# Banff Challenge 3: Systematic Uncertainties

Tom Junk

*Fermilab*

BIRS Workshop 23w5096 - Systematic Effects and Nuisance Parameters in Particle Physics Data Analyses
April 25, 2023

# Systematic Error Nomenclature (For this talk)

The **"Good":** Nuisance parameter values are constrained by measurements.

- May be *in situ* measurements, or possibly external – other experiments, possibly decades old.
- These are essentially statistical uncertainties that get classified as systematic uncertainties
- In the case of external measurements that cannot be repeated, these errors may not dissipate with more data.

The **"Bad":** Nuisance parameter values are theory predictions, or (educated) guesses.

- Priors on these nuisance parameters are also (educated) guesses
- Experimenters often rely on detailed domain knowledge to make these guesses.
- If big, can be showstoppers for experiments (e.g. P5 and PINGU).
- "Good" uncertainties can have "Bad" components (such as extrapolation factors; more on this later)

The **"Ugly":** Not thought of, incorrectly dismissed, or otherwise unknown sources of error.

These are sadly not uncertainties!

- Uglier than Pekka Sinervo's Type 3.
  https://inspirehep.net/literature/637578
- Famous examples: OPERA's loose cable causing a measurement of the speed of neutrinos to exceed *c*.
  False discovery of the top quark by UA1 (40 GeV top quark). Problem was inadequate modeling of W+jets
  17 keV neutrino false discovery
  More examples in Sheldon Stone's "Pathological Science", hep-ph/0010295

# A Little History

**Banff Challenge 1, Upper Limits.** BIRS meeting in July 2006. Joel Heinrich constructed the challenge.

N independent measurements ("bins", "channels") per repetition.
Main measurement data, and subsidiary background and acceptance measurements:

$$\begin{aligned} n_i &\sim \mathrm{Pois}(\epsilon_i s + b_i) && \text{(main measurement)} \\ y_i &\sim \mathrm{Pois}(t_i b_i) && \text{(subsidiary background measurement)} \\ z_i &\sim \mathrm{Pois}(u_i \epsilon_i) && \text{(subsidiary acceptance measurement)} \end{aligned}$$

Joel summarized the results at Phystat-LHC, 2007

http://cds.cern.ch/record/1021125

Specifically, Joel's article:

http://cds.cern.ch/record/1099980?ln=en

# Banff Challenge 1

$$\begin{aligned} n_i &\sim \mathrm{Pois}(\epsilon_i s + b_i) && \text{(main measurement)} \\ y_i &\sim \mathrm{Pois}(t_i b_i) && \text{(subsidiary background measurement)} \\ z_i &\sim \mathrm{Pois}(u_i \epsilon_i) && \text{(subsidiary acceptance measurement)} \end{aligned}$$

Joel provided (n, y, z, t, u) for each repetition of the measurement. One of these for the one-bin case, and a ten-tuple per repetition for the ten-bin case.

*t* and *u* were varied but were not uncertain on a repetition.

$\epsilon_i$ and $b_i$ are nuisance parameters. "Nuisance parameter uncertainties are about 30%" (priors? Usually the subsidiary is enough).

Joel computed coverage and credibility for the intervals of *s*, the parameter of interest (the "signal rate")

# Banff Challenge 1

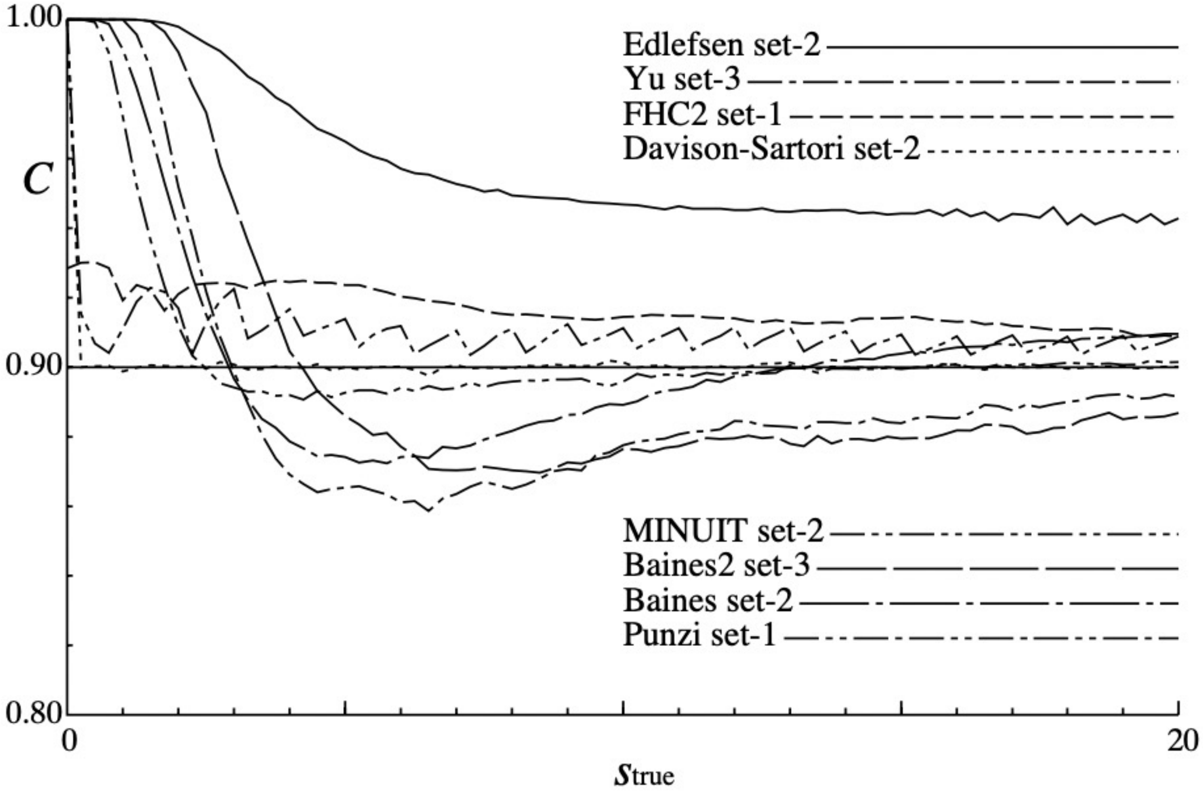
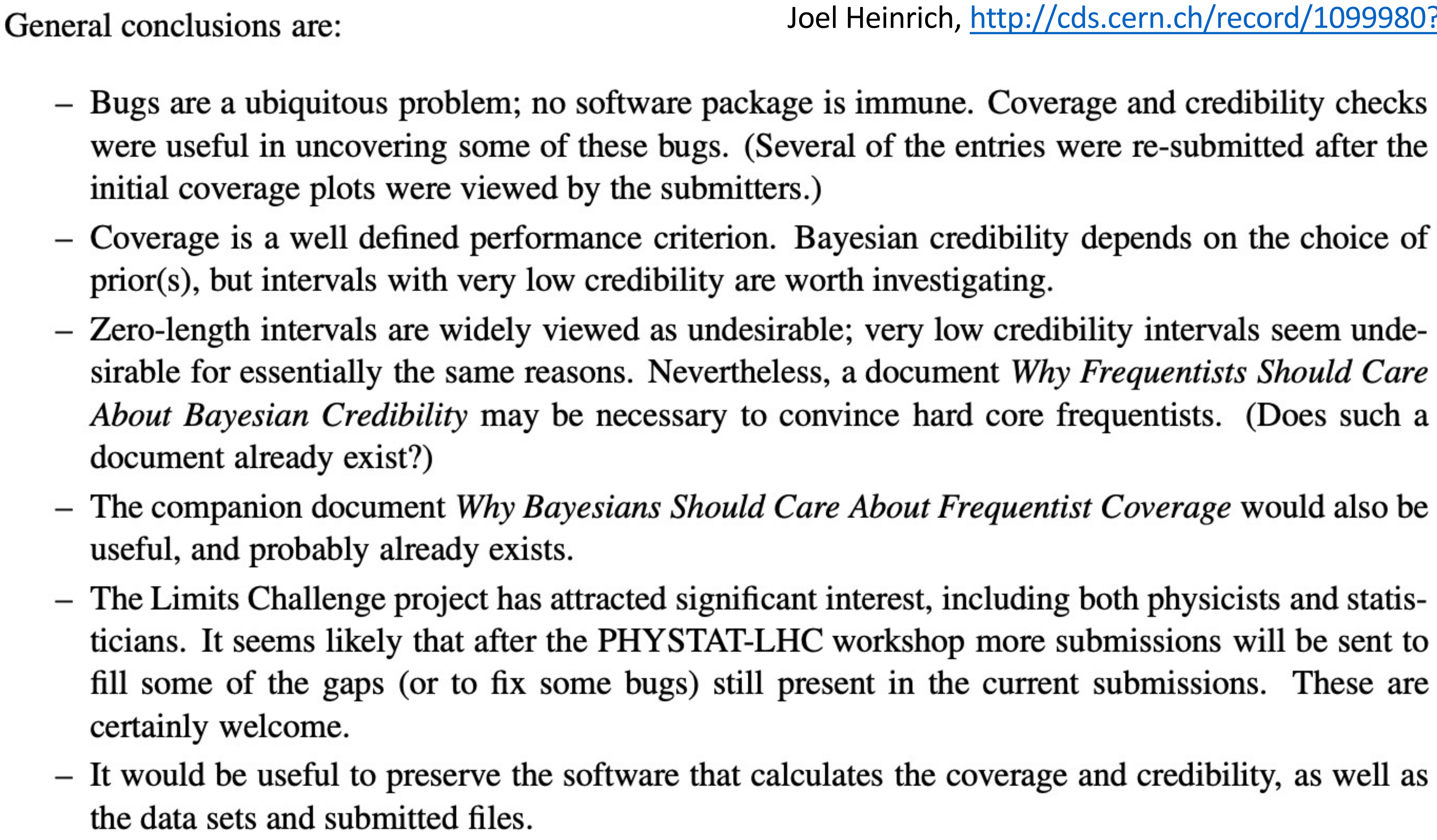


**Fig. 1:** Coverage of selected 90% intervals as a function of the true value of $s$.

Joel Heinrich, http://cds.cern.ch/record/1099980?ln=en

Joel Heinrich, http://cds.cern.ch/record/1099980?ln=en



General conclusions are:

- Bugs are a ubiquitous problem; no software package is immune. Coverage and credibility checks were useful in uncovering some of these bugs. (Several of the entries were re-submitted after the initial coverage plots were viewed by the submitters.)
- Coverage is a well defined performance criterion. Bayesian credibility depends on the choice of prior(s), but intervals with very low credibility are worth investigating.
- Zero-length intervals are widely viewed as undesirable; very low credibility intervals seem undesirable for essentially the same reasons. Nevertheless, a document *Why Frequentists Should Care About Bayesian Credibility* may be necessary to convince hard core frequentists. (Does such a document already exist?)
- The companion document *Why Bayesians Should Care About Frequentist Coverage* would also be useful, and probably already exists.
- The Limits Challenge project has attracted significant interest, including both physicists and statisticians. It seems likely that after the PHYSTAT-LHC workshop more submissions will be sent to fill some of the gaps (or to fix some bugs) still present in the current submissions. These are certainly welcome.
- It would be useful to preserve the software that calculates the coverage and credibility, as well as the data sets and submitted files.

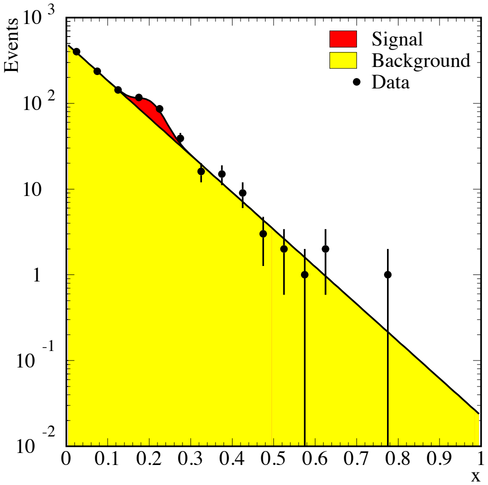
# A Little History – Banff Challenge 2

Two stabs at it – the one that was completed was called Banff Challenge 2a for a while.
https://www.birs.ca/events/2010/5-day-workshops/10w5068
*"Statistical issues relevant to significance of discovery claims"*

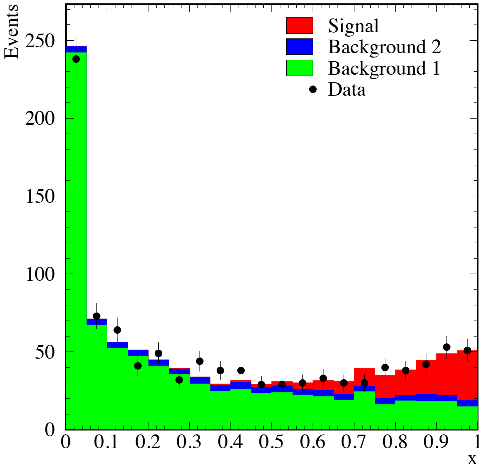
Two problems:

1) Classic "bump on a smooth background" problem
2) A more arbitrary distribution, meant to mimic the scores from a neural net classifier

Problem 1



Problem 2



Search and discovery analyses often made final interpretations based on distributions like these.

# Banff Challenge 2

## Deliverables:

1) Power of test: Correct discovery probability estimate assuming a Type-1 error rate of 1%. Probs 1 & 2
2) Writeup
3) For each experimental repetition,
    a) Yes-no discovery claim. Desired Type-I error rate is 1%   Probs 1 & 2. Prob. 1 claims test LEE.
    b) Null-hypothesis test *p* value, Bayes factor, or something equivalent. Probs 1 & 2
    c) Location parameter estimate and 68% CL interval for Prob 1.
    d) Extra credit: Signal strength parameter point estimate and 68% CL interval. Probs 1 & 2

Proceedings of Phystat 2011:

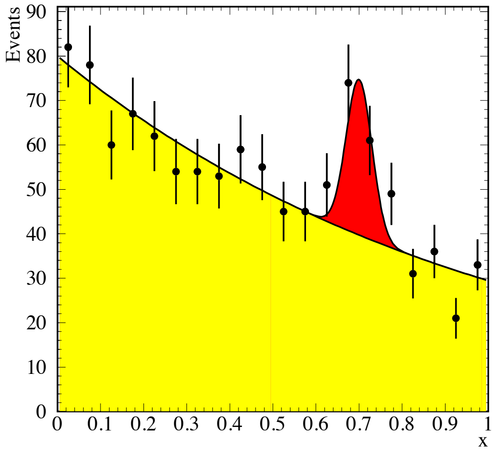
http://cds.cern.ch/record/1306523

Specifically, Tom's contribution:

http://cds.cern.ch/record/2203235?ln=en

# Banff Challenge 2 Problem 1

Participants were told the background was exponential and the signal was Gaussian. The width of the Gaussian signal was fixed and told to participants. All other parameters varied from repetition to repetition.



$$B(x) + S(x) = Ae^{-Cx} + De^{-(x-E)^2/2\sigma^2}$$

$\sigma$ was provided, and $D$ and $E$ were parameters of interest

Parameters of the exponential background and the Gaussian signal were varied from repetition to repetition.

20000 unbinned datasets provided.

Most data sets had zero true signal. These were needed to calculate the Type-I error rate.

Most signals were "just barely discernable"

V. Niess

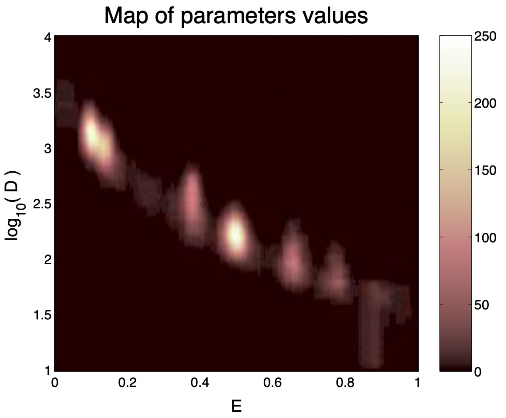
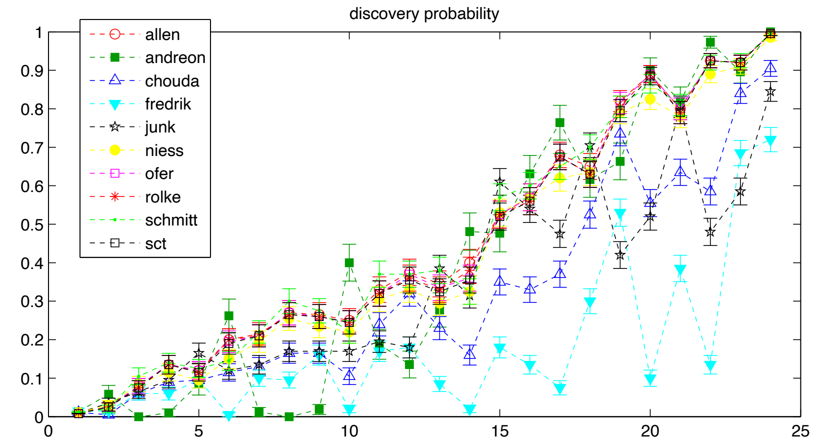
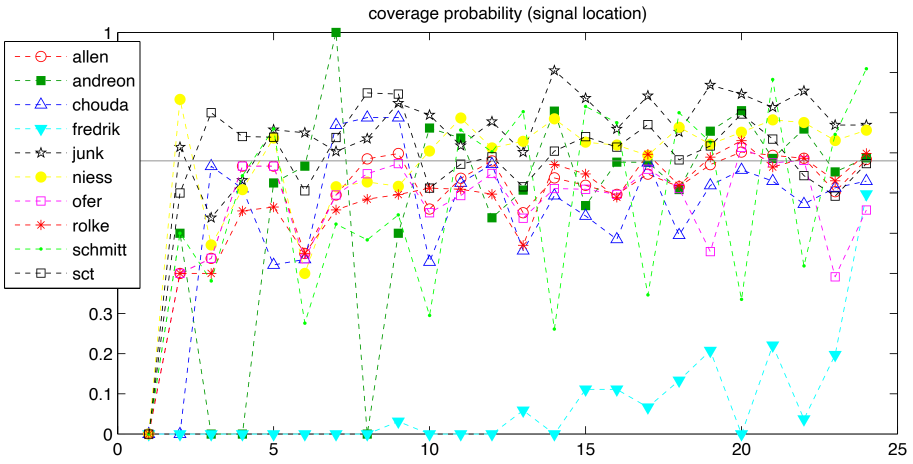
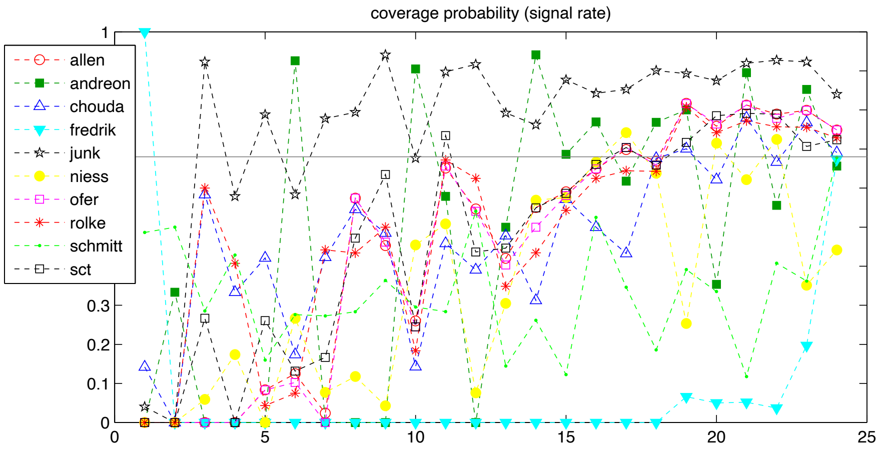


FIG. 1. Map of signal parameter values. The intensity on the 2D map reads as the number of signal candidates for which the individual confidence belts include the point of coordinates $(E, D)$

# BC2 Summary plots from Ofer Vitells

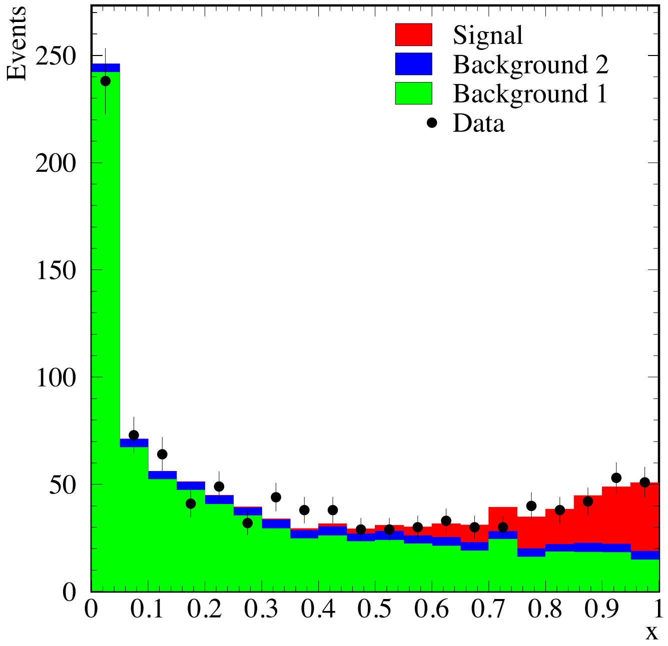

discovery probability


coverage probability (signal location)


coverage probability (signal rate)

# Problem 2 Had More Undercoverage in the Results

Data were unbinned, and the true distribution functions were hidden.  Instead, finite-size "Monte Carlo" samples were given, one for the signal, and one for each of two background components.



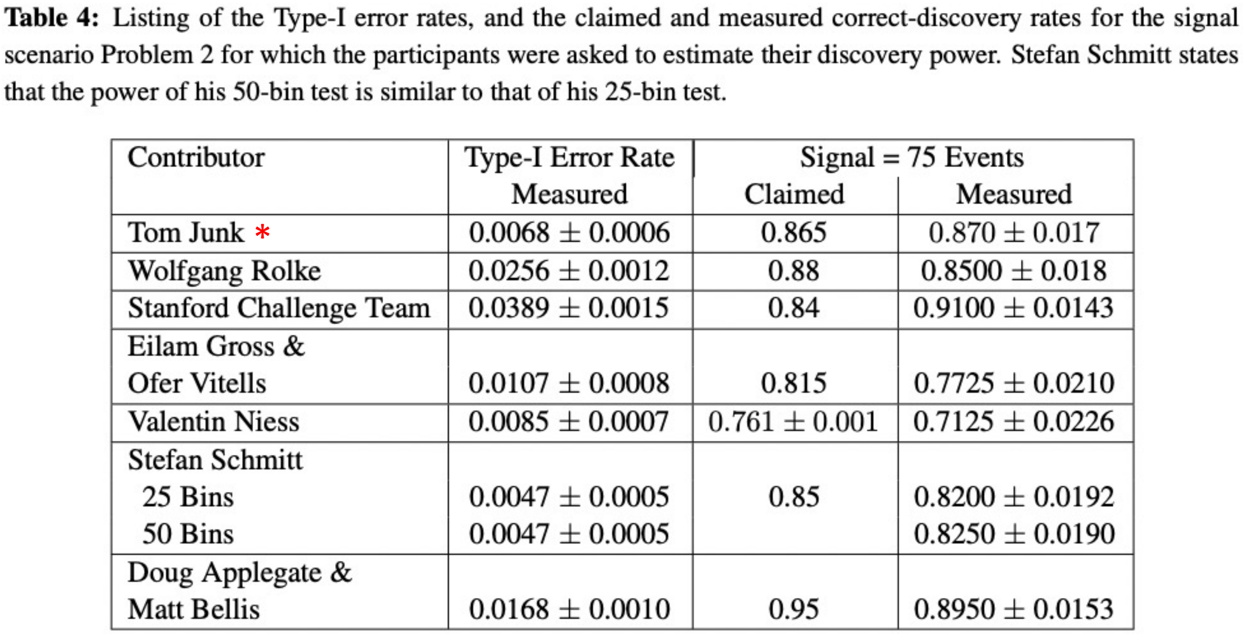


**Table 4:** Listing of the Type-I error rates, and the claimed and measured correct-discovery rates for the signal scenario Problem 2 for which the participants were asked to estimate their discovery power. Stefan Schmitt states that the power of his 50-bin test is similar to that of his 25-bin test.

| Contributor | Type-I Error Rate | Signal = 75 Events | |
|---|---|---|---|
| | Measured | Claimed | Measured |
| Tom Junk * | 0.0068 ± 0.0006 | 0.865 | 0.870 ± 0.017 |
| Wolfgang Rolke | 0.0256 ± 0.0012 | 0.88 | 0.8500 ± 0.018 |
| Stanford Challenge Team | 0.0389 ± 0.0015 | 0.84 | 0.9100 ± 0.0143 |
| Eilam Gross & Ofer Vitells | 0.0107 ± 0.0008 | 0.815 | 0.7725 ± 0.0210 |
| Valentin Niess | 0.0085 ± 0.0007 | 0.761 ± 0.001 | 0.7125 ± 0.0226 |
| Stefan Schmitt | | | |
| 25 Bins | 0.0047 ± 0.0005 | 0.85 | 0.8200 ± 0.0192 |
| 50 Bins | 0.0047 ± 0.0005 | | 0.8250 ± 0.0190 |
| Doug Applegate & Matt Bellis | 0.0168 ± 0.0010 | 0.95 | 0.8950 ± 0.0153 |

*My entry doesn't count.  All participants' Type-I error rates were supposed to be < 1%

# Banff Challenge 2 Criteria for "Winning"

Measured Type-I error rate could not exceed 1%.

Measured true discovery rate must be at least the claimed discovery rate.

Highest claimed discovery rate is the winner.

These criteria were not announced at the time the challenge was issued, so we were very generous in declaring winners.

# Banff Challenge 3: Systematic Uncertainties

BC1 and BC2 already explored analyses with nuisance parameters.

But they were all **Good**. Statistically constrained *in situ*.

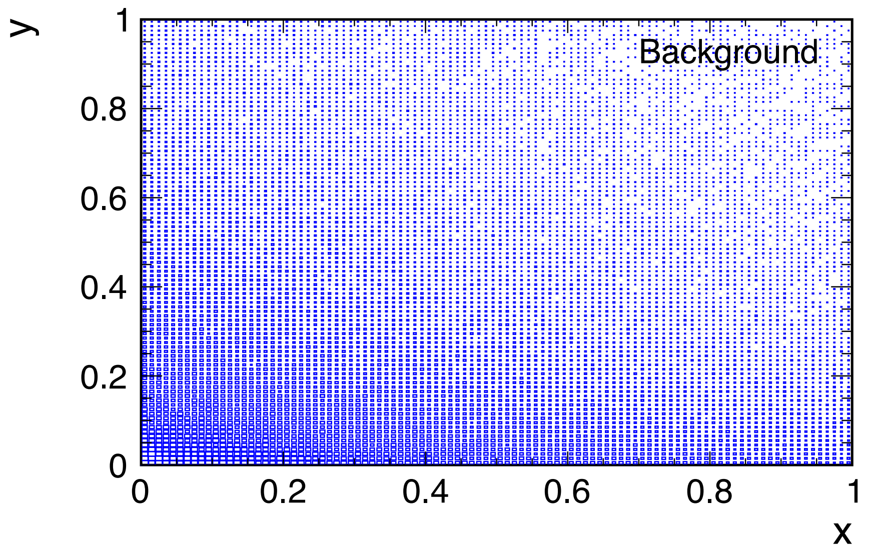
All arbitrary parameters were provided with zero uncertainty.

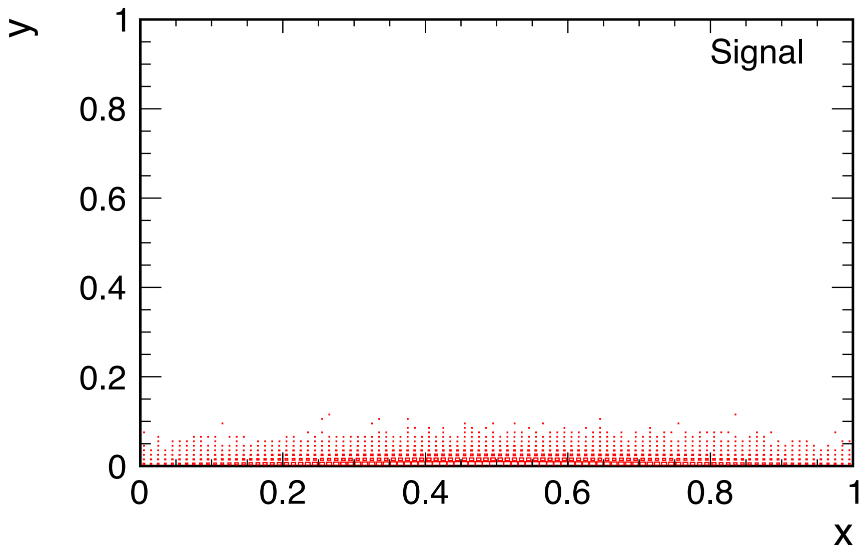
How do we explore something new?

- Add **Bad** and **Ugly** systematic errors. A step towards realism.
- How to do this without devolving into a guessing game?
- Simplest case of guess-the-hidden-offset is realistic, but not instructional.
  - Need domain knowledge to make informed guesses.

# Banff Challenge 3:  2D Problem

Two feature variables per interaction, *x* and *y*.   Patterned on MET and ISO in CDF's old W cross section analysis. See Pekka Sinervo's description:  [https://inspirehep.net/literature/637578](https://inspirehep.net/literature/637578)

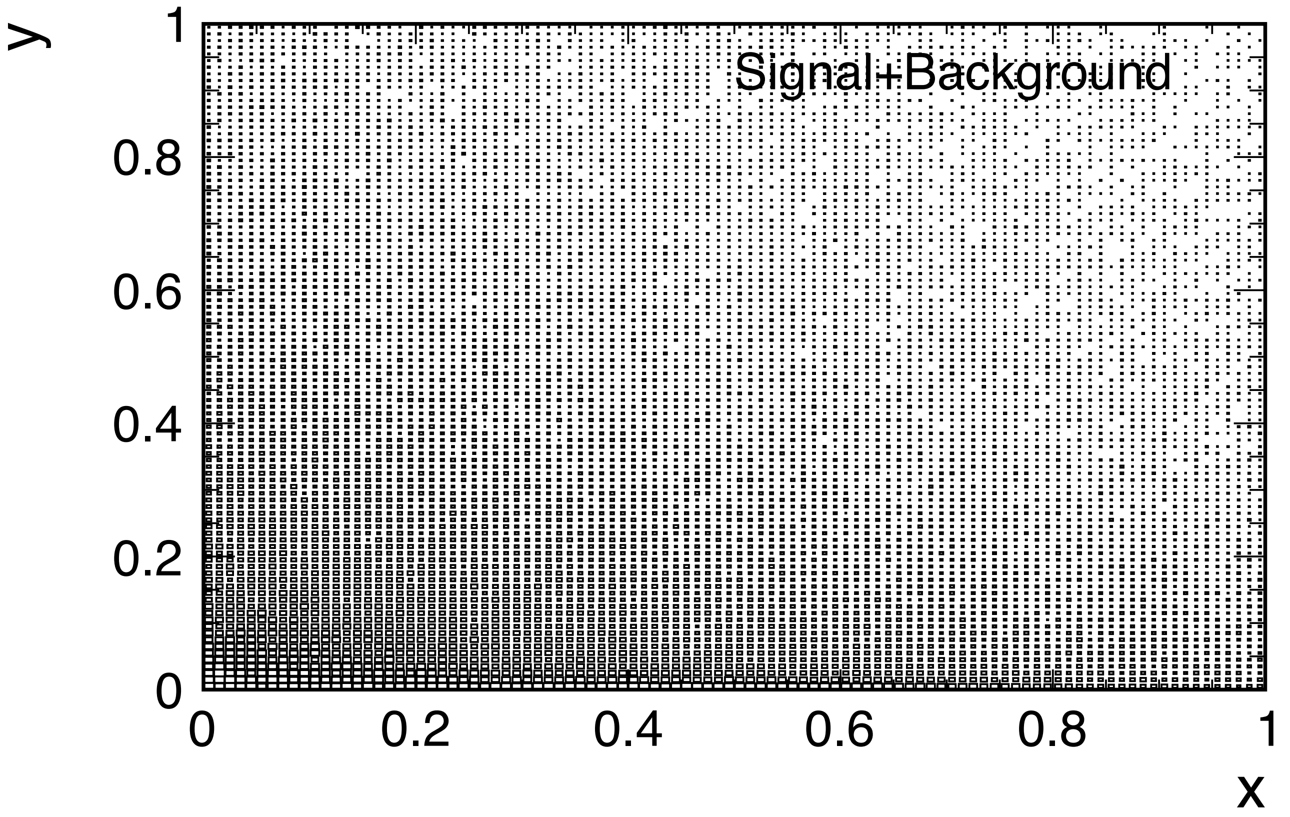




*x* and *y* are assumed to be independent in the background sample.  And independent in the signal sample.  But the sum of signal+background, they are not independent.

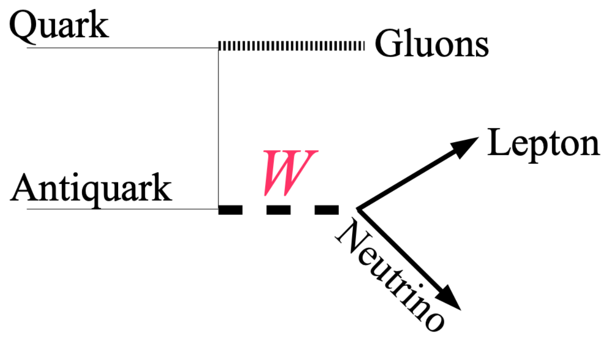
# Banff Challenge 3:  2D Problem

Experimental data are an unlabeled mixture of signal and background contributions.

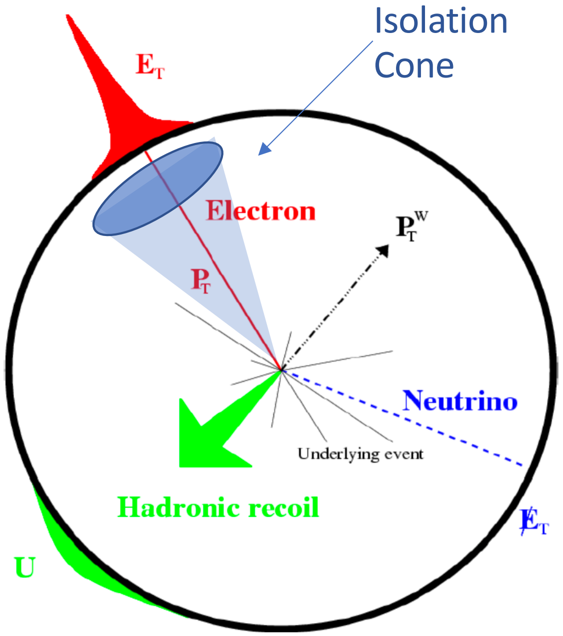


# The Physics Mechanism Inspiring the Distributions

## W Boson Production at the Tevatron



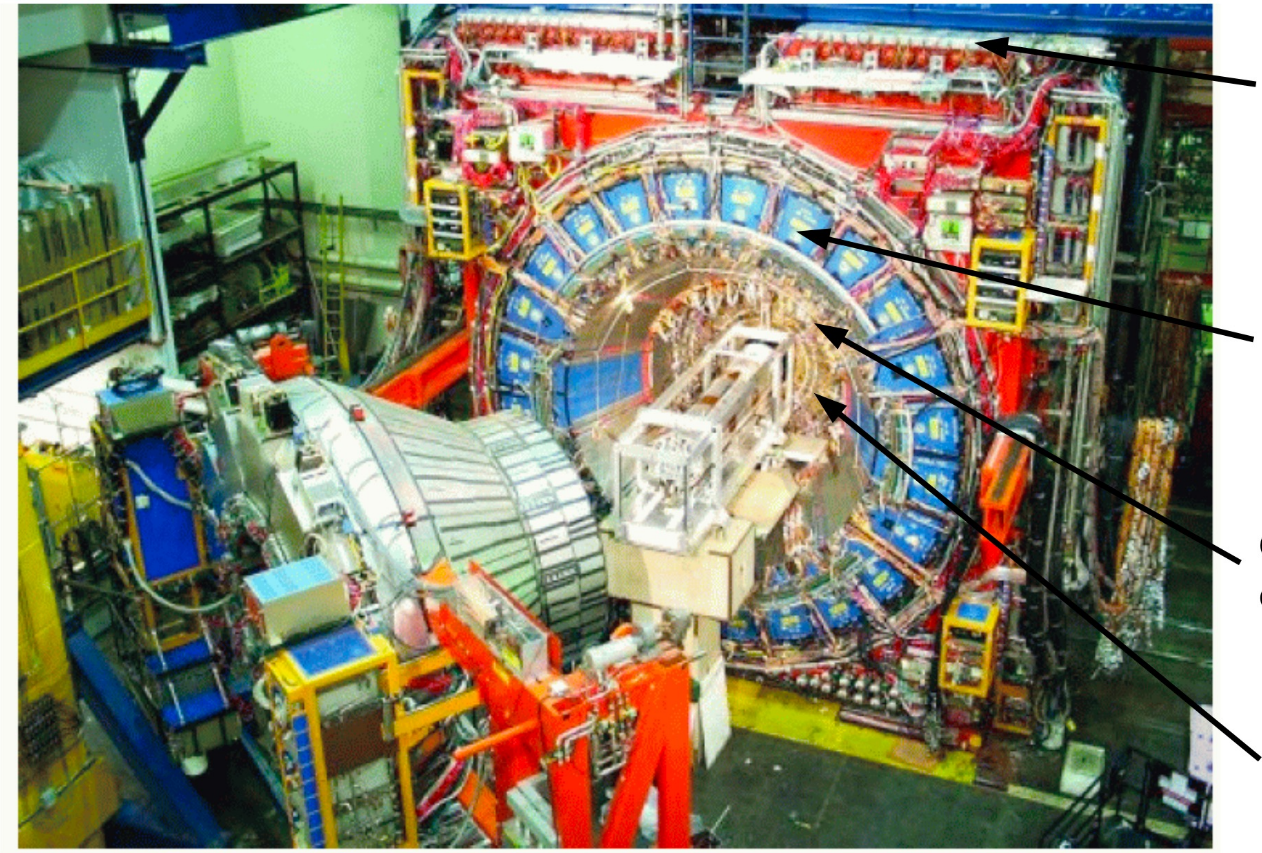
Quark-antiquark annihilation dominates (80%)



Lepton $p_T$ carries most of *W* mass information, can be measured precisely (achieved 0.004%)

Background: QCD multijets. No true missing energy, but fake missing energy pointing along a mismeasured jet. Fake lepton in QCD background tends to have other particles nearby, so the isolation variable *y* has larger values on average for QCD multijets than for signal.

# Collider Detector at Fermilab (CDF)



Muon detector

Central hadronic calorimeter

Central EM calorimeter

Central outer tracker (COT)

# Banff Challenge 3: What's Provided

- 100 sets of unlabeled "data". ASCII text files containing (*x*,*y*) pairs. Of the order 100k interactions per data set.
- Seven sets of simulated background "Monte Carlo" interactions
  - Exactly 100k interactions in each sample.
  - One "central" sample
  - One "alternate MC generator" sample
  - Three pairs of "up" and "down" systematic samples.
- Six sets of simulated signal "Monte Carlo" interactions
  - Three pairs of "up" and "down" systematic interactions
  - Nuisance parameters considered independent between signal and background. They correspond to different features of the models anyway.
- Not provided: true rate or shape information for any of the data samples. MC sample pairs may only cover some of the unknown parameters.
- Data samples are to be analyzed in isolation of each other – they were generated with different values of the parameters.

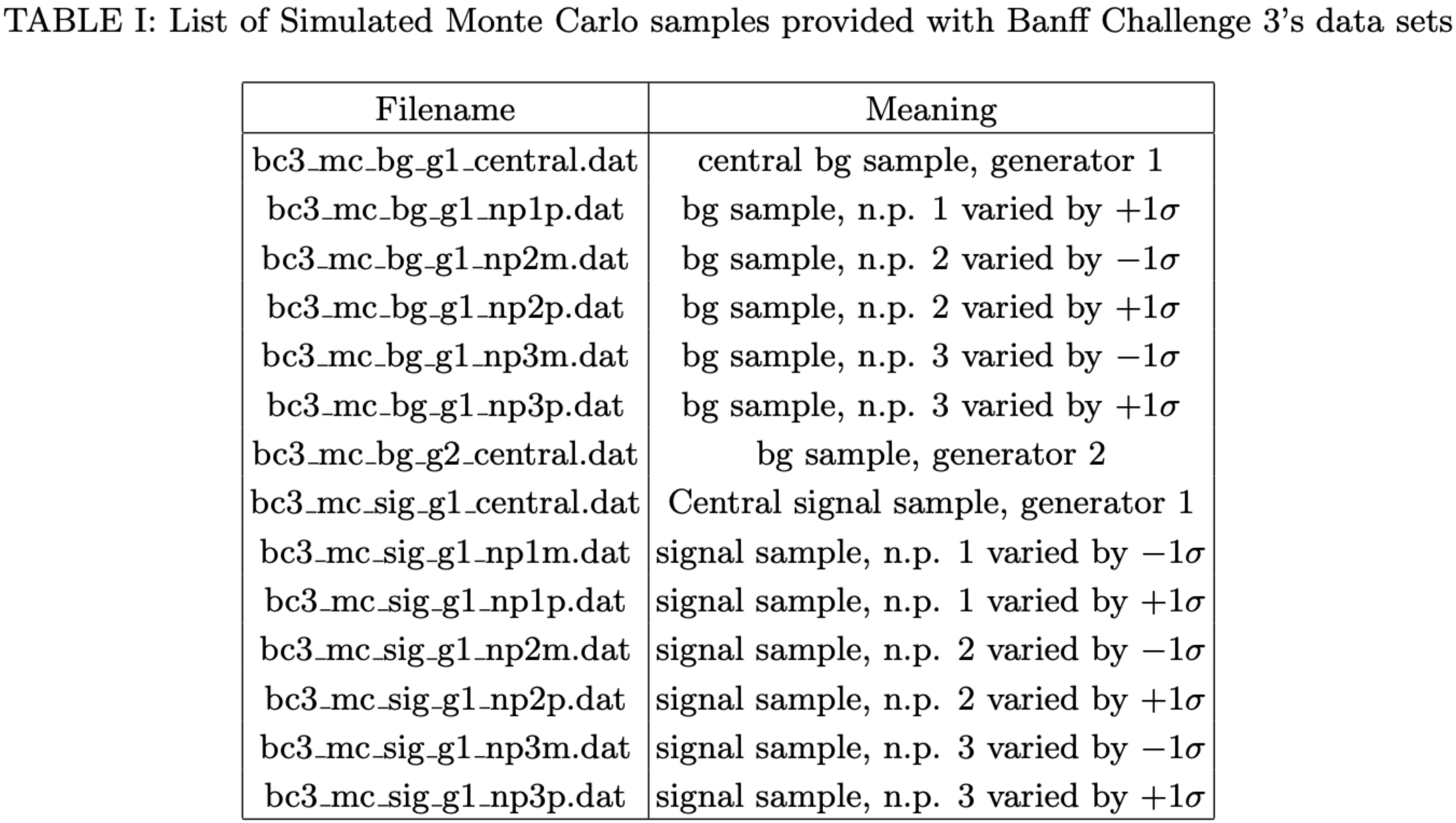


TABLE I: List of Simulated Monte Carlo samples provided with Banff Challenge 3's data sets

| Filename | Meaning |
|---|---|
| bc3_mc_bg_g1_central.dat | central bg sample, generator 1 |
| bc3_mc_bg_g1_np1p.dat | bg sample, n.p. 1 varied by $+1\sigma$ |
| bc3_mc_bg_g1_np2m.dat | bg sample, n.p. 2 varied by $-1\sigma$ |
| bc3_mc_bg_g1_np2p.dat | bg sample, n.p. 2 varied by $+1\sigma$ |
| bc3_mc_bg_g1_np3m.dat | bg sample, n.p. 3 varied by $-1\sigma$ |
| bc3_mc_bg_g1_np3p.dat | bg sample, n.p. 3 varied by $+1\sigma$ |
| bc3_mc_bg_g2_central.dat | bg sample, generator 2 |
| bc3_mc_sig_g1_central.dat | Central signal sample, generator 1 |
| bc3_mc_sig_g1_np1m.dat | signal sample, n.p. 1 varied by $-1\sigma$ |
| bc3_mc_sig_g1_np1p.dat | signal sample, n.p. 1 varied by $+1\sigma$ |
| bc3_mc_sig_g1_np2m.dat | signal sample, n.p. 2 varied by $-1\sigma$ |
| bc3_mc_sig_g1_np2p.dat | signal sample, n.p. 2 varied by $+1\sigma$ |
| bc3_mc_sig_g1_np3m.dat | signal sample, n.p. 3 varied by $-1\sigma$ |
| bc3_mc_sig_g1_np3p.dat | signal sample, n.p. 3 varied by $+1\sigma$ |

# Banff Challenge 3 Deliverables and Criterion for Winning

Participants should provide, for each challenge dataset,

1) Point estimate for the signal strength
2) 68% CL interval for the signal strength

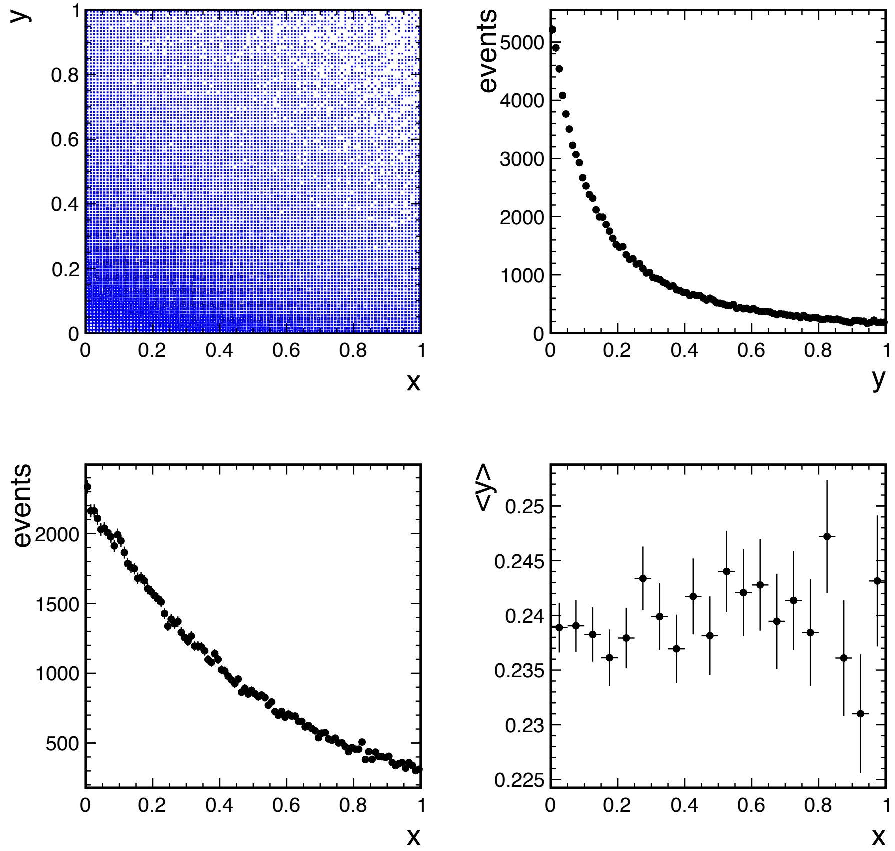
Winning requires 68% coverage of the intervals for the true signal strength, and among entries that cover, the winner will have the smallest average interval length, when the average is taken over the 100 simulated **data** sets *.

* In a real experiment, you use the MC to estimate sensitivity, but here we know too much about the MC (i.e. the true signal rate), and thus it is easy to submit a "too good" sensitivity estimate if we use the MC.  So we use the (simulated) data instead.

*Not* required:  null-hypothesis test *p* values, or GOF *p* values.

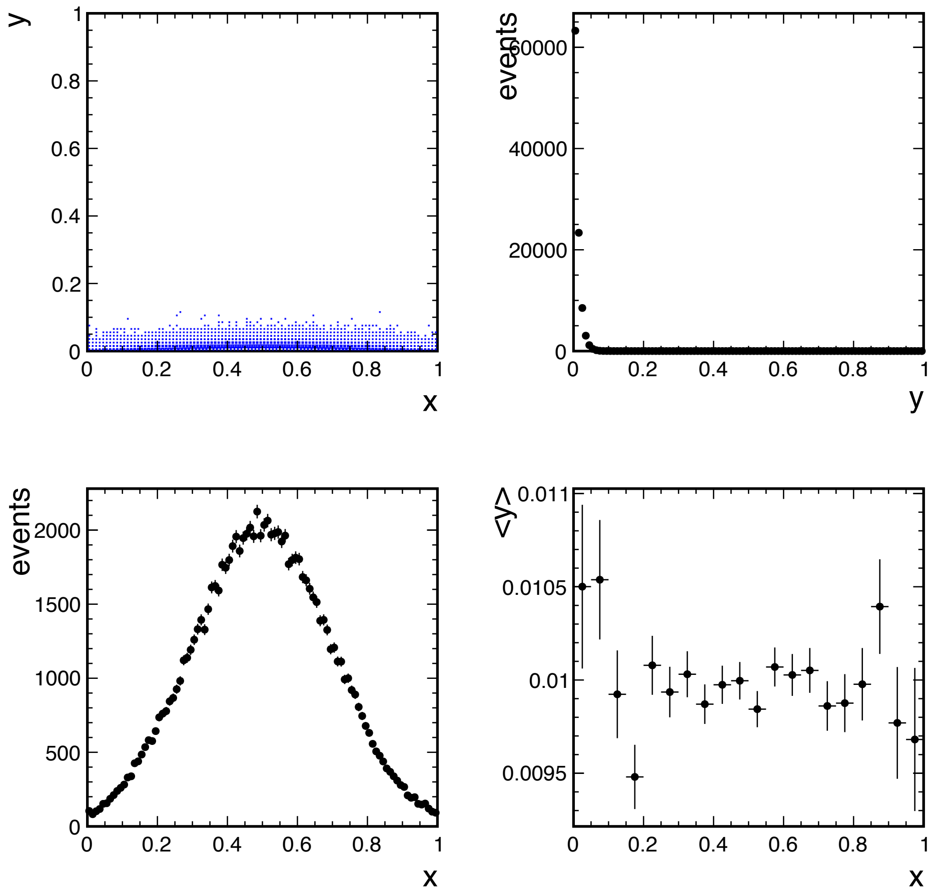
# Some Ways of Looking at the Data

Nominal
Background
MC Sample



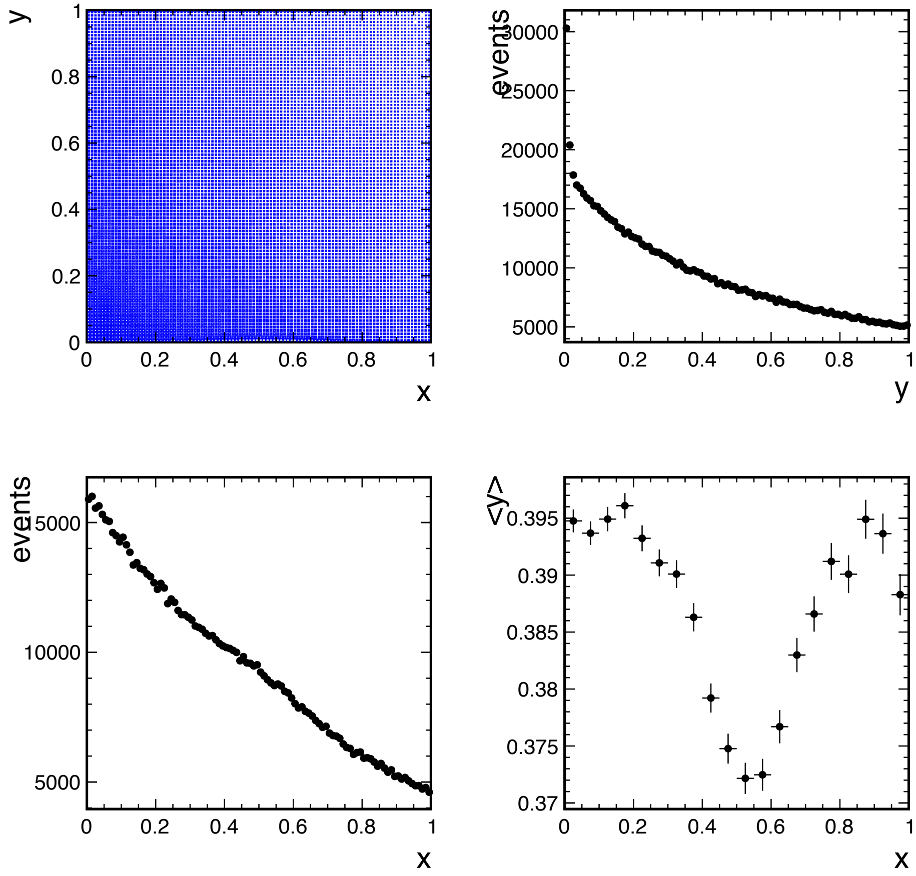
# Some Ways of Looking at the Data

Nominal
Signal
MC Sample



# Some Ways of Looking at the Data

Data Sample #4
-- has an obvious
signal in it.

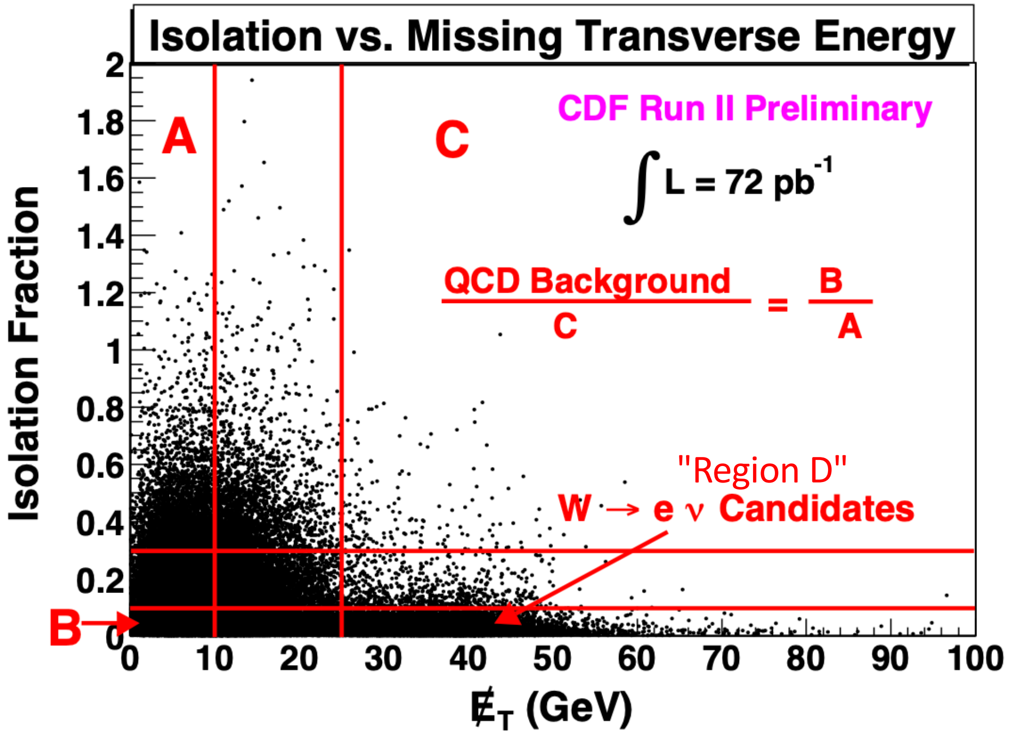


Variables are
not independent
for the sum of s+b

# ABCD Methods

"QCD Background" to be subtracted from the measured counts in Region D "W→ev Candidates"

$$\sigma_W = \frac{N_{\text{obs}} - N_{\text{bkg}}}{\epsilon L}$$



Pekka Sinervo, from Phystat 2003

https://inspirehep.net/literature/637578

Perform measurement in Region D.
Use ABCD formula to evaluate background

Watch out for signal contamination in A, B and C!
Gaps between cuts are meant to improve the purity of the samples but it's never 100%.

Efficiency $\epsilon$ is the probability for a signal interaction to be in region D. Total efficiency for a signal interaction to be anywhere in the plot = 100%, with no uncertainty.

*L* is the integrated luminosity – chosen to be 1 here.

**You don't have to use an ABCD method if you don't want to.**
**We encourage innovative techniques!**

# Differences from Banff Challenge 1

*A*/*B* plays the role of *t*. *C* plays the role of *y*.

Everything's still measured! Just an exercise like BC1 and BC2. But ...

Assumptions are needed

1) Independence of *x* and *y* in the signal and background samples separately
2) Amount of signal that leaks out of D and into A, B and C is known.

Both of these assumptions are broken in the challenge datasets.

Simulated Monte Carlo samples are provided. Seven sets for the background, six for the signal.
They provide "up" and "down" variations for nuisance parameters.
One background MC set corresponds to a one-sided model comparison
(such as another generator).

# Trivial Multiplicative Uncertainties are Not Included in BC3

The usual formula for a cross section: $\sigma_W = \frac{N_{\text{obs}} - N_{\text{bkg}}}{\epsilon L}$

Uncertainties on $\epsilon$ and $L$ are "interesting" only if there are subsidiary measurements, and those are covered in BC1 and BC2.

If instead they have priors, they just become a task of propagating uncertainties, or become guessing games if the true values of $\epsilon$ and $L$ are hidden.

So we set $\epsilon$ = 1 and $L$ = 1.  Report signal rate in number of events.

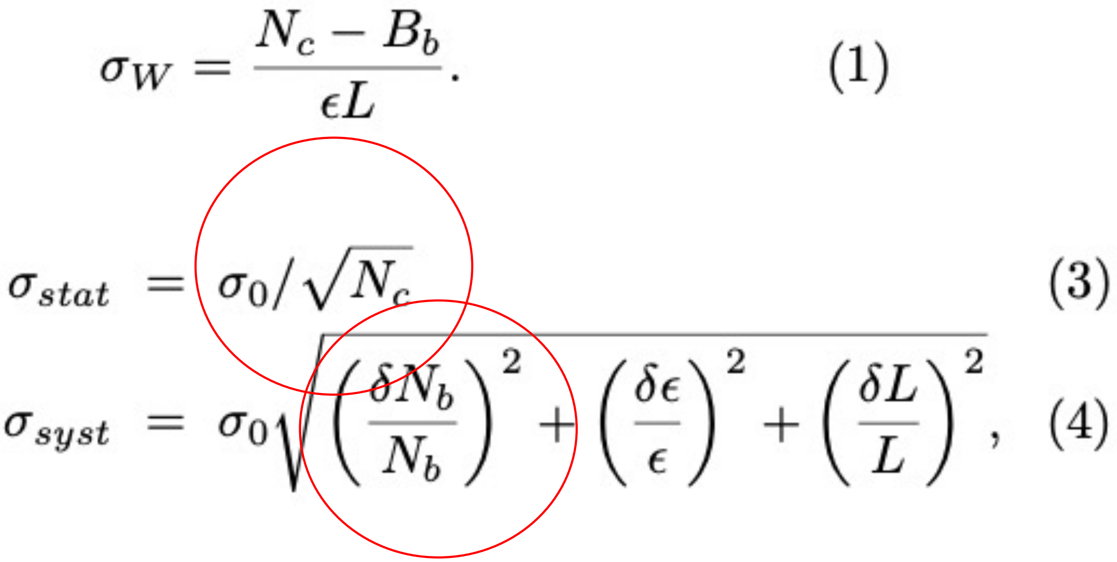
Here $\epsilon$ is the total efficiency for a signal interaction to be recorded.  If you select a subset by cutting on *x* and/or *y*, you have to estimate your $\epsilon$ and compute uncertainties.

# Domain Knowledge is Crucial for Real Analyses

Perhaps the most challenging aspect of estimating systematic uncertainties is to define in a consistent manner all the relevant sources of systematic uncertainty. This requires a comprehensive understanding of the nature of the measurement, the assumptions implicit or explicit in the measurement process, and the uncertainties and assumptions used in any theoretical models used to interpret the data.

P. Sinervo

# A Word of Caution about Pekka's Note



$$\sigma_W = \frac{N_c - B_b}{\epsilon L}. \tag{1}$$

$$\sigma_{stat} = \sigma_0/\sqrt{N_c} \tag{3}$$

$$\sigma_{syst} = \sigma_0\sqrt{\left(\frac{\delta N_b}{N_b}\right)^2 + \left(\frac{\delta \epsilon}{\epsilon}\right)^2 + \left(\frac{\delta L}{L}\right)^2}, \tag{4}$$

The circled terms are not right.  They treat additive uncertainties as multiplicative.

# Ideas for a More Challenging Challenge

*p* value calculation including systematic uncertainties

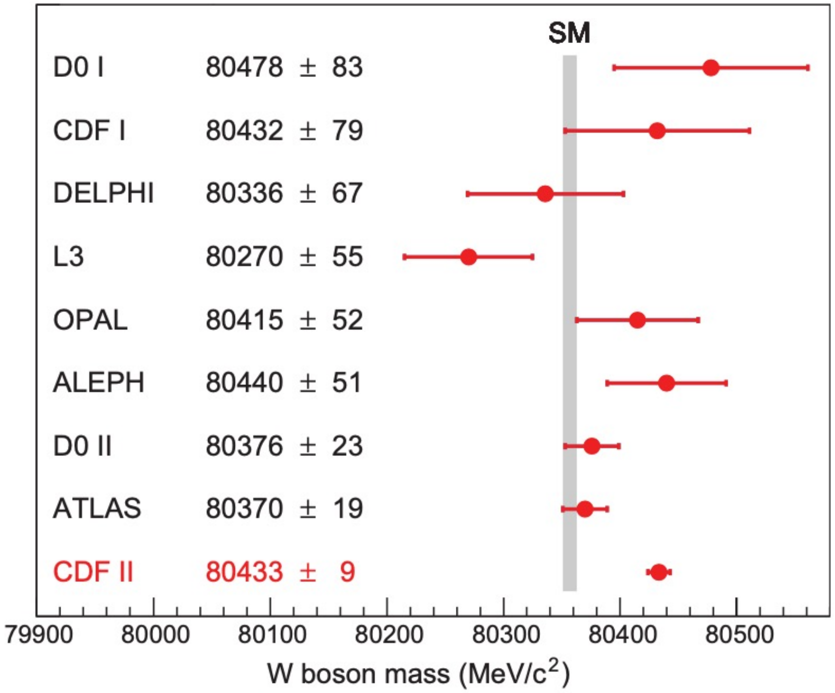
- Would need to provide tens of thousands of data sets to measure an error rate of 0.01 well
- This is more easily possible with binned data.  Otherwise 100s of GB of data need to be exchanged

Something with a bit more domain knowledge

- Example:  W mass measurement with a Z calibration sample
    - Lepton energy scale
        - Might need a little special relativity to compute $m_{ll}$.  Or one could just provide data from a known distribution (say Gaussian) and one fits the mean.
        - Not different enough from BC1's systematic uncertainty perhaps?

    - $P_T$ spectrum of Z is well measured.  Extrapolate to W.  Nuisance parameters are parton distribution function parameters

**Fig. 5. Comparison of this CDF II measurement and past $M_W$ measurements with the SM expectation.** The latter includes the published estimates of the uncertainty (4 MeV) due to missing higher-order quantum corrections, as well as the uncertainty (4 MeV) from other global measurements used as input to the calculation, such as $m_t$. *c*, speed of light in a vacuum.



CDF Collaboration, Science 376, 170–176 (2022)

Combination status, with lots of domain-specific studies:
http://cds.cern.ch/record/2815187?ln=en

# BC3 Data Sets and the Note

It's on my Google Drive:

https://drive.google.com/drive/folders/1i2yDyiQo7wQOw0hGv2guwSPwAgIuCfdo?usp=sharing

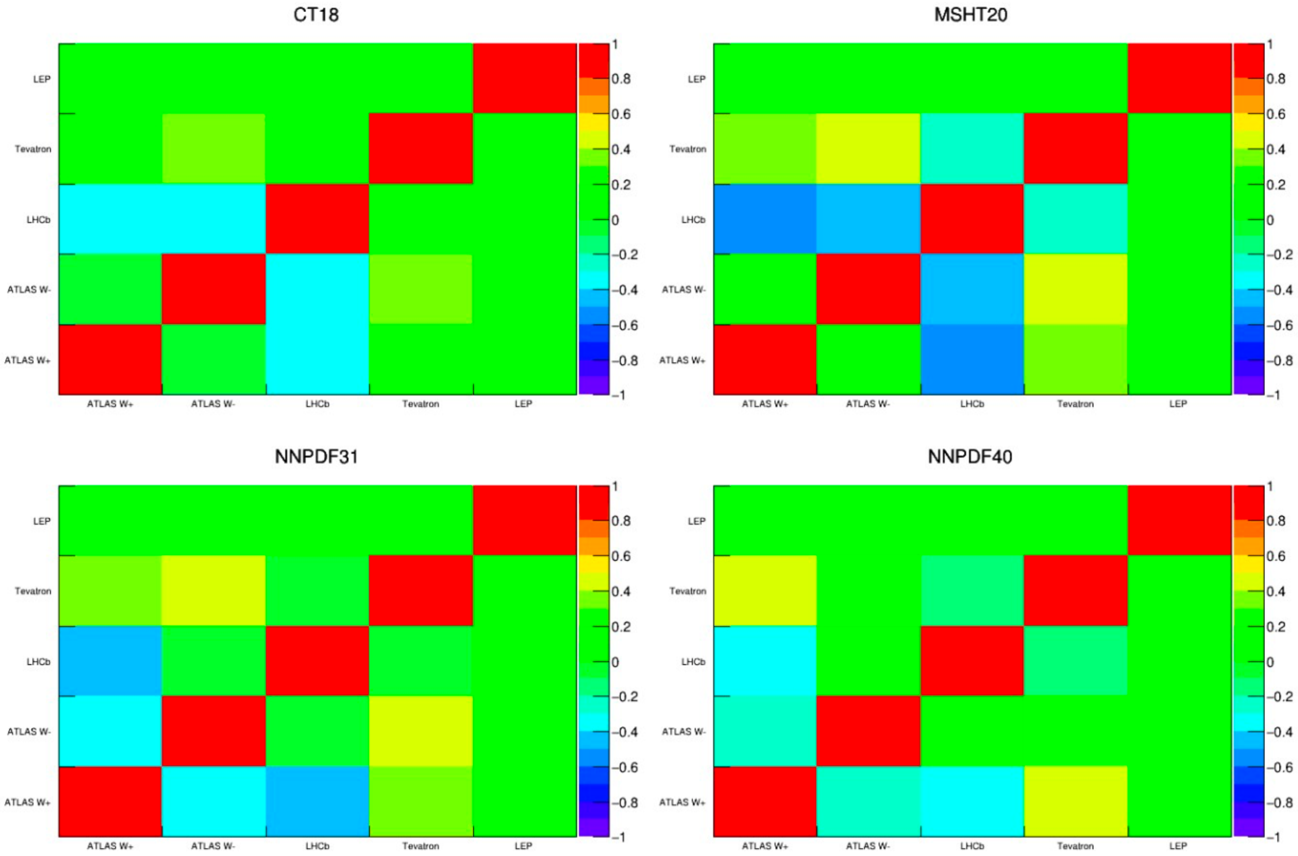
It contains:

1) Document with problem description and instructions.
2) Bzip2'd tarball containing ASCII data and MC sets
   Unpack this with tar -xjf bc3_challenge_sets.bz2
3) A copy of Pekka Sinervo's Phystat 2003 proceedings

Good luck and have fun!

# Extras

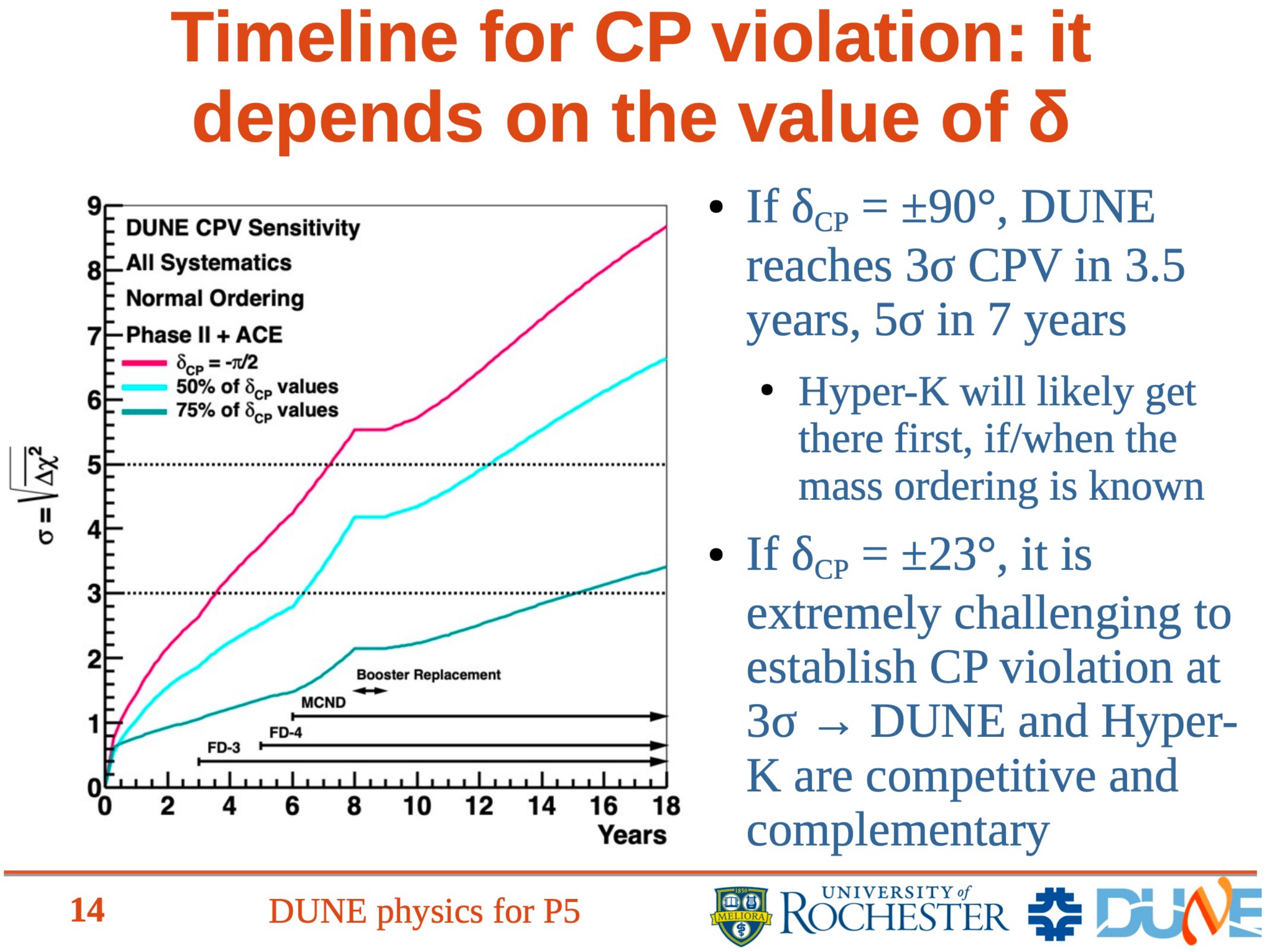
# W Mass PDF Correlations



Maarten Boonekamp

https://indico.cern.ch/event/1251919/contributions/5336989/attachments/2630101/4548847/mWdays_170423.pdf

# Slide from Chris Marshall on DUNE's CP Sensitivity



## Timeline for CP violation: it depends on the value of δ

- If $\delta_{CP}$ = ±90°, DUNE reaches $3\sigma$ CPV in 3.5 years, $5\sigma$ in 7 years
  - Hyper-K will likely get there first, if/when the mass ordering is known
- If $\delta_{CP}$ = ±23°, it is extremely challenging to establish CP violation at $3\sigma$ → DUNE and Hyper-K are competitive and complementary

14 DUNE physics for P5