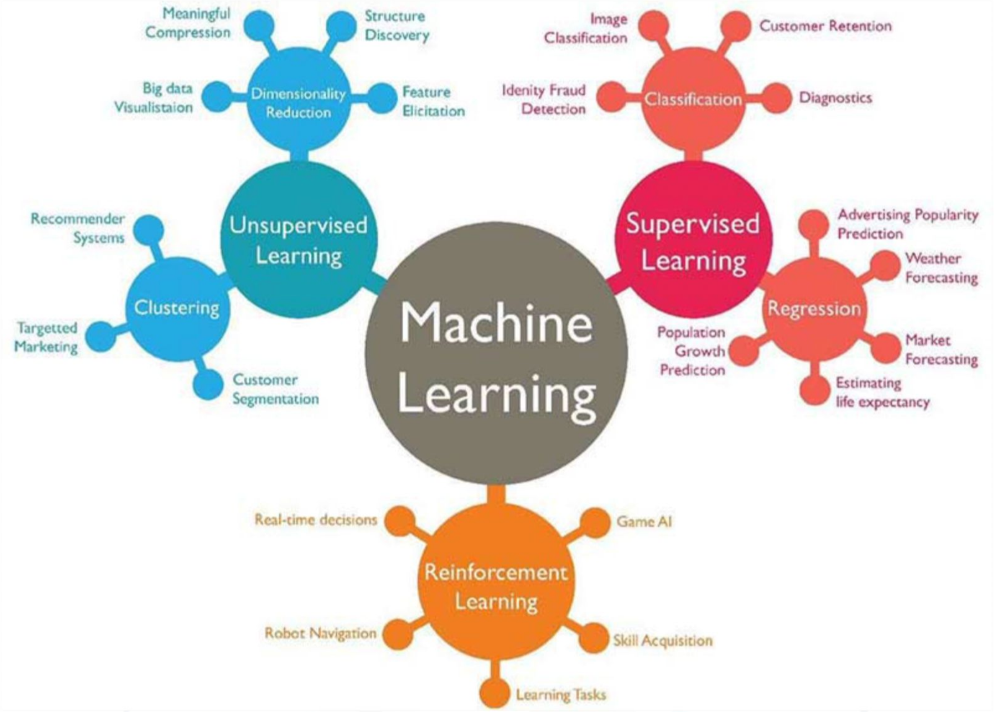


Noah Wurfel, University of Michigan

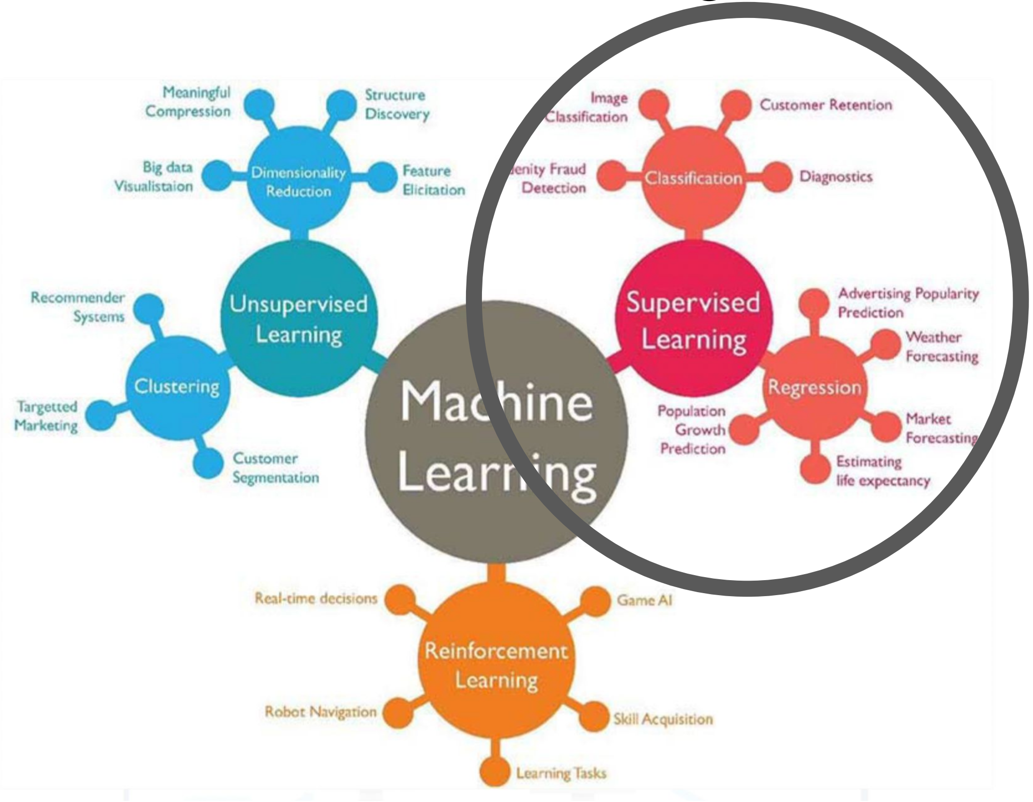
Another Possible Solution: Machine Learning



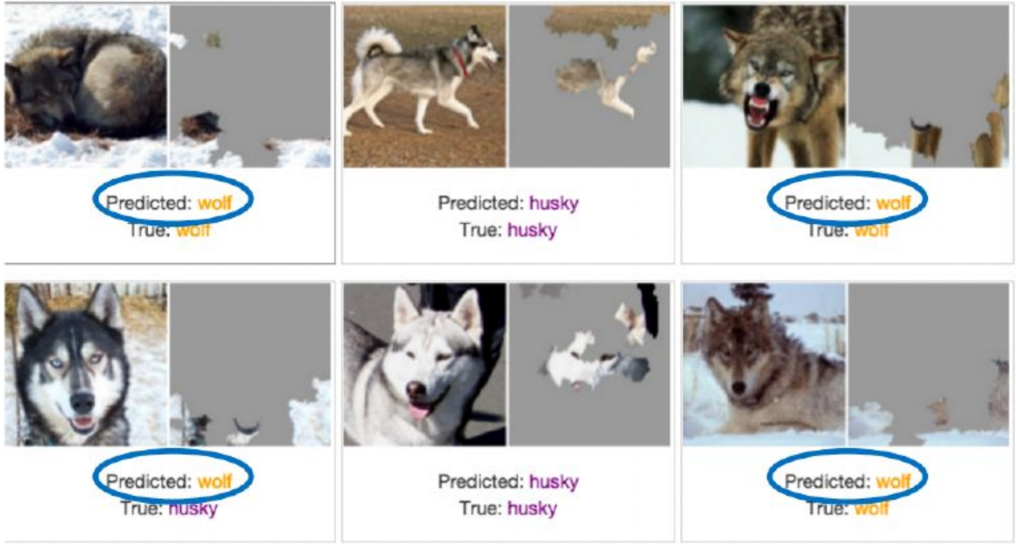
The many flavors of machine learning

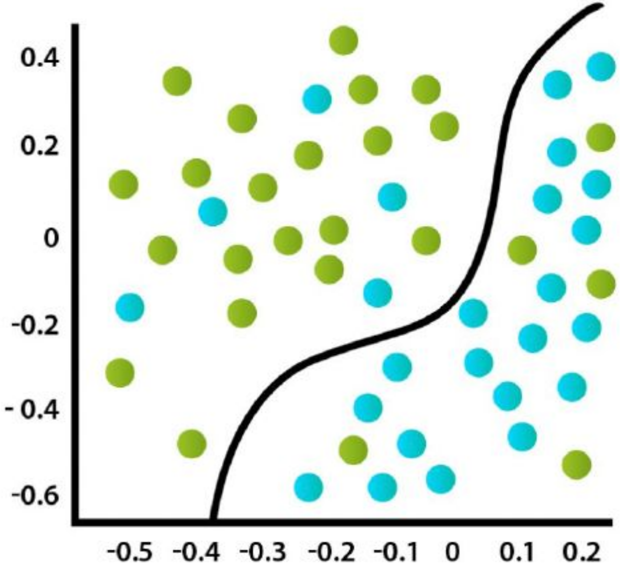


The many flavors of machine learning

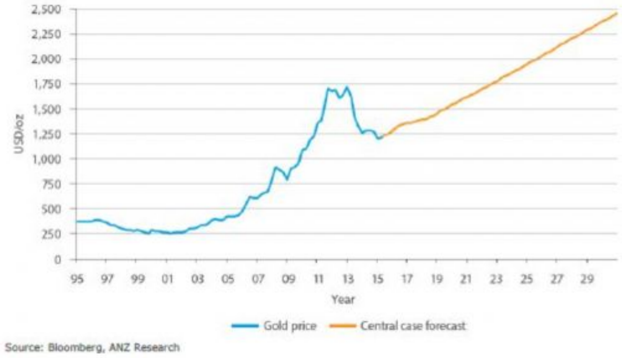


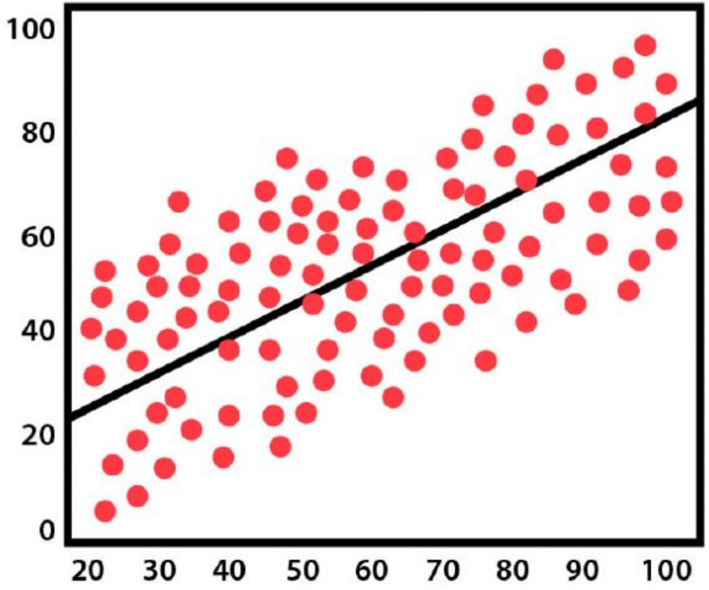
Machine Learning: Classification



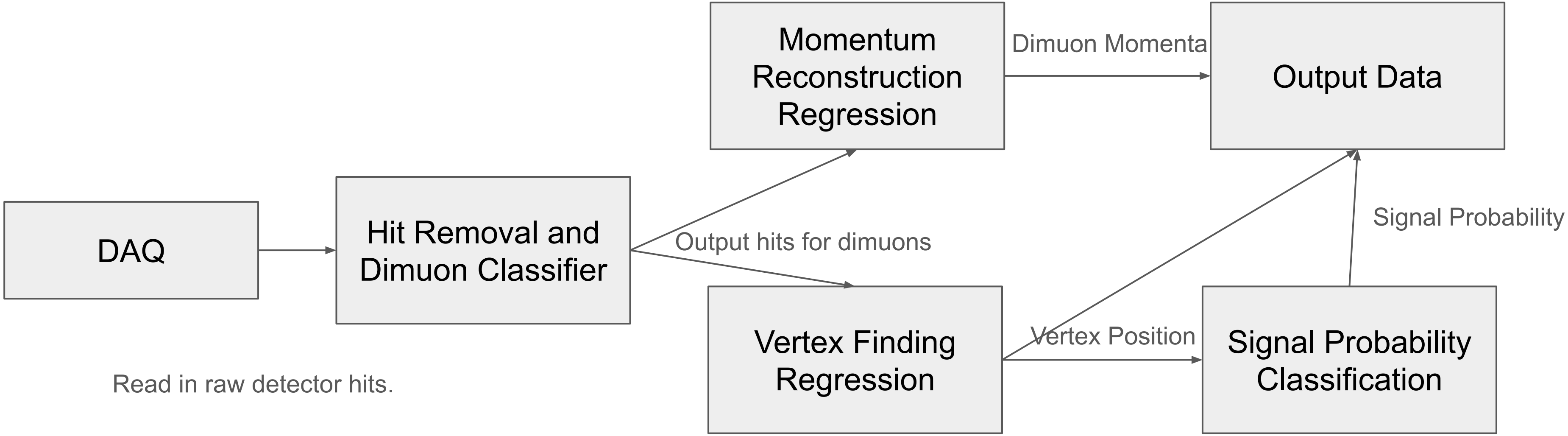


Machine Learning: Regression

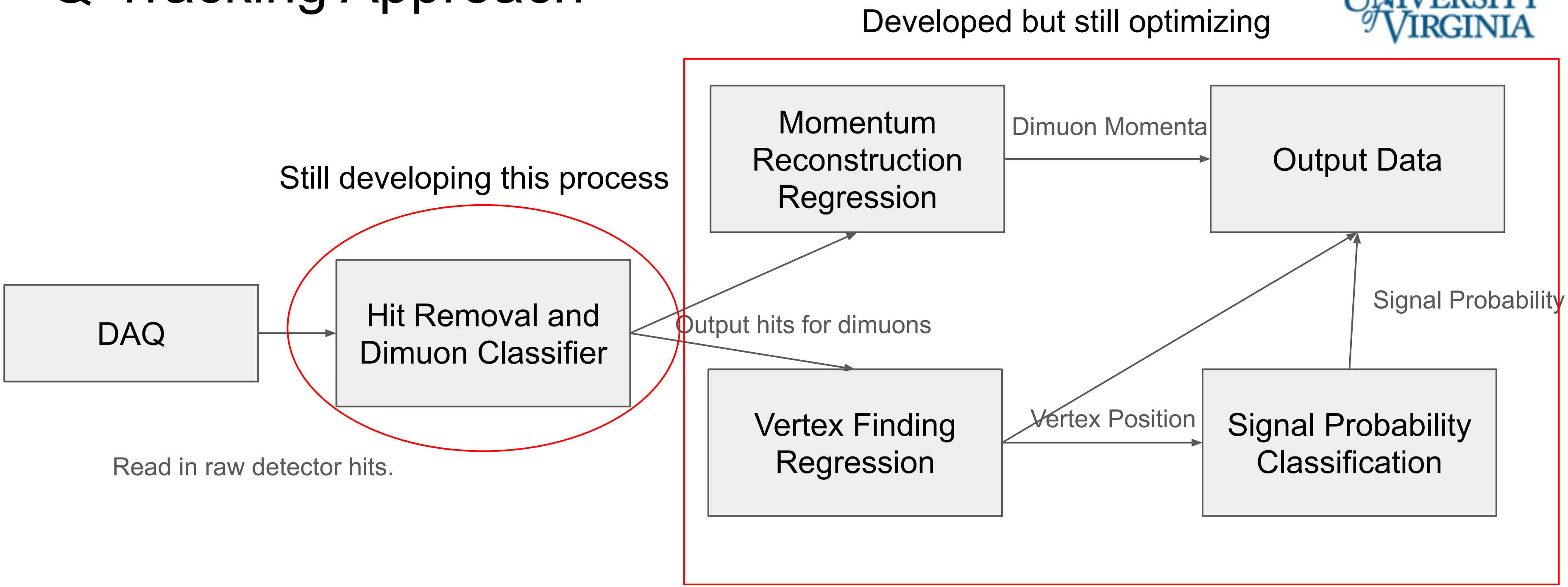




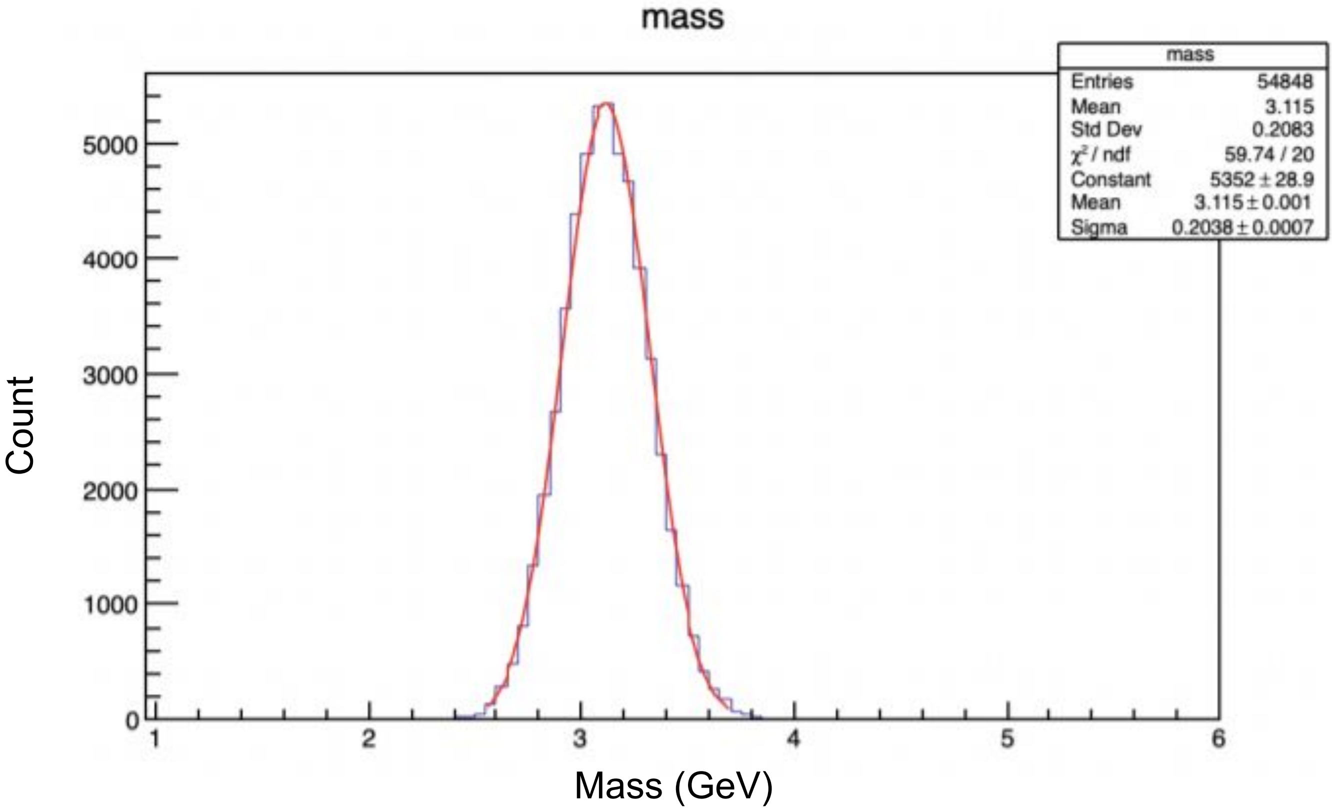
Q-Tracking Approach

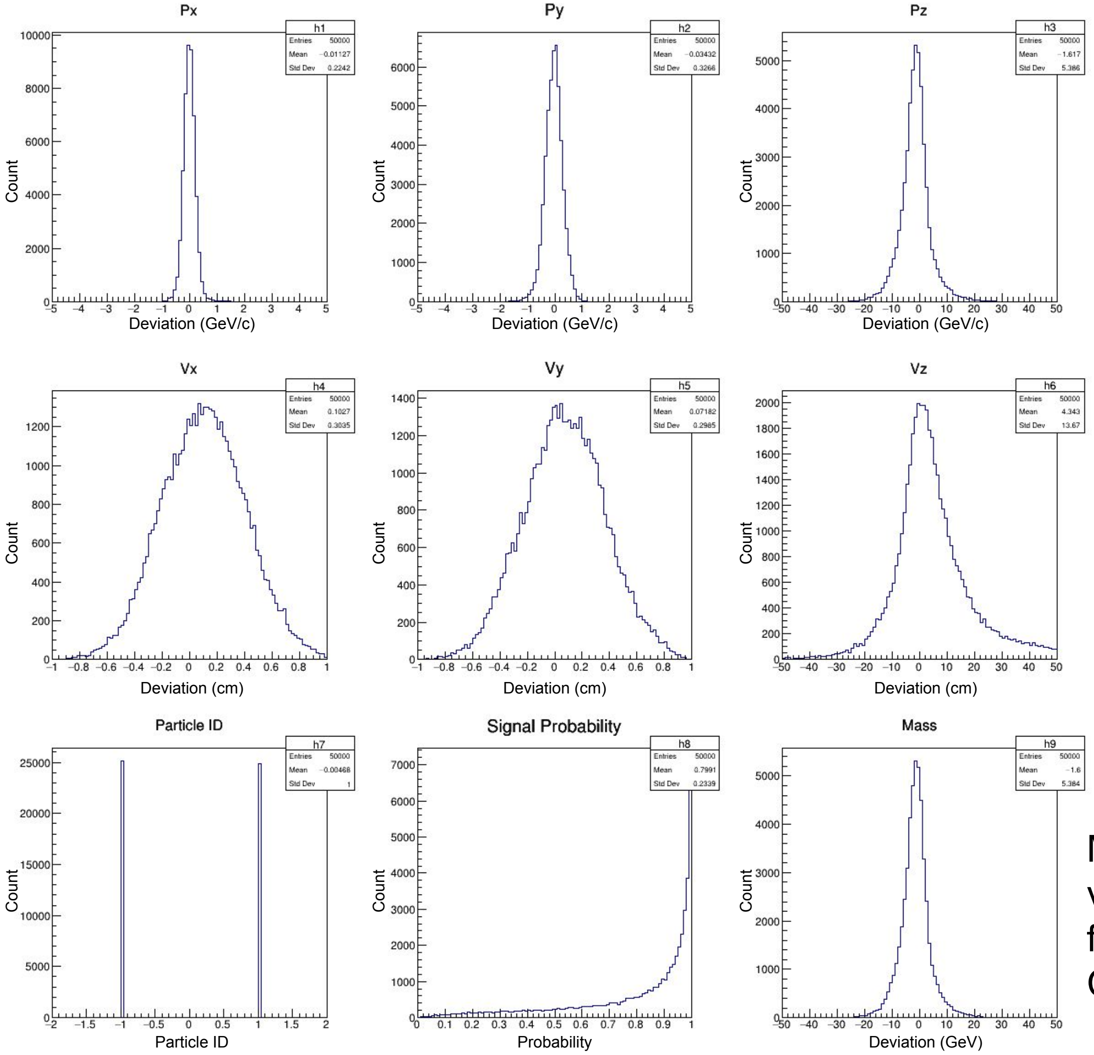


Q-Tracking Approach



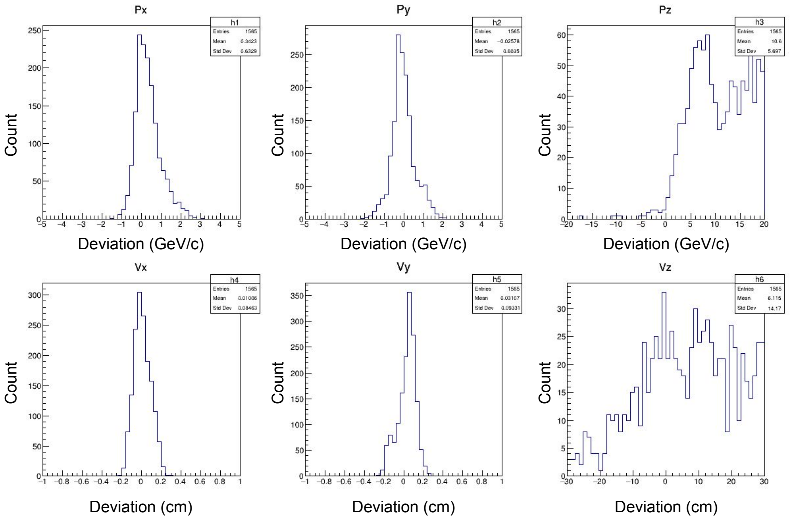
J/Psi Monte Carlo Peak Reconstruction

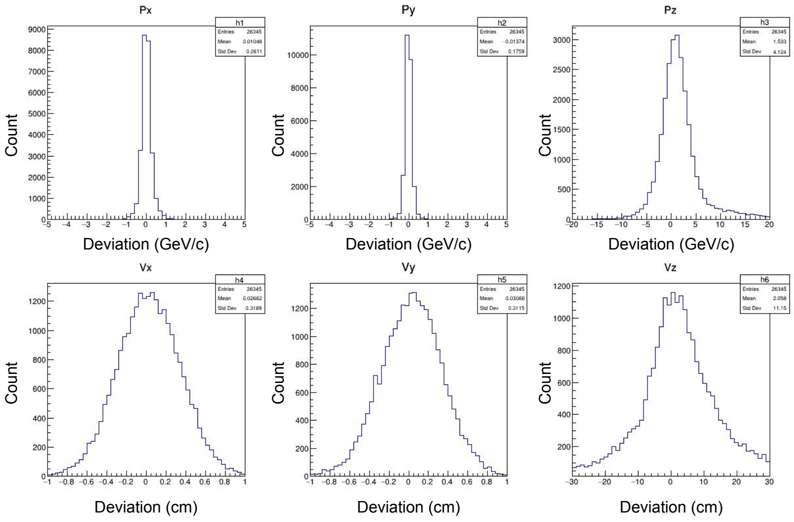




Momentum, mass, and vertex plots are deviation from true values of Monte Carlo data.

Evaluating data trained with a different process





Accelerating Neural Networks

- Performing regression and classification on a neural network is a series of matrix operations.
- GPUs allow for much more parallelization than CPUs. A typical CPU will have 8-16 threads, while a GPU can have 1,000+ threads.
- Different types of cores on GPU: Cuda Cores and Tensor Cores. Tensor Cores allow for even faster processing, since multiple operations can be done in a single clock cycle.
- Trade-off is memory allocation and loading data onto the GPU. This adds a latency time, so processes that trade back and forth between CPU and GPU can be bogged down.

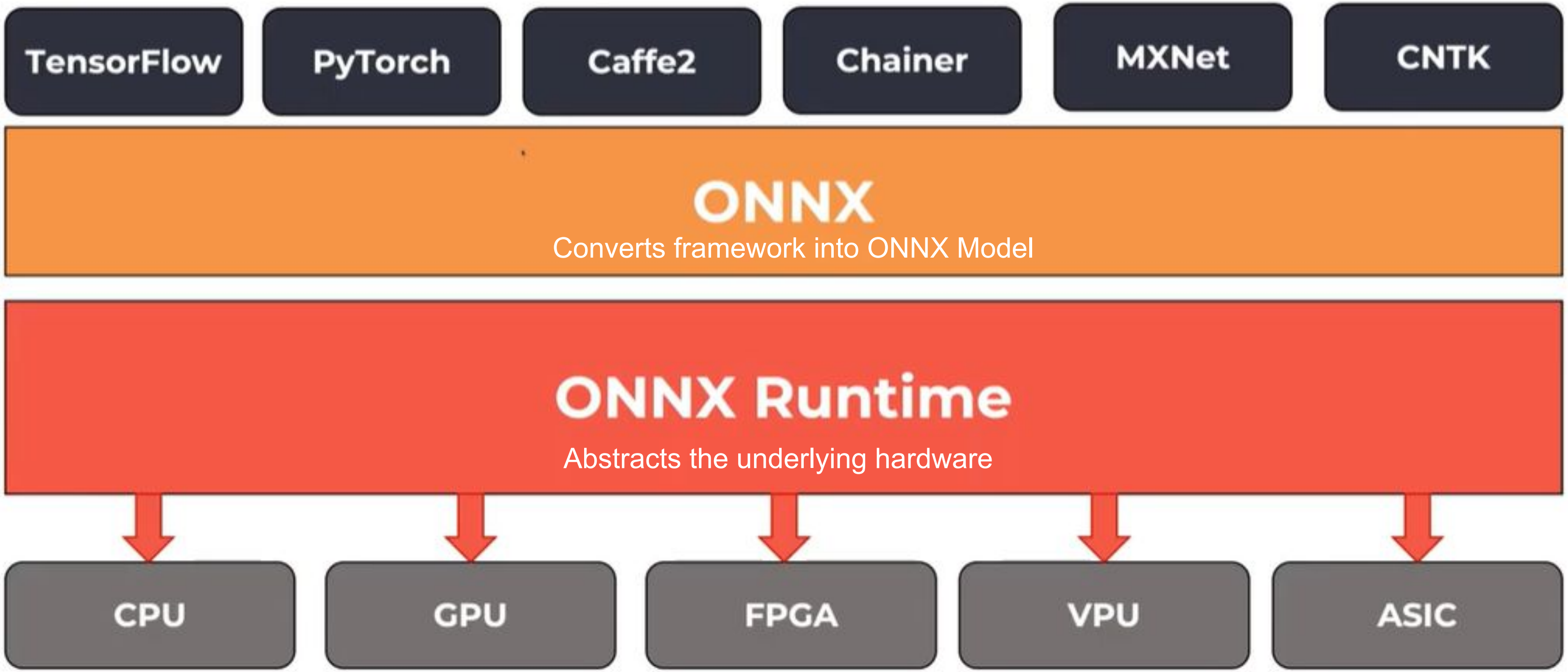
Why GPUs

- Machine learning is “embarrassingly parallel”
- GPUs have dedicated VRAM, which allows other operations to run on the CPU concurrently.
- Cuda Cores vs Tensor Cores
- Cost of consumer grade vs data center grade

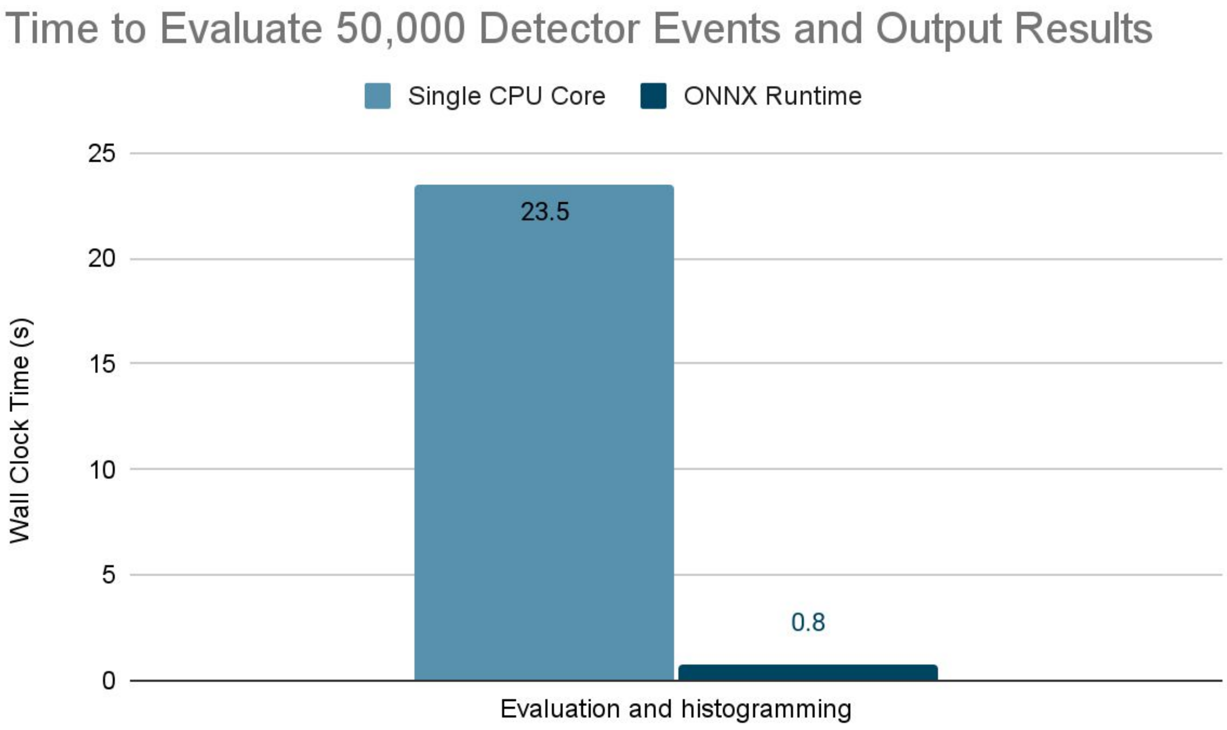
ONNX (Open Neural Network eXchange)

- ONNX takes a trained neural network (from a variety of frameworks) as an input and outputs an ONNX model.
- That model can then be run using ONNX Runtime.
- ONNX Runtime uses an extensible architecture, which allows it to use local optimizers and hardware accelerators.
- This allows inference to happen upwards of 15x faster than with non-optimized frameworks.

How ONNX Runtime Works



Comparison of analysis time (after filtering)

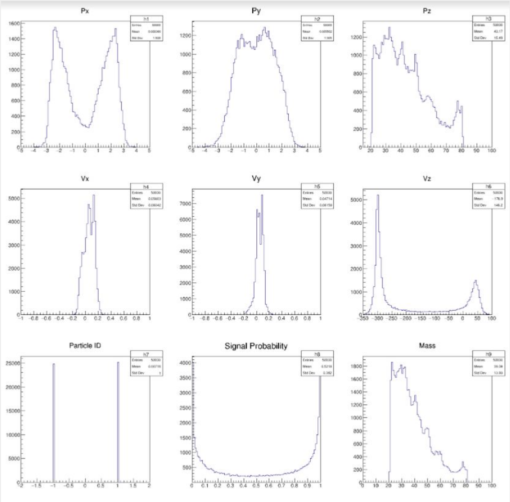


Plans for Online Monitoring System

- Output the reconstructed kinematic data between spills each minute.
- Use reconstructed kinematic data to detect any false asymmetries in the data. Asymmetries should not be measurable on a spill-by-spill basis.
- Generate images of the path of dimuons through detector arrays.
- Train an additional model to detect unexpected changes in detected events, as they could be a sign of target damage or other problems that need to be addressed.

Online Monitoring Mock-Up

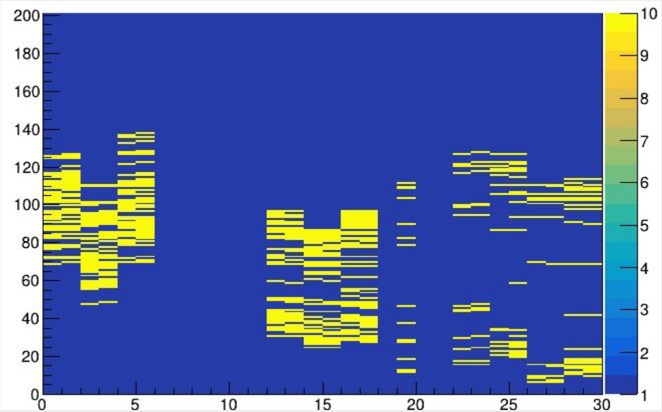
Kinematics



OPERATING NORMALLY

NO ASYMMETRY DETECTED

Dimuon Tracks



DY Detected: 4
J/Psi Detected: 26
Mean muon mass: 3.43 GeV

Summary

- SpinQuest online monitoring offers unique challenges that will require new, faster approaches.
- One of a few methods being pursued is to utilize neural networks to aid with filtering and reconstruction.
- This method shows promising results, but work is ongoing to fine-tune.
- Methods are available to accelerate evaluation, letting us perform the online monitoring within the time constraints while not sacrificing accuracy and precision.

SpinQuest Collaboration

Contact Spokespersons: Kun Liu (liuk@fnal.gov) - LANL

Dustin Keller (dustin@virginia.edu) - UVA

More information: https://spinquest.fnal.gov/

Schedule/Status:

- Ongoing since summer 2018: Equipment commissioning
- Winter 2021-22: Beam commissioning planned start
- 2022-2024: Experiment runs

University of Virginia

Fermilab