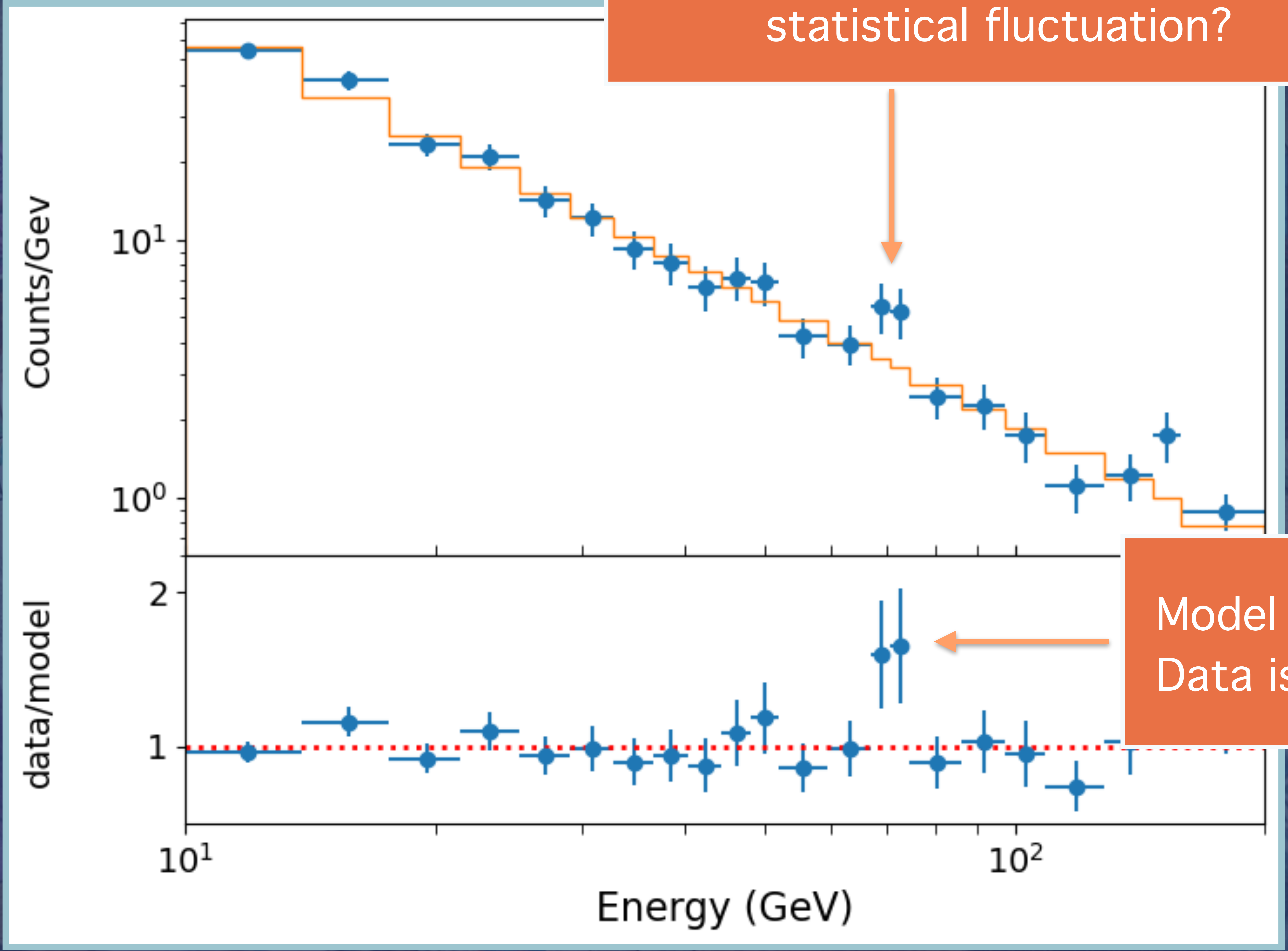
Model Comparison

DANIELA HUPPENKOTHEN

Universiteit van Amsterdam

d.huppenkothen@uva.nl

Is there an emission line? Or is it a statistical fluctuation?



Model misspecification? Data issues?

All models are wrong, but some are useful.

— George Box

Model comparison: which of models A or B was more probable to have generated the data?

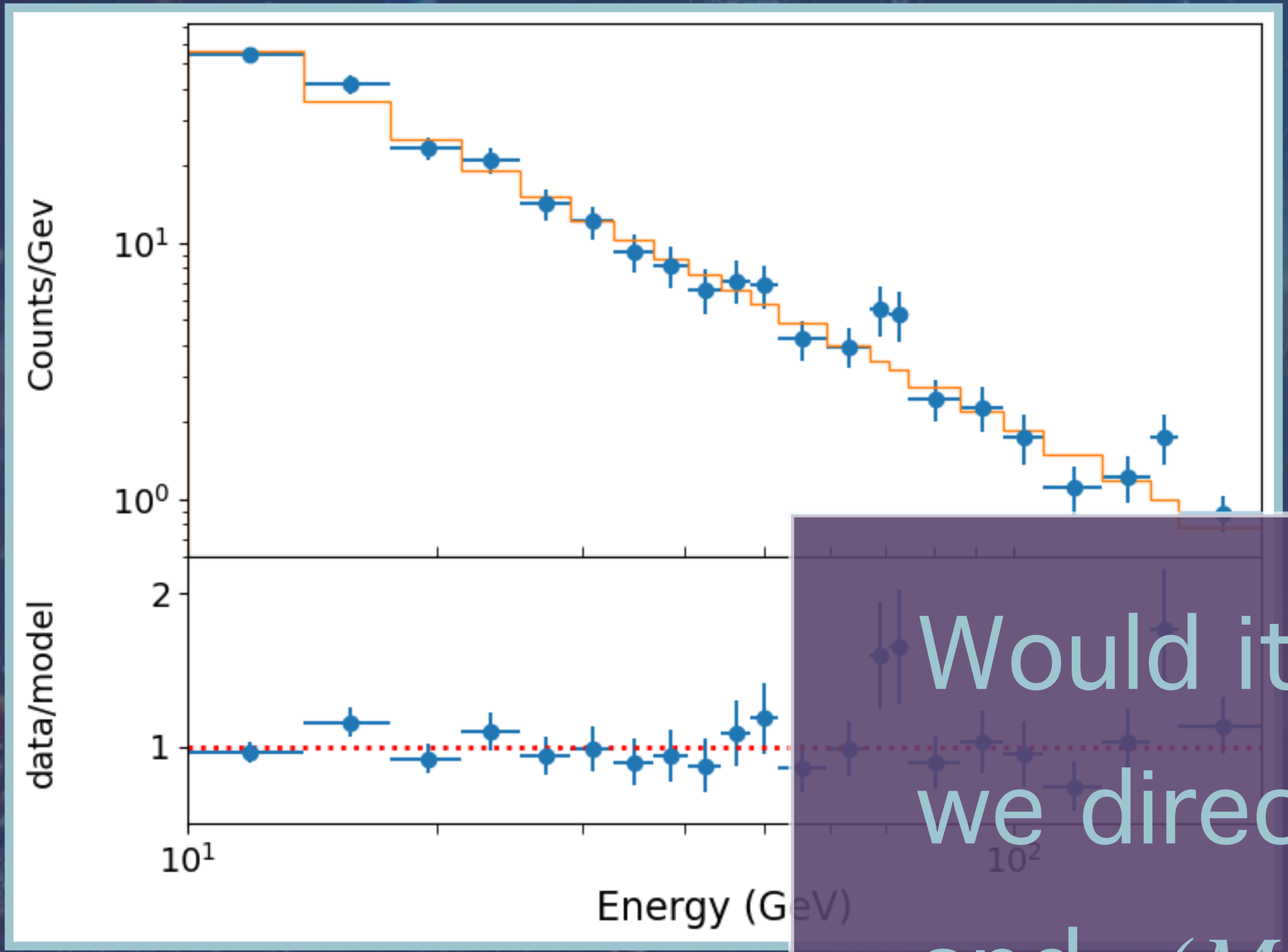
Selected Model Comparison Methods

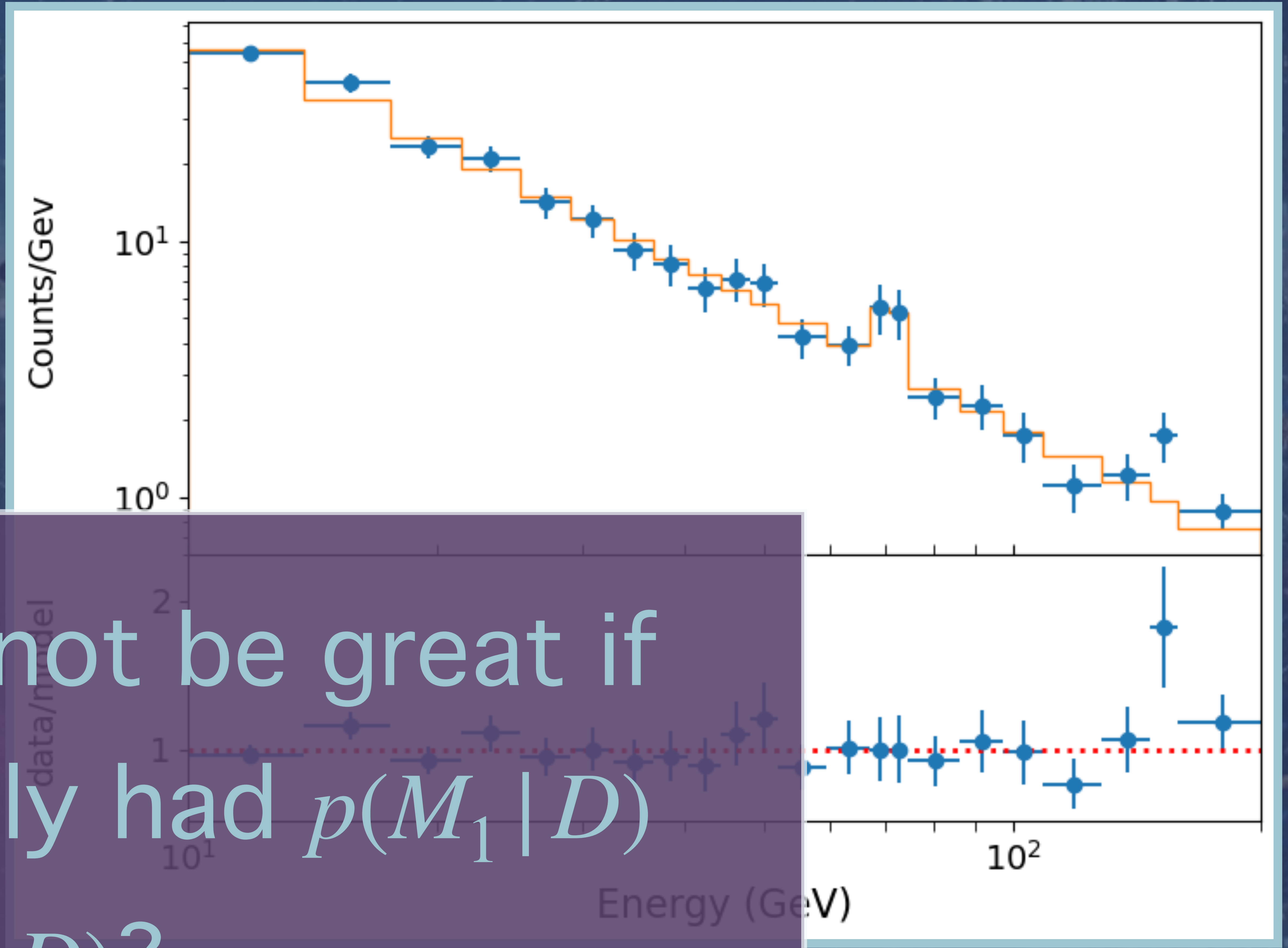
- **Hypothesis testing**
- **Information Criteria**
- **Posterior Odds (today)**

👻 Model comparison is really hard! 👻

(If you can reframe your problem as a parameter estimation problem, you should!)

Which is better?





Would it not be great if we directly had $p(M_1 | D)$ and $p(M_2 | D)$?

$$\boxed{P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}}$$

Bayes rule

We kind of cheated here!

$\theta =$ parameters

$D =$ data

$$\boxed{P(\theta \mid D, I) = \frac{P(D \mid \theta, I)\,P(\theta \mid I)}{P(D \mid I)}}$$

$\theta =$ parameters

$D =$ data

$I =$ other assumptions

$$P(\theta \mid D, M, I_{\text{rest}}) = \frac{P(D \mid \theta, M, I_{\text{rest}})\, P(\theta \mid M, I_{\text{rest}})}{P(D \mid M, I_{\text{rest}})}$$

$\theta =$ parameters

$D =$ data

$I = (M, I_{\text{rest}})$ other assumptions

$$P(\theta | D, M) = \frac{P(D | \theta, M) P(\theta | M)}{P(D | M)}$$

$\theta =$ parameters

$D =$ data

$M =$ model choice

Likelihood of the model, integrated over all parameters!

$$P(M \mid D, I) = \frac{P(D \mid M, I) P(M \mid I)}{P(D \mid I)}$$

$D =$ data

$M =$ model choice

$$\frac{p(M_1 \mid D, I)}{p(M_0 \mid D, I)} = \frac{p(D \mid M_1, I)p(M_1 \mid I)}{p(D \mid M_0, I)p(M_0 \mid I)}$$

Posterior odds

Bayes factor

$D =$ data

$M_1, M_0 =$ The models to be compared

Bayesian Model Comparison

… allows you to directly compare the relative odds of two hypotheses

… also enables you to put sensible priors on both models

But: Bayes factors are very sensitive to the parameter priors of each model

Bayes factors involve intractable integrals in high dimensions and are thus very challenging to compute!

$$p(D|M) = \int_{-\infty}^{\infty} p(D|\theta, M)p(\theta|M)d\theta$$

$\theta =$ parameters

$D =$ data

$M =$ model choice

We've been ignoring this during MCMC! 😬

Nested Sampling

Further reading:

- https://arxiv.org/abs/2101.09675
- https://www.nature.com/articles/s43586-022-00121-x

Sampling method to compute compute integral $p(D|M)$:

1) Draw points from prior
2) Pick point(s) with lowest likelihood
3) Draw contours at that likelihood surface
4) Sample new point(s) from within those contours
5) Repeat from step 2

https://chi-feng.github.io/mcmc-demo/app.html