Mateus F. Carneiro

Oregon State University

Nu-Print 2018
Mar 12, 2018

Data Release and preservation in MINERvA

Outline

General Motivation

How does MINERvA analysis work

Current MINERvA Data release

MINERvA Data preservation plans

Outline

General Motivation

How does MINERvA analysis work

Current MINERvA Data release

MINERvA Data preservation plans

MINERvA Run Plan and why we are talking about this

MINERVA continue running in **FY2019** to accumulate **12e20 POT antineutrino**

Expect to have SCD data analysis resources similar to what we have now for two years after the end of data taking

There is broad interest in the collaboration in continuing data analysis beyond 3-4 years from now

Would like to **preserve the ability to analyze data** many years in the future

e.g. to **update** our existing measurements given new information about neutrino interaction models or NuMI fluxes

And to even to do **new analyses**

MINERvA Run Plan and why we are talking about this

We had a data preservation workshop in September

We discussed the many meanings of “data preservation” and various paths that we could pursue:

Start	Title	Author(s)	Topic(s)	File(s)	Length	Edit
09:00	Workshop Introduction	Laura Fields	Software	Fields_DataPre...pdf	00:30	Edit
09:30	DZero Experience with Data Preservation	Non-MINERvA Author	Software	R2DP_Minerva_0...pdf	00:20	Edit
09:50	CDF Data Preservation	Non-MINERvA Author	Software	090817-minerva...pdf	00:20	Edit
10:10	Experience Using Neutrino Data Releases	Patrick Stowell	Physics	NUISANCE-Miner...pdf	00:20	Edit
10:30	Break	None	None	None	00:20	Edit
10:50	Data preservation efforts at CERN	Gabriel N Perdue	Software	data_pres_at_c...pdf	00:20	Edit
11:10	DSTs for Data Preservation	Kevin McFarland	Offline Infrastructure	Presentation (pdf) Presentation (pptx)	00:30	Edit
11:40	Releasing Data as an Image	Jonathan A Miller	Collaboration	ReleasingDataa...pdf	00:20	Edit

Outline

General Motivation

How does MINERvA analysis work

Current MINERvA Data release

MINERvA Data preservation plans

MINERvA from data taking to data release

1. “Keep-up” Production (reformatting of the binary-raw data)
2. Calibrations (deriving the constants for Data)
3. Physics Simulations (for MC)
4. MINOS Inputs to Production
5. Calibration (applying constants) and Reconstruction Production

Done by different groups and provided to users

6. User Analysis Ntuple Production
7. End-user Analysis Production (Results Histograms Production)
8. The data handling spreadsheet

User specific

MINERvA from data taking to data release

1. "Keep-up" Production (reformatting of the binary-raw data)
2. Calibrations (deriving the constants for Data)
3. Physics Simulations (for MC)
4. MINOS Inputs to Production
5. Calibration (applying constants) and Reconstruction Production

Done by different groups and provided to users

Users start from files with **tracks/clusters**, and then do their own reconstruction techniques and event selection using a **Gaudi Tool**

6. User Analysis Ntuple Production
7. End-user Analysis Production (Results Histograms Production)
8. The data handling spreadsheet

User specific

MINERvA from data taking to data release

1. "Keep-up" Production (reformatting of the binary-raw data)
2. Calibrations (deriving the constants for Data)
3. Physics Simulations (for MC)
4. MINOS Inputs to Production
5. Calibration (applying constants) and Reconstruction Production

Done by different groups and provided to users

Users use their produced tuples and use the **PlotUtils** package to get to a cross section while dealing with the systematic errors

6. User Analysis Ntuple Production
7. End-user Analysis Production (Results Histograms Production)
8. The data handling spreadsheet

User specific

Outline

General Motivation

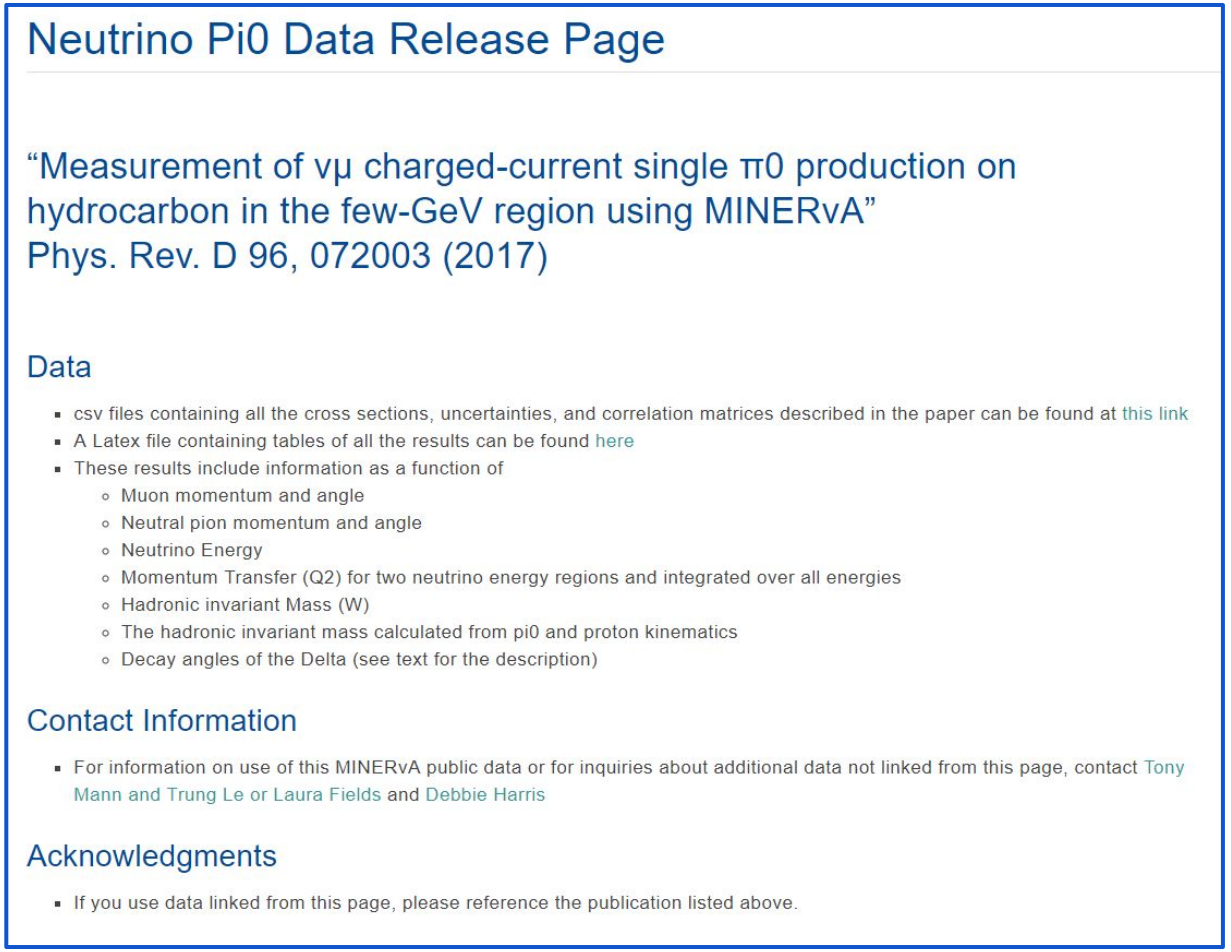
How does MINERvA analysis work

Current MINERvA Data release

MINERvA Data preservation plans

Current MINERvA data release

MINERvA currently aim to have available **csv files containing all the cross sections, uncertainties, and correlation matrices described in the paper**



Neutrino Pi0 Data Release Page

"Measurement of νμ charged-current single π0 production on hydrocarbon in the few-GeV region using MINERvA"
Phys. Rev. D 96, 072003 (2017)

Data

- csv files containing all the cross sections, uncertainties, and correlation matrices described in the paper can be found at this link
- A Latex file containing tables of all the results can be found here
- These results include information as a function of
  - Muon momentum and angle
  - Neutral pion momentum and angle
  - Neutrino Energy
  - Momentum Transfer (Q2) for two neutrino energy regions and integrated over all energies
  - Hadronic invariant Mass (W)
  - The hadronic invariant mass calculated from pi0 and proton kinematics
  - Decay angles of the Delta (see text for the description)

Contact Information

- For information on use of this MINERvA public data or for inquiries about additional data not linked from this page, contact Tony Mann and Trung Le or Laura Fields and Debbie Harris

Acknowledgments

- If you use data linked from this page, please reference the publication listed above.

Current MINERvA data release

During the workshop we had feedback from Patrick Stowell (part of the NUISANCE group) on “The good, The bad and The ugly” of our data release

The good:

Easy access

MINERvA always releases **full covariance information** (including nu/nubar cross-correlations for some results)

Current MINERvA data release

During the workshop we had feedback from Patrick Stowell (part of the NUISANCE group) on “The good, The bad and The ugly” of our data release

The bad:

Signal definition needs to be crystal clear

Data release + paper should be **self contained** and enough

Updates on already available datasets need to be more clear and consistent

Providing a table of **MC points** is also important in the data release.

Always provide both **restricted and full phase space** distributions

Current MINERvA data release

During the workshop we had feedback from Patrick Stowell (part of the NUISANCE group) on “The good, The bad and The ugly” of our data release

The Ugly:

“No-one has ever really given us a **covariance** that covers **multiple distributions** from the same measurement”

“Experiments finding a proper way to release **N-tuple data** that we can apply parameterized detector smearing to could be very powerful.”

”**HepData** project at Durham wants to also start including neutrino cross-section data.”

Outline

General Motivation

How does MINERvA analysis work

Current MINERvA Data release

MINERvA Data preservation plans

Data preservation plan Goals

1. Data & simulation **DST-style ntuples** + **tools to reproduce existing MINERvA analyses or perform new analyses** with those Ntuples, publicly available and documented sufficiently that a trained experimental neutrino physicist can successfully use them

2. **Tools for the general public** to use the data/simulation in specific, well defined ways

3. A mechanism for people to produce **new ntuples to replace all or part of the simulation** ntuples given new models

4. Tools for **turning tuples into 'images'** for Machine Learning research

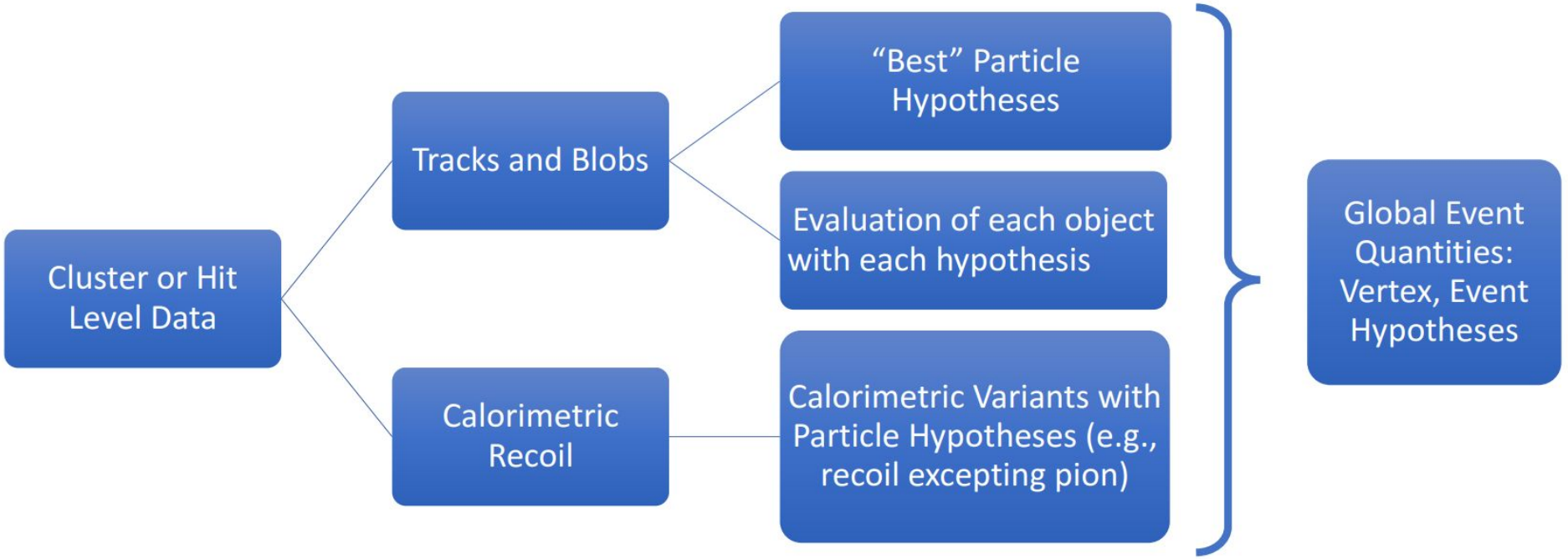
Data preservation current plan steps

1. Finish our current reconstruction work plan in **Gaudi**
   - Forward MINOS tracking
   - Neutrons
   - Move framework to SL7 (?)

2. Develop **one analysis ntuple** to rule them all
   - Would include sufficient information to reproduce all of our **active analyses**
   - Would also have **full event record** (hits, clusters, tracks, MC Truth, etc…)
   - Would ideally be readable in **stand-alone root**, and not require MINERvA-specific libraries that we have to maintain

Data preservation current plan steps

2. Develop **one analysis ntuple** to rule them all



But how to deal with changes in the "standard" MC model in the field?

reweight? regenerate?

Data preservation current plan steps

3. Divorce **PlotUtils** from framework

It becomes a set of tools for transforming tuples into cross sections

PlotUtils will have to be updated to deal with future versions of root, or we have to force anyone using the tuples to use a fixed version of root

4. Make **tuples** for our full data sample, **simulation, PlotUtils libraries, executable** publicly available

Along with documentation: step-by-step examples of a few analyses

Data preservation current plan steps

5. Proposed plan for ML data preservation

HDF5 is the standard format for data for use in python

Metadata contain information **beam, weight, configuration, event, dataset and etc**

Images would contain **standard images** of time and energy

MINOS would contain the **MINOS information** as an array of floats.

Simulation would contain all of the **produced particles and kinematic information**.

Classes would contain a **short array of interesting classes** for the ML community

Current Status

Refine some of the details

Computing resources to host datasets, libraries

People to execute the plan

Need grid resources for making analysis ntuples

Submit proposal(s) requesting funds for what we need Likely a multi-institution proposal

Discussion time!

Is there a need to adopt a **standard data release format**? E.g. Hep data used by many collider experiments: https://hepdata.net/

What **issues** have been found with past data releases, and how were they resolved?

What **technical challenges** have been faced by those releasing data?

As data releases become more complex, what can be done to ensure data is used correctly?

How do we best ensure that **comparisons to data are done rigorously and correctly**, e.g., with correct treatment of correlated uncertainties and final state definitions?