



# Report from PhyStat-nu @ CERN

Alex Himmel

Neutrino Seminar

Fermilab

March 28<sup>th</sup>, 2019

# Introduction

- These workshops are really interesting (I think) and I highly recommend them.
- Link to everything:
  - <https://indico.cern.ch/event/735431/timetable/>
- I of course can't touch on everything from the whole workshop in just an hour!
- I will focus on a few things, biased towards what I found most interesting.
  - Feldman-Cousins and Confidence Intervals, Generally
  - Testing the Mass Hierarchy
  - The Challenge of Unfolding
  - Uncertain Uncertainties

# Feldman-Cousins and Confidence Intervals, Generally

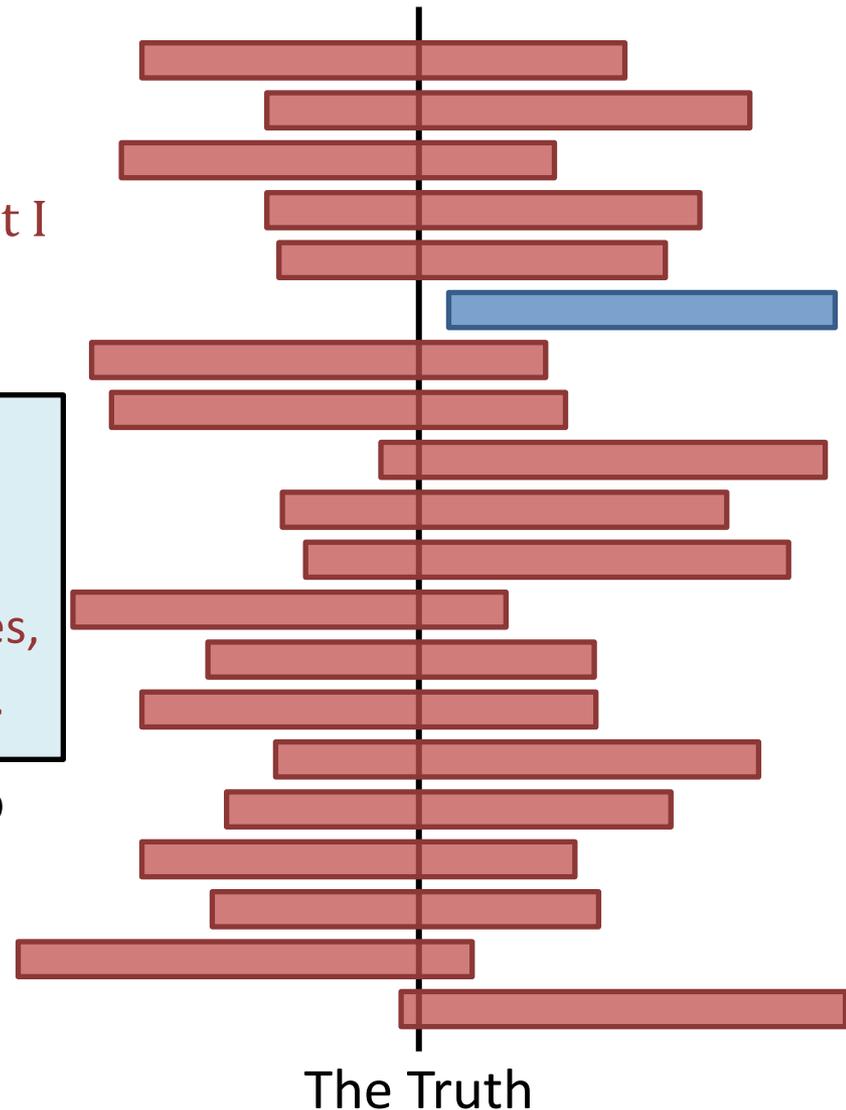
# Building Confidence Intervals

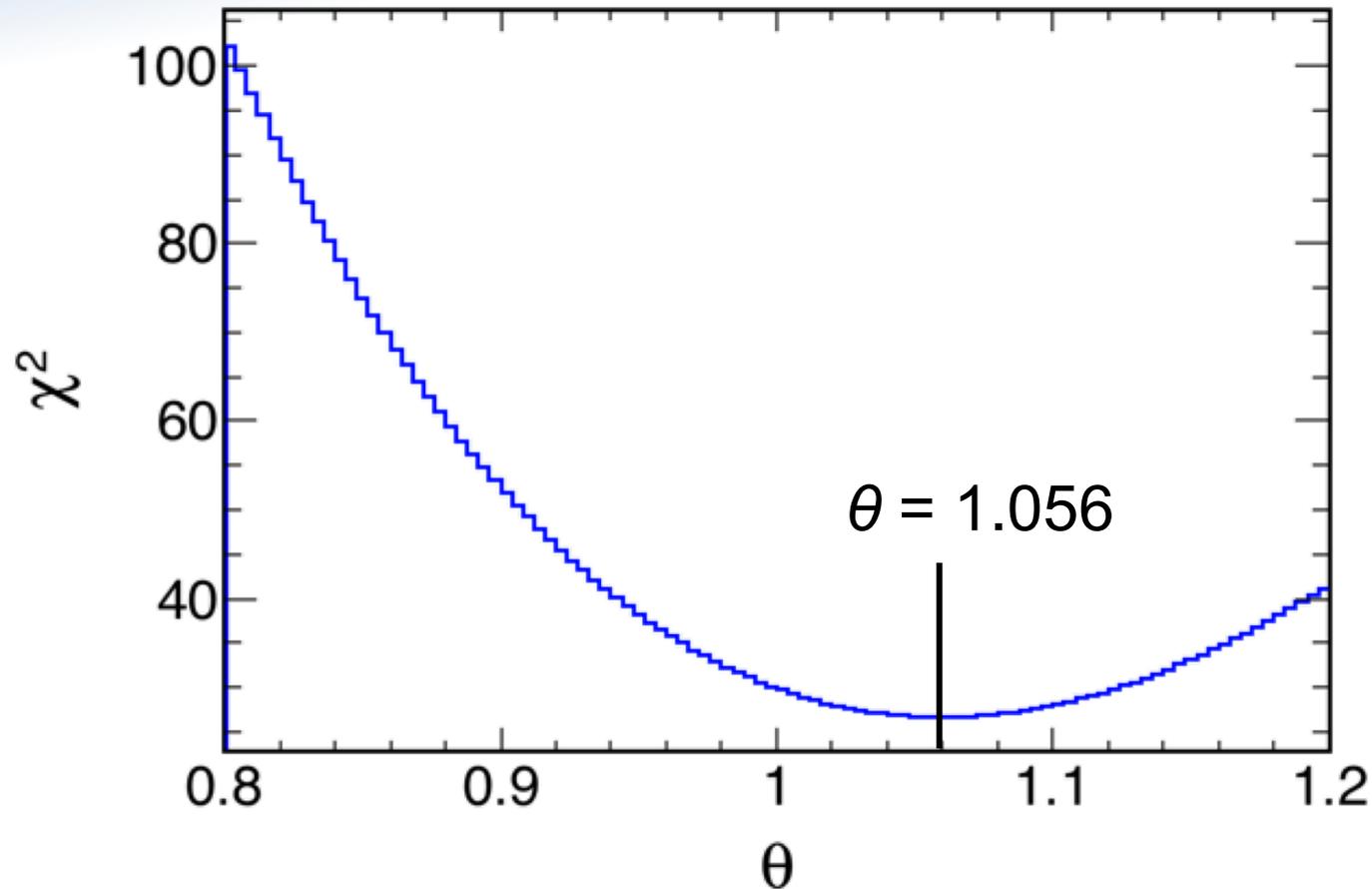
- A *brief* introduction “frequentist” confidence intervals.
  - Not actually from the workshop, but I think it’s important background to have.

Definition of an Confidence Interval  
at level  $\alpha$ :

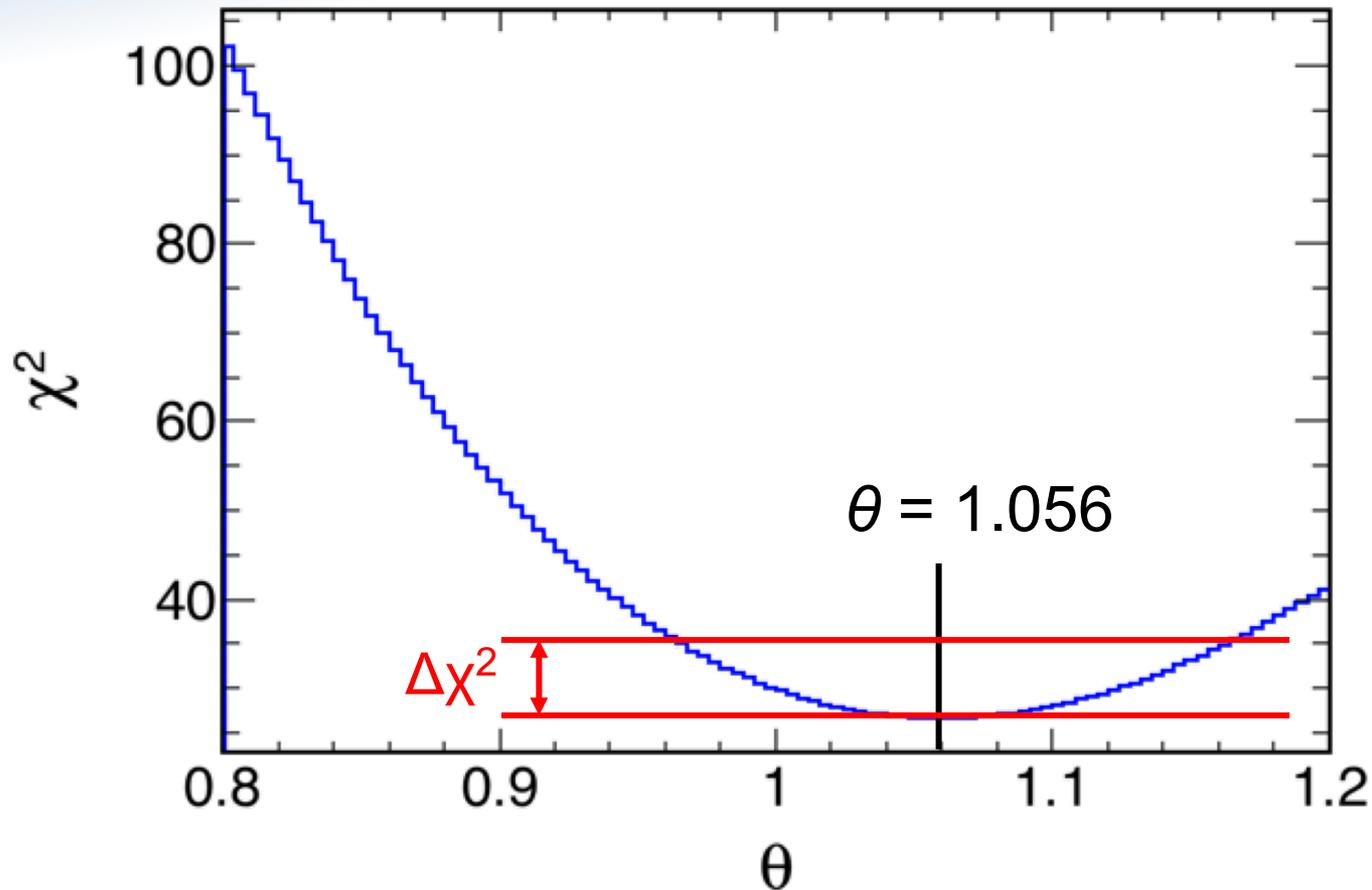
If we repeat the experiment numerous times,  
 $\alpha$  of the intervals will cover the true value.

- This isn’t really what you wanted to know, but it has been rigorously defined.
- There are many ways to construct CI’s depending on the circumstance.

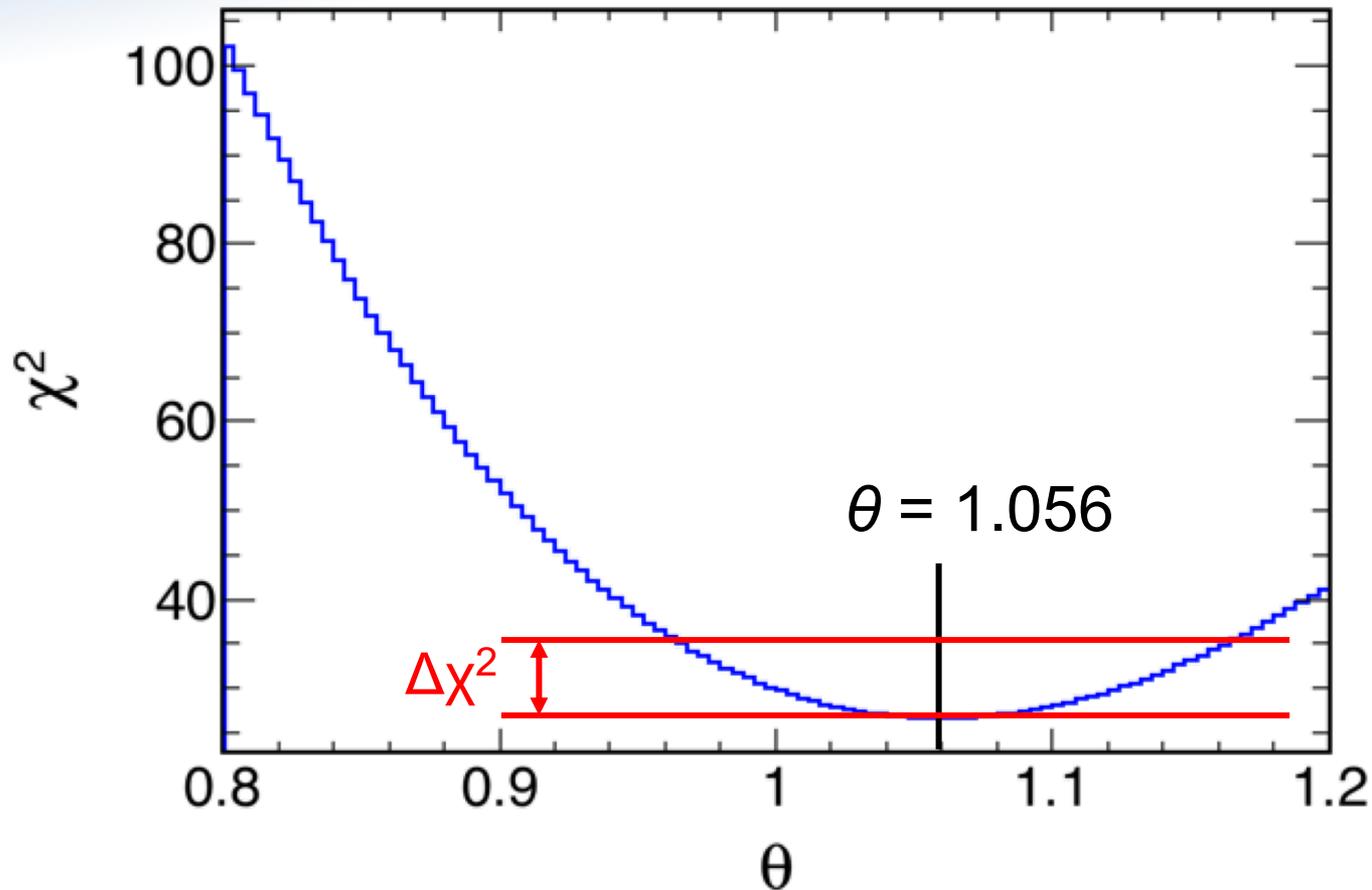




- If your problem has all Gaussian errors, then the distribution of the estimator of the parameter is *also* Gaussian.
  - Presented without proof, since that's what the PDG does, too.
  - This is the case for our example, too.



- We will use the likelihood distribution to draw the CI.
- We allow inside our CI any values of  $\theta$  with small values  $\Delta\chi^2$  relative to the best fit, and we exclude values of  $\theta$  with larger values of  $\Delta\chi^2$ .



- The question you should be asking:
- How do I know what “up value” or “critical value” to choose to know which  $\theta$ 's are in and which are out?

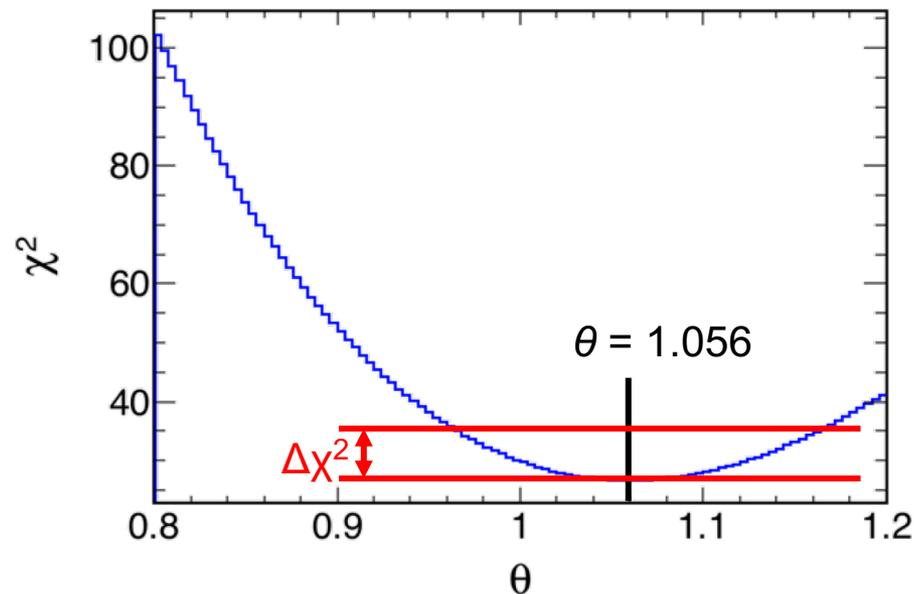
# Wilks' Theorem

- The  $\Delta\chi^2$  between your best fit and other points will follow a standard  $\chi^2$

## IF conditions are met:

- Large statistics
- Well away from parameter boundaries
- *Nested hypotheses.*
  - Testing one value against a continuous set of alternatives

- Frequently, these are not met. So now what?



$(1 - \alpha)$ (%)	$m = 1$	$m = 2$	$m = 3$
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

## Nested Hypothesis Testing: Duality with Intervals

In classical/frequentist formalism (but *not* Bayesian formalism), the theory of these hypo tests maps to that of confidence intervals:

Having observed data  $x_{\text{obs}}$ , suppose the 90% C.L. confidence interval for  $\mu$  is  $[\mu_1, \mu_2]$ .

This contains all values of  $\mu$  for which observed  $x_{\text{obs}}$  is ranked in the *least extreme* 90% of possible outcomes  $x$  according to  $p(x|\mu)$  and the ordering principle in use.

Now suppose we wish to test  $H_0$  vs  $H_1$  at Type I error prob  $\alpha = 10\%$ . We reject  $H_0$  if  $x_{\text{obs}}$  is ranked in the *most extreme* 10% of  $x$  according to  $p(x|\mu)$  and the ordering principle in use.

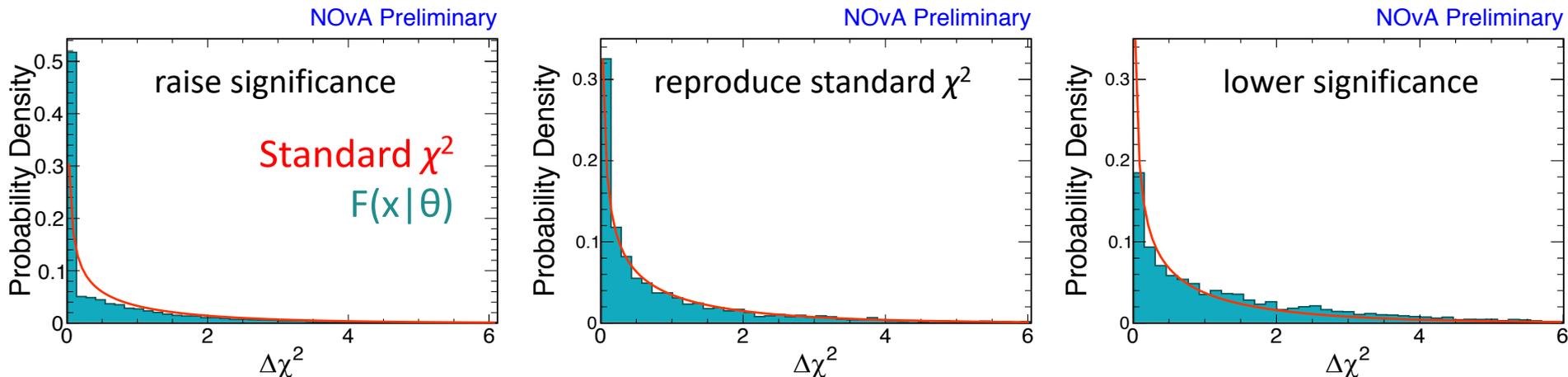
Comparing the two procedures, we see:

**Reject  $H_0$  at  $\alpha=10\%$  iff  $\mu_0$  is *not* in 90% C.L. conf. interval  $[\mu_1, \mu_2]$ .**

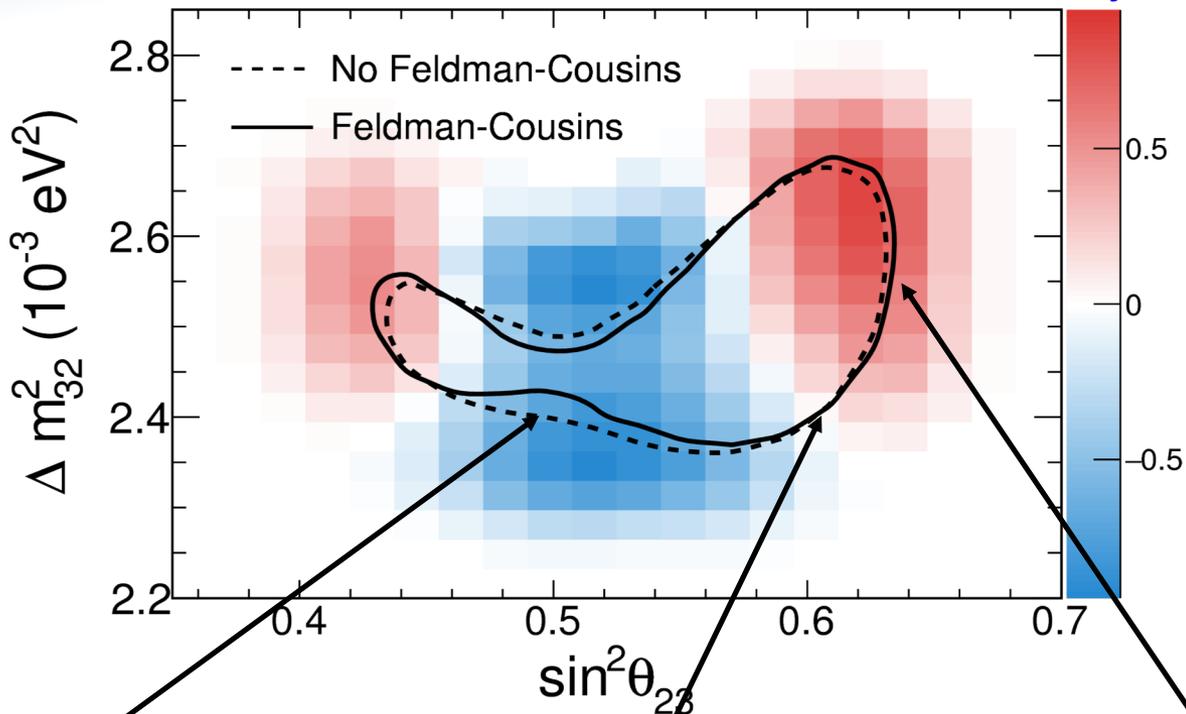
Use of the duality is referred to as “**inverting a test**” to obtain confidence intervals, and vice versa. (Section 7.4)

# Inverting the Test

- At each point in parameter space, we conduct a hypothesis test.
  - The parameter we test becomes  $H_0$ , and everything else is  $H_1$ .
- We ask: how likely is the observed data assuming  $H_0$ ?
  - With Wilk's theorem, we could just compare the  $\Delta\chi^2$  from our data to the  $\chi^2$  distribution.
  - Instead we **generate pseudo-experiments** to answer this question.



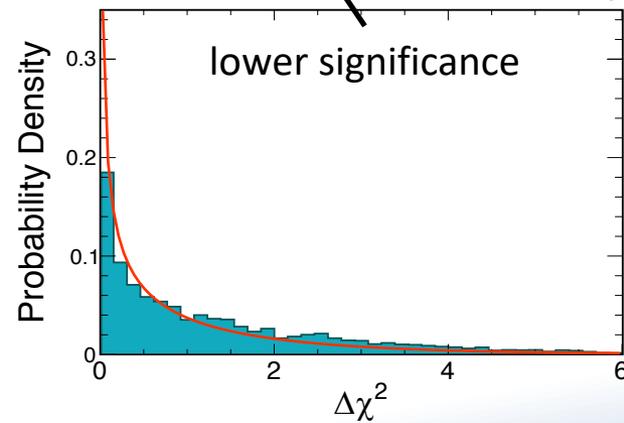
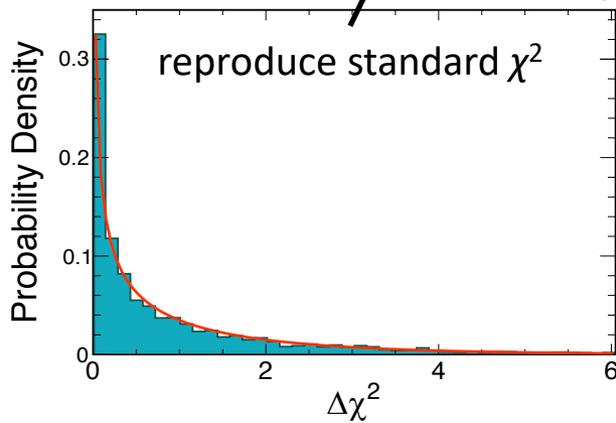
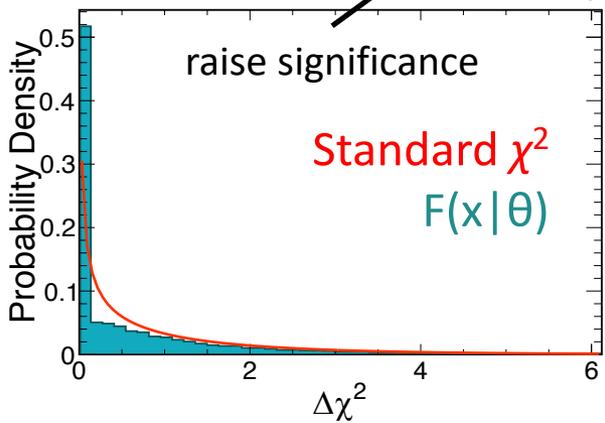
- If we would reject  $H_0$  at level  $(1-\alpha)$  in the hypothesis test, we also exclude this parameter from the  $\alpha$  CI.



NOvA Preliminary

NOvA Preliminary

NOvA Preliminary



# Feldman-Cousins (or the “Unified Approach”)

PHYSICAL REVIEW D

VOLUME 57, NUMBER 7

1 APRIL 1998

## **Unified approach to the classical statistical analysis of small signals**

Gary J. Feldman<sup>\*</sup>

*Department of Physics, Harvard University, Cambridge, Massachusetts 02138*

Robert D. Cousins<sup>†</sup>

*Department of Physics and Astronomy, University of California, Los Angeles, California 90095*

(Received 21 November 1997; published 6 March 1998)

We give a classical confidence belt construction which unifies the treatment of upper confidence limits for null results and two-sided confidence intervals for non-null results. The unified treatment solves a problem (apparently not previously recognized) that the choice of upper limit or two-sided intervals leads to intervals which are not confidence intervals if the choice is based on the data. We apply the construction to two related problems which have recently been a battleground between classical and Bayesian statistics: Poisson processes with background and Gaussian errors with a bounded physical region. In contrast with the usual classical construction for upper limits, our construction avoids unphysical confidence intervals. In contrast with some popular Bayesian intervals, our intervals eliminate conservatism (frequentist coverage greater than the stated confidence) in the Gaussian case and reduce it to a level dictated by discreteness in the Poisson case. We generalize the method in order to apply it to analysis of experiments searching for neutrino oscillations. We show that this technique both gives correct coverage and is powerful, while other classical techniques that have been used by neutrino oscillation search experiments fail one or both of these criteria.

[S0556-2821(98)00109-X]

PACS number(s): 06.20.Dk, 14.60.Pq

## Duality in Nested Hypothesis Testing

While F-C was “in proof”, Gary realized that “our” intervals were simply those obtained by “inverting” the classic “exact” LR hypothesis test (which specifies LR ordering) in Kendall and Stuart.

It was all on 1¼ pages, plus profiling nuisance parameters!

See Gary’s Fermilab talk, “Journeys of an Accidental Statistician”, <http://users.physics.harvard.edu/~feldman/Journeys.pdf>

This was of course good !  
It led to rapid inclusion in PDG RPP.

CHAPTER 22

### LIKELIHOOD RATIO TESTS AND TEST EFFICIENCY

#### The LR statistic

22.1 The ML method discussed in Chapter 18 is a constructive method of obtaining estimators which, under certain conditions, have desirable properties. A method of test construction closely allied to it is the likelihood ratio (LR) method, proposed by Neyman and Pearson (1928). It has played a role in the theory of tests analogous to that of the ML method in the theory of estimation. As before, we have the LF

$$L(x|\theta) = \prod_{i=1}^n f(x_i|\theta),$$

where  $\theta = (\theta_r, \theta_s)$  is a vector of  $r + s = k$  parameters ( $r \geq 1, s \geq 0$ ) and  $x$  may also be a vector. We wish to test the hypothesis

$$H_0 : \theta_r = \theta_{r0}, \quad (22.1)$$

which is composite unless  $s = 0$ , against

$$H_1 : \theta_r \neq \theta_{r0}.$$

We know that there is generally no UMP test in this situation, but that there may be a UMPU test – cf. 21.31.

The LR method first requires us to find the ML estimators of  $(\theta_r, \theta_s)$ , giving the unconditional maximum of the LF

$$L(x|\hat{\theta}_r, \hat{\theta}_s), \quad (22.2)$$

and also to find the ML estimators of  $\theta_s$ , when  $H_0$  holds,<sup>1</sup> giving the conditional maximum of the LF

$$L(x|\hat{\theta}_{r0}, \hat{\theta}_s), \quad (22.3)$$

$\hat{\theta}_s$  in (22.3) has been given a double circumflex to emphasize that it does not in general coincide with  $\hat{\theta}_s$  in (22.2). Now consider the likelihood ratio<sup>2</sup>

$$l = \frac{L(x|\hat{\theta}_{r0}, \hat{\theta}_s)}{L(x|\hat{\theta}_r, \hat{\theta}_s)}. \quad (22.4)$$

Since (22.4) is the ratio of a conditional maximum of the LF to its unconditional maximum, we clearly have

$$0 \leq l \leq 1. \quad (22.5)$$

Intuitively,  $l$  is a reasonable test statistic for  $H_0$ : it is the maximum likelihood under  $H_0$  as a fraction of its largest possible value, and large values of  $l$  signify that  $H_0$  is reasonably acceptable. The critical region for the test statistic is therefore

$$l \leq c_\alpha, \quad (22.6)$$

where  $c_\alpha$  is determined from the distribution  $g(l)$  of  $l$  to give a size- $\alpha$  test, that is,

$$\int_0^{c_\alpha} g(l) dl = \alpha. \quad (22.7)$$

Neither maximum value of the LF is affected by a change of parameter from  $\theta$  to  $\tau(\theta)$ , the ML estimator of  $\tau(\theta)$  being  $\tau(\hat{\theta})$  – cf. 18.3. Thus the LR statistic is invariant under reparametrization.

# Feldman-Cousins

**Long email to me from Gary on June 5, 1998, detailing widespread interest in F-C, noting:**

“

**Most people seem to have heard about our paper, or, if not, are starting to ask about it.**

***The most disconcerting thing is that I keep getting introduced as ‘Feldman, of Feldman and Cousins.’***

”

# Feldman-Cousins in Neutrinos

## 20 years of experience with F-C

Lots of experience in HEP, many find it useful, especially when:

- ★ A model parameter is bounded (mass, cross section, sin/cosine of an angle, etc.); and/or
- ★ When log-likelihood is non-Gaussian (so Wilks's Theorem is inaccurate); multiply connected confidence regions; and/or
- ★ The interesting parameter space is  $>1D$ , where LR ordering a la F-C and K&S is particularly useful, and other orderings are poorly defined (metric dependent)

Neutrino community gets three gold stars, so major user!

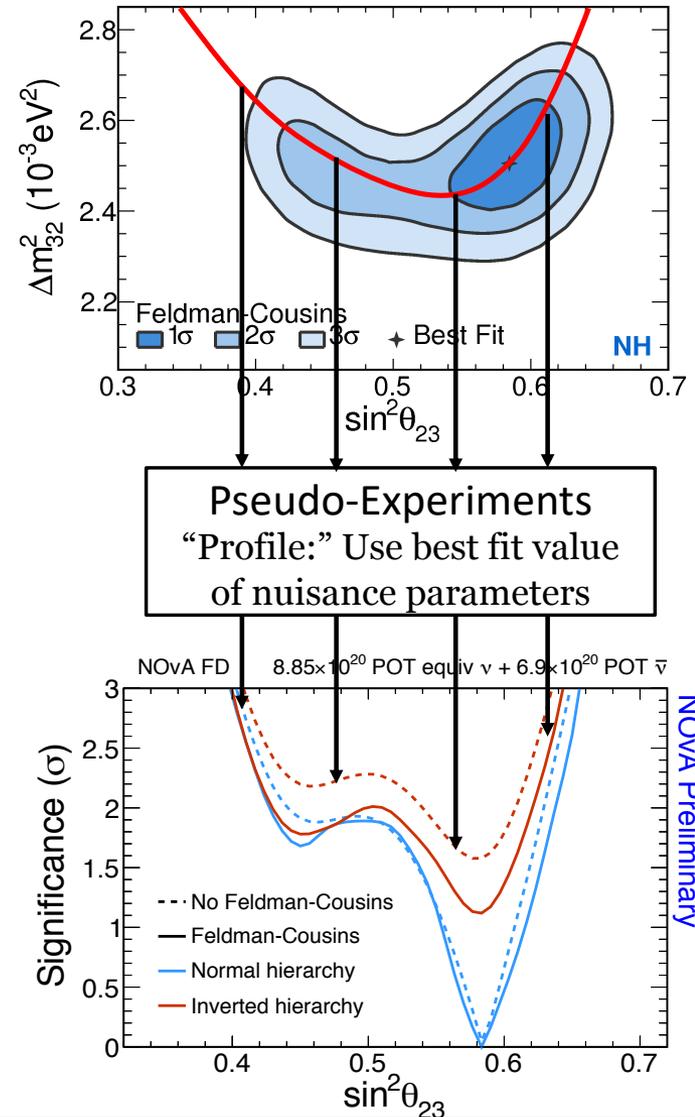
(And in fact F-C were working on the NOMAD neutrino experiment at CERN in 1998.)

BTW, for data with a “5-sigma discovery”, the F-C “unified approach” reproduces same answer as usual one-tailed test.

## Feldman-Cousins Pseudo-experiments: NOvA

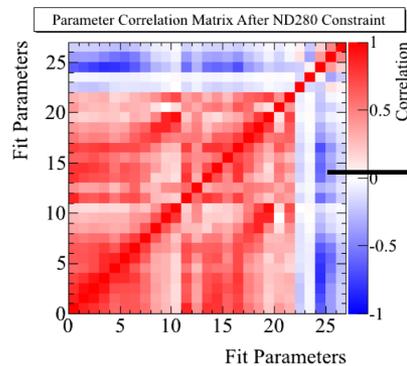
NOvA Preliminary

- Fit the data and extract parameters with all possible values of  $\theta$ .
- When generating experiments, always use the best fit to the nuisance parameters from the fit to data for each  $\theta$ .
  - Minimizes over-coverage of all methods we examined while still never under-covering.
- Tested coverage with a method from Berger and Boos for handling  $p$ -values with unknown nuisance parameters.
  - R. L. Berger and D. D. Boos, *J. Amer. Statist. Assoc.*, 89, 1012 (1994)
  - Tested coverage at a variety of choices of oscillation nuisance parameters within  $3\sigma$ 
    - Reducing quoted significance by a very small amount
  - In all cases, the other choices of nuisance parameters produced stronger rejection than the quoted rejection at the nominal profiled values.
  - This is as expected if everything is working correctly since the profiled point should give the widest CIs or lowest significance.

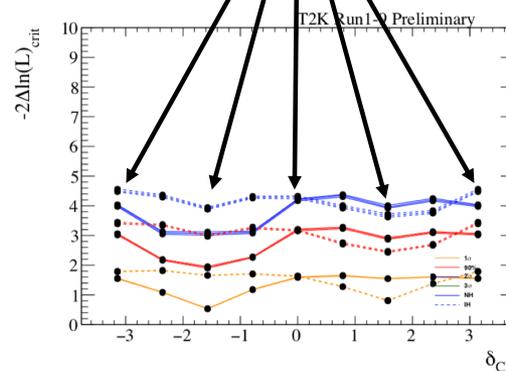
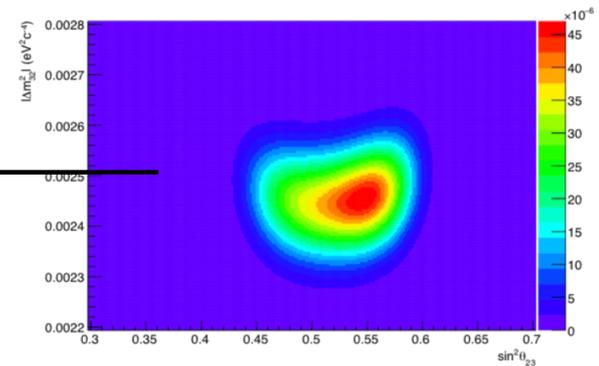


## Feldman-Cousins Pseudo-experiments: T2K

- For  $\theta_{13}$  and systematics:
  - Draw from prior distributions (PDG or output of ND fit)
- For  $\sin^2\theta_{23}$  and  $\Delta m^2$ :
  - Generate an Asimov dataset at best fit values, and construct a likelihood for this simulated dataset.
  - Convert the likelihood to a PDF, and draw values for the experiments from that distribution.

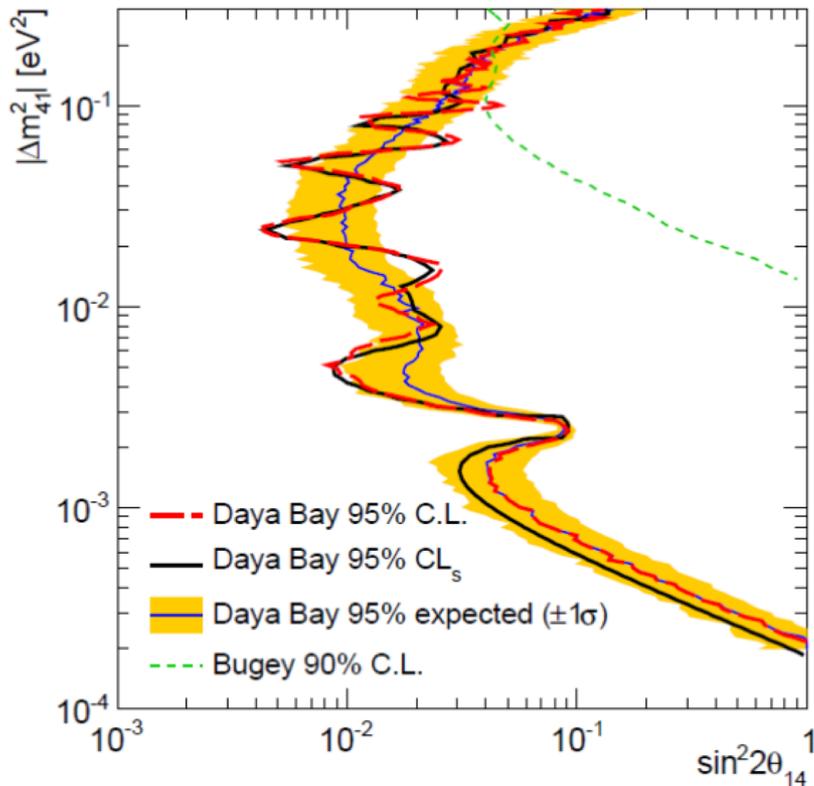


Pseudo-Experiments  
Sample from these  
prior/posterior  
distributions



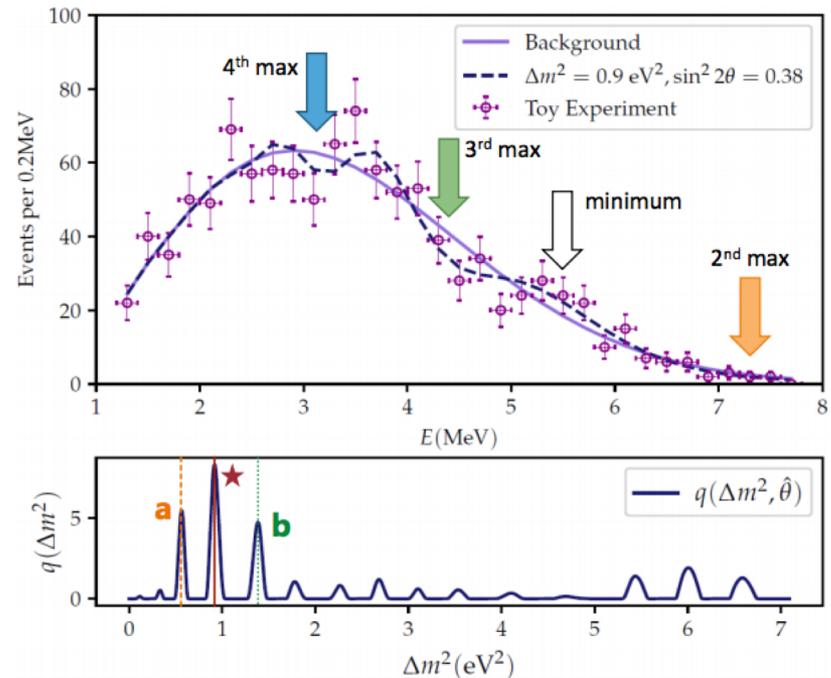
# FC (and alternatives) for Sterile Neutrinos

- Daya Bay used 2 different approaches in their sterile search: Feldman-Cousins and CLs\*
  - \*We'll come back to CLs in a few slides.



Daya Bay PRL 113, 141802, 2014

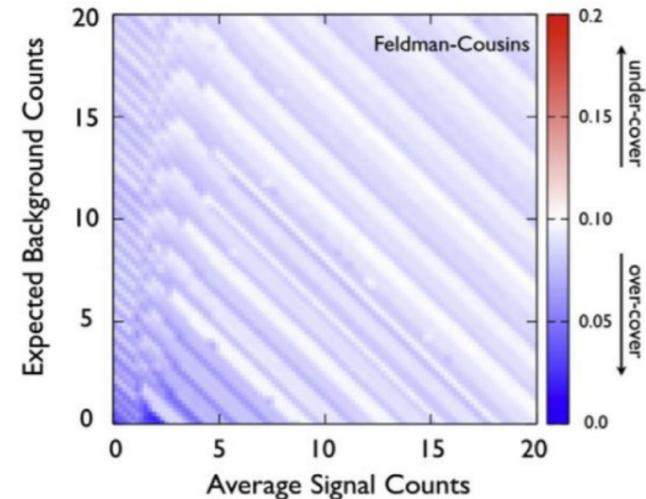
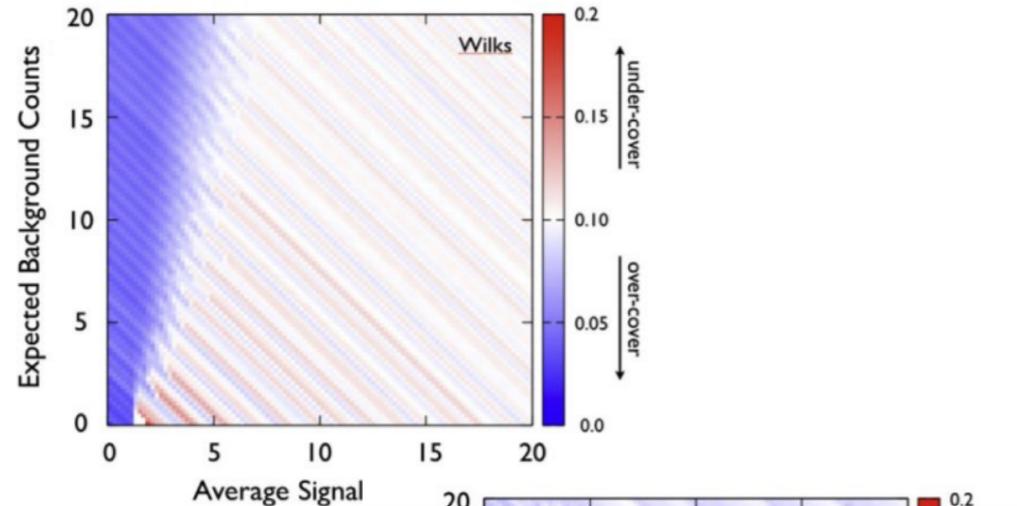
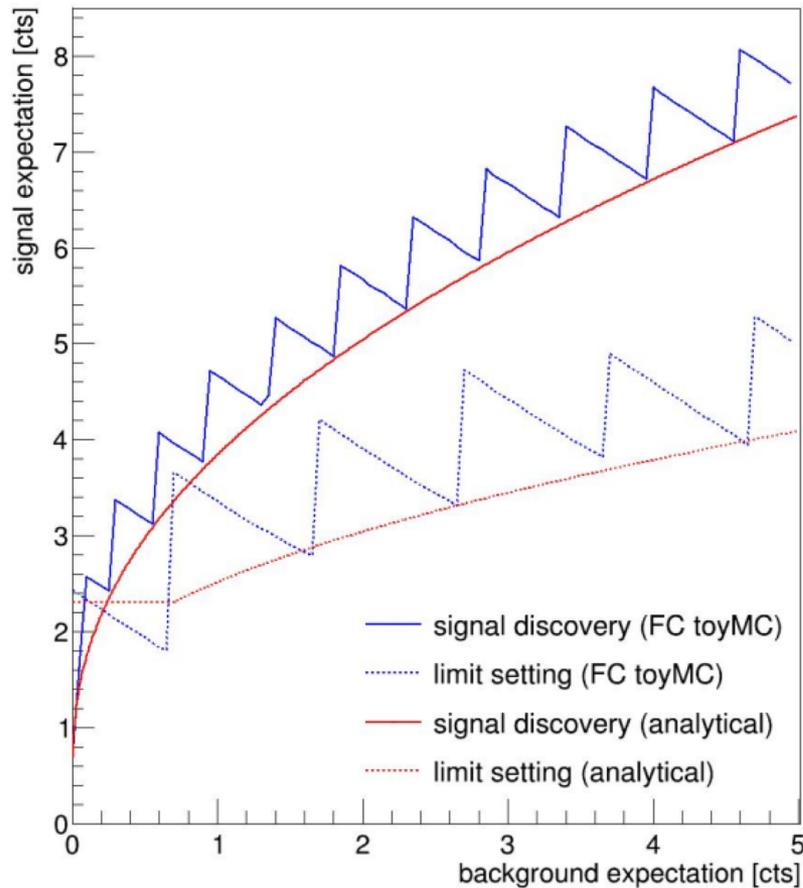
- Litchfield/Waldron developing an analysis based on accounting for the “Look Elsewhere Effect” in sterile searches.
  - Calculate “local significance” and then correct back to global significance.
  - Many challenges applying this to oscillations like harmonic local minima (below).



[https://indico.cern.ch/event/735431/contributions/3268131/attachments/1784220/2904149/PhyStatNu\\_LEEandNeutrinos\\_tutorial120190124.pdf](https://indico.cern.ch/event/735431/contributions/3268131/attachments/1784220/2904149/PhyStatNu_LEEandNeutrinos_tutorial120190124.pdf)

# FC (and alternatives) for $0\nu\beta\beta$

- Small, discrete numbers make the statistical problem difficult.
  - A variety of statistical techniques giving different answers.
- Also creates some paradoxes:
  - In the FC analysis, sensitivity gets better with larger background.
  - Why? To avoid under-coverage.
  - Bayesian? Low-stats  $\rightarrow$  strong priors



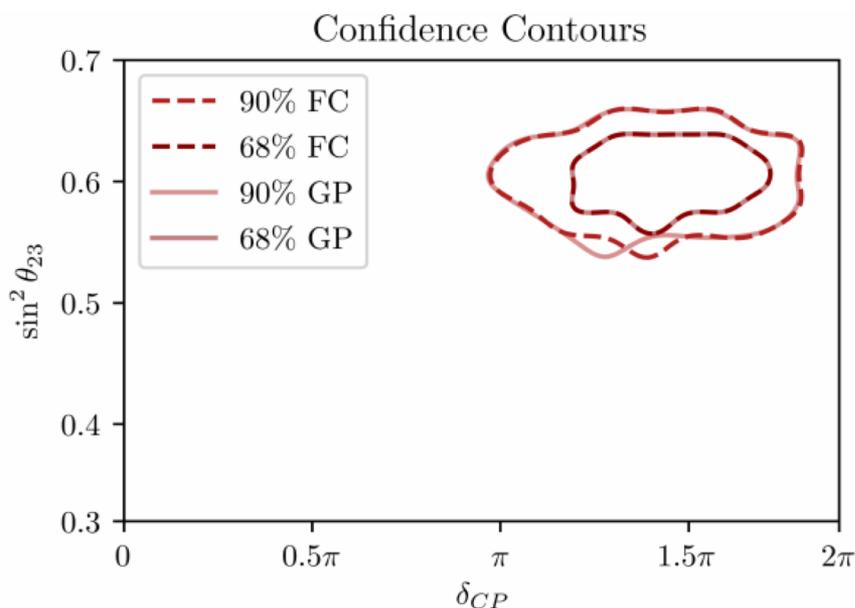
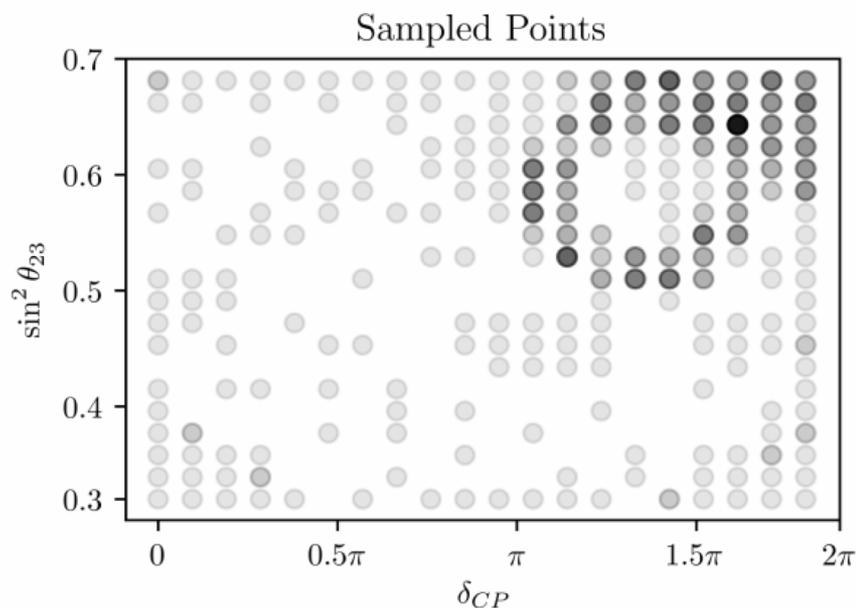
# FC (and alternatives) for $0\nu\beta\beta$

- A variety of techniques used, often multiple in the same experiment.
  - MJD is the winner with **5 (!)** different statistical analyses
  - Generally experiments using Wilks' theorem have still used toy MC to test their coverage.

Experiment	Frequentist	Bayesian
<b>MAJORANA Demonstrator</b>	Counting Unbinned $\mathcal{L}$ Unbinned $\mathcal{L}$	FC FC CLs Flat Prior Jeffreys Prior
<b>KamLAND-ZEN</b>	Multivariate $\mathcal{L}$	Wilks'
<b>EXO-200</b>	Multivariate $\mathcal{L}$	Wilks'
<b>CUORE</b>	Bounded profile $\mathcal{L}$	Wilks' Flat Prior
<b>GERDA</b>	Extended unbinned $\mathcal{L}$ Profile $\mathcal{L}$	FC Flat Prior

# Speeding up FC with Bayesian Methods

- One big challenge with FC is that you need to throw and fit a very large number of fake experiments.
  - **NOvA used 33 million CPU hours on a supercomputer!**
- The idea: use a Bayesian technique (Gaussian Process), to optimally choose which fake experiments to throw.
- Can speed things up by an order of magnitude.



# Testing the Mass Hierarchy

# Testing the Mass Hierarchy

- This can sometimes feel like a moving target.
  - Slides from the same statistician in 2016 and 2019.

## Normal Hierarchy versus Inverted Hierarchy

### Non-nested parameterized models

$H_0$  : normal hierarchy    i.e.,  $\Delta m_{32}^2 \leq 0$   
 $H_1$  : inverted hierarchy    i.e.,  $\Delta m_{32}^2 > 0$

### Computing a p-value using LRT

- Non-nested models. If no unknown parameters in either model
  - LRT follows a Gaussian distribution under  $H_0$  or  $H_1$ .
- With unknown parameters (e.g.,  $\Delta m_{32}^2, \delta_{CP}, \theta_{23}$ ):
  - Std theory (Wilks, Chernoff) does not apply: dist'n of LRT unknown
  - What is null distribution of  $\hat{\delta}$  when fitting  $H_1$ ?
  - Some results, but strong assumptions (Blennow, et al. arXiv:1311. Apply to reactor neutrino experiments, not accelerator experiments involving  $\delta_{CP}$  (Emilio Ci...))
  - Low power owing to degeneracy.
  - What about uncertainty in  $|\Delta m_{32}^2|$ ?

*Are we back to Monte Carlo (toys)? at  $5\sigma$ ?*

## Nested I: One-sided Tests

### Nested I: One-sided Tests

(JIM B)

- $H_0 : \theta \leq \theta_0$  versus  $H_1 : \theta \geq \theta_0$ .
- E.g.,  $H_0 : \Delta m_{32}^2 \leq 0$  versus  $H_1 : \Delta m_{32}^2 > 0$ .

**P-values:**  $p\text{-value} = \sup_{\theta \leq \theta_0} \Pr(T(y) \geq T(y_{\text{obs}}) | \theta)$  (Use Wilks Thm.)

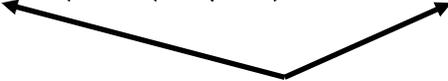
**Bayesian:** Avoid  $p_0(y)$  and  $p_2(y)$ :  $\Pr(H_0 | y) = \Pr(\theta \leq \theta_0 | y)$ .

- Requires only one model and one prior specification.
- Can incorporate external knowledge into Bayesian analysis via prior, e.g.,  $|\Delta m_{32}^2| = 2.43 \pm 0.13$ .
- Mass hierarchy can be handled this way (frequency or Bayesian)
  - ...much easier than non-nested model comparison.

*Again methods give consistent results.*

# Testing for the Mass Hierarchy

- I'm actually fairly sure that neither of those framings are right.
  - First, **non-nested**: this generally refers to two totally different models, often with different numbers of parameters.
  - Second, **one-sided**: all log-likelihood tests are one-sided (“more extreme than...”)
    - Two-sided tests only make sense when the test statistic is drawn from a symmetric distribution under the  $H_0$ .
- What we have is a likelihood ratio test with a **composite hypothesis**:
  - $H_0$ :  $\theta$  is in a subset  $\Theta_0$  of all allowed values  $\Theta$
  - $H_1$ :  $\theta$  is in the complement of  $\Theta_0$
- The likelihood ratio test statistic then is:

$$T(x) = \frac{\sup(\mathcal{L}(\theta|x) : \theta \in \Theta_0)}{\sup(\mathcal{L}(\theta|x) : \theta \in \Theta)}$$


Equiv. to profile over  $\theta$  in  $\Theta$  to maximize  $L$ .

# Testing for the Mass Hierarchy

$$T(x) = \frac{\sup(\mathcal{L}(\theta|x) : \theta \in \Theta_0)}{\sup(\mathcal{L}(\theta|x) : \theta \in \Theta)}$$

Applying ln...

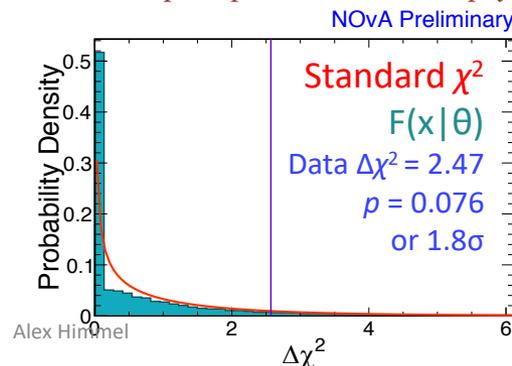
$$F(x) = \chi_{\text{IH}}^2(x) - \chi_{\text{BestFit}}^2(x)$$

- This corresponds exactly to how we construct our mass hierarchy test on NOvA.
- Still need to worry about the distribution of T.
  - In some limiting cases they can be normally distributed (*not*  $\chi^2$ ).
  - In practice, usually still need p-experiments.
- There is still an additional subtlety...

## FC for Mass Hierarchy in NOvA

- Deciding if any individual point  $\theta_0$  is outside a CI is equivalent to a hypothesis test where  $H_0$  is  $\theta = \theta_0$ .
  - Same FC technique used for setting CI's can be used for this hypothesis test.
- Since our best fit is in the NH, we want to know how strongly we reject the IH.
  - So  $H_0$  is IH and we generate pseudo-experiments at our best fit in the IH.
  - Follow the FC procedure with:
 

$\chi^2(\text{test point}) - \chi^2(\text{best fit}) \rightarrow \chi^2(\text{IH}) - \chi^2(\text{best fit})$
  - If an experiment has a best fit in the IH, then the difference is 0.
  - This pile-up at 0 behaves like a physical boundary: it increases significance.



### Limiting Case: No sensitivity

- Half of experiments in each hierarchy and  $\Delta\chi^2 = 0$
- $p = 0.5$
- 50% for either NH or IH
- All "prior"

## Careful with $p$ -values

- A  $p$ -value can only exclude the null hypothesis, it cannot accept the alternative.
  - Practical example: can get a small  $p$ -value for the IH because the NH is true, but can also get a small  $p$ -value because the fit is just bad.
  - In the latter case, would *also* get a small  $p$ -value for the NH.
- This is what CLs set out to solve, but a ratio of  $p$ -values is not a well-defined thing among statisticians.
- Good advice from Bob Cousin's: report both.

Read (2000) suggested excluding  $H_1$  only if

$$CL_s = \frac{1 - p_1}{1 - p_0} = \frac{\Pr(T < t_{\text{obs}} | H_1)}{\Pr(T < t_{\text{obs}} | H_0)} \leq \alpha.$$

Exclude  $H_1$  if  $T < t_{\text{obs}}$  much less likely under  $H_1$  than under  $H_0$

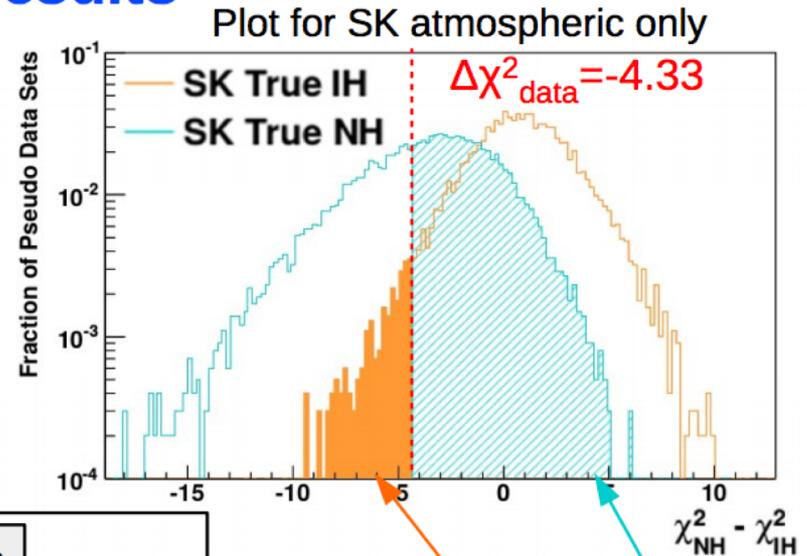
**Bob C:** Better to report both  $p$ -values.

**DvD:** Three parameter sets: no sensitivity, excluded, not excluded.

## Mass hierarchy significance Super-K results

11

- Used CLs to report significance: not truly frequentist, but conservative
- Computed p-values and CLs for lower/upper edges of the 90% CL intervals for  $\sin^2(\theta_{23})$  and  $\delta$
- Quoted a range of CLs-based significance in the paper



P-values and CLs for IH exclusion

P-values	Lower	Best fit	Upper
SK only	0.012	0.027	0.020
SK+T2K model	0.004	0.023	0.024

CLs	Lower	Best fit	Upper
SK only	0.181	0.070	0.033
SK+T2K model	0.081	0.075	0.056

$$CL_s = \frac{p_0(IH)}{1 - p_0(NH)}$$

# Bayesian Mass Hierarchy

## Bayesian for Mass Hierarchy in T2K

- Binary parameters are no problem: simply integrate the posterior within each choice.
  - This is my favorite feature of doing a Bayesian analysis.

	$\sin^2\theta_{23} < 0.5$	$\sin^2\theta_{23} > 0.5$	Sum
Normal	0.184	0.705	0.889
Inverted	0.021	0.090	0.111
Sum	0.205	0.795	1

- Bayes factor of 8 preferring the Normal Hierarchy...
  - ...but most physicists have no instinct for Bayes factors.

# Bayesian Testing

## Two Types of Bayesian Problems

**I. Estimation (confidence limit) problems:** In principle, they are straightforward.

- There are optimal prior distributions for most problems (e.g., *reference priors* - although their derivation can be difficult).
- Implementation of Bayes is usually easy, through MCMC.

**II. Hypothesis testing or model uncertainty problems:** Not so easy.

- Sometimes one can use the optimal estimation priors, but often not.
- In the latter case, the answers can be quite sensitive to the choice of prior,
  - so that one often seeks a *robust* conclusion over the choice.
- Computations can be much more difficult.

3

- What saves the day for T2K?
- NH vs. IH is more like an estimation problem than testing two different models against one another.
  - Both NH and IH have the same prior so in the ratio (Bayes factor), the priors cancel.

<https://indico.cern.ch/event/735431/contributions/3137764/attachments/1783073/2901754/phystat2018.pdf>

# Unfolding

## The unfolding problem

- Any differential cross section measurement is affected by the finite resolution of the particle detectors
  - This causes the observed spectrum of events to be “smeared” or “blurred” with respect to the true one
- The *unfolding problem* is to estimate the true spectrum using the smeared observations
- Ill-posed inverse problem with major methodological challenges

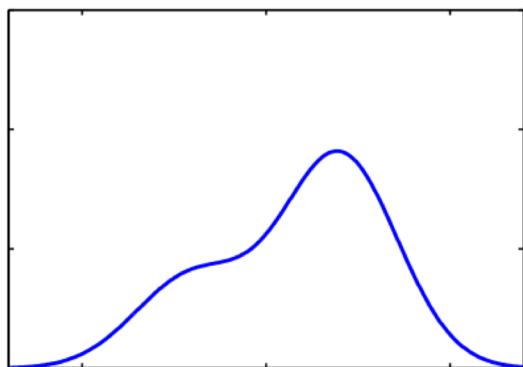


Figure: Smeared spectrum

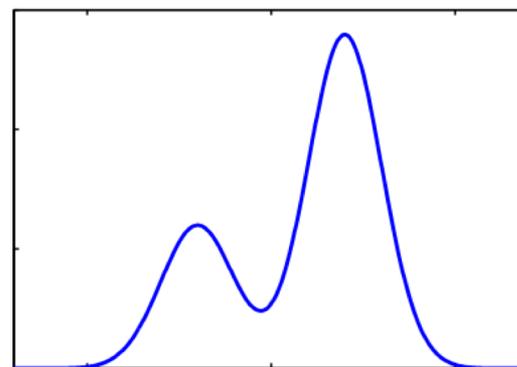
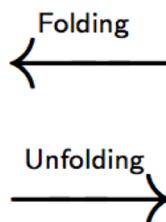
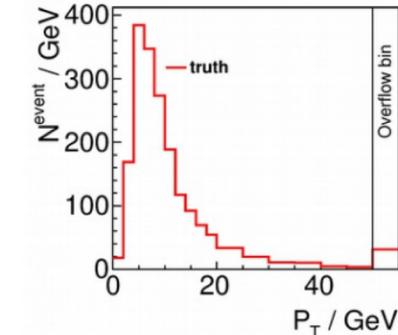


Figure: True spectrum

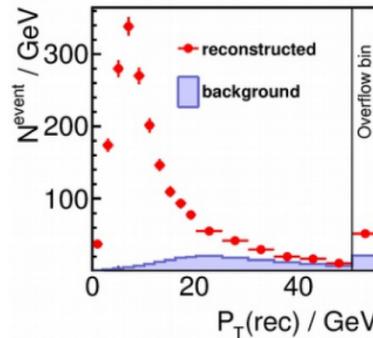
# What do you mean, “ill-posed?”



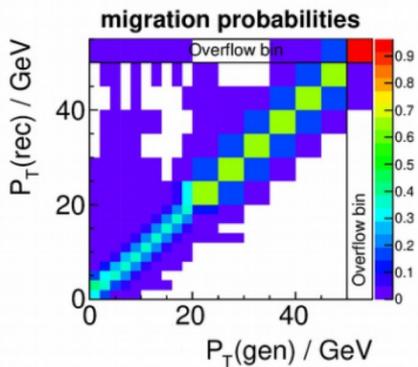
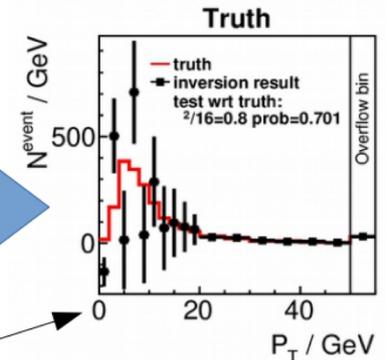
## Matrix inversion example



folding



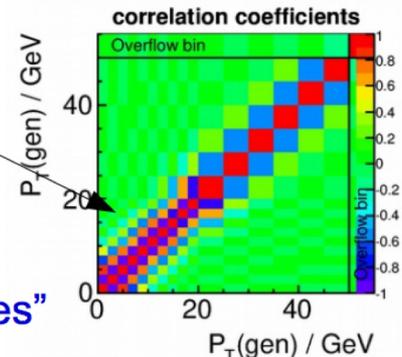
Matrix inversion



Result shows “oscillating” structures.  
Large (anti-)correlations between bins.

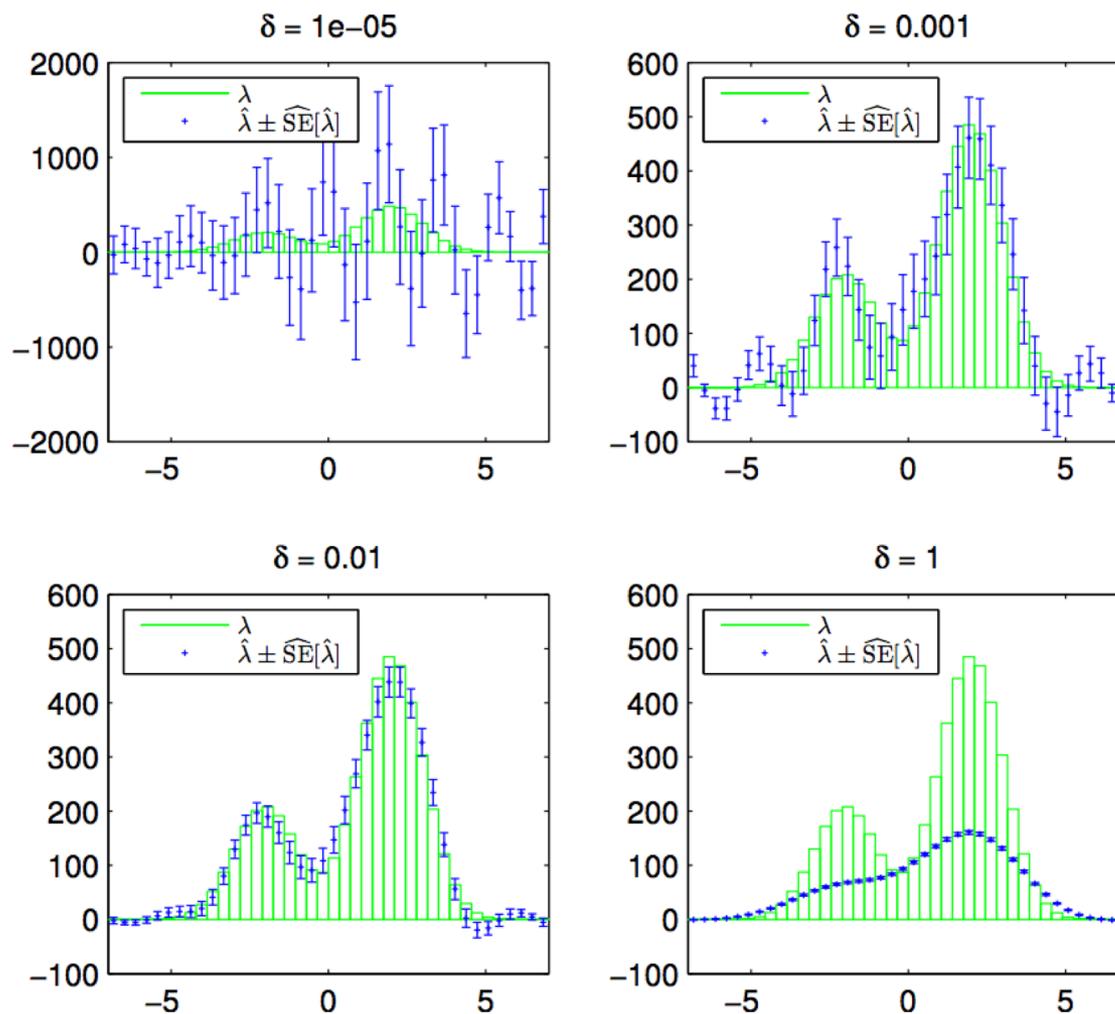
Qualitative explanation:

Finite detector resolution  $\sigma$  washes out differences between bins. Unfolding “replaces” the missing information by statistical “noise”  
→ statistical fluctuations are amplified



# The Regularization "Solution"

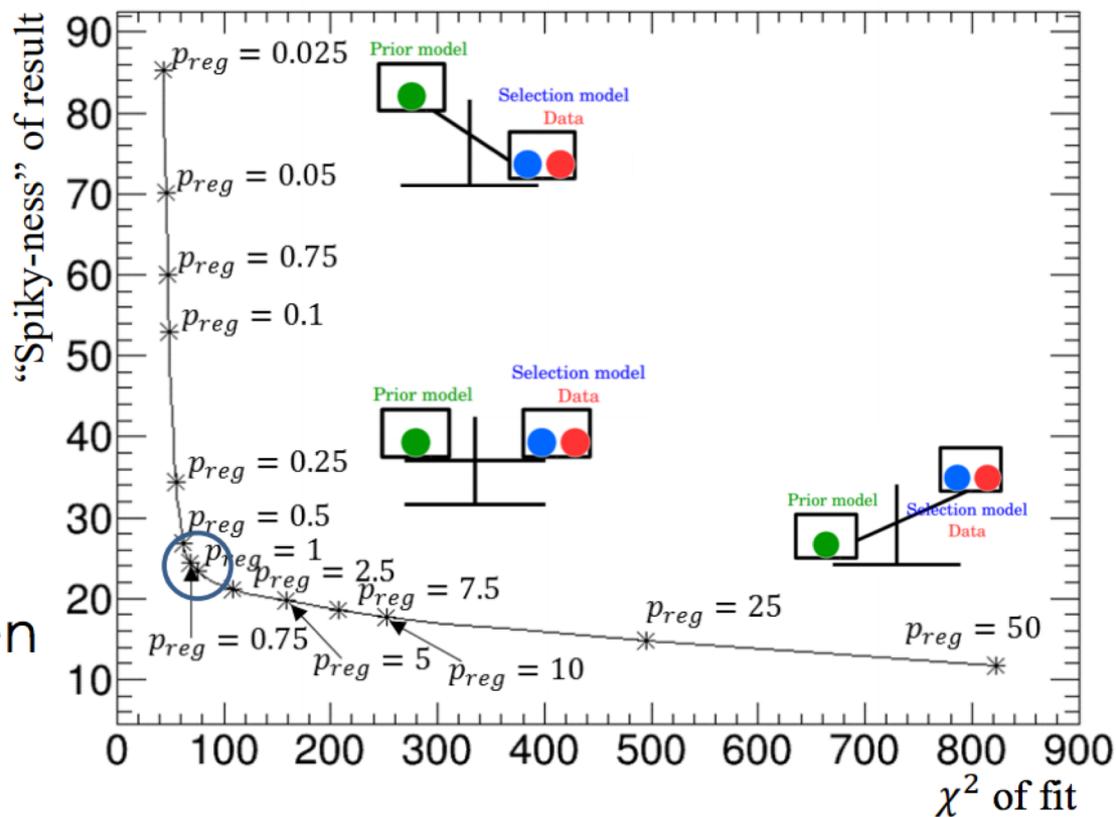
Tikhonov regularization,  $P(\lambda) = \|\lambda\|^2$ , varying  $\delta$



# How much to regularize?

## Regularisation optimisation: The L-curve

- Balance regulation with bias by **choosing the “kink” in the curve**
- **L-curve can be formed on real data** – data driven regularisation
- Well established method to **select the smoothest of many almost degenerate solutions:**



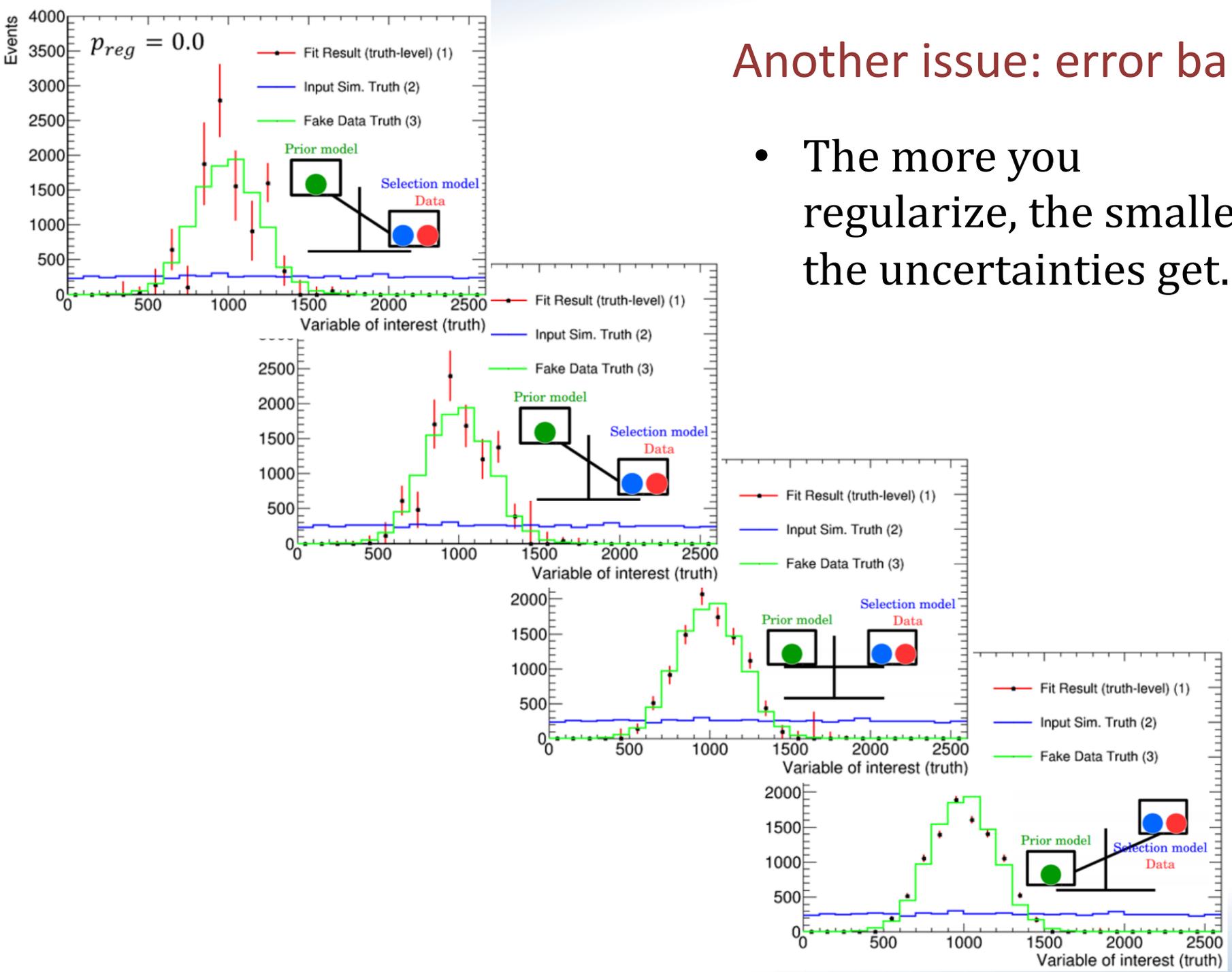
<http://epubs.siam.org/doi/abs/10.1137/1034115>

<http://epubs.siam.org/doi/abs/10.1137/0914086>

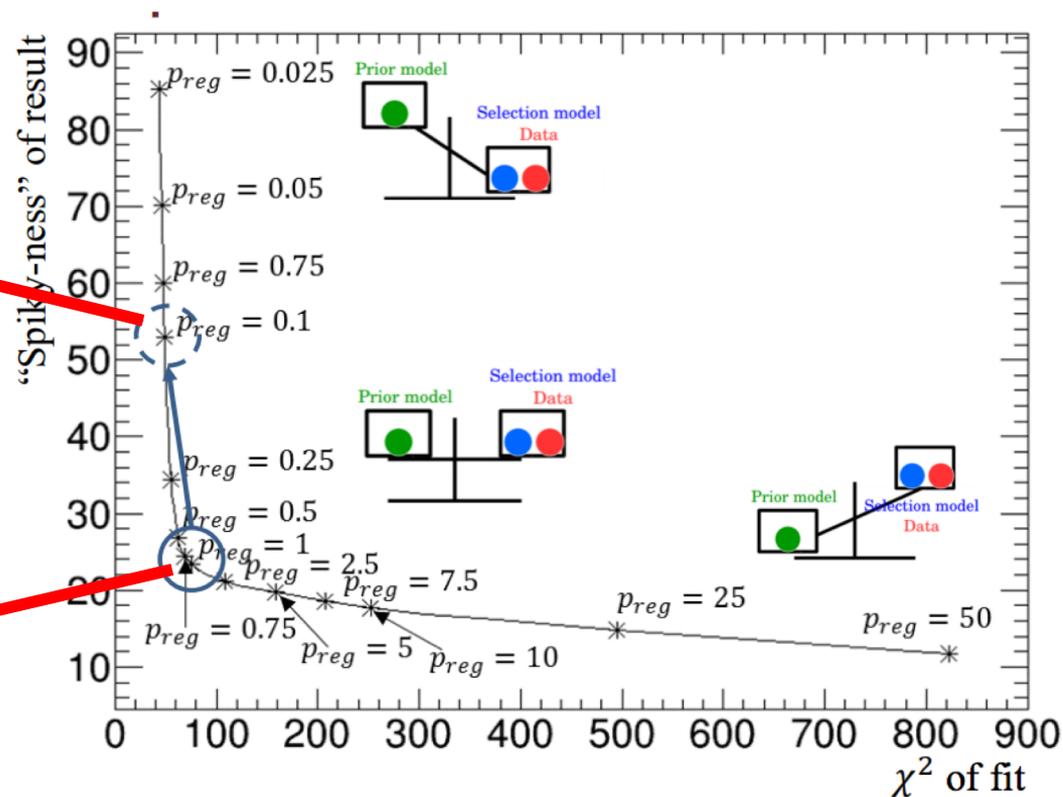
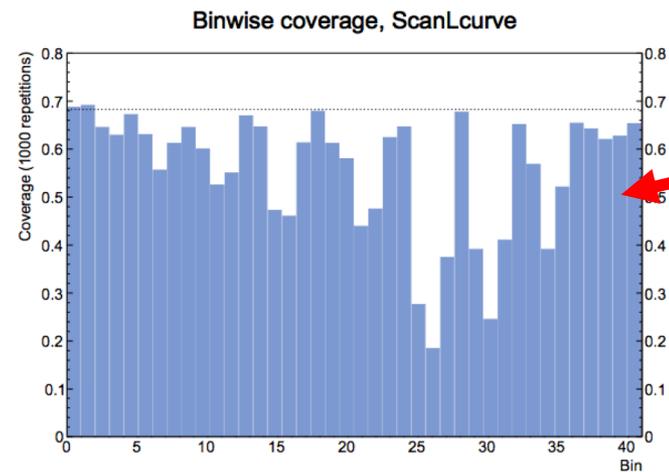
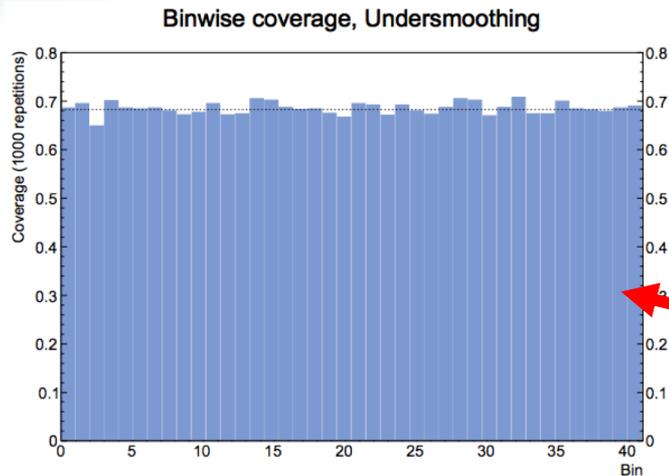
<http://arxiv.org/pdf/1205.6201v4.pdf> - use in TUnfold

# Another issue: error bars

- The more you regularize, the smaller the uncertainties get.



# Solution 1: Undersmoothing



- In general, the optimum point from the  $L$ -curve under-reports errors.
- By under-smoothing, you get more spikiness, but better represent the uncertainty.

# Solution 2: Regularizing with a Wiener Filter

1/23/2019

PHYSTAT-nu 2019

18

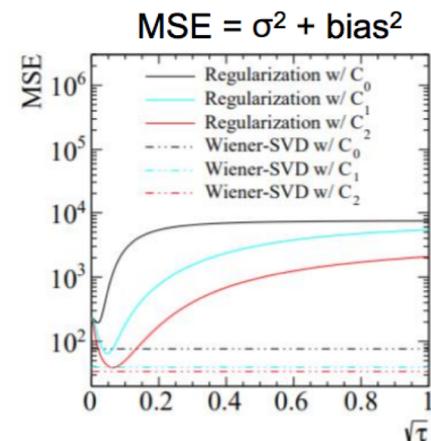
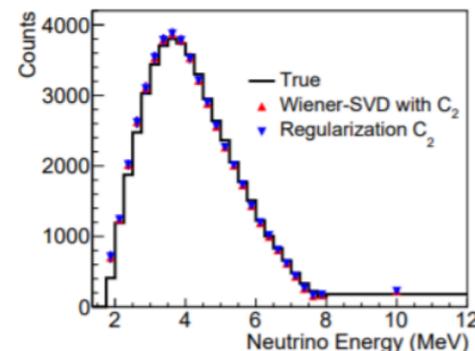
## Wiener-SVD Unfolding

- Inspired by the popular **Wiener Filter** used in digital signal processing to **maximize signal-to-noise ratio**
- Apply Wiener Filter to SVD spectrum (“effective frequency domain”)
  - Construct WF based on expected signal (event counts) and noise (fluctuations)
- Advantage
  - Avoids the scanning of regularization strength
  - Naturally balances bias vs. variance, leading to a small MSE

$$R = U \cdot D \cdot V^T \quad W_i = \frac{d_i^2 \cdot s_i^2}{n_i^2 + d_i^2 \cdot s_i^2}$$

$R$ : smearing matrix      $W$ : Wiener filter  
 $D$ : smearing in the effective frequency domain  
 $s_i$ : expected signal in effective freq. domain  
 $\overline{n_i^2} \equiv 1$ : (white) noise in effective freq. domain

X. Li et al. [JINST 12, P10002](#)



Also see: X. Qian's talk on Thursday

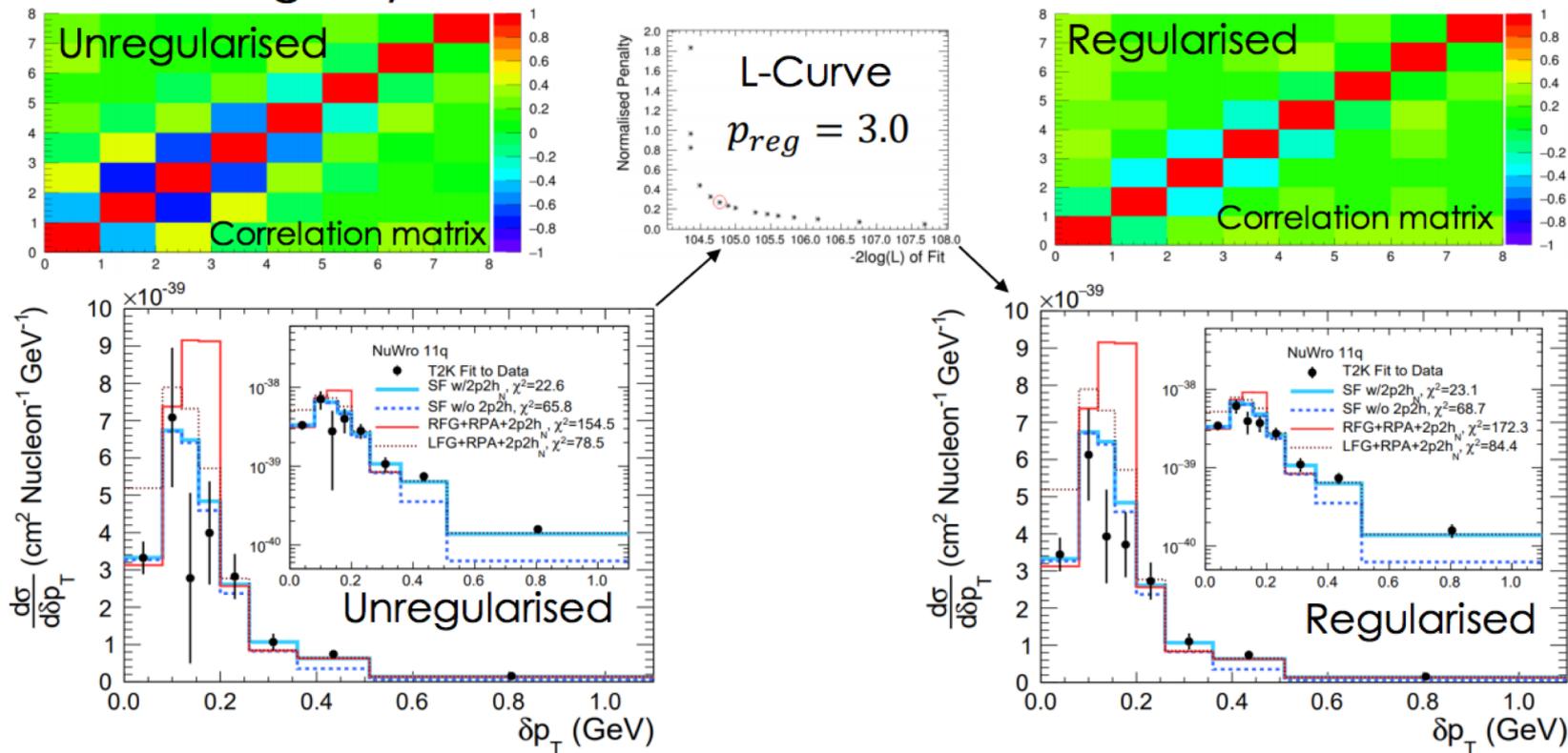
- This is the solution used by Daya Bay for their reactor spectrum measurements.

# Solution 3: Report Unregularized Results

## Case study: $CC0\pi$ in $\delta p_T$

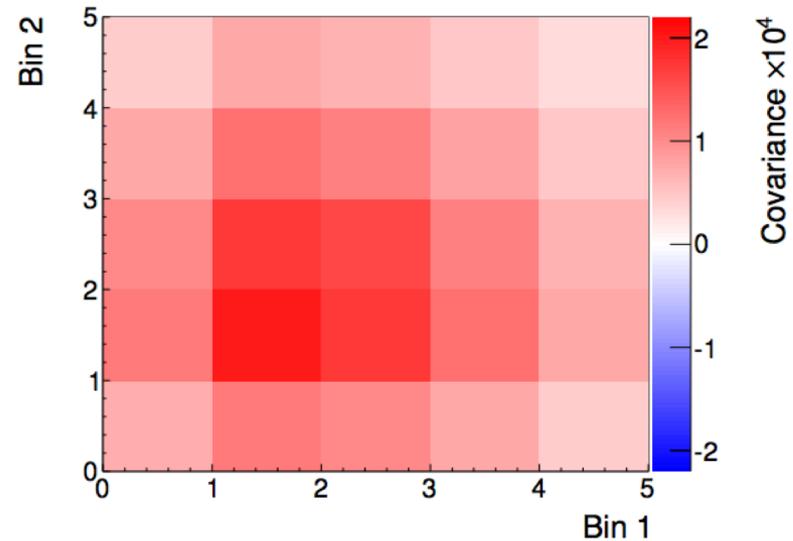
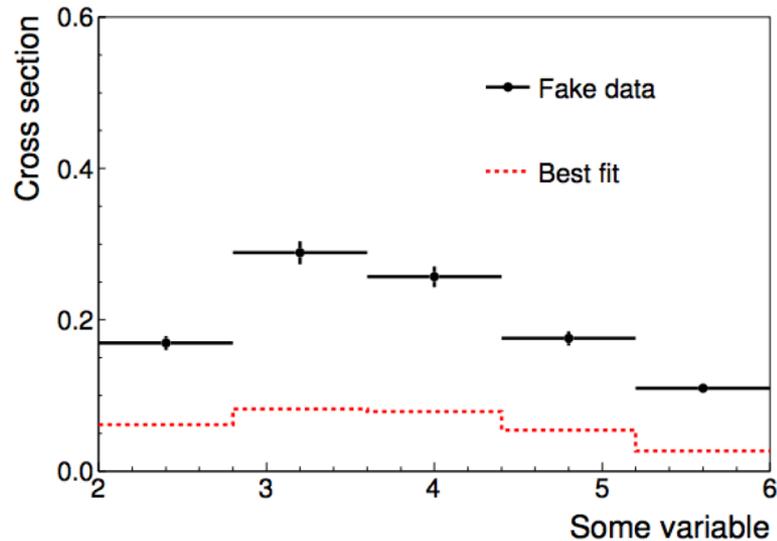
- Measure  $CC0\pi$ +protons cross section in missing transverse momentum ( $\delta p_T$ )
- Unregularised best for  $\chi^2$ , regularised best for actually showing anywhere

Phys. Rev. D **98**, 032003 (2018)



# But...Peelle's Pertinent Puzzle

## Fits to strongly-correlated data 2



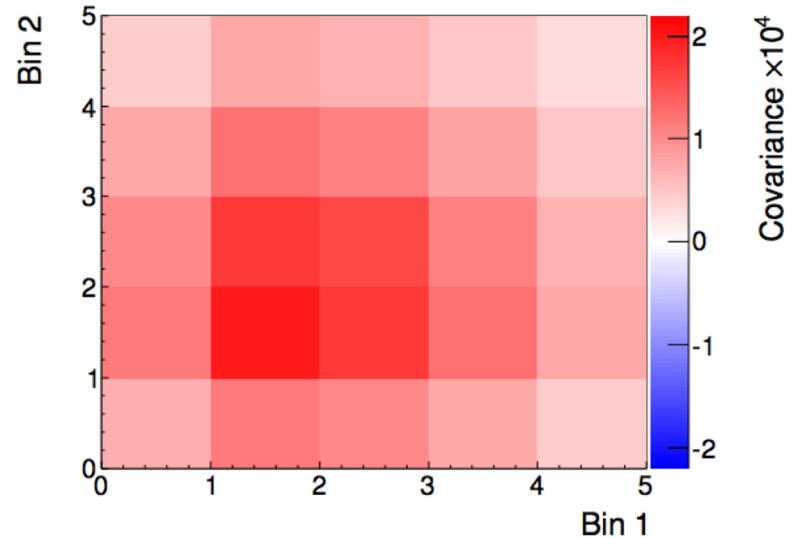
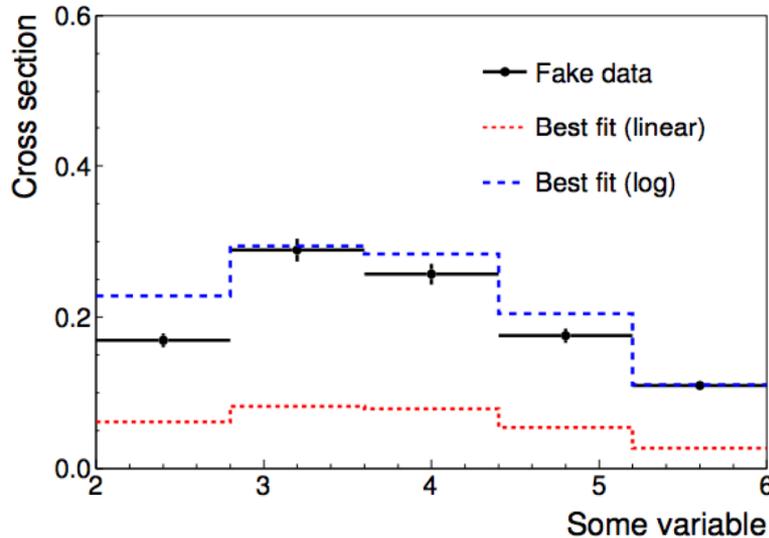
$$\chi^2 = (\mathbf{D} - \mathbf{M})^T \mathbf{C}^{-1} (\mathbf{D} - \mathbf{M})$$

$$C_{ij} = \sum_{\text{universe } k} (y_i^{(k)} - y_i^*) (y_j^{(k)} - y_j^*)$$

- ▶ “Multi-universe”: throw random systematic universes, re-extract result

# But...Peelle's Pertinent Puzzle

## Fits to strongly-correlated data 2



$$\chi^2 = (\mathbf{D} - \mathbf{M})^T \mathbf{C}^{-1} (\mathbf{D} - \mathbf{M})$$

$$C_{ij} = \sum_{\text{universe } k} (y_i^{(k)} - y_i^*) (y_j^{(k)} - y_j^*)$$

- ▶ “Multi-universe”: throw random systematic universes, re-extract result
- ▶ Empirically,  $y \rightarrow \log(y)$ , ameliorates the issue,  $\Rightarrow$  log-normal uncertainties on  $y(?)$

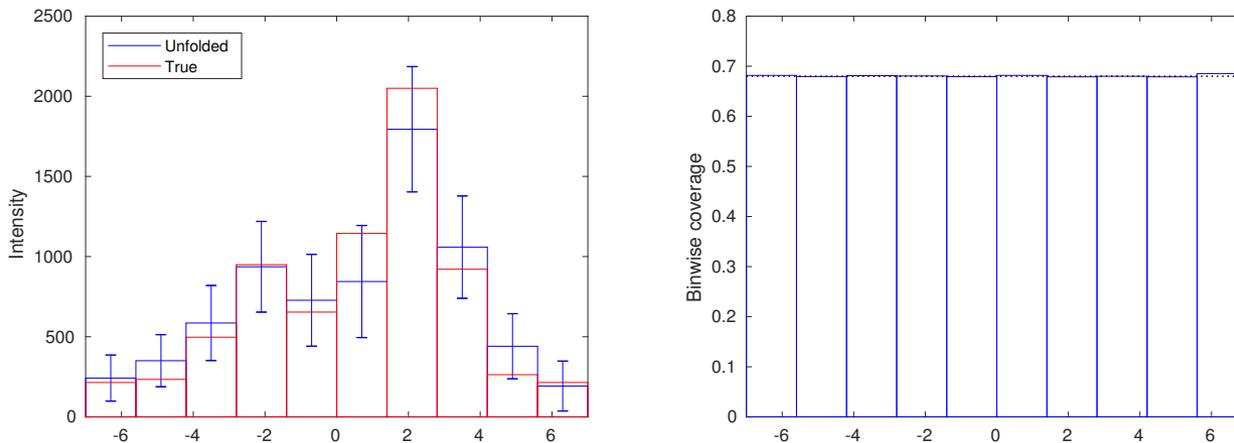
“Box-Cox transformation for resolving the Peelle's Pertinent Puzzle in curve fitting”, Oh and Seo 2004

- ▶ Is this the best way to communicate our systematics?

# Solution 4: Re-smearing instead of Regularizing

- Regularization not necessary if you re-smear after you unfold.
  - Unfold in narrow bins, and then rebin into wide bins.
  - Get both central values *and* errors right.

## Wide bins via fine bins, perturbed MC



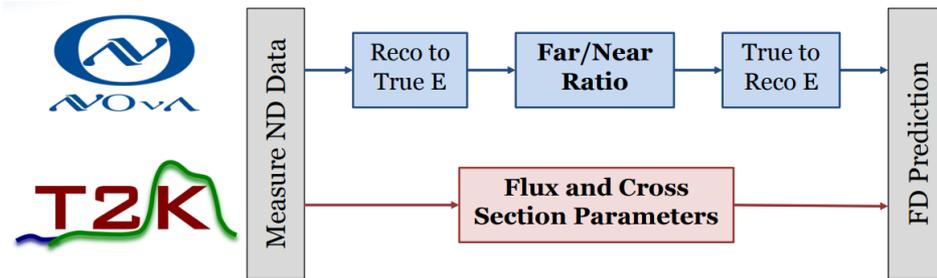
Wide bins via fine bins gives both correct coverage and intervals with reasonable length<sup>2</sup>

<sup>2</sup>More unfolded realizations given in the [backup](#).

# Solution 4: Re-smearing instead of Regularizing

- Might be interesting to apply to extrapolation, too.

## Application to long-baseline neutrino experiments?



*From A. Himmel's talk yesterday*

- Notice the steps in NOvA: ND unfolding  $\rightarrow$  ND/FD mapping  $\rightarrow$  FD smearing
- This means that we are really interested in functionals of the form

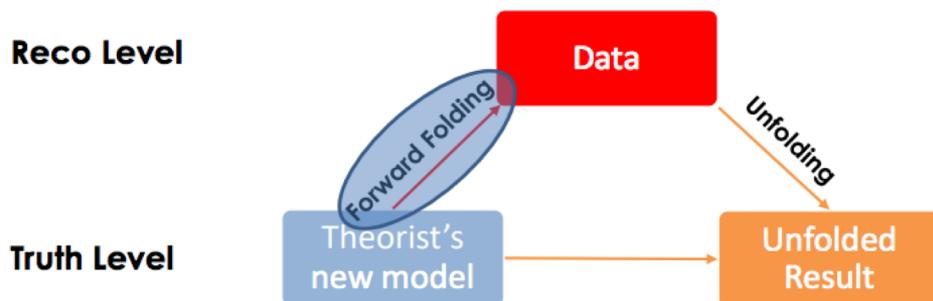
$$H_i[f] = \int_{S_i} \int_T k_{\text{FD}}(s, t) r_{\text{ND} \rightarrow \text{FD}}(t) f_{\text{ND}}(t) dt ds$$

- This should be a well-behaved functional since it is *resmearing* the unfolded spectrum
- Hence the previous discussion should apply: first unfold in ND using fine bins and no regularization, map to FD, resmear

# Solution 5: Forward Folding

## Just don't unfold!

- Producing a good *unfolded* result that can be interpreted by-eye with is **hard!** But maybe there's another way ...



Sounds simple, in practice quite complicated.

- Too few true variables:
  - Model-dependent true-to-reco behavior.
- Too many true variables:
  - Become prohibitive to quantify all the uncertainties.

## Response Matrix Utils (Lukas Koch)

- A tool for forward folding analyses
  - Builds response matrix
  - Tests model dependence
  - Evaluates uncertainties
  - Compare model to data (likelihoods, p-values, MCMC)
- More information: <https://remu.readthedocs.io/>



# Uncertain Uncertainties

**HOWEVER:** the main question is to make sure one does not misevaluate what one knows (information) or does not know

Two quotes to conclude:

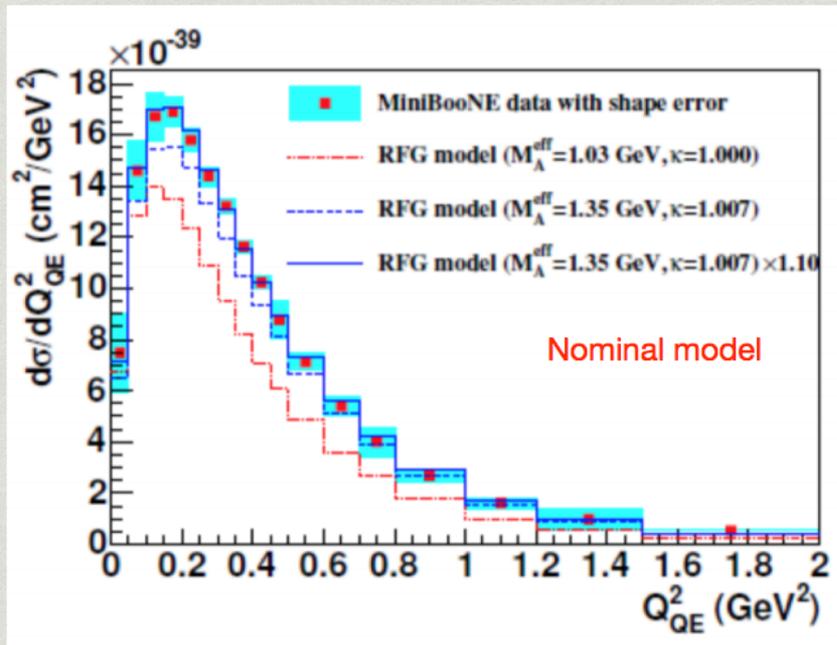
- OK to use a parametric fit to a well known problem ( $\nu$  oscillation, Z line shape etc.)
- It is however not recommended (i.e. should be forbidden really) to fit some data with a convenient but arbitrary or unsure or model-dependent function (i.e. fit looks good) and act as if the error matrix of the fit represents the uncertainty on the fit data. It does not, -- and this can go very wrong!

## Experiments confronting data/MC discrepancies

Experiments need a model that describes their data

However, often, **data/MC agreements are handled in a non-satisfactory way**

- Overemphasising own data - breaking consistency with other neutrino data
- Largely ignoring complementary constraints from charged-lepton and hadron scattering



A typical (and conveniently old and non-controversial) example comes from the MiniBooNE experiment:

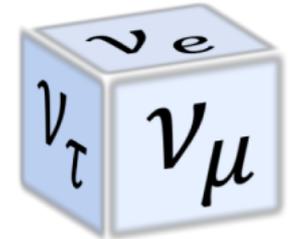
*Tweaking axial form factor parameter*

- Axial mass 1.03 → 1.35 GeV
- Not consistent with bubble chamber results

*Tweaking Pauli blocking*

- Not consistent with textbook physics

**Good description of own data.  
But wrong physics!**



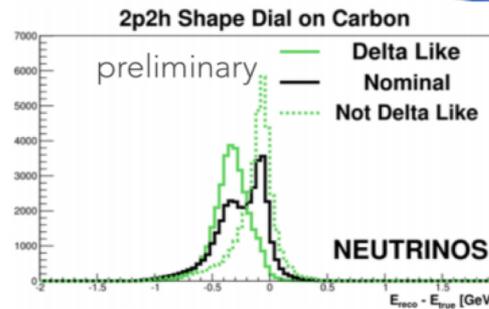
## Have we learned from this?

- Yes, we have learned.
- Our models for neutrino interactions have improved their description of the data.
- We include more possibilities in trying to describe differences between data and interaction models.

- No, we haven't learned.

It is however not recommended (i.e. should be forbidden really) to fit some data with a convenient but arbitrary or unsure or model-dependent function (i.e. fit looks good) and act as if the error matrix of the fit represents the uncertainty on the fit data. It does not, -- and this can go very wrong!

M. Datkiewicz,  
k, 36 A. Blondel, 12, \*  
Ruizza Avanzini 10



- T2K fits an *ad hoc*  $f_{\text{“delta like”}}$  to ND data.

25 January 2019

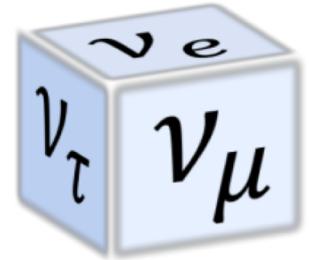
K. McFarland, Neutrino Summary

31

- Clear potential for “over-fitting”:
  - Data used to establish need for a fix, the form of that fix, and the parameters.
- But, without theoretical guidance, it is hard to avoid.

# Uncertainties from Theory

- That being said, theory isn't necessarily a panacea...



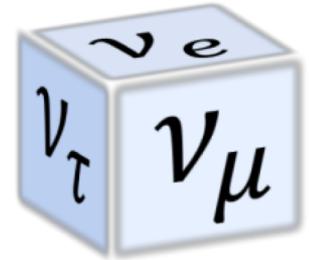
## Uncertainty estimation by survey of theory models

- This has an obvious and fatal failure mode.
- Dave Soper compared this method to attempting to measure the width of a valley...



# Uncertainties from Theory

- That being said, theory isn't necessarily a panacea...



## Uncertainty estimation by survey of theory models

- This has an obvious and fatal failure mode.
  - Sheep read each others' papers.
- It's just wrong.
- But we continue to do it because often there are not straightforward alternatives.
  - In the PDF community, this was addressed by fitters explicitly producing uncertainties as an output.
- Dave Soper compared this method to attempting to measure the width of a valley...



... by studying the variance of the position of sheep grazing in it.

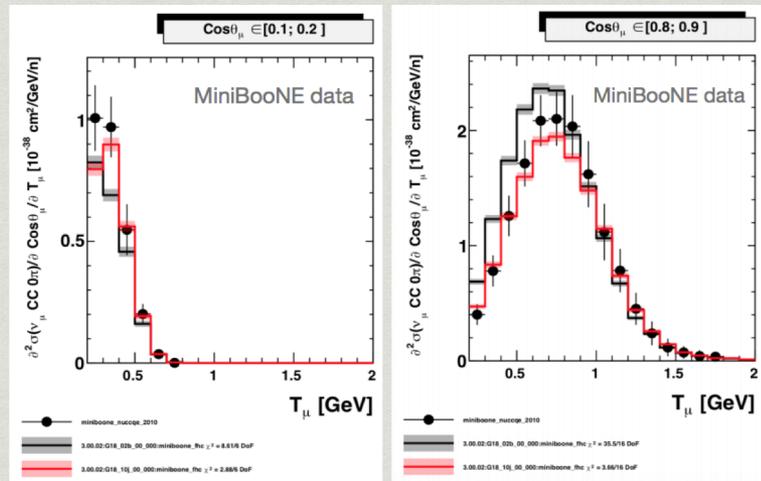
# Tensions in Datasets

- What to do when trying to fit theory to data, but datasets disagree?
- Picked one cross section example at right, but brought up numerous times.
  - Another common example: sterile neutrinos.
- Genie's solution: "partial tunes"
  - Several tunes which only fit consistent datasets.
  - Up to the user to decide among them.

# Tensions in neutrino interaction data

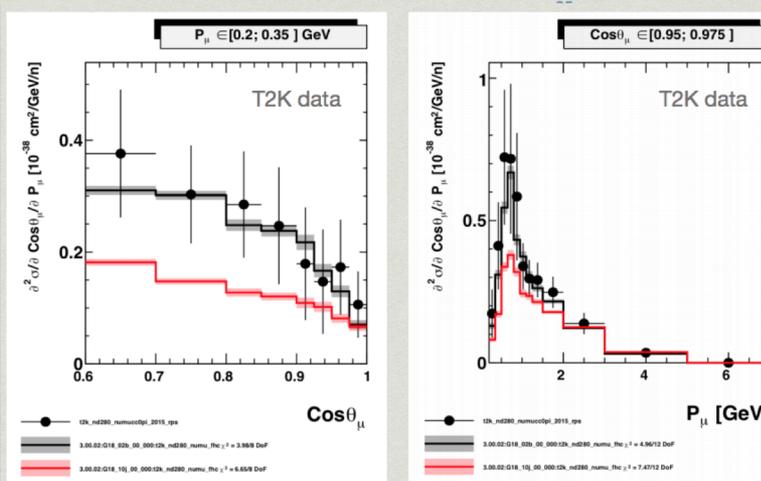
Example: Tensions between data from modern experiments

- \* **G18\_02b**: An improved empirical model in GENIE3.
- \* **G18\_10j**: A more theory-based model configuration in GENIE3.



GENIE tune	$\chi^2/ndf$
<b>G18_02b_00_000</b>	330 / 137
<b>G18_10j_00_000</b>	63.7 / 137

MiniBooNE data  
prefers G18\_10j

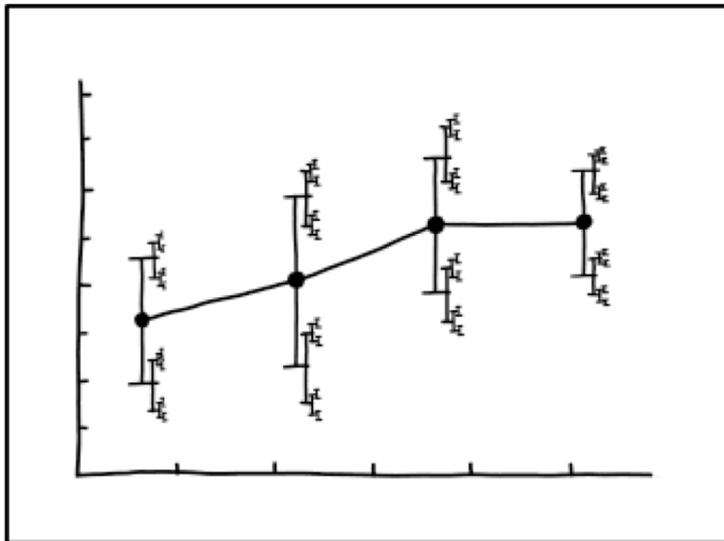


GENIE tune	$\chi^2/ndf$
<b>G18_02b_00_000</b>	73.9 / 80
<b>G18_10j_00_000</b>	80.4 / 80

T2K data  
prefers G18\_02b

[https://indico.cern.ch/event/735431/contributions/3137831/attachments/1785482/2906728/Neutrino\\_Summary\\_PHYSTATnu\\_2019\\_final.pdf](https://indico.cern.ch/event/735431/contributions/3137831/attachments/1785482/2906728/Neutrino_Summary_PHYSTATnu_2019_final.pdf)

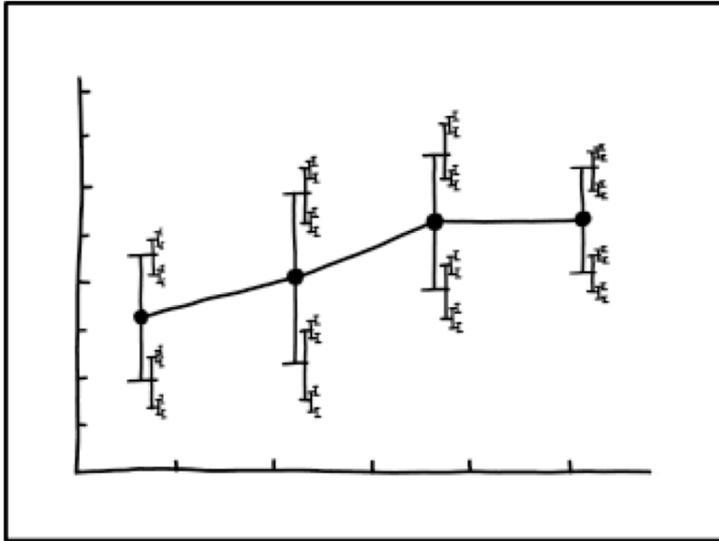
# Errors on Errors



I DON'T KNOW HOW TO PROPAGATE  
ERROR CORRECTLY, SO I JUST PUT  
ERROR BARS ON ALL MY ERROR BARS.

<https://xkcd.com/2110/>

# Errors on Errors



I DON'T KNOW HOW TO PROPAGATE ERROR CORRECTLY, SO I JUST PUT ERROR BARS ON ALL MY ERROR BARS.

<https://xkcd.com/2110/>

- Let the true  $\sigma_i^2$  be unknown.
- $v_i$  is our estimate based on the data
  - Equiv. of a “pull term”
- $r_i$  is the fractional uncertainty on  $\sigma_i^2$

$$\ln L(\mu, \theta) = \ln P(y|\mu, \theta) - \frac{1}{2} \sum_{i=1}^N \frac{(u_i - \theta_i)^2}{\sigma_{u_i}^2}$$

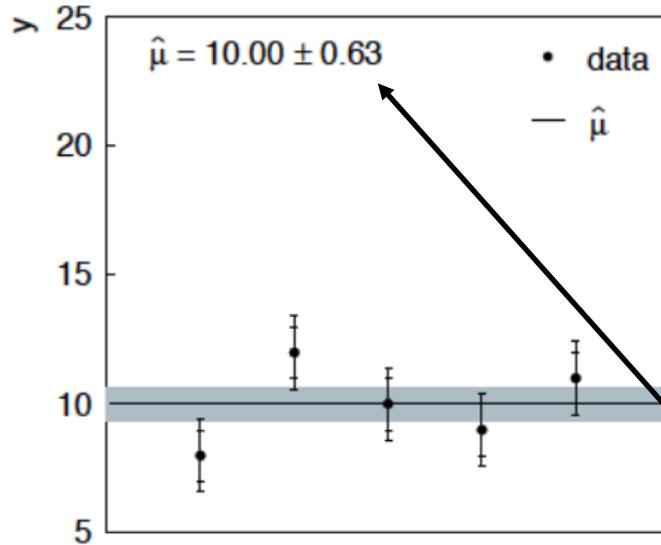


$$\begin{aligned} \ln L'(\mu, \theta) &= \ln L(\mu, \theta, \widehat{\sigma}_{u,i}^2) \\ &= \ln P(y|\mu, \theta) - \frac{1}{2} \sum_{i=1}^N \left( 1 + \frac{1}{2r_i^2} \right) \ln \left[ 1 + 2r_i^2 \frac{(u_i - \theta_i)^2}{v_i} \right] \end{aligned}$$

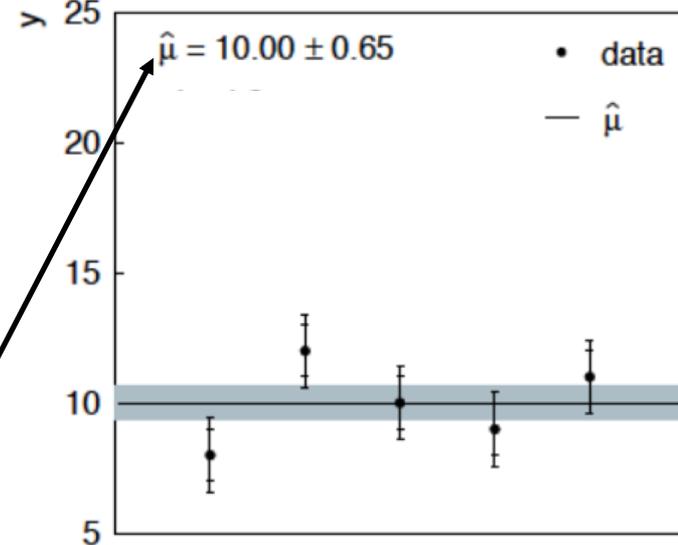
In limit of small  $r_i$ ,  $v_i \rightarrow \sigma_{u,i}^2$  and the log terms revert back to the quadratic form seen with known  $\sigma_{u,i}$ .

# Errors on Errors Example: Find a Mean

Known Systematic  
( $r_i = 0.01$ )



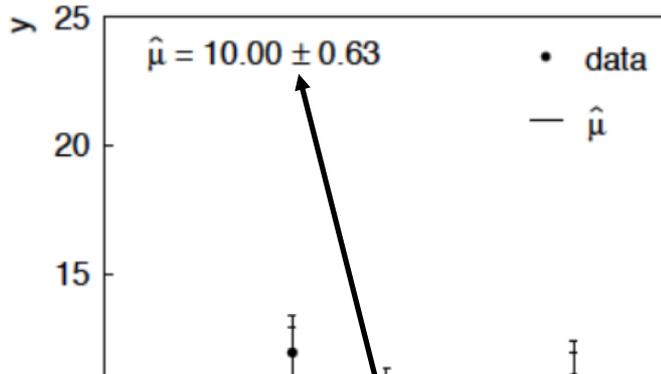
Uncertain Systematic  
( $r_i = 0.2$ )



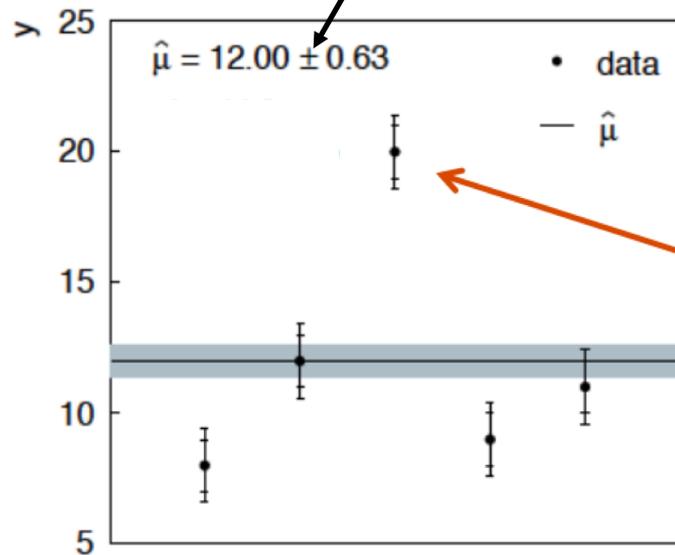
If the data is well-behaved,  
they give ~the same mean  
and confidence interval.

# Errors on Errors Example: Find a Mean

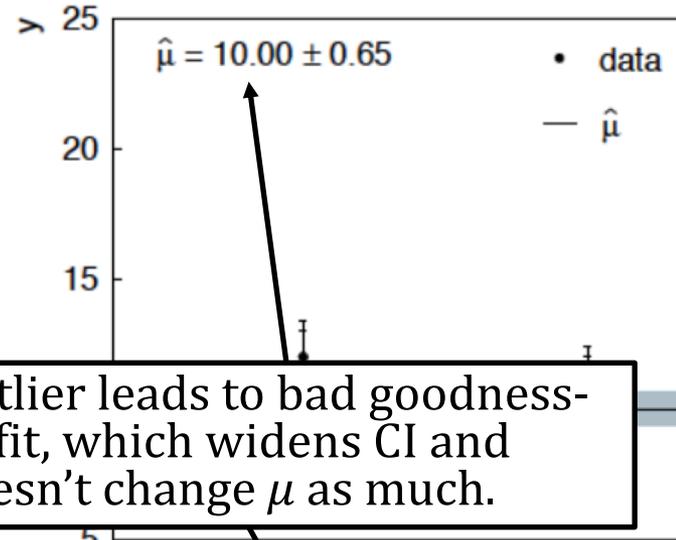
Known Systematic  
( $r_i = 0.01$ )



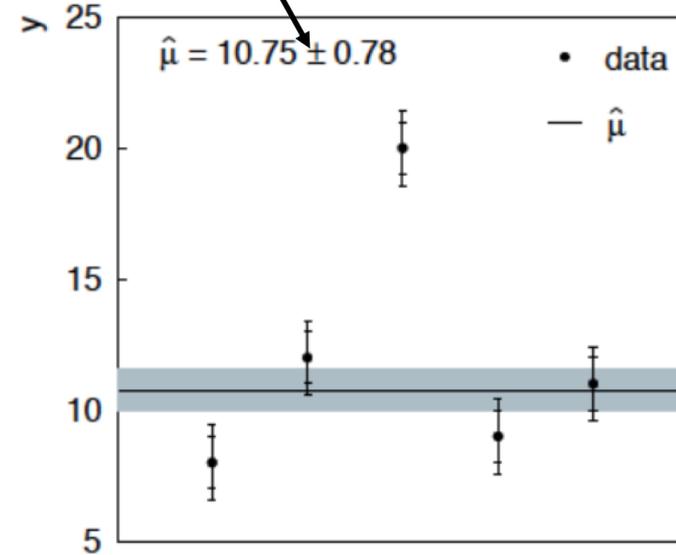
Outlier makes a big change in  $\mu$  and no difference to CI.



Uncertain Systematic  
( $r_i = 0.2$ )



Outlier leads to bad goodness-of-fit, which widens CI and doesn't change  $\mu$  as much.



# Conclusions

- The space of possible statistical analysis is much larger than it seems.
  - The right statistical technique is generally **problem-dependent**.
- Often, it seems like there's 1 right way, because it's been done before
  - ...or it's in the PDG
  - Not necessarily bad! If a technique is well-known, less justification is needed.
- But, be careful of assumptions!
  - Everything that seems “simple,” assumes a certain kind of problem.
  - If you don't meet the assumptions (small stats – looking at you), then the answer isn't valid even if the technique is familiar.
- Is it time for neutrino statistics committees?
  - The big LHC experiments have standing committees to discuss these issues.
  - Most neutrino experiments are too small by themselves (notable exception: DUNE), but maybe if we join forces?