What's **prediction** got to do with **adaptation?**

Dave Kleinschmidt // Princeton Neuroscience Institute

Kavli Summer Institute // 26 June 2017

Adaptation is ubiquitous

- In the **brain**
  - Less BOLD signal from repeated stimuli
  - Lower single neuron firing from repeated stimuli
  - Changes in preferred stimuli based on stimulus ensemble
  - Smaller ERPs from repeated stimuli
- In **behavior**

What does it Mean??

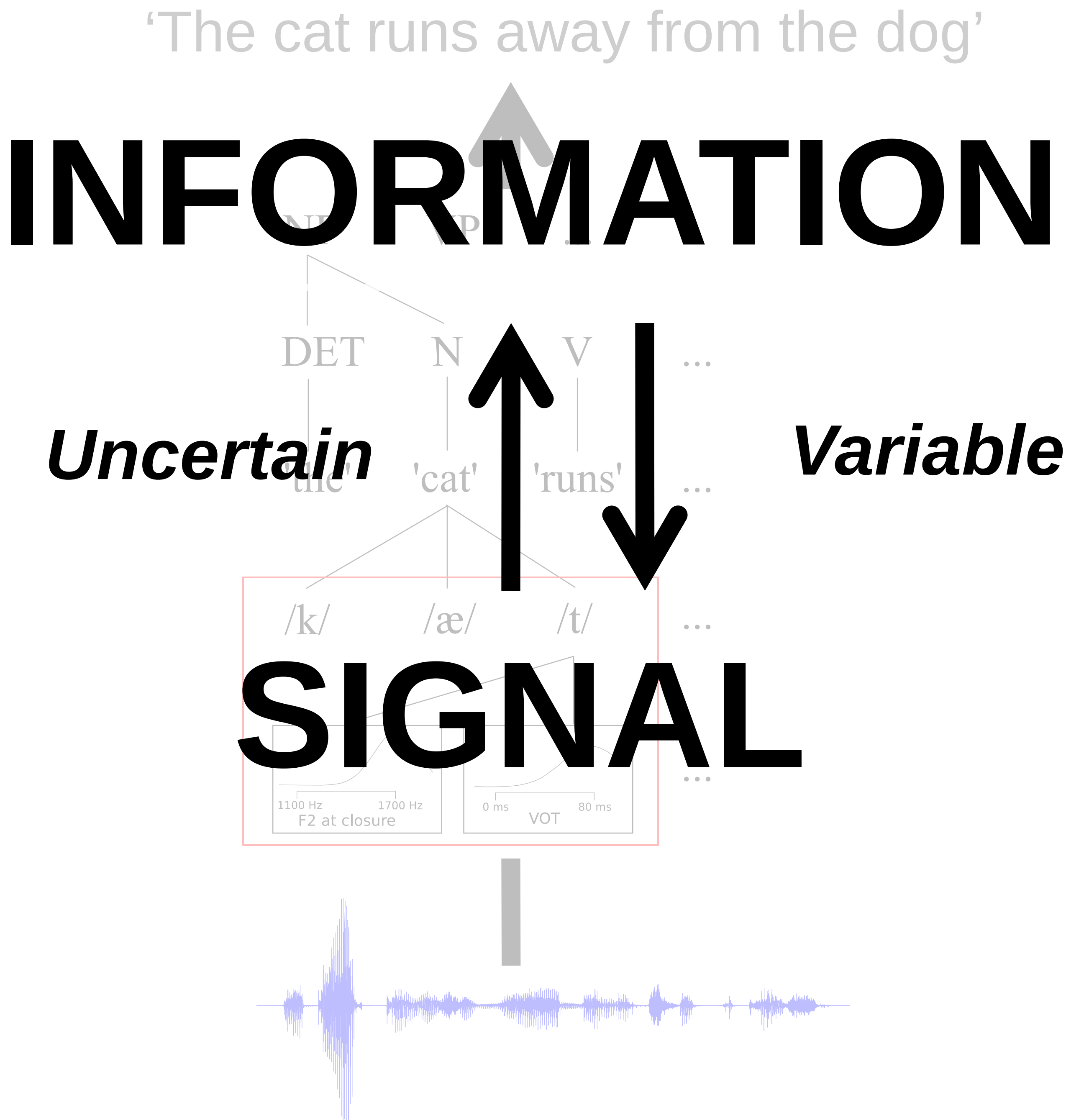
Computational level

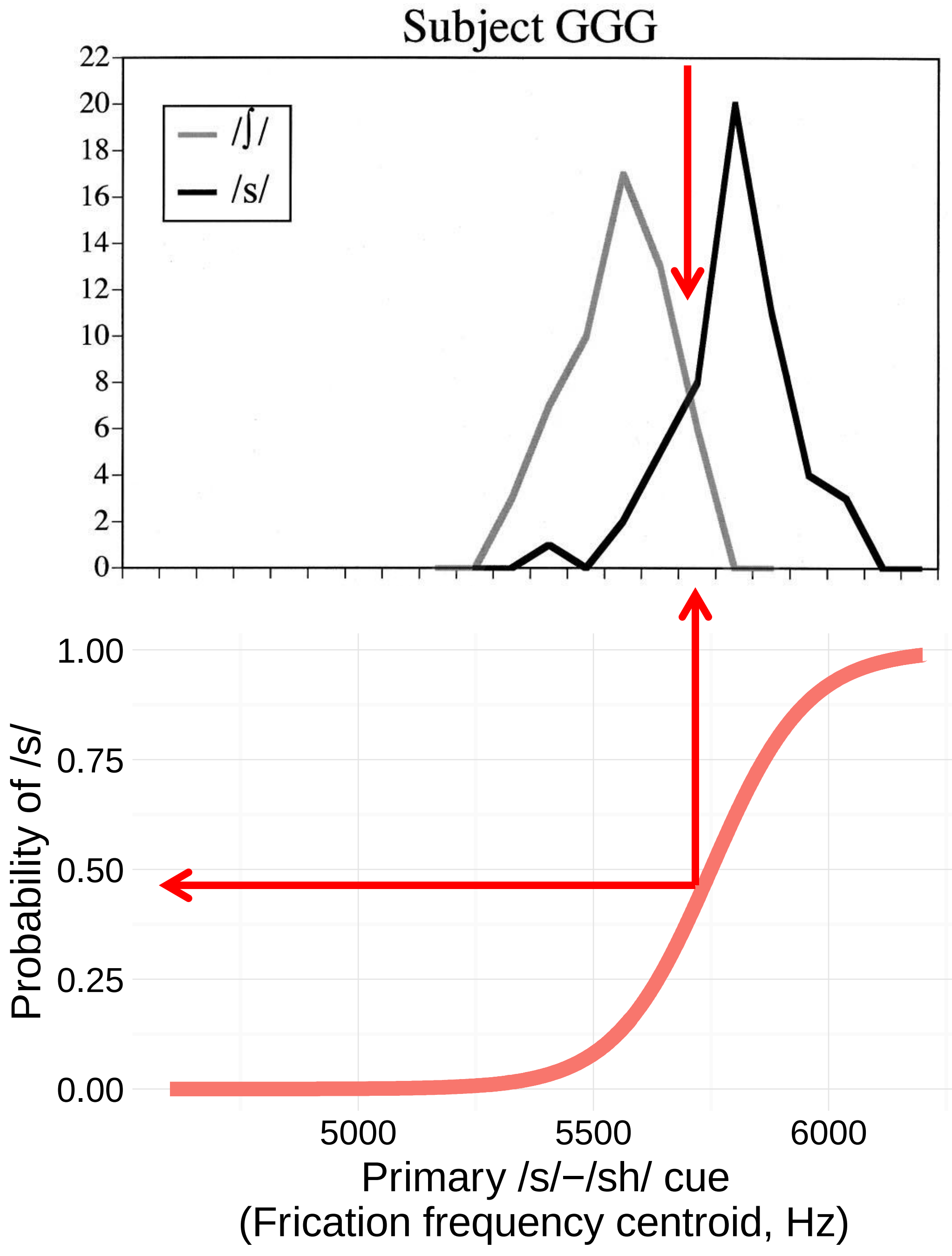
Why should the brain adapt?

What is adaptation?

Goals of the organism

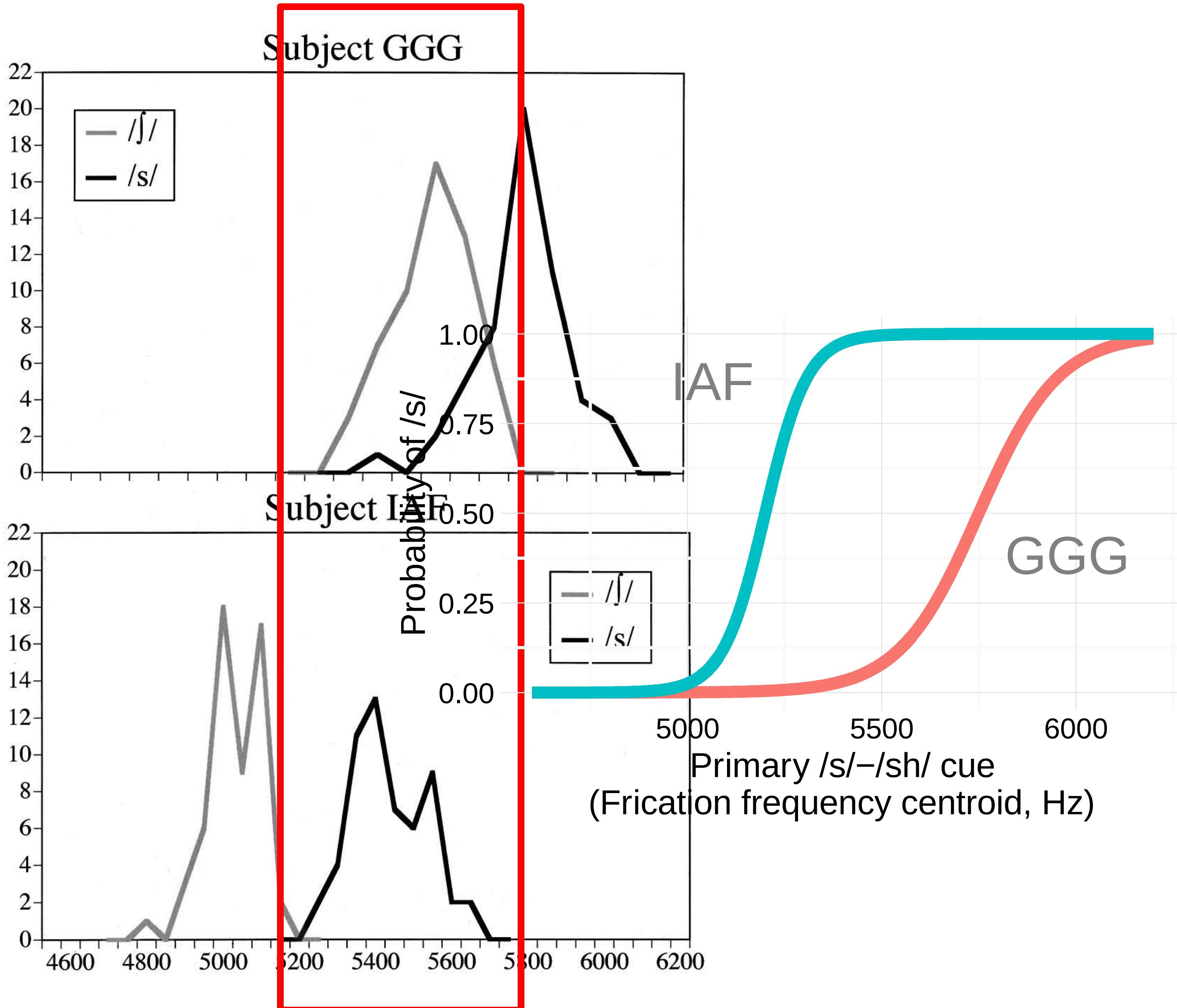
Constraints from the brain (meat computer)


‘The cat runs away from the dog’
INFORMATION
NP
VP
DET
N
V
...
Uncertain
Variable
'the'
'cat'
'runs'
...
/k/
/æ/
/t/
...
SIGNAL
1100 Hz
1700 Hz
F2 at closure
0 ms
80 ms
VOT
...



Likelihood:
p(cue | category)

Posterior:
p(category | cue)



[Newman, Clouse, and Burnham, 2001]

Coping with **variation**

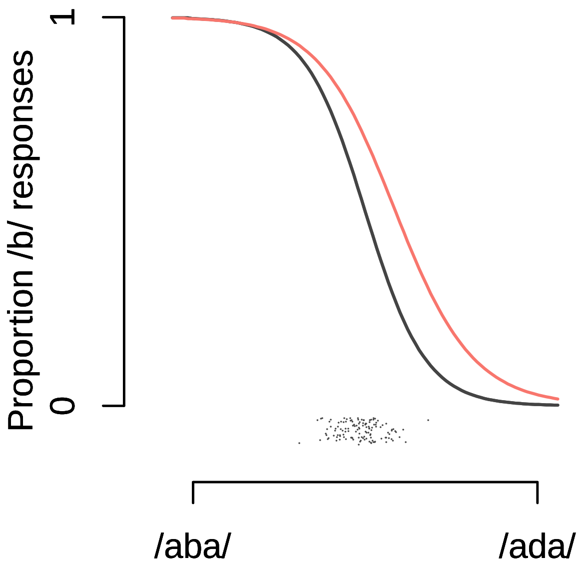
- Listeners **rapidly adapt** to accented speech
- With just a **few minutes** of exposure, comprehension gets much better
  [Clarke and Garrett, 2004]
- **How** do people adjust category boundaries so fast?

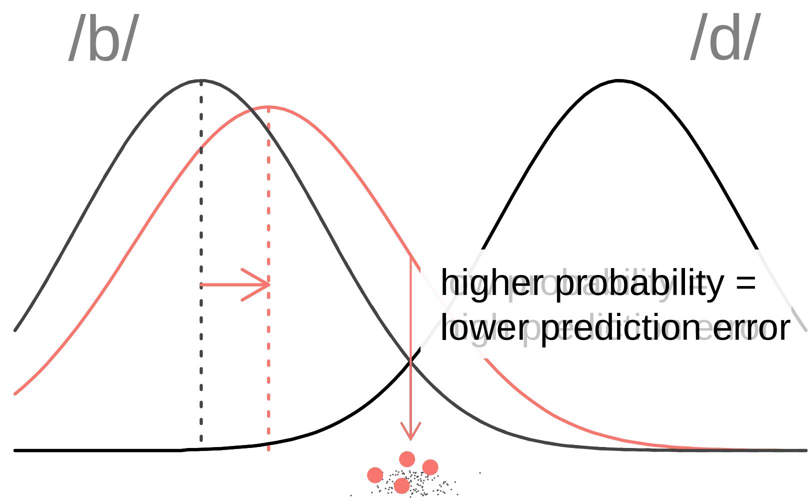
Distributional learning!!

Distributional learning

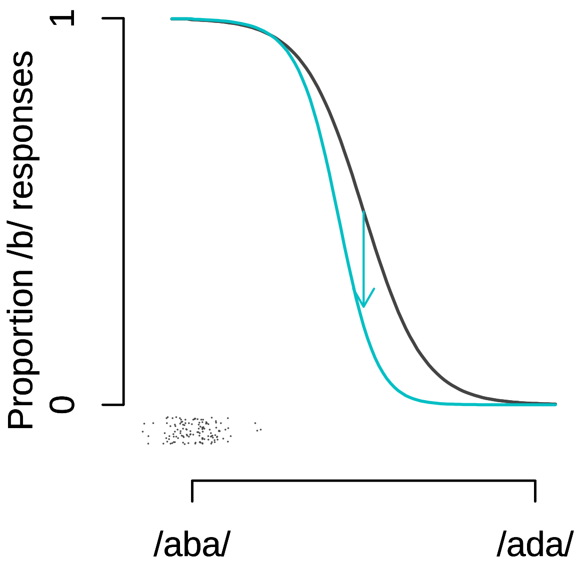
- Try to **reduce prediction error** for the data you've seen
- Makes your predictions better in the future
- ...and inferences more accurate

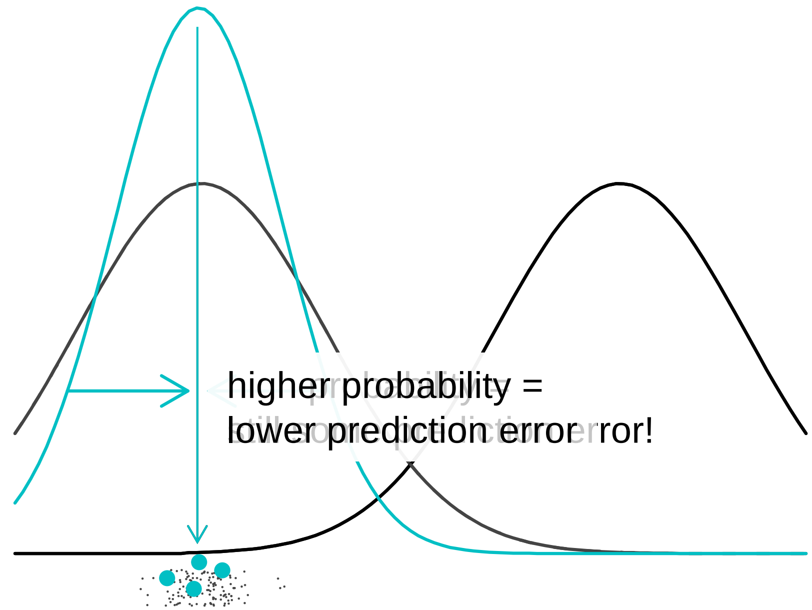
distributional learning for **ambiguous /b/**





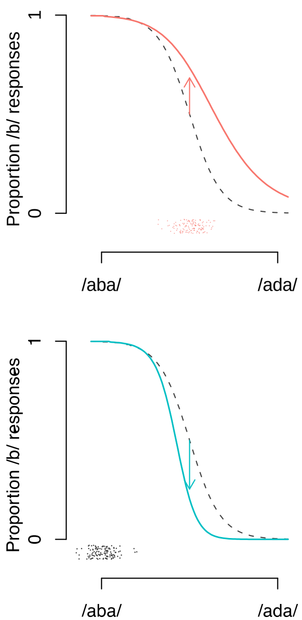
distributional learning for **prototypical** /b/

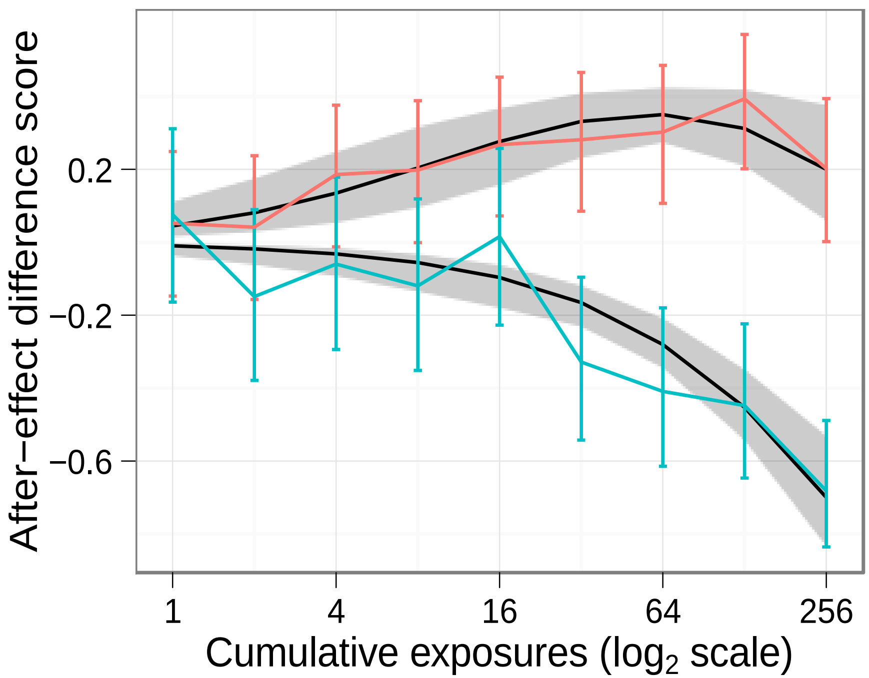


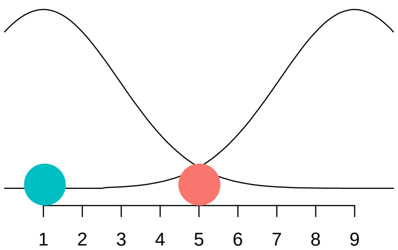


Model results

[Data: Vroomen et al. 2007; Model: Kleinschmidt & Jaeger 2015]

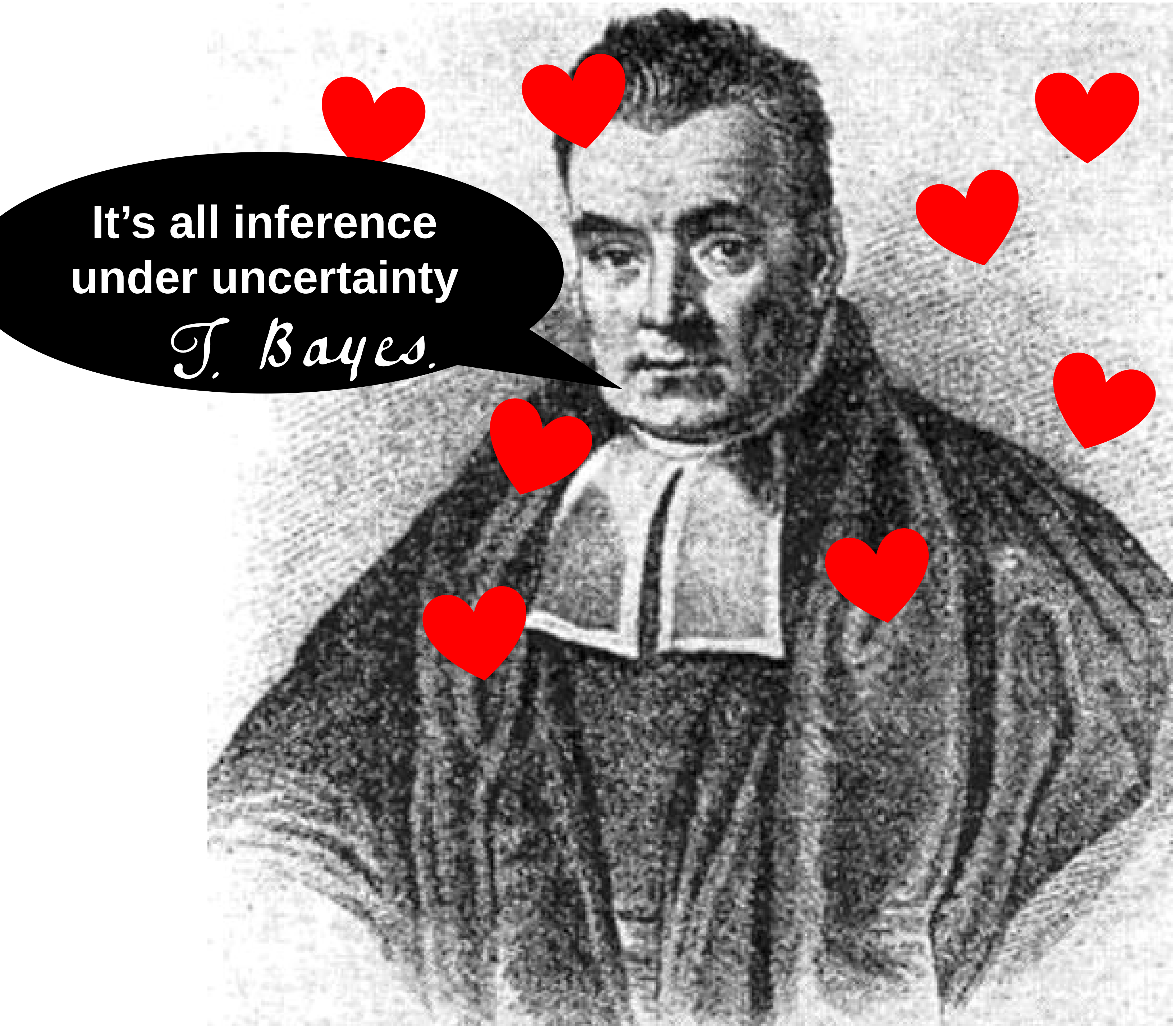






$r^2 = 0.93$

But what **is** distributional learning?


It’s all inference under uncertainty
T. Bayes.

But what **is** distributional learning?

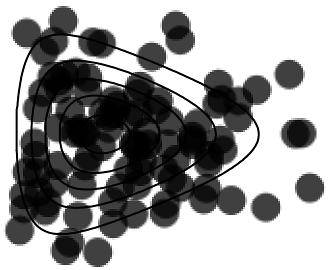
Infer categories and **generative model** (e.g. category means/variances)

$$\underbrace{p(\mu, \sigma^{\mathbf{2}}, c|x)}_{\text{posterior (updated) beliefs}} \propto \underbrace{p(x|c, \mu, \sigma^{\mathbf{2}})}_{\text{likelihood}} \underbrace{p(\mu, \sigma^{\mathbf{2}}) p(c)}_{\text{prior beliefs}}$$

Hypotheses

$p(\mu, \sigma^2)$

μ



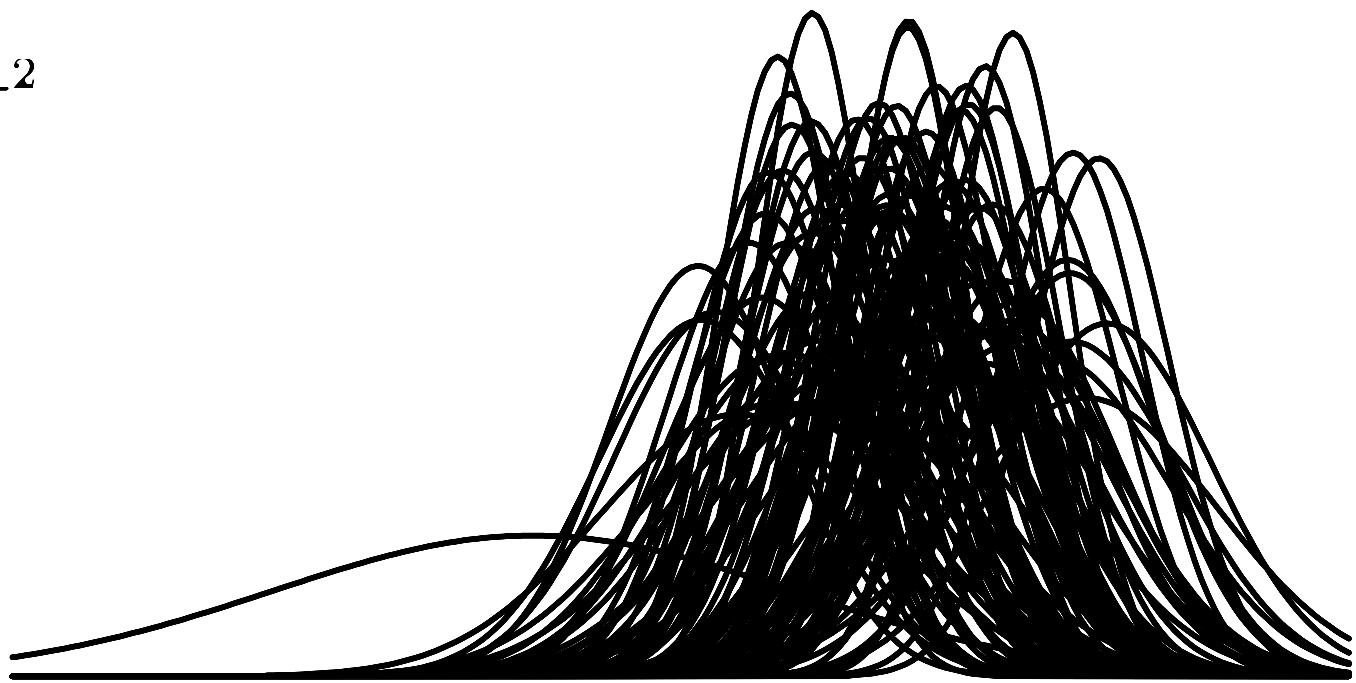
$1/\sigma^2$

(prior) probability

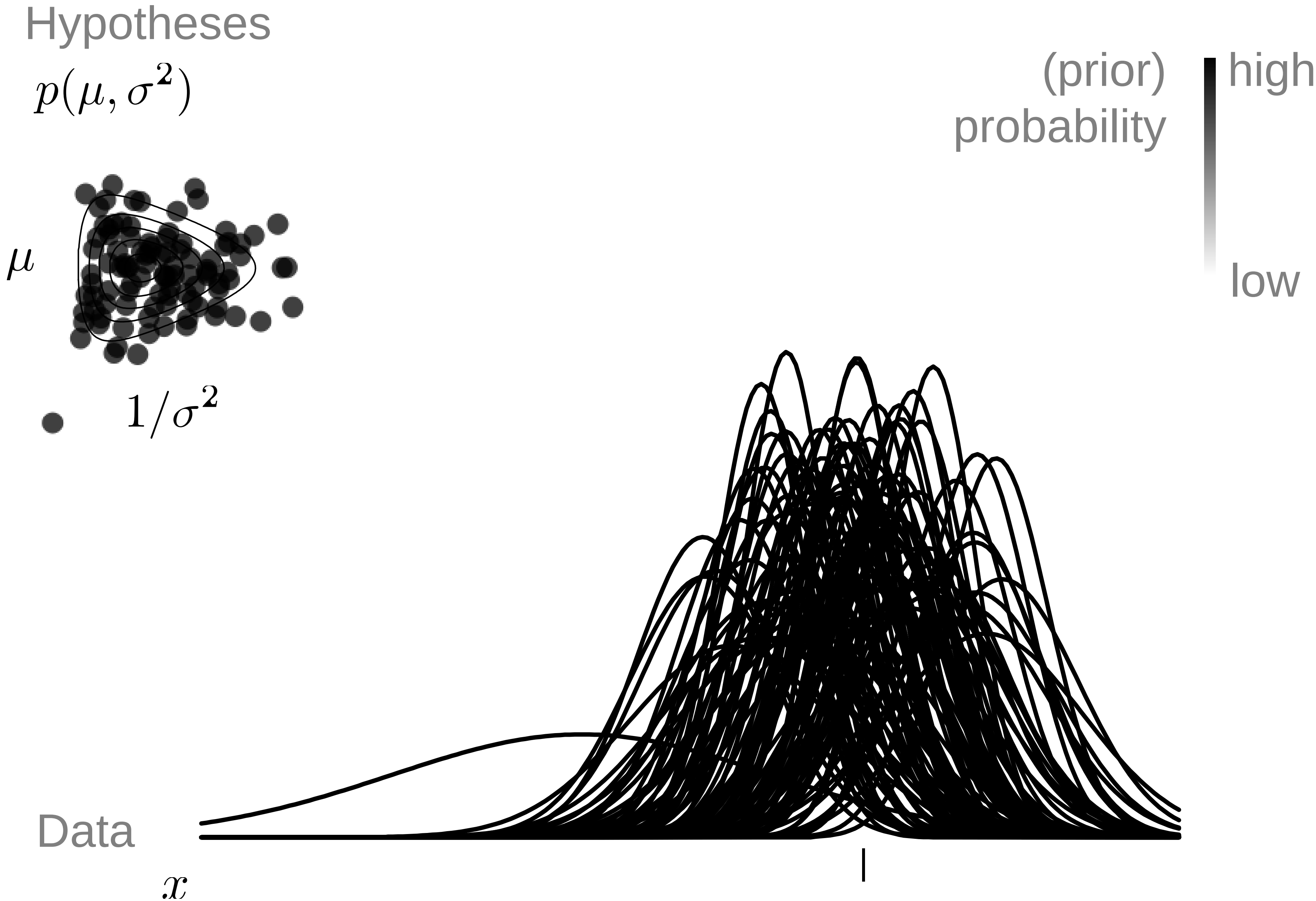
high

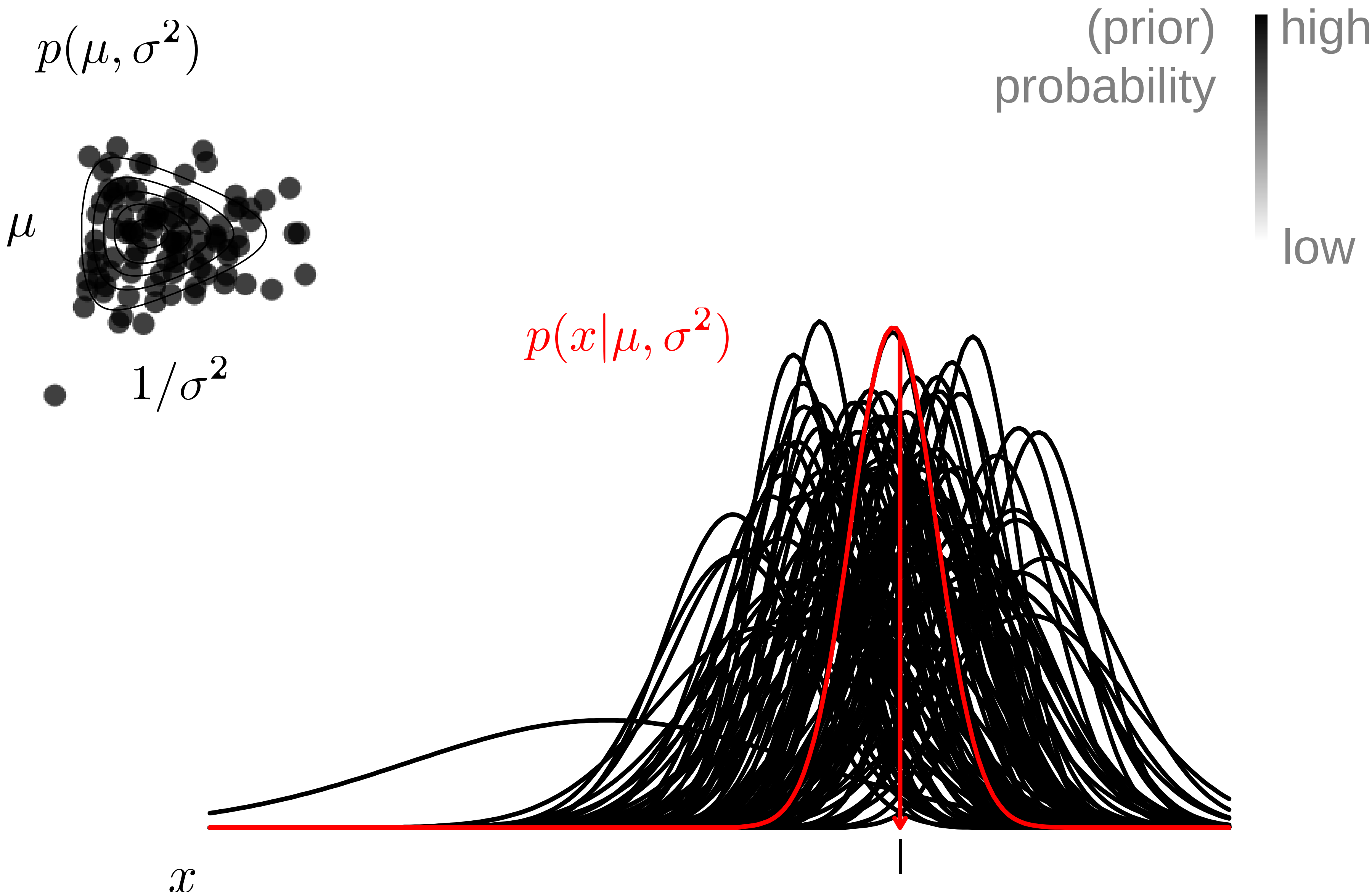
low

Data

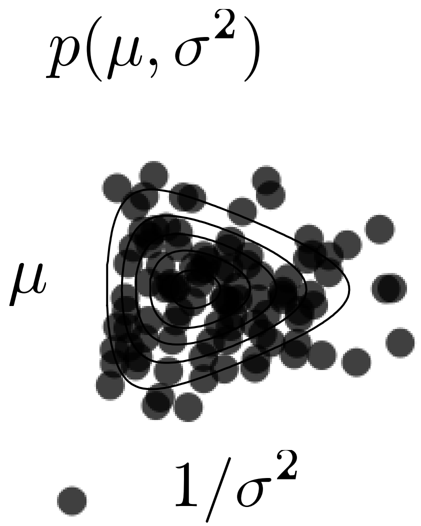


x

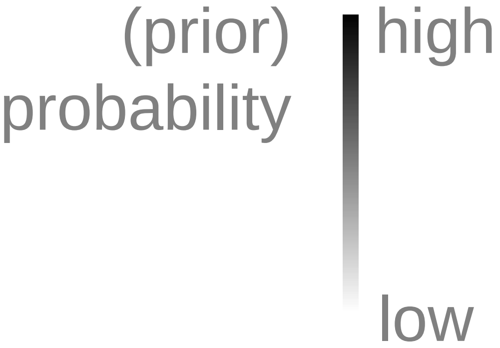

Hypotheses
$p(\mu, \sigma^2)$
(prior) probability
high
low
μ
$1/\sigma^2$
Data
x


$p(\mu, \sigma^2)$
μ
$1/\sigma^2$
(prior)
probability
high
low
$p(x|\mu, \sigma^2)$
x

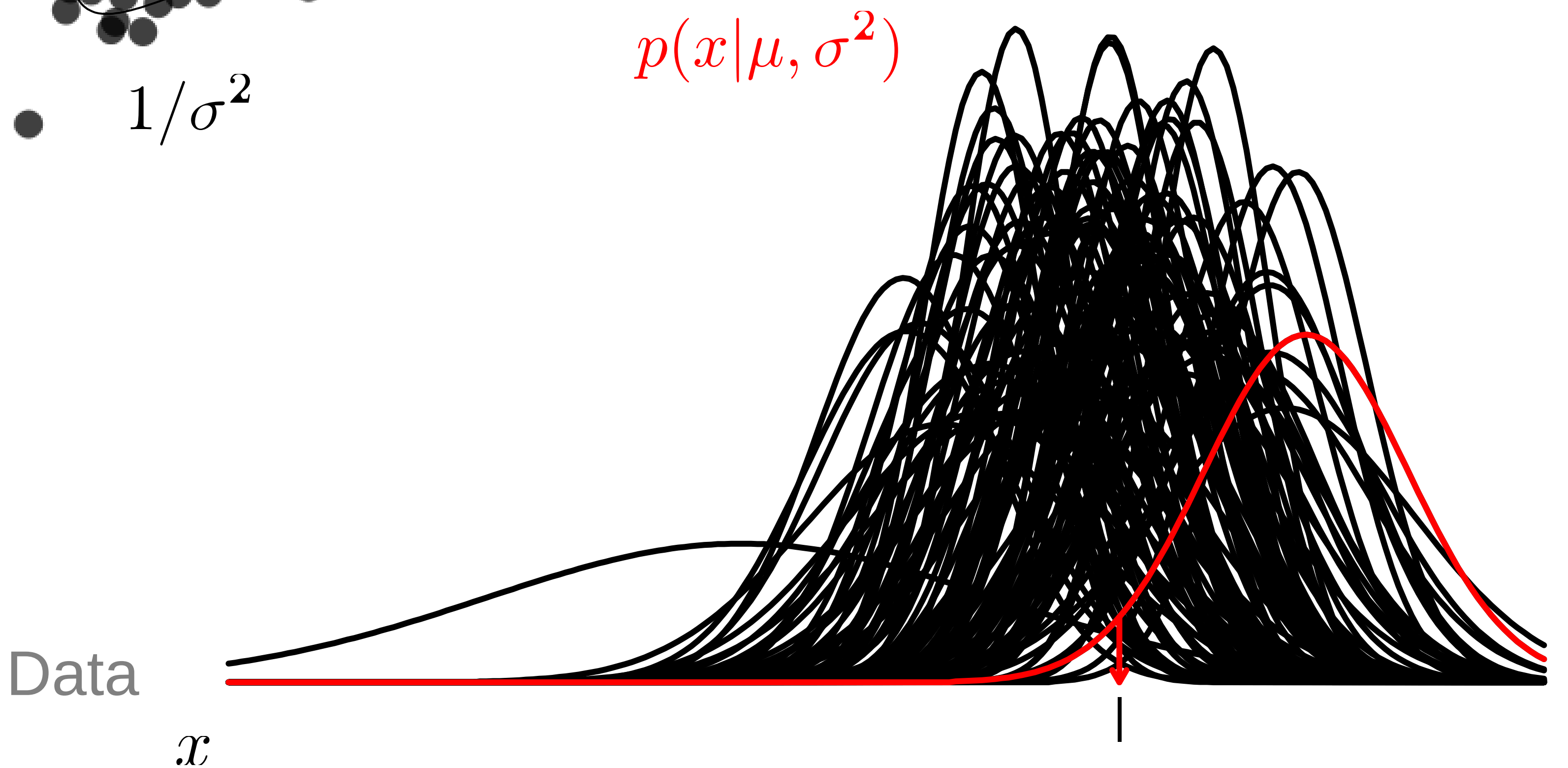
Hypotheses



$p(\mu, \sigma^2)$
μ
$1/\sigma^2$



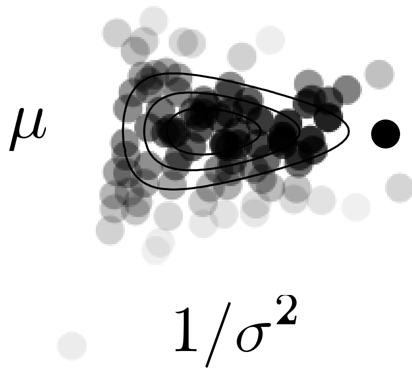
(prior) probability
high
low

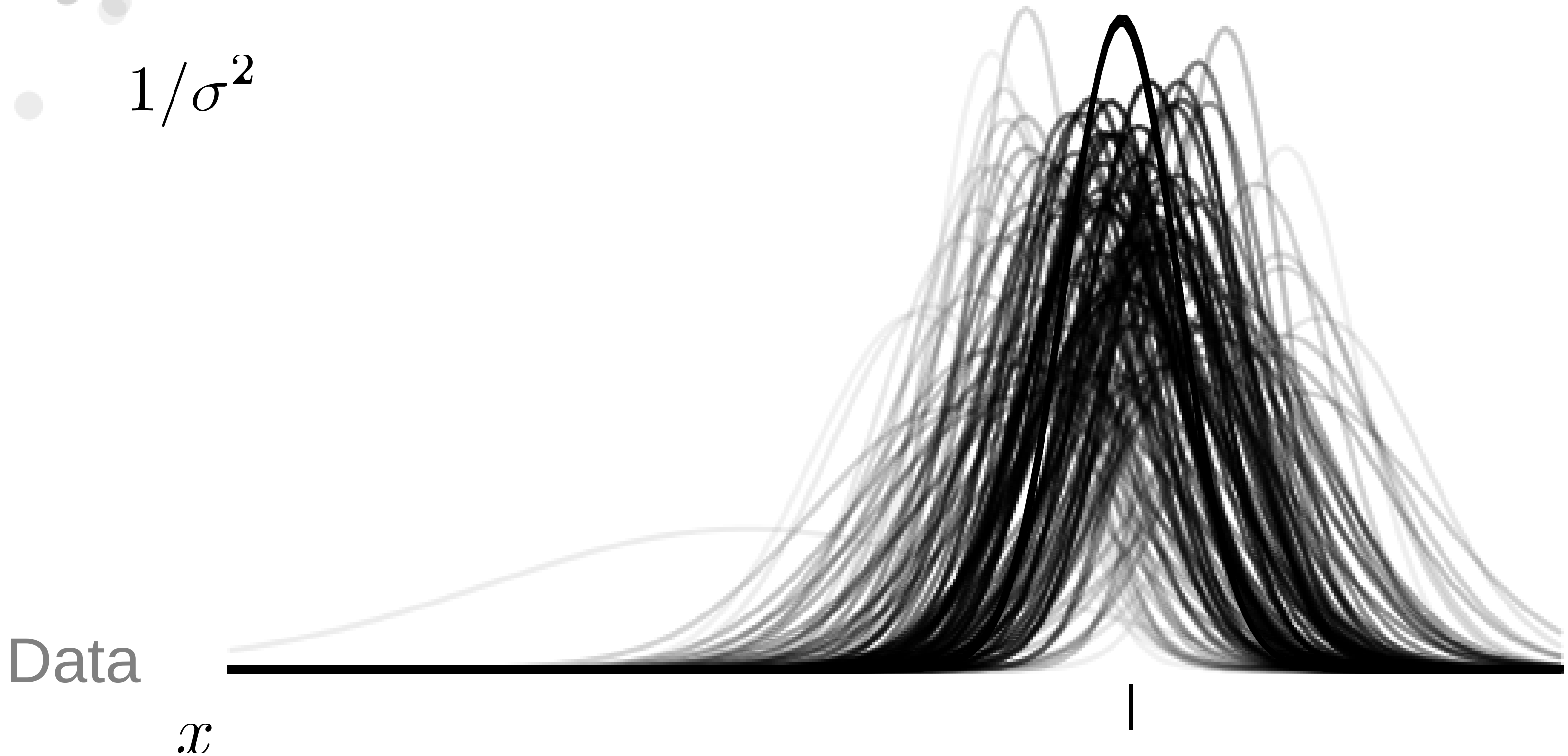


$p(x|\mu, \sigma^2)$
Data
x

Hypotheses

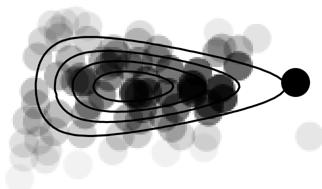
$$p(\mu, \sigma^2 | x) \propto p(x|\mu, \sigma^2) p(\mu, \sigma^2)$$

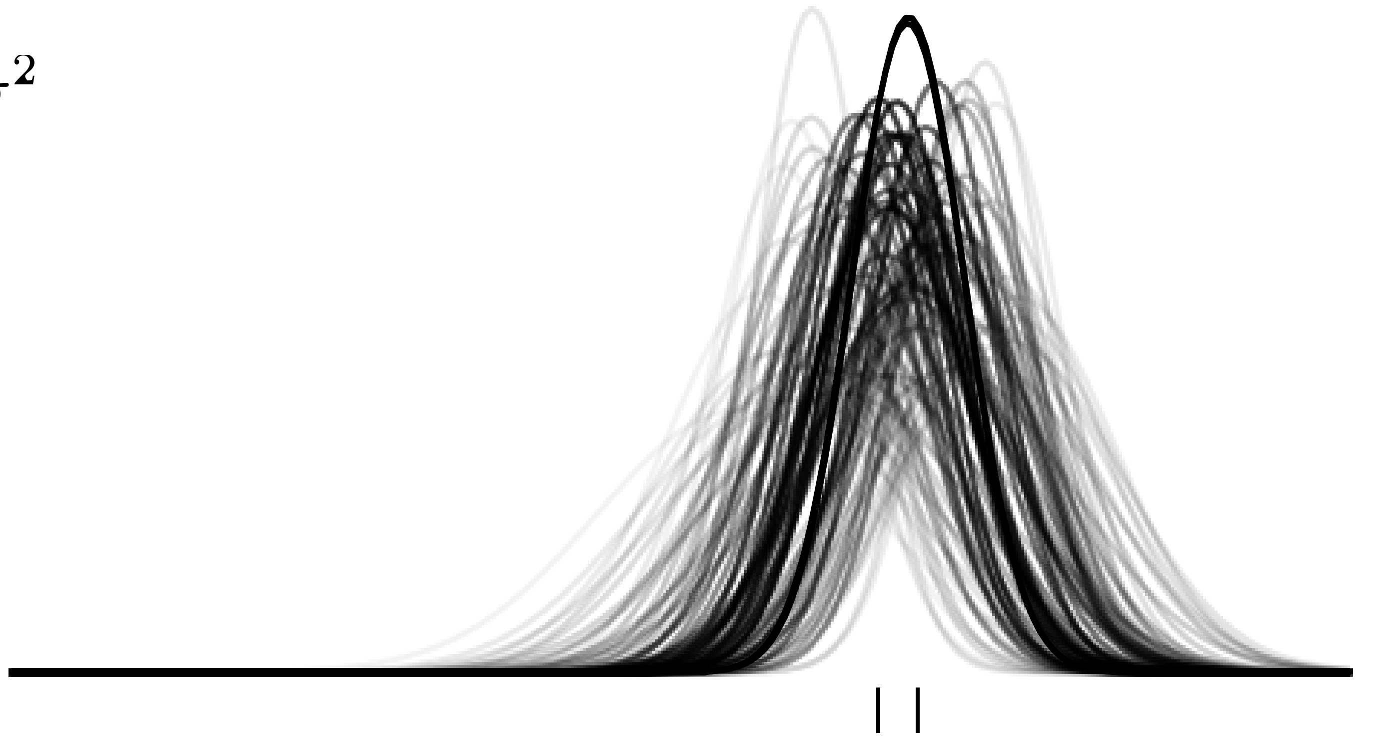




Hypotheses

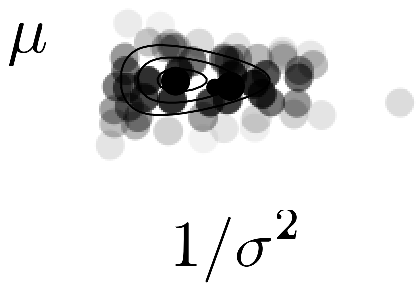
$$p(\mu, \sigma^2 | x) \propto p(x | \mu, \sigma^2) p(\mu, \sigma^2)$$

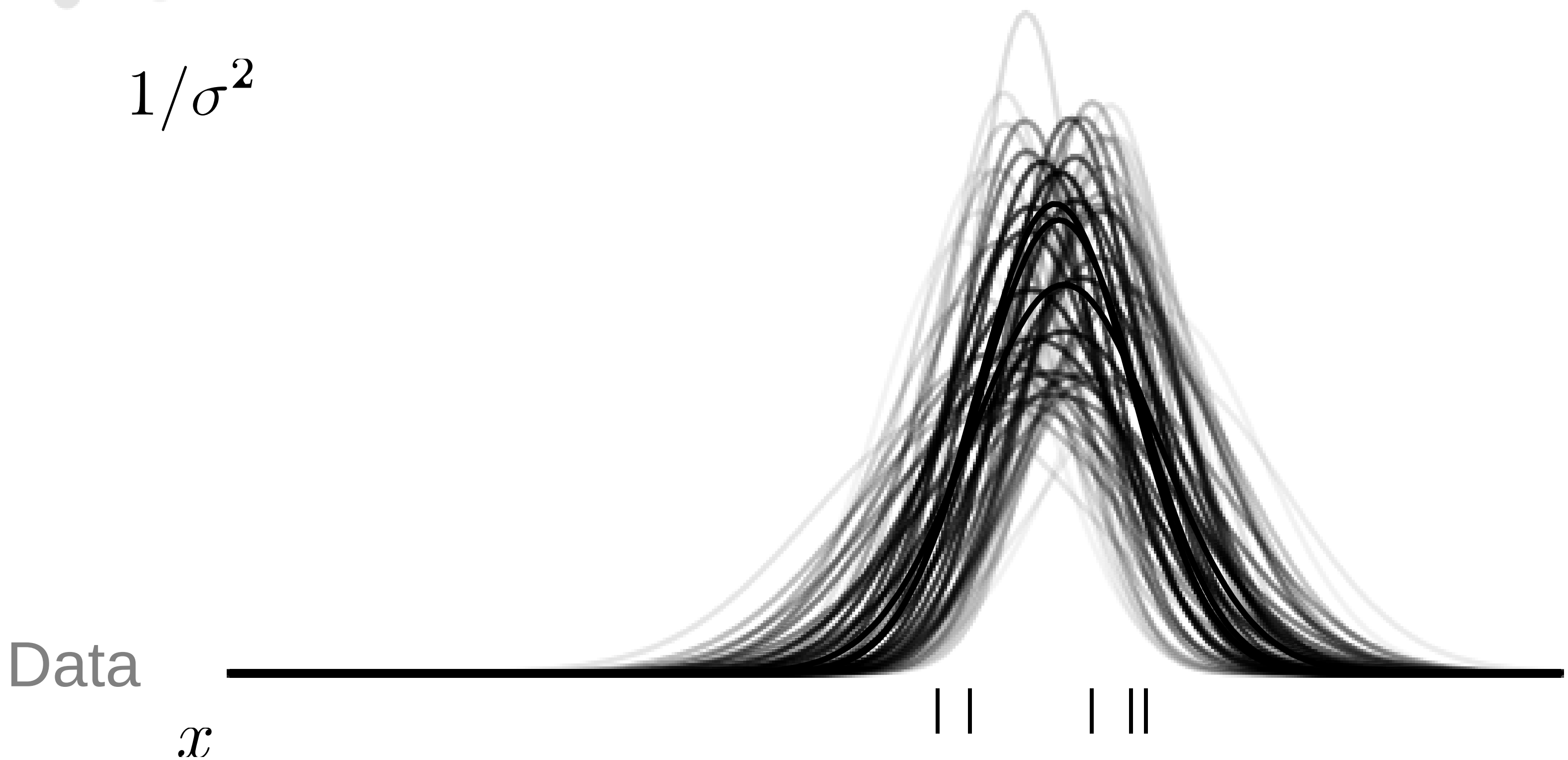




Hypotheses

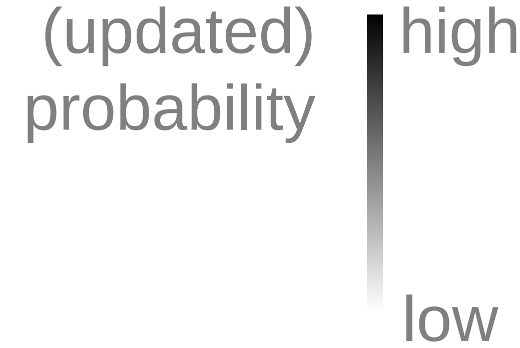
$$p(\mu, \sigma^2 | x) \propto p(x | \mu, \sigma^2) p(\mu, \sigma^2)$$

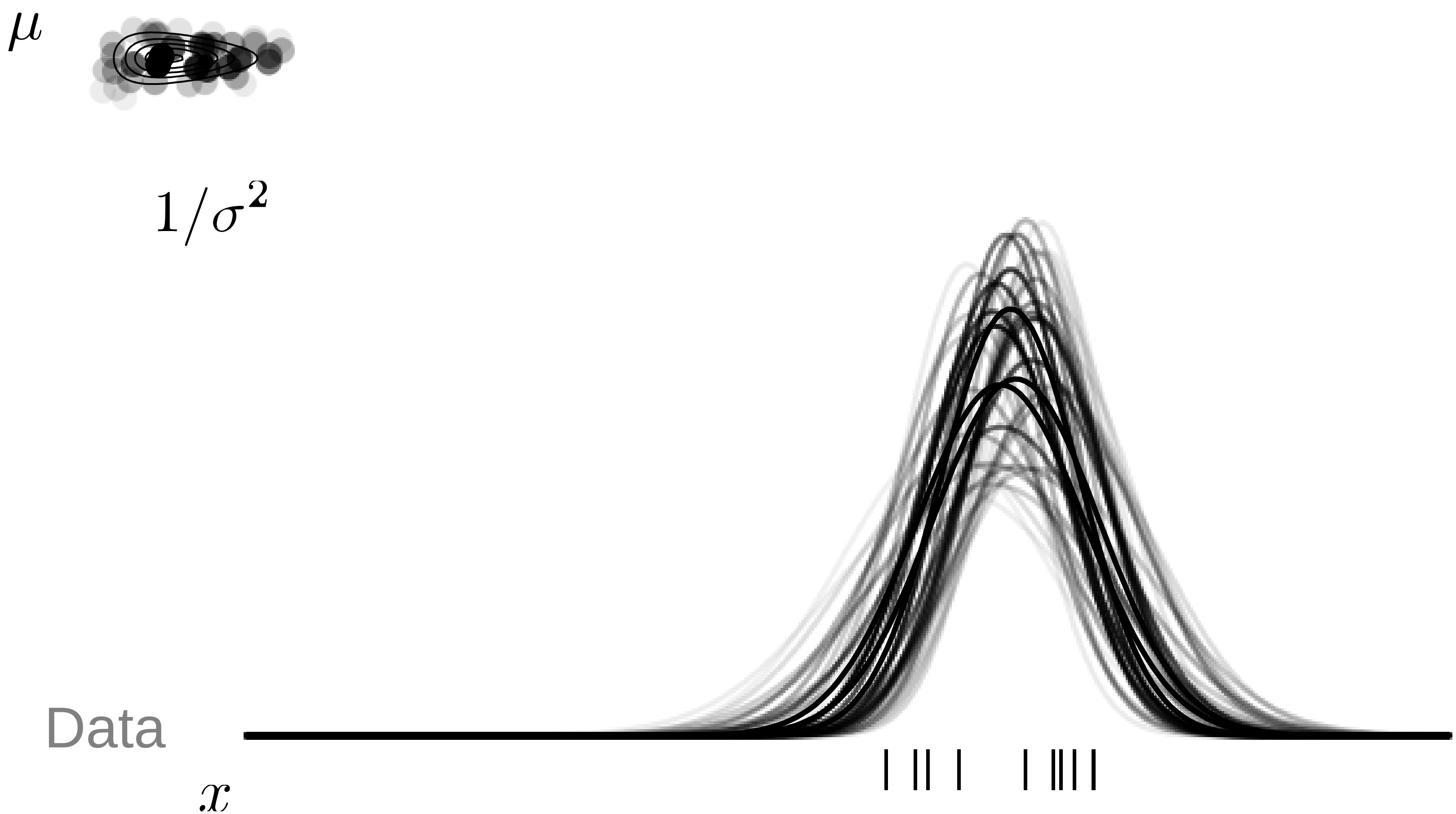




Hypotheses

$$p(\mu, \sigma^2 | x) \propto p(x | \mu, \sigma^2) p(\mu, \sigma^2)$$





Hypotheses

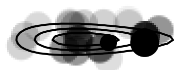
$$p(\mu, \sigma^2 | x) \propto p(x|\mu, \sigma^2) p(\mu, \sigma^2)$$

(updated) probability

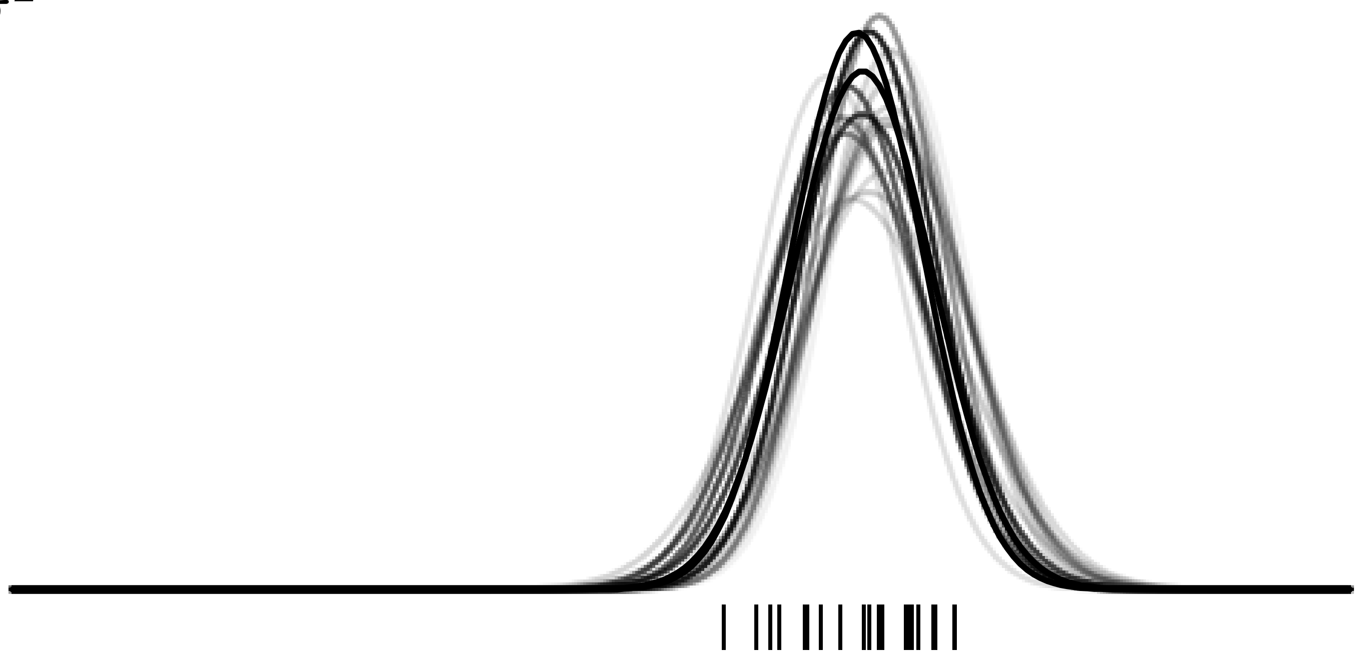
high

low

μ



$1/\sigma^2$

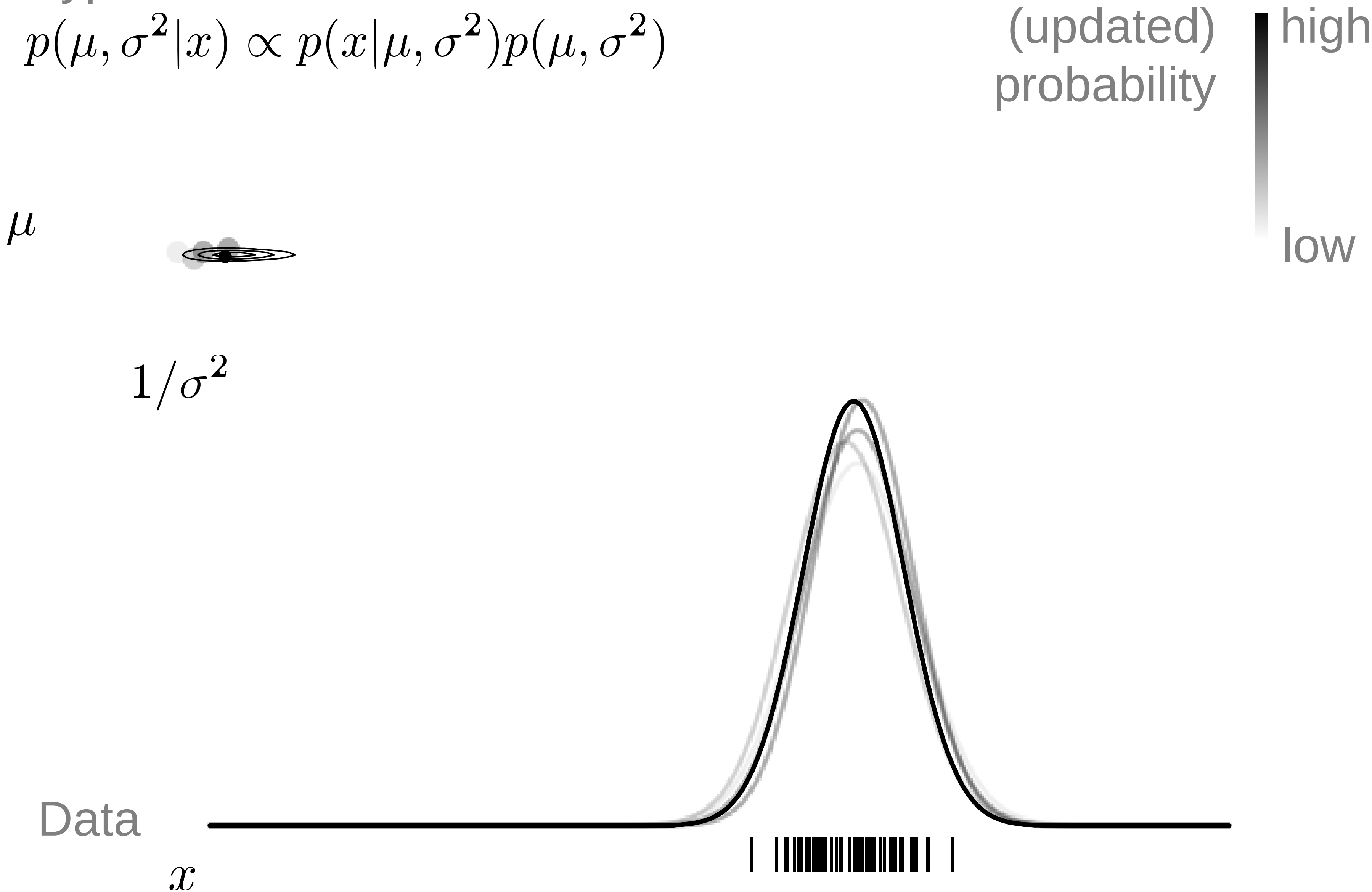


Data

x

Hypotheses

$$p(\mu, \sigma^2 | x) \propto p(x | \mu, \sigma^2) p(\mu, \sigma^2)$$



Hypotheses

$$p(\mu, \sigma^2 | x) \propto p(x|\mu, \sigma^2) p(\mu, \sigma^2)$$

(updated) probability

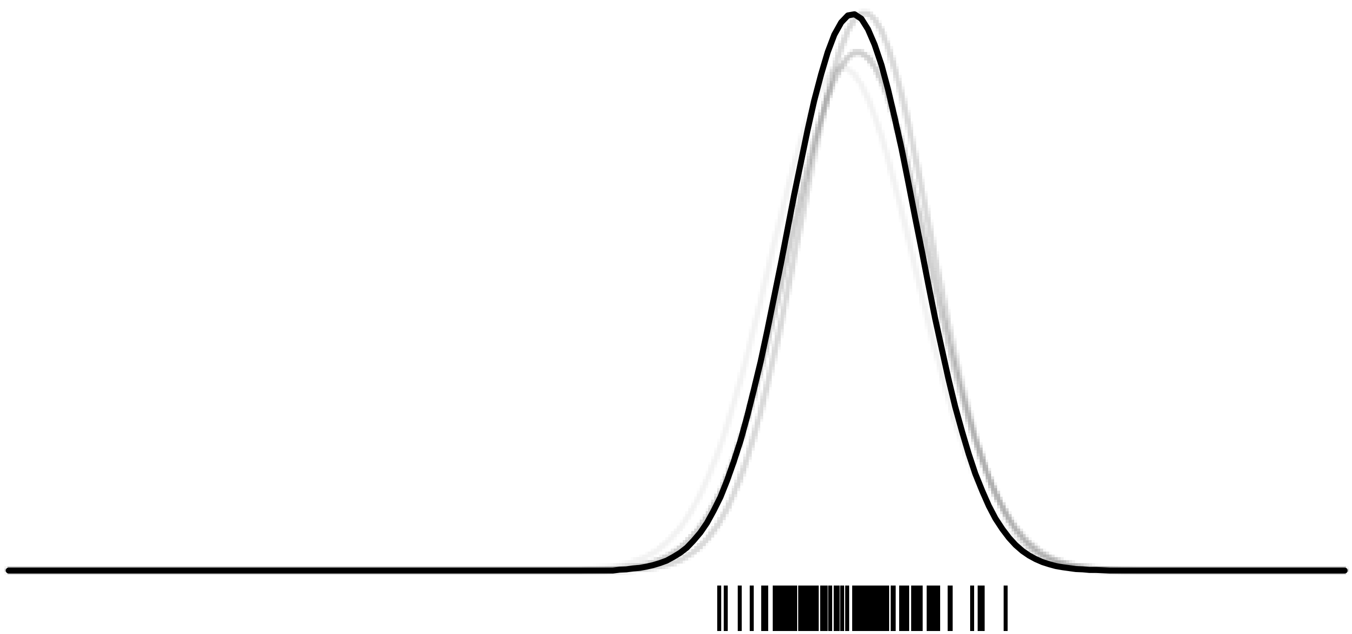
high

low

μ

$1/\sigma^2$

Data



x

Where do you get the **prior**?

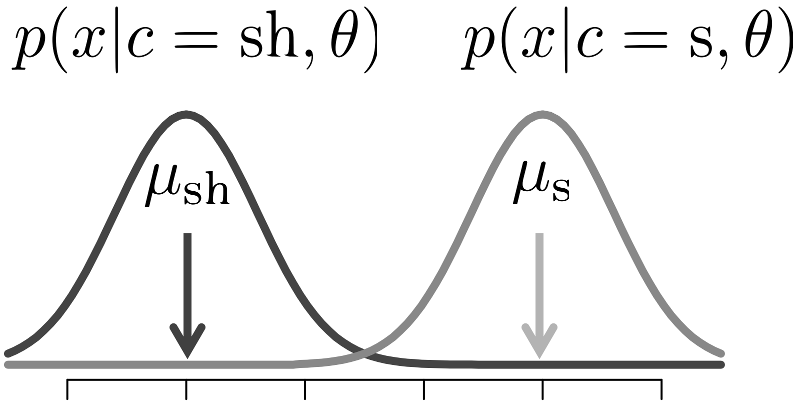

Distributional
learning

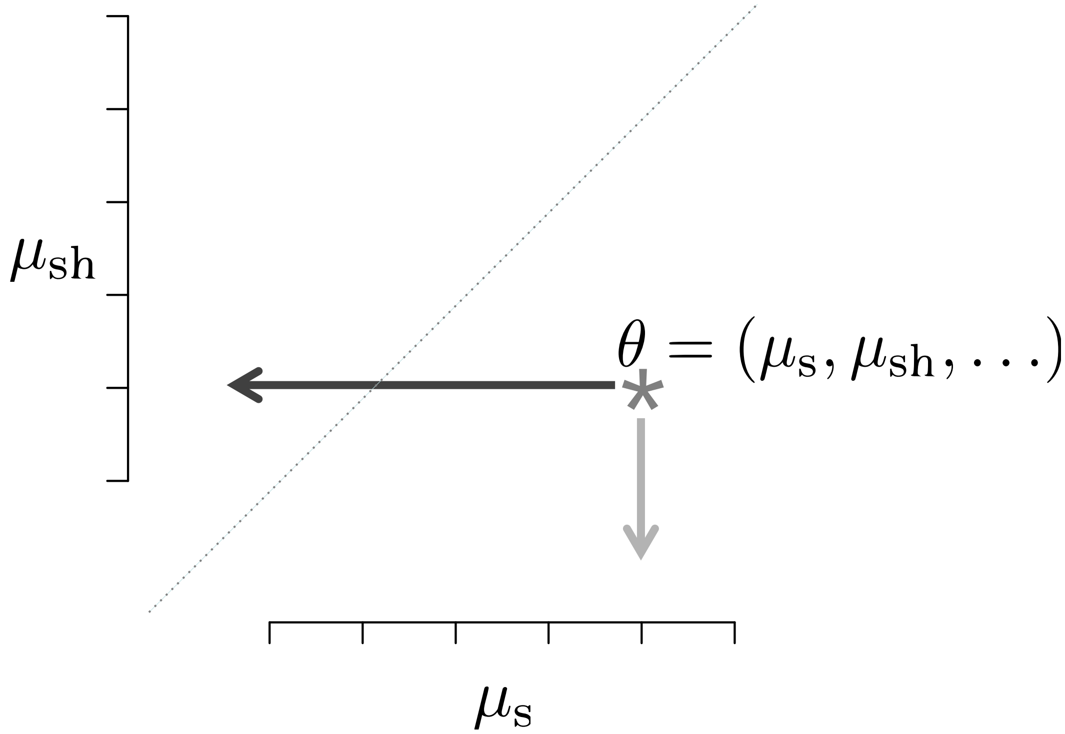
Where do you get the **prior**?

- Prior is a **prediction**
- Utility of prediction depends on **distribution** of **things in the world**
  - Predict things you **never** get: inefficient (takes longer to adapt)
  - Rule out things you **do** get: can't adapt
- Good prior depends on **learning those distributions** (of distributions)

$$p(\mu, \sigma^2, c|x) \propto p(x|c, \mu, \sigma^2) p(\mu, \sigma^2) p(c)$$

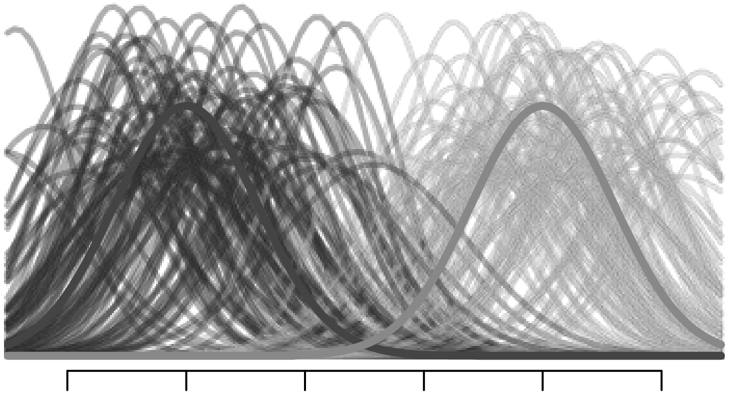
Observations Prior experience



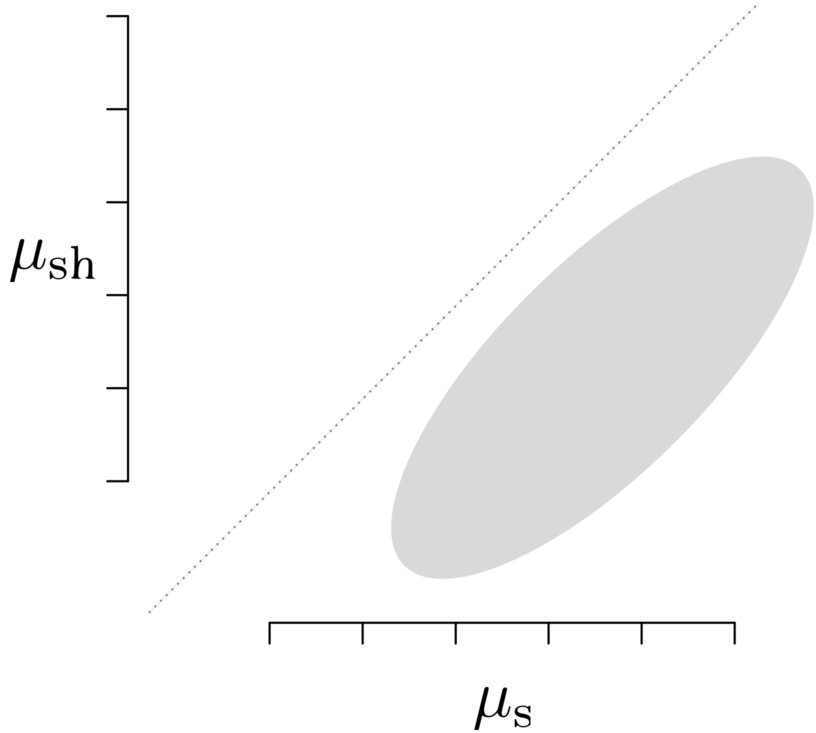


$$p(\mu, \sigma^2, c|x) \propto \boxed{p(x|c, \mu, \sigma^2)}\boxed{p(\mu, \sigma^2)}p(c)$$

$\boxed{p(x|c, \theta)}$

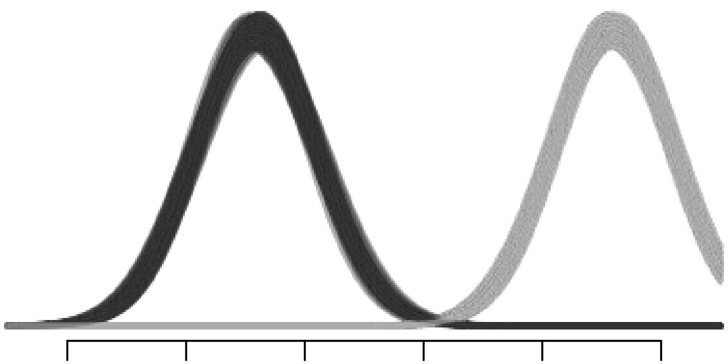


$\boxed{p(\theta)}$

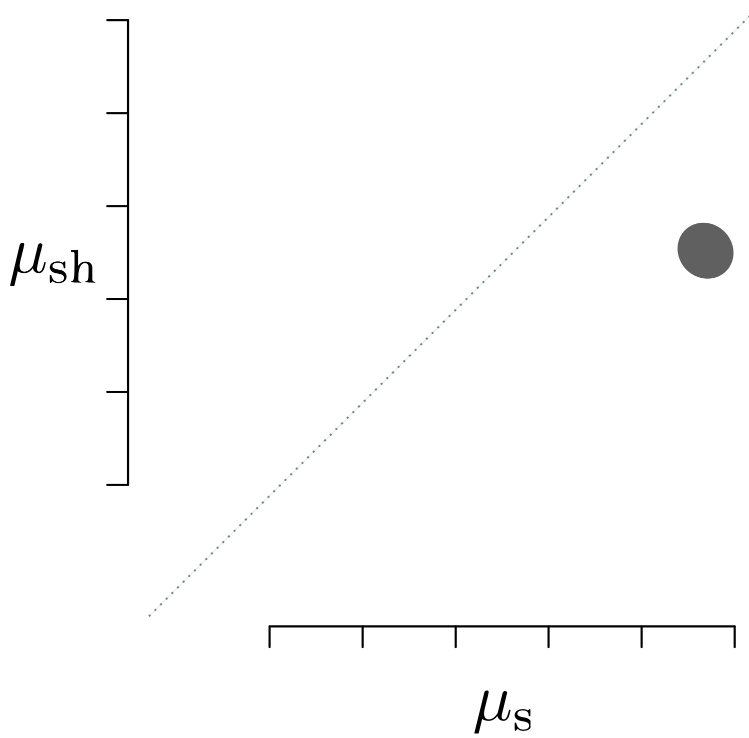


$$p(\mu, \sigma^2, c|x) \propto \boxed{p(x|c, \mu, \sigma^2)}\boxed{p(\mu, \sigma^2)}p(c)$$

$\boxed{p(x|c, \theta)}$

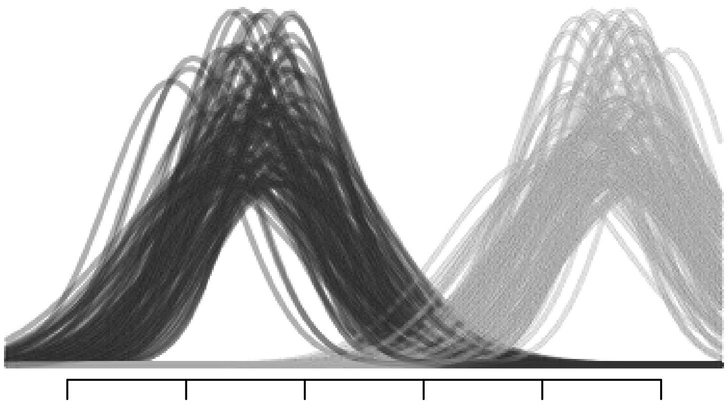


$\boxed{p(\theta)}$

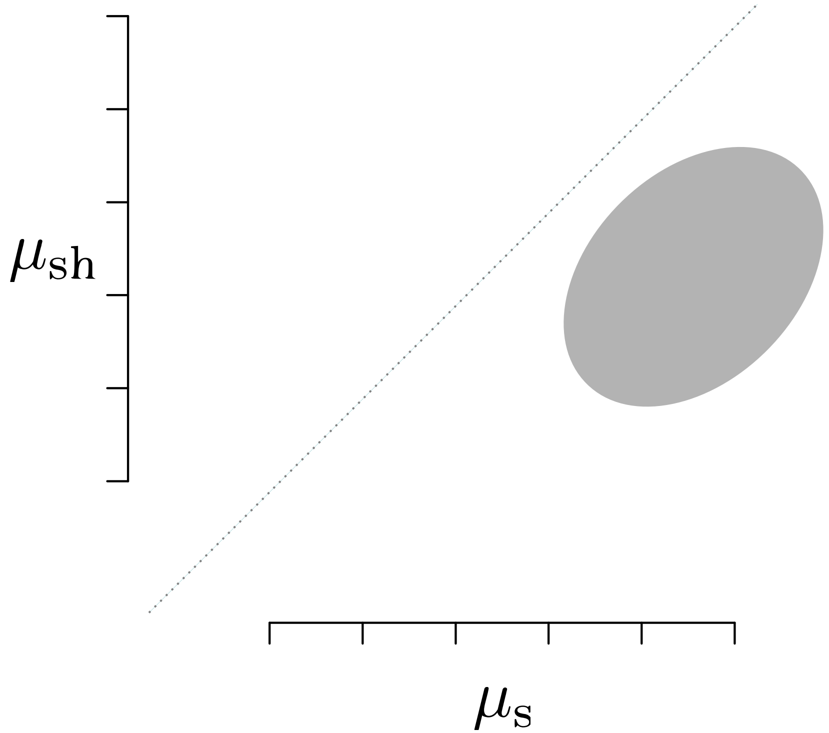


$$p(\mu, \sigma^2, c|x) \propto \boxed{p(x|c, \mu, \sigma^2)}\boxed{p(\mu, \sigma^2)}p(c)$$

$p(x|c, \theta)$



$p(\theta)$

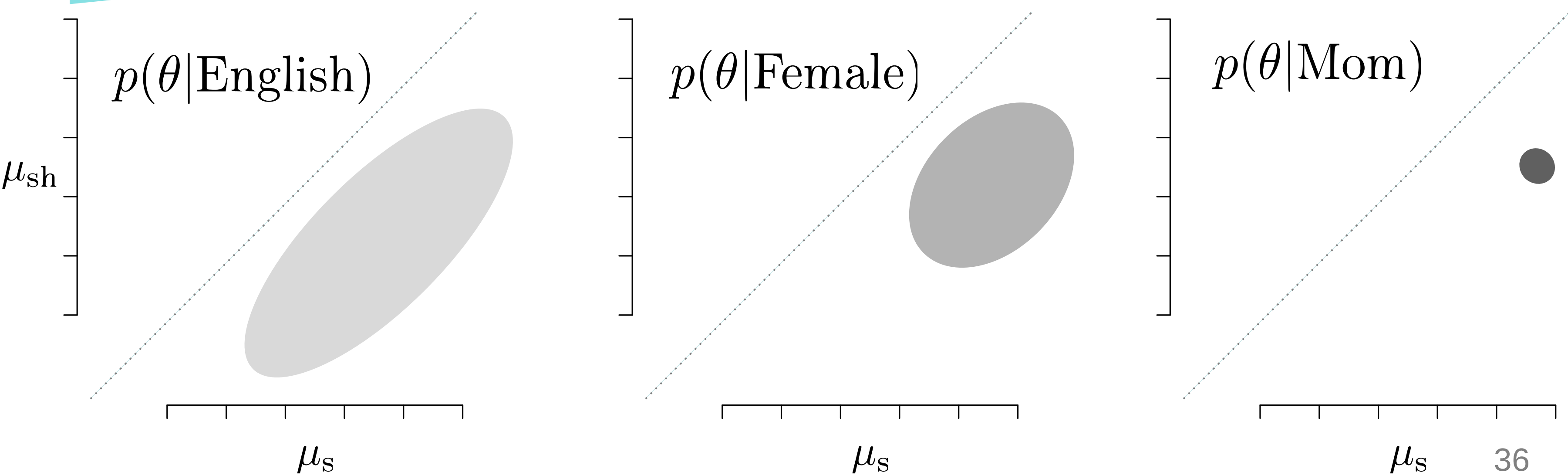


$$p(\mu, \sigma^2, c|x) \propto p(x|c, \mu, \sigma^2) p(\mu, \sigma^2|c) p(c)$$

How relevant is prior experience right now?

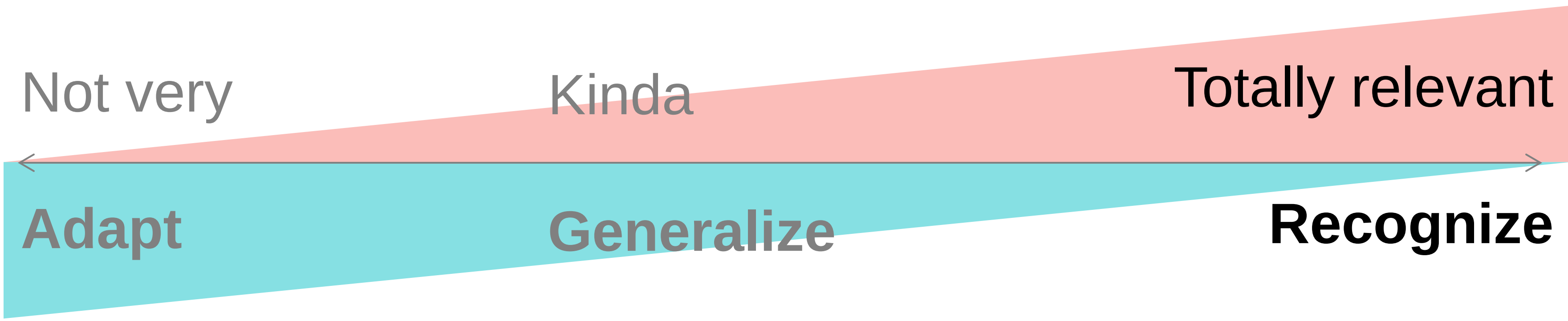
Not very | Kinda | Totally relevant

Adapt | **Generalize** | **Recognize**



$$p(\mu, \sigma^2, c, \boldsymbol{t} | x) \propto p(x | c, \mu, \sigma^2) p(\mu, \sigma^2 | \boldsymbol{t}) p(\boldsymbol{t}) p(c)$$

How relevant is prior experience right now?



Talker-specificity

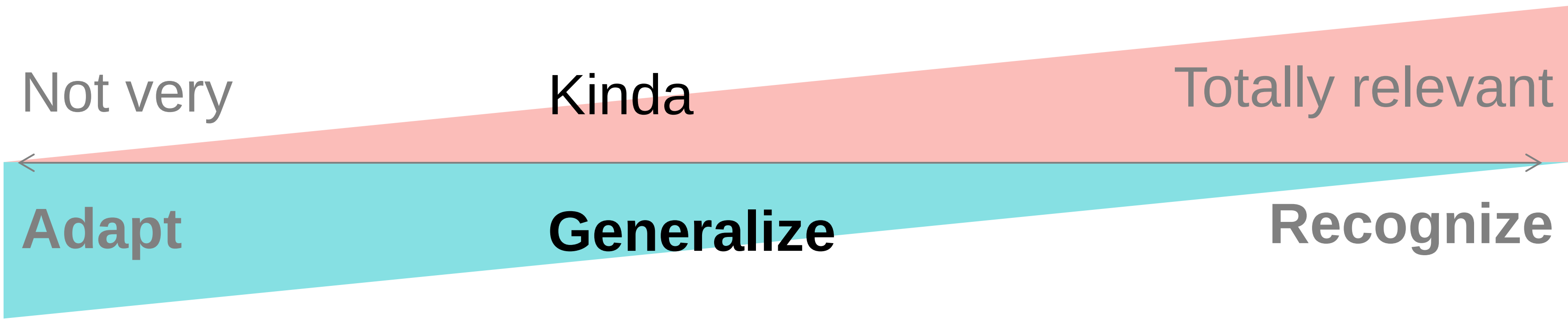
[e.g. Goldinger; Nygaard; Pisoni; Sumner; many others]

Persistent adaptation

[Eisner & McQueen, 2006; Kraljic & Samuel 2005]

$$p(\mu, \sigma^2, c, \boldsymbol{t} | x) \propto p(x | c, \mu, \sigma^2) p(\mu, \sigma^2 | \boldsymbol{t}) p(\boldsymbol{t}) p(c)$$

How relevant is prior experience right now?



Top down cues (gender, dialect, accent)

[e.g. Hay & Drager, 2010; B. Munson 2011, Niedzielski, 1999; Staum Casasanto, 2008; Strand & Johnson, 1996]

Accent adaptation

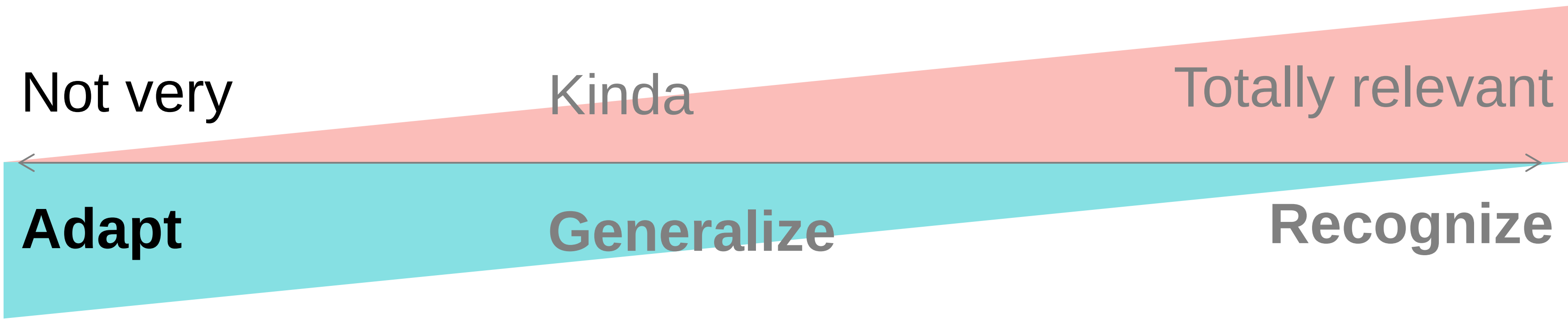
[Bradlow & Bent, 2008; Baese-Berk et al. 2013]

Generalization of recalibration

[Kraljic & Samuel 2005, 2007; C. Munson, 2011; Reinisch & Holt, 2013]

$$p(\mu, \sigma^2, c, \boldsymbol{t} | x) \propto p(x | c, \mu, \sigma^2) p(\mu, \sigma^2 | \boldsymbol{t}) p(\boldsymbol{t}) p(c)$$

How relevant is prior experience right now?



Perceptual learning/Recalibration

[e.g., Bertelson et al. 2003; Norris et al. 2003; Maye, Aslin, & Tanenhaus, 2008]

Selective adaptation*

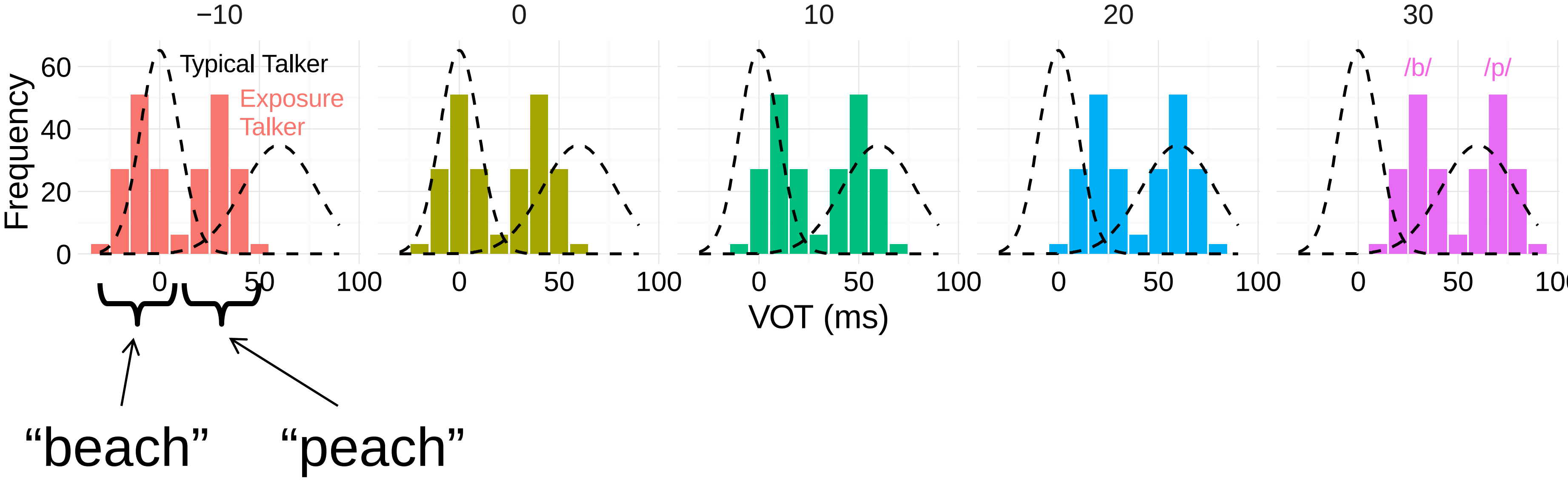
[e.g., Eimas and Corbit, 1973; Samuel 1986]

Language-level prior

[e.g., Idemaru & Holt, 2011; Schertz et al., 2016; Kleinschmidt & Jaeger, *submitted*]

Prior beliefs constrain adaptation

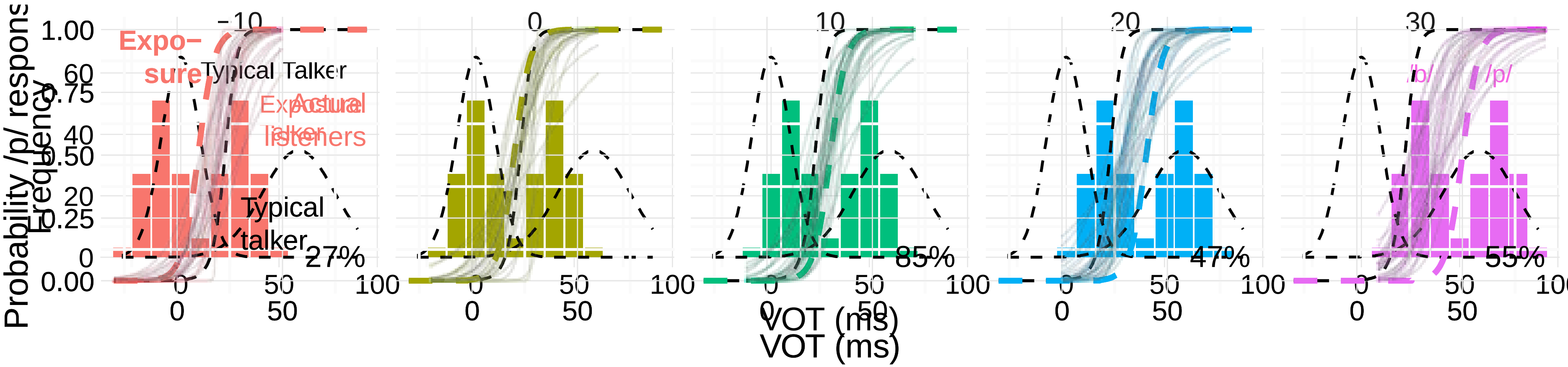
Expose to **different accents** (*n* = 138 on MTurk):



[Based on Clayards et al. 2008; Kleinschmidt & Jaeger, CogSci2016]

Prior beliefs constrain adaptation

Predict classification curves …and **compare** with behavior

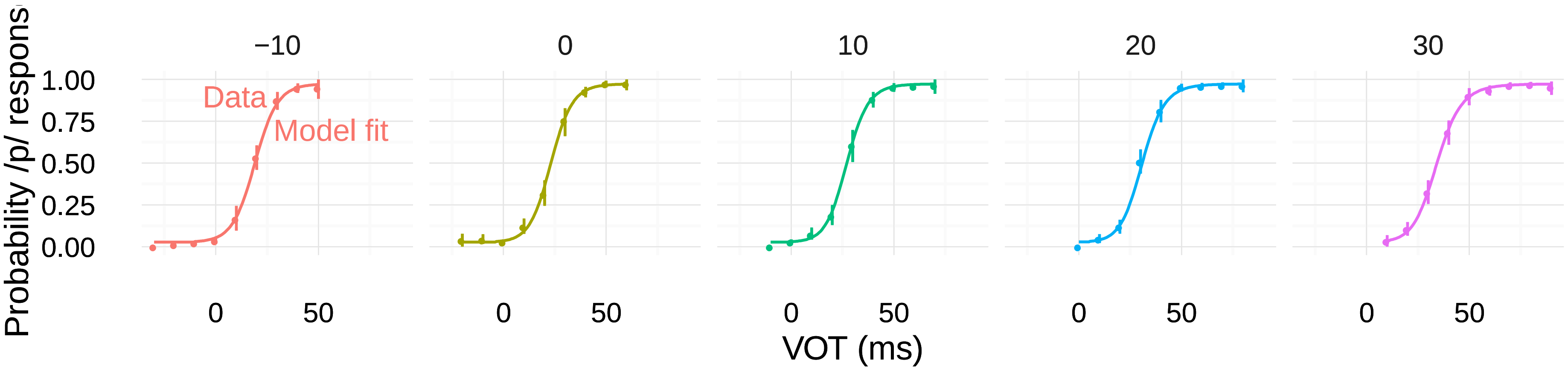


[Based on Clayards et al. 2008; Kleinschmidt & Jaeger, CogSci2016]

Adaptation tells us something about **prior beliefs**

Fit “inverted belief updating” model

Prior beliefs are **unknown**, inferred from adaptation data



How does the **brain** do this?

We have no idea!!

How does the **brain** do this?

- **Predictive coding** in posterior STS [Blank & Davis, 2016 PLoS biology]
- Neural correlates of **recalibration** in frontal, temporal, and parietal areas [Kilian Hutten et al. 2011; Myers & Mesite, 2014]

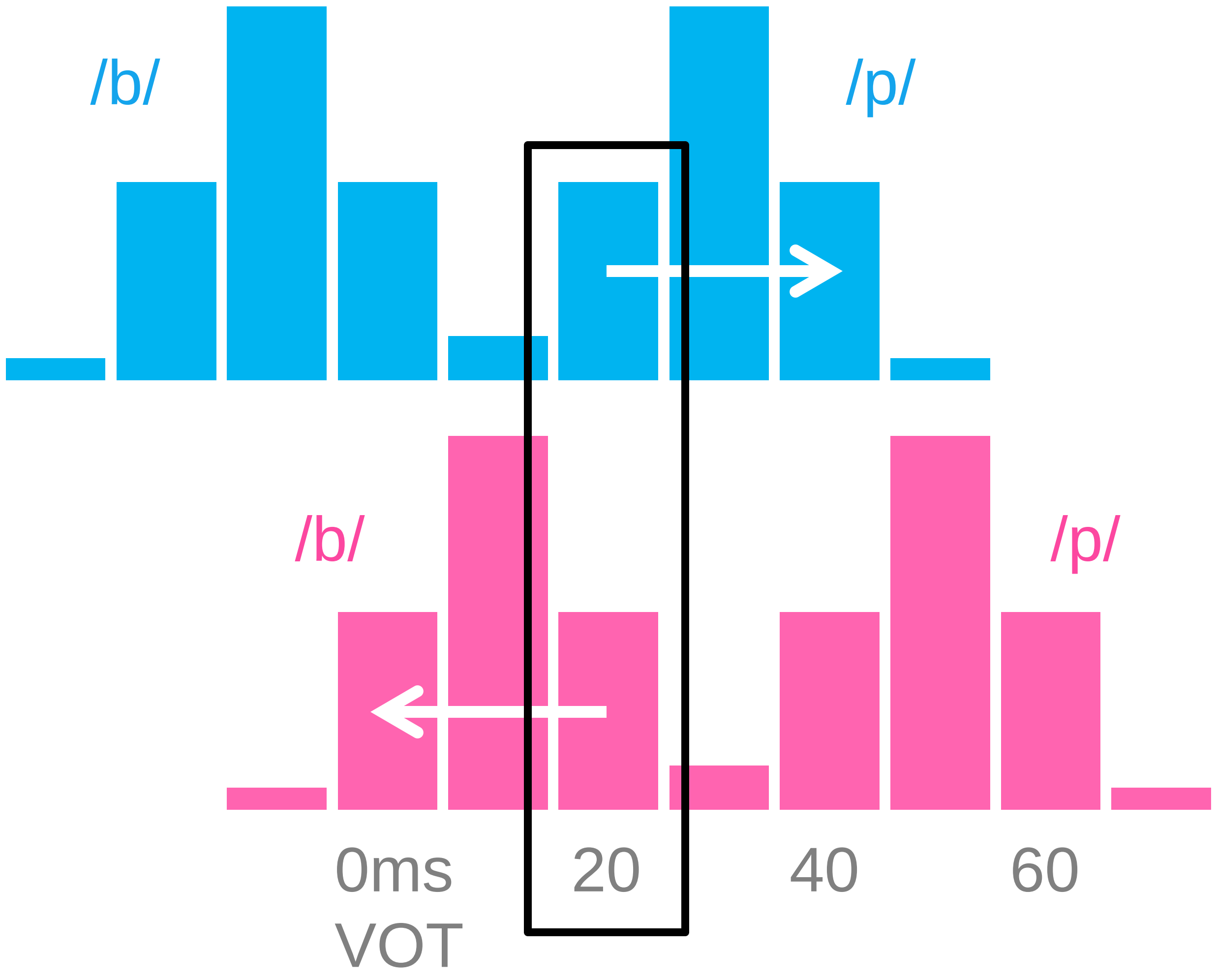
Open questions: **adaptation**

- How does the brain **learn new models**?
- ...store, retrieve, and update **existing** models?
- ...learn, represent, and deploy knowledge about the **structure** of variation (in e.g. groups of talkers)?

Learning **new models**

- What is the locus of updated beliefs?
  - Low level representation of sounds
  - Higher level categorical read out

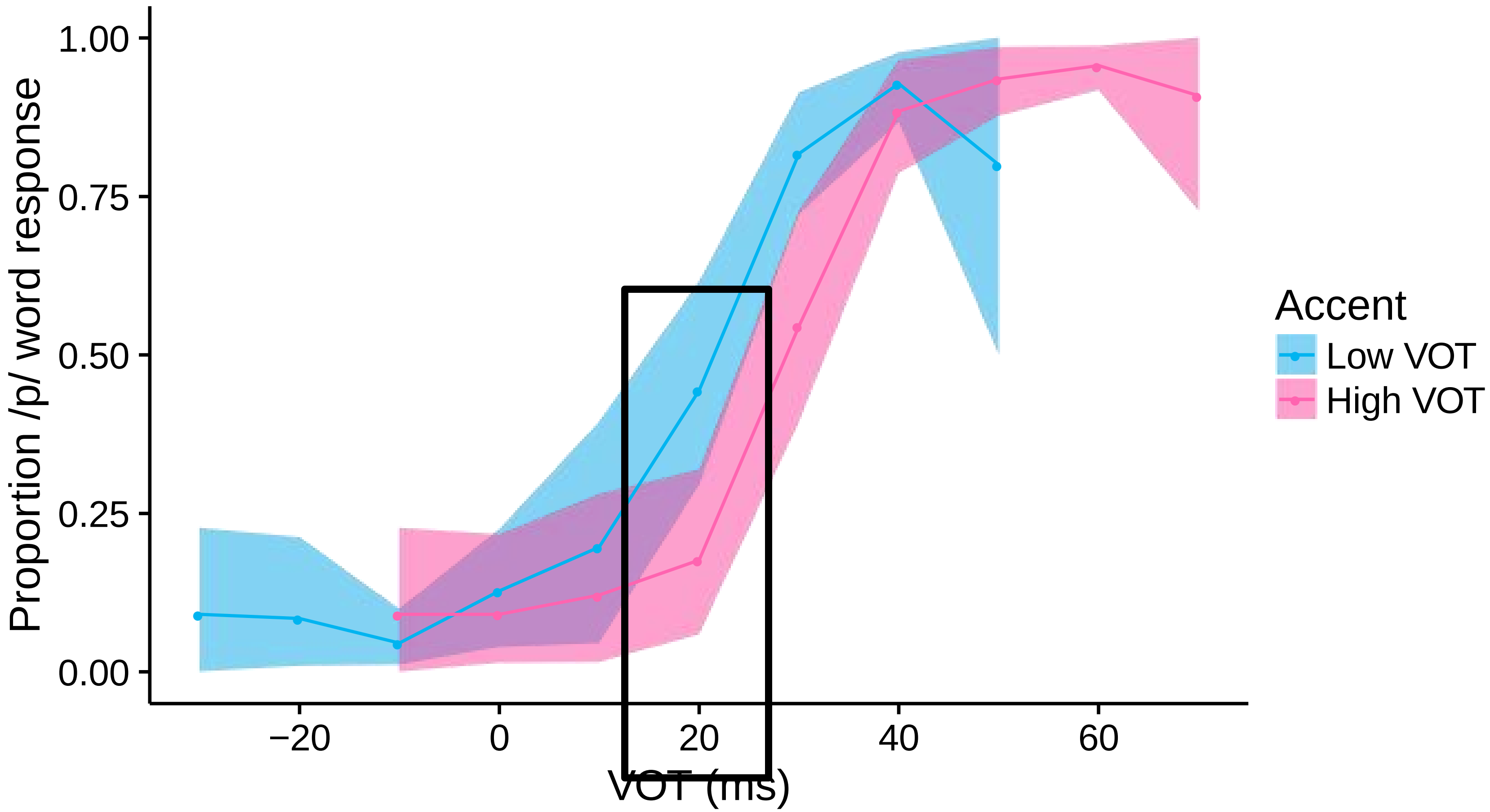
1. **Expose** each subject to one of two accents



Low VOT
n = 10 subjects

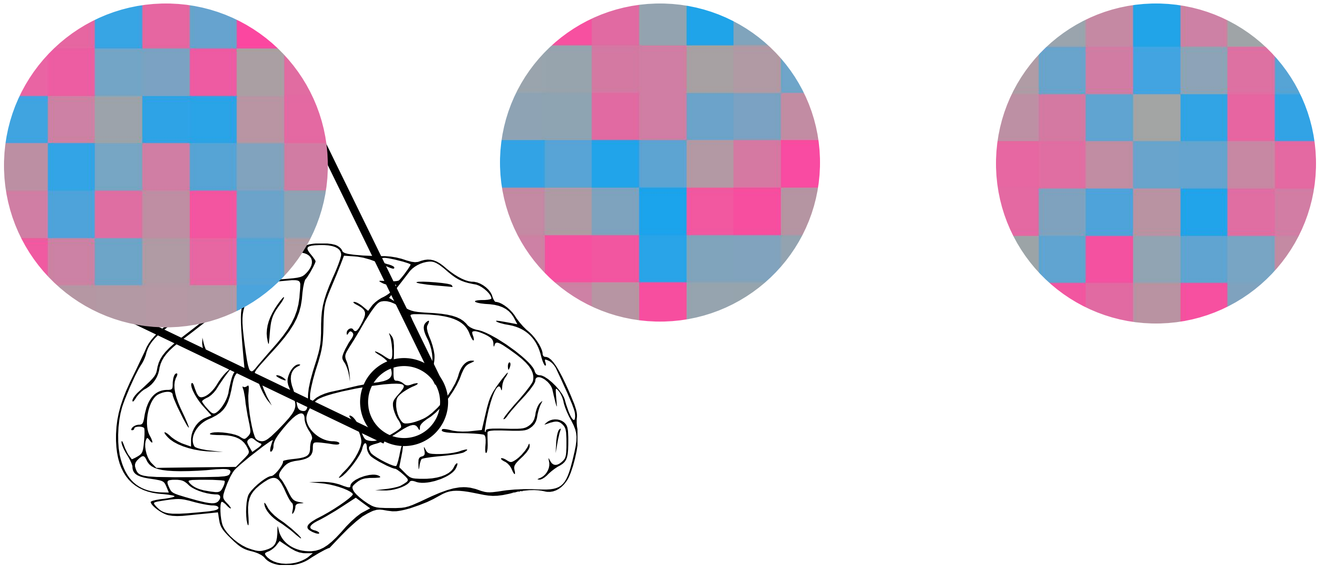
High VOT
n = 10 subjects

1½. Verify **distributional learning** behaviorally

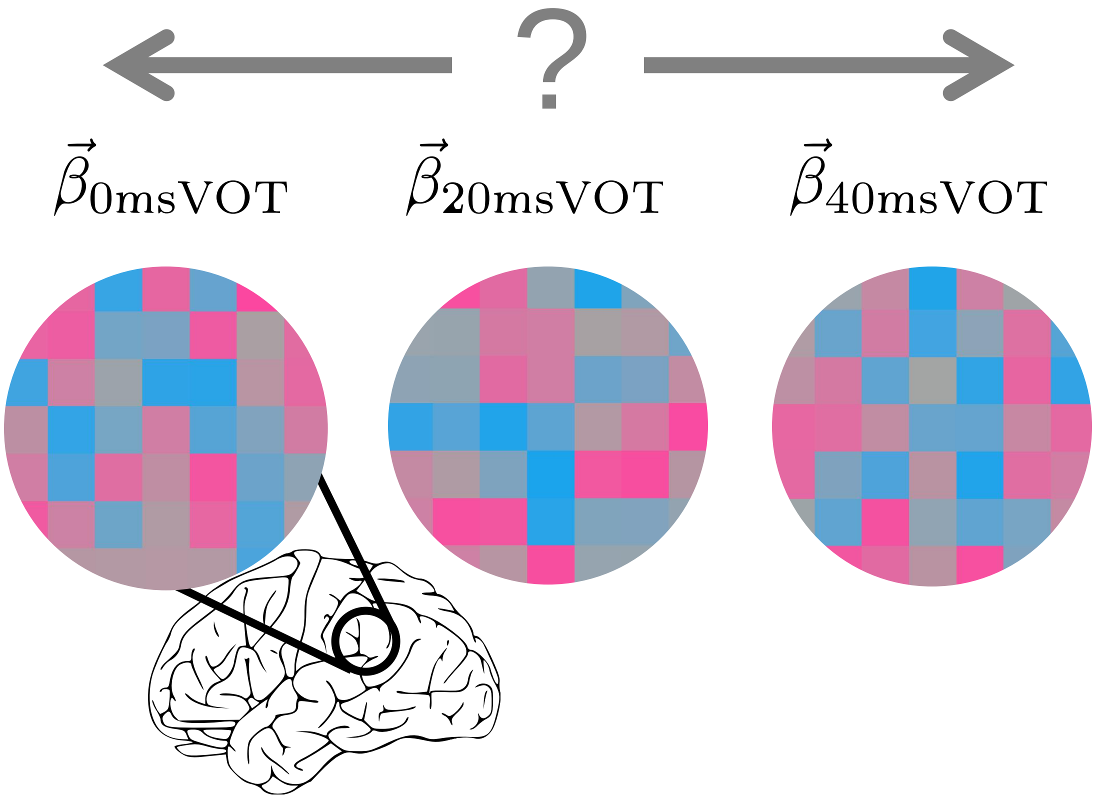


2. **Measure** multivoxel patterns with fMRI

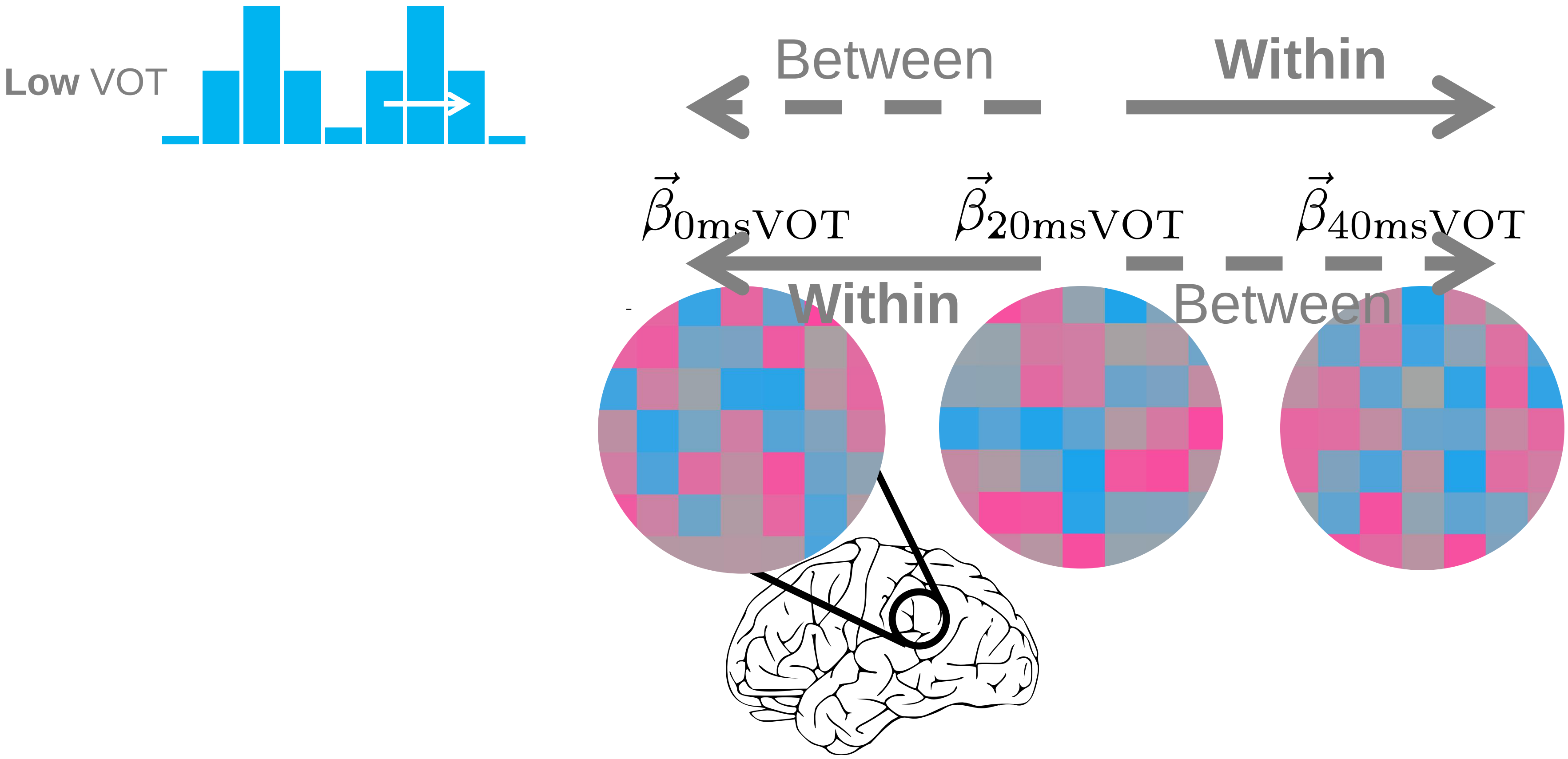
$\vec{\beta}_{\mathrm{0msVOT}}$ $\vec{\beta}_{\mathrm{20msVOT}}$ $\vec{\beta}_{\mathrm{40msVOT}}$



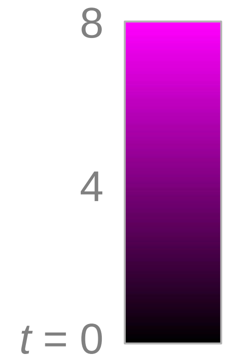
3. **Compare** within- vs. between-category pattern similarity



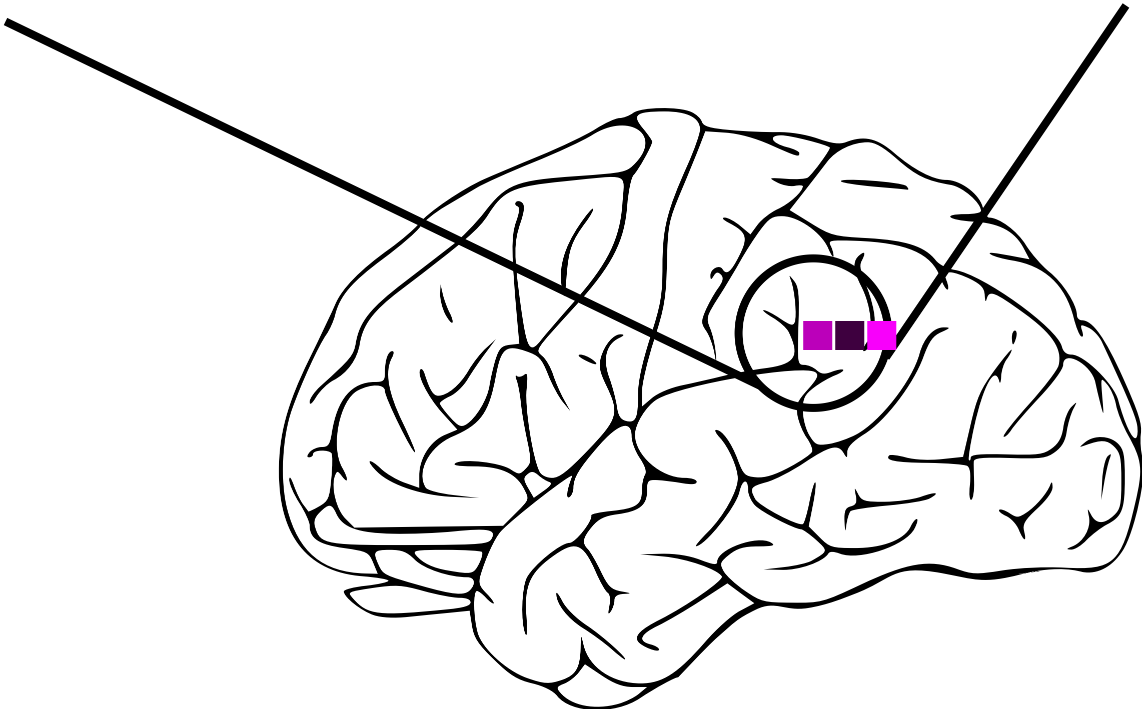
3. **Compare** within- vs. between-category pattern similarity



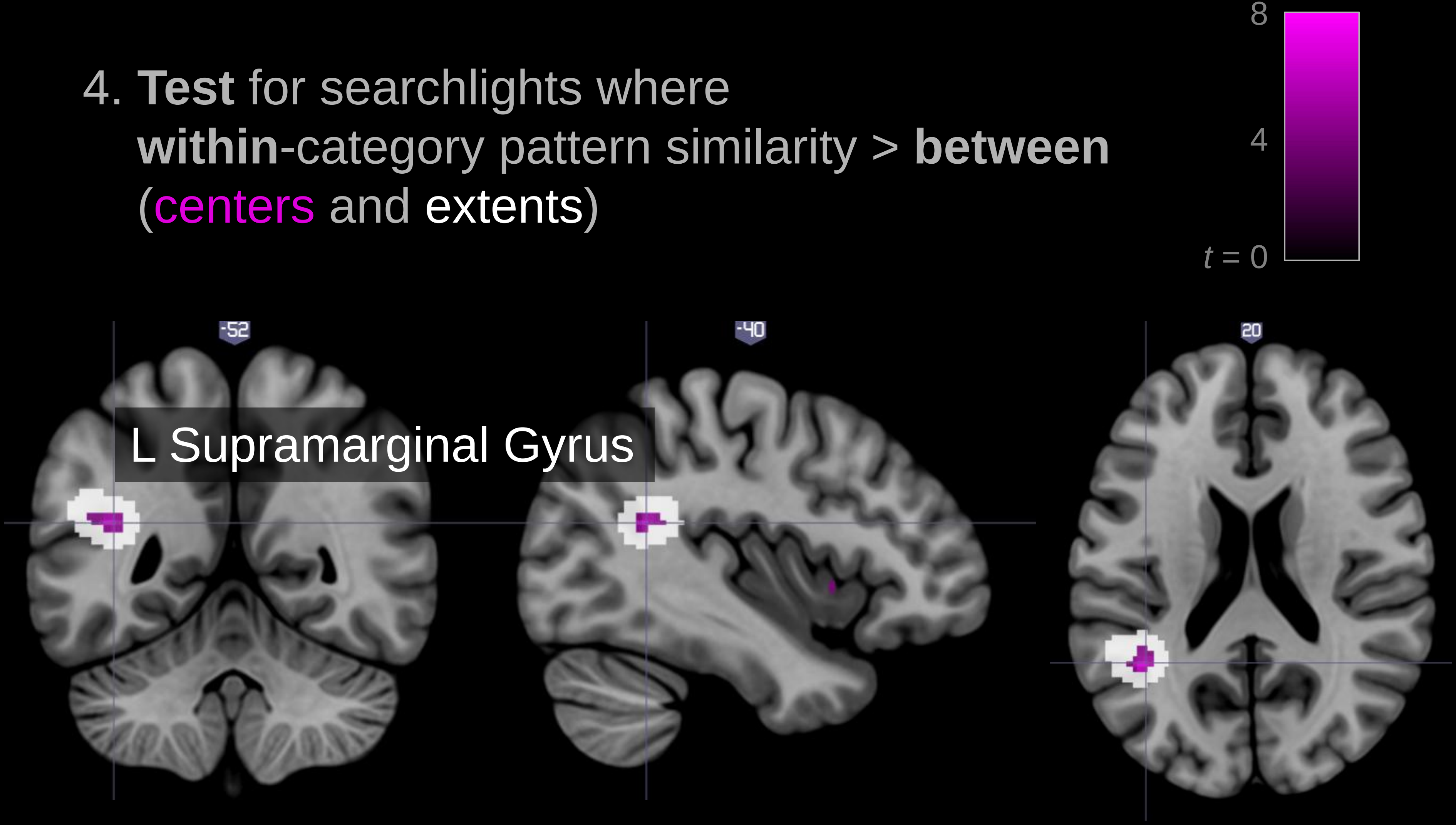
4. **Test** for searchlights where **within**-category pattern similarity > **between**



$$\boldsymbol{t}\left(\mathrm{sim}(\vec{\beta}_{\mathbf{2}0\mathrm{ms}}, \vec{\beta}_{\mathrm{within}}) - \mathrm{sim}(\vec{\beta}_{\mathbf{2}0\mathrm{ms}}, \vec{\beta}_{\mathrm{between}})\right)$$



4. **Test** for searchlights where **within**-category pattern similarity > **between** (centers and extents)



searchlight cluster size ($t > 4$) $p < 0.05$, peak $p < 0.001$ by sign-flipping permutation test

Supramarginal gyrus

- **Categorical** processing of speech.
  [Raizada & Poldrack, 2007]
- Integrates auditory + articulatory representations (+ other senses).
  [Hickok & Poeppel, 2007]
- Implicated in **phonetic recalibration**, another rapid adaptation effect that requires a **teaching signal.**
  [Killian-Hütten et al. 2011; Myers et al. 2014]

A unified mechanism?

- Recalibration is a form of distributional learning. [Kleinschmidt & Jaeger, 2015]
- Unified **computational** principle might correspond to unified **neural** mechanism in SMG.

Wrapping up

- Good inference requires accurate **predictions**
- Good **predictions** require **adaptation** (when distributions change)
- **Adaptation** is **inference** (at a higher level)
- ...so adaptation requires good **predictions** (at a higher level)

Effective inference in a **variable world** involves prediction (errors) at **many levels**:

- About **stimuli** (cues)
- About **distributions** (accents)
- About **contexts** (type of talker)