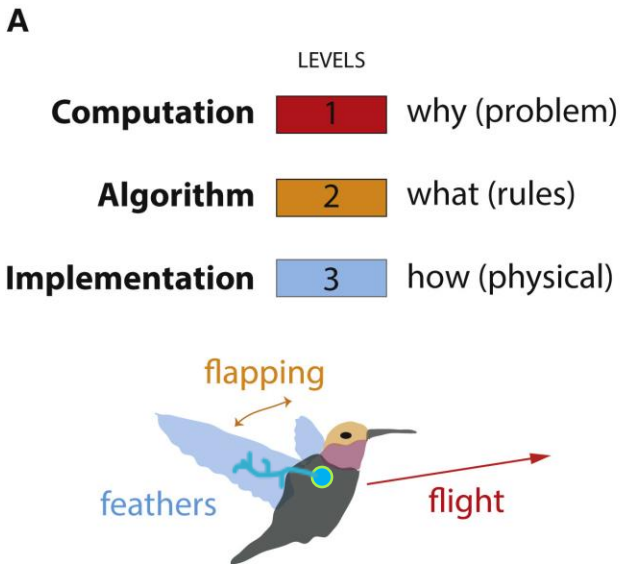
Rewards, decisions & RL in the brain

Mackenzie Mathis, PhD

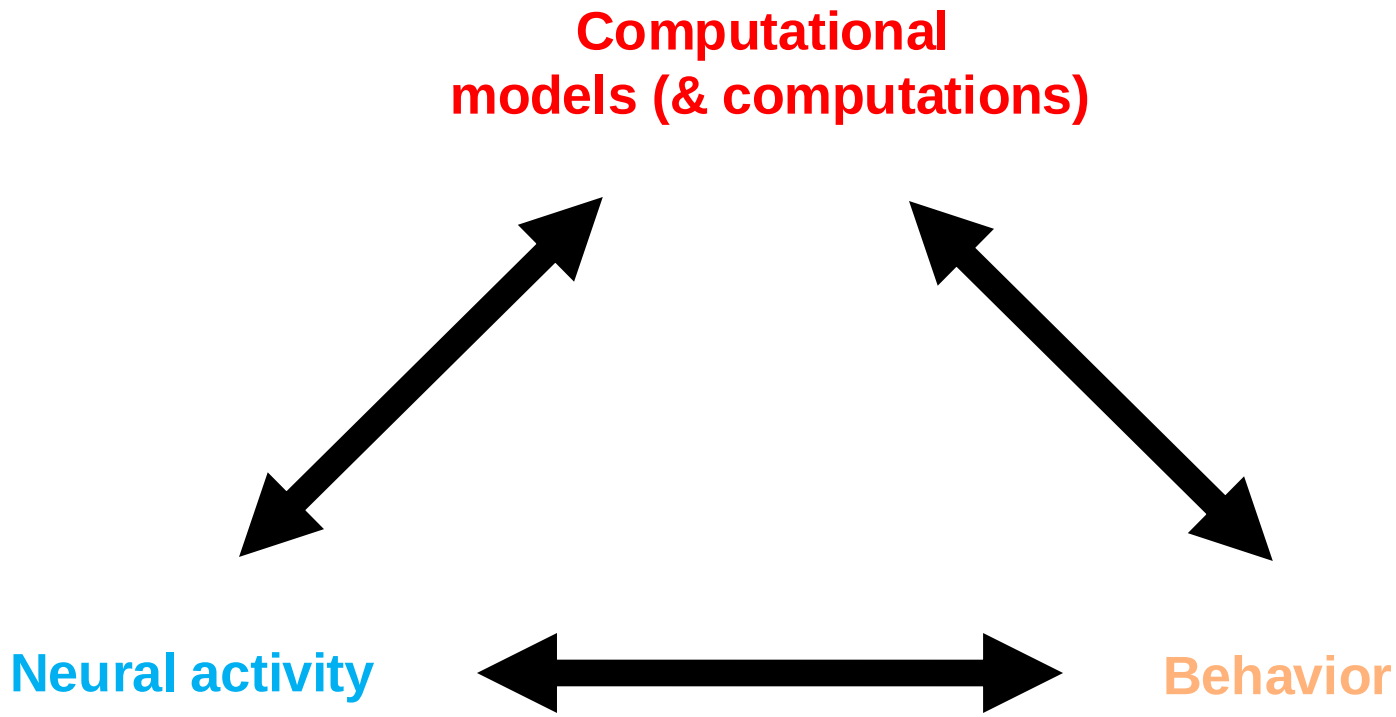
NX-435

image from https://neurosciencenews.com/decision-making-brain-19504/

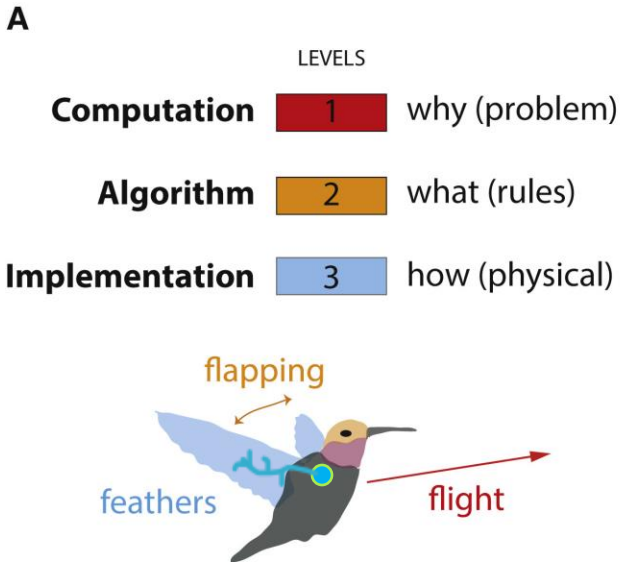
Towards closing the gap between the neuron & behavior



David Marr's three levels,
Krakauer et al. Neuron 2017

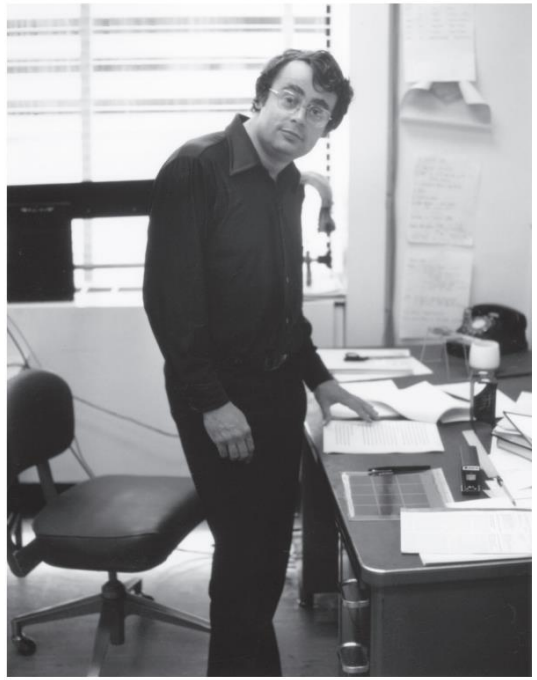


Foundations of computational neuroscience



David Marr's three levels,
Krakauer et al. Neuron 2017

David Marr (1945-1980) proposed three levels of analysis:
1.the problem (Computational Level)
2.the strategy (Algorithmic Level)
3.how its actually done by networks of neurons (Implementational Level)



David Marr at MIT. Photo by the author.

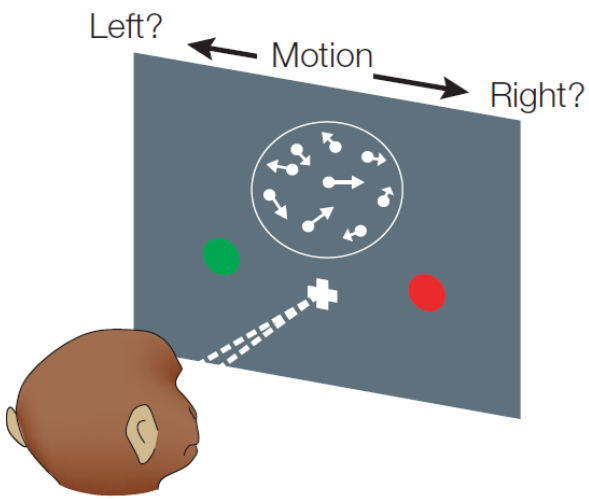
Decision-making & behavior

A problem we all face in our daily lives is how to make optimal decisions
(maximize reward, minimize punishment)

To ski or to not ski?

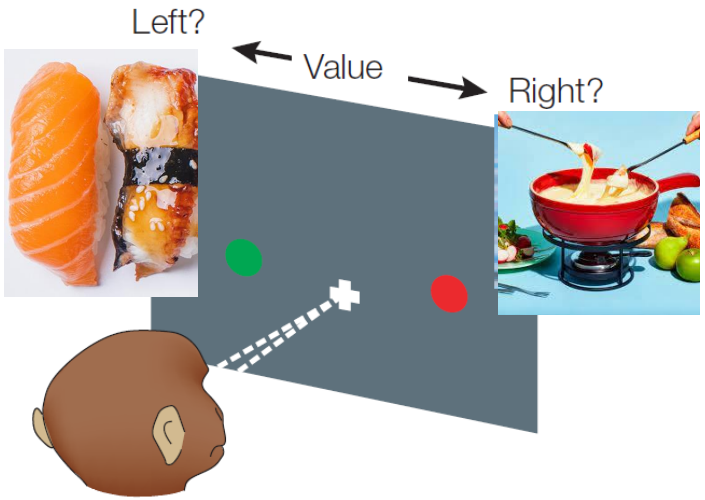
Decision-making & behavior

Perceptual judgment



- Sensory evidence

Value-based decisions



- Cost/benefit
- Value (utility)

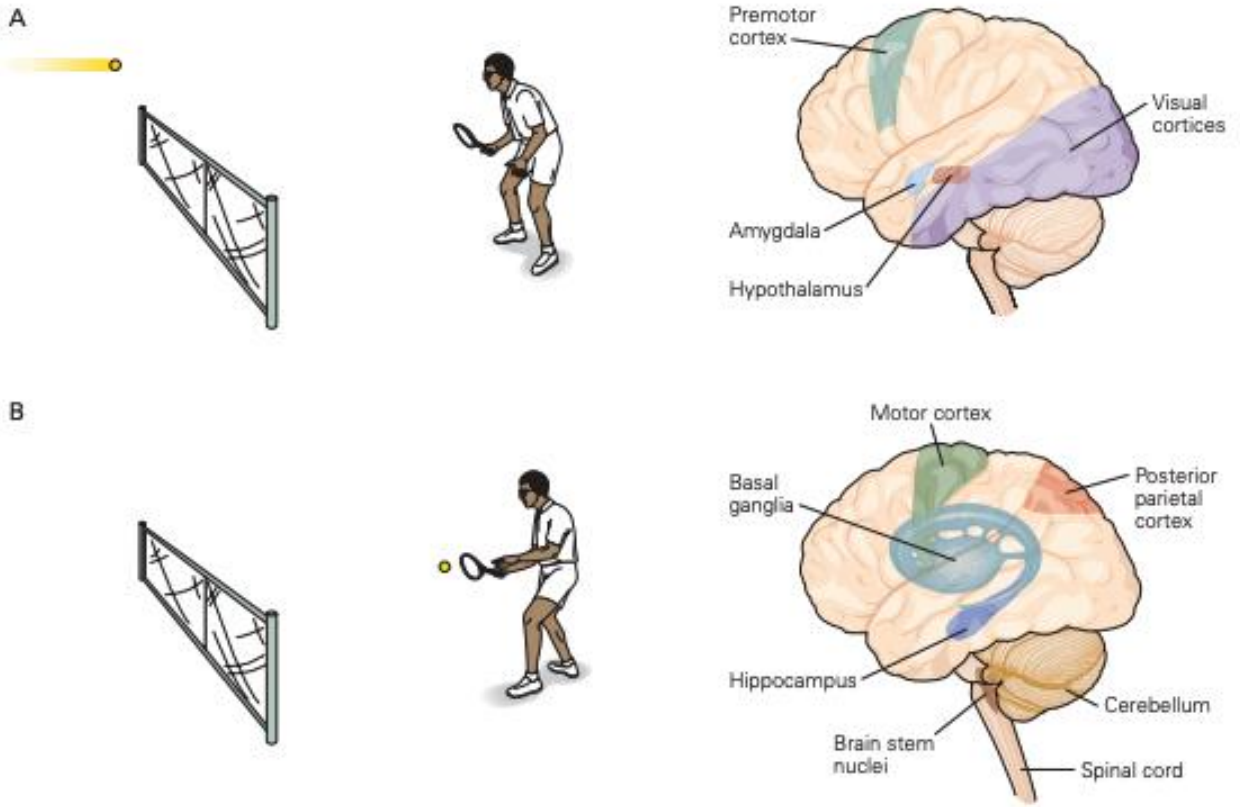
Making decisions can be challenging…

Why is decision-making difficult?

- Reward/punishment may be delayed (mins, days, weeks, years)

- Outcomes may depend on a series of actions
  ⇒ "**credit assignment problem**" (Marvin Minsky 1961 & Sutton, 1978)

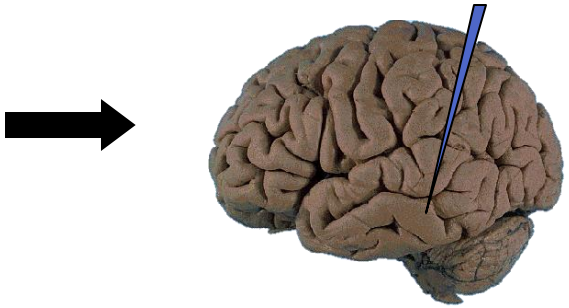
i.e., the agent needs to determine which action(s) will lead to a given outcome



Which action leads to each outcome? (what is the body position, the swing, the follow-through, and/or your visual acuity?)

Encoding of information in neural firing patterns

Stimulus

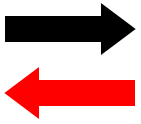


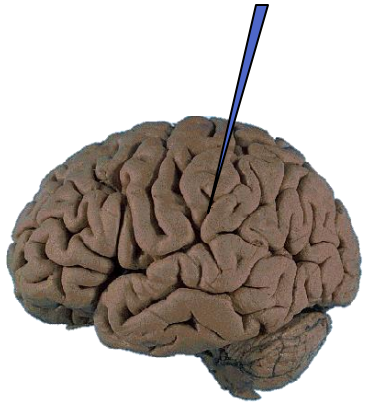
Spikes

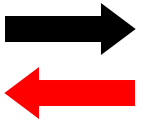
- How do neurons respond to a certain stimulus?

Encoding and decoding

Stimulus





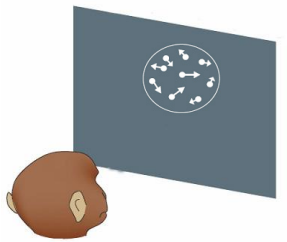


Spikes

- Our brain needs to determine what is going on in the real world from patterns of spikes.

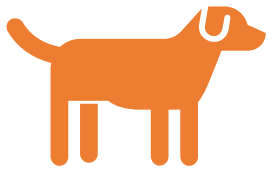
Rewards (punishment) and decision making

- **Perceptual & Value-based Decision Making**

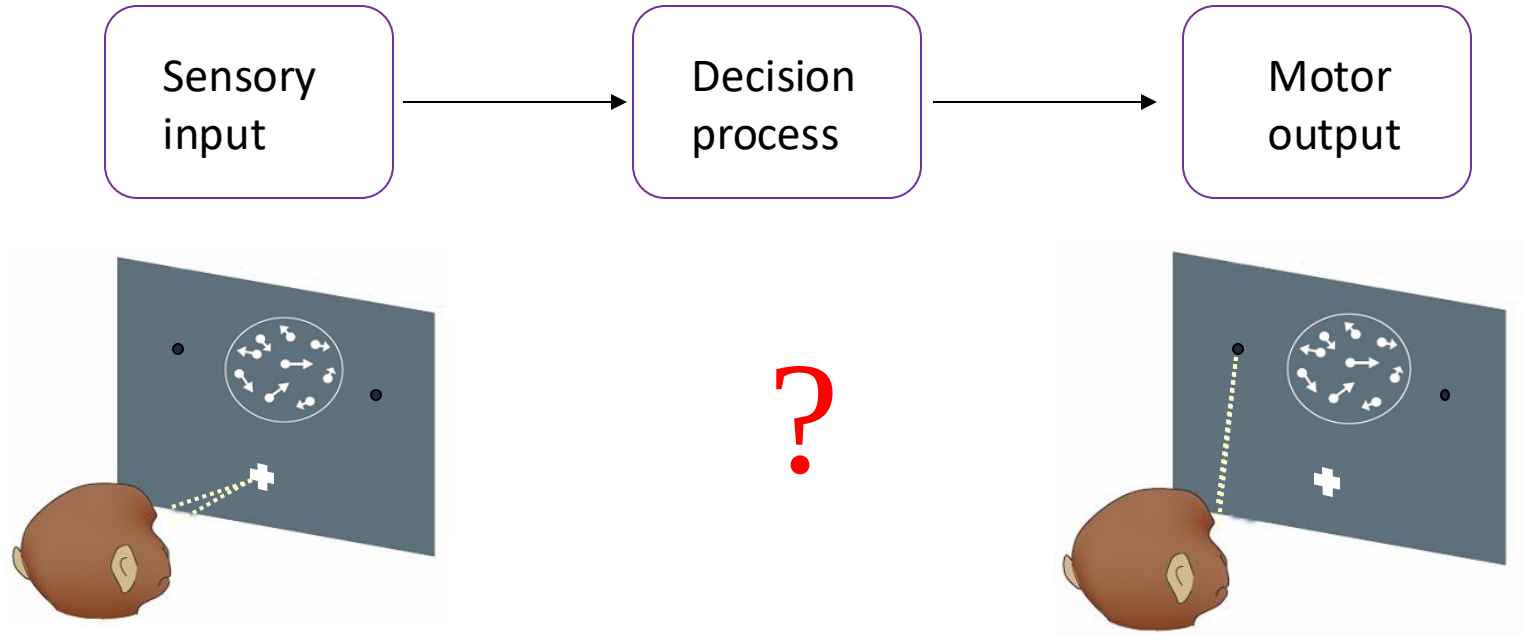


- Operant Conditioning

- Classical Conditioning



Perceptual decision making



Random dot motion task

- It takes up to 1-2 seconds to decide
- Decisions unfold gradually by accumulating noisy evidence.

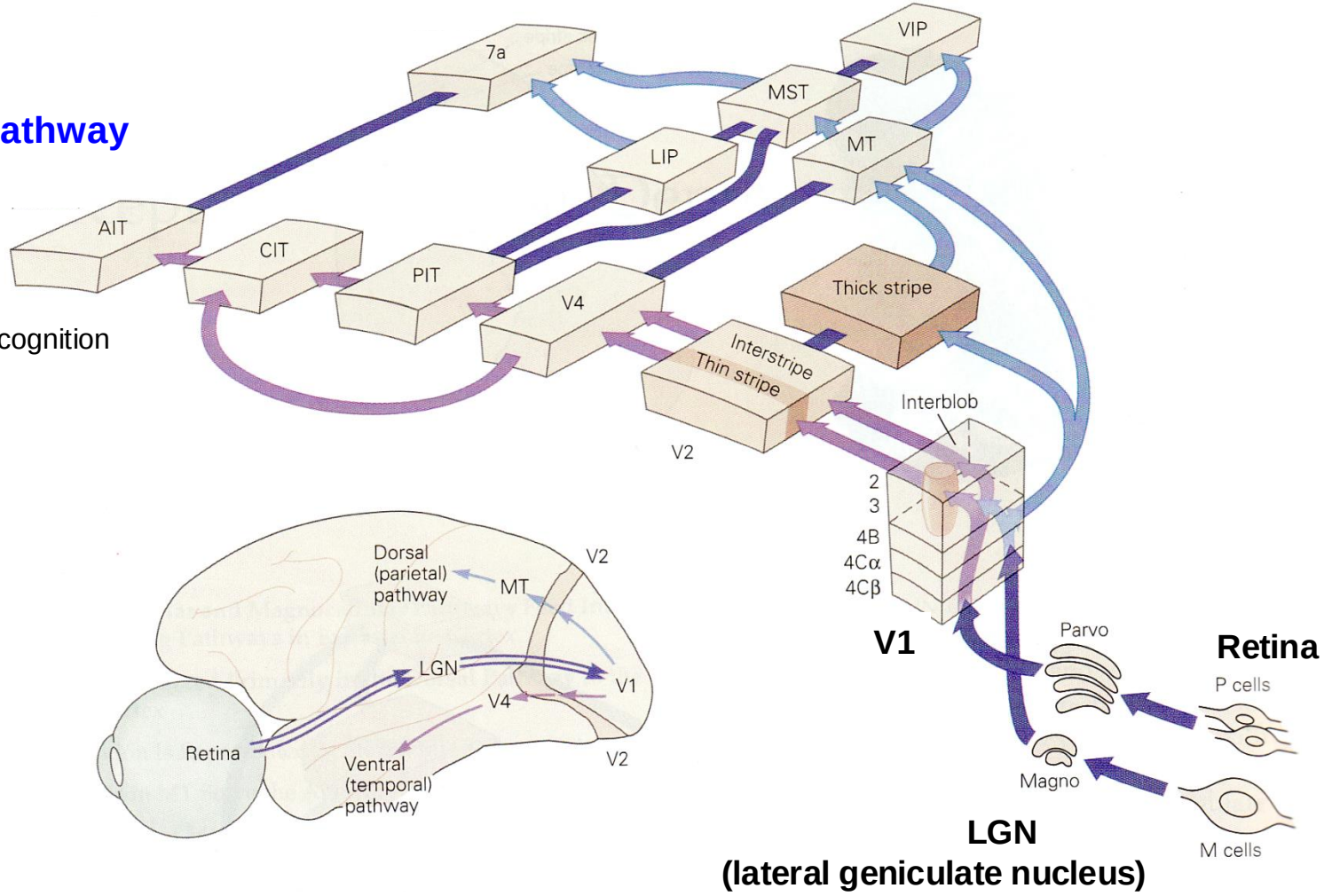
The Visual pathway

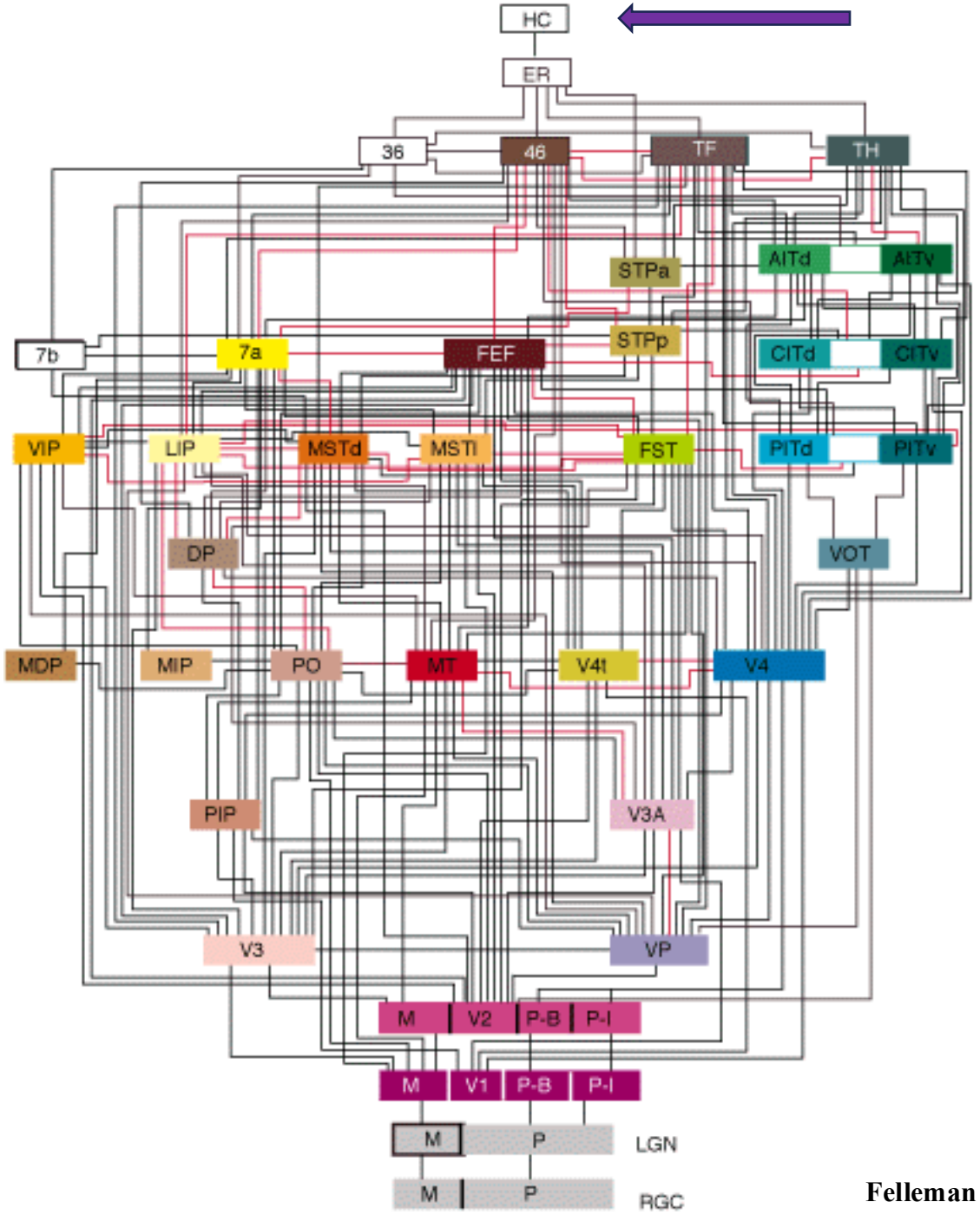
Dorsal pathway ("where")

- location, motion, depth
- color-blind

Ventral pathway ("what")

- object recognition
- color

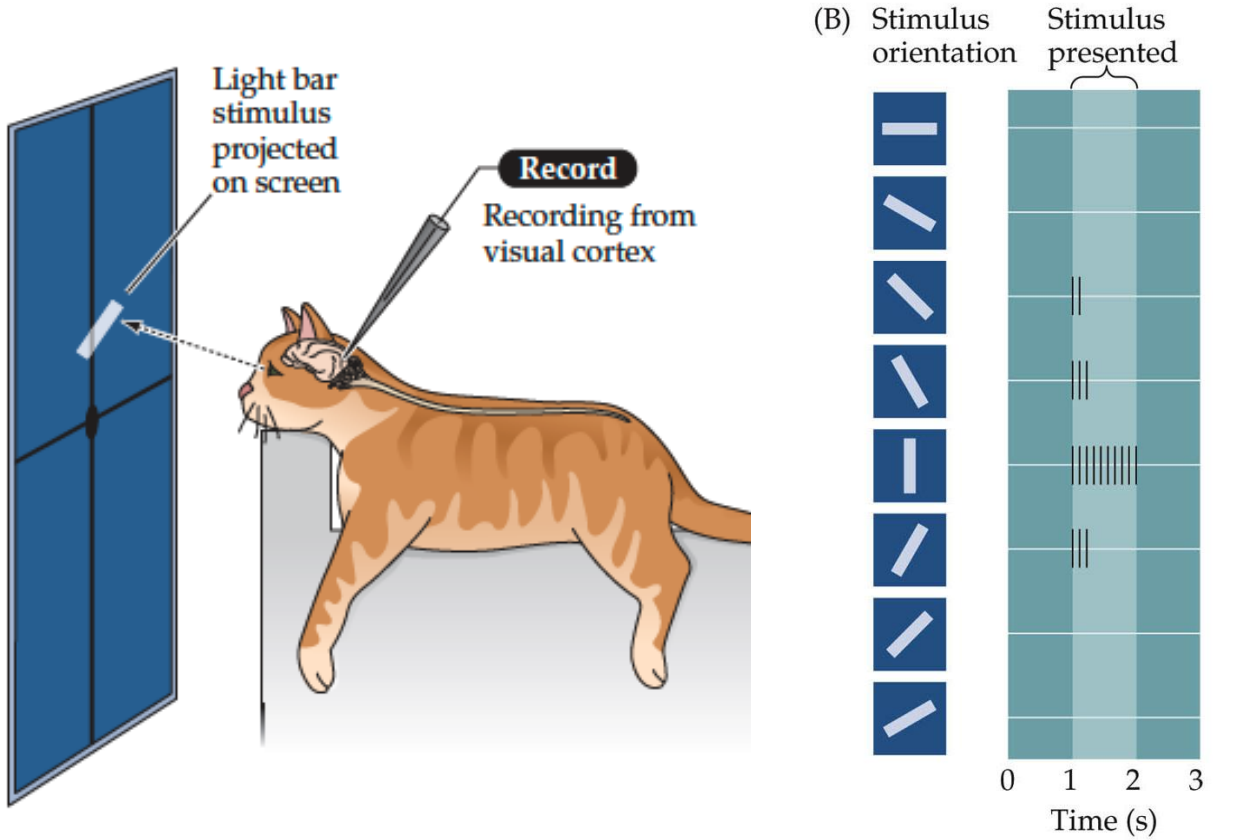




Reminder: vision is tightly integrated into decision-making, memory *(and movement areas)*

Felleman and Van Essen *Cerebral Cortex* 1991

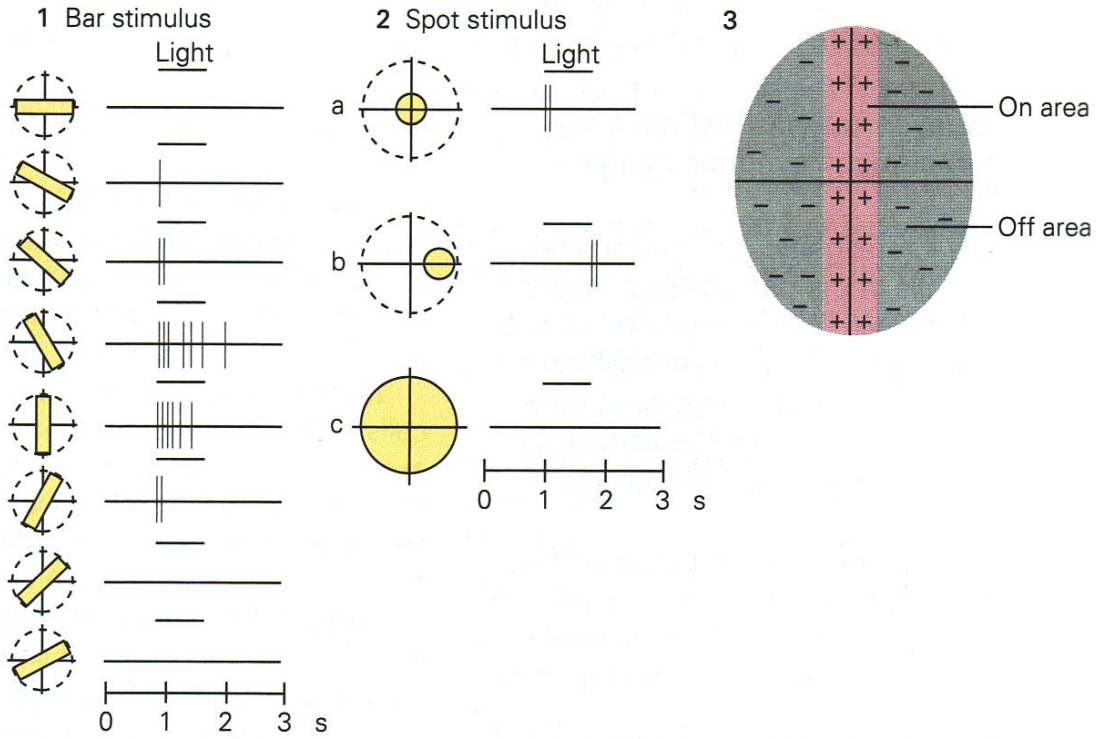
Organization of receptive field of neurons in the primary (V1) visual cortex



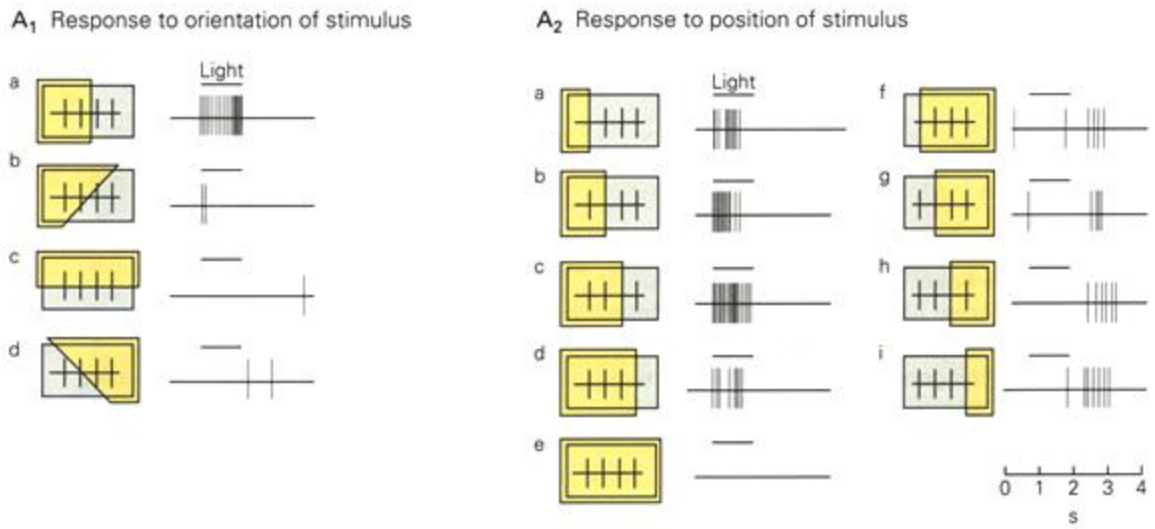
Example for a V1 neuron with a "simple" (bar-like) receptive field

Purves Fig. 12.8

V1 simple & complex cells

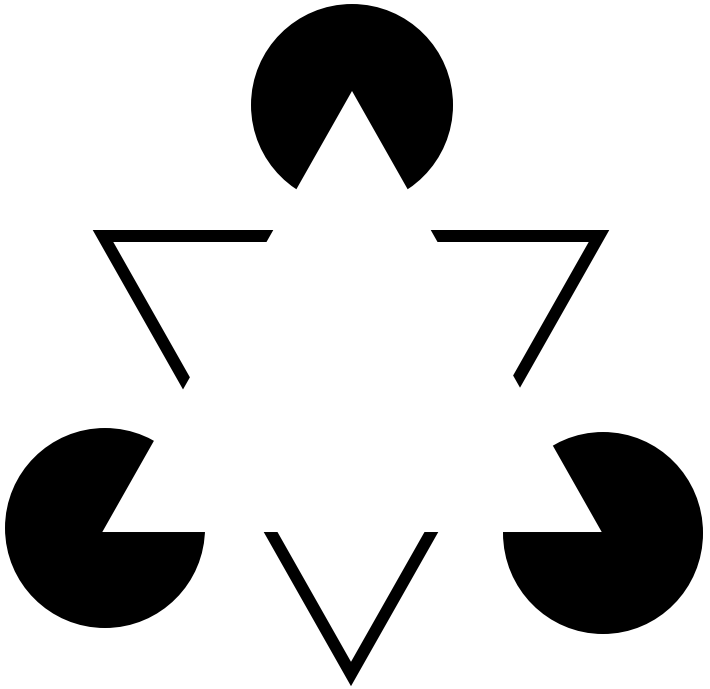


- Orientation specificity!



- Orientation specificity!
- Less sensitive to exact locations

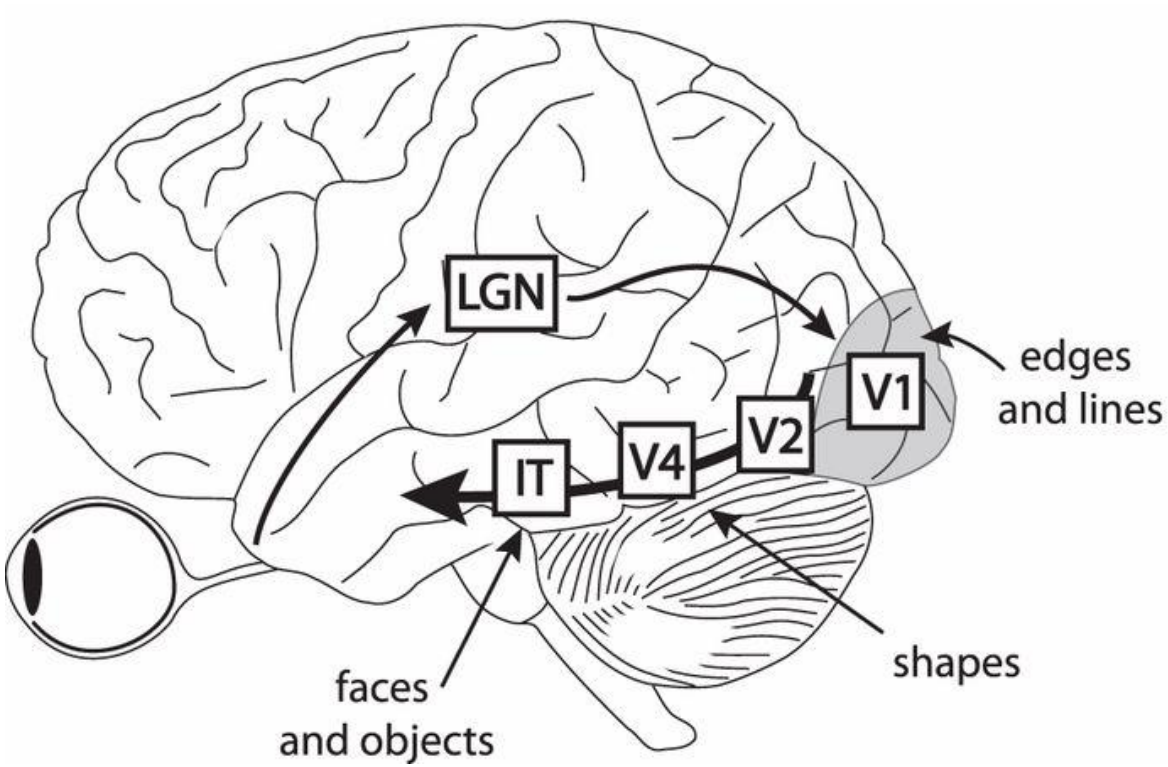
V2 neurons: edge detection

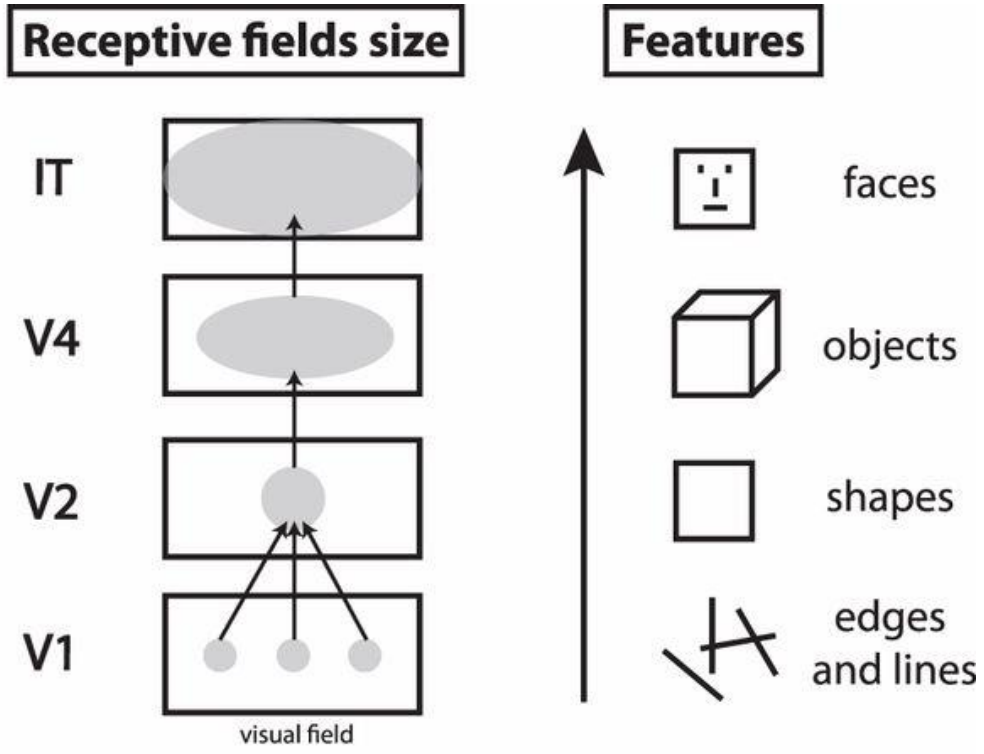


(Kanizsa, 1955)

Slide courtesy of Prof. Uchida

Hierarchical visual processing

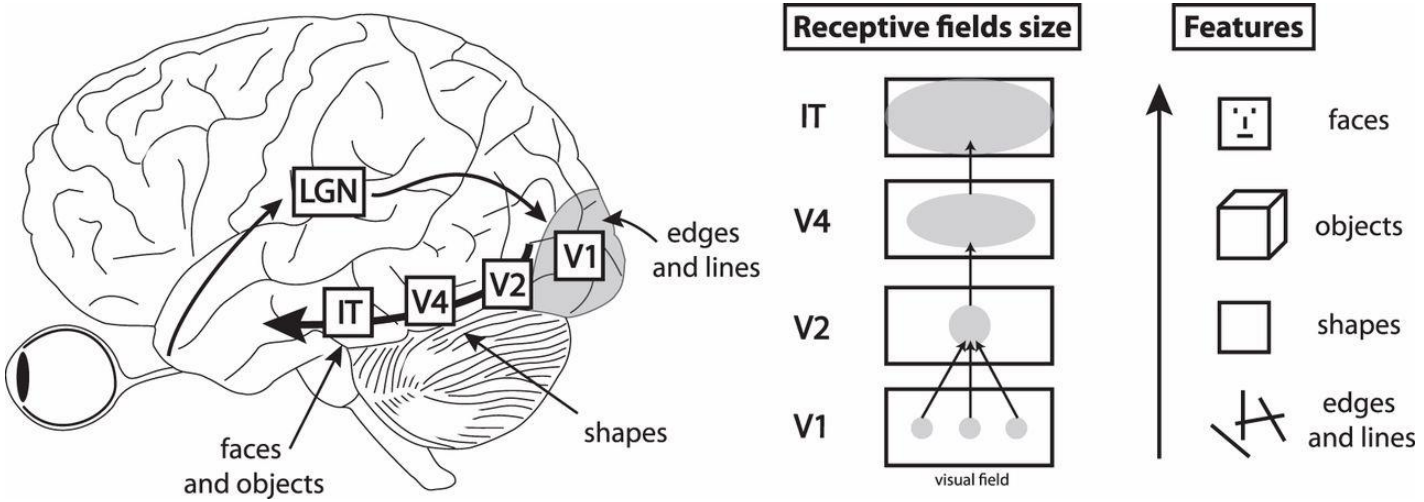


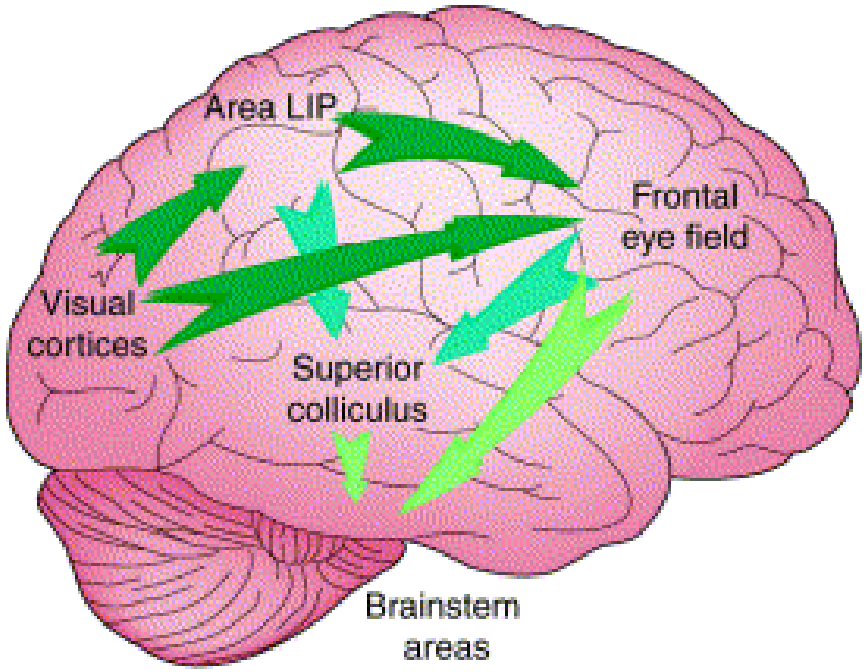


- V1 neurons are most sensitive to low-level features, such as edges and lines.
- In higher visual areas, like V4 and IT, receptive fields are larger, and neurons are sensitive to complex features, such as shapes and objects.
- Responses of high-level neurons are fully determined by the neural firing of lower-level neurons. For example, the neural firing to a square is determined by the neural firing for two vertical and two horizontal lines.

From: When crowding of crowding leads to uncrowding
Journal of Vision. 2013;13(13):10. doi:10.1167/13.13.10

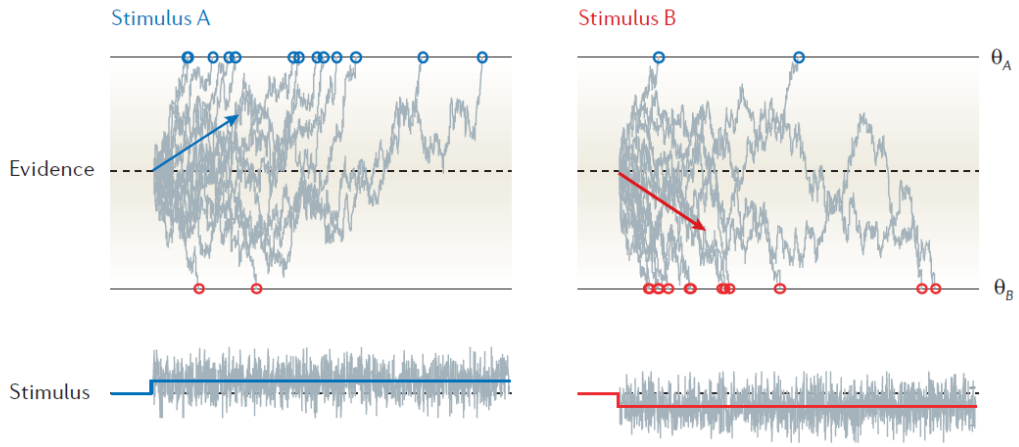
Hierarchical visual processing → decision areas



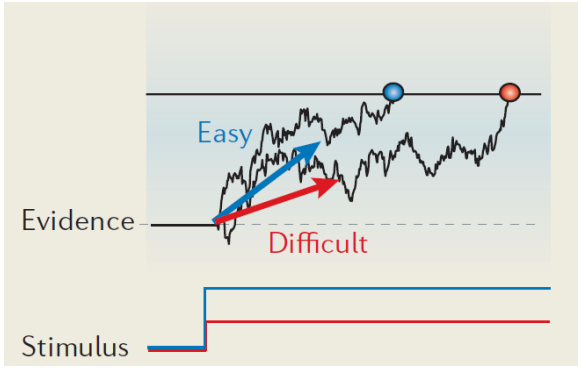


Drift diffusion models: accumulating noisy evidence

- Variability in response times and judgments

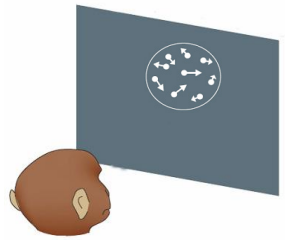


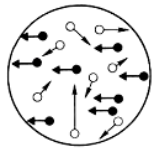
- Effect of difficulty on response times

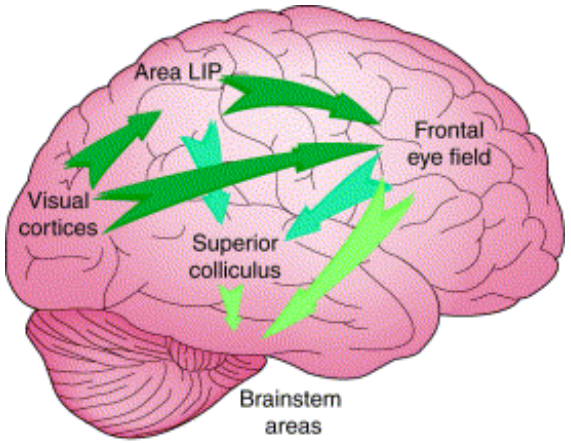


(Uchida et al., 2006)

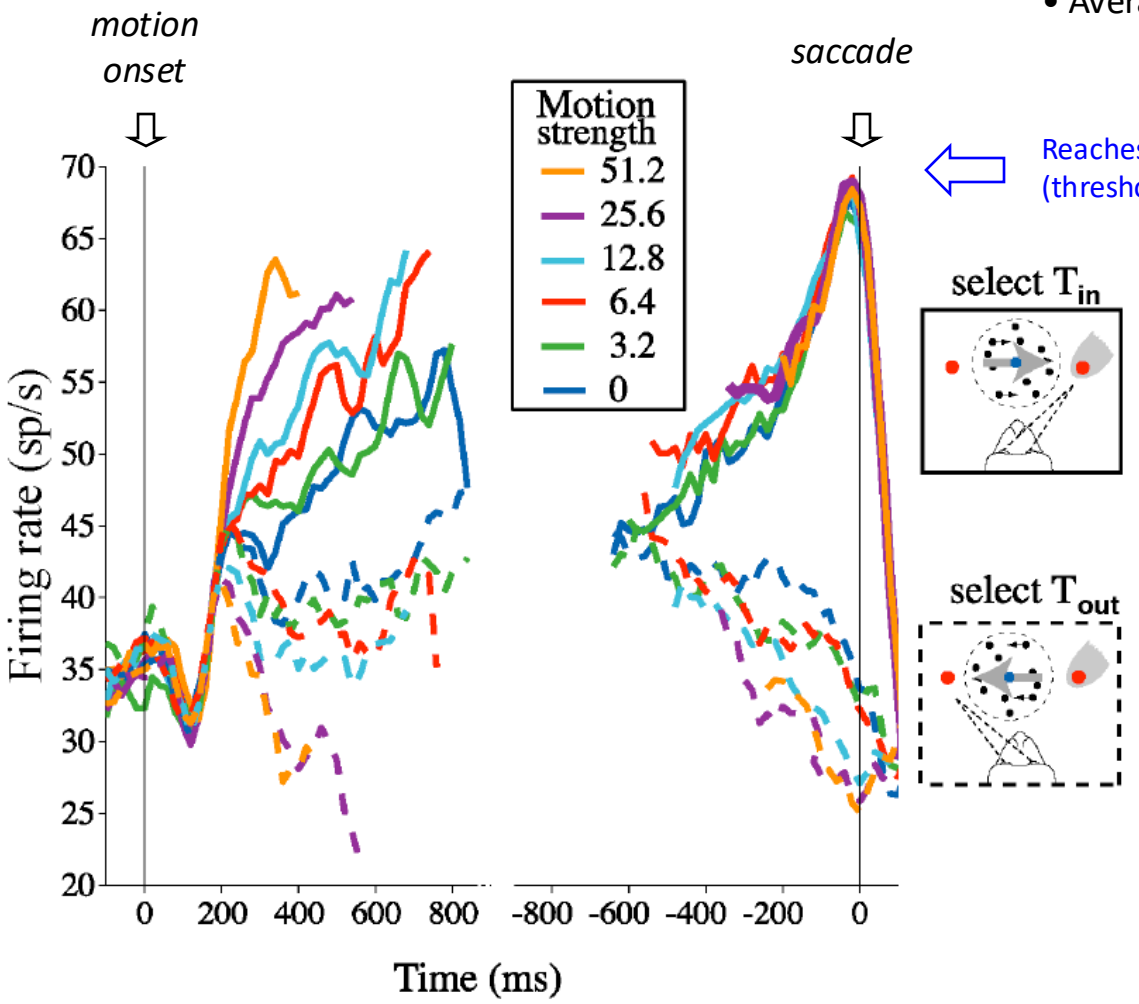
LIP neurons: accumulating evidence







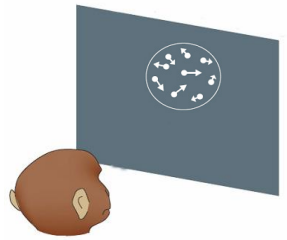
- Average of 54 neurons



Reaches to the same level (threshold)

(Roitman & Shadlen, 2002)

- Strong signal ⟹ faster increase in firing rate



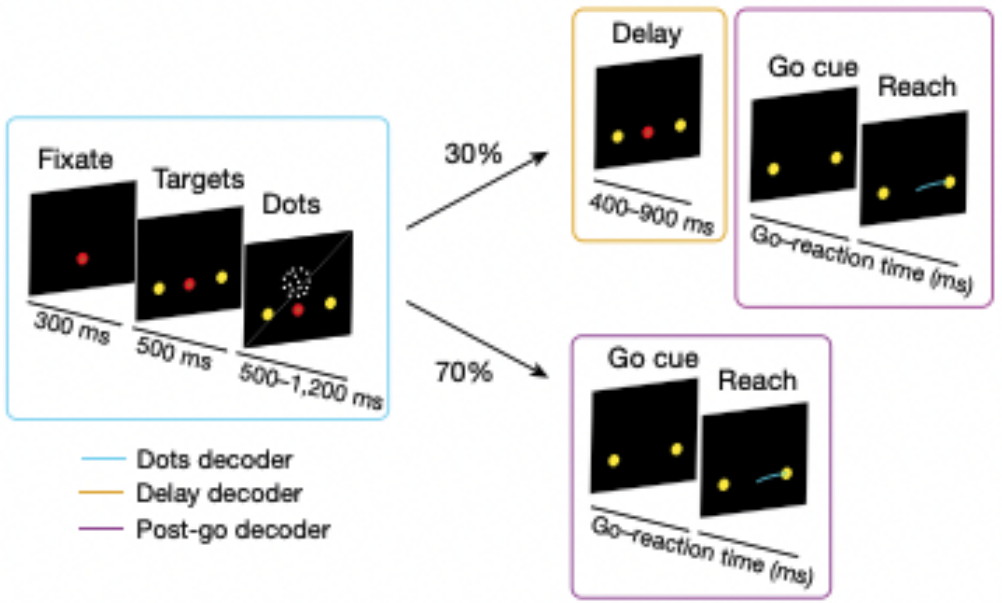
Article | Published: 20 January 2021

Decoding and perturbing decision states in real time

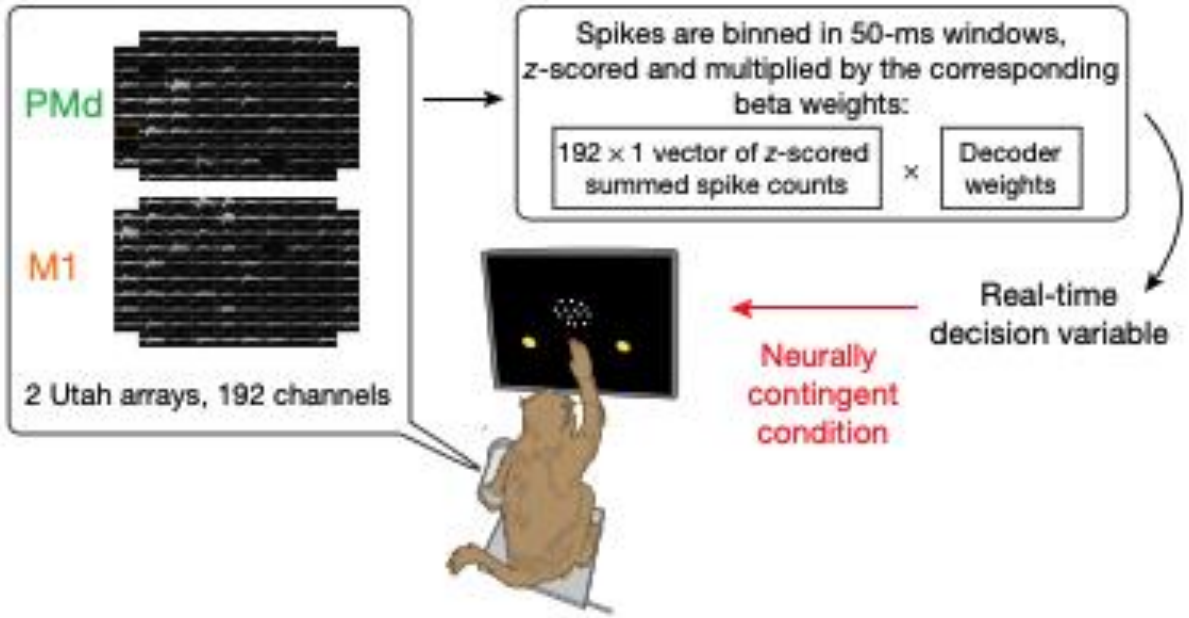
Diogo Peixoto, Jessica R. Verhein, Roozbeh Kiani, Jonathan C. Kao, Paul Nuyujukian, Chandramouli Chandrasekaran, Julian Brown, Sania Fong, Stephen I. Ryu, Krishna V. Shenoy & William T. Newsome

Nature **591**, 604–609 (2021) | Cite this article

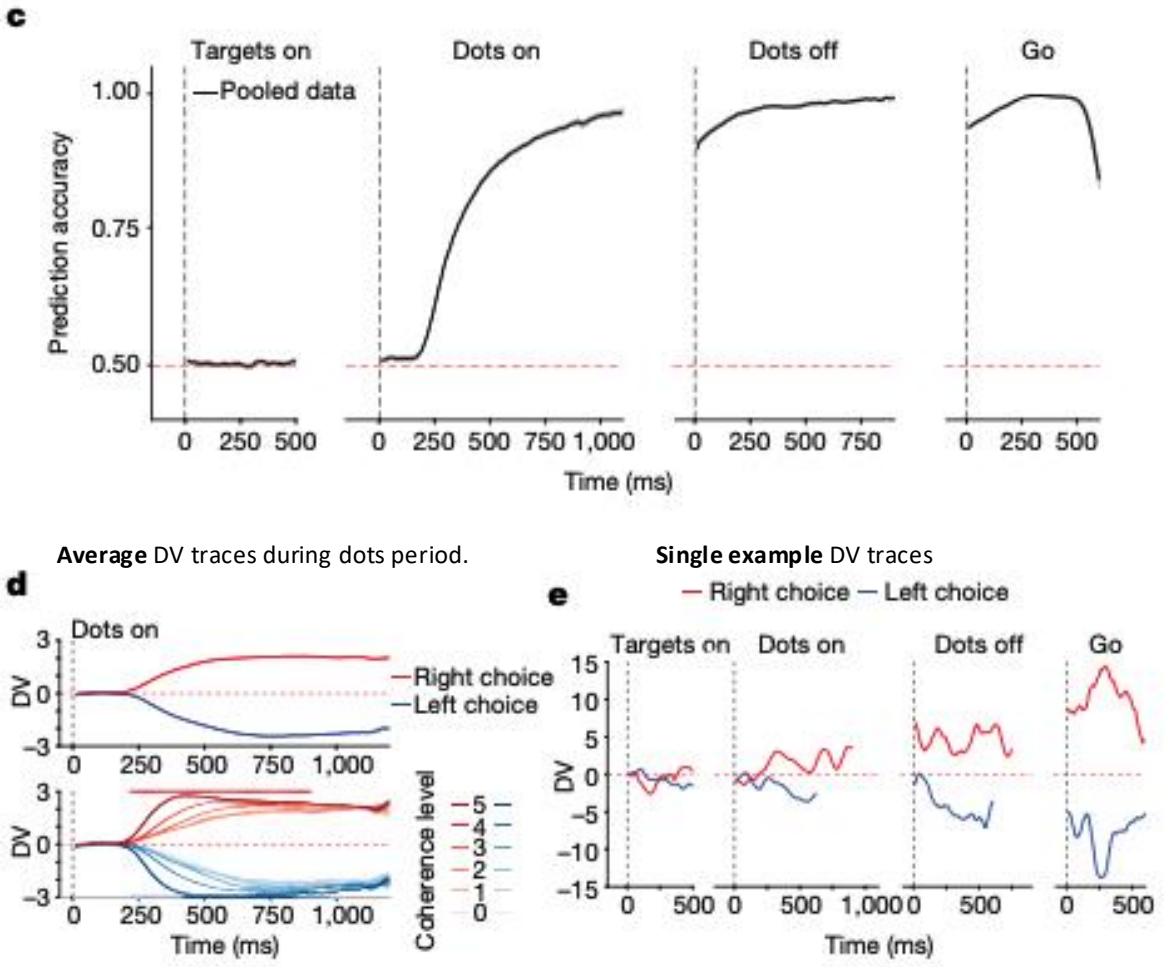
In dynamic environments, subjects often integrate multiple samples of a signal and combine them to reach a categorical judgment**. The process of deliberation can be described by a time-varying decision variable (DV), decoded from neural population activity, that predicts a subject's upcoming decision. Within single trials, however, there are large moment-to-moment fluctuations in the DV, the behavioral significance of which is unclear.** Here, using real-time, neural feedback control of stimulus duration, we show that within-trial DV fluctuations, decoded from motor cortex, are tightly linked to decision state in macaques, predicting behavioral choices substantially better than the condition-averaged DV or the visual stimulus alone. Furthermore, robust changes in DV sign have the statistical regularities expected from behavioral studies of **changes of mind**. Probing the decision process on single trials with weak stimulus pulses, we find evidence for time-varying absorbing decision bounds, enabling us to distinguish between specific models of decision making.



Real-time readout of decision states during a motion discrimination task



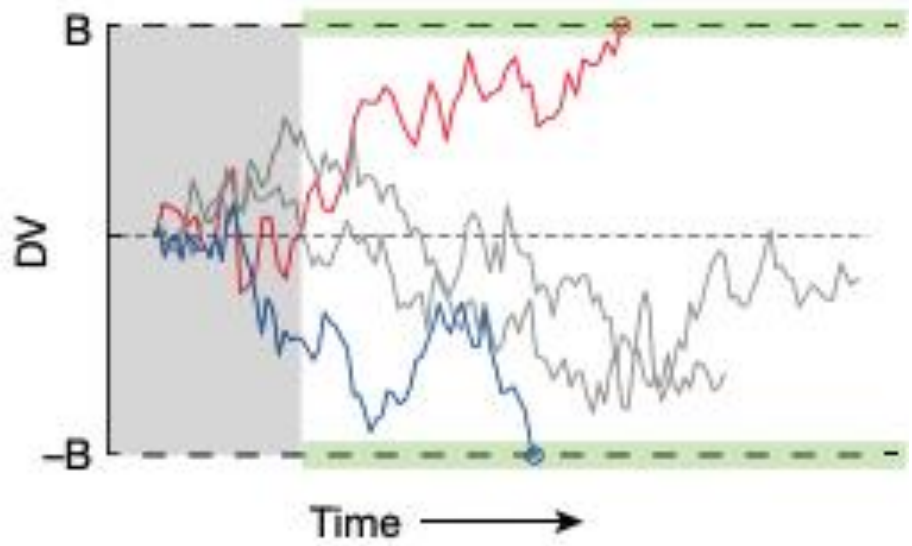
- they could decode the correct choice starting at ~250 ms
- the "DV" decodability in time correlates with the strength of the motion coherence
- *single trials are noisy!*



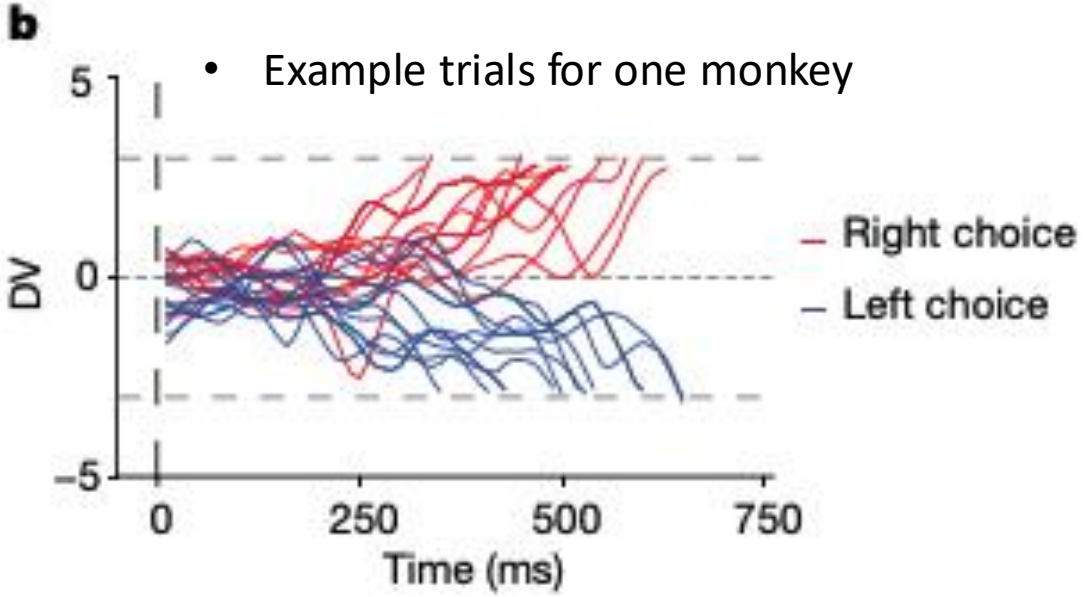
*DV: a continuous readout of the strength of the model's prediction for the subject's choice. They calculated the logistic model's log odds ratio for the two choices for each time point on every trial.

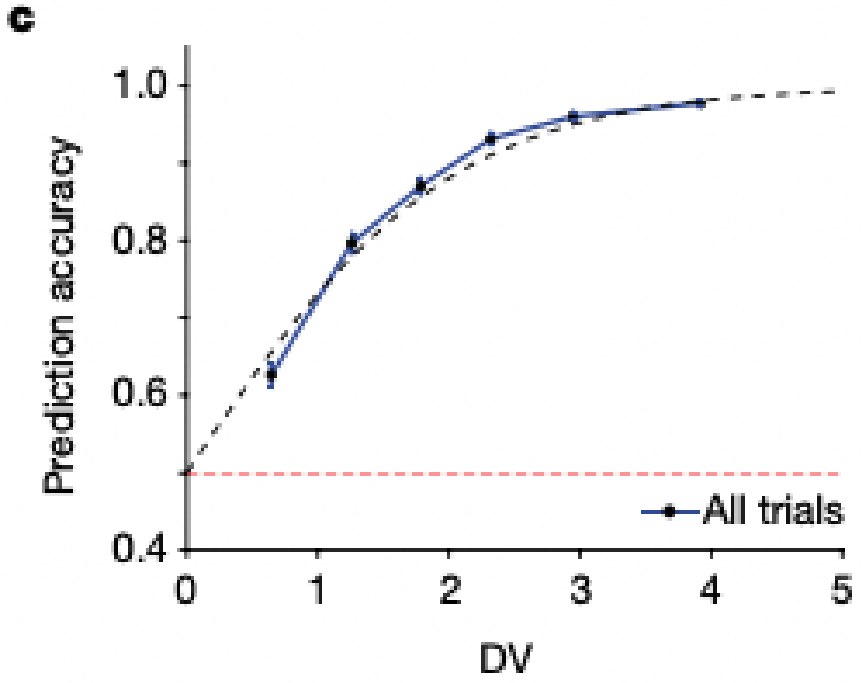
DV fluctuations track evolving decisions

- Virtual boundary experiment schematic



- Example trials for one monkey

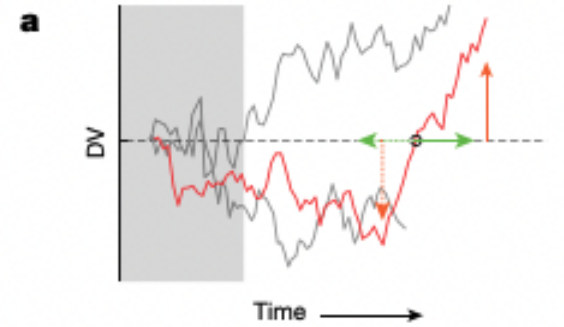




- Amazingly, they can very well predict the choice outcome depending on the DV strength

Closed-loop experiment to test: is it a change a mind (CoM)?

- They established neural criteria for a candidate CoM that, when met in real time, led to stimulus termination and the monkey's decision (i.e., they needed to rapidly decide!)



the relationship between prediction accuracy and DV at stimulus termination was very similar for CoM and non-CoM trials – and only ~2% error rate! Meaning yes, the state of the DV at the time of termination DID predict the choice! ✅

(more than correlation!)

- CoM criteria:
  - (1) CoMs are more frequent for low- and intermediate-coherence trials compared with high-coherence trials
  - (2) more likely to be corrective than erroneous
  - (3) CoMs are more frequent early in the trial than later in the trial

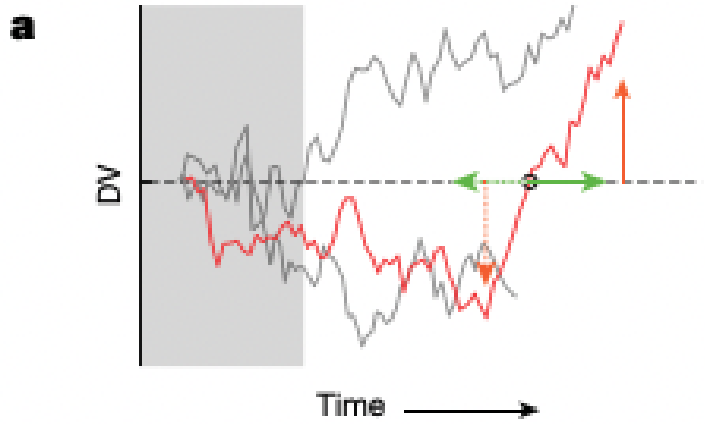
Making decisions can be challenging…

Why is decision-making difficult?

- Reward/punishment may be delayed (mins, days, weeks, years)

- Outcomes may depend on a series of actions
  ⇒ "**credit assignment problem**" (Marvin Minsky 1961 & Sutton, 1978)

i.e., the agent needs to determine which action(s) will lead to a given outcome



How (else) can we formalize studying decisions in the brain?

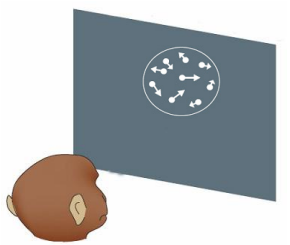
- An algorithm (to test): reinforcement learning (RL)

- Experimental paradigms: classical conditioning & operant conditioning

- Aim: to understand the neural basis – can RL be implemented in the brain, and if so, where?

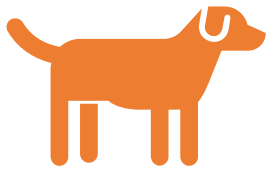
Rewards (punishment) and decision making

- Perceptual & Value-based Decision Making



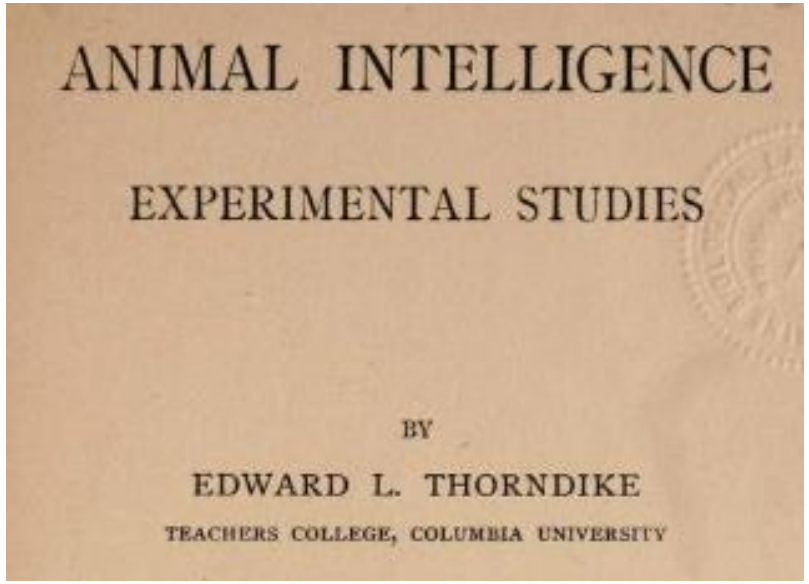
- **Operant Conditioning**

- **Classical Conditioning**

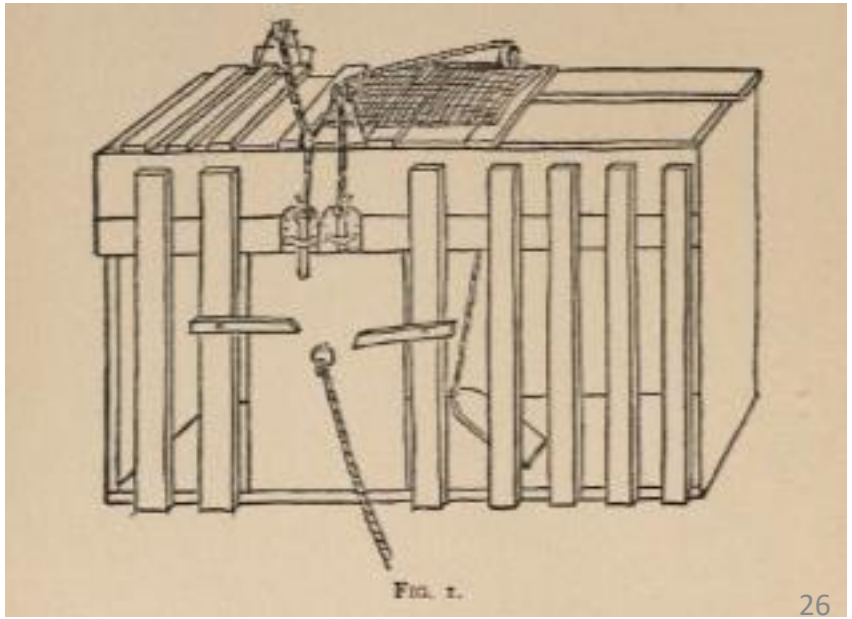


Operant Conditioning (also called trial-and-error learning)

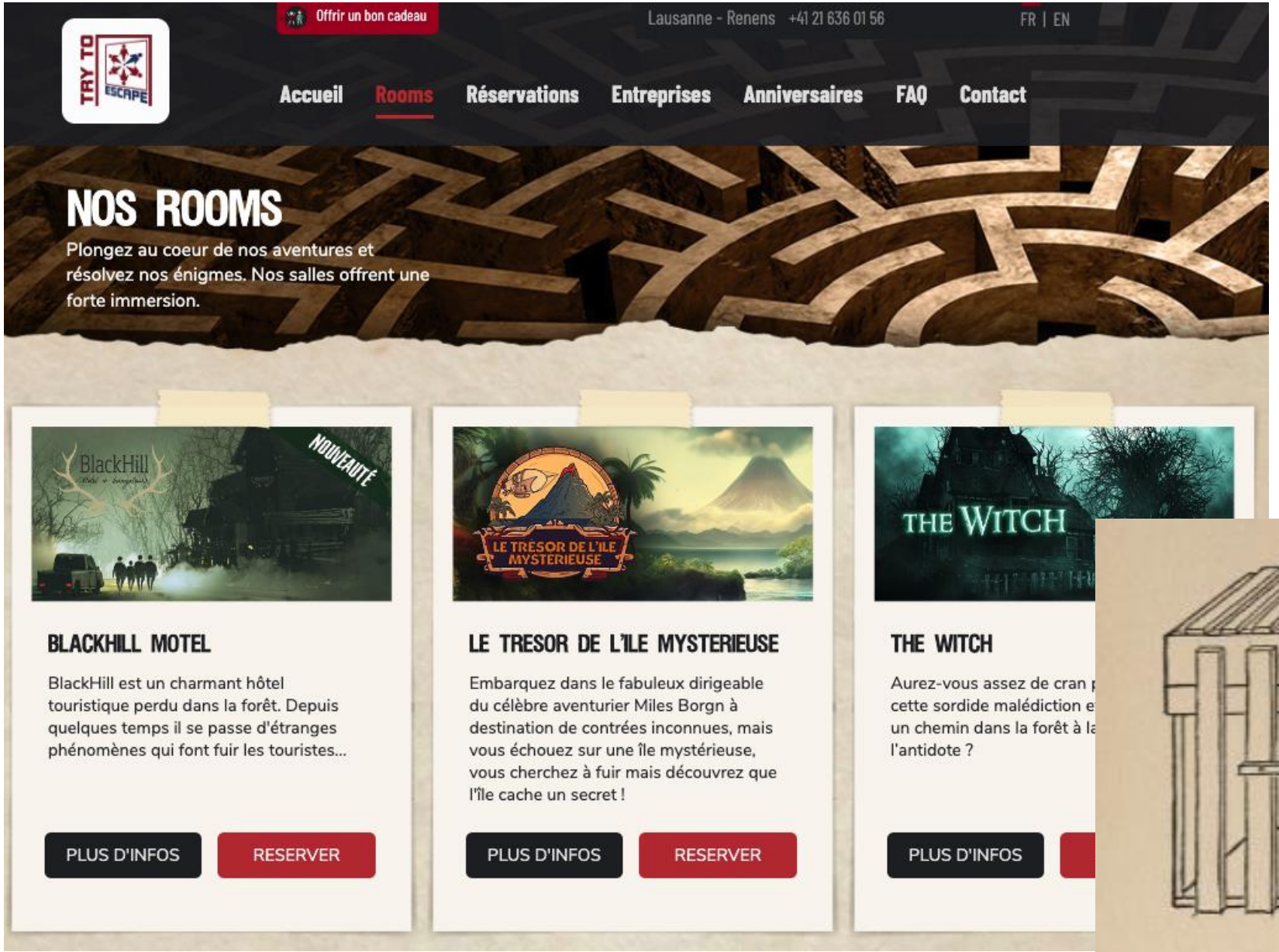
Edward Thorndike
(Wikipedia, 1912)



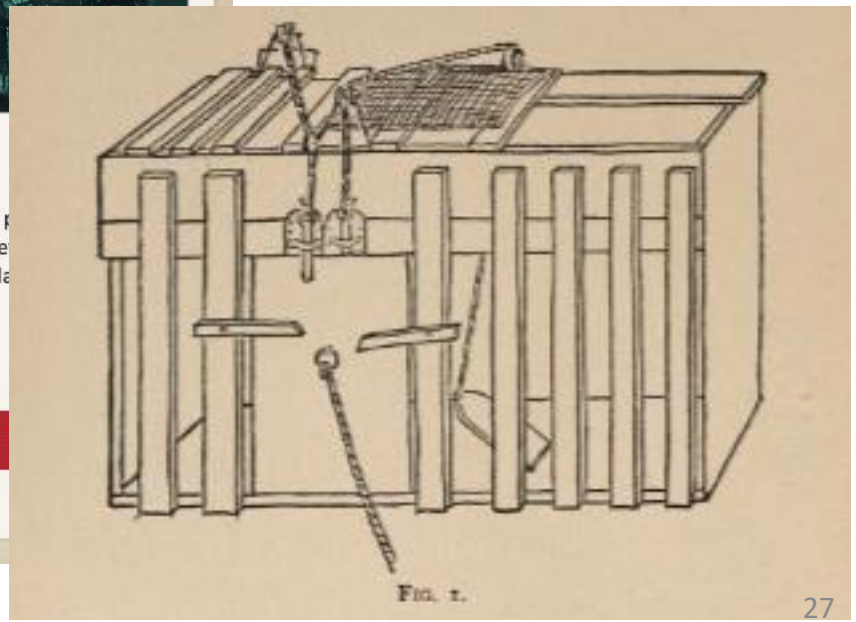
- operant conditioning can be considered as the formation of a **predictive relationship between an action and an outcome**
- **classical conditioning is the formation of a predictive relationship between two stimuli (the CS and the US)*



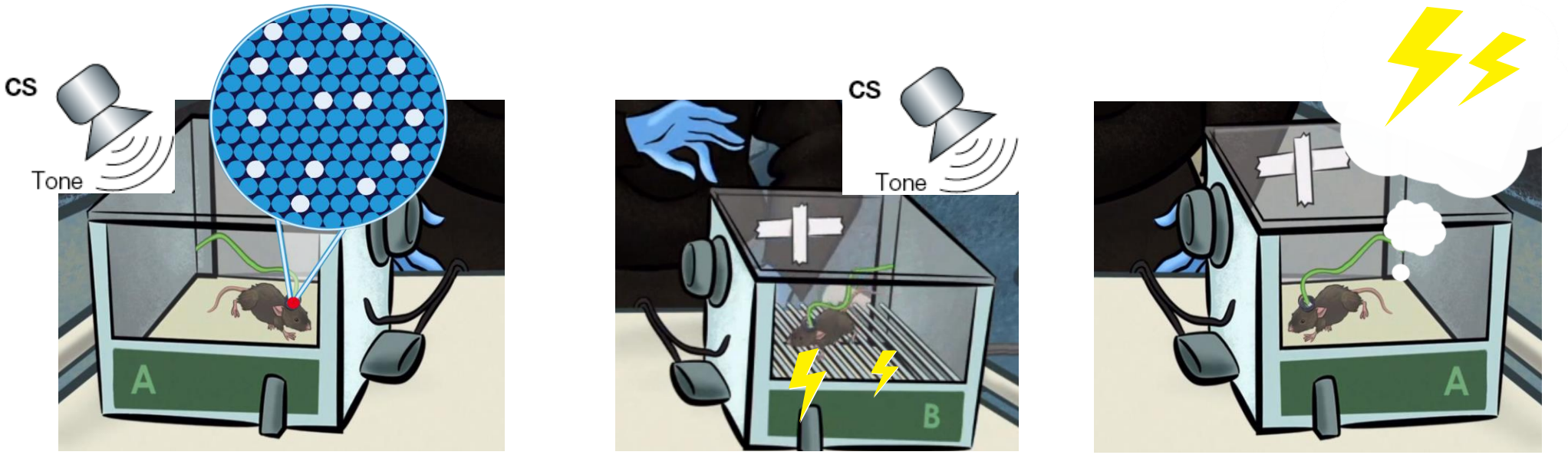
https://archive.org/details/animalintelligen00thor/page/20/mode/2up



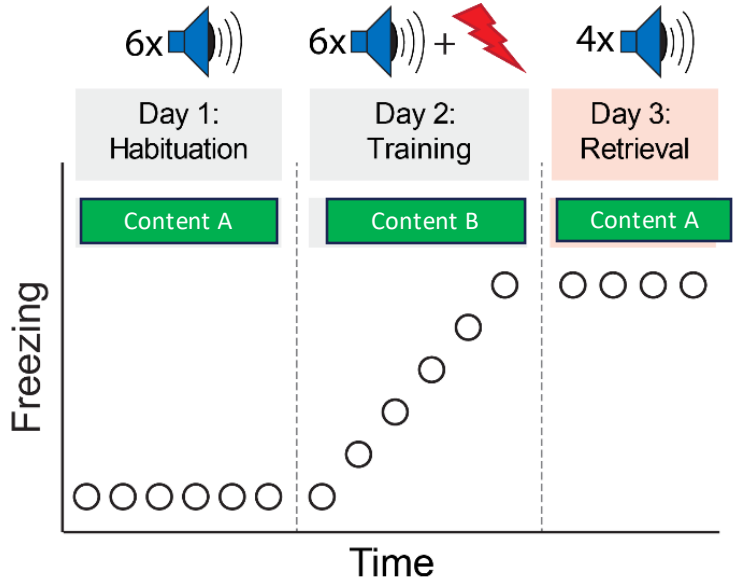
Thorndike: the OG escape room...



Aversive classical conditioning: "fear learning"



Cartoons by Prof. Steve Ramirez (BU)



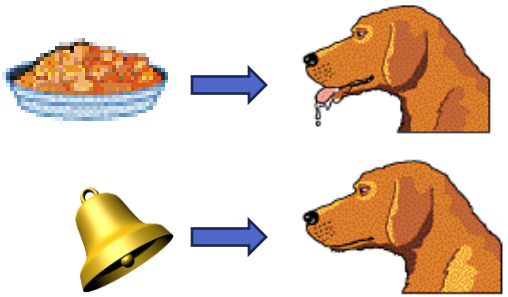
CS: An innocuous sensory stimulus
(tone of 7 kHz, 80 dB, 0.1 s, 30x)

US: A mild electric footshock
Activates nociceptors on the feet and probably other low-threshold mechanoreceptors

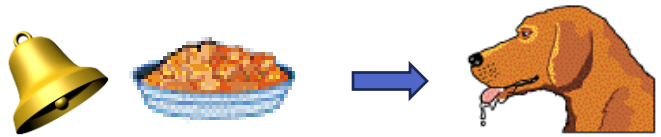
CR: freezing (behavioral immobility)
An evolutionary useful response in the presence of a not clearly present threat

Pavlov's classical conditioning

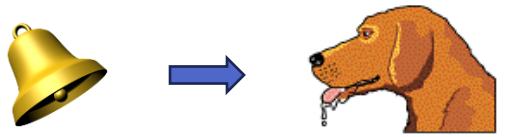
Before Conditioning

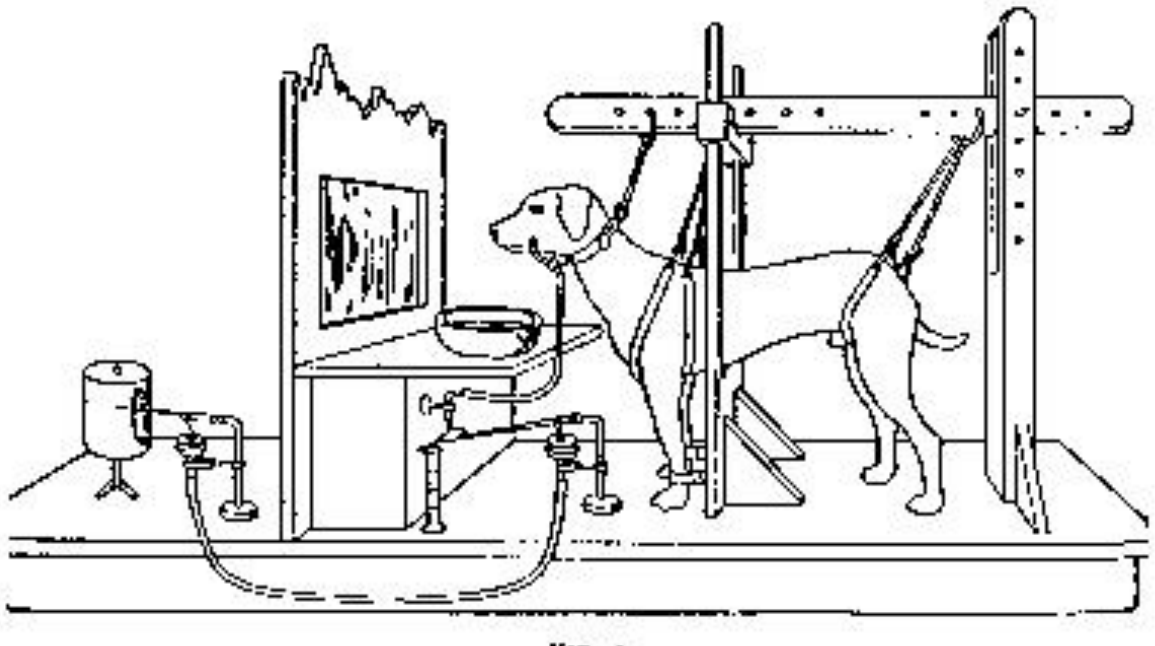


During Conditioning



After Conditioning



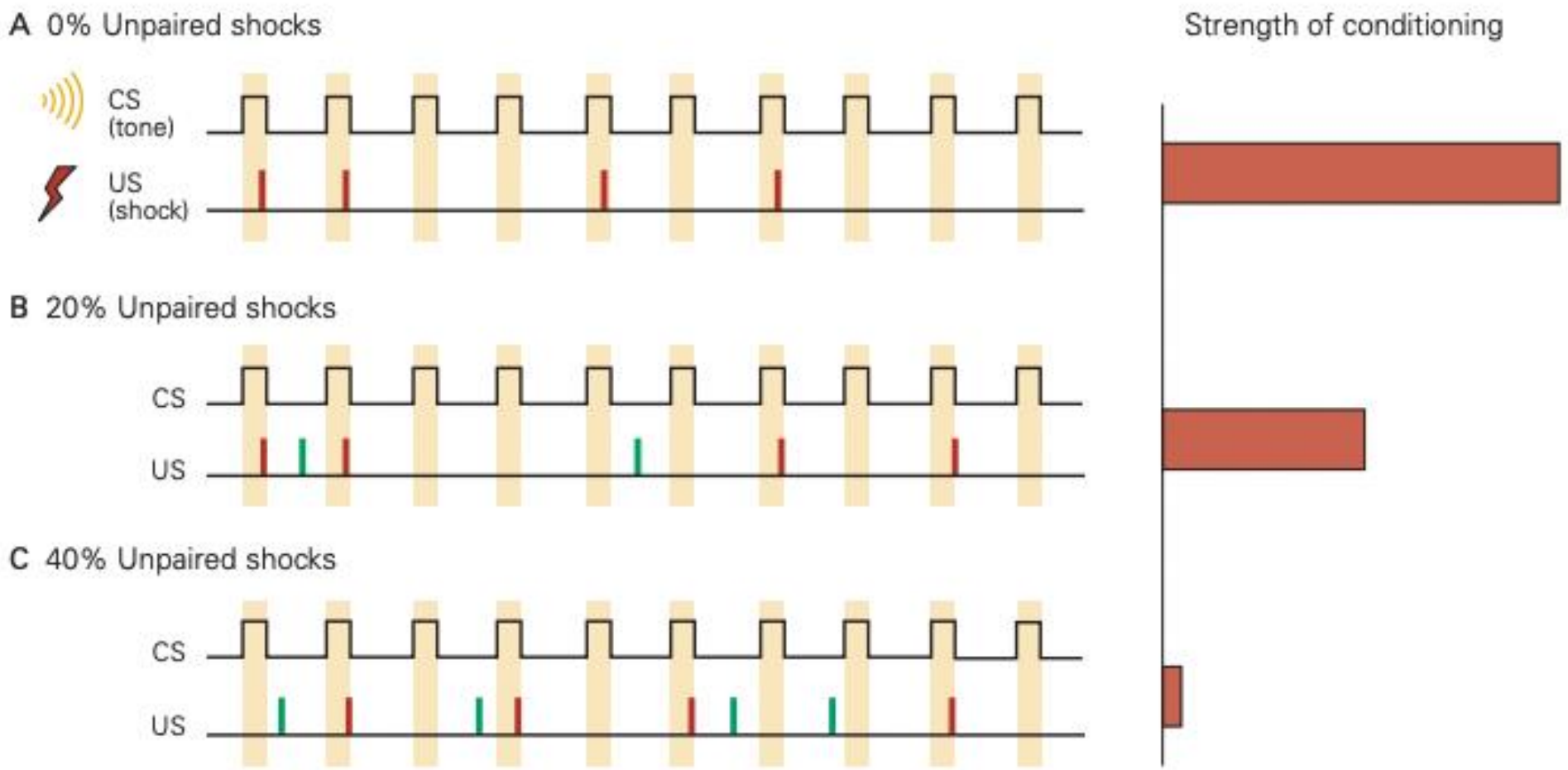


https://en.wikipedia.org/wiki/Classical_conditioning#/media/File:Ivan_Pavlov_research_on_dog's_reflex_setup.jpg

Ten of the more photogenic of Pavlov's dogs. Krasavietz (upper left), Beck, Milkah, Ikar, Joy, Tungus, Arleekin, Ruslan, Toi and Murashka (bottom right). The rest of Pavlov's dogs and their corresponding *Drosophila* memory mutants can be found on the author's webpage at www.cshl.org.

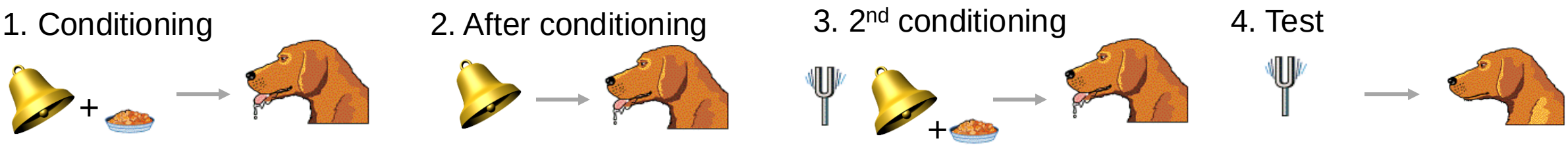
https://www.sciencedirect.com/science/article/pii/S0960982203000666

Classical conditioning depends on degree of stimulus-outcome correlation

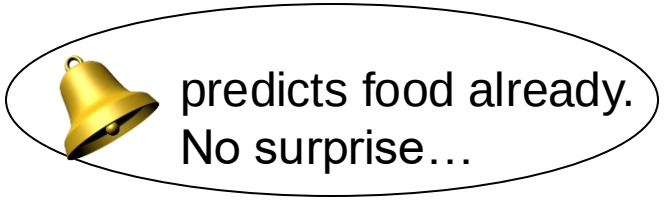


Kandel. Figure 65–12

Kamin's blocking experiment



Kamin, L. J. (1969). Predictability, Surprise, Attention, and Conditioning. In B. A. Campbell, & R. M. Church (Eds.), Punishment Aversive Behavior (pp. 279-296). New York: Appleton- Century-Crofts



"Blocking"

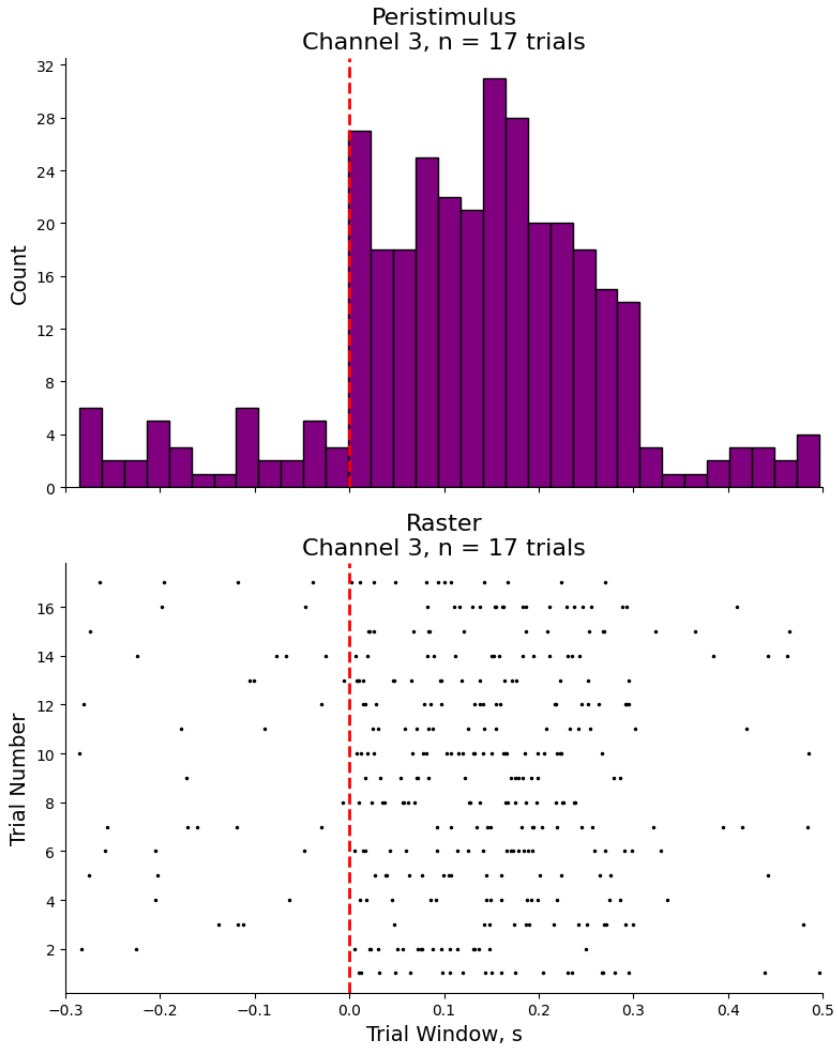
- **Learning occurs only when expectation is violated!**
- *What is the neural basis of this?*

Key concept: peri-stimulus time histogram

The Peri-Stimulus Time Histogram (PSTH) plots the average firing rate of a neuron over time relative to the onset of a stimulus. Here's how it's typically calculated:

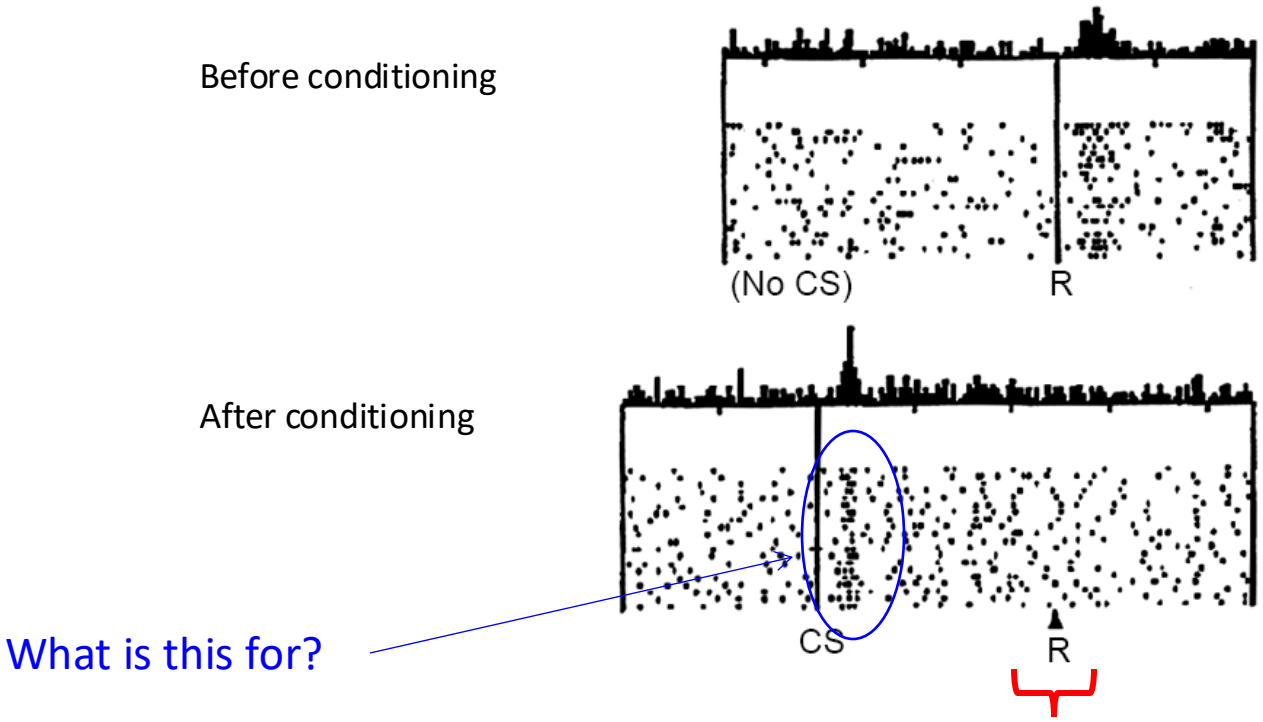
1. Define a time window around the onset of the stimulus.
2. Divide this time window into small bins.
3. Count the number of spikes (action potentials) that occur within each bin across multiple trials.
4. Average the spike counts across trials for each bin.
5. Plot the average spike count (firing rate) for each bin as a function of time.

https://github.com/MMathisLab/Nx-435_EPFL



https://colab.research.google.com/github/MMathisLab/Nx-435_EPFL/blob/main/Notebooks/Demo_PSTH.ipynb

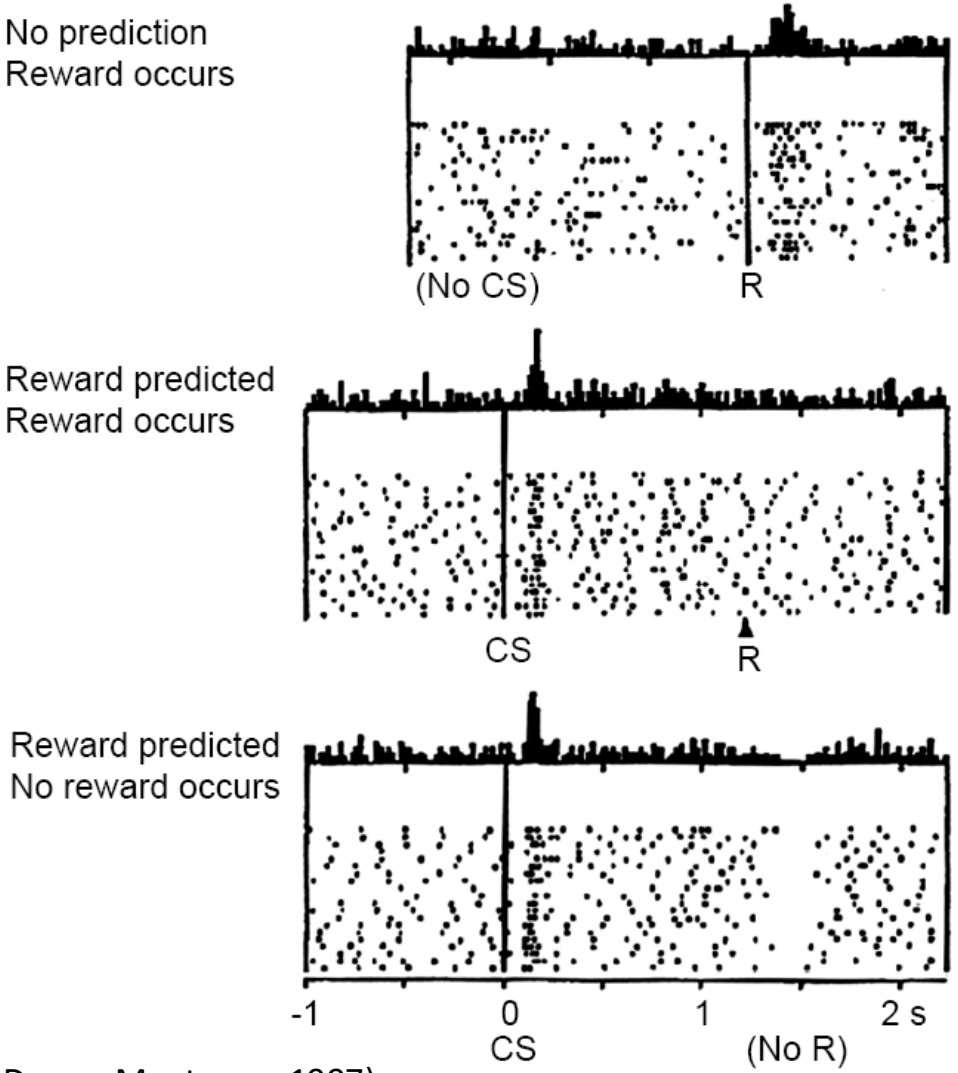
Dopamine neurons in the ventral tegmental area



• Lack of reward responses when the reward was fully predicted

(Schultz, Dayan, Montague, 1997)

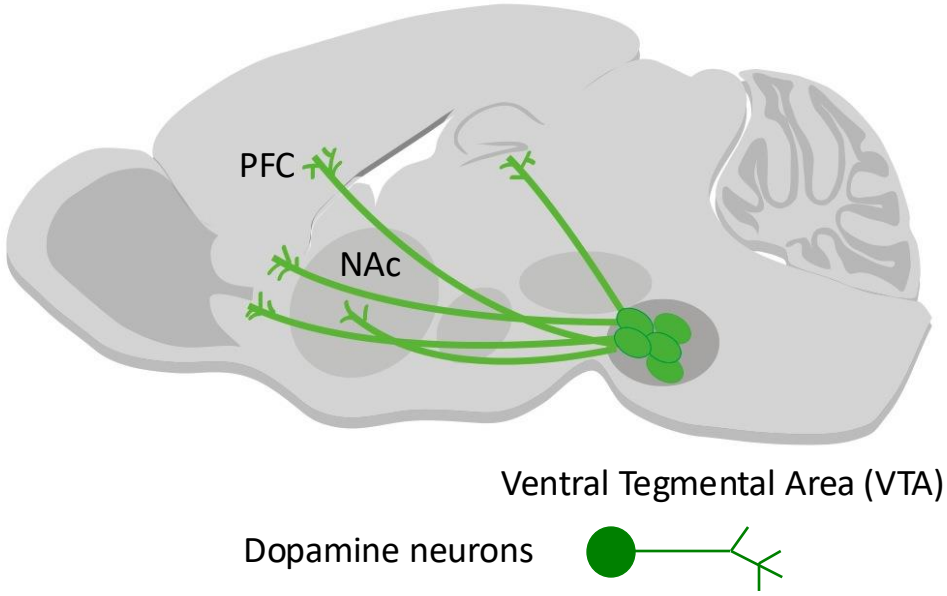
Dopamine as ~~reward~~ temporal difference (TD) error: reward prediction errors!

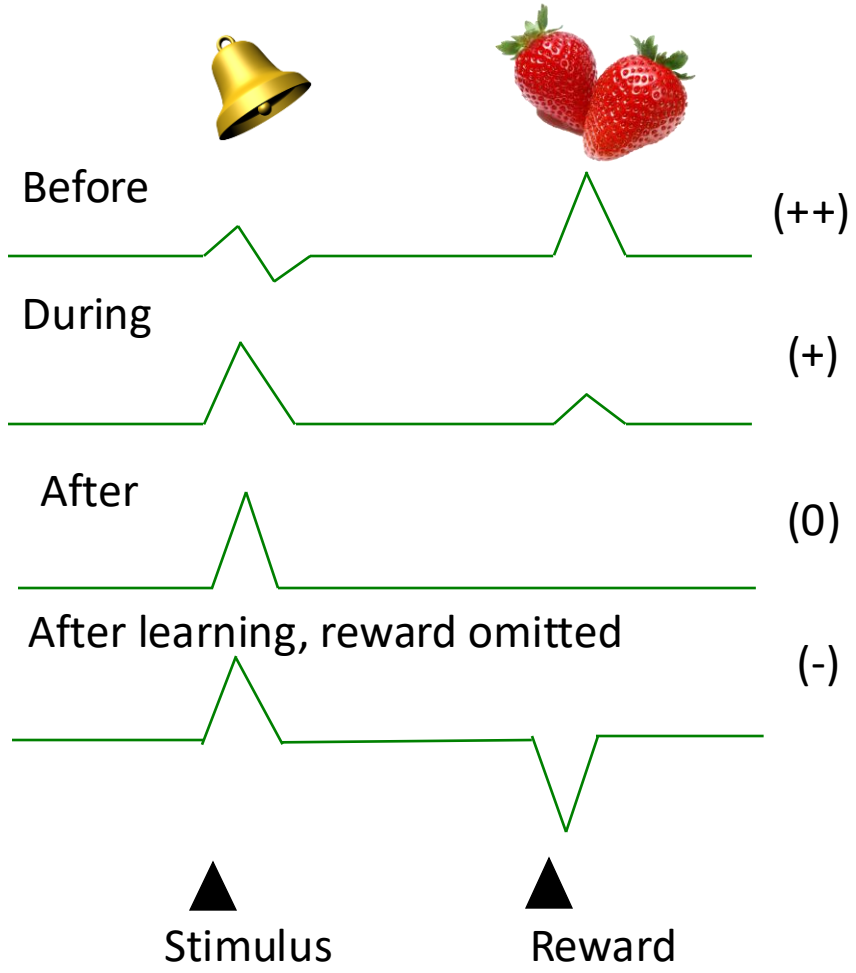


- Dopaminergic (DA) neurons fire phasically (100–500 ms) after unpredicted rewards or cues that predict reward.
- Their response to reward is reduced when a reward is fully predicted (the phasic firing happens at cue presentation).
- DA activity is suppressed when a predicted reward is omitted (negative prediction error).

(Schultz, Dayan, Montague, 1997)

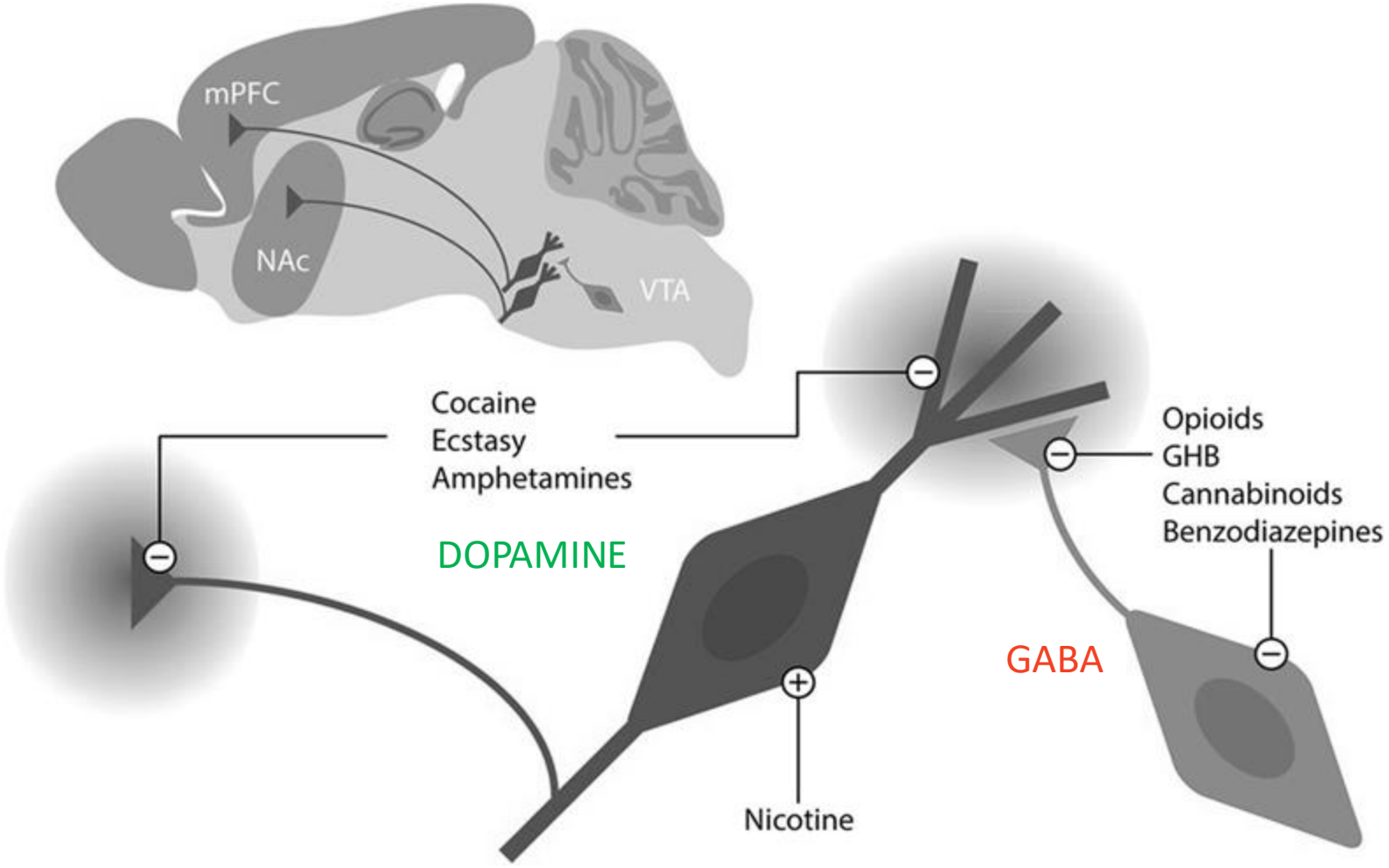
Dopamine circuitry of the brain





- Dopaminergic neurons are ~55–65% of VTA neurons
- The rest are mostly GABAergic inhibitory neurons or Glutamatergic neurons

Dopamine circuitry of the brain: drugs have strong effects



Drug-Evoked Synaptic Plasticity Causing Addictive Behavior

Christian Lüscher

Journal of Neuroscience 6 November 2013, 33 (45) 17641-17646; DOI: https://doi.org/10.1523/JNEUROSCI.3406-13.2013

Addictive drugs cause an increase in mesocorticolimbic dopamine through three distinct cellular mechanisms:

- (1) direct activation of dopamine neurons (e.g., nicotine)
- (2) indirect disinhibition of dopamine neurons [opioids, gamma-hydroxybutyric acid (GHB), cannabinoids, and benzodiazepines]
- (3) interference with dopamine reuptake (cocaine, ecstasy, and amphetamines).

Dopamine neurons in VTA are also involved in motor learning & Parkinson's disease

Dopaminergic meso-cortical projections to M1: role in motor learning and motor cortex plasticity

*Jonas A. Hosp[1,2] and Andreas R. Luft[1,2,3]**

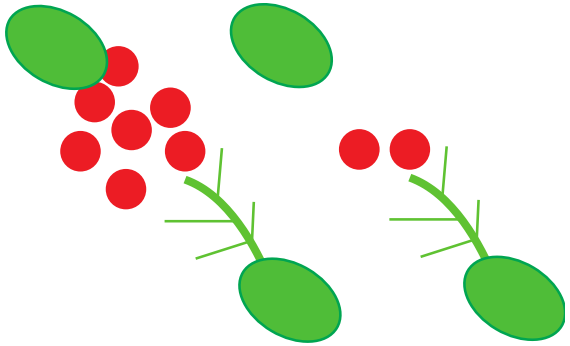
Hosp & Luft, Frontiers in Neurology, 2013

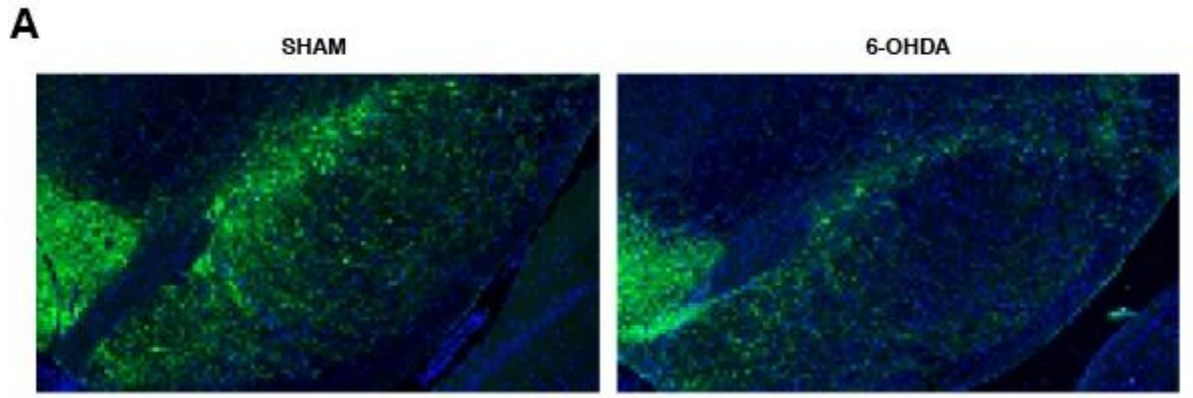
Dopaminergic Projections from Midbrain to Primary Motor Cortex Mediate Motor Skill Learning

Jonas A. Hosp,[1,2*] Ana Pekanovic,[1,2*] Mengia S. Rioult-Pedotti,[1,2,3] and Andreas R. Luft[1,2,4]

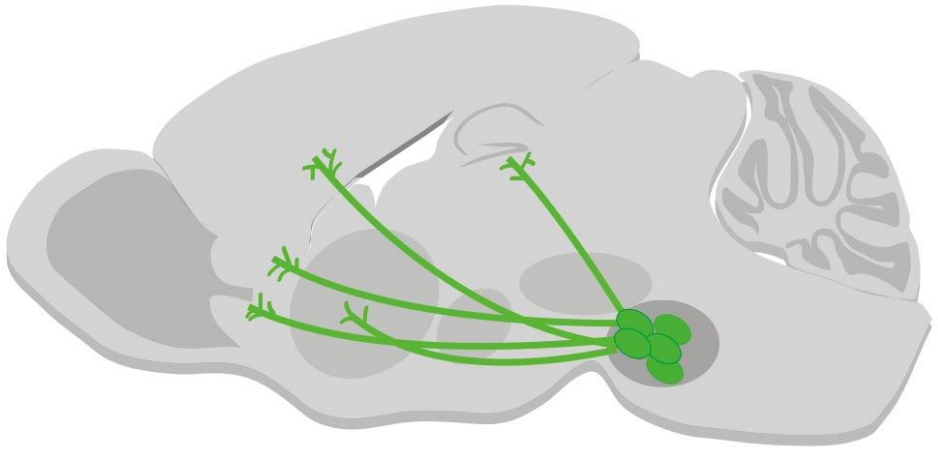
Hosp et al, J. Neuroscience, 2011

Parkinson's disease: reduced DA





https://www.criver.com/products-services/discovery-services/pharmacology-studies/neuroscience-models-assays/parkinsons-disease-studies/vivo-models-parkinsons-disease?region=3696

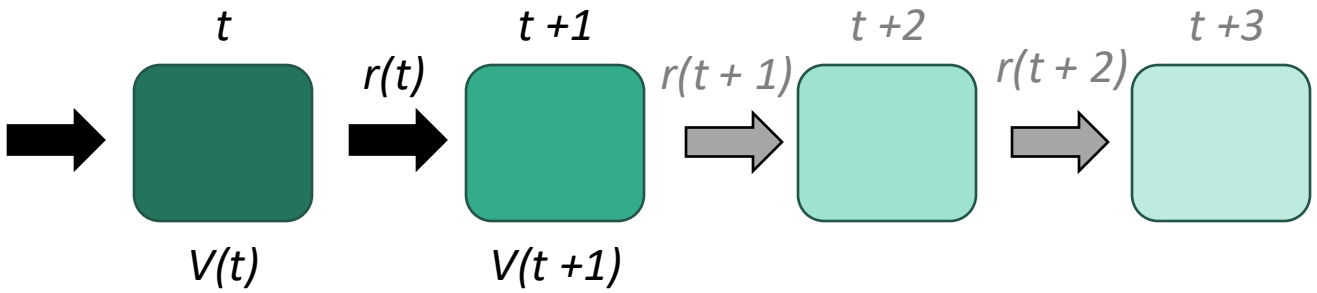


Thus, the role of dopamine in the brain is complex, and there are many hypotheses about its various roles:

- Reward Prediction Errors ✅ *(but of course this can be implemented in many ways…)*
- Role in saliency/attentional modulation
- Uncertainty estimation
- Energizing/motivating behavior

Temporal difference (TD) error reinforcement learning (RL)

States



t = time
r = reward
V(t) = value

Value is the discounted sum of all future rewards!

$$V(t) = r(t) + \gamma * r(t + 1) + \gamma^2 * r(t + 2) + ...$$

$$V(t) = r(t) + \gamma * V(t +1)$$

Discounting factor

$0 < \gamma < 1$

Temporal difference prediction error:

$$\delta(t) = r(t) + \gamma * \hat{V}(t +1) - \hat{V}(t)$$

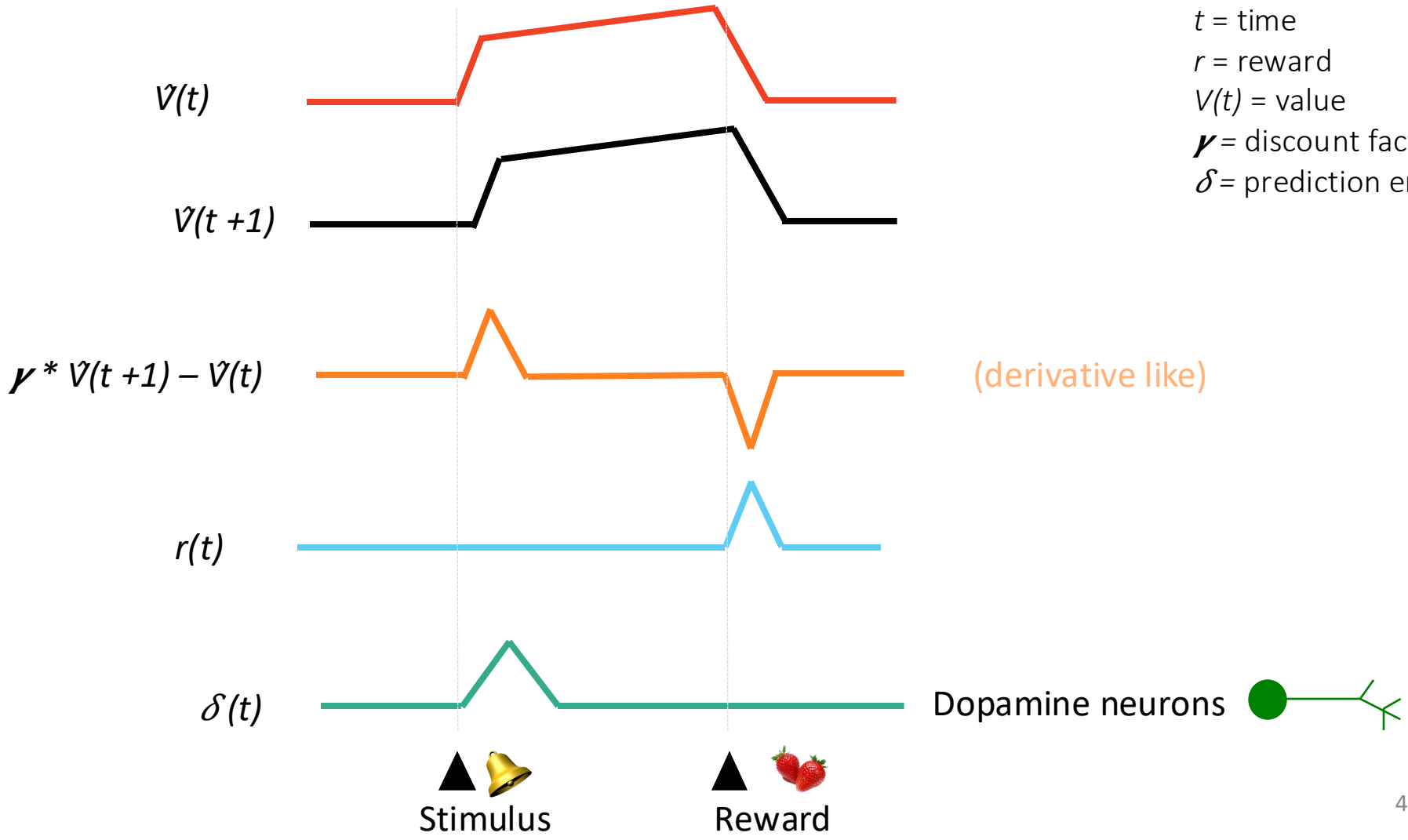
Update *V(t)*

$$\hat{V}(t) \leftarrow \hat{V}(t) + \alpha * \delta$$

Sutton & Barto 1996, Schultz et al. 1997, Cohen et al. 2012

How could a system encode a temporal difference (TD) error

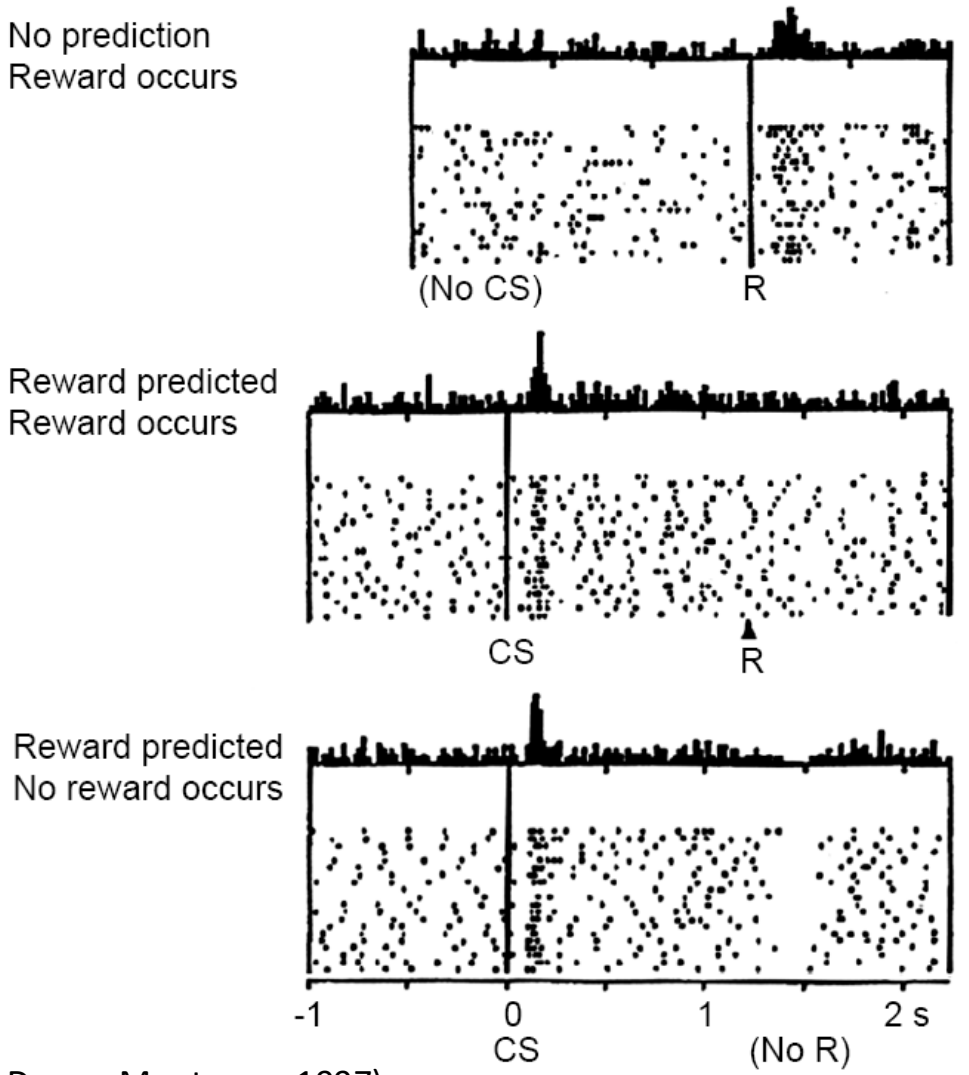
TD error as a derivative-like computation: $\delta(t) = r(t) + \gamma * \hat{V}(t+1) - \hat{V}(t)$
(neurally doable!)



t = time
r = reward
$V(t)$ = value
γ = discount factor
δ = prediction error

REMINDER: receiving a reward with a magnitude below the mean reward will elicit a negative RPE, whereas larger magnitudes will elicit a positive RPE!

(if a LARGER reward was given it would elicit a DA response at the time of reward)



(Schultz, Dayan, Montague, 1997)

Letter | Published: 18 January 2012

Neuron-type-specific signals for reward and punishment in the ventral tegmental area

Jeremiah Y. Cohen, Sebastian Haesler, Linh Vong, Bradford B. Lowell & Naoshige Uchida

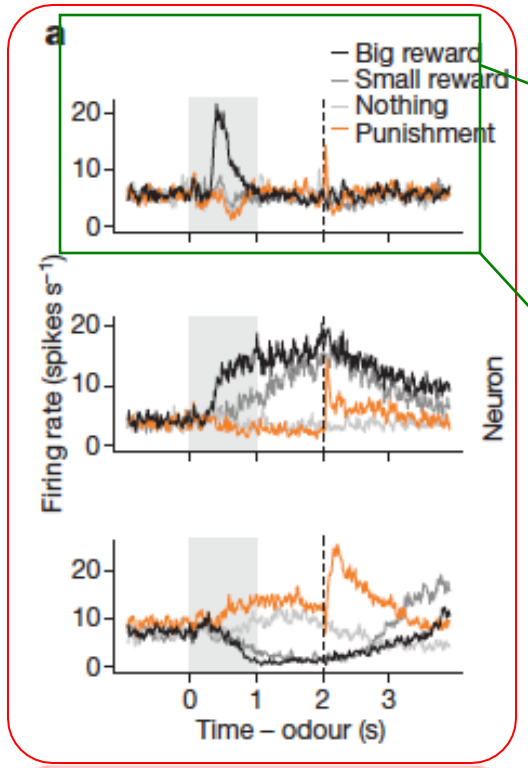
Nature **482**, 85–88 (2012) | Cite this article

3 functional "types"

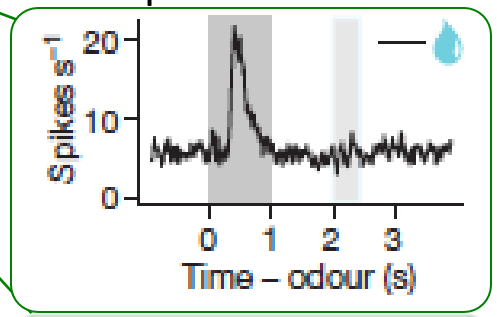
52%

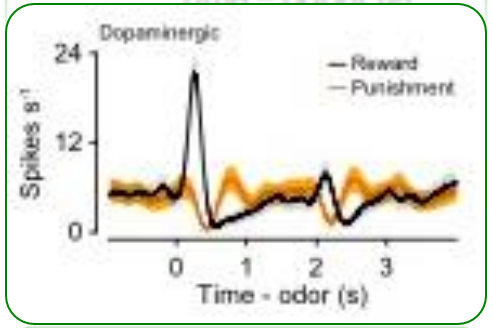
31%

18%



Optogenetically identified Dopamine Neuron





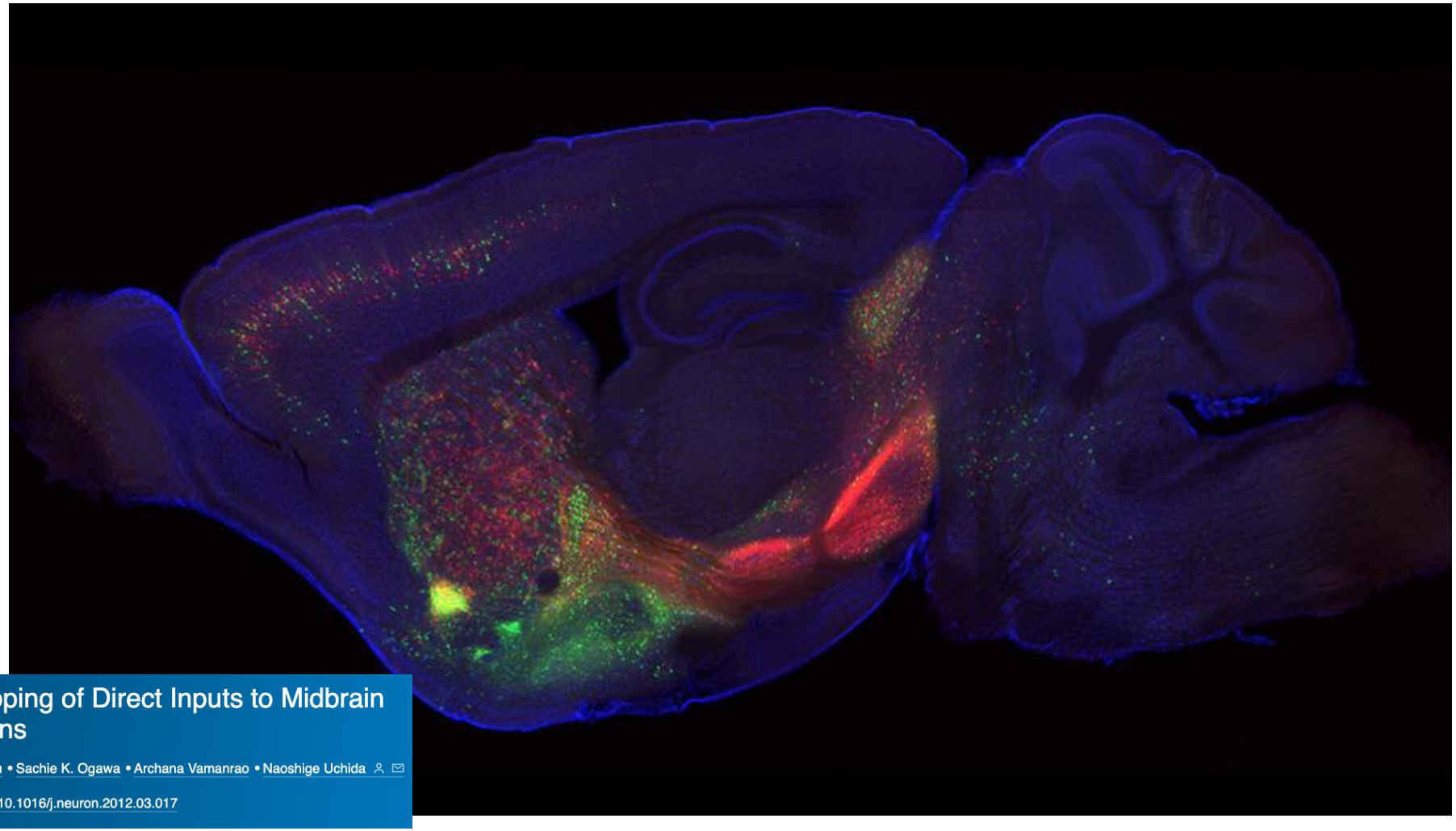
Light-identified dopamine neurons encode RPE in mice

Cohen*, Haesler*, et al, Nature 2012

Cohen, Amoroso, Uchida, eLife 2015

What cells could enable this computation?

- **Hypothesis:** Neurons within the VTA could encode all needed features or neurons that project TO DA neurons could contribute

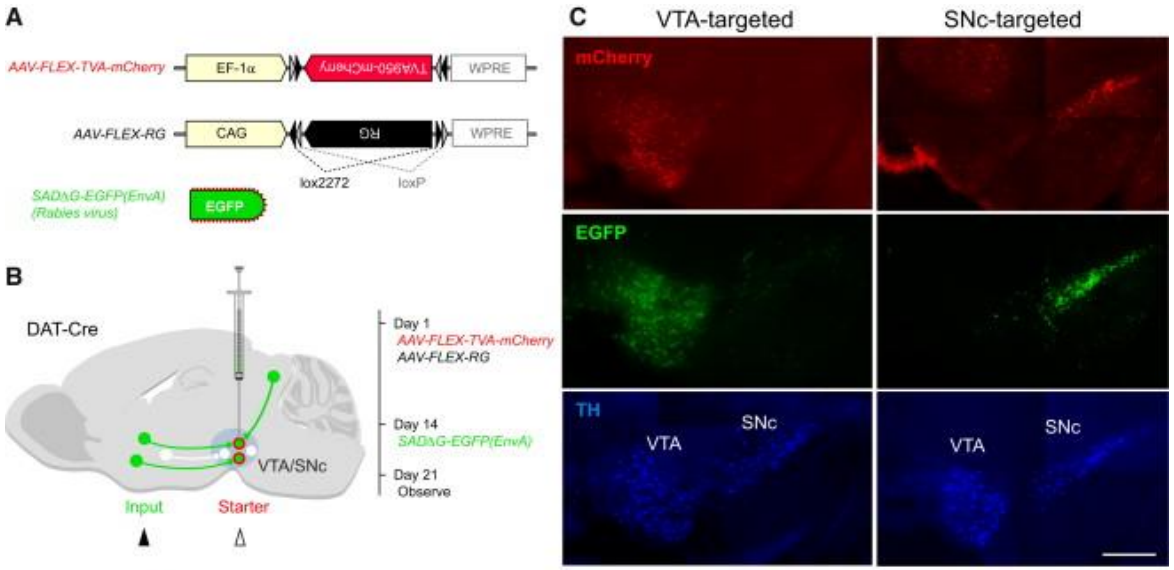


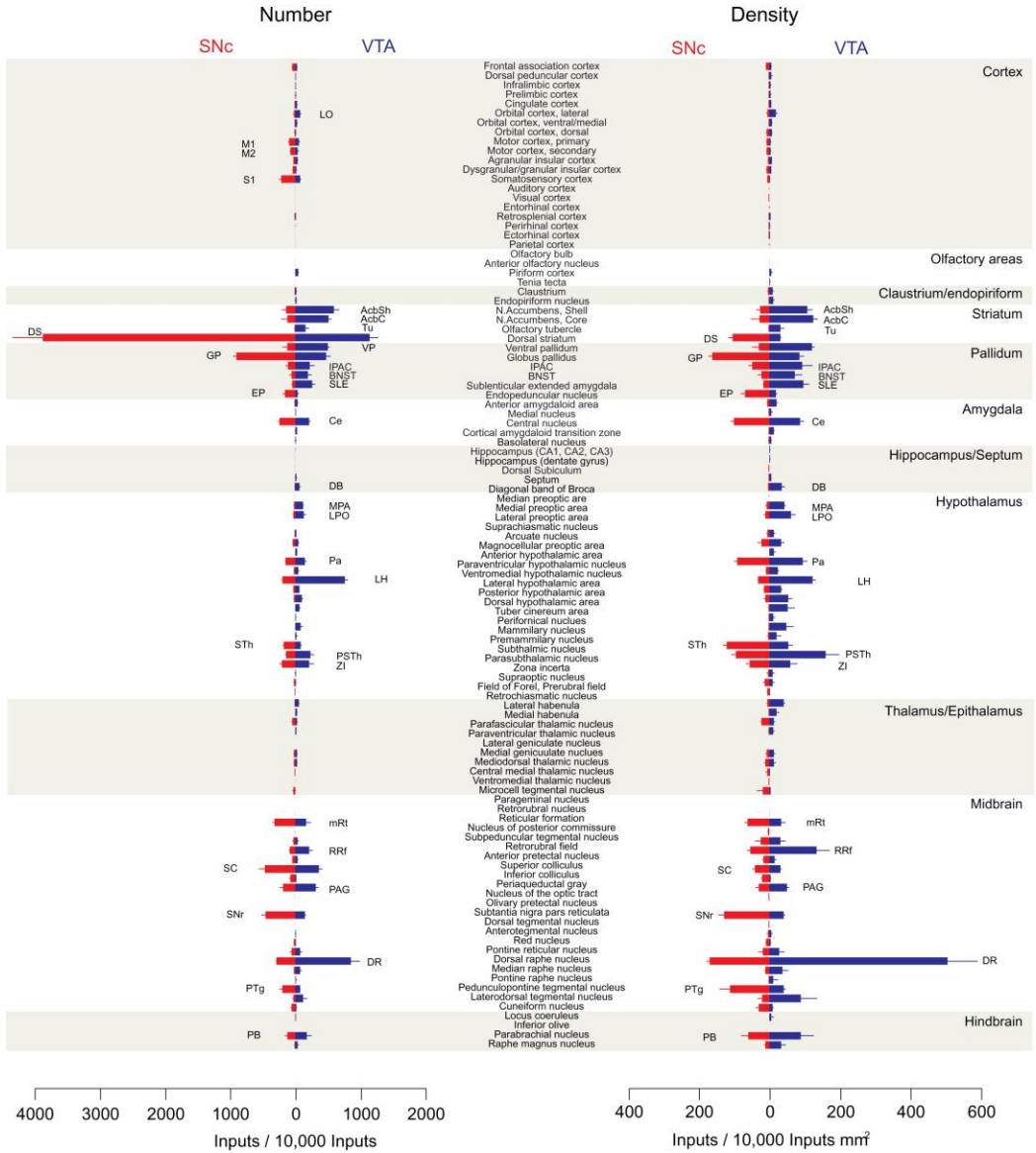
Whole-Brain Mapping of Direct Inputs to Midbrain Dopamine Neurons

Mitsuko Watabe-Uchida • Lisa Zhu • Sachie K. Ogawa • Archana Vamanrao • Naoshige Uchida

Open Archive • DOI: https://doi.org/10.1016/j.neuron.2012.03.017

Comparison of Input Areas between VTA and SNc Dopamine Neurons



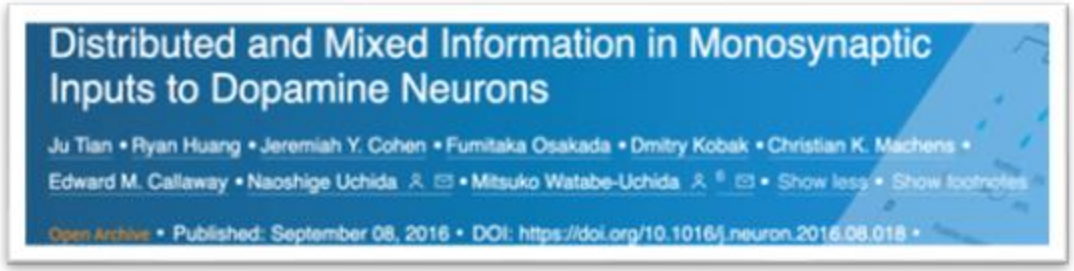


Whole-Brain Mapping of Direct Inputs to Midbrain Dopamine Neurons

Mitsuko Watabe-Uchida • Lisa Zhu • Sachie K. Ogawa • Archana Vamanrao • Naoshige Uchida

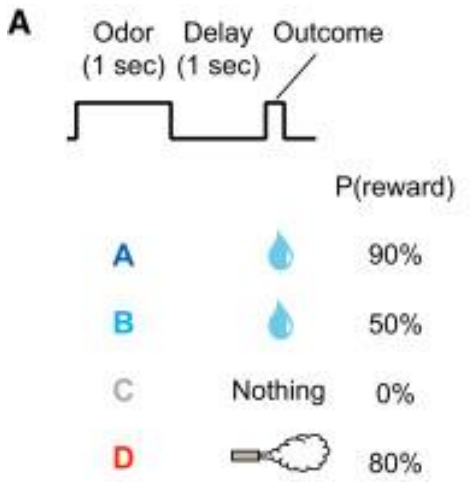
Open Archive • DOI: https://doi.org/10.1016/j.neuron.2012.03.017

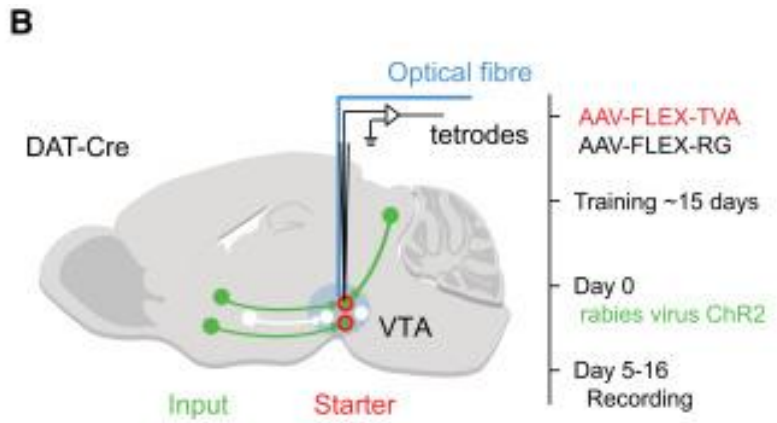
What are the inputs to Dopamine neurons encoding?

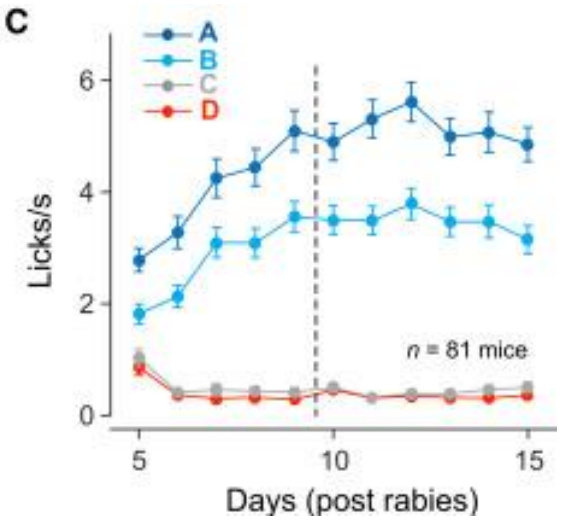


- Electrophysiological recording from monosynaptic inputs of dopamine neurons was performed
- Rabies virus tracing was combined with optogenetic tagging in awake recording

→They recorded from input-dense areas along this ventral stream that have most often been used in RPE models:
→ the ventral and dorsal striatum
→Ventral
→ lateral hypothalamus
→ subthalamic nucleus
→rostromedial tegmental nucleus (RMTg)
→pedunculopontine tegmental nucleus (PPTg)







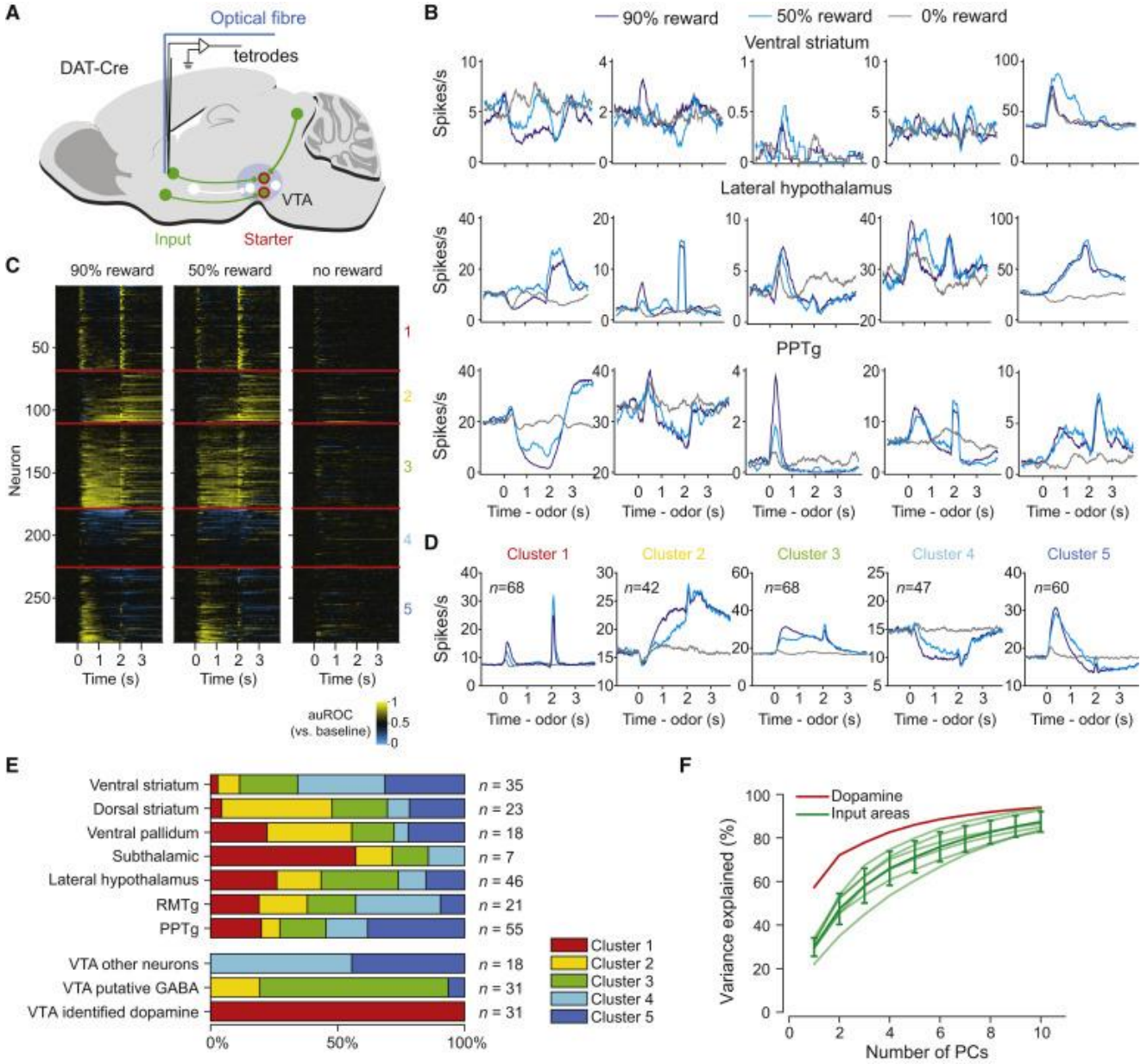
What are the inputs to Dopamine neurons encoding?

Distributed and Mixed Information in Monosynaptic Inputs to Dopamine Neurons

Ju Tian • Ryan Huang • Jeremiah Y. Cohen • Fumitaka Osakada • Dmitry Kobak • Christian K. Machens • Edward M. Callaway • Naoshige Uchida • Mitsuko Watabe-Uchida • Show less • Show footnotes

Open Archive • Published: September 08, 2016 • DOI: https://doi.org/10.1016/j.neuron.2016.08.018 •

- Information required to compute reward prediction errors (RPEs) was distributed!
- There are mixed representations of variables and partially computed RPEs in input neurons



Article

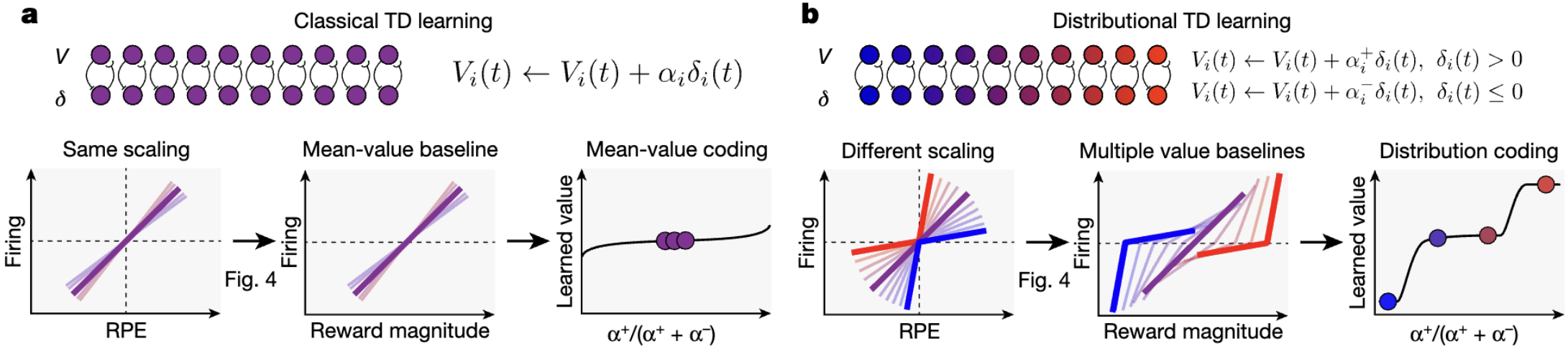
A distributional code for value in dopamine-based reinforcement learning

https://doi.org/10.1038/s41586-019-1924-6
Received: 3 January 2019
Accepted: 19 November 2019
Published online: 15 January 2020

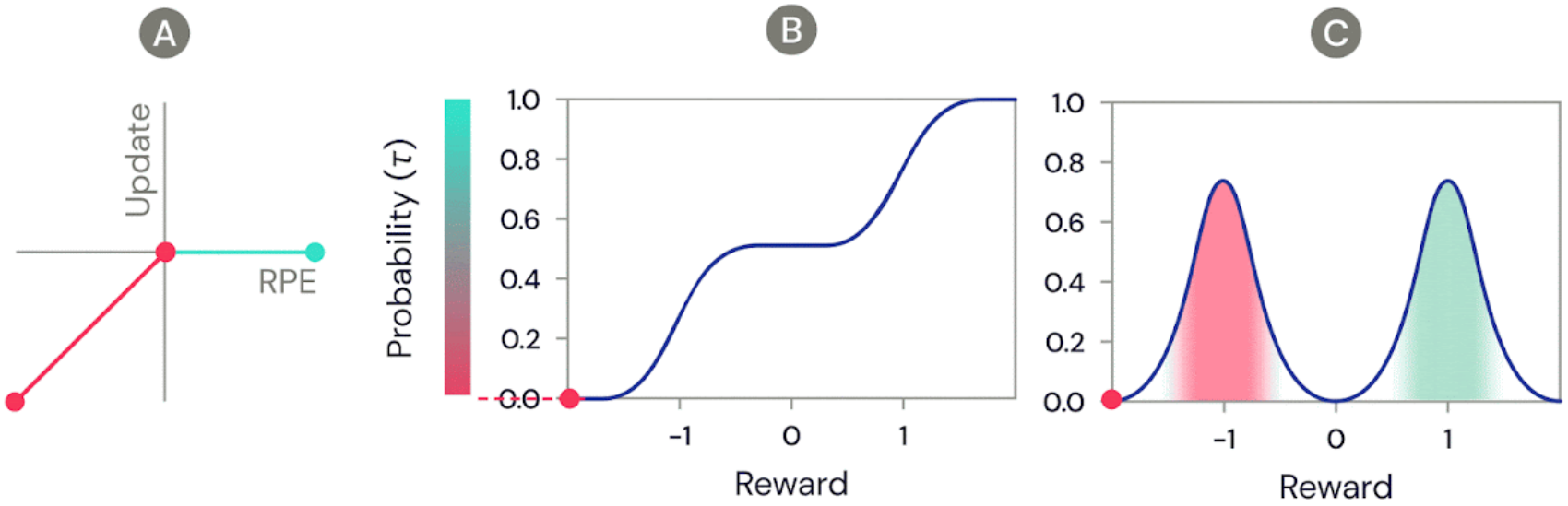
Will Dabney[1,6*], Zeb Kurth-Nelson[1,2,5], Naoshige Uchida[3], Clara Kwon Starkweather[3], Demis Hassabis[1], Rémi Munos[1] & Matthew Botvinick[1,4,5]

Since its introduction, the reward prediction error theory of dopamine has explained a wealth of empirical phenomena, providing a unifying framework for understanding

In contrast to classical temporal-difference (TD) learning, distributional RL posits a diverse set of RPE channels, each of which carries a different value prediction, with varying degrees of optimism across channels!



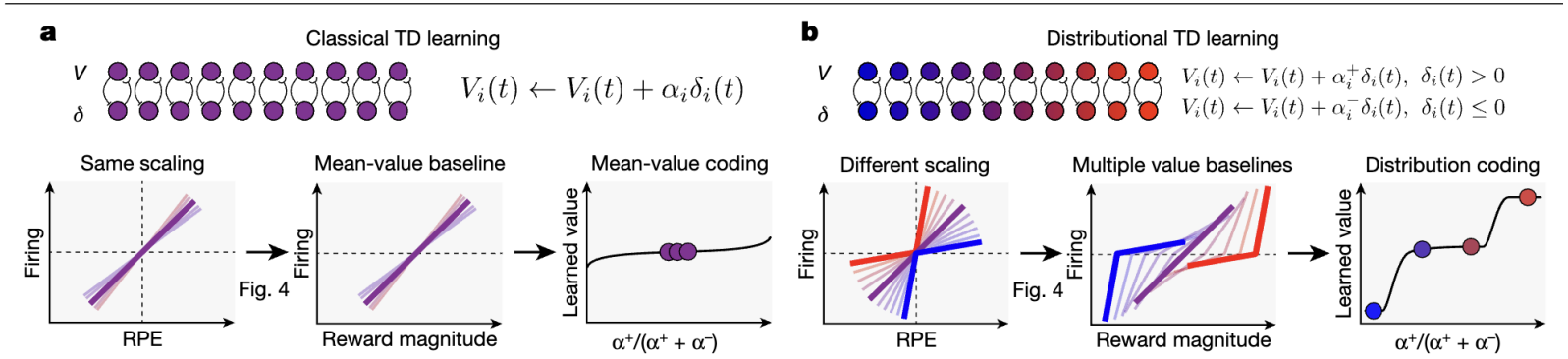
*In standard TD learning, all value predictors learn the same value V. Each dopamine cell is assumed to have the same relative scaling for positive and negative RPEs. This causes each value prediction (or value baseline) to be the mean of the outcome distribution.



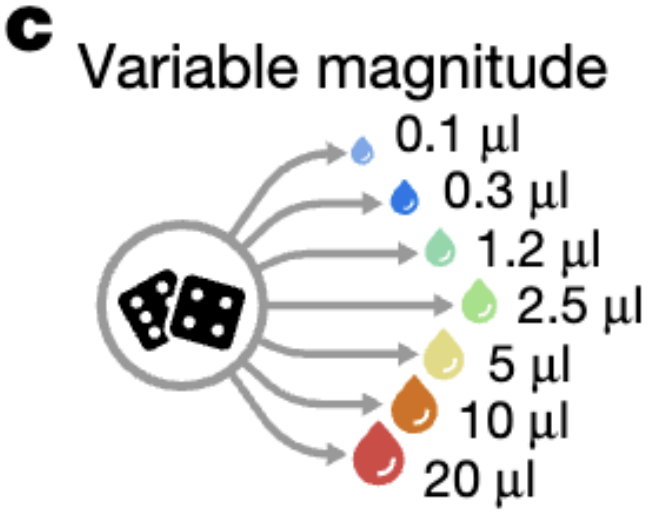
Distributional TD learns value estimates for many different parts of the distribution of rewards. Which part a particular estimate covers is determined by the type of asymmetric update applied to that estimate. (A) A 'pessimistic' cell would amplify negative updates and ignore positive updates, an 'optimistic' cell would amplify positive updates and ignore negative updates. (B) This results in a diversity of pessimistic or optimistic value estimates, shown here as points along the cumulative distribution of rewards, that capture (C) the full distribution of rewards.

https://deepmind.google/discover/blog/dopamine-and-temporal-difference-learning-a-fruitful-relationship-between-neuroscience-and-ai/

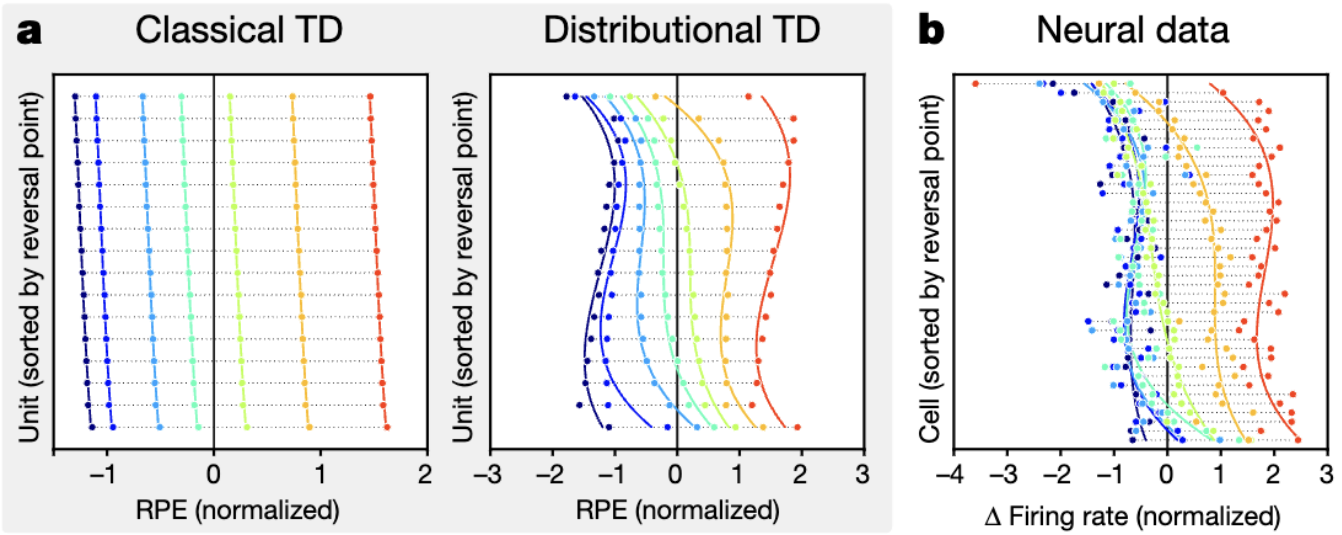
Hypothesis: DA neurons firing during complex tasks are better fit by D-TD



On each trial, a variable reward size is given:



They simulate C-TD vs. D-TD and compare to real data:



Summary:

- Marr's 3 levels provide a computational formulation for studying computations in the brain
- Decision-making is hard: the "credit assignment problem", delayed rewards, uncertain outcomes
- Perceptual and value-based decision-making can help refine how to study and where in the brain to study
  - → reminder for the neuro-anatomy that supports visually guided decisions
  - → Encoding & decoding is critical
- Decision variables DV), evidence accumulation, and how to use decoding to closed-loop test how DV are related to actions
- → Change of mind in decisions – how did they test this?
- Operant and classical conditioning
- PSTH
- Dopamine (DA) neurons in VTA
- RPEs
- RL & TD learning
- How to formalize finding computations: mapping TD to DA
- Iputs to DA neurons show distributed information and even (possibly) partially computed RPEs
- Distributional RL in the DA population better fits the data