

# **Models of decision making**

**Lucas Pinto  
Assistant Professor  
Department of Neuroscience  
Northwestern University**

**NUIN 443, 04/08/2024**

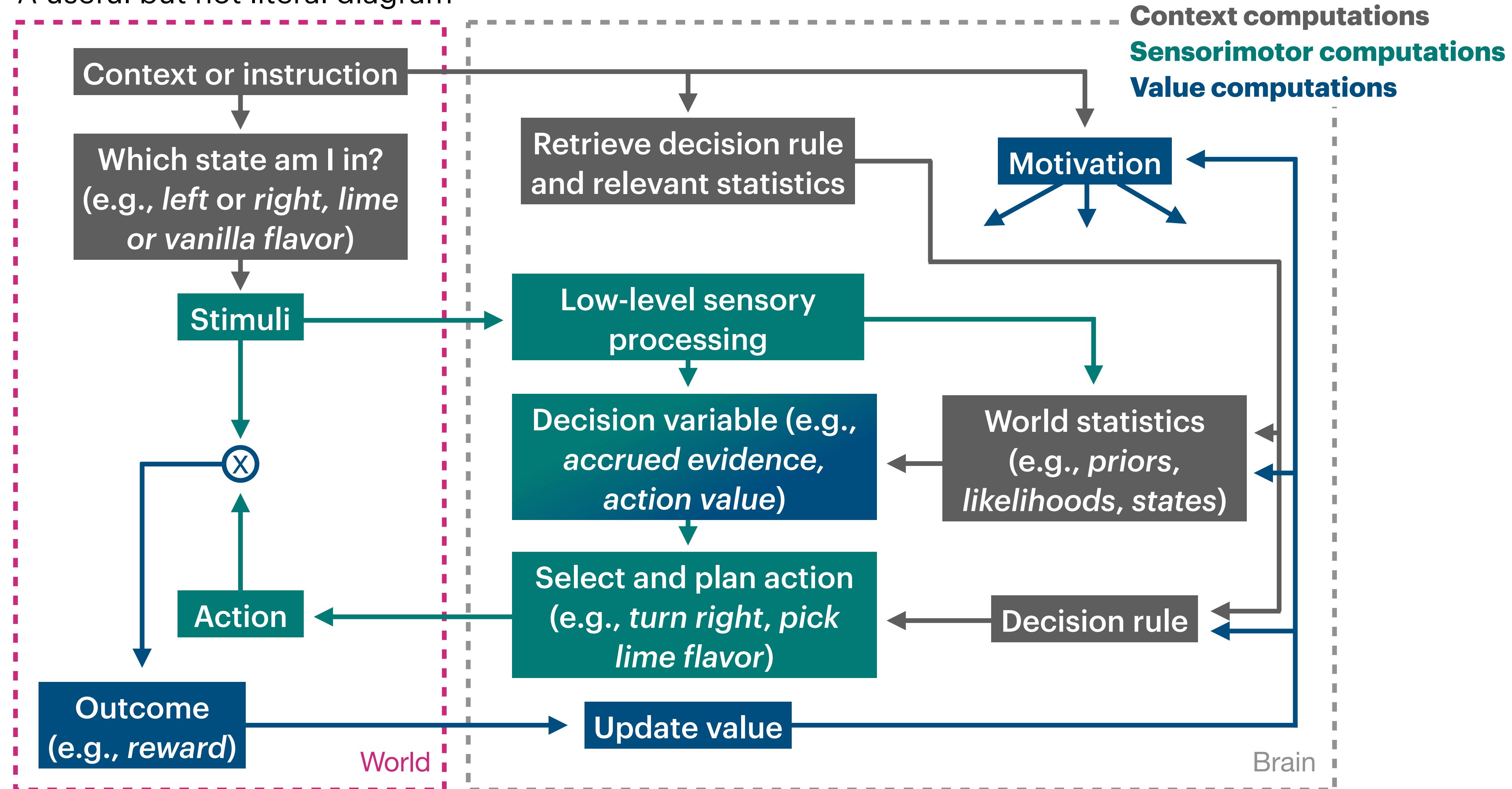
# Outline

- ▶ Intro: decision making
- ▶ Perceptual decision making
  - ▶ Signal detection theory (SDT)
  - ▶ Sequential sampling: DDMs and friends
- ▶ Value-based decision making
  - ▶ Value, utility
  - ▶ RL

# Intro: decision making

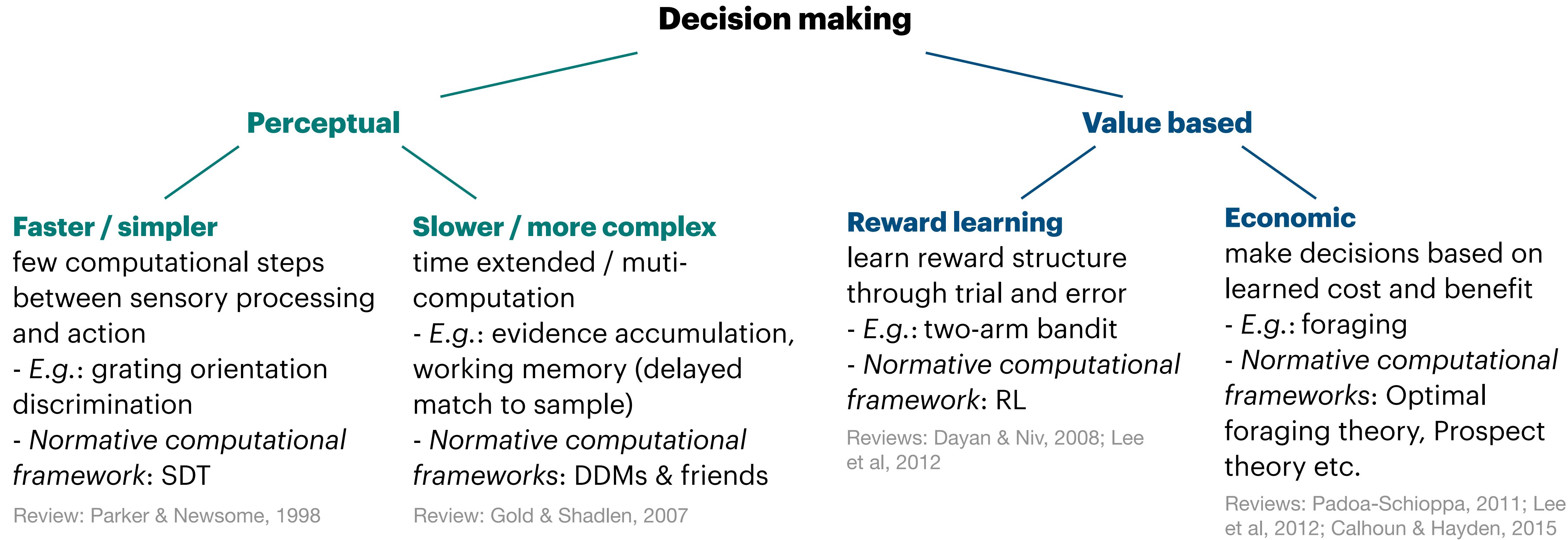
# Introduction

A useful but not literal diagram



after Gold & Shadlen, 2007

# Decision making in the lab: unofficial taxonomy



## Common task types by response:

**Go/no-go**: make or omit a response

**Two- (n-) alternative forced choice**: choose one of n alternatives

## Common task types by timing:

**Trial start**: Self-paced vs. Enforced trials

**Decision**: Reaction time vs. Fixed-time interrogation

# Decisions are noisy



[www.shutterstock.com](http://www.shutterstock.com)

Most formal decision-making frameworks start from the same assumption: *decisions are noisy*.

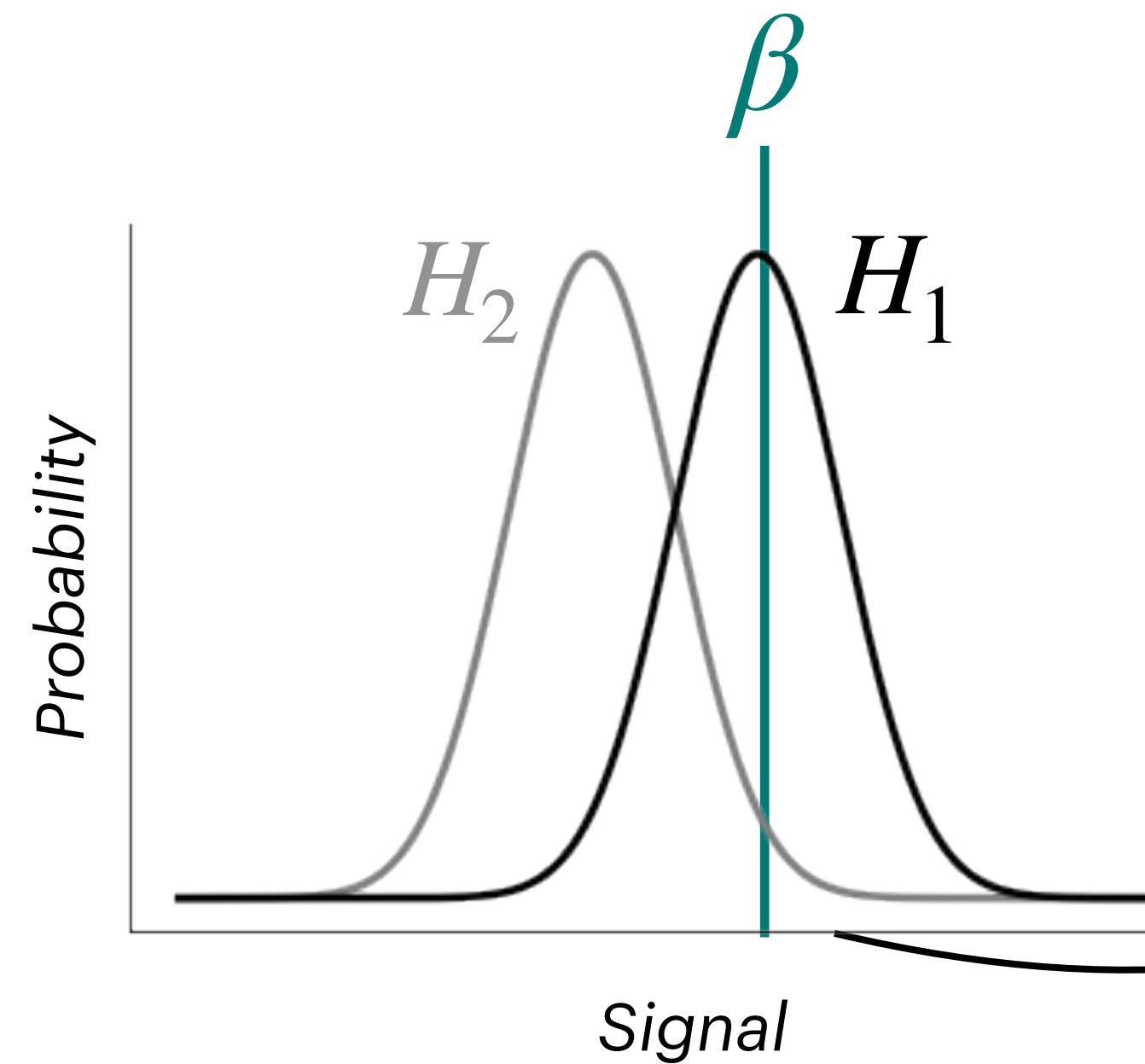
Take a visual decision:

- ▶ External noise:
  - ▶ Photon emission is Poisson
  - ▶ Environmental conditions vary (rain water is scattering)
- ▶ Internal noise
  - ▶ Eye optics scatter photons
  - ▶ Synaptic release, transduction, temperature
  - ▶ Ongoing brain dynamics (may not be noise, but still leads to behavioral variability)

# **Perceptual decision making: Signal detection theory (SDT)**

# SDT: decisions from single samples

We have two distributions,  $H_1$  and  $H_2$ , for the magnitude of a signal (stimuli, spike counts etc)



Now we draw a single piece of noisy evidence,  $E$ . How do we decide which distribution it came from?

We first compute the likelihood ratio  $L_{12}$  between the two conditional probabilities

$$L_{12}(E) = \frac{P(E|H_1)}{P(E|H_2)}$$



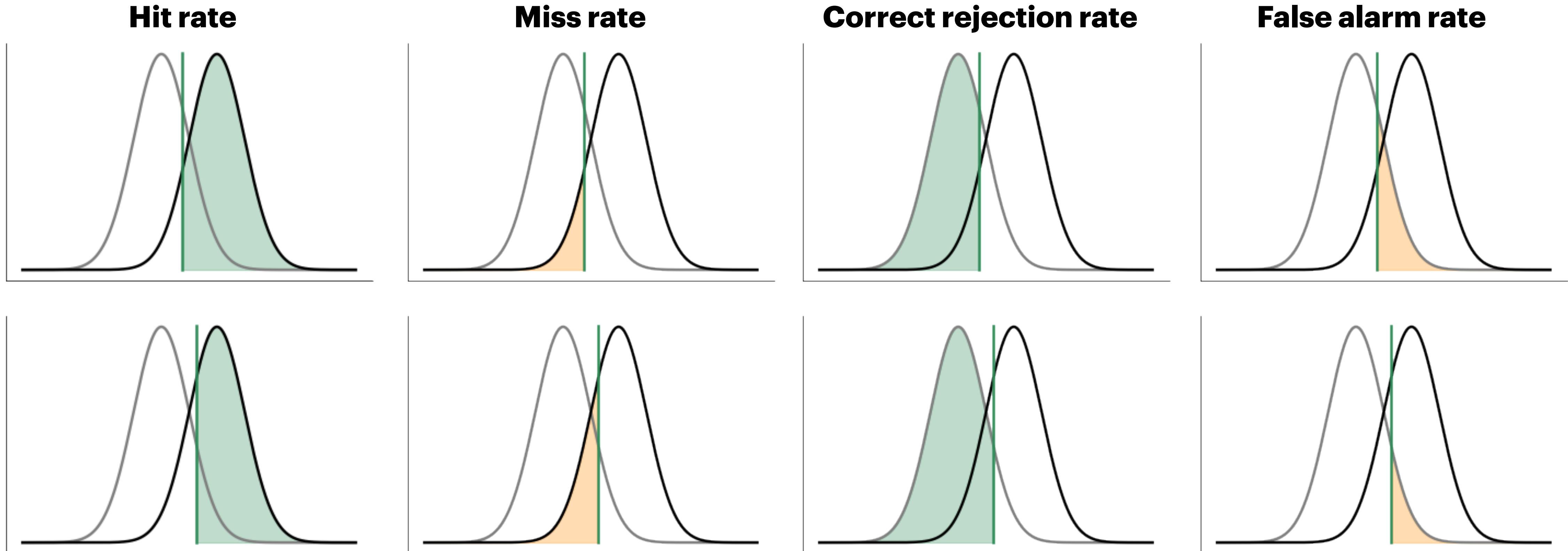
We choose  $H_1$  if

$$L_{12}(E) > \beta \text{ (criterion, i.e. threshold)}$$

- $H_1$  and  $H_2$  can represent different things depending on the task. E.g., for detection tasks  $H_2$  is a noise distribution, but for discrimination  $H_1$  and  $H_2$  each correspond to a stimulus class

# Choosing beta

Take a go/no-go decision



- ▶ If the goal is overall accuracy and  $H_1$  and  $H_2$  are equally likely, it can be shown that the optimal is  $\beta = 1$
- ▶ If the goal is overall accuracy and  $H_1$  and  $H_2$  have different prior probability  $P$ , the the optimal is  $\beta = P(H_1)/P(H_2)$

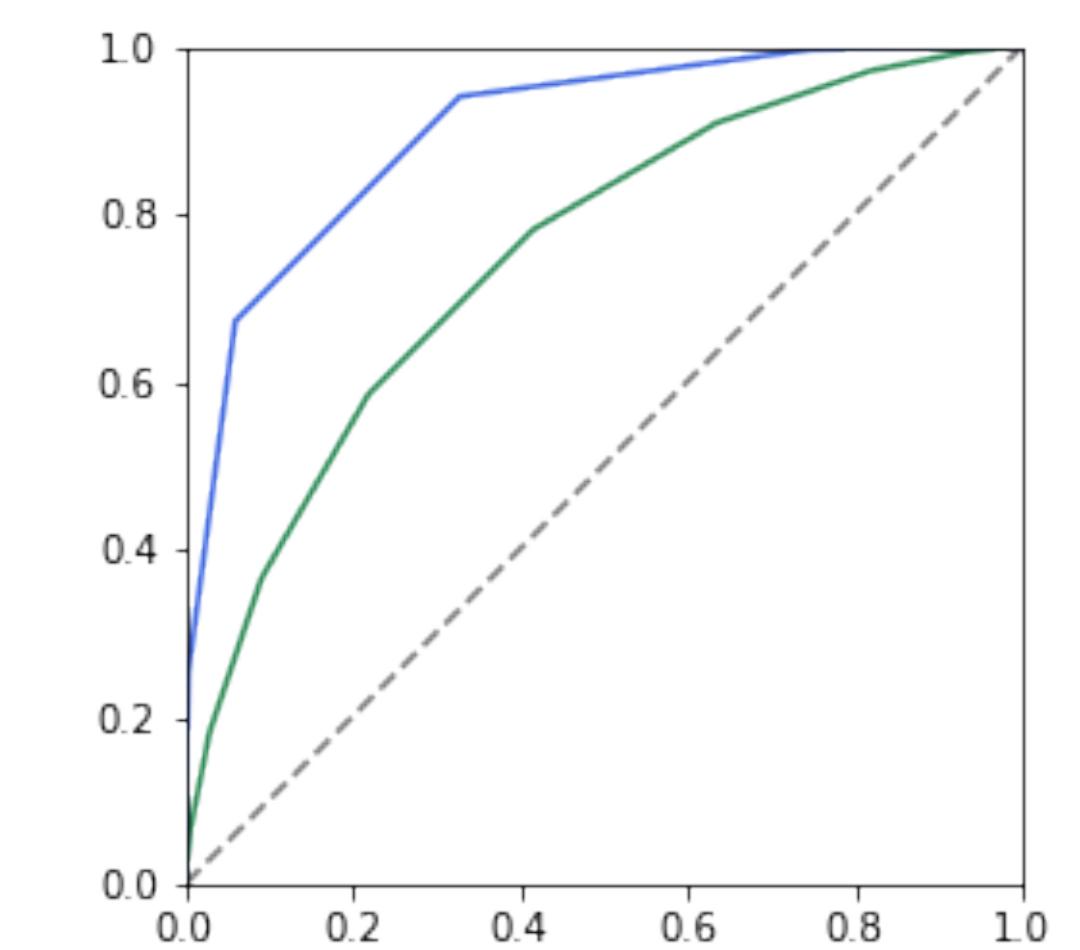
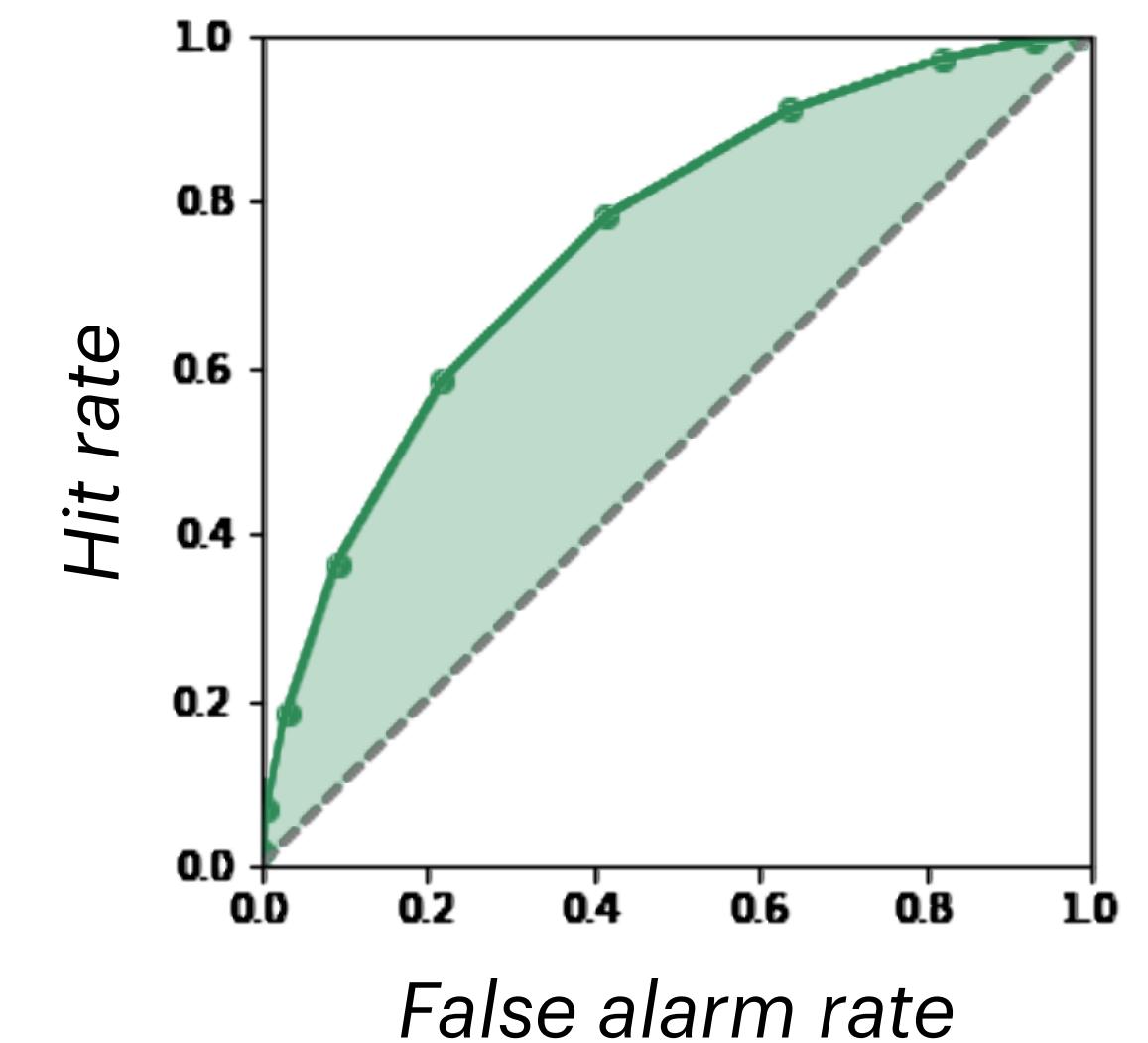
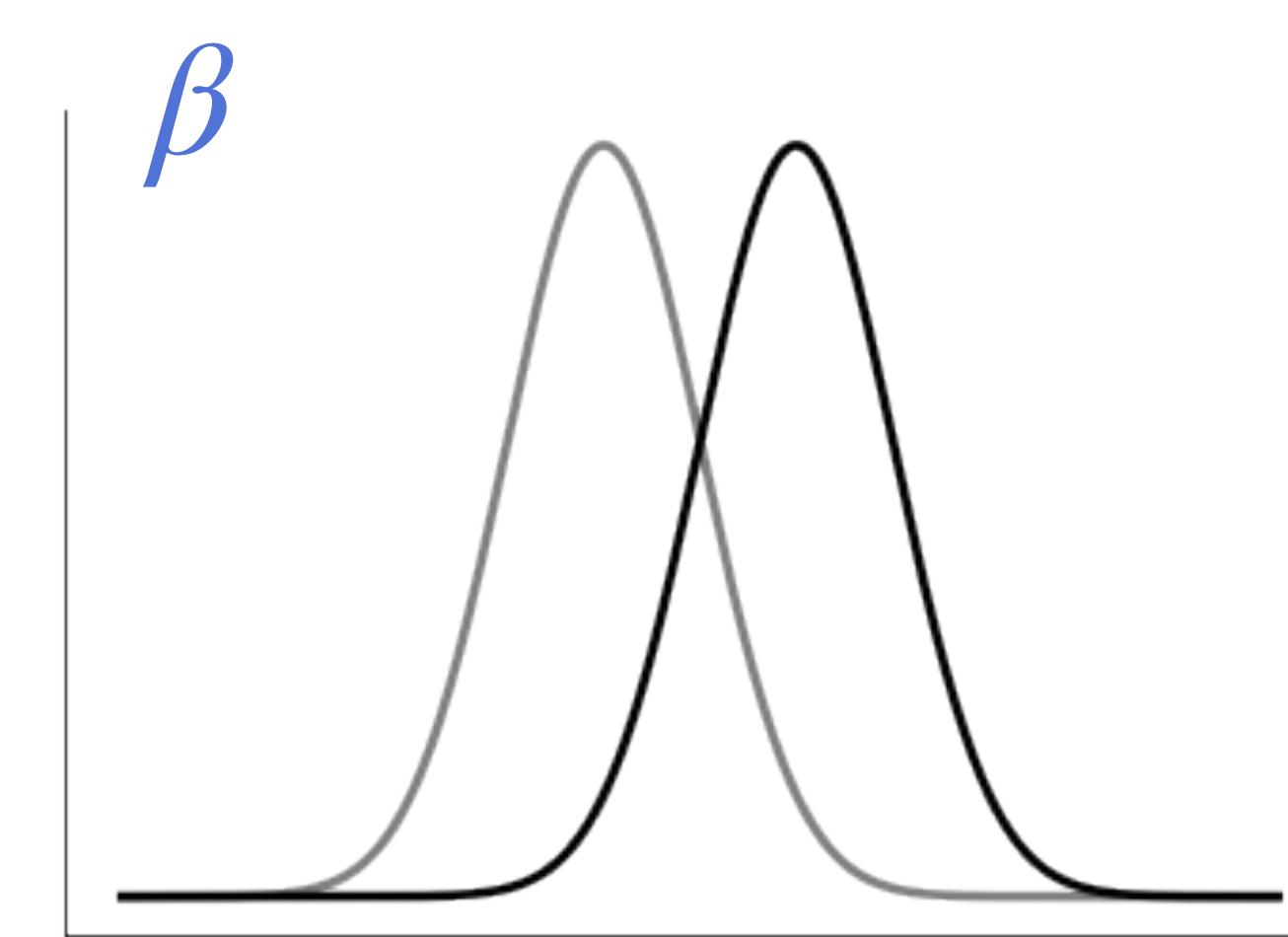
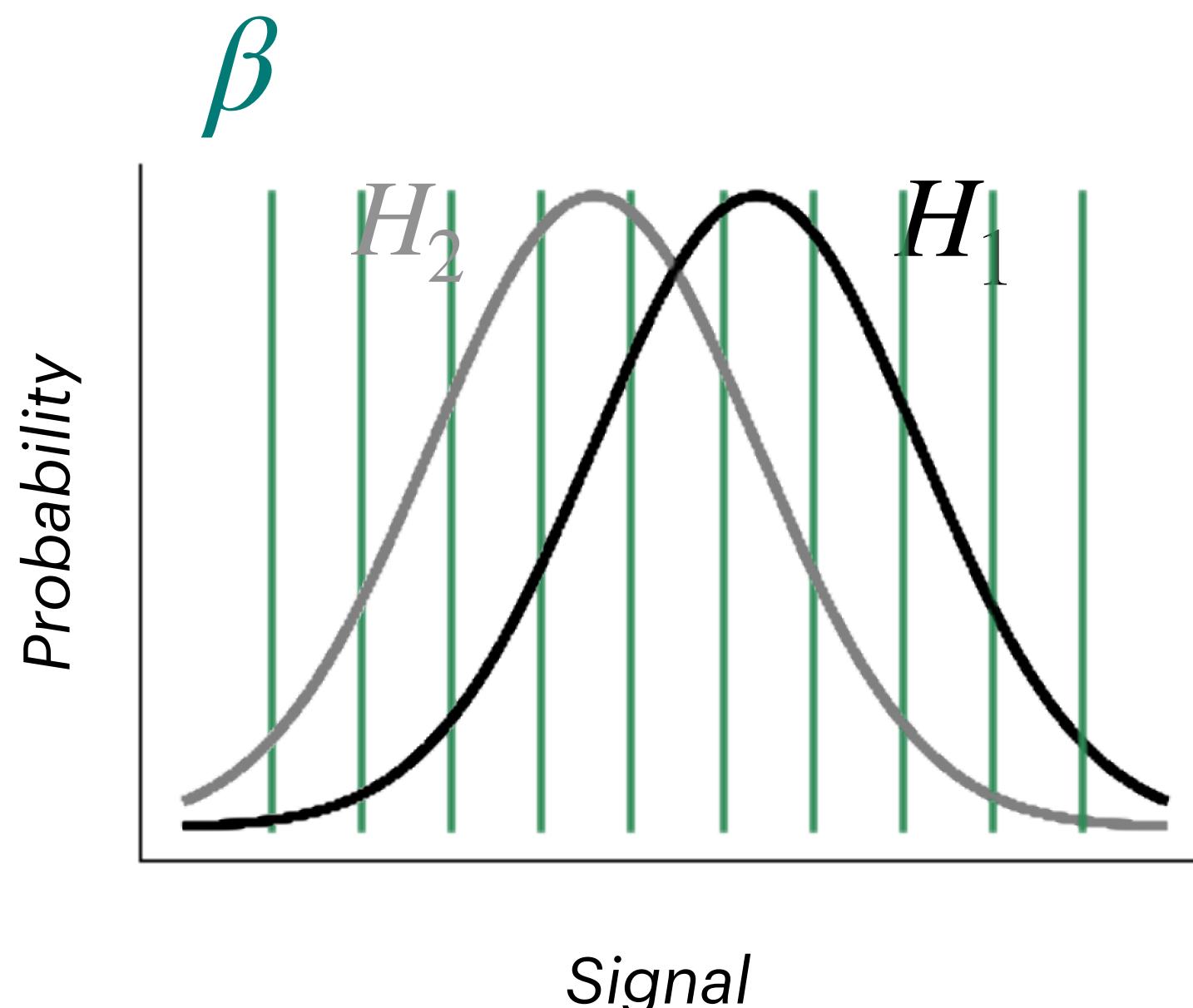
# The receiver operating characteristic (ROC) curve

The ROC curve captures the full  $H_1$  and  $H_2$  by systematically shifting  $\beta$  and asking how the classification probabilities change  
 $P(E \in H_1 | \beta), P(E \in H_2 | \beta)$

The area under the ROC curve corresponds to the classification accuracy

E.g., distributions with the same mean but smaller  $\sigma$  will be pushed towards the top left corner, and indistinguishable distributions will yield curves on the unity line

Note that this is a non-parametric measure (does not assume any shape of the distribution), so it's broadly useful for different types of data



# Discriminability: $d'$

$d'$  is a commonly used summary metric of how discriminable  $H_1$  and  $H_2$  are

$$d' = \frac{\text{separation}}{\text{spread}}$$

Which is given by the difference in mean divided by the pooled standard deviation

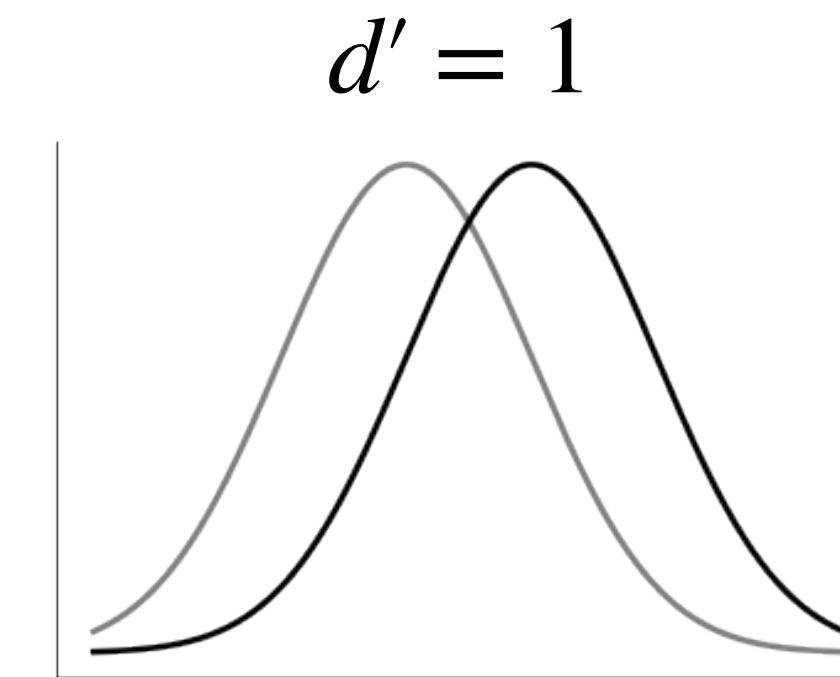
$$d' = \frac{\mu_2 - \mu_1}{\sqrt{(\sigma_1^2 + \sigma_2^2)/2}}$$

Which reduces to

$$d' = \frac{\mu_2 - \mu_1}{\sigma} , \text{ for } \sigma_1 = \sigma_2$$

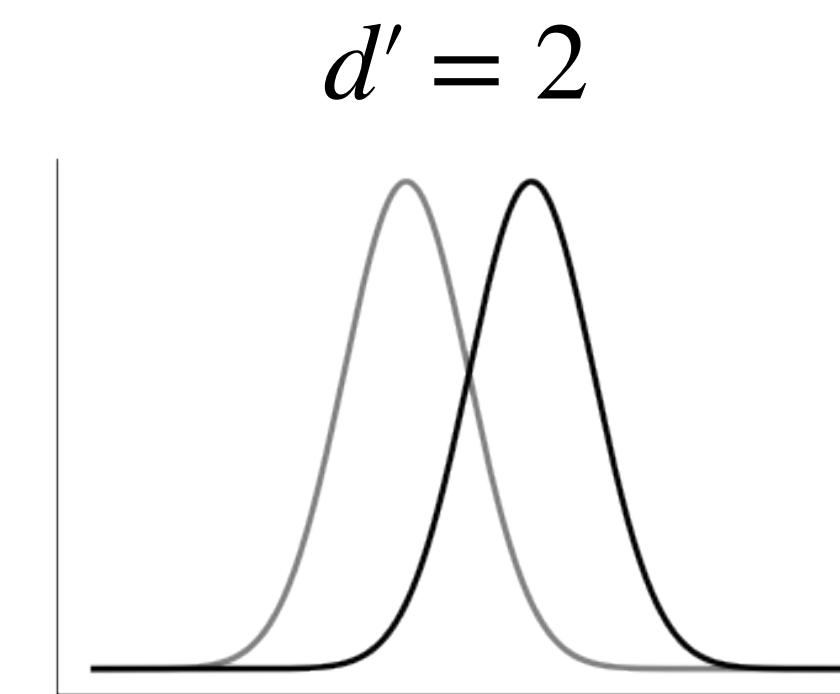
In practice, we usually assume normal data and unit  $\sigma$  and compute the inverse Gaussian (z score)

$$d' = z(HR) - z(FA)$$



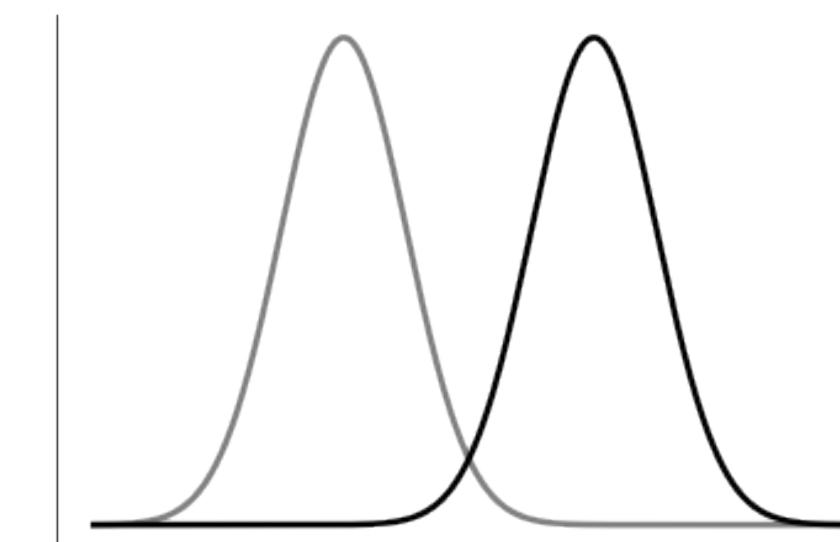
$$d' = 1$$

$$\mu_2 = 1, \mu_1 = -1, \sigma = 2$$



$$d' = 2$$

$$\mu_2 = 1, \mu_1 = -1, \sigma = 1$$



$$d' = 4$$

$$\mu_2 = 2, \mu_1 = -2, \sigma = 1$$

# Perceptual decision making: Sequential sampling models

# Sequential sampling

Sequential sampling is a generalization of SDT where we take multiple (independent) discrete samples  $E_i$  to decide if they come from  $H_1$  or  $H_2$ . In this case, the likelihood ratio becomes

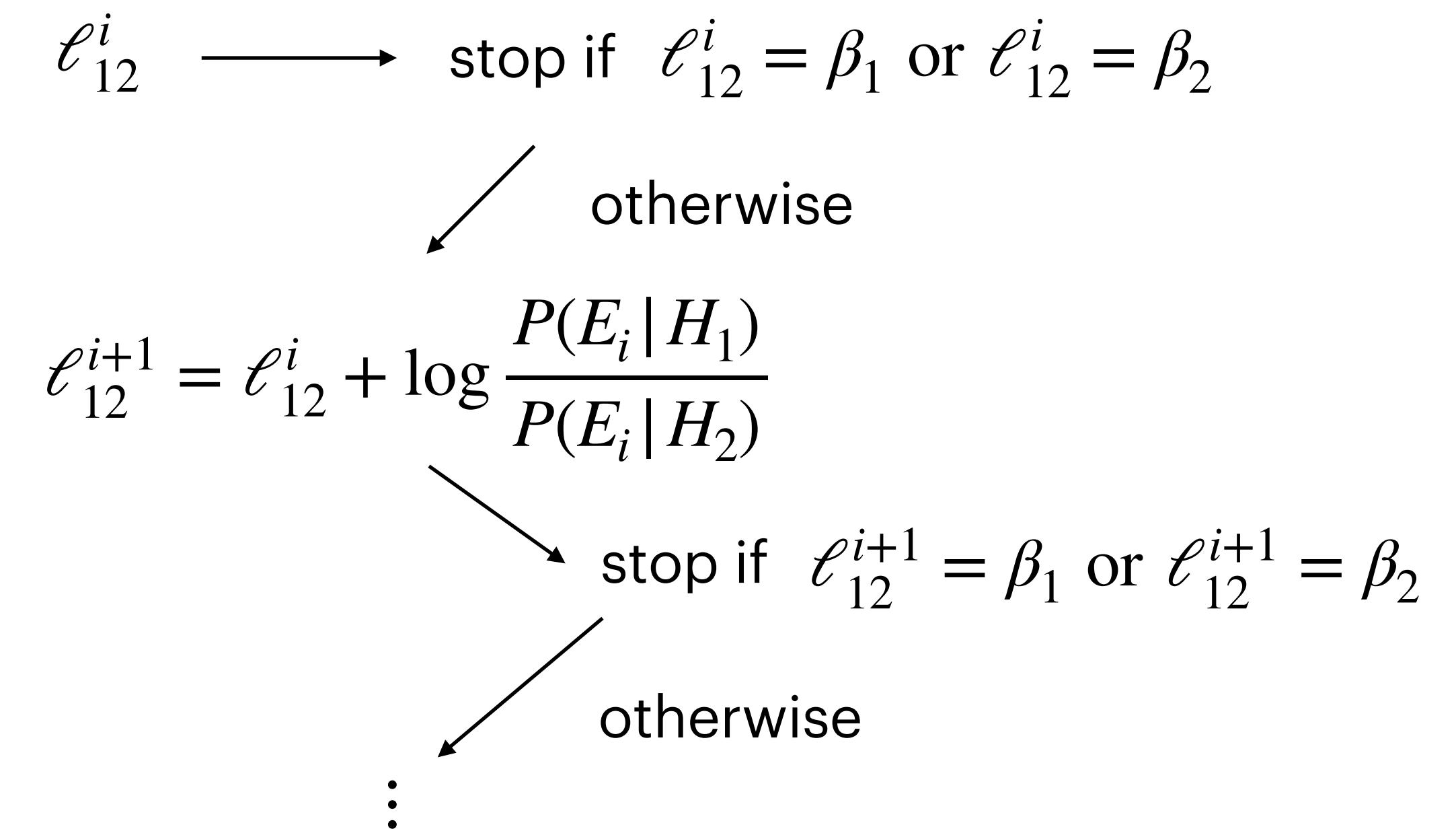
$$L_{12}(E_1, E_2, \dots, E_n) = \frac{P(E_1, E_2, \dots, E_n | H_1)}{P(E_1, E_2, \dots, E_n | H_2)} = \prod_{i=1}^n \frac{P(E_i | H_1)}{P(E_i | H_2)}$$

Or, for mathematical convenience, the log likelihood ratio

$$\ell_{12} = \sum_{i=1}^n \log \frac{P(E_i | H_1)}{P(E_i | H_2)}$$

The log likelihood ratio is bound by two thresholds  
 $\beta_2 \leq \ell_{12} \leq \beta_1$

This is such that with the addition of each sequential sample we approach one of them, and make a decision when we hit either threshold



This is known as the *sequential probability ratio test*. It can be shown to be optimal in the sense that, for a given threshold, it gives the best accuracy. But with *speed-accuracy tradeoff* (more on this soon).

# DDMs: continuous sequential sampling

Drift diffusion models are essentially the continuous-time limit of sequential sampling

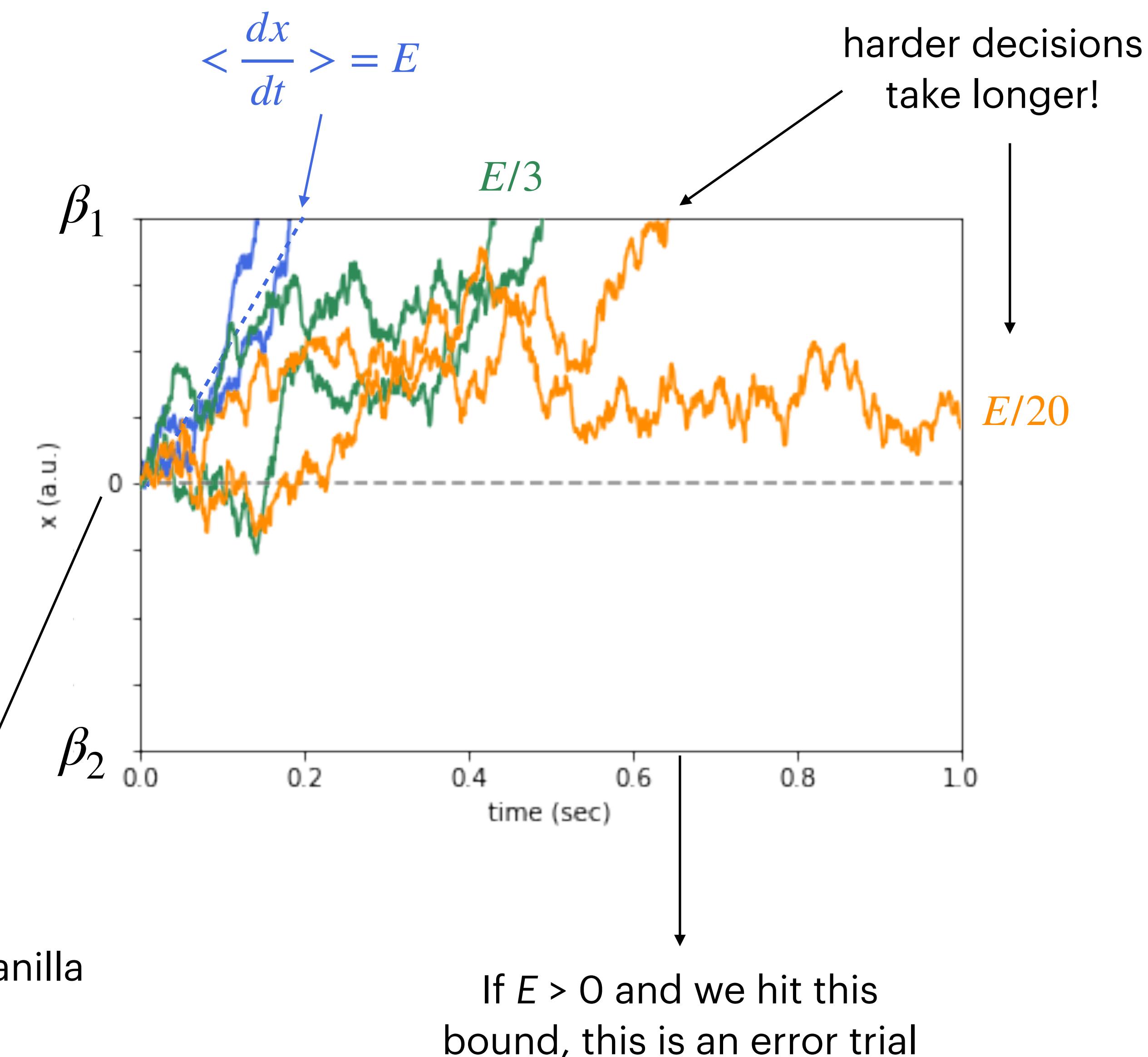
$$dx = Edt + cdW$$

latent decision variable  
(accumulated evidence)

**“Drift”:**  $E$  dictates how quickly  $x$  moves towards the decision threshold (a.k.a. bound). Can be thought of as the strength of sensory evidence

Wiener process,  
a.k.a. brownian motion, a.k.a.  
**“diffusion”:**  
Independent gaussian noise  $\sim \mathcal{N}(0, cdt)$

$x$  starts at 0 in vanilla DDMs



# Decisions with DDMs

Again, a vanilla DDM is

$$dx = E dt + c dW, \quad x_{t=0} = 0$$

Ignoring decision thresholds for now, the probability density of  $x$  at time  $t$  is

$$p(x, t) = \mathcal{N}(Et, c\sqrt{t})$$

If I need to make a decision between  $H_1$  and  $H_2$  at time  $T$ , if  $x > 0$  I decide  $H_1$

The average error rate,  $ER$ , at time  $T$  is the probability that  $x < 0$ , which we get by integrating the probability density

$$ER = \Phi\left(-\frac{E}{c}\sqrt{T}\right), \quad \Phi(y) = \int_{-\infty}^y \frac{1}{\sqrt{2\pi}} e^{-(x^2/2)} dx$$

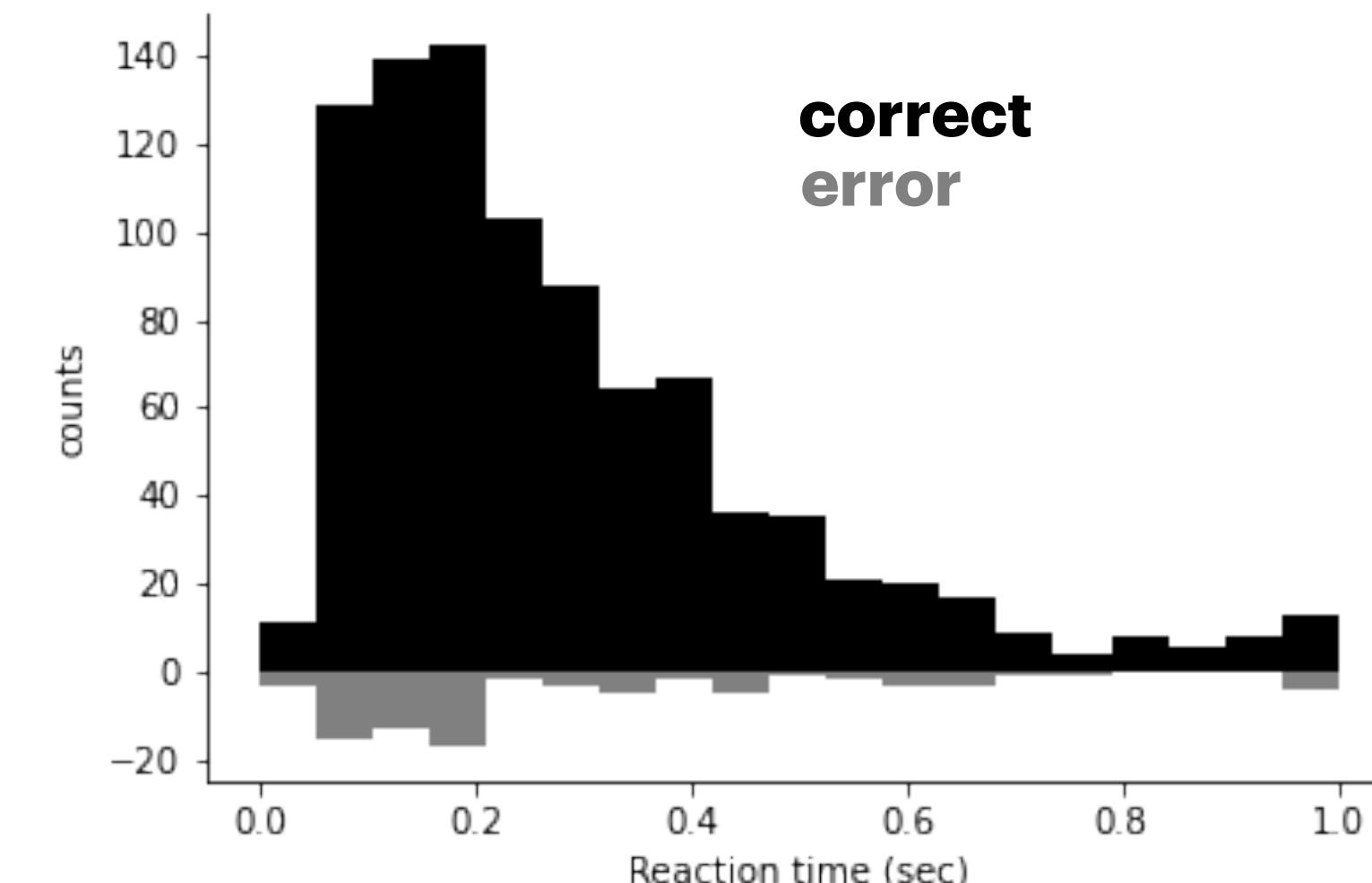
For full derivation see Bogacz et al, 2006

In a free response task, we stop deciding when  $x$  equals one of the thresholds  $\beta$ , with decision (reaction) time  $RT$ . In this case

$$ER = \frac{1}{1 + e^{\frac{2E\beta}{c^2}}} \quad RT = \frac{\beta}{E} \tanh\left(\frac{E\beta}{c^2}\right)$$

higher thresholds yield smaller ER but longer RT: speed-accuracy tradeoff

This probabilistic nature of the model captures well reaction time distributions during perceptual decision making, including in error trials



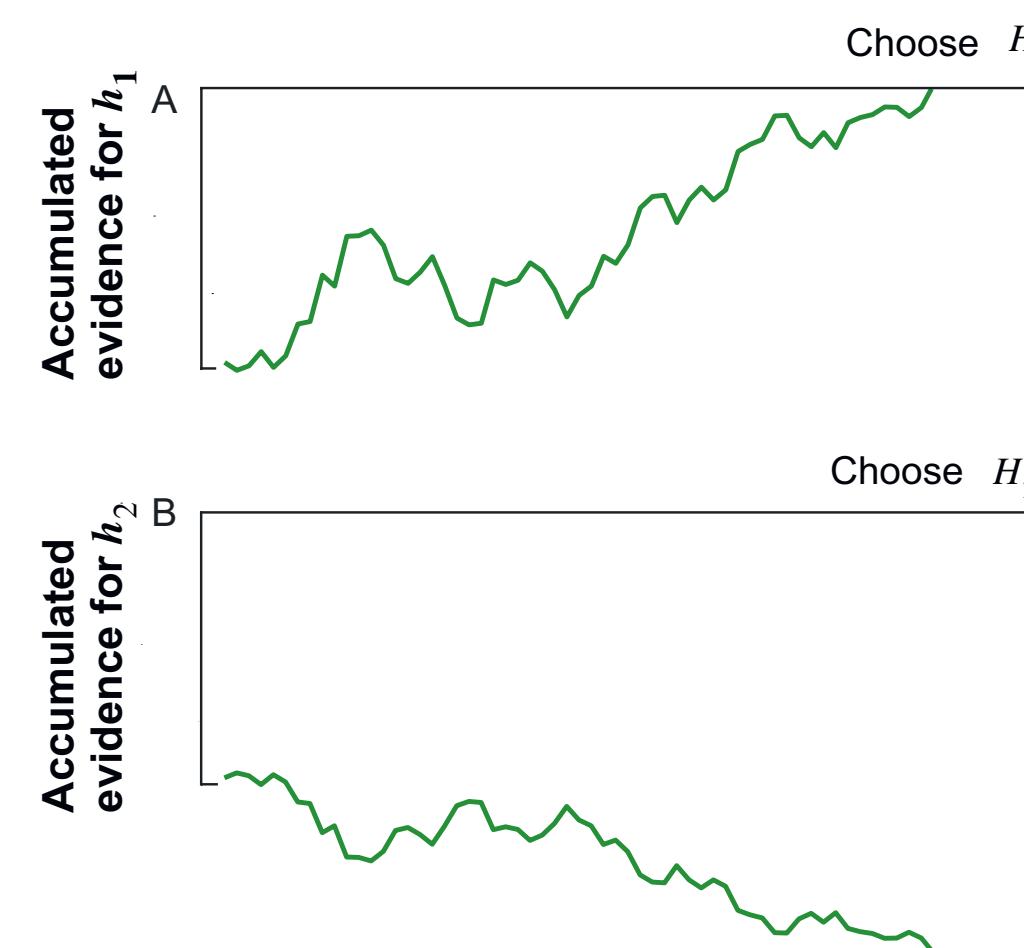
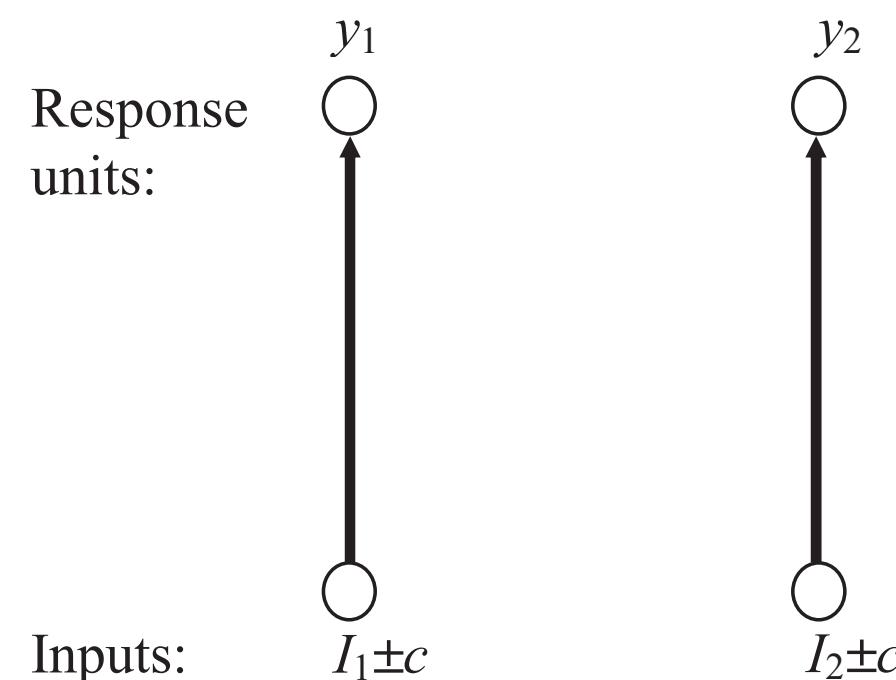
# DDM variations

## 1. Vanilla DDM modifications.

- Starting point  $x_{t=0}$  can become a parameter, e.g. to capture bias
- Making  $E$  stochastic (i.e. the mean of a Gaussian, drawn on each trial) better captures RTs
- Time-varying bounds can capture different strategies, e.g. primacy/recency

## 2. Race models.

- Two separate accumulators for  $H_1$  or  $H_2$ , decision is which one finishes first
- Do not reduce to DDMs
- Struggle to capture RT distributions, esp. in error trials

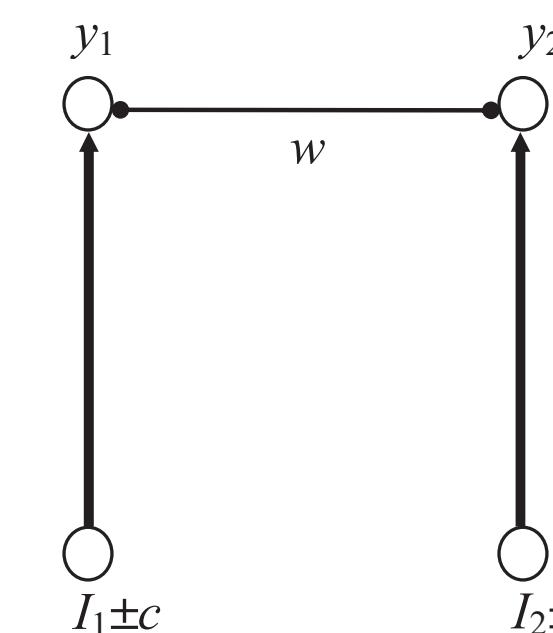


## 3. Leaky integrators with inhibition.

- Related to the Ornstein-Uhlenbeck model (next)
- Connectionist; inspired by circuit wiring diagrams
- Reduce to DDMs under certain parameter regimes

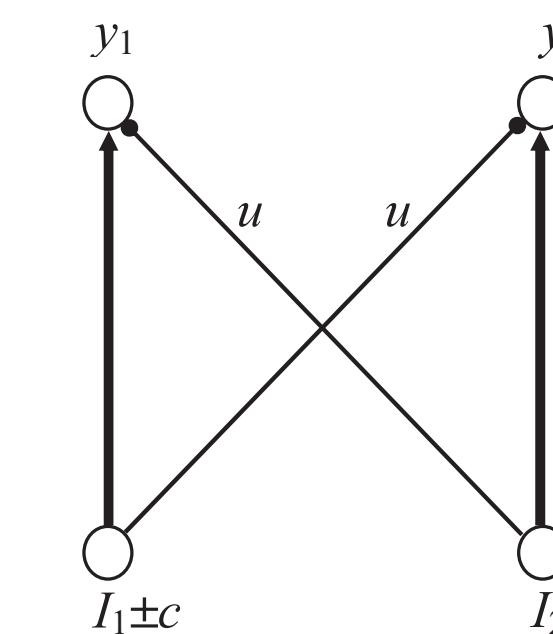
### Mutual inhibition

(Usher & McClelland, 2001)



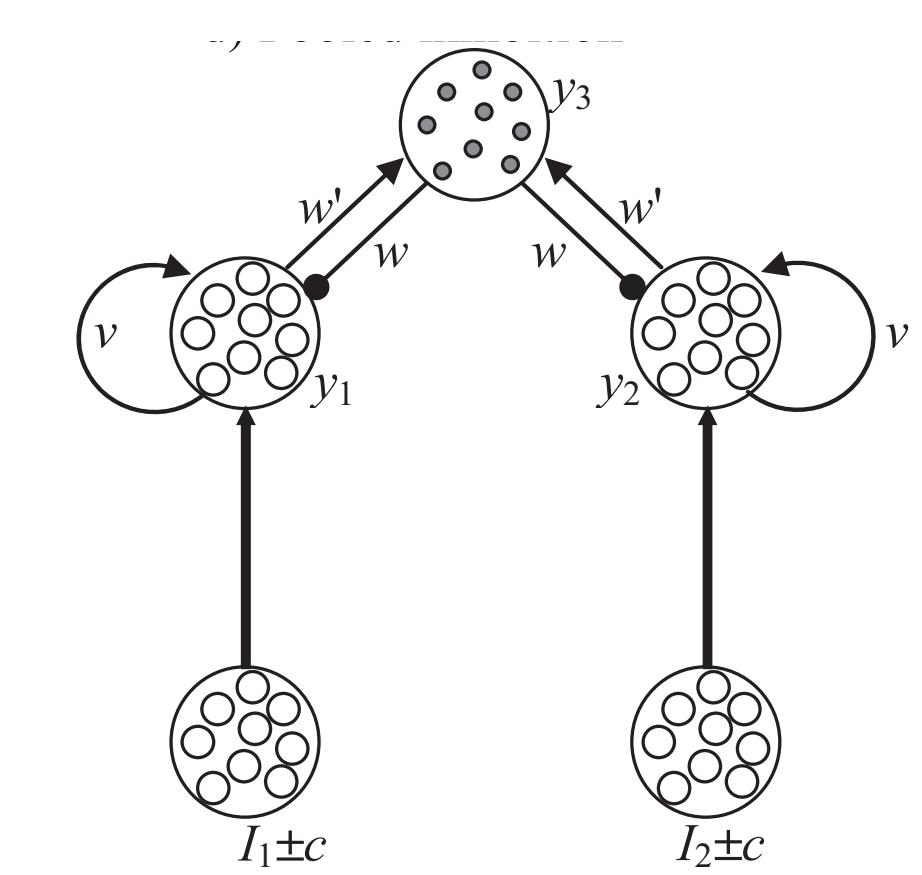
### Feedforward inhibition

(Shadlen & Newsome, 2001)



### Pooled inhibition

(Wang, 2002)



Bogacz et al, 2006

Gold & Shadlen, 2007

# The Ornstein-Uhlenbeck (O-U) model

O-U models have an additional term that make  $dx$  also depend on its current value, with magnitude controlled by  $\lambda$

$$dx = (\lambda x + E)dt + c dW, \quad x_{t=0} = 0$$

Depending on the sign of  $\lambda$ ,  $x$  accelerates or decelerates towards the bound

$\lambda < 0$  : growing  $x$  leads to decreasing  $dx/dt$

$\lambda > 0$  : growing  $x$  leads to increasing  $dx/dt$

This is said to be an *unstable integrator* (i.e. evidence progressively leads to “impulsive” decisions)

$\lambda = 0$  : vanilla DDM

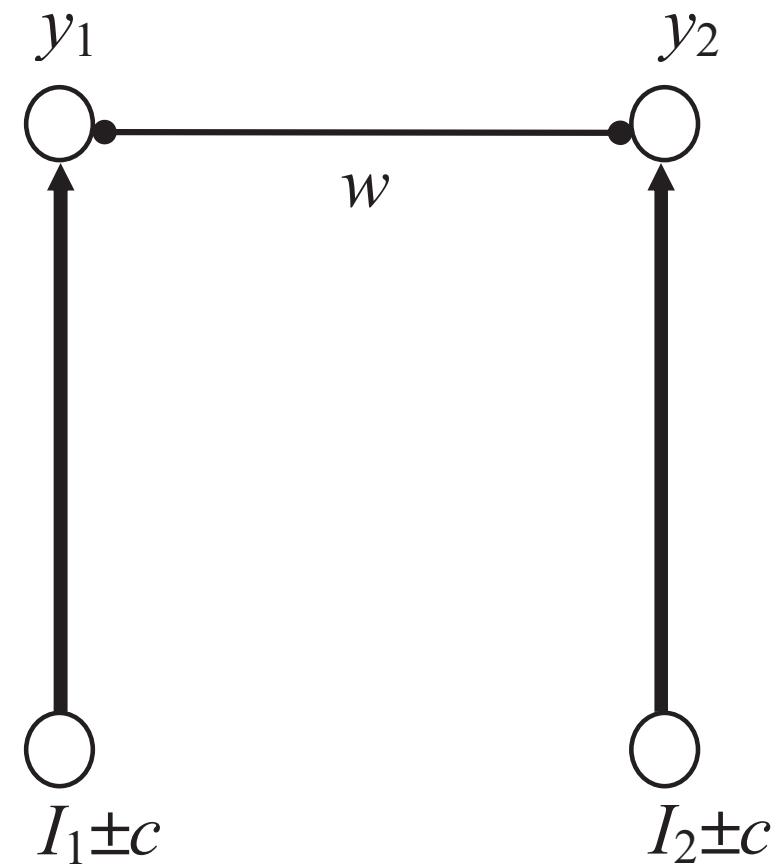
The time-dependent probability density converges to

$$p(x) = \mathcal{N}\left(-\frac{E}{\lambda}, \frac{c}{\sqrt{-2\lambda}}\right)$$

This is said to be a *leaky integrator* (i.e. evidence is lost over time)

# Mutual inhibition model

Mutually inhibiting units:



$$\begin{cases} dy_1 = (-ky_1 - wy_2 + I_1)dt + cdW_1 \\ dy_2 = (-ky_2 - wy_1 + I_2)dt + cdW_2 \end{cases}, y_1^{t=0} = y_2^{t=0} = 0$$

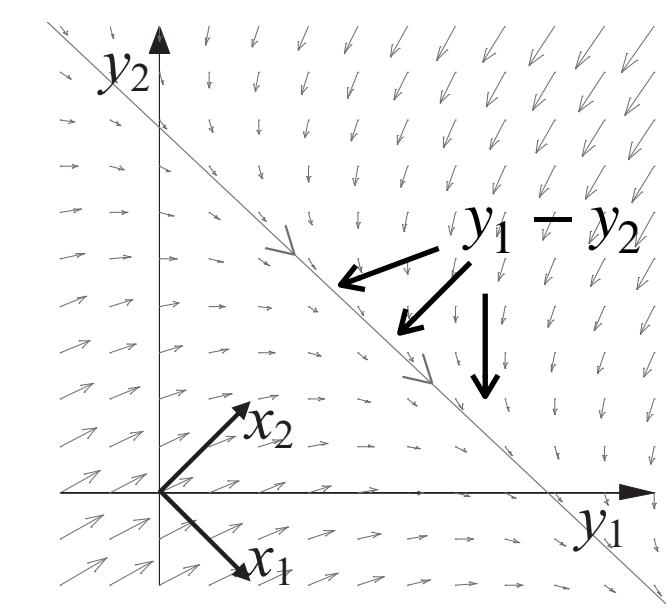
↓  
 Activity decay  
 ↓  
 Inhibition from other unit

↓  
 Input (evidence)

**Model state space** ( $I_1 > I_2$ )

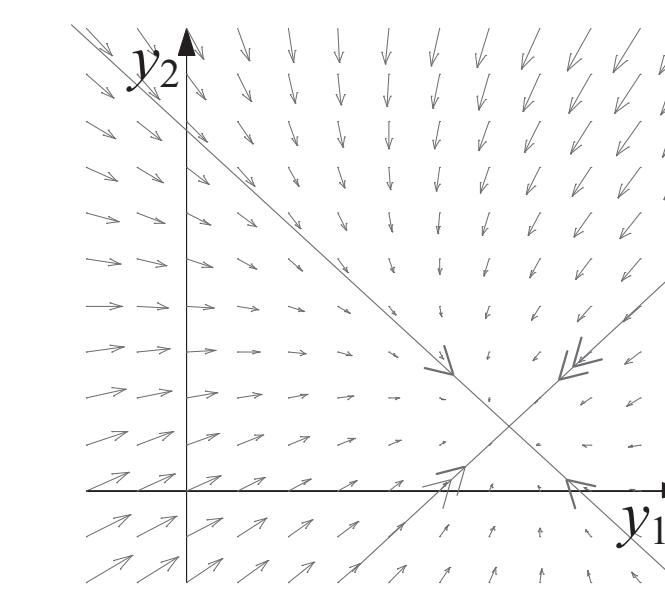
It can be shown that this behaves like an O-U model with  $\lambda = w - k$

$$k = w \equiv \lambda = 0$$



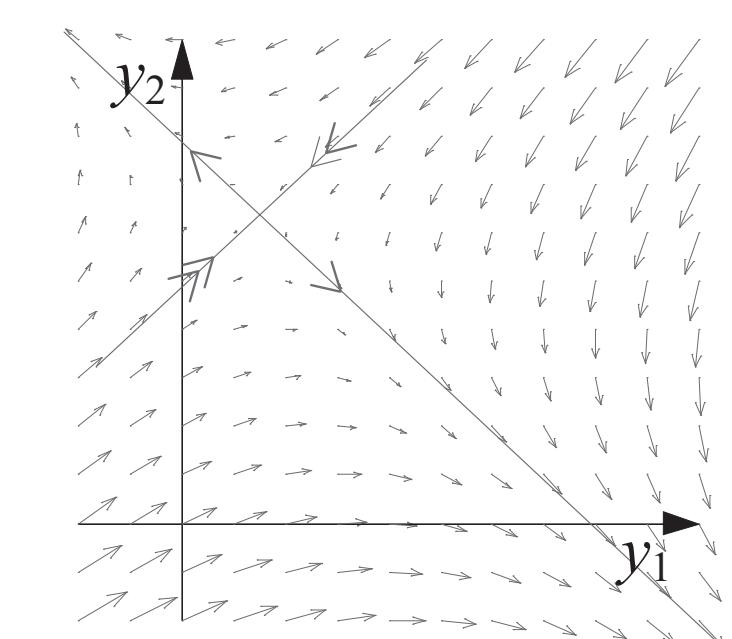
Line attractor,  
Equivalent to DDM

$$k > w \equiv \lambda < 0$$



Attracts towards  
fixed point,  
Equivalent to  
leaky integrator

$$k < w \equiv \lambda > 0$$



Repelled from  
fixed point,  
Equivalent to  
unstable integrator

- ▶ Note that this is a linear approximation of the original model, where activity is rectified

# Fitting DDMs

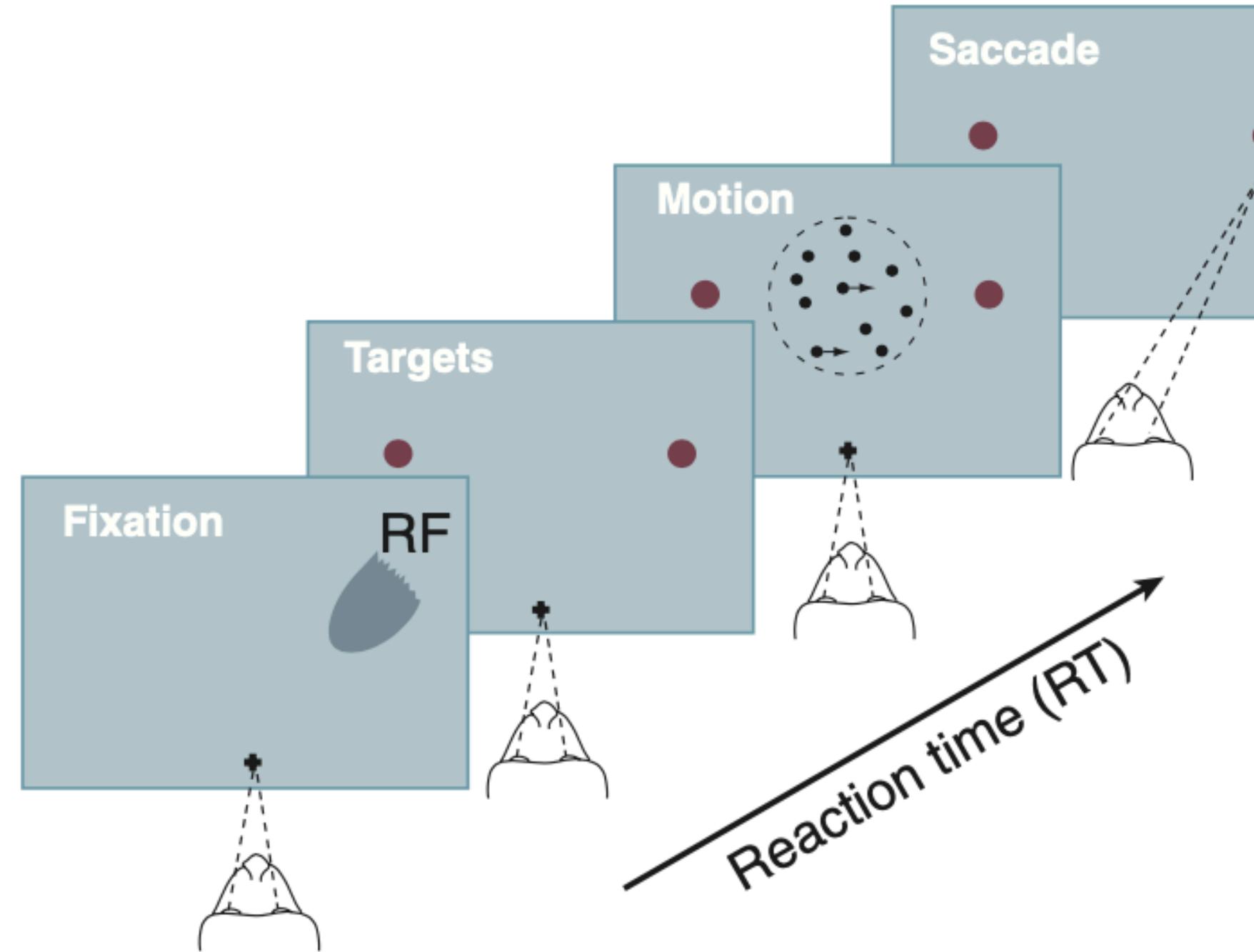
Vanilla DDMs have known analytical solutions

Extended DDMs and other variants don't, we need to iteratively propagate the distributions forward in time, usually by solving the Fokker-Planck equation, which describes the temporal evolution of the probability density function  $p$

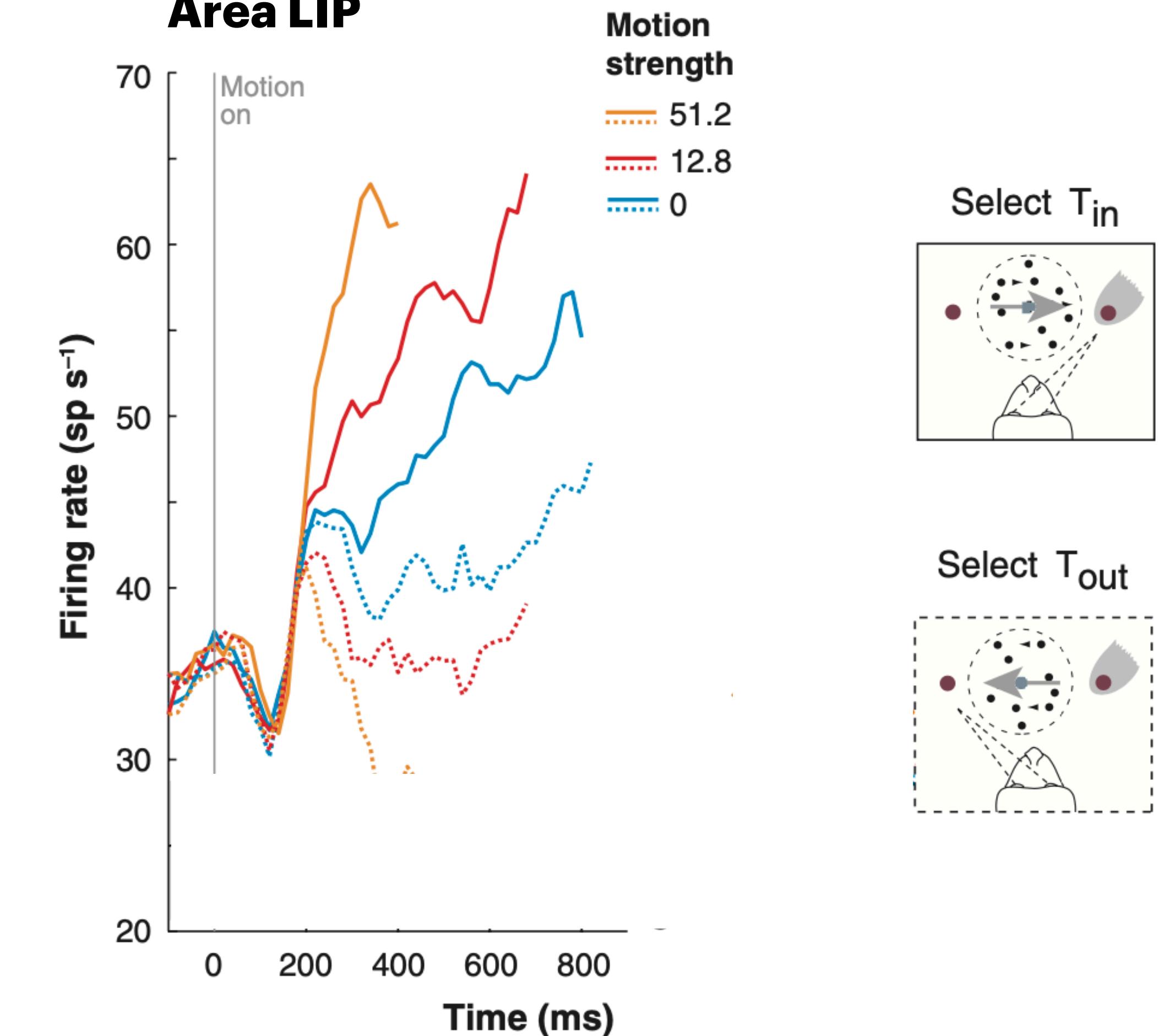
$$\frac{\partial}{\partial t} p(x, t) = - \frac{\partial}{\partial x} [E(t)p(x, t)] + \frac{\partial^2}{\partial x^2} \left[ \frac{c^2(t)}{2} p(x, t) \right]$$

# DDM and neural dynamics

## Random dot motion task



## Area LIP

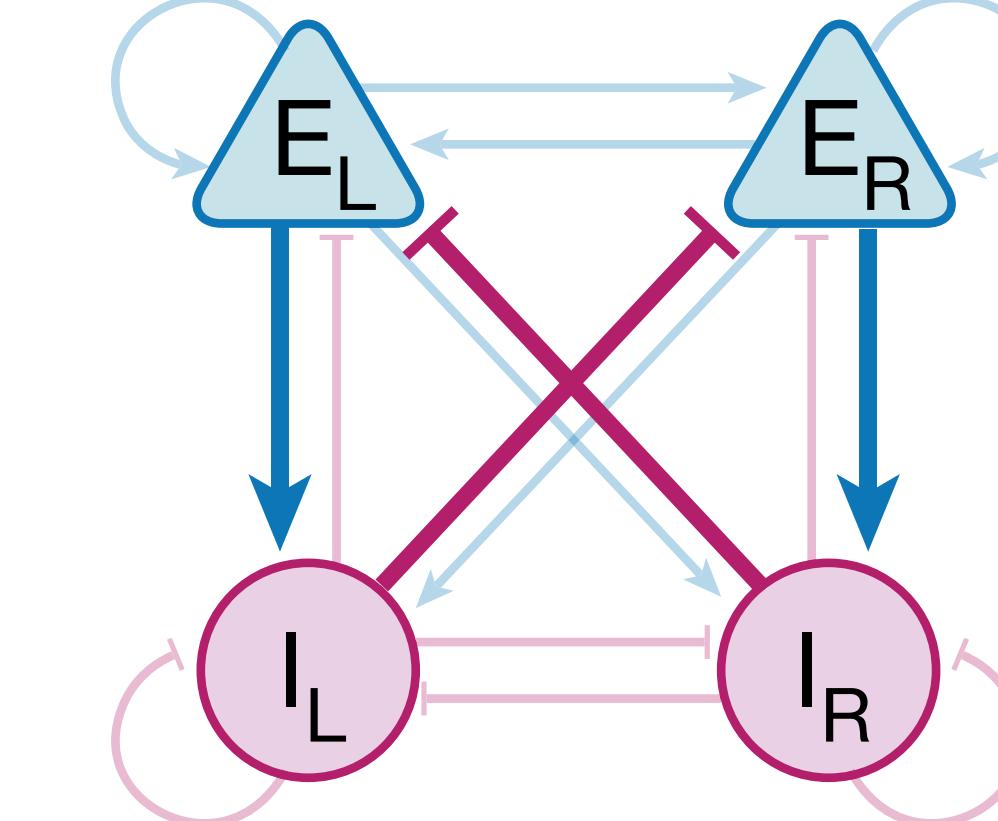
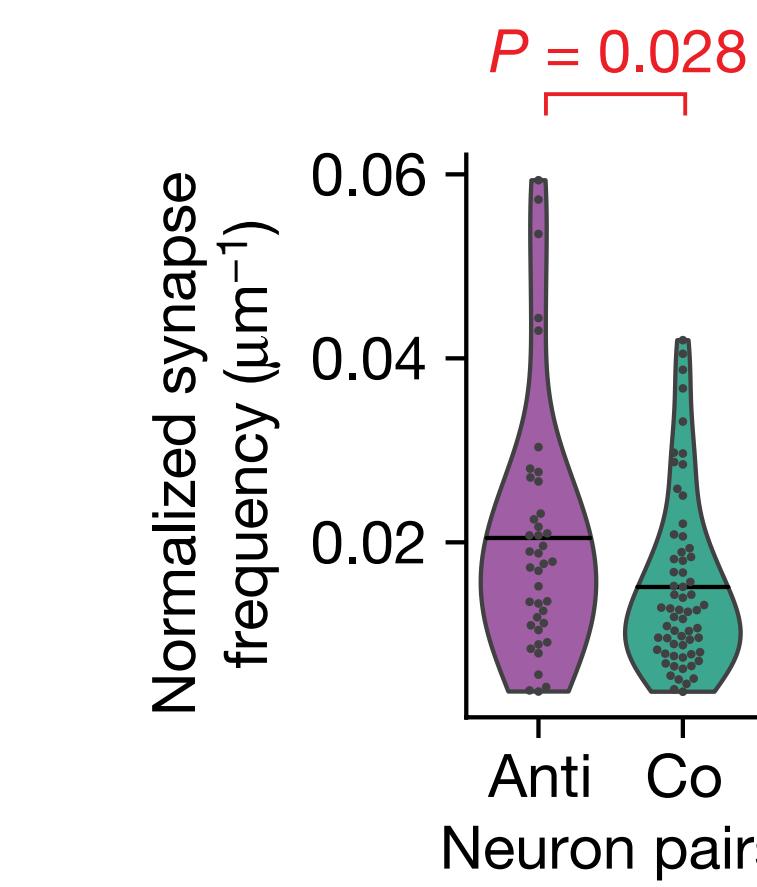
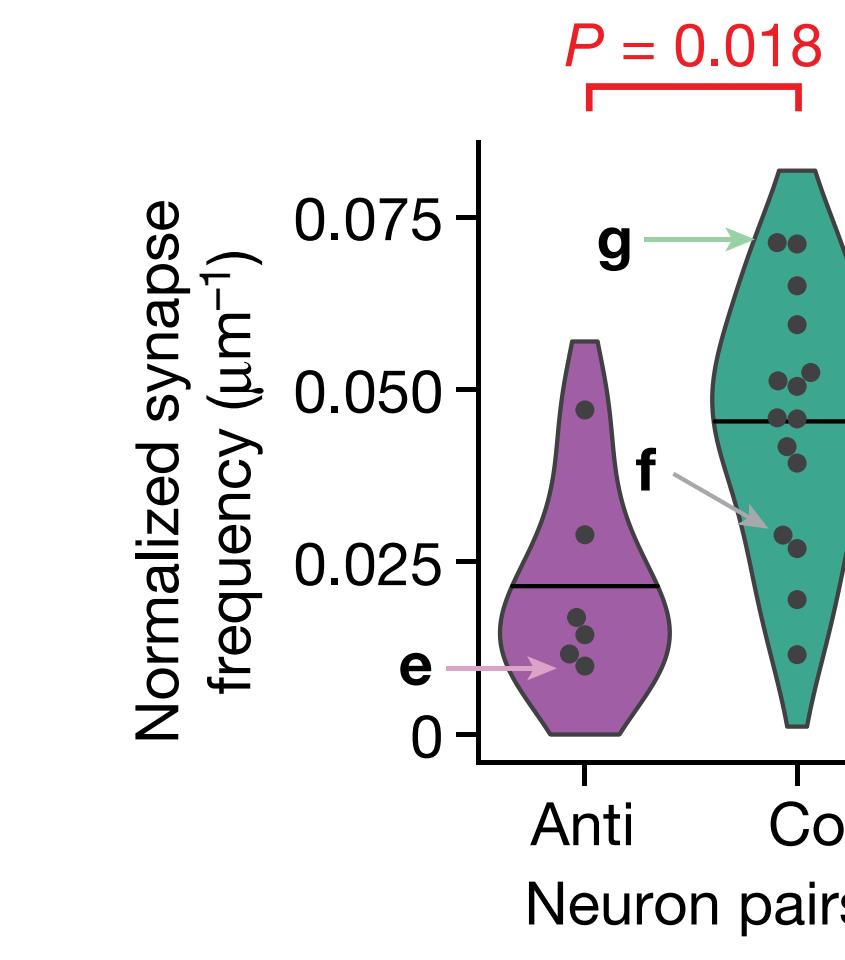
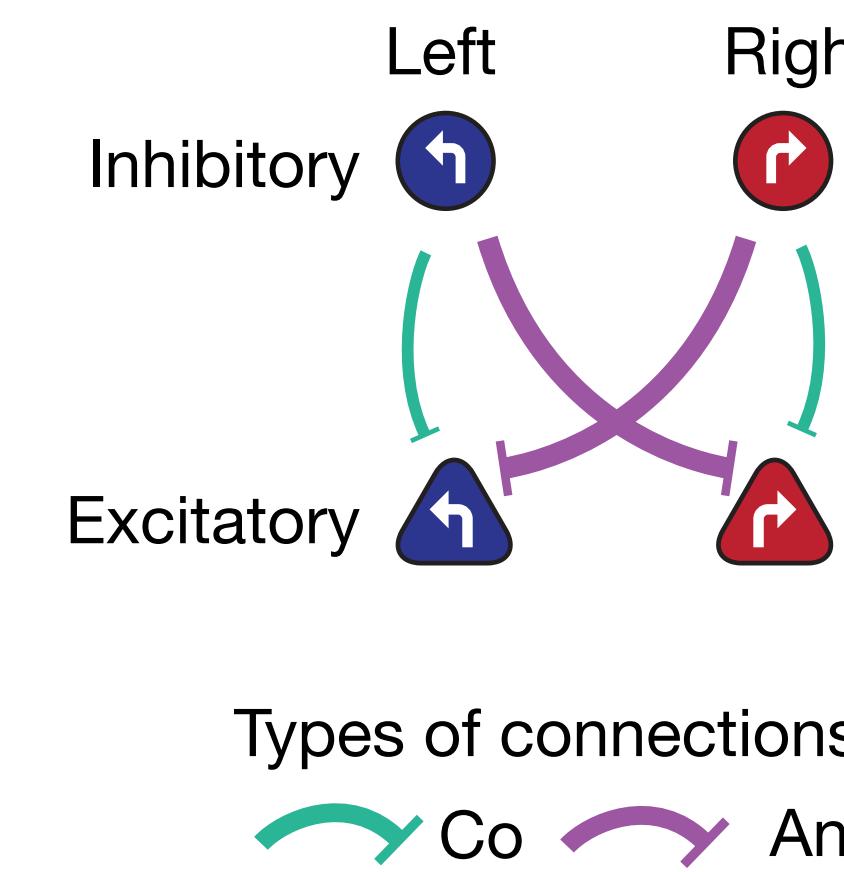
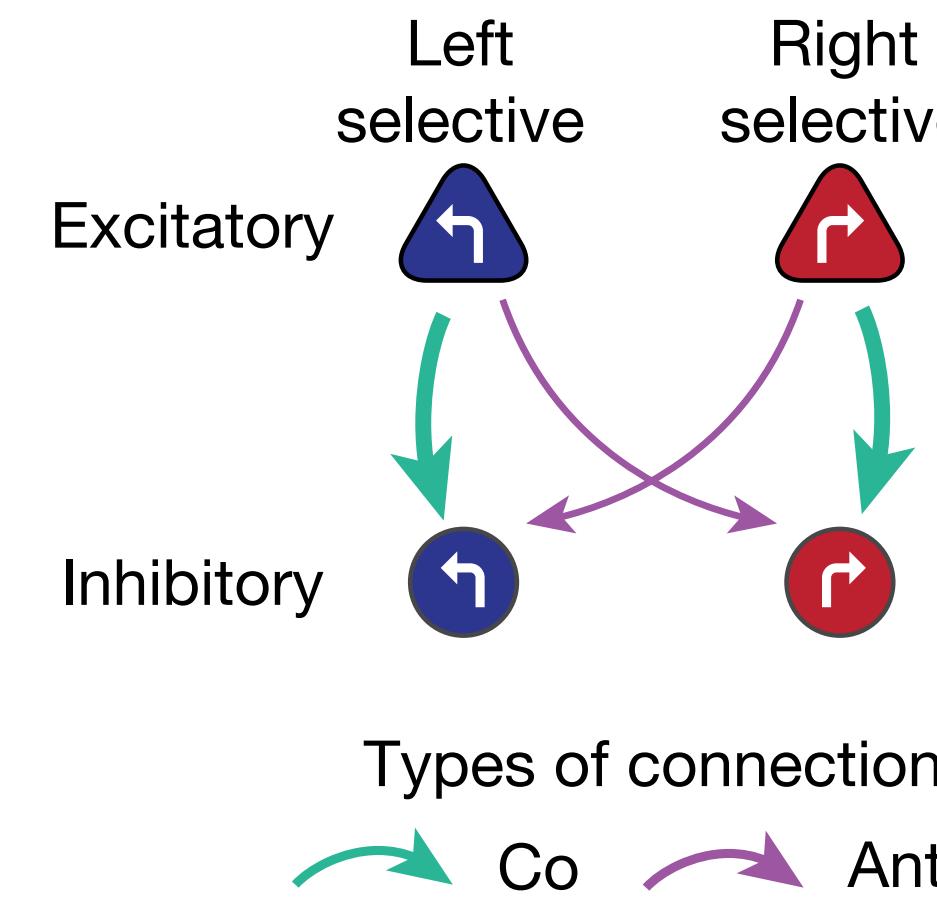
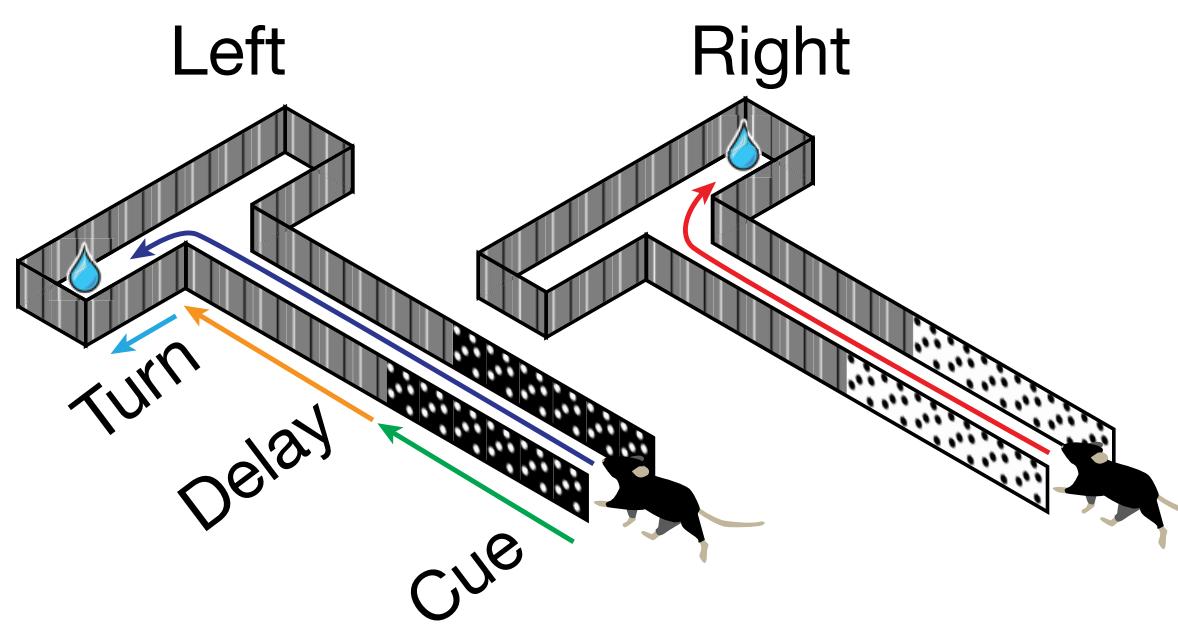


- ▶ These average ramping responses have been also observed in a number of other cortical and subcortical structures, suggesting distributed computation
- ▶ Single-trial population dynamics appear more complicated than DDM-like ramps

Shadlen, Newsome, Britten, Gold ... c.f.  
Gold & Shadlen, 2007, *Annu Rev Neurosci*

See also: Hanks & Brody, 2016, *Curr Opin Neurobiol*  
But see, e.g.: Latimer ... Pillow, 2015, *Science*

# Mutual inhibition and cortical wiring



# Value-based decision making

# A few key concepts

A *reward* is an event that satisfies a motivational drive

- It produces a positive affective experience
- It is salient; attracts attention
- It drives learning:
  - Pavlovian: *prediction*. Learn about events that predict reward
  - Operant: *positive reinforcement*. Increase the frequency of actions that lead to reward

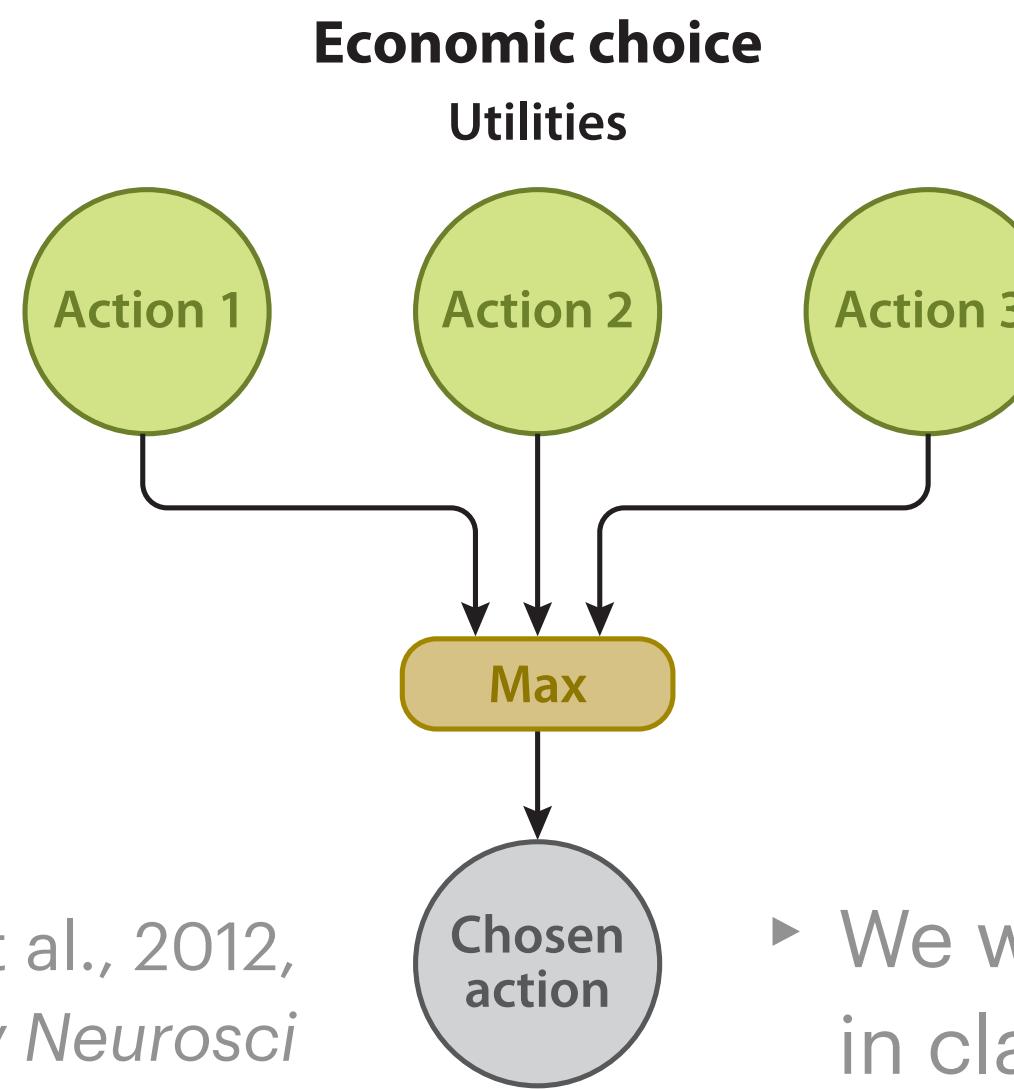
A *punishment* is an event that generates withdrawal or avoidance

- It produces a negative affective experience
- It is salient; attracts attention
- It drives learning:
  - Pavlovian: *prediction*. Learn about events that predict punishment
  - Operant: *negative reinforcement*. Decrease the frequency of actions that lead to punishment

*Cost* is a loss or expenditure associated with an action, e.g.: metabolic (effort), opportunity (time, forgone alternatives)

Loosely speaking, value-based decision making refers to decisions that are based on expected reward, punishment and cost

# Economic decisions vs. RL



► We won't cover these  
in class any further

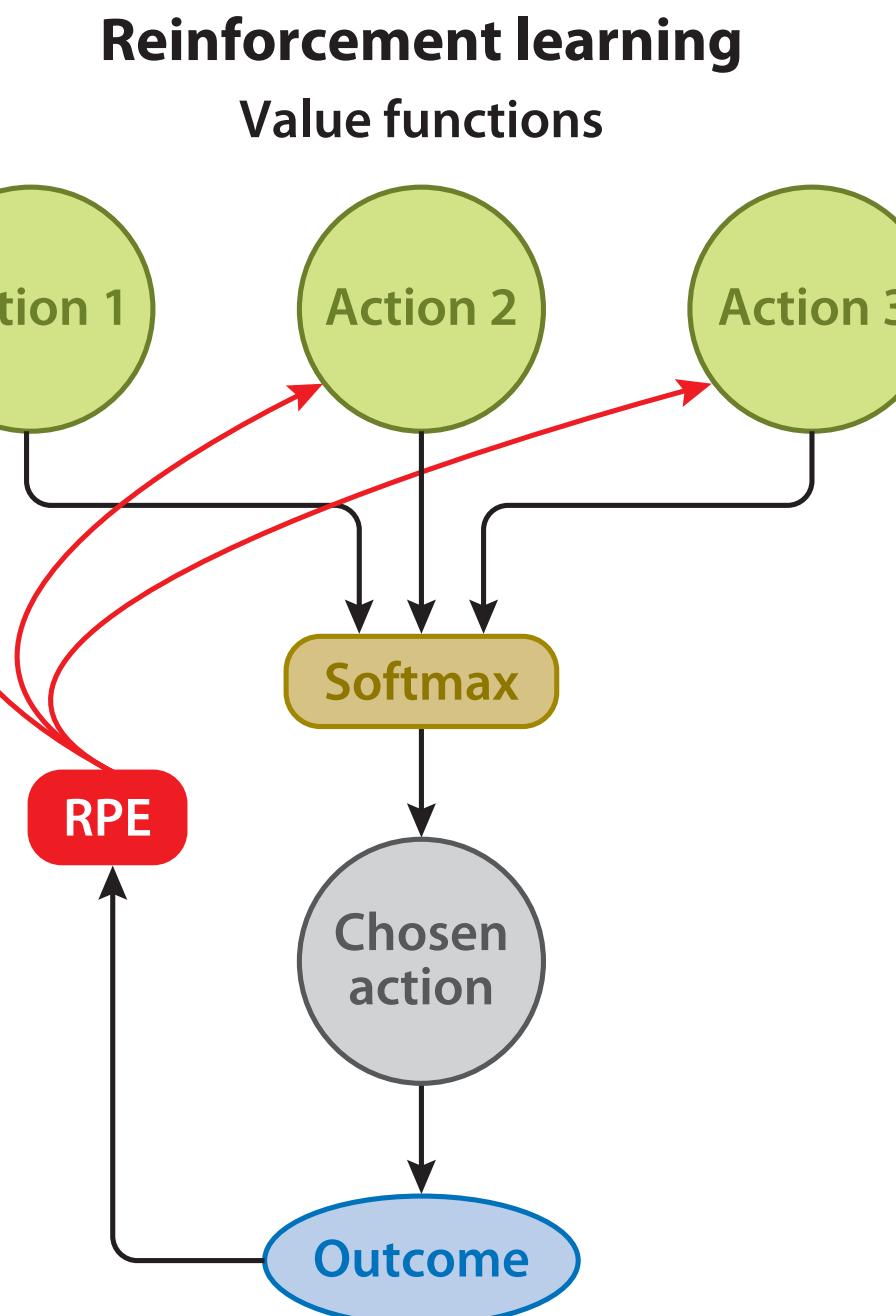
Economic decisions are made to maximize **utility**,  
defined as the total satisfaction an animal derives from  
a state (action). Utility is therefore **subjective**.

**Prospect theory** has been successful in describing  
these kinds of decisions

For details see Kahneman & Tversky, 1979; Padoa-Schioppa, 2011

**Foraging** is a ubiquitous form of economic decision in  
which animals sequentially decide to accept or reject  
offers based on utility (value) and cost. **Optimal foraging  
theory** is a normative framework describing this.

For details see Pulliam, 1974; Calhoun & Hayden, 2015



In RL, probabilistic decisions are made to maximize  
**value**,  $V$ , defined as the expected sum of future rewards  $r$   
for a given **policy**  $\pi$ , which maps a **state**  $S$  to an **action**  $A$

$$V_\pi = E \left[ \sum_{t=0}^{\infty} r_t \right], \pi = S \times A = P(A = a | S = s)$$

The policy is updated by a **Reward Prediction Error**

# RL: the framework

Classic RL is a Markovian process where an agent transitions between different states  $s$  depending on their action  $a$  in  $s$

$$S = \{s_0, s_1, \dots, s_n\}$$

$$A = \{a_0, a_1, \dots, a_n\}$$

$$P_a(s, s') = P(S_{t+1} = s' | S_t = s, A_t = a)$$

where  $a$  depends on a policy

$$\pi(s, a) = P(A_t = a | S_t = s)$$

and retrieve some reward  $r$  immediately after this transition

$$r_a(s, s') = P(A_t = a | S_t = s)$$

Each state is associated with a value function  $V$

$$V_\pi(s) = E[G_0 | s_0 = s] = E \left[ \sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s \right], \quad 0 \leq \gamma^t \leq 1$$

$G$  is thus the discounted reward function given that we started in state  $s$ , where typically  $\gamma$  decreases with  $t$ , i.e. distant future rewards have less weight on  $V$ .

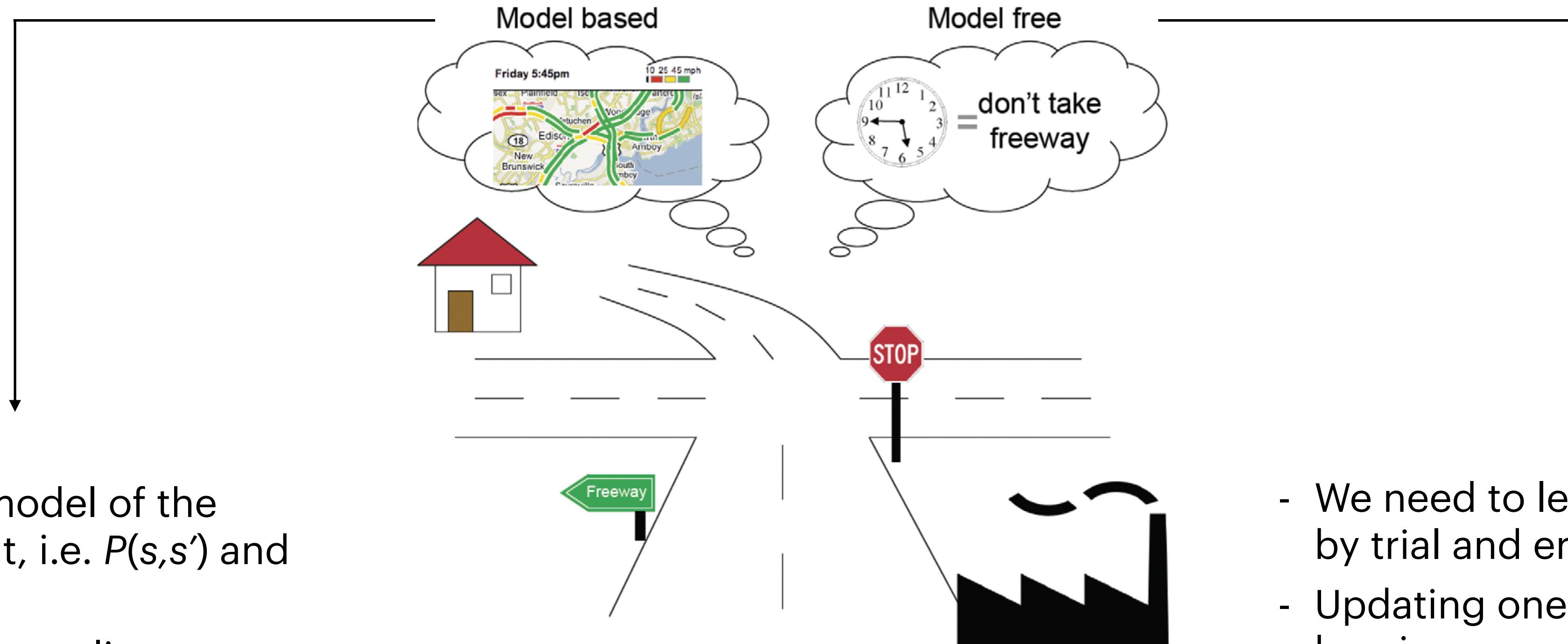
Similarly, each action within a state is also associated with a value function  $Q$

$$Q_\pi(s, a) = E \left[ \sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s, a_0 = a \right]$$

Our goal is to optimize our policy to maximize action value

$$\pi_*(s) = \arg \max_a Q_*(s, a)$$

# Model-based vs. model-free RL



- We have a model of the environment, i.e.  $P(s,s')$  and  $r(s,s')$
- Updating one policy can update others via this model
- More efficient / generalizable
- Lots of evidence that animals do this (e.g. reward devaluation, Tollman's cognitive maps)

Dayan & Niv, 2008, *Curr Opin Neurobiol*

- We need to learn every  $Q(s,a)$  by trial and error, using RPE
- Updating one policy has no bearing on another
- Thus, very inefficient
- Still thought to describe some types of decision making well, e.g. habitual actions

# The Q-learning algorithm

Recall that

$$Q(s, a) = E \left[ \sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s, a_0 = a \right]$$

From Bellman's optimality equations, this is equivalent to the sum of current reward and discounted Q for future states

$$Q(s, a) = r_t + \gamma E \left[ \sum_{t=1}^{\infty} \gamma^t r_t \mid s_1 = s', a_1 = a' \right]$$

For simplicity, in the neuroscience context we can often think of states as time or behavioral trial

$$Q_t = r_t + \gamma Q_{t+1}$$

To learn Q, we define a *prediction error*,  $\delta$ , to update our estimate of Q

$$\delta_t = r_t + \gamma \hat{Q}_{t+1} - \hat{Q}_t$$

Now we can iteratively update our estimate of Q with a learning rate  $\alpha$

$$\hat{Q}_t \leftarrow \hat{Q}_t + \alpha \delta_t, \quad 0 \leq \alpha \leq 1$$

$$\hat{Q}_t \leftarrow \hat{Q}_t + \alpha(r_t + \gamma \hat{Q}_{t+1} - \hat{Q}_t)$$

$$\hat{Q}_t \leftarrow \hat{Q}_t + \alpha \gamma \hat{Q}_{t+1} + \alpha(r_t - \hat{Q}_t)$$

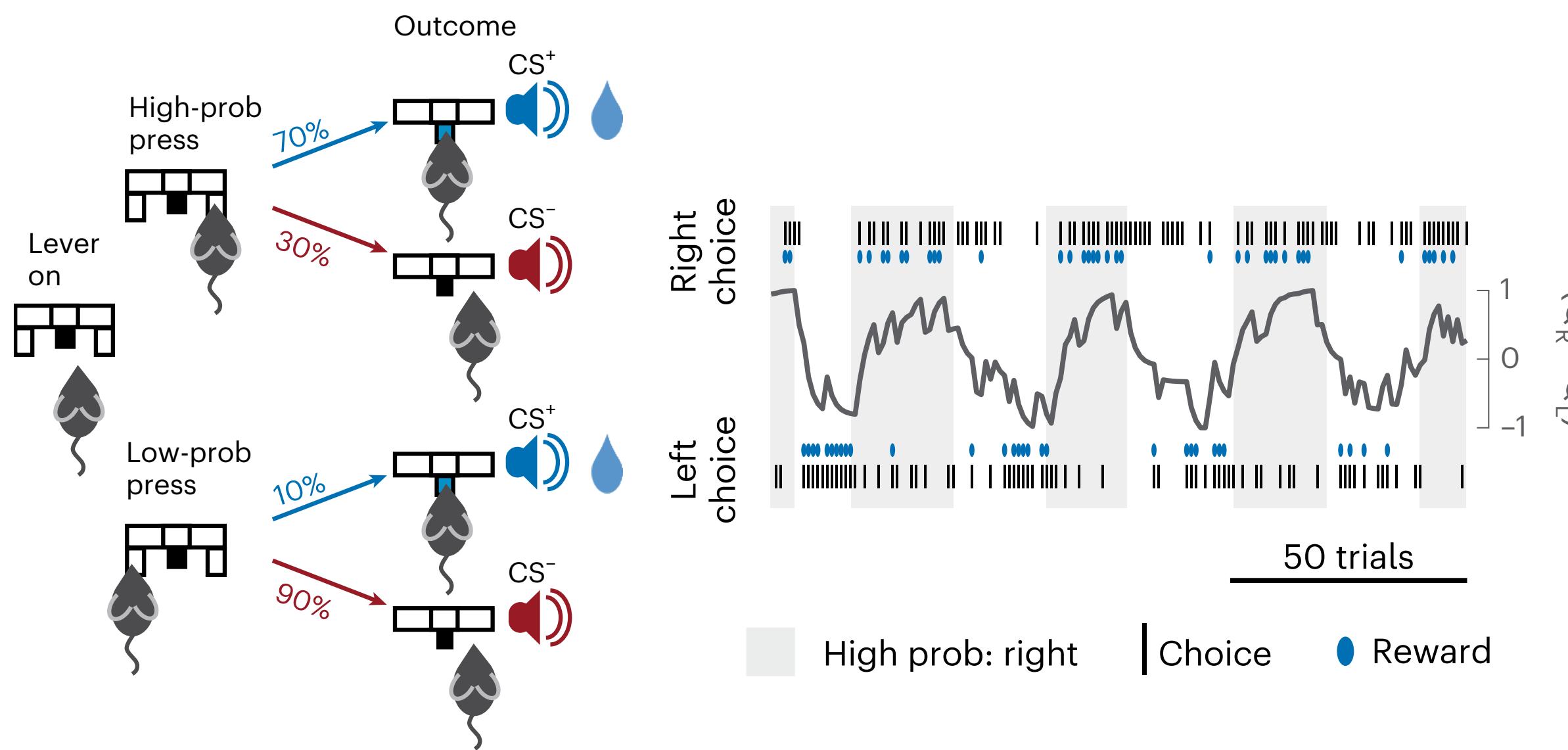
**RPE**

Rearranging this we get another common way of expressing the Q-learning rule

$$\hat{Q}_t \leftarrow (1 - \alpha) \hat{Q}_t + \alpha(r_t + \gamma \hat{Q}_{t+1})$$

# Fitting Q-learning to behavioral data

In many value-based decision making (i.e. operant) tasks, animals need to constantly update estimates of the value of their actions. E.g., *two-arm bandit*:



Cox et al., 2023, *Nature Neurosci*

Equivalently from the previous slide, the outcome of an action, e.g. getting a reward when choosing left, triggers an update in the value of that action

$$Q_{t+1} = Q_t + \alpha \delta_t$$

For example, if the animal chose left

$$Q_{t+1}^L = Q_t^L + \alpha(r_t - Q_t^L)$$

$$Q_{t+1}^R = Q_t^R$$

We then predict the probability of a left choice with a logistic function

$$P(L) = \frac{1}{1 + e^{\beta_Q(Q_L - Q_R)}}$$

(Additional behavioral parameters can also go in the exponential term)

- Our parameters are  $\alpha$  and  $\beta$ . Because  $P(L)$  has a defined likelihood function, we can fit this with MLE, given some initial value of Q's
- Some authors alternatively use Markov chain Monte Carlo methods, which Josh Glaser will cover

# Primer: temporal difference (TD) learning

In Pavlovian conditioning, there is no action, so we typically want to predict reward from a state. In this case, instead of estimating  $Q$ , we want to estimate the state value  $V$ . The math holds

$$\hat{V}_t \leftarrow \hat{V}_t + \alpha(r_t + \gamma \hat{V}_{t+1} - \hat{V}_t)$$

A useful extension of regular TD (and Q-learning) is the addition of *eligibility traces*, since in the real world reward is often delayed with respect to a state or action.

- The difficulty in assigning a reward to a past events is known as the *credit assignment problem*.

This is done by estimating the value function over a window of  $n$  time points.

$TD(\lambda)$  algorithm :

$$G_{t:t+1} = r_t + \gamma V_{t+1}$$

$$G_{t:t+n} = r_t + \sum_{i=1}^n \gamma^i V_{t+i}$$

$$\hat{V}_t \leftarrow \hat{V}_t + \alpha(G_{t:t+n} - \hat{V}_t)$$

But rather than picking a specific  $n$ , we can take a weighted average of all possible  $n$ 's with a single parameter  $\lambda$

$$G_t^\lambda = (1 - \lambda) \sum_{n=1}^{\infty} \lambda^{n-1} G_{t:t+n}$$

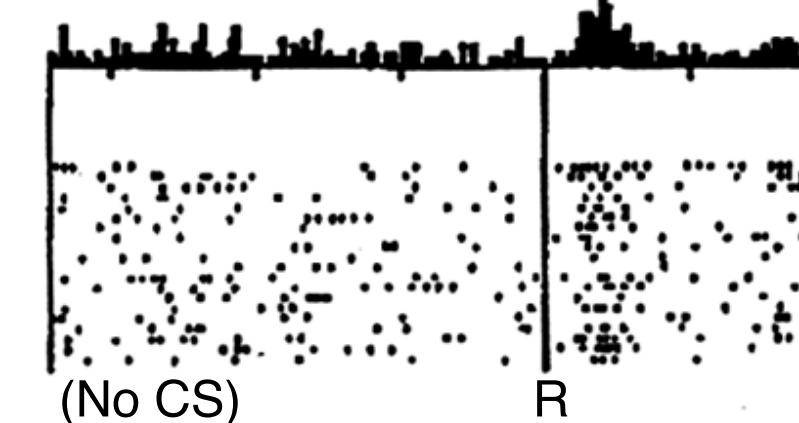
$$\hat{V}_t \leftarrow \hat{V}_t + \alpha(G_t^\lambda - \hat{V}_t)$$

# Value-base decisions in the brain

- ▶ In Pavlovian conditioning, the activity of dopamine neurons looks like RPE

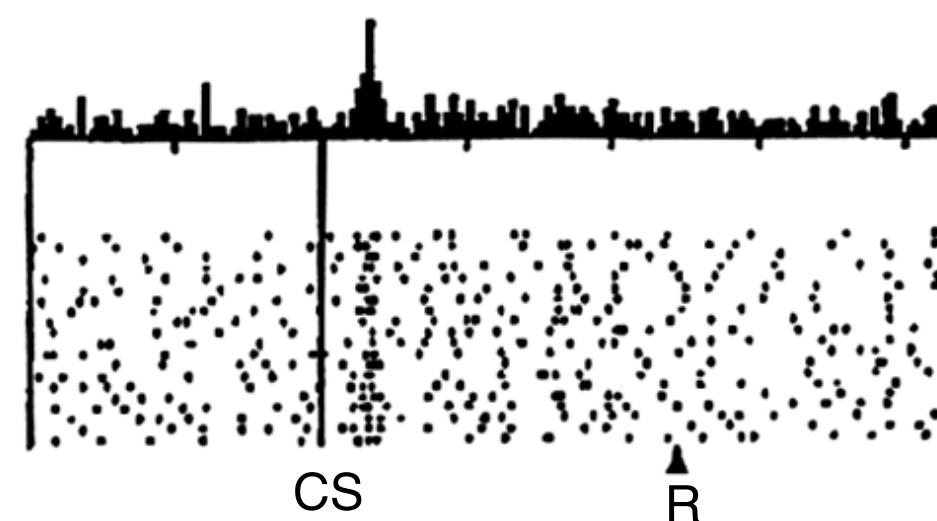
## Before learning

Reward response



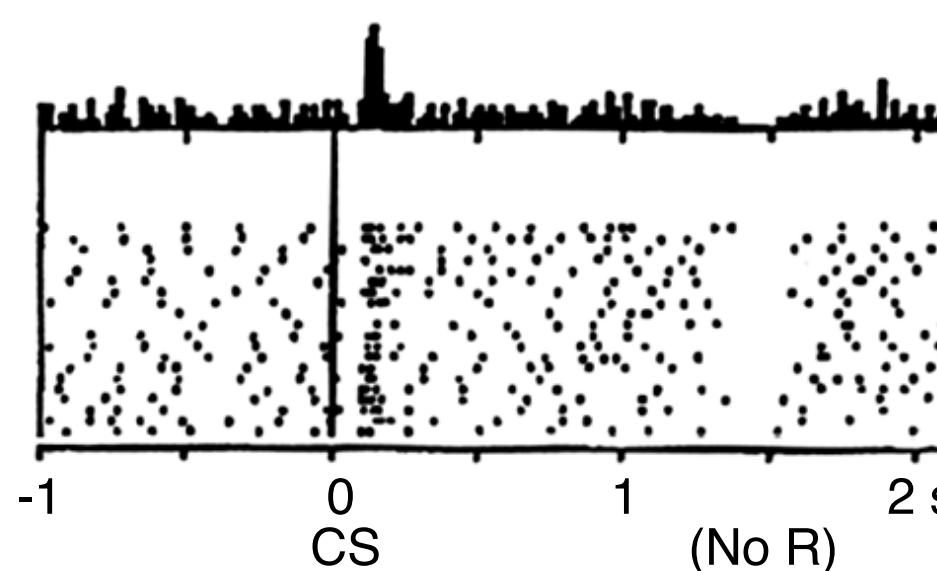
## After learning

Reward prediction response



## After learning

Reward prediction + RPE responses



- ▶ Things get messier in more complex tasks, but many results are still compatible with the RL framework (though still lots of debate)
- ▶ Like perceptual decisions, neural correlates of action value, stimulus value, effort, motivation, reward etc seem to be distributed (see reviews in suggested reading)

# Summary

- ▶ Decisions can be broadly divided into perceptual and value based
- ▶ SDT is the optimal way to decide based on a single sample from noisy data
- ▶ Sequential sampling is a generalization of SDT for multiple independent samples
- ▶ DDMs are the continuous-time limit of sequential sampling, and explain well behavioral data in perceptual tasks (choice and reaction time)
- ▶ Other models related to DDMs implement leaky integrators and inhibitory interactions inspired by neural-circuit architecture (e.g. mutual inhibition)
- ▶ Value-based decisions can be divided into economic (deterministically maximize utility) or reinforcement learning (probabilistically learn value)
- ▶ RL algorithms iteratively update value estimates based on future expected reward

# Suggested reading

- ▶ Parker & Newsome (1998). Sense and the Single Neuron: Probing the Physiology of Perception. *Annu Rev Neurosci* 21:227-277
- ▶ Ratcliff (1978). A Theory of Memory Retrieval. *Psych Rev* 85:59-108.
- ▶ Bogacz et al. (2006). The Physics of Optimal Decision Making: A Formal Analysis of Models of Performance in Two-Alternative Forced-Choice Tasks. *Psych Rev* 113:700-765.
- ▶ Gold & Shadlen (2007). The Neural Basis of Decision Making. *Annu Rev Neurosci* 30:535-574
- ▶ Usher & McClelland (2001). On the Time Course of Perceptual Choice: The Leaky Competing Accumulator Model. *Psych Rev* 108:550-592.
- ▶ Shadlen & Newsome (2001). Neural Basis of a Perceptual Decision in the Parietal Cortex (Area LIP) of the Rhesus Macaque. *J Neurophysiol* 86:1916-1936.
- ▶ Wang (2002). Probabilistic Decision Making by Slow Reverberation in Cortical Circuits. *Neuron* 36:1-20.
- ▶ Shinn et al (2020). A Flexible Framework for Simulating and Fitting Generalized Drift-Diffusion Models. *eLIFE* 9:e56938
- ▶ Kuan et al (2024). Synaptic Wiring Motifs in Posterior Parietal Cortex Support Decision-Making. *Nature*, 627:367-373
- ▶ Lee et al (2012). Neural Basis of Reinforcement Learning and Decision Making. *Annu Rev Neurosci* 35:287-308
- ▶ Dayan & Niv (2008). Reinforcement learning: The Good, The Bad, and The Ugly. *Curr Open Neurobiol* 18:185-189
- ▶ Kahneman & Tversky (1979). An Analysis of Decision under Risk. *Econometrica* 47:263-292
- ▶ Padoa-Schioppa (2011). Neurobiology of Economic Choice: a Good-Based Model. *Ann Rev Neurosci* 34:333-359
- ▶ Pulliam (1974). On the Theory of Optimal Diets. *The American Naturalist* 108:59-74
- ▶ Calhoun & Hayden (2015). The Foraging Brain. *Curr Opin Behav Sci* 5:24-31
- ▶ Murphy (2023). Probabilistic Machine Learning: Advanced Topics. Chapters 34 and 35.
- ▶ Journal club papers on canvas

The end