# The Confidence Game

Steve Biller

Oxford University

# In Conclusion:

## Frequentist confidence intervals

- Do not reliably indicate the relevance of individual measurements nor provide a robust basis for the comparison of different results;
- Are prone to frequent misinterpretation that can result in incorrect conclusions and seemingly paradoxical behaviours;
- Have no self-consistent method for error propagation or for producing single parameter bounds from a multi-dimensional parameter space;
- Do a poor job of conveying the objective information content of the data in a useful form.

# There is a better way!

*Another look at confidence intervals: proposal for a more relevant and transparent approach*
Biller and Oser, http://arxiv.org/abs/1405.5010 (2014)

*"In many cases you get similar looking bounds with either approach, so then it doesn't matter so much"*

*"In many cases you get similar looking bounds with either approach, so then it doesn't matter so much"*

**If it matters:**
You should use the correct formalism to answer the relevant question.

**If it doesn't matter:**
Why not use the correct formalism to answer the relevant question?

"A Frequentist uses impeccable logic to answer the wrong question, while a Bayesian answers the right question by making assumptions that nobody can fully believe in." P.G. Hamer

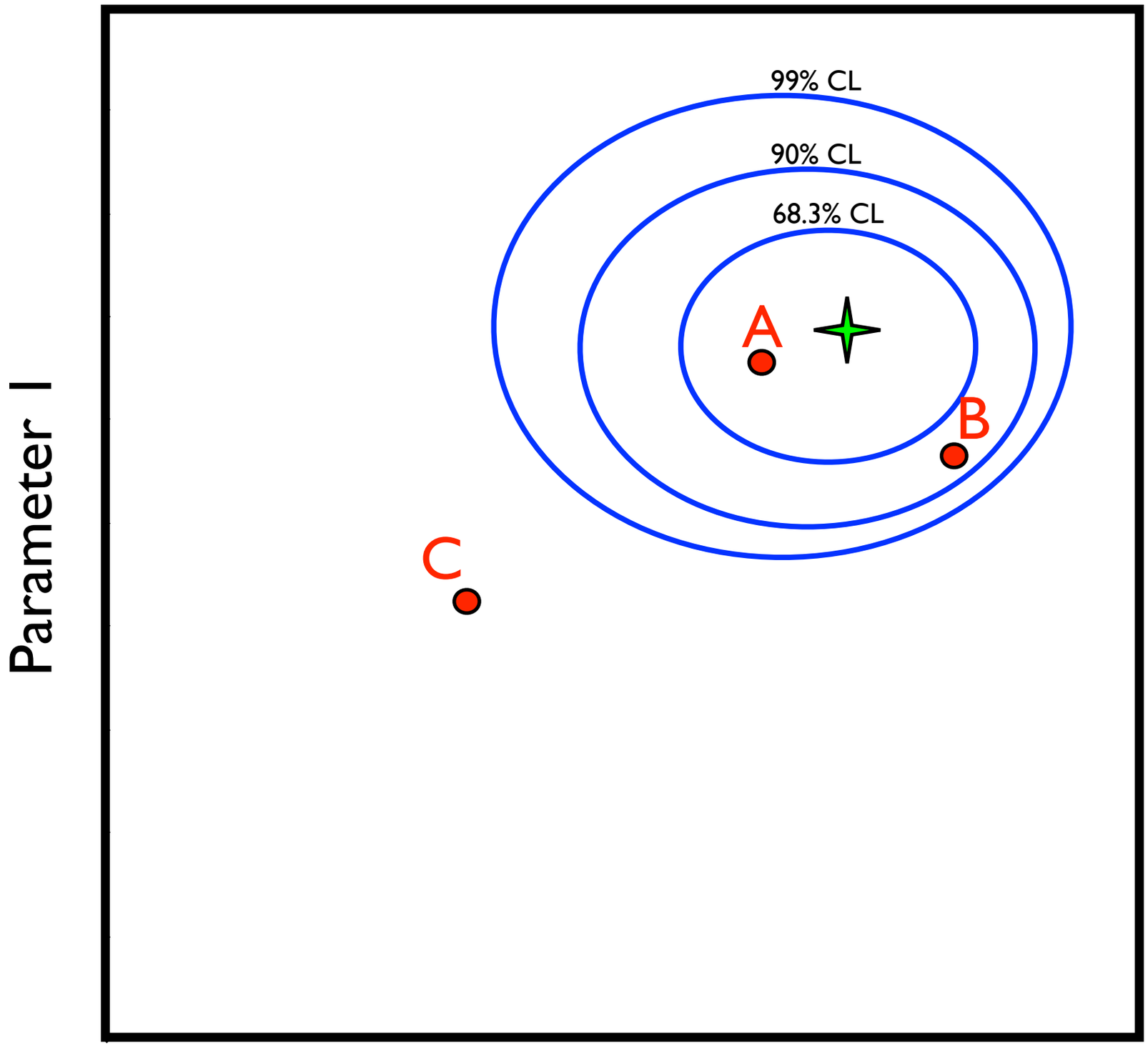
**This gets you nowhere!!**

"A Frequentist uses impeccable logic to answer the wrong question, while a Bayesian answers the right question by making assumptions that nobody can fully believe in." P.G. Hamer

**ALWAYS** ask the right question, even if the answer isn't necessarily straight-forward!!
(this is what being a scientist is all about)


THE SET-UP

Consider a single experiment in which 2 parameters are measured (✦) and compared with predictions from 3 different theoretical models (A, B, C)



# Different Definitions of Probability in Relation to Models:

**Bayesian:**

Degree of belief. Given a single measurement, ascribe "betting odds" to the phase space of possible models. Requires an assumed context for the comparison of these models (prior). There is no relevance to the "statistical coverage of a confidence interval," because there is only one measurement.

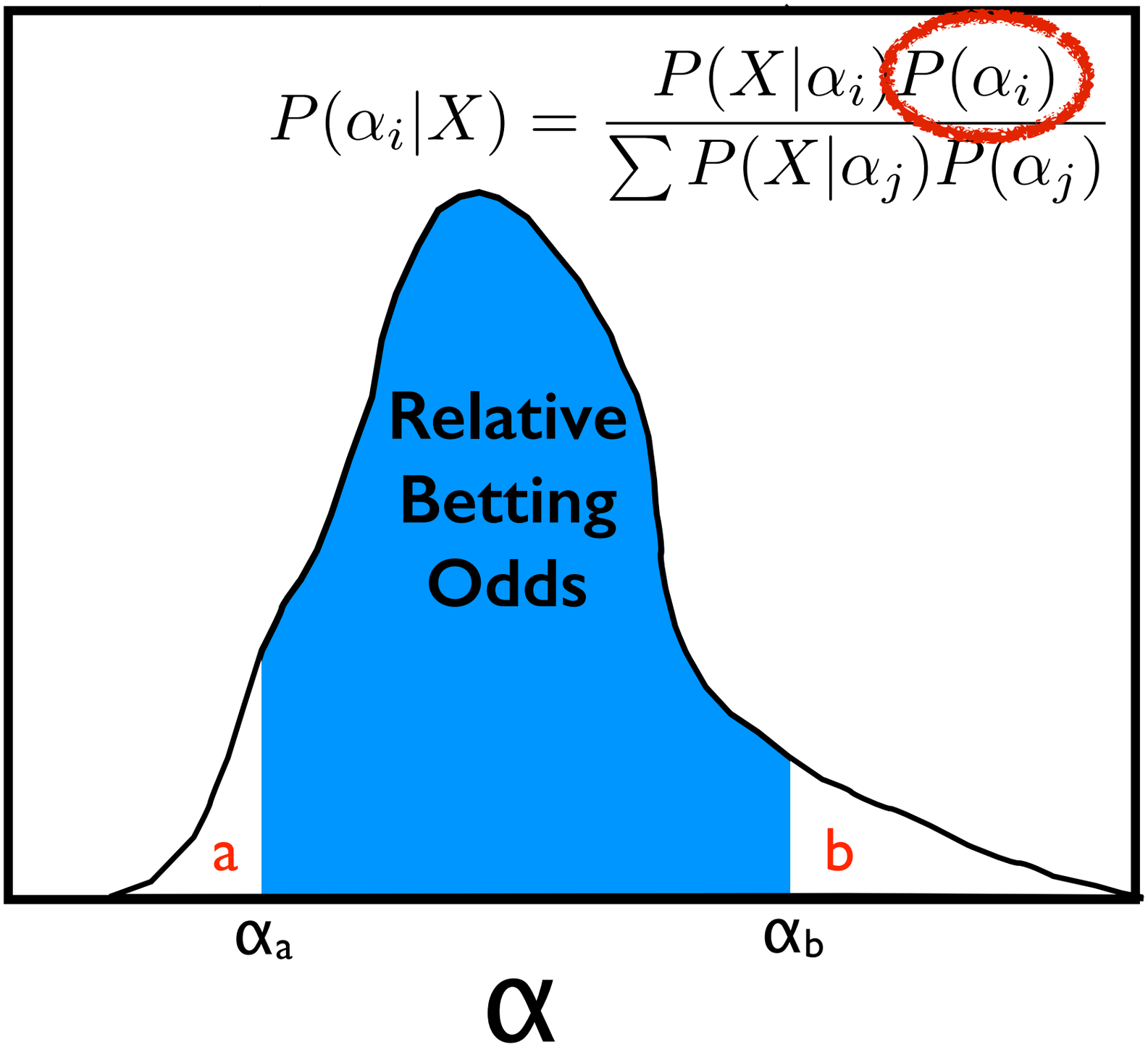
**Frequentist:**

Frequency of occurrence given a hypothetical ensemble of 'identical' experiments. Individual measurements are not used to assess the validity of a model. There is no such thing as a "probability" for a model parameter to lie within derived bounds - either it does or it doesn't. However, if everyone played the same game, the correct model would be bounded a known fraction of the time.

# Bayesian Credible Intervals

PDF for model parameter of interest, "α"

$$P(\alpha_i|X) = \frac{P(X|\alpha_i)P(\alpha_i)}{\sum P(X|\alpha_j)P(\alpha_j)}$$




a = fraction of PDF < $\alpha_a$

b = fraction of PDF > $\alpha_b$

"Credibility"
(fraction contained in the interval):

CI = 1- a - b

Blue and Black

White and Gold

Blue and Black

White and Gold

Context is necessary to relate data to models
(which are of central importance)

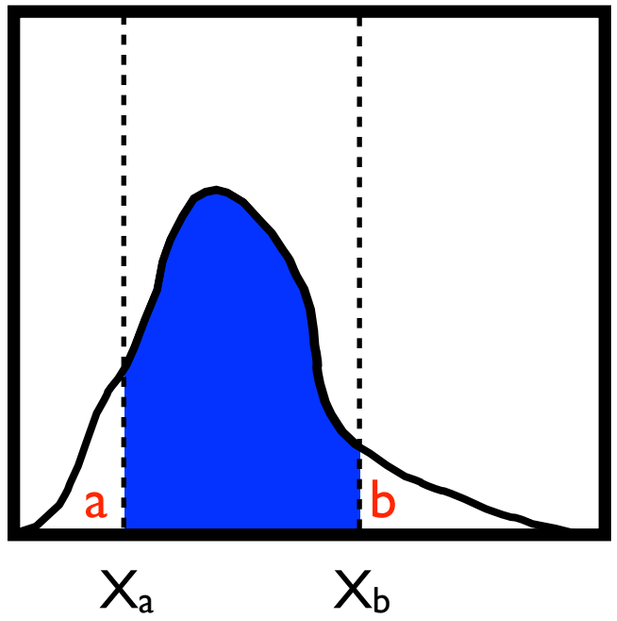
The resulting ambiguity is real!
(*i*.e. an uncertainty that should be conveyed)


THE TALE

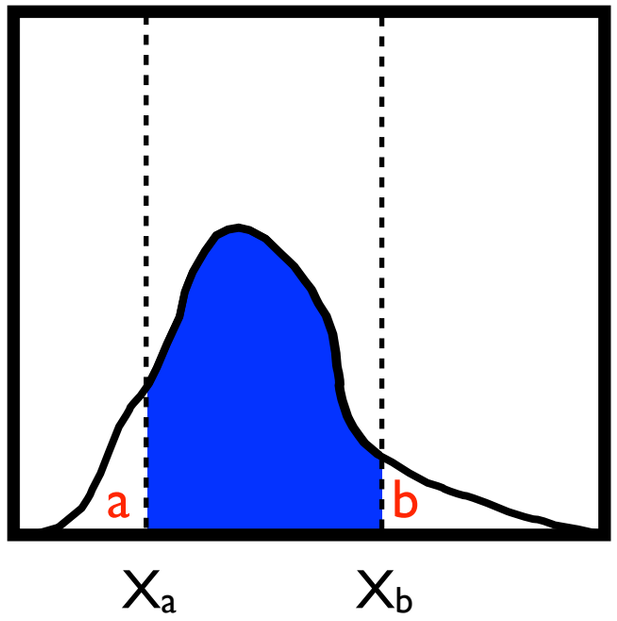
# Neyman Construction of Frequentist Intervals

"Standard"

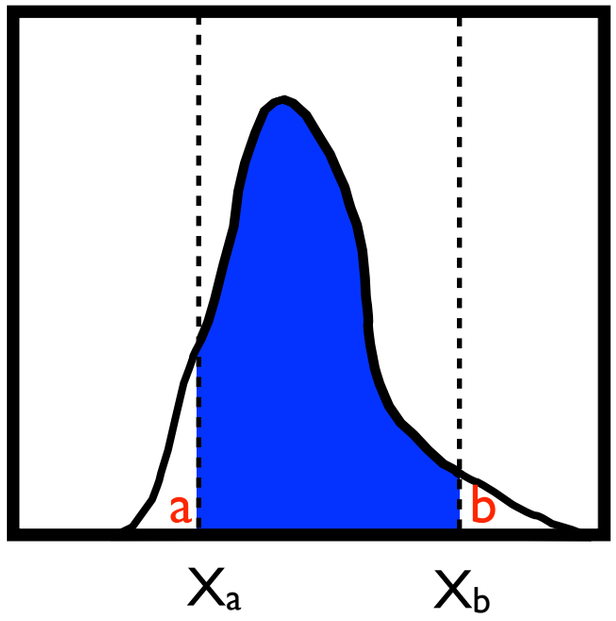
PDF for X assuming the model $\alpha_1$




PDF for X assuming model $\alpha_2$



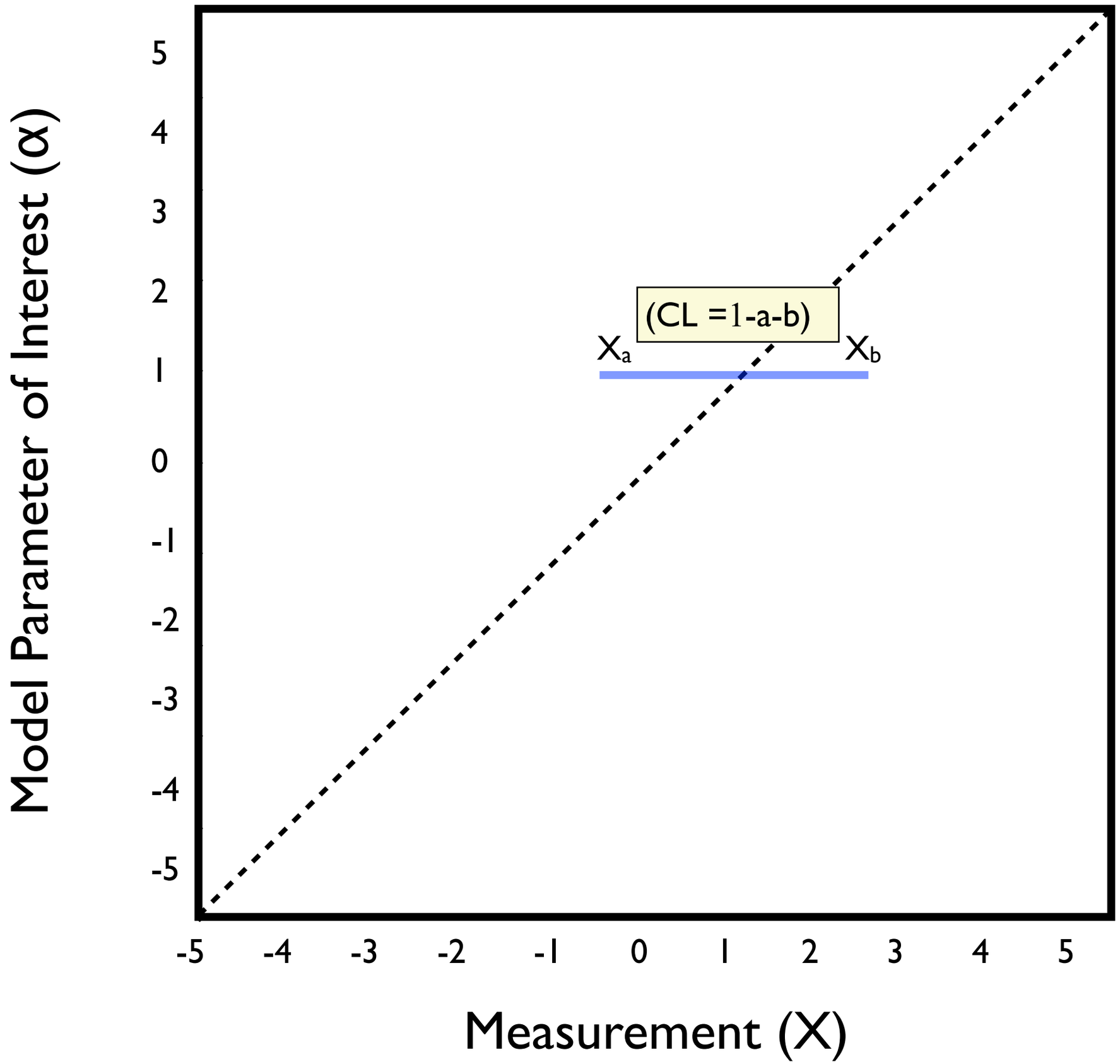

PDF for X assuming model $\alpha_3$



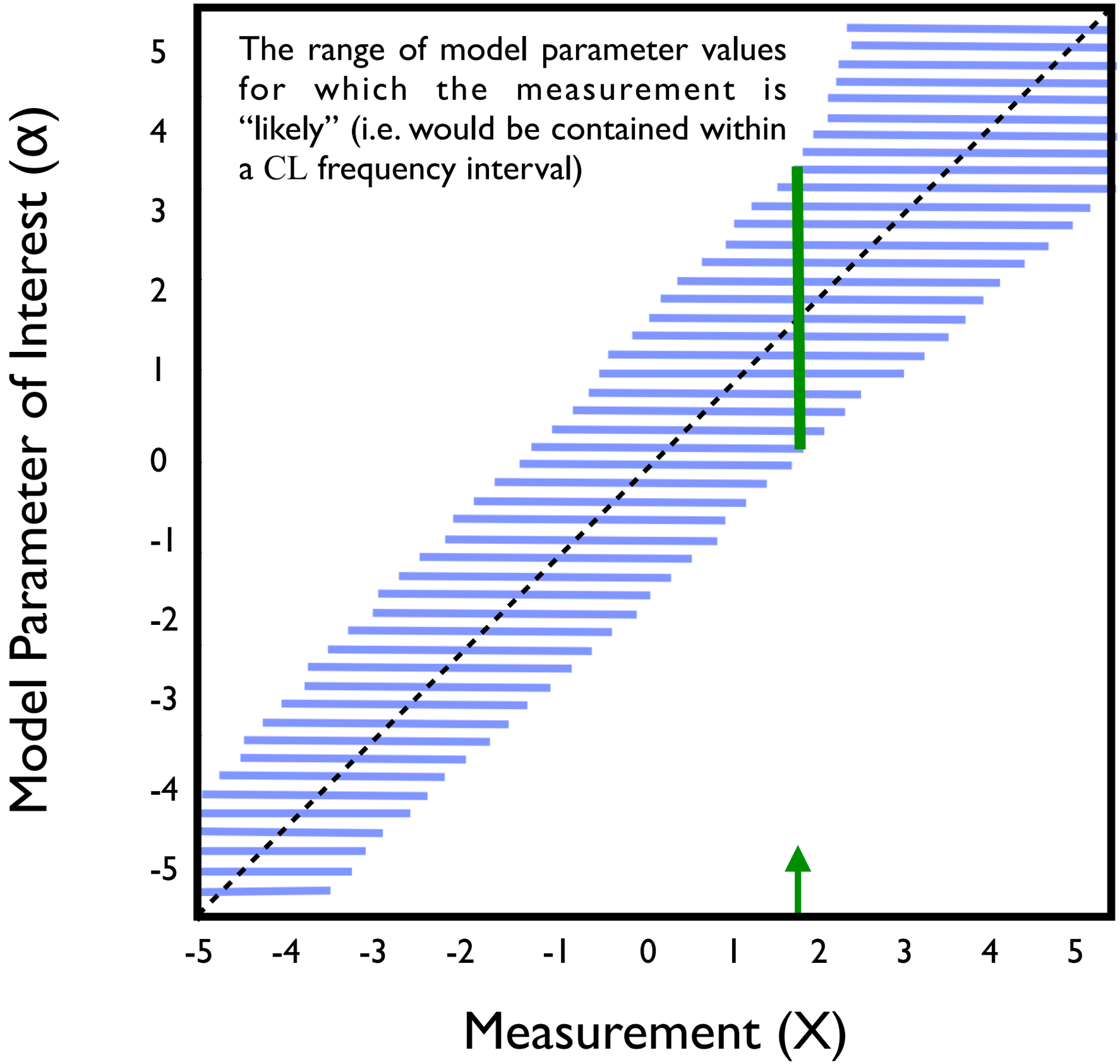

... etc.

$$CL = 1 - a - b$$

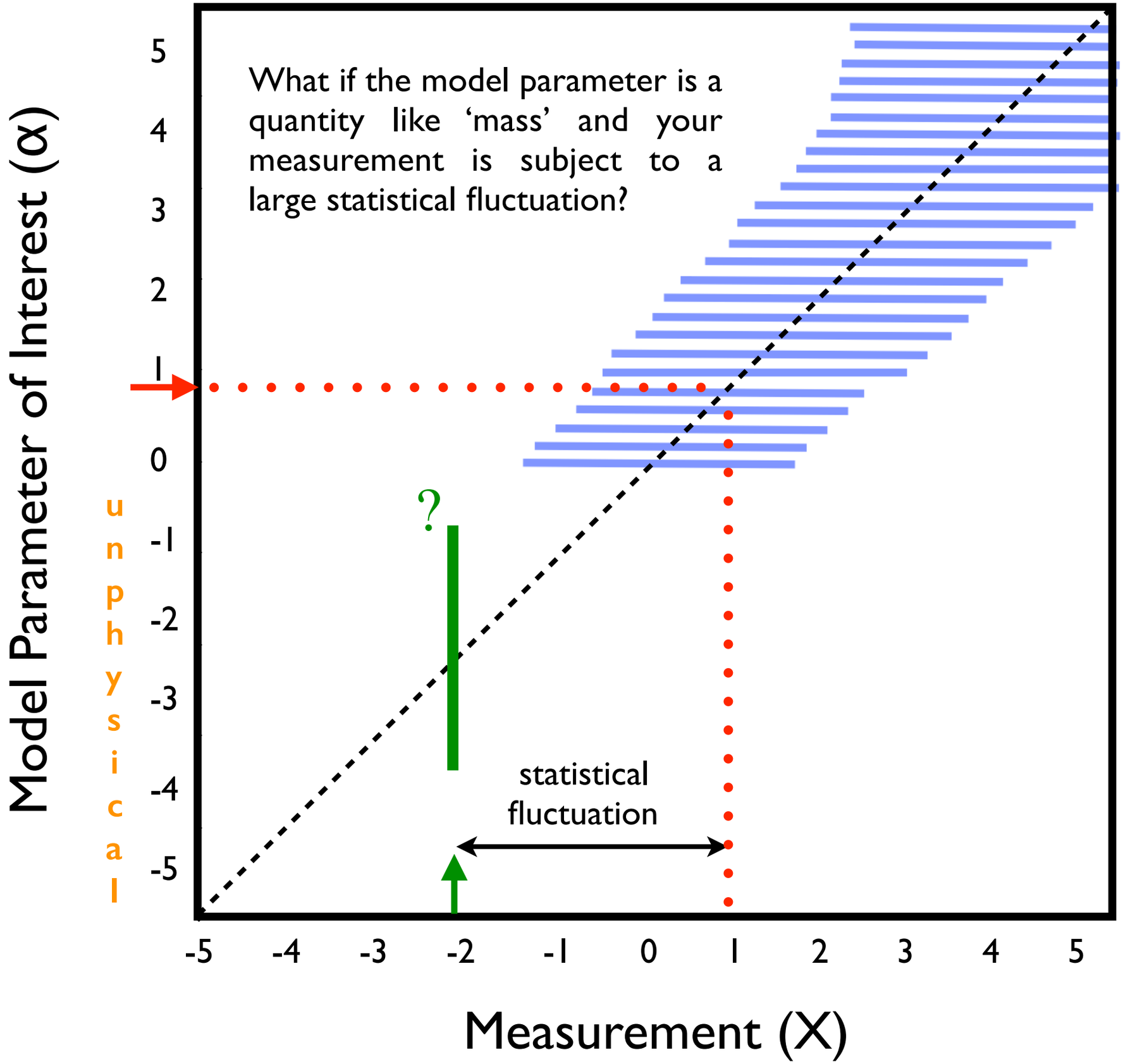
(where "Confidence Level" here refers to the frequency of measurements for a given model)



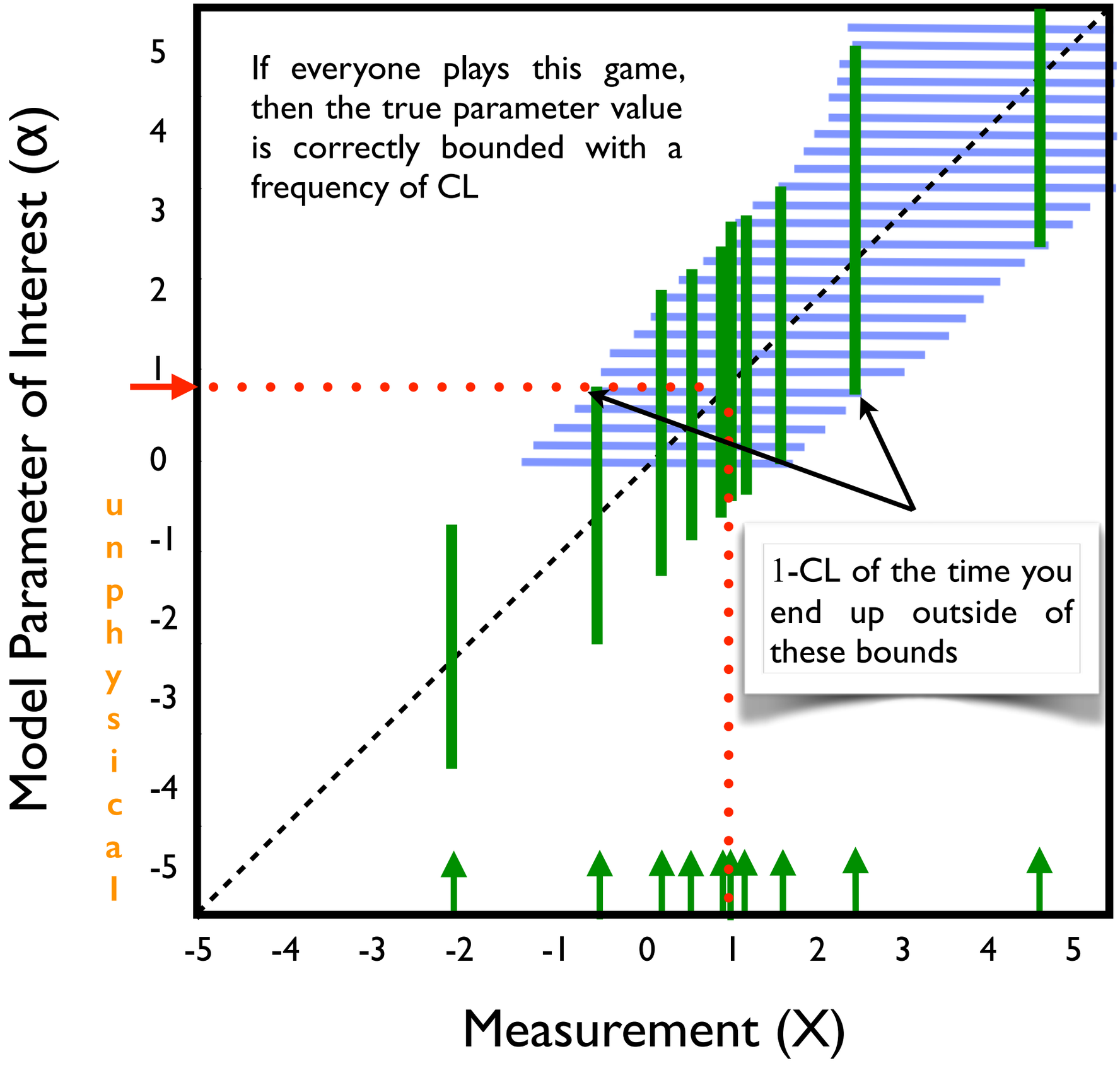

Assume that the measurement X is an unbiased estimator for the model parameter α




Assume that the measurement X is an unbiased estimator for the model parameter α




# What's gone wrong?




**Nothing!**

Frequentists don't care about you, only about the ensemble of many experiments

which are not a problem for true frequentists!

though these are not bounds on physical models, so this appearance is misleading!

To avoid empty intervals and give bounds that appear more physical, Feldman and Cousins proposed a *Unified Approach to the Classical Statistical Analysis of Small Signals* (Phys.Rev.D 57:3873-3889,1998):

which are not a problem for true frequentists!

though these are not bounds on physical models, so this appearance is misleading!

To avoid empty intervals and give bounds that appear more physical, Feldman and Cousins proposed a *Unified Approach to the Classical Statistical Analysis of Small Signals* (Phys.Rev.D 57:3873-3889,1998):

Prescription: The content of the Neyman interval (e.g. where you place the upper and lower boundaries) for an assumed true value of the model parameter, μ, is determined by ordering these according to the likelihood ratio:

$$R = \frac{P(x|\mu)}{P(x|\mu_{best})}$$

where $x$ is the candidate measurement value to be added to the interval and $\mu_{best}$ is the model value that maximises the likelihood for $x$ in the physically allowed region.

Note: this normalisation does not represent a comparison with the "most likely model." For this interval construction, we assume μ is actually the correct model.

The comparison across different models is not intuitive!

*"The difficulties with such an approach are, as before, the lack of a frequency interpretation for (the function that defines the interval content) or, indeed, any direct interpretation for the function."*

(A. Stuart and J.K. Ord, *Kendall's Advanced Theory of Statistics*, Vol. 2, 5th Ed.)

Situations where conventional frequentist intervals are empty are often situations where the Feldman-Cousins procedure returns a value that is prone to be misinterpreted as an unduly strict bound on a model.

Feldman & Cousins:

Accompany each limit by the average expected sensitivity *"in order to provide information that will assist in (a Bayesian) assessment."*

Situations where conventional frequentist intervals are empty are often situations where the Feldman-Cousins procedure returns a value that is prone to be misinterpreted as an unduly strict bound on a model.

Feldman & Cousins:

Accompany each limit by the average expected sensitivity *"in order to provide information that will assist in (a Bayesian) assessment."*

- The expected sensitivity is not a measurement (does not indicate relevance of the observation);
- ***All*** negative fluctuations give smaller values than the expected sensitivity. The approach should **never** be trusted to bound a model (there's no 'threshold');
- No guidelines for how to use this quantitatively to correct or compare with other measurements;

Experiment A looks at 1000 possible exotic branching ratios. The most significant result is 4σ, which becomes ~2σ post-trial.

Experiment A looks at 1000 possible exotic branching ratios. The most significant result is 4σ, which becomes ~2σ post-trial.

Experiment B is less sensitive, but can do a specific test of this hypothesis with no trials, allowing it to obtain a >3σ result if the hypothesis is correct. It sees nothing and, in fact, observes a negative fluctuation below the expected background level.

Experiment A looks at 1000 possible exotic branching ratios. The most significant result is 4σ, which becomes ~2σ post-trial.

Experiment B is less sensitive, but can do a specific test of this hypothesis with no trials, allowing it to obtain a >3σ result if the hypothesis is correct. It sees nothing and, in fact, observes a negative fluctuation below the expected background level.

If the experimental results are compared using 90% CL FC intervals, the value for experiment B in the face of a negative fluctuation can be misinterpreted as providing an inappropriately strict bound on the model. But comparing the expected sensitivity curves would suggest that experiment B is less sensitive than A in terms of constraining the hypothesis!

(A comparison of Bayesian bounds with a common prior would correctly convey the relevant context)

# THE SHUT-OUT

Frequentist intervals are ***NEVER*** statements that there is a particular probability that the true value is bounded by your measurement!


THE STING

90% CL/CI upper bounds on a possible average signal level from a simple counting experiment

| | Initial Test: B=5, n=2 |
|---|---|
| Standard Frequentist | 0.32 |
| Feldman-Cousins | 1.73 |
| Bayesian (prior uniform in rate) | 3.13 |

90% CL/CI upper bounds on a possible average signal level from a simple counting experiment

| | Initial Test: B=5, n=2 |
|---|---|
| Standard Frequentist | 0.32 |
| Feldman-Cousins | 1.73 |
| Bayesian (prior uniform in rate) | 3.13 |

New analysis technique: suppresses backgrounds by a factor of 10 with no loss in signal efficiency!

90% CL/CI upper bounds on a possible average signal level from a simple counting experiment

| | Initial Test: B=5, n=2 | Improved Cuts: B=0.5, n=0 |
|---|---|---|
| Standard Frequentist | 0.32 | |
| Feldman-Cousins | 1.73 | |
| Bayesian (prior uniform in rate) | 3.13 | |

New analysis technique: suppresses backgrounds by a factor of 10 with no loss in signal efficiency!

90% CL/CI upper bounds on a possible average signal level from a simple counting experiment

| | Initial Test: B=5, n=2 | Improved Cuts: B=0.5, n=0 |
|---|---|---|
| Standard Frequentist | 0.32 | 1.8 (worse) |
| Feldman-Cousins | 1.73 | 1.94 (worse) |
| Bayesian (prior uniform in rate) | 3.13 | 2.3 (better) |

New analysis technique: suppresses backgrounds by a factor of 10 with no loss in signal efficiency!

90% CL/CI upper bounds on a possible average signal level from a simple counting experiment

| | Initial Test: B=5, n=2 | Improved Cuts: B=0.5, n=0 |
|---|---|---|
| Standard Frequentist | 0.32 | 1.8 (worse) |
| Feldman-Cousins | 1.73 | 1.94 (worse) |
| Bayesian (prior uniform in rate) | 3.13 | 2.3 (better) |

Can appear to be overly strict bounds on the average signal strength

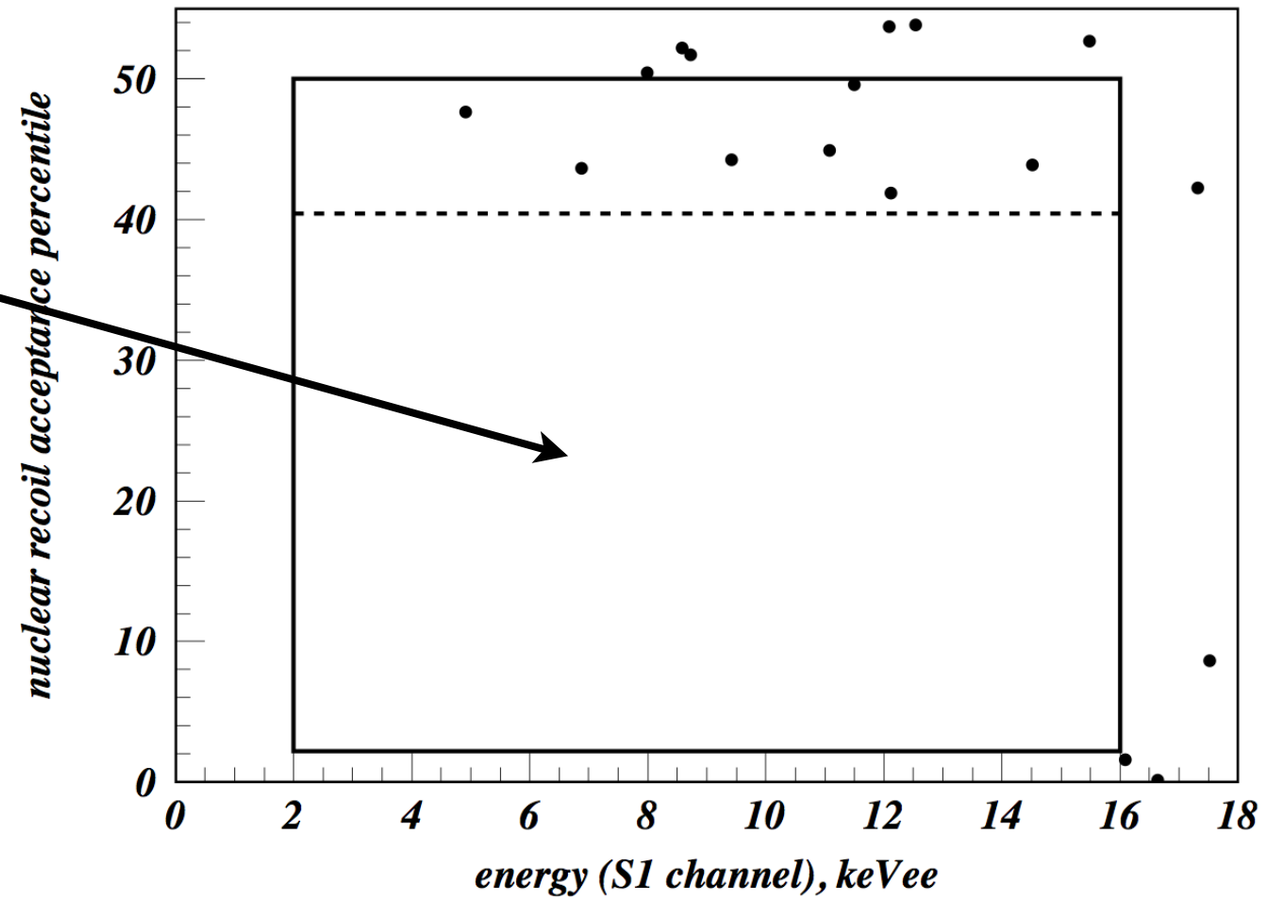
New analysis technique: suppresses backgrounds by a factor of 10 with no loss in signal efficiency!

# ZEPLIN 3

## Dark Matter Search

Limit were based on a region with zero observed events, where the extrapolated non-zero background expectation has a large uncertainty.

(V.N. Lebedenko *et al.*, Phys. Rev. D80 052010 2009)

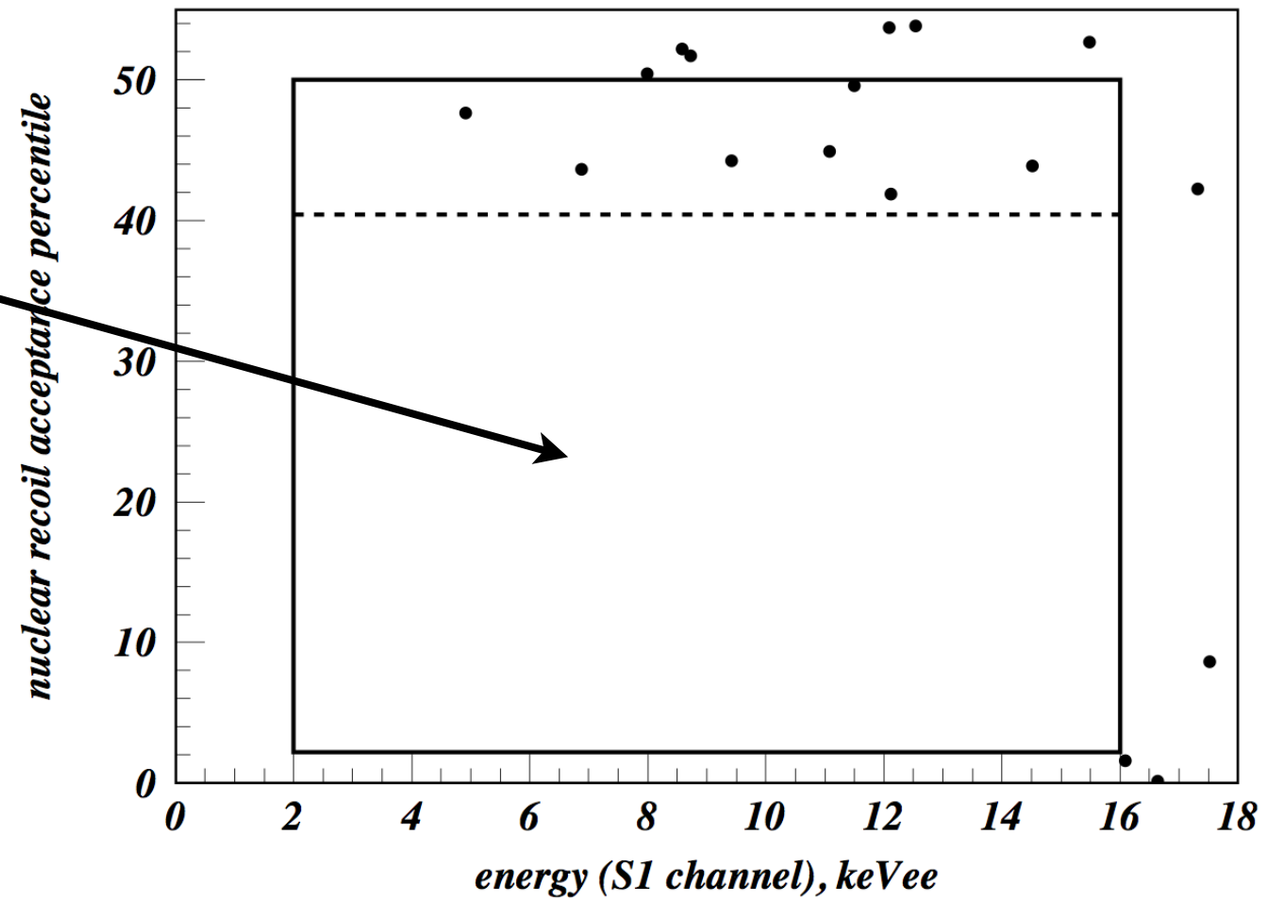


# ZEPLIN 3

## Dark Matter Search

Limit were based on a region with zero observed events, where the extrapolated non-zero background expectation has a large uncertainty.

(V.N. Lebedenko *et al.*, Phys. Rev. D80 052010 2009)




Frequentist bounds for such cases are not well-defined and conventional approaches can, paradoxically, actually lead to **more** restrictive confidence intervals as the background uncertainty gets larger!

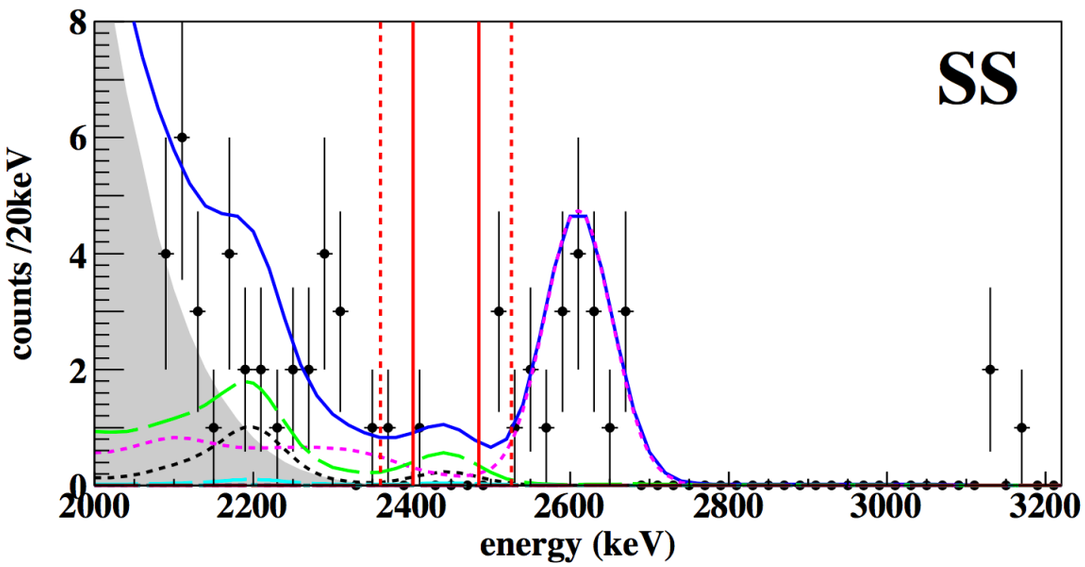
Consequently, authors simply quote the largest possible numerical value for a "Feldman-Cousins-ish" interval.

The Bayesian bound **IS** well-defined here, and using a prior uniform in event rate would give a value numerically comparable to that in the paper.

# EXO 200

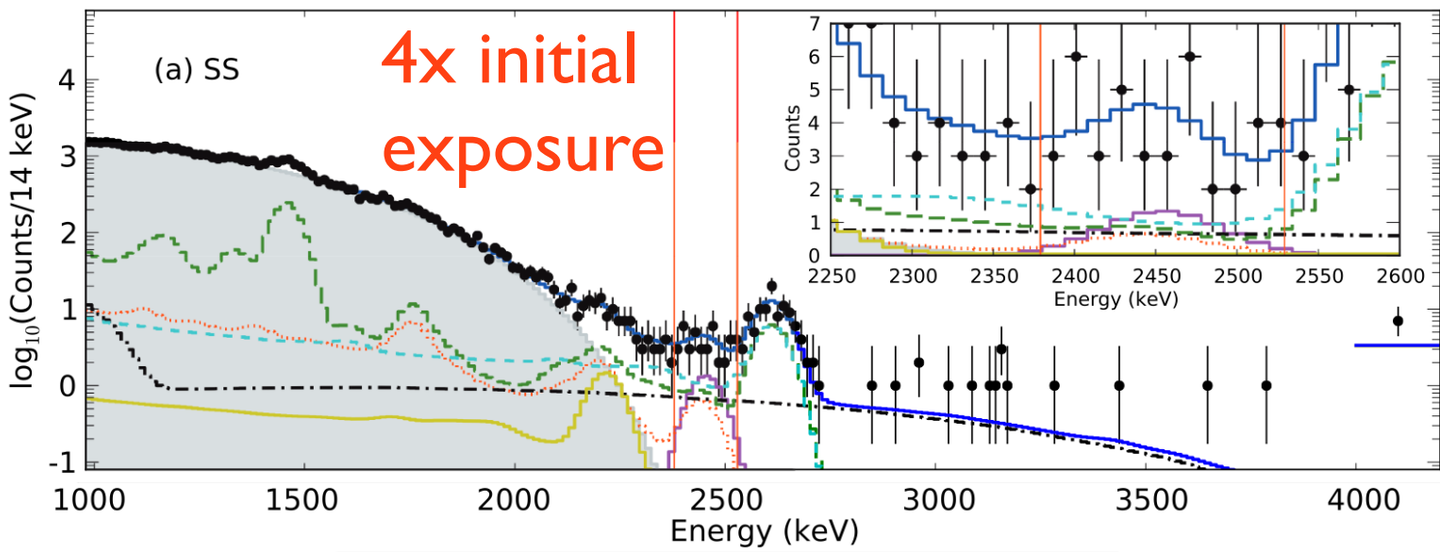
# Neutrinoless Double Beta Decay

(M. Auger *et al.*, Phys. Rev. Lett. **109** 032505, 2012)




±1σ bin: B=4.1±0.3, n=1

(The EXO-200 collaboration, Nature **510**, 229-234, 2014)

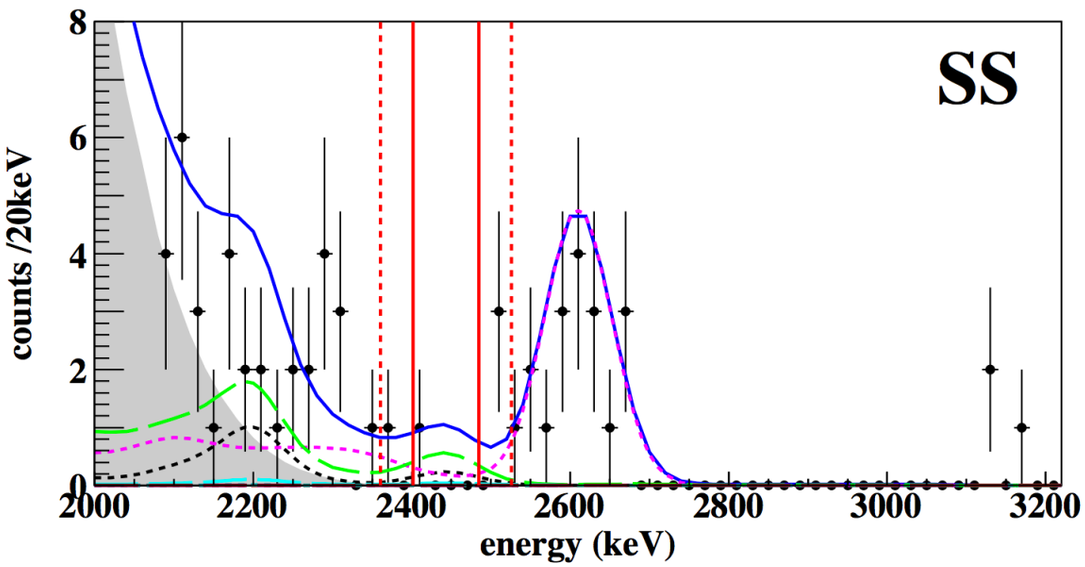



±1σ bin: B=16±2, n=21

# EXO 200

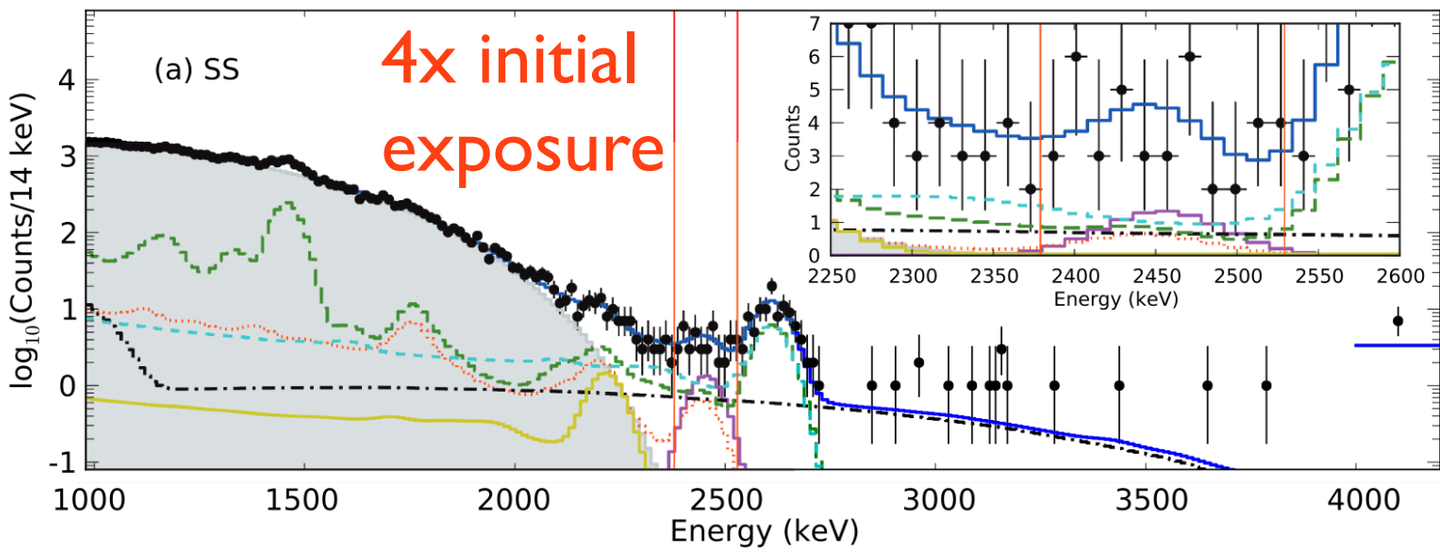
# Neutrinoless Double Beta Decay

(M. Auger *et al.*, Phys. Rev. Lett. **109** 032505, 2012)




±1σ bin: B=4.1±0.3, n=1

(The EXO-200 collaboration, Nature **510**, 229-234, 2014)




±1σ bin: B=16±2, n=21

90% CL/CI upper bounds on $T_{1/2}$ (units of $10^{25}$ yrs)

| | Spectrum + Wilks' | F-C ($\pm1\sigma$ bin) | Bayesian ($\pm1\sigma$ bin) |
|---|---|---|---|
| Data Set 1 | >1.6 | >2.2 | >1.1 |
| Data Set 2 | >1.1 | >1.3 | >1.4 (integration: >1.2) |

For larger data set: 45% worse | 70% worse | ~10-20% better

# Search for Neutrinoless Double-Beta Decay of $^{130}$Te with CUORE-0

(Dated: April 10, 2015)

We report the results of a search for neutrinoless double-beta decay in a 9.8 kg·yr exposure of $^{130}$Te using a bolometric detector array, CUORE-0. The characteristic detector energy resolution and background level in the region of interest are $5.1 \pm 0.3$ keV FWHM and $0.058 \pm 0.004$ (stat.) $\pm$ 0.002 (syst.) counts/(keV·kg·yr), respectively. The median 90 % C.L. lower-limit sensitivity of the experiment is $2.9 \times 10^{24}$ yr and surpasses the sensitivity of previous searches. We find no evidence for neutrinoless double-beta decay of $^{130}$Te and place a Bayesian lower bound on the decay half-life, $T^{0\nu}_{1/2} > 2.7 \times 10^{24}$ yr at 90 % C.L. Combining CUORE-0 data with the 19.75 kg·yr exposure of $^{130}$Te from the Cuoricino experiment we obtain $T^{0\nu}_{1/2} > 4.0 \times 10^{24}$ yr at 90 % C.L. (Bayesian), the most stringent limit to date on this half-life. Using a range of nuclear matrix element estimates we interpret this as a limit on the effective Majorana neutrino mass, $m_{\beta\beta} < 270 - 760$ meV.

# Kamland-Zen II

(A. Gando *et al.*, arXive:1605.02889v1, 2016)

## Phase II 90% CL Limit:

$$T^{0\nu}_{1/2} > 9.6 \times 10^{25}\,\mathrm{yr}$$

*"A MC of an ensemble of experiments assuming the best-fit background spectrum without a 0νββ signal indicates a sensitivity of $4.9 \times 10^{25}$yr, and the probability of obtaining a limit stronger than the presented result is 11%.*

# Kamland-Zen II

(A. Gando *et al.*, arXive:1605.02889v1, 2016)

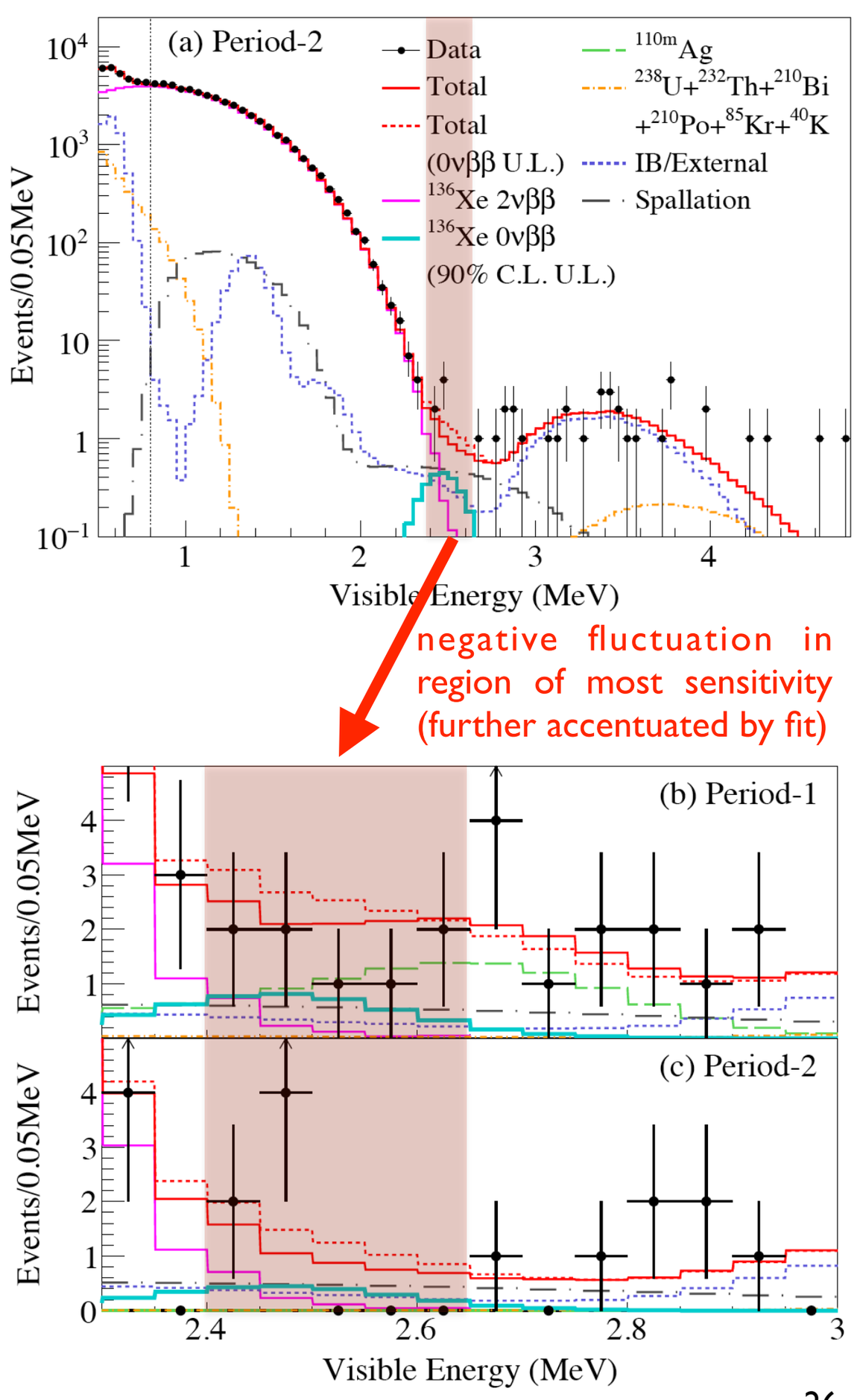
## Phase II 90% CL Limit:

$$T^{0\nu}_{1/2} > 9.6 \times 10^{25}\,\mathrm{yr}$$

*"A MC of an ensemble of experiments assuming the best-fit background spectrum without a 0νββ signal indicates a sensitivity of $4.9 \times 10^{25}$ yr, and the probability of obtaining a limit stronger than the presented result is 11%.*

**Not a bound on the physical (model) lifetime!**

numerical value will not be robust to further data taking and improvements; comparisons with others will be misleading.



# Comparison of Experimental Results

Frequentism only cares about the ensemble, so if your data set has fluctuated away from expectation in term of background etc., the formalism is less concerned with getting the context right for that "unlikely" occurrence.

Frequentist bounds are not a robust basis for the comparison of experimental results! Comparing Bayesian bounds using a common prior tends to be much better in this regard.

# Propagation of Experimental Uncertainties

There is no rigorous, self-consistent method to propagate uncertainties in the frequentist paradigm! Typical compromise approaches involve Bayesian integration over parts of the phase space, treating all systematic uncertainties as having a statistical origin, and sacrificing the strict definition of statistical coverage.

Instead of seeking compromises and work-arounds, perhaps we should be asking why such steps are necessary in the first place! An inconsistent framework may suggest inconsistent thinking that is prone to result in problems.

# The "Flip-Flop"

If experimenters choose for themselves when to quote a given type of interval based on the result, this can lead to a small statistical bias in frequentist coverage.

The F-C "Unified Method" automatically transitions from a 1-sided to a 2-sided bound for a given CL

# The "Flip-Flop"

**(Only an issue for frequentist intervals!)**

If experimenters choose for themselves when to quote a given type of interval based on the result, this can lead to a small statistical bias in frequentist coverage.

**<u>Worst case (at borderline of CL):</u>**
**a 90% CL might only have 85% coverage;**
**a 99% CL might only have 98.5% coverage**

(concern over unfiltered surveys of largely borderline results)

The F-C "Unified Method" automatically transitions from a 1-sided to a 2-sided bound for a given CL

• Conflicts with scientifically well-motivated convention to quote 90% or 95% CL upper/lower bounds for results consistent with the null hypothesis, but only claim a 2-sided discovery interval when the null hypothesis is rejected at a considerably higher confidence level;

• Can't easily cope with look-elsewhere effects (trials factors): Search for gamma-ray emission from 1000 different astrophysical sources results in no event excess above 3σ, consistent with statistical fluctuations. Most appropriate to quote upper bounds on the possible emission from each source, but unified approach forces 3σ detection interval;

• Even for a clear detection, it may still be relevant to also quote upper and lower bounds in the context of different models. Different interval constructions can be simultaneously valid and relevant for the same results, they simply address different questions.

# Towards a More Transparent and Relevant Approach

# Four Typical Goals in Data Presentation

**1)** Present the measured value of a direct observable and an assessment of systematic uncertainties that could bias the measurement. Straightforward presentation of 'raw' observation.

**2)** Address "How often would a measurement `like mine' occur under the null hypothesis?" GoF consistency test (p-value etc.)

**3)** Address "What constraints do my measurements of direct observables place on models?" Bayesian

**4)** Objectively convey the relevant information content of the data to allow the impact of alternative assumptions to be evaluated, facilitate the testing of different models, and permit information to be effectively combined with that from other experiments. Likelihood

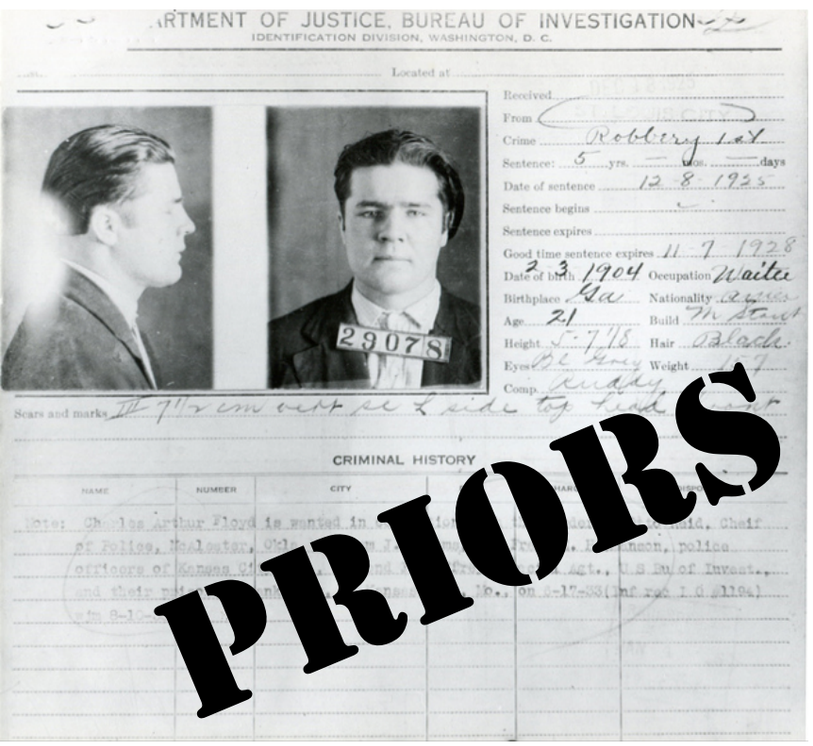
# Four Typical Goals in Data Presentation

**1)** Present the measured value of a direct observable and an assessment of systematic uncertainties that could bias the measurement. Straightforward presentation of 'raw' observation.

**2)** Address "How often would a measurement `like mine' occur under the null hypothesis?" GoF consistency test (p-value etc.)

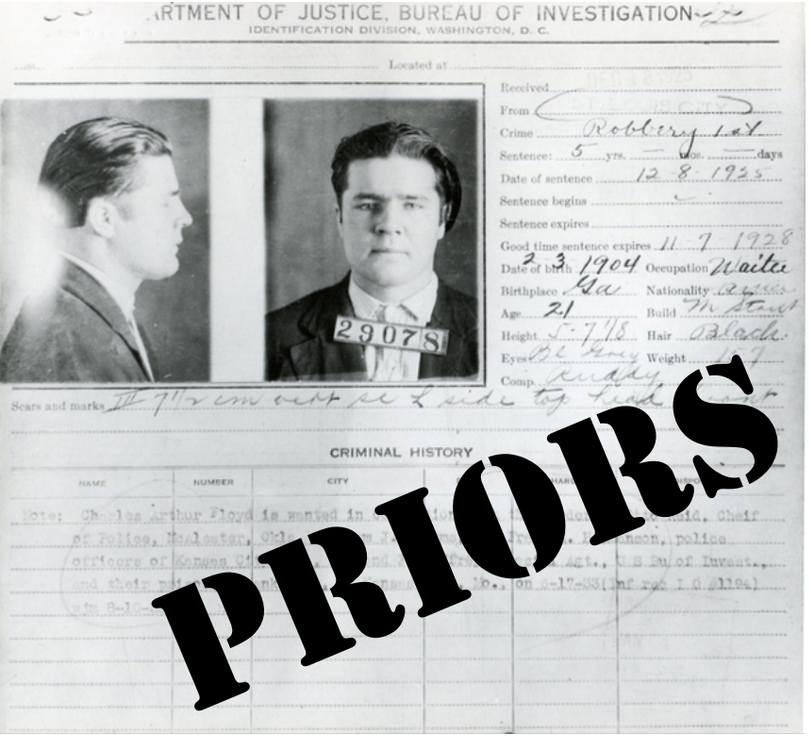
**3)** Address "What constraints do my measurements of direct observables place on models?" Bayesian

**4)** Objectively convey the relevant information content of the data to allow the impact of alternative assumptions to be evaluated, facilitate the testing of different models, and permit information to be effectively combined with that from other experiments. Likelihood




# 1. Informative:

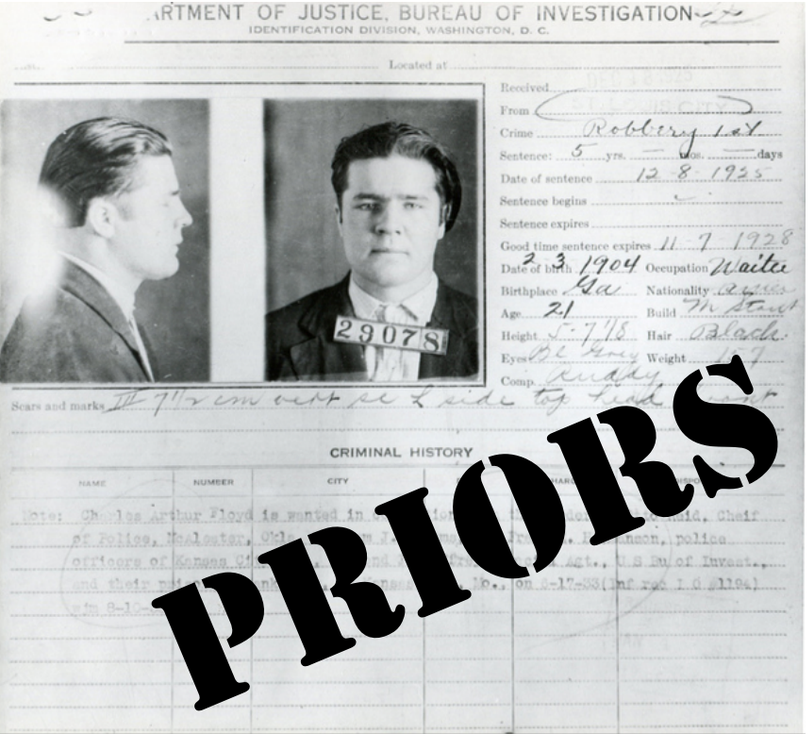
Permits known, physical constraints to be imposed (*e.g.* energies and masses must be greater than zero; the position of observed events must be inside the detector, *etc.*) and allows known attributes of the physical system to be taken into account (*e.g.* energies are being sampled from a particular spectrum; the relative probabilities for different event classes are drawn from a given distribution, *etc.*).




# 2. Non-Informative:

When the relative weighting of models is not obvious, but you still need to provide a context for comparison

(A Case of Too Much History?)




# 2. Non-Informative:

When the relative weighting of models is not obvious, but you still need to provide a context for comparison

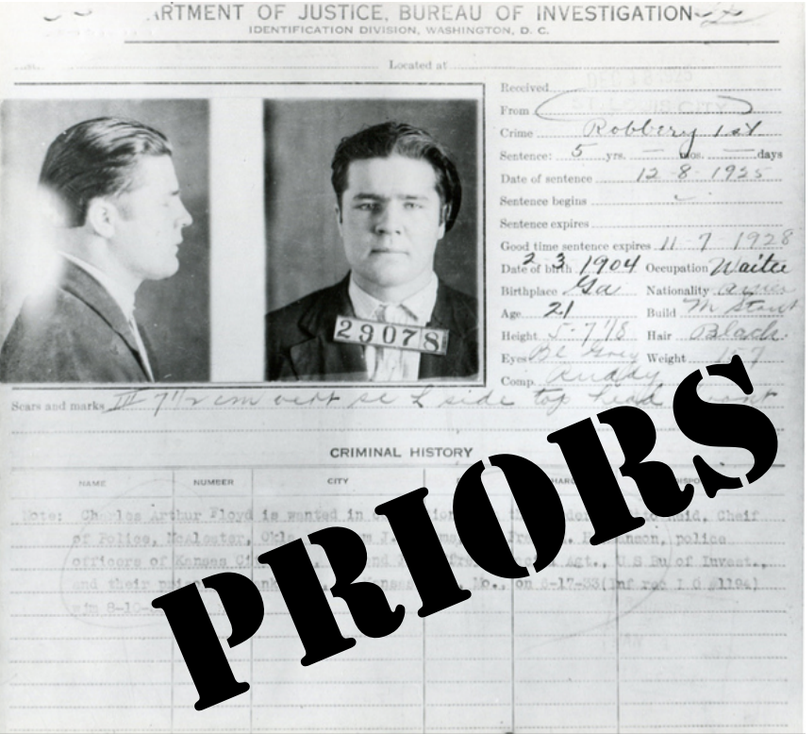
(A Case of Too Much History?)

## Very Briefly:

**"Principle of Indifference":** equal probabilities for outcomes believed to be equally likely.

**Jeffreys Priors:** use likelihood itself to weight according to Fisher Information
(transformationally invariant "objective Bayesian")

**Reference Priors**: maximise divergence between posterior and prior
(alternative "objective Bayesian")




# 2. Non-Informative:

When the relative weighting of models is not obvious, but you still need to provide a context for comparison

(A Case of Too Much History?)

## Very Briefly:

**"Principle of Indifference":** equal probabilities for outcomes believed to be equally likely.

**Not "objective" or necessarily transformationally invariant**

**Jeffreys Priors:** use likelihood itself to weight according to Fisher Information

(transformationally invariant "objective Bayesian")

**Form is not always intuitive; disconcertingly depends on exact data set & analysis procedure; can be complex and not easy to compute for real-world multi-dimensional analyses; doesn't always have good convergence behaviour;** ***(also violates strong version of Likelihood principle)***

**Reference Priors**: maximise divergence between posterior and prior

(alternative "objective Bayesian")

**Similar issues as for Jeffreys plus not necessarily transformationally invariant, though can be easier to compute and often better behaved**

# What's the way out??

**Pragmatism!**

**There is no "correct" choice of prior!**

# What's the way out??

**Pragmatism!**

**There is no "correct" choice of prior!**

- Where possible, use informative priors or follow standard conventions (if they exist);
- Otherwise, choose simple prior forms that are easy to understand and visualise (*e.g.* uniform);
- Use common (*i.e.* standardised) parameter choices that "make sense" for these priors;
- If there's an ambiguity that leads to a non-conservative bound, show the sensitivity to the choice of prior!

- If there's an ambiguity that leads to a non-conservative bound, show the sensitivity to the choice of prior!

*The ambiguity due to the choice of prior is **real**! Choosing a frequentist construction does not avoid the ambiguity, but merely hides it, potentially leading to false conclusions regarding the robustness of implications for models.*

Science & Environment

# Cosmic inflation: 'Spectacular' discovery hailed

By Jonathan Amos
Science correspondent, BBC News

17 March 2014 | Science & Environment

**Scientists say they have extraordinary new evidence to support a Big Bang Theory for the origin of the Universe.**

Researchers believe they have found the signal left in the sky by the super-rapid expansion of space that must have occurred just fractions of a second after everything came into being.

It takes the form of a distinctive twist in the oldest light detectable with telescopes.

**The work will be scrutinised carefully**, but already there is talk of a Nobel.

"This is spectacular," commented Prof Marc Kamionkowski, from Johns Hopkins University.

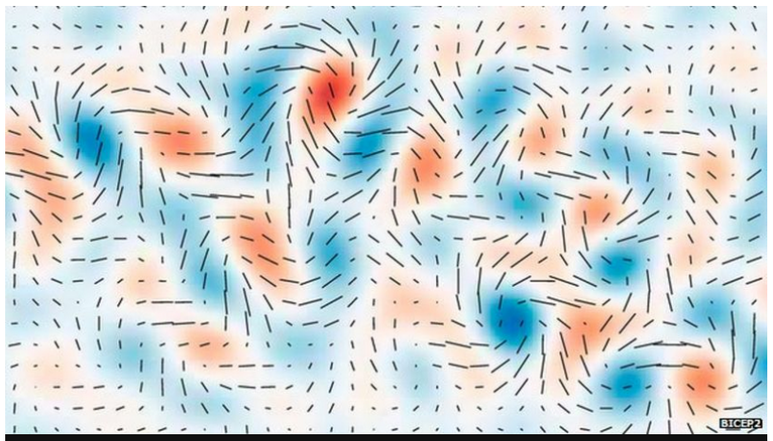
"I've seen the research; the arguments are persuasive, and the scientists involved are among the most careful and conservative people I know," he told BBC News.

> **Nature did not have to be so kind and the theory didn't have to be right**
> Prof Alan Guth, Inflation pioneer

The breakthrough was announced by an American team working on a project known as **BICEP2**.

This has been using a telescope at the South Pole to make detailed observations of a small patch of sky.

The aim has been to try to find a residual marker for "inflation" - the idea that the cosmos experienced an exponential growth spurt in its first trillionth, of a trillionth of a trillionth of a second.



Gravitational waves from inflation put a distinctive twist pattern in the polarisation of the CMB

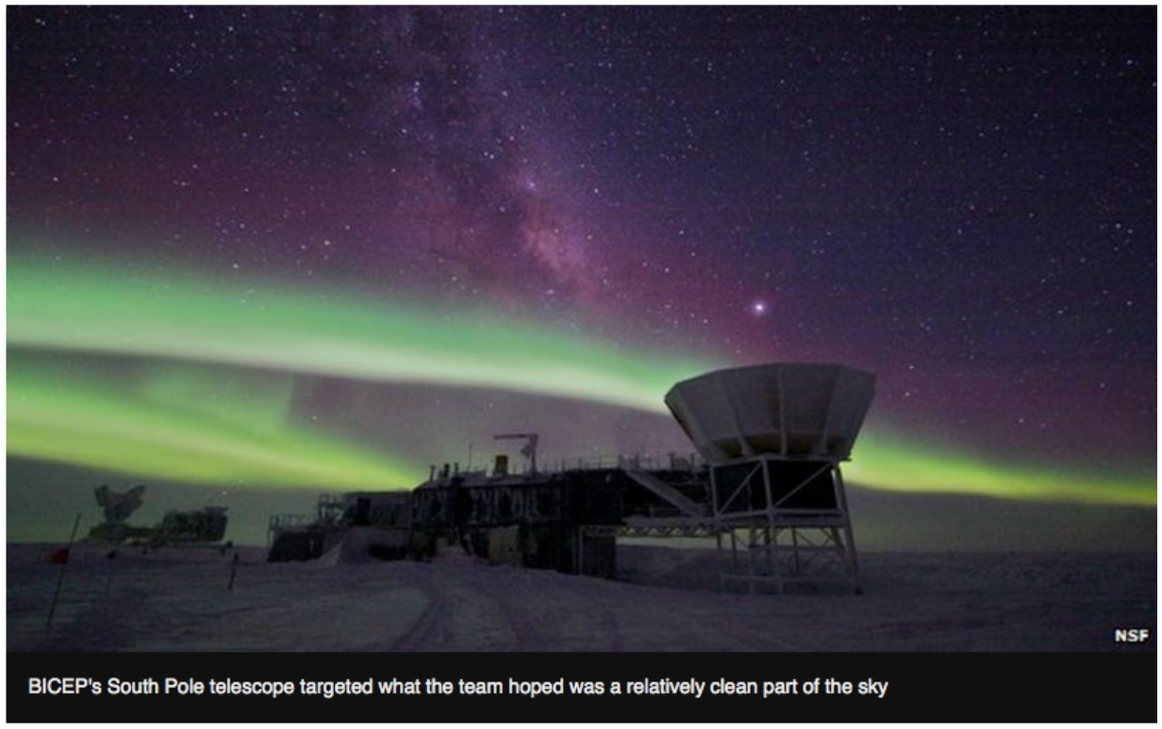


Science & Environment

# Cosmic inflation: BICEP 'underestimated' dust problem

By Jonathan Amos
Science correspondent, BBC News

22 September 2014 | Science & Environment



BICEP's South Pole telescope targeted what the team hoped was a relatively clean part of the sky

**One of the biggest scientific claims of the year has received another set-back.**

# Strategy in the Absence of Clear Guidance:
# Start with intuition - typically 2 types of parameters:

## Variables of Magnitude

~equally likely values spanning same general order of magnitude

(e.g. *unknown phase angle; the value of a spectral index; precision measurement of a quantity whose rough magnitude is constrained)*

⇒ prior uniform in the magnitude of the quantity

## Variables of Scale

values ~equally likely to span several orders of magnitude

(e.g. *energy scale for new physics, cross section for some non-standard interaction; first measurement of a quantity whose rough magnitude is not constrained.)*

⇒ prior uniform in the log of the quantity

**These tend to bound the range of "plausible" priors**

# More Pragmatism:

Unless a clear physical argument exists, for parameters proportional to a counting rate, choose a prior that is uniform in the magnitude

- Observable rates for a newly discovered process are unlikely to span many orders of magnitude;
- Upper bounds for a non-observation will be conservative;
- For simplest case of Poisson statistics, limits also provide good statistical coverage (convenient even if not essential)

# More Pragmatism:

Unless a clear physical argument exists, for parameters proportional to a counting rate, choose a prior that is uniform in the magnitude

- Observable rates for a newly discovered process are unlikely to span many orders of magnitude;
- Upper bounds for a non-observation will be conservative;
- For simplest case of Poisson statistics, limits also provide good statistical coverage (convenient even if not essential)

## Improper Priors?

*Integral not finite,*
*Weights S=1 the same as S=$10^{10}$*

Likelihood function kills impact of prior far from ROI, so precise form there doesn't matter (uniform = approximation to a prior that tails off in a manner that doesn't need to be specified).

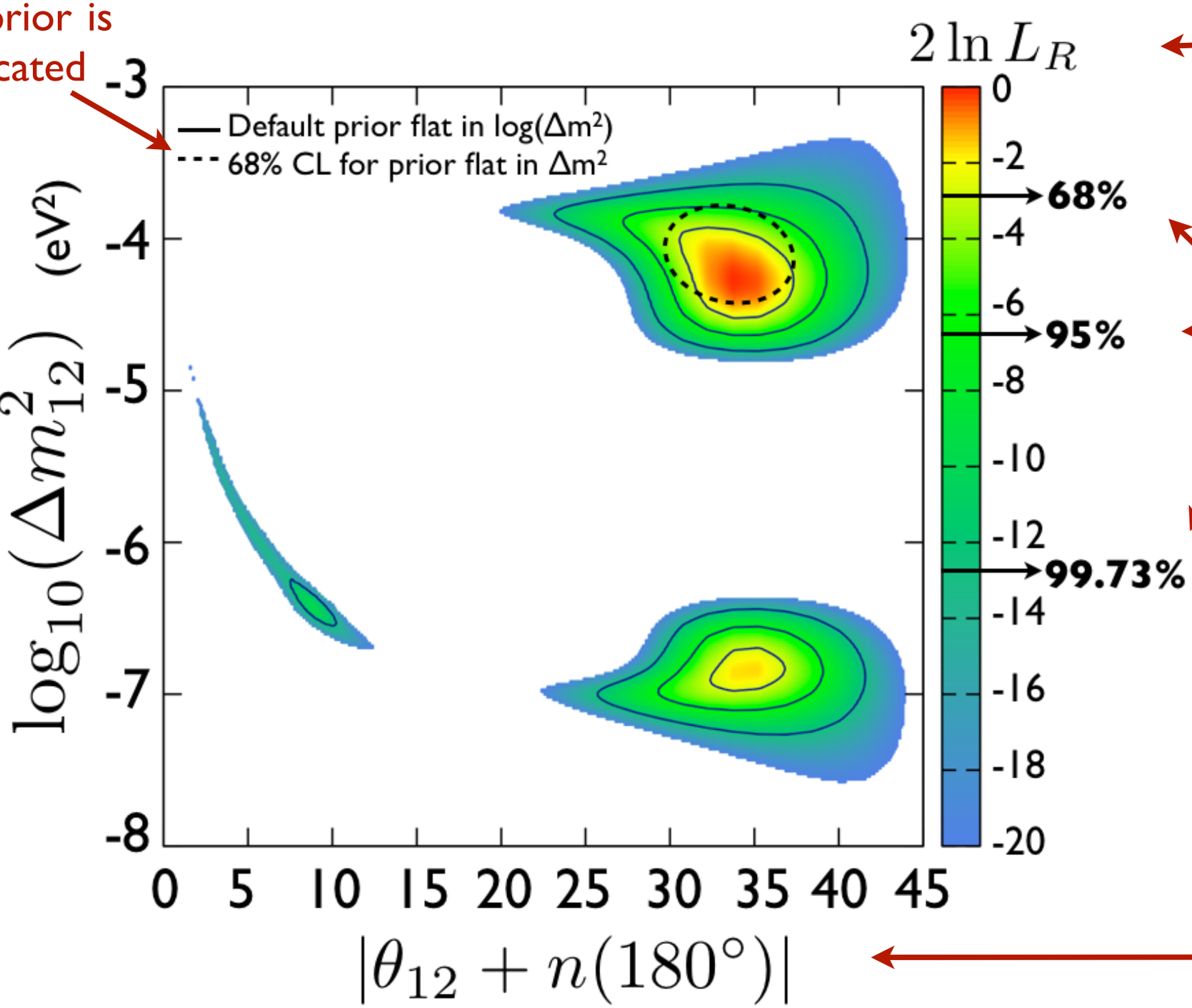
# Cashing In!


The display of Bayesian constraints in terms of "flat" variable forms (*i.e.* forms for which a uniform prior "make sense") simultaneously provides objective Likelihood contours from which alternative prior choices can be applied or frequentist bounds can be derived via Wilks (if you still really want to do that!)

# Example of "Unified" Likelihood Map: SNO salt phase solar ν data

(using publicly available data associated with Phys. Rev. Lett. **101**, 111301, 2008)

Sensitivity to prior is indicated

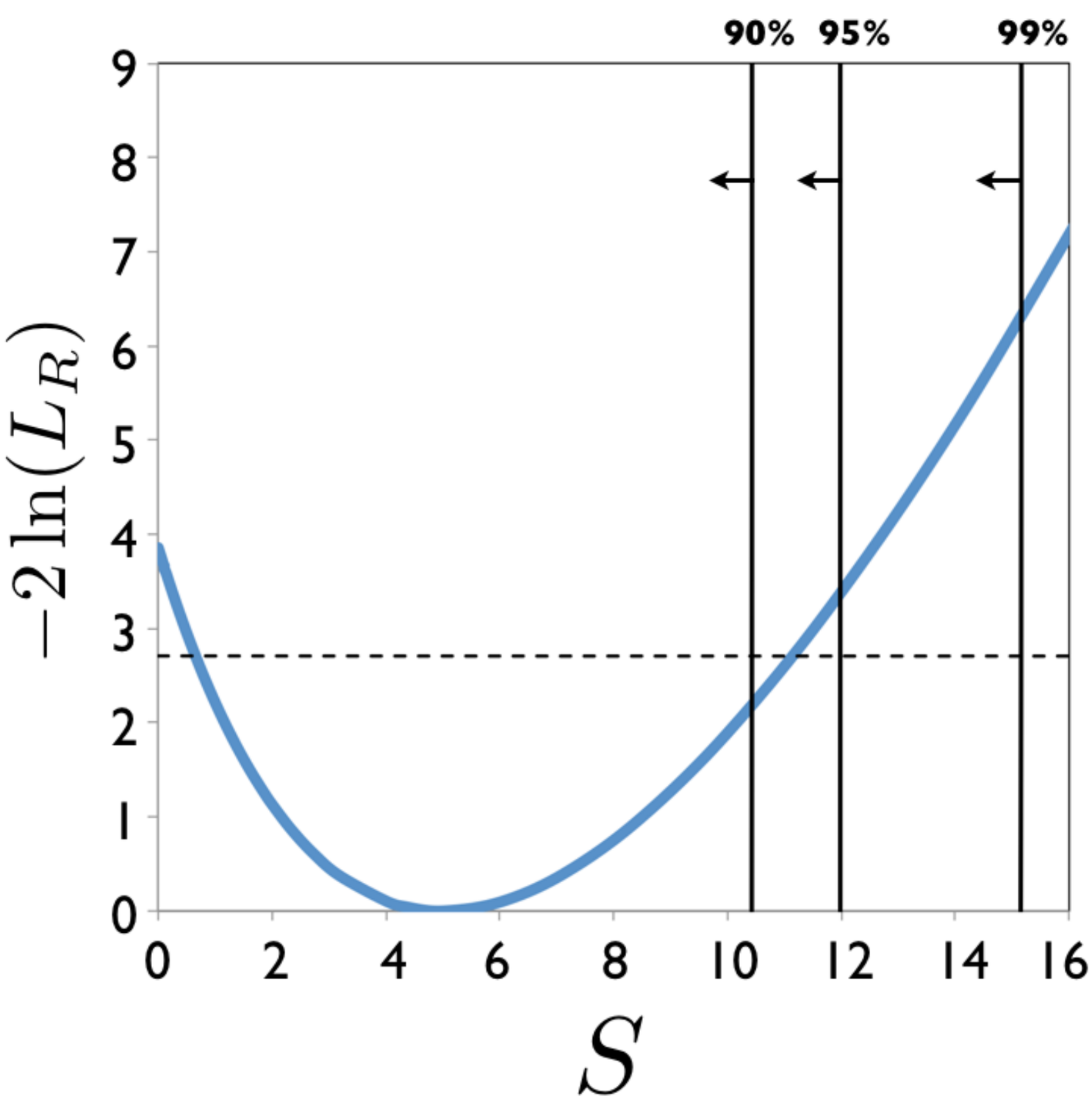



Approximate $\Delta\chi^2$ value from Wilks

Bayesian contours from integration of likelihood assuming priors uniform in θ and log(Δm²)

Form for fundamental angle accounts for quadrant ambiguity

# Example 2: Rare Event Search Counting Experiment (B=5, n=10)

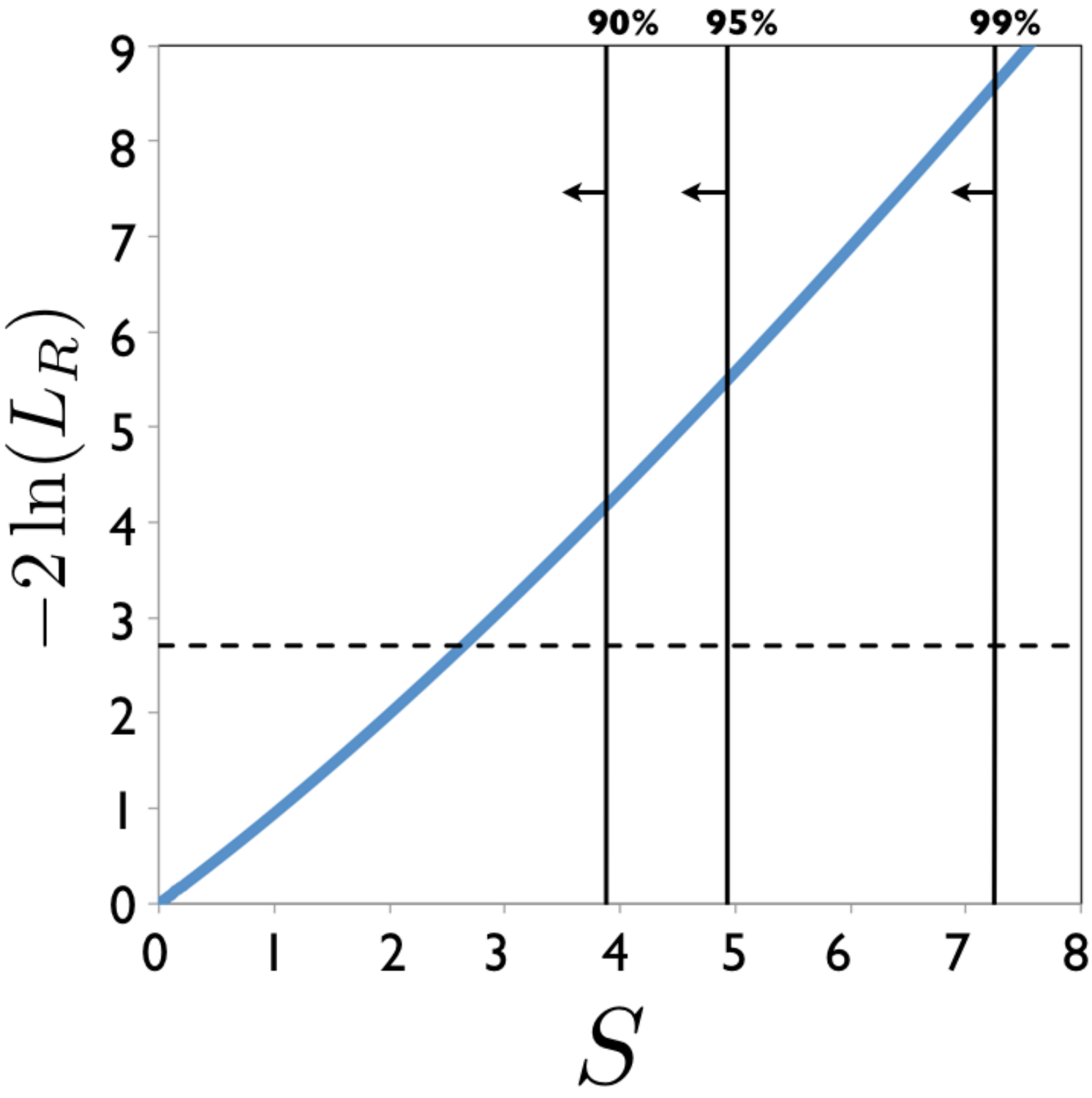



Bayesian upper bounds from integration of likelihood assuming priors uniform in S ('conservative')

Approximate 2-sided 90% CL frequentist bound derived from Wilks' theorem (comparable to FC)

S < 10.4(15.2) at 90%(99%) CI

# Example 3: Rare Event Search Counting Experiment (B=9, n=5)




Bayesian upper bounds from integration of likelihood assuming priors uniform in S ('conservative')

Approximate 2-sided 90% CL frequentist bound derived from Wilks' theorem (comparable to FC)

S < 3.88(7.25) at 90%(99%) CI

# In Conclusion:

## Frequentist confidence intervals

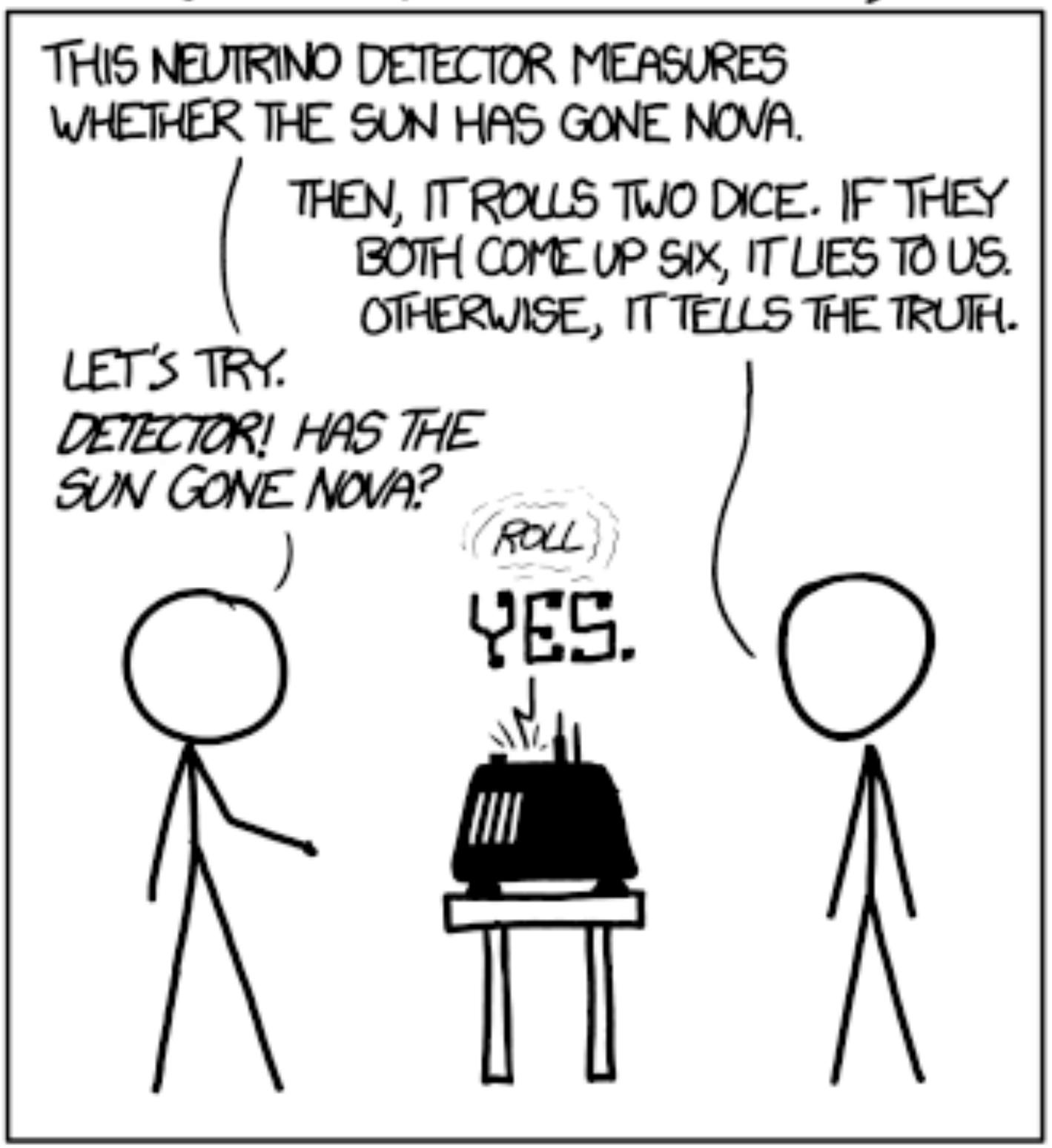
- Do not reliably indicate the relevance of individual measurements nor provide a robust basis for the comparison of different results;
- Are prone to frequent misinterpretation that can result in incorrect conclusions and seemingly paradoxical behaviours;
- Have no self-consistent method for error propagation or for producing single parameter bounds from a multi-dimensional parameter space;
- Do a poor job of conveying the objective information content of the data in a useful form.

# There is a better way!

*Another look at confidence intervals: proposal for a more relevant and transparent approach*
Biller and Oser, http://arxiv.org/abs/1405.5010 (2014)

http://xkcd.com/1132/

# DID THE SUN JUST EXPLODE?
## (IT'S NIGHT, SO WE'RE NOT SURE.)



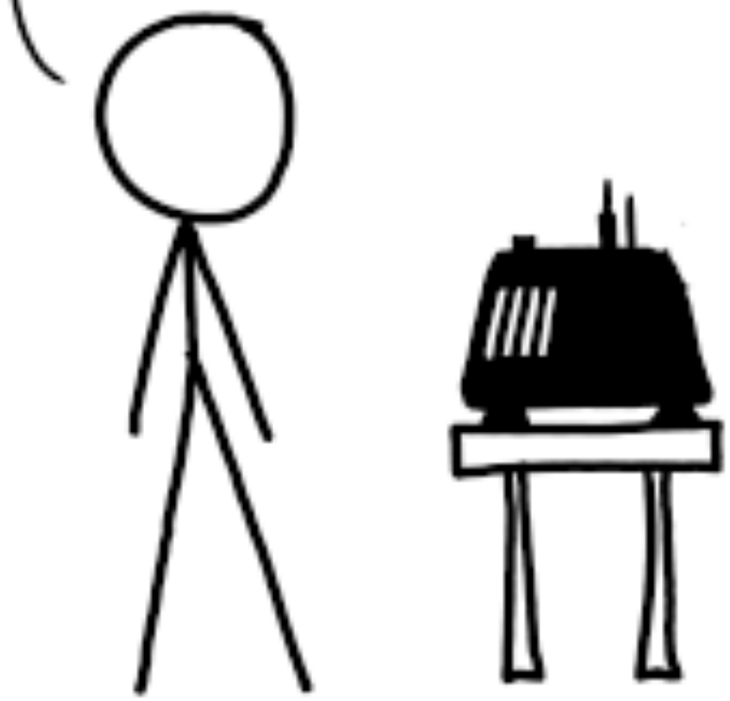

FREQUENTIST STATISTICIAN:

BAYESIAN STATISTICIAN:

FREQUENTIST STATISTICIAN:

THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS $\frac{1}{36} = 0.027$.

SINCE $p < 0.05$, I CONCLUDE THAT THE SUN HAS EXPLODED.



BAYESIAN STATISTICIAN:

BET YOU $50 IT HASN'T.

# Bayesians vs Frequentists

# THE HOOK

Feldman-Cousins intervals only **APPEAR** to be more physical for negative fluctuations, assisting in the misinterpretation.

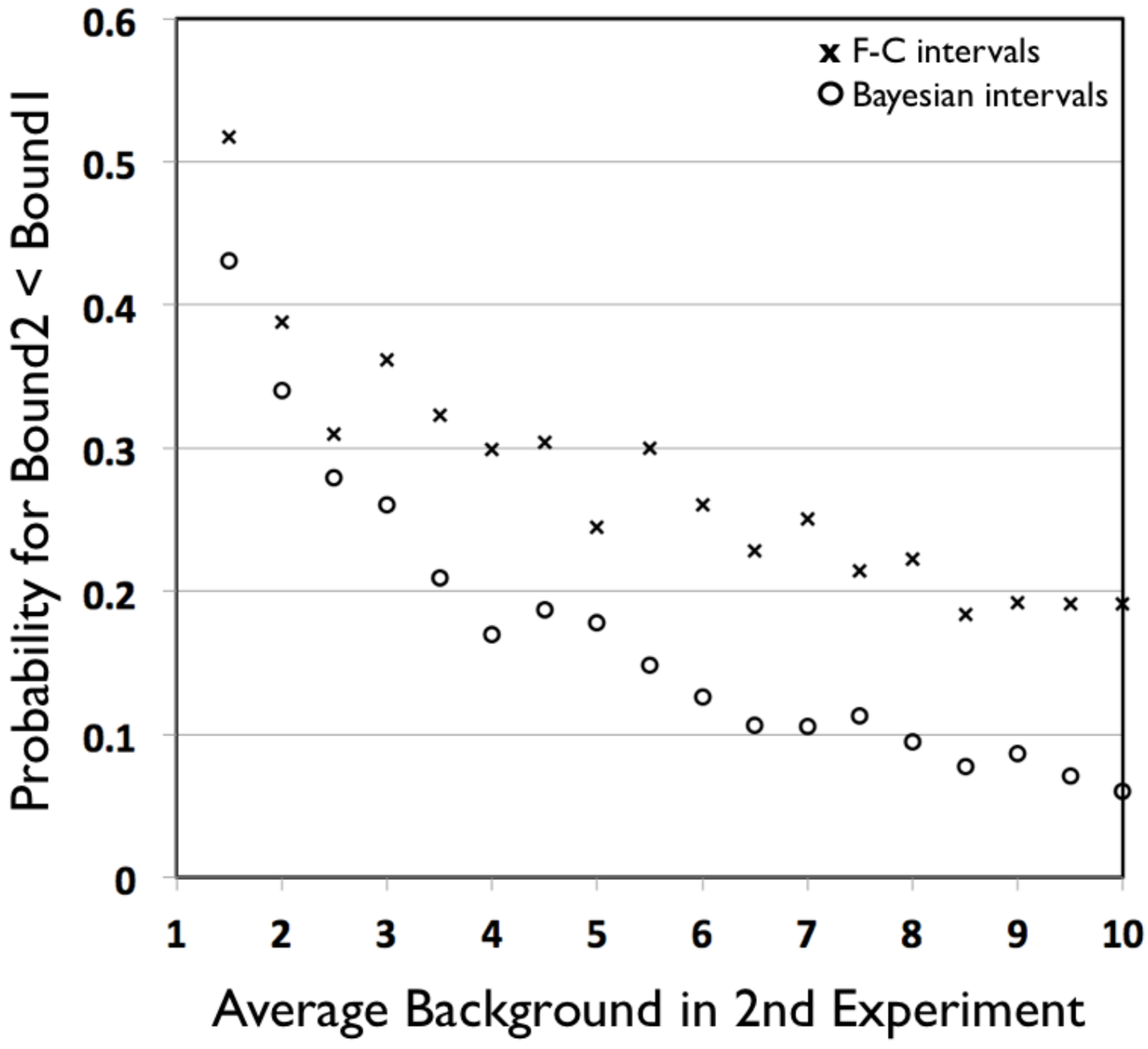
A 90% CL Feldman-Cousins interval has exactly the same coverage as a 90% CL standard frequentist interval!

# 90% CL Upper Bounds for Counting Experiments Under the Null Hypothesis

Experiment 1: Expected background of 1 count
Experiment 2: Higher expected background level

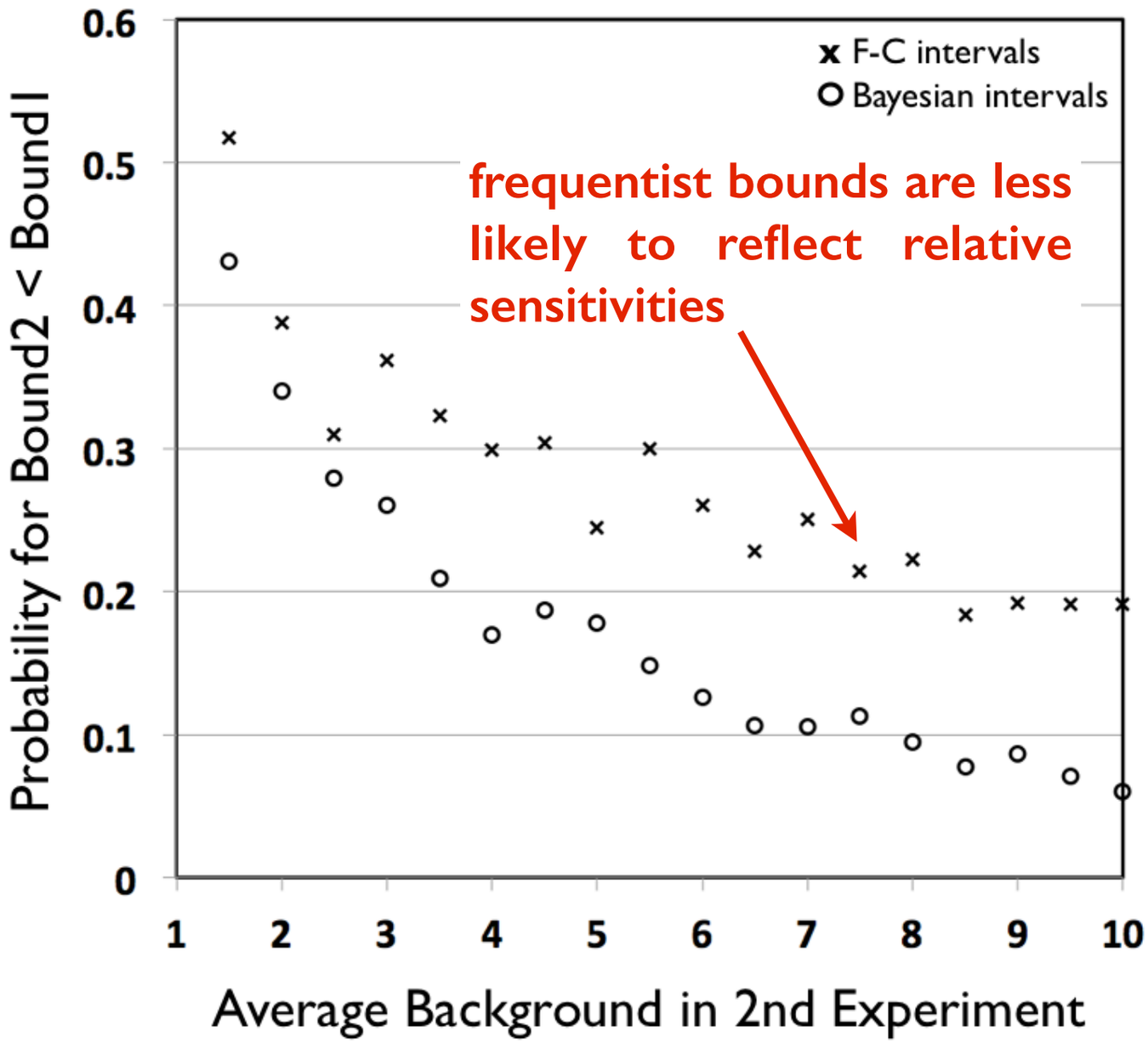
# 90% CL Upper Bounds for Counting Experiments Under the Null Hypothesis

Experiment 1: Expected background of 1 count
Experiment 2: Higher expected background level



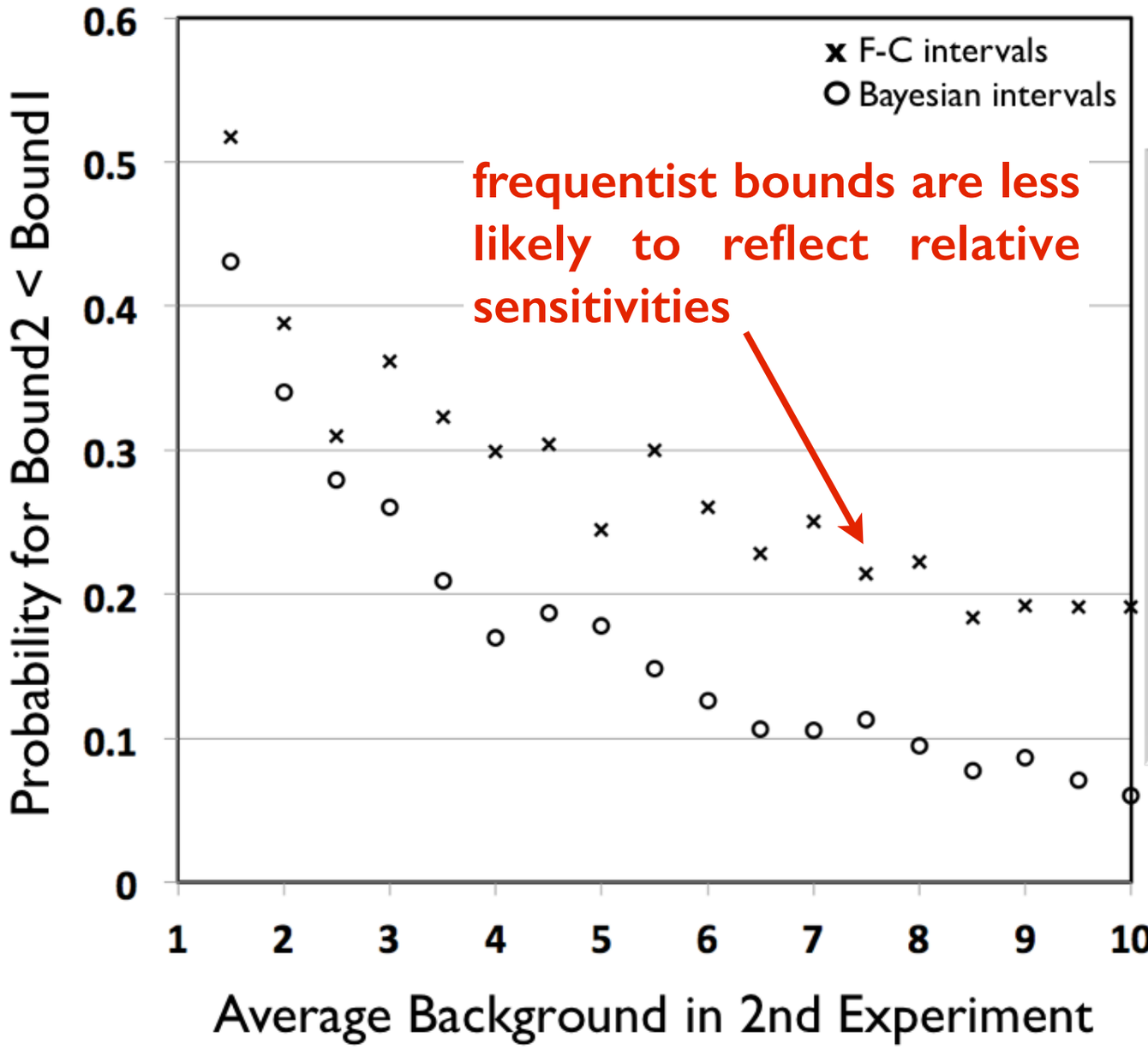
# 90% CL Upper Bounds for Counting Experiments Under the Null Hypothesis

Experiment 1: Expected background of 1 count
Experiment 2: Higher expected background level



# 90% CL Upper Bounds for Counting Experiments Under the Null Hypothesis

Experiment 1: Expected background of 1 count
Experiment 2: Higher expected background level

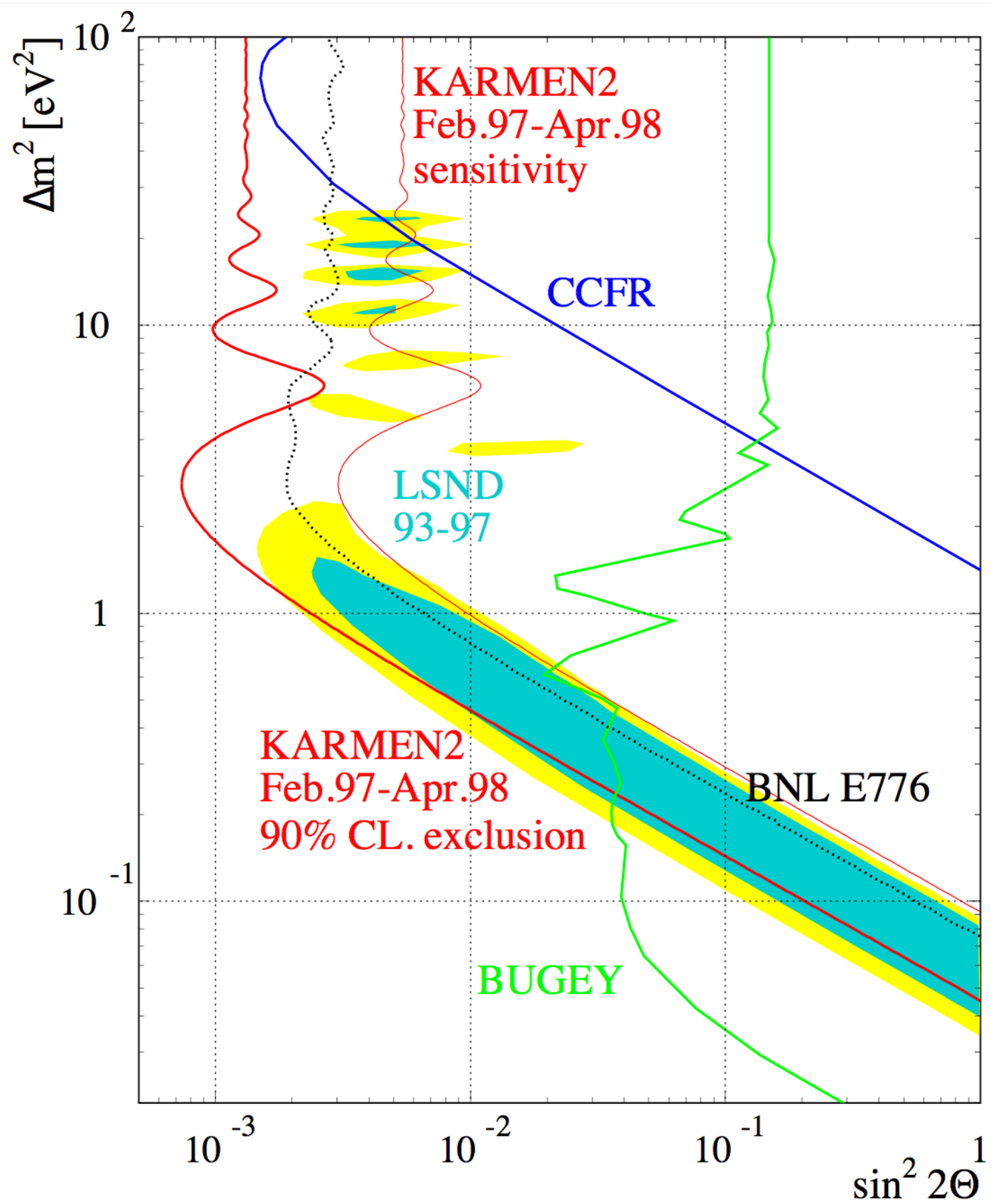



Comparisons of frequentist bounds among different experiments are **misleading**!

# KARMEN 2

**Avg Expected**
**Background: 2.88±0.13**
**Observed: zero**

## Neutrino Oscillation



# KARMEN 2

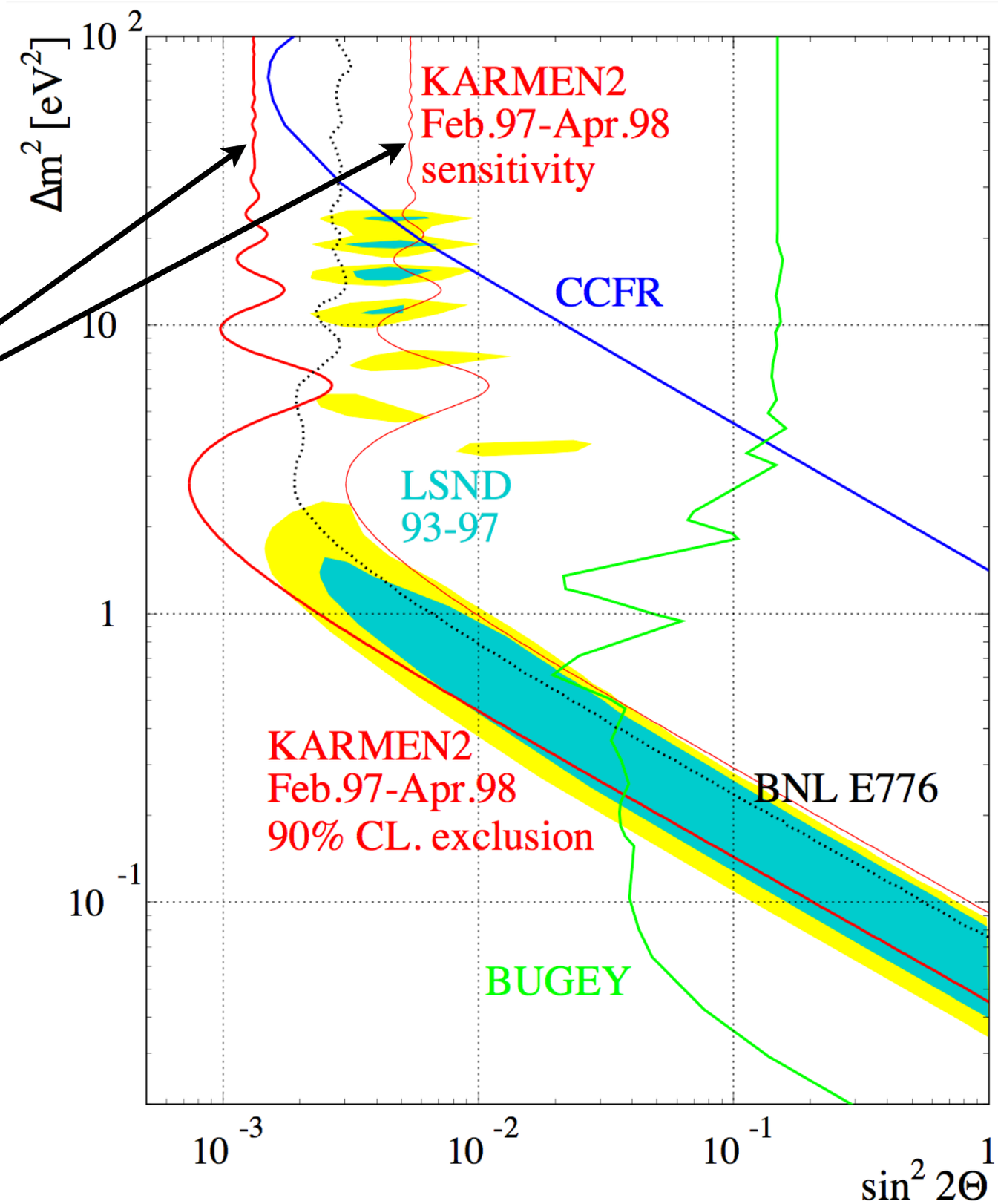
**Avg Expected Background: 2.88±0.13**

**Observed: zero**

Feldman-Cousins CL gives substantially better "exclusion" region than expected sensitivity

(K. Eitel et al., Nucl. Phys. Proc. Suppl. **77**, 1999)

## Neutrino Oscillation



# KARMEN 2

**Avg Expected Background: 2.88±0.13**

**Observed: zero**

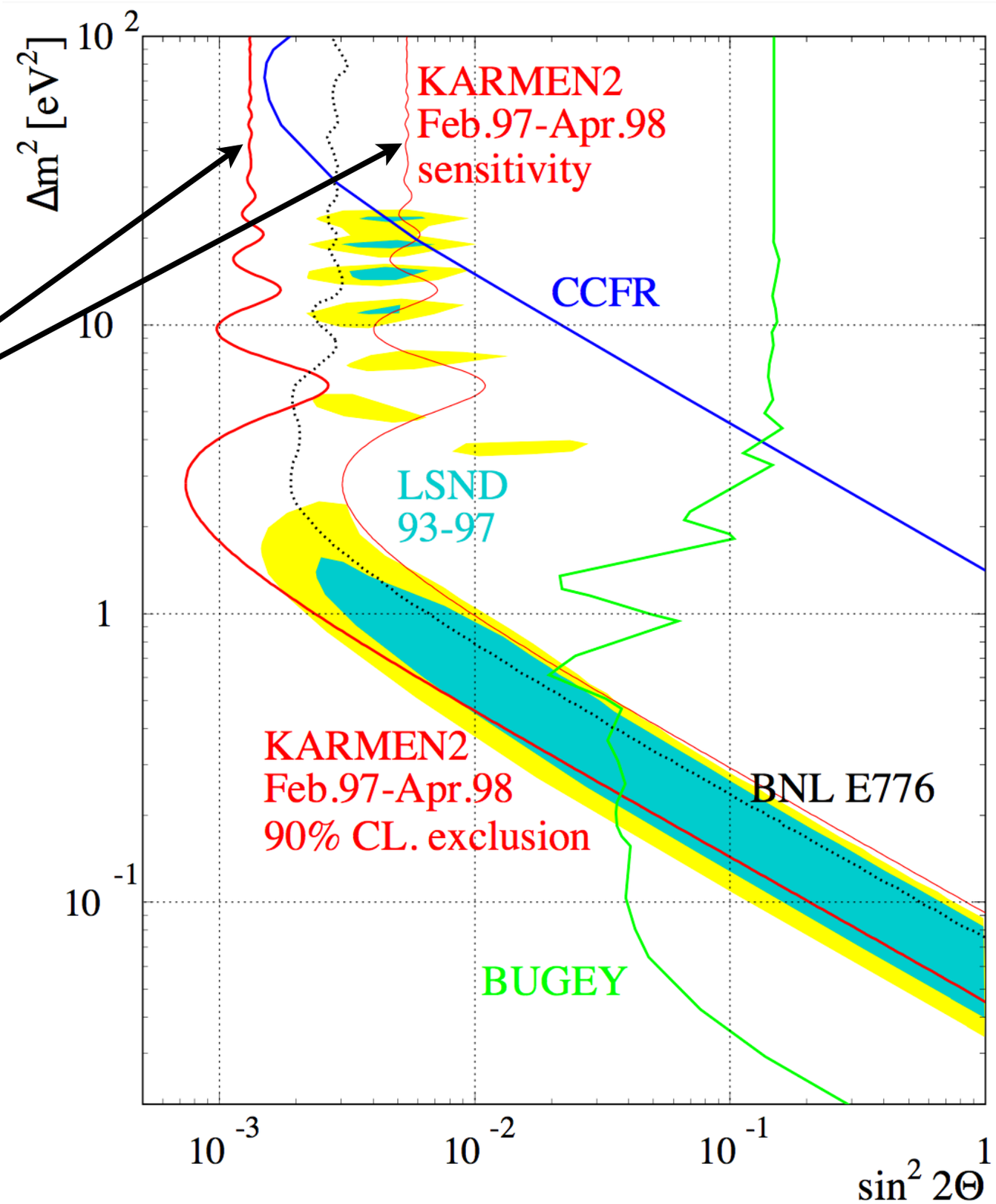
Feldman-Cousins CL gives substantially better "exclusion" region than expected sensitivity

(K. Eitel et al., Nucl. Phys. Proc. Suppl. **77**, 1999)

When statistics were quadrupled with longer running, the bounds obtained were nearly identical

(K. Eitel et al., Nucl. Phys. Proc. Suppl. **91**, 2001)

## Neutrino Oscillation



# KARMEN 2

**Avg Expected Background: 2.88±0.13**

**Observed: zero**

Feldman-Cousins CL gives substantially better "exclusion" region than expected sensitivity

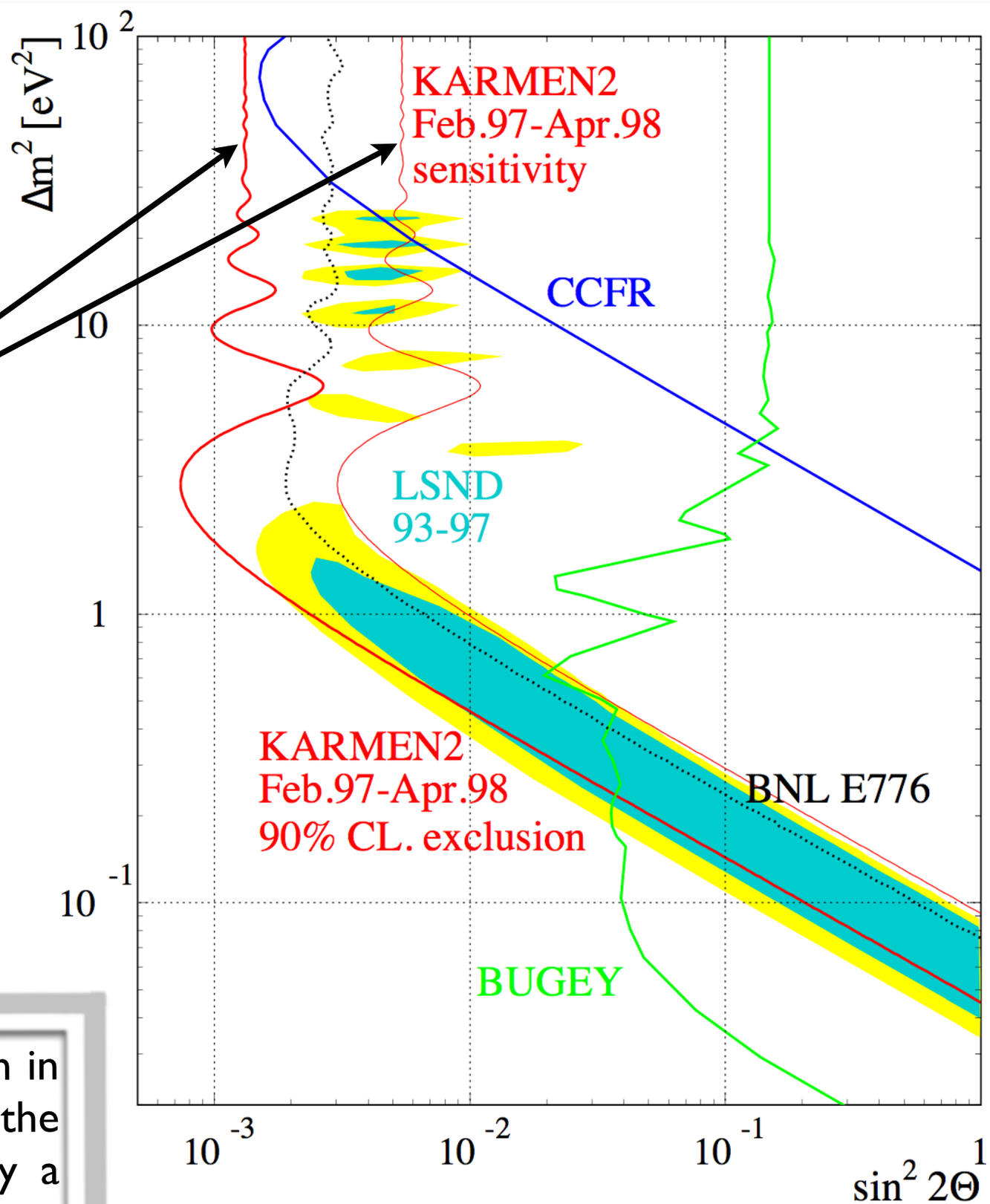
(K. Eitel et al., Nucl. Phys. Proc. Suppl. **77**, 1999)

When statistics were quadrupled with longer running, the bounds obtained were nearly identical

(K. Eitel et al., Nucl. Phys. Proc. Suppl. **91**, 2001)

Using Bayesian bounds with a prior uniform in rate, the initial bound is much closer to the expected sensitivity, and then improves by a factor of ~2 when the statistics are quadrupled

## Neutrino Oscillation



# LEP & LHC

*"(use of frequentist intervals) may lead to the undesired possibility that a large downward fluctuation of the background would allow hypotheses to be excluded for which the experiment has no sensitivity due to the small expected signal rate."*

# Particle Searches at Accelerators

# LEP & LHC

*"(use of frequentist intervals) may lead to the undesired possibility that a large downward fluctuation of the background would allow hypotheses to be excluded for which the experiment has no sensitivity due to the small expected signal rate."*

## Particle Searches at Accelerators

Consequence of misinterpreting frequentist bounds or, equivalently, using the wrong construction to answer a Bayesian question.

# LEP & LHC

*"(use of frequentist intervals) may lead to the undesired possibility that a large downward fluctuation of the background would allow hypotheses to be excluded for which the experiment has no sensitivity due to the small expected signal rate."*

## "$CL_{S+b}$" Procedure

Re-define "frequentist" upper bounds based on the ratio of the chance probability for the observation under a given signal +background hypothesis to that under the hypothesis of background alone.

If result is above some level of significance, then quote 2-sided Feldman-Cousins bound.

## Particle Searches at Accelerators

Consequence of misinterpreting frequentist bounds or, equivalently, using the wrong construction to answer a Bayesian question.

# LEP & LHC

*"(use of frequentist intervals) may lead to the undesired possibility that a large downward fluctuation of the background would allow hypotheses to be excluded for which the experiment has no sensitivity due to the small expected signal rate."*

## "$CL_{S+b}$" Procedure

Re-define "frequentist" upper bounds based on the ratio of the chance probability for the observation under a given signal +background hypothesis to that under the hypothesis of background alone.

If result is above some level of significance, then quote 2-sided Feldman-Cousins bound.

# Particle Searches at Accelerators

Consequence of misinterpreting frequentist bounds or, equivalently, using the wrong construction to answer a Bayesian question.

Equivalent to Bayesian for:
Poisson (prior uniform in mean rate);
Gaussian (prior uniform in mean);
Some others (different prior forms)

# LEP & LHC

*"(use of frequentist intervals) may lead to the undesired possibility that a large downward fluctuation of the background would allow hypotheses to be excluded for which the experiment has no sensitivity due to the small expected signal rate."*

## "$CL_{S+b}$" Procedure

Re-define "frequentist" upper bounds based on the ratio of the chance probability for the observation under a given signal +background hypothesis to that under the hypothesis of background alone.

If result is above some level of significance, then quote 2-sided Feldman-Cousins bound.

# Particle Searches at Accelerators

Consequence of misinterpreting frequentist bounds or, equivalently, using the wrong construction to answer a Bayesian question.

Equivalent to Bayesian for:
Poisson (prior uniform in mean rate);
Gaussian (prior uniform in mean);
Some others (different prior forms)

But not in all cases, such as likelihood ratios used for particle searches!

# LEP & LHC

*"(use of frequentist intervals) may lead to the undesired possibility that a large downward fluctuation of the background would allow hypotheses to be excluded for which the experiment has no sensitivity due to the small expected signal rate."*

## "$CL_{S+b}$" Procedure

Re-define "frequentist" upper bounds based on the ratio of the chance probability for the observation under a given signal +background hypothesis to that under the hypothesis of background alone.

If result is above some level of significance, then quote 2-sided Feldman-Cousins bound.

# Particle Searches at Accelerators

Consequence of misinterpreting frequentist bounds or, equivalently, using the wrong construction to answer a Bayesian question.

Equivalent to Bayesian for:
Poisson (prior uniform in mean rate);
Gaussian (prior uniform in mean);
Some others (different prior forms)

But not in all cases, such as likelihood ratios used for particle searches!

**No consistent meaning:**
Not frequentist (coverage not guaranteed);
Not generally Bayesian (if so, hidden prior)

# LEP & LHC

*"(use of frequentist intervals) may lead to the undesired possibility that a large downward fluctuation of the background would allow hypotheses to be excluded for which the experiment has no sensitivity due to the small expected signal rate."*

## "$CL_{S+b}$" Procedure

Re-define "frequentist" upper bounds based on the ratio of the chance probability for the observation under a given signal +background hypothesis to that under the hypothesis of background alone.

If result is above some level of significance, then quote 2-sided Feldman-Cousins bound.

# Particle Searches at Accelerators

Consequence of misinterpreting frequentist bounds or, equivalently, using the wrong construction to answer a Bayesian question.

Equivalent to Bayesian for:
Poisson (prior uniform in mean rate);
Gaussian (prior uniform in mean);
Some others (different prior forms)

But not in all cases, such as likelihood ratios used for particle searches!

**No consistent meaning:**
Not frequentist (coverage not guaranteed);
Not generally Bayesian (if so, hidden prior)

Violates principles of F-C approach

# LEP & LHC

*"(use of frequentist intervals) may lead to the undesired possibility that a large downward fluctuation of the background would allow hypotheses to be excluded for which the experiment has no sensitivity due to the small expected signal rate."*

## "$CL_{S+b}$" Procedure

Re-define "frequentist" upper bounds based on the ratio of the chance probability for the observation under a given signal +background hypothesis to that under the hypothesis of background alone.

If result is above some level of significance, then quote 2-sided Feldman-Cousins bound.

# Particle Searches at Accelerators

Consequence of misinterpreting frequentist bounds or, equivalently, using the wrong construction to answer a Bayesian question.

Equivalent to Bayesian for:
Poisson (prior uniform in mean rate);
Gaussian (prior uniform in mean);
Some others (different prior forms)

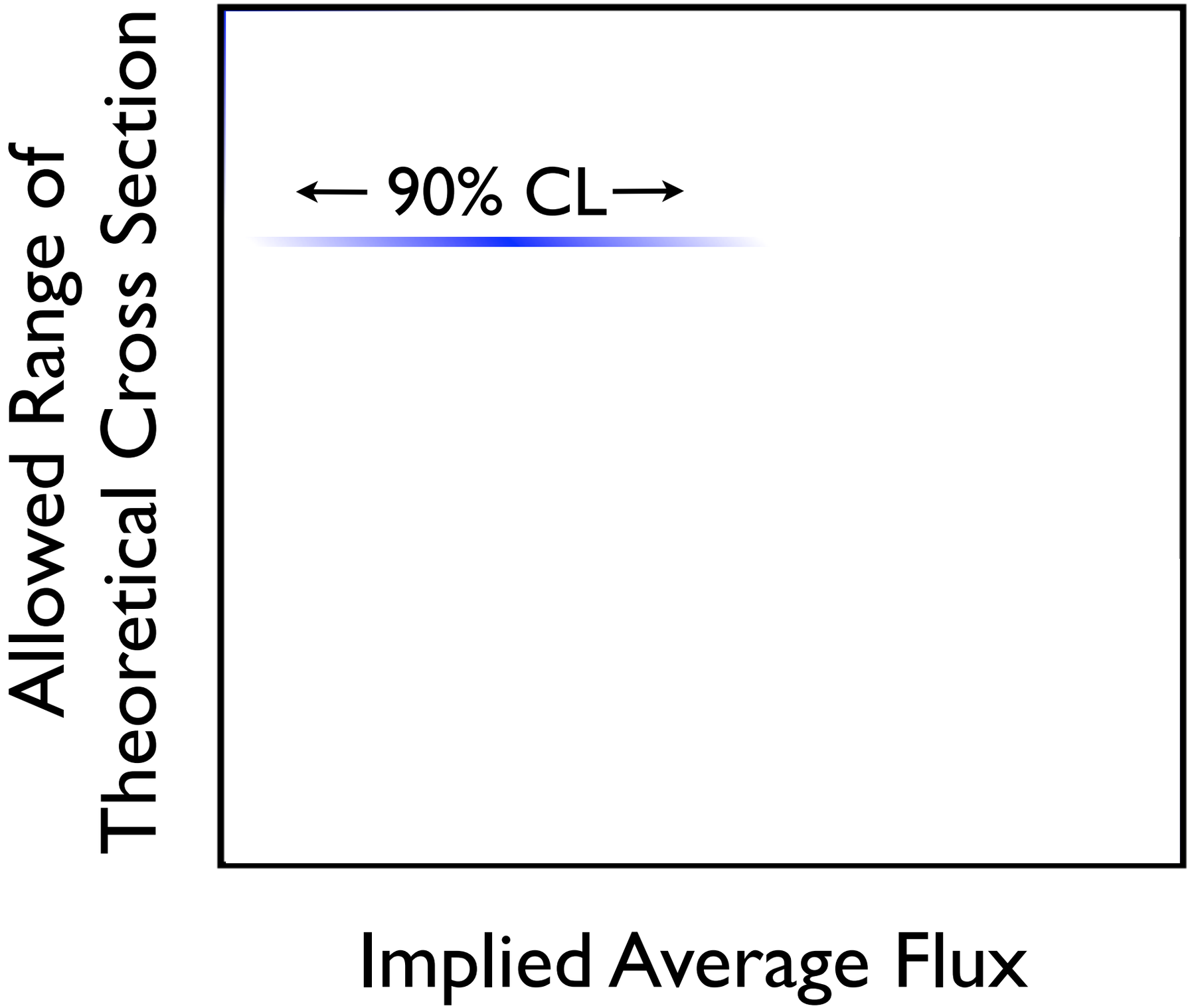
But not in all cases, such as likelihood ratios used for particle searches!

**No consistent meaning:**
Not frequentist (coverage not guaranteed);
Not generally Bayesian (if so, hidden prior)

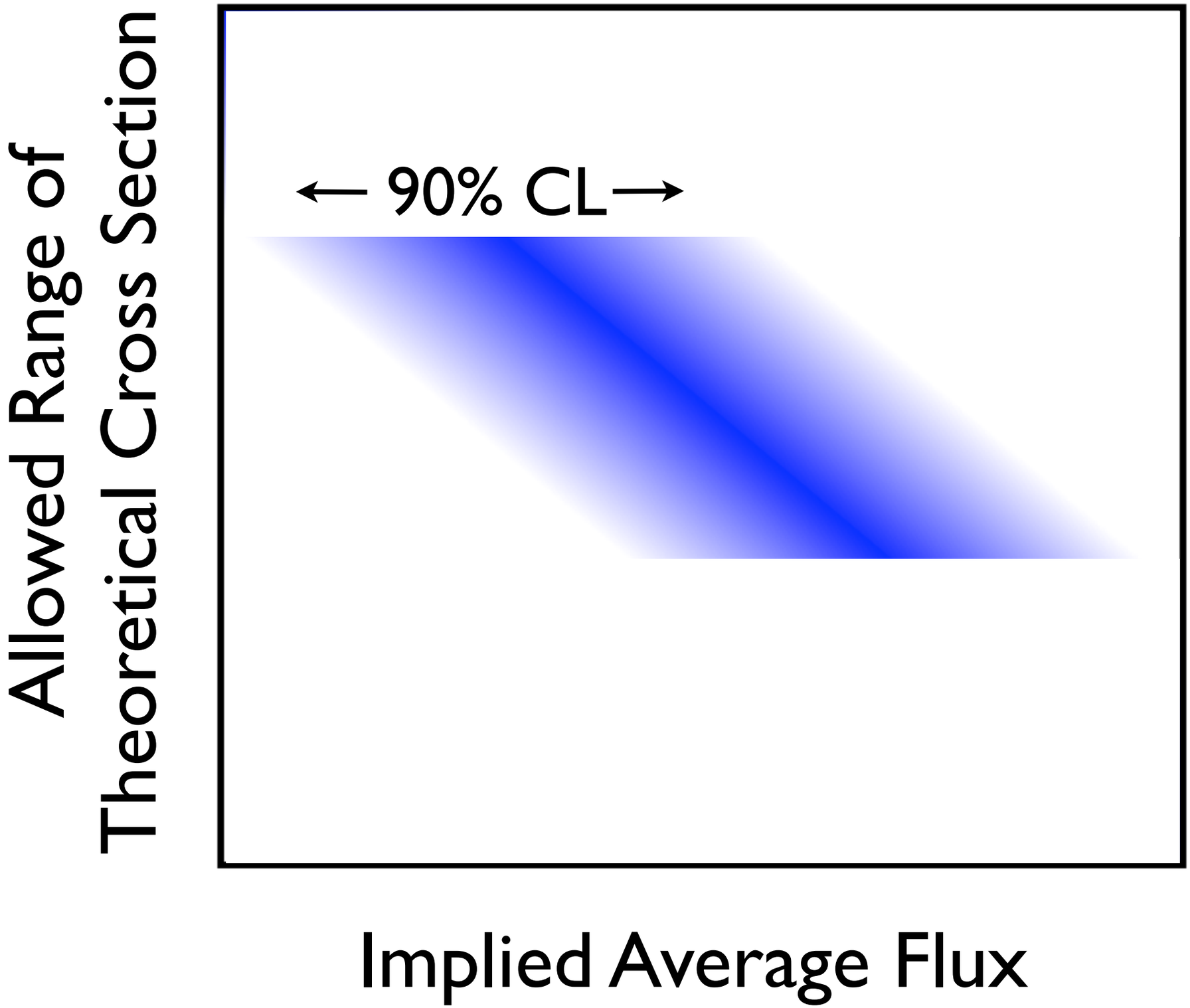
Violates principles of F-C approach

Alternative: Confront question explicitly & apply consistent mathematical formalism, such as p-values and Bayesian bounds with transparent priors

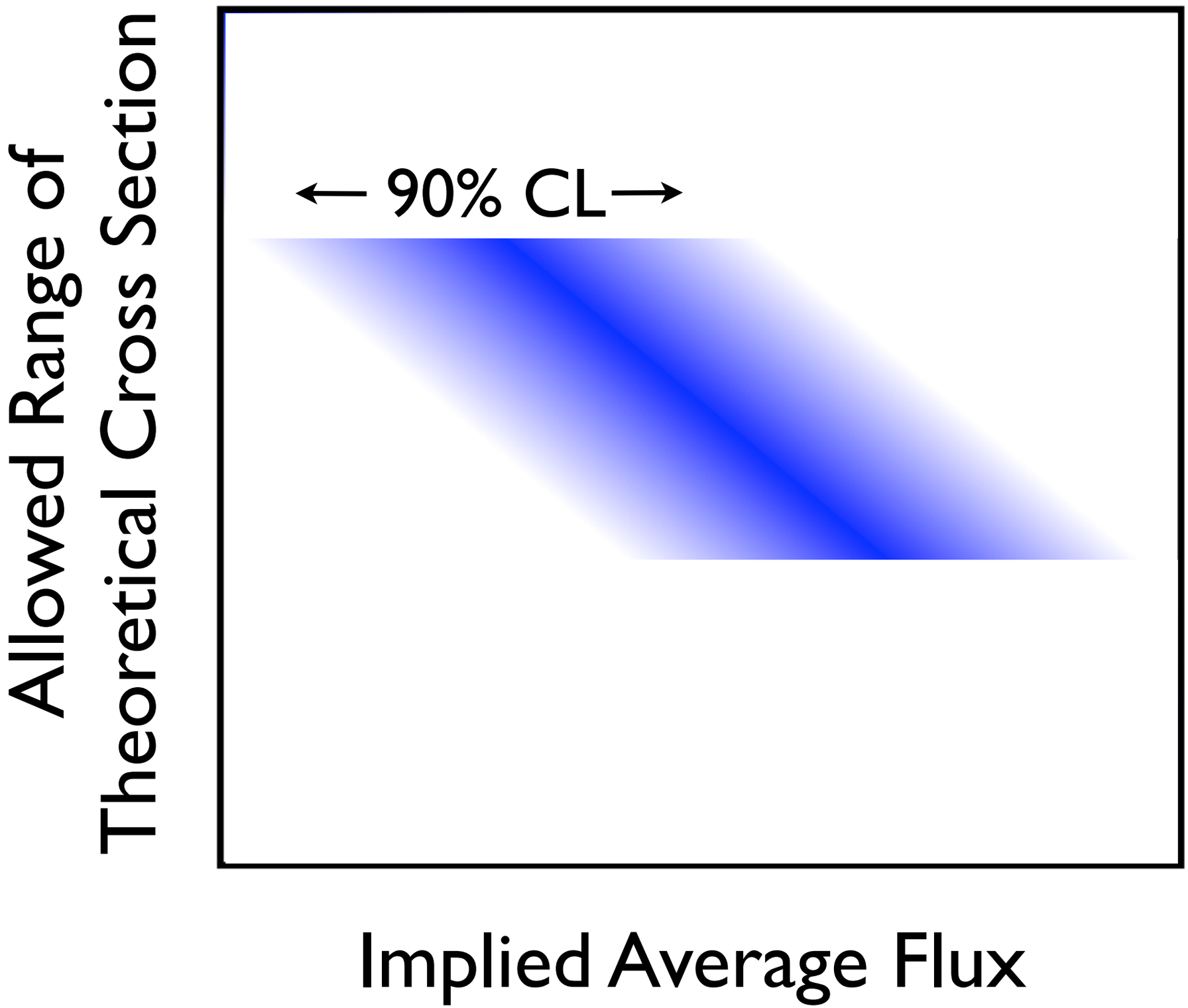
# Propagation of Experimental Uncertainties



# Propagation of Experimental Uncertainties



# Propagation of Experimental Uncertainties




**No rigorous self-consistent purely frequentist method to propagate systematic uncertainties!**

integration over $\xi$

Allowed Range of Detector Energy Scale $\xi$
(statistical calibration)

joint likelihood of a given data set
assuming $\xi$ is statistically distributed

$\theta$

Bayesian integration over $\xi$ assuming uniform prior

Allowed Range of Detector Energy Scale $\xi$
(statistical calibration)

joint likelihood of a given data set
assuming $\xi$ is statistically distributed

$\theta$

Allowed Range of
Detector Energy Scale ξ
(statistical calibration)

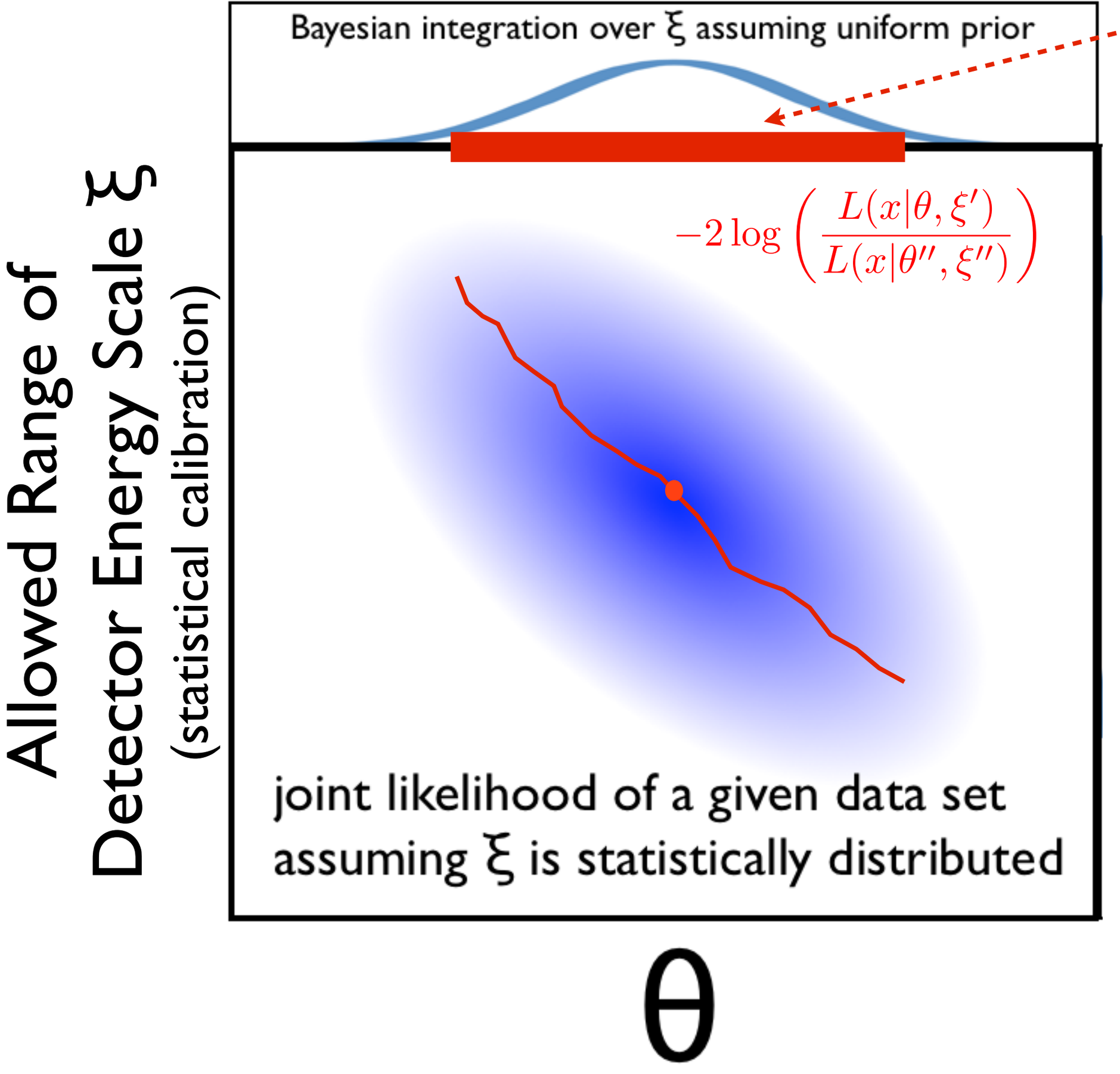
Bayesian integration over ξ assuming uniform prior

Coverage not insured for all ξ, unless you greatly over-cover

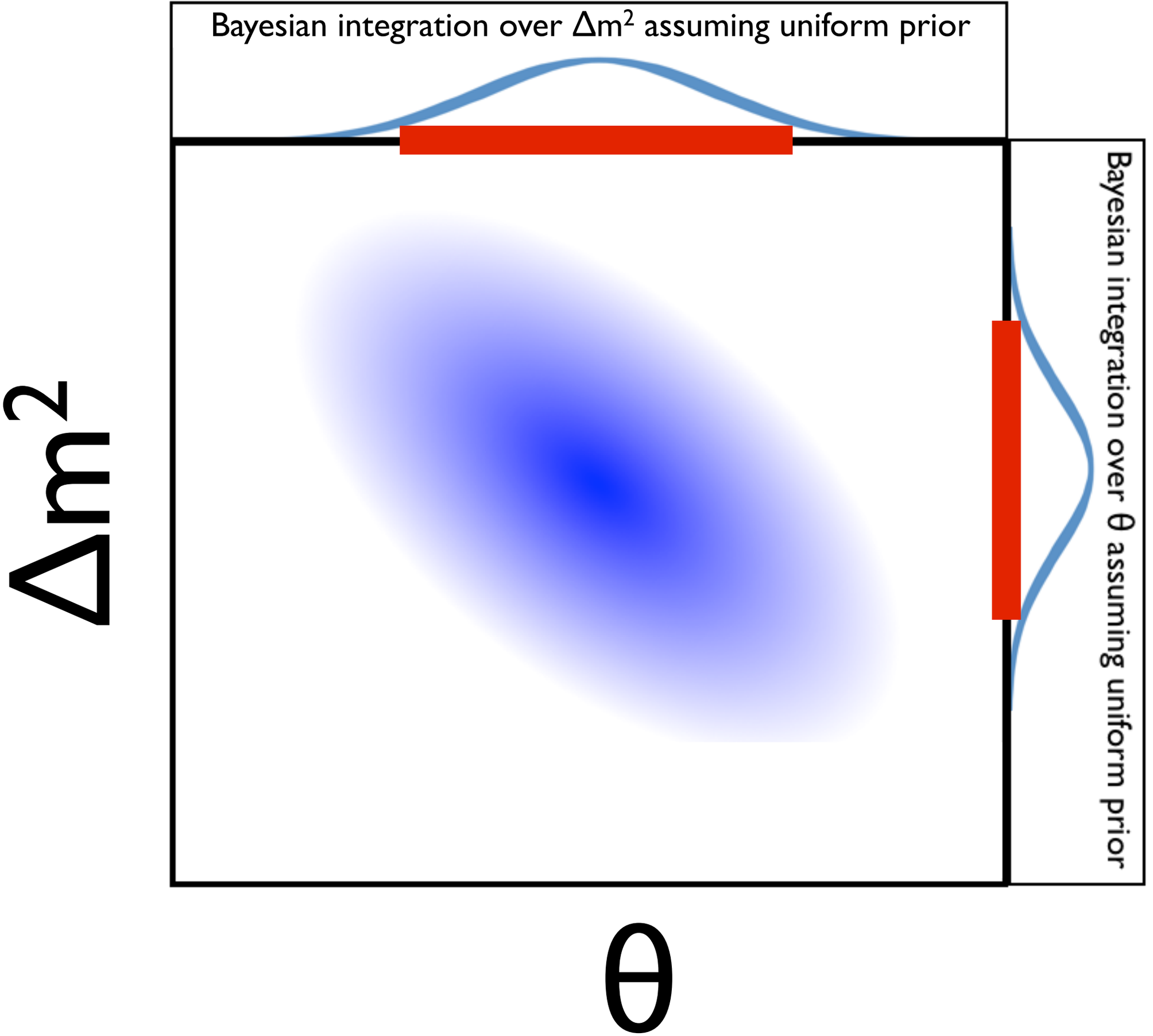
$$-2\log\left(\frac{L(x|\theta,\xi')}{L(x|\theta'',\xi'')}\right)$$

joint likelihood of a given data set
assuming ξ is statistically distributed

θ




**No rigorous self-consistent purely frequentist method to propagate statistical uncertainties!**




**No rigorous self-consistent purely frequentist method to get lower dimensional bounds!**