Lecture 9:

- Blind Analysis
- Bifurcated Side-Band Analysis
- Data "Correction"
- Statistical Optimisation
- Redundancy

"Blind" Analysis Techniques

Goal: To remove the ability to unconsciously tune on statistical fluctuations and/or adjust analyses towards a particular outcome by hiding the final result until the full analysis (incl. assessment of uncertainties) is fixed.

At which point you then "open the box" and take what life brings you!

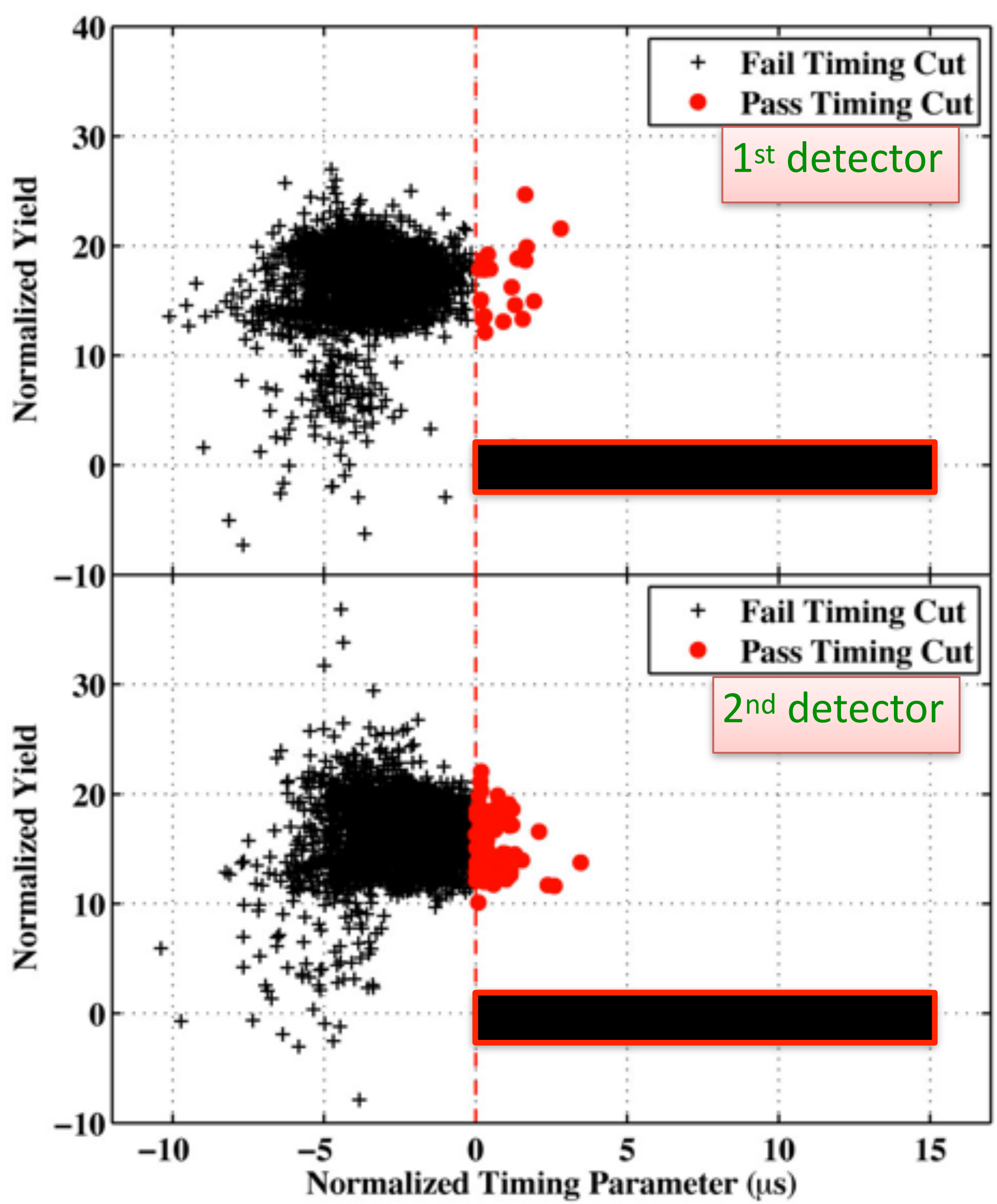
Rules of the Game

- **Agree on an appropriate blindness scheme in advance**
- **Make sure no one breaks it**
- **Agree on the criteria necessary to "open the box"**
- **State the blindness scheme up front in any publication**
- **Agree to show exactly what results from box-opening and then justify any alterations**

Signal Box Method

CDMS results on search for Dark Matter (Dec, 2009)

Expected summed background in both detectors: **0.9 ± 0.2**



RESULTS:

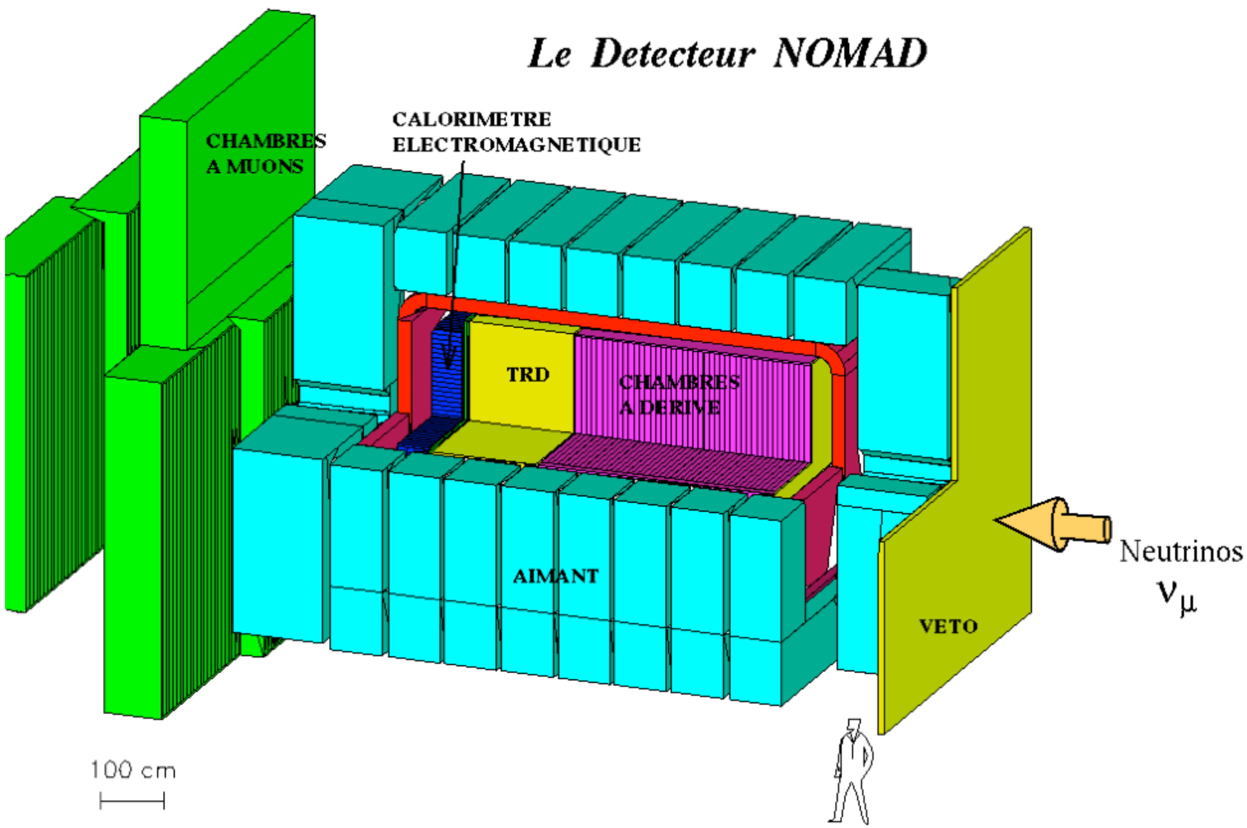
[redacted]

Divided Data Sample

NOMAD Search for

ν_μ - ν_τ oscillations

(Feb, 1999)



Used 20% of data to confirm background predictions and define search window, then impose signal box method on remaining 80% of the data

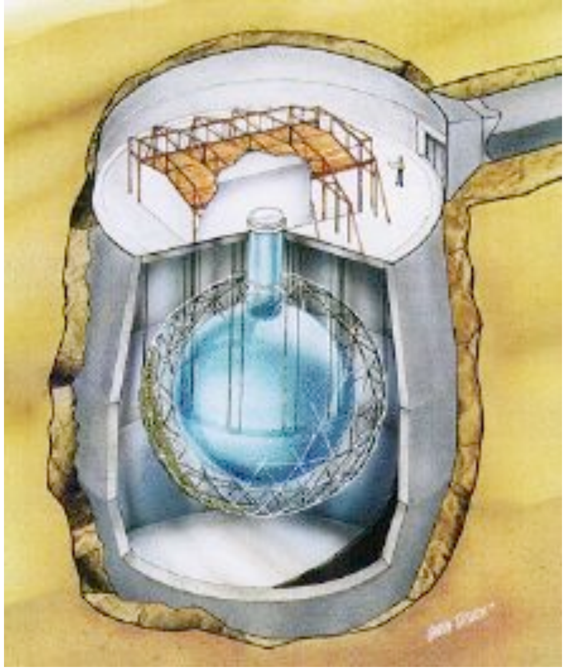
Expected background in signal box: **6.5 ± 1.1**

RESULTS:

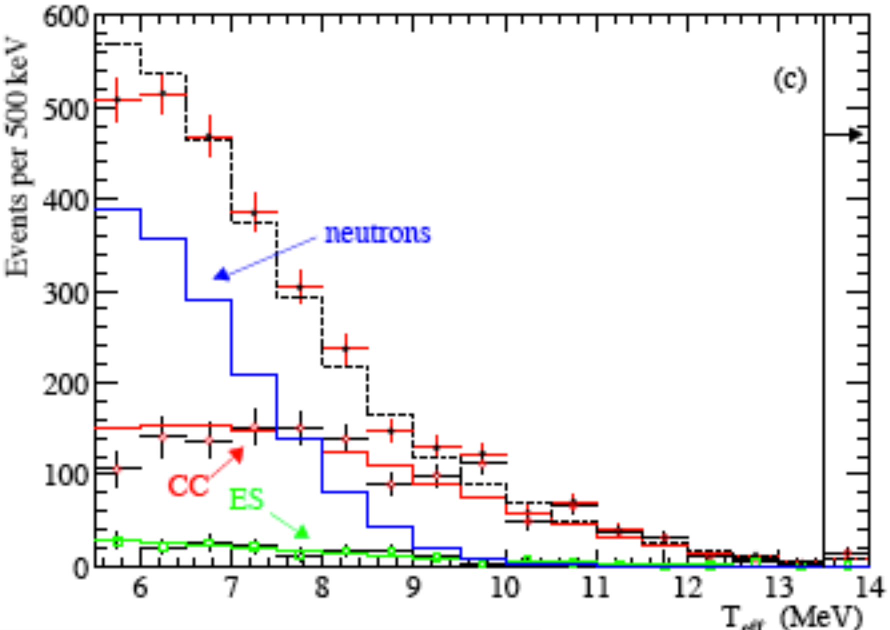
[redacted]

Hidden Parameters

SNO Measurement of total solar neutrino flux (Sept, 2003)



Excluded a hidden fraction of the final data set (unknown flux normalisation), included hidden admixture of tagged background neutrons, scaled simulation NC cross section by hidden factor

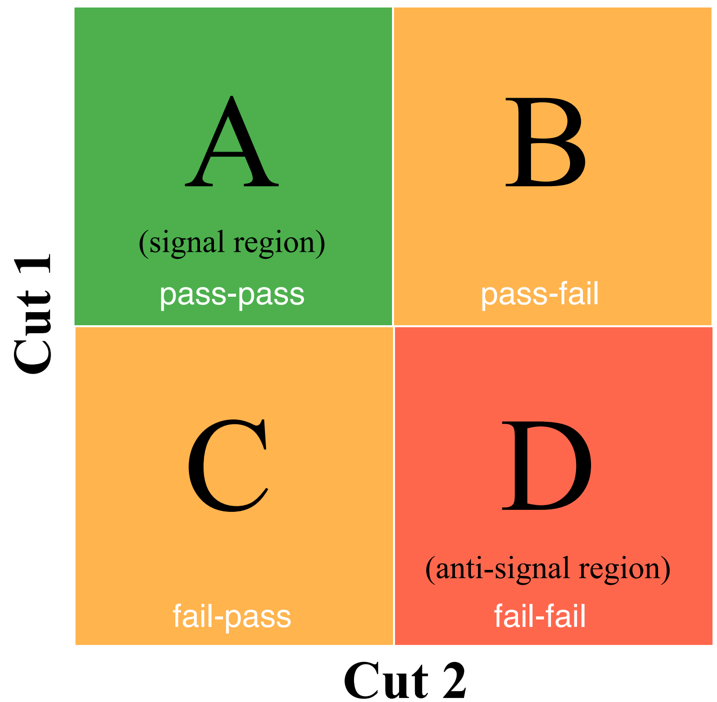


RESULTS:

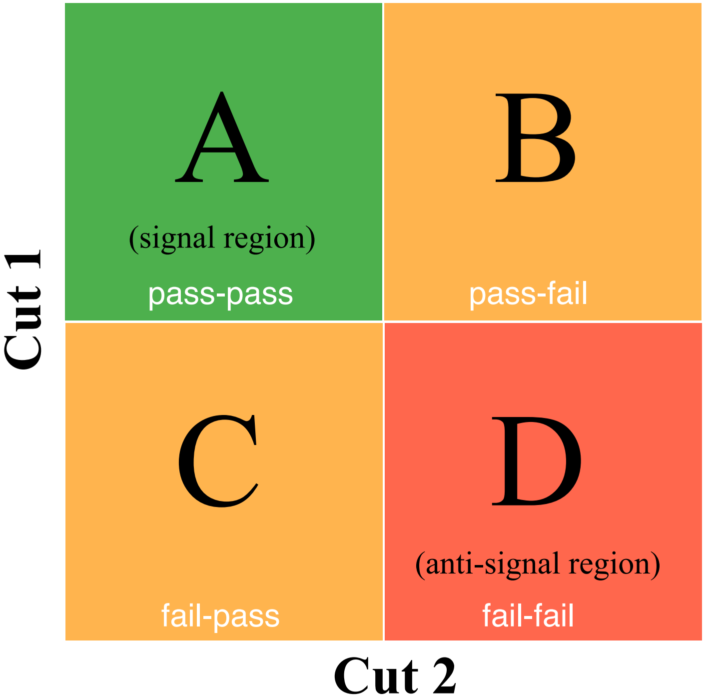
[redacted]

Bifurcated Side-Band Analysis*

Assume we have a data set with a total number of signal S and a total number of background B. Further assume that we have two independent parameters (for example, energy and fiducial volume) that can be used to cut out some number of unknown background while maintaining high signal efficiency (based on simulations of the signal). We wish to estimate the background contamination in the signal region:



*Generalisation of Adler *et al.*, PRL 79, 12 1997 and Nix *et al.*, NIM A615, 2, 2010 to account for signal efficiencies



Take the efficiency of retaining signal from each cut in the signal region to be **ε_1** and **ε_2**, respectively. Similarly, take the fractions of background rejected by each cut in this region to be **r_1** and **r_2**, respectively.

$$N_A = S\epsilon_1\epsilon_2 + Br_1r_2 \qquad \equiv s + b$$

$$N_B = S\epsilon_1(1-\epsilon_2) + Br_1(1-r_2)$$

$$N_C = S\epsilon_2(1-\epsilon_1) + Br_2(1-r_1)$$

$$N_D = S(1-\epsilon_1)(1-\epsilon_2) + B(1-r_1)(1-r_2)$$

To simplify the algebra a bit, let's redefine variables:

$$n_A \equiv \frac{N_A}{\epsilon_1\epsilon_2} = S + B\left(\frac{r_1r_2}{\epsilon_1\epsilon_2}\right)$$

$$n_C \equiv \frac{N_C}{\epsilon_2(1-\epsilon_1)} = S + B\left(\frac{r_2(1-r_1)}{\epsilon_2(1-\epsilon_1)}\right)$$

$$n_B \equiv \frac{N_B}{\epsilon_1(1-\epsilon_2)} = S + B\left(\frac{r_1(1-r_2)}{\epsilon_1(1-\epsilon_2)}\right)$$

$$n_D \equiv \frac{N_D}{(1-\epsilon_1)(1-\epsilon_2)} = S + B\left(\frac{(1-r_1)(1-r_2)}{(1-\epsilon_1)(1-\epsilon_2)}\right)$$

$$n_A - S = B\left(\frac{r_1 r_2}{\epsilon_1\epsilon_2}\right) \quad n_B - S = B\left(\frac{r_1(1-r_2)}{\epsilon_1(1-\epsilon_2)}\right) \quad n_C - S = B\left(\frac{r_2(1-r_1)}{\epsilon_2(1-\epsilon_1)}\right) \quad n_D - S = B\left(\frac{(1-r_1)(1-r_2)}{(1-\epsilon_1)(1-\epsilon_2)}\right)$$

$$(n_C - S)(n_B - S) = (n_A - S)(n_D - S)$$

$$n_C n_B - n_C S - S n_B + S^2 = n_A n_D - n_A S - S n_D + S^2$$

$$S = \frac{n_A n_D - n_C n_B}{n_A + n_D - n_C - n_B}$$

re-expanding:

$$\boxed{S = \frac{N_A N_D - N_C N_B}{N_A(1-\epsilon_1)(1-\epsilon_2) + N_D\epsilon_1\epsilon_2 - N_C\epsilon_1(1-\epsilon_2) - N_B\epsilon_2(1-\epsilon_1)}}$$

$$s = S\epsilon_1\epsilon_2 \qquad b = N_A - S\epsilon_1\epsilon_2$$

Do not need to look inside the signal region,
nor necessarily know details about r_1 and r_2 !

$$S = \frac{N_A N_D - N_C N_B}{N_A(1-\epsilon_1)(1-\epsilon_2) + N_D\epsilon_1\epsilon_2 - N_C\epsilon_1(1-\epsilon_2) - N_B\epsilon_2(1-\epsilon_1)}$$

$$s = S\epsilon_1\epsilon_2 \qquad\qquad b = N_A - S\epsilon_1\epsilon_2$$

note: as $\epsilon_1, \epsilon_2 \to 1$ $\qquad b \to \frac{N_B N_C}{N_D}$

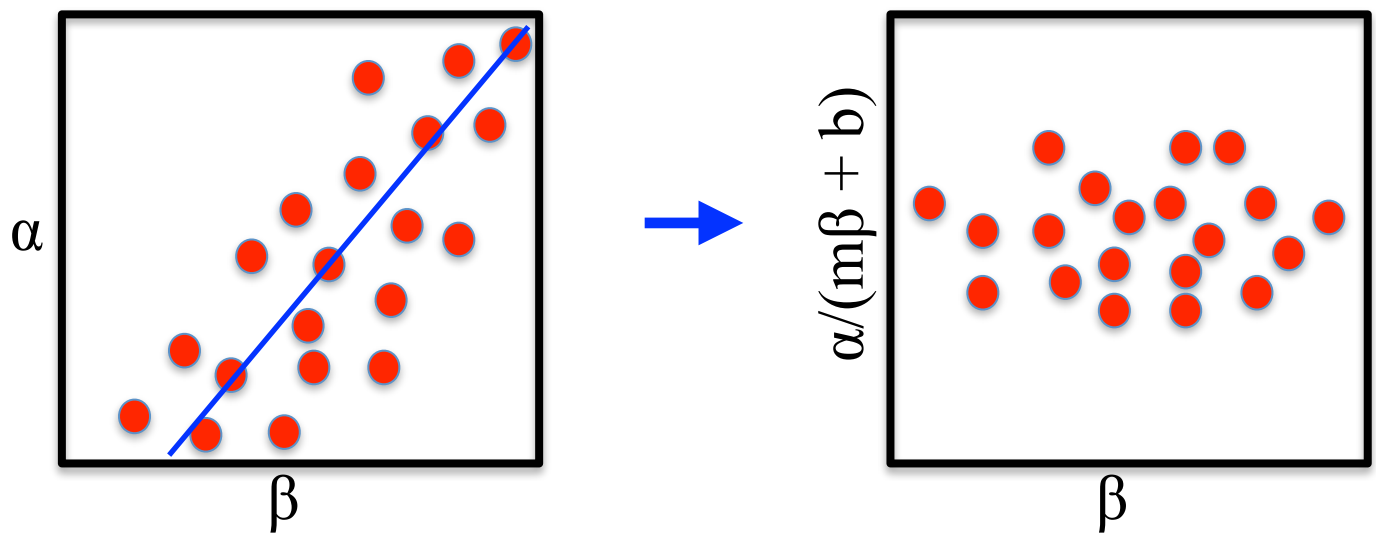
So, for large efficiencies, the variance in the estimated background contamination, $\mathbf{b}$, is approximately:

$$\sigma_{var}^2 \simeq N_B\left(\frac{N_C}{N_D}\right)^2 + N_C\left(\frac{N_B}{N_D}\right)^2 + N_D\left(\frac{N_B N_C}{N_D^2}\right)^2$$

Remember, this assumes cut parameters are uncorrelated! Note that a mixed background model can inadvertently produce correlations if, for example, both **r1** and **r2** are notably different between background components: then a particular cut value could favour a particular background, which could then produce a correlated rejection for the second cut.

In general, should look for possible correlations by plotting one cut parameter versus another, for example, in the anti-signal cut region (*i.e.* box D).

If a correlation is present, you may be able to redefine your parameters to remove this to first order. For example:



Alternatively, we can first define the background model as the sum of various components. Now assume that we can decompose these into a set of backgrounds that are **well-modelled and potentially sub-dominant,** plus a background with the highest uncertainty that we most wish to evaluate:

$$\sum_i B_i r_1^i r_2^i + B r_1 r_2$$

background we most want to evaluate

Then, similar to before, we can define the following quantities:

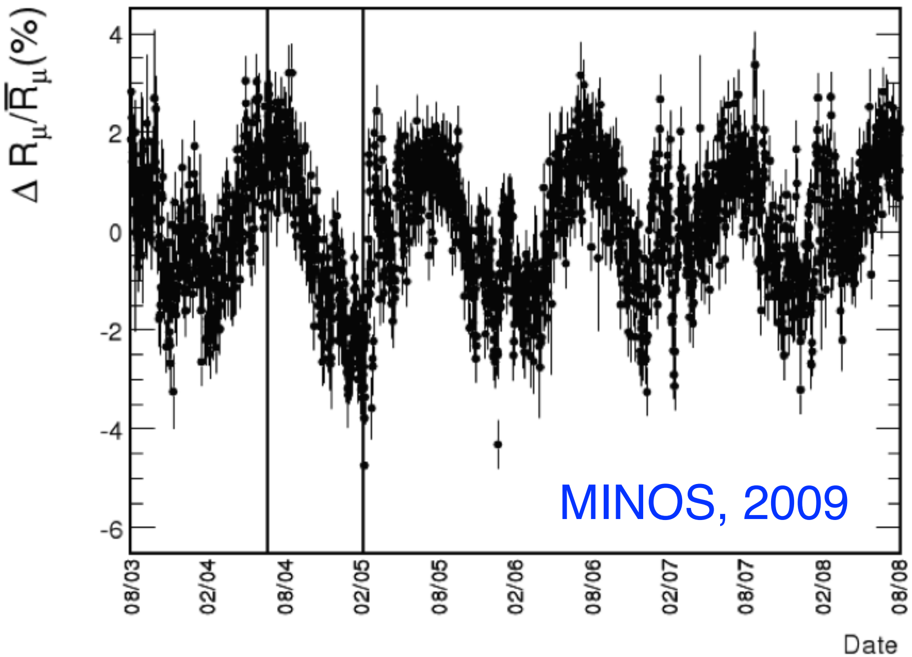
$$\eta_A \equiv \frac{1}{\epsilon_1\epsilon_2}\left(N_A - \sum_i B_i r_1^i r_2^i\right) = S + B\left(\frac{r_1 r_2}{\epsilon_1\epsilon_2}\right)$$

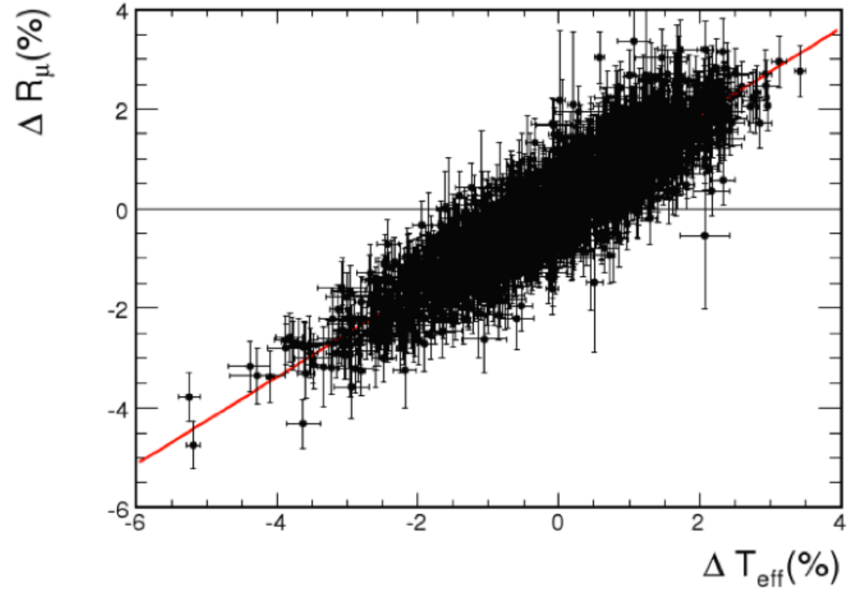
$$\eta_B \equiv \frac{1}{\epsilon_1(1-\epsilon_2)}\left(N_B - \sum_i B_i r_1^i (1-r_2^i)\right) = S + B\left(\frac{r_1(1-r_2)}{\epsilon_1(1-\epsilon_2)}\right)$$

$$\eta_C \equiv \frac{1}{\epsilon_2(1-\epsilon_1)}\left(N_C - \sum_i B_i r_2^i (1-r_1^i)\right) = S + B\left(\frac{r_2(1-r_1)}{\epsilon_2(1-\epsilon_1)}\right)$$

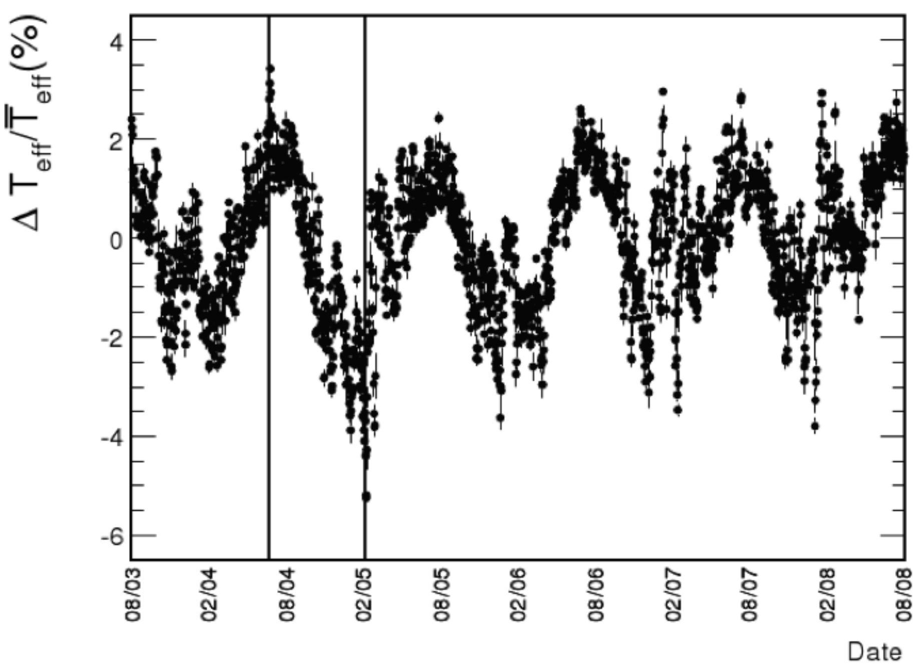
$$\eta_D \equiv \frac{1}{(1-\epsilon_1)(1-\epsilon_2)}\left(N_D - \sum_i B_i (1-r_1^i)(1-r_2^i)\right) = S + B\left(\frac{(1-r_1)(1-r_2)}{(1-\epsilon_1)(1-\epsilon_2)}\right)$$

$$\boxed{S = \frac{\eta_A\eta_D - \eta_C\eta_B}{\eta_A + \eta_D - n_C - n_B}}$$





Correlation can be used
to correct model prediction



Working Backwards

Assume that both the signal and background levels are proportional to the detector mass, M, and running time, T. Find an expression for the maximum background level that can be tolerated to achieve a 3σ detection as a fraction of the expected signal for a given model. How does the sensitivity change as a function of M and T ?

$$B = fS$$

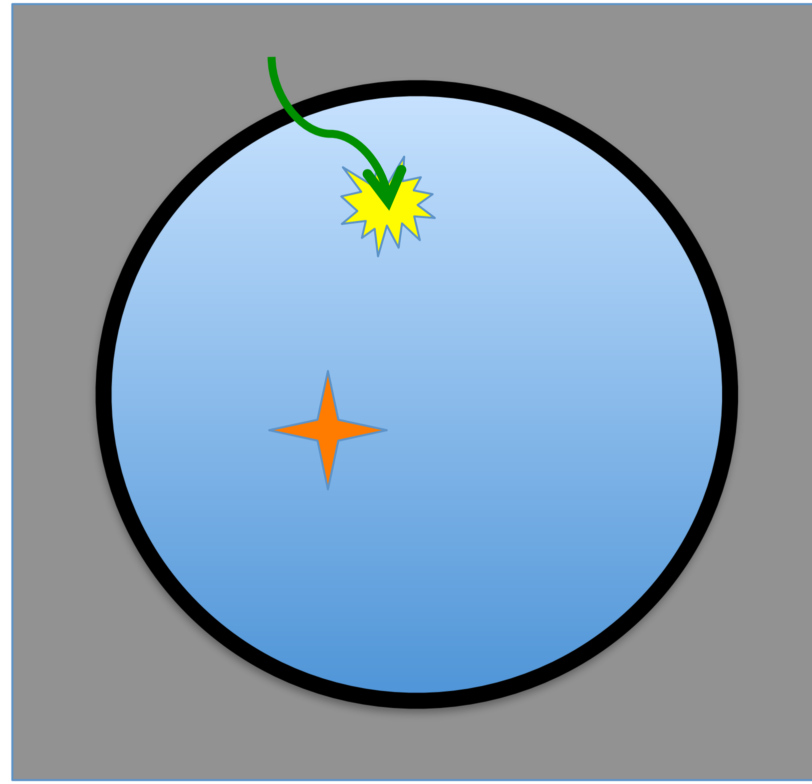
$$1\sigma = \sqrt{B} = \sqrt{fS}$$

under H0

Thus, for a 3σ signal: $$3\sqrt{fS} = S$$

(able to tolerate more background for larger signal) $$f = \frac{S}{9}$$

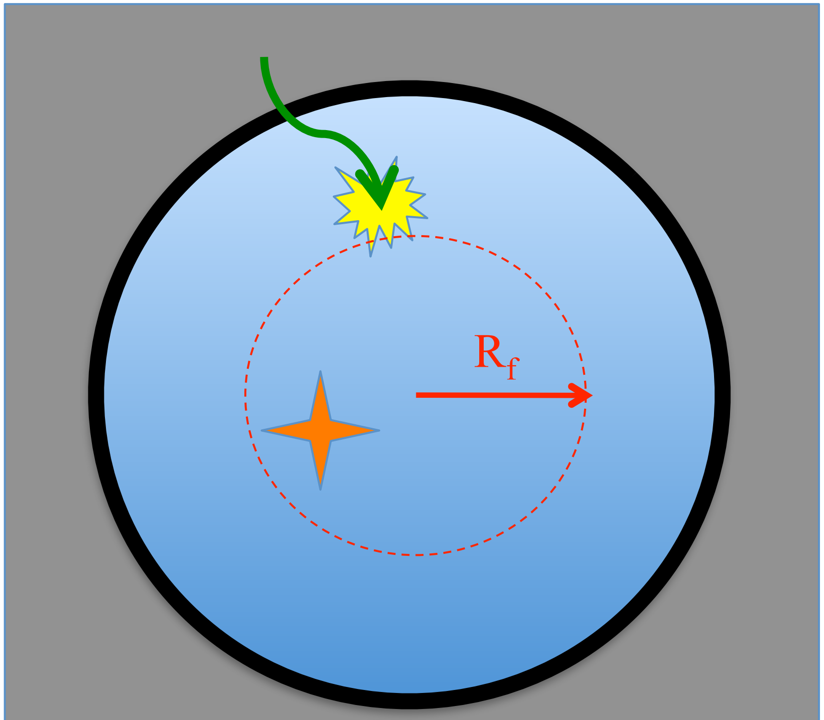
or $$B = \frac{S^2}{9}$$

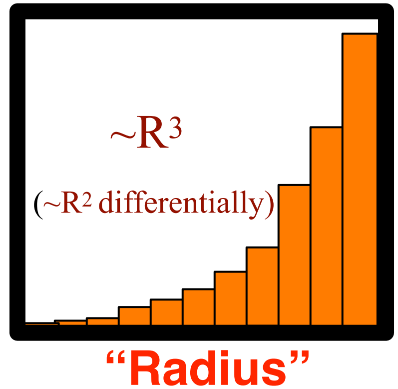


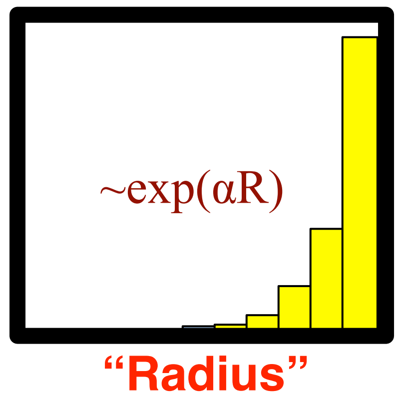
$$\text{Significance } (\sigma\text{'s}) = \frac{S}{\sqrt{B}}$$

$$= \frac{\alpha MT}{\sqrt{f\alpha MT}} \propto \sqrt{MT}$$

Example of Statistical Optimisation







$$\frac{S}{\sqrt{B}} \sim \frac{R^3}{\sqrt{\exp(\alpha R)}} = R^3 e^{-\alpha R/2}$$

Assume that we are in the "large N" limit and expected the number of counts to be dominated by background events.

We wish to exclude the worst of the background by choosing a radius to define a "fiducial volume," within which will look for an excess of events as evidence of a signal.

What choice of fiducial radius will give the best sensitivity for the search?

maximise:

$$3R^2 e^{-\alpha R/2} - \frac{\alpha}{2} R^3 e^{-\alpha R/2} = 0$$

$$3R^2 = \frac{\alpha}{2} R^3 \qquad R = \frac{6}{\alpha}$$

From the plot, it looks like backgrounds fall by ~1/e when R changes by 10% of the detector radius... so α ~10

$$R_f = 0.6 R_d$$

Sudbury Neutrino Observatory (SNO)

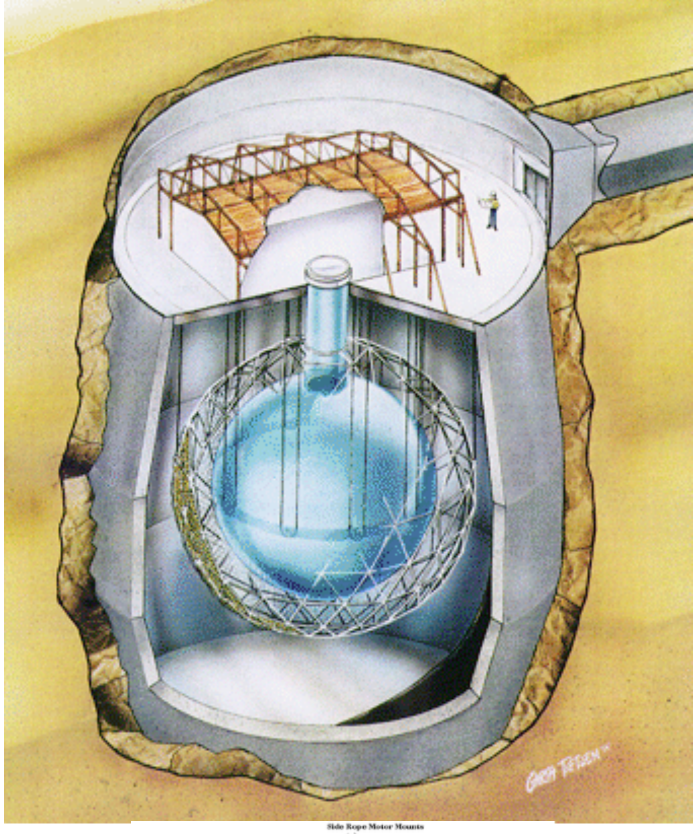
3 Different Operational Phases

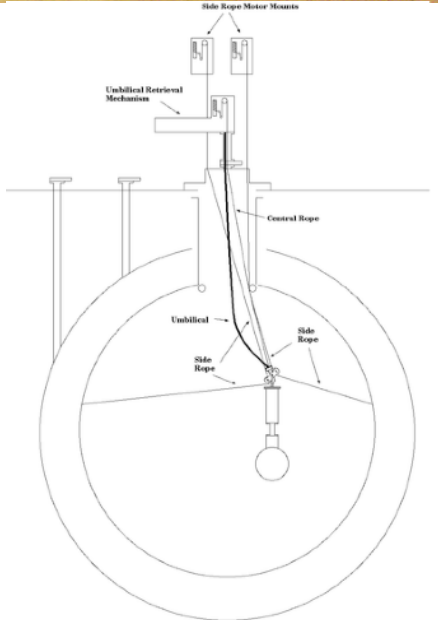
Found that estimated systematic uncertainty in possible position-dependent energy resolution was larger for the 2nd phase, which should have performance at least as good as 1st phase(?!)

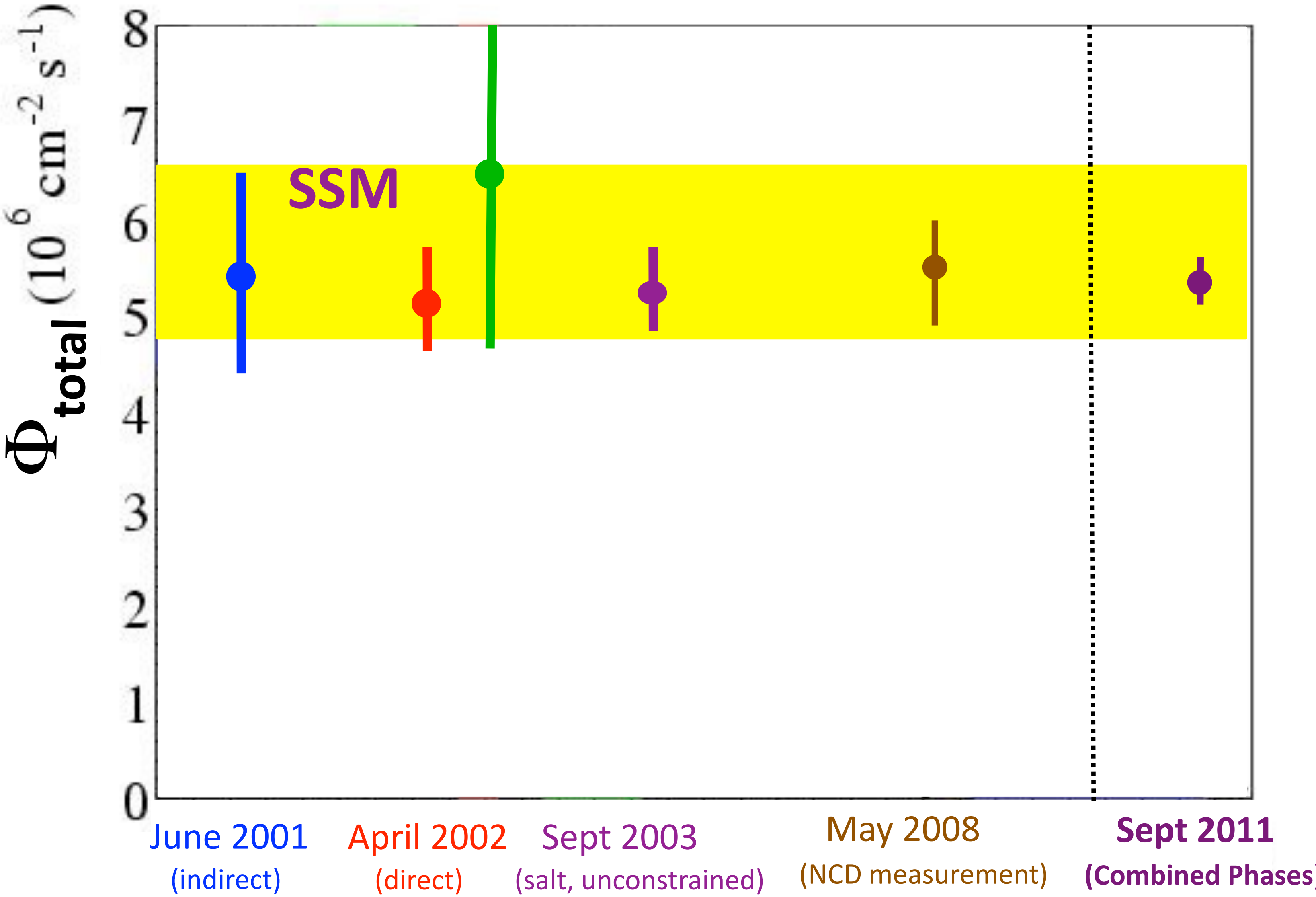
Realised that fewer calibrations had been done in 1st phase, so there was less data to compare!

If you don't look, you don't see!!

(Some groups seem to have elevated this to a strategy for getting small errors!)







3 Experimental Techniques,
at Least 2 Analyses/Technique + Combined Cross-checks