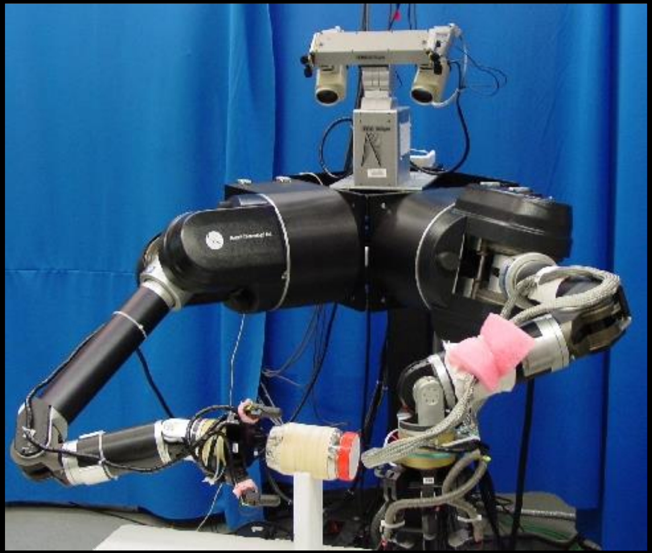


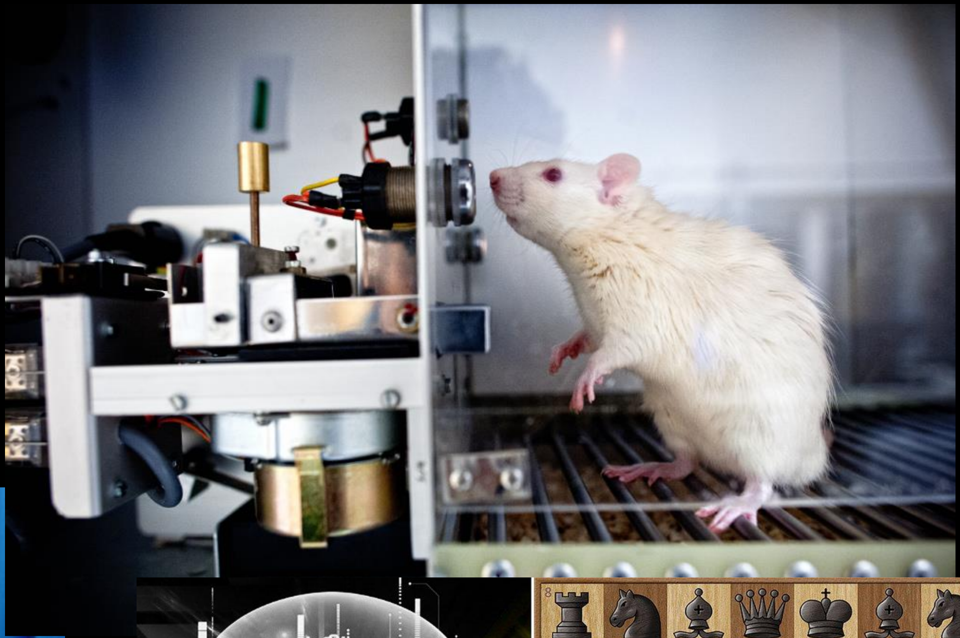
Reinforcement Learning

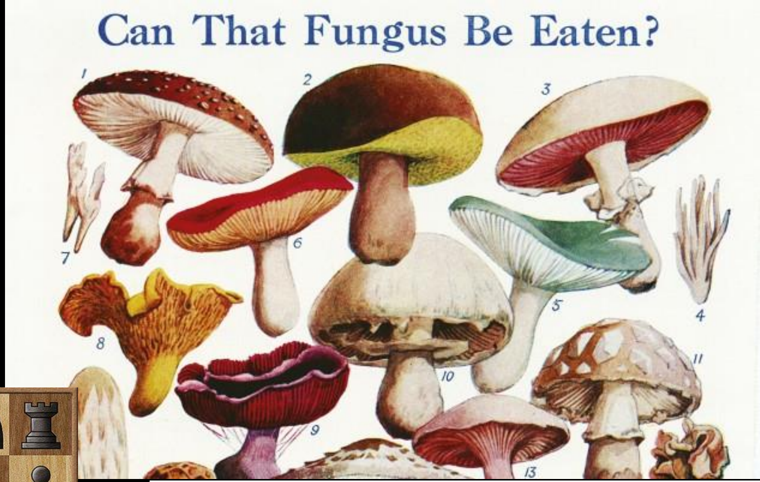
Guest Lecture, 10/15/2019

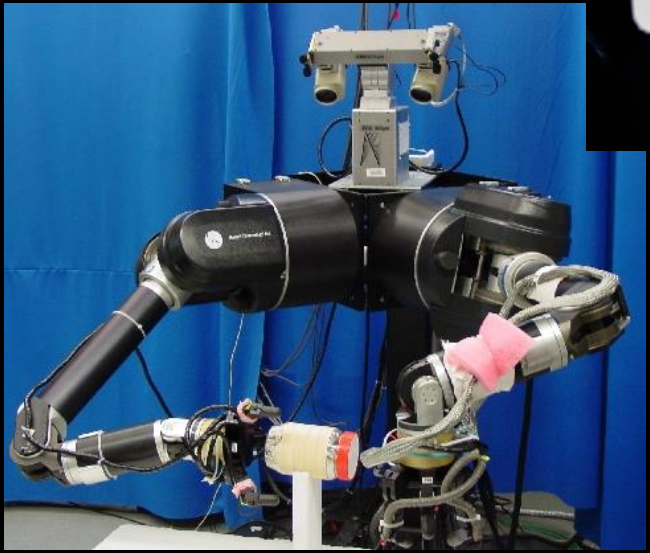
Andra Geana

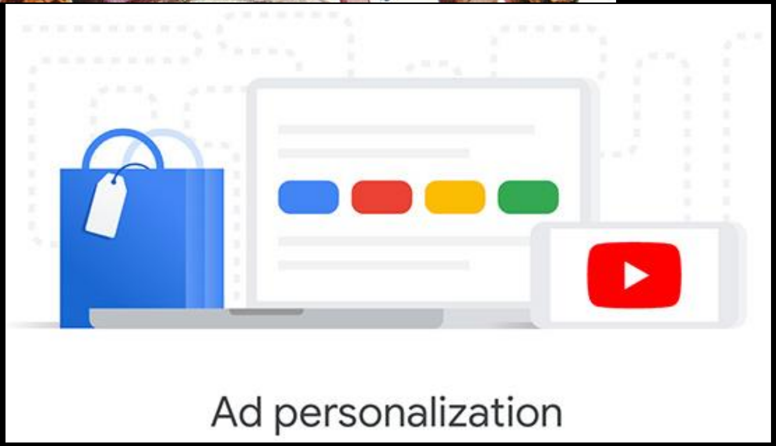
What is reinforcement learning?

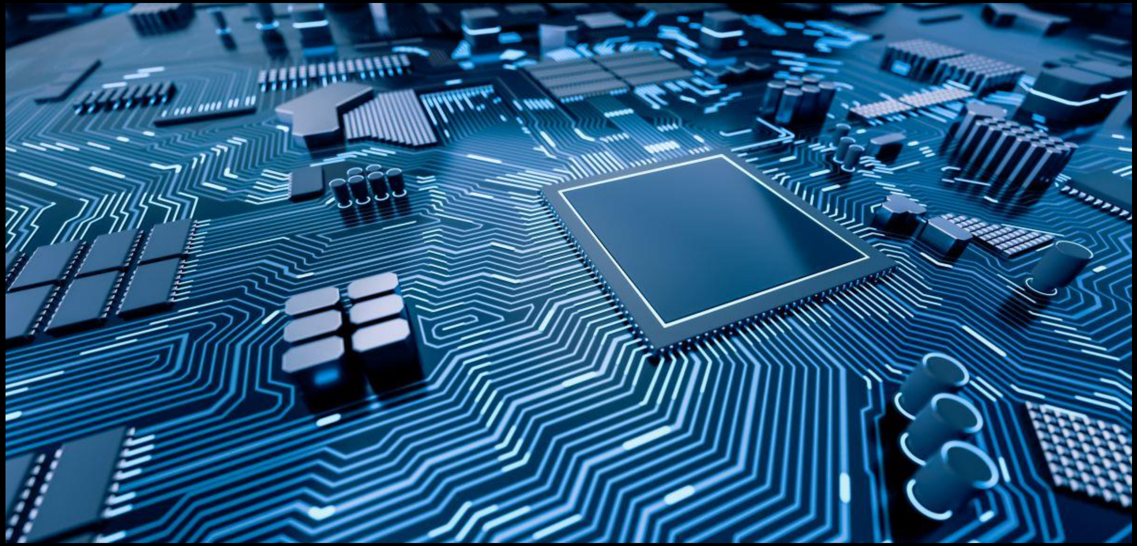
What is reinforcement learning?











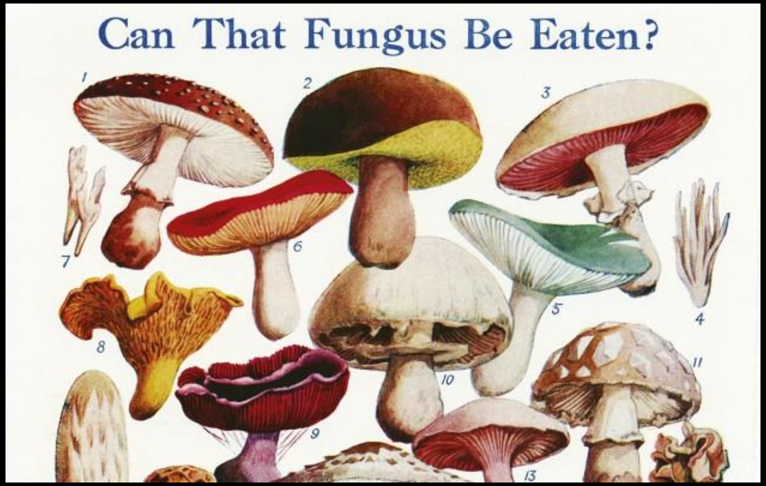
how does a system learn to do things on its own?

Why do people act & behave they do?

Why do we care about learning behaviors?

Why do we care about learning behaviors?

most living organisms need to learn what decisions will keep them alive!



Why do we care about learning behaviors?

most living organisms need to learn what decisions will keep them alive!

Reinforcement Learning principles help us describe and quantify how the learning happens, and how it leads to decisions

What is reinforcement learning?

Using trial-and-error to learn how to map situations to actions so as to maximize a numerical reward signal

Sutton & Barto, 2nd Ed, 2018

What is reinforcement learning?

Using trial-and-error to learn how to map situations to actions so as to maximize a numerical reward signal

Sutton & Barto, 2nd Ed, 2018

Reinforcement Learning

Reinforcement Learning

- Reinforcement (term from operant conditioning)
  - Something (e.g. food, money, game points, social/legal consequences) that makes it more likely that a certain response will re-occur

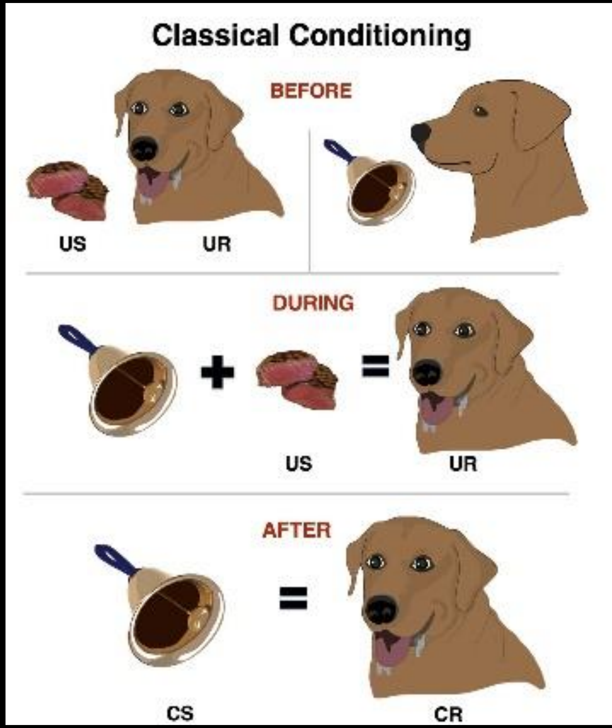
Reinforcement Learning

- Reinforcement (term from operant conditioning)
  - Something (e.g. food, money, game points, social/legal consequences) that makes it more likely that a certain response will re-occur


Reinforcement Learning

- Reinforcement (term from operant conditioning)
  - Something (e.g. food, money, game points, social/legal consequences) that makes it more likely that a certain response will re-occur

- Conditioning: training an organism to respond in specific ways to certain stimuli (thus potential to shape behavior)



Classical (Pavlovian): train involuntary responses (CR)

Instrumental (operant): train voluntary responses (actions)

Reinforcement learning: aimed to elicit voluntary responses (actions) in response to current situations (states), with the goal of maximizing rewards

Reinforcement learning: aimed to elicit voluntary responses (actions) in response to current situations (states), with the goal of maximizing rewards

Reinforcement learning: aimed to elicit voluntary responses (actions) in response to current situations (states), with the goal of maximizing rewards

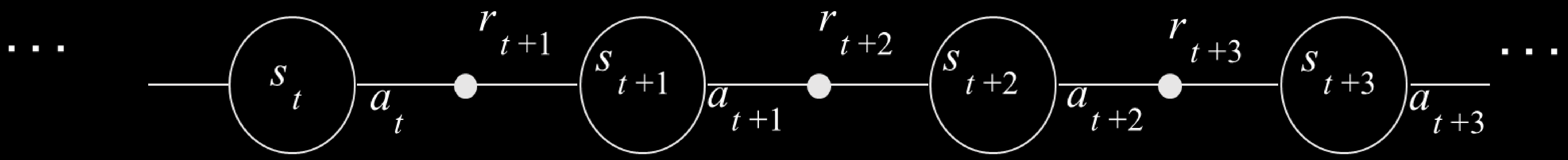
Agent

Environment
(e.g. the world)

Learning process: interaction b/w agent and environment

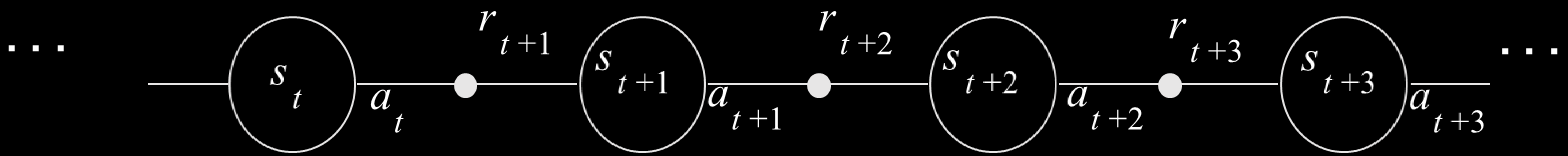
Agent

Environment
(e.g. the world)



Agent observes reward from current environment (e.g. state)
Based on observations, it updates its knowledge of the state
Based on updated knowledge, agent chooses action
Action moves agent into next state

repeat this until goals have been achieved …



LEARNING

Agent observes reward from current environment (e.g. state)
Based on observations, it updates its knowledge of the state
Based on updated knowledge, agent chooses action
Action moves agent into next state

DECISION

repeat this until goals have been achieved …

Reinforcement learning describes

a **learning** component: how do we use past experience to build our knowledge of the world?

a **decision** component: how do we use our knowledge of the world to choose actions?

A slightly more in-depth example

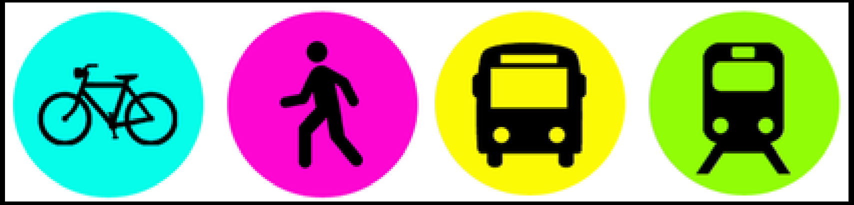
Real-world decision scenario: You are at home (in a major city) and have to meet a friend for coffee downtown in 20 minutes.

How do you get to your meeting?

- ENVIRONMENT STATE Where am I?

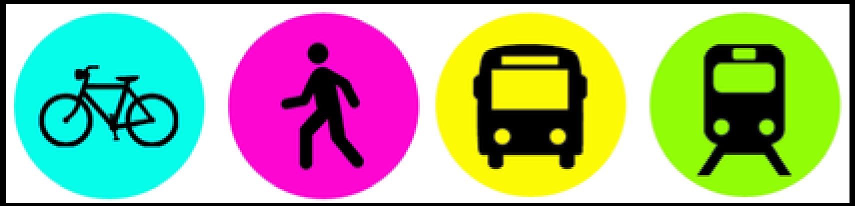
- ENVIRONMENT STATE Where am I?

- ACTION What can I do?



- ENVIRONMENT STATE Where am I?

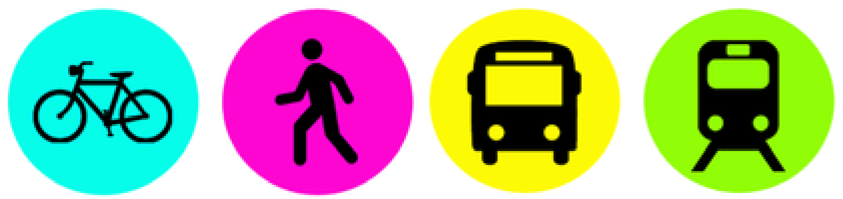
- ACTION What can I do?



- REWARD What benefit do I get from taking this action now?

- ENVIRONMENT STATE Where am I?

- ACTION What can I do?

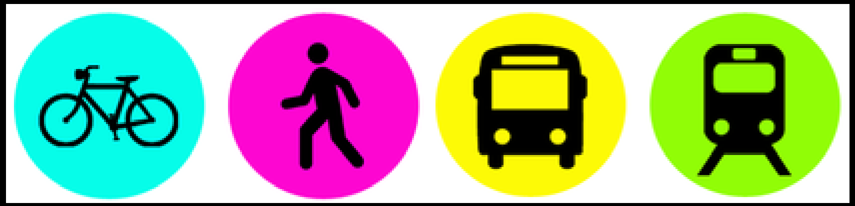


- REWARD What benefit do I get from taking this action now?

- VALUE What is this action worth to me?

- ENVIRONMENT STATE Where am I?

- ACTION What can I do?



- REWARD What benefit do I get from taking this action now?

- VALUE What is this action worth to me?

- CHOICE What action do I end up taking right now?

- ENVIRONMENT STATE — Where am I?
- ACTION — What can I do?
- REWARD — What benefit do I get from taking this action now?
- VALUE — What is this action worth to me?
- CHOICE — What action do I end up taking right now?

Learning

Decision

- ENVIRONMENT STATE Where am I?
- ACTION What can I do?
- REWARD What benefit do I get from taking this action now?
- VALUE What is this action worth to me?
- CHOICE What action do I end up taking right now?

We decide our next actions based on what we've learned from interacting with the environment in the past

- ENVIRONMENT STATE — Where am I?
- ACTION — What can I do?
- REWARD — What benefit do I get from taking this action now?
- VALUE — What is this action worth to me?
- CHOICE — What action do I end up taking right now?

- ENVIRONMENT STATE — Where am I?
- ACTION — What can I do?
- $\boxed{\text{REWARD}}$ — What benefit do I get from taking this action now?
- $\boxed{\text{VALUE}}$ — What is this action worth **to me**?
- CHOICE — What action do I end up taking right now?

Values are subjective: one person may value time costs above $$ or physical effort, and thus choose driving, while another may value exercise benefits over time costs and choose biking etc

- ENVIRONMENT STATE Where am I?
- ACTION What can I do?
- REWARD What benefit do I get from taking this action now?
- VALUE What is this action worth **to me**?
- CHOICE What action do I end up taking right now?

Reward/value space is a "numerical signal" (e.g. impose a scalar value on whatever your physical reward space is—you have to do this for RL to work)

Questions so far?

How do we quantify these processes?

Rescorla & Wagner (1972): A basic Reinforcement Learning algorithm

Rescorla & Wagner (1972): A basic Reinforcement Learning algorithm

Basic assumption: error-driven learning

change in value is proportional to the difference between actual and predicted outcome

Rescorla & Wagner (1972): A basic Reinforcement Learning algorithm

Basic assumption: error-driven learning

change in value is proportional to the difference between actual and predicted outcome

e.g. "surprise drives learning"

The more surprised we are by an outcome, the more we use that outcome to update our knowledge of the world

Goal: Maximize reward!!

Rescorla & Wagner (1972): A basic Reinforcement Learning algorithm

"surprise drives learning": The more surprised we are by an outcome, the more we use that outcome to update our knowledge of the world

Current knowledge: Traffic downtown is bad. H Street is usually the least jammed

Action: I will choose H Street. That's the best one to take.

Rescorla & Wagner (1972): A basic Reinforcement Learning algorithm

"surprise drives learning": The more surprised we are by an outcome, the more we use that outcome to update our knowledge of the world

Current knowledge: Traffic downtown is bad. H Street is usually the least jammed

Action: I will choose H Street. That's the best one to take.

Case 1 Outcome:

Rescorla & Wagner (1972): A basic Reinforcement Learning algorithm

"surprise drives learning": The more surprised we are by an outcome, the more we use that outcome to update our knowledge of the world

Current knowledge: Traffic downtown is bad. H Street is usually the least jammed

Action: I will choose H Street. That's the best one to take.

Case 1 Outcome:

Value updating: ¯_(ツ)_/¯H Street is sometimes jammed, but (probably) still not as bad as the rest. Still probably the best one to take.

Rescorla & Wagner (1972): A basic Reinforcement Learning algorithm

"surprise drives learning": The more surprised we are by an outcome, the more we use that outcome to update our knowledge of the world

Current knowledge: Traffic downtown is bad. H Street is usually the least jammed

Action: I will choose H Street. That's the best one to take.

Case 1 Outcome

Case 2 Outcome:

Rescorla & Wagner (1972): A basic Reinforcement Learning algorithm

"surprise drives learning": The more surprised we are by an outcome, the more we use that outcome to update our knowledge of the world

Current knowledge: Traffic downtown is bad. H Street is usually the least jammed

Action: I will choose H Street. That's the best one to take. .

Case 1 Outcome

Case 2 Outcome:

Rescorla & Wagner (1972): A basic Reinforcement Learning algorithm

"surprise drives learning": The more surprised we are by an outcome, the more we use that outcome to update our knowledge of the world

Current knowledge: Traffic downtown is bad. H Street is usually the least jammed

Action: I will choose H Street. That's the best one to take.

Case 1 Outcome

Case 2 Outcome:

Value updating: H STREET IS THE WORST EVER!!!!!

Rescorla & Wagner (1972): A basic Reinforcement Learning algorithm

PREDICTION ERROR (SURPRISE) DRIVES LEARNING

change in value is proportional to the difference between actual and predicted outcome

Rescorla & Wagner (1972): A basic Reinforcement Learning algorithm

PREDICTION ERROR (SURPRISE) DRIVES LEARNING

change in value is proportional to the difference between actual and predicted outcome

$$V_{k+1}(s_k) = V_k(s_k) + \alpha\delta_k$$

Value at time k+1 of state k (e.g. street X)

Previous value of Sk (e.g. predicted outcome)

Learning rate: how much do we care about each new data point?

Rescorla & Wagner (1972): A basic Reinforcement Learning algorithm

PREDICTION ERROR (SURPRISE) DRIVES LEARNING

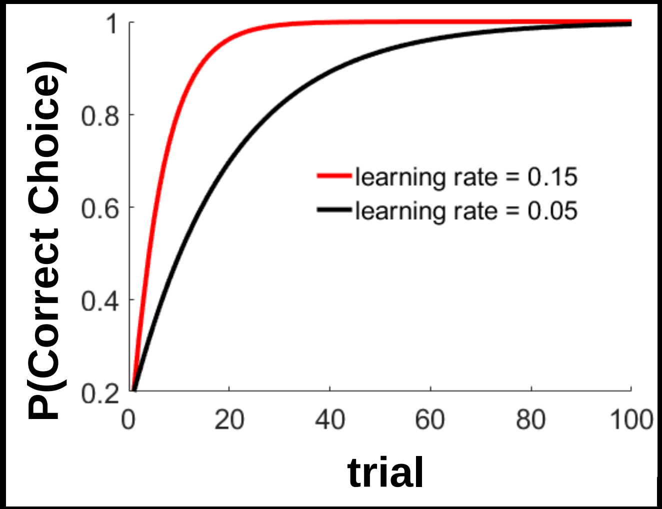
change in value is proportional to the difference between actual and predicted outcome

$$V_{k+1}(s_k) = V_k(s_k) + \alpha\delta_k$$

Value at time k+1 of state k (e.g. street X)

Previous value of Sk (e.g. predicted outcome)

Learning rate: how much do we care about each new data point?



Higher learning rates means we learn faster!

Rescorla & Wagner (1972): A basic Reinforcement Learning algorithm

PREDICTION ERROR (SURPRISE) DRIVES LEARNING

change in value is proportional to the difference between actual and predicted outcome

$$V_{k+1}(s_k) = V_k(s_k) + \alpha\delta_k$$

Actual outcome

$$\delta_k = r_k - V_k(s_k)$$

Value at time k+1 of state k (e.g. street X)

Previous value of Sk (e.g. predicted outcome)

Learning rate: how much do we care about each new data point?

Prediction error: difference b/w what we predicted and what actually happened

Rescorla & Wagner (1972): A basic Reinforcement Learning algorithm

Surprise drives learning: change in value is proportional to the difference between actual and predicted outcome

Learning

$$V_{k+1}(s_k) = V_k(s_k) + \alpha\delta_k$$

$$\delta_k = r_k - V_k(s_k)$$

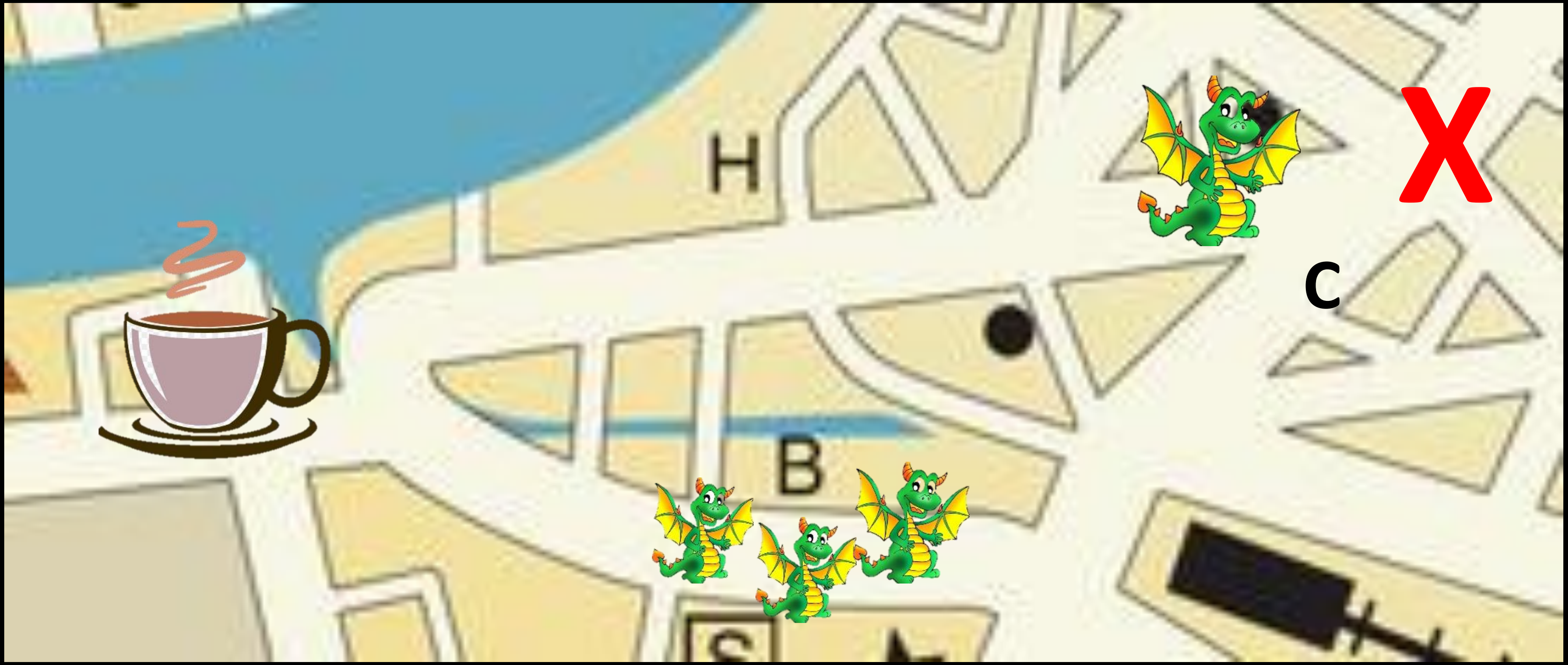
Decision:

$$Choice_k = \max(V_k)$$

e.g. choose action that will put you in the highest-value state

But what if…

But what if…



But what if…

Maximize (final/total) reward over a sequence of actions → need to look in the future

TD-Learning: R-W's cousin that looks at future rewards

$$V_t = E\left[\sum_{i=t}^{T} r_i\right]$$

Value of a state at time t is the expected value of all future rewards from time t on

(compare to R-W that estimates V_t based on weighted average of past rewards $V_t = \eta \sum_{i=1}^{t} (1-\eta)^{t-i} r_i$)

$$\begin{aligned}
V_t &= E\left[r_t + r_{t+1} + r_{t+2} + \ldots + r_T\right] \\
&= E\left[r_t\right] + E\left[r_{t+1} + r_{t+2} + \ldots + r_T\right] \\
&= E\left[r_t\right] + V_{t+1} \\
\delta_t &= E\left[r_t\right] + V_{t+1} - V_t
\end{aligned}$$

(compare to R-W $\delta = r_t - V_t$)

TD-Learning: R-W's cousin that looks at future rewards

$$V(S_t) \leftarrow V(S_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]$$

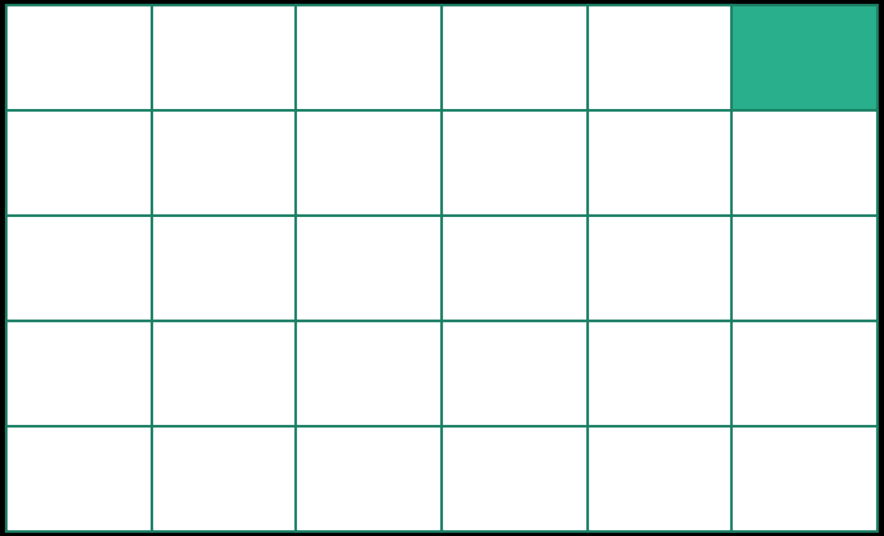
Learn values for states based on all (discounted) future rewards

How do we know future rewards?

- We don't, at first
- That's why we iterate over these tasks and refine our model of the world

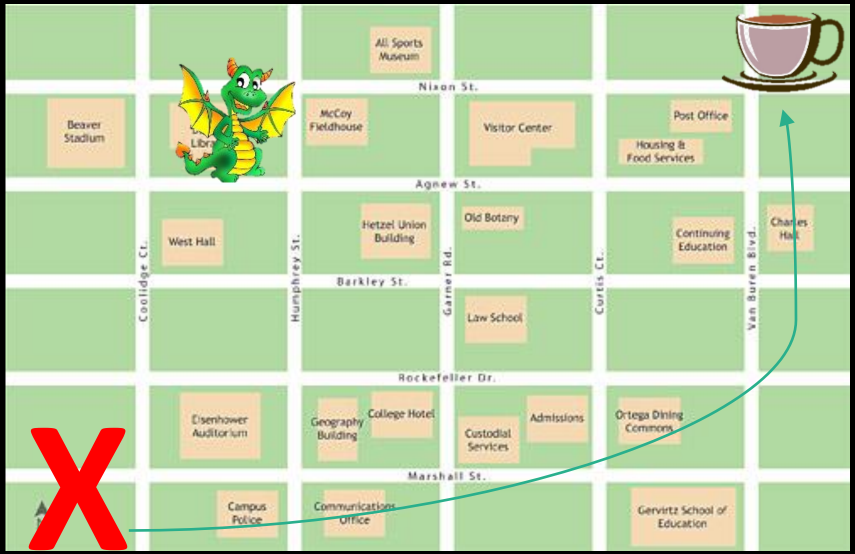
TD-Learning: R-W’s cousin that looks at future rewards

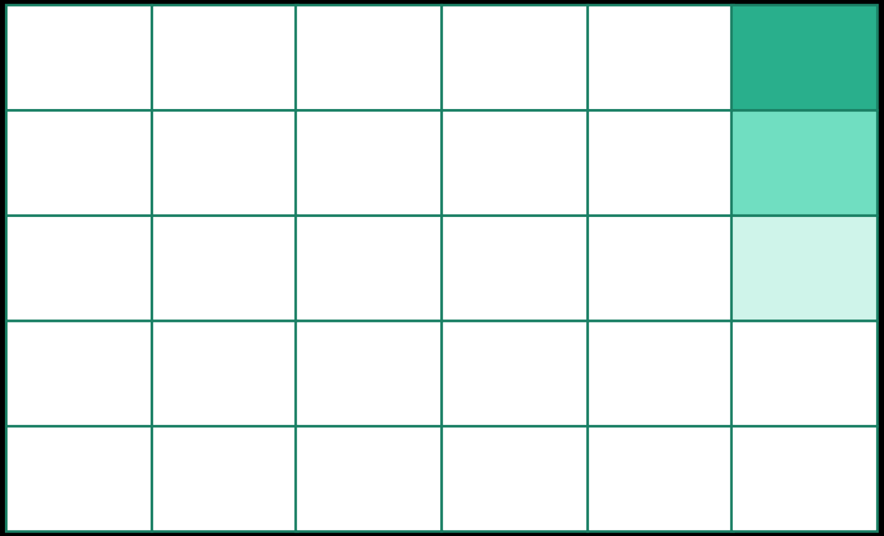
$$V(S_t) \leftarrow V(S_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]$$



TD-Learning: R-W's cousin that looks at future rewards

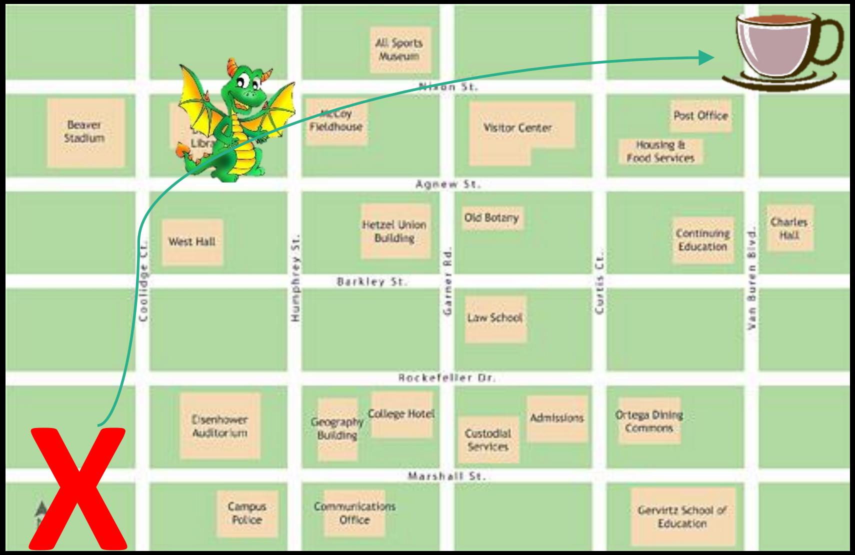
$$V(S_t) \leftarrow V(S_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]$$

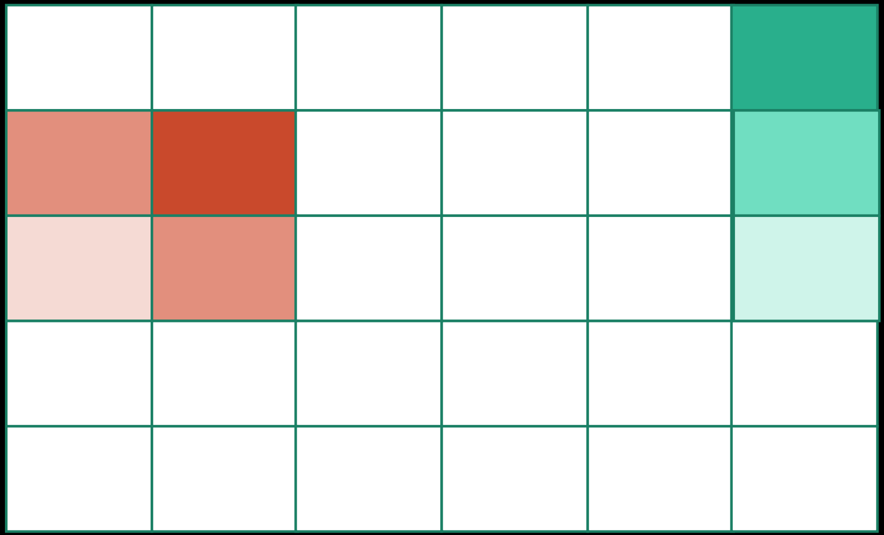




TD-Learning: R-W's cousin that looks at future rewards

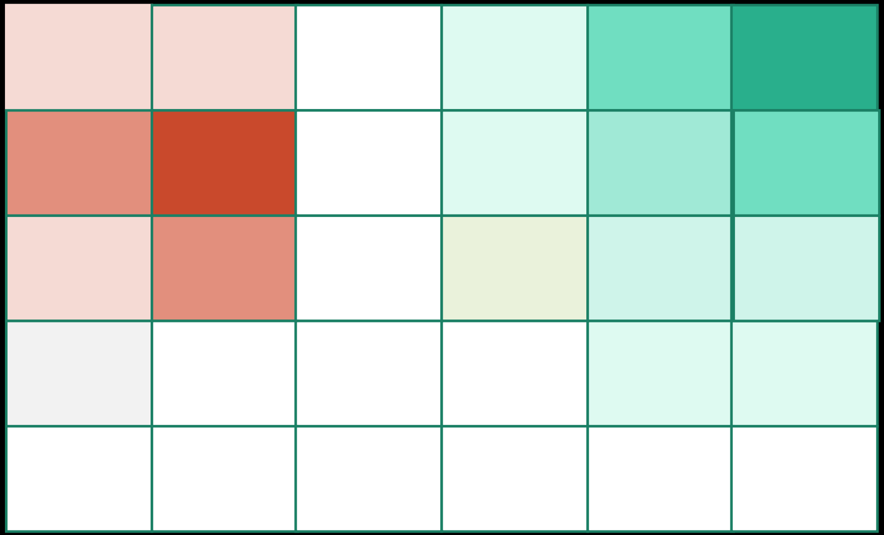
$$V(S_t) \leftarrow V(S_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]$$





TD-Learning: R-W’s cousin that looks at future rewards

$$V(S_t) \leftarrow V(S_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]$$



Many flavors of TD learning

- Q-learning, SARSA: assign values to state-action combinations (rather than states alone)

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)]$$

- On-policy: take into account what agent's current policy is
- Off-policy: always assume maximizing action

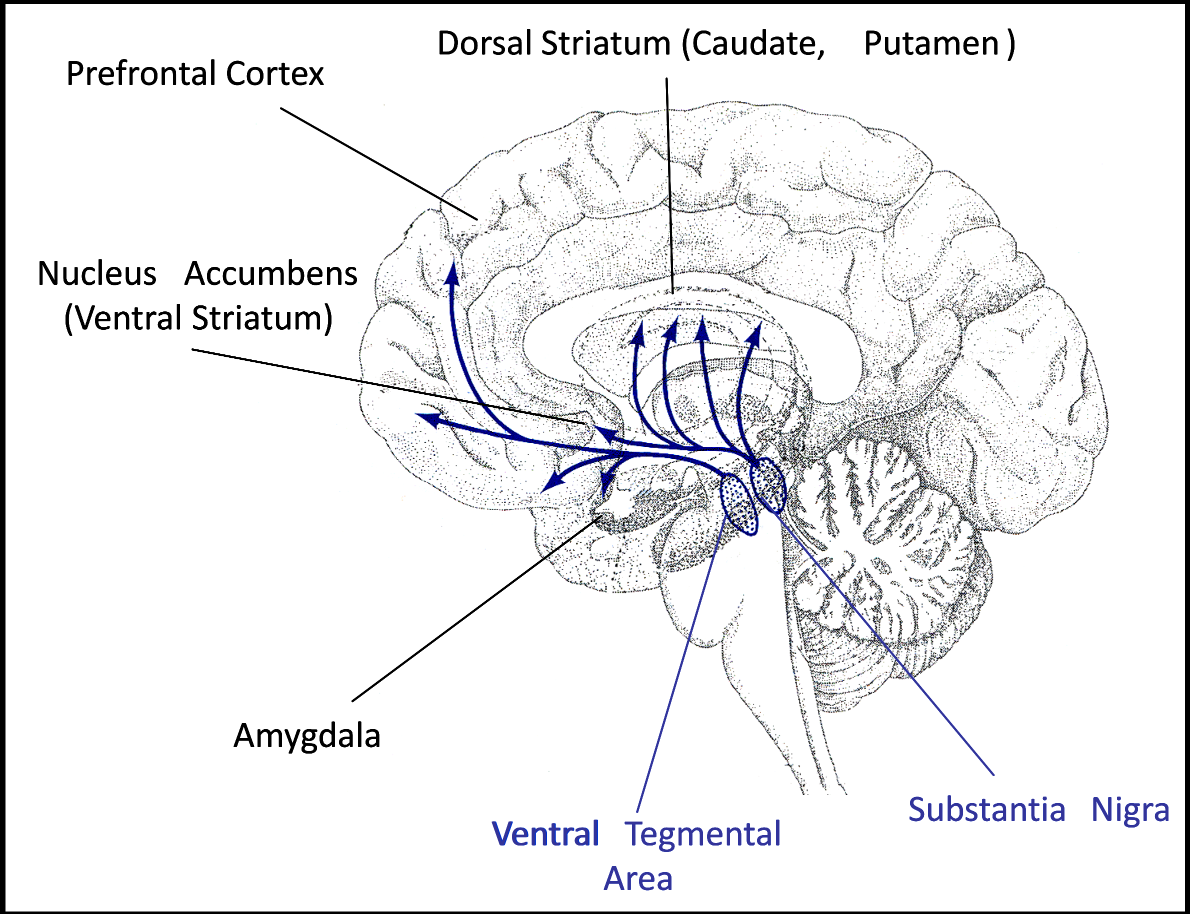
TD Learning: learn policy (mapping state to actions) to maximize sum of future rewards

Reinforcement Learning maps onto the Brain!

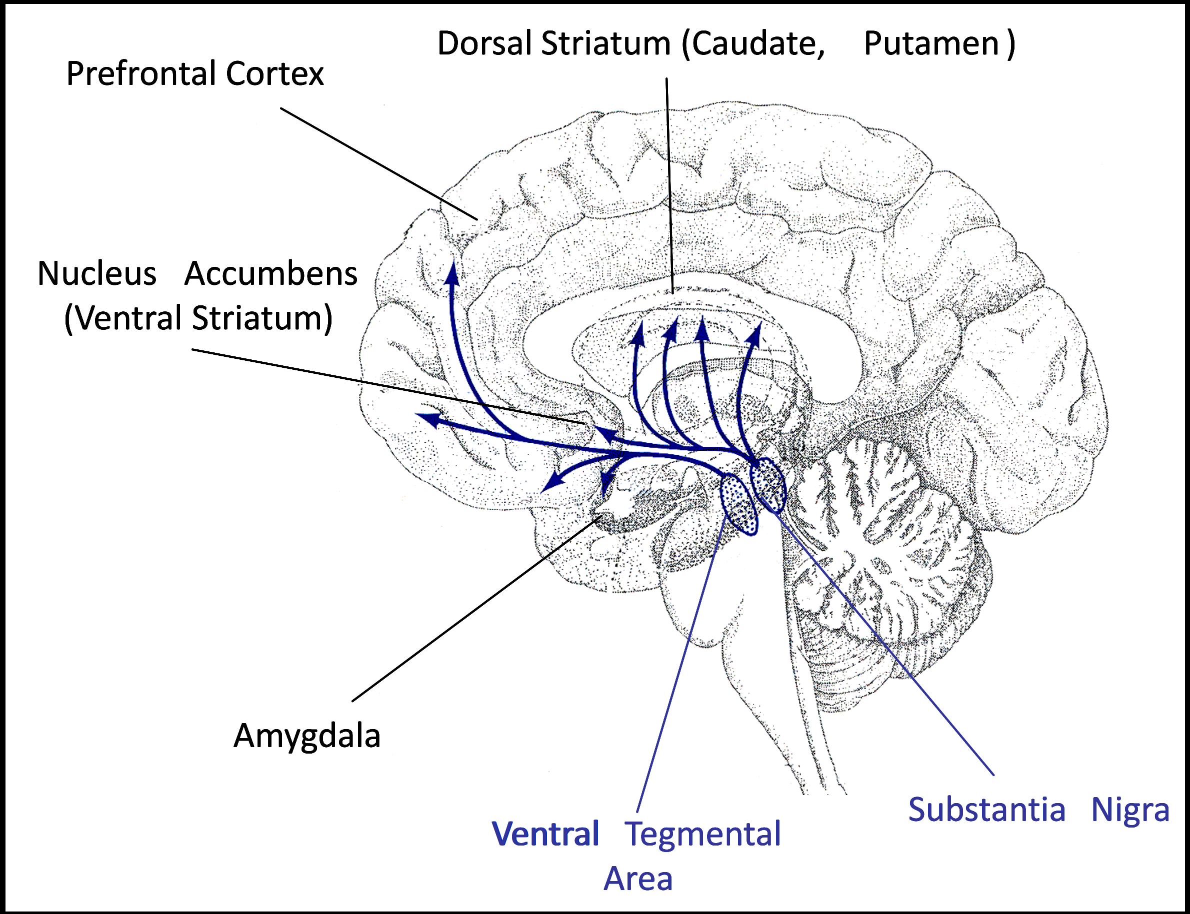
Neural implementation: (how) Does the brain do reinforcement learning?

- One of the advantages of RL framework is that research has found evidence for neural implementation

Dopamine


Dorsal Striatum (Caudate, Putamen)
Prefrontal Cortex
Nucleus Accumbens
(Ventral Striatum)
Amygdala
Ventral Tegmental
Area
Substantia Nigra

Dopamine

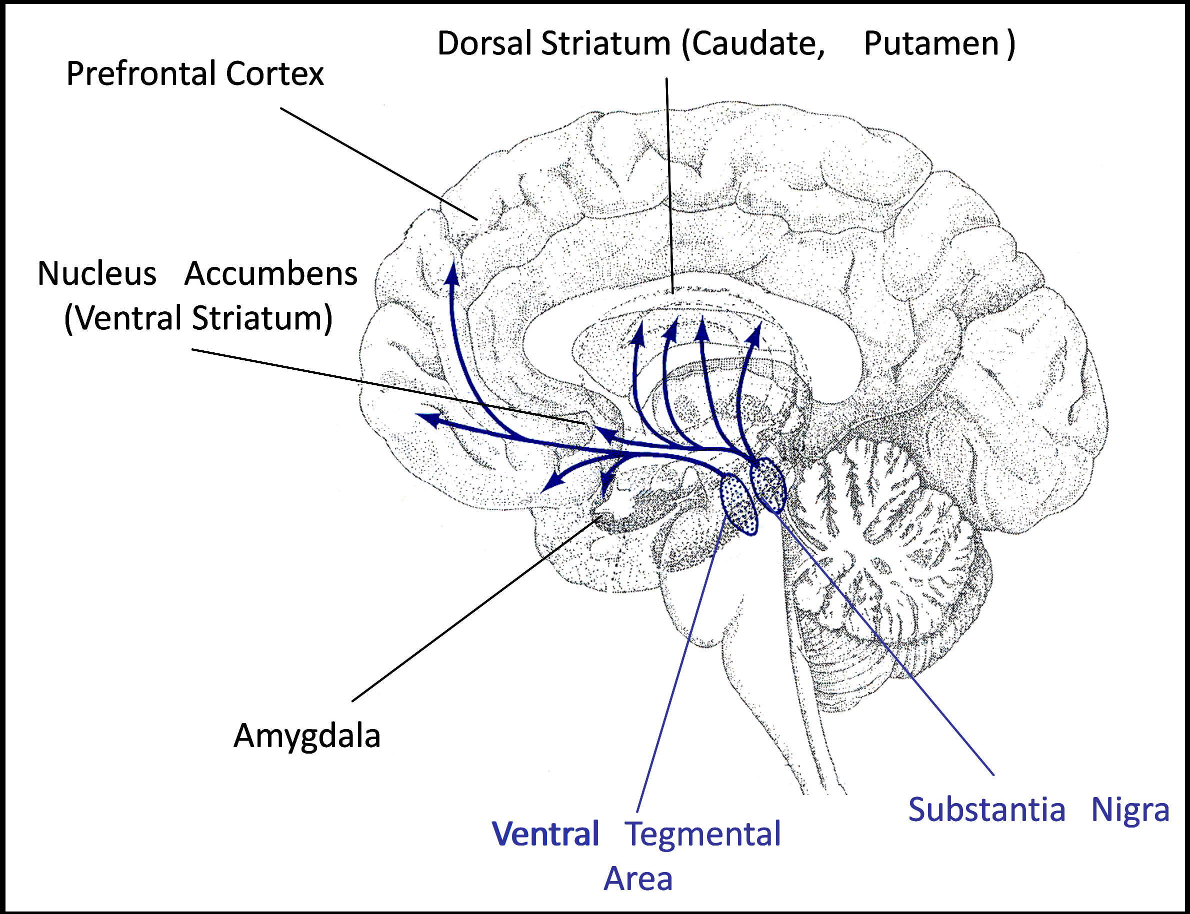


Several dopamine pathways:

- Reward (VTA→NAcc)
- Executive (VTA→PFC)
- Motor (SNc→striatum)

Each linked to different cognitive and psychopathology phenomena

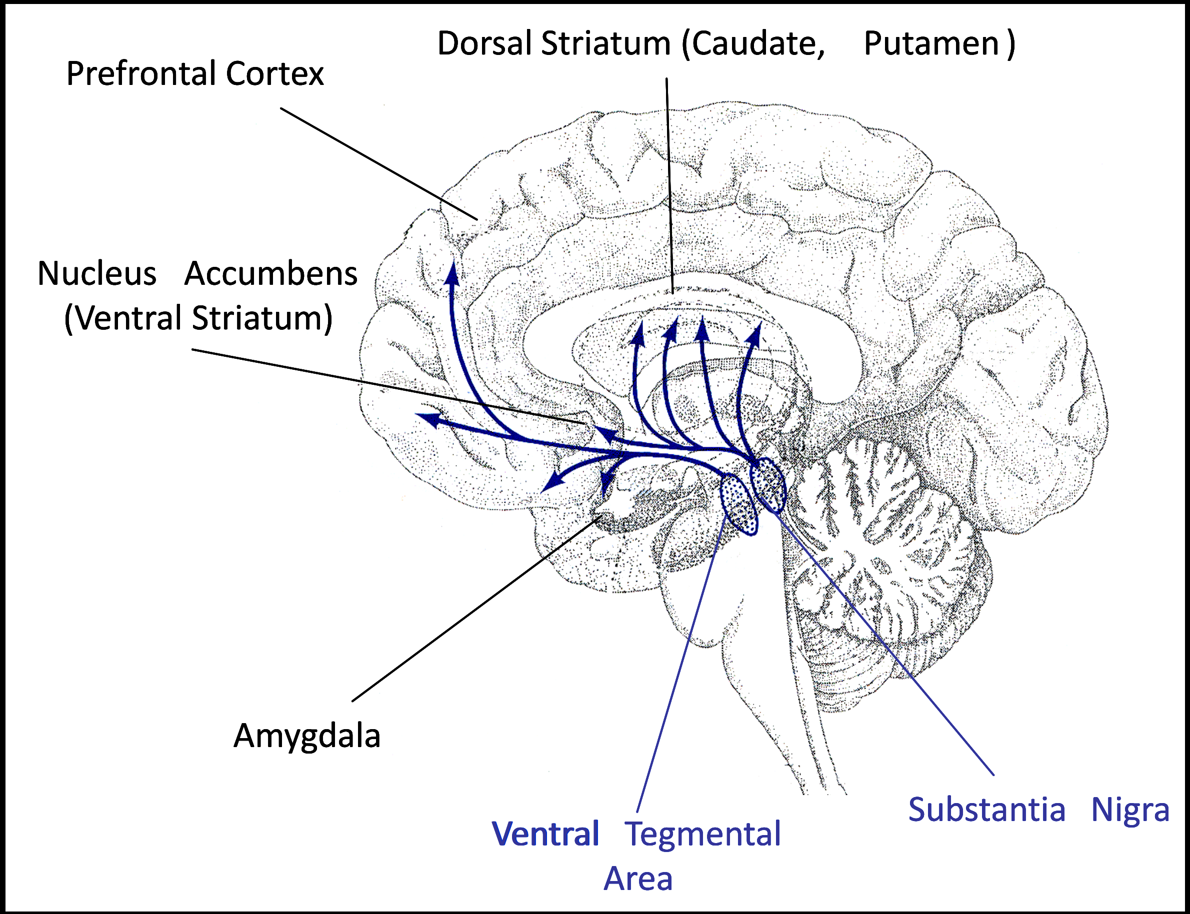
Dopamine



"It turns out that dopamine is a chemical on double duty in the brain. Along with its role in motor commands, it also serves as the main messenger in the reward systems, guiding a person toward food, drink, mates, and all things useful for survival.

…imbalances in dopamine can trigger gambling, overeating, and drug addiction - behaviors that result from a reward system gone awry."

Dopamine



Parkinson's Disease

→ Motor control + initiation?

Intracranial self-stimulation;

Drug addiction;

Natural rewards

→ Reward pathway?

→ Learning?

Also involved in:

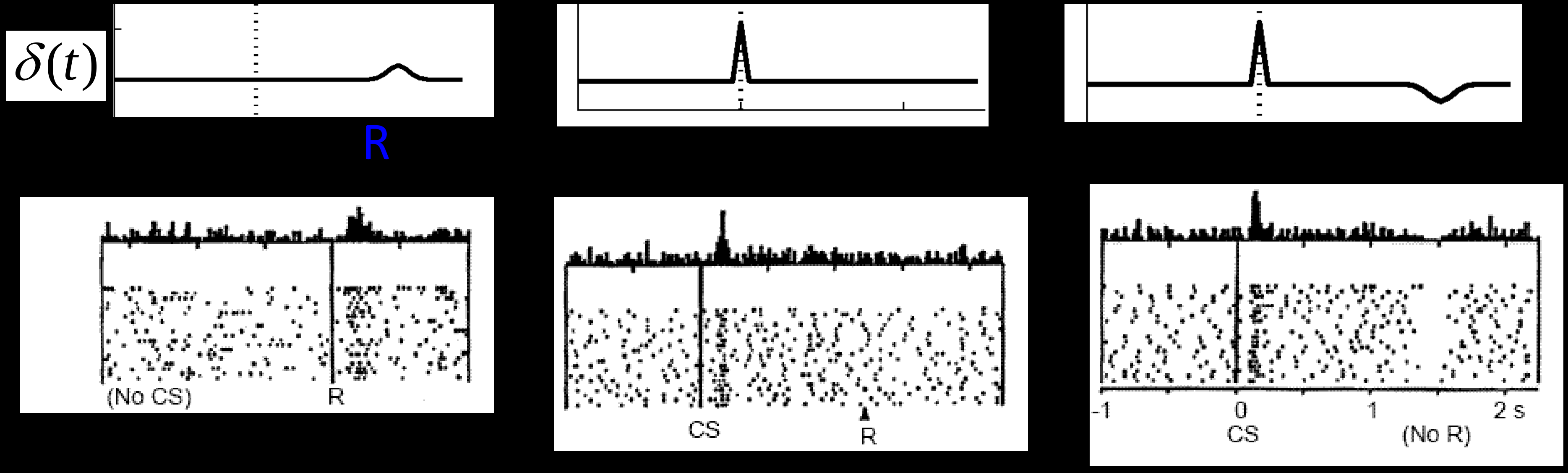
- Working memory
- Novel situations
- ADHD
- Schizophrenia
- …

Role of dopamine: Many hypotheses

- Avolition/anhedonia hypothesis
- Learning, action selection
- Salience/attention
- Uncertainty
- Cost/benefit computation
- Energizing/motivating behavior

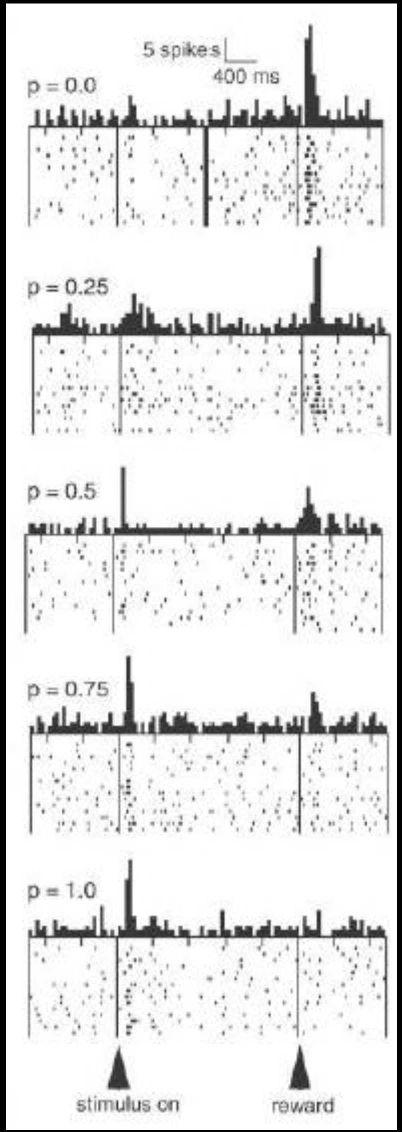
dopamine and prediction error

$$\delta_t = r_t + V_{t+1} - V_t$$



Schultz 1997; via A. Guez

prediction error hypothesis of dopamine

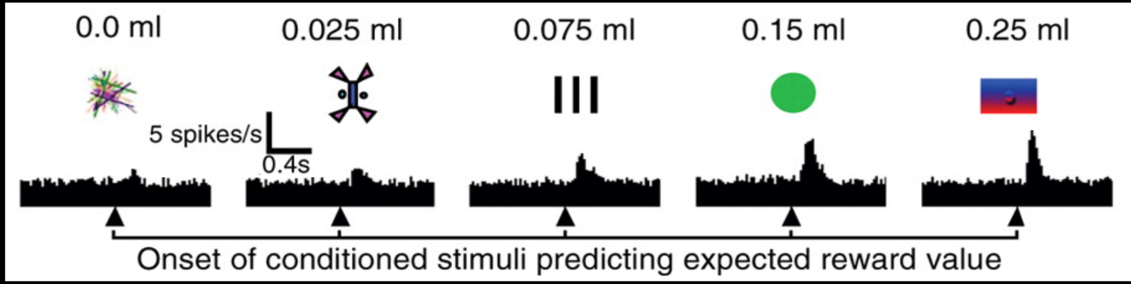


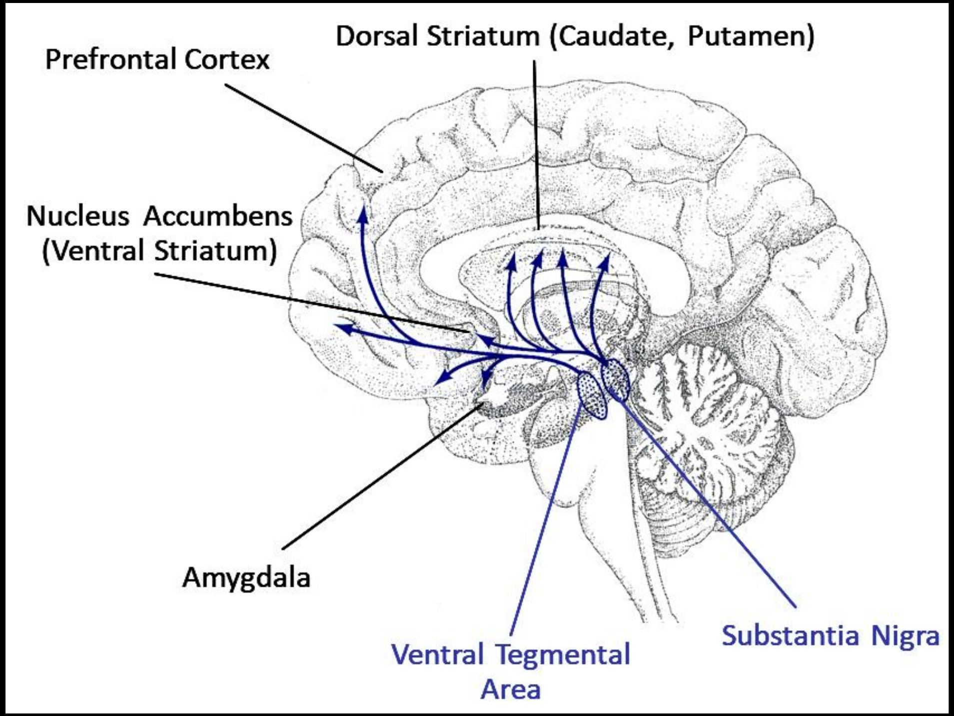
Fiorillo et al, 2003

The idea: Dopamine encodes a reward prediction error

This can be sensitive to probability of reward, magnitude of reward, overall range (and others)

Tobler et al, 2005





Recap: activity of striatal dopamine neurons matches predictions from RL algorithms

- dip with negative prediction error (omitted reward)
- boost with positive prediction error (unexpected reward)

Recap

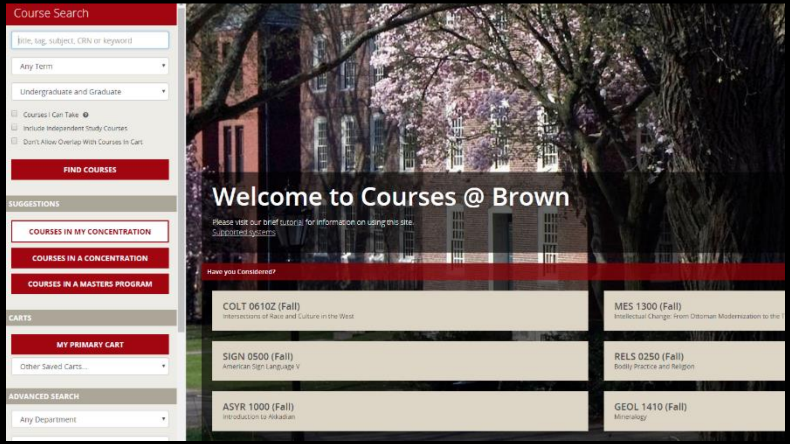
RL frames decision scenarios as based on past learning experiences + choice rules

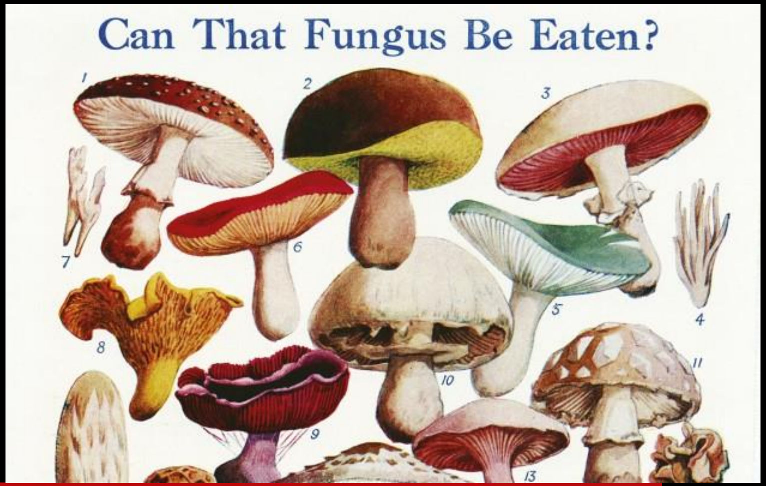
Core concepts: states, actions, values, rewards, choices

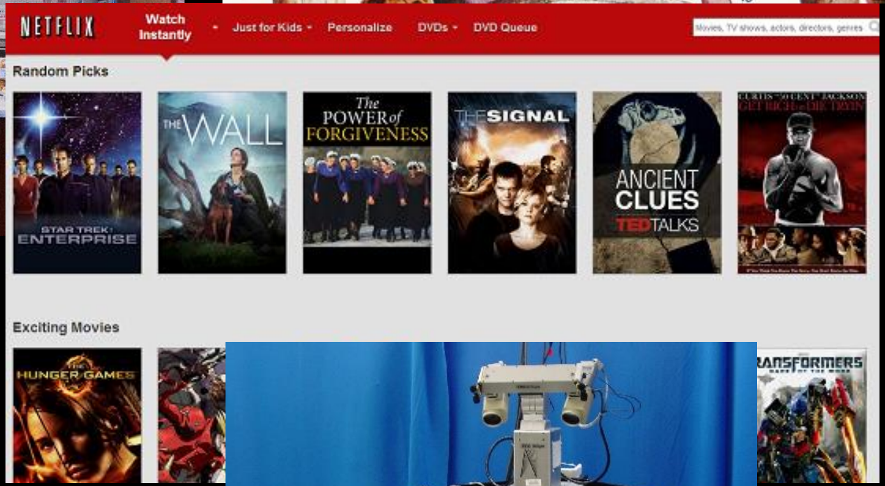
Relies on (operant) conditioning principles

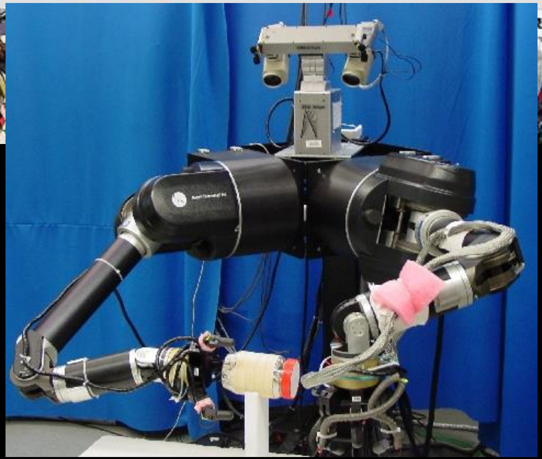
Evidence of neural implementation in dopamine (DA) neurons

Why do we care?



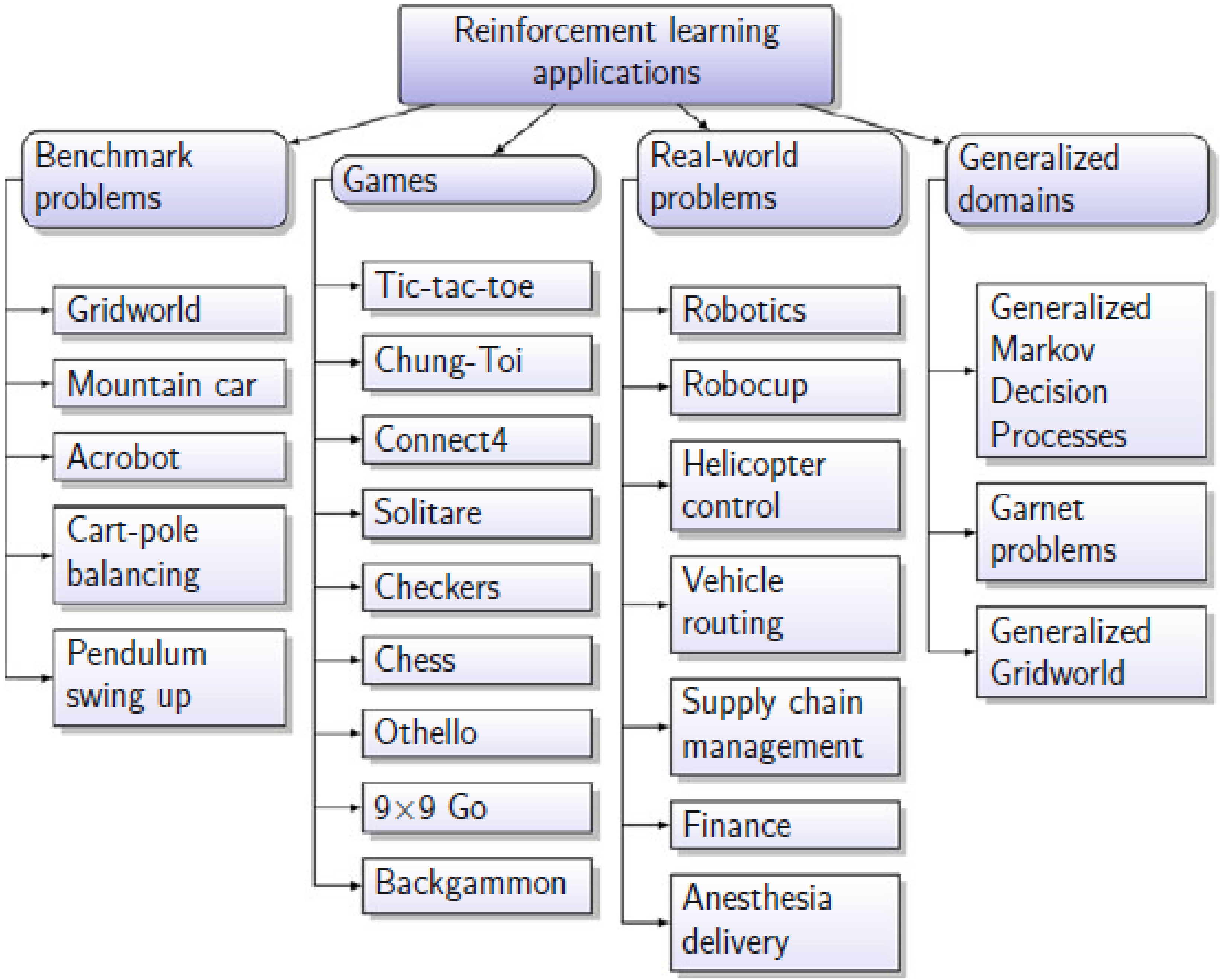




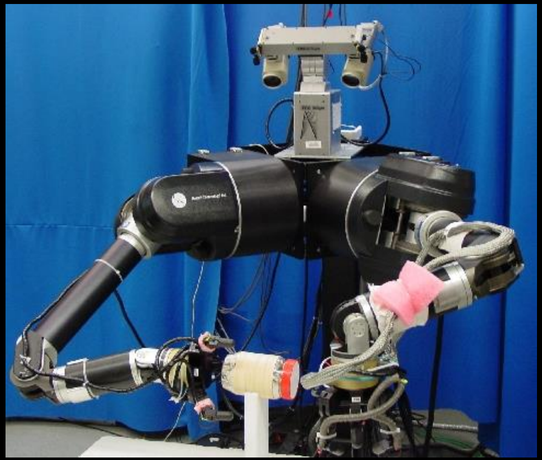


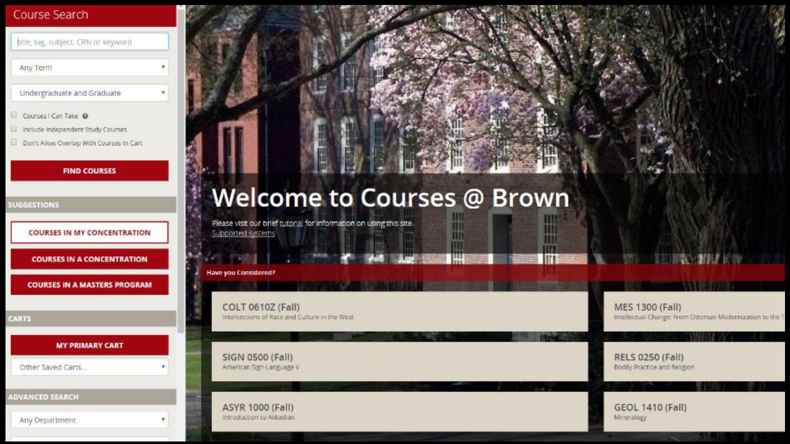
Most decision problems can be framed as RL, by tweaking definition of states, actions, and rewards

- Mathematically, this is useful b/c RL provides already-solved algorithms for making optimal decisions (e.g. “the knapsack problem”)
- Also v helpful for training artificial intelligence to perform various tasks (e.g. laundry-folding robot) or play games



Gatti & Embrechts (2012)

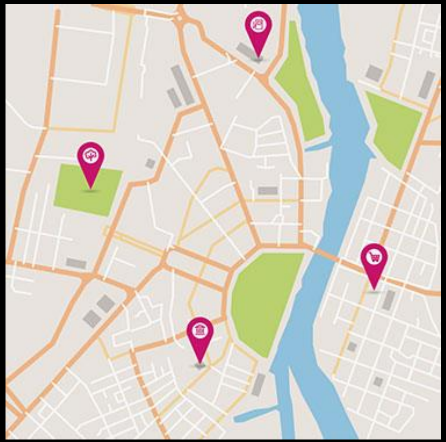



Course Search
Any Term
Undergraduate and Graduate
FIND COURSES
SUGGESTIONS
COURSES IN MY CONCENTRATION
COURSES IN A CONCENTRATION
COURSES IN A MASTERS PROGRAM
CARTS
MY PRIMARY CART
ADVANCED SEARCH
Any Department
Welcome to Courses @ Brown
Have you Considered?
COLT 0610Z (Fall)
MES 1300 (Fall)
SIGN 0500 (Fall)
RELS 0250 (Fall)
ASYR 1000 (Fall)
GEOL 1410 (Fall)


DeepMind

THE END


Boston Dynamics




Cakes
chocolate layer
cupcakes
coconut layer
Boston cream pie
red velvet
chocolate raspberry
cookies 'n cream
Strawberry layer
carrot
German chocolate
Buttercream layer
Devils food