7. Motor Control and Reinforcement Learning

Outline

A. Action Selection and Reinforcement
B. Temporal Difference Reinforcement Learning
C. PVLV Model
D. Cerebellum and Error-driven Learning

Sensory-Motor Loop

- Why animals have nervous systems but plants do not: *animals move*
  - — a nervous system is needed to coordinate the movement of an animal's body
  - — movement is fundamental to understanding cognition
- Perception conditions action
- Action conditions perception
  - — profound effect of action on structuring perception is often neglected

Overview

- Subcortical areas:
  - basal ganglia
    - reinforcement learning (reward/punishment)
    - connections to “what” pathway
  - cerebellum
    - error-driven learning
    - connections to “how” pathway
  - disinhibitory output dynamic
- Cortical areas:
  - frontal cortex
    - connections to basal ganglia & cerebellum
  - parietal cortex
    - maps sensory information to motor outputs
    - connections to cerebellum

Learning Rules Across the Brain

	Learning Signal			Dynamics		
Area	**Reward**	**Error**	**Self Org**	**Separator**	**Integrator**	**Attractor**
Primitive						
Basal Ganglia	+++	- - -	- - -	++	-	- - -
Cerebellum	- - -	+++	- - -	+++	- - -	- - -
Advanced						
Hippocampus	+	+	+++	+++	- - -	+++
Neocortex	++	+++	++	- - -	+++	+++

+ = has to some extent … +++ = defining characteristic – definitely has
- = not likely to have … - - - = definitely does not have

(slide < O'Reilly)

Primitive, Basic Learning…

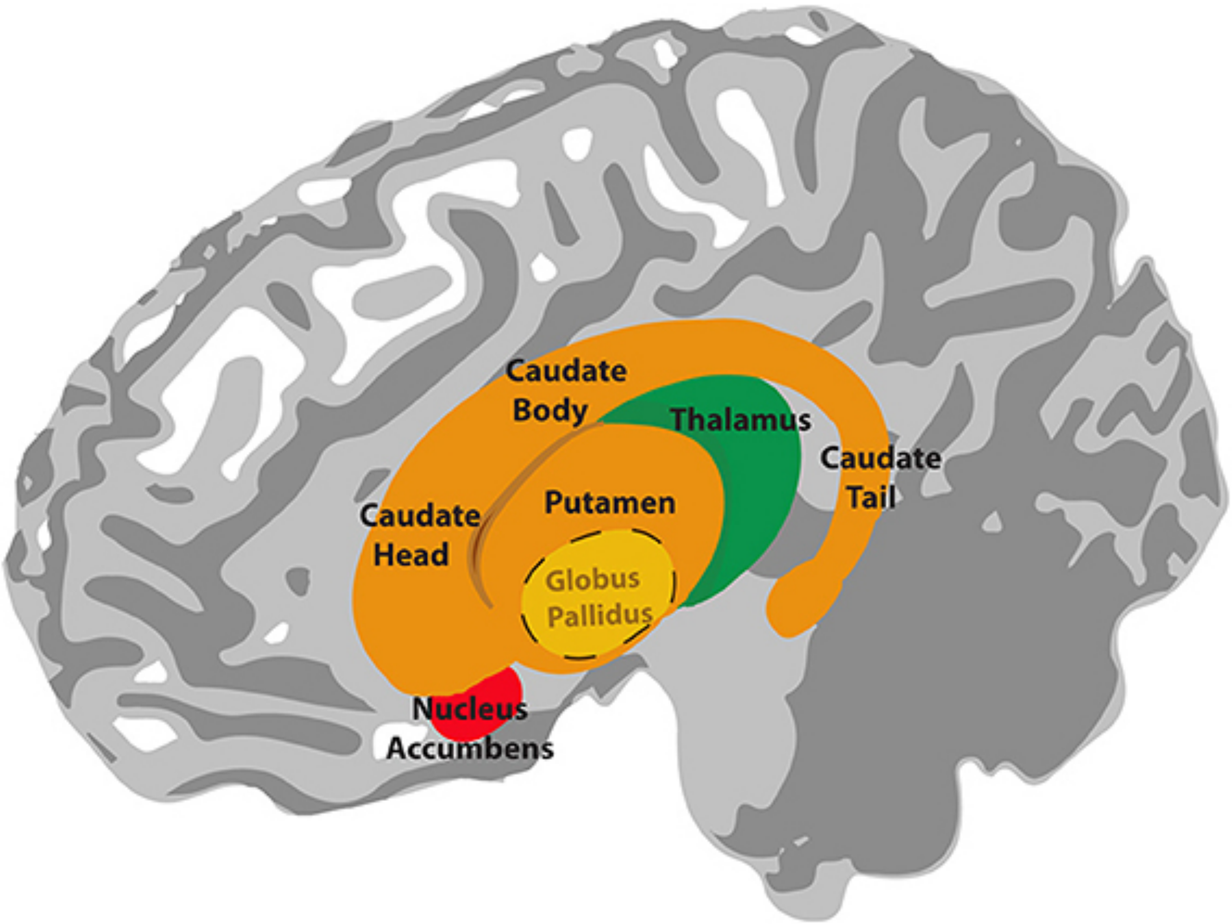
	Learning Signal			*Dynamics*		
Area	**Reward**	**Error**	**Self Org**	**Separator**	**Integrator**	**Attractor**
Primitive Basal Ganglia	+++	- - -	- - -	++	-	- - -
Cerebellum	- - -	+++	- - -	+++	- - -	- - -

- Reward & Error = most basic learning signals (self organized learning is a luxury…)
- Simplest general solution to any learning problem is a *lookup table* = separator dynamics

(slide < O'Reilly)

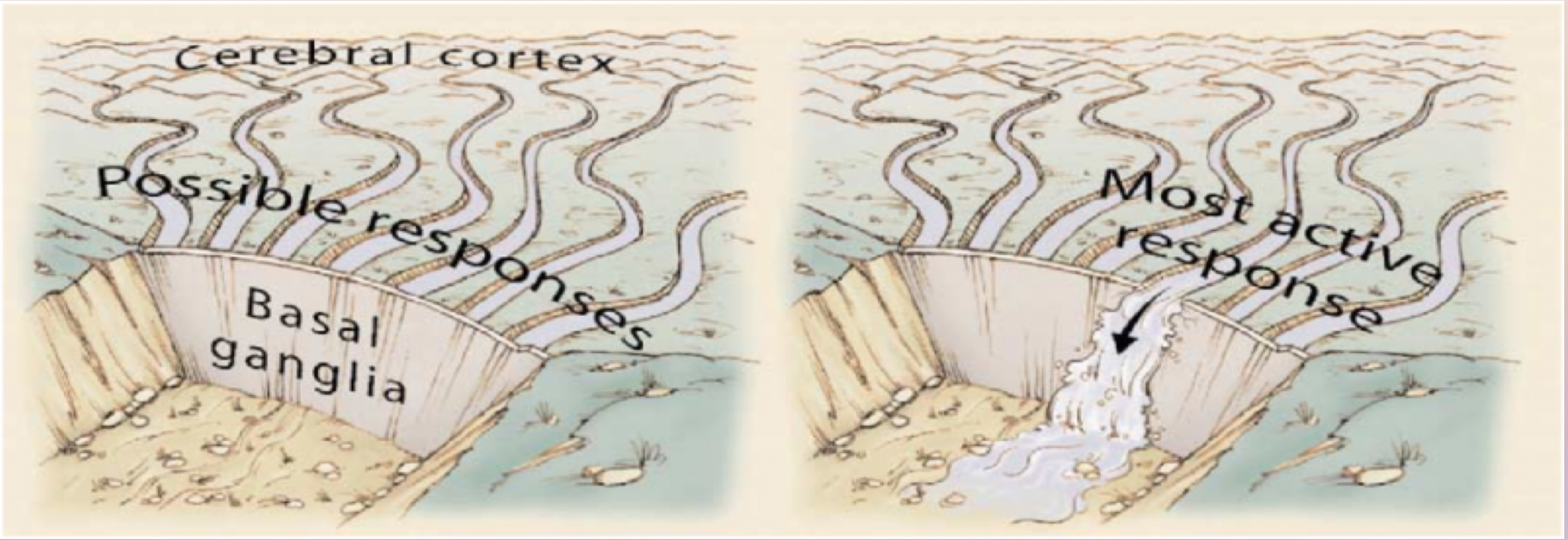
A. Action Selection and Reinforcement

Anatomy of Basal Ganglia



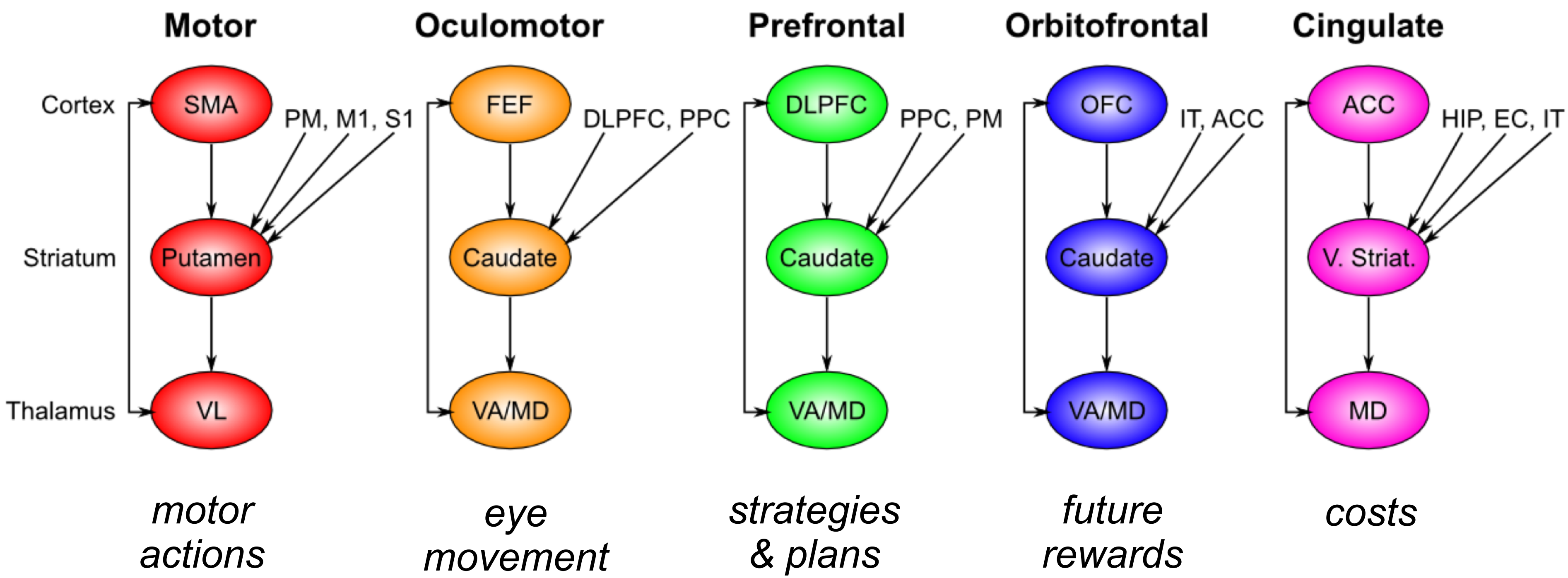
Lim S-J, Fiez JA and Holt LL - Lim S-J, Fiez JA and Holt LL (2014) How may the basal ganglia contribute to auditory categorization and speech perception? Front. Neurosci. 8:230. doi: 10.3389/fnins.2014.00230
http://journal.frontiersin.org/article/10.3389/fnins.2014.00230/full

Basal Ganglia and Action Selection



(slide < O'Reilly)

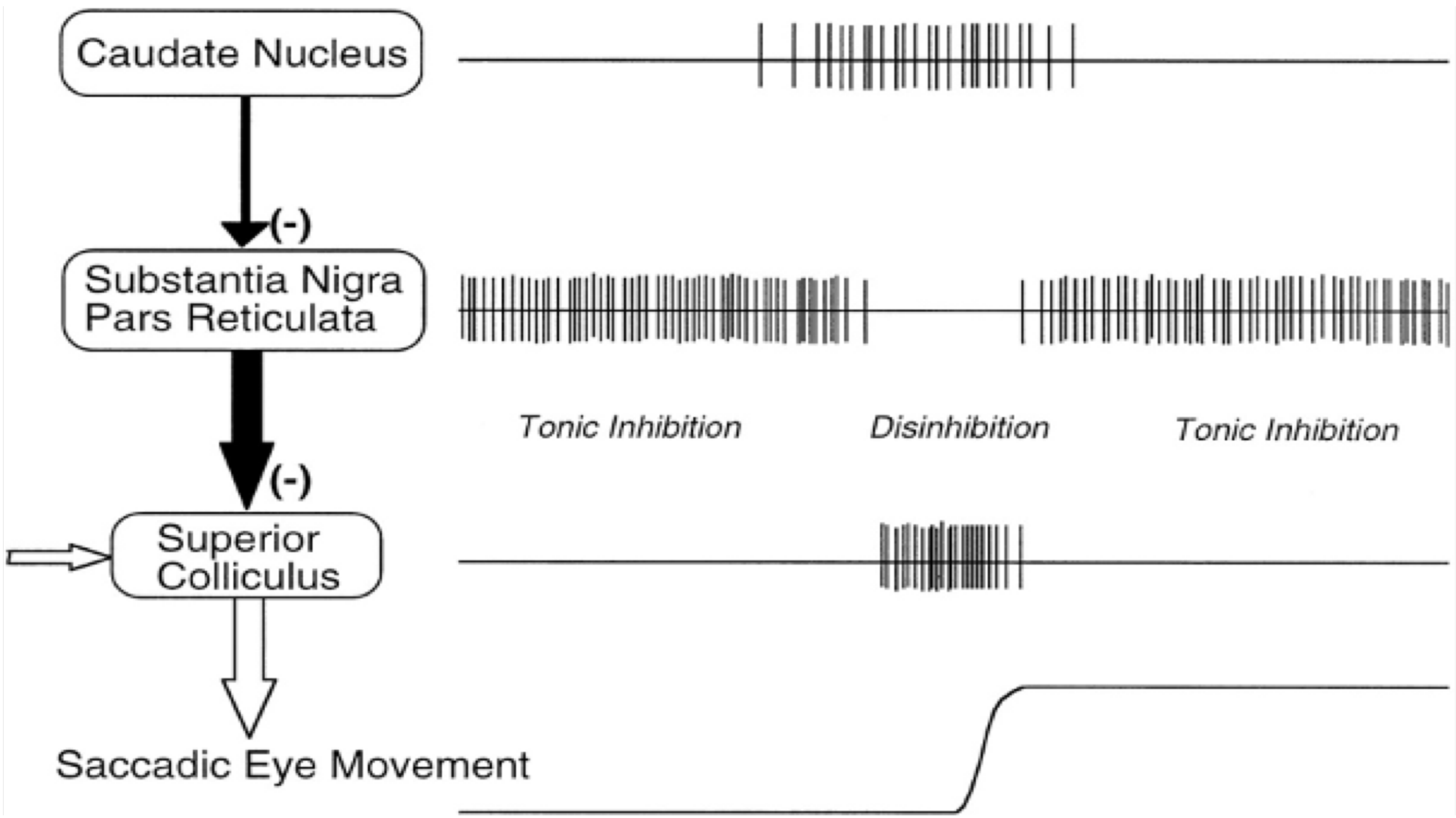
Basal Ganglia: Action Selection



- Parallel circuits select motor actions and “cognitive” actions across frontal areas

(slide based on O’Reilly)

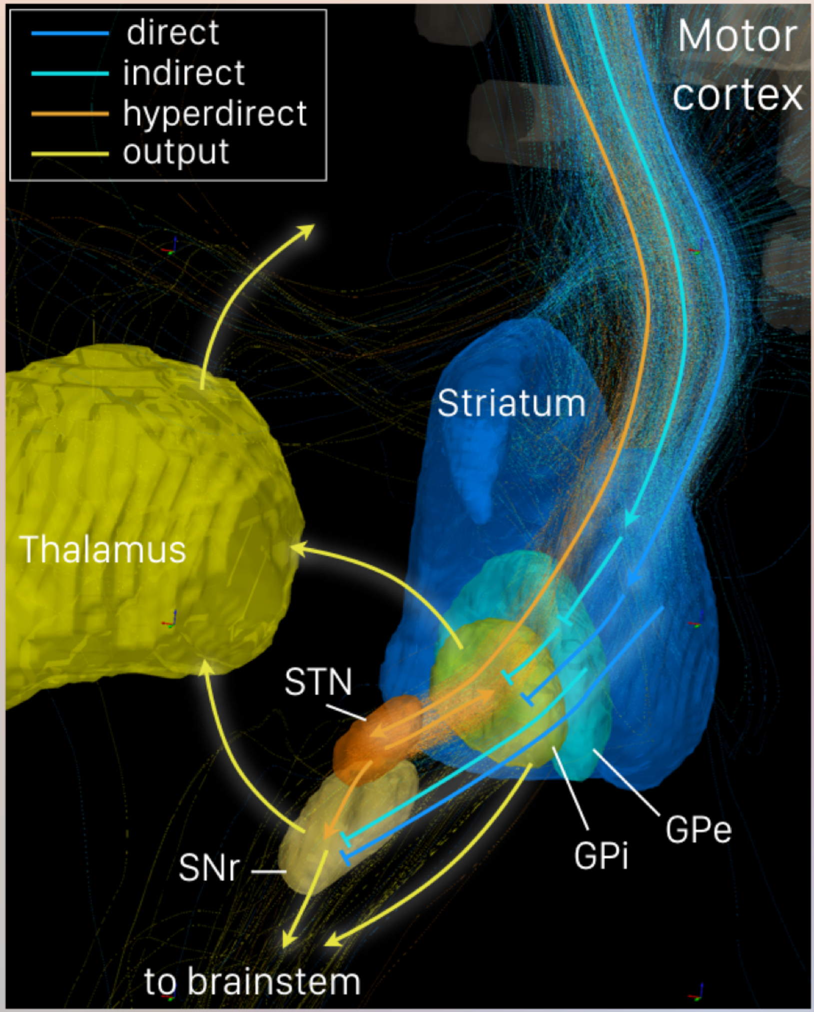
Release from Inhibition



(slide < O'Reilly)

Motor Loop Pathways

- Direct: striatum inhibits GPi (and SNr)
- Indirect: striatum inhibits GPe, which inhibits GPi (and SNr)
- Hyperdirect: cortex excites STN, which diffusely excites GPi (and SNr)
- GPi inhibits thalamus, which opens motor loops



Basal Ganglia System

- Striatum
  - matrix clusters (inhib.)
    - direct (Go) pathway ⊣ GPi
    - indirect (NoGo) path ⊣ GPe
  - patch clusters
    - to dopaminergic system
- Globus pallidus, int. segment (GPi)*
  - tonically active
  - inhibit thalamic cells
- Globus pallidus, ext. segment (GPe)
  - tonically active
  - inhibits corresponding GPi neurons

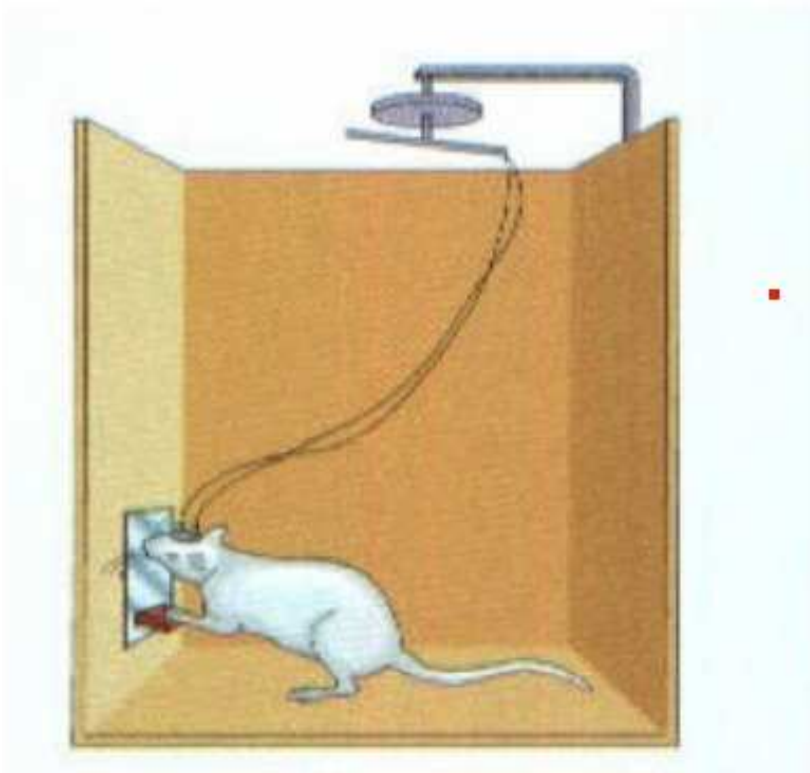
*and substantia nigra pars reticulata (SNr)

- Thalamus*
  - cells fire when both:
    - excited (cortex)
    - disinhibited (GPi)
  - disinhibits FC deep layers
- Substantia nigra pars compacta (SNc)
  - releases dopamine (DA) into striatum
  - excites D1 receptors (Go)
  - inhibits D2 receptors (NoGo)
- Subthalamic nucleus (STN)
  - hyperdirect pathway
  - input from cortex
  - diffuse excitatory output to GPi
  - global NoGo delays decision

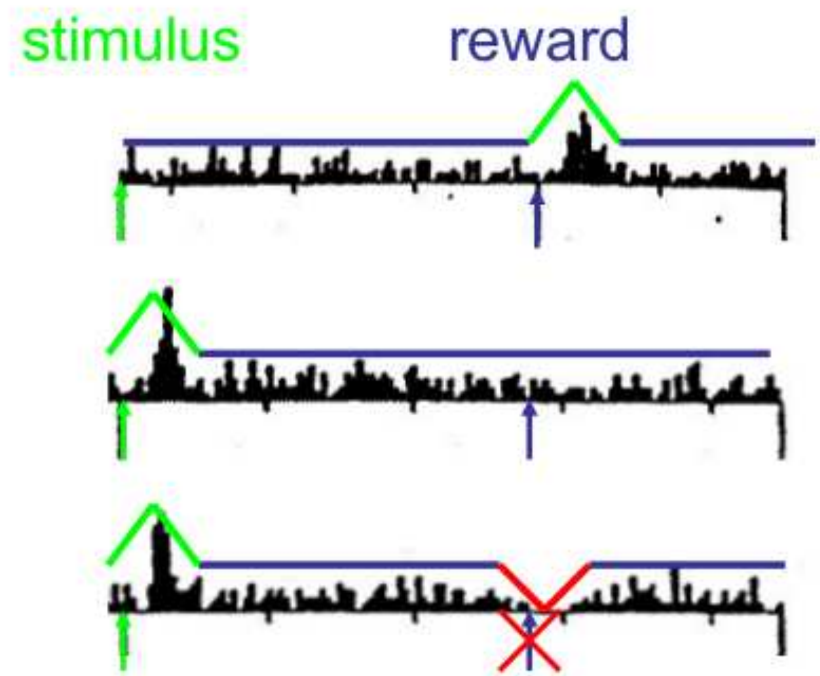
*and superior colliculus (SC)

What is Dopamine Doing?

Dopamine carries the brain's ~~reward~~ signal
reward prediction error



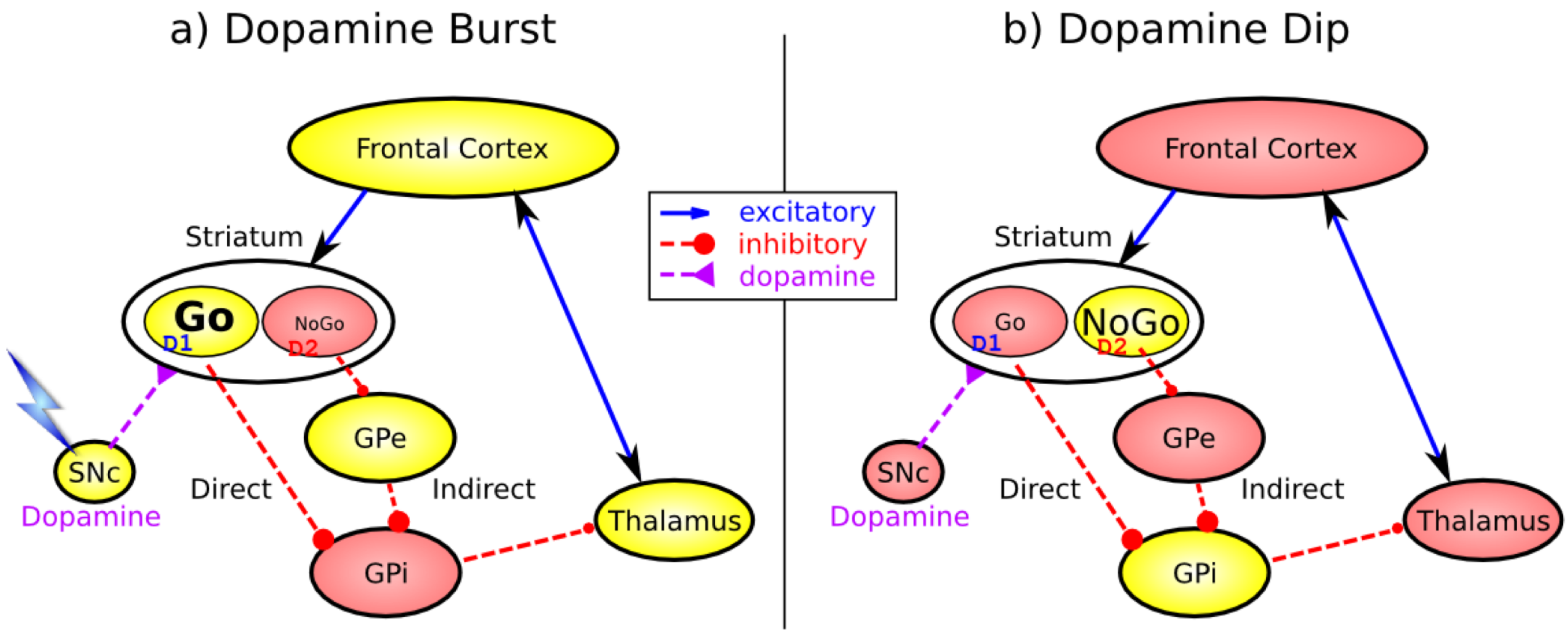
Wise & Romper, 89



Schultz et. al, 98

Basal Ganglia Reward Learning

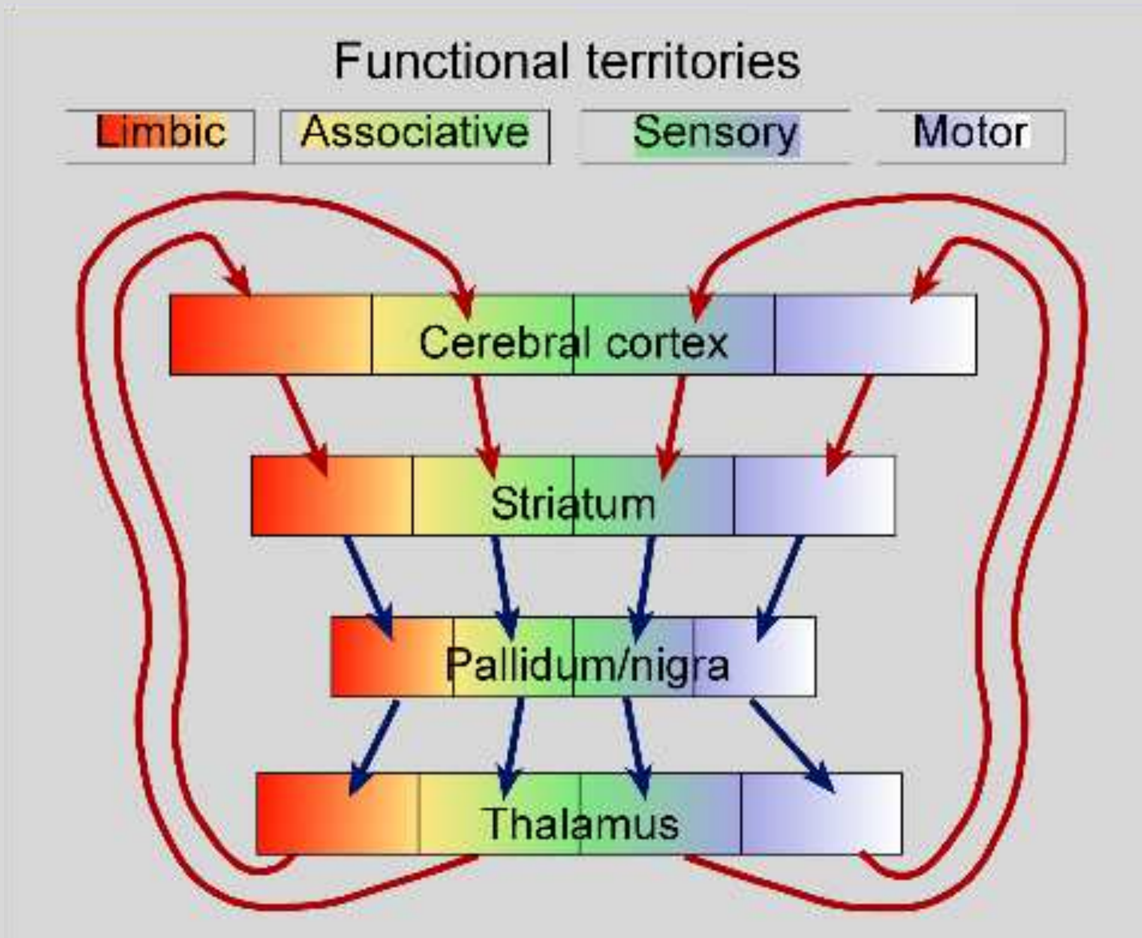
(Frank, 2005…; O'Reilly & Frank 2006)



- Feedforward, modulatory (disinhibition) on cortex/motor (same as cerebellum)
- Co-opted for higher level cognitive control ⟶ PFC

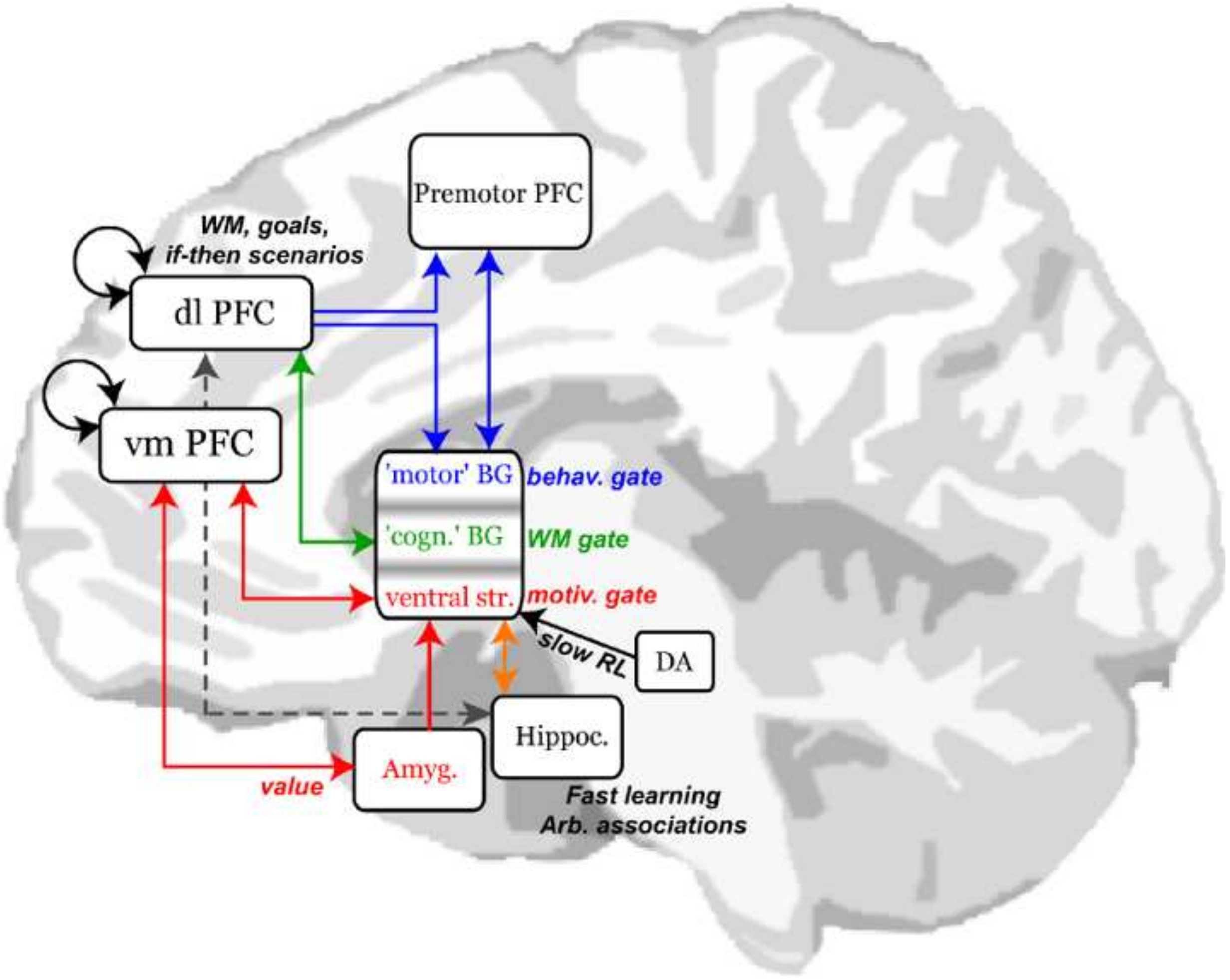
(slide < O'Reilly)

Basal Ganglia Architecture: Cortically-based Loops



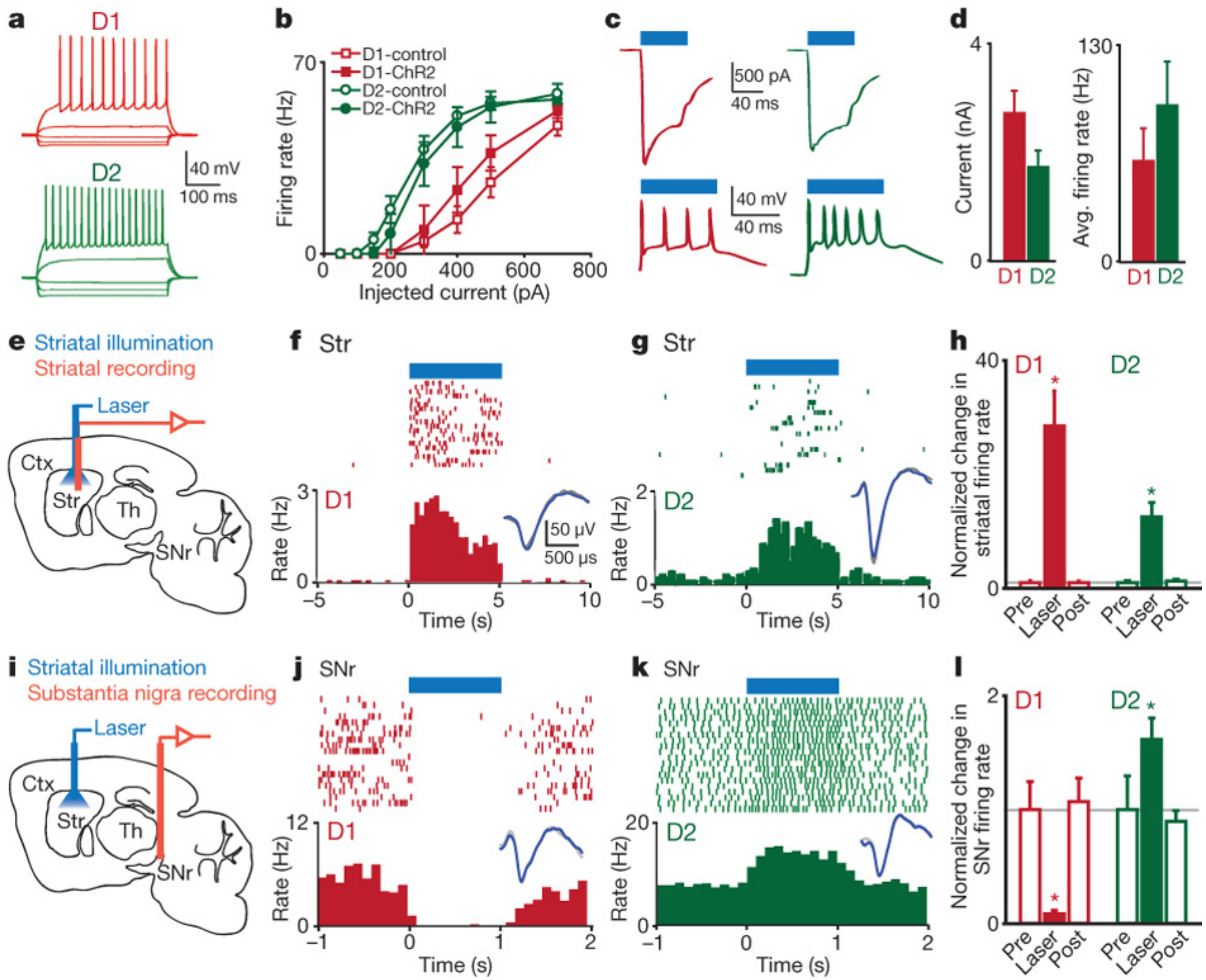
(slide < Frank)

Fronto-basal Ganglia Circuits in Motivation, Action, & Cognition



(slide < Frank)

ChR2-mediated excitation of direct- and indirect-pathway MSNs *in vivo* drives activity in basal ganglia circuitry



nature

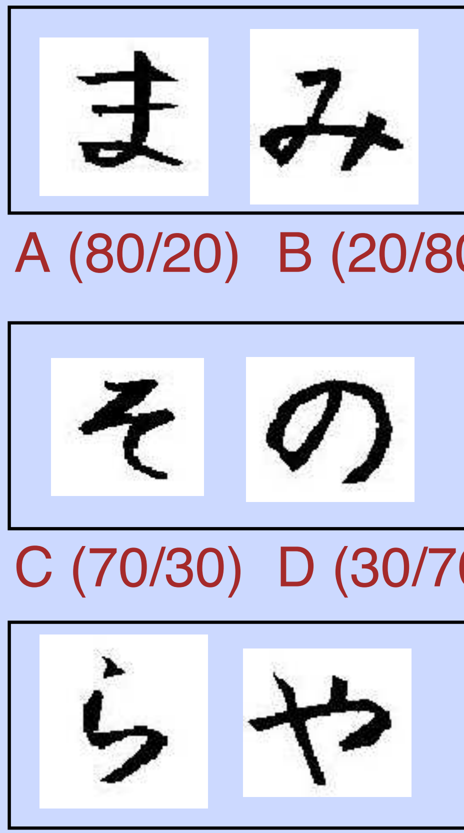
AV Kravitz *et al. Nature* 466(7306):622-6 (2010) doi:10.1038/nature09159

Human Probabilistic Reinforcement Learning

- Patients with Parkinson's disease (PD) are impaired in cognitive tasks that require learning from positive and negative feedback
- Likely due to depleted dopamine
- But dopamine medication actually worsens performance in some cognitive tasks, despite improving it in others

Frank, Seeberger & O'Reilly (2004)

Train



A (80/20) B (20/80)

C (70/30) D (30/70)

E (60/40) F (40/60)

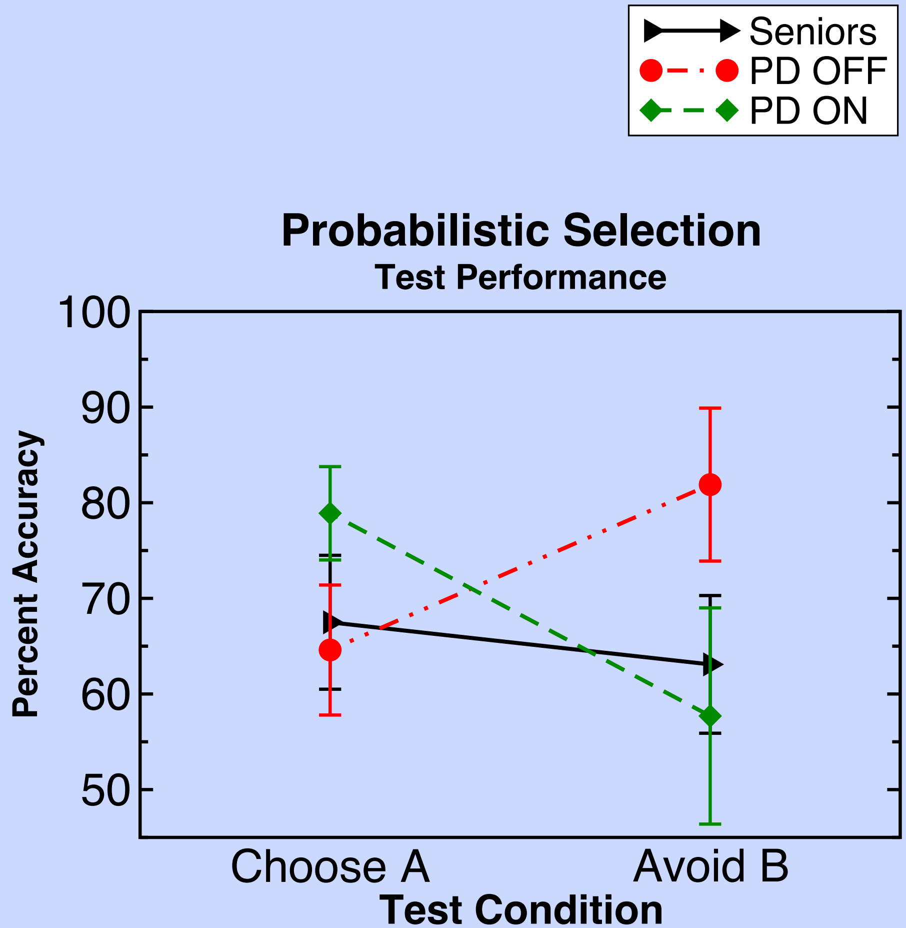
Test

A > CDEF **Choose A?**

B < CDEF **Avoid B?**

(slide based on Frank)

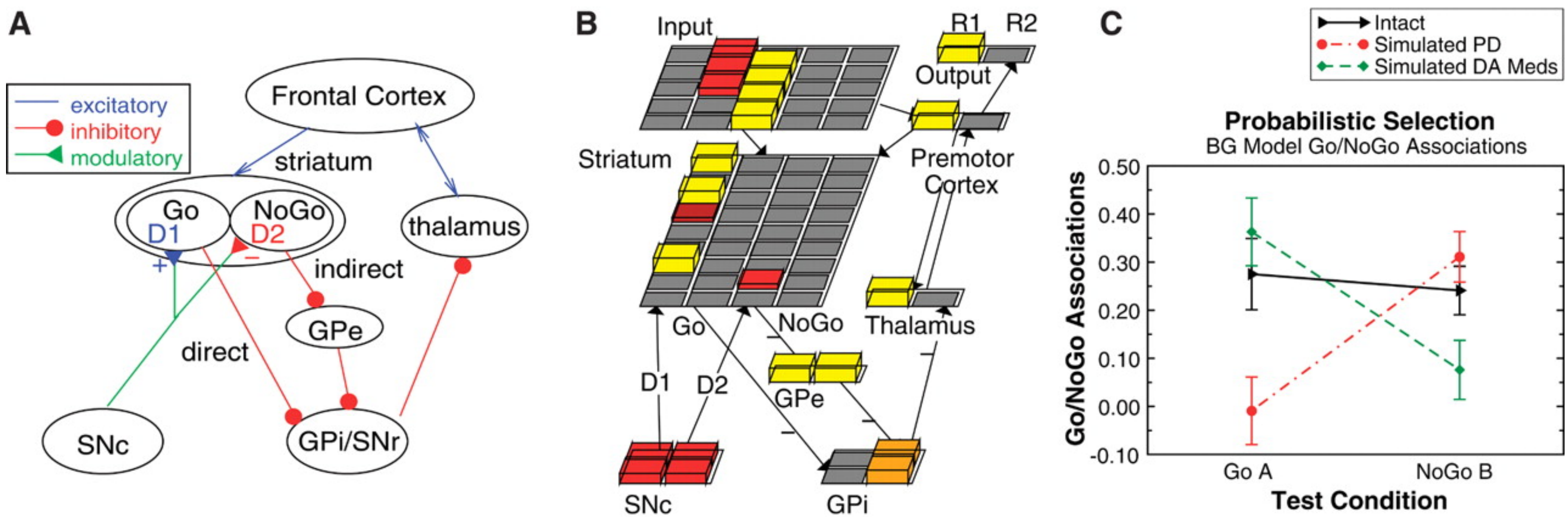
Testing the Model: Parkinson's and Medication Effects



Frank, Seeberger & O'Reilly (2004)

(slide < Frank)

BG Model: DA Modulates Learning from Positive/Negative Reinforcement



(A) The corticostriato-thalamo-cortical loops, including the direct (Go) and indirect (NoGo) pathways of the basal ganglia.
(B) M. Frank's neural network model of this circuit.
(C) Predictions from the model for the probabilistic selection task

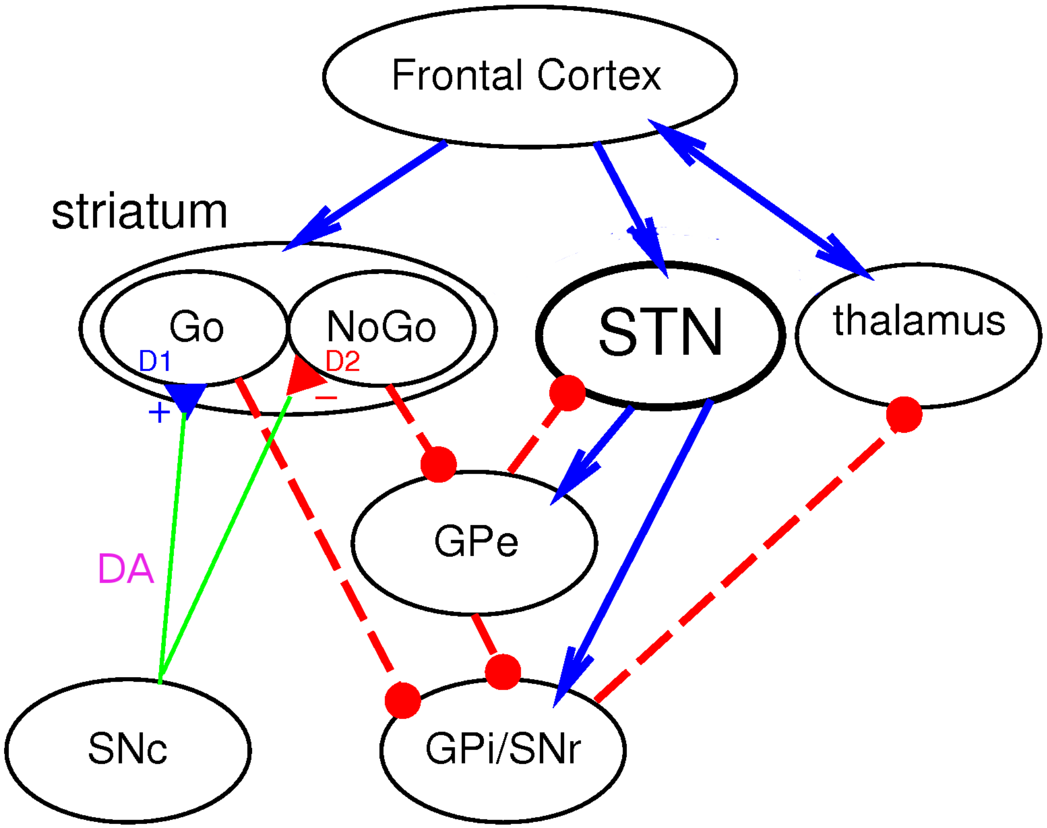
Michael J. Frank et al. Science 2004;306:1940-1943

Published by AAAS

emergent Demonstration: BG

A simplified model compared to Frank, Seeberger, & O'Reilly (2004)

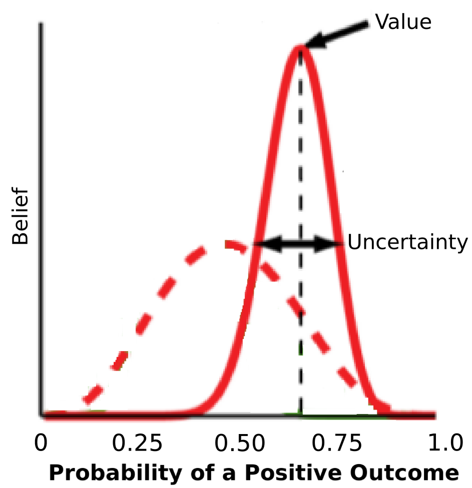
Anatomy of BG Gating Including Subthalamic Nucleus (STN)

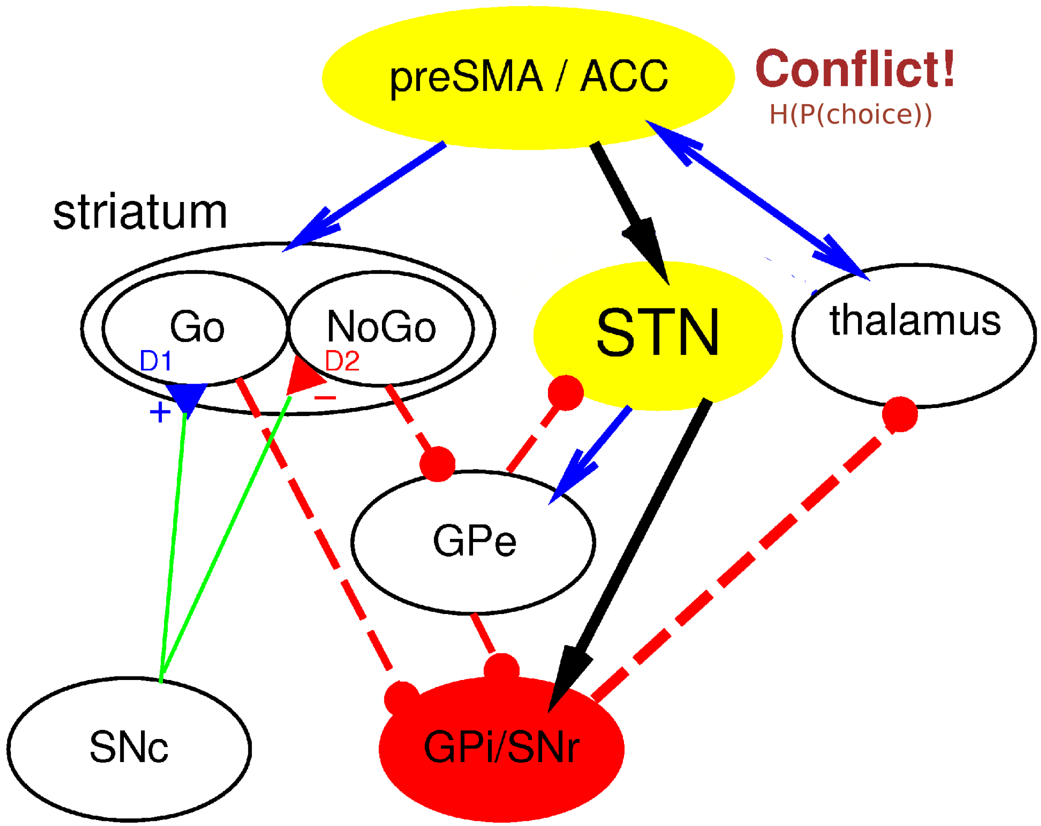


PFC-STN provides an override mechanism

(slide < Frank)

Subthalamic Nucleus: Dynamic Modulation of Decision Threshold



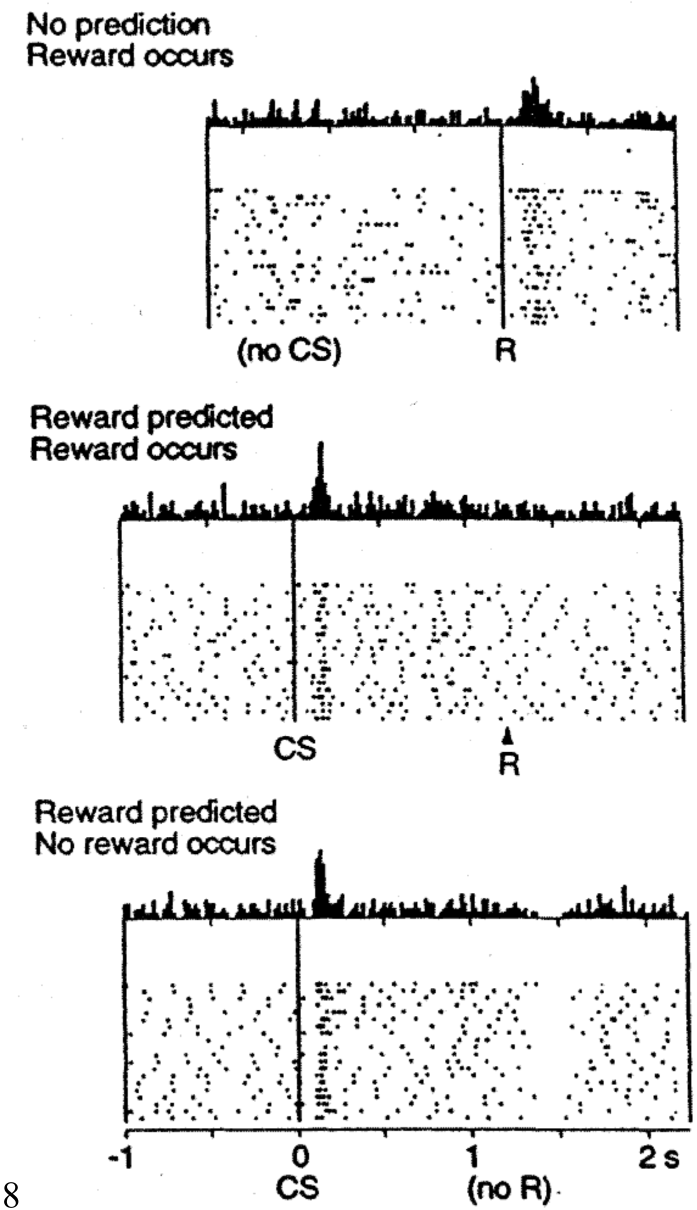


Conflict (entropy) in choice prob ⇒ delay decision!

(slide < Frank)

B. Temporal Difference Reinforcement Learning

Reinforcement Learning: Dopamine



Rescorla-Wagner / Delta Rule:

- $\delta = r - \hat{r}$
- $\delta = r - \sum xw$

But no CS-onset firing – need to anticipate the future!

- $\delta = (r + f) - \hat{r}$

CS-onset = future reward = f

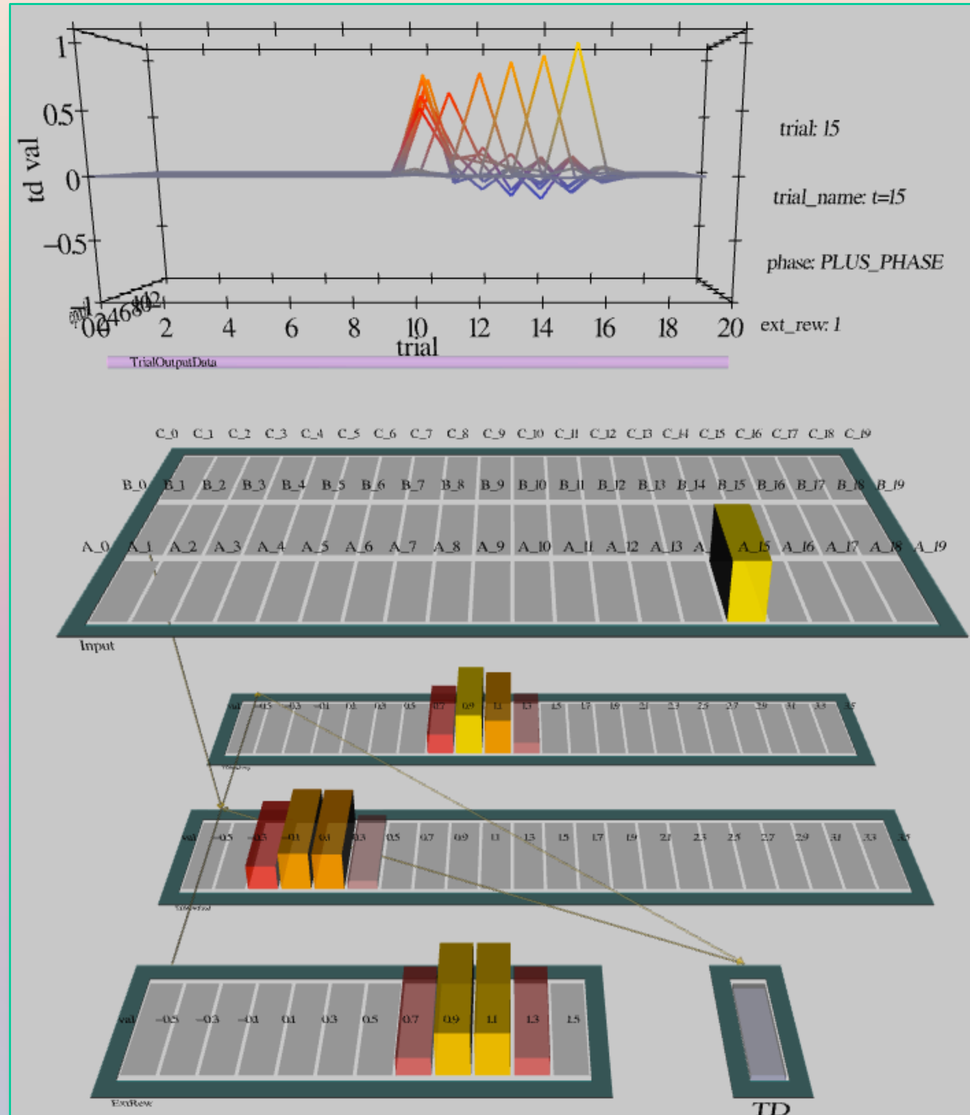
(slide < O'Reilly)

Temporal Differences Learning

- $V(t) = r(t) + \gamma^1 r(t+1) + \gamma^2 r(t+2) + \cdots$
  $= r(t) + \gamma[r(t+1) + \gamma^1 r(t+2) + \cdots]$
- $\hat{V}(t) = r(t) + \gamma\hat{V}(t+1)$
- $0 = \left(r(t) + \gamma\hat{V}(t+1)\right) - \hat{V}(t)$
- $\delta = \left(r(t) + \gamma\hat{V}(t+1)\right) - \hat{V}(t)$
- $f = \gamma\hat{V}(t+1)$ ☜ this is the future!

(slide based on O'Reilly)

Network Implementation



(slide < O'Reilly)

The RL-cond Model

- ExtRew: external reward $r(t)$ (based on input)
- TDRewPred: learns to predict reward value
  - minus phase = prediction $V(t)$ from previous trial
  - plus phase = predicted $V(t+1)$ based on Input
- TDRewInteg: Integrates ExtRew and TDRewPred
  - minus phase = $V(t)$ from previous trial
  - plus phase = $V(t+1) + r(t)$
- TD: computes temporal dif. delta value ≈ dopamine signal
  - compute plus – minus from TDRewInteg

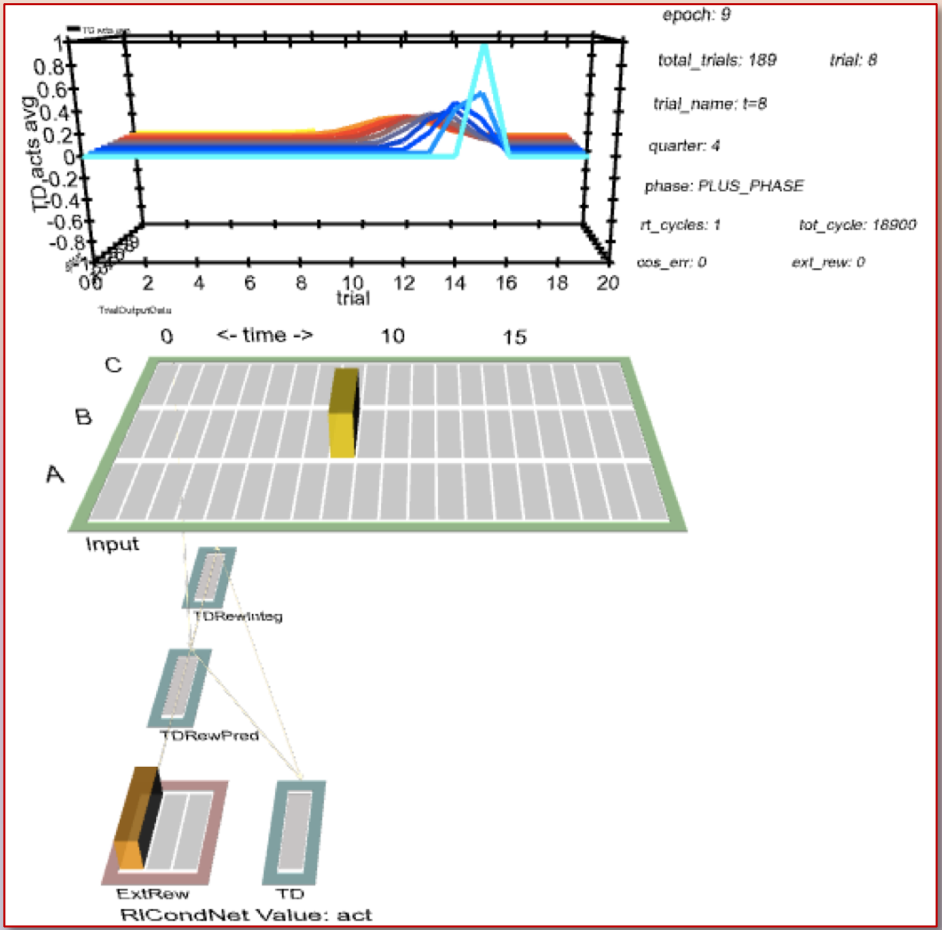
Classical Conditioning

- Forward conditioning
  - — unconditioned stimulus (US): doesn’t depend on experience
  - — leads to unconditioned response (UR)
  - — preceding conditioned stimulus (CS) becomes associated with US
  - — leads to conditioned response (CR)
- Extinction
  - — after CS established, CS is presented repeatedly without US
  - — CR frequency falls to pre-conditioning levels
- Second-order conditioning
  - — CS1 associated with US through conditioning
  - — CS2 associated with CS1 through conditioning, leads to CR

CSC Experiment

- A serial-compound stimulus has a series of distinguishable components
- A complete serial-compound (CSC) stimulus has a component for every small segment of time before, during, and after the US
  - Richard S. Sutton & Andrew G. Barto, “Time-Derivative Models of Pavlovian Reinforcement,” *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, M. Gabriel and J. Moore, Eds., pp. 497–537. MIT Press, 1990
- RL-cond.proj implements this form of conditioning
  - somewhat unrealistic, since the stimulus or some trace of it must persist until the US

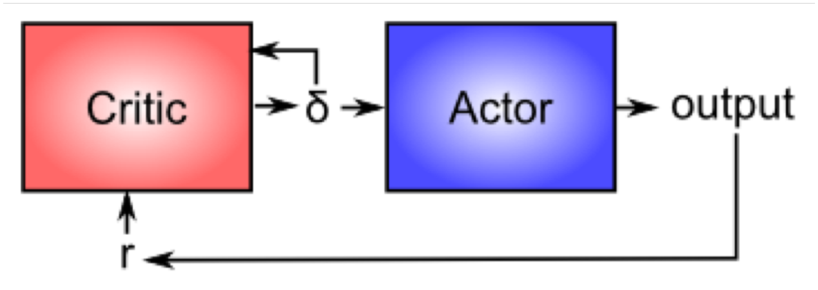
RL-cond.proj

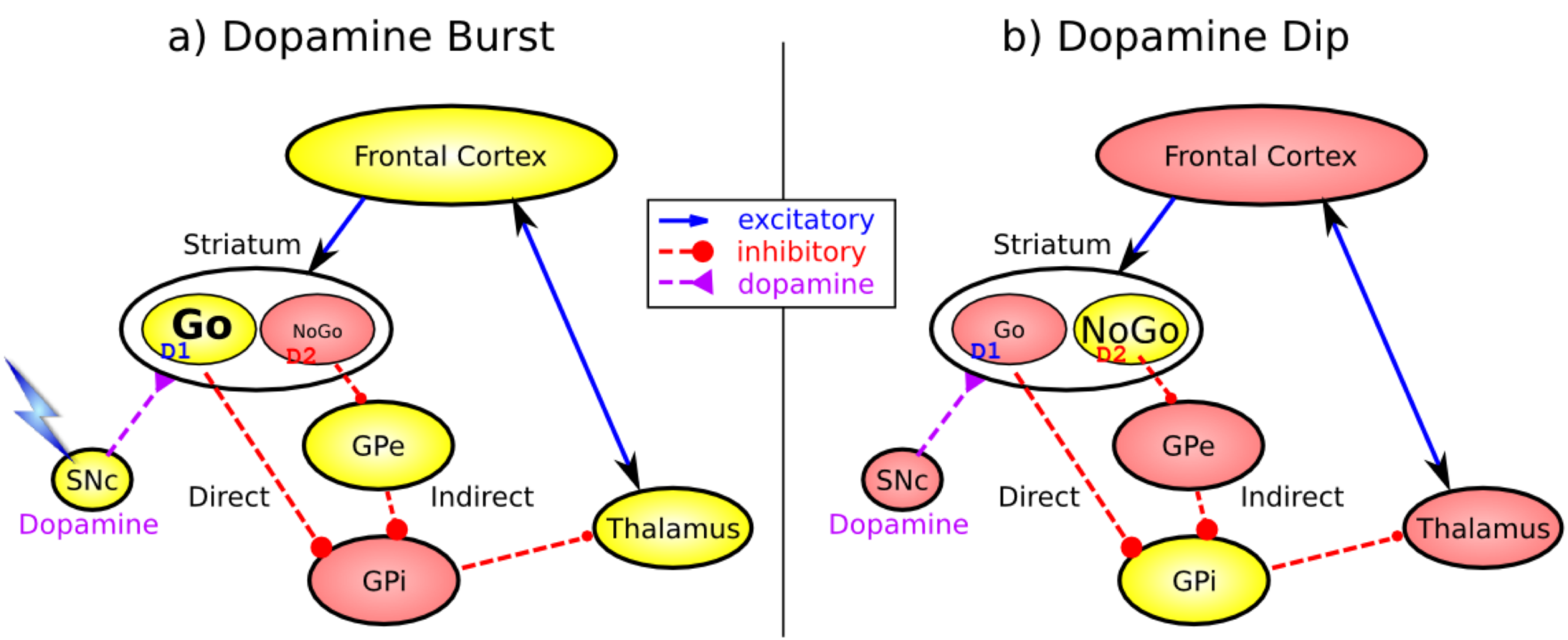

epoch: 9
total_trials: 189
trial: 8
trial_name: t=8
quarter: 4
phase: PLUS_PHASE
rt_cycles: 1
tot_cycle: 18900
cos_err: 0
ext_rew: 0
TD_acts avg
trial
<- time ->
C
B
A
Input
TDRewInteg
TDRewPred
ExtRew
TD
RlCondNet Value: act

emergent Demonstration: RL

A simplified model of temporal difference reinforcement learning

Actor - Critic

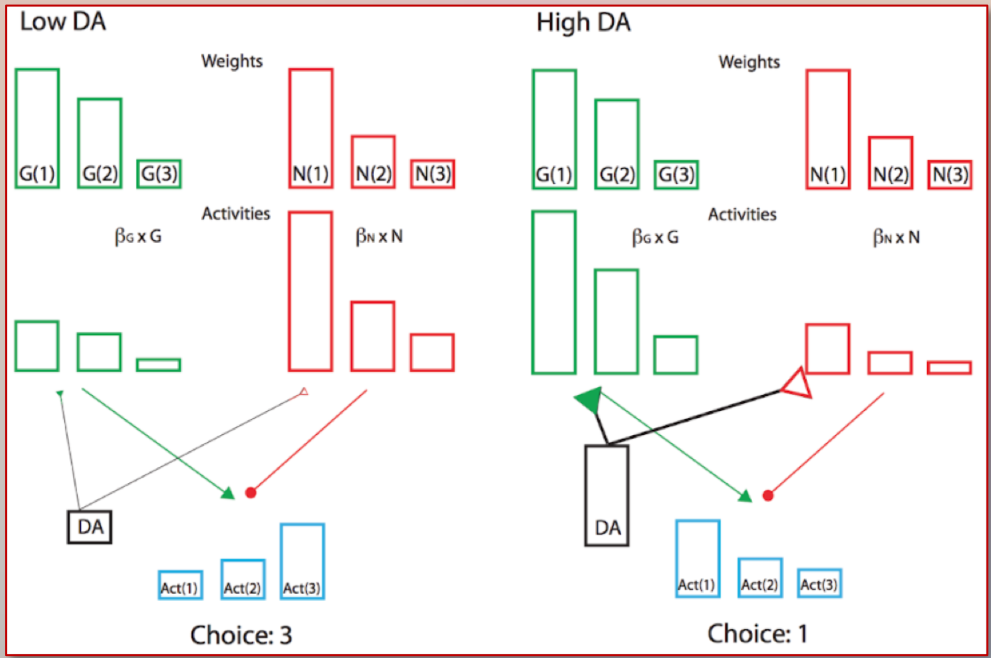




(slide < O'Reilly)

Opponent-Actor Learning (OpAL)

- Actor has independent G and N weights
- Scaled by dopamine (DA) levels during choice
- Choice based on relative activation levels
- Low DA: costs amplified, benefits diminished ⇒ choice 1
- High DA: benefits amplified, costs diminished ⇒ choice 3
- Moderate DA ⇒ choice 2
- Accounts for differing costs & benefits



C. PVLV Model of DA Biology

A model of dopamine firing in the brain

Brain Areas Involved in Reward Prediction

- Lateral hypothalamus (LHA): provides a primary reward signal for basic rewards like food, water etc.
- Patch-like neurons in ventral striatum (VS-patch)
  - have direct inhibitory connections onto dopamine neurons in VTA and SNc
  - likely role in canceling influence of primary reward signals when they're successfully predicted
- Central nucleus of amygdala (CNA)
  - important for driving dopamine firing at the onset of conditioned stimuli
  - receives input broadly from cortex
  - projects directly and indirectly (via VS-patch) to the VTA and SNc (DA neurons)
  - neurons in the CNA exhibit CS-related firing

PVLV Model of Dopamine Firing

- Two distinct systems: Primary Value (PV) and Learned Value (LV)
- DA signal at time of external reward (US):

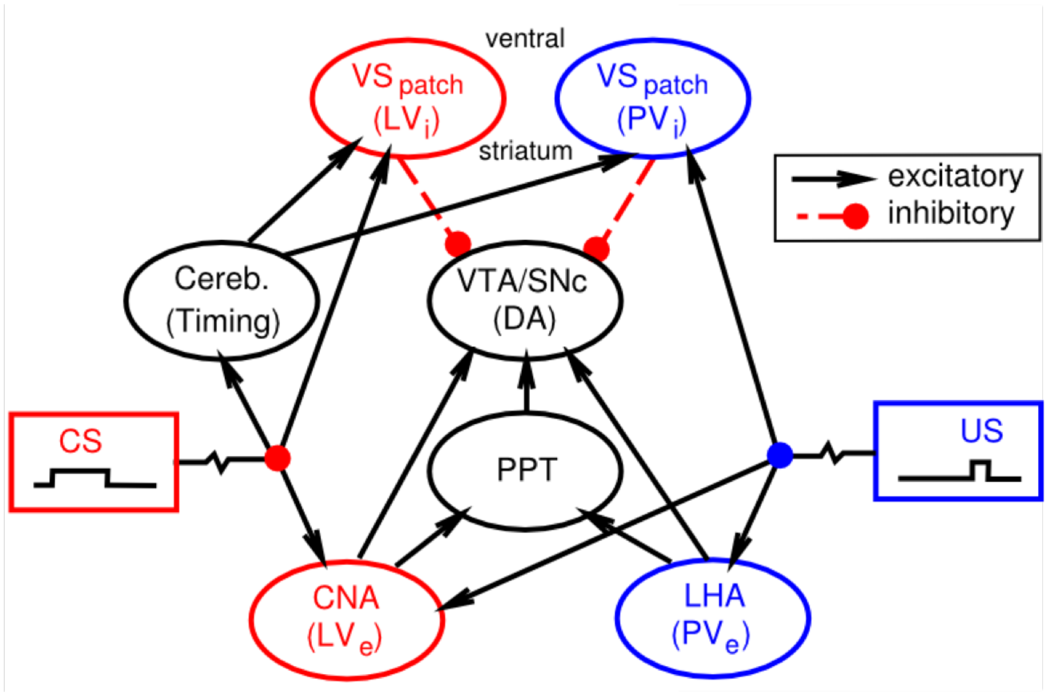
$$\delta_{\mathrm{pv}} = \mathrm{PV_e} - \mathrm{PV_i} = r - \hat{r}$$

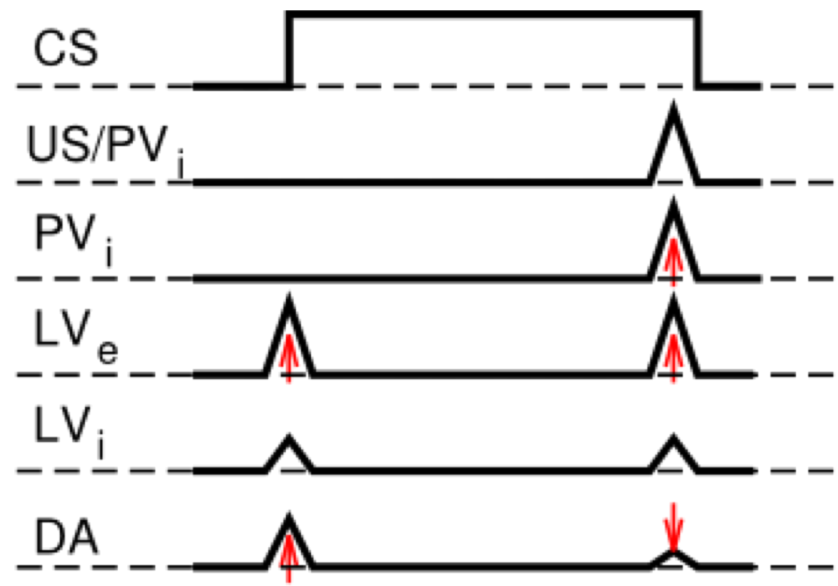
- DA signal for LV when PV not present/expected:

$$\delta_{\mathrm{lv}} = \mathrm{LV_e} - \mathrm{LV_i}$$

- LV_e is excitatory drive from CNA responding to CS (eventually canceled by LV_i)
- LV_e and LV_i values learned from PV_e when rewards present/expected
- Hence, CS (or some trace) must still be present when US occurs
- CNA supports 1st order conditioning, but not 2nd order (that's in BLA)

Biology of Dopamine Firing





(slide < O'Reilly)

More Detailed Description of PVLV

- Major issue: Which of PV/LV systems should be in charge of overall dopamine system?
- PV and LV learning occur when PV present or expected (indicated by $PV_r > \boldsymbol{\Theta}_{pv}$)
- PVr system learns: $\delta w_{\mathrm{pvr}} = r_{\mathrm{present}} - \mathrm{PV_r}$ (improves prediction)
- Recall alternative DA signals:
$$\delta_{\mathrm{pv}} = \mathrm{PV_e} - \mathrm{PV_i}, \qquad \delta_{\mathrm{lv}} = \mathrm{LV_e} - \mathrm{LV_i}$$
- Novelty Value (NV) signal reflects stimulus novelty
- Overall dopamine signal:
$$\delta = \begin{cases} \delta_{\mathrm{pv}}(t) - \delta_{\mathrm{pv}}(t-1) & \text{if } \mathrm{PV_r} > \Theta_{\mathrm{pv}} \\ [\delta_{\mathrm{lv}}(t) - \delta_{\mathrm{lv}}(t-1)] + [\mathrm{NV}(t) - \mathrm{NV}(t-1)] & \text{otherwise} \end{cases}$$
- Note DA burst is phasic (ceases after CS onset)

More Detailed Description (ctu'd)

- Learning $\mathrm{PV_i}$ weights:

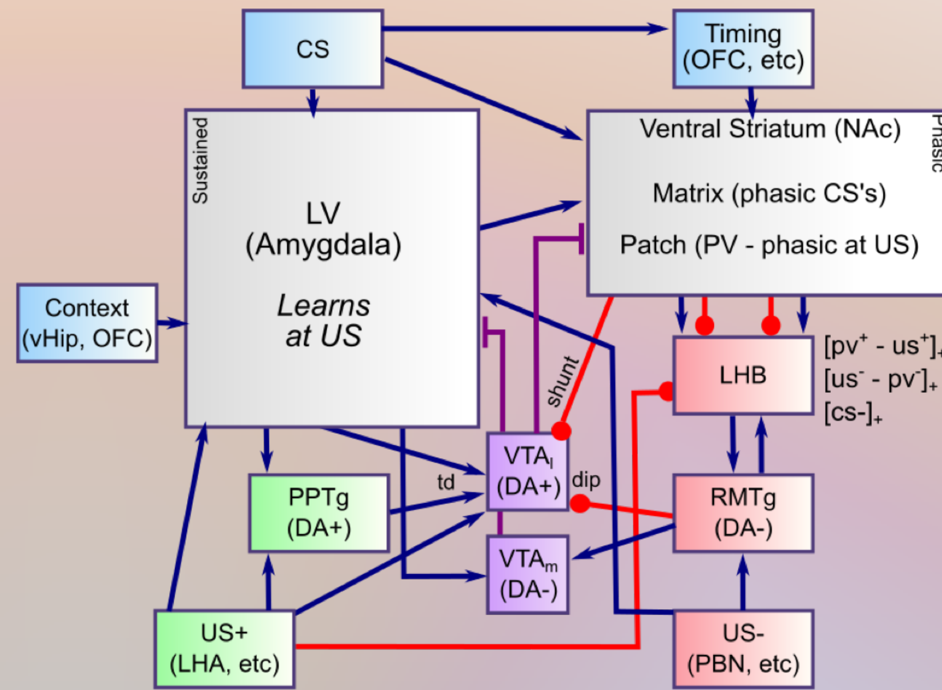
$$\delta w_{\mathrm{pv}} = \varepsilon(\mathrm{PV_e} - \mathrm{PV_i})x$$

- Learning LV weights is conditional on PV filter:

$$\delta w_{\mathrm{lv}} = \begin{cases} \varepsilon(\mathrm{PV_e} - \mathrm{LV_e})x & \text{if } \mathrm{PV_r} > \Theta_{\mathrm{pv}} \\ 0 & \text{otherwise} \end{cases}$$

PVLV.proj Model

- PV in Ventral Striatum system
- LV in Amygdala system
- VTA_l and VS adapt to US+
- Eventually VTA_l bursts for CS onset
- LHB+RMTg and VS adapt to US–
- VTA_m and VS adapt to US–
- Eventually DA dip for CS



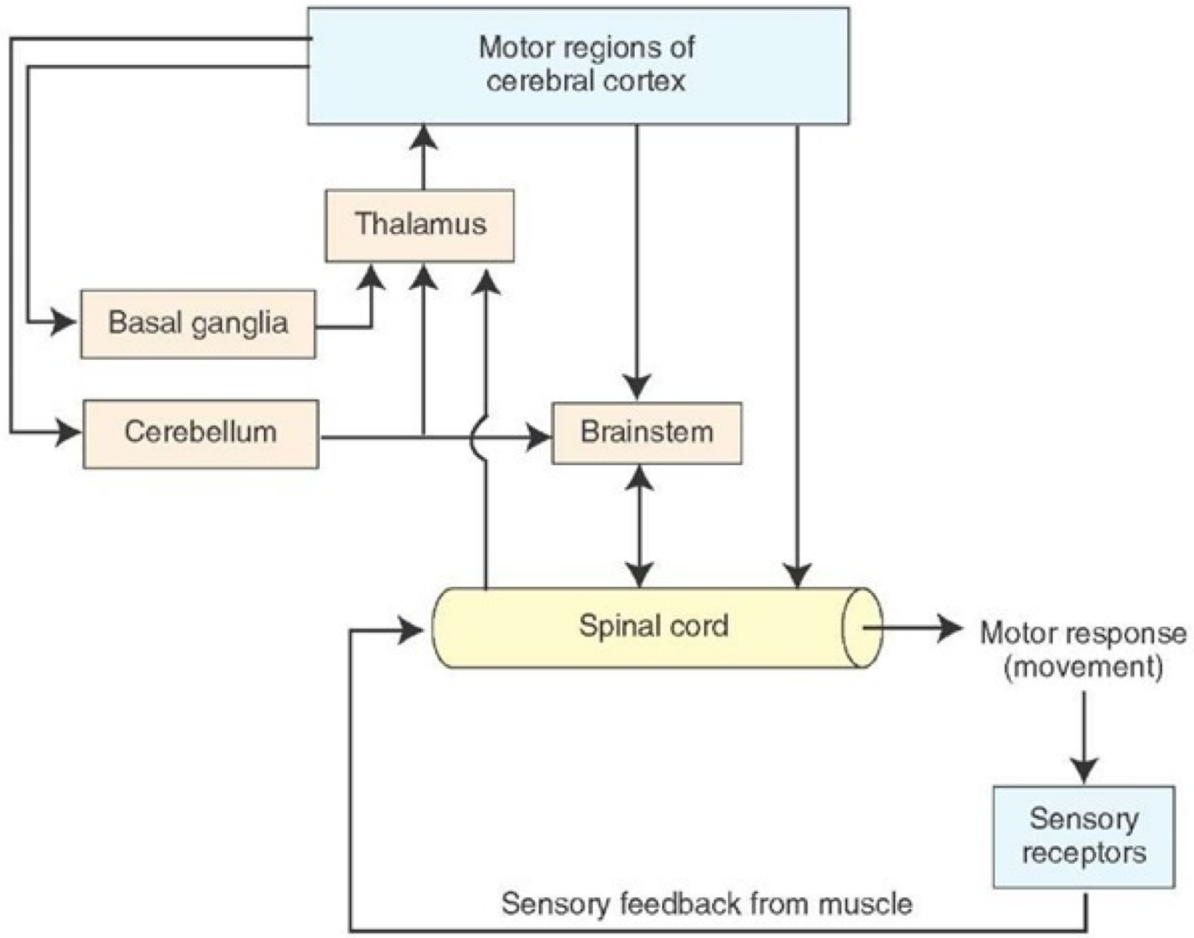
simplified!

emergent Demonstration: PVLV

D. Cerebellum and Error-driven Learning

"The blessing of dimensionality"

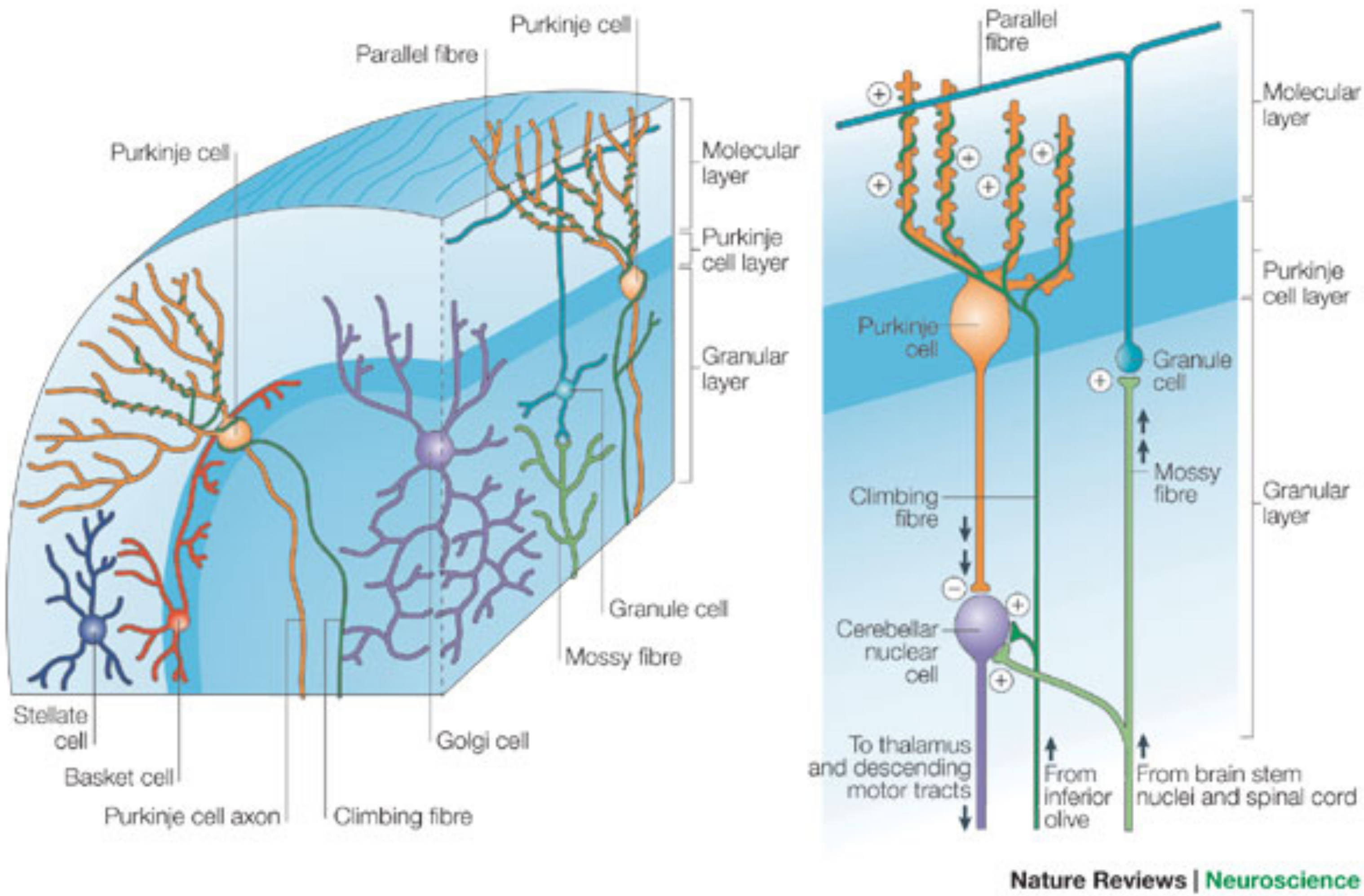
The Motor Control System



Functions of Cerebellum

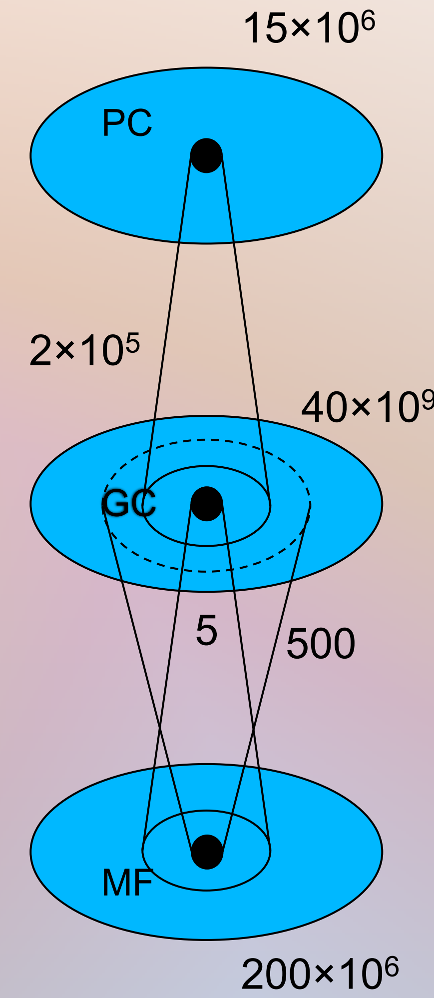
- Maintenance of equilibrium and posture
- Timing of learned, skilled motor movement
  - — any motor movement that improves with practice
  - — timing, fluency, rhythm, coordination
  - — involved in cognitive processes too
- Correction of errors during the execution of movements
  - — error-driven learning
- Many inputs from cortical motor and sensory areas
- Influences cortical motor outputs to spinal chord

Cerebellar Microstructure

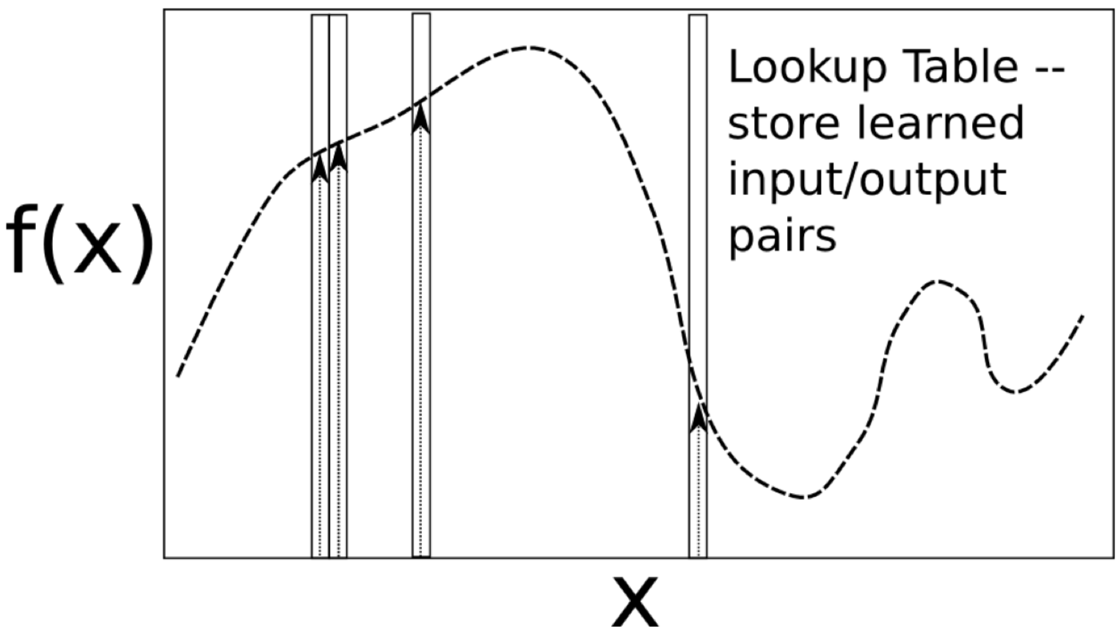


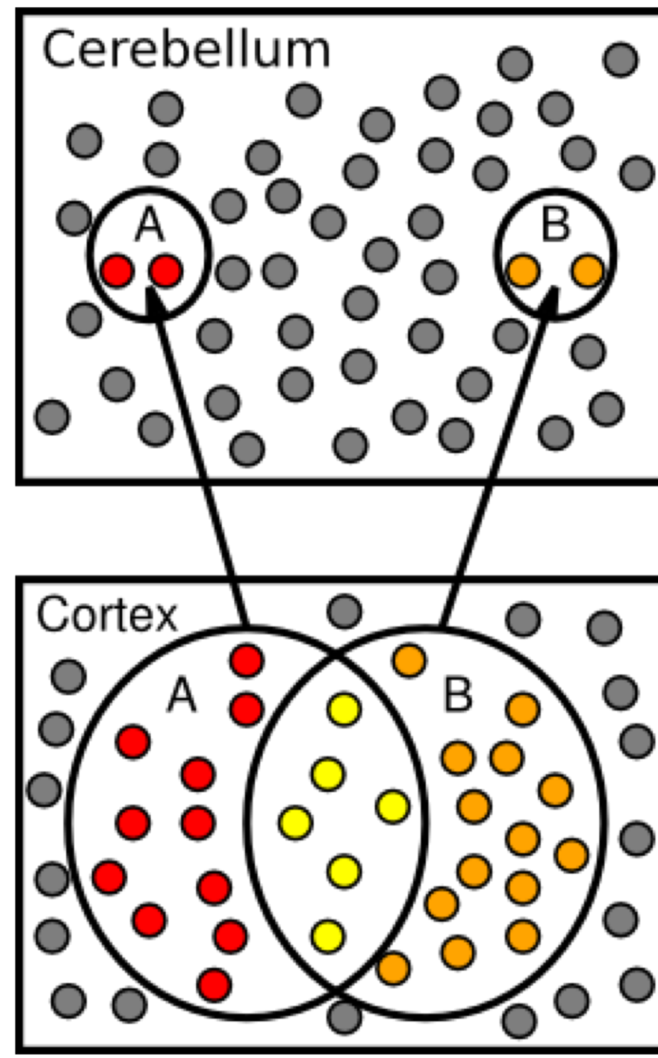
Cerebellum

- Inputs from parietal cortex and motor areas of frontal cortex
- Three layers, very many cortical maps
- Single basic circuit replicated throughout
- 200 million mossy fiber inputs (each to 500 granule cells)
    - — projection of input into hyperdimensional space
    - — separator learning and dynamics
- 40 billion granule cells (input from 4–5 mossy fibers)
- 15 million Purkinje cells (input from 200,000 granule cells)
    - — matrix organization
    - — enormous integration and cross connection
- Climbing fibers (one per Purkinje, from inferior olive)



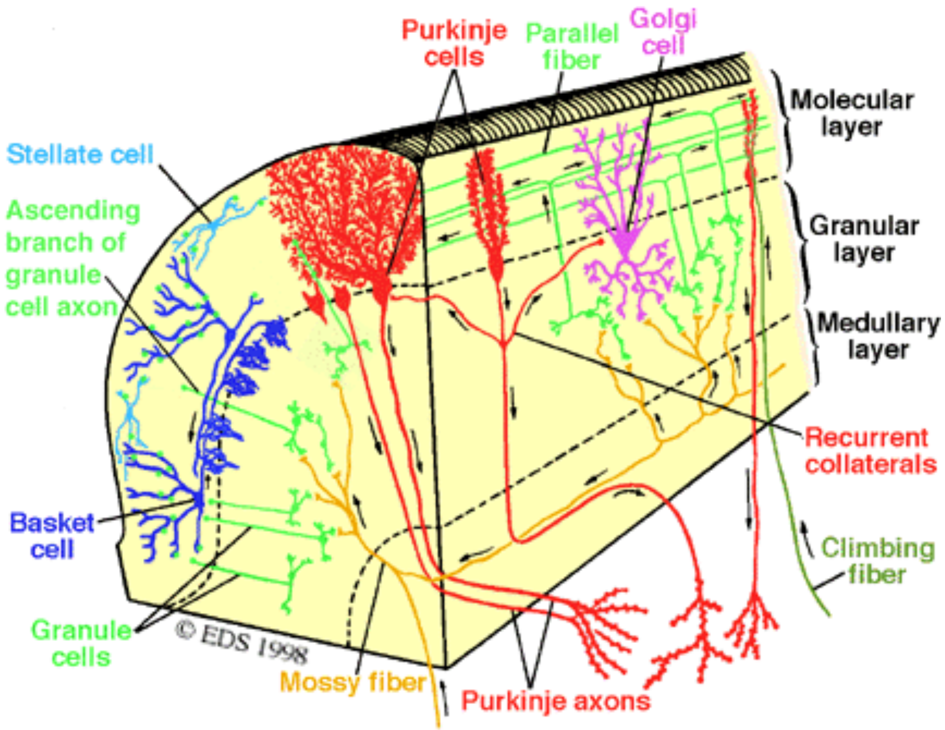
Lookup Table & Pattern Separation

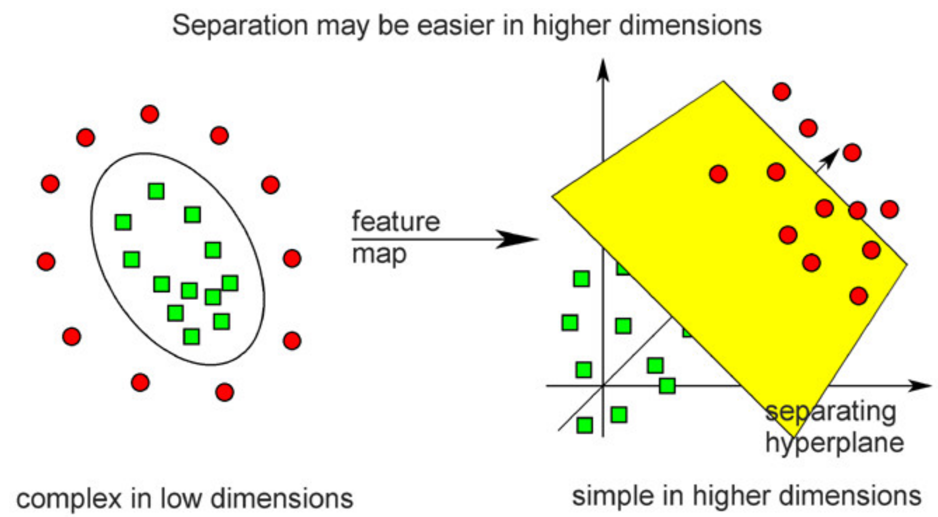




(slide < O'Reilly)

Cerebellar Error-driven Learning





Cerebellum =
Support Vector Machine

- Granule cells = high-dimensional encoding (separation)
- Purkinje/Olive = delta-rule error-driven learning
- Classic ideas from Marr (1969) & Albus (1971)

(slide < O'Reilly)

Cerebellum is Feed Forward

Feedforward circuit:

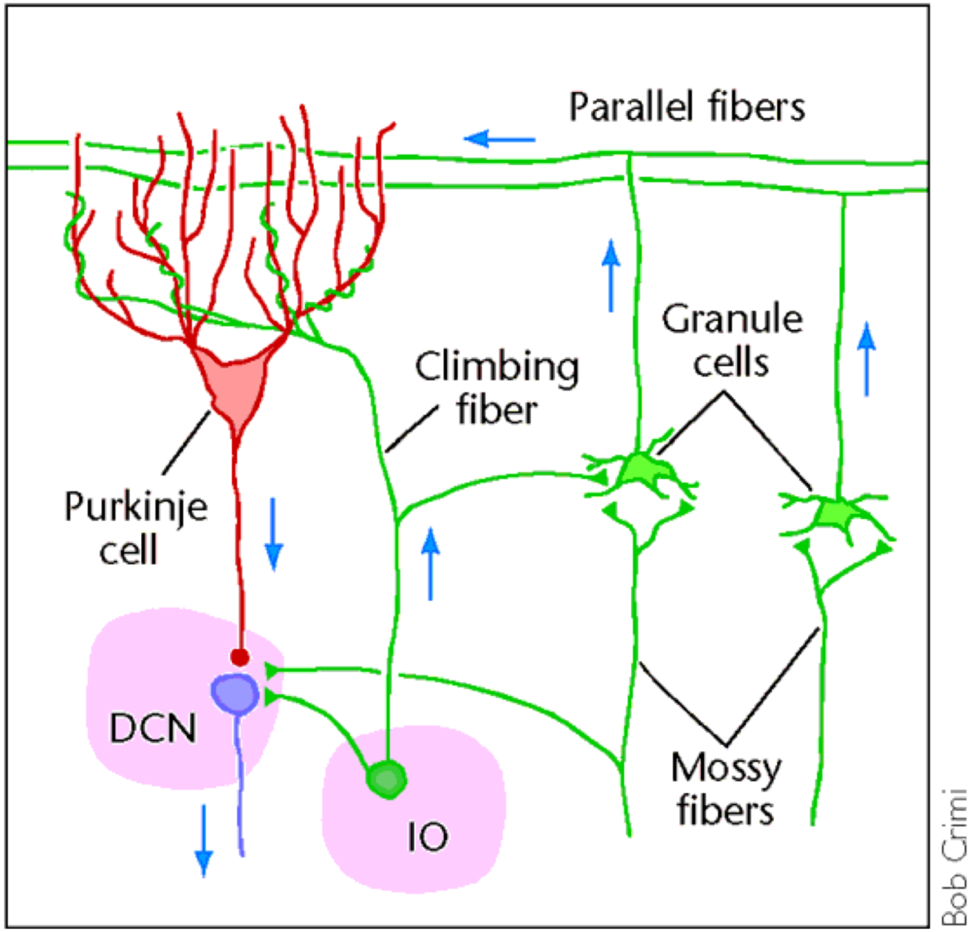
Input (PN) ⟶ granules ⟶ Purkinje ⟶ Output (DCN)

Inhibitory interactions – no attractor dynamics

<u>Key idea</u>: does delta-rule learning bridging small temporal gap:

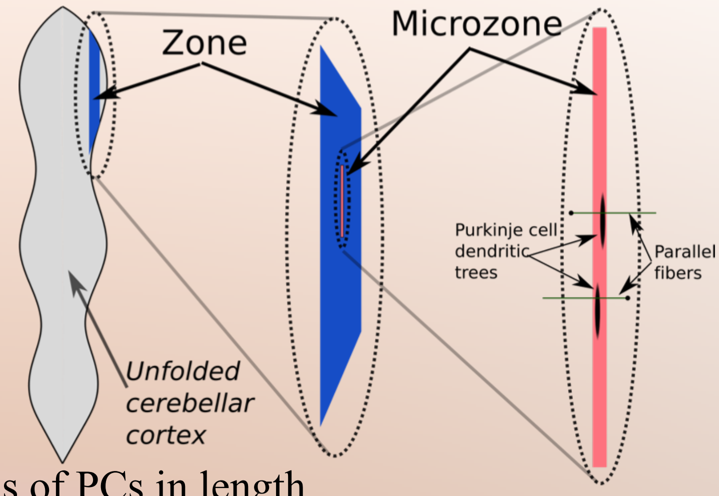
$S(t-100) \longrightarrow R(t)$

$\uparrow$ Error(t+100)



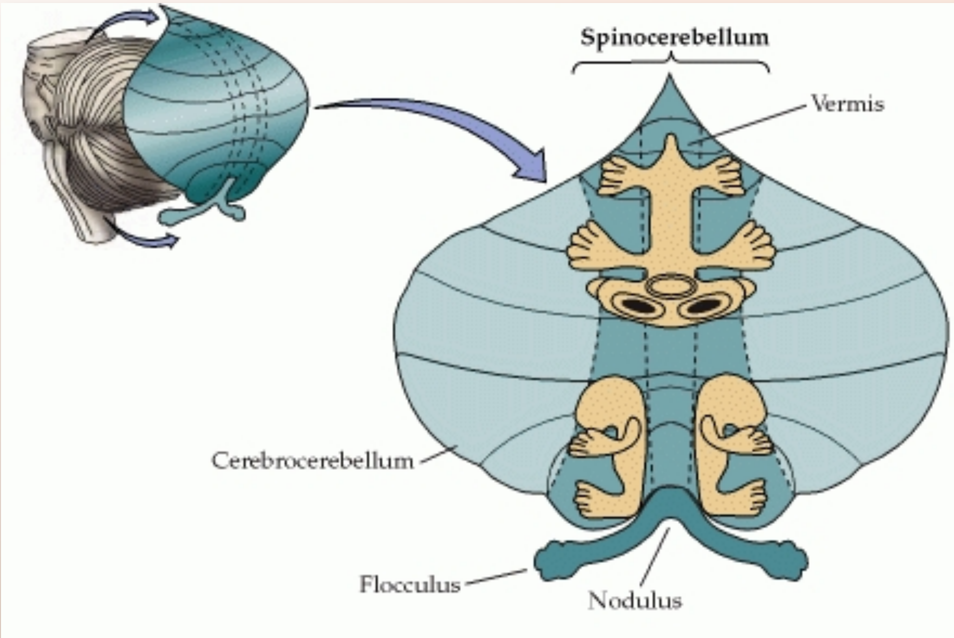
(slide < O'Reilly)

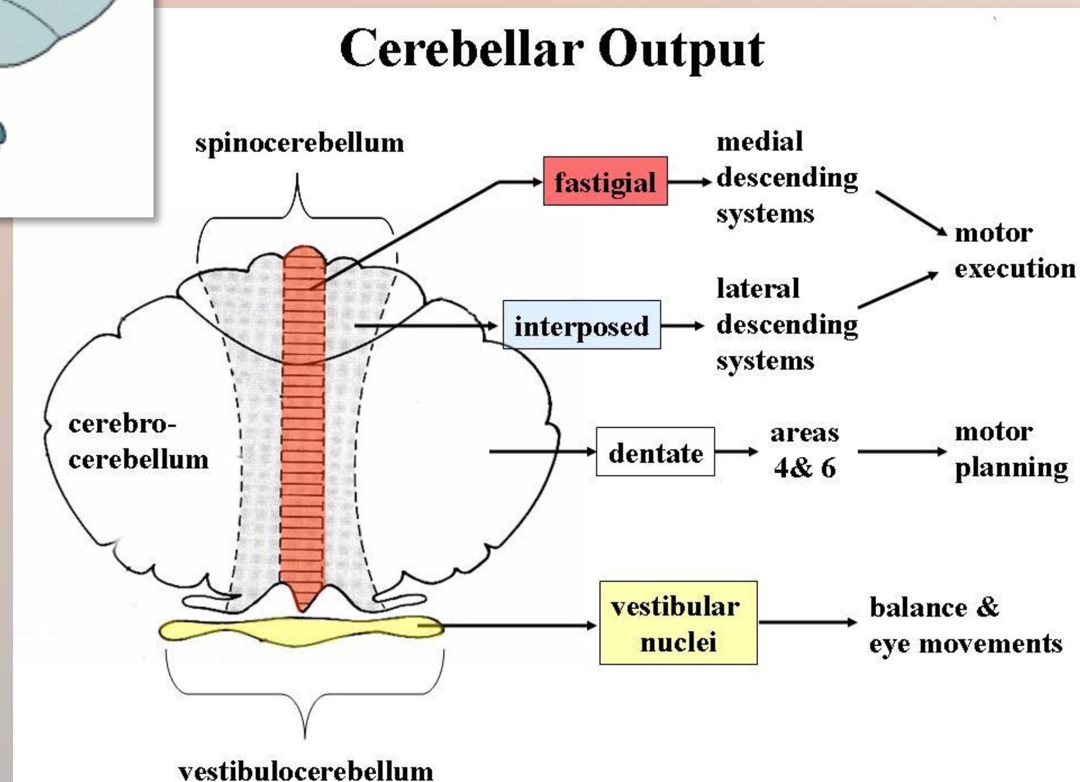
Mesostructure



- Microzone: defined by group of adjacent PCs contacted by CFs with same receptive profiles
  - comprises hundreds of PCs and several hundreds of thousands of other neurons
  - shaped as narrow strips a few PCs wide and several dozens of PCs in length
  - a fraction of a millimeter in width and several millimeters in length
  - parallel fibers (PFs) extend for several millimeters, crossing width of microzone and extending into neighbors
  - estimated that cat has about 5000 microzones, human has several hundred thousand
- Multizonal micro-complexes (MZMCs): basic functional units of cerebellar cortex
  - each comprises several microzones receiving common CF input and delivering their PC output to the same region of the cerebellar nuclei
  - seem to have an integrated function
  - constituent microzones may be in different regions of the cortex, which receive different MF input and may be associated with different aspects of motor control
  - MZMCs may provide for parallel processing and integration of inputs

Macrostructure





Properties of Hyperdimensional Spaces

- Hyperdimensional spaces = spaces of very high dimension
- Consider vectors of 10,000 bits
  - measure distance by Hamming distance (HD)
  - or normalized Hamming distance (NHD)
- Mean HD = 5000, SD = 50 (binomial distribution)
- $< 10^{-9}$ of space closer than NHD = 0.47 or farther than 0.53 (±300 = ±6 SD)
- Therefore random vectors almost surely have NHD = 0.5±0.03
- Vectors with < 3000 changed bits still accurately recognized
- Ref: Pentti Kanerva (2009), Hyperdimensional Computing: An Introduction to Computing in Distributed Representation with High-Dimensional Random Vectors, *Cognitive Computation*, **1**(2)

Orthogonality of Random Hyperdimensional Bipolar Vectors

- 99.99% probability of being within 4σ of mean
- It is 99.99% probable that random n-dimensional vectors will be within $\varepsilon = 4/\sqrt{n}$ orthogonal
- $\varepsilon = 4\%$ for $n = 10{,}000$
- Probability of being less orthogonal than ε decreases exponentially with n
- The brain gets approximate orthogonality by using random high-dimensional vectors

$$|\mathbf{u} \cdot \mathbf{v}| < 4\sigma$$

$$\text{iff } \|\mathbf{u}\| \; \|\mathbf{v}\| \; |\cos\theta| < 4\sqrt{n}$$

$$\text{iff } n|\cos\theta| < 4\sqrt{n}$$

$$\text{iff } |\cos\theta| < 4/\sqrt{n} = \varepsilon$$

$$\Pr\{|\cos\theta| > \varepsilon\} = \operatorname{erfc}\left(\frac{\varepsilon\sqrt{n}}{\sqrt{2}}\right)$$

$$\approx \frac{1}{6}\exp\left(-\varepsilon^2 n/2\right) + \frac{1}{2}\exp\left(-2\varepsilon^2 n/3\right)$$

Hyperdimensional Pattern Associator

- Suppose $\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_P$ are a set of random hyperdimensional bipolar vectors (inputs)
- Let $\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_P$ be arbitrary bipolar vectors (outputs)
- Define Hebbian linear associator matrix

$$\mathrm{M} = \frac{1}{N} \sum_{k=1}^{P} \mathbf{q}_k \mathbf{p}_k^{\mathrm{T}}$$

- Then $\mathrm{M}\mathbf{p}_k \approx \mathbf{q}_k$ (table lookup)
- To encode a sequence of random vectors $\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_P$:

$$\mathrm{M} = \frac{1}{N} \sum_{k=1}^{P-1} \mathbf{p}_{k+1} \mathbf{p}_k^{\mathrm{T}}$$

- Then $\mathrm{M}\mathbf{p}_k = \mathbf{p}_{k+1}$ (sequence readout)

Some math…

- Suppose $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_P$ are random hyperdimensional bipolar vectors
- Suppose $\mathrm{M} = \frac{1}{N}\sum_{j=1}^{P} \mathbf{q}_j \mathbf{p}_j^T$
- Then, $\mathrm{M}\mathbf{p}_k = \left(\frac{1}{N}\sum_{j=1}^{P} \mathbf{q}_j \mathbf{p}_j^T\right) \mathbf{p}_k$
  $= \frac{1}{N}\left(\mathbf{q}_k \mathbf{p}_k^T + \sum_{j \neq k} \mathbf{q}_j \mathbf{p}_j^T\right)\mathbf{p}_k$
  $= \frac{1}{N}\mathbf{q}_k \mathbf{p}_k^T \mathbf{p}_k + \frac{1}{N}\sum_{j \neq k} \mathbf{q}_j \mathbf{p}_j^T\, \mathbf{p}_k$
  $= \mathbf{q}_k + \frac{1}{N}\sum_{j \neq k} \mathbf{q}_j \mathbf{p}_j^T\, \mathbf{p}_k$
- For random hyperdimensional vectors, $\mathbf{p}_j^T \mathbf{p}_k \approx 0$
- Therefore, $\mathrm{M}\mathbf{p}_k \approx \mathbf{q}_k$

BG + Cerebellum Capacities

- Learn what satisfies basic needs, and what to avoid (BG reward learning)
  - And what information to maintain in working memory (PFC) to support successful behavior
- Learn basic Sensory ⟶ Motor mappings accurately (Cerebellum error-driven learning)
  - Sensory ⟶ Sensory mappings? (what is going to happen next)

(slide < O'Reilly)

BG + Cerebellum Incapacities

- Generalize knowledge to novel situations
  - Lookup tables don't generalize well…
- Learn abstract semantics
  - Statistical regularities, higher-order categories, etc
- Encode episodic memories (specific events)
  - Useful for instance-based reasoning
- Plan, anticipate, simulate, etc…
  - Requires robust working memory

(slide < O'Reilly)

emergent Demonstration: Cereb