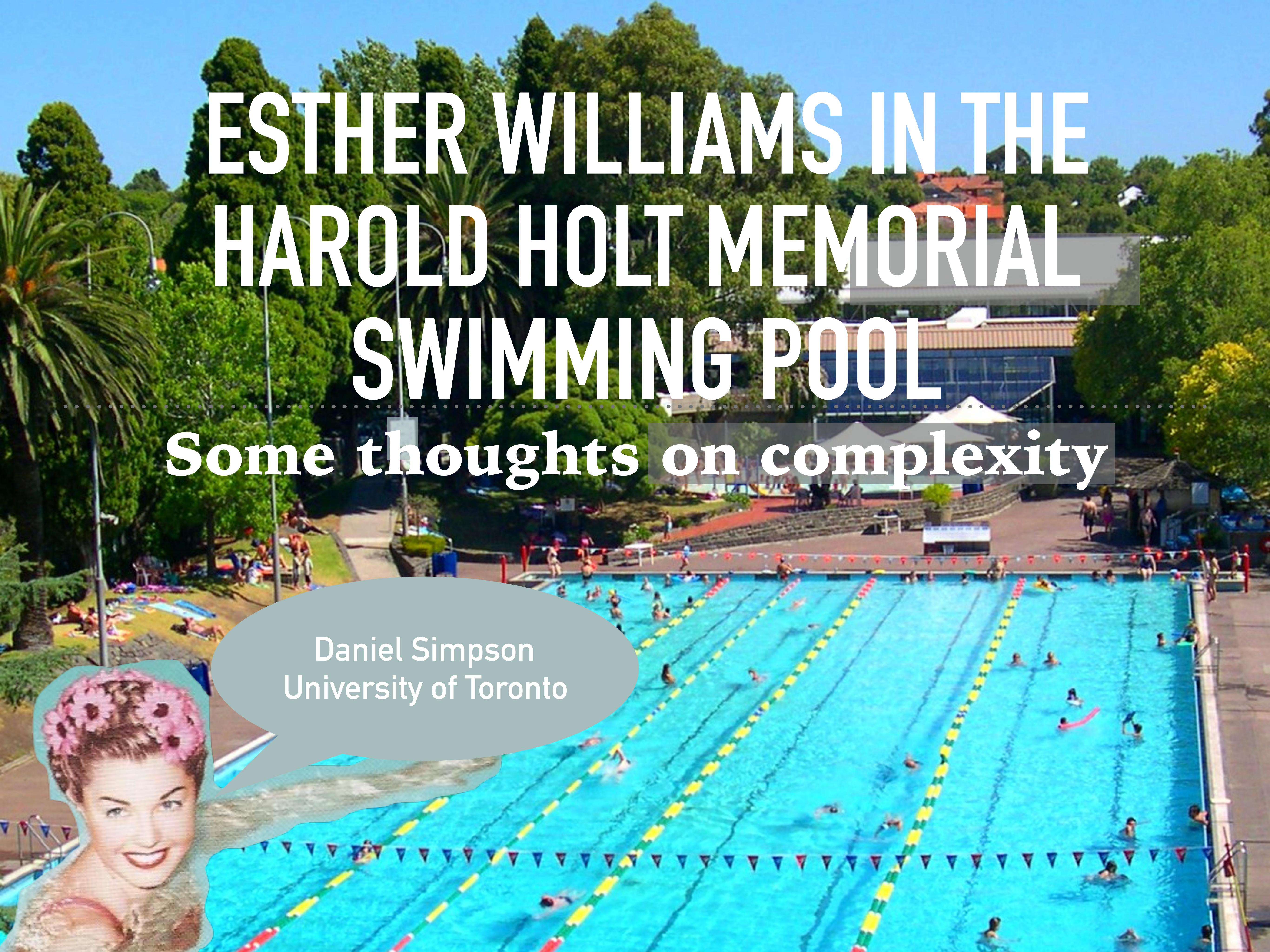
ESTHER WILLIAMS IN THE HAROLD HOLT MEMORIAL SWIMMING POOL

Some thoughts on complexity



Daniel Simpson
University of Toronto

HAROLD HOLT, THE MUSICAL (DRAMATIS PERSONAE)

- Harold Holt (17th Prime Minister of Australia)
  - Our metaphor for statisticians

HAROLD HOLT, THE MUSICAL (DRAMATIS PERSONAE)

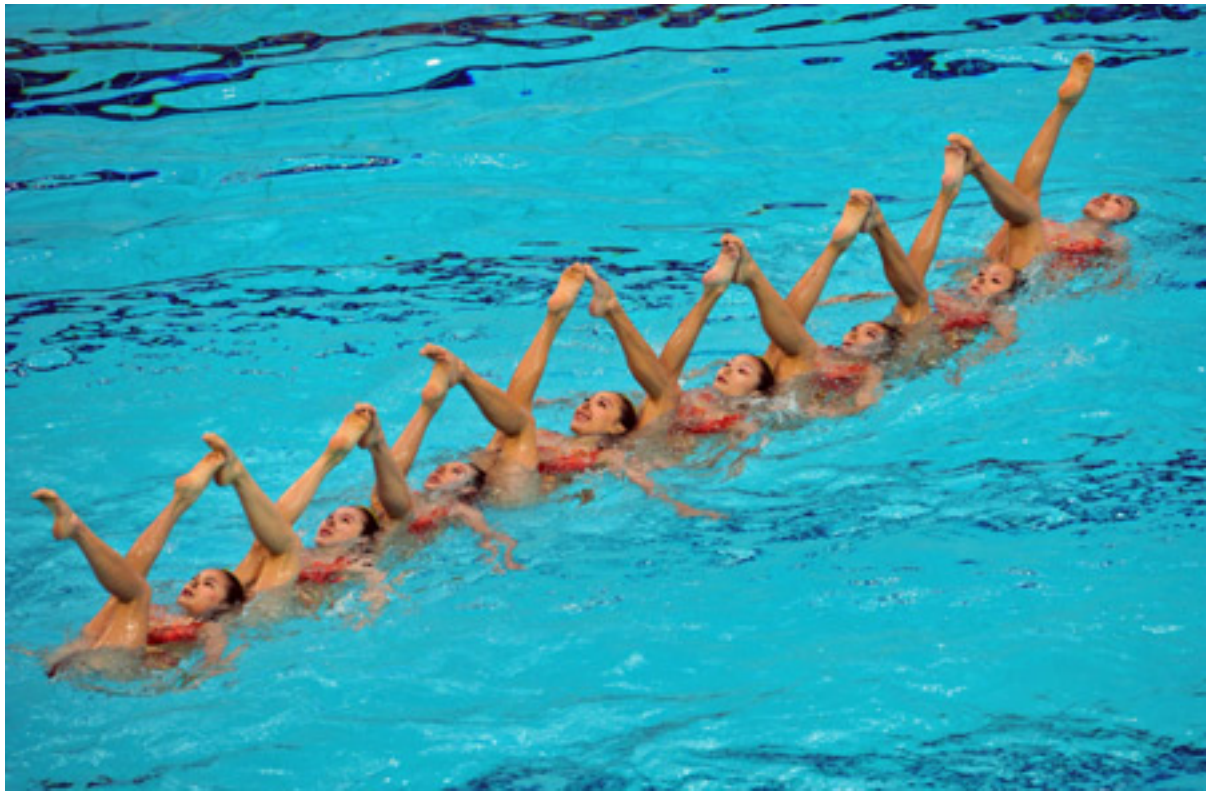
- Harold Holt (17th Prime Minister of Australia)
  - (Our metaphor for *statisticians*)
- Harold Holt Memorial Swimming Pool (A swimming pool)
  - (*Things we report from a statistical analysis)*
- The Bass Strait (A large body of water)
  - *(A dangerous sea of statistical methods)*
- Esther Williams (Esther Williams)
  - *(A synchronized swimmer)*

GENERALLY SPEAKING, THINGS HAVE GONE ABOUT AS FAR AS THEY CAN POSSIBLY GO, WHEN THINGS HAVE GOTTEN ABOUT AS BAD AS THEY CAN REASONABLY GET.

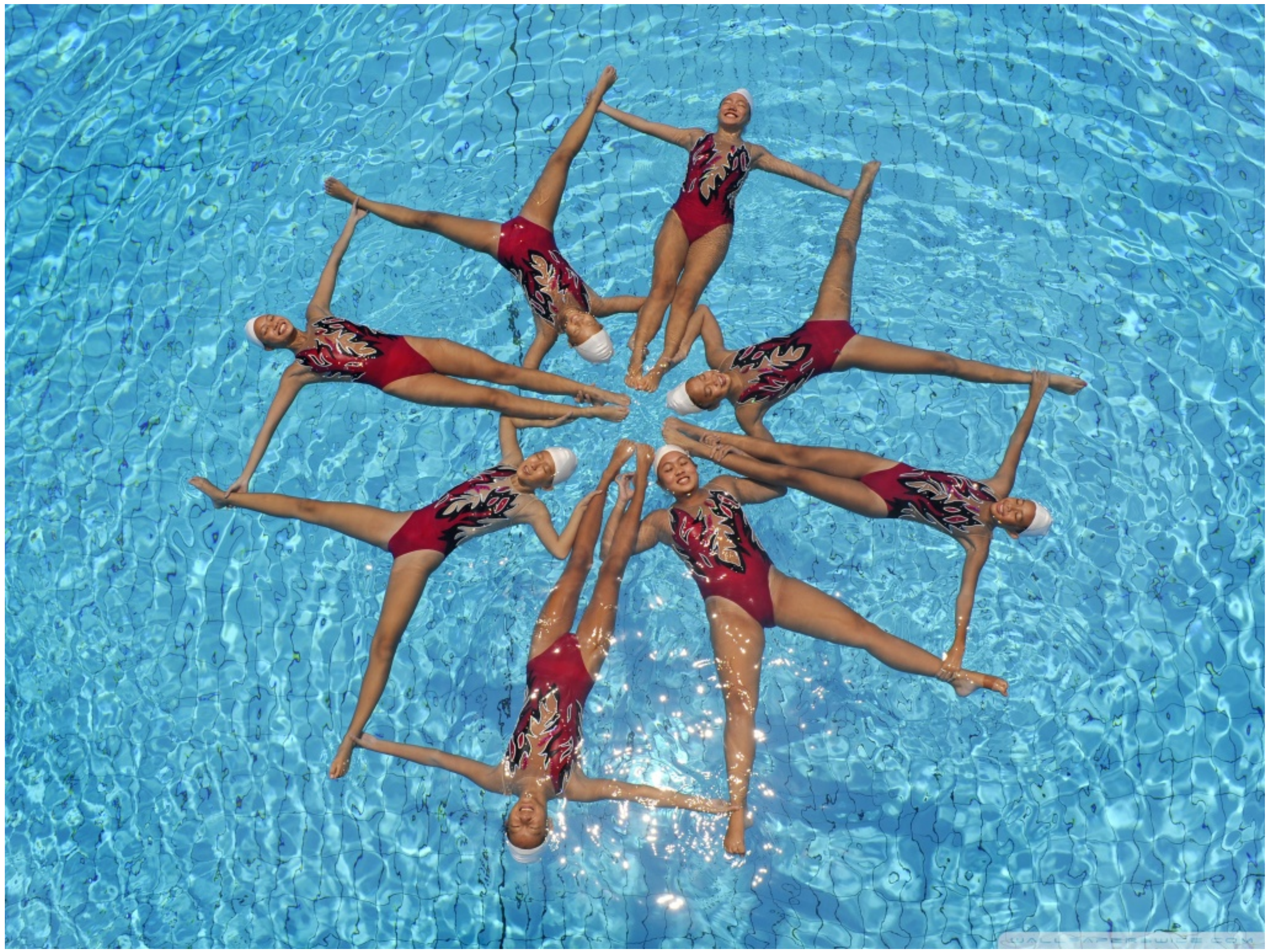
(Tom Stoppard)

WE USED TO JUST ESTIMATE MEANS OF GAUSSIANS

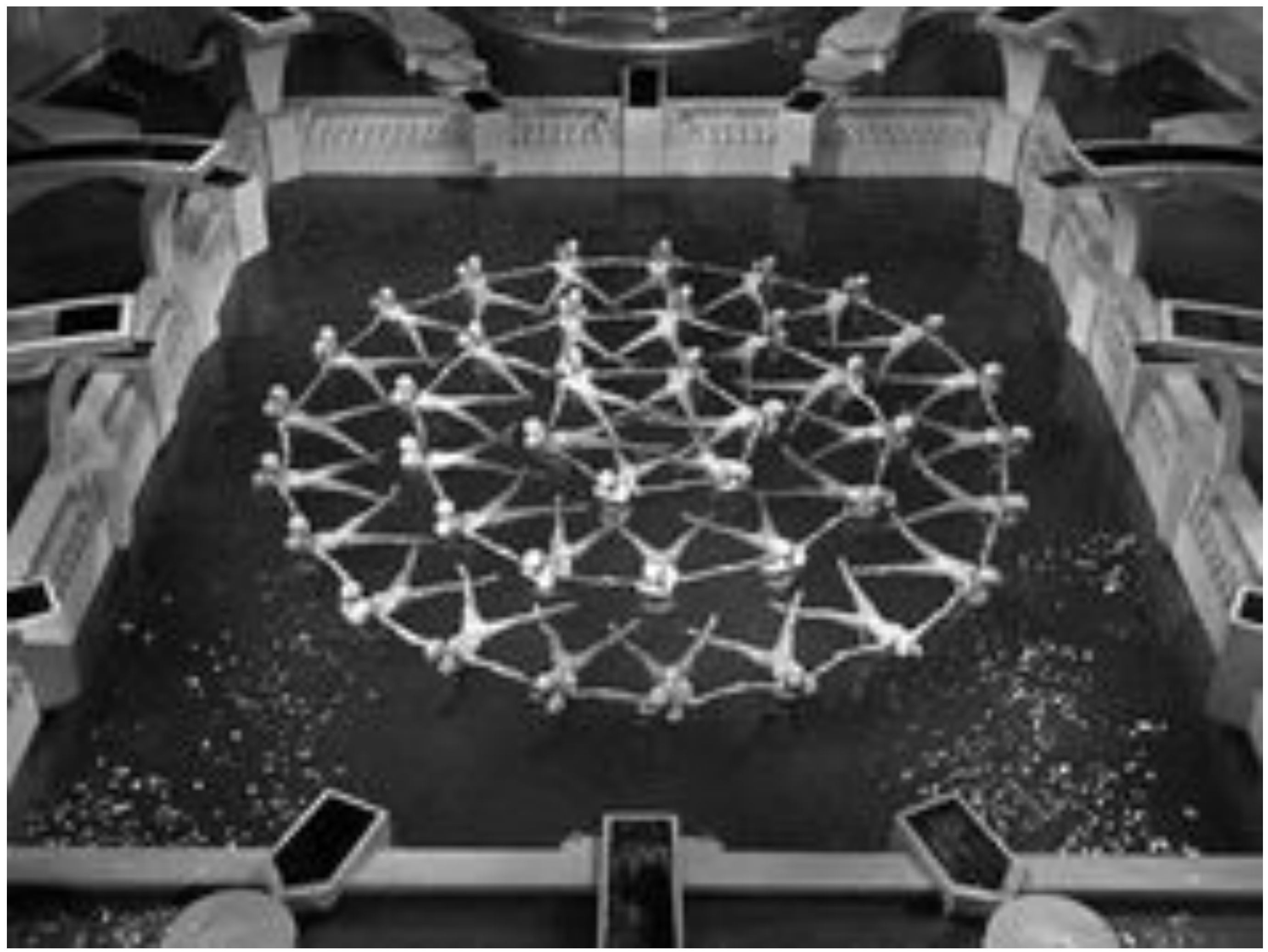
THEN THE MCMC REVOLUTION CHANGED EVERYTHING



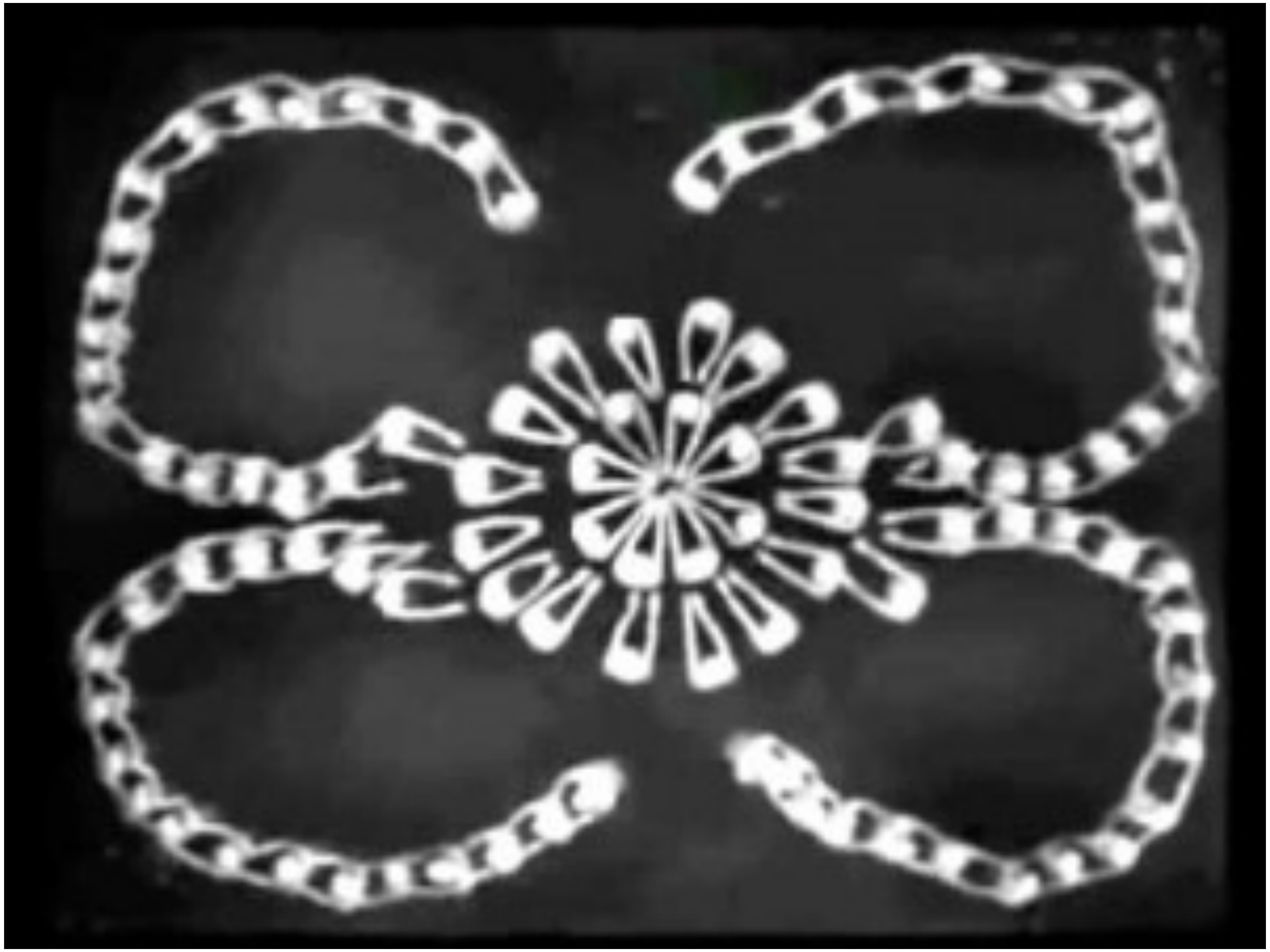
BUGS CAME ALONG AND REDEFINED THE POSSIBLE



METHODS LIKE INLA HELPED US SCALE UP



BUT THEN STAN CAME ALONG



GOD IS PRESENT IN THE SWEEPING GESTURES, BUT THE DEVIL IS IN THE DETAILS

A PARTIAL ORDER OF MASSIVE ASSUMPTIONS

Data gathering

Asymptotic regime

Model evaluation criteria

Likelihood

Prior

Computation

THE GREAT LIE OF STATISTICS

- Once the models get complex, we don't really know much about how they work.
- We can sometimes say some things about how things will work "eventually"
- But even that is limited to either essentially useless qualitative statements or very simple models

THEOREM 2.1. *Suppose that for a sequence ε_n with $\varepsilon_n \to 0$ and $n\varepsilon_n^2 \to \infty$, a constant $C > 0$ and sets $\mathscr{P}_n \subset \mathscr{P}$, we have*

$$\log D(\varepsilon_n, \mathscr{P}_n, d) \le n\varepsilon_n^2, \tag{2.2}$$

$$\Pi_n(\mathscr{P} \setminus \mathscr{P}_n) \le \exp(-n\varepsilon_n^2(C+4)), \tag{2.3}$$

$$\Pi_n\left(P:\ -P_0\left(\log\frac{p}{p_0}\right) \le \varepsilon_n^2,\ P_0\left(\log\frac{p}{p_0}\right)^2 \le \varepsilon_n^2\right) \ge \exp(-n\varepsilon_n^2 C). \tag{2.4}$$

Then for sufficiently large M, we have that $\Pi_n(P: d(P, P_0) \ge M\varepsilon_n | X_1, \dots, X_n) \to 0$ in P_0^n-probability.

Ghosal, Ghosh, and van der Vaart
Convergence rates of posterior distributions (2000)

THE GREAT LIE OF COMPUTATIONAL STATISTICS

- To do Bayesian statistics is to have long practical experience of pre-asympototic behaviour
- This was especially true with BUGS and JAGS, but is also true with Stan
- Because MCMC methods only ever converge asymptotically, so we are typically drawing inference from a biased chain

PICTURES AND FEAR

- So if we don't really have sharp enough theory to understand how our inference works, and we don't really have sharp enough theory to guarantee our computation works, what do we have?

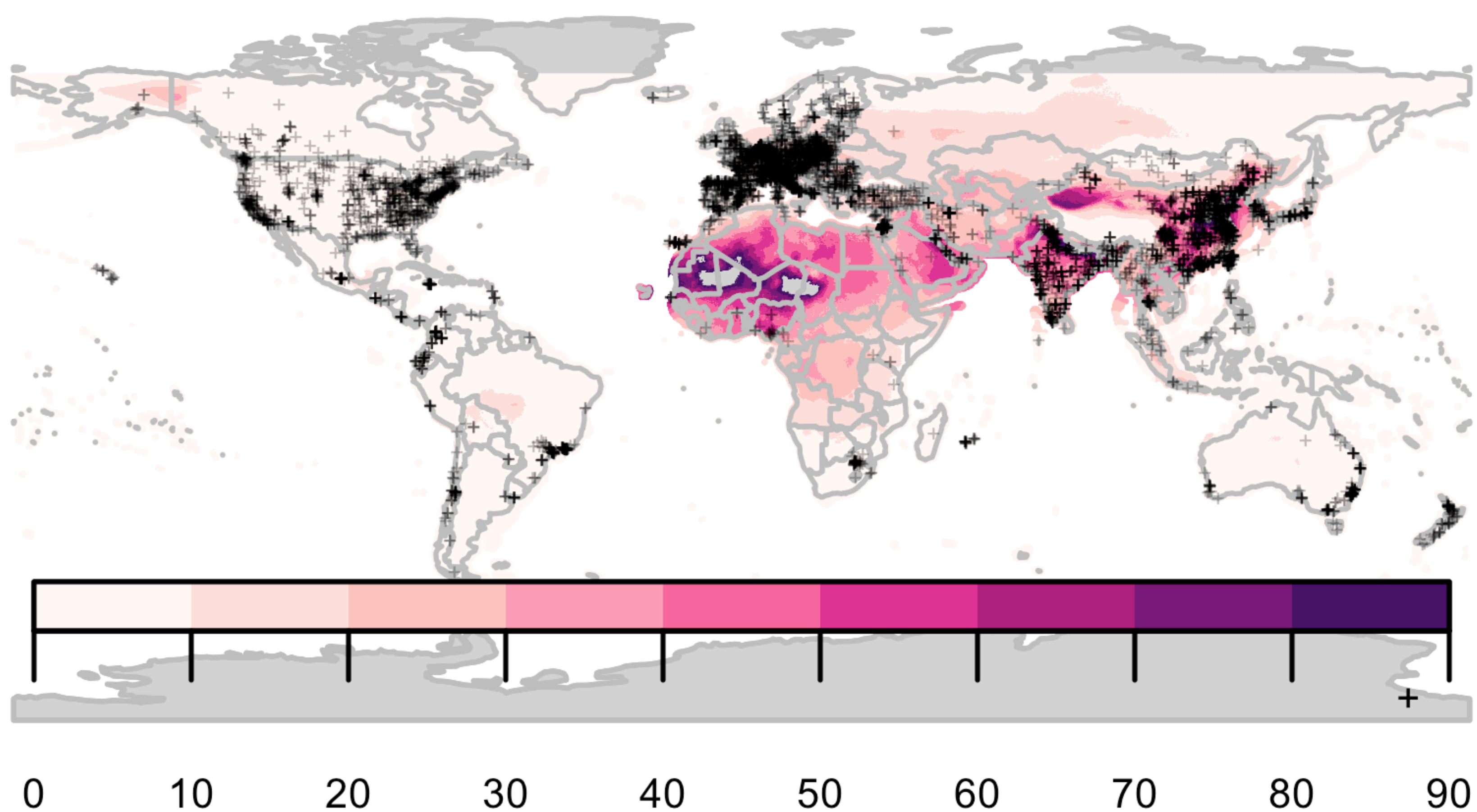
COMPUTING THE WRONG THING PERFECTLY IS NOT AS USEFUL AS YOU'D THINK

AS ALWAYS, BRITNEY SPEARS WAS AHEAD OF THE GAME

WHEN KYLIE SAID "BREATHE" THIS WASN'T WHAT SHE WANTED

Goal Estimate global PM2.5 concentration

Problem Most data from noisy satellite measurements (ground monitor network provides sparse, heterogeneous coverage)



Satellite estimates of PM2.5 and ground monitor locations

ARIANISM WAS A HERESY FOR A REASON

- Many are taught that the likelihood is the fundamental building block of a Bayesian model and the prior is a secondary object
- This is a very limiting view.
- In reality, we build a **joint distribution** for the data and the likelihood
- People who don't do this (like people who use reference priors) are making some heavy assumptions
- (and, in this analogy, are heretics but don't worry so much about that)

Gelman, A., Simpson, D., and Betancourt, M. (2017).
The prior can often only be understood in the context of the likelihood.
arXiv preprint: arxiv.org/abs/1708.07487

THE MAJESTY OF GENERATIVE MODELS

- If we disallow improper priors, then Bayesian modelling is generative.
- In particular, we have a simple way to simulate from $p(y)$:
  - Simulate $\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta})$
  - Simulate $\mathbf{y}^* \sim p(\mathbf{y} \mid \boldsymbol{\theta}^*)$
  - (Repeat for each sample)

PRIOR PREDICTIVE CHECKING

What do vague/non-informative priors imply about the data our model can generate?

$$\log(\mathrm{PM}_{2.5})_i = \alpha_i + \beta_i \log(\mathrm{sat}_i) + \epsilon_i$$

$$\alpha_j \sim N(\alpha_0, \tau_\alpha^2)$$

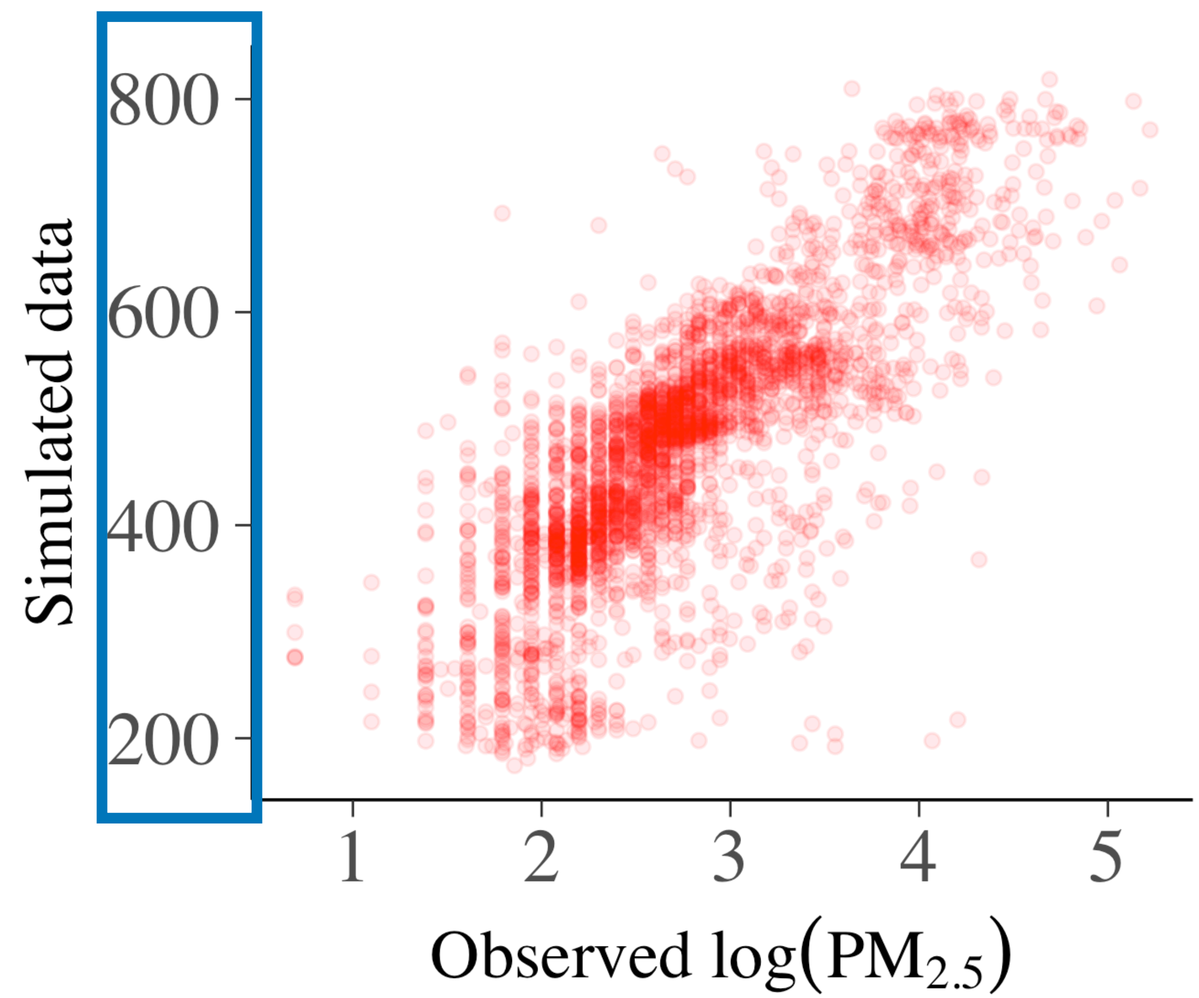
$$\beta_j \sim N(\beta_0, \tau_\beta^2)$$

$$\alpha_0 \sim N(0, 100)$$

$$\beta_0 \sim N(0, 100)$$

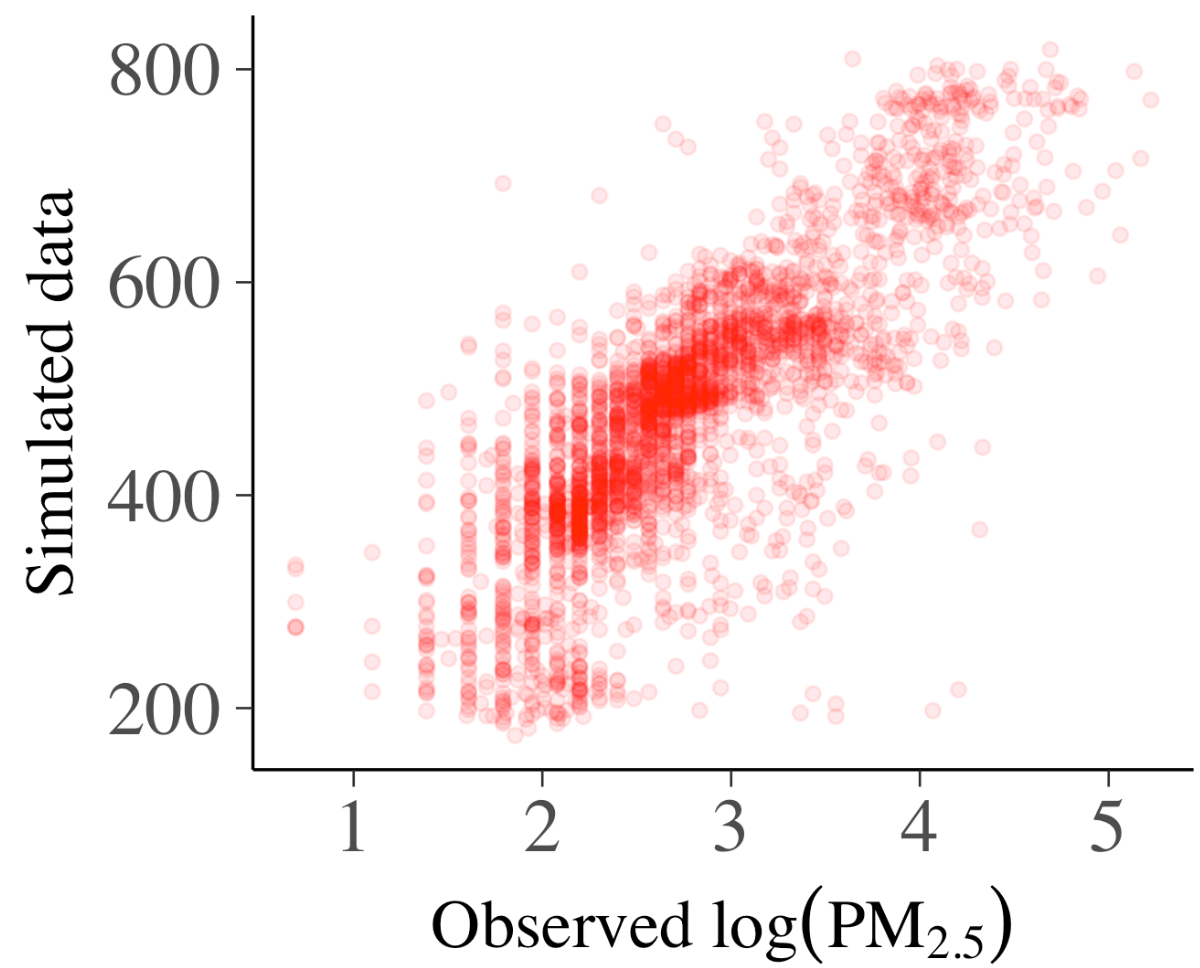
$$\tau_\alpha^2 \sim \mathrm{InvGamma}(1, 100)$$

$$\tau_\beta^2 \sim \mathrm{InvGamma}(1, 100)$$



WAIT! WHAT?

- The prior model is two orders of magnitude off the real data
- Two orders of magnitude on the log scale!
- Log density of neutron star only 60 $\mu g m^{-3}$!!
- What does this mean practically?
- The data will have to overcome the prior…



IT CAN GUIDE YOUR CHOICE OF PRIOR

What are better priors for the global intercept and slope and the hierarchical scale parameters?

$$\log(\mathrm{PM}_{2.5})_i = \alpha_i + \beta_i \log(\mathrm{sat}_i) + \epsilon_i$$

$$\alpha_j \sim N(\alpha_0, \tau_\alpha^2)$$

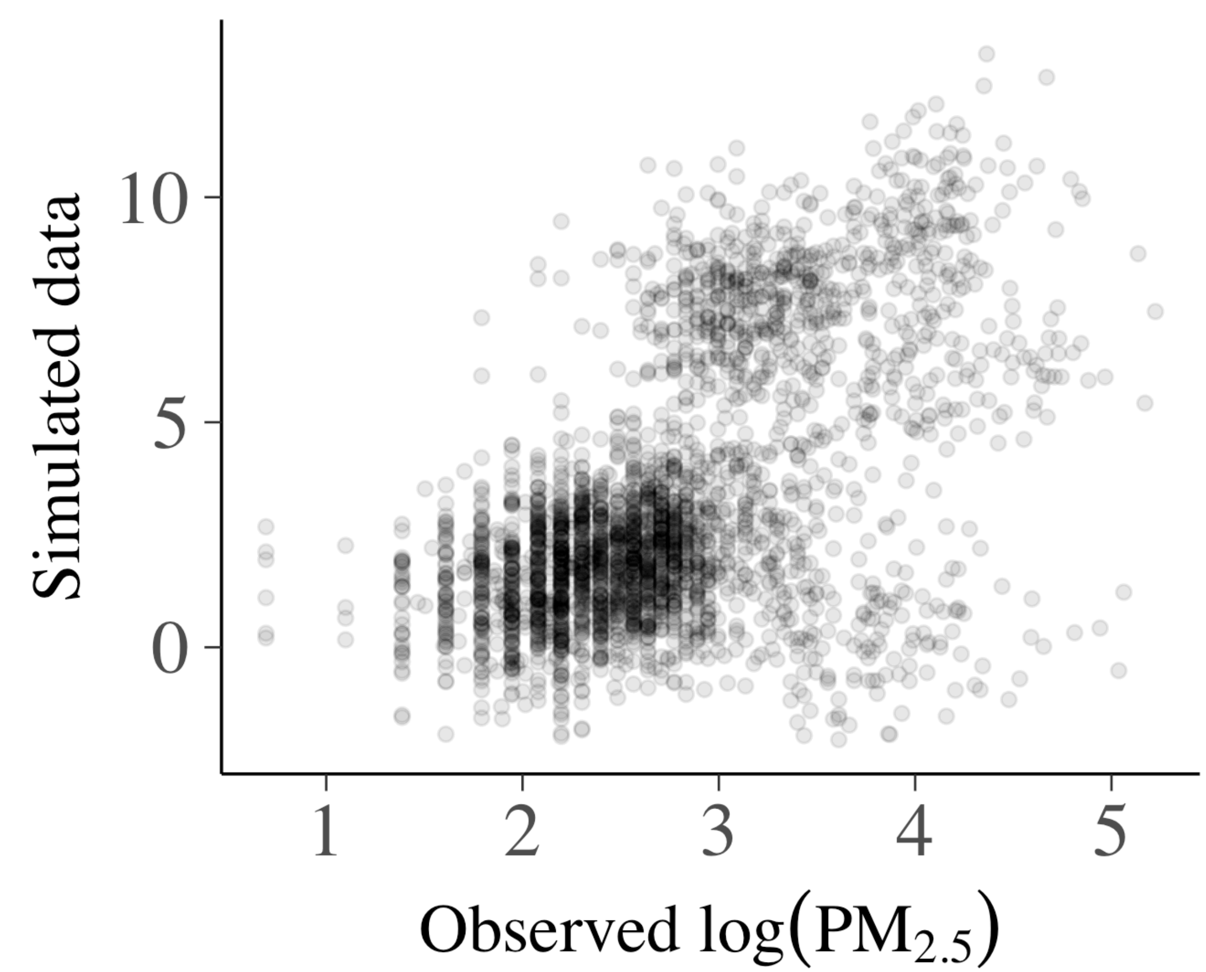
$$\beta_j \sim N(\beta_0, \tau_\beta^2)$$

$$\alpha_0 \sim N(0, 1)$$

$$\beta_0 \sim N(1, 1)$$

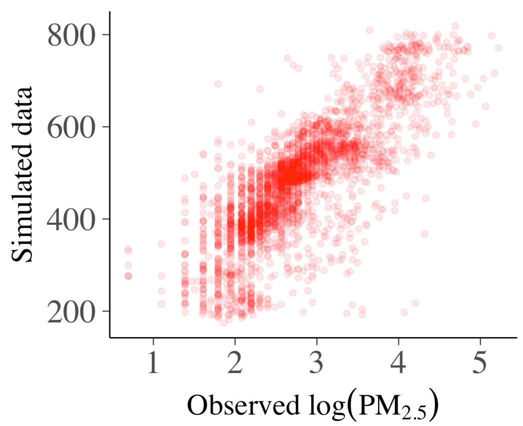
$$\tau_\alpha \sim N_+(0, 1)$$

$$\tau_\beta \sim N_+(0, 1)$$

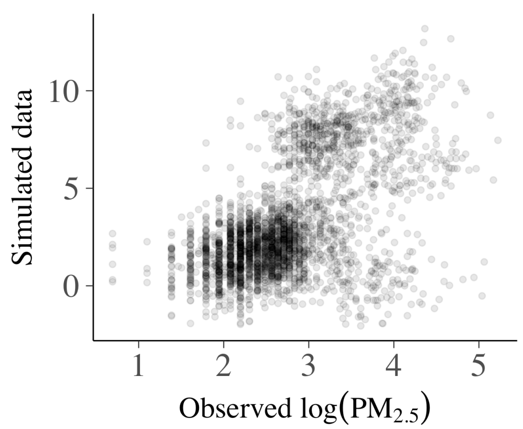


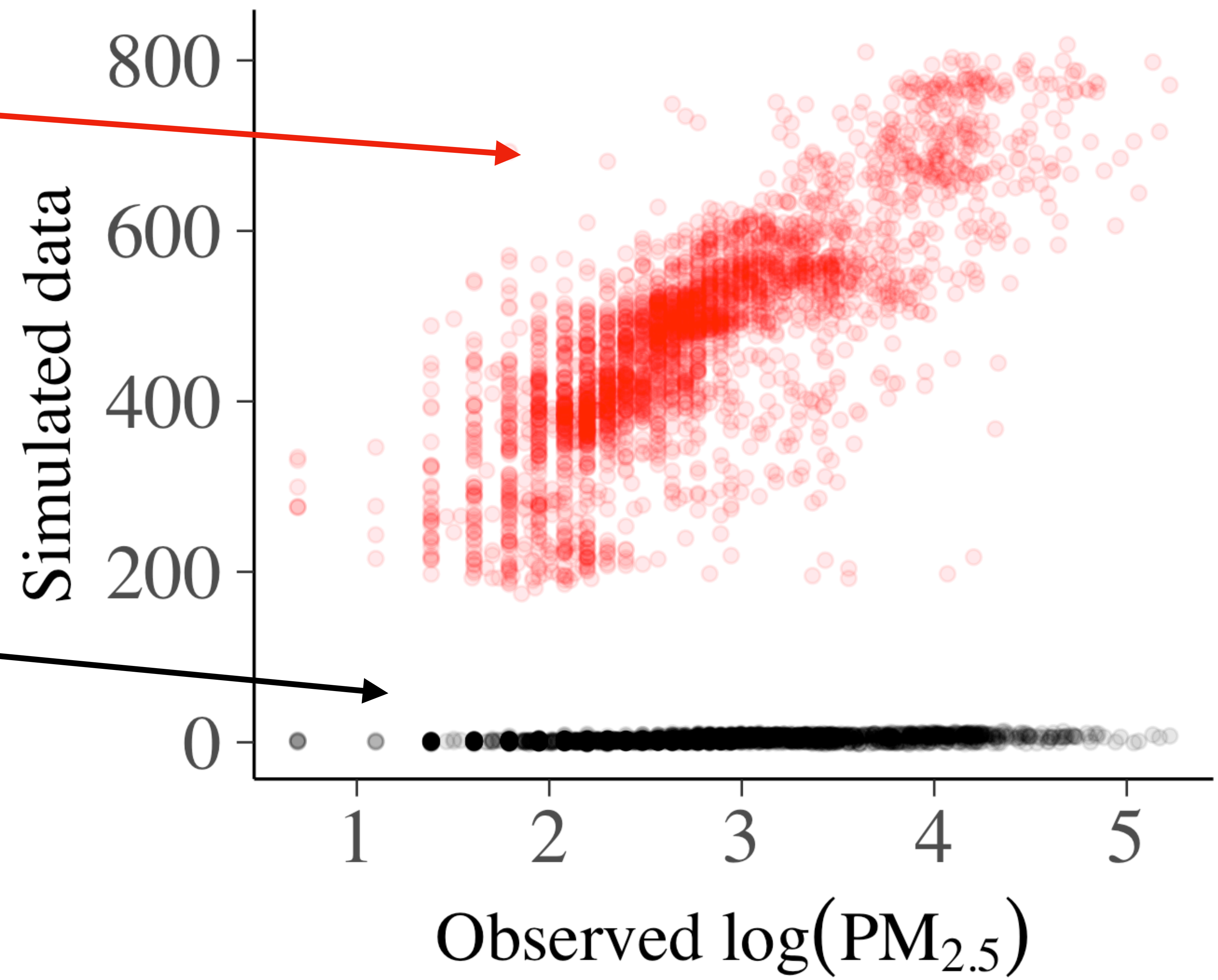
AND MAKE IT EASIER TO DEFEND YOUR MODELLING CHOICES

Non-informative

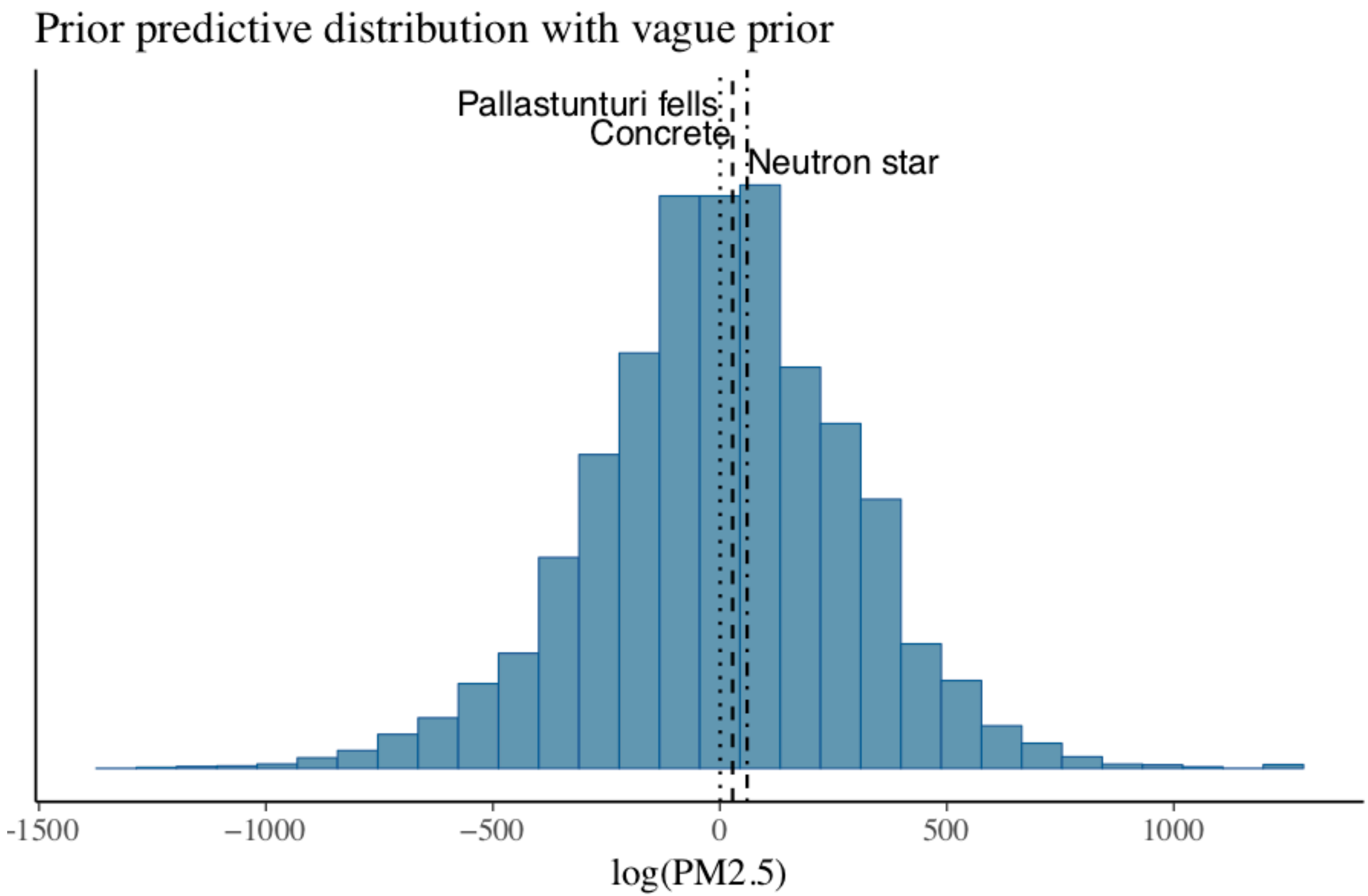


Weakly informative

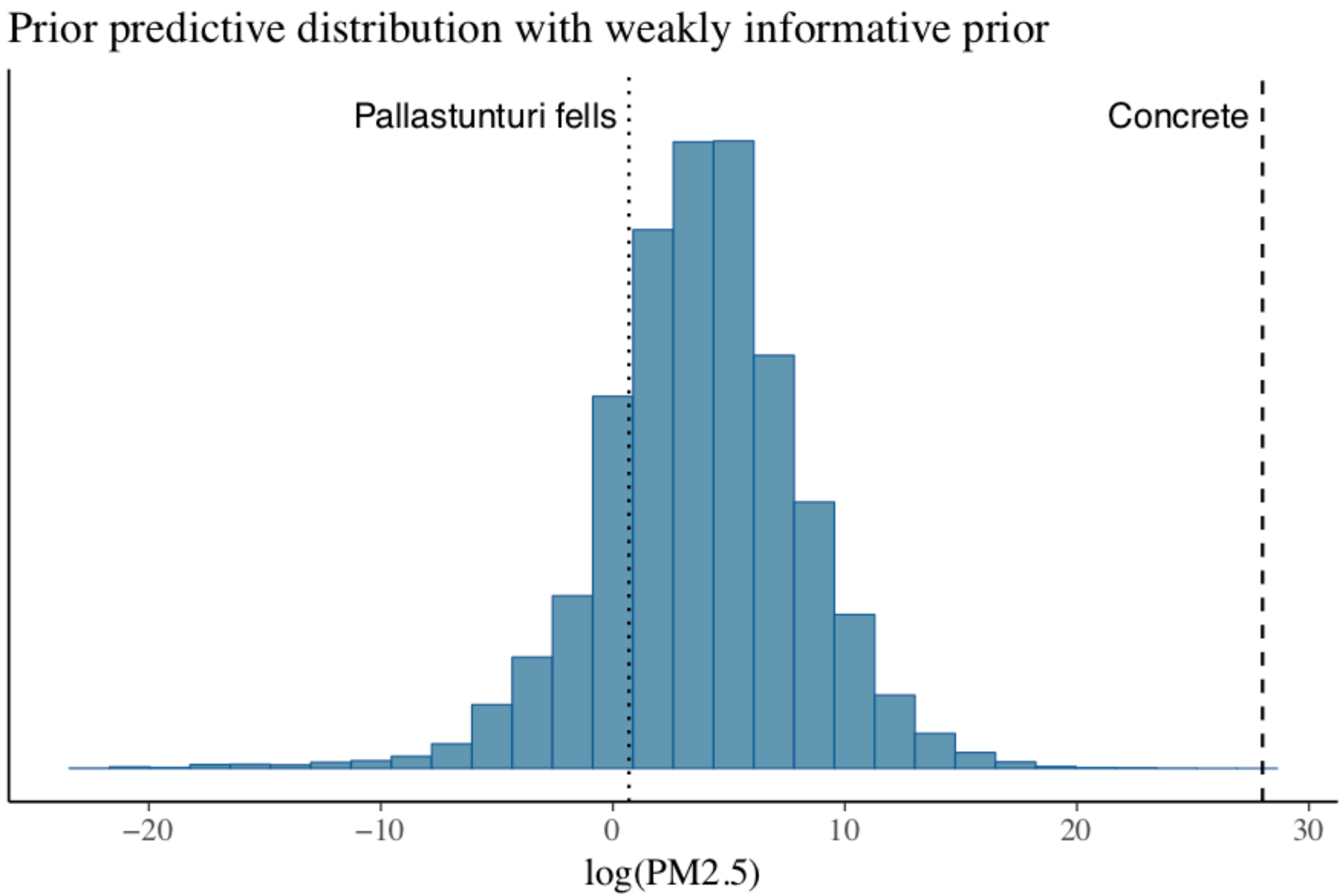




AND NOW, WITHOUT THE DATA



MORE REASONABLE PRIORS



SOME THOUGHTS

- We are very bad at reasoning about logarithms. Always check the natural scale!
- This is a GLM, so the natural summary of the problem that we can reason about is the observation
- For more complex models, a lot more substantive knowledge is needed
- Wang, Nott, Drovndi, Mengersen, Evans (2018) use a numerical summary of the predictive distribution as a way to choose priors ("history matching").

PRE-EXPERIMENT PROPHYLAXIS

A THING YOU SHOULD ALWAYS DO

- Just because you think your prior is a good idea, doesn't mean that it will be
- So you have to check!
- Looking at the implied data generating mechanism is just one way to do this
- The other way is to do this is to fit the model to fake data with the features that you think your model can pick up

THE BAYESIAN LASSO - A MODEL THAT DOES NOT WORK

- A nice, clean, safe example of this is the Bayesian Lasso

$$\beta_j \sim \text{Laplace}(\lambda)$$

- Despite it's name, it bares essentially no relationship to the frequentist Lasso and is a terrible sparsity prior
- I know this because I am the sort of person who reads papers written by Dutch asymptoticists
- But there's an easy way

IF I WERE WRITING AN EXAM QUESTION

- Well if we get a sparse signal we need most of the entries to be small ($< \epsilon$) and a few to be large ($> \epsilon$).
- What is the probability of that happening under a Lasso prior?
- Well, if we have p covariates, the number of non-zero entries is *a priori* a $\text{Bin}\left[p, \Pr(|\beta| > \epsilon)\right] = \text{Bin}\left[p, 2e^{-\lambda\epsilon}\right]$ random variable
- So if we want, on average, s_0 non-zeros, we need

$$\lambda \approx \epsilon^{-1} \log\left(\frac{p}{s_0}\right)$$

SO WHAT IS EPSILON?

- Well, if I don't want the "zero" terms to effect the RMSE, I will need $\epsilon = o(p^{-1})$
- So that means $\lambda = o(p^{-1}\log(p))$ is required for the Bayesian Lasso to have *a priori* mass on sparse signals
- But with this λ, $\Pr(|\beta| > 1) = \exp(-p\log p) = p^{-p}$ which is **very** small.
- So this suggests that the prior doesn't support signals that are mostly zero but have some larger values, which makes it inappropriate for sparsity.

WHAT AN ENVELOPE!

- Now this back of the envelope calculation was possible because the Laplace prior is easy to work with.
- It's very hard to do in general, but by the power of Mathematica and a lot of time with Abramowitz and Stegun, you can show that the following prior will pass the "back of the envelope test"

$$\beta_j \sim N(0, \tau_j^2)$$
$$\tau_j \sim p(\tau)$$

as long as τ_j has fewer than 2 moments.

BUT WHY BOTHER WITH MATHS?

- We have computers!
- And we have pictures!
- So maybe we can assess this without all the hard maths.

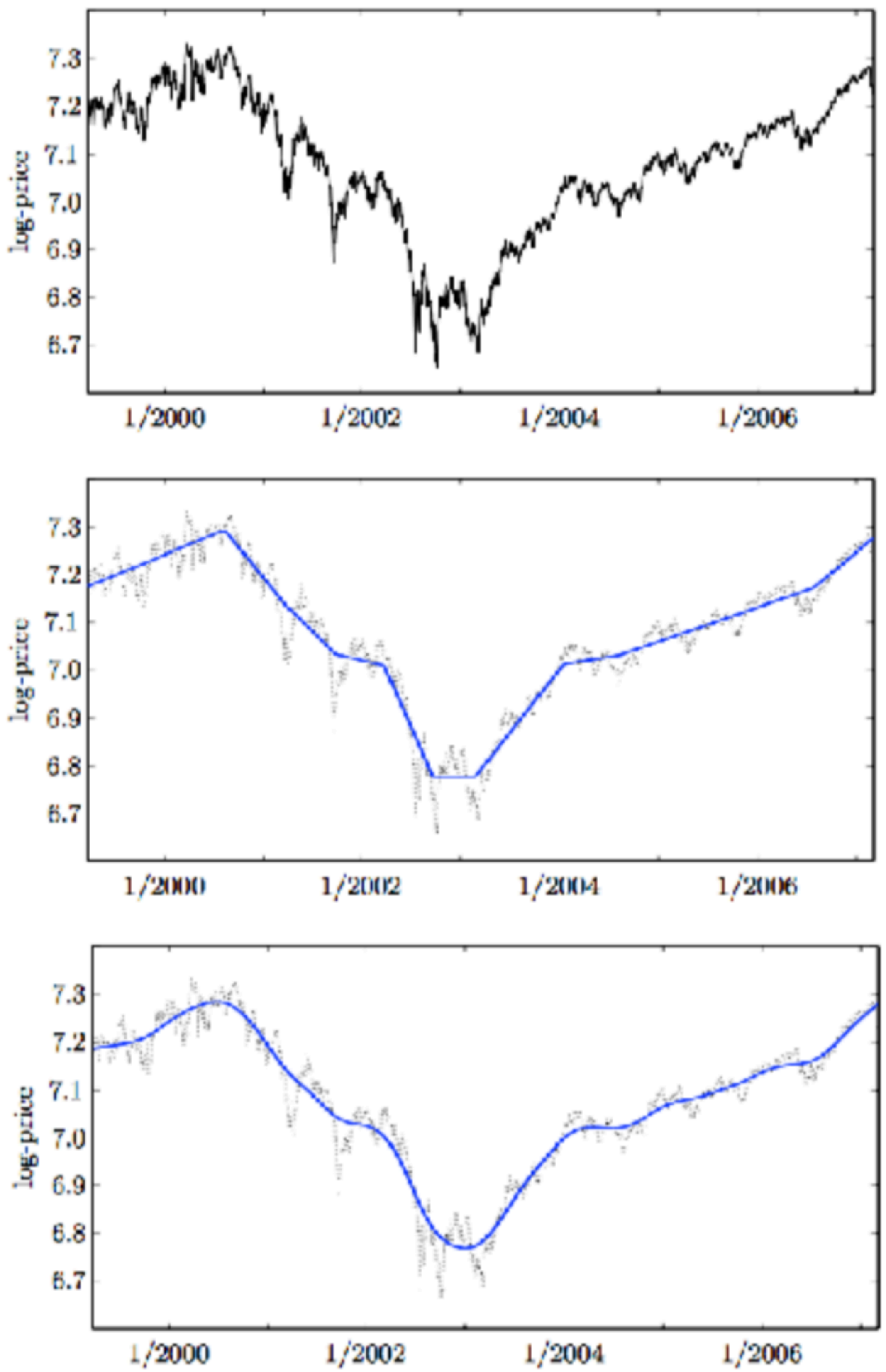
BAYESIAN TREND FILTERING

- Just for fun, let's actually look at a slight extension to the Bayesian Lasso.
- Let's assume that our underlying signal $\mathbf{x}$ is piecewise constant, so we'll put a Bayesian Lasso on it's increments

$$x_i - x_{i-1} \sim \text{Laplace}(\lambda)$$

- It will surprise you not a bit that this also does not work
- But how can we know?

FITTING A PIECEWISE LINEAR FUNCTION



- Sometimes a non-linear effect / Gaussian Process is too smooth
- Piecewise linear functions tend not to over-fit (in theory anyway)
- A model of this is called ℓ_1 trend filtering.

Figure: Kim et al. (2009), SIAM Review, 51(2), pp. 339-360.

HOW WILL I KNOW IF HE REALLY LOVES ME

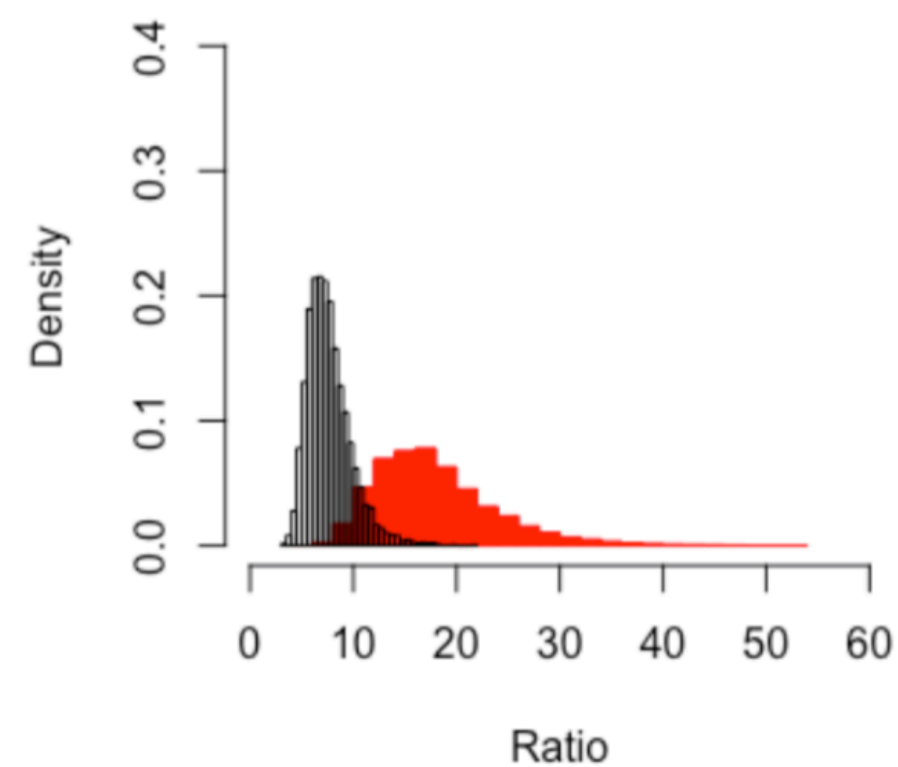
- Well, if we want a piecewise constant curve we need most of the increments to be almost zero and a few to be really big
- One way to check this is to simulate from the prior and see if it has this feature
- The trick is to find some "cartoon" version of the model we want to fit and ask if has prior support.
- What's our trend filtering cartoon? A step function.

IT'S IN HIS KISS (THAT'S WHERE IT IS)

- The statistic I chose was

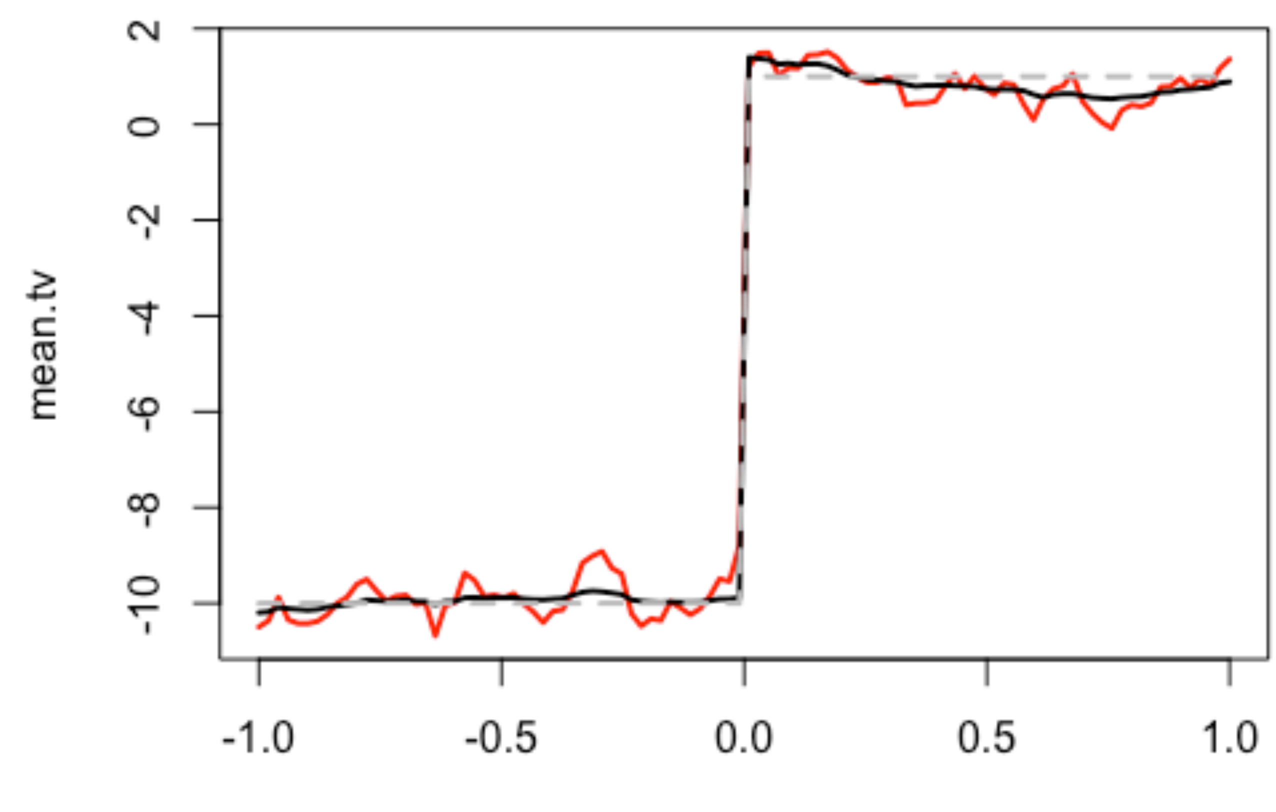
$$\frac{\max_i |x_i - x_{i-1}|}{\text{median}^*(|x_i - x_{i-1}|}$$

- That is, the largest jump divided by the median of all the other jumps
- (Median because the jump distribution hopefully has a heavy tail!)
- If the model works, this should have a long tail…
- Black is the Bayesian Lasso
- Red is the Horseshoe, which does work



NOW THAT'S WHAT I CALL EVIDENCE

- But we can do better.
- Let's simulate data from the simplest case: a step function
- Here Black is the Horseshoe, Red is the Bayesian Lasso (I know!)
- The narrowness of the jump distribution for the Lasso shows in it over-fitting the noise here



NO EXCUSES

- There really isn't any excuse not to check your model before you see data
- (Or to use the Bayesian Lasso!)
- You don't need fancy theory to show that these things don't work
- You can just use your computer and a bit of thought!
- Pre-experiment prophylaxis prevents poorly performing posteriors.

OF COURSE, YOU SHOULD LOOK AT YOUR RESULTS

POSTERIOR PREDICTIVE CHECKING

The *posterior predictive distribution* is the average data generation process over the entire model

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)\, p(\theta|y)\, d\theta$$

POSTERIOR PREDICTIVES CAN BE USEFUL FOR MODEL COMPARISON

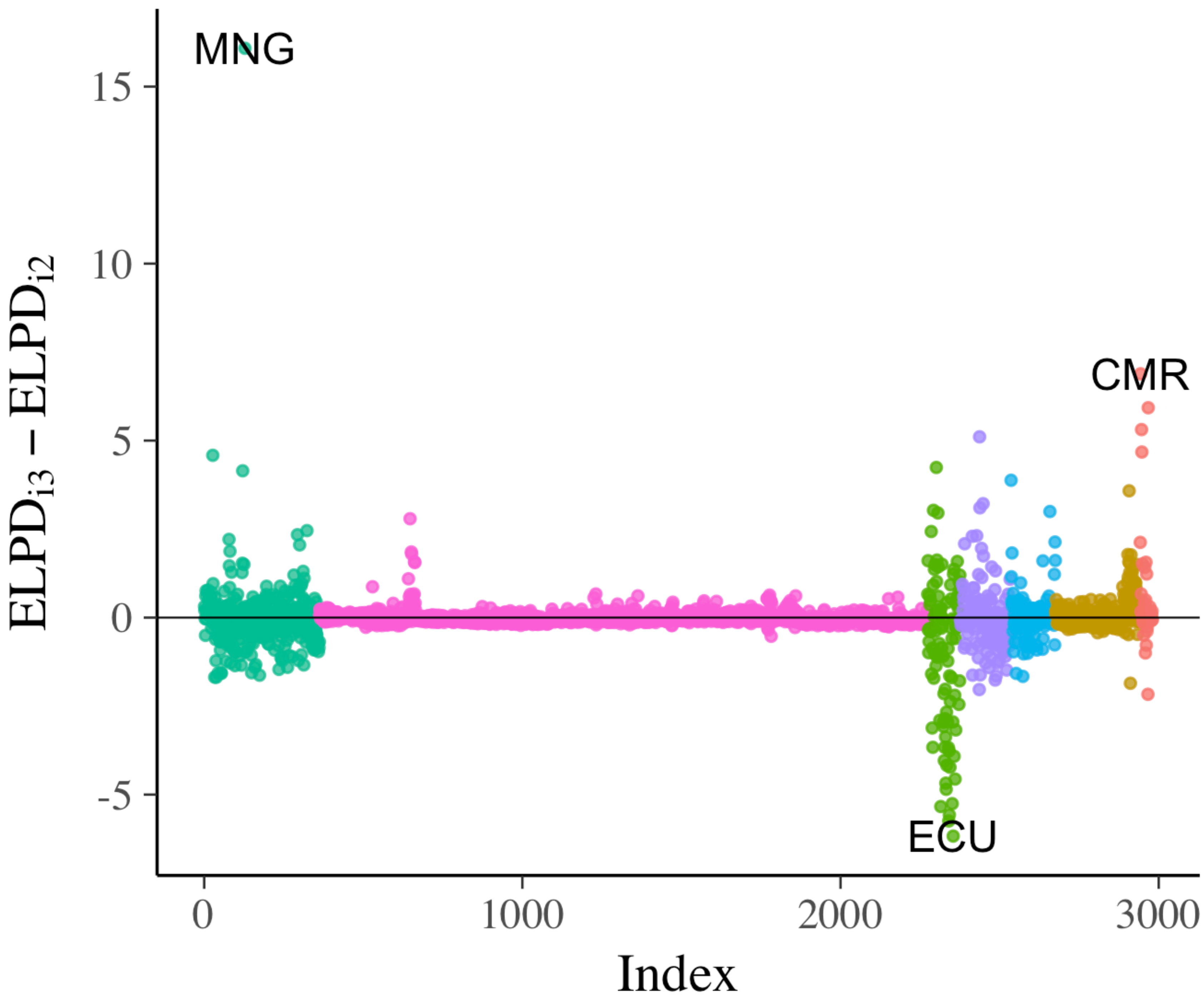
- One thing that can be worth looking at is the predictive distribution we would've had if one observation was missing

$$p(\tilde{y} \mid y_{-i}) \propto \int p(\tilde{y} \mid \theta) p(\theta \mid y_{-i})\, d\theta$$

- This can be computed with self-normalized importance sampling with proposal distribution $g(\theta) = p(\theta \mid y)$ and importance ratios

$$r(\theta) = \frac{1}{p(y \mid \theta)} \propto \frac{p(\theta \mid y_{-i})}{p(\theta \mid y)}$$

MORE THAN JUST COMPUTING A STATISTIC



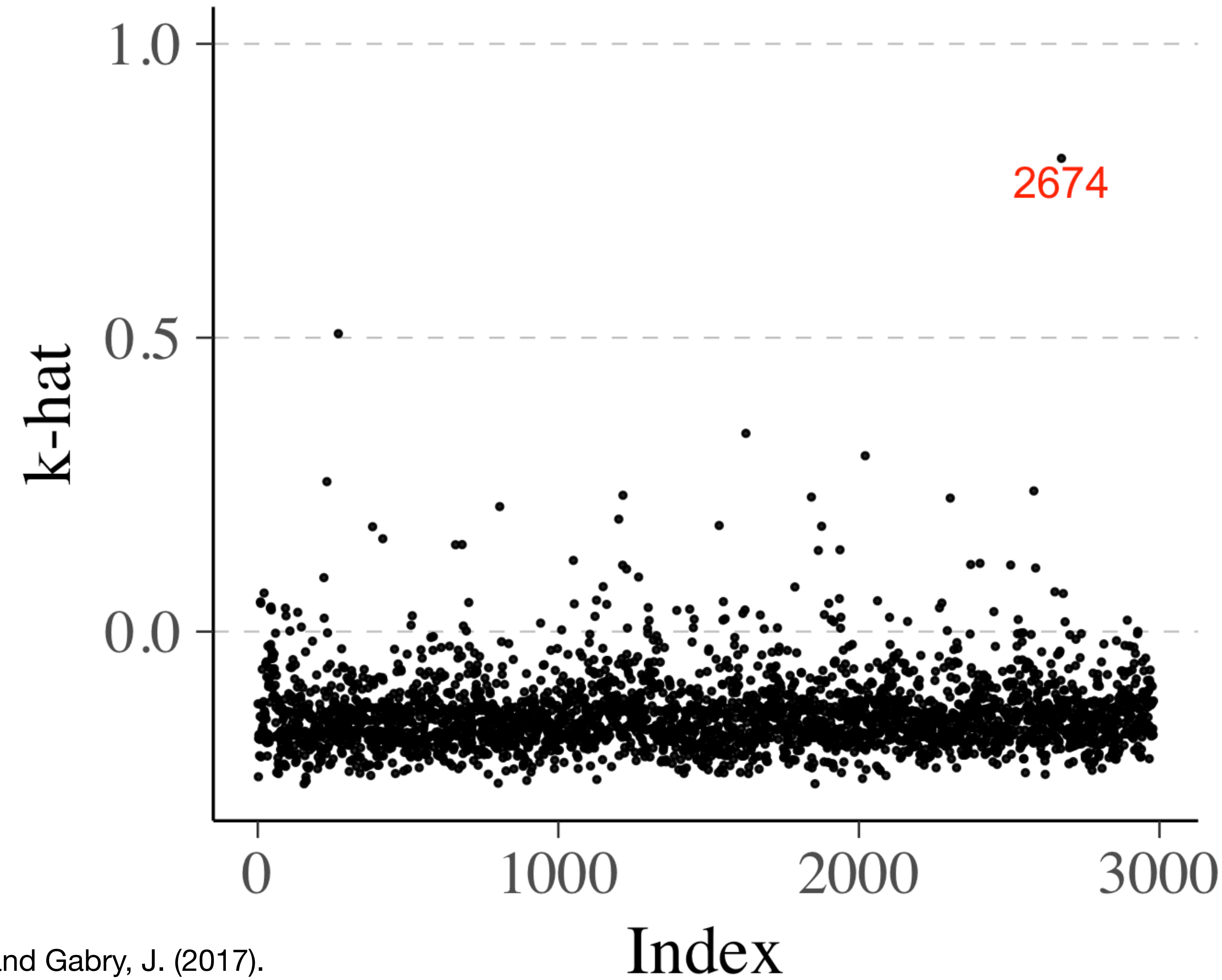
IDEA: HOW MUCH DOES THE PREDICTIVE CHANGE?

- One thing that is useful to look at is how much the posterior predictive distribution changes when a single data point is left out
- We can do this by looking at k-hat for

$$r(\theta) \propto \frac{p(\theta \mid y_{-i})}{p(\theta \mid y)}$$

- If k-hat is large, this means that adding the *i*th point greatly changes the posterior, so the inference is sensitive to this observation
- It is strongly related to leverage for linear models (Peruggia, 1997)

DIAGNOSTICS (K-HAT: A PREDICTIVE LEVERAGE)



Mongolia

Vehtari, A., Gelman, A., and Gabry, J. (2017).
Pareto smoothed importance sampling.
working paper arXiv: arxiv.org/abs/1507.02646/

THE HAROLD HOLT MEMORIAL SWIMMING POOL

STATISTICS IS HARD

- As tempting as it is, there is no way to avoid thinking of all of the aspects of the model simultaneously
- Think of the aspects of your data gathering, modelling, computation, and model evaluation as all being made of the same substance
- And right now, I'm not sure there are any good ways to keep track of anything at once

THERE WON'T BE TRUMPETS

- Sometimes there are loud warnings that things have gone badly:
    - Divergences
    - R-hat (kinda)
    - Simulation Based Calibration (expensive)
    - Prior predictive simulations (if you're clever)
    - Posterior predictive checks (watch your assumptions)
- But really, we need to build careful simulation studies and meaningful checks of the pre-observation joint distribution of the parameters and the data.

HAROLD HOLT'S HUBRIS

- Harold Holt went swimming in dangerous surf and drowned.
- No amount of synchronized swimming would not have saved him.
- So make sure you focus on the right things and stop just building memorial swimming pools.

This has been joint work with Michael Betancourt, Jonah Gabry, Andrew Gelman, and Aki Vehtari.

JOBS! JOBS! JOBS!

- Statistics (Full Professor)
- Data Science (100% Stats)
- Teaching Stream (100% Stats)
- Causal Inference (100% Stats)
- With Philosophy (49% Stats, 51% Phil)
- With School of Environment (51% Stats, 49% Phil)
- With Computer Science (51% Stats, 49% CS)
- With Information Science (51% Stats, 49% iScience)
- With Psychology (66% Stats, 34% Psych)
- With CS on Data Visualization
- Statistical Genetics and Genomics (100% Stats)