Data and Analysis Preservation

Tom Junk

Fermilab

Path to Dark Sector Discoveries at Neutrino Experiments

June 5, 2023

T. Basaglia *et al.*, Data Preservation in High Energy Physics – DPHEP Global Report 2022
https://arxiv.org/abs/2302.03583

T. Junk and L. Lyons, Reproducibility and Reproduction of Experimental Particle Physics Results
Harvard Data Science Review, Fall 2021
https://hdsr.mitpress.mit.edu/pub/1lhu0zvn/release/4?readingCollection=c6cf45bb

Definitions

- *Reproduction* of a result:

  Start with the experimental data, simulation, assumptions, analysis tools ("digital artifacts") and recompute the results. "Computational Reproducibility" A necessary but insufficient criterion for reliability.

- *Replication* of a result:

  Obtaining consistent results across studies aimed at answering the same scientific question. Examples: re-run the experiment and collect new data, build a similar experiment.

Following the convetion of National Academies of Sciences, Engineering, and Medicine. (2019). *Reproducibility and replicability in science*. The National Academies Press. https://doi.org/10.17226/25303

Definitions

- *Recasting* of an analysis:

  Analysis artifacts are almost complete – stacked histograms with systematic uncertainties already evaluated. Add a new signal model and re-do the exclusion calculation and/or *p* value calculation

  https://iris-hep.org/projects/recast.html

- *Re-Use* of data:

  This is what experiments do to publish different physics analyses with the same data. Often different trigger streams are used by different analyses – these do not constitute re-use.

  Example: measure the ttbar cross section at the LHC, and use the same data to measure ttH production rates.

  New students and postdocs starting an analysis on a collaboration usually spend some time reproducing earlier work.

FAIR Principles

- Findable
- Accessible
- Interoperable
- Re-Usable

Data, metadata, databases, software, adequate environments and knowledge are required for data and analysis preservation and re-use.

Documentation helps with knowledge transfer, but it can be imperfect. Even perfect documentation can be ignored.

Steps and Ingredients in an Analysis

Not including experiment and analysis design

- **Input Raw Data** -- usually high-dimensional data from multiple detectors and subsystems with different technologies – Input.

- **Monte Carlo Simulations** of Detector Response for signals and backgrounds

- **Data calibrations and Monte Carlo adjustments**
  - Data are calibrated to make them more useful
  - Monte Carlo samples are adjusted to make them match the data better
    - You don't change data to match MC!
    - It is easier to smear MC than to unsmear it. Though the latter is possible and has been done!
    - It is easier to smear MC than to generate and simulate new MC samples.

- **Event selection** – signal and control samples

- **Statistical analysis**
- **Systematic uncertainty estimation and propagation to final results.**

} Often people are interested in reproducing just these steps. Perhaps that's all that *can* be done with available data.

Different Levels of Reproduction/Re-Use

1) Start from scratch – Raw data → Final Results

- The gold standard – exercises the entire data analysis and scientific result chain
- Experiments have to do this routinely anyway. Graduate students graduate, and postdocs move on. New personnel must pick up where they left off.
- Usually this is limited to collaboration members – I'll explain why
- It would be great if collaborations could be this transparent all the way through!

2) Start with processed, calibrated raw data

- Easier than 1), but still challenging
- Within collaborations, the calibrations and initial reco are shared among all experimenters so these tedious steps do not have to be re-done by non-experts.
- Models (generators, simulation, reconstruction, adjustments, tuning) still needs to be reproduced.
- Lots of data may be irrelevant to an analysis and simply is cut out. Don't have to understand cosmic rays when measuring H→bb for example.

Different Levels of Reproduction/Re-Use

3) Analyze processed and selected data and MC events provided by experiments as open data

- A better distillation of physically-relevant information
- This step can be lossy. Acceptance for previously unforeseen signals can be low.
- People with new ideas will need new MC samples, at least for signals.
- Many instances of this in the literature

4) Read data off of published histograms and analyze them

- Rather common
- Also lossy – need signal models and detector simulation
- Requires detailed understanding of systematic uncertainties

Different Levels of Reproduction/Re-Use

5) Use Published Likelihood Functions

- Advocated in the proceedings of PhyStat 2000, the First Workshop on Confidence Limits

https://cds.cern.ch/record/411537?ln=en

- Assumes a signal model under test.
- Needs to be a function of all considered nuisance parameters (sources of systematic uncertainty)
- Maximizing ("profiling") or integrating out ("marginalizing") systematic uncertainties are lossy steps. Need to retain this information in order to combine with similar measurements.
- Limited to signal models considered when making the results.
- Still, not happening frequently enough

Different Levels of Reproduction/Re-Use

6) Exchange Likelihood Functions between experiments and combine without publishing the internals.

Tevatron Higgs combinations were done this way.

Really a digital version of **#4** with standard representations of systematic uncertainties

- Signal, background and data histograms for each channel
- Systematic uncertainties: rate and shape uncertainties on each model component, identified by named source

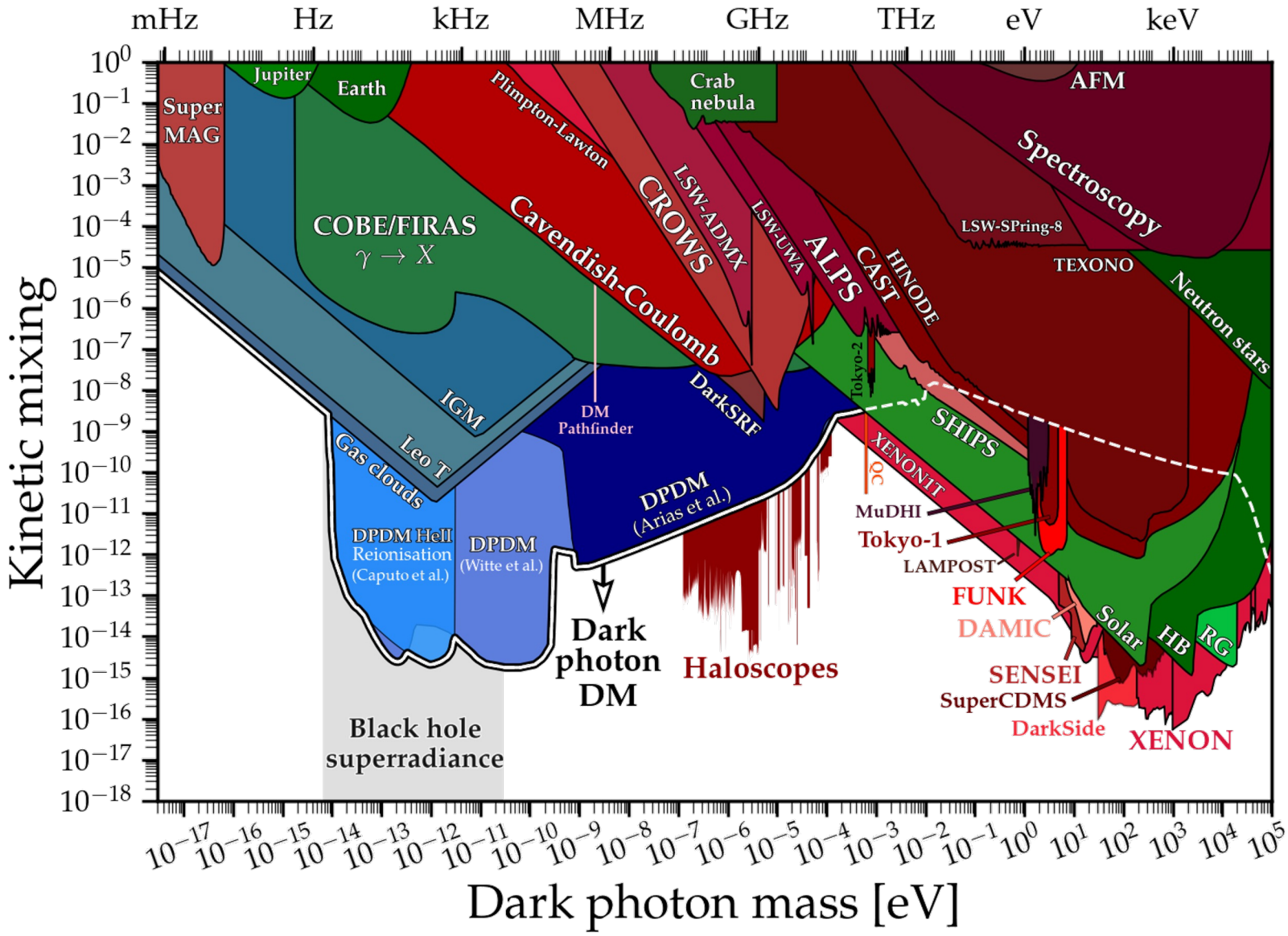
Experiments must agree on conventions/names/central values for shared nuisance parameters

Different Levels of Reproduction/Re-Use

7) Use Published Unfolded Cross Sections and Confidence Intervals

- This is what experimenters usually aim to produce
- Justification is that consumers of results ought not to be bothered with things like detector acceptance, resolution, or individual systematics
- Very lossy and model dependent!
- Usually restricted to low-dimensional representations

Example of #7 – overlay exclusions. Slide shown by B. Giaccone at the FNAL JETP Seminar, May 26, 2023



https://cajohare.github.io/AxionLimits/docs/dp.html

Caputo et al., https://arxiv.org/abs/2105.04565

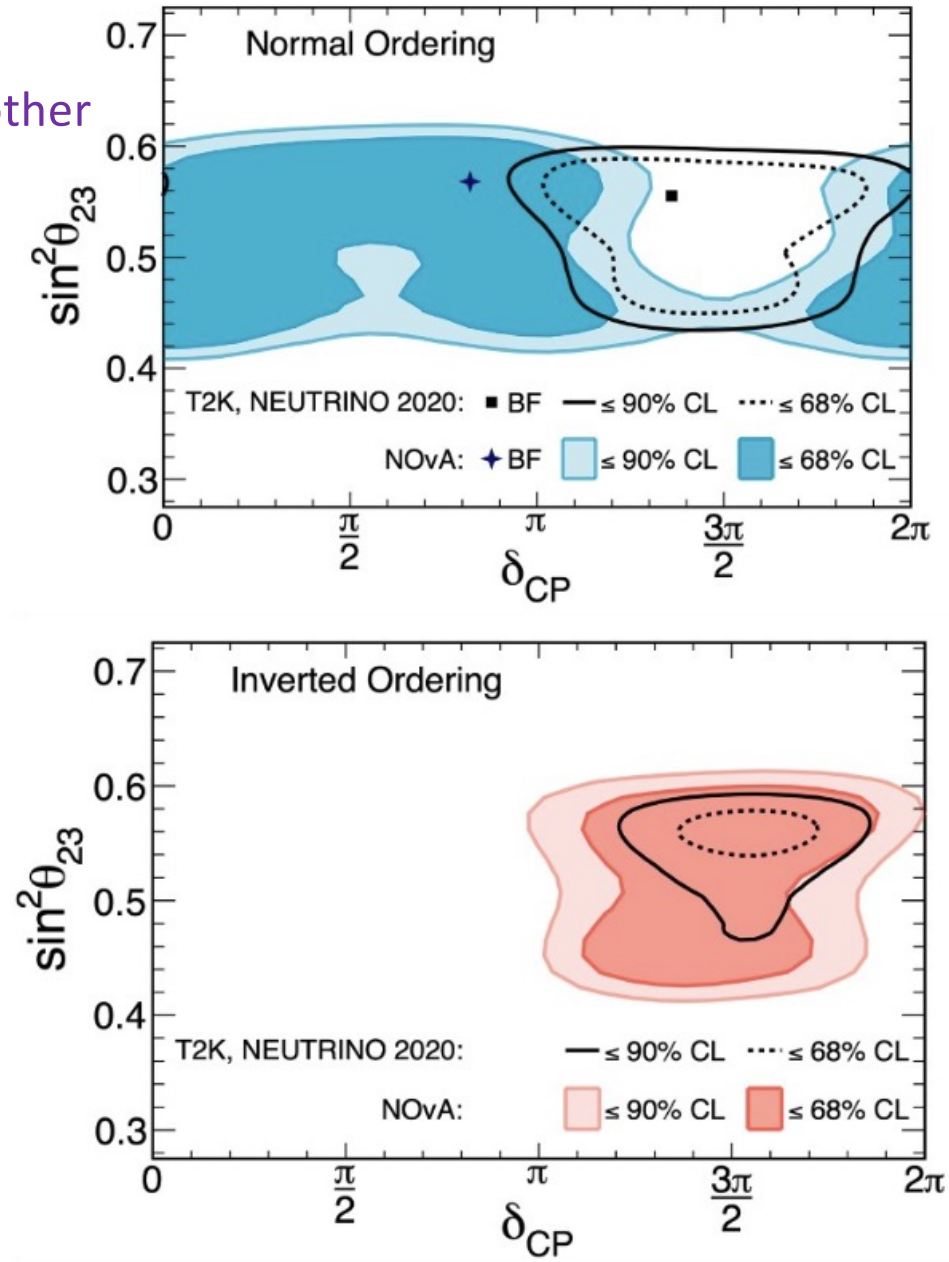
Example of #7 – plot the published intervals from different experiments on top of one another

T2K and NOvA

$\sin^2\theta_{23}$ vs δ_{CP} confidence regions produced by both experiments

Simplest visualization – just overlay the regions.

Lots of questions arise when doing this.



P. Dunne, Latest neutrino oscillation results from T2K, 10.5281/zenodo.3959558(2020).

A. Himmel, New oscillation results from the NOvA experiment, 10.5281/zenodo.3959581(2020).

T2K-NOvA Joint Analysis Workshop, Nov. 2021 https://indico.fnal.gov/event/51305/

The 2020 Global Reassessment of the Neutrino Oscillation Picture

P. F. de Salas, D. V. Forero, S. Gariazzo, P. Martínez-Miravé, O. Mena, C. A. Ternes, M. Tórtola, J. W. F. Valle
J. High Energ. Phys. 2021, 71 (2021)
https://doi.org/10.1007/JHEP02(2021)071

Example of #4

In order to perform our analysis, we extract the relevant data for each experiment from the corresponding reference. We simulate the signal and background rates using the GLoBES software [76, 77]. For the energy reconstruction we assume Gaussian smearing. We include bin-to-bin efficiencies, which are adjusted to reproduce the best-fit spectra reported in the corresponding references. Finally, for our statistical analysis we include systematic uncertainties, related to the signal and background predictions, which we minimize over.

Example joint result:

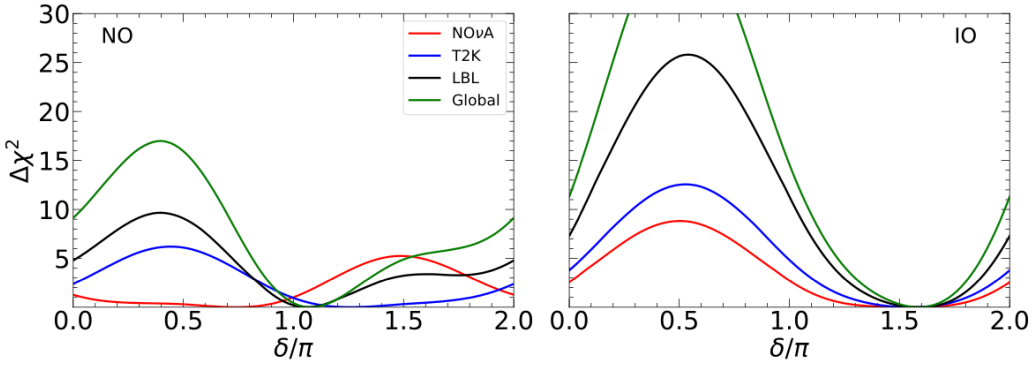


Figure 8. $\Delta\chi^2$ profiles for δ obtained from the analysis of NOνA (red), T2K (blue), all long-baseline data (black) and from the global fit (green).

Jesse Thaler's Colloquium at Fermilab on Sep. 30, 2020

Example of #3: Use reconstructed, selected open data

https://events.fnal.gov/colloquium/events/event/no-colloquium-9/

Includes video! Slides are available at:

https://indico.cern.ch/event/882586/contributions/4042612/attachments/2112093/3555143/jthaler_2020_09_OpenData_FermilabColloquium.pdf

CMS Open Data Workshop for Theorists at the LPC, Sep. 30, 2020

https://indico.cern.ch/event/882586/timetable/?view=standard

A. Tripathee et al., *Phys.Rev.D* 96 (2017) 7, 074003 e-Print: 1704.05842 [hep-ph]

A slide from Jesse Thaler's Sep 30, 2020 Colloquium

In Backup

"Researching physics in and beyond the Standard Model"

All ~~13~~ 19 papers (thus far) using CMS Open Data

Standard Model Analyses

[Tripathee, Xue, Larkoski, Marzani, JDT, PRL 2017, PRD 2017]
[Apyan, Cuozzo, Klute, Saito, Schott, Sintayehu, JINST 2020]

BSM Searches

[Cesarotti, Soreq, Strassler, JDT, Xue, PRD 2019]
[Lester, Schott, JHEP 2019]

Machine Learning Studies

[Fernández Madrazo, Heredia Cacha, Lloret Iglesias, Marco de Lucas, EPJWoC 2019]
[Andrews, Paulini, Gleyzer, Poczos, CSBS 2020]
[Andrews, Alison, An, Bryant, Burkle, Gleyzer, Narain, Paulini, Poczos, Usai, NIM 2020]
[Moreno, Nguyen, Vlimant, Cerri, Newman, Periwal, Spiropulu, Duarte, Pierini, PRD 2020]
[Knapp, Dissertori, Cerri, Nguyen, Vlimant, Pierini, arXiv 2020]

And More!

[Pata, Spiropulu, arXiv 2019]
[Paktinat Mehdiabadi, Fahim, JPG 2019]
[Komiske, Mastandrea, Metodiev, Naik, JDT, PRD 2020]

Please contact me if I missed your CMS Open Data study!

Contact Jesse

The MIT Group's Analysis of Jet Substructure using Open CMS Data

Publicly available AOD Files – Analysis Object Data. 2.0 TB of Jet Primary Dataset data.

Data available via XRootD

Contains low-level reco objects such as tracks and clusters, and high-level objects such as jets.

MIT group ran CMSSW in a virtual machine, extracting only the items needed for their analysis. More convenient to extract data in private format.

Text-file format used by the MIT group called "MOD". MODProducer software is available on GitHub.

https://github.com/ tripatheea/MODProducer

The Regulatory Landscape (at least in the United States)

https://www.science.gov/publicAccess.html

https://new.nsf.gov/public-access

https://www.energy.gov/datamanagement/doe-policy-digital-research-data-management

https://energy.gov/downloads/doe-public-access-plan.pdf

Example implementations:

DUNE Data Management Plan: DUNE-Doc-5759-v3

CMS Data Management Plan:

https://uscms.org/uscms_at_work/data_computing/data_management/index.shtml

https://www.energy.gov/datamanagement/doe-policy-digital-research-data-management

Requirements and Guidance from DOE Sponsoring Offices

All DMPs submitted to any DOE sponsoring office should meet the following requirements:

•DMPs should describe whether and how data generated in the course of the proposed research will be shared and preserved and, at a minimum, describe how data sharing and preservation will enable validation of results, or how results could be validated if data are not shared or preserved.

•DMPs should provide a plan for making all research data displayed in publications resulting from the proposed research open, machine-readable, and digitally accessible to the public at the time of publication. This includes data that are displayed in charts, figures, images, etc. In addition, the underlying digital research data used to generate the displayed data should be made as accessible as possible to the public in accordance with the Principles stated above. The published article should indicate how these data can be accessed.

•DMPs should consult and reference available information about data management resources to be used in the course of the proposed research. In particular, DMPs that explicitly or implicitly commit data management resources at a facility beyond what is conventionally made available to approved users should be accompanied by written approval from that facility. In determining the resources available for data management at DOE Scientific User Facilities, researchers should consult the published description of data management resources and practices at that facility and reference it in the DMP.

•DMPs must protect confidentiality, personal privacy, Personally Identifiable Information, and U.S. national, homeland, and economic security; recognize proprietary interests, business confidential information, and intellectual property rights; avoid significant negative impact on innovation and U.S. competitiveness; and otherwise be consistent with all applicable laws, regulations, agreement terms and conditions, and DOE orders and policies.

A Typical Experiment's Data Management Policy

- **Raw Data:**
  - DAQ files, metadata and conditions/configuration database values
  - Precious. Replicated for safety.
  - Access limited to collaboration members.
  - Some experiments even restrict access to a subset of collaborators for purely practical reasons
  - Raw data can be disseminated with collaboration approval
  - Retained until the dissolution of the collaboration, and further retained to meet funding agency requirements.
- **Analysis Data:**
  - Includes Monte Carlo (with truth labels!)
  - Calibration databases
  - Processed data files – signal processing, hit finding, pattern recognition and further steps
  - Data to be made available to all collaborators
  - Distribution of Analysis Data requires collaboration approval
- **Published results:**
  - Properly archived
  - Machine-readable versions of plot data provided along with the publications.

Many Neutrino Analyses Use Raw Data (or something close to it)

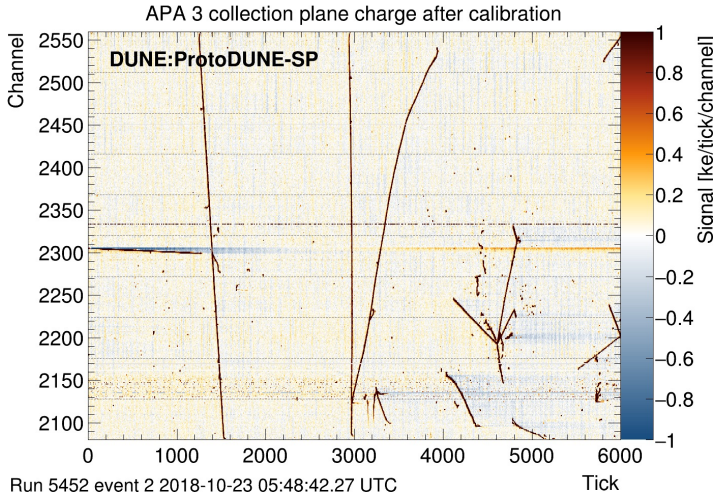
- Neutrino detectors produce fine-grained 2D or 3D images of particle interactions
- Neutrino detectors are also the target material.
  - They are simultaneously trackers and calorimeters (liquid argon, water, oil, steel and plastic are common materials)
  - Many are not magnetized, but some are
  - Images are mostly empty, but have locally high-density regions where several particles are emitted together.
  - Showers can make a very thick spray of particles

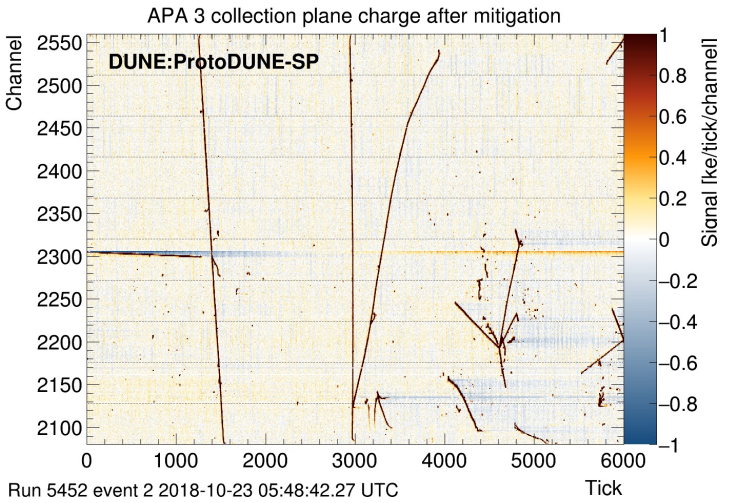
- Convolutional Neural Networks (CNNs) are being used for many headline neutrino analyses

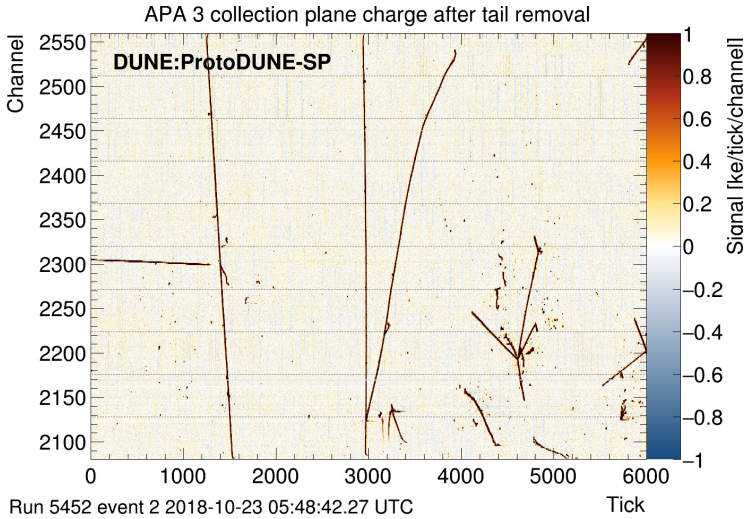
See, for example,
DUNE Collab: *Phys.Rev.D* 102 (2020) 9, 092003 e-Print: 2006.15052 [physics.ins-det]
DUNE Collab: *Eur.Phys.J.C* 82 (2022) 10, 903 e-Print: 2203.17053 [physics.ins-det]

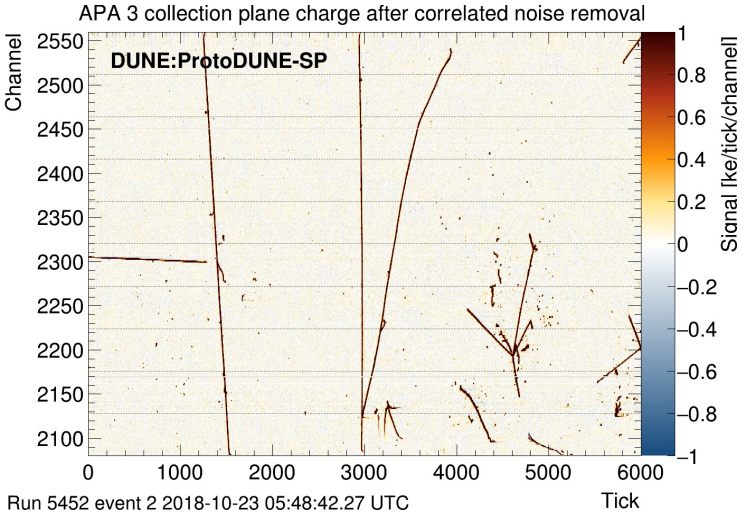
for nueCC identification and track/shower separation using CNNs.

Example Data Preparation: DUNE's ProtoDUNE-SP





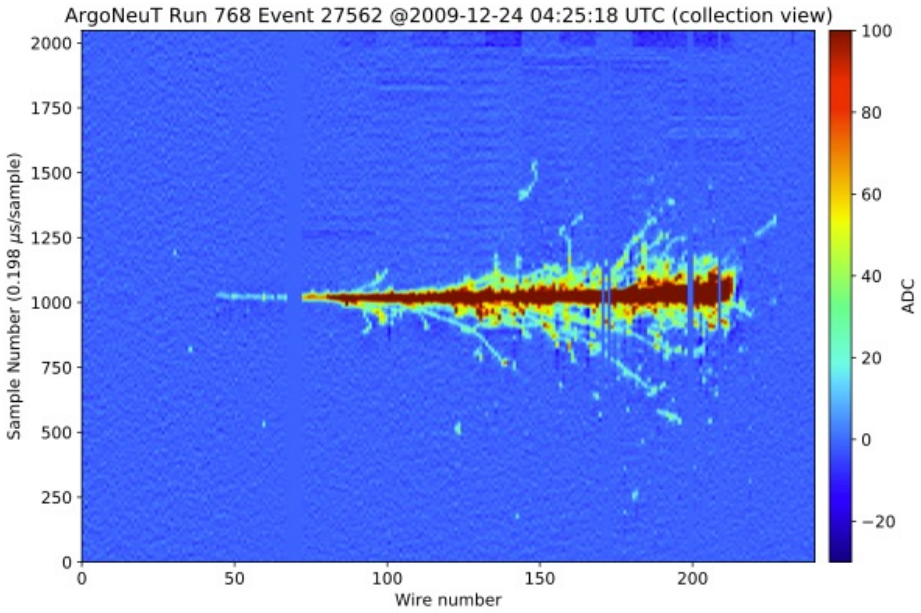


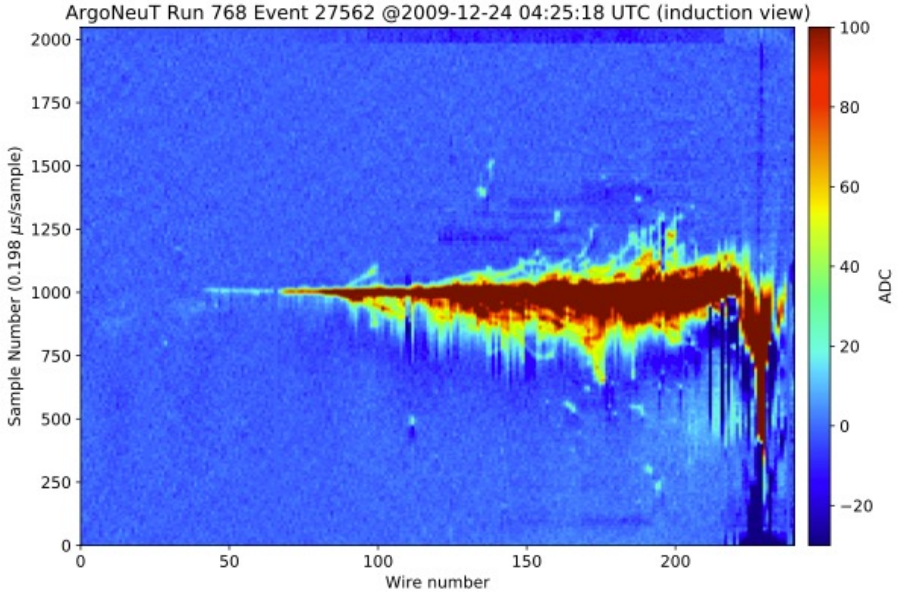


DUNE Collab., *JINST* 15 (2020) 12, P12004 e-Print: 2007.06722

Published Images of Raw Data

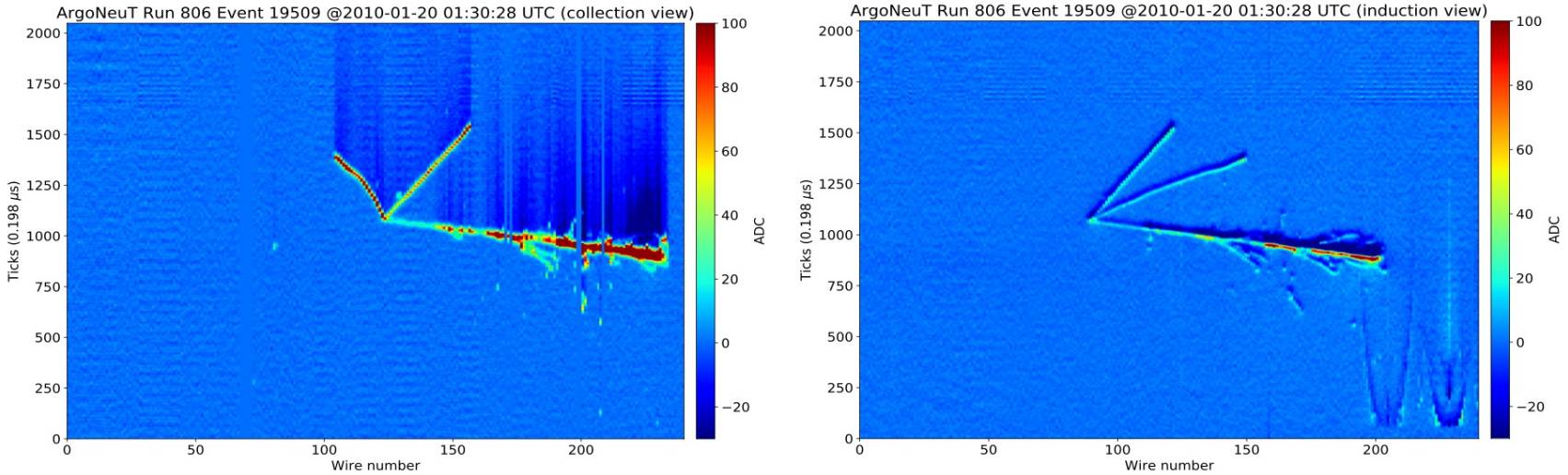
ArgoNeuT Collab., *Phys.Rev.D* 102 (2020) 1, 011101 e-Print: 2004.01956 [hep-ex]



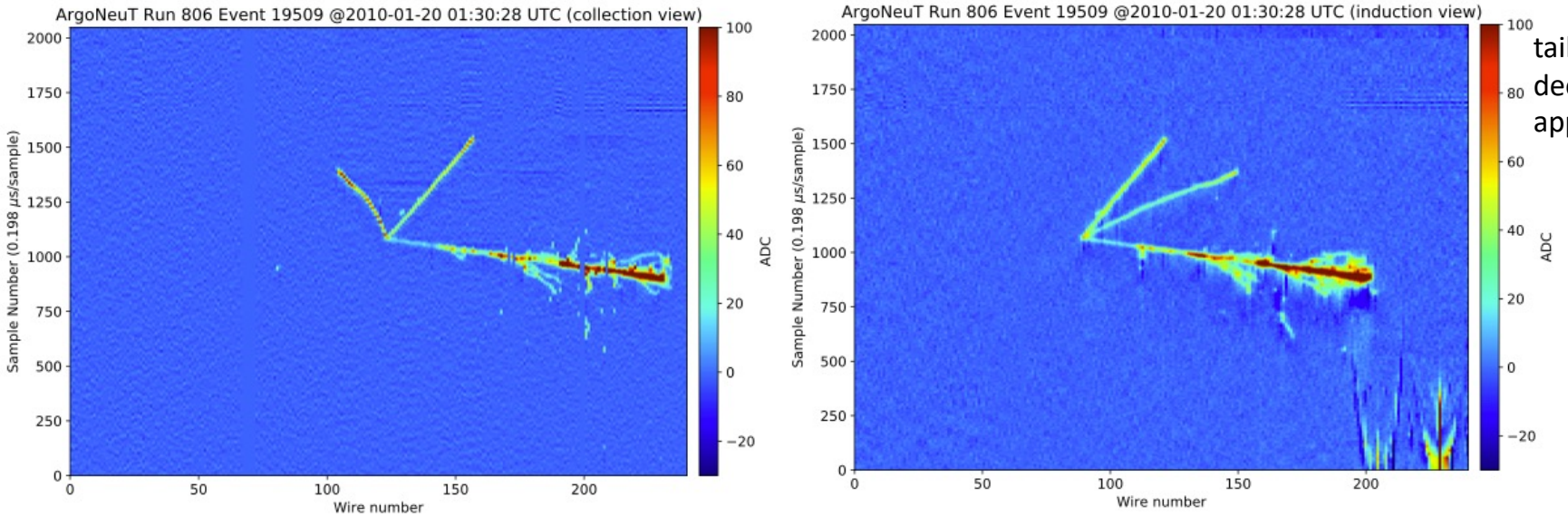


Various artifacts are visible: Noise, field distortions, long-range induction, non-containment of shower, and early collection of charge (bias cards external to the field cage are suspected sources of stray charge collected on induction plane channels)

Two Published Versions of the Same ArgoNeuT Event



The ArgoNeuT Collaboration, 2022 JINST 17 P01018

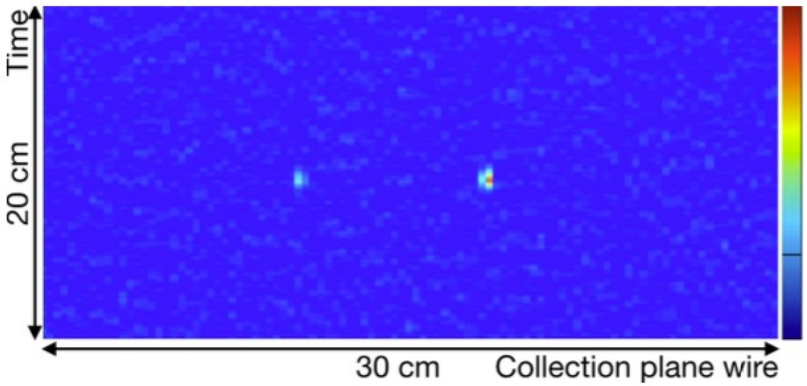


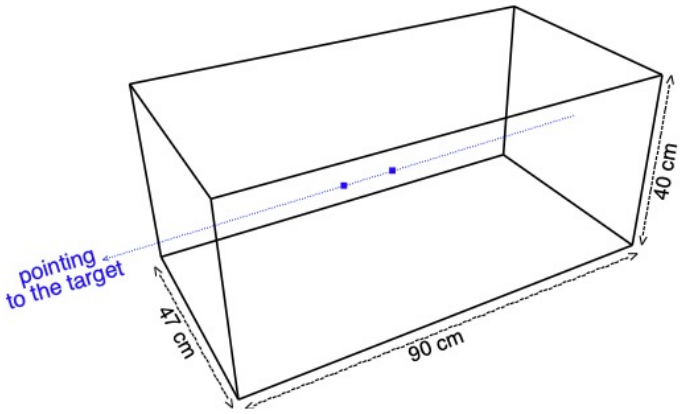
tail removal, deconvolution applied

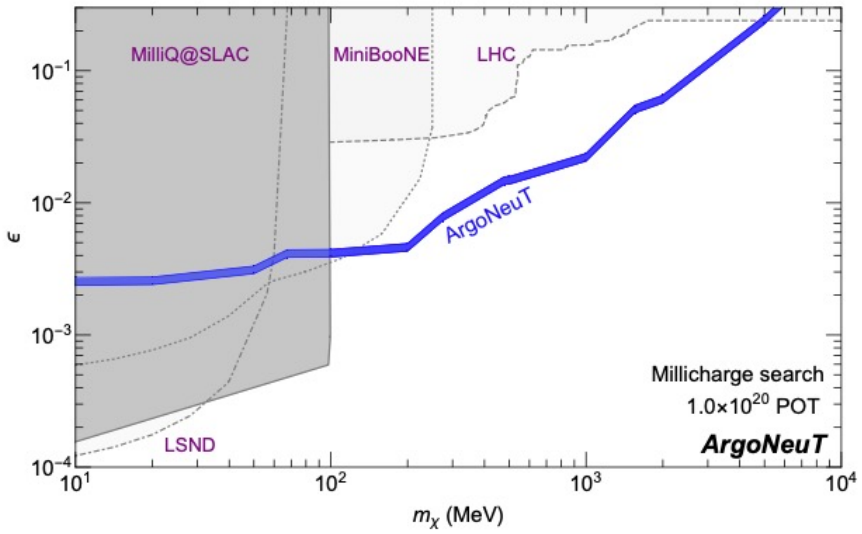
ArgoNeuT Collab., *Phys.Rev.D* 102 (2020) 1, 011101 e-Print: 2004.01956 [hep-ex]

An Interesting Analysis – Millicharged Particles in ArgoNeuT

Phys.Rev.Lett. 124 (2020) 13, 131801 e-Print: 1911.07996 [hep-ex]







The candidate signal event. Double-blip event. LArTPCs have *lots* of blips, only some doubles that point properly at the beam source.

MicroBooNE's Public Data Sets

See Giuseppe Cerati's presentation at CHEP 2023
https://indico.jlab.org/event/459/contributions/11677/

Dataset definitions

Each HDF5 sample comes in two flavors: with and without wire information (waveform).
Due to size requirements, sample with this information contain less events.

Sample	DOI	N events	N HDF5 files	HDF5 size	N artroot files	artroot size
Inclusive, NoWire	10.5281/zenodo.7261798	141,260	20	34 GB	3400	787 GB
Inclusive, WithWire	10.5281/zenodo.7262009	24,332	18	44 GB	720	136 GB
Electron neutrino, NoWire	10.5281/zenodo.7261921	89,339	20	31 GB	2151	761 GB
Electron neutrino, WithWire	10.5281/zenodo.7262140	19,940	20	39 GB	540	170 GB

Open BNB inclusive sample is a subsample of what internally available. We may open a larger sample upon request and if technically feasible.

Challenges Faced When Preserving Data and Analyses

Storing digital artifacts has never been easier!

Preserving analyses also benefits from software containers – you can restore a decades-old computing environment on modern hardware without additional worries about security.

Even providing access to data (in small amounts, of order TB or less) is quite easy. Bigger samples require some coordination.

Some real issues:

- Collaboration review and approval
- The need to provide full documentation and bulletproof tools
- Tools should be convenient enough that non-experts who may get frustrated easily can still use them.
  - patience + stubbornness = perseverance
  - graduate students and postdocs usually have the most time to devote to developing expertise
- Expert hand-holding when things go wrong (and they will!)

Detector and Analysis Complexity are Challenges

- Great need to automate procedures to make sure nothing gets forgotten

- Detectors are imperfect
  - broken, shorted or simply missing channels
  - aging – time-dependent response
  - Detectors are also upgraded and improved over time!
  - Look for a signal that requires hermeticity – such as a missing energy
    measurement – need to know where all the cracks and holes are in the detector
  - Some "cracks" come from physics – unmeasured neutrinos and neutrons for example.
- I remember a CHAMP analysis on OPAL that mistakenly selected some Z^0 calibration events that were mixed in with higher-sqrt(s) data.
  These things are obvious when you know what you're looking for.

Blind Analyses and HARKing

- A collaboration may be more comfortable if an analysis is done "blind"
  J. Klein and A. Roodman, *Ann.Rev.Nucl.Part.Sci.* 55 (2005) 141-163
- Several techniques, most of which involve hiding the "important" parts of the data until the analysis is finalized.
- Adjusting the hypothesis after data and features have been selected invalidates classical inference. Todd Kuffner at Phystat-Nu 2016: https://indico.fnal.gov/event/11906/contributions/10634/attachments/6998/9048/talk_2016_PhyStat.pdf

- HARKing – Hypothesizing After Results are Known is a real problem
  - the real problem with HARKing is the suppression of the original hypothesis and presentation of the post-hoc hypothesis as the prior hypothesis
  - Original publications remain valid
  - https://pubmed.ncbi.nlm.nih.gov/15647155/

Blind Analysis Pitfalls

Blind analyses sometimes run into unforeseen troubles.

Sometimes there is a mistake that evades the blinded review.

Or some very rare background process contributes to a rare-particle search – it may be included in the simulation model, but not enough simulation was run in order to predict it reliably.

Often an obvious cut is easier to apply than an enormous simulation campaign.

Example: OPAL acoplanar dilepton event with a hadronic shower – a SUSY candidate until it wasn't.

Or an OPAL acoplanar dilepton event with the photon hitting a steel support wheel. It was in the MC, just not sampled enough.

Benefits of Reinterpretation

- A discovery of something not included in the SM prediction may still be ambiguous – What is it?
- HARKing is a problem when *post hoc* hypotheses are presented as *a priori* hypotheses.
  - Example – drawing ever-smaller boxes in high-dimensional kinematic space around observed events
  - The probability of observing each event exactly the way it was is vanishingly small.
- Reinterpretation of exclusion contours for one model in terms of another model gets more physics out with less effort.

Reasons Why an Analysis May Fail to be Reproducible

- Missing digital artifacts
- Improper packaging of digital artifacts
- Version mismatches
- Incomplete documentation
- Misunderstanding of the documentation that is there
- Mistakes in the original analysis
- Mistakes in the reproduction attempt
- Use of random numbers in the analysis of the data
- Computational non-reproducibility (radiologicals, bit errors, failing storage)
  - sometimes a tape gets stuck in a tape drive
  - or a job crashes for reasons that have nothing to do with what the job was doing. e.g., power cut, node runs out of memory due to other jobs, etc.
  - Analyzers must get good at detecting and recovering failed jobs.

Recasting Analyses

- If data histograms corresponding signal and background histograms are preserved, they can be re-used to test models beyond the ones motivating the original search.

- If the acceptance of the experimental apparatus and analysis cuts, as well as resolutions can be computed for a new signal, then an entirely new interpretation of the data can be achieved.

- Even if there are uncertainties in the process
  - parameterized simulation and estimates of reconstruction effects
  - different kinematic distributions of the new signal and the one trained on

  one can still assign systematic uncertainty and proceed. How to make everyone happy?

- Multivariate analyses (NN, BDT, deep networks, etc) are the hardest to recast.
  - They are highly optimized to select a specific signal.
  - Some may be so specific that even "adjacent" signals would not populate "interesting" bins.

Unfolding

Production of differential cross section results from observed data counts as functions of reconstructed quantities (energies, angles, number and type of particles, etc)

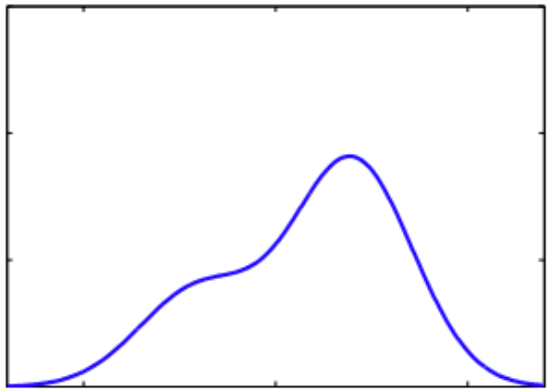
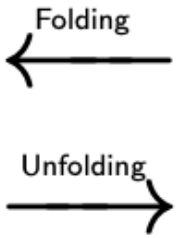


Figure : Smeared spectrum



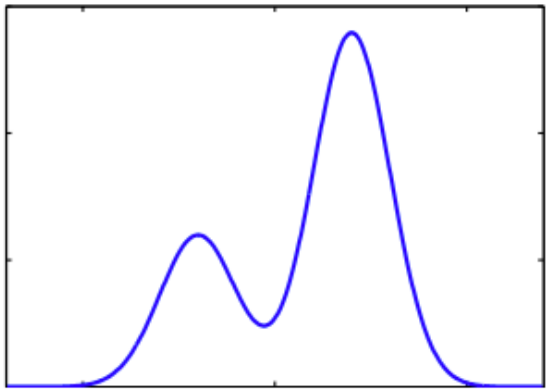


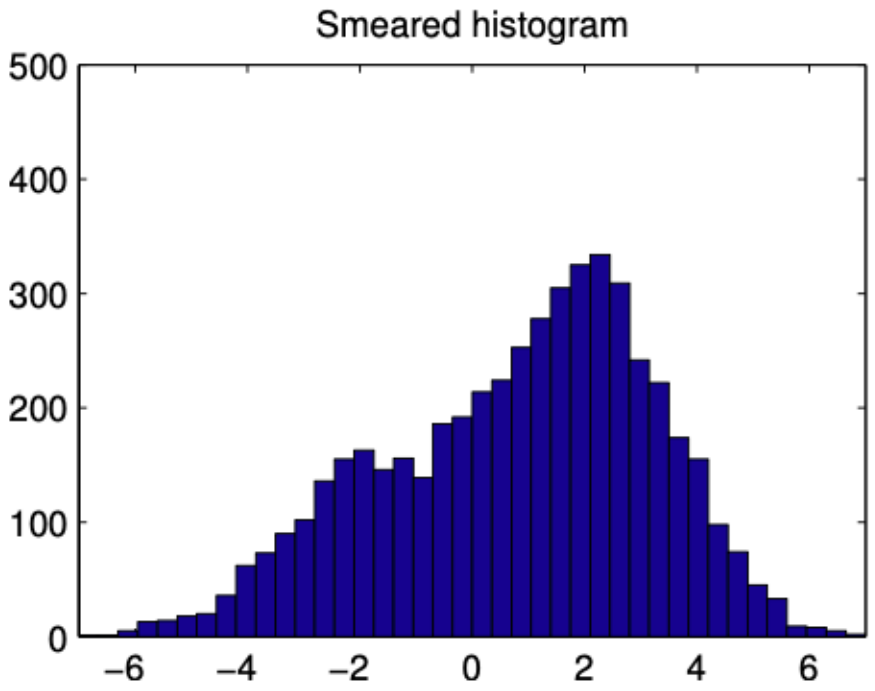
Figure : True spectrum

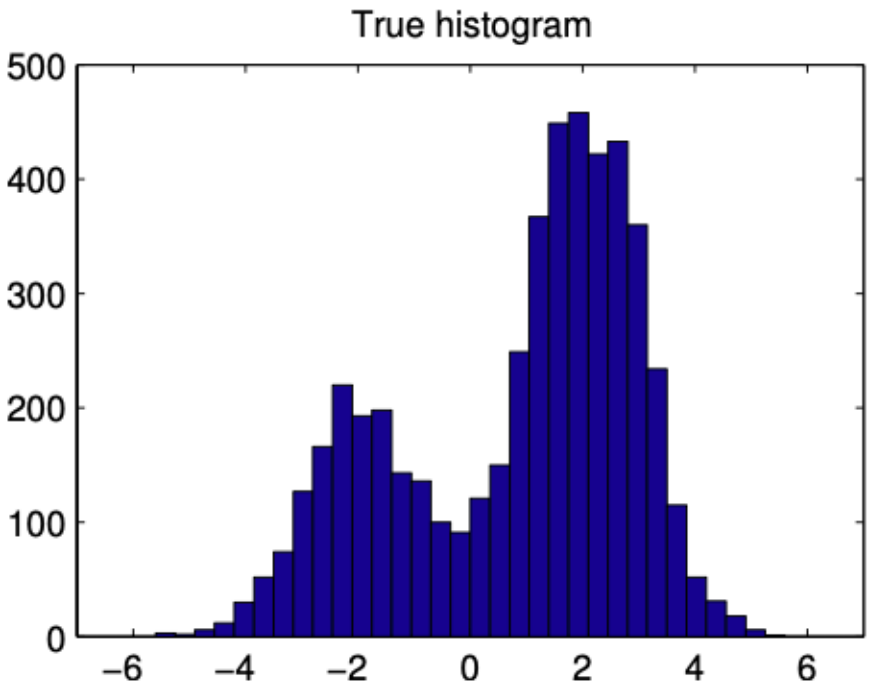
Mikael Kuusela (UChicago) September 21, 2016 2 / 21

See for example, Mikael Kuusela's Ph.D. thesis,
"Uncertainty quantification in unfolding particle spectra at the Large Hadron Collider" Ecole Polytechnique, Lausanne, 2016
https://inspirehep.net/literature/1762535

Slides from M. Kuusela, Phystat-Nu Fermilab 2016

Demonstration of ill-posedness





$$\boldsymbol{\mu} = \mathbf{K}\boldsymbol{\lambda}, \quad \mathbf{y} \sim \text{Poisson}(\boldsymbol{\mu}) \quad \overset{??}{\Longrightarrow} \quad \hat{\boldsymbol{\lambda}} = \mathbf{K}^{-1}\mathbf{y}$$

D'Agostini demo, $k = 0$

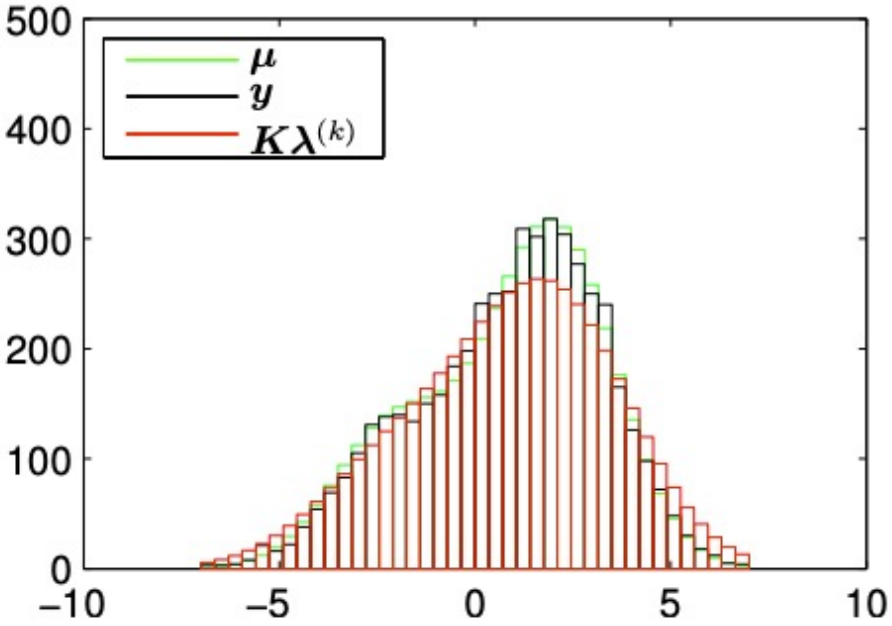


Figure : Smeared histogram

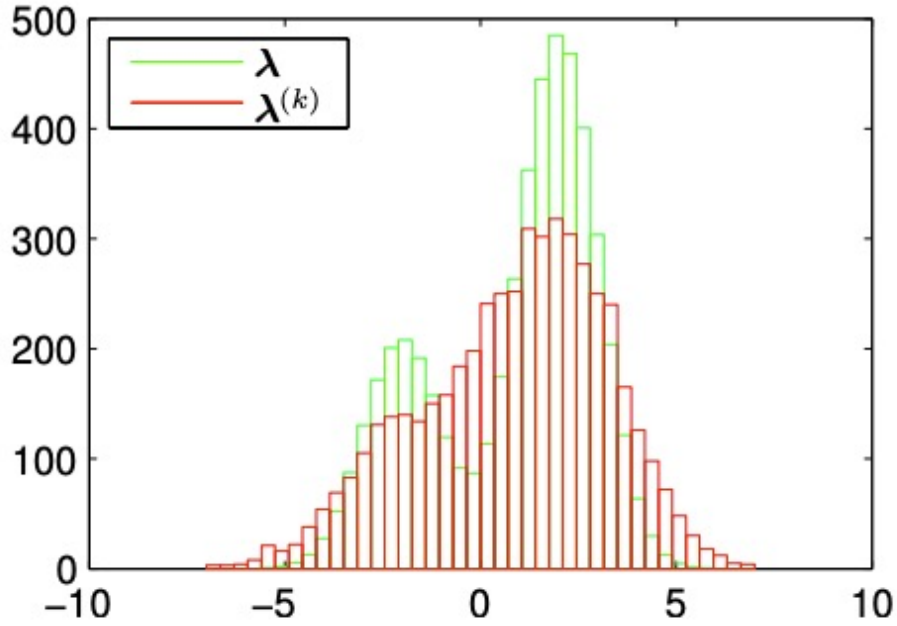


Figure : True histogram

D'Agostini demo, $k = 100$

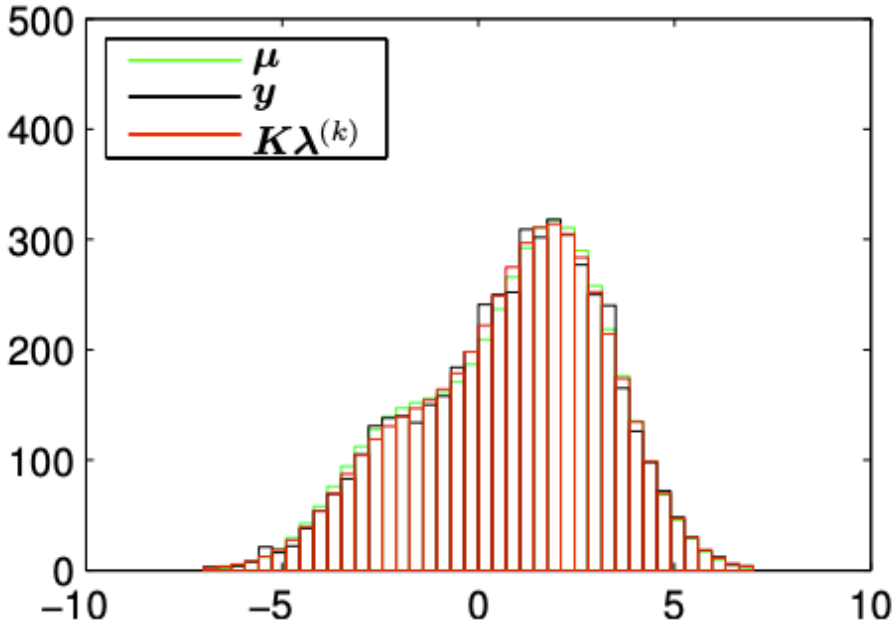


Figure : Smeared histogram

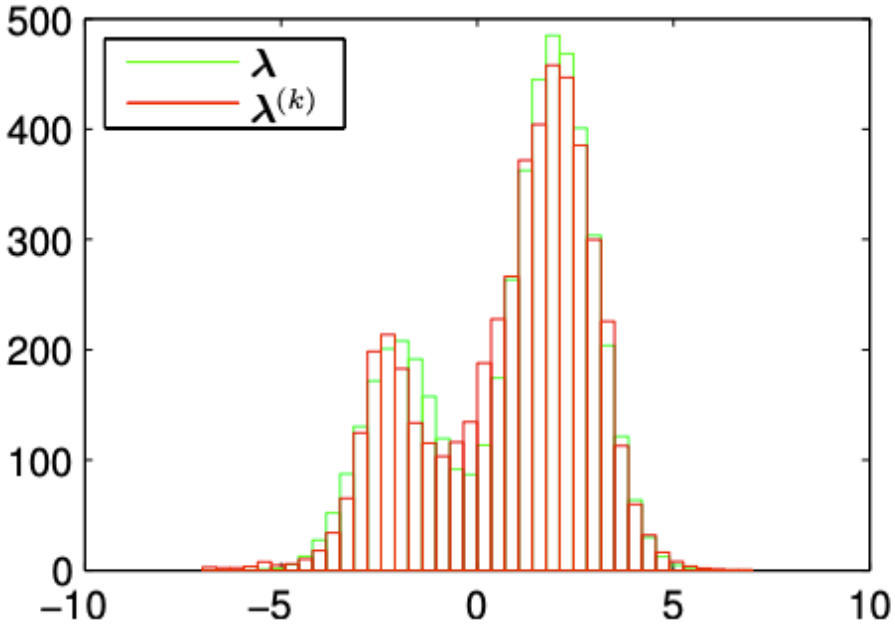


Figure : True histogram

D'Agostini demo, $k = 10000$

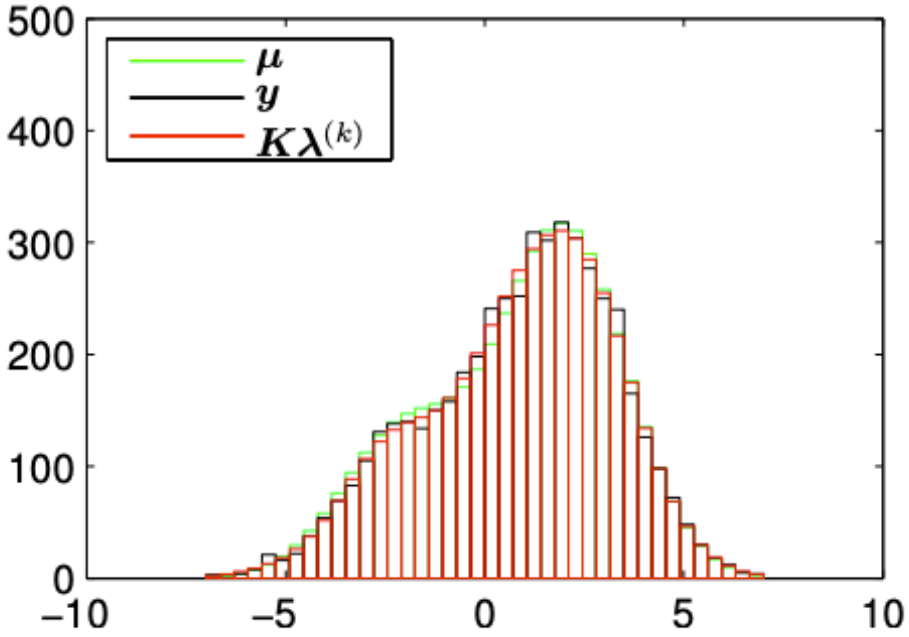


Figure : Smeared histogram

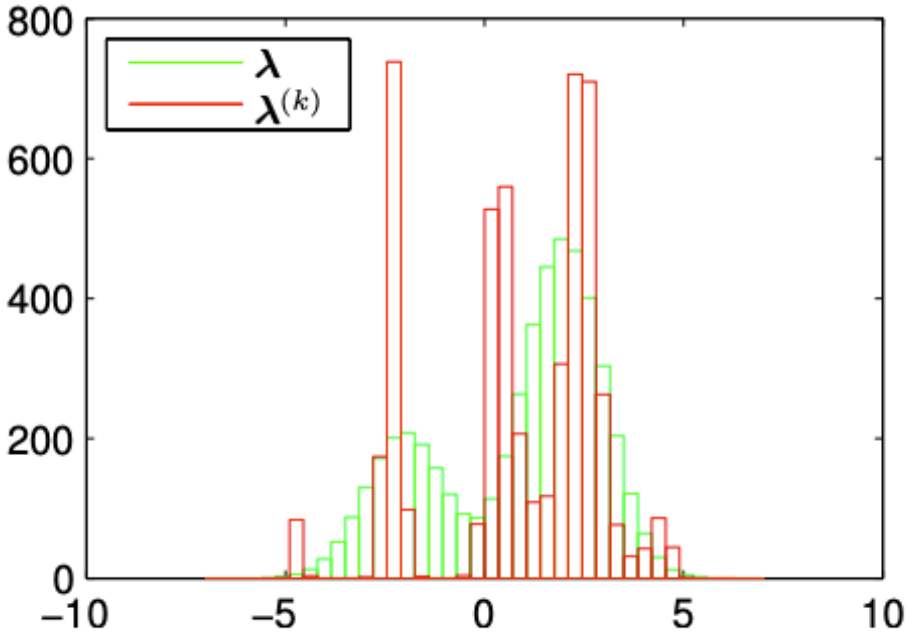


Figure : True histogram

D'Agostini demo, $k = 100000$

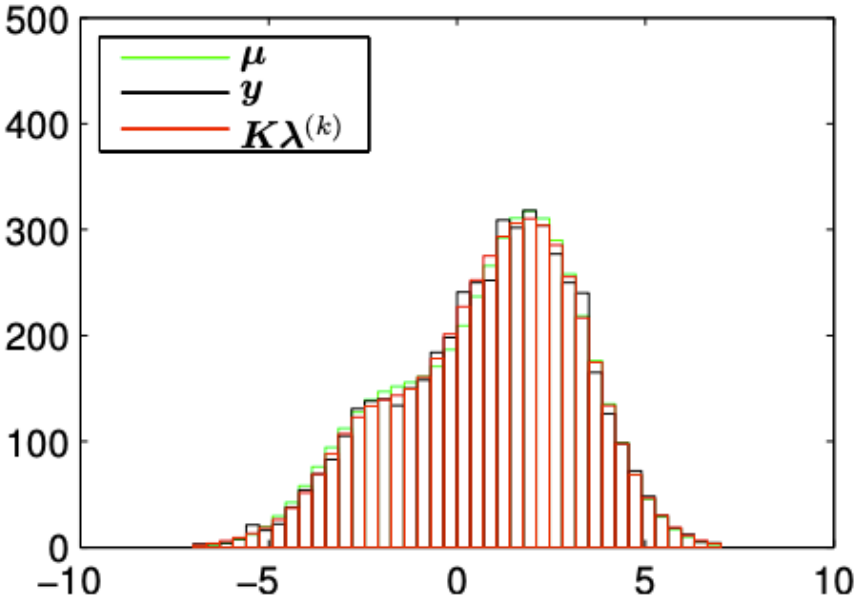


Figure : Smeared histogram

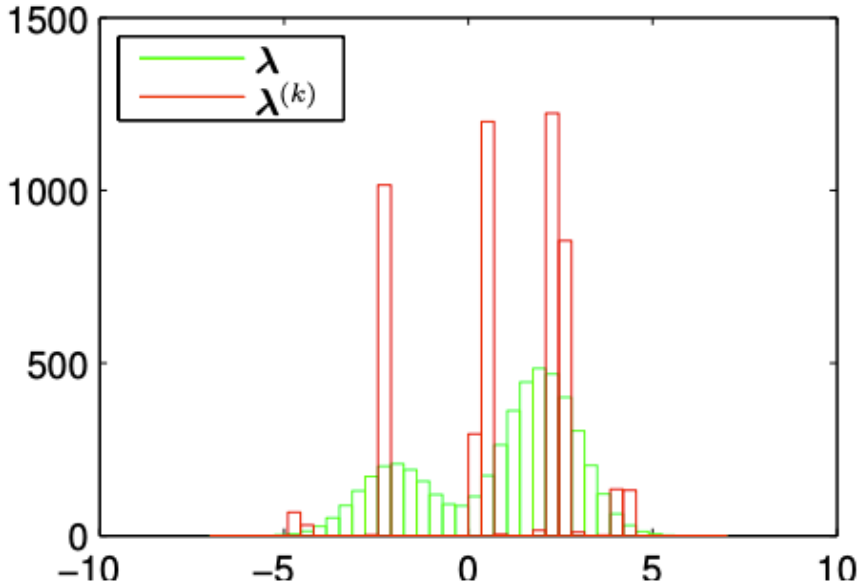


Figure : True histogram

Unfolding – Regularization and Uncertainties

Often, in RooUnfold, the number of iterations *k* is used to ensure well-behaved output behavior.

The choice is arbitrary, but the end goal is to produce uncertainties as small as possible on the results.

Uncertainties come with correlations – change one bin's measured cross section, you may need to change many other bins' values too in order to remain consistent with the measurements.

I have been asked by theorists to provide advice on interpreting old LEP-1 histograms of QCD observables where no correlation information was provided. One solution is simply to leave the correlations uncertain.

In the extreme case of D'Agostini with $k = 10000$ and above, Gaussian uncertainties with an error matrix is inadequate.

Some experimenters would like to provide the raw data, background histogram, and the folding matrix *K*.

Summary

- It is good to think about data and analysis preservation while designing an experiment and building the collaboration
- Some aspects of D & A preservation are required for basic functioning of a collaboration producing results
- Great strides have been made to make it easier to distribute and document digital artifacts that can be ported to a number of computing platforms
- Doing D & A preservation well requires time and effort (= money), not only to do the initial work, but to approve it and support it.
- Funding agencies are requiring D & A preservation plans
- The goal should be to enable great science

Extras

Reanalysis of Deuterium Bubble-Chamber Data from the 1980's

A. Meyer, M. Betancourt, R. Gran and R. Hill Phys. Rev. D 93, 113015 (2016)
https://arxiv.org/abs/1603.03048