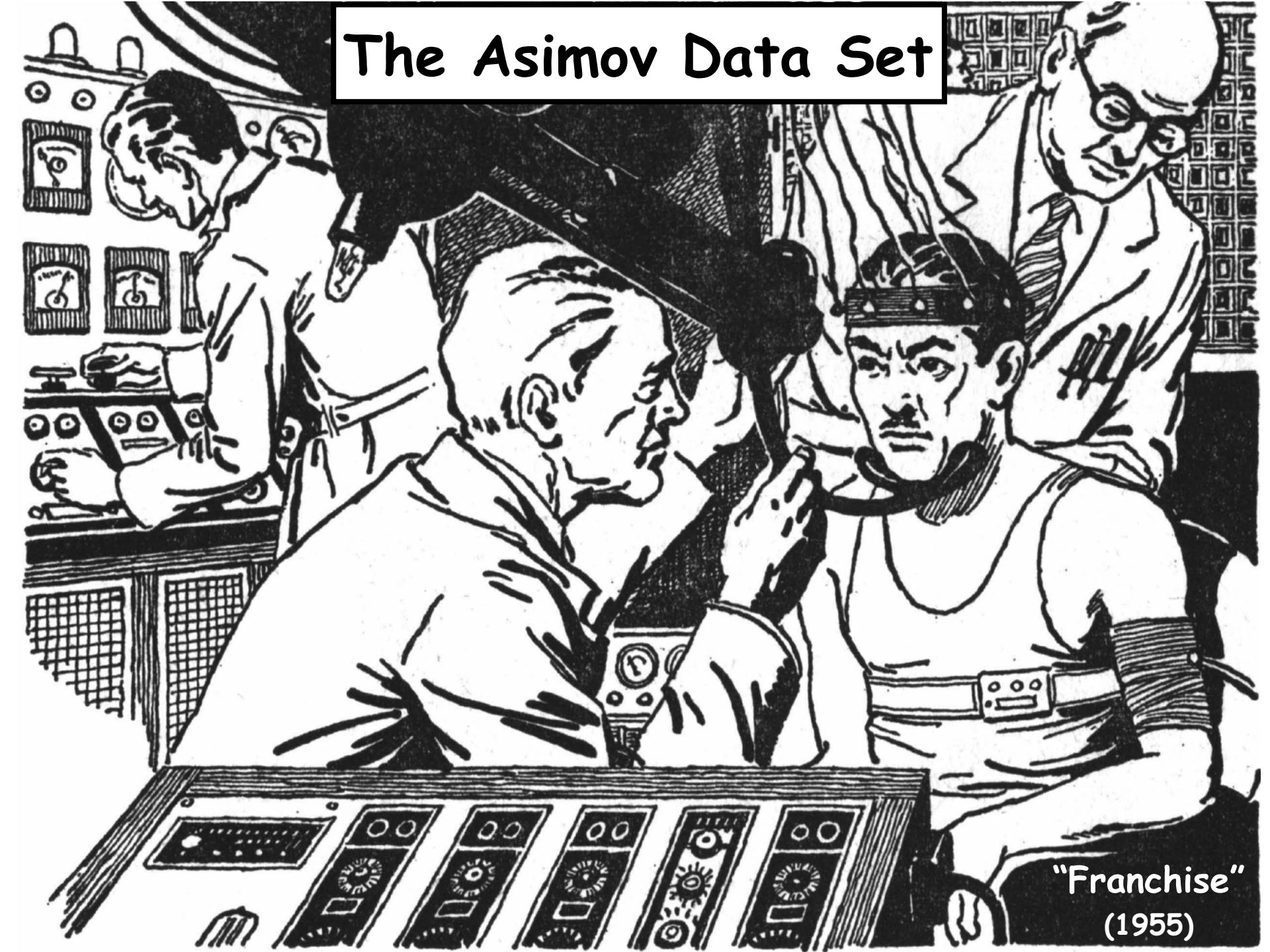
Lecture 6:

- Asimov
- Barlow-Beeston-Conway
- Bayes

The Asimov Data Set



"Franchise"
(1955)

Assume that we have a set of multi-dimensional PDF defined by an arbitrary number of bins (N_{bins}) that can be combined under a particular model ($\boldsymbol{m}$, defined by $\boldsymbol{q}$ parameters) to yield a predicted mean number of observed counts ($\mathbf{n}$) in each bin ($\boldsymbol{i}$) from a given data set. The likelihood can then be expressed as:

$$\mathscr{L} = \prod_{i=1}^{N_{bins}} \left[\frac{m_i(\mathbf{q})^{n_i} e^{-m_i(\mathbf{q})}}{n_i!} \right]$$

$$\log \mathscr{L} = \sum_{i=1}^{N_{bins}} \left[n_i \log m_i(\mathbf{q}) - m_i(\mathbf{q}) - \log n_i! \right]$$

The log-likelihood ratio with respect to some nominal model, $\boldsymbol{m(q_0)}$, is then given by:

$$\log \frac{\mathscr{L}}{\mathscr{L}_0} \equiv \log \mathscr{L}_R$$

$$= \sum_{i=1}^{N_{bins}} \left[n_i \log m_i(\mathbf{q}) - m_i(\mathbf{q}) - n_i \log m_i(\mathbf{q_0}) + m_i(\mathbf{q_0}) \right]$$

$\boldsymbol{q_0}$ might, for example, represent the null hypothesis or could just be the point where the likelihood is maximum

Say we're interested in what to expect on average for the log-likelihood ratio as a function of 'test' parameter values:

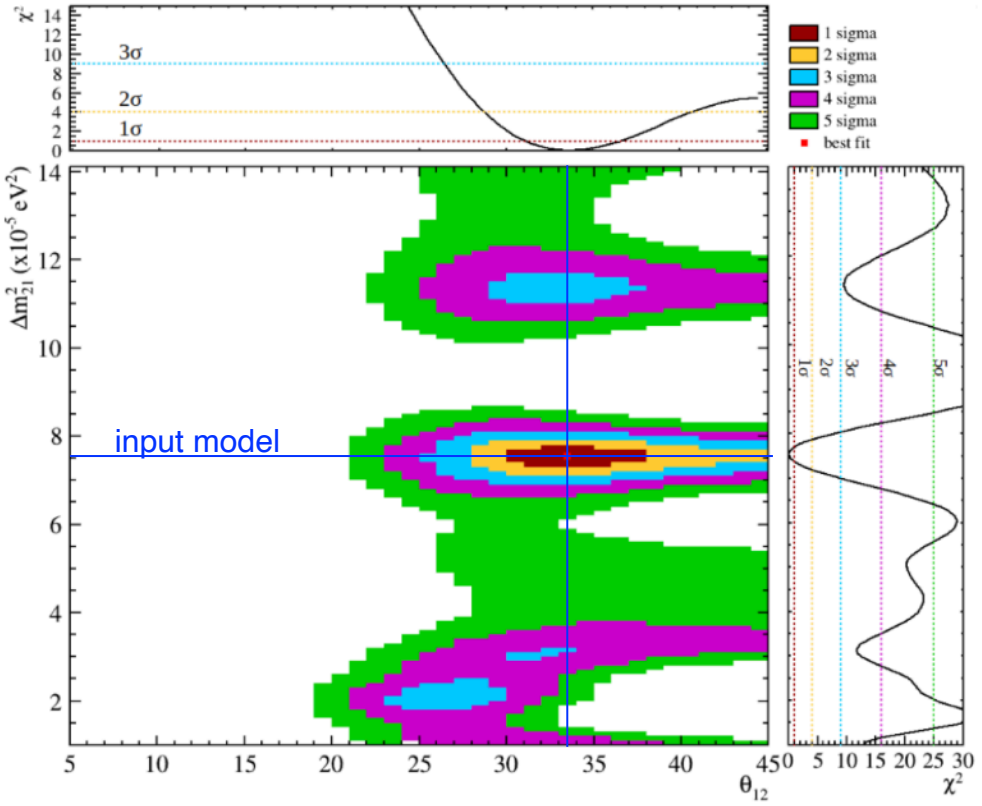
$$\left\langle \log \mathscr{L}_R \right\rangle = \left\langle \sum_{i=1}^{N_{bins}} \left[n_i \log m_i(\mathbf{q}) - m_i(\mathbf{q}) - n_i \log m_i(\mathbf{q_0}) + m_i(\mathbf{q_0}) \right] \right\rangle$$

$$= \sum_{i=1}^{N_{bins}} \left\langle \left[n_i \log m_i(\mathbf{q}) - m_i(\mathbf{q}) - n_i \log m_i(\mathbf{q_0}) + m_i(\mathbf{q_0}) \right] \right\rangle$$

$$= \sum_{i=1}^{N_{bins}} \left[\left\langle n_i \right\rangle \log m_i(\mathbf{q}) - m_i(\mathbf{q}) - \left\langle n_i \right\rangle \log m_i(\mathbf{q_0}) + m_i(\mathbf{q_0}) \right]$$

So we just need to substitute in "perfect, un-fluctuated" expectation values for a representative data set. This could, for example, be taken from scaling the PDF model for some particular set of parameters to the size of a typical data set.

Can be used to find the expected sensitivity for discovering a particular phenomenon, or the expected power to discriminate between different model, or the expected accuracy in constraining model parameters.



Expected ability to constrain oscillation parameters after 5 years of reactor anti-neutrino data from SNO+

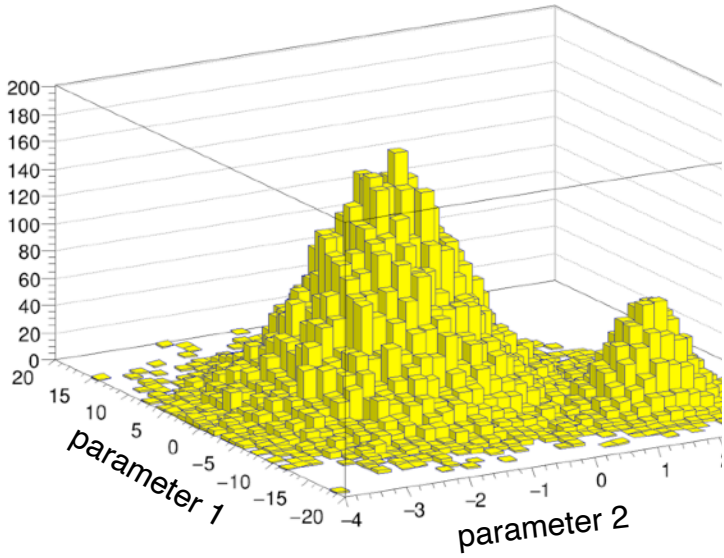
(thanks to Iwan Morton-Blake)

Perfect data ought to give perfect results if you're doing things right!

Incredibly useful! Also an excellent way to check if your code is doing the right thing and understanding basic characteristics without having to run the full analysis chain thousands of times!

Accounting for Statistical Uncertainties in PDFs

Let's say you create a set of PDFs for some parameters by running lots of simulations, binning the resulting distributions of parameter values and then normalising the areas of each histogram to one.



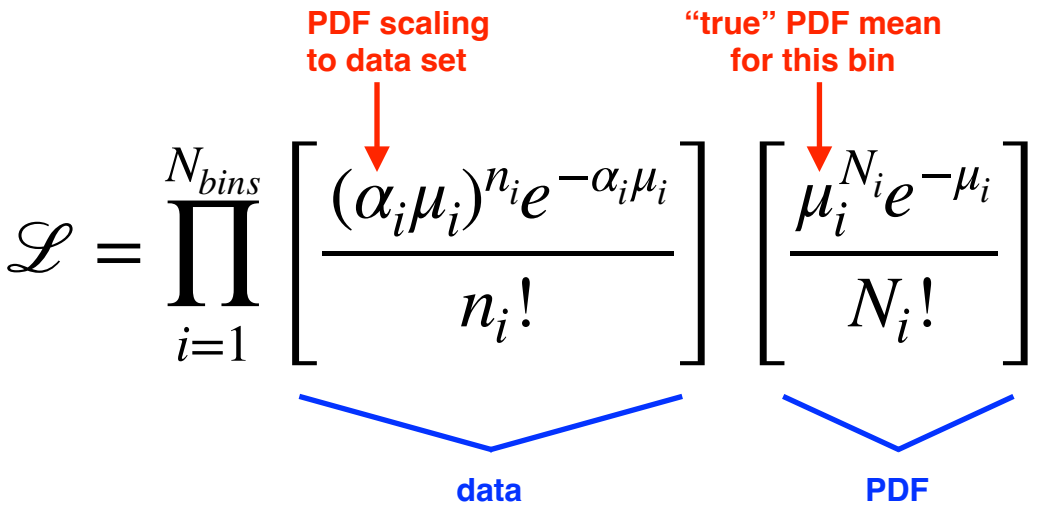
How do you deal with the statistical uncertainties in the constructed PDFs?

Could smooth PDFs, but can be tricky in multiple dimensions and has the potential to produce artefacts

$$P(Data \mid Model) \rightarrow \cancel{P(Data \mid PDF\ Histograms)}$$

$$\searrow$$

$$P(Data, PDF\ Histograms \mid Model)$$



Want to maximise overall likelihood, so maximise here over the set of "true" PDF means

Complicated by model correlations between bins and multi-component models

First analysed by Barlow and Beeston (Comp. Phys. Comm. 77, 219, 1993)

A much more practical approximation by Conway (PHYSTAT 2011, arXiv:1103.0354), which is what we'll follow here.

For the ith bin in the data and PDF histogram, the contribution to the extended likelihood is:

$$-\ln \mathscr{L}_i = -n_i \ln \mu_i + \mu_i$$

where n_i = number observed and μ_i = model prediction based on the PDFs

Make Two Simplifying Assumptions:

1) Take model systematics to be uncorrelated between bins to allow bin-by-bin error propagation (conservative);
2) Assume the uncertainty in μ due to statistical fluctuations in the contributing PDFs can be approximated by a single Gaussian scaling.

We can then drop the bin subscript for simplicity incorporate the Gaussian uncertainty scaling into the likelihood for that bin as follows:

$$-\ln \mathscr{L} = -n \ln \beta\mu + \beta\mu + \frac{(\beta - 1)^2}{2\sigma^2}$$

We want to maximise the likelihood (minimise $-\ln\mathscr{L}$), which can be explicitly done bin-by-bin in the parameter β by differentiation:

$$\beta^2 + (\mu\sigma^2 - 1)\beta - n\sigma^2 = 0$$

Solve for β in each bin and calculate the likelihood…

What is σ for the bin?

Assume we are using k PDFs to model the total number of events predicted in this bin:

$$\mu = \sum_{j=1}^{k} \mu_j$$

where

$$\mu_j = f_j \left(\frac{m_j}{N_j} \right)$$

simulated events for this PDF that fall in this bin

PDF normalisation

total # simulated events for this PDF

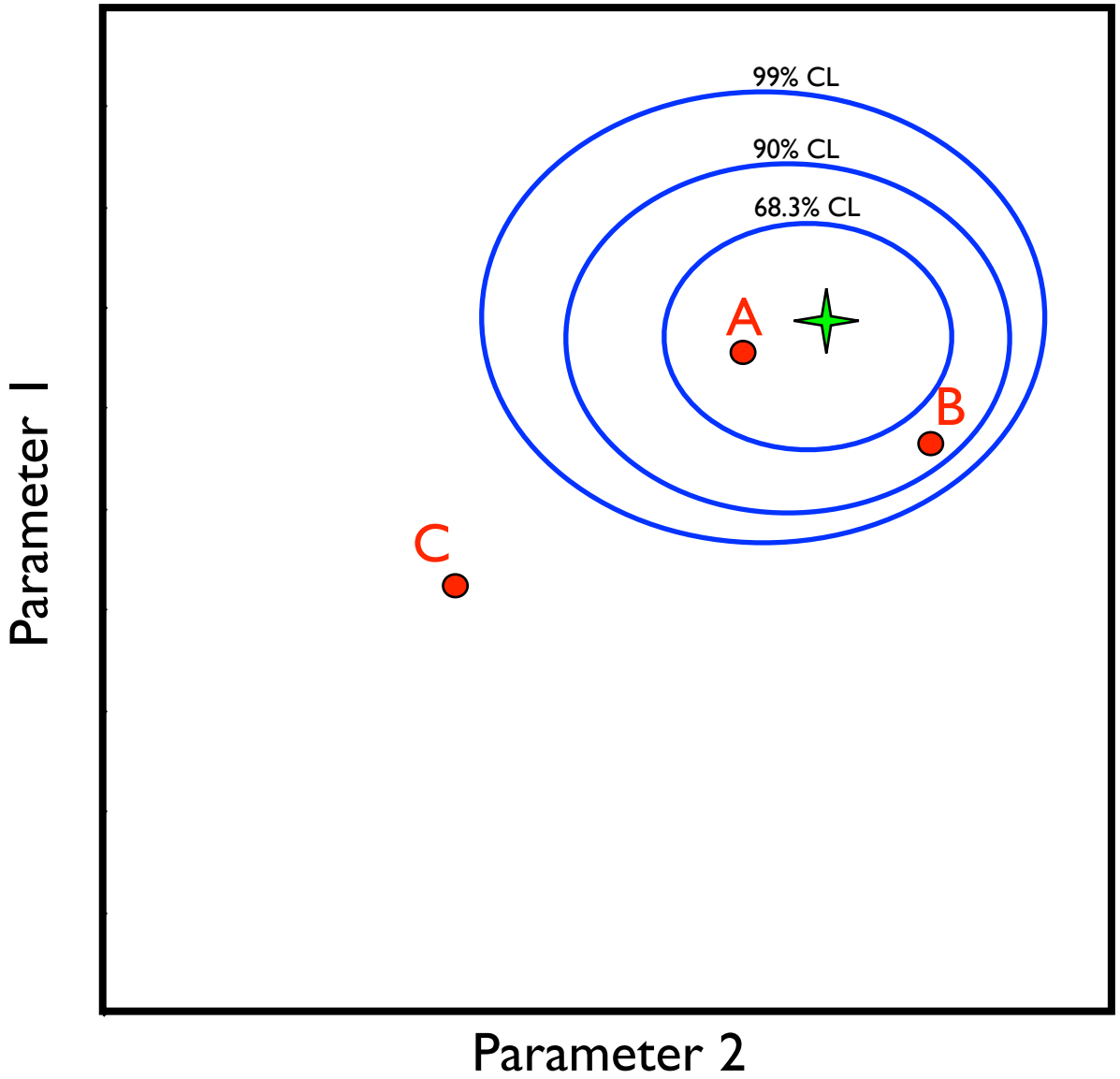
$$\sigma_j \equiv \frac{\Delta \mu_j}{\mu_j} \simeq \frac{\Delta m_j}{m_j} \simeq \frac{1}{\sqrt{m_j}} = \sqrt{\frac{f_j}{\mu_j N_j}}$$

So just need to remember this

$$\sigma^2 = \sum_{j=1}^{k} \sigma_j^2$$

Still issues for $m_j \sim 0$, (hard to get around)

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 (which is not repeated over and over again).

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.

Bayes' Theorem

$$P(\text{A and B}) = P(\text{B})P(\text{A}|\text{B}) = P(\text{A})P(\text{B}|\text{A})$$

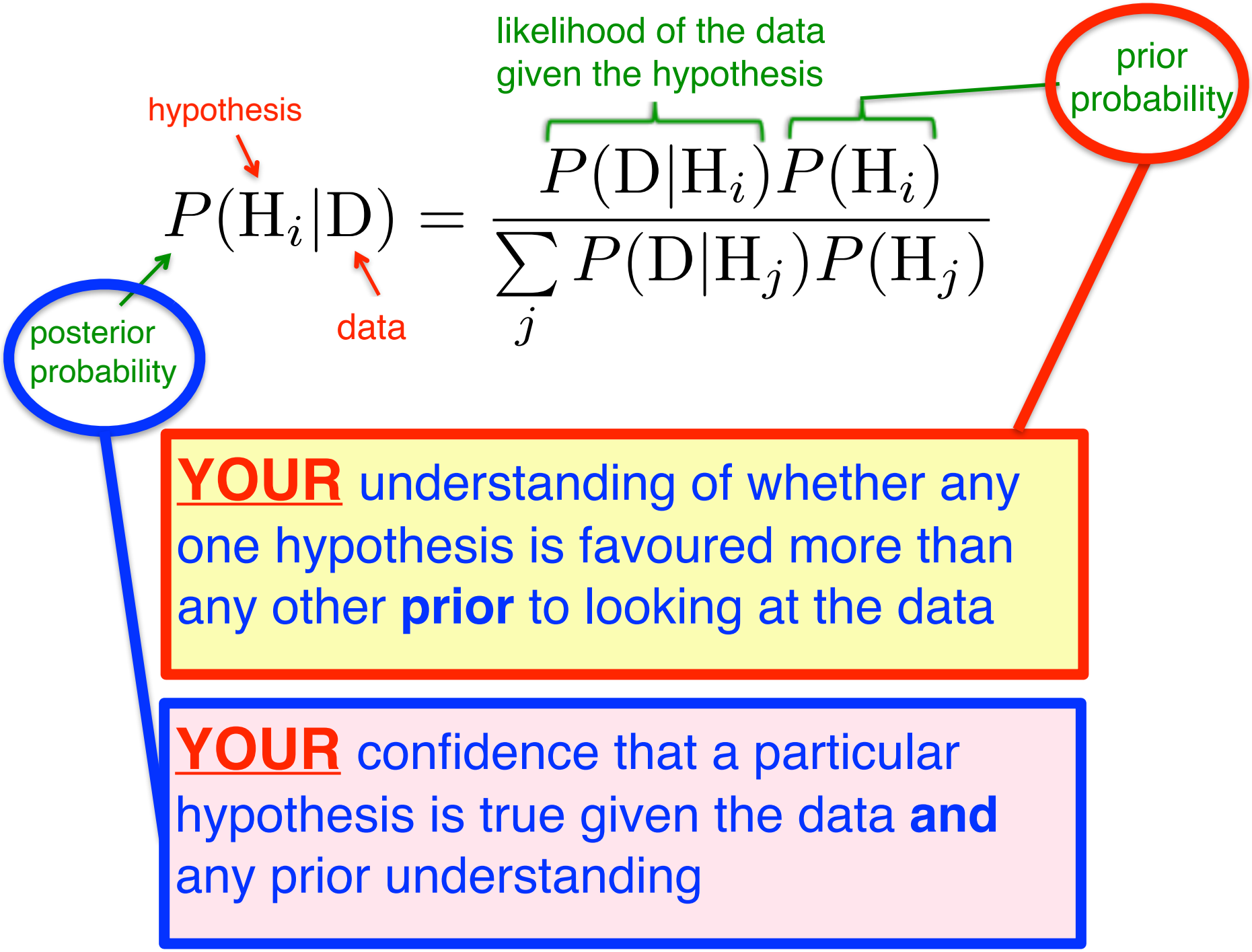
$$P(A\,|\,B) = \frac{P(B\,|\,A)P(A)}{P(B)}$$

note:
$P(A\,|\,B) \neq P(B\,|\,A)$

If there are multiple versions of A to choose from, then

$$P(B) = \sum_j P(B\,|\,A_j)P(A_j)$$

$$P(A_i\,|\,B) = \frac{P(B\,|\,A_i)P(A_i)}{\sum_j P(B\,|\,A_j)P(A_j)}$$


likelihood of the data
given the hypothesis
prior
probability
hypothesis
$P(\mathrm{H}_i|\mathrm{D}) = \frac{P(\mathrm{D}|\mathrm{H}_i)P(\mathrm{H}_i)}{\sum_j P(\mathrm{D}|\mathrm{H}_j)P(\mathrm{H}_j)}$
posterior
probability
data
YOUR understanding of whether any one hypothesis is favoured more than any other prior to looking at the data
YOUR confidence that a particular hypothesis is true given the data and any prior understanding

Relative probability ratio between two different hypothesis given the same observed data

$$\frac{P(H_i|D)}{P(H_k|D)} = \frac{P(D|H_i)}{P(D|H_k)} \frac{P(H_i)}{P(H_k)}$$

likelihood ratio

"odds" ratio

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 some particular spectrum; the relative probabilities for different event classes are drawn from some given distribution, *etc.*).

The probabilities of different hypotheses are the same in what metric?

All values of **A** are equally likely $\neq$ All values of $\mathbf{A}^2$ are equally likely

2. Non-Informative:

(A Case of Too Much History!)

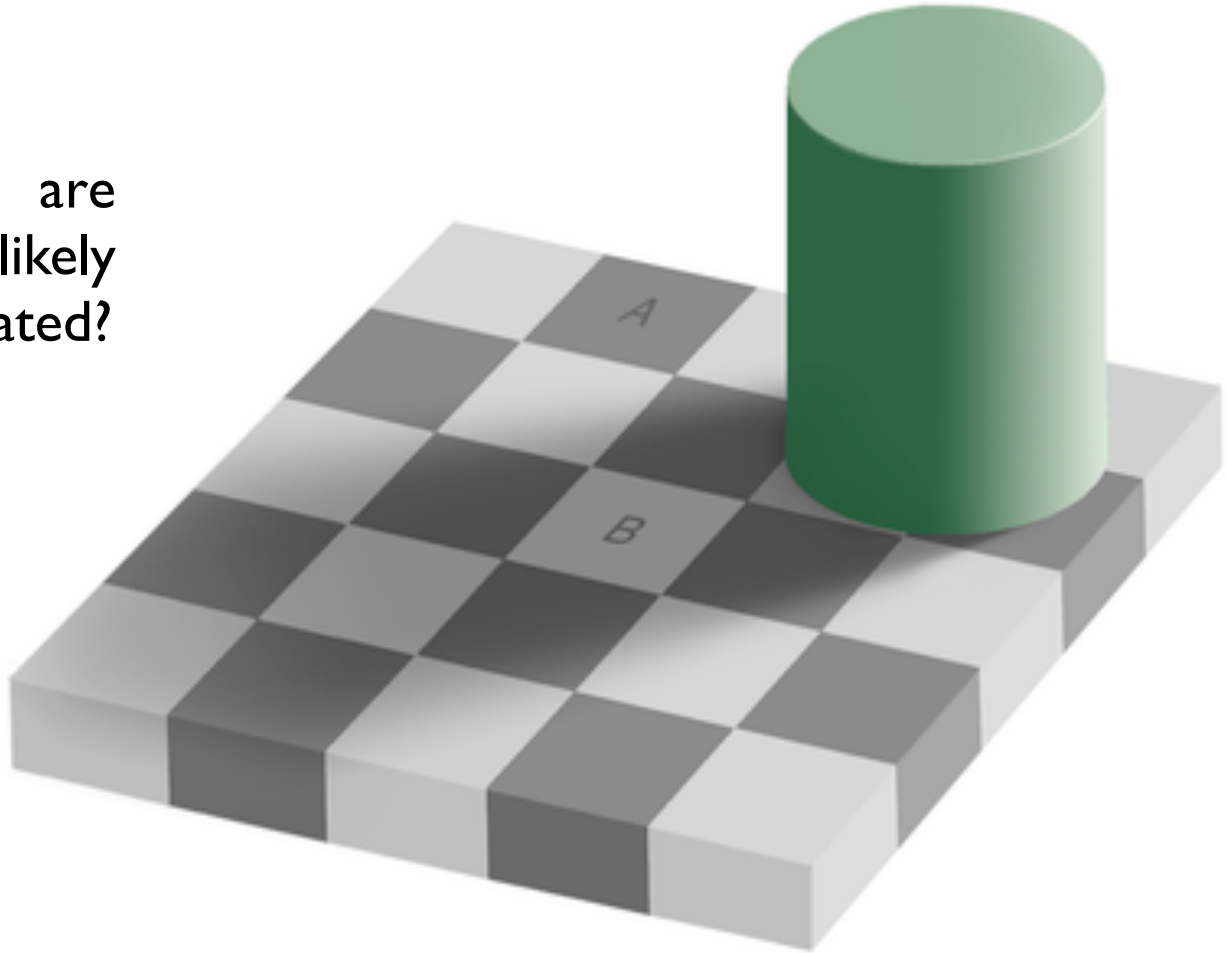
When there is no clear *a priori* preference, you must still choose a context to be used for comparing models.

Your brain inherently makes Bayesian inferences:

Context is necessary to relate data to model parameters

(visual observation) (optical properties of surface)

Prior: How are the squares likely being illuminated?



The <u>model</u> is of central importance to enable predictions

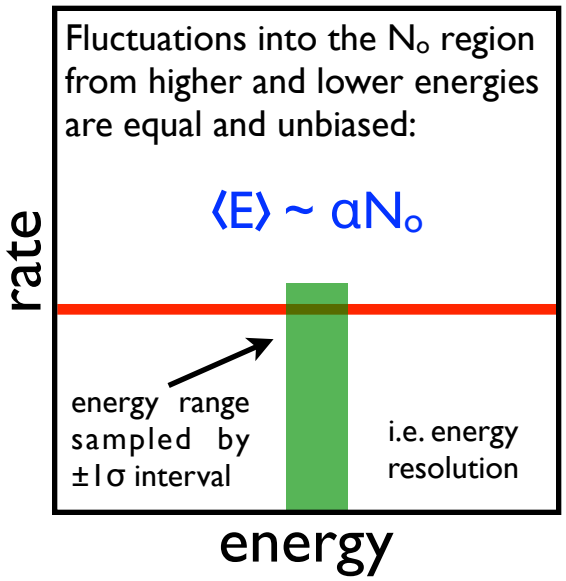
Blue and Black

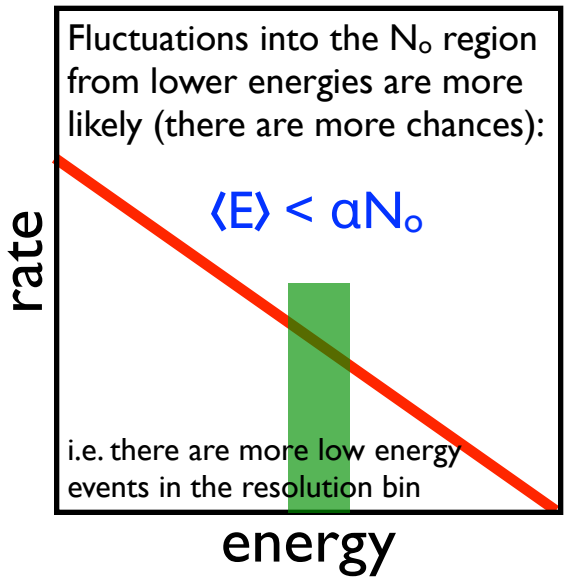
White and Gold

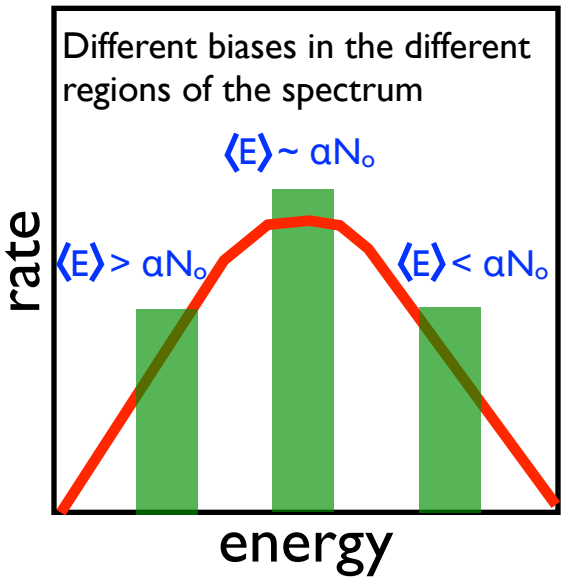
Another example:

Charged particles produce light as they pass through plastic scintillators, which can be detected by photomultiplier tubes and used as an estimator for the energy deposition. Say that that you calibrate such an instrument using known gamma line energies from various radioactive sources and determine that the energy can be very well described by taking the mean number (N) of detected photons (drawn from a Gaussian distribution of width σ) and multiplying it a proportionality constant, α.

Now you measure emission from some continuous spectrum and detect N_o photons from an interaction. What is the best estimate of the gamma ray energy?







Relating data to model parameters **<u>requires</u>** a context (i.e a prior)!

Any inference about models based on an observation is an inherently Bayesian undertaking as it requires an assessment of the posterior probability $P(H_i|D)$ and, thus, **requires** the choice of a prior!

This is rarely ~~often not~~ appreciated! The assumption that the relative likelihoods for two hypotheses alone is the same as the betting odds for which hypothesis is correct tacitly assumes an odds ratio of 1.

If there is an ambiguity in the choice of prior that can lead to notably different conclusions, you should show this!

Example:

As the result of a random blood test, you are diagnosed with "Saturday Night Fever," a disease suffered by 0.5% of the population that results in convulsions when exposed to anything associated with John Travolta. The blood test reliably diagnoses the disease in 80% of cases and yields a false positive 5% of the time. Should you avoid listening to BeeGees albums?

$$P(SNF \mid B) = \frac{P(B \mid SNF)P(SNF)}{P(B \mid SNF)P(SNF) + P(B \mid no\ SNF)P(no\ SNF)}$$

$$= \frac{(0.8)(0.005)}{(0.8)(0.005) + (0.05)(0.995)} = 0.074$$

What if the reason you went to your GP for a blood test was that you got splitting headaches whenever someone mentioned the word "Grease?"

These are basically the same numbers as for COVID-19 (early Oct 2020).

What if you feel ill and get a positive test?

Say the average person is typically ill 10 days per year, so the odds of currently being ill from the common cold is ~10/365 = 0.027. With social distancing, reduce this by a factor of ~10 to 0.0027. So, the fraction of people feeling ill that have COVID-19 is perhaps something like 0.005/(0.005+0.0027) = 0.65 (this, then, is the prior instead of 0.005).

$$P(CV19|+T) = \frac{P(+T|CV19)P(CV19)}{P(+T|CV19)P(CV19) + P(+T|no\ CV19)P(no\ CV19)}$$

$$= \frac{(0.8)(0.65)}{(0.8)(0.65) + (0.05)(0.35)} = 0.97$$

Priors are important!

Example 2:

Atmospheric neutrinos result from the decay of charged pions produced by hadronic interactions in the atmosphere. The characteristic decay sequences are:

$$\pi^- \rightarrow \mu^- + \overline{\nu}_\mu$$
$$\hookrightarrow e^- + \nu_\mu + \overline{\nu}_e$$

$$\pi^+ \rightarrow \mu^+ + \nu_\mu$$
$$\hookrightarrow e^+ + \overline{\nu}_\mu + \nu_e$$

You are detecting these neutrinos coming from directly overhead with an underground water Cherenkov detector. From the fuzziness of the ring pattern of observed light from a particular event, simulations tell you that 70% of ν_e's will produce a ring at least this fuzzy, whereas only 50% of ν_μ's will do this. What is the probability that this event is a ν_e ?

$$P(\nu_e|R) = \frac{P(R|\nu_e)P(\nu_e)}{P(R|\nu_e)P(\nu_e) + P(R|\nu_\mu)P(\nu_\mu)}$$

$$= \frac{(0.7)(1/3)}{(0.7)(1/3) + (0.5)(2/3)} = 0.41$$

Bernstein – von Mises Theorem

In the limit of an infinitely large data set, the posterior probability is independent of the exact form of the prior probability.

(the likelihood function that multiplies the prior crushes it's impact away from the region of interest)

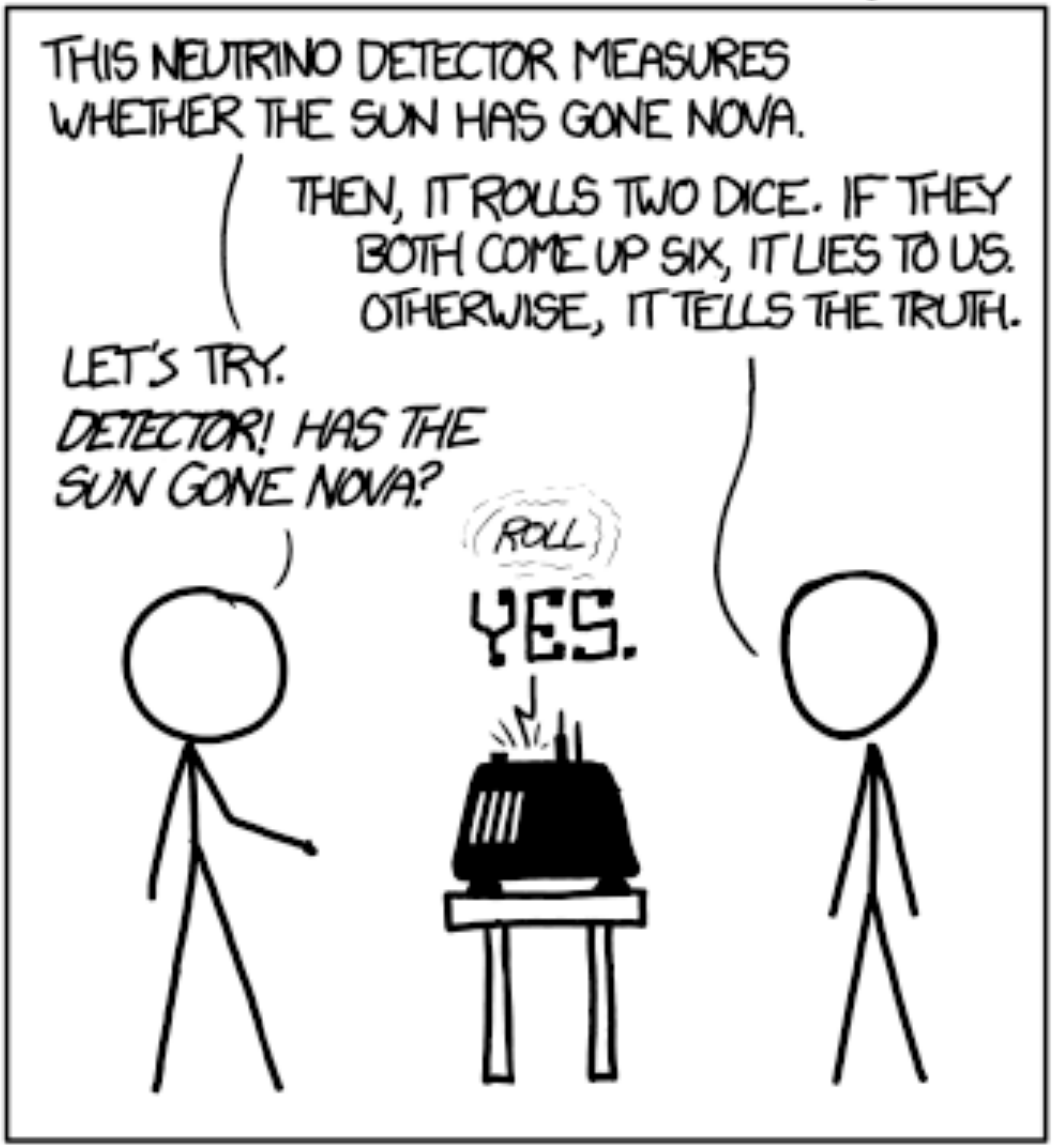
For example, if you instead asked for the probability for a large number Cherenkov events to be ν_e out of a big data set, the information contained in the distribution of ring fuzziness within the data itself carries more weight than the form of any previously assumed prior.

Priors carry greater weight for weaker data sets

http://xkcd.com/1132/

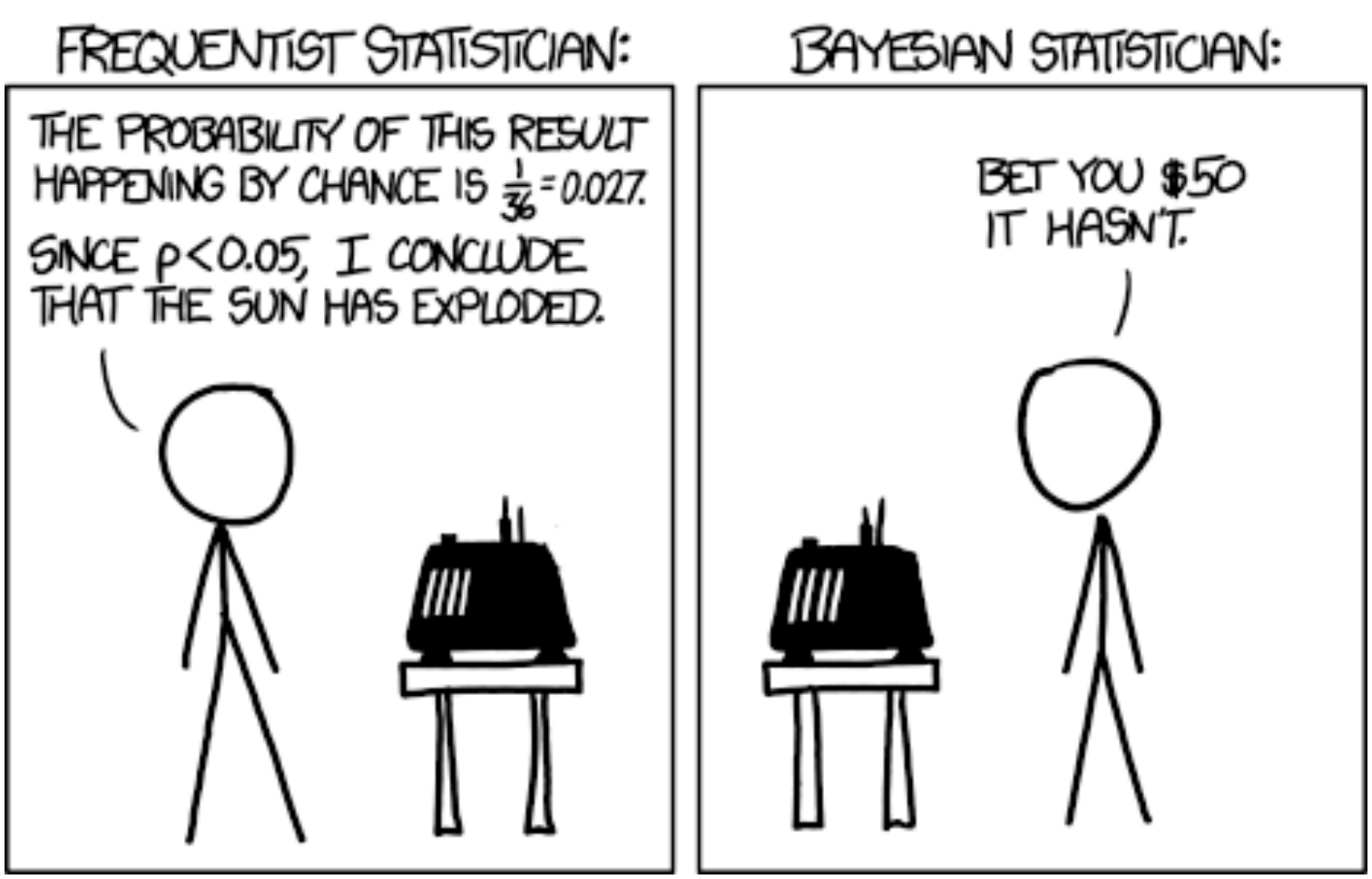
DID THE SUN JUST EXPLODE?

(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.