Bayesian statistics & MCMC

Tanvi Karwal

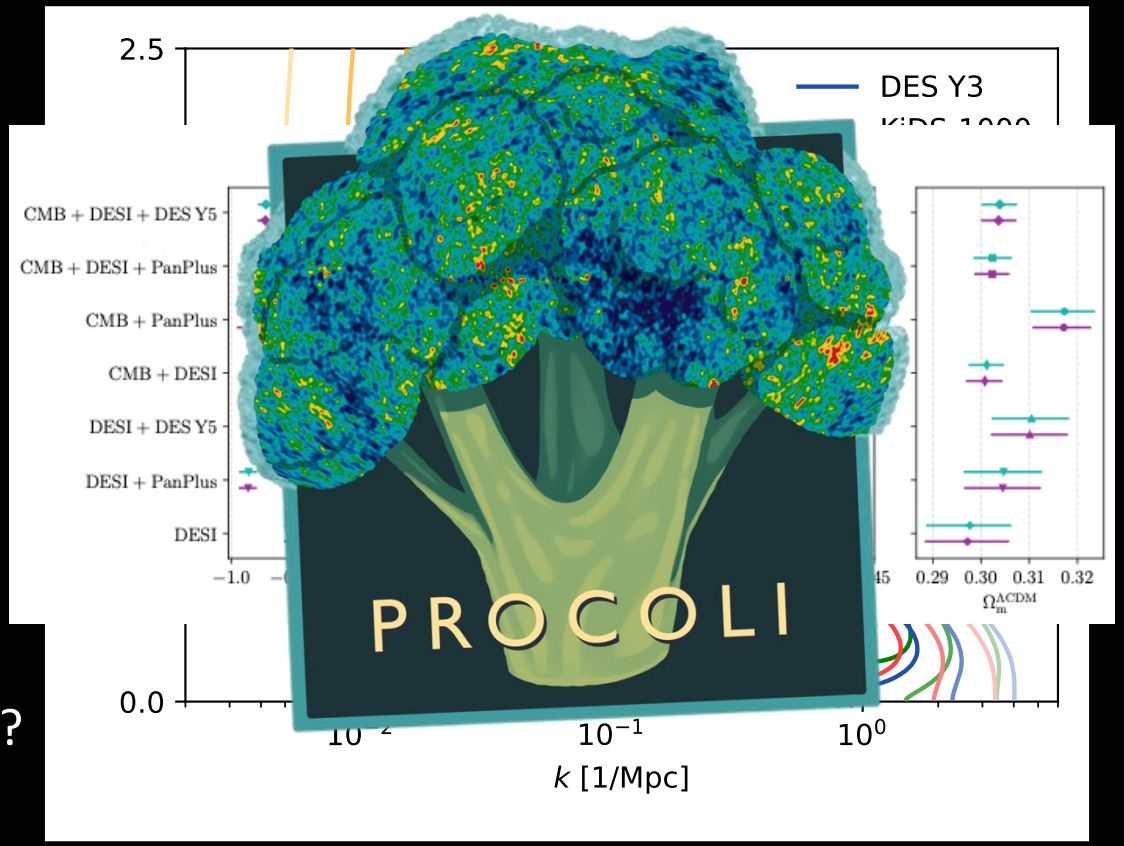
THE UNIVERSITY OF CHICAGO KICP

GGI workshop: Exploring New Frontiers in Cosmology
Parameter estimation in cosmology I
8th July 2026

Phenomenological cosmology

- Early dak energy to solve the Hubble tension
- S_8 tension
- Late dark energy / preference for evolving dark energy
- Building profile likelihood infrastructure and applications

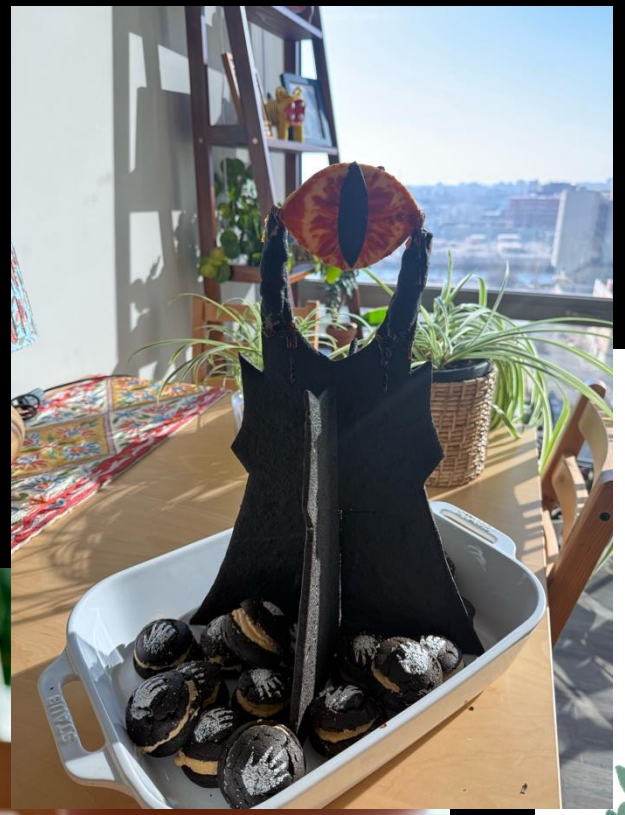
What are data sensitive to?
What degeneracies are they susceptible to?
What freedoms exist in data where new physics might hide?
How do we constrain the parameters of our Universe?

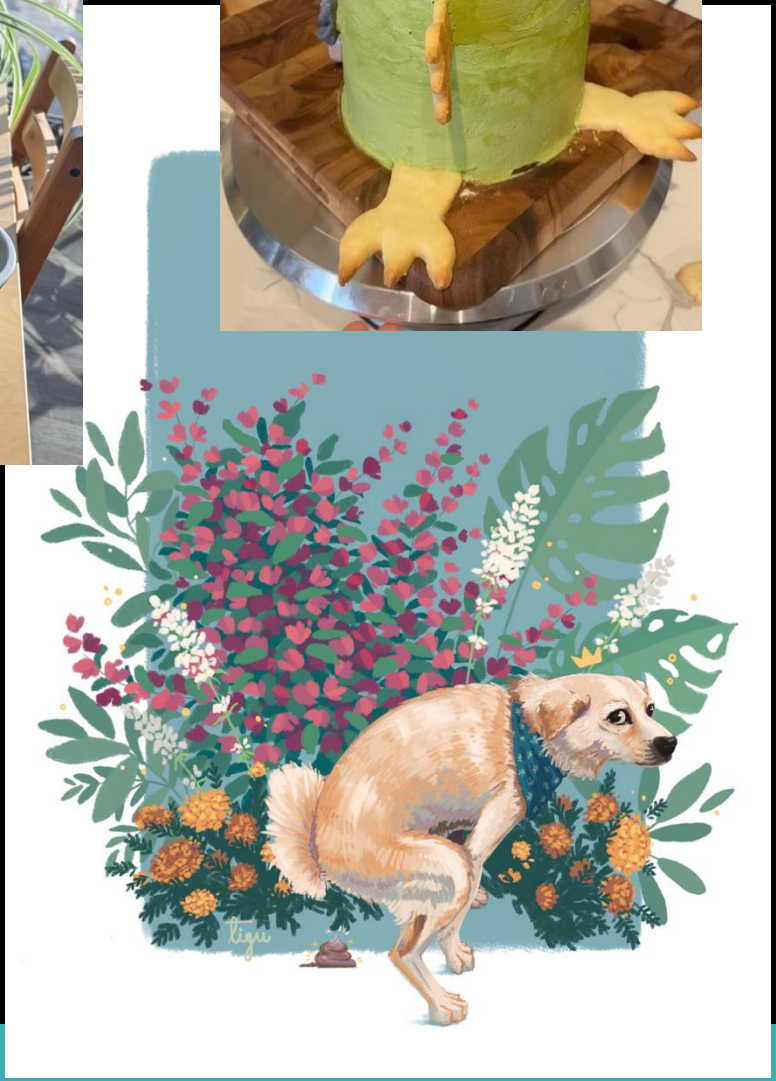


I have strong opinions about dessert

I draw things

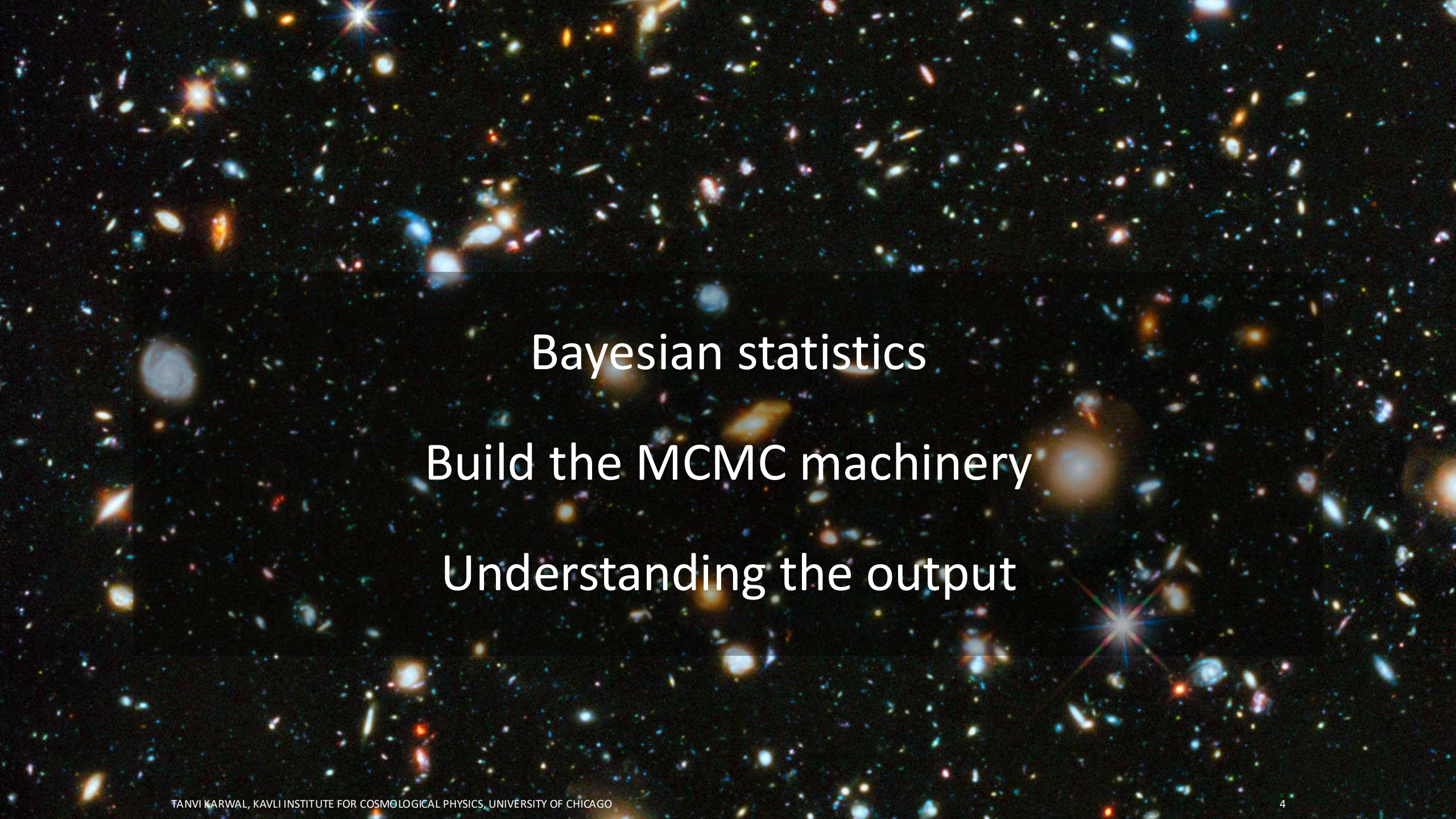
I run slowly,
I climb badly,
I have too many plants





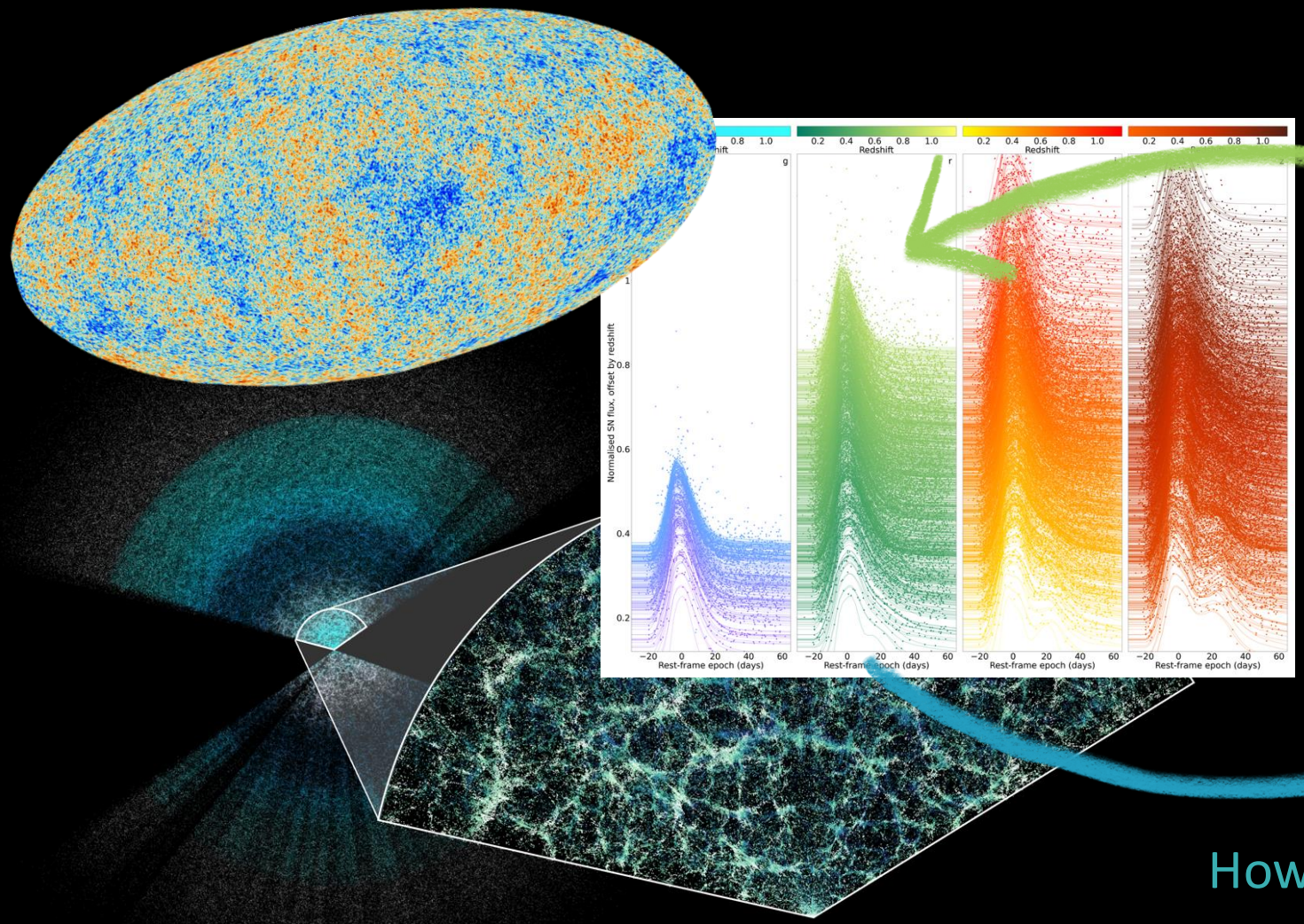
Open the notebook now — we'll code as we go

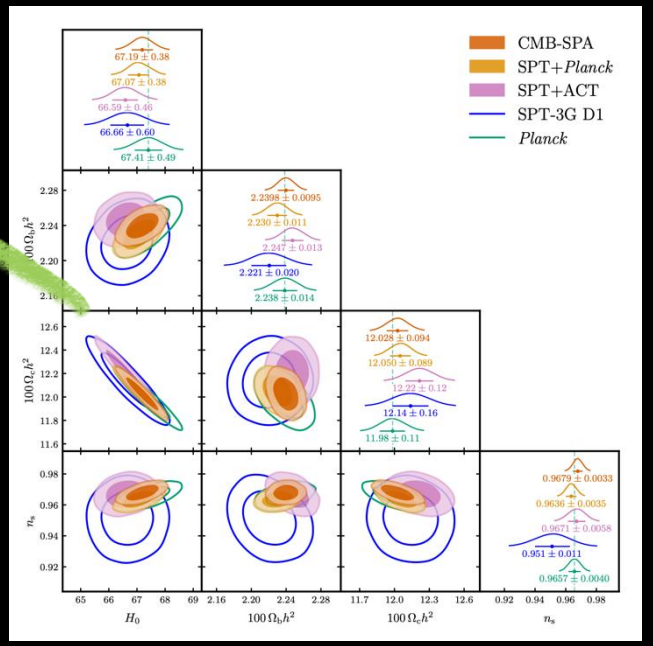
1. Scan the QR → opens the repo
2. Open `ggi_write_MH_MCMC.ipynb` in Google Colab
3. Run the first cell (imports) to check your kernel works
4. Leave it open — I'll tell you which section to jump to


Bayesian statistics
Build the MCMC machinery
Understanding the output


Bayesian statistics
Build the MCMC machinery
Understanding the output

We don’t directly measure parameters — only data





Inference is inverting the arrow

How do we invert it, and how much do we trust the answer?

Bayesian inference ingredients

For data D and parameters θ

- **Likelihood** $\mathcal{L}(\boldsymbol{D} \mid \boldsymbol{\theta})$

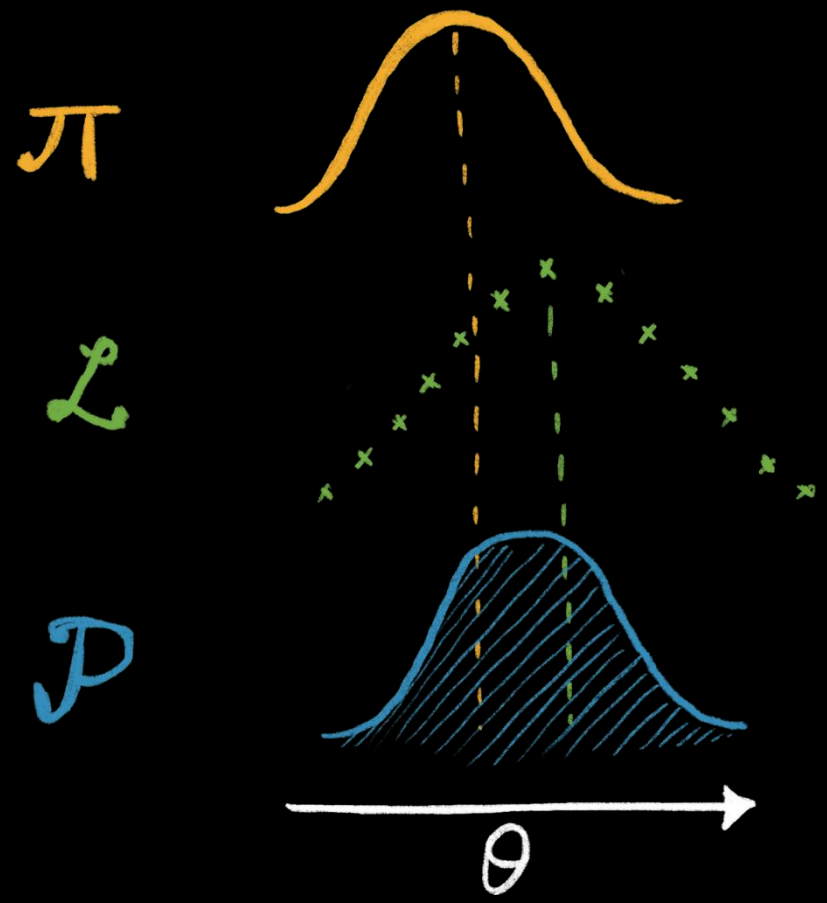
probability of the data given the parameters

- **Prior** $\boldsymbol{\pi}(\boldsymbol{\theta})$

what we knew before this dataset

- **Posterior** $\boldsymbol{P}(\boldsymbol{\theta} \mid \boldsymbol{D})$

our updated knowledge after seeing the data



prior × likelihood → posterior

Bayesian inference ingredients

For data D and parameters θ

- **Likelihood** $\mathcal{L}(\boldsymbol{D} \mid \boldsymbol{\theta})$

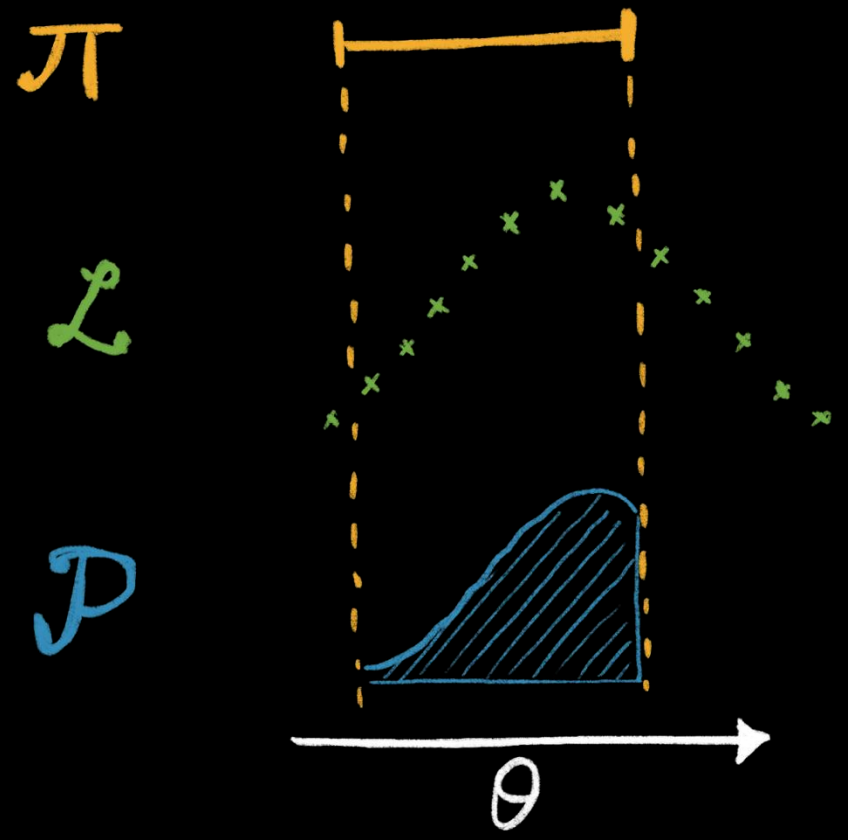
probability of the data given the parameters

- **Prior** $\boldsymbol{\pi}(\boldsymbol{\theta})$

what we knew before this dataset

- **Posterior** $\boldsymbol{P}(\boldsymbol{\theta} \mid \boldsymbol{D})$

our updated knowledge after seeing the data



prior × likelihood → posterior

Bayes' theorem

$$\boldsymbol{P(\theta \mid D)} = \frac{\boldsymbol{\pi(\theta)} \times \boldsymbol{\mathcal{L}(D \mid \theta)}}{\mathcal{E}(D)}$$

In the notebook this line is literally log_posterior = log_like + log_prior.

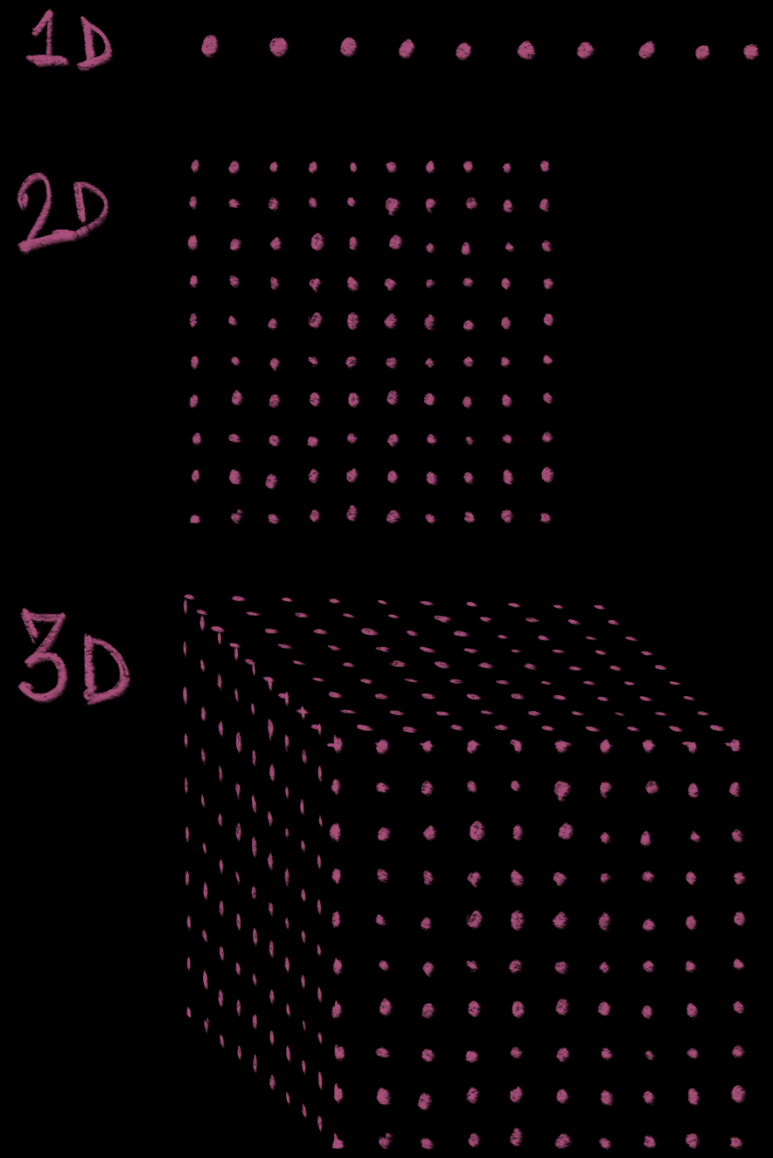
The Bayesian philosophy
Probability = degree of belief

A parameter has a probability distribution representing our state of knowledge.

$$\boldsymbol{P(\theta \mid D)} \propto \pi(\theta) \times \mathcal{L}(D \mid \theta)$$

"Given the data and what I knew before, here is how strongly I believe each value of θ."

Why Bayesian methods took over cosmology



- High-dimensional parameter spaces

grid cost = k^d where d = dimensions

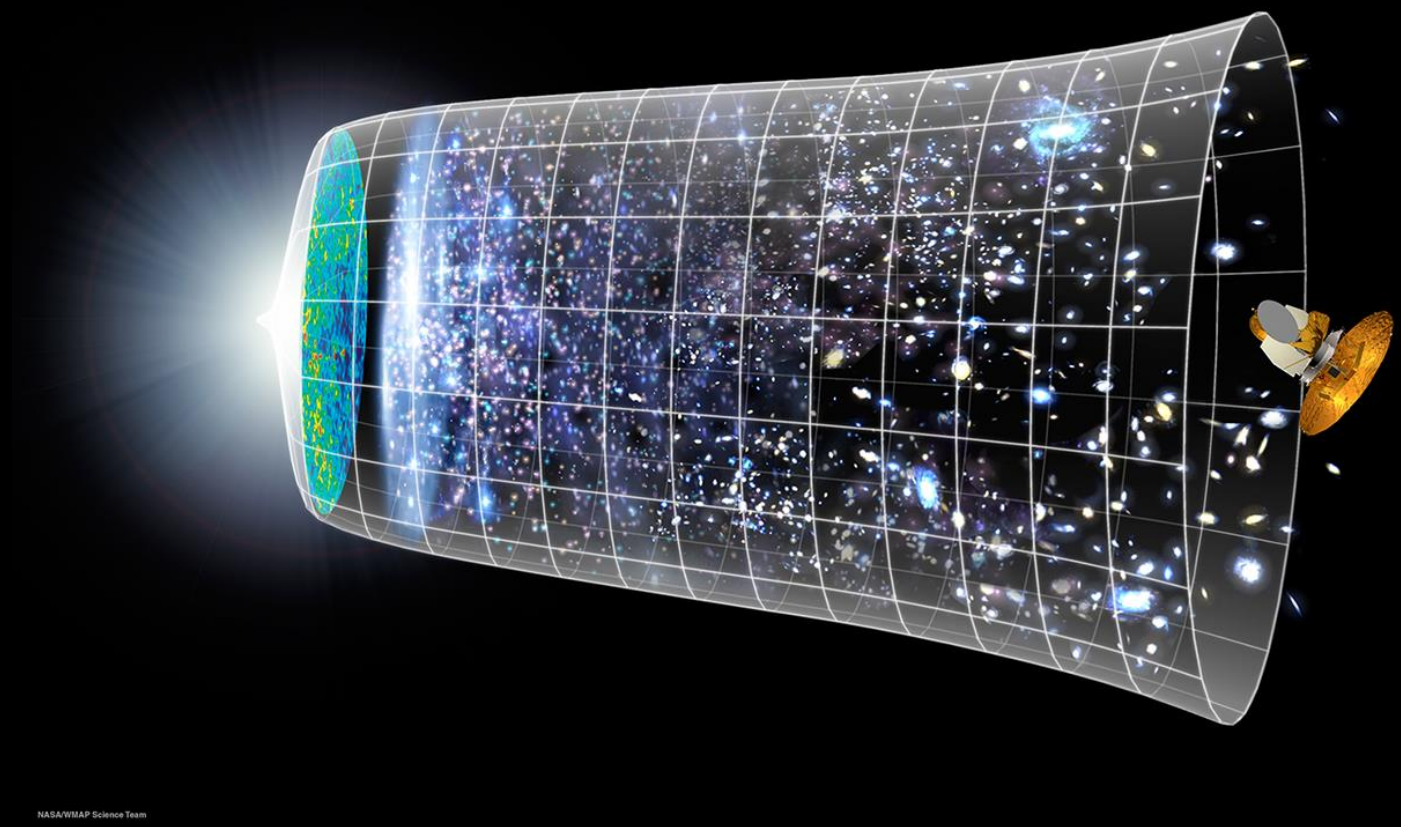
$d = 6$ (ΛCDM), $k = 20 \rightarrow 6.4 \times 10^7$ likelihood calls

$d = 27$ (with nuisances) → absurd

Section 1

Why Bayesian methods took over cosmology

- High-dimensional parameter spaces
- We genuinely have prior information



Why Bayesian methods took over cosmology

- High-dimensional parameter spaces
- We genuinely have prior information
- Mature, public infrastructure

MontePython

emcee

CosmoMC


Bayesian statistics
Build the MCMC machinery
Understanding the output

The idea behind MCMC:
Sample density approximates posterior probability

We can't evaluate the posterior everywhere (grid), so draw samples whose density IS the posterior.

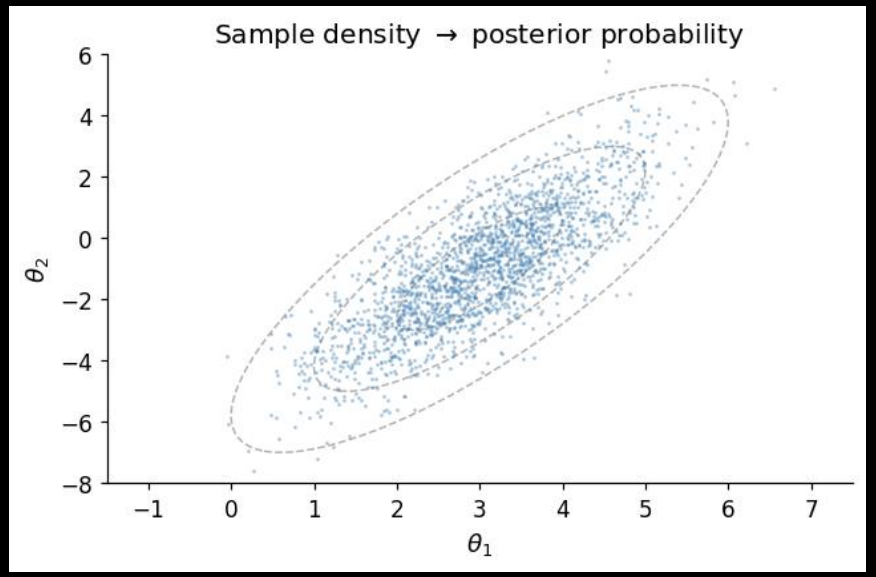
Thanks Elsa for sharing this meme

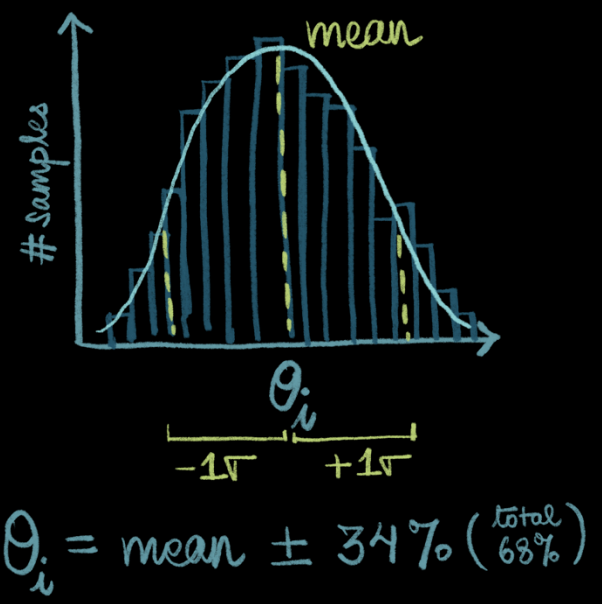
The idea behind MCMC:
Sample density approximates posterior probability

We can't evaluate the posterior everywhere (grid), so draw samples whose density IS the posterior.

Then, where samples pile up = high probability.

The sampler is handed a black box $\mathcal{L} \times \pi$ and just returns points. It never needs to know the physics.





Posteriors give a clearer picture of the landscape than any feasible gridding

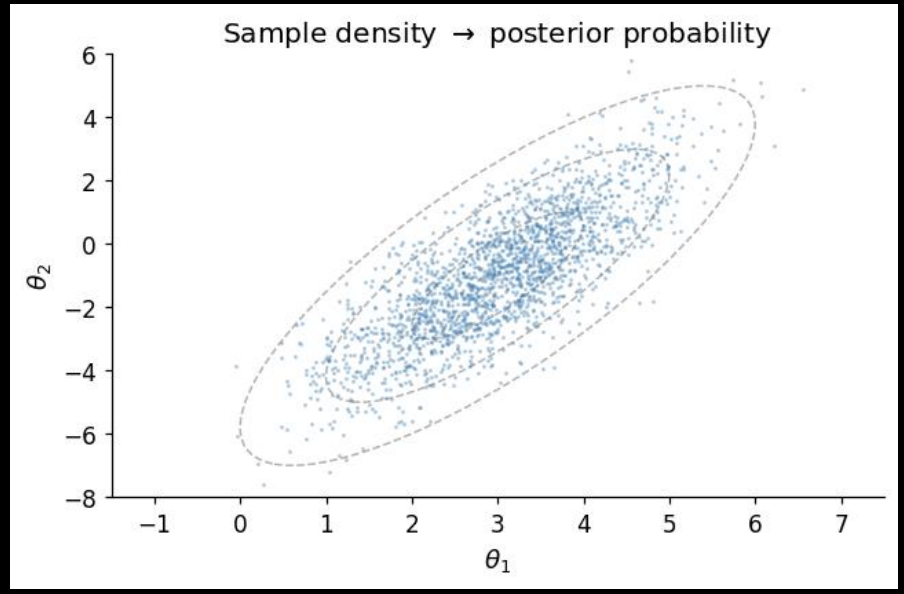
Bayesian posteriors

Grid sampling in 27 dimensions

Meme in collaboration with Tristan

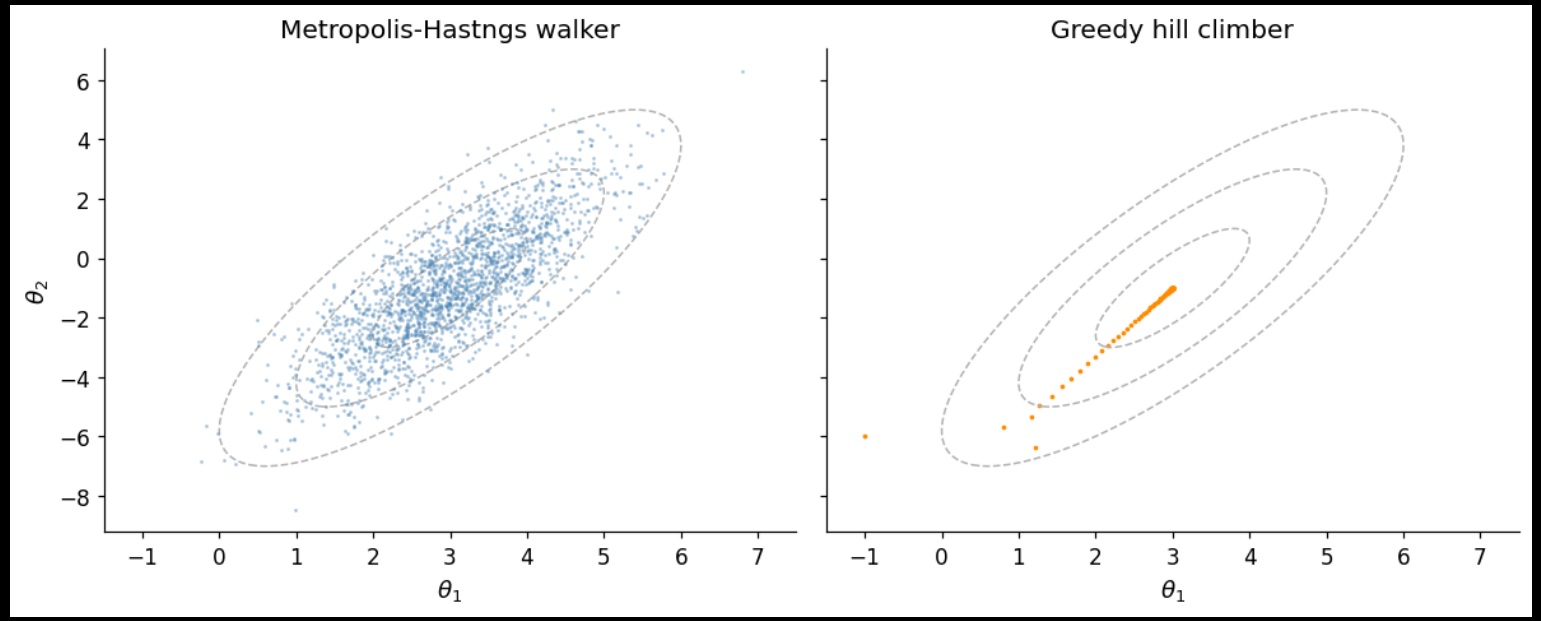
Metropolis–Hastings MCMC

1. start at θ_current
2. propose θ* ~ Normal(θ_current, step)
3. log_α = log_post(θ*) − log_post(θ_current)
4. draw u ~ Uniform(0,1)
5. if log(u) < log_α : ACCEPT → θ_current = θ*
   else : REJECT → record θ_current again



One counter-intuitive rule

Sometimes accept a worse point.



Downhill moves are what let the chain explore and map the full distribution — not just find the peak.

“accept-if-better-only” is a hill-climber, not a sampler.

To the notebook: build, validate on the known input Gaussian

Sections 3 — you write three functions:

`acceptance_probability` · `propose` · `mh_step`

We will test with the analytic 2D correlated Gaussian:

$$\mu = [3, -1], \qquad \sigma = [1, 2], \qquad \rho = 0.8$$

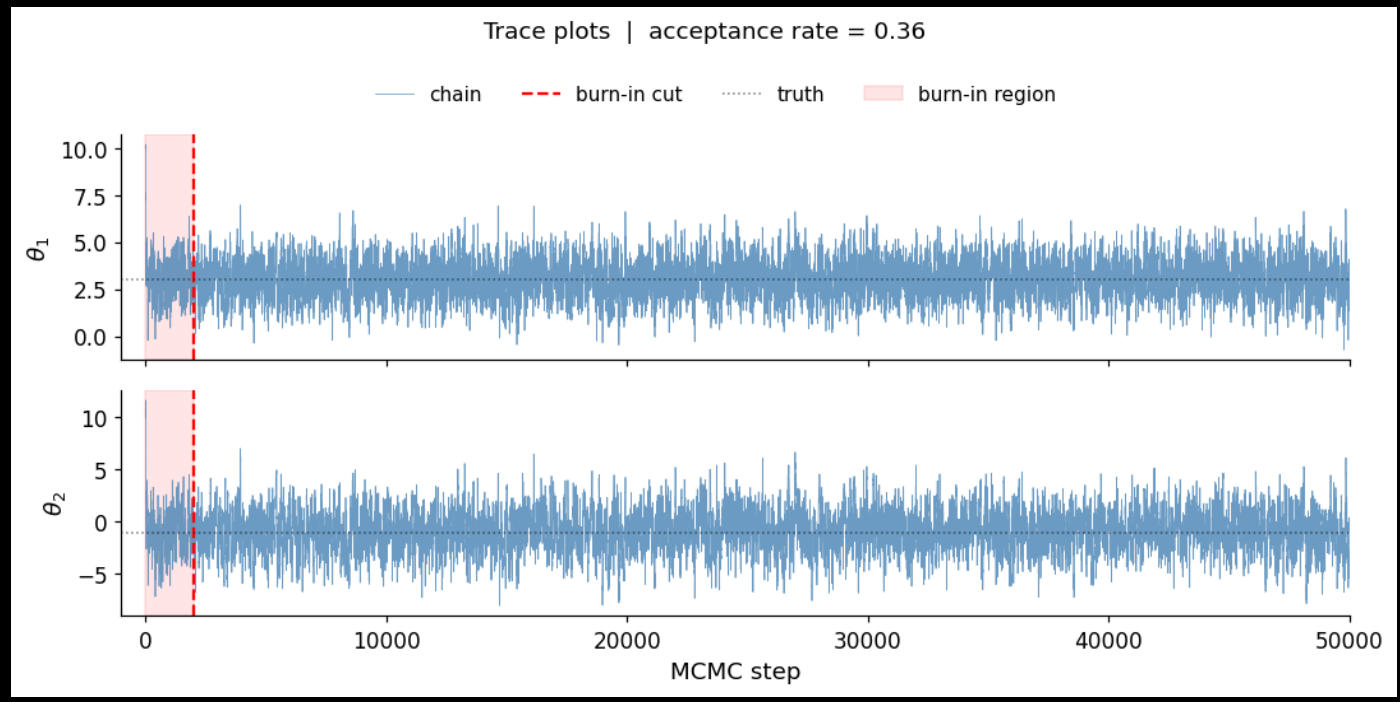
```
MCMC: 100% ████████████████████ 9999/9999 [00:00<00:00, 116694.85step/s]

Chain shape:      (10000, 2)
Acceptance rate: 0.36   (target: 0.25 – 0.45)
```

We will do Section 4 together over the next few slides

Diagnostics I — burn-in & initial guesses

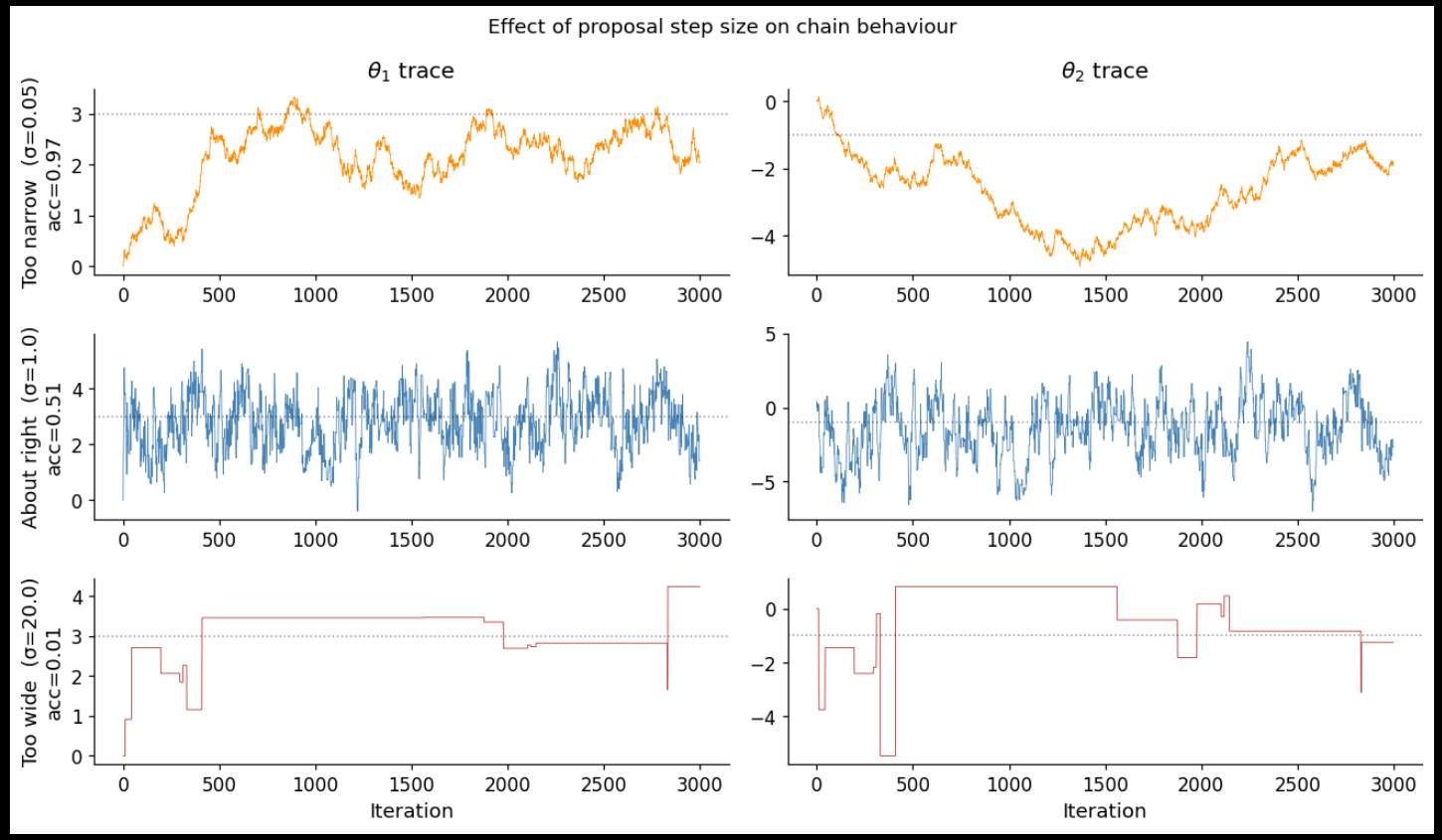
- A healthy chain looks like white noise after burn-in
- Discard the early transient where the walker is still finding the bulk
- Poor start → longer burn-in
  good start → shorter



Diagnostics II — step size & acceptance rate

A high acceptance rate is not a good thing

- Too small → 95% acceptance, but the chain crawls (correlated).
- Just right → acceptance ≈ 0.2 – 0.4.
- Too large → almost everything rejected; chain barely moves.



Diagnostics III — autocorrelation & effective samples

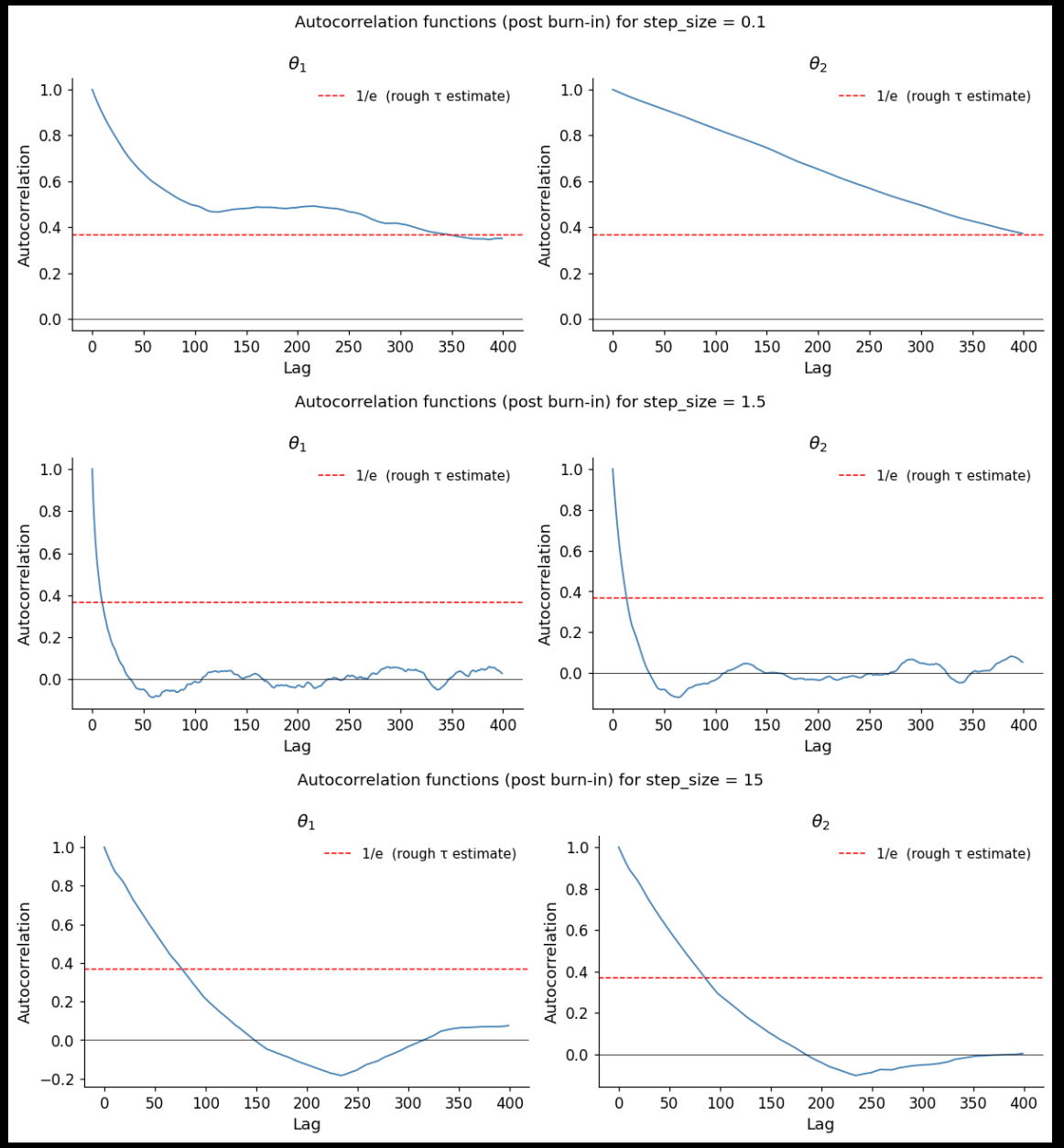
- Correlated samples ≠ independent samples
- τ = lag where the autocorrelation drops below $1/e$
- N_{steps} is a lie; N_{eff} is the truth about how much you actually learned

$N_{eff} \simeq 20$

$N_{eff} \simeq 300$

$N_{eff} \simeq 50$

For the same $N_{steps} = 8000$



Diagnostics IV — convergence

One chain can't certify its own convergence.
We usually run several from dispersed starts.

Diagnostics IV — convergence (Gelman–Rubin)

One chain can't certify its own convergence.
We usually run several from dispersed starts.

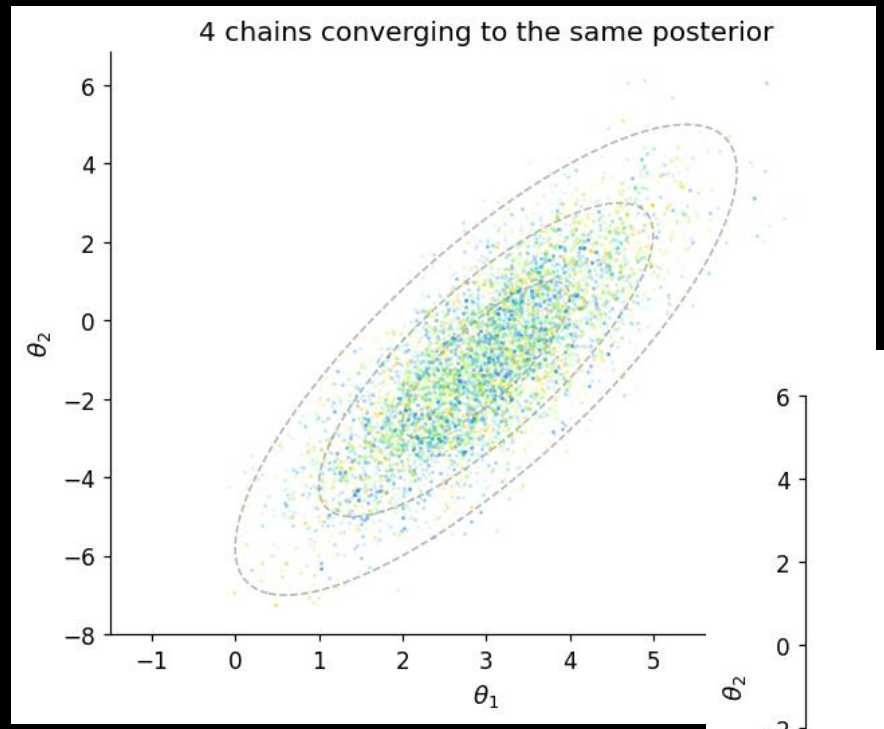
Compare within-chain vs between-chain variance → R

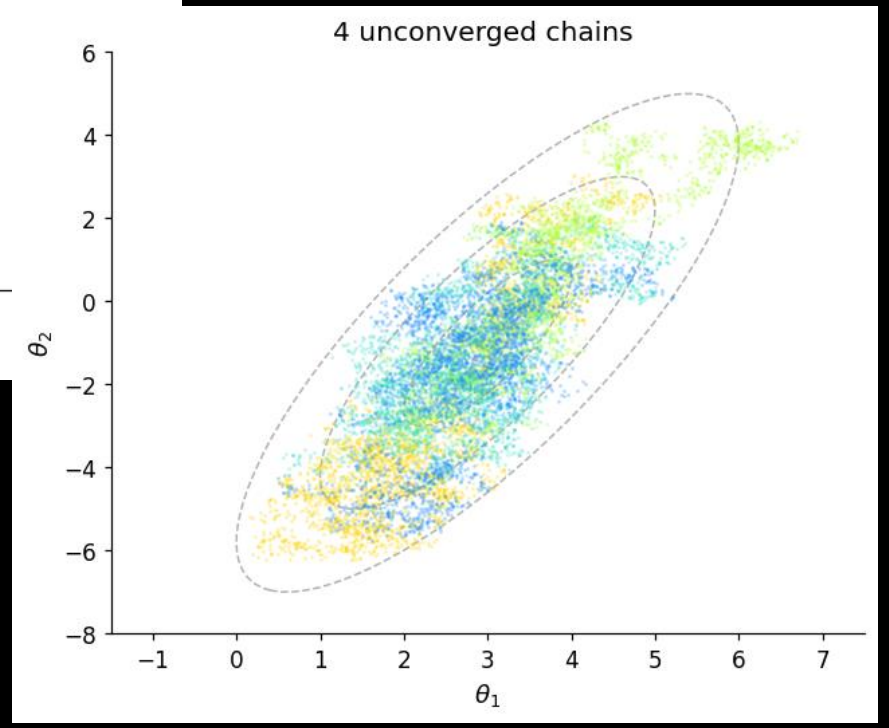
Convergence requires

$$R - 1 \lesssim 0.05$$

but exact constraint varies by taste.

Checks robustness to initial-point bias and local minima.





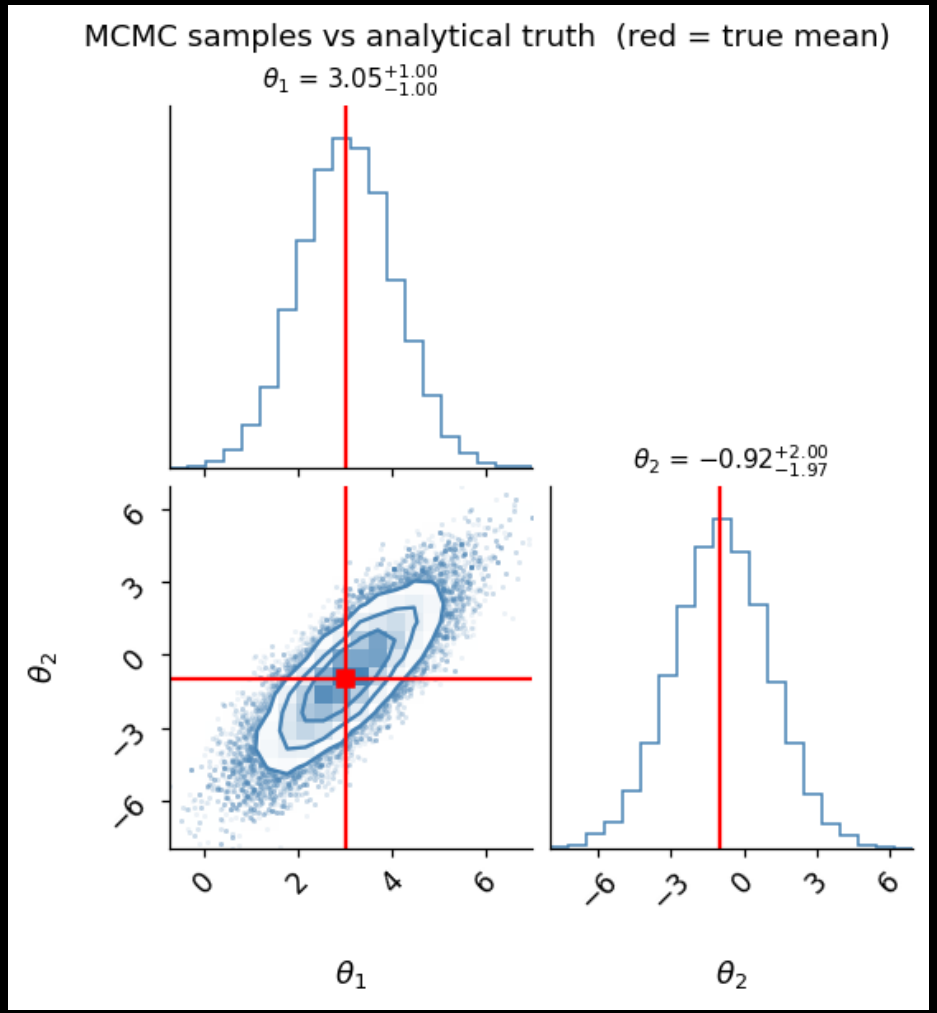
Validate on the known input Gaussian

Sections 5

Recover the analytic 2D correlated Gaussian:

$$\mu = [3, -1], \qquad \sigma = [1, 2], \qquad \rho = 0.8$$

- Checkpoint: your samples must reproduce input numbers.



Why one step size fails in real cosmology

Section 7 — real ΛCDM CMB TT likelihood, 6 parameters

- Posterior widths span two orders of magnitude:

  $\omega_b \sim 10^{-4}$ while $h \sim 10^{-2}$

Even more with $A_s \sim 10^{-9}$ and $H_0 \sim 10^2$

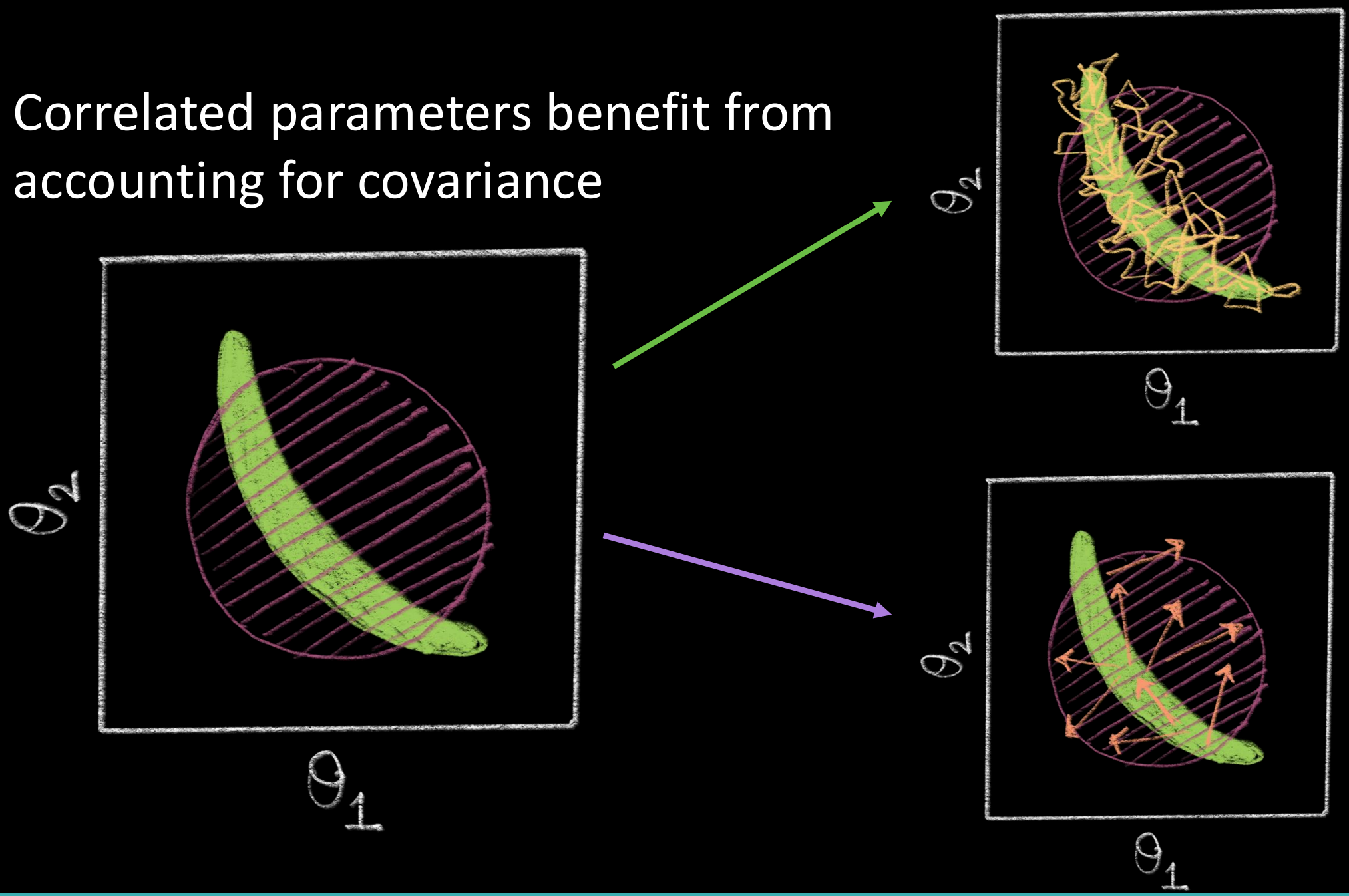
- And they're correlated at $|\rho| \sim 0.95$

A single scalar step size cannot serve all five.

Parameter	Plik best fit	Plik [1]
$\Omega_b h^2$	0.022383	0.02237 ± 0.00015
$\Omega_c h^2$	0.12011	0.1200 ± 0.0012
$100\theta_{MC}$	1.040909	1.04092 ± 0.00031
τ	0.0543	0.0544 ± 0.0073
$\ln(10^{10} A_s)$	3.0448	3.044 ± 0.014
n_s	0.96605	0.9649 ± 0.0042

Planck 2018

Correlated parameters benefit from accounting for covariance

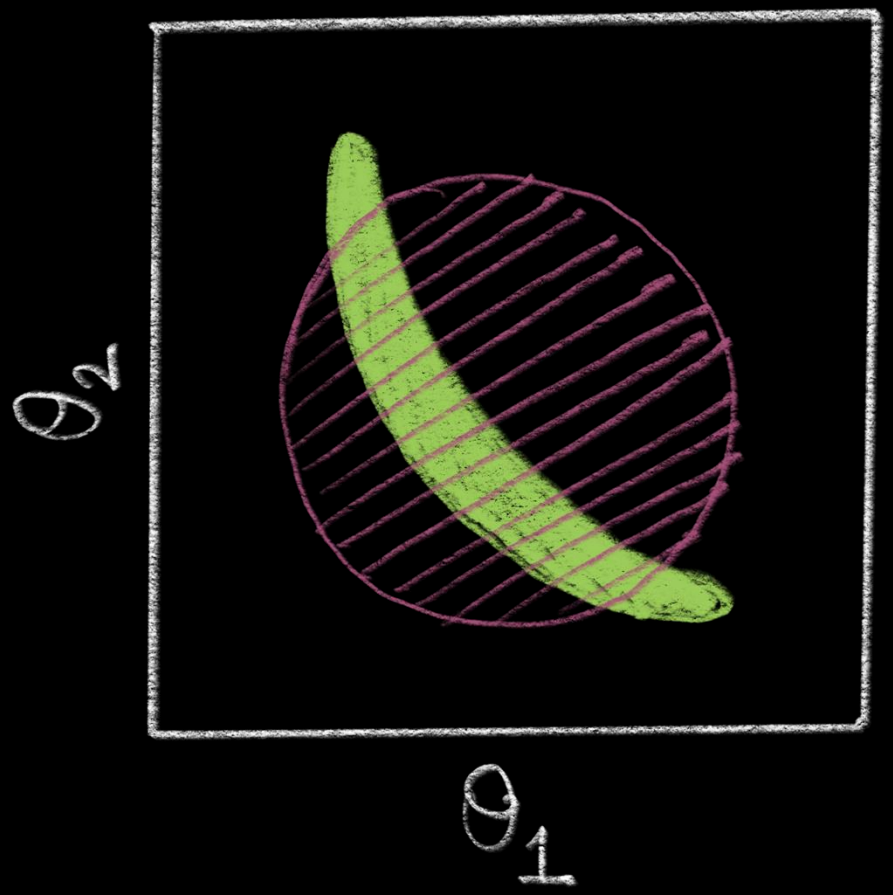


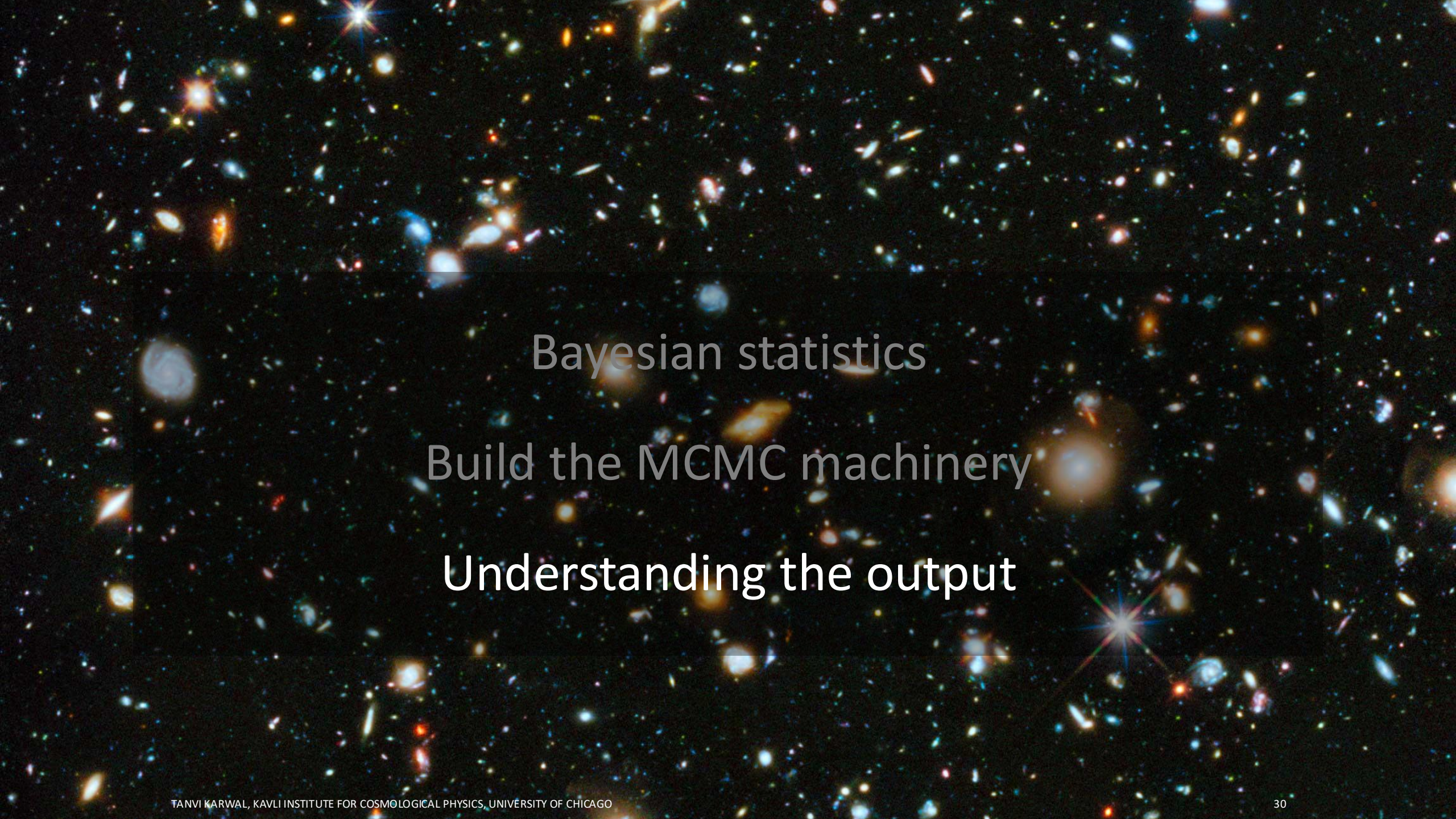
From step size to a proposal covariance

$$\Delta\theta \sim \mathcal{N}(0, f^2\mathbb{C}) \quad f = \frac{2.38}{\sqrt{d}}$$

- Diagonal of covariance $\mathbb{C}$ fixes the per-parameter scales.
- Off-diagonals follows the degeneracy directions.

- Same `run_mh_chain` — only `propose` changes
  → ~10 × more independent samples.
- This is how Cobaya / MontePython actually work:
  start from a Fisher or pilot covariance,
  keep adapting and ‘learning’ it.



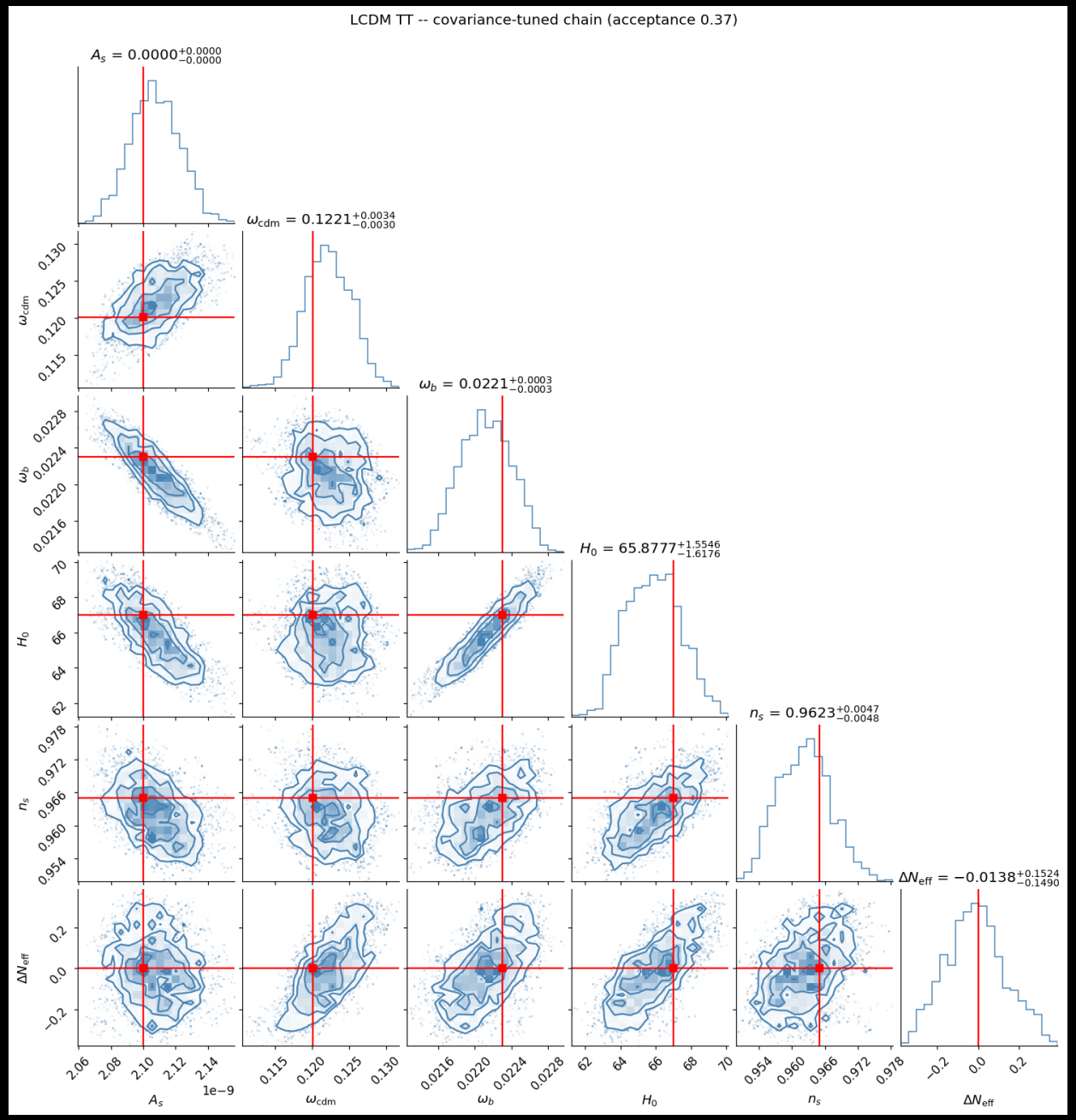

Bayesian statistics
Build the MCMC machinery
Understanding the output

Extracting constraints from the corner plot

This Bayesian approach automatically gives you

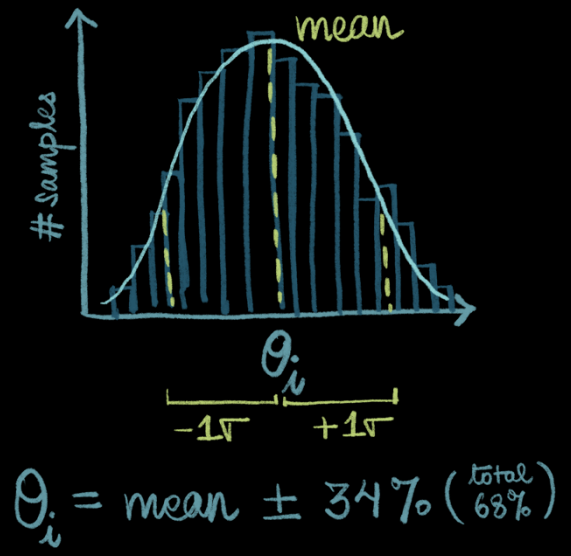
- Easily marginalisation by simply ignoring rows / columns
- 1D histograms (each parameter)
  + 2D contours (each pair)
- Posterior mean, peak (or MAP)
- Interval: central 68% by sample density / upper or lower limits

Constraints come from the density of samples.

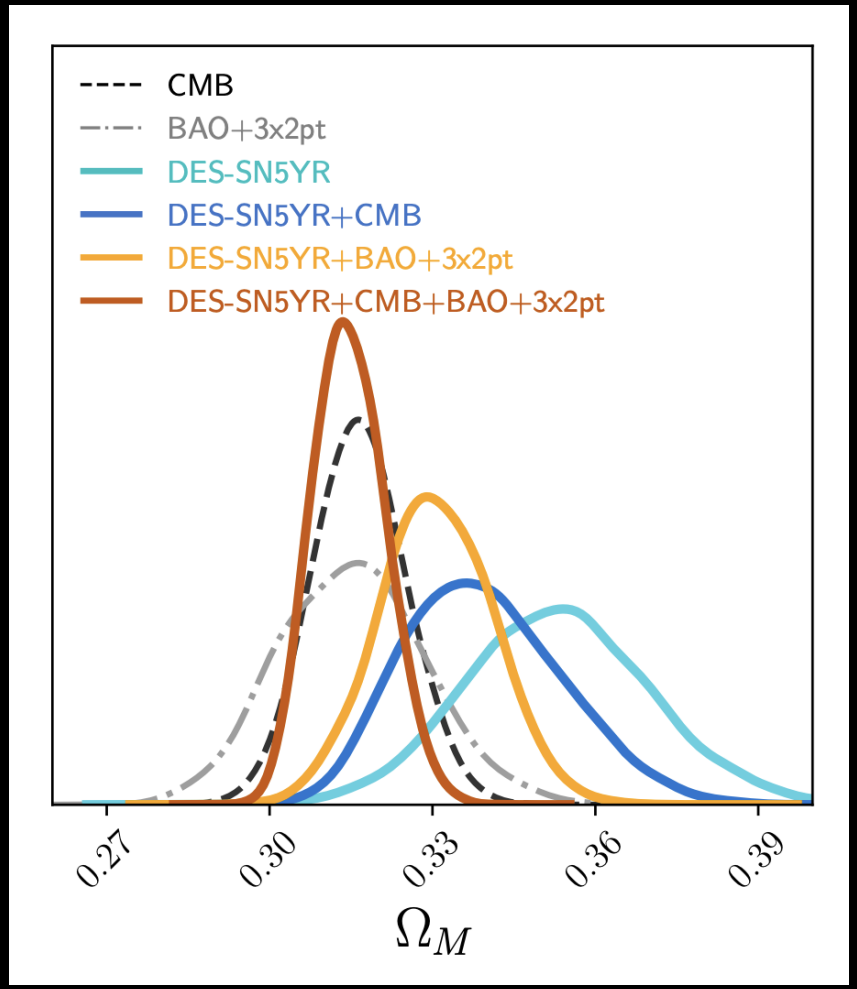


What a credible interval mean

"Given the data and my prior, I believe at 68% that θ lies in this range."



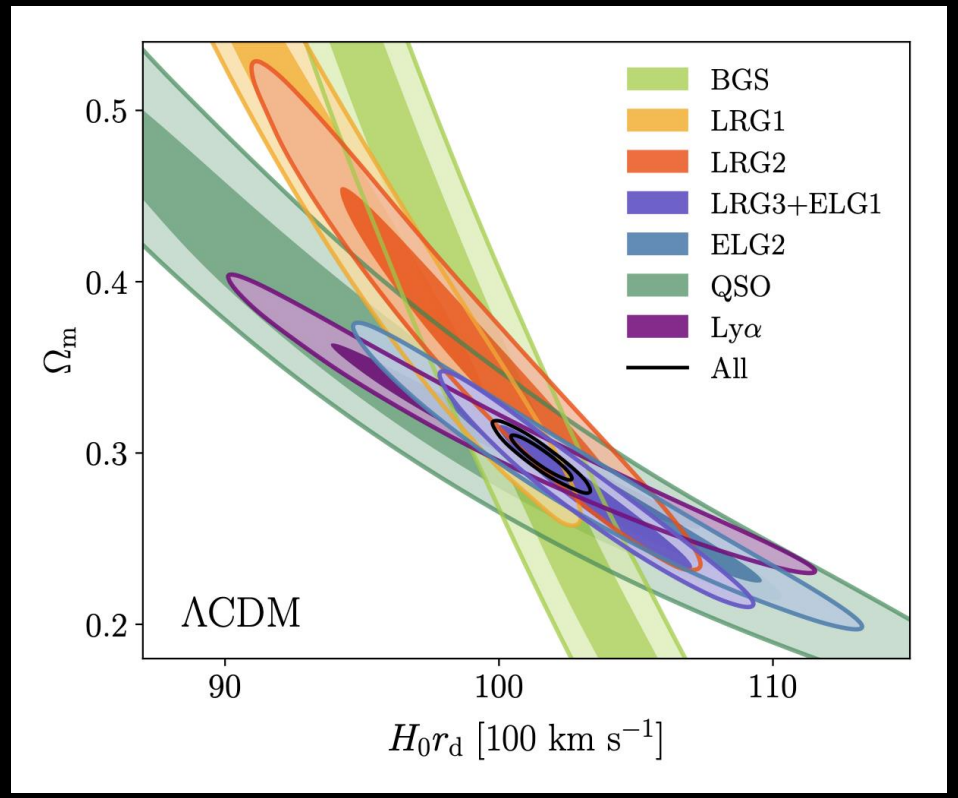
It is a statement about belief — conditioned on the prior



DESY5 Supernovae

Where Bayesian methods shine

- When priors carry real information
- When you want the full joint distribution and all marginals
- Combining datasets: today's posterior = tomorrow's prior
- Model comparison via the evidence



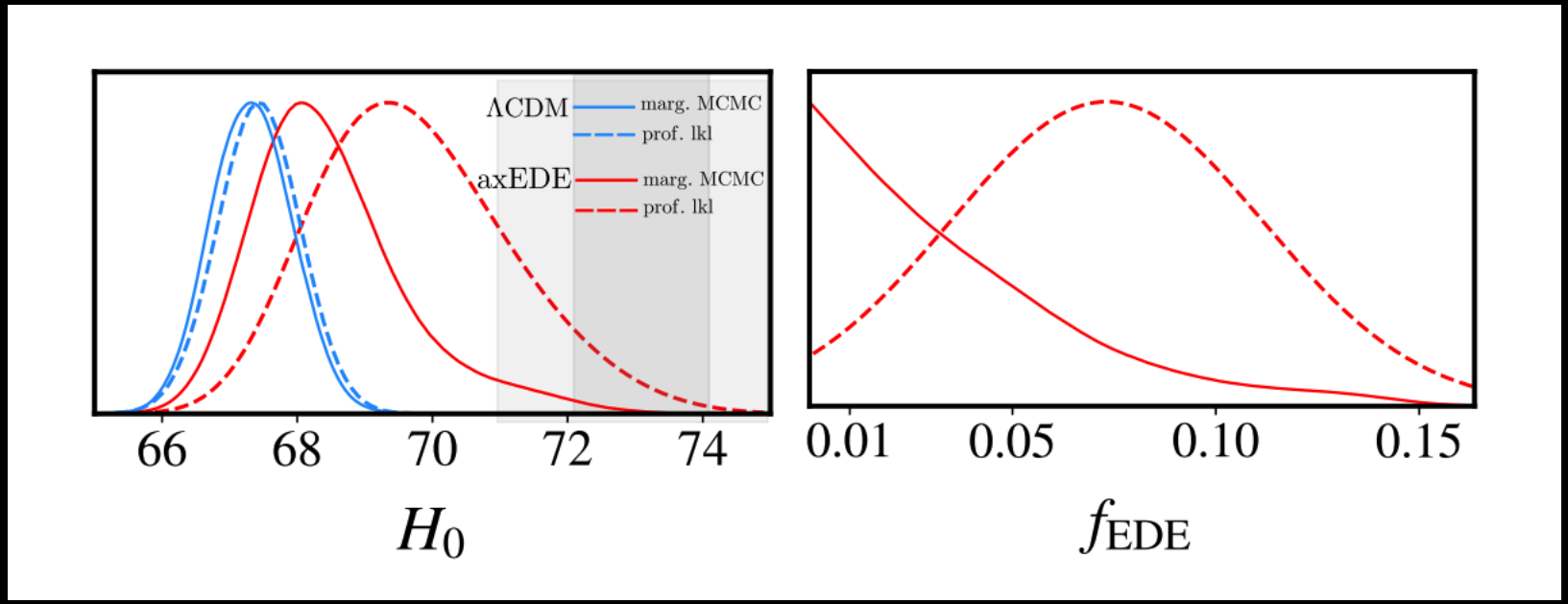
DESI DR1 BAO

Where Bayesian methods fall short

- When data are weak, the posterior leans on the prior
- Even with FLAT priors, marginalising over nuisance directions can shift a constraint — the prior-volume / projection effects

Is that a bug or a feature?

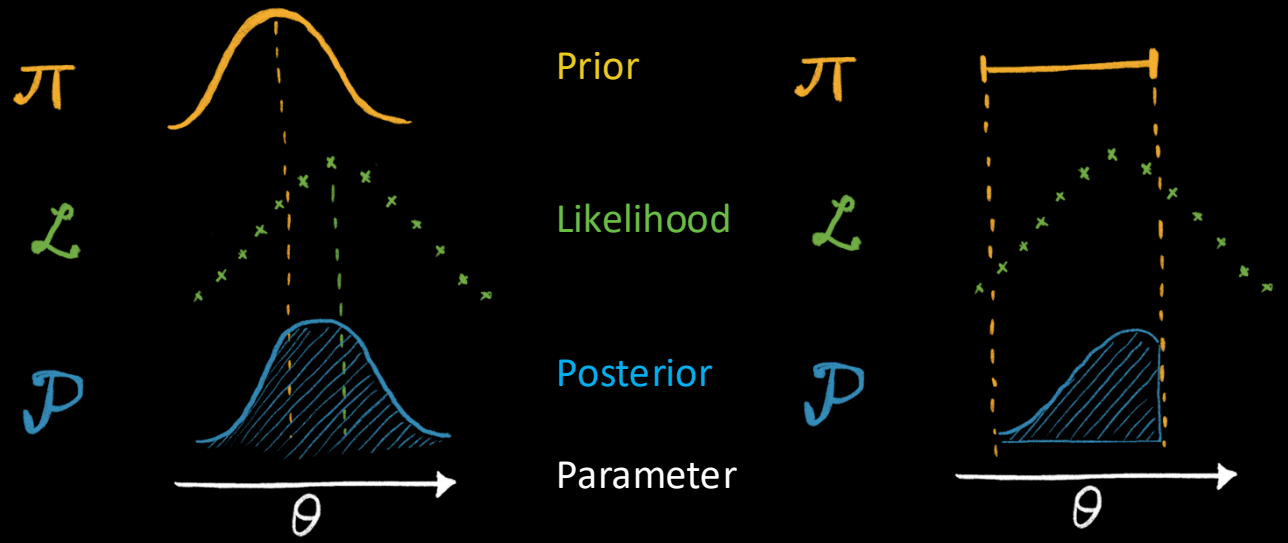
→ Lecture 2: frequentist profiles as a prior-free check

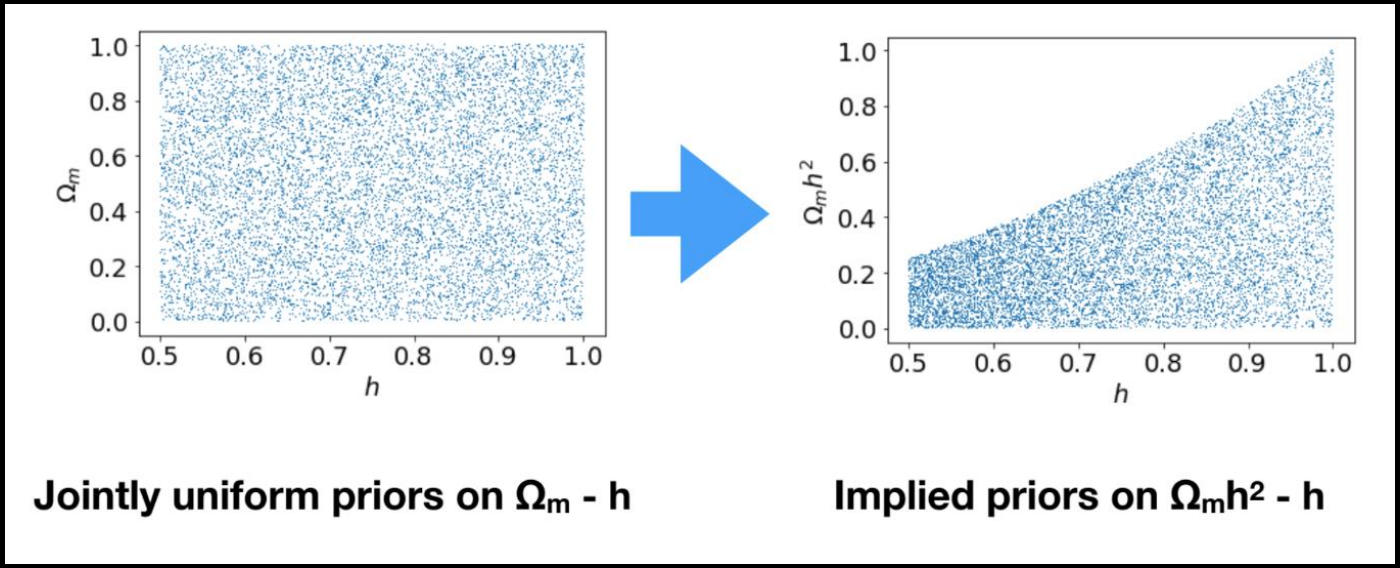


EDE review [2302.09032]

Priors are a modelling choice

‘Flat uninformative priors’ are parameterisation-dependent
eg. flat in θ vs flat in $\log\theta$





Graphic from Elisabeth Krause

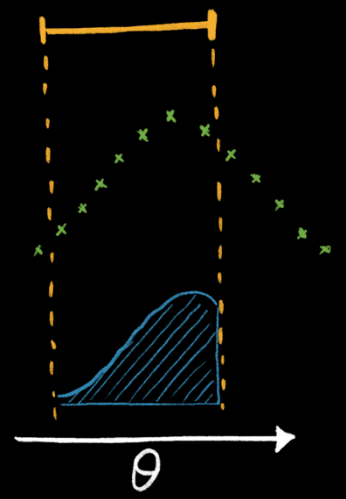
Choosing a prior is a modelling choice – state it and test sensitivity to it.

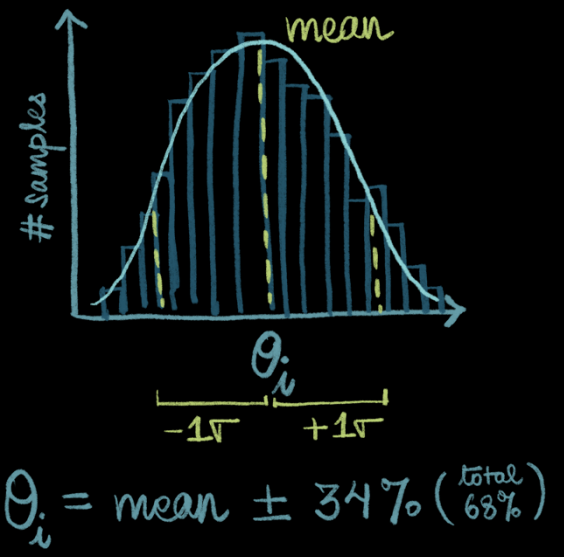
Prior π

Likelihood $\mathcal{L}$

Posterior $\mathcal{P}$

Parameter





Bayes theorem

$P(\theta|d) = \mathcal{L}(d|\theta)\Pi(\theta)$

Prior dependence

Sampling algorithm

Metropolis-Hastings MCMC

Existing infrastructure

Diagnostics

- Burn-in
- Initial guesses
- Step-sizes
- Acceptance rates
- Autocorrelation
- Effective number of steps
- Convergence
- Covariant proposals

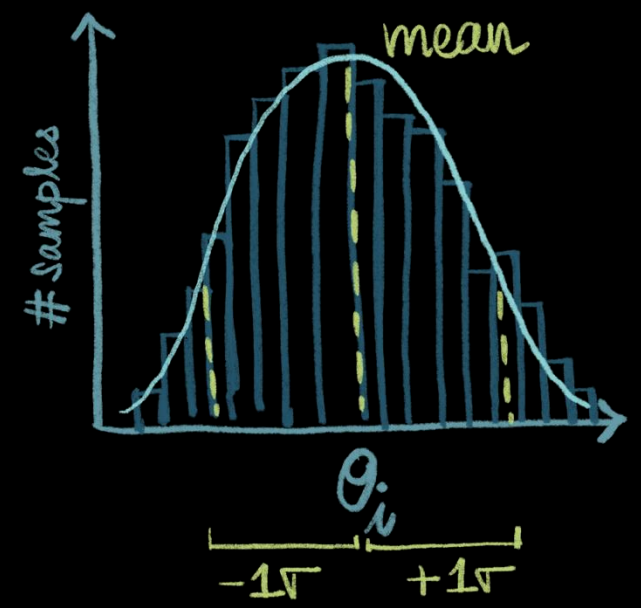
Constraints

Density of samples

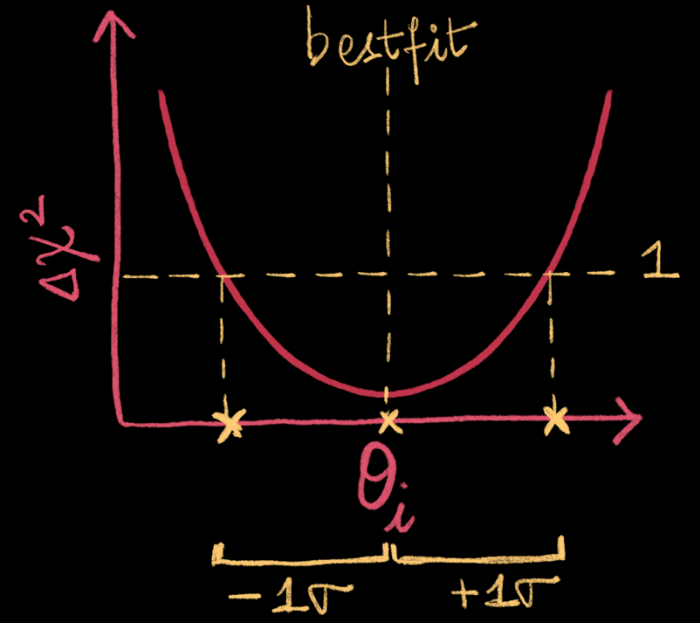
Statement about belief

Next: frequentist statistics & profile likelihoods

Same likelihood · different philosophy · a prior-free check



θ_i = mean $\pm$ 34% (total 68%)



θ_i = bestfit $\pm$ $\Delta\chi^2 = 1$ crossing

Beyond MH MCMCs — two families to know

- Nested sampling

Computes the evidence $\mathcal{E}(d)$ for model comparison, and handles multimodal posteriors.
Tools: MultiNest, PolyChord, dynesty

- Gradient-based samplers (HMC / NUTS)

Use the gradient of the log-posterior to move far with high acceptance — win big in high dimensions.
Tools: NumPyro, Stan, blackjax

Four more terms worth knowing

- Importance sampling / reweighting

reuse an existing chain under a new prior or added dataset — no full rerun.

- Posterior predictive checks (PPDs)

simulate mock data from the posterior and ask whether it looks like the real data.

- Tension metrics

quantify whether two datasets actually disagree. Toolkit: tensiometer.