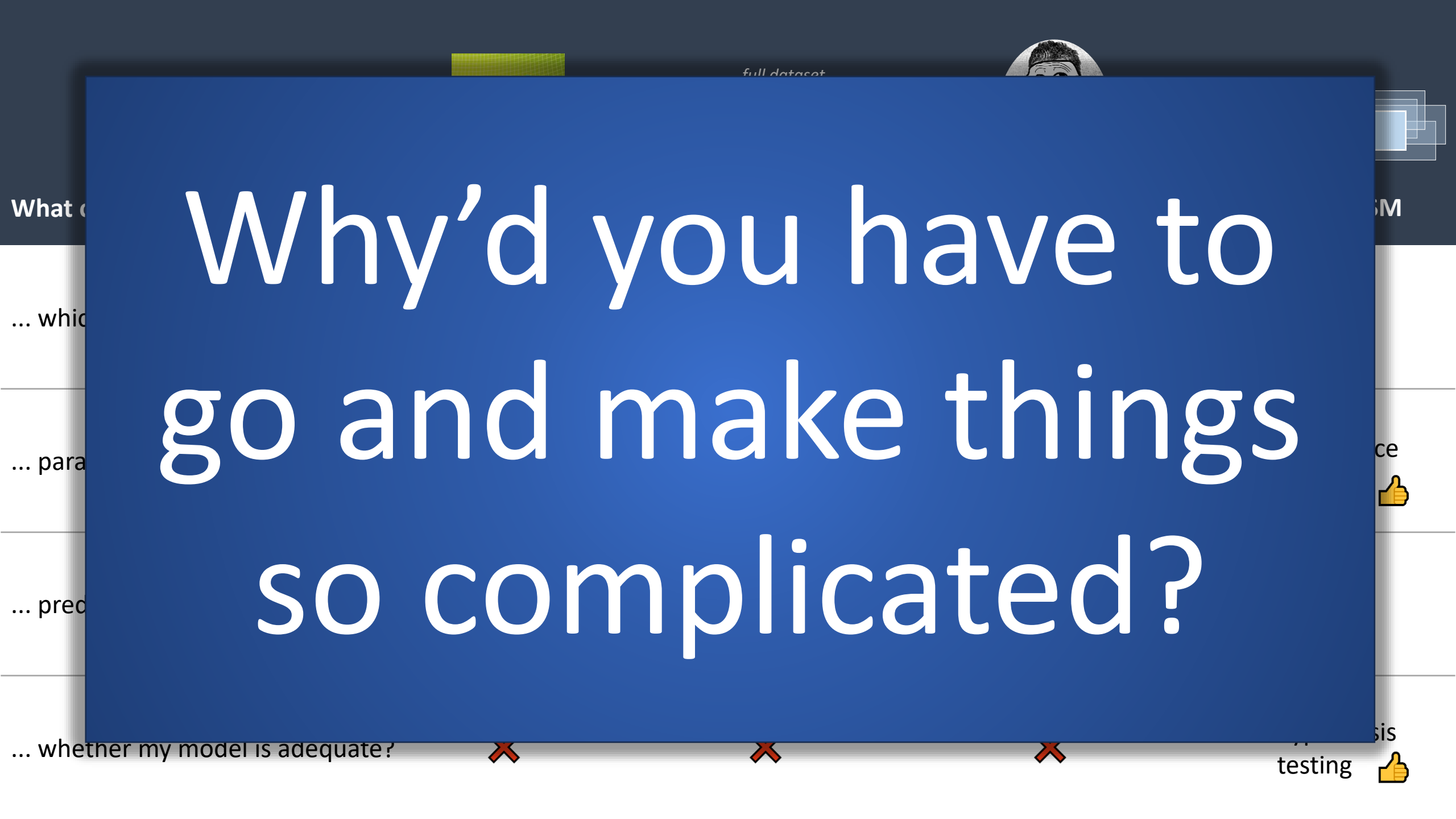








What can I say about ...	FITTING	HOLDOUT	BAYESIANISM	FREQUENTISM
... which model to choose?	≈	holdout log likelihood 👍	it's all just posterior distributions 👍	
... parameter values?	≈	≈	👍	confidence intervals 👍
... predicted outcomes?			👍	
... whether my model is adequate?	✗	✗	✗	hypothesis testing 👍


full dataset
What
... whic
... para
... pred
... whether my model is adequate?
testing
Why'd you have to go and make things so complicated?

All the swans I have seen are white.

Can I deduce that all swans are white?



The problem of induction



I have data, and I've learned something from it. What can I say about the future?
"I trained my model on a dataset. Will it work well on in-the-wild data?"

Philosophers call this *The Problem of Induction*, and it's what we've been studying.
It is IMPOSSIBLE to mathematically deduce conclusions about the future from a dataset,
so don't hanker after rigorous proofs or strict procedures for the right way to do it!



" Induction is the glory of Science and the scandal of Philosophy."

C.D. Broad
1926

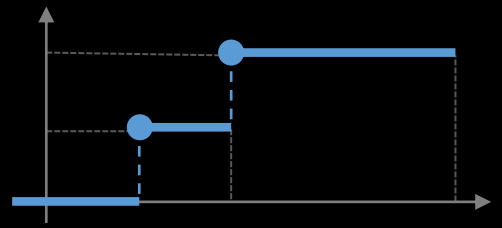
deduction : rigorous maths
induction : the practice of science

Naive fitting does not justify inductive claims

dataset

MODEL

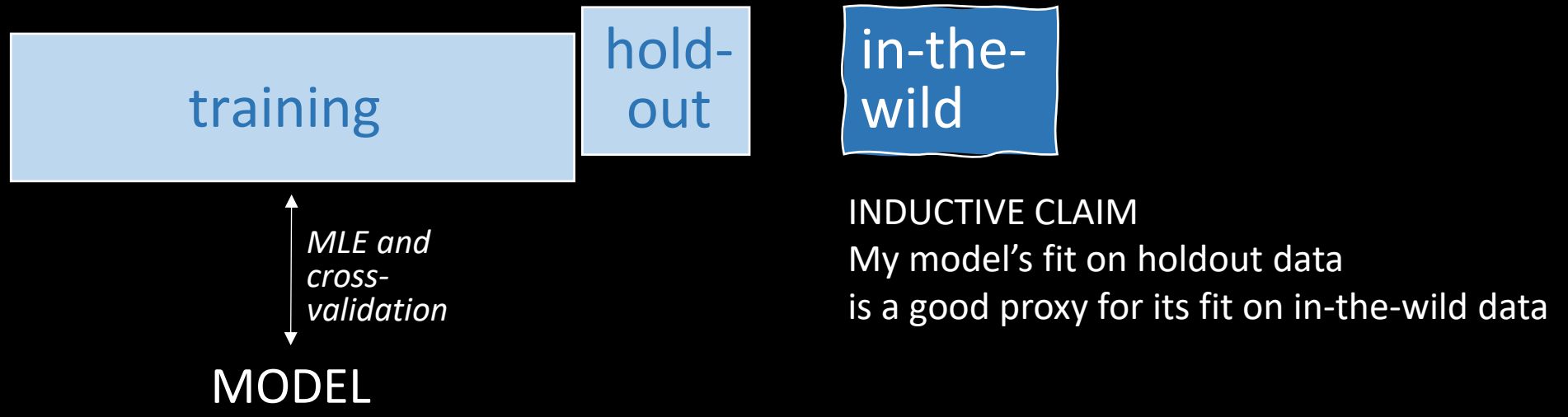
I can fit a model to the data, perfectly.



in-the-wild

But I have no grounds for believing that my model-fitted-to-dataset will fit equally well to in-the-wild data.

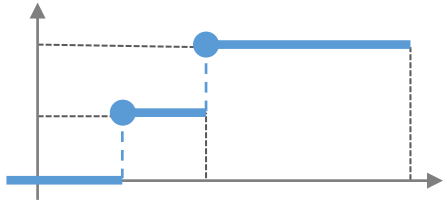
Holdout testing can support an inductive claim



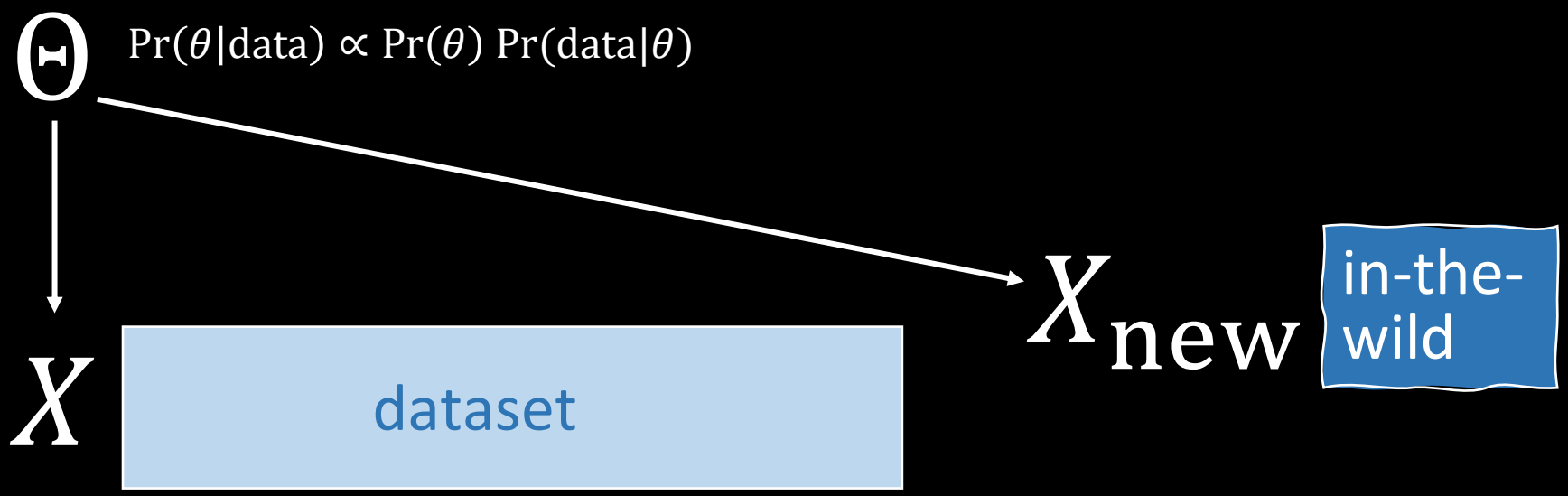
INDUCTIVE CLAIM
My model's fit on holdout data
is a good proxy for its fit on in-the-wild data





What can I say about ...	DEDUCTION	EMPIRICAL INDUCTION	SCHOLASTIC INDUCTION
... which model to choose?	I can measure how well a model fits my dataset, and I can choose the model that fits best. *(I won't make inductive claims about it — it overfits.)* 	*training data* *hold-out* *in-the-wild* Empirical evidence is what justifies my inductive claim about the model	*prior* *dataset* *in-the-wild* Prior beliefs are what justify my inductive claim about the model
... parameter values?			
... predicted outcomes?		Once we have a model, we can bootstrap it to answer questions about confidence and model adequacy *simulated dataset*	
... whether my model is adequate?			∅

Bayesianism makes an inductive claim:
that the same parameter governs both the dataset and in-the-wild data



Suppose I simplistically pick the most likely θ (called the MAP, Maximum A Posteriori).
In other words, I compute

$$\max_{\theta} \log \Pr(\theta|\text{data}) = \max_{\theta} \left\{ \log \Pr(\text{data}|\theta) + \log \Pr(\theta) \right\}$$

maximum likelihood estimation

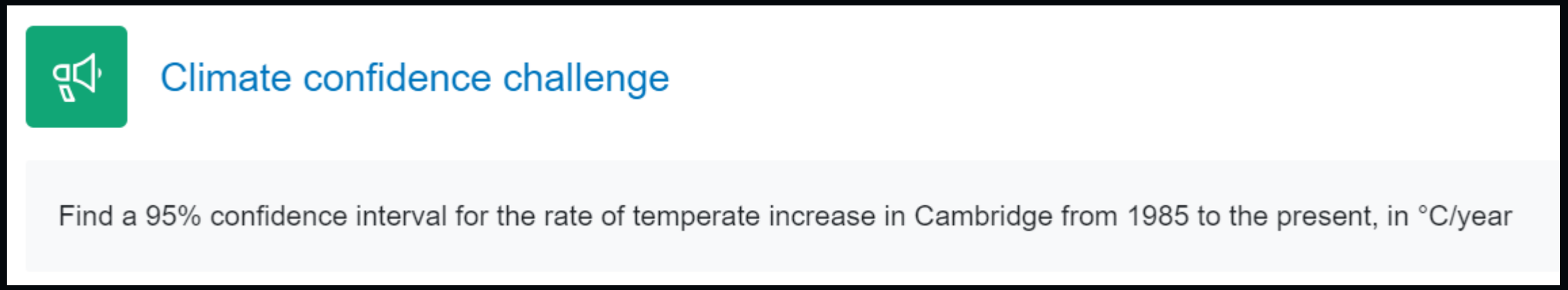
with a penalty for straying onto θ I don't believe in

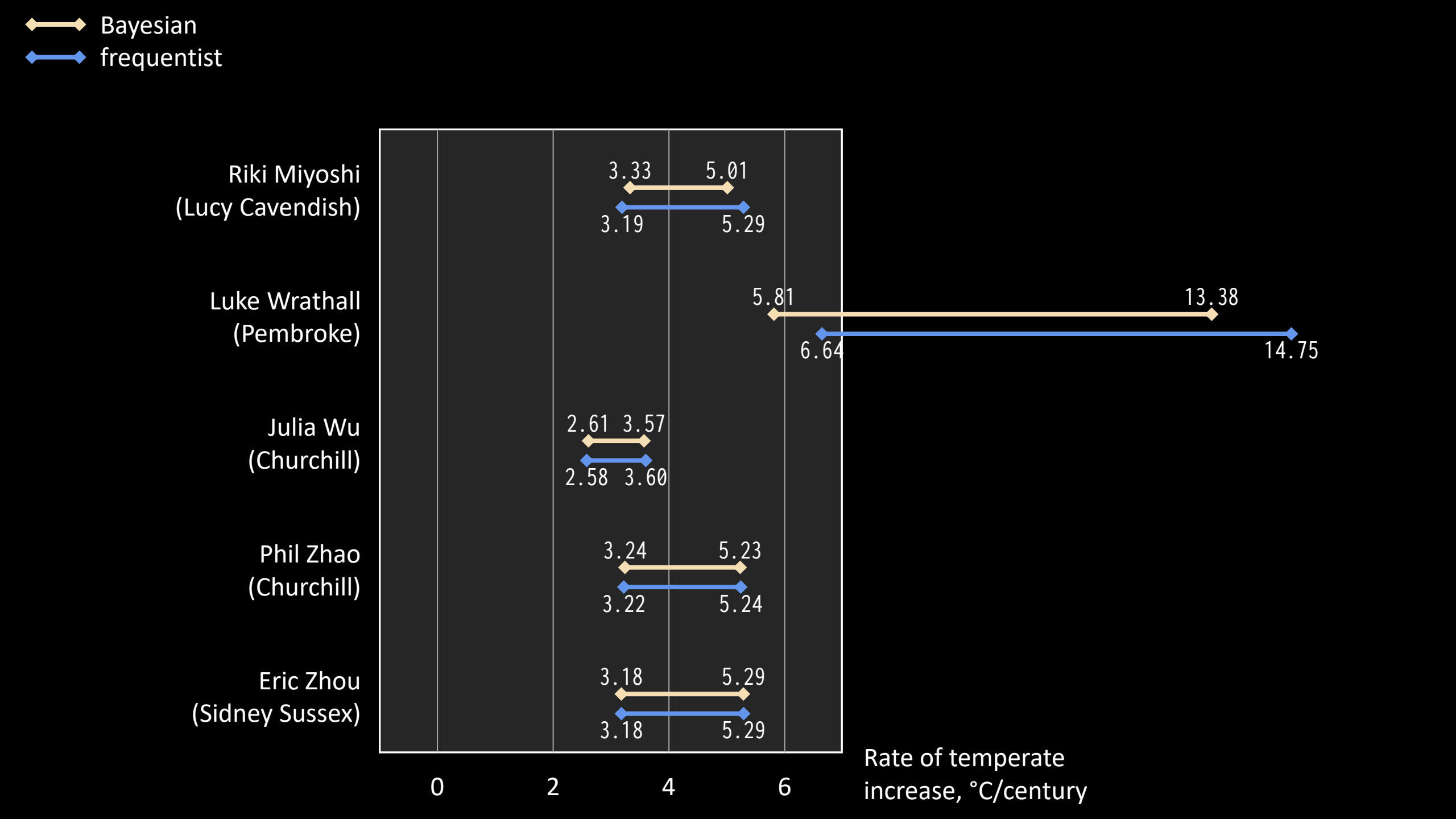
In effect, Bayesianism guards against overfitting by applying a *regularization penalty.*

THE GREAT GENERALIZATION

SMACKDOWN

BAYESIANIST v FREQUENTIST




Bayesian
frequentist
Riki Miyoshi
(Lucy Cavendish)
3.33
5.01
3.19
5.29
Luke Wrathall
(Pembroke)
5.81
13.38
6.64
14.75
Julia Wu
(Churchill)
2.61
3.57
2.58
3.60
Phil Zhao
(Churchill)
3.24
5.23
3.22
5.24
Eric Zhou
(Sidney Sussex)
3.18
5.29
3.18
5.29
0
2
4
6
Rate of temperate
increase, °C/century

Which side is ChatGPT on?

I have a dataset of student marks on an exam (numbers in the range 0 to 20). I also have a record of the gender of the student (male / female / other). I want to learn if there is gender bias. Give me code to do this.

I would start with differences in mean marks, using heteroskedasticity-robust inference and correcting for multiple comparisons.

$$H_0: \mu_{\text{male}} = \mu_{\text{female}} = \mu_{\text{other}}$$

The pairwise tests estimate differences in average marks, with 95% confidence intervals and Holm-adjusted p-values.

I would pay at least as much attention to the estimated differences and their confidence intervals as to statistical significance. For example, an estimated difference of 0.15 marks on a 20-point scale might be statistically significant in a large dataset but of little practical importance.