What is reinforcement learning (RL)?

"Reinforcement learning is an area of machine learning inspired by behaviorist psychology, concerned with how an agent can learn from interactions with an environment."

- Wikipedia / Sutton & Barto / Phil 1998

Agent - Environment Diagram

Environment: World, lab, software environment.

Blood glucose (sugar) vs. time (1 day) shows hyperglycemia and hypoglycemia.
Neuroscience: How do animals learn? A specific agent or set of agents.

- The study of some examples of learning and intelligence.

R2: How can we make an agent that learns?

- The study of learning & intelligence.

Dopamine & Temporal Difference Error (in general)
Two most related fields
- Operations research
- Control (adaptive / classical)

Main difference is that these fields typically assume the environment (Plant) can and should be directly approximated.
When should you use RL?

→ As a last resort.

key properties:

Next time!
Key properties:

1) Evaluative feedback
   (not instructional)

   "This is how good the outcome was; not "this is what you should have done."

2) Sequential.

   No data set.
MENACE (Donald Michie, 1961)

<table>
<thead>
<tr>
<th>X</th>
<th>O</th>
<th>X</th>
<th>O</th>
<th>O</th>
<th>O</th>
</tr>
</thead>
</table>

304 matchboxes.
4 bead colors.
6 possible moves.

Win: Put back bead, +3 more of same color.
Lose: Remove beads.
Tie: Put back bead, +1 more of same color.
(Draw)

Idea:
- Act at least slightly random.
- Good outcome
  - take chosen actions more often
- Bad outcome
  - take chosen actions less often.
Act randomly? ??

- "Necessary" for learning!
- "Exploration" vs " exploitation".

Choosing actions that think are suboptimal.

$t$: time step.
$t \in \{0, 1, 2, \ldots, \infty\}$

$S_t$: State at time $t$.
$A_t$: Action at time $t$.
$R_t$: Reward at time $t$.

$A_t \in \text{POLICY}$

A policy $\pi$ is a way of selecting actions.

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

$A_t \sim \pi(S_t)$

- Many policies - some good, some bad.
Agent's goal: find a policy that maximizes the expected amount of reward the agent receives. (not deterministic)

Objective function: \[ J(\pi) = \mathbb{E}\left[ \sum_{t=0}^{\infty} R_t \mid \pi \right] \]

- Could be infinite, \( J(\pi) \)
- Doesn't discount based on time
- cares about short term
- care about long term. Typically, \( \gamma < 1 \)

YAP: yet another hyperparameter. \( \beta \in [0, 1] \)
If we care about

agent learns quickly

agent learns slowly.

Agent's goal is to find or approximate

\[ \pi^* = \arg \max_{\pi} J(\pi) \]

\[ \pi^* \] is called an optimal policy.