er : 27404 | Run number : 8115 | Spill : 51004 | Time : Mon 2012-01-23 06:04:28 JST |Trigger

State of the Nu-tion

平成29年 06月 24日

Data releases & complications Do we need unfolding?

Morgan O. Wascko

m.wascko@imperial.ac.uk

Imperial College London

Outline

- Introduction from the oscillation analysis perspective
- Common analysis methods
- Sermon on model independent measurements
- Personal perspectives on alternate analysis ideas:

1. Alternate approach to the issue of unfolding
2. Providing maximal information
3. Generator-free MC?
4. Making best use of complex data

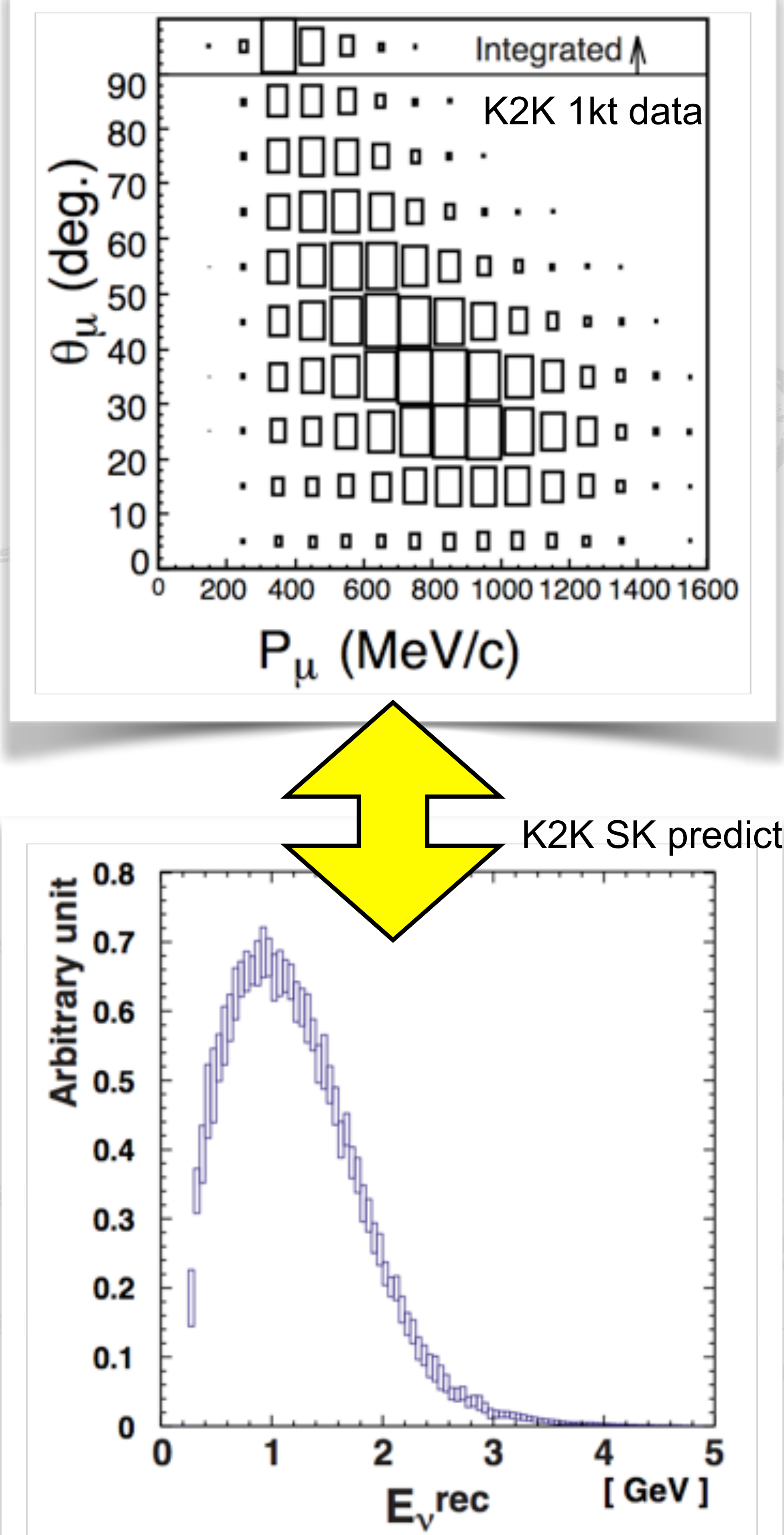
- Conclusion

What does OA need?

- Predictions:
  - Event rates
  - final state particle kinematics
- Need to accurately calculate inferred (physics) variables from our observed variables
  - For oscillations, need to E$_\nu$
    - **different ways to do this**
      - *All methods need good xsecs!*
      - all beams are relatively wideband
      - all detectors are relatively poor at neutron detection
- Need to accurately predict background contamination

➡Need to understand neutrino-nucleus cross-sections precisely

➡Need good models



K2K SK prediction

Phys.Rev.D **74** 072003 (2006)

Xsecs and Oscillations

- Cross section **models** used by experiments do not describe observations by: K2K, MiniBooNE, SciBooNE, Argoneut, MINERvA, T2K, …
  - Leads to inflation of systematic uncertainties
- Model dependence often injected into data analysis
  - Inferred variables
    - Energy, Q^2 reconstruction
  - Background subtraction
- Using discrepant models will always give such uncertainties.
  - Need to use better models!

Systematic uncertainties reported by the world's most sensitive ν_e appearance experiments vs. time

Experiment	xsec err (%)	total err (%)
MiniBooNE (2007)	12.3	17.6
T2K (2012)	7.5	10.3
T2K (2016)	4%	6%

How are we improving the errors?
Using better models!
How are we improving models?
Tuning with better data!

Analysis Methods

Method 1: template fit

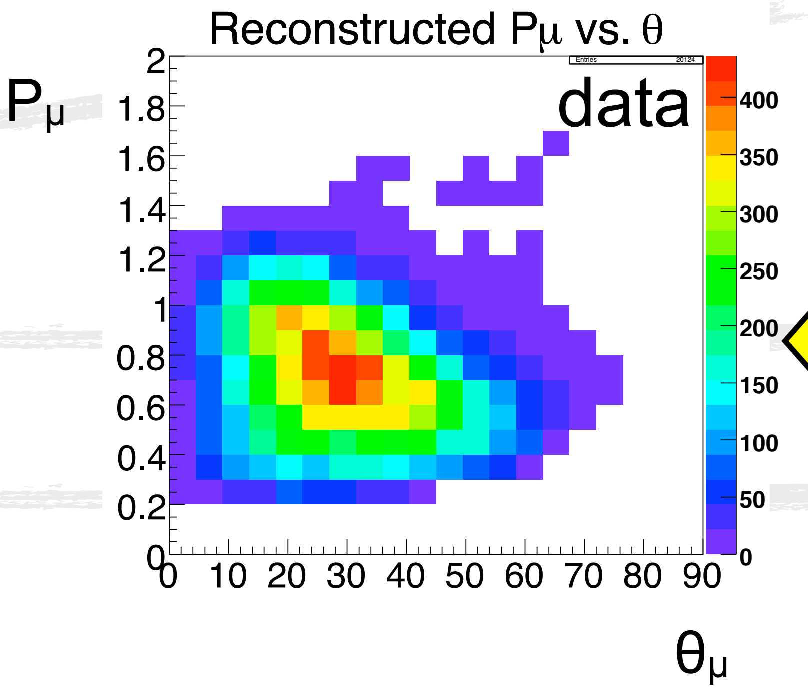
Compare data to MC, within context of a model, using templates.

$$\sigma_i = f_i < \sigma^{pred} >_i = \frac{f_i N_i^{pred} P_i}{\epsilon_i T \Phi_i}$$

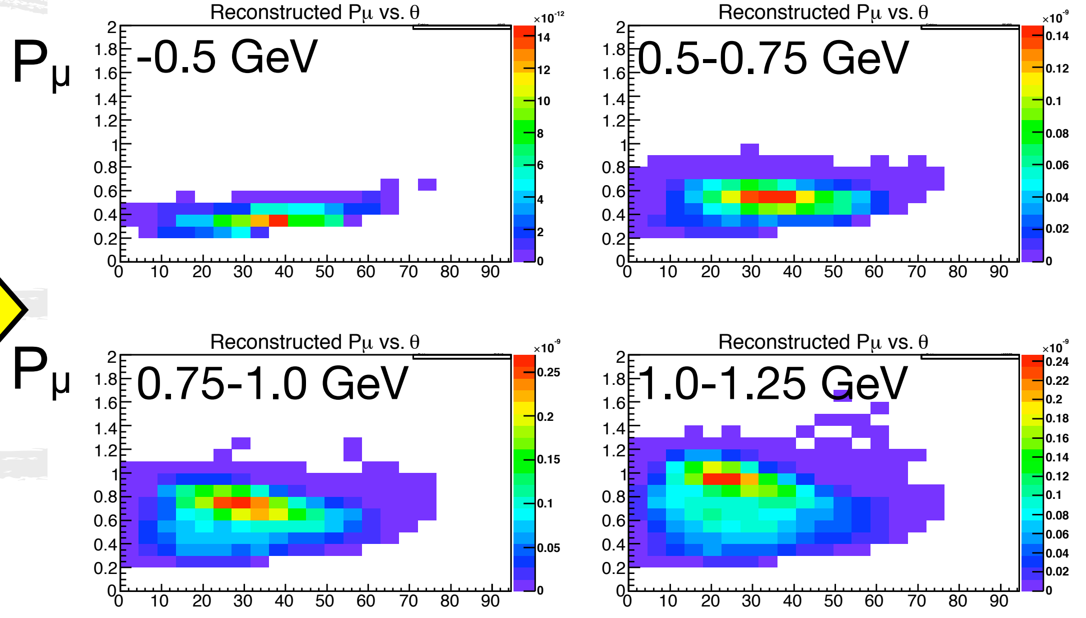
i = bin of xsec variable
f_i = normalisation factor
N_i^{pred} = predicted # of events
P_i = purity
ϵ_i = efficiency
T = number of nuclear targets
Φ_i = neutrino flux per bin

Other variants are used, for example the T2K off-axis CCQE and on-axis E_ν analysis

Templates can be produced in observed dynamical variables, different from xsec variable.



MC template (MRD-stop)

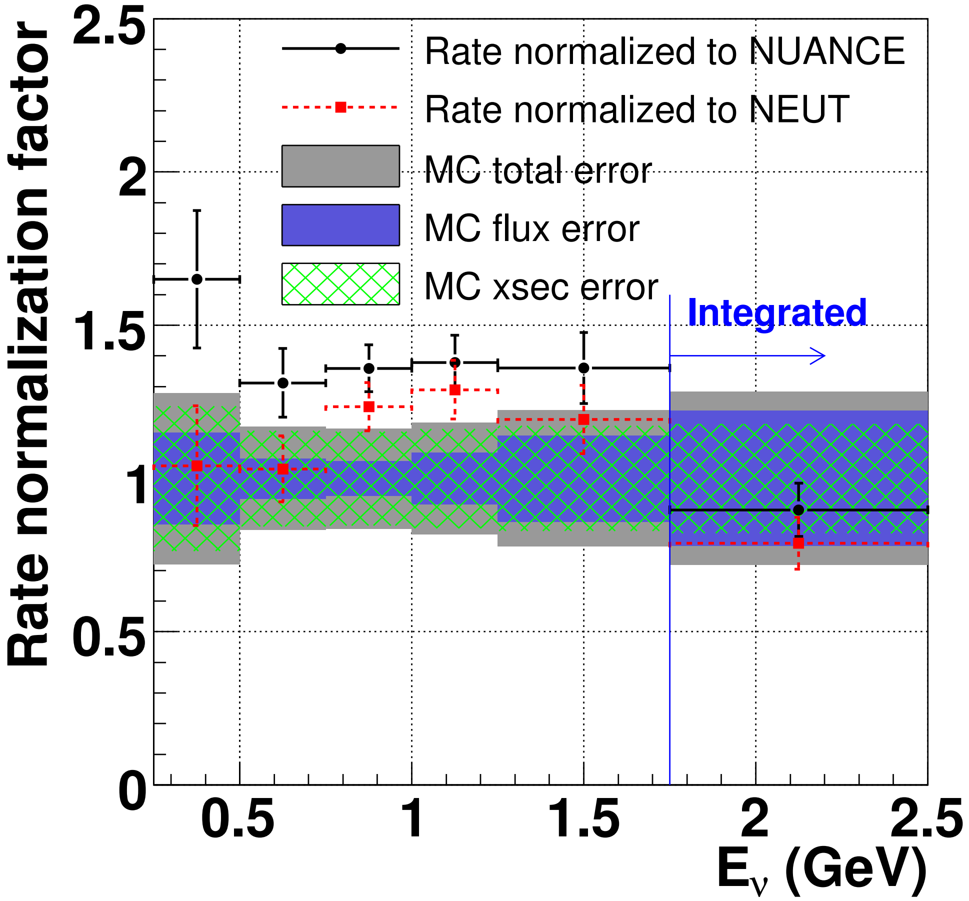


Method 1: template fit

Fit for values of f_i that minimise some GoF parameter (χ^2, likelihood), and use MC to infer the measured value of cross section.

$$\chi^2 = \sum_{j,k}^{Nbins} (N_j^{obs} - f_j N_j^{pred})(V^{sys} + V^{stat})^{-1}_{jk}(N_k^{obs} - f_k N_k^{pred})$$

Uncertainties: estimated with fake data studies by repeating the template fit with MC variants.



Advantage: This method is especially useful for measuring cross sections as functions of inferred variables, like the input variable E_ν or internal variable Q^2, and for parameter tuning.

Drawback: This method is susceptible to model bias. If your MC model differs from nature in some important way, you can easily infer the wrong answer!

Method 2: Matrix unfolding

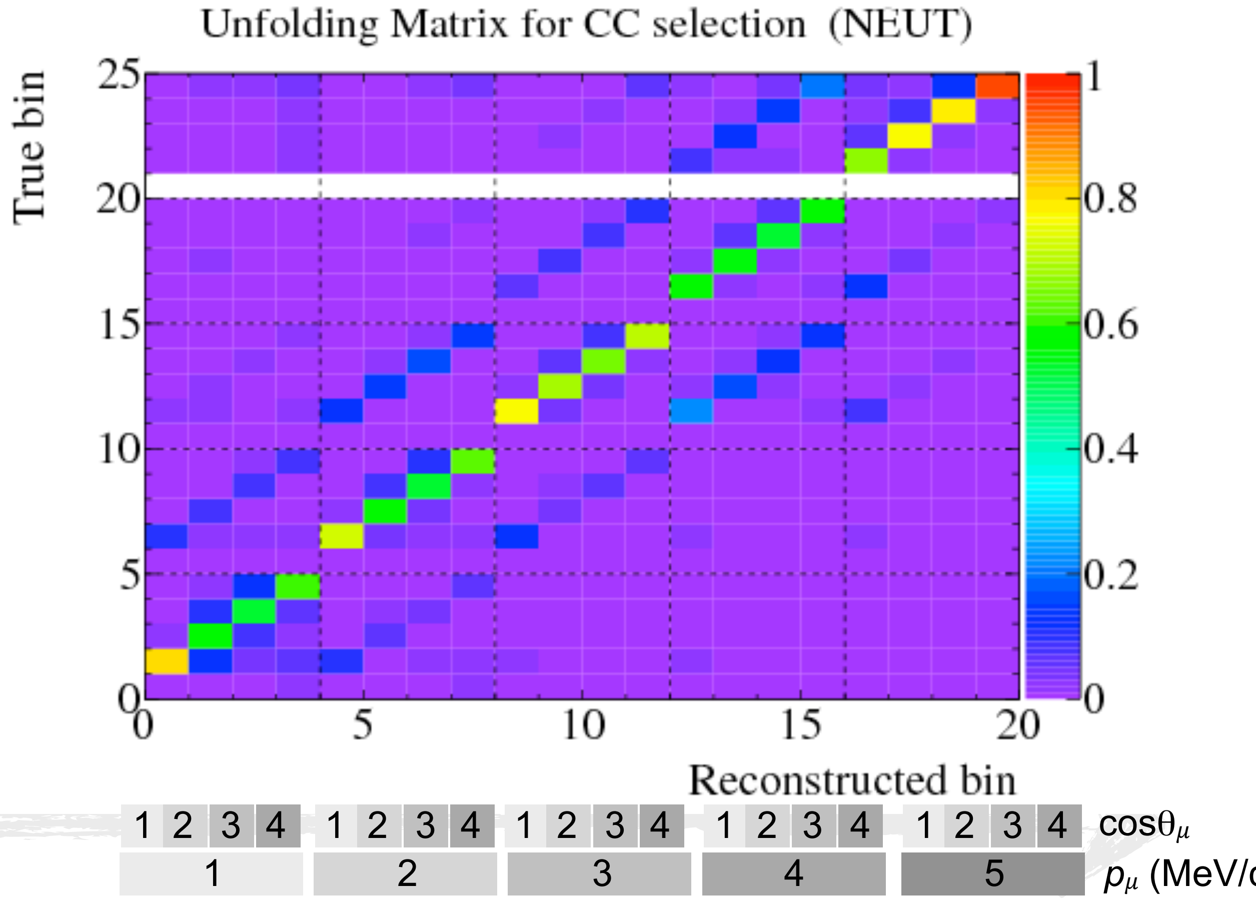
- Calculate cross section directly from number of events

$$\hat{N}_i = \frac{\sum_j U_{ij}(N_j^{obs} - B_j)}{\epsilon_i}$$

1. Apply BG/purity correction
2. Unfold to correct detector smearing
   - Different methods available
3. Apply efficiency correction
4. Normalise with neutrino flux and number of nuclear targets to get cross section

➡Result is flux averaged differential cross section

j = bin of reconstructed variable
i = bin of corrected (“true”) variable
N_j^{obs} = observed # of events
B_j = background events
(could use purity correction η_j)
U_{ij} = unfolding matrix
ϵ_i = efficiency



$$\frac{d\sigma}{dx_i} = \frac{1}{T\Phi_\nu}\frac{\hat{N}_i}{\Delta x_i}$$

T = number of nuclear targets
Φ_ν = total integrated neutrino flux
Δx_i = bin width

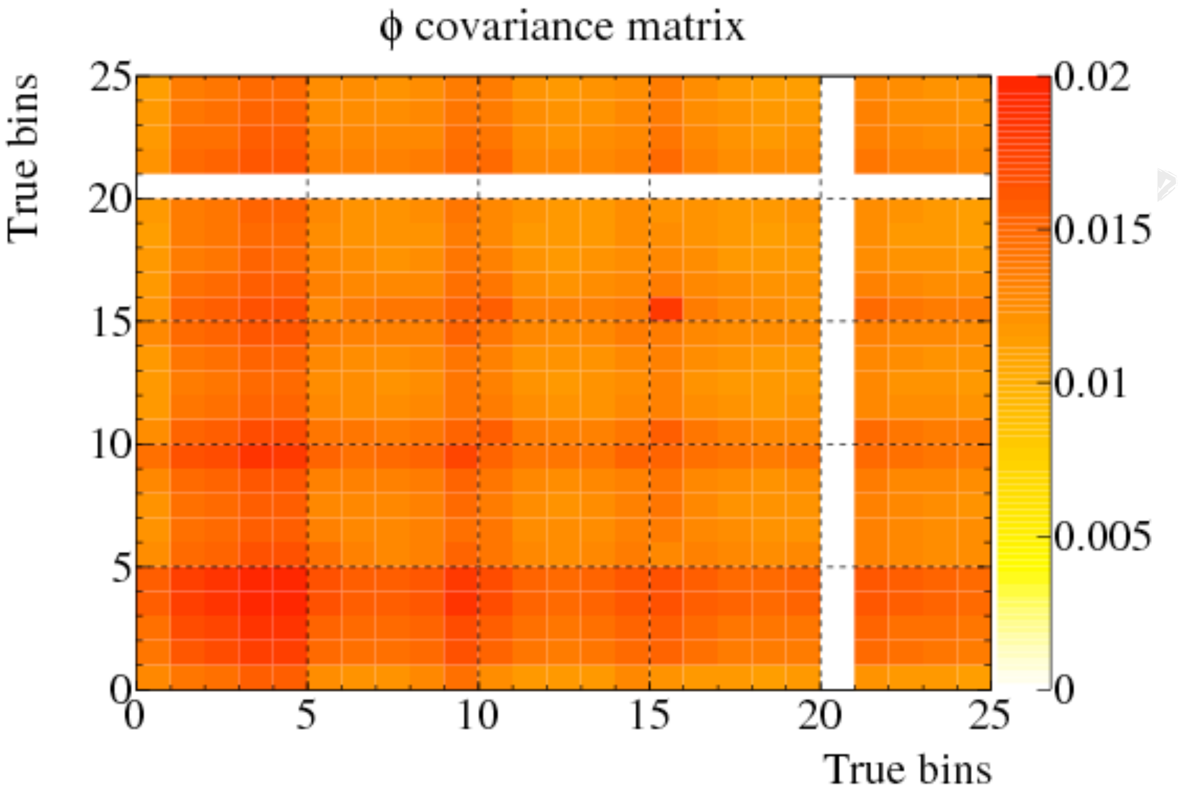
Method 2: Matrix unfolding

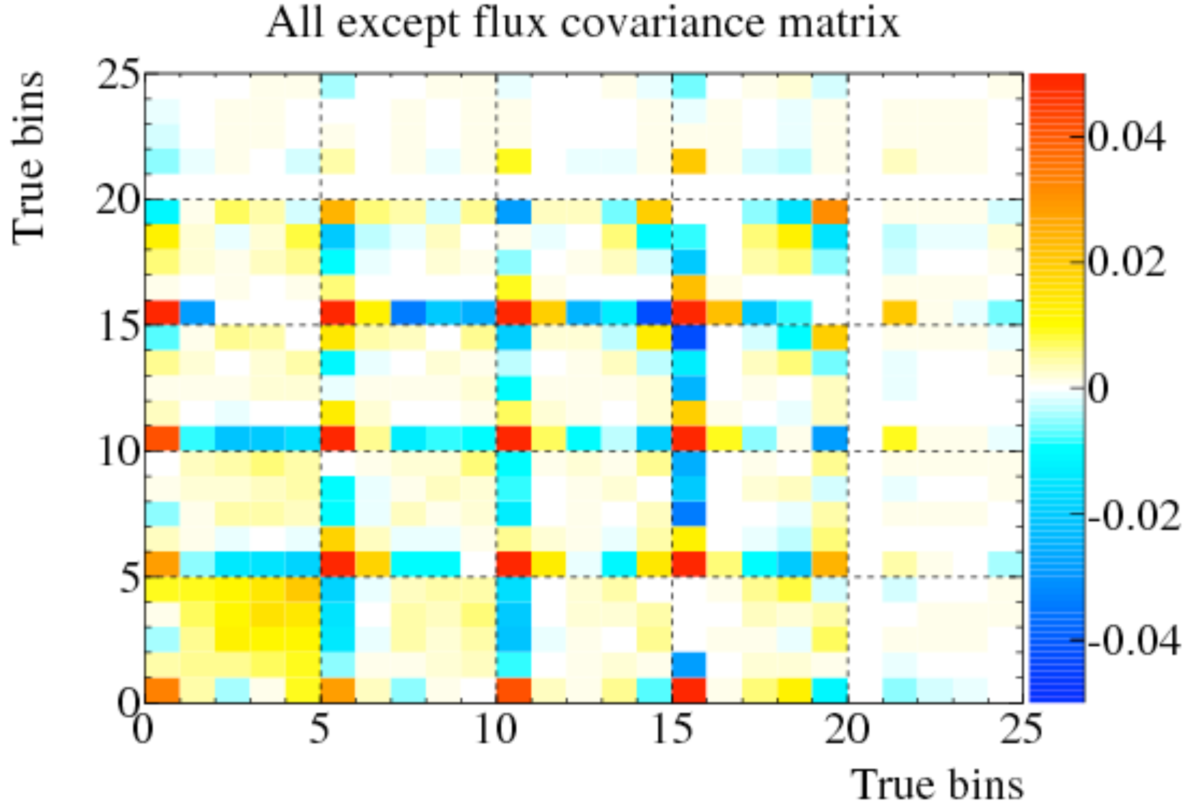
$$V_{kl} = \frac{1}{N_{var}} \sum_{s}^{N_{var}} (\sigma_k^s - \sigma_k^{nom})(\sigma_l^s - \sigma_l^{nom})$$

- Use MC variants to create covariance matrix
- Neutrino flux is (usually) just a normalisation error
  - We do, of course, propagate the full shape covariance
- Very useful to separate out the flux error

➡Potential for reducing model dependence with this method

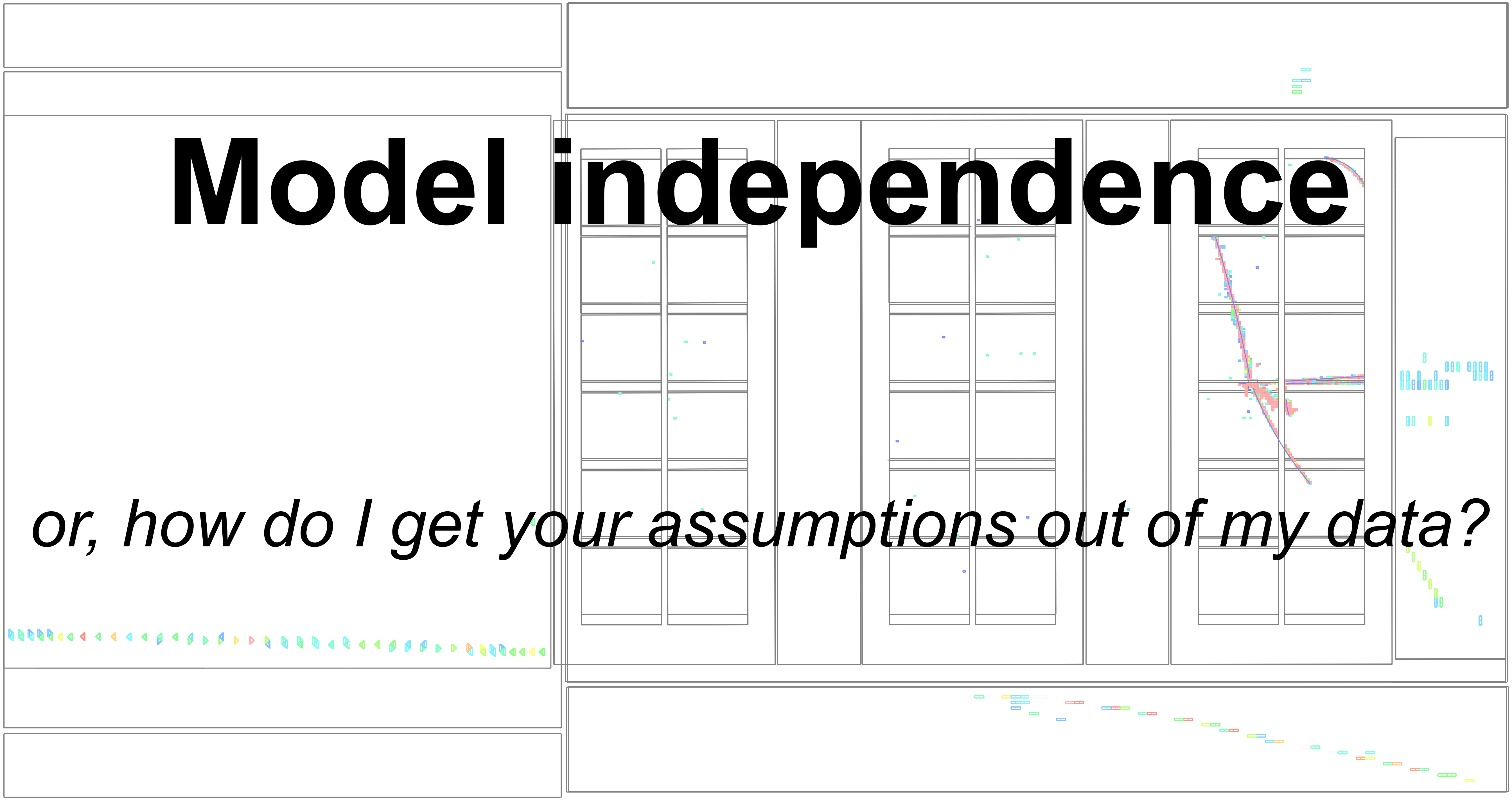
- But, issues with unsmearing…





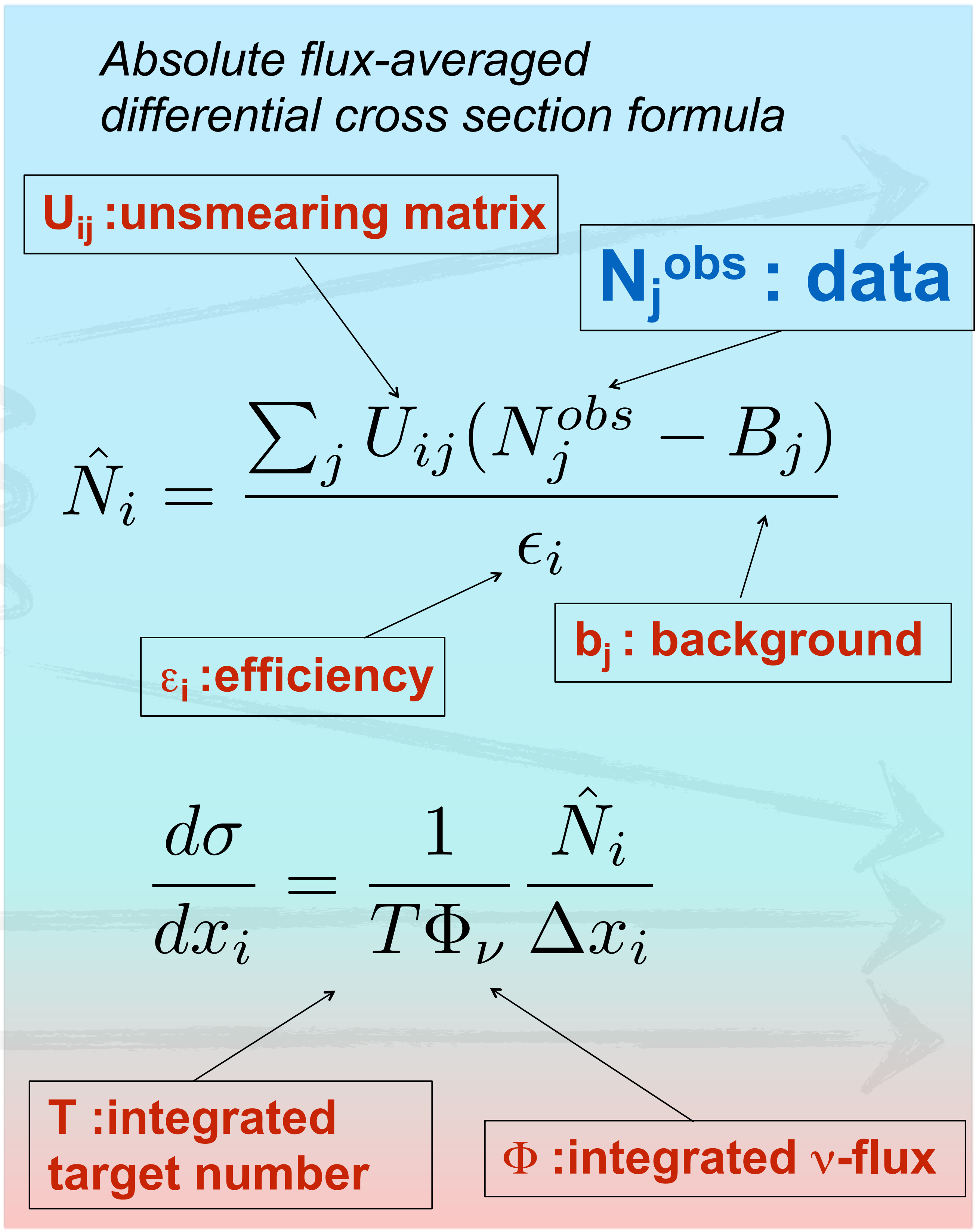
Model independence

or, how do I get your assumptions out of my data?



What does model dependence mean?

- Distinguish between σ model and detector model
  - Any MC-derived quantity is, of course, model-dependent
- Restricting corrections (unsmearing, BGs, efficiencies) to detector MC quantities—**not cross section processes**—is probably the best we can do
- This is why we should publish final state particle cross sections, in addition to process measurements, etc.



Absolute flux-averaged differential cross section formula

U_{ij} :unsmearing matrix

N_j^{obs} : data

$$\hat{N}_i = \frac{\sum_j U_{ij}(N_j^{obs} - B_j)}{\epsilon_i}$$

ε_i :efficiency

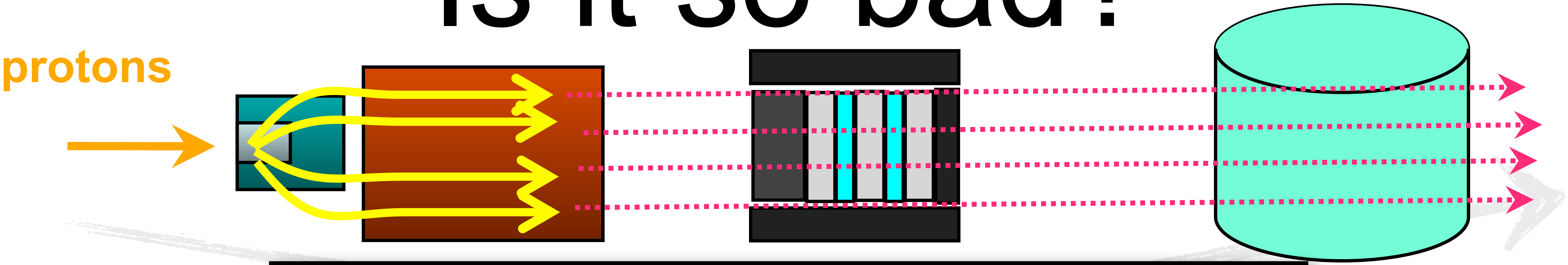
b_j : background

$$\frac{d\sigma}{dx_i} = \frac{1}{T\Phi_\nu}\frac{\hat{N}_i}{\Delta x_i}$$

T :integrated target number

Φ :integrated ν-flux

Is it so bad?



$$\Phi_\nu^{near}(E)\cdot\sigma^{near}(E)\cdot\varepsilon^{near}(E) \Leftrightarrow \Phi_\nu^{far}(E,\theta,\Delta m^2,\delta)\cdot\sigma^{far}(E)\cdot\varepsilon^{far}(E)$$

- Interaction models are useful:
  - Relate final state particles to neutrino energy, estimate systematic errors.
  - Cannot do neutrino oscillation analysis without a model!
    - However, error cancellation only works if the model matches Nature!
- Ulrich Mosel's observation, NuInt11:
  - Theorist's paradigm: "A good generator MC does not have to fit the data, provided its model is correct"
  - Experimentalist's paradigm: "A good generator MC does not have to be correct, provided it fits the data"

Can we do it differently?

Insanity is repeating the same mistakes and expecting different results.

- What do we really want to do with a cross section measurement?
  - Let's provide enough info for later analysts to cleanly use data with a new model.

➡ We are creating crucibles for proving models with precise data.

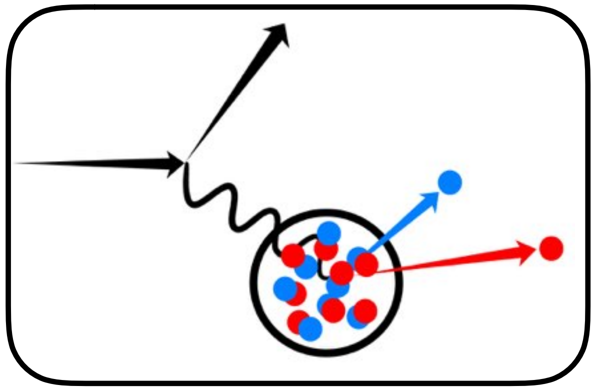
" I would do some things differently if I had a chance to do it all over again. '

What do we (experimenters) do?

(A recap of the earlier section of this talk…)

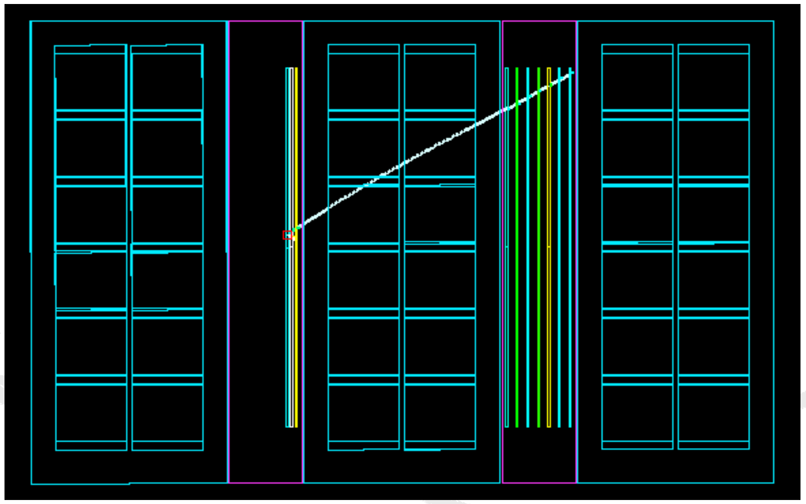
- To get better models, experimenters need theorists to use our data effectively
  - **It’s in our best interest to make that as easy and effective as possible**

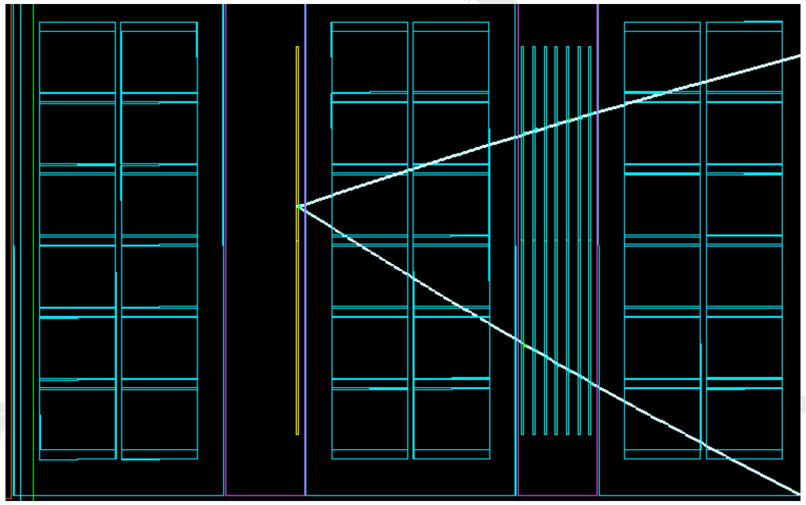
- Typically, our goal is to produce cross section measurements
  - We use the detector MC to model the efficiency and smearing,
  - We then correct those effects with unfolding matrices and efficiency functions

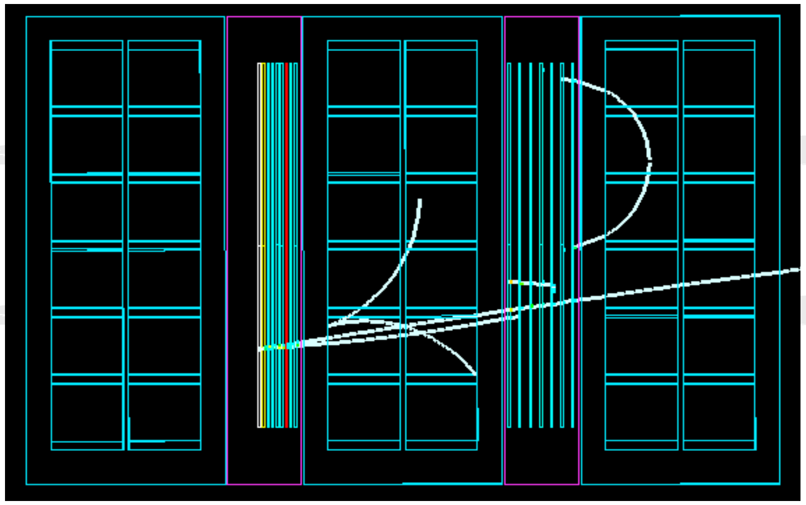


NEUTRINOS
@DETECTOR

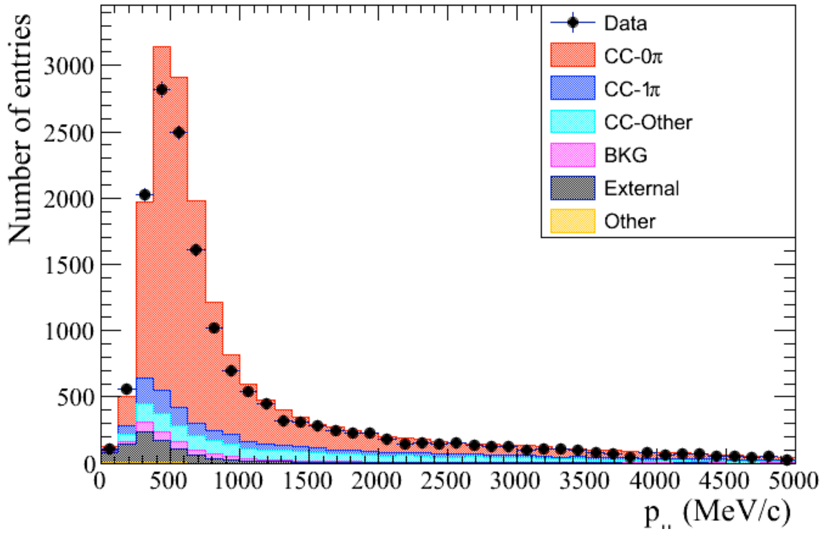
INTERACTIONS

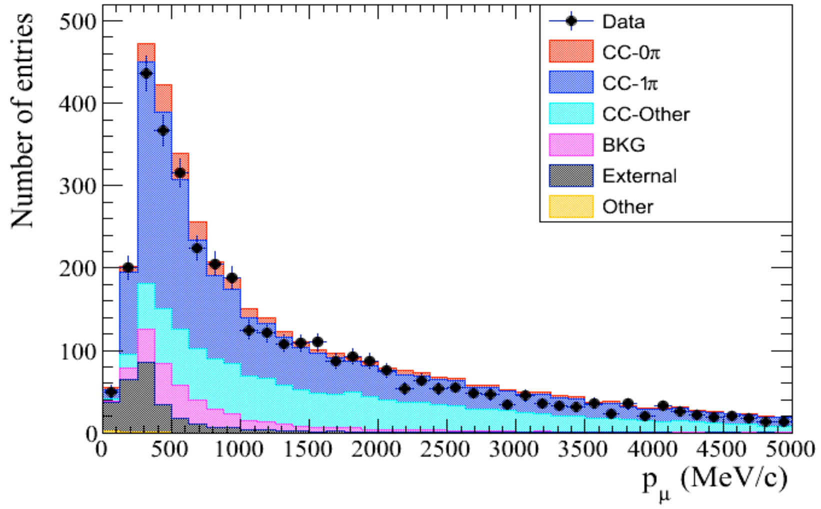


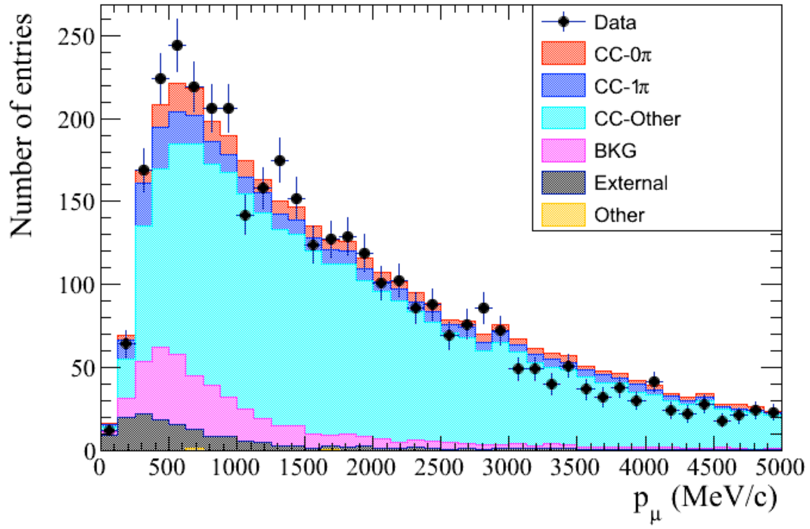


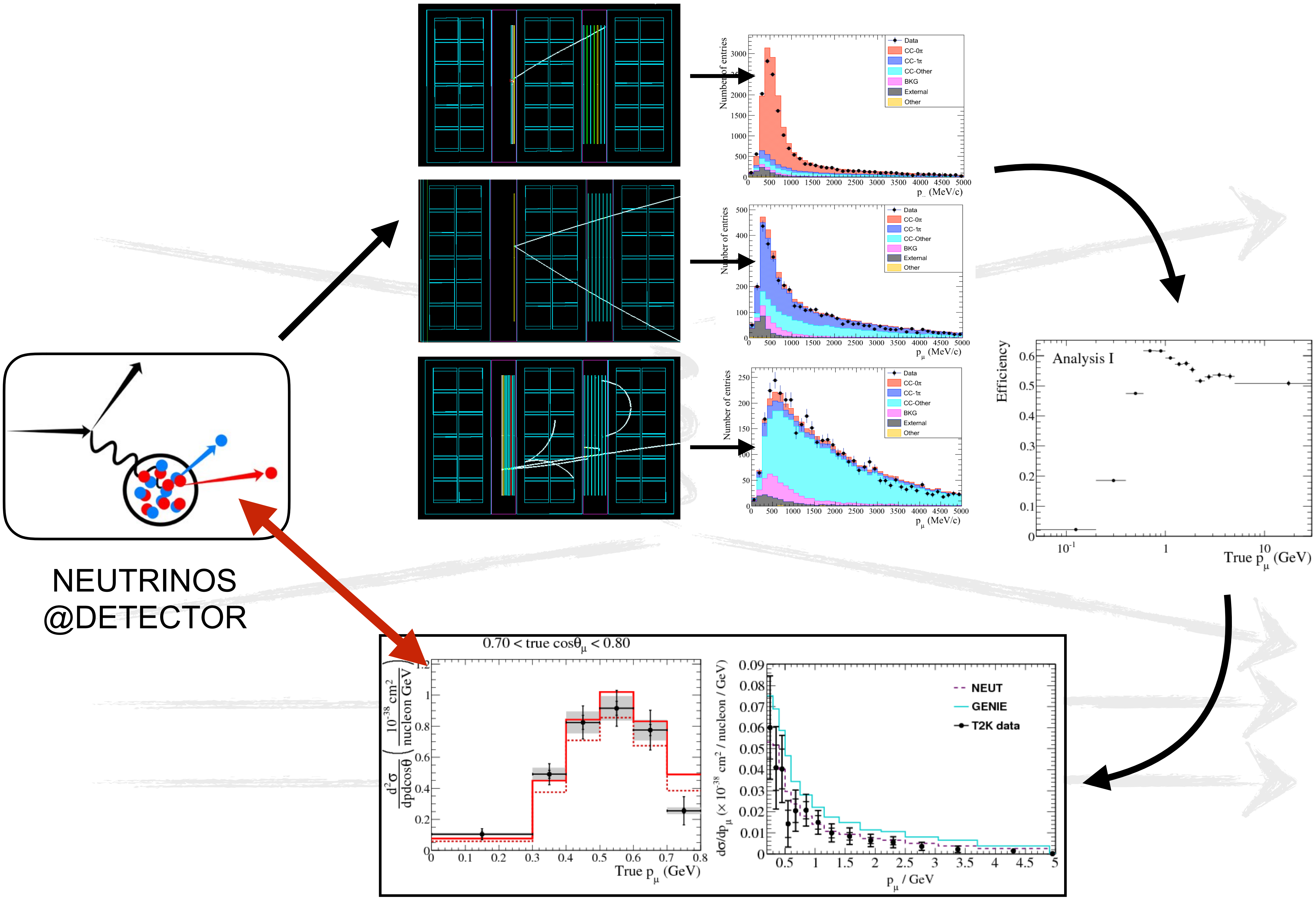


DATA SAMPLES








Number of entries
Data
CC-0π
CC-1π
CC-Other
BKG
External
Other
p_μ (MeV/c)
Efficiency
Analysis I
True p_μ (GeV)
0.70 < true cosθ$_\mu$ < 0.80
$\frac{d^2\sigma}{dp d\cos\theta}\left(\frac{10^{-38}\,\mathrm{cm}^2}{\mathrm{nucleon\ GeV}}\right)$
True p_μ (GeV)
$d\sigma/dp_\mu$ ($\times 10^{-38}$ cm^2 / nucleon / GeV)
NEUT
GENIE
T2K data
p_μ / GeV

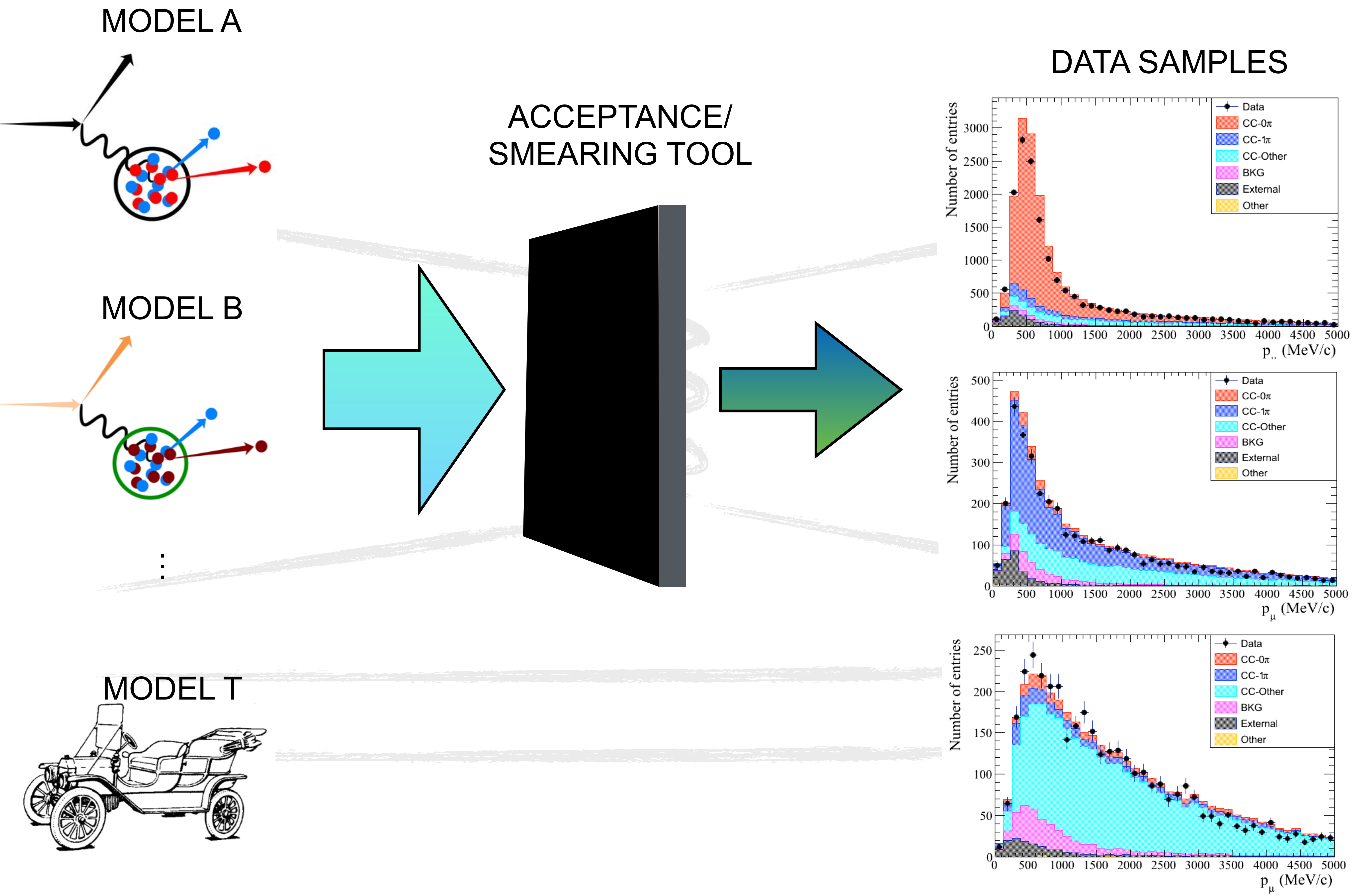
NEUTRINOS
@DETECTOR

Best way to present data?

- We could alternatively provide theorists with the tools to analyze our data the way that we do
    - We don’t use tools like NUISANCE for internal data fits!
    - Usually: numbers of events in bins of p_μ, θ_μ
- It would not be productive to just dump detector MC code in a theorist’s home directory!
    - But we could provide efficiency functions (including smearing) with systematics and our measured data
    - The efficiency function could be applied to inclusive simulated data samples, allowing theorists to perform analysis the way we do
    - Obviously need to provide neutrino fluxes, too, **but we already do that.**

(maybe not so) Crazy idea 1

- Internal data analysts use *uncorrected* data to perform parameters fits.
  - We use detector MC to naturally handle efficiencies and smearing,
    - by comparing smeared MC samples after cuts to data after cuts.
  - We use MC samples/reweighting techniques to adjust MC until it matches data.
- **Why not publish uncorrected data along with appropriate smearing and efficiency functions?**
  - "Should unfolded histograms be used to test hypotheses?" by Cousins, May, Sun. arXiv:1607.07038 [physics.data-an]



MODEL A

MODEL B

⋮

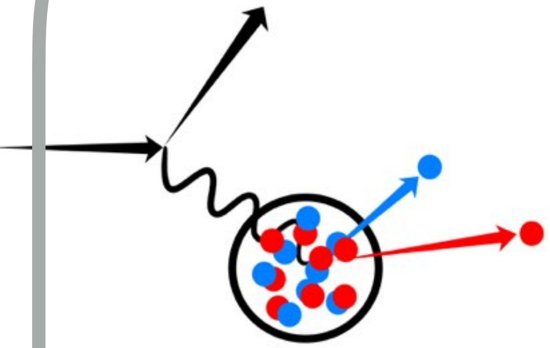
MODEL T

ACCEPTANCE/
SMEARING TOOL

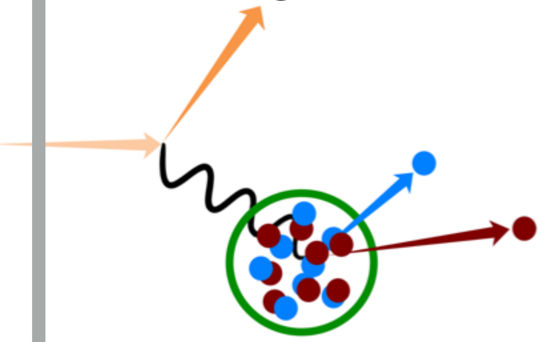
DATA SAMPLES

This is various ways of describing Nature

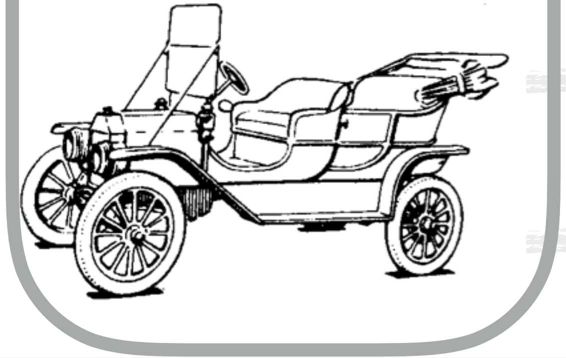
MODEL A



MODEL B



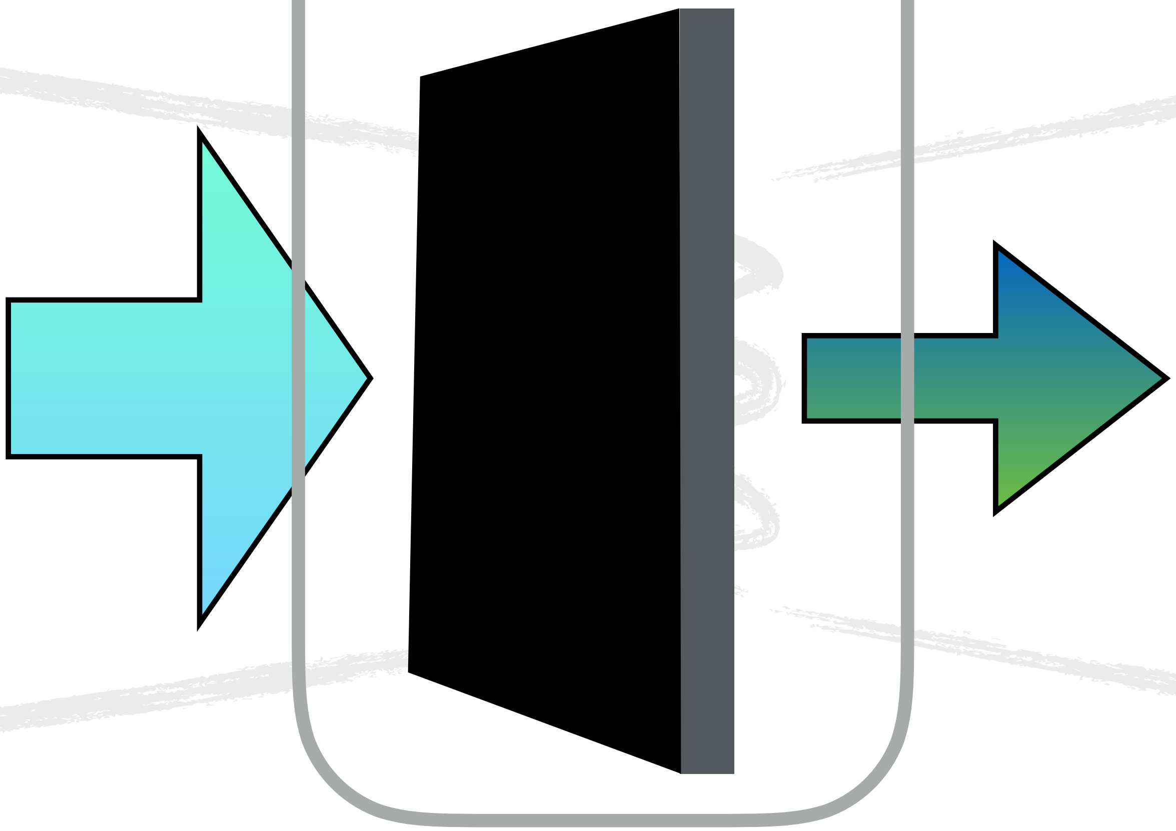
MODEL T



We'd like to choose the best one!

This is our best guess at describing our apparatus

ACCEPTANCE/ SMEARING TOOL

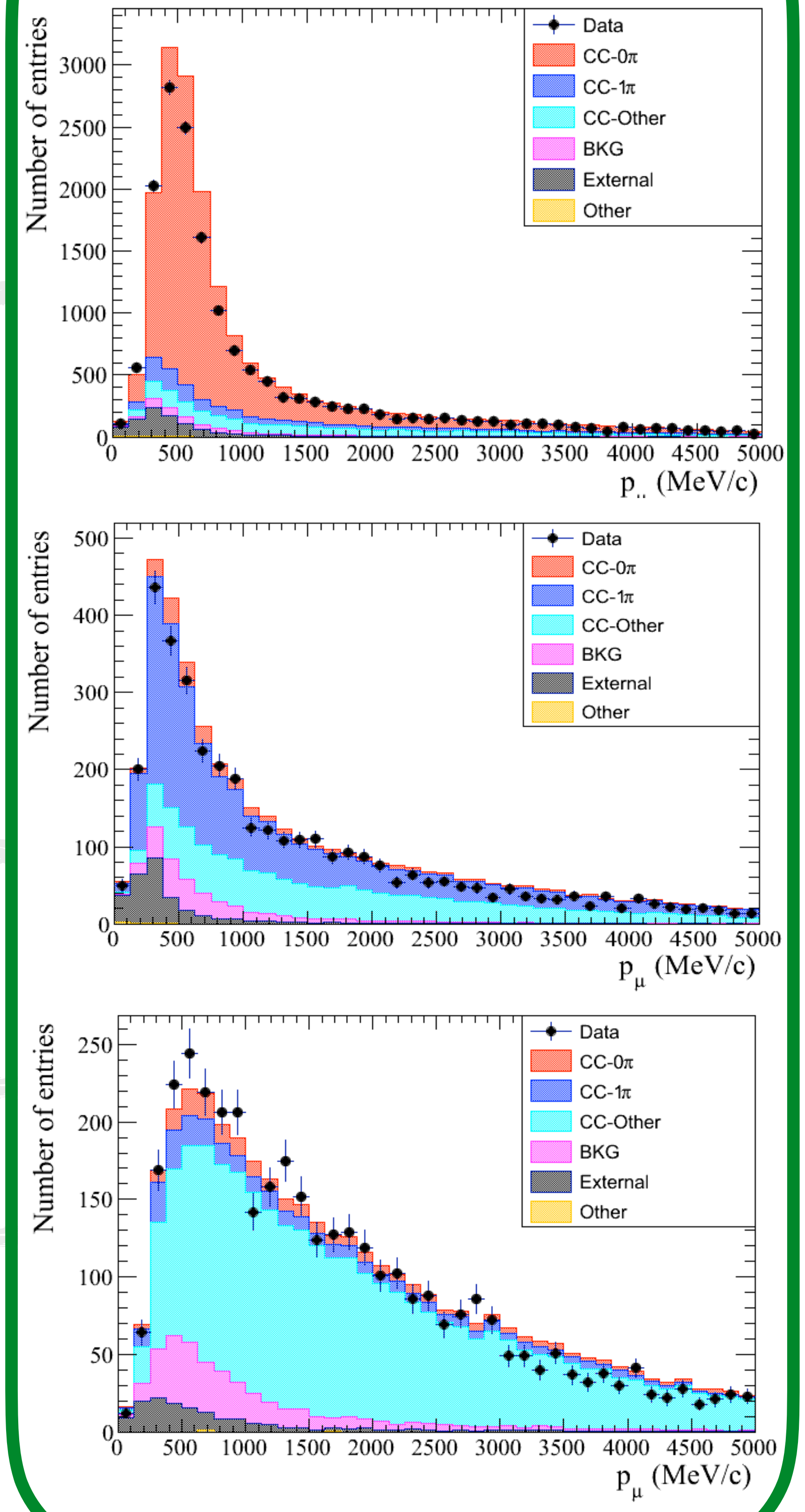


Only we, the experimenters, can provide this middle step,

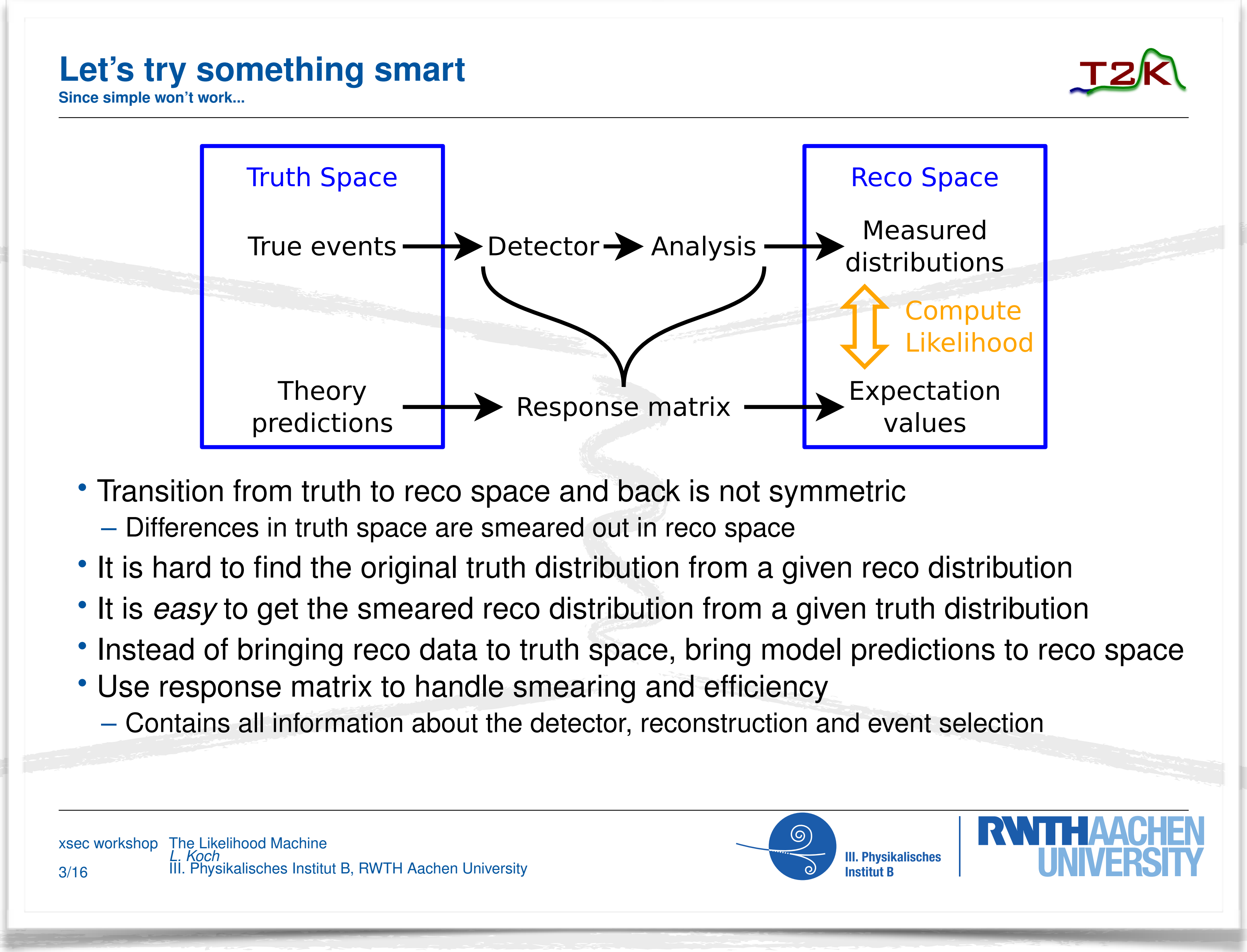
which is a crucial part of understanding the data—our true contribution to the world.

This is something unique that actually occurred!

DATA SAMPLES



Didn't we just hear this?



Let’s try something smart

Since simple won’t work...

Truth Space

True events → Detector → Analysis → Measured distributions

Theory predictions → Response matrix → Expectation values

Reco Space

Compute Likelihood

- Transition from truth to reco space and back is not symmetric
  - Differences in truth space are smeared out in reco space
- It is hard to find the original truth distribution from a given reco distribution
- It is *easy* to get the smeared reco distribution from a given truth distribution
- Instead of bringing reco data to truth space, bring model predictions to reco space
- Use response matrix to handle smearing and efficiency
  - Contains all information about the detector, reconstruction and event selection

xsec workshop The Likelihood Machine
L. Koch
III. Physikalisches Institut B, RWTH Aachen University

3/16

III. Physikalisches Institut B

RWTH AACHEN UNIVERSITY

- If this sounds familiar, it’s because Lukas is already doing it.
- You can too!
  - Be a pioneer like Lukas!

Didn't we just hear this?

- If g
  it

- You can too!
  - Be a pioneer like Lukas!

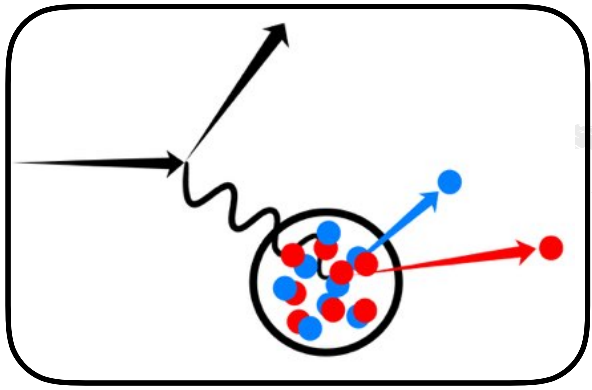
(maybe not so) Crazy idea 2

- Apply efficiency/smearing corrections to individual events.
- **Publish an ntuple of events: reconstructed final state particles.**
  - Each event comes with a cross-section weight derived with POT numbers, flux & detector MC.
  - Allows one to make a plot giving cross sections instead of number of events.

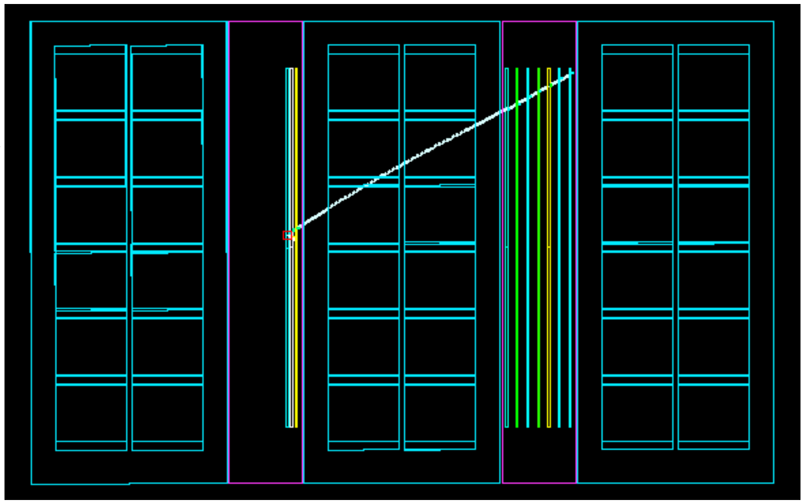
➡ Gives unprecedented knowledge to future analysers since it would allow analysis of <u>new variables</u>

INTERACTIONS

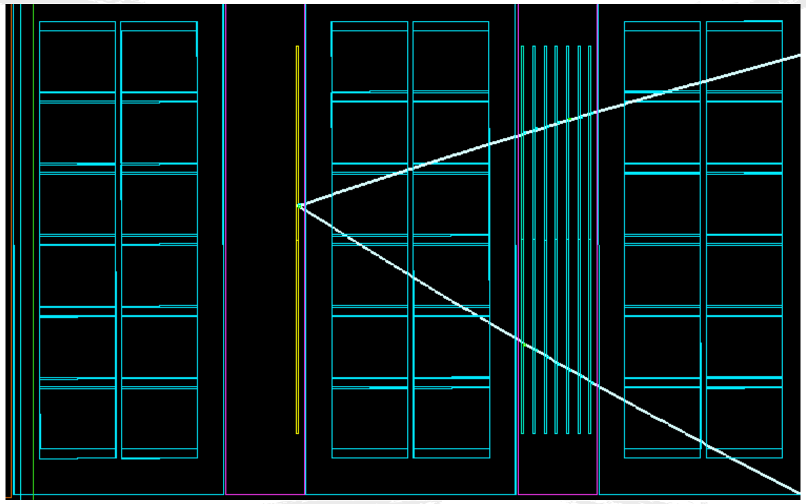
DATA



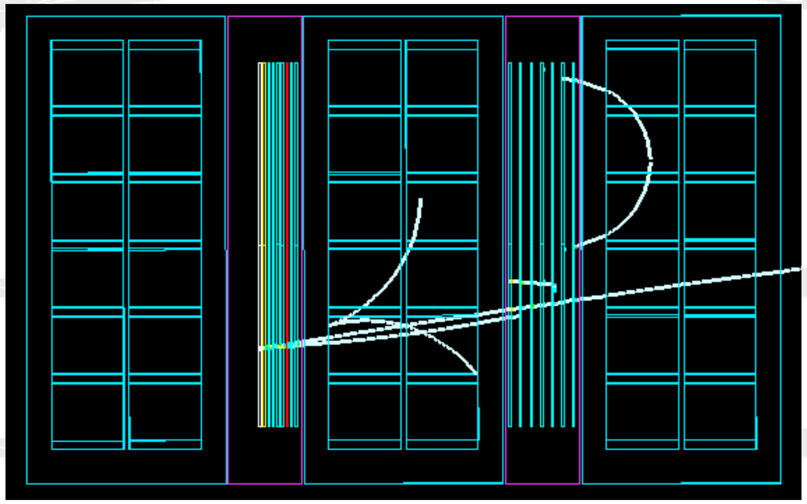
NEUTRINOS
@DETECTOR



Vertex
(x,y,z)
mu–
(px,py,pz) MeV/c



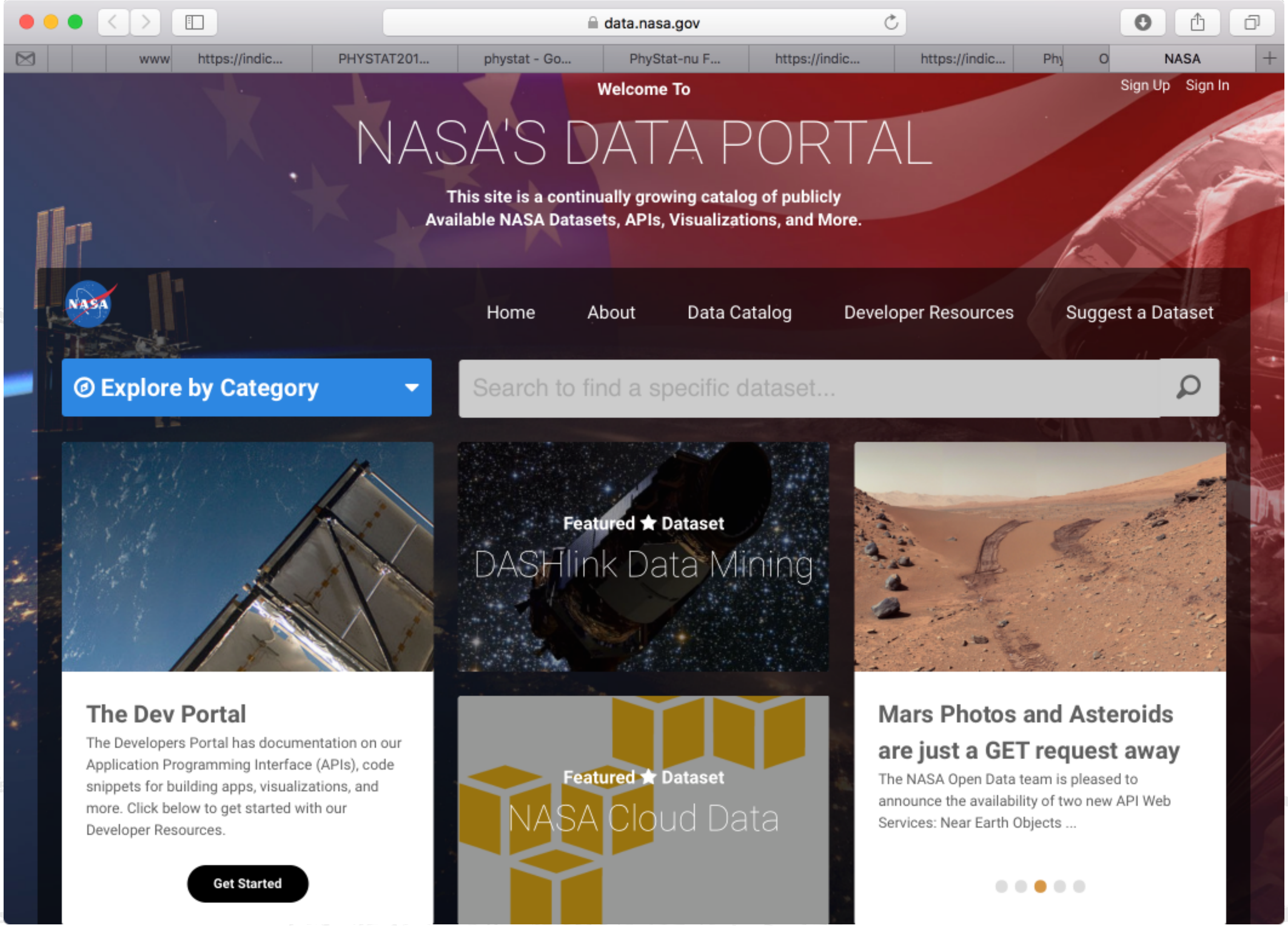
Vertex
(x,y,z)
mu–
(px,py,pz) MeV/c
pi+
(px,py,pz) MeV/c



Vertex
(x,y,z)
mu–
(px,py,pz) MeV/c
p+
(px,py,pz) MeV/c
pi0
(px,py,pz) MeV/c
⋮

⋮

Could include experts-only info too, like PID pulls



data.nasa.gov

www | https://indic... | PHYSTAT201... | phystat - Go... | PhyStat-nu F... | https://indic... | https://indic... | Phy | O | NASA

Sign Up Sign In

Welcome To

NASA'S DATA PORTAL

This site is a continually growing catalog of publicly Available NASA Datasets, APIs, Visualizations, and More.

Home About Data Catalog Developer Resources Suggest a Dataset

Explore by Category

Search to find a specific dataset...

The Dev Portal

The Developers Portal has documentation on our Application Programming Interface (APIs), code snippets for building apps, visualizations, and more. Click below to get started with our Developer Resources.

Get Started

Featured ★ Dataset

DASHlink Data Mining

Featured ★ Dataset

NASA Cloud Data

Mars Photos and Asteroids are just a GET request away

The NASA Open Data team is pleased to announce the availability of two new API Web Services: Near Earth Objects ...

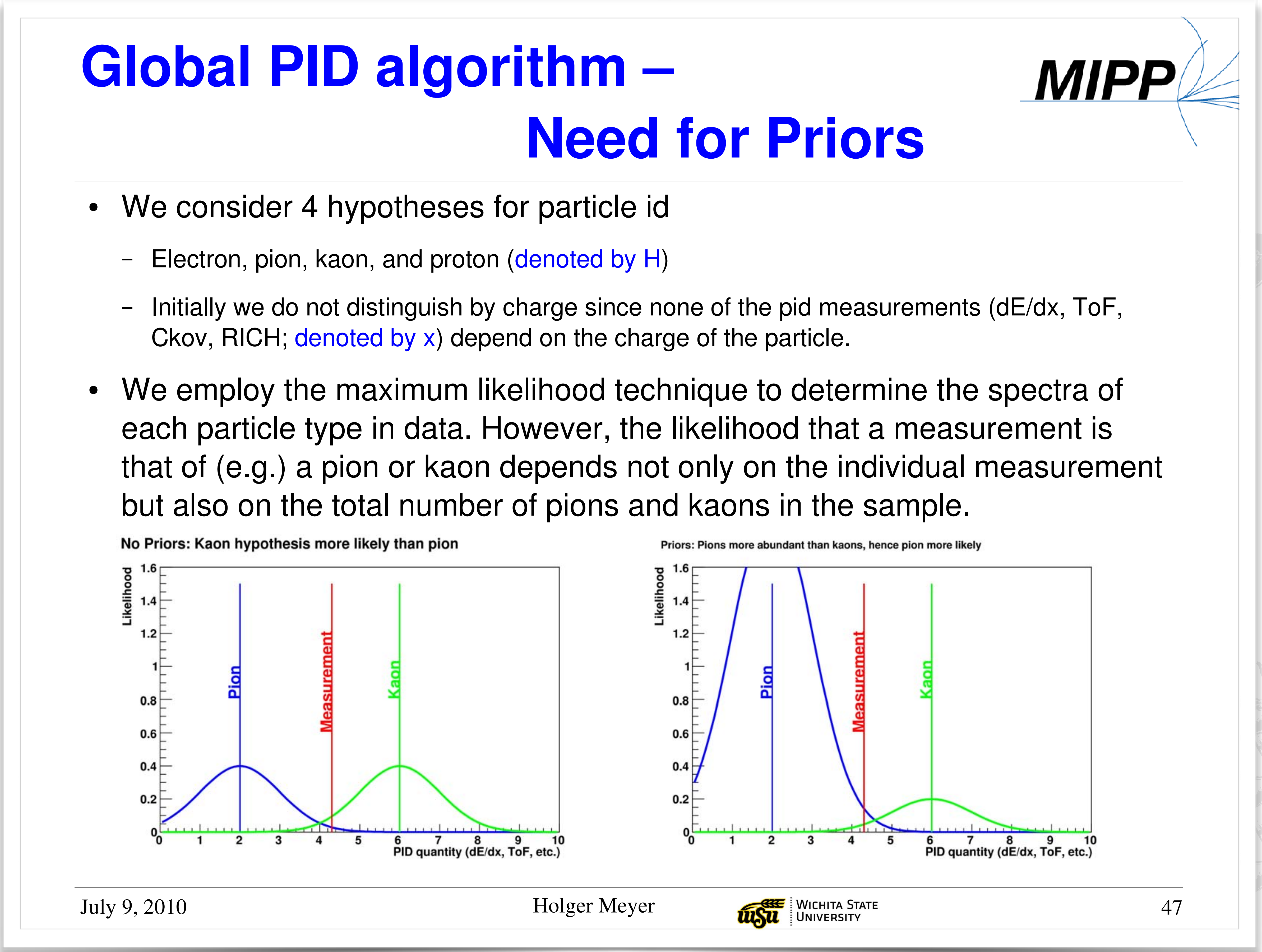
NASA releases data to the public

(fairly) Crazy idea 3

Generator-free analysis

- Can we do an xsec analysis without using a generator at all?
- In principle, can generate events flat in all phase space
  - “particle blizzard”* for efficiencies and purities
- Still need to turn flat phase space into a model for PID, systematic studies, etc.
- This job is usually done by the generator, but can a completely data-driven method be developed?

* This needs a good name. “Particle bomb” is not a good choice.



Global PID algorithm – Need for Priors

MIPP

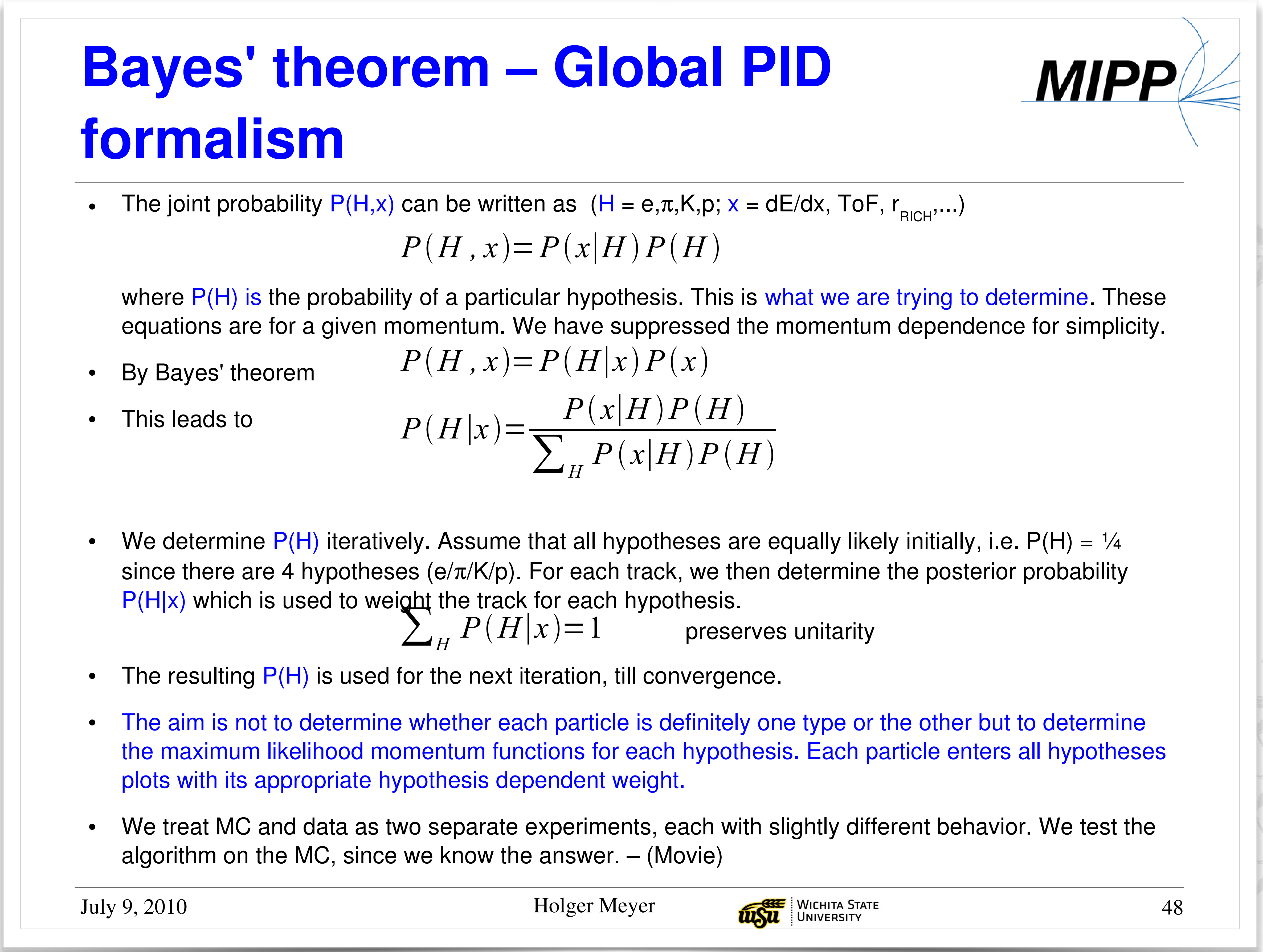
- We consider 4 hypotheses for particle id
  - Electron, pion, kaon, and proton (denoted by H)
  - Initially we do not distinguish by charge since none of the pid measurements (dE/dx, ToF, Ckov, RICH; denoted by x) depend on the charge of the particle.
- We employ the maximum likelihood technique to determine the spectra of each particle type in data. However, the likelihood that a measurement is that of (e.g.) a pion or kaon depends not only on the individual measurement but also on the total number of pions and kaons in the sample.

No Priors: Kaon hypothesis more likely than pion

Priors: Pions more abundant than kaons, hence pion more likely

July 9, 2010 Holger Meyer Wichita State University 47

http://theory.fnal.gov/jetp/talks/JETP9Jul2010.pdf



Bayes' theorem – Global PID formalism

MIPP

- The joint probability P(H,x) can be written as (H = e,π,K,p; x = dE/dx, ToF, r_{RICH},...)

$$P(H,x)=P(x|H)P(H)$$

where P(H) is the probability of a particular hypothesis. This is what we are trying to determine. These equations are for a given momentum. We have suppressed the momentum dependence for simplicity.

- By Bayes' theorem $$P(H,x)=P(H|x)P(x)$$
- This leads to $$P(H|x)=\frac{P(x|H)P(H)}{\sum_H P(x|H)P(H)}$$
- We determine P(H) iteratively. Assume that all hypotheses are equally likely initially, i.e. P(H) = ¼ since there are 4 hypotheses (e/π/K/p). For each track, we then determine the posterior probability P(H|x) which is used to weight the track for each hypothesis.

$$\sum_H P(H|x)=1$$ preserves unitarity

- The resulting P(H) is used for the next iteration, till convergence.
- The aim is not to determine whether each particle is definitely one type or the other but to determine the maximum likelihood momentum functions for each hypothesis. Each particle enters all hypotheses plots with its appropriate hypothesis dependent weight.
- We treat MC and data as two separate experiments, each with slightly different behavior. We test the algorithm on the MC, since we know the answer. – (Movie)

July 9, 2010 Holger Meyer WSU Wichita State University 48

http://theory.fnal.gov/jetp/talks/JETP9Jul2010.pdf

(maybe not so) Crazy idea 4

- What’s the best way to use all the information?
- High dimensionality presents challenges
- Example: to nail down 2p2h interactions, we’d like to measure p,θ of 1μ, 1μ1p, and 1μ2p events.
  - that’s up to 6 variables, with fewer events in the higher multiplicity samples
  - binning those samples will remove lots of information
    - Why publish event-by-event if you just have to bin the data later?

➡ **Let’s enjoy un-binned analysis methods!**

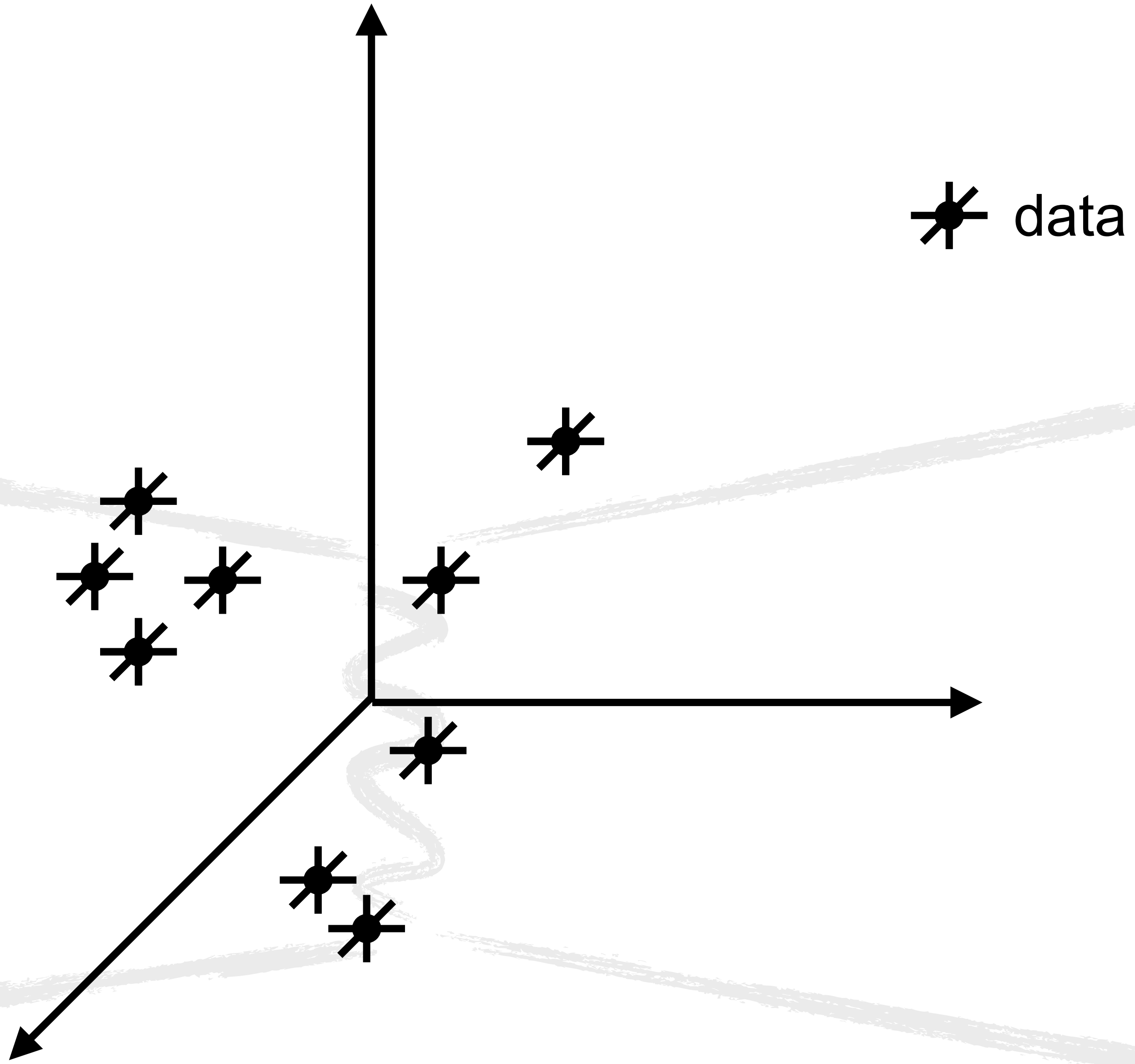
(maybe not so) Crazy idea 4

M. Williams, “How good are your fits?” arXiv:1006.3019v2 [hep-ex]

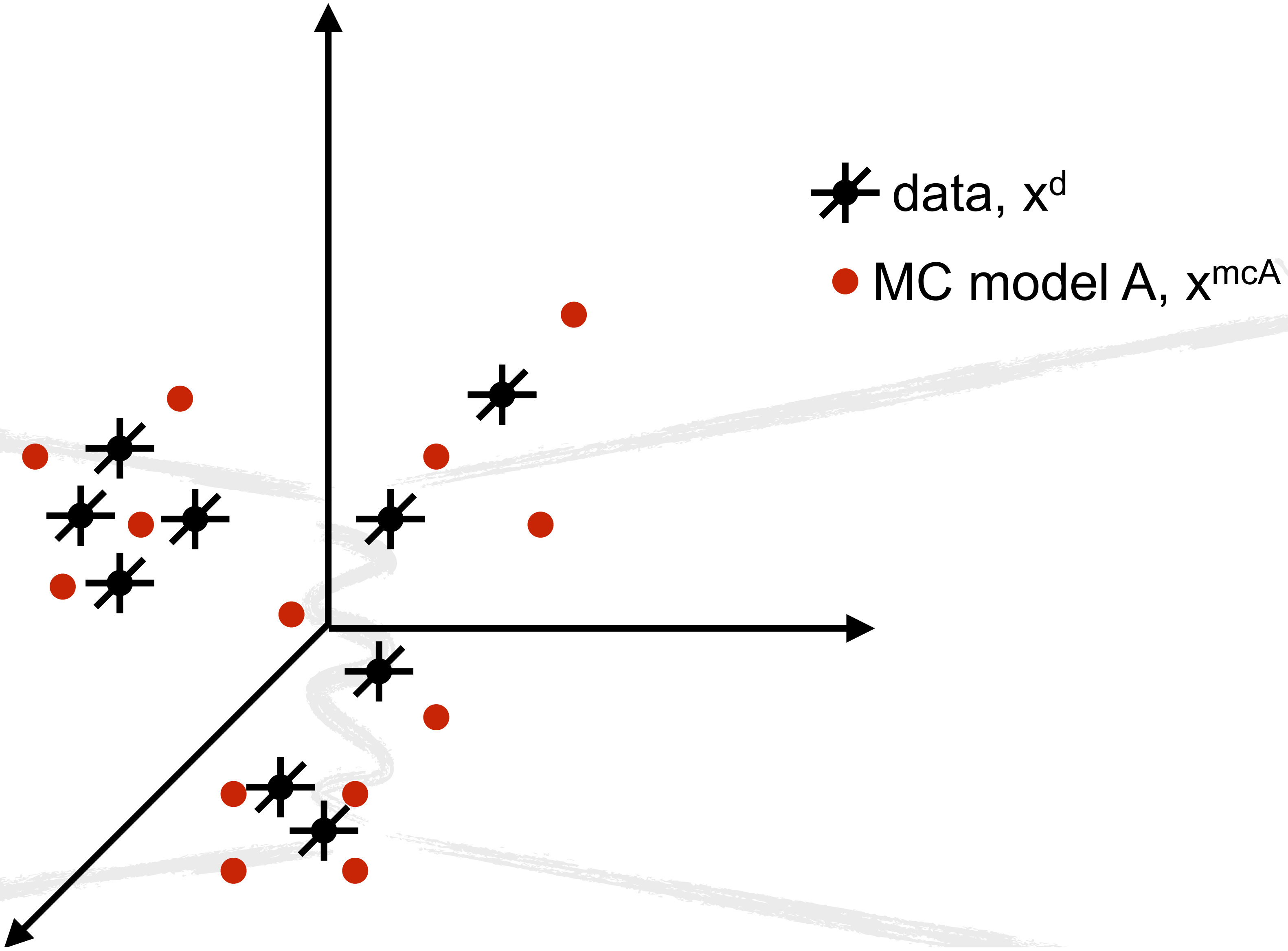
- There are many ways to calculate goodness-of-fit parameters in un-binned analyses
- An interesting class is point to point dissimilarity methods
  - based on measuring the absolute distance between the points in two sample distributions
  - similar to electrostatic energy calculation
- Let x^d be your data and x^{mc} be your MC, then an interesting GOF test statistic is:

$$T = \frac{1}{n_d^2} \sum_{i,j>i}^{n_d} \psi(|\vec{x}_i^d - \vec{x}_j^d|) - \frac{1}{n_d n_{mc}} \sum_{i,j}^{n_d, n_{mc}} \psi(|\vec{x}_i^d - \vec{x}_j^{mc}|)$$

(Can experiment with different forms of Ψ)

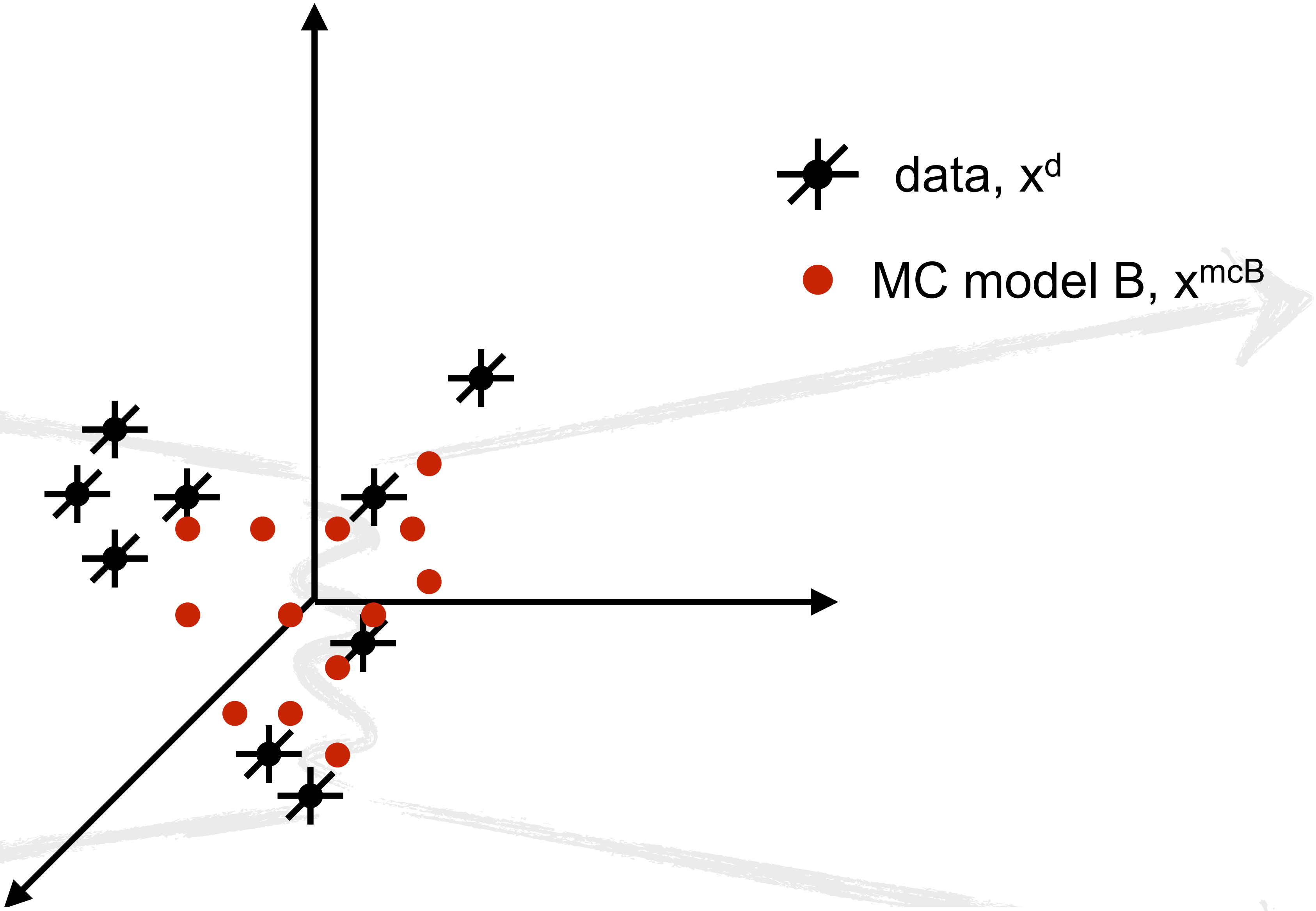


This is your multidimensional data



$$T = \frac{1}{n_d^2}\sum_{i,j>i}^{n_d}\psi(|\vec{x}_i^d - \vec{x}_j^d|) - \frac{1}{n_d n_{mc}}\sum_{i,j}^{n_d,n_{mc}}\psi(|\vec{x}_i^d - \vec{x}_j^{mc}|)$$

Calculate T for your data and MC points



$$T = \frac{1}{n_d^2}\sum_{i,j>i}^{n_d}\psi(|\vec{x}_i^d - \vec{x}_j^d|) - \frac{1}{n_d n_{mc}}\sum_{i,j}^{n_d,n_{mc}}\psi(|\vec{x}_i^d - \vec{x}_j^{mc}|)$$

This MC model is not as close to the data, resulting in larger T

Can be used with weighting techniques for MC?
Can be implemented into a regression algorithm?

What to take away…

- If you are already working on a cross-section analysis…
  - I do not want to imply what you are doing is wrong!
  - Don’t stop your work—keep going, write a paper, implement your data release in NUISANCE**!**
  - If you want to fit Q^2_{QE} data for M_A (or to measure F_A), then go for it!
    - But let’s do the model independent stuff too!
- If you are looking to start an analysis now, why not try one of these crazy ideas?

Summary

- Measuring cross sections is a tricky business
- We've been discussing these analysis issues for a while now...
- It's time for us to learn our own lessons!
  - Having used external data sets to constrain cross section models, we've learned a lot about what not to do
- The world has lots of good data analysis ideas
  - Let's try some!

> "Human progress has always been driven by a sense of adventure and unconventional thinking."
> –Andre Geim, 2010 Nobel Prize for Physics

Thank you for your attention!

ご清聴いただきまして、ありがとうございました

水戸の梅の花

Best way to publish data?

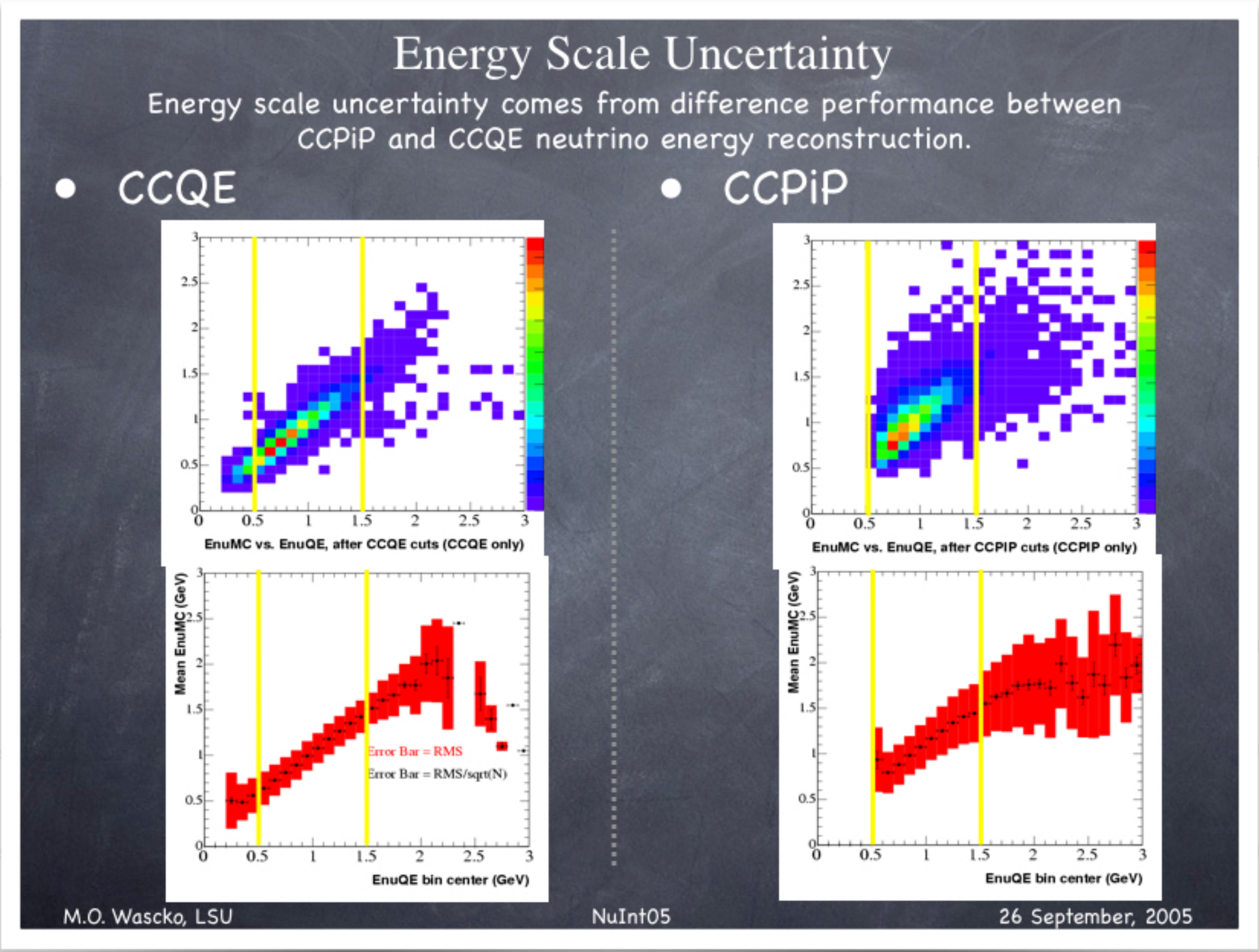
- *Measured quantities vs inferred quantities:*
  - *How to interpret unfolded data? Generator dependencies?*

- Can we provide the tools to allow theorists to fit our data in the same way we do?
- Other fields publish data in different (and creative) ways—maybe we should consider some of these.

Show the whole story

- All neutrino data samples are rife with backgrounds.
  - MiniBooNE: CCQE $\leftrightarrow$ CC1π^+ $\leftrightarrow$ CC1π^0 $\rightarrow ?$ NC1π^0
  - but really: μ,μ+p $\leftrightarrow$ $\mu+\pi^+$ $\leftrightarrow$ $\mu+\pi^0$ $\rightarrow ?$ π^0
- Knowing what we do (e.g. about 2p2h from e-A experiments), we cannot have confidence in one sample without seeing all the others.
  - We’ve already learned that seeing each of them isn’t enough!
- For example, to extract M_A from neutrino data:
  - Requires nuclear model & background predictions match Nature.
  - Predicated on assumption that F_A is a dipole.

Full disclosure: I’ve unfolded before, *but why repeat my mistakes?*



Energy Scale Uncertainty

Energy scale uncertainty comes from difference performance between CCPiP and CCQE neutrino energy reconstruction.

- CCQE
- CCPiP

M.O. Wascko, LSU

NuInt05

26 September, 2005

Full full disclosure: I inverted the matrix—no feaux-Bayesian mumbo jumbo