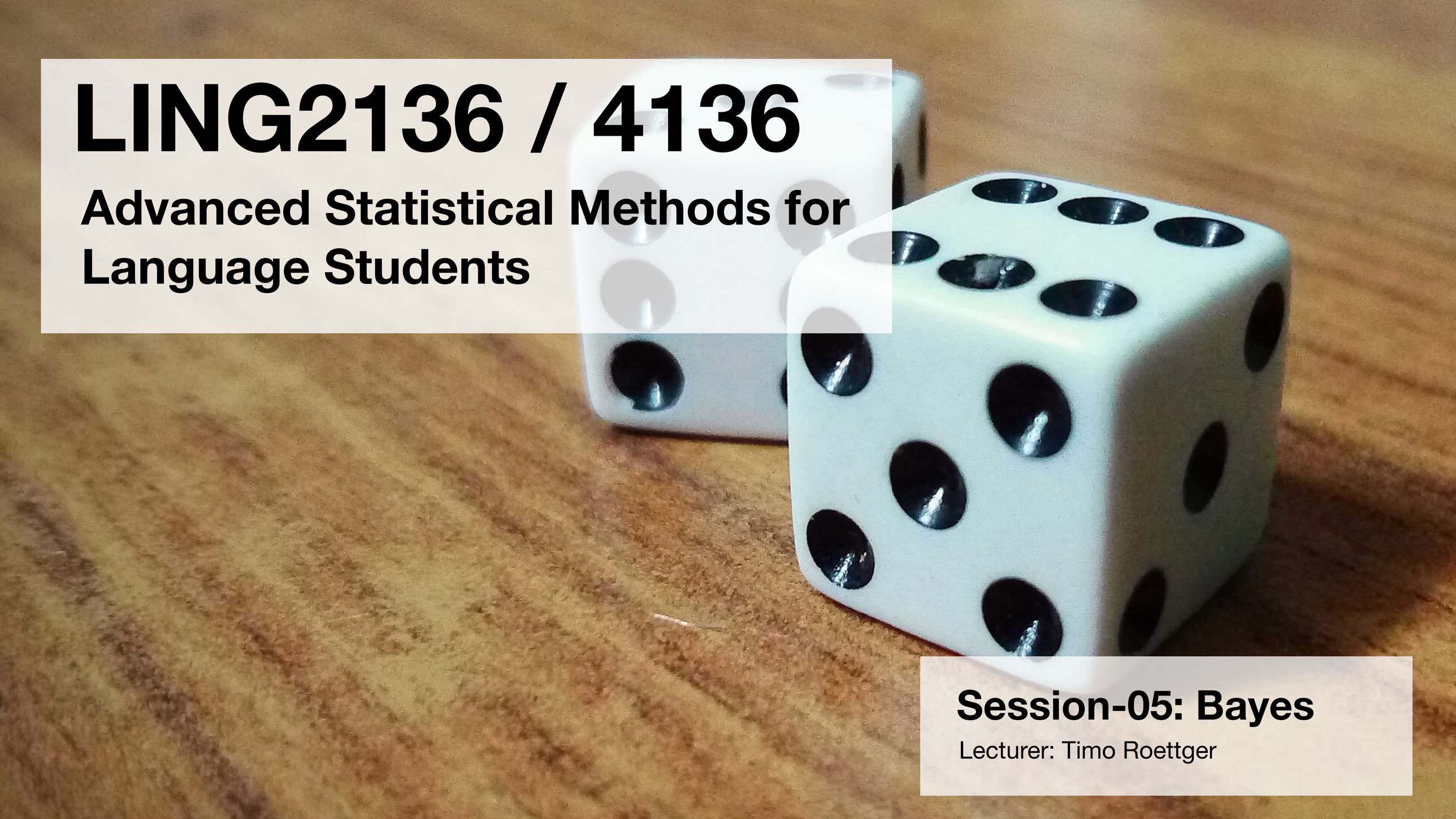
LING2136 / 4136

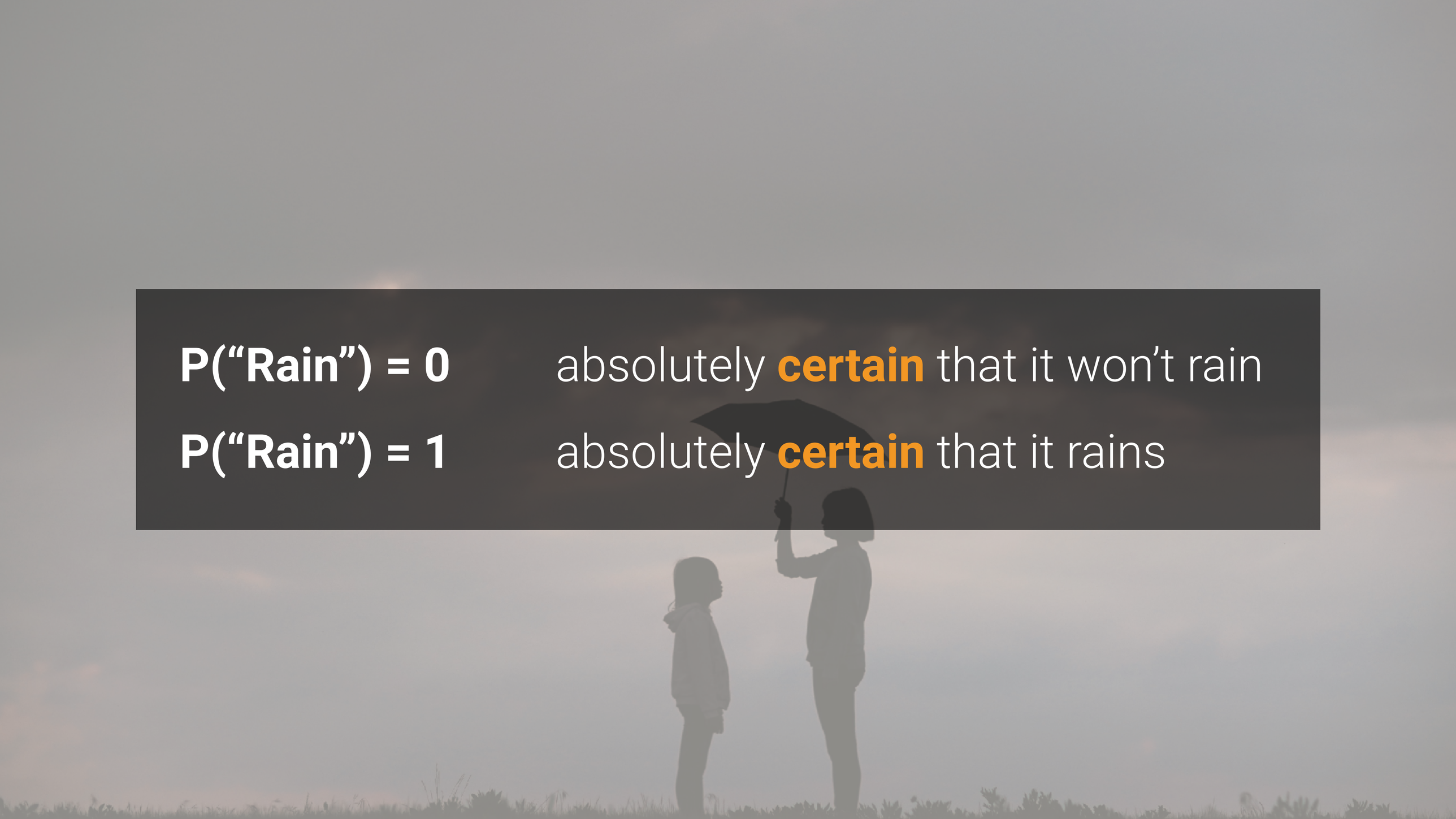
Advanced Statistical Methods for Language Students



Session-05: Bayes

Lecturer: Timo Roettger

REASONING ABOUT PROBABILITY



P(“Rain”) = 0 absolutely **certain** that it won’t rain

P(“Rain”) = 1 absolutely **certain** that it rains


P(Rain) = 0.3
P(¬Rain) = 0.7

I observe evidence: 

P(Rain) = 0.3

P(¬Rain) = 0.7

P(Clouds | Rain) = 0.8

P(Clouds | ¬Rain) = 0.2


DO I CHANGE MY BELIEF?

I observe evidence: 

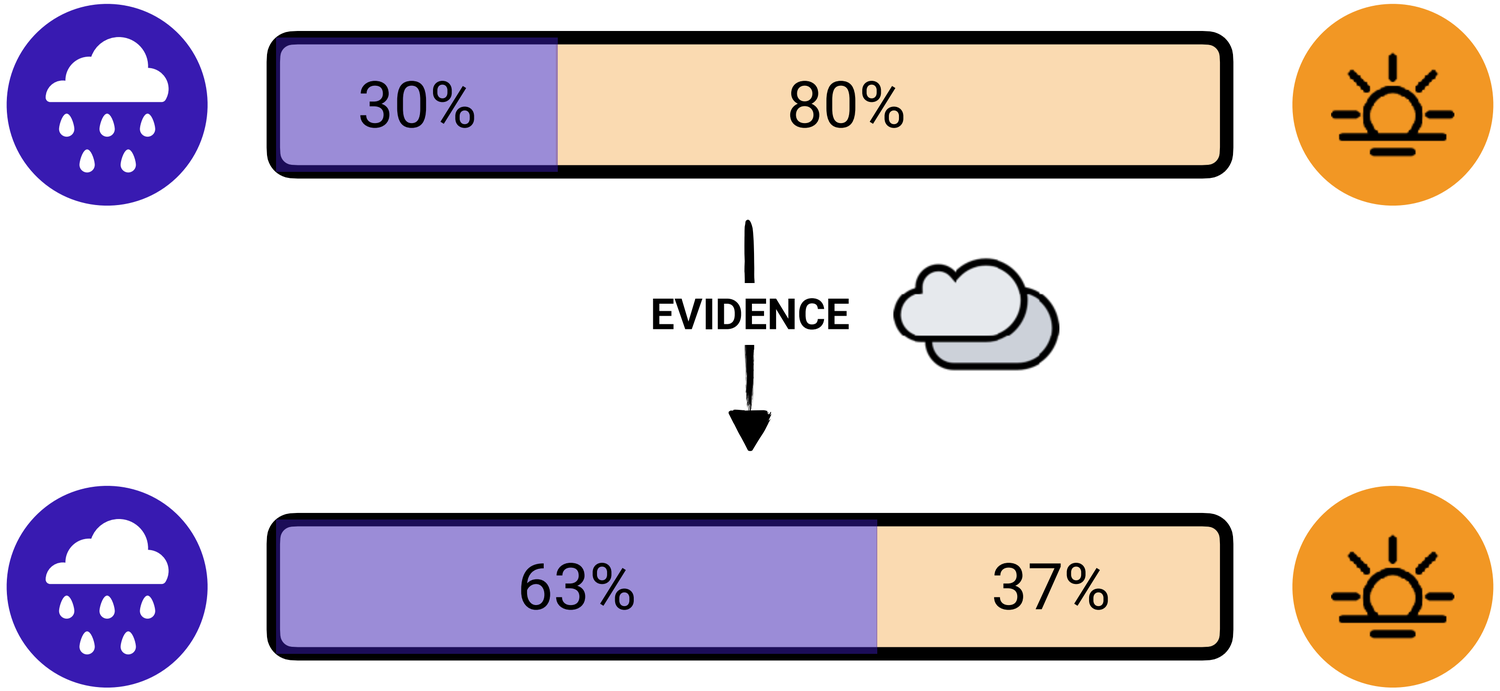
$$P(\text{Rain} \mid \text{Clouds}) = \frac{24}{24 + 14} \approx 0.63$$

P(Rain) = 0.3

P(¬Rain) = 0.7

P(Clouds | Rain) = 0.8

P(Clouds | ¬Rain) = 0.2


30%
80%
EVIDENCE
63%
37%

When to use Bayes' theorem

You have a hypothesis:

It's gonna rain

You have observed evidence:

There are dark clouds

You want to know:

$$P(\text{H} \mid \text{E})$$

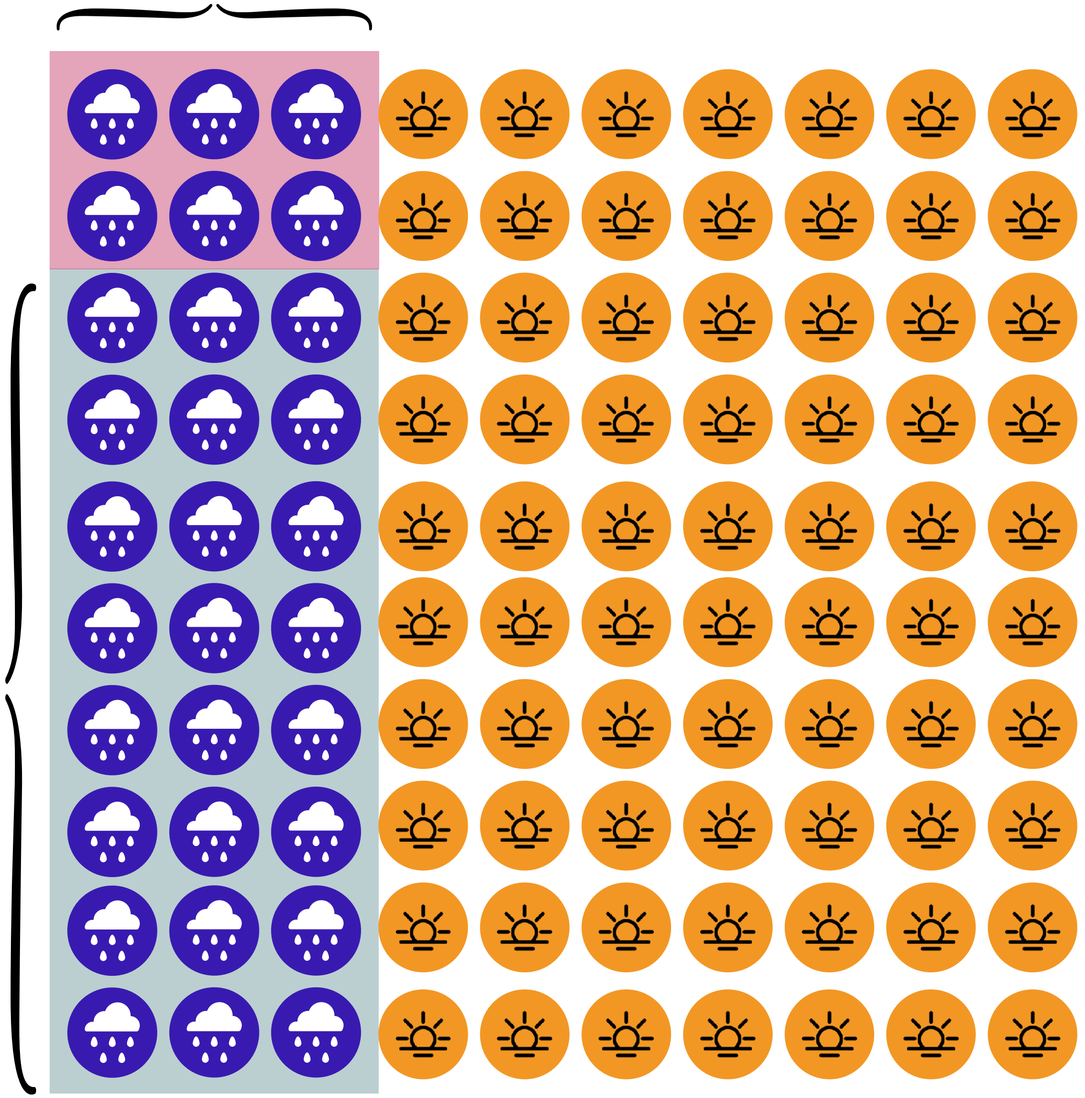
$$P\left(\begin{array}{c}\text{hypothesis}\\ \text{given}\\ \text{evidence}\end{array}\right)$$

“Prior” → $P(\text{Hypothesis}) = 30 / 100 = 0.3$

“Prior” → $P(\text{Hypothesis}) = 30 / 100 = 0.3$

“Likelihood”

$P(\text{Evidence} \mid \text{H}) = 0.8$



“Prior” → $P(\text{Hypothesis}) = 30 / 100 = 0.3$

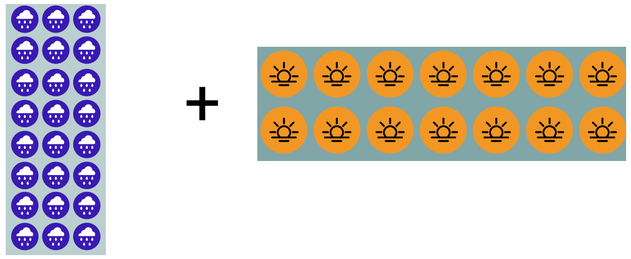
“Likelihood”

$P(\text{Evidence} \mid \text{H}) = 0.8$

¬ means “not”

$P(\text{E} \mid \neg\text{H}) = 0.2$

$$P(H \mid E) = \frac{0.3 * 0.8 = 0.24}{0.3 * 0.8 + 0.7 * 0.2 = 0.38} = \frac{P(H)\ P(E \mid H)}{P(H)\,P(E \mid H) + P(\neg H)\,P(E \mid \neg H)}$$



“Prior” **“Likelihood”**

$P(H) = 0.3$

$P(E \mid H) = 0.8$

$P(\neg H) = 0.7$

$P(E \mid \neg H) = 0.2$

“Prior” “Likelihood”

$$P(H \mid E) = \frac{P(H)\ P(E \mid H)}{P(E)} = \frac{P(H)\ P(E \mid H)}{P(H)\,P(E \mid H) + P(\neg H)\,P(E \mid \neg H)}$$

“Posterior”

$P(H) = 0.3$

$P(E \mid H) = 0.8$

$P(\neg H) = 0.7$

$P(E \mid \neg H) = 0.2$

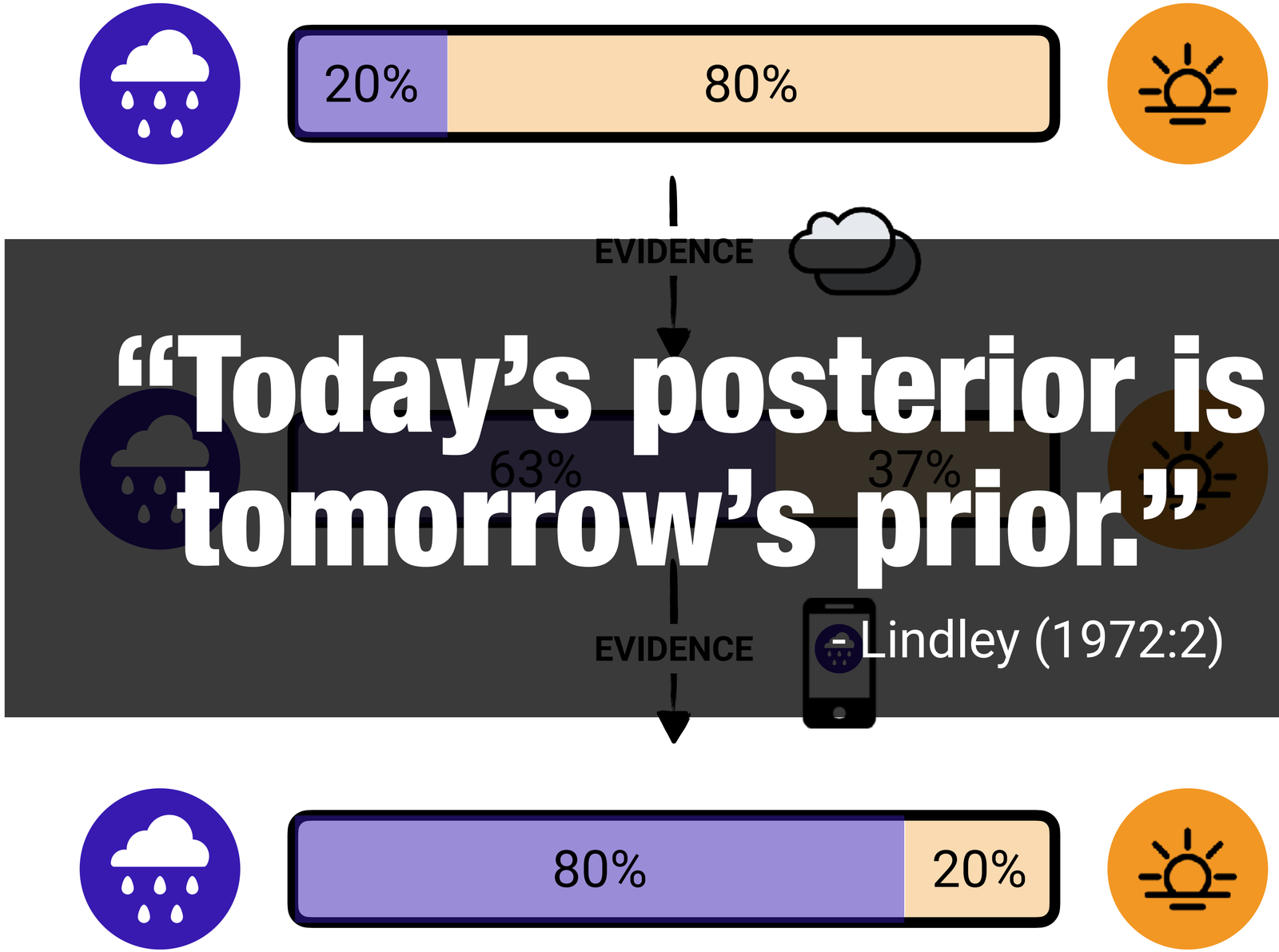
This is Bayes Theorem

prior: initial degree of belief in hypothesis

likelihood: the probability of the evidence given the hypothesis

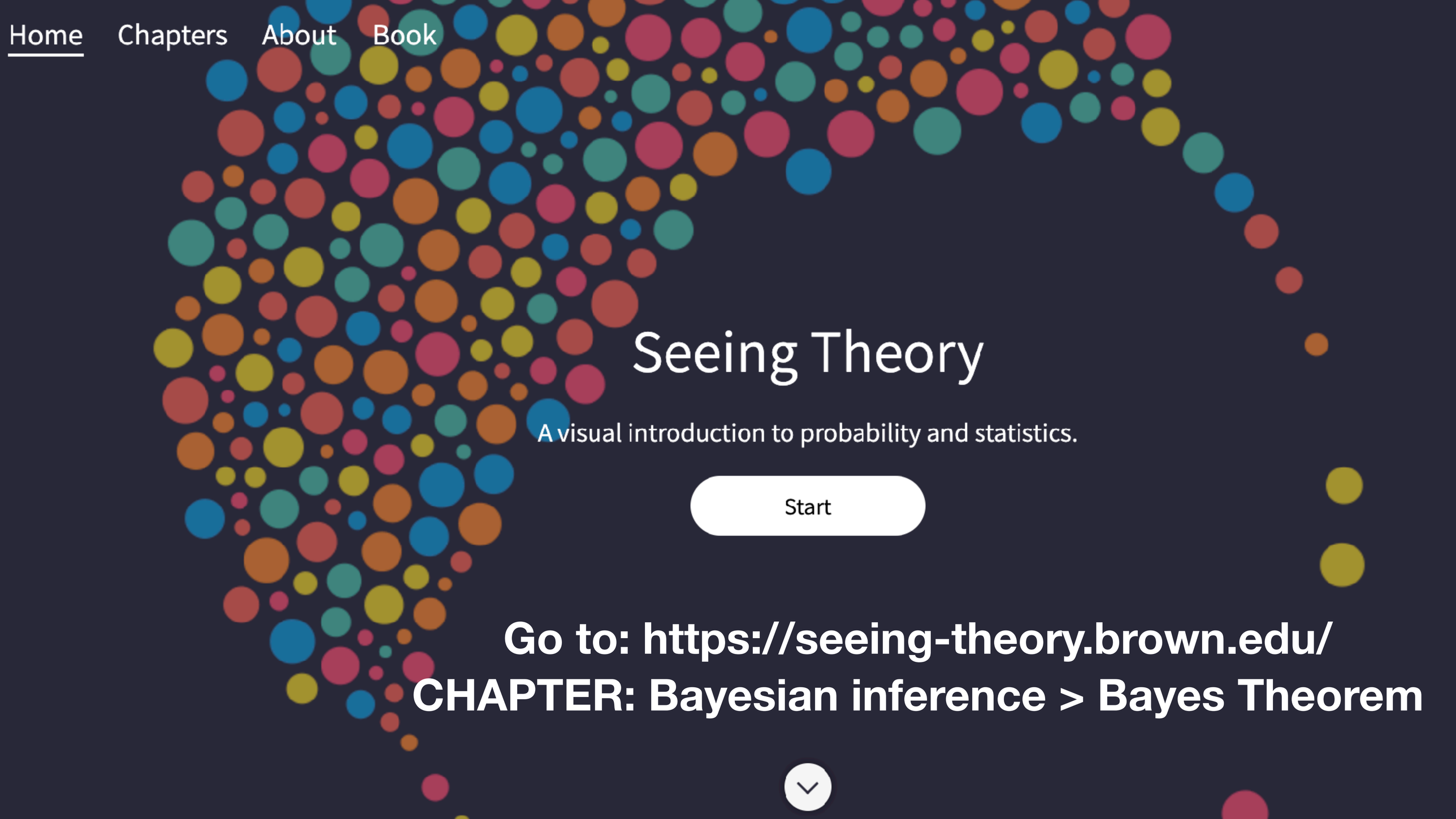
$$P(\mathbf{H} \mid \mathbf{E}) = \frac{P(\mathbf{H})\ P(\mathbf{E} \mid \mathbf{H})}{P(\mathbf{E})}$$

posterior: degree of belief in hypothesis, after seeing evidence



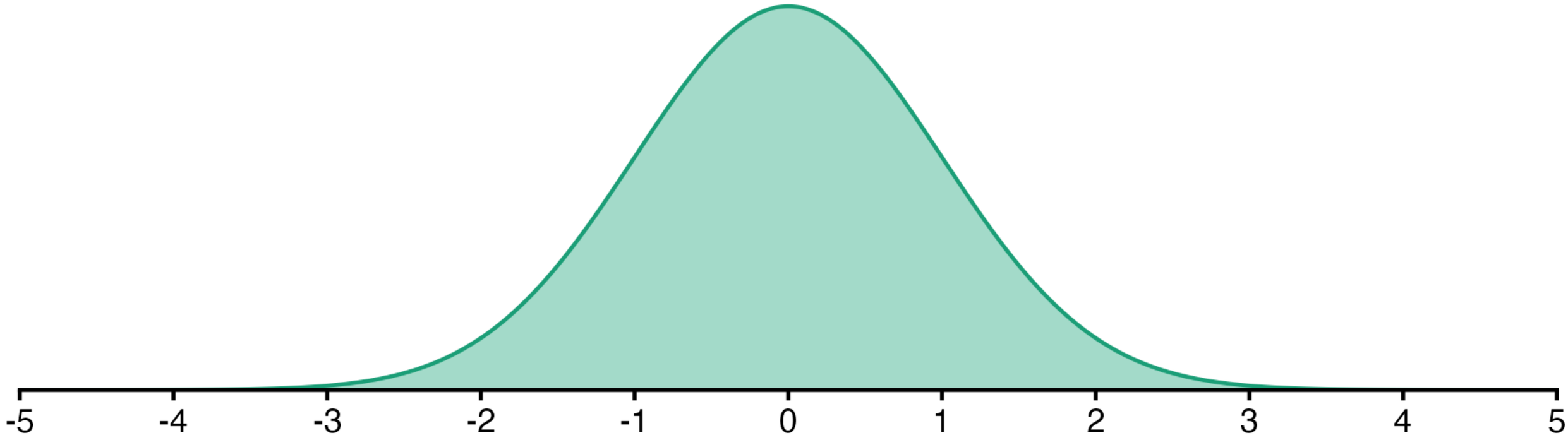
“Today’s posterior is tomorrow’s prior.”

– Lindley (1972:2)


Home
Chapters
About
Book
Seeing Theory
A visual introduction to probability and statistics.
Start
Go to: https://seeing-theory.brown.edu/
CHAPTER: Bayesian inference > Bayes Theorem

The Gaussian (normal) distribution

has parameters: mean (μ) & standard deviation (σ)



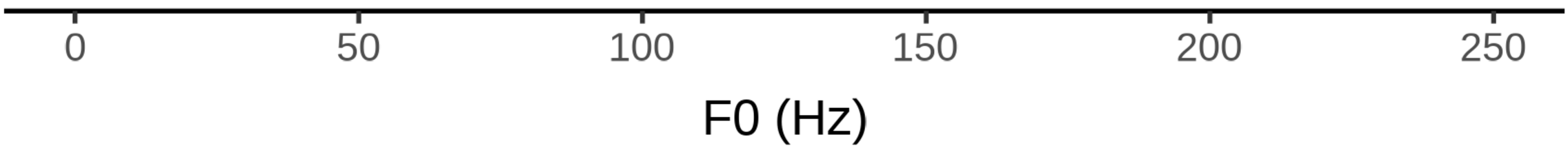
Demo with fundamental frequency (F0) data
(corresponds to ‘voice pitch’ in perception)

data: average F0 from 32 male speakers

Winter et al. 2021

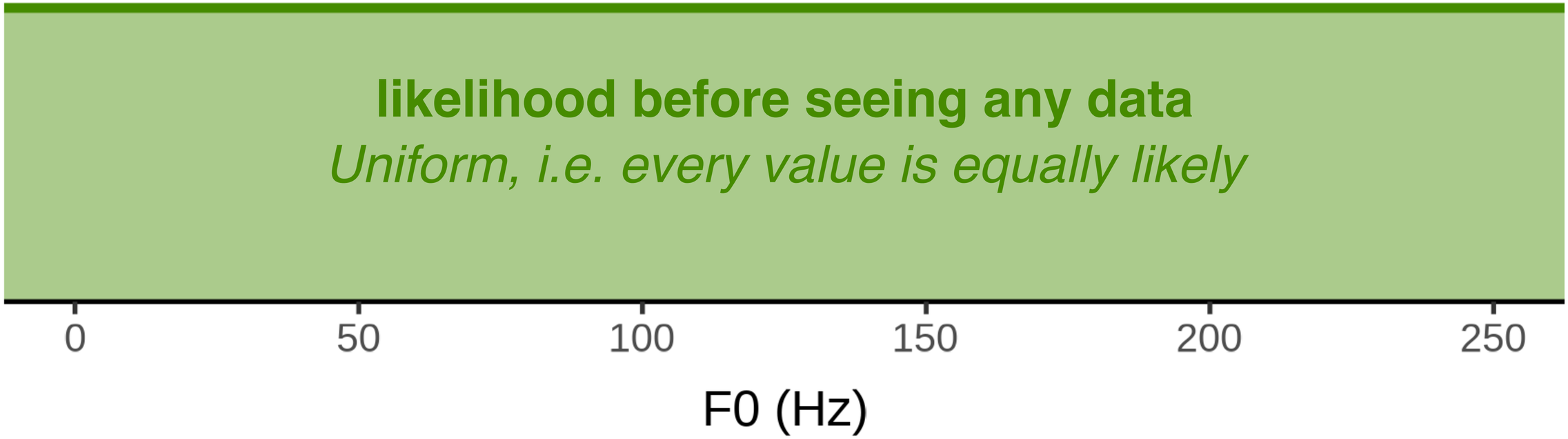
f0 ~ 1

What we want to know:
given some F0 measurements,
what is the average F0 of a male speaker



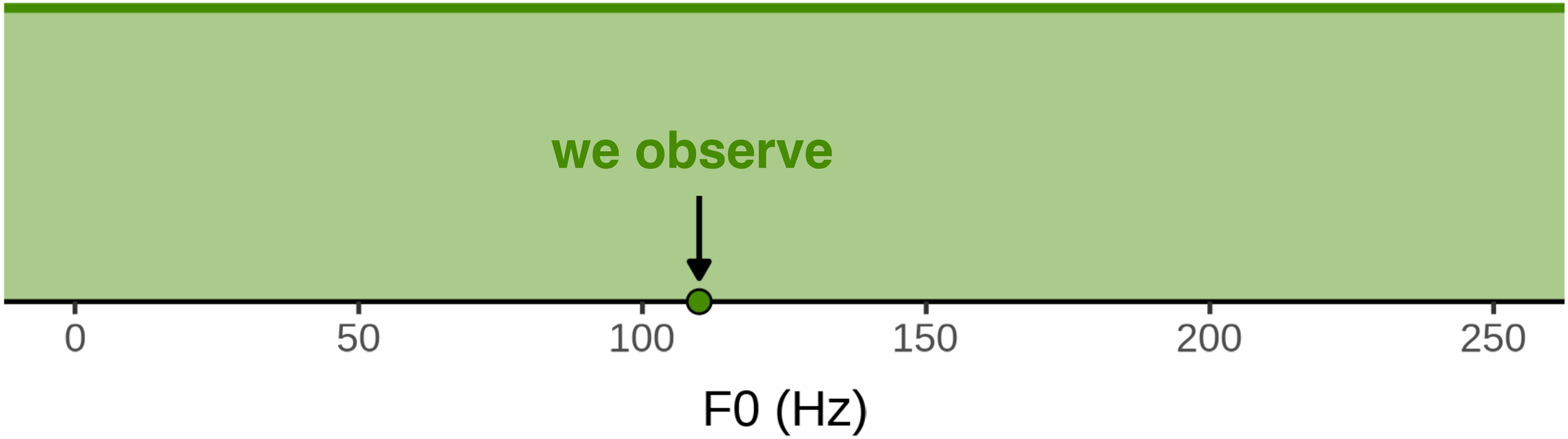
Likelihood functions

Knowing the data,
what is the **likelihood** of
the μ parameter?



Likelihood functions

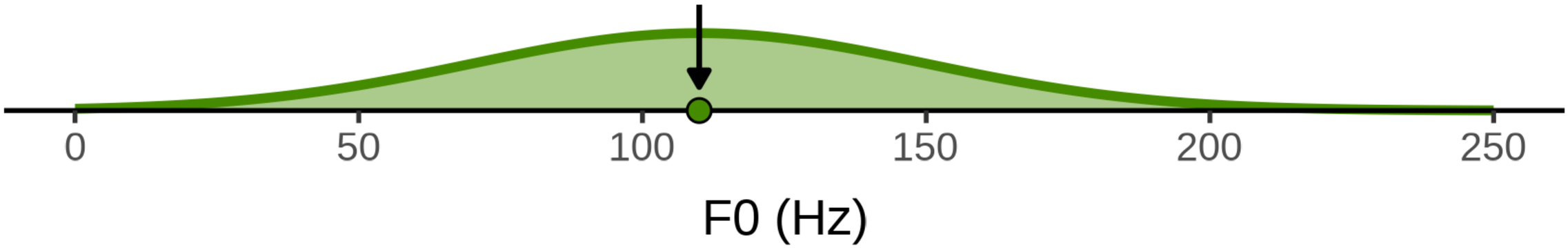
Knowing the data,
what is the **likelihood** of
the μ parameter?



Likelihood functions

Knowing the data,
what is the **likelihood** of
the μ parameter?

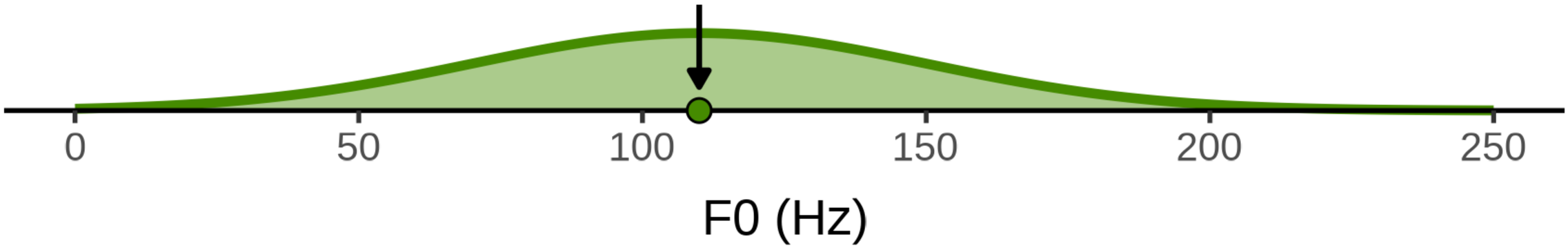
The likelihood function for μ



Likelihood functions

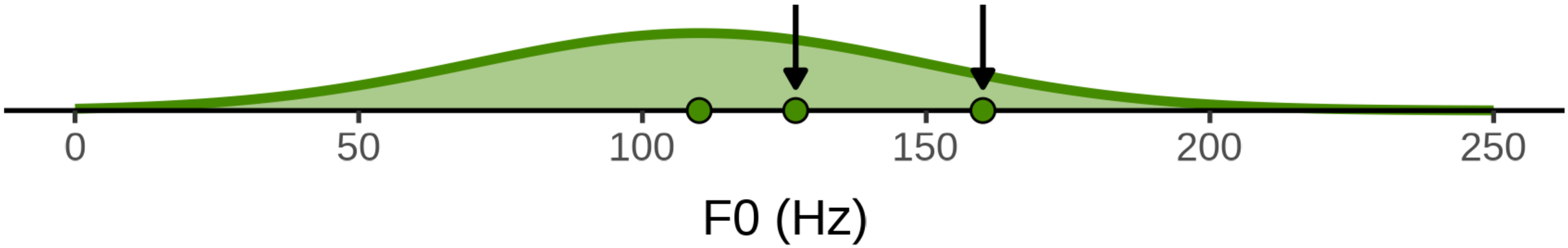
Knowing the data,
what is the **likelihood** of
the μ parameter?

This one data point is **compatible** with
a **whole range** of values for **μ**



Likelihood functions

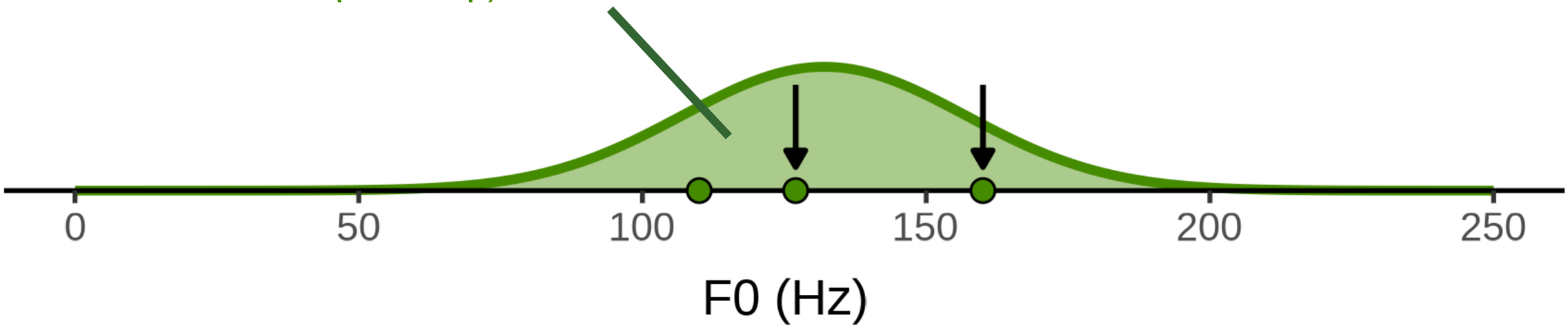
Knowing the data,
what is the **likelihood** of
the μ parameter?



Likelihood functions

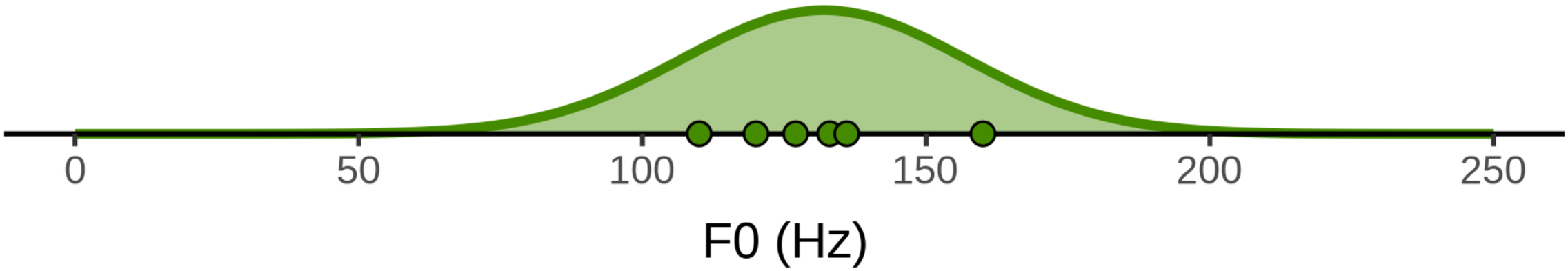
Knowing the data,
what is the **likelihood** of
the μ parameter?

more data make the likelihood function **narrower** (you are accumulating evidence for a specific μ)



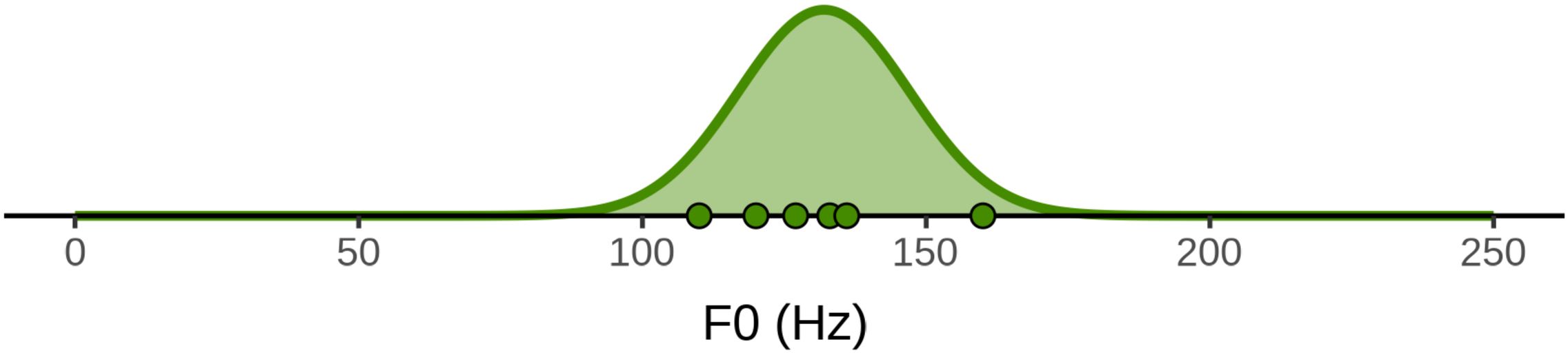
Likelihood functions

Knowing the data,
what is the **likelihood** of
the μ parameter?



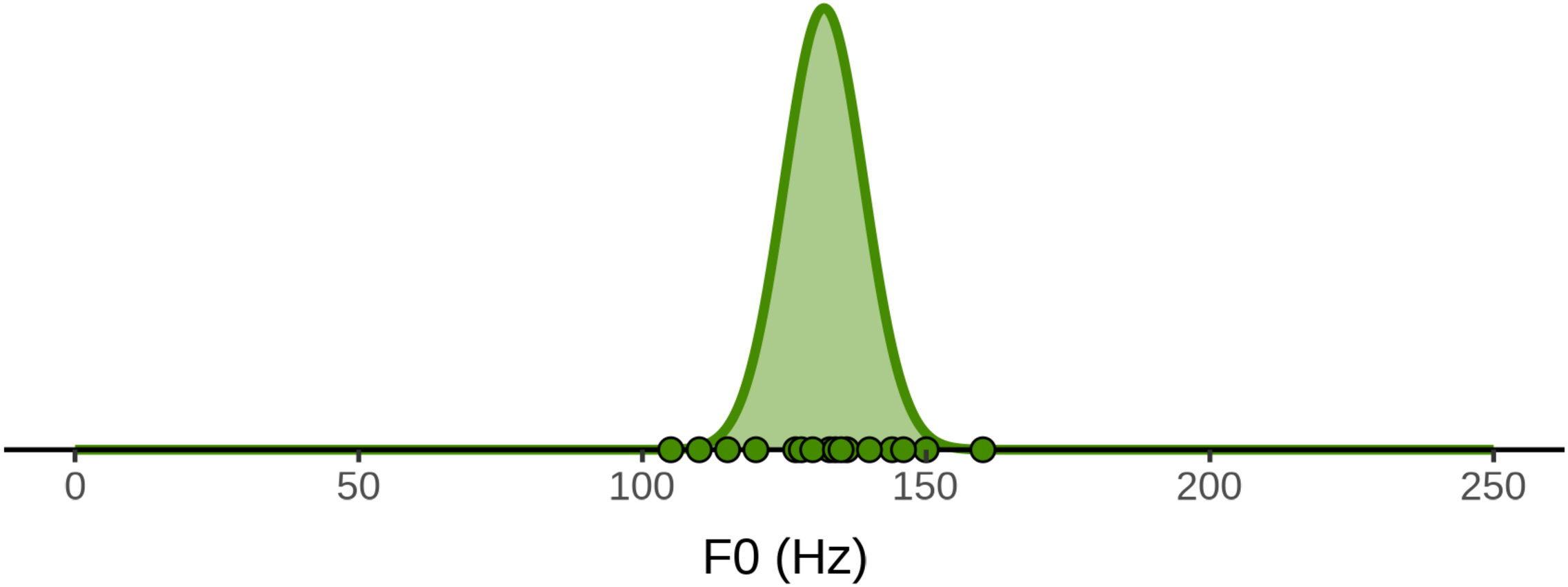
Likelihood functions

Knowing the data,
what is the **likelihood** of
the μ parameter?



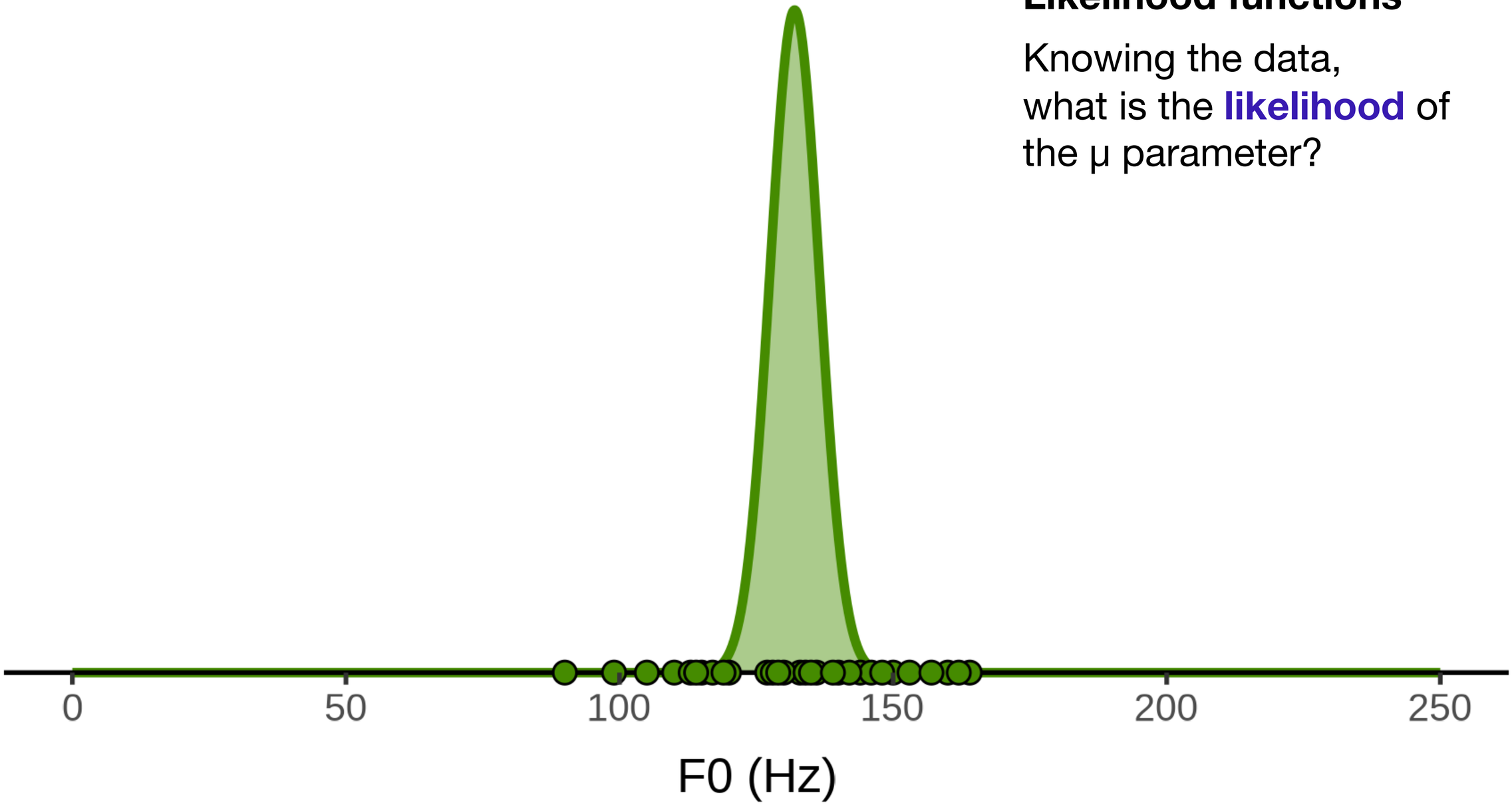
Likelihood functions

Knowing the data,
what is the **likelihood** of
the μ parameter?



Likelihood functions

Knowing the data,
what is the **likelihood** of
the μ parameter?



This is Bayes Theorem

prior: initial degree of belief in hypothesis

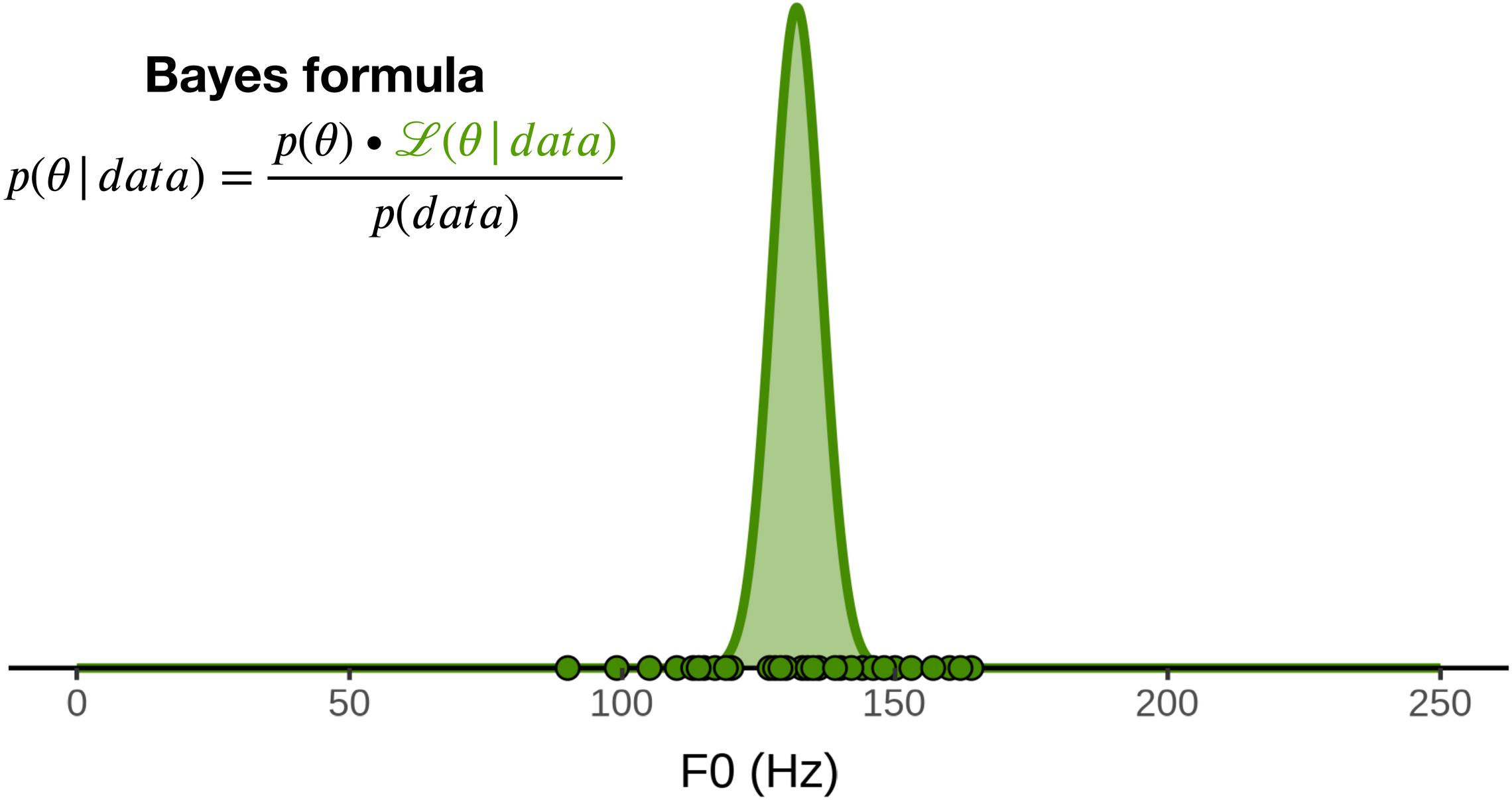
likelihood: the probability of the evidence given the hypothesis

$$P(\mathbf{H} \mid \mathbf{E}) = \frac{P(\mathbf{H})\ P(\mathbf{E} \mid \mathbf{H})}{P(\mathbf{E})}$$

posterior: degree of belief in hypothesis, after seeing evidence

Bayes formula

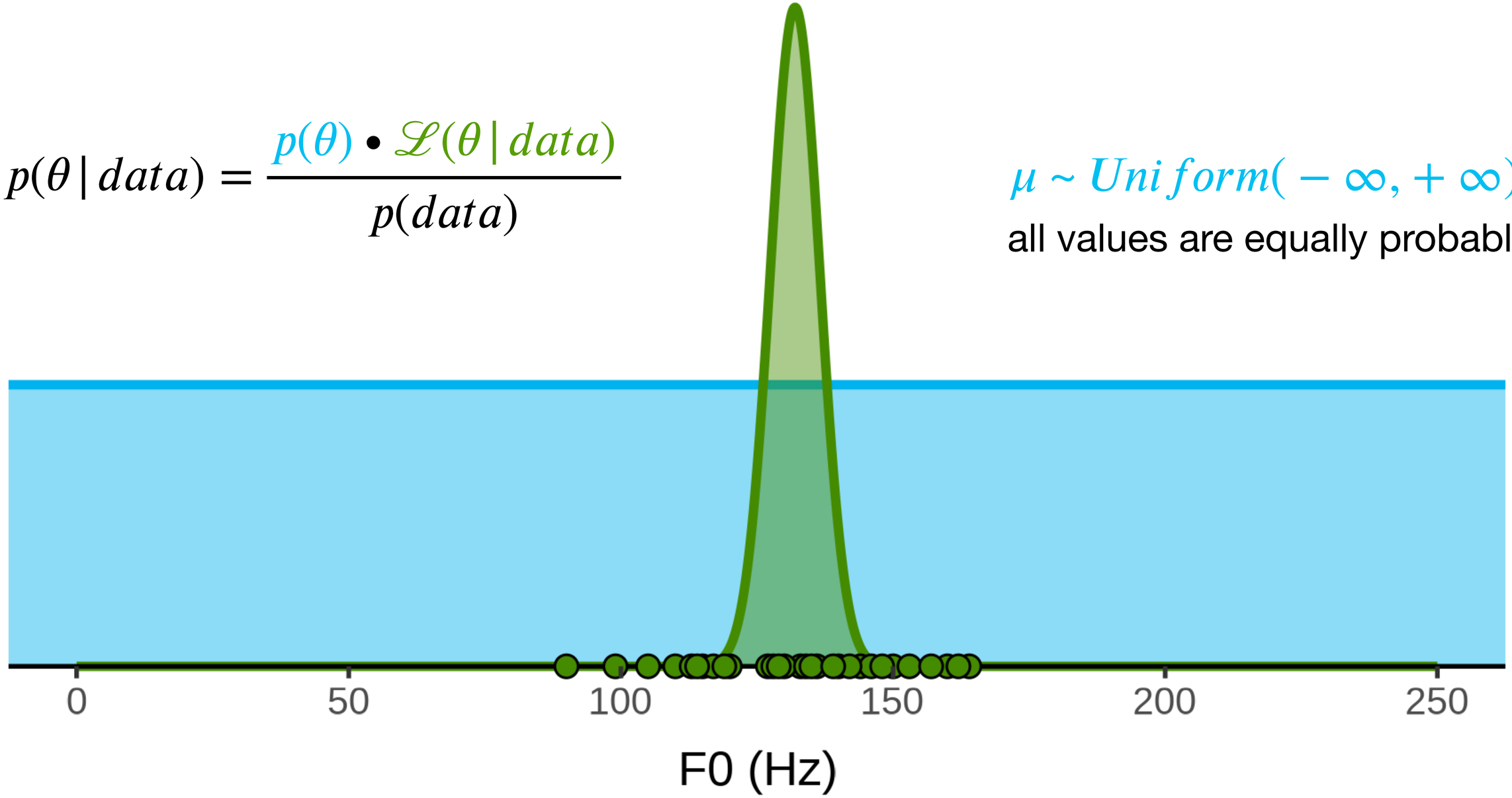
$$p(\theta \mid data) = \frac{p(\theta) \bullet \mathcal{L}(\theta \mid data)}{p(data)}$$



$$p(\theta \mid data) = \frac{p(\theta) \bullet \mathcal{L}(\theta \mid data)}{p(data)}$$

$$\mu \sim Uniform(-\infty, +\infty)$$

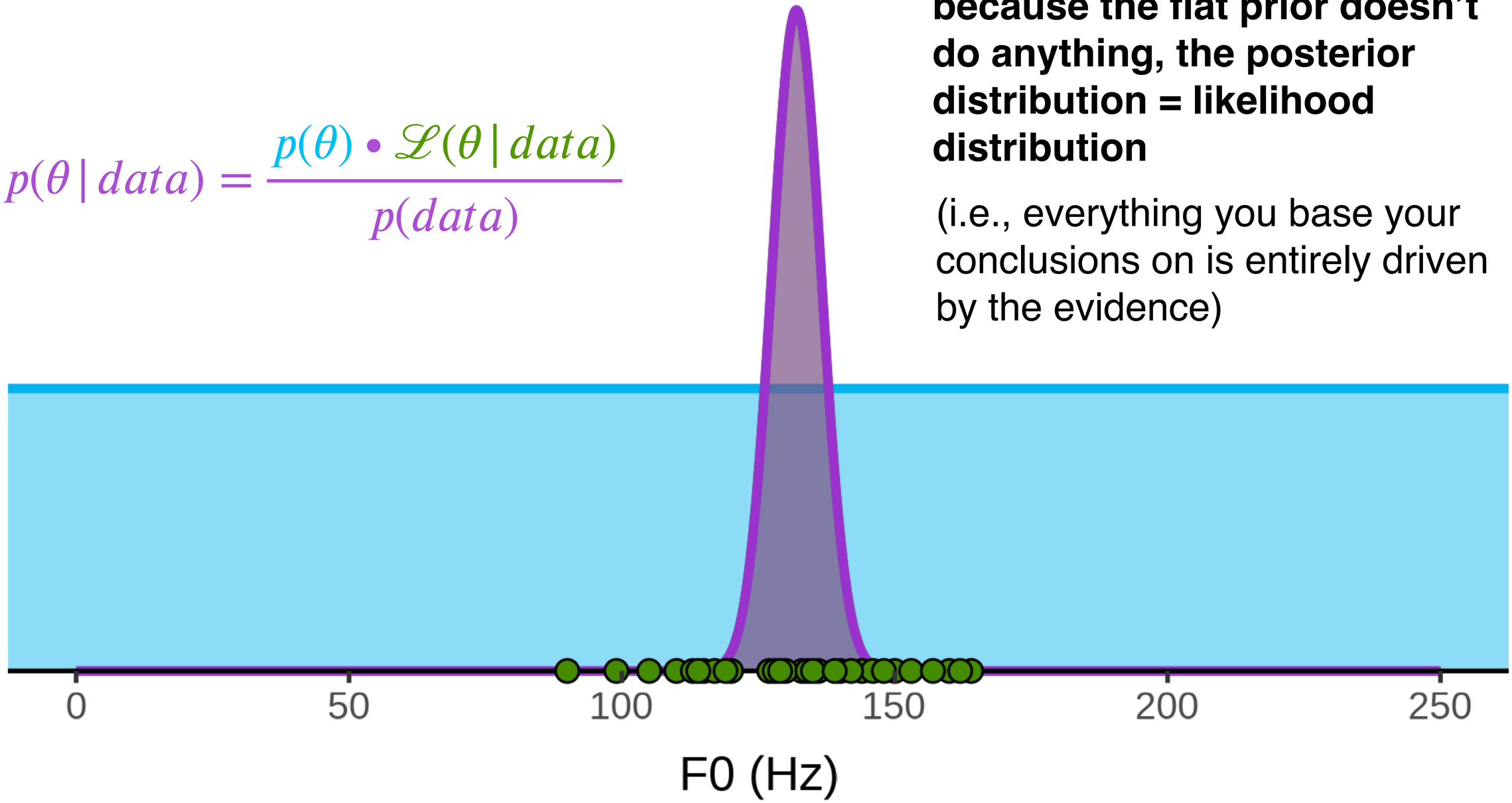
all values are equally probable



$$p(\theta \mid data) = \frac{p(\theta) \bullet \mathscr{L}(\theta \mid data)}{p(data)}$$

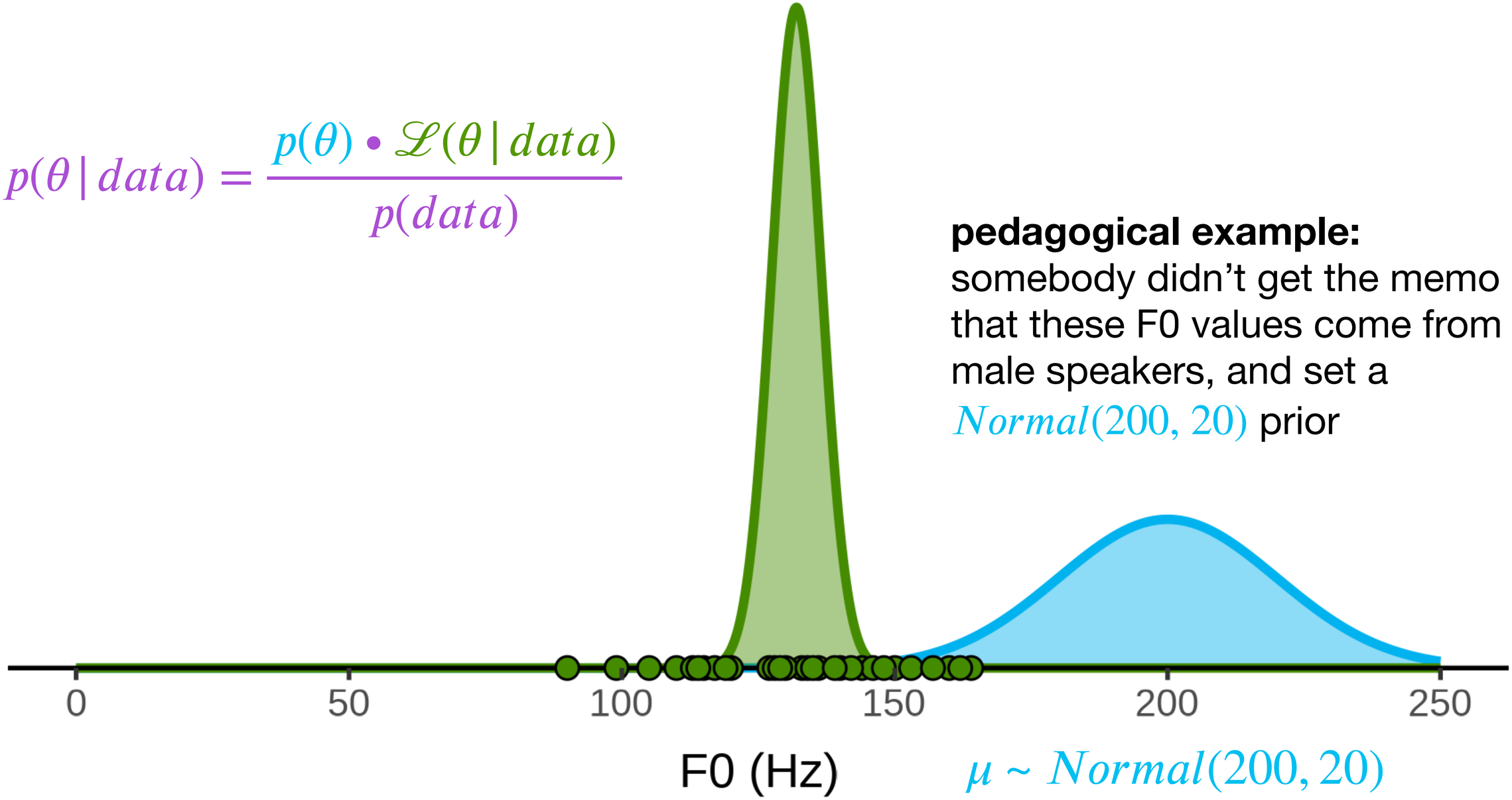
because the flat prior doesn't do anything, the posterior distribution = likelihood distribution

(i.e., everything you base your conclusions on is entirely driven by the evidence)



$$p(\theta \mid data) = \frac{p(\theta) \bullet \mathscr{L}(\theta \mid data)}{p(data)}$$

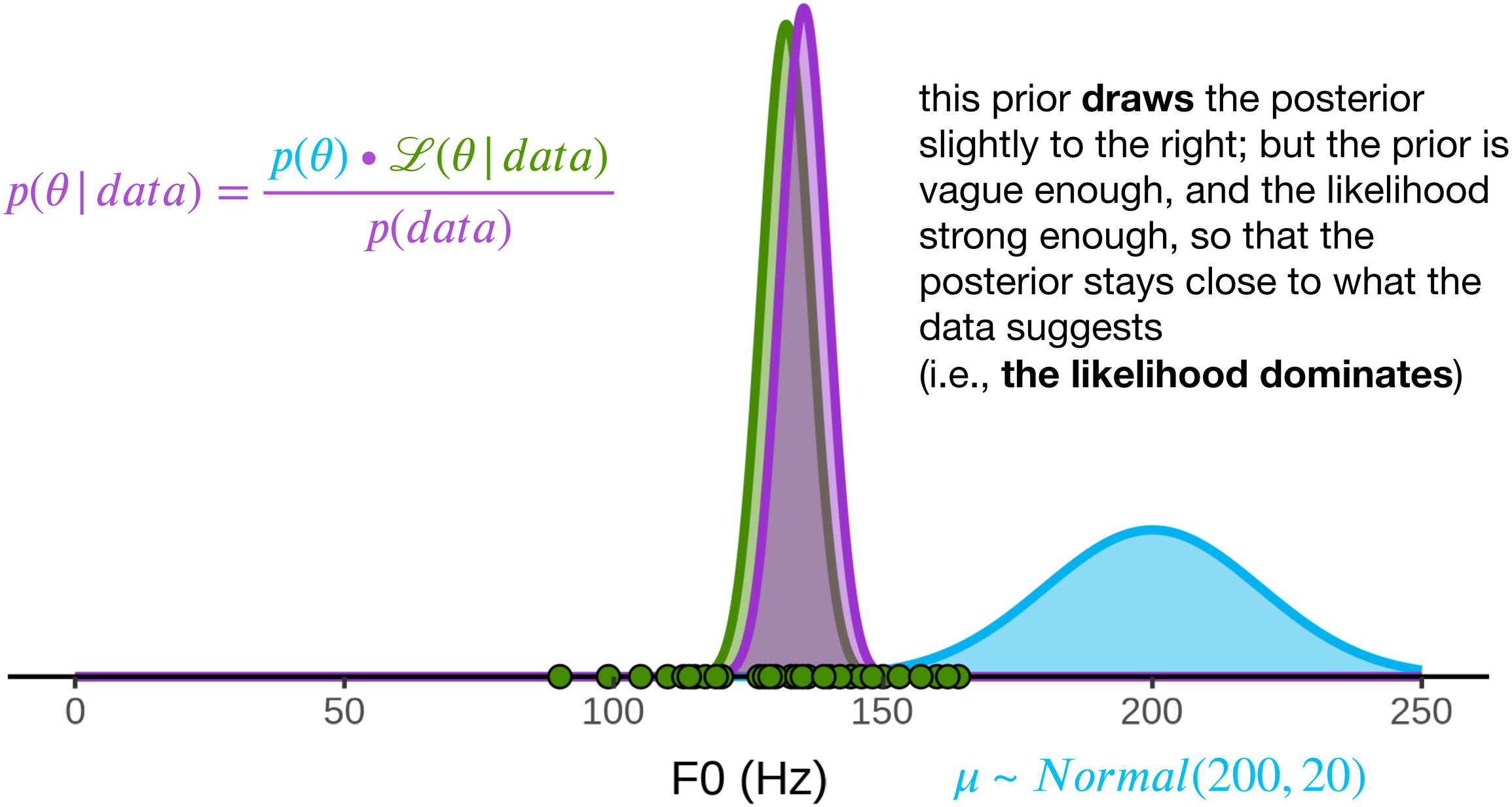
pedagogical example:
somebody didn't get the memo that these F0 values come from male speakers, and set a $Normal(200, 20)$ prior



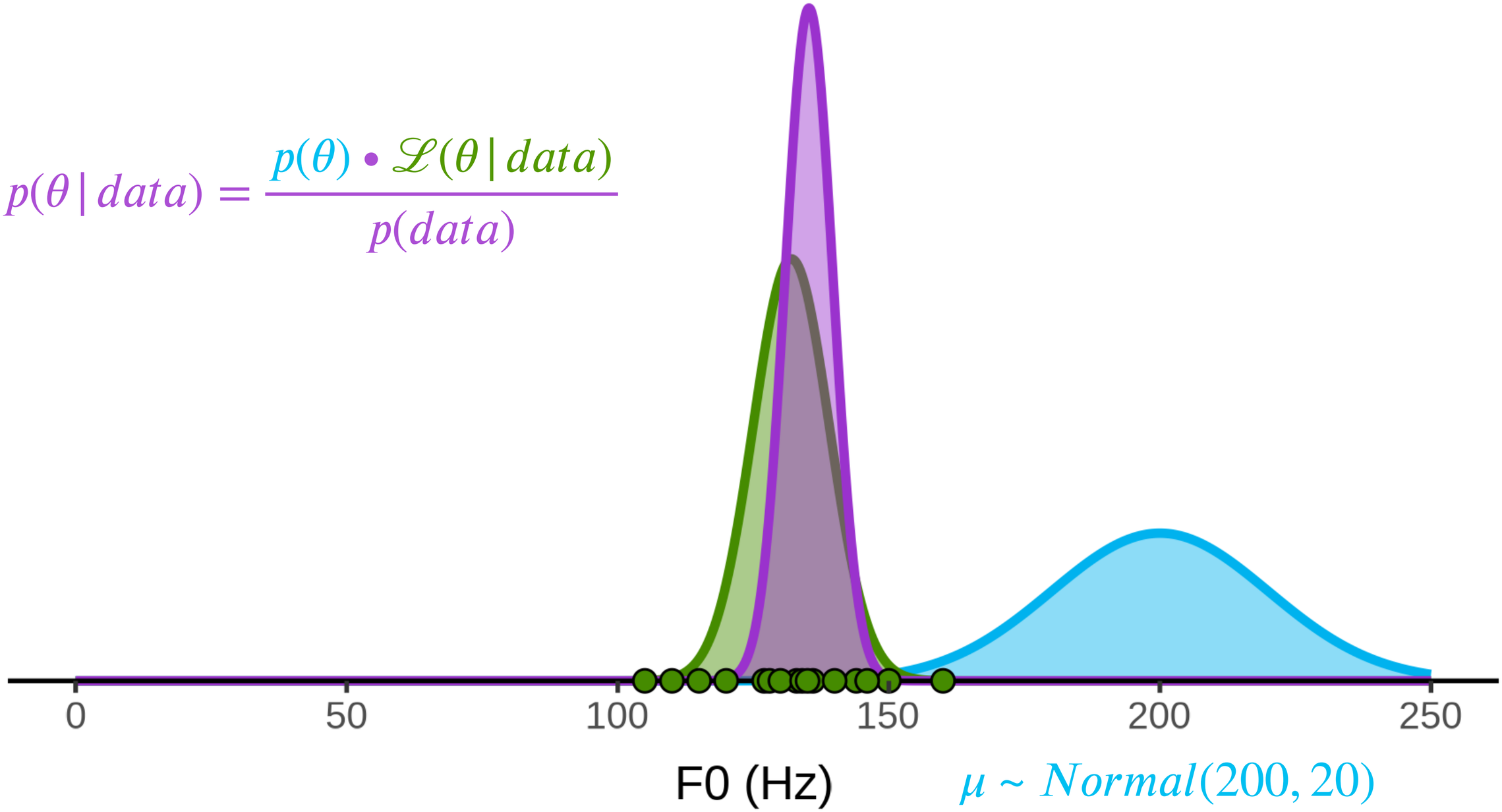
$$\mu \sim Normal(200, 20)$$

$$p(\theta \mid data) = \frac{p(\theta) \bullet \mathscr{L}(\theta \mid data)}{p(data)}$$

this prior **draws** the posterior slightly to the right; but the prior is vague enough, and the likelihood strong enough, so that the posterior stays close to what the data suggests
(i.e., **the likelihood dominates**)



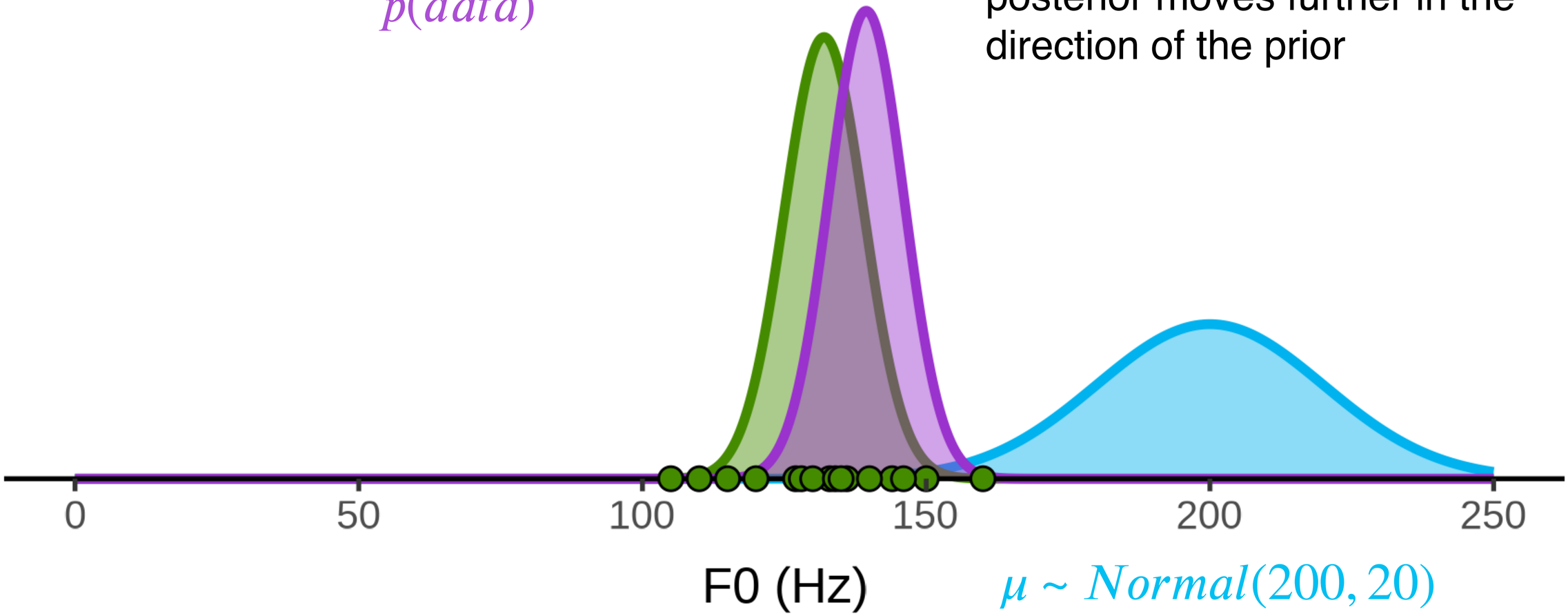
$$p(\theta \mid data) = \frac{p(\theta) \bullet \mathscr{L}(\theta \mid data)}{p(data)}$$



$\mu \sim Normal(200, 20)$

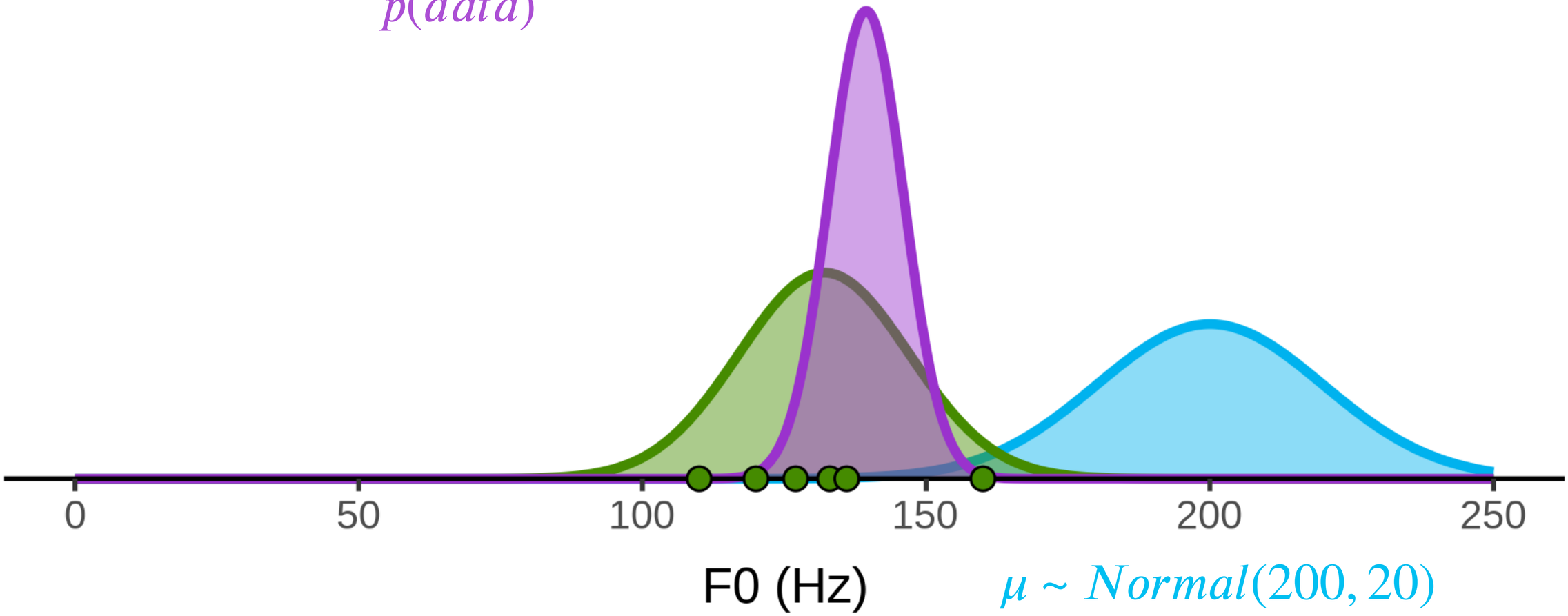
$$p(\theta \mid data) = \frac{p(\theta) \bullet \mathscr{L}(\theta \mid data)}{p(data)}$$

with **less data**, the likelihood becomes weaker, and the posterior moves further in the direction of the prior



$$\mu \sim Normal(200, 20)$$

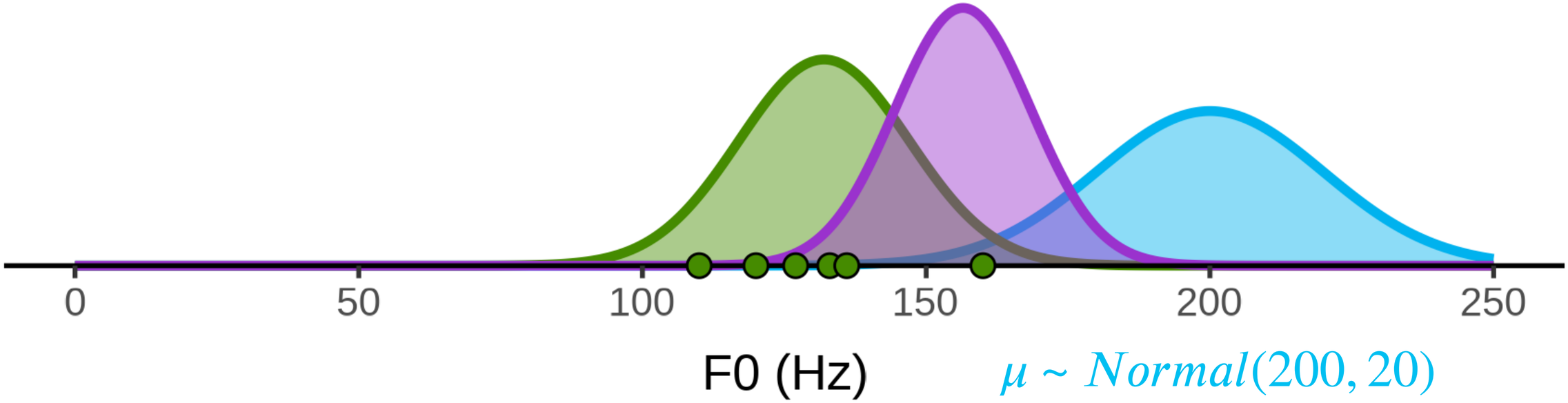
$$p(\theta \mid data) = \frac{p(\theta) \bullet \mathscr{L}(\theta \mid data)}{p(data)}$$



$\mu \sim Normal(200, 20)$

$$p(\theta \mid data) = \frac{p(\theta) \bullet \mathscr{L}(\theta \mid data)}{p(data)}$$

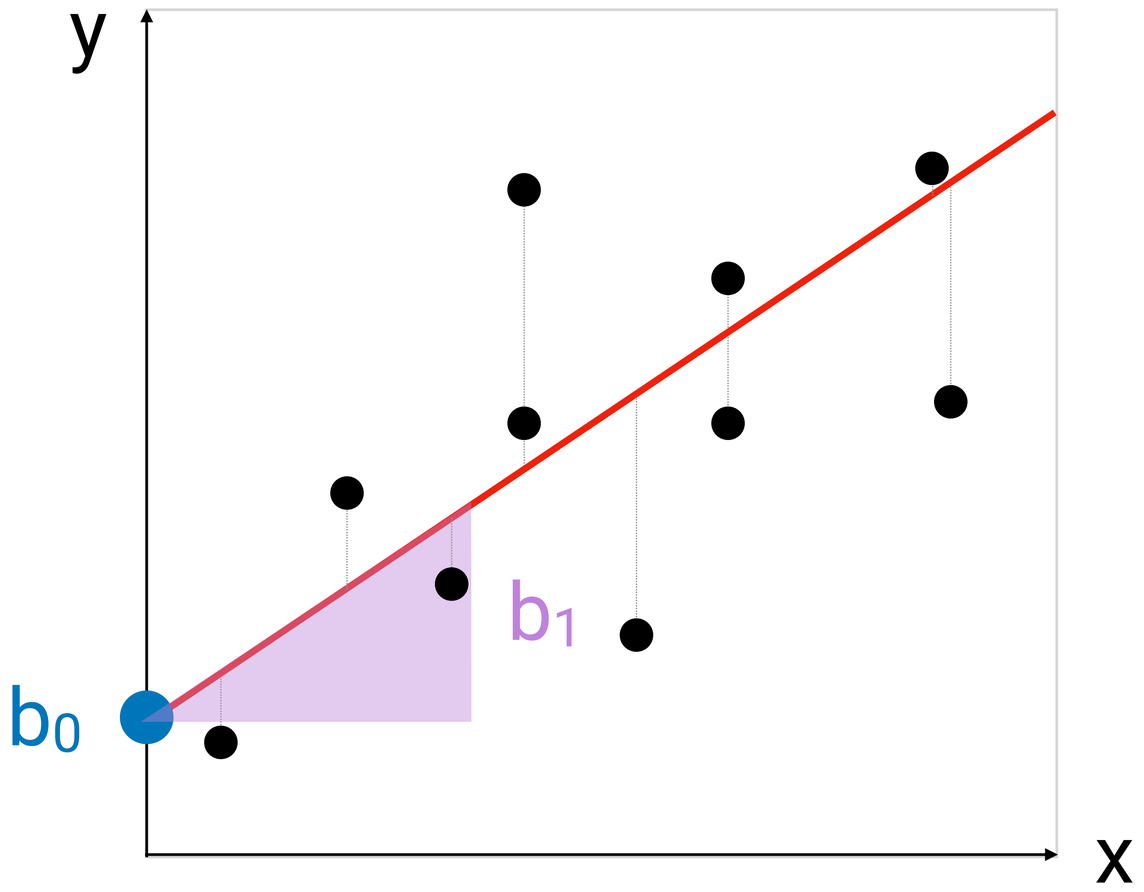
with even weaker likelihood,
the prior exerts an even
stronger pull on the posterior



$\mu \sim Normal(200, 20)$

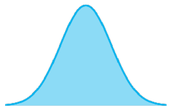
Priors in linear models

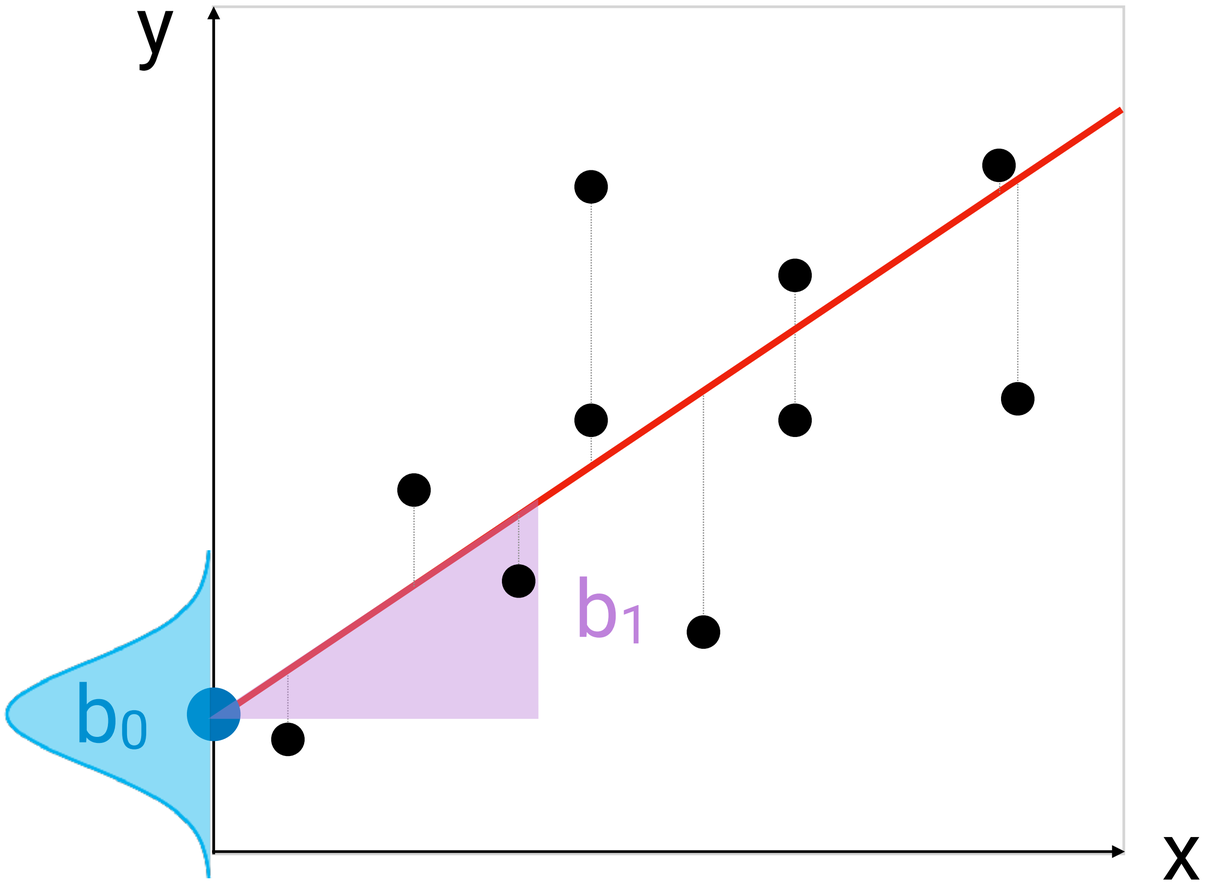
$$y = b_0 + b_1 * x + \sigma$$



Priors in linear models

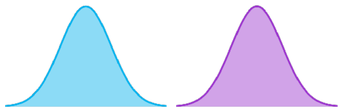
$y = b_0 + b_1 * x + \sigma$

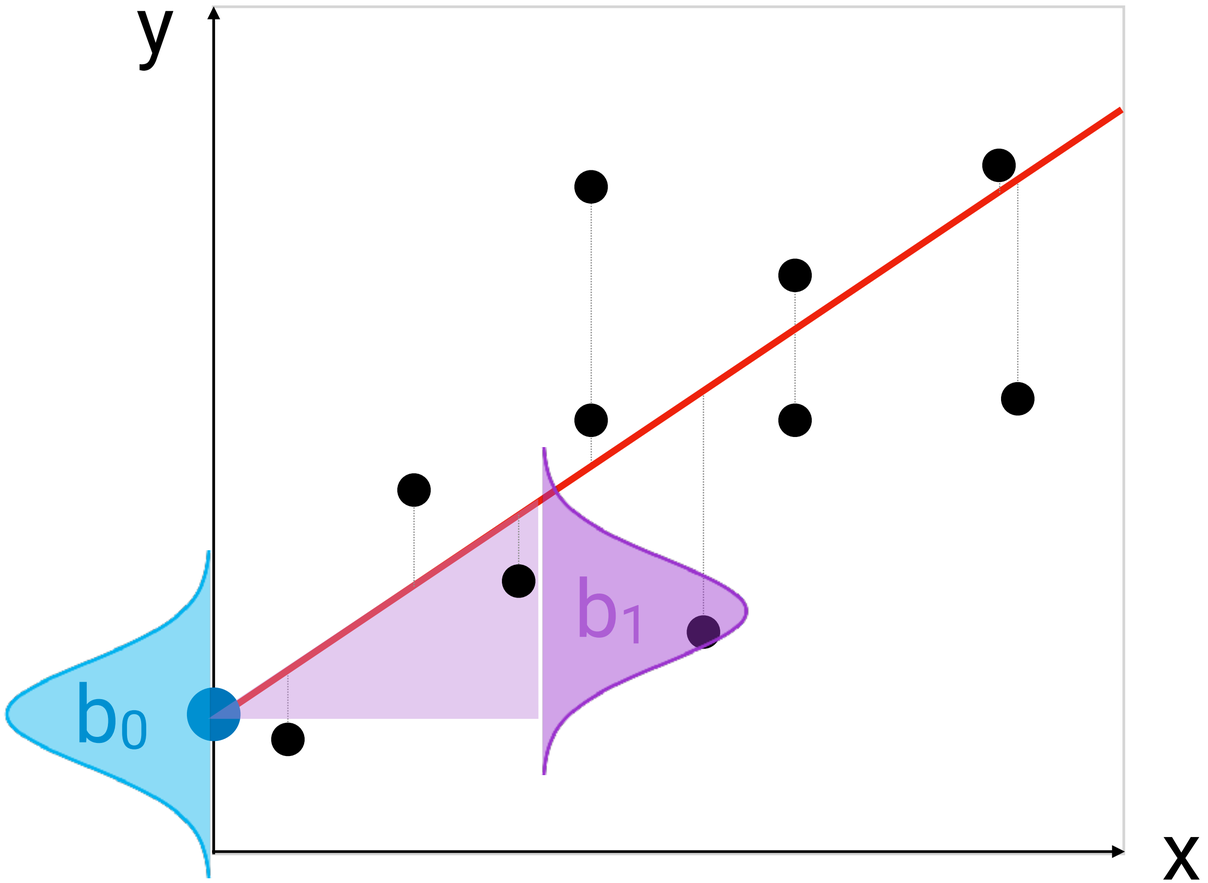




Priors in linear models

$y = b_0 + b_1 * x + \sigma$

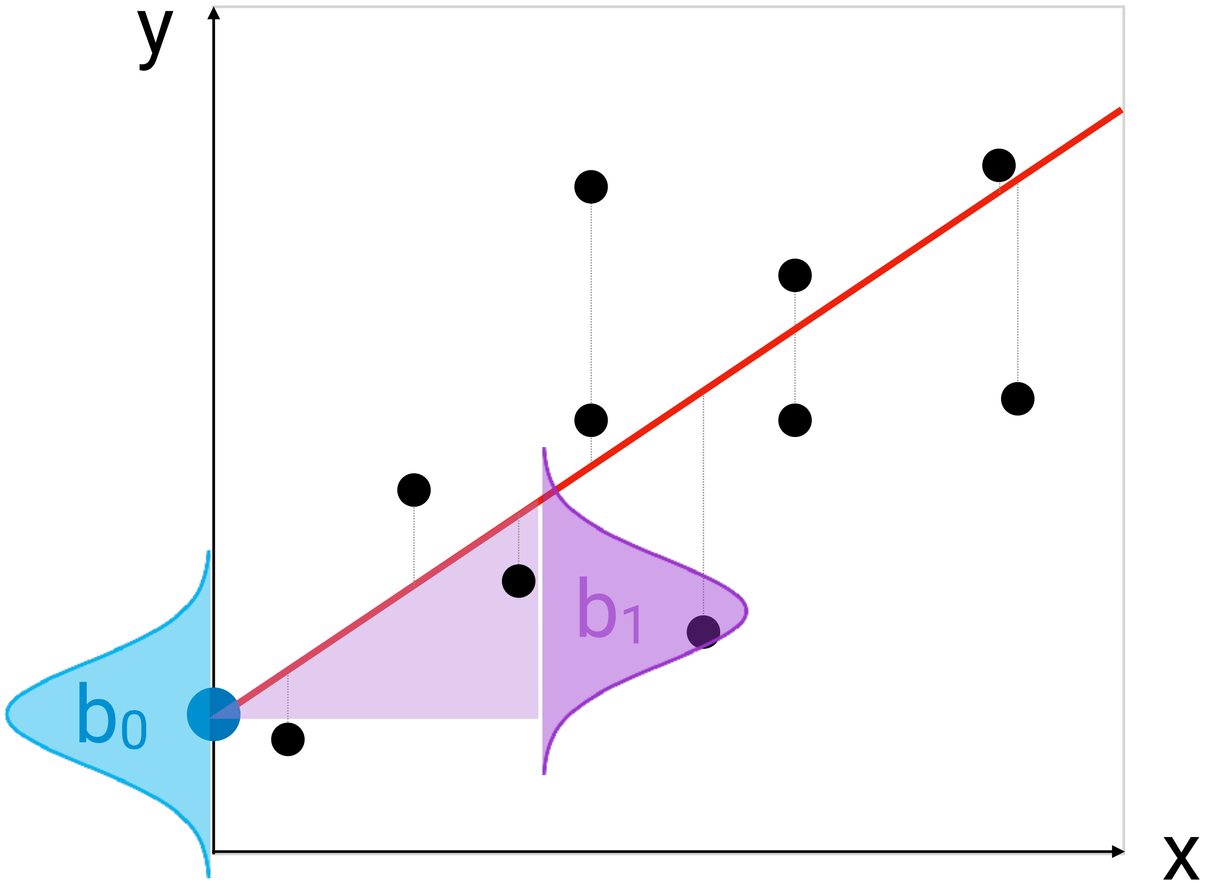




Priors in linear models

$$y = b_0 + b_1 * x + \sigma$$

every parameter is modelled
as a probability distribution
and receives a prior

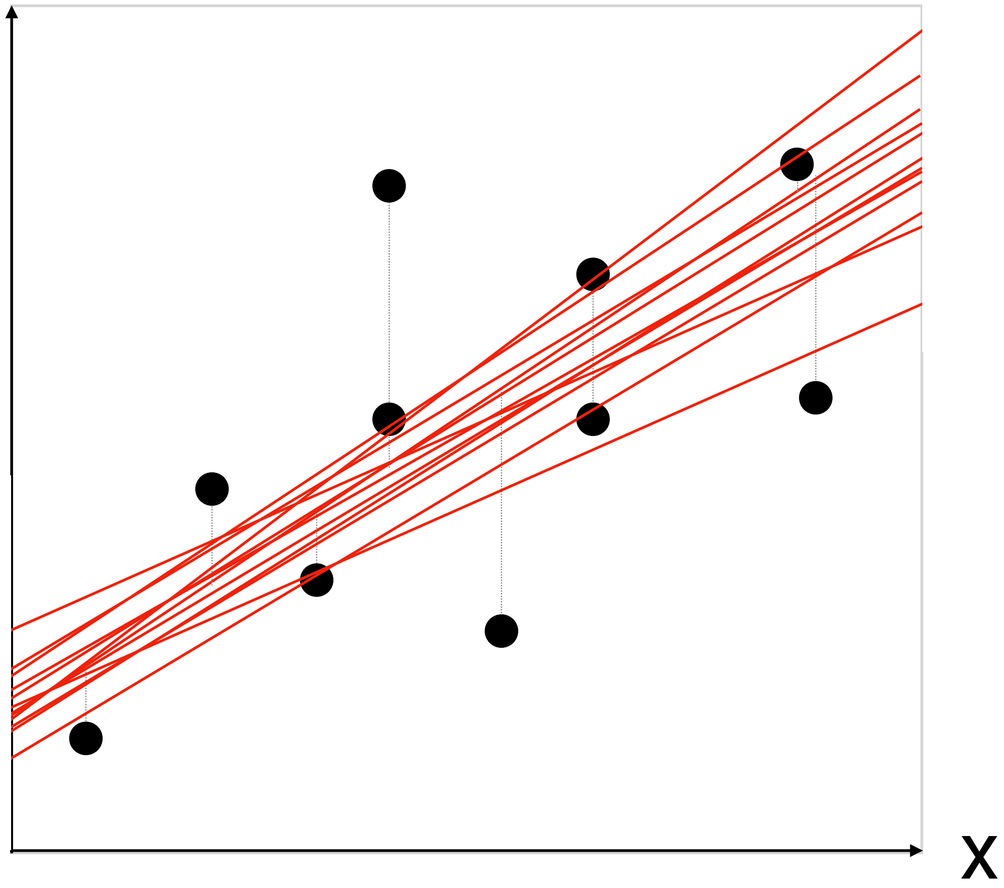


Priors in linear models

$$y = b_0 + b_1 * x + \sigma$$

every parameter is modelled
as a probability distribution
and receives a prior

the model predicts a range
of plausible intercepts and
slopes, i.e. a range of
plausible linear models

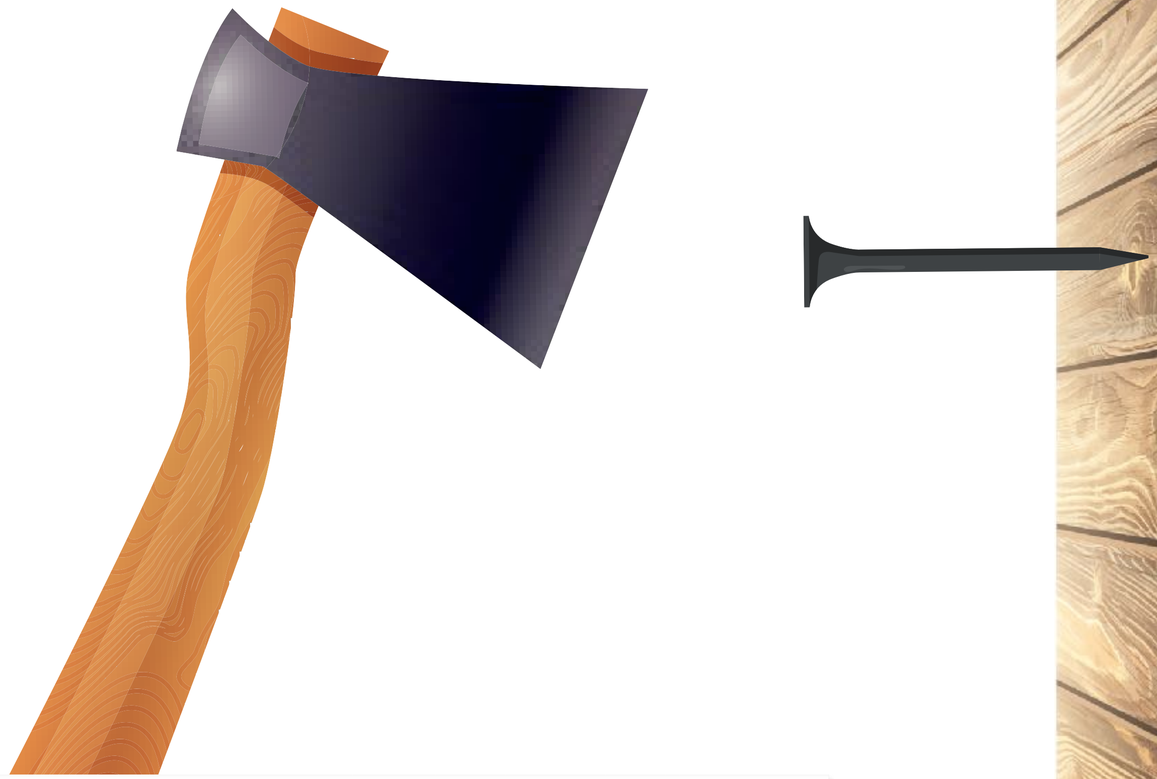



NHST vs Bayes


CRAFTSMAN
NICKEL PLATED

Null hypothesis significance testing

- often does **not** allow **appropriate** use,



```
Warning message:
In checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv,  :
  Model failed to converge with max|grad| = 0.0139723 (tol = 0.002, component 1)
```

```
boundary (singular) fit: see help('isSingular')
```

Null hypothesis significance testing

- often does **not** allow **appropriate** use,
- is **not intuitive**,

89% of books that covered statistical significance defined or explained it incorrectly

Cassidy et al. (2019)

Null hypothesis significance testing

- often does **not** allow **appropriate** use,
- is **not intuitive**,
- and **cannot** provide an **answer** to the **questions** we are interested in.

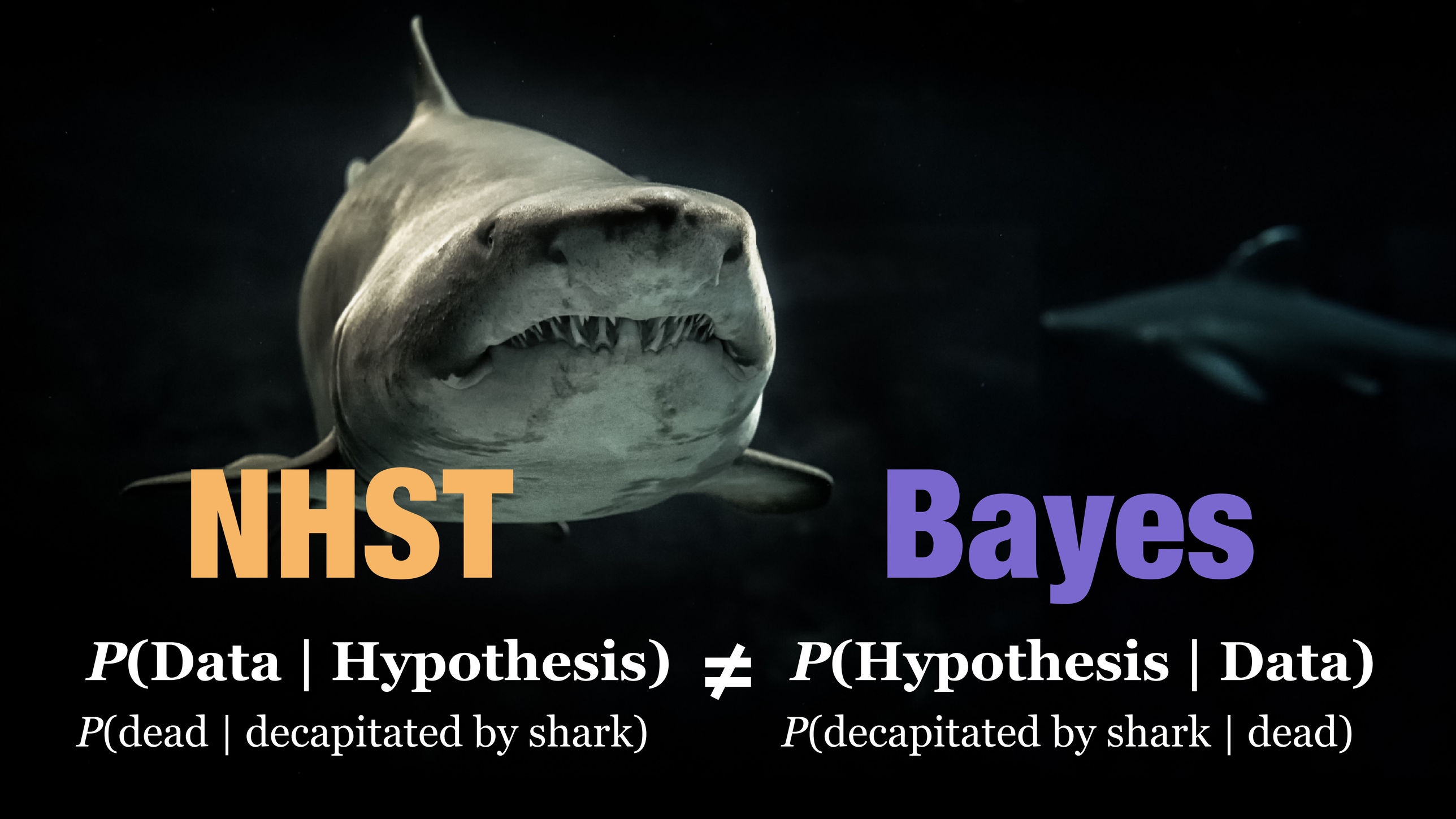

CRAFTSMAN
NICKEL PLATED

Bayesian Inference

- robust inference
- intuitive
- flexible

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of statistical software*, *80*(1), 1-28.

`brms`


NHST
Bayes
P(Data | Hypothesis) ≠ P(Hypothesis | Data)
P(dead | decapitated by shark)
P(decapitated by shark | dead)

NHST	≠	Bayesian
P(Data | Theory)		*P*(Theory | Data)
no prior knowledge		incorporates prior knowledge
quantifies long-run probability of finding a false positive		quantifies uncertainty around possible parameter values
hard cut-off decisions		gradual assessment of evidence
yields an estimate of one true parameter value		yields a distribution of plausible parameter values

NHST	≠	Bayesian
P(Data | Theory)		*P*(Theory | Data)
no prior knowledge		incorporates prior knowledge
quantifies long-run probability of finding a false positive		quantifies uncertainty around possible parameter values
hard cut-off decisions		gradual assessment of evidence
yields an estimate of one true parameter value		yields a distribution of plausible parameter values

Bayesian

- 👍 very **flexible** in terms of model architecture
- 👍 not limited by optimization constraints (no “**convergence failures**”)
- 👍 not limited to categorical decision procedure

- 👎 computationally expensive
- 👎 one more layer of researcher degrees of freedom
- 👎 **more “thinking”** required

flexibility 👍

one and the same framework for everything you need

types of **error distributions**

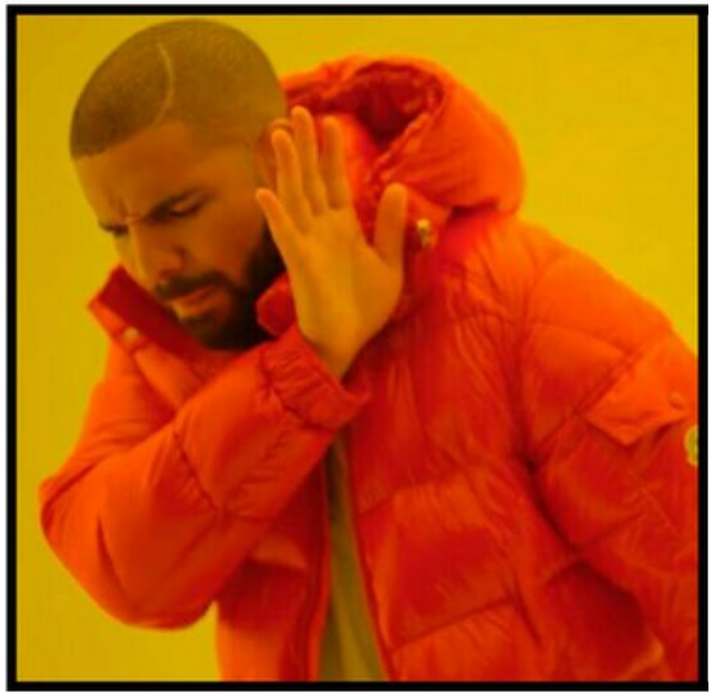
gaussian, binomial, ordinal, multinomial, etc.

types of **fitting procedures**

univariate, multivariate, mixture, etc.

levels of **covariance**

simple regression, multiple regression, mixed-effect regression, etc.



convergence issues?

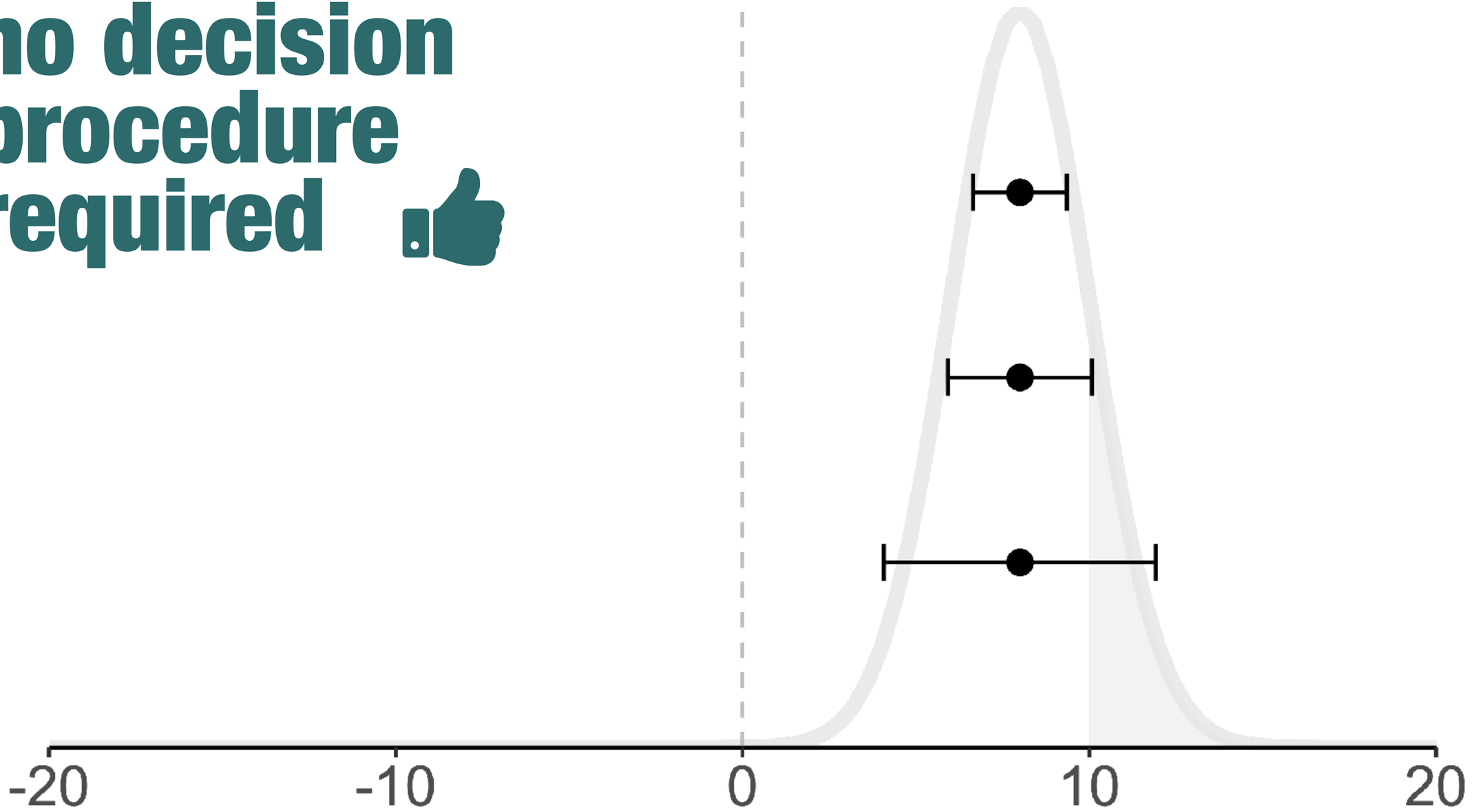
lmer()

```
Warning message:
In checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv,  :
  Model failed to converge with max|grad| = 0.0139723 (tol = 0.002, component 1)
```

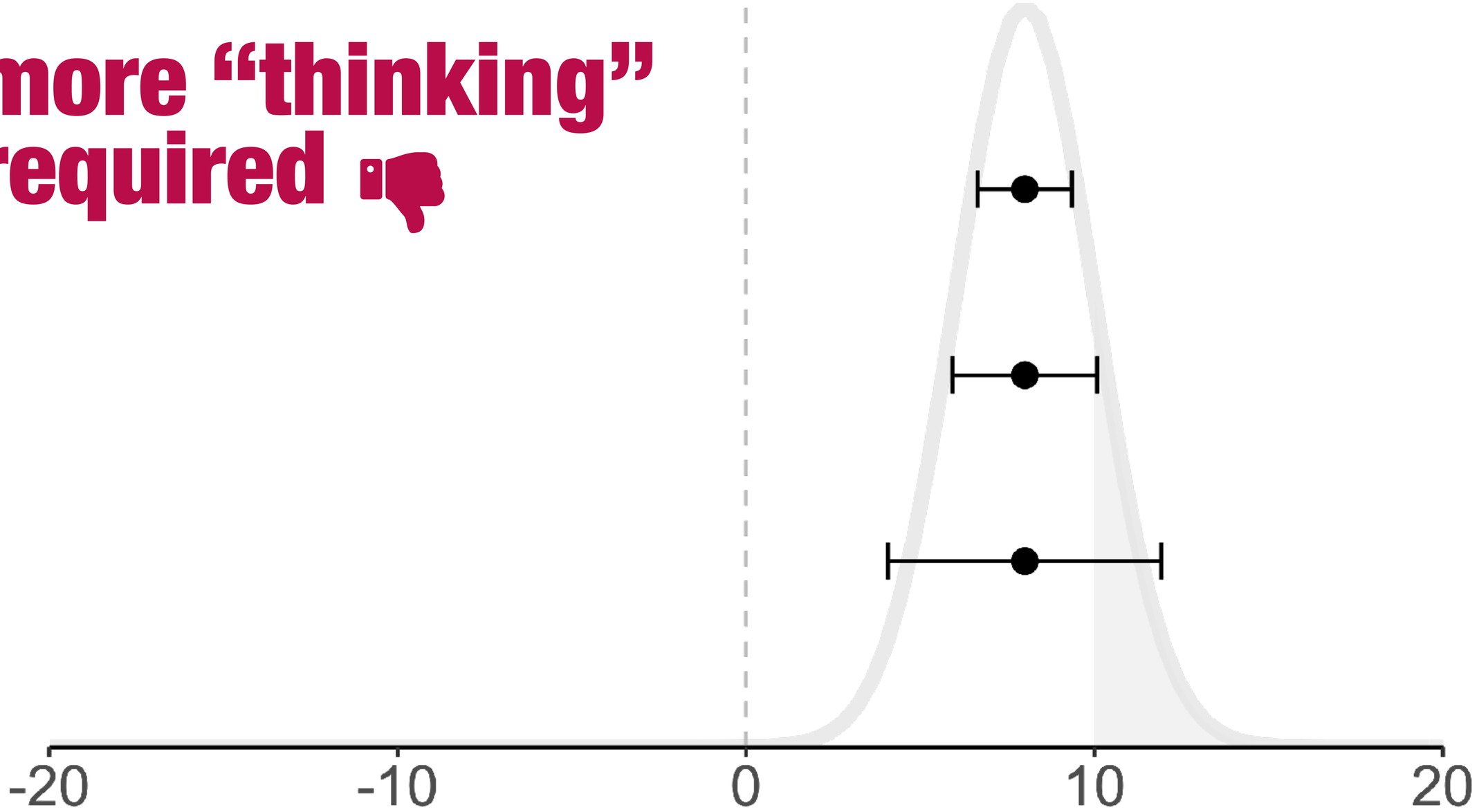

DO I BRING MY UMBRELLA?


ARE MY RESULTS SIGNIFICANT?

no decision procedure required



more “thinking” required 👎



computationally expensive

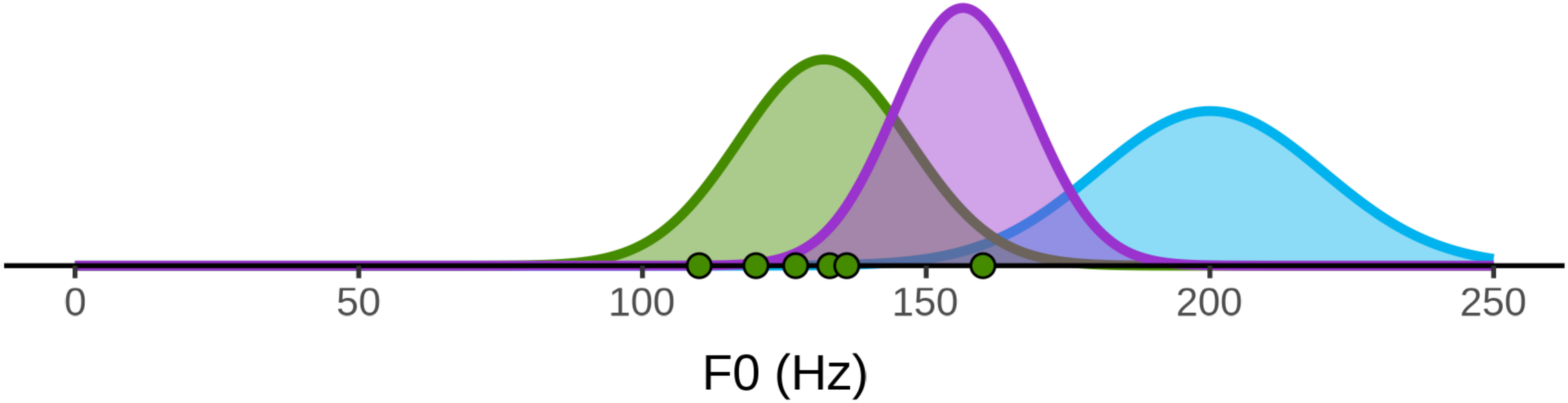
$$Pr(\text{Theory} \mid \text{Data}) = \frac{Pr(\text{Data} \mid \text{Theory}) \times Pr(\text{Theory})}{Pr(\text{Data})}$$

$$Pr(\text{Data}) = \int Pr(\text{Data}, \text{Theory})\, d_{\text{Theory}}$$

can be **intractable** to solve, but...

can be **approximated** with clever algorithms

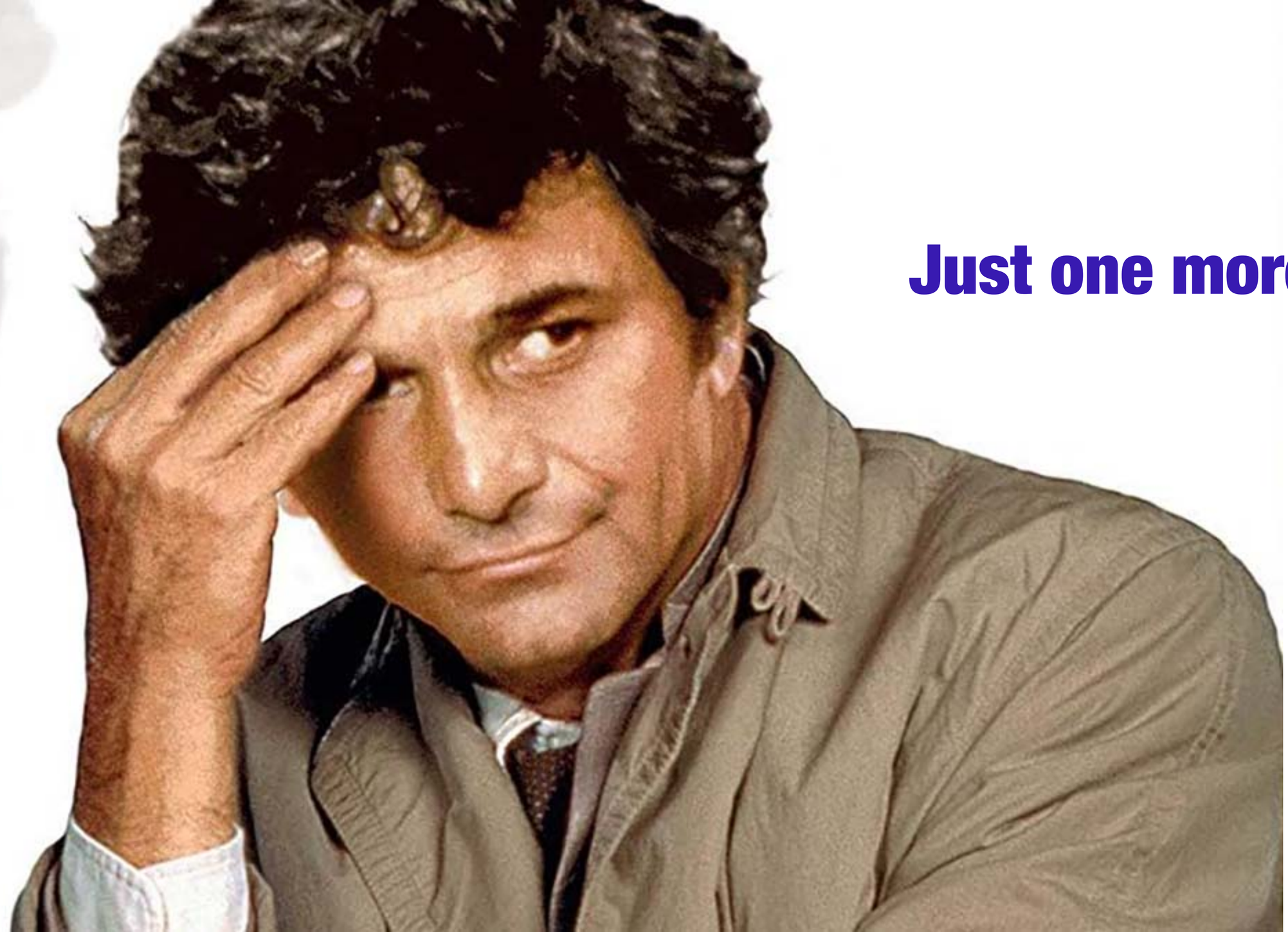
more degrees of freedom? 👎



Bayesian

- 👍 very **flexible** in terms of model architecture
- 👍 not limited by optimization constraints (no “**convergence failures**”)
- 👍 not limited to categorical decision procedure

- 👎 computationally expensive
- 👎 one more layer of researcher degrees of freedom
- 👎 **more “thinking”** required



Just one more thing…

The technical implementation is easy

```
xlm <- lm(F0 ~ 1, data = pitch)
summary(xlm)
```

```
xblm <- brm(F0 ~ 1, data = pitch)
summary(xblm)
```

```
Regression Coefficients:
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept   124.12      4.15   115.73   132.21 1.00     2713     2508

Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma    23.32      2.99    18.37    29.96 1.00     2722     2271
```