Fermilab | U.S. DEPARTMENT OF ENERGY | Office of Science

Neutrino Analyses and Computing Looking Towards 2026+

A. Norman

DUNE Software and Computing

Workshop on UK/FNAL Computing

Edinburgh 2018

Introduction

When we talk about computing for DUNE what do we mean?

There are a lot of questions

- What is the scope?
- How does this tie into the physics?
- How are the requirements, workflows, techniques different from….
  - Colliders?
  - Other neutrino experiments?
  - Other physics programs?
- How do we begin mapping this out?
- How does this connect back to the “projects” we need to start?

The Physics

- Primary mission of DUNE is to constrain/measure δ_{CP}
- Why?
  - This is the phase that manifestly can violate CP conservation.
  - It is not the phase directly responsible for the matter/anti-matter asymmetry
    - But if it is non-zero, then other related phases can be the source
- To do this means measuring $\nu_\mu \rightarrow \nu_e$ and $\nu_\mu \rightarrow \nu_\mu$ spectra and looking for deficits.
  - This is fundamentally different from most collider measurements and uses different computing techniques

Where we are today with δ_{CP}

The best data to date (summer 2018) comes from NOvA.

- They make separate and combined measurements of appearance and disappearance in neutrino and anti-neutrino modes:

$$P(\nu_\mu \rightarrow \nu_e) \quad P(\bar{\nu}_\mu \rightarrow \bar{\nu}_e)$$
$$P(\nu_\mu \rightarrow \nu_\mu) \quad P(\bar{\mu}_\nu \rightarrow \bar{\nu}_\mu)$$

- From each set you try to extract the PMNS mixing parameters
- Best indication is that we are NOT in a world with $\delta_{CP}=3\pi/2$
  - This means the DUNE baseline and wide-band beam configuration is required to resolve the phase.
  - Bet…everything is low stats and can change

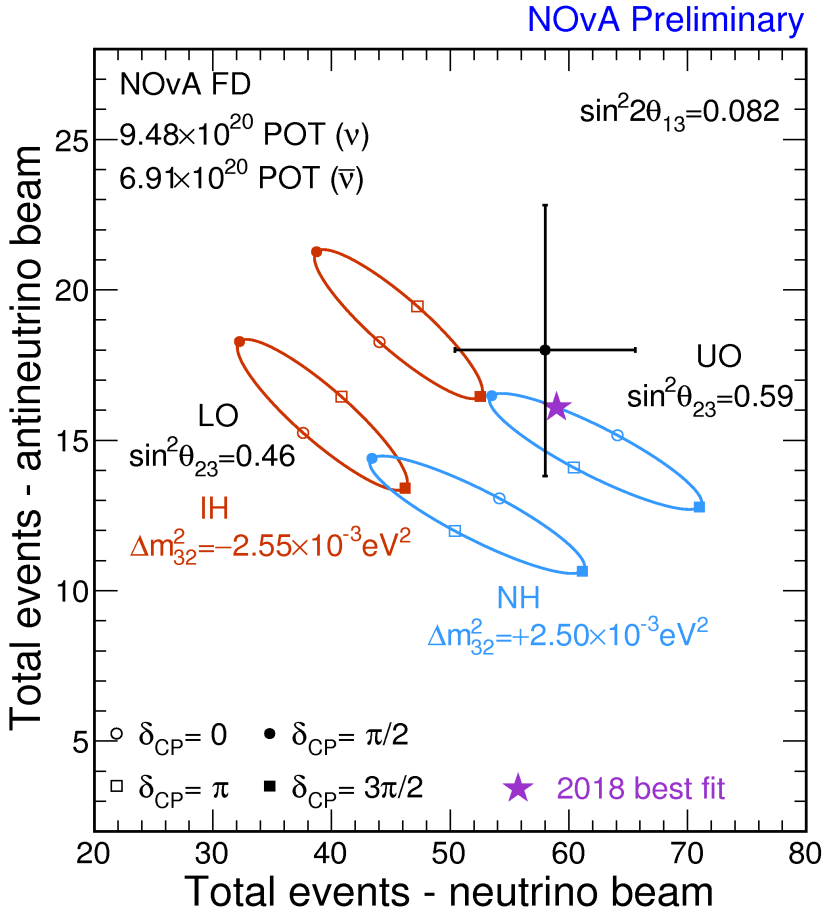
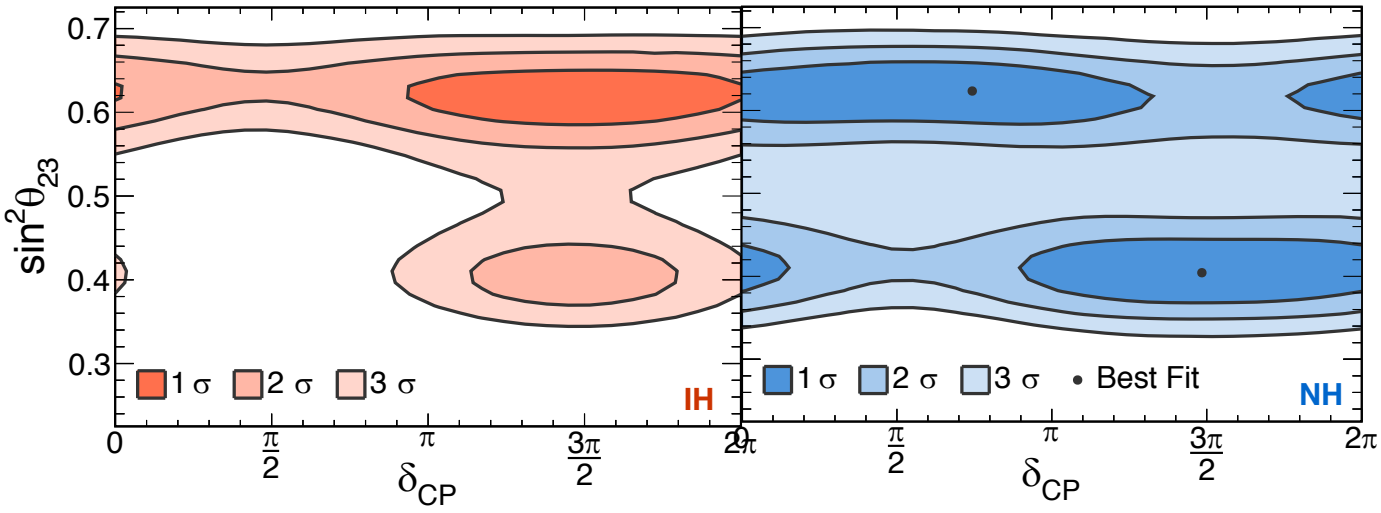


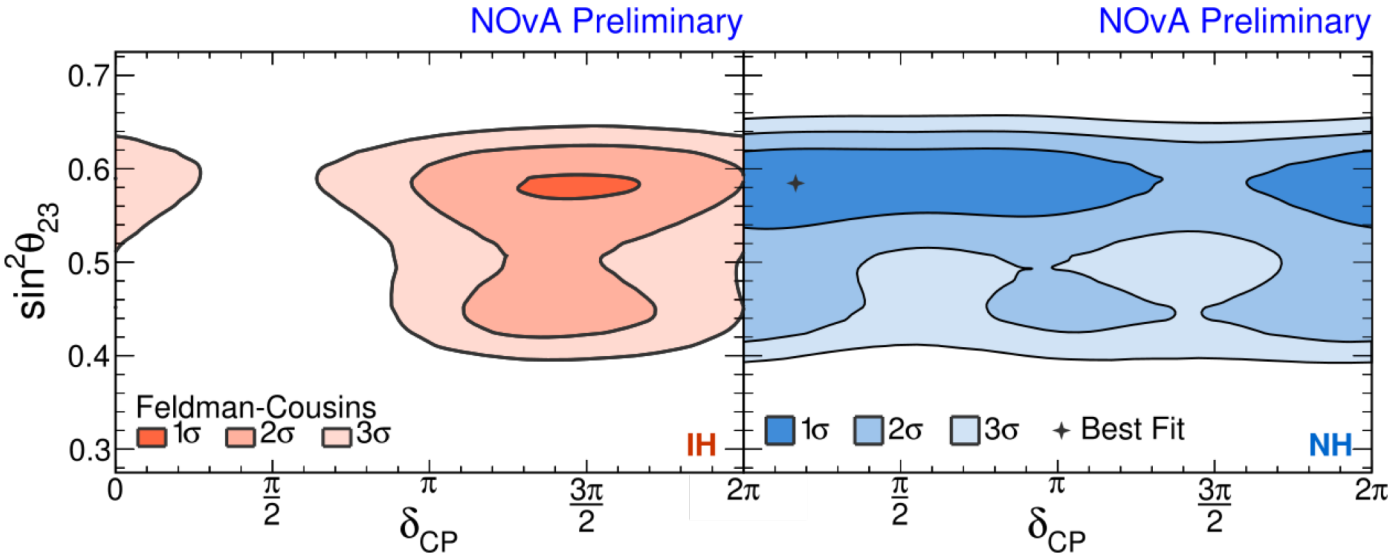
FIG. 1: Neutrino/antineutrino oscillations bi-probability ellipses for NOvA data with 2018 best fit solution. Blue/red ellipses are for the Normal/Inverted neutrino mass hierarchy. Solutions for the upper ($\theta_{23}>\pi/4$) and lower octant ($\theta_{23}<\pi/4$) are shown.

Data

2017

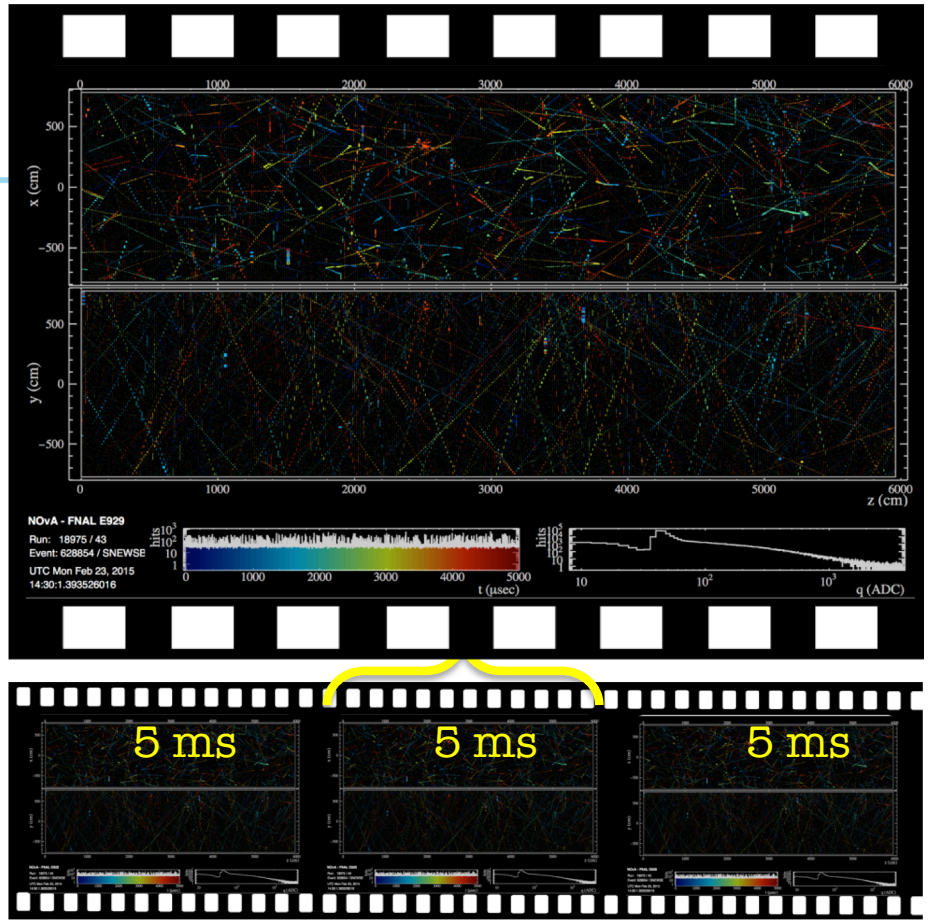


2018



The Neutrino Data

- Detector is effectively “always on”
  - Particles crossing the volume leave ionization (tracks) with ms scale lifetimes
- Very different from collider or fixed targets
- Effectively acquire a continuous “movie” of the activity in the detector
  - ”Frame” corresponds to the drift across the volume
  - Need some way to determine which time “slices” to save.
- Min-bias/Zero-bias just takes time region
- “Triggered” readout uses some external/internal system to identify a likely region of activity to extract (asynchronously) from the systems.
  - Photon system detects light, causes readout (semi-traditional)
  - Front ends produce “trigger primitives” that can be aggregated (SN trigger)
  - Beam signal comes from FNAL
  - Other external inputs, other calibration source, etc…
- A “trigger” becomes a t_0 and a Δt window of interest



The Neutrino Data

- Because the data is a movie, it's useful to talk about the "fraction of total live" that is recorded.
  - Beam spills currently are 10μs in length and occur at ~.83 Hz
    - This is 1/120,482nd of total live.
    - Requiring a full drift makes this 1/240th of total live
  - A supernova readout is 100% of total live

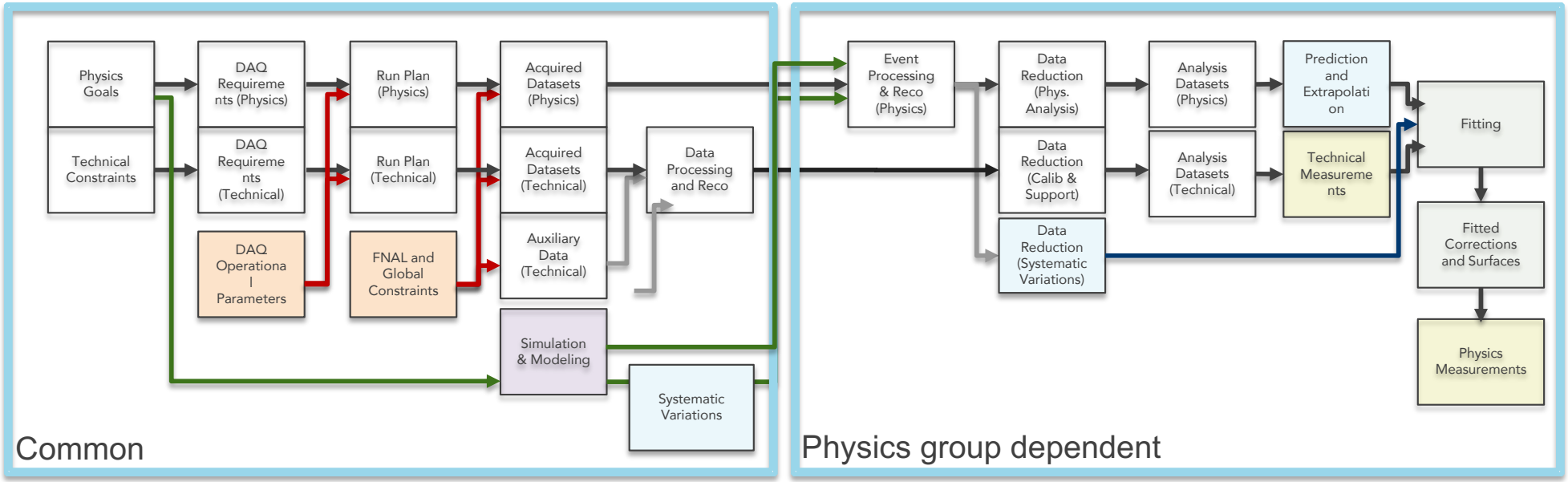
- These are vastly different time scales.
- These are vastly different data volumes.

- Computing needs to be able to deal with the full range.

- Today we talk about a data volume of 30 PB/yr as a design target

Analysis Path

- How do you analyze this data?
  - There needs to be a well defined chain the flows from the measurements being made (effectively stages of simulation, data processing, reco, event classification, fitting.)



- For neutrinos, we know how to do all of these at some level†
  - NOvA, MicroBooNe, Minos, have demonstrated these for data volumes about 1/10th of DUNE

†Ignore calibration for now, this is hard

Analysis Chain

- So for a “typical” year of data what would this look like?

- Some Assumptions
  - Analysis Cycle is 18-24 month
    (i.e. tied to Neutrino conferences)
  - Target raw rate is 30 PB/yr and runs constantly
    (i.e. still runs during shutdowns)
  - Shutdowns are annual (12-16 weeks)
  - Beam rep gives ~2-2.4E7 spills/year
  - Analysis model looks similar to NOvA/Minos

Simulation vs. Data processing

- Incident neutrinos are invisible
  (we only observe the interaction not the actual particle).
  - But oscillations are driven by their energy

- Knowledge of the incident spectra are driven by simulation
  - Highly dependent on hadron production
- Knowledge of the interaction are driven by simulation
  - Highly dependent on nuclear models
- Knowledge of the detector response are driven by simulation
  - Highly dependent on external data

- **All neutrino measurements require high Monte Carlo to Data ratios**
  - **Typical is 10:1 with goals of 50:1 and 100:1**
  - **With reuse some experiments are getting 20:1 currently**

Data Processing

- Raw data is big (~30 PB/yr) but the processing models typically only spin through the lowest stages of its processing one time.
  - The first “raw2root” stage is the most IO intensive, but is typically done in a “keep-up” mode
  - Prior experiments have had very stable first stages
    - NOvA has to date not changed or re-run the keep-up. (this would require the restaging of multiple petabytes of data from tape)
    - Possible because algorithms were specifically decoupled from early stages (i.e. no hit finding, tracking)
  - For DUNE this may not be the case.
    - Hit finding and noise suppression are needed to reduce the event data size down from 6.2 GB/evt
    - May result in a need for respins of raw data
  - Hit dropping is NOT currently performed by most experiments (they try to preserve all hit info for ML applications)
    - DUNE may not have this luxury

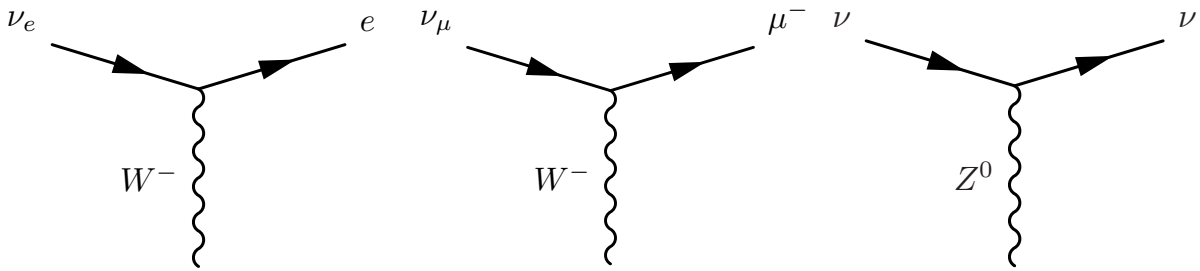
Reconstruction

- Reconstruction is tricky.
  - To understand what reco needs to do accomplish, you need to understand the measurements

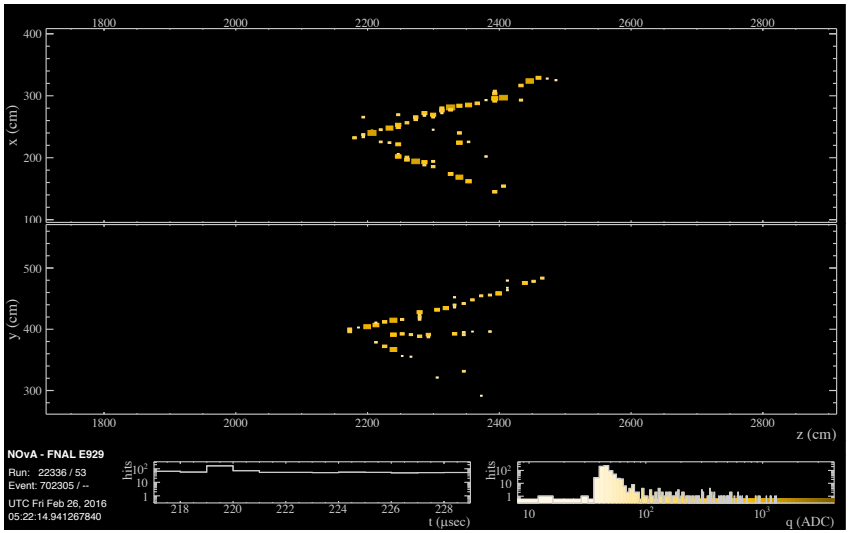
- Reconstruction is also very expensive to run
  - Typical LAr reco takes 10's of minutes per event
  - Driven by the complexity of the events and raw hit multiplicities -- many algorithms are worse than $O(N^2)$
  - Machine learning exacerbates this

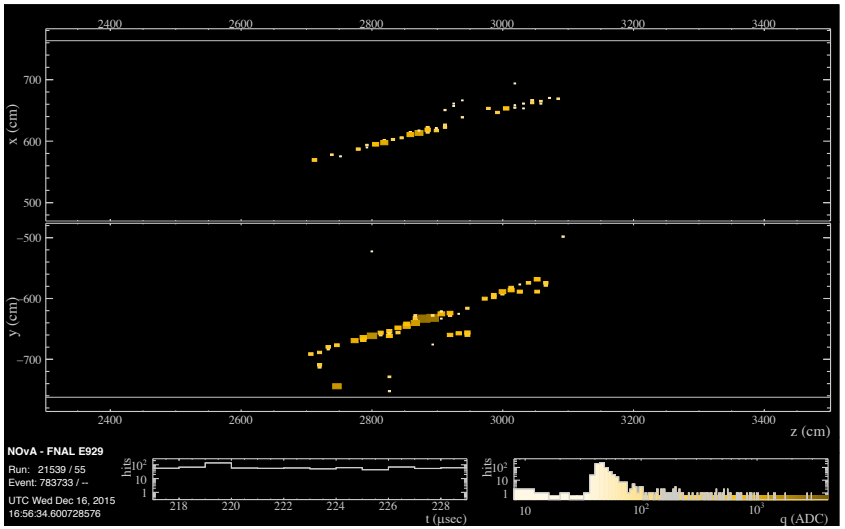
The Analysis

- Neutrino Oscillation analyses are about classifying three topologies, ν_e-CC, ν_μ-CC, ν-NC



- Estimating the energy of the incident neutrino
- Counting in bins of energy how many of each type we see

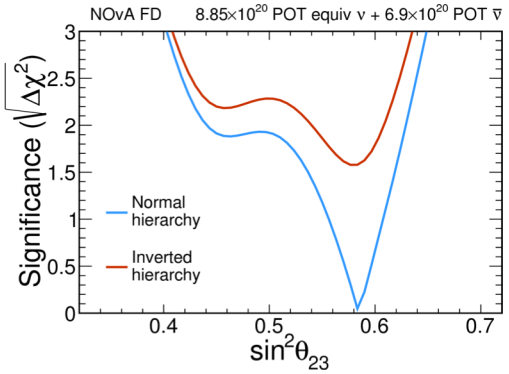


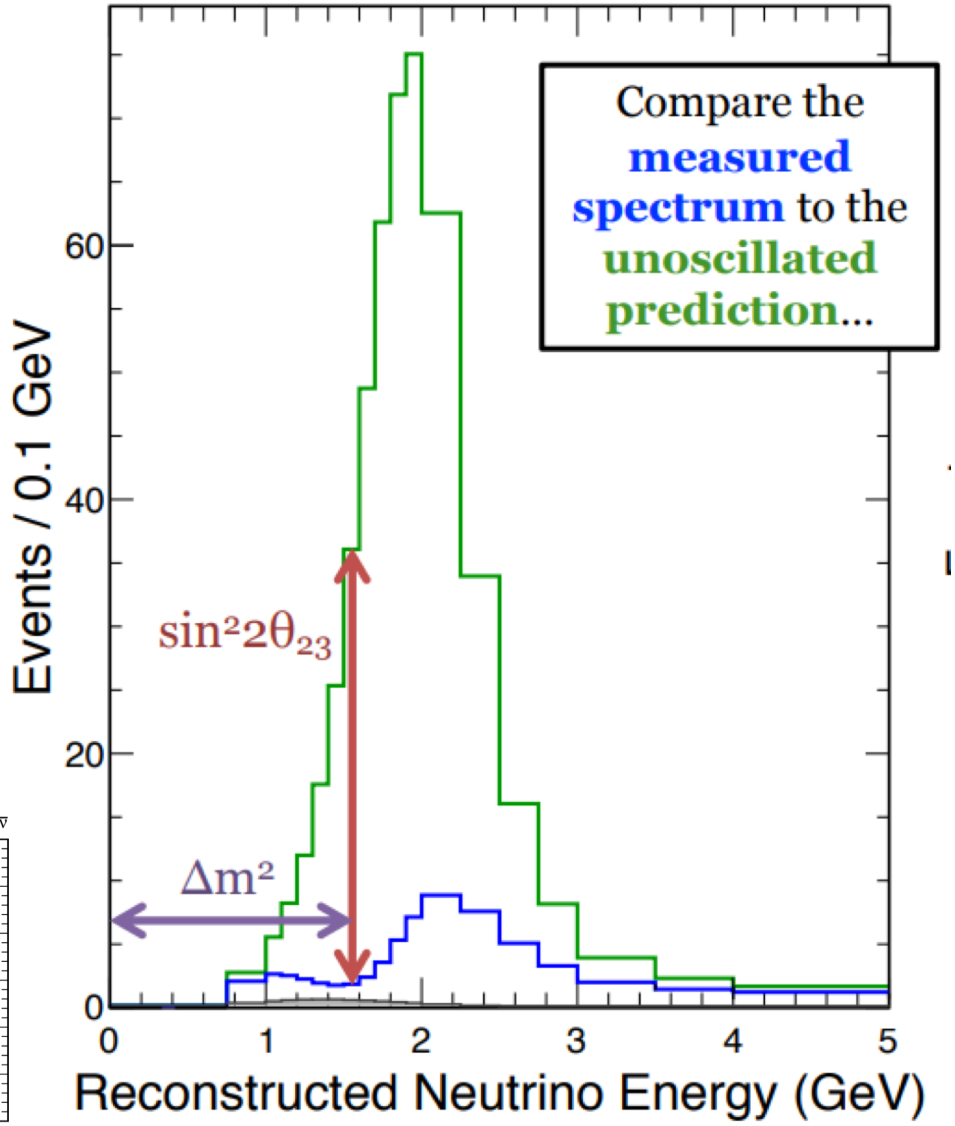


Selected ν_e-CC Topologies

Measurement Technique

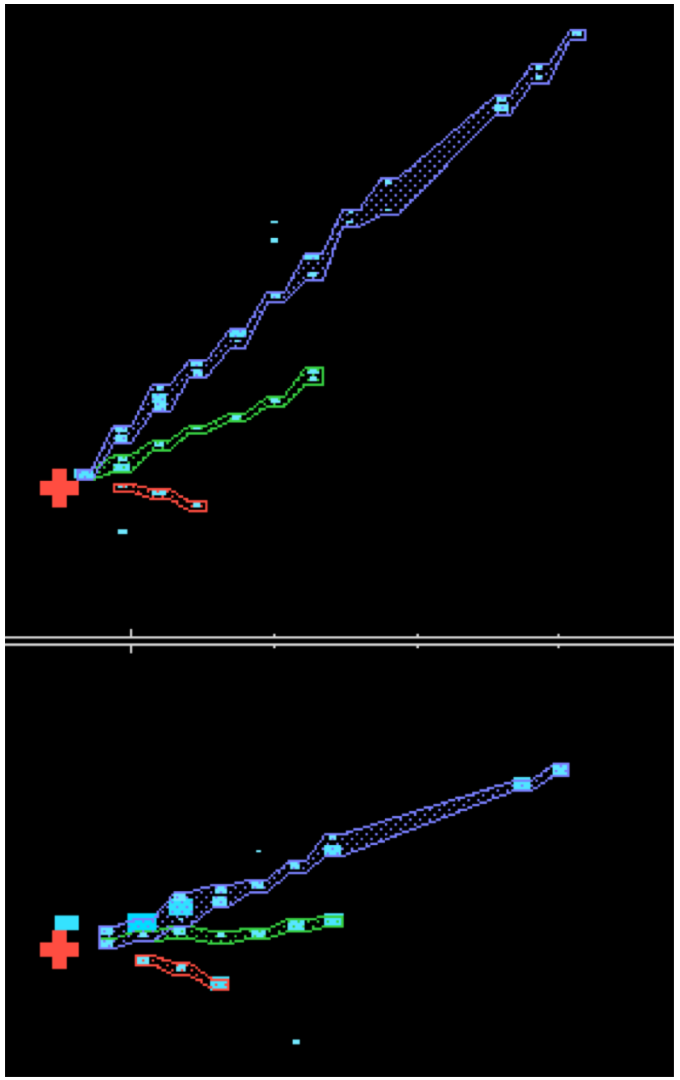
- Measure near detector neutrino spectra and predict the far detector spectra before and after convolution with an oscillation model.
- Fit for the oscillation model parameters using a likelihood method and determine the confidence intervals for the measurement.
- Key is that neutrino energy must be reconstructed and that event types are classified correctly
- Shape fits across very large spaces
- Current fits are 60+ parameters
- Limited by observables and free parameters





Classifying Events (Traditional)

- Traditional Event classification relies on individual particle recognition and reconstruction
- Computationally:
  - Clustering/Pattern recognition algorithm
  - “Fitting” algorithm (for track-like objects)
  - Vertexing algorithm
  - Energy estimator algorithm (per particle)
  - Energy estimator algorithm (global)
- Then
  - Run all the kinematics through a series of selection cuts
  - ANN, BDT or other multivariate selection
- Works well….when you can reconstruct ALL particles
- But…LAr events have such high hit multiplicities and low energy “stubby” tracks (deltas) that this doesn’t always scale



Classifying Events (Deep Learning)

- What if you didn't have to actually disentangle particle by particle?

- Instead you classify each event as a whole based on image recognition technology
  - This is the Convolutional Visual Network (CVN) approach first used by NOvA in 2016.
  - It achieves a 30% higher signal efficiency
    - Effectively you need to through out less of your data due to un-reconstruct-able particles
  - Comes at a cost
    - Requires GPUs, special computing, etc.....
- Here's how it works

Image Filtering

- Traditional Image Filtering is convolution with a given kernel.
  - Reduces to a series of matrix multiplications

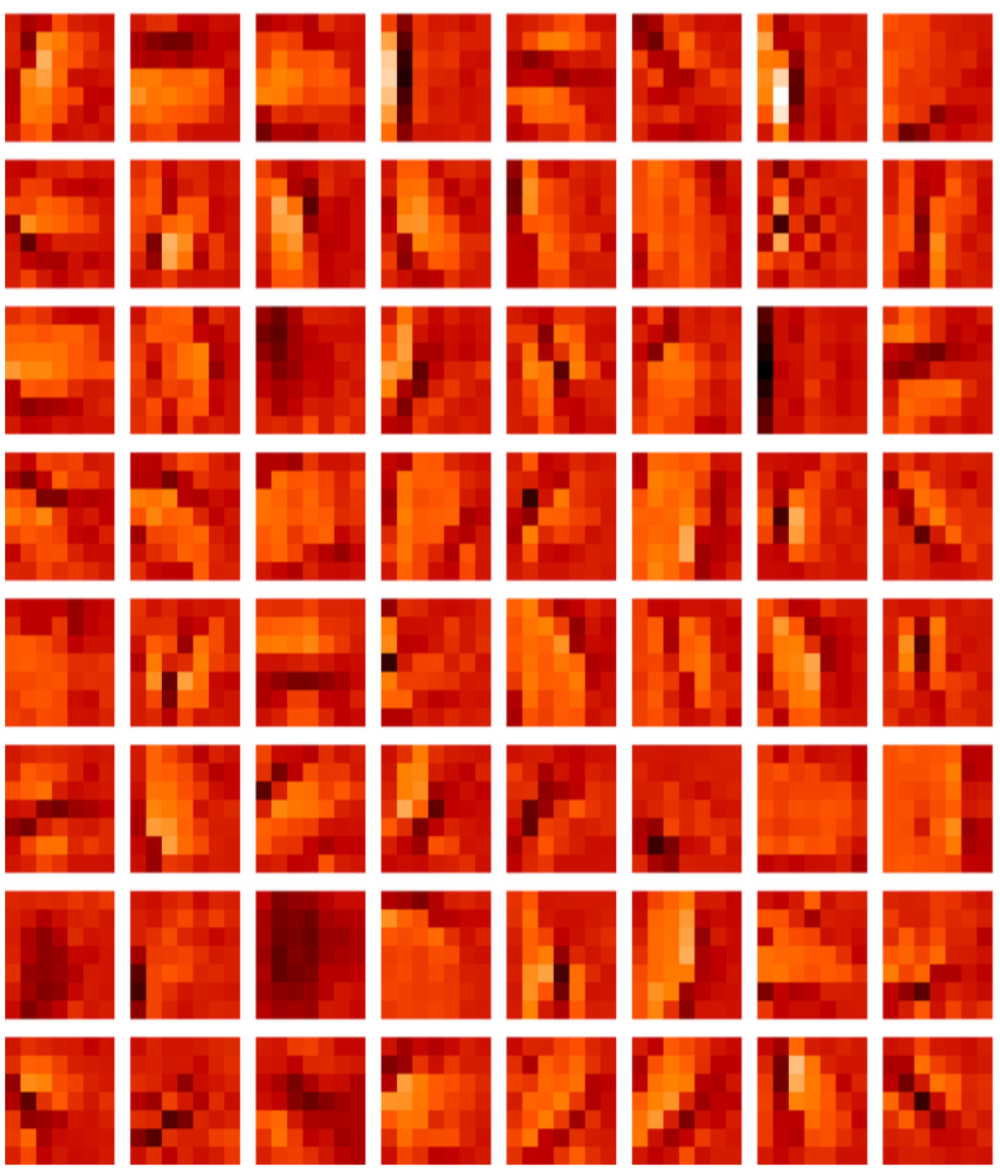
X/Y Kernels

$$\begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{pmatrix}$$

Horiz Edge Vert. Edge

- Goal is to find a series of kernels that "sees" different neutrino interactions

Neutrino Detection Kernels



These are made by the machine through convolutional network training. Requires GPUs.

Large University resources currently. Targeting HPC facility's in the 2020's

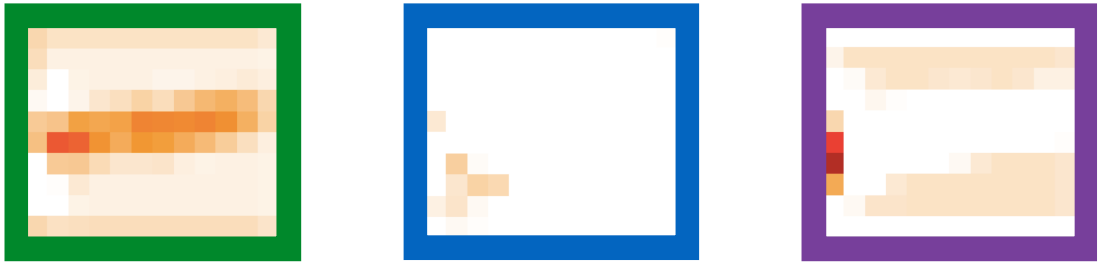
NOvA Convolution Kernels Kernels for Neutrino Detection (2017/2018)

Fermilab

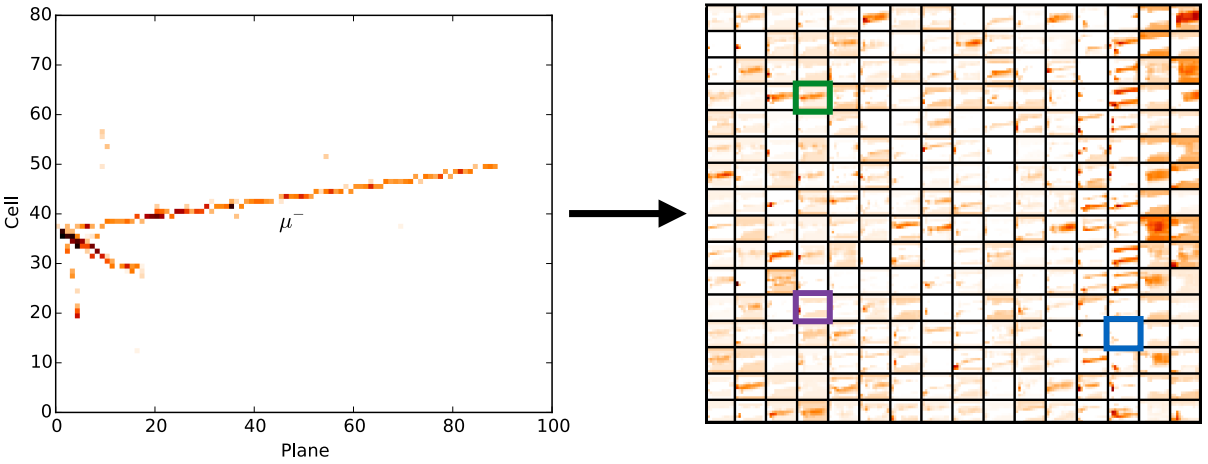
Convolutional Networks & Deeplearning

- Train through a neural network, so that the network learns which kernels are important to a signal/background sample
  - Feed back in and let the kernels vary so that the features maps become tailored to the problem

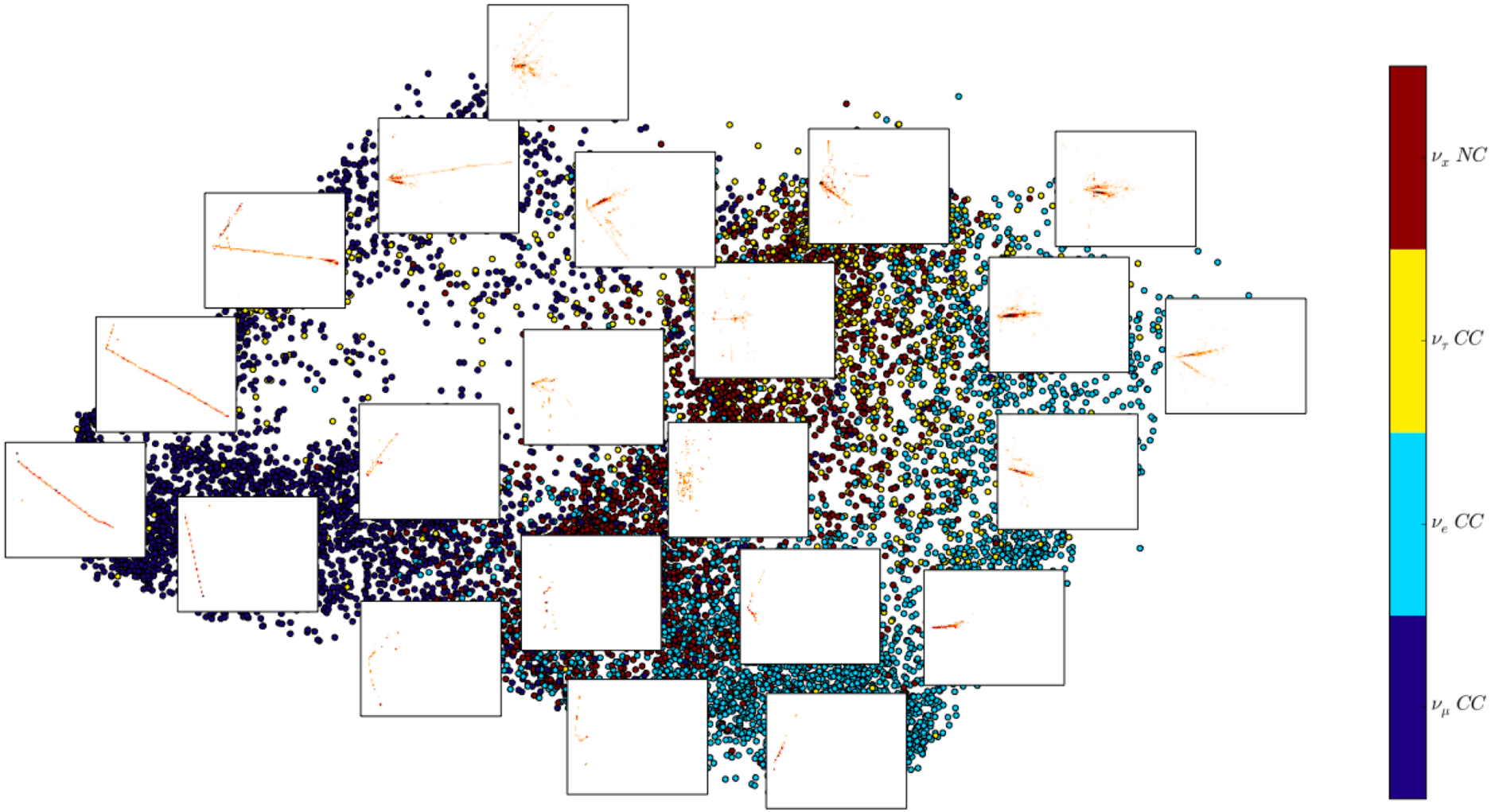
ν_μ-CC interaction and resulting feature maps



Training requires GPUs evaluation on events is near constant time and reduces to large matrix multiplications (ideal for HPC environ.)



Selection Space

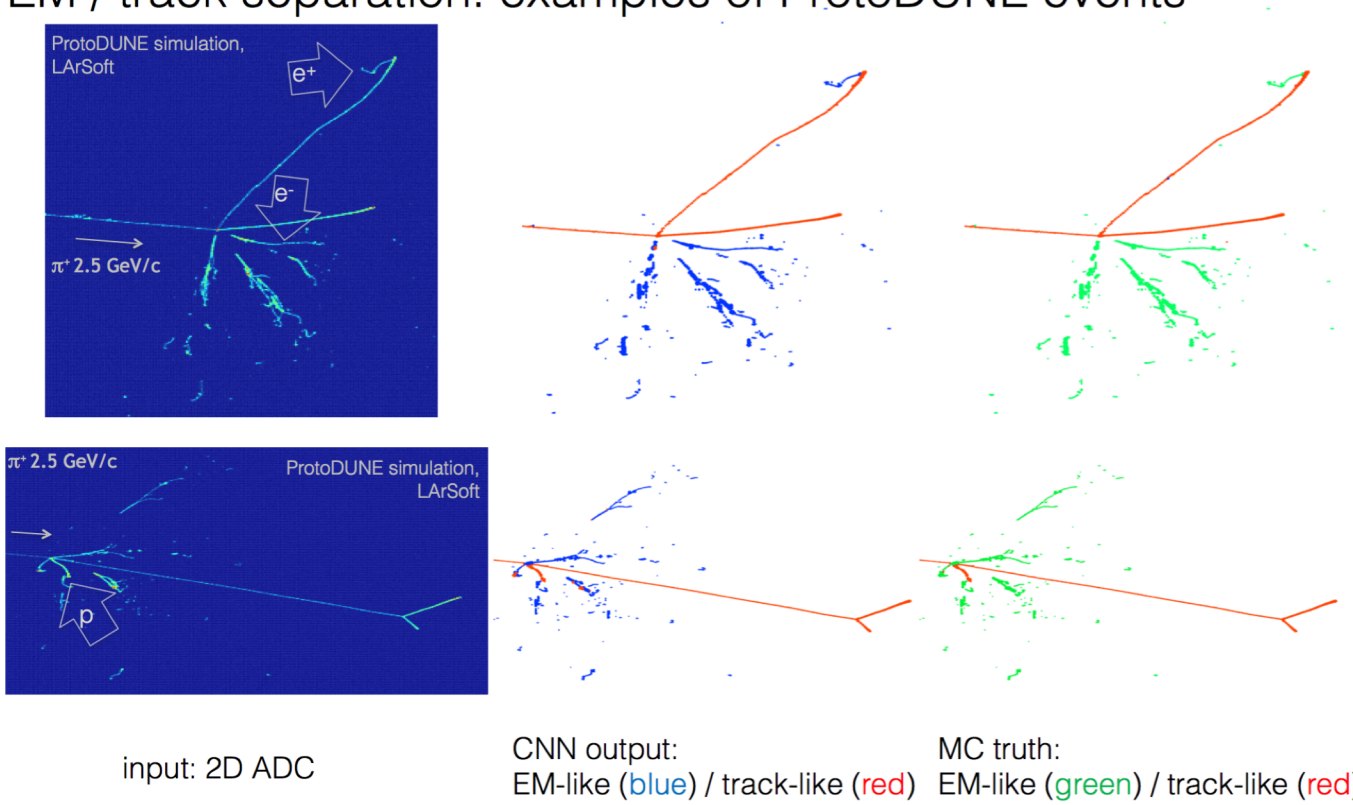


Different interaction topologies cluster in the selection space

Or for particle reconstruction

- Semantic Segmentation (i.e. pixel by pixel deep learning)

EM / track separation: examples of ProtoDUNE events



input: 2D ADC

CNN output:
EM-like (blue) / track-like (red)

MC truth:
EM-like (green) / track-like (red)

Event displays: R.Sulej, Connecting The Dots / Intelligent Trackers, May 2017, LAL-Orsay, France

Works to separate out individual tracks and showers.

Allows better energy estimators

Need combo GPUs/CPUs at large scales

Reco + Event Selection

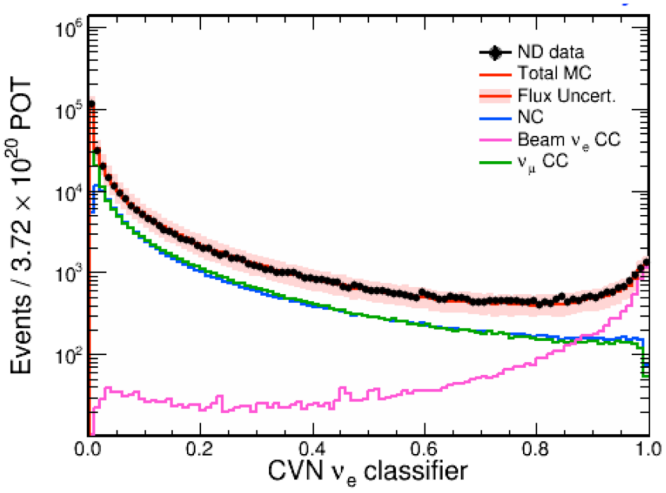
- After reconstruction you have higher level quantities that can be used to select samples

- This allows for the creation of “reduced” or “n-tuple” style analysis files.
  - Typical reduction of 100x—1000x in data size.
  - Example: NOvA 13 PB (raw+sim+MC) -> 25 TB of CAF
- These n-tuple files are run over repeatedly

- However…there is a feedback loop. Selection/analysis techniques push refined reco, require re-reco, require regeneration of n-tuples…repeat….repeat…repeat…
  - This stage ends up consuming more resources than initial processing
  - Also scales w/ systematic studies (i.e. each systematic requires it’s own set of reco/tuple runs)

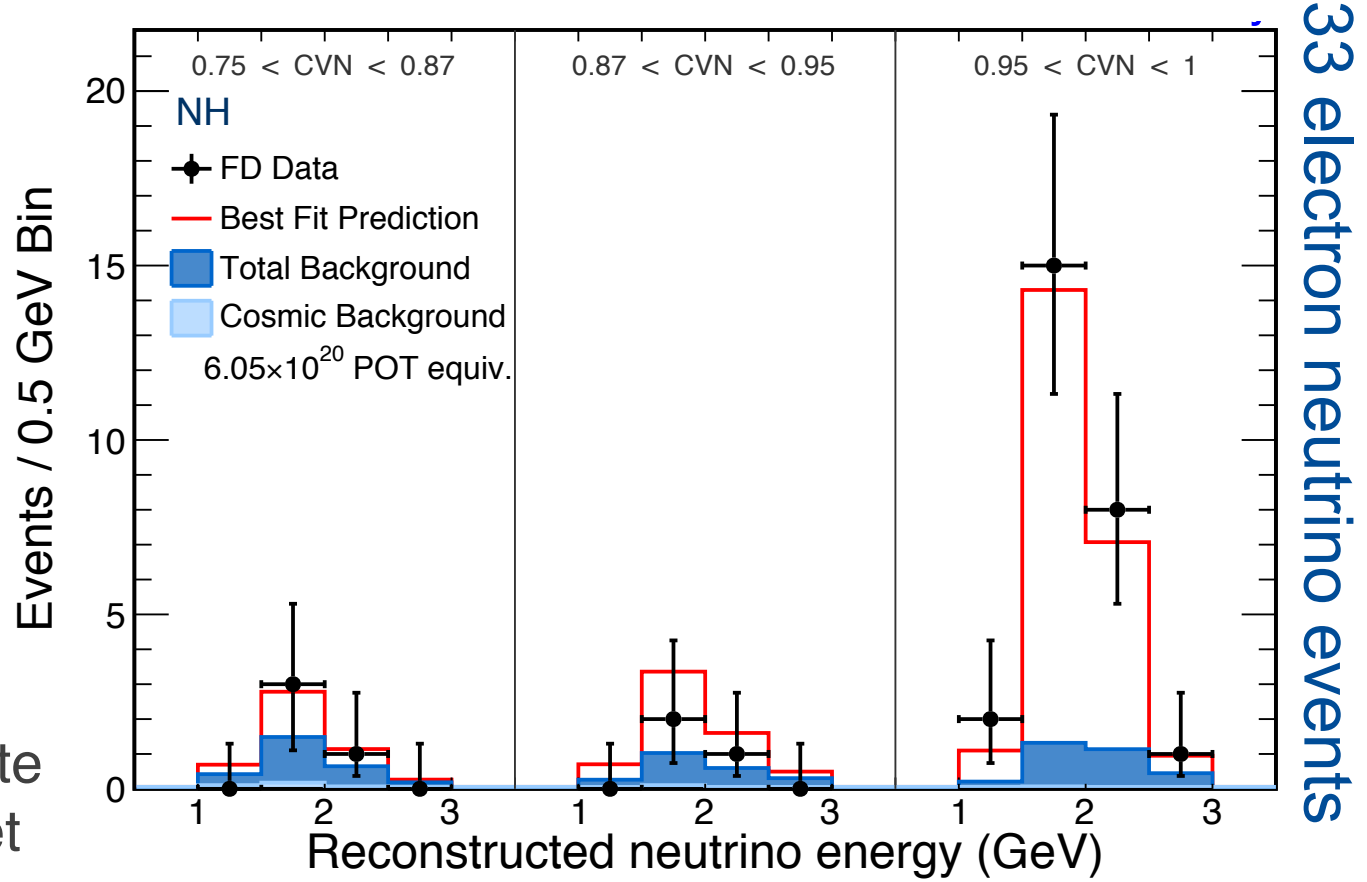
Result

- Eventually….After applying your standard and ML techniques, you have a normalize selector suite



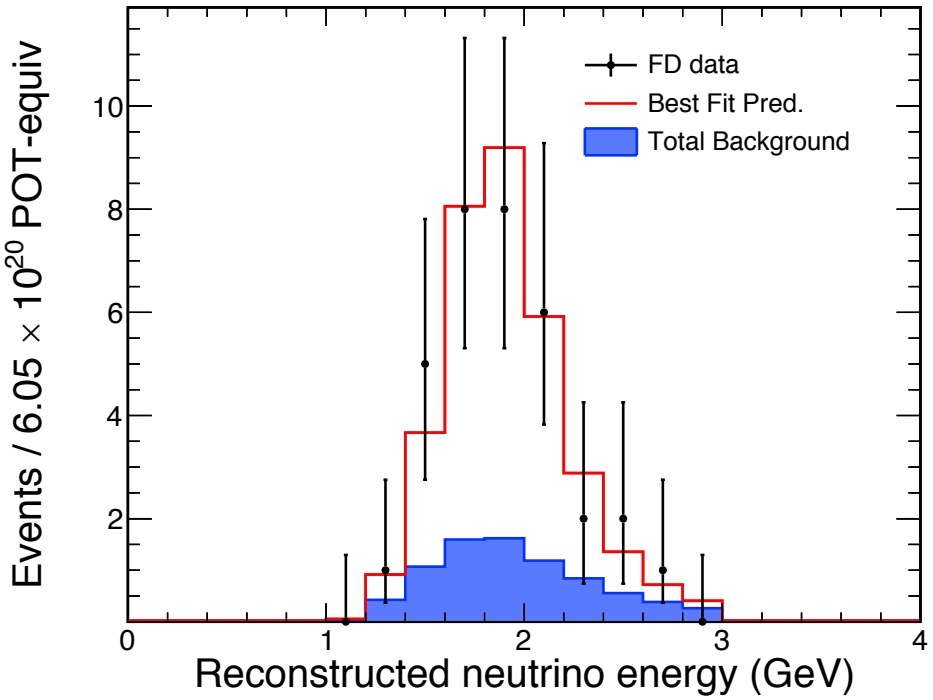
- You classify your data AND you have to predict your spectra from your Monte Carlo and background samples

- Monte Carlo and background samples (data driven) dominate the computing budget

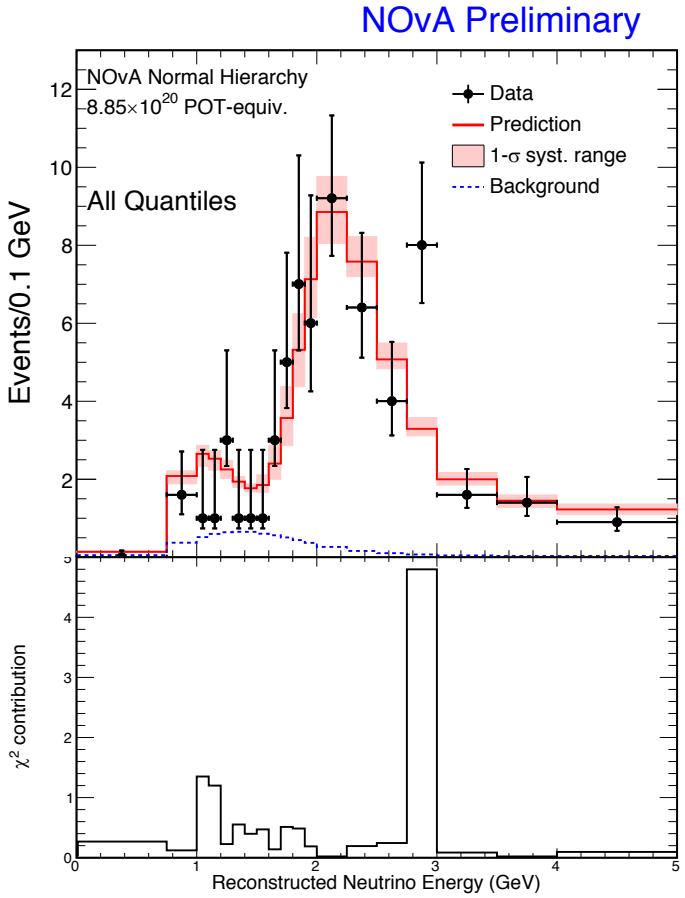


The Experimental Observation

- You observes ν_e appearance and ν_μ disappearance



NOvA ν_e appearance spectrum as a function of observed energy

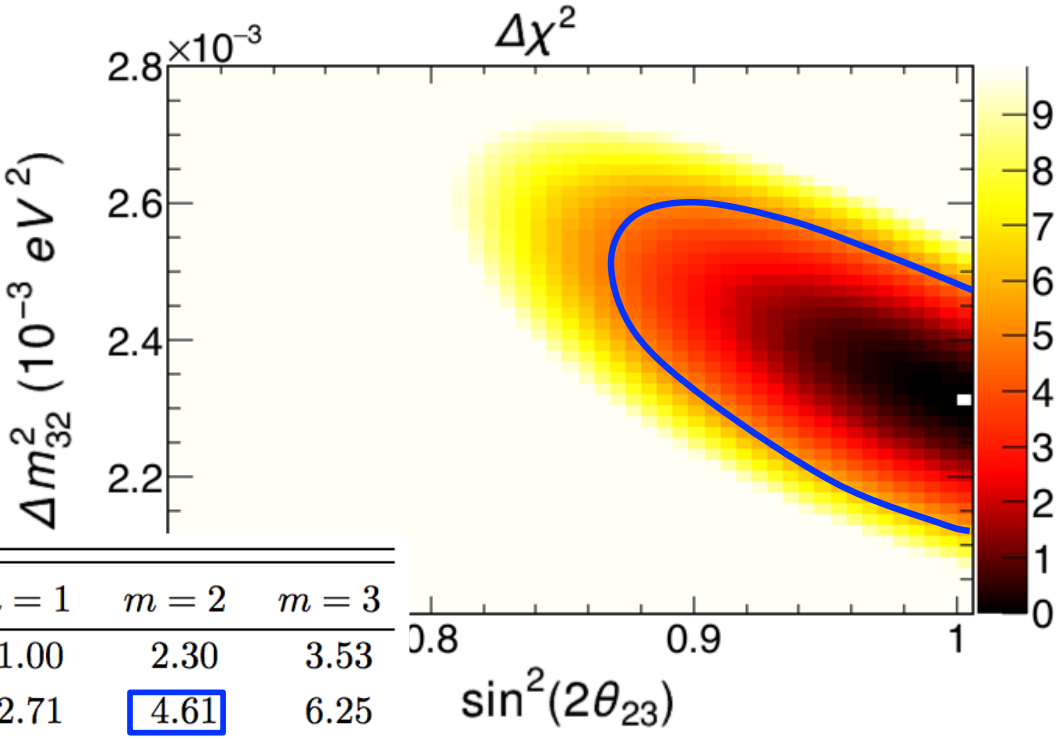


NOvA ν_μ disappearance spectrum as a function of observed energy for quantile 1

Parameter Estimation

- Once the observation is made, the actual physics extraction requires a solving of the inverse problem.
  - The inverse problem is actually very hard…but…

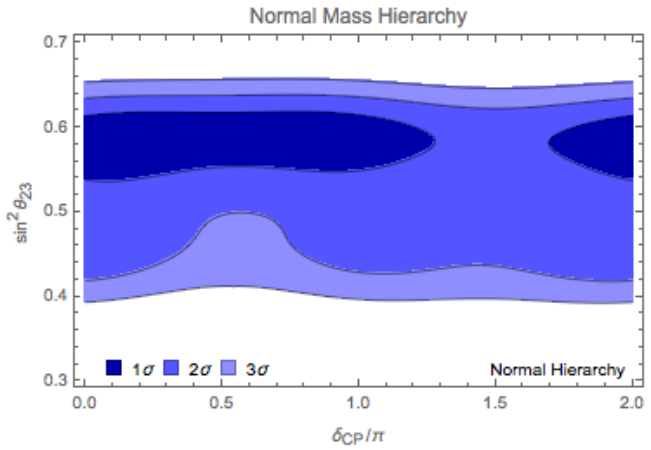
- To first order this is easy
  - Just fit your results using some likelihood function and get a χ^2 surface…go to your Erf table.....
    draw a contour

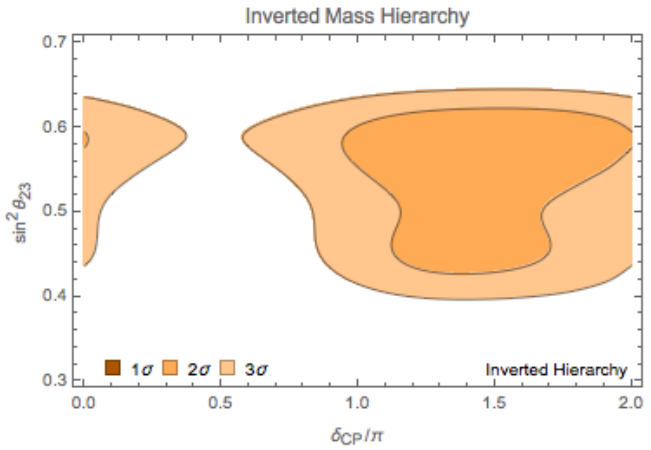


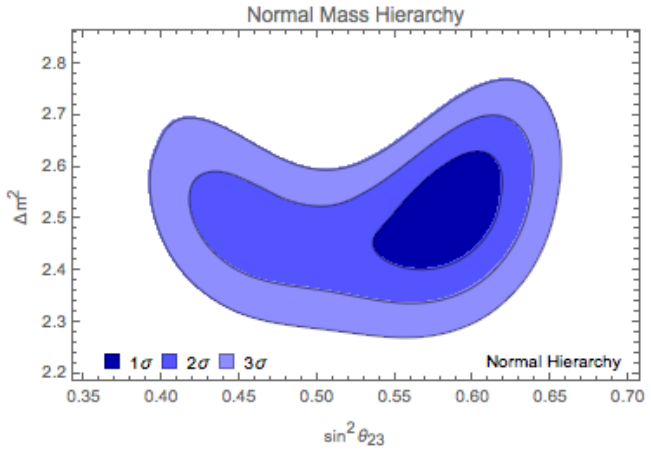
$(1-\alpha)$ (%)	$m=1$	$m=2$	$m=3$
68.27	1.00	2.30	3.53
90.	2.71	4.61	6.25
95.	3.84	5.99	7.82
95.45	4.00	6.18	8.03
99.	6.63	9.21	11.34
99.73	9.00	11.83	14.16

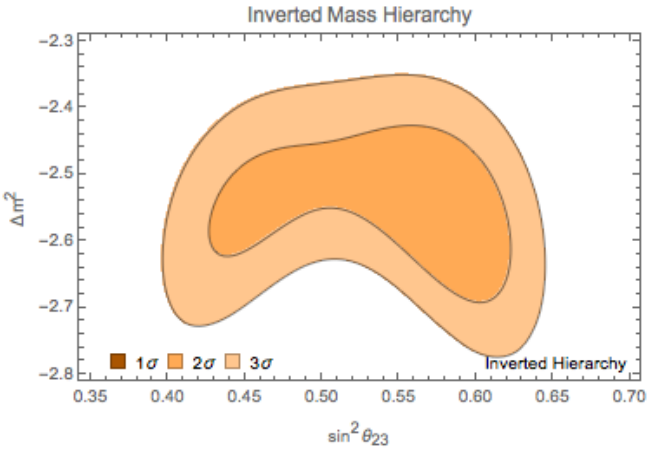
Guassian Contours

- These are the baseline results







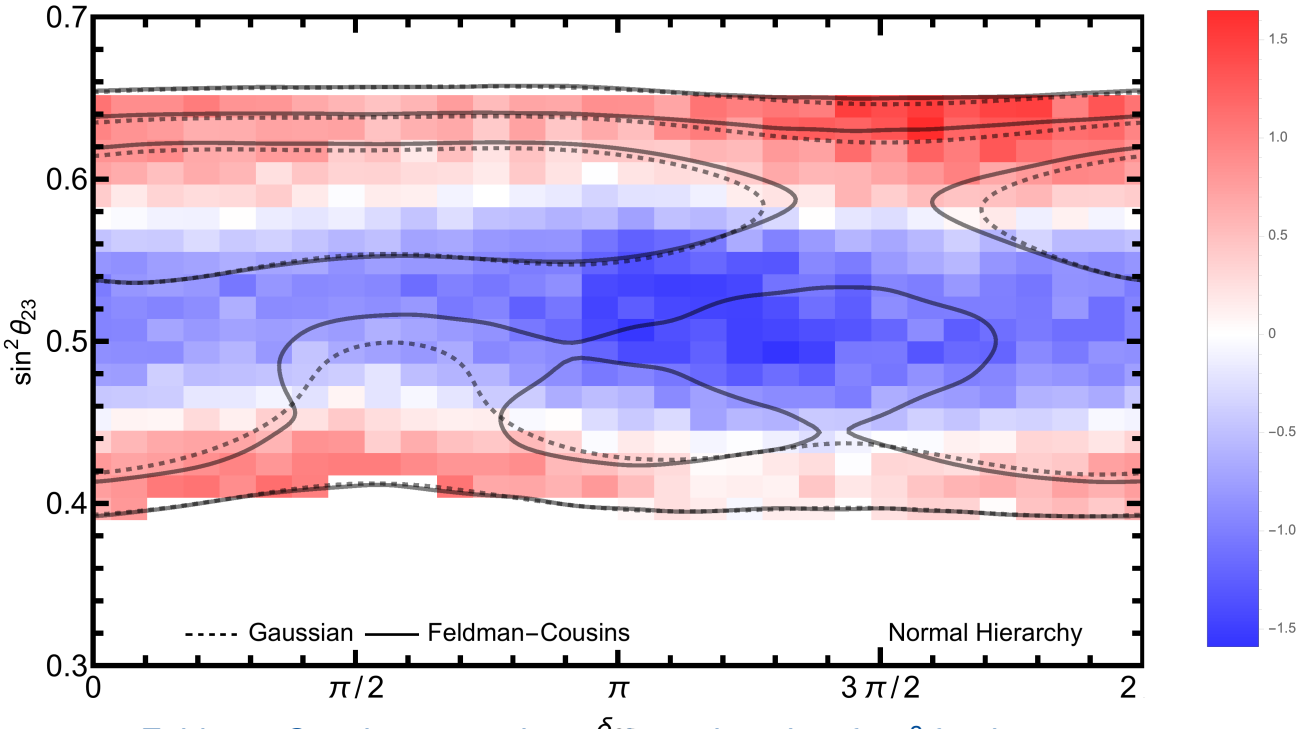


- They may be wrong (very wrong)

Parameter Estimation

- But you need to profile or marginalize over nuisance parameters….
- Understand how correlated your systemmatics really are
- Take into account over/under coverage near physical boundaries

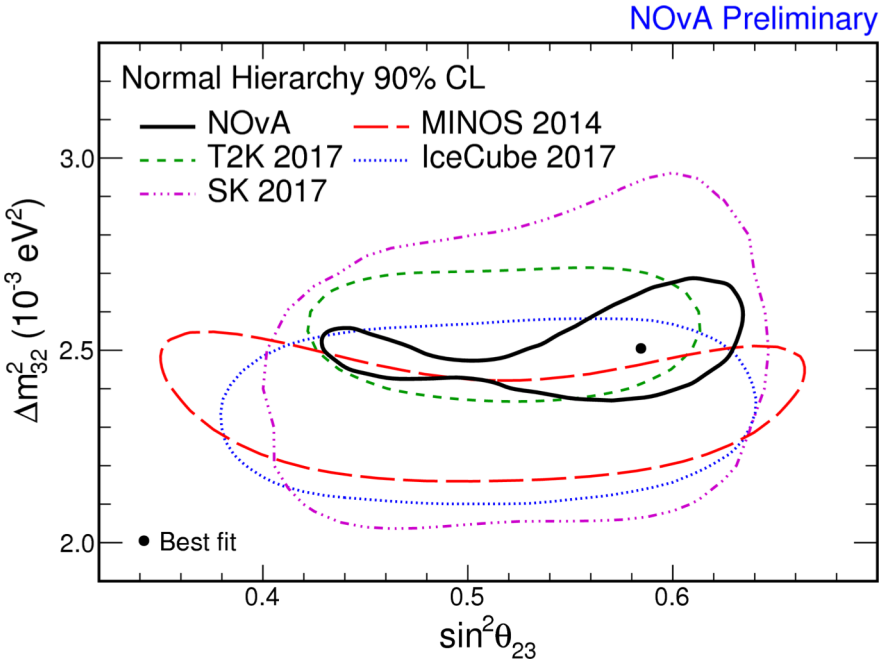
- In general you need to simulate gausian variations in your parameters and Monte Carlo integrate to find corrections your coverage contours
- In “multi-universe” approaches this can consume millions of calculations per grid point
- This one step ≥ all previous computation
  - Solving the analysis at each point in a multi-dim space

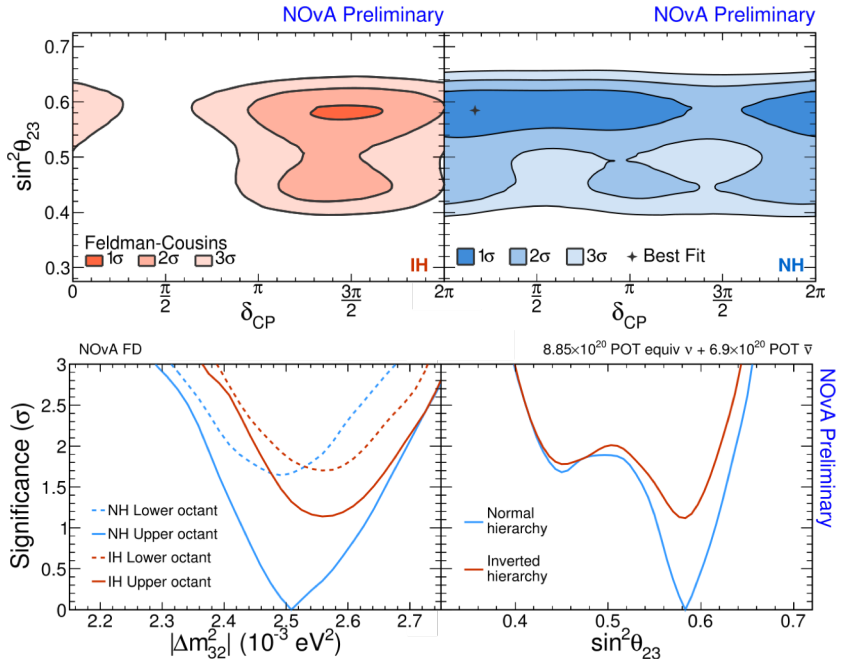


Feldman-Cousins correction surfaces in units of $\Delta\chi^2$ for the 2σ (95%) confidence bands in δ_{CP} and Δm^2_{32}. The 1, 2, and 3 σ Gaussian contours (dashed lines) are shown along with the statistically corrected contours (solid lines).

The Result

- Ultimately you get this:





- Each contour represents ~ 30-50M hours of computing when integrated across each stage. (excluding GPU training)
  - Full long baseline analysis is 200-300M hours of CPU
  - This is spread in across multiple years of calendar time
  - Typical cycle is 2 years for data, 3-4 years for MC (i.e. MC is partially reused)

Take aways…

- Each stage of analysis is similar in “cost”
  - Each stage is different in the type of cost
    - Some are high IO, others CPU, others GPU
  - i.e. there is no single dominate item (so no silver bullets)

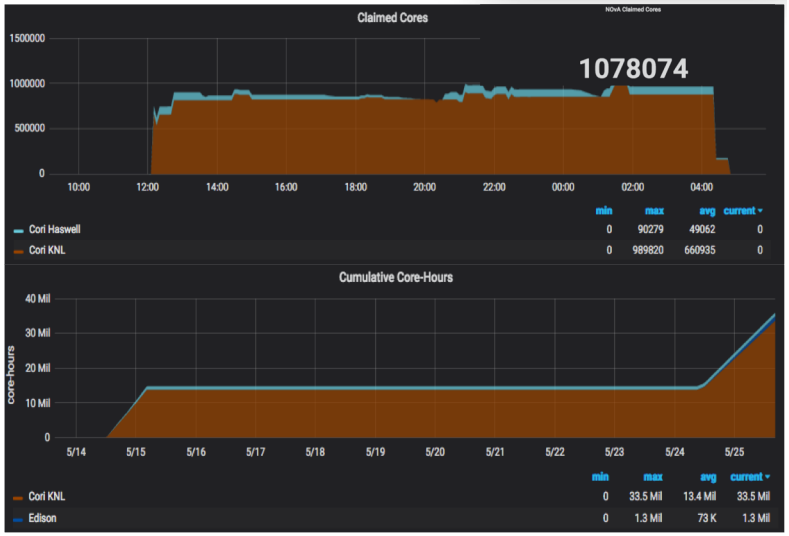
- Most expensive stage currently is reconstruction
  - Expect this to get more costly as new tech is used
- But…driving stage is user level analysis & systematics
  - Effectively duplicates the chain for each variation
  - Most potential for requiring re-reco/alternate reco

- Final analysis stages are extremely expensive

The 2020+ Era

- How will we do this?
- Single technology solutions won't work
  - Need hybrid approach:
    grid + GPU + HPC + ???
    - Traditional grid resources
      are needed for initial data processing
    - GPUs are needed for
      ML techniques
    - HPC are needed for
      large scale selection/fitting/sim
- For DUNE this will work.
  - DUNE is a green field
  - We can do what we need to do

CORI and Edison Supercomputer at NERSC were able to reproduce the 4-week long 2017 NOvA Analysis in a few hours



Over a million cores on CORI and Edison were provisioned and run through the **HEPCloud** facility in two analysis runs in May 2018.

The future