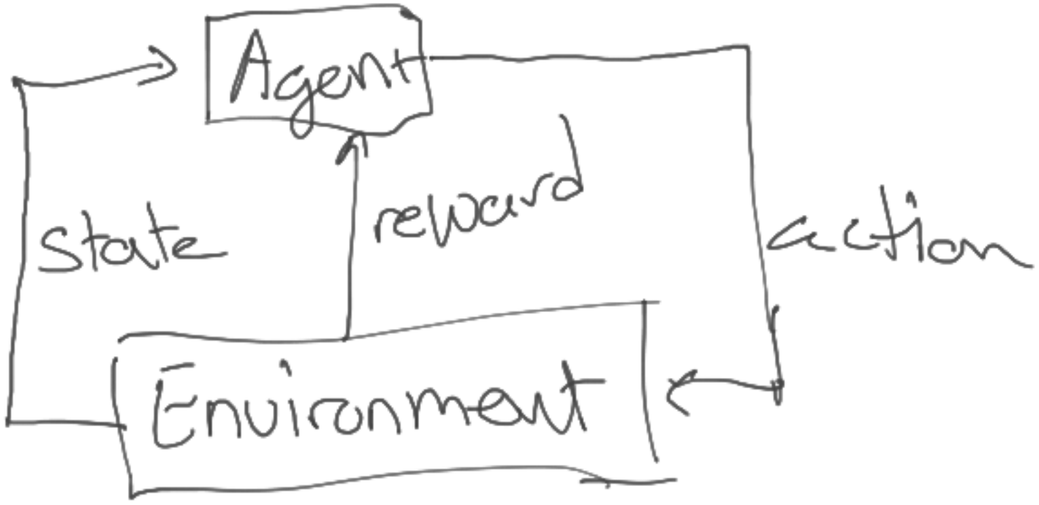
What is reinforcement learning (RL)?

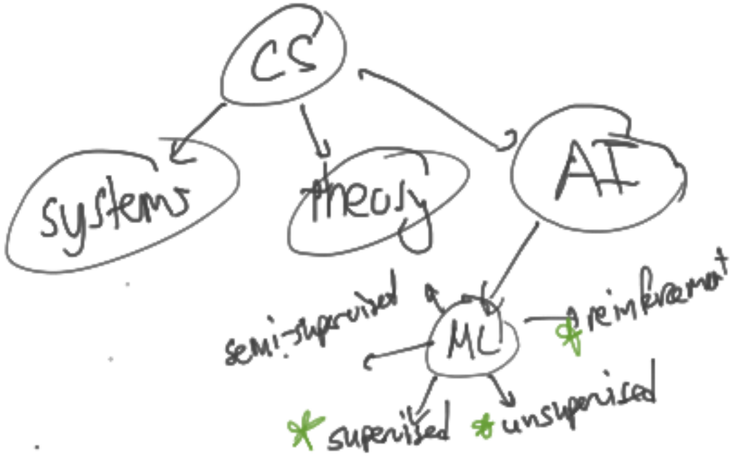
"Reinforcement learning is an area of machine learning, inspired by behaviorist psychology, concerned with how an agent can learn from interactions with an environment."

– Wikipedia / Sutton & Barto / Phil

1998

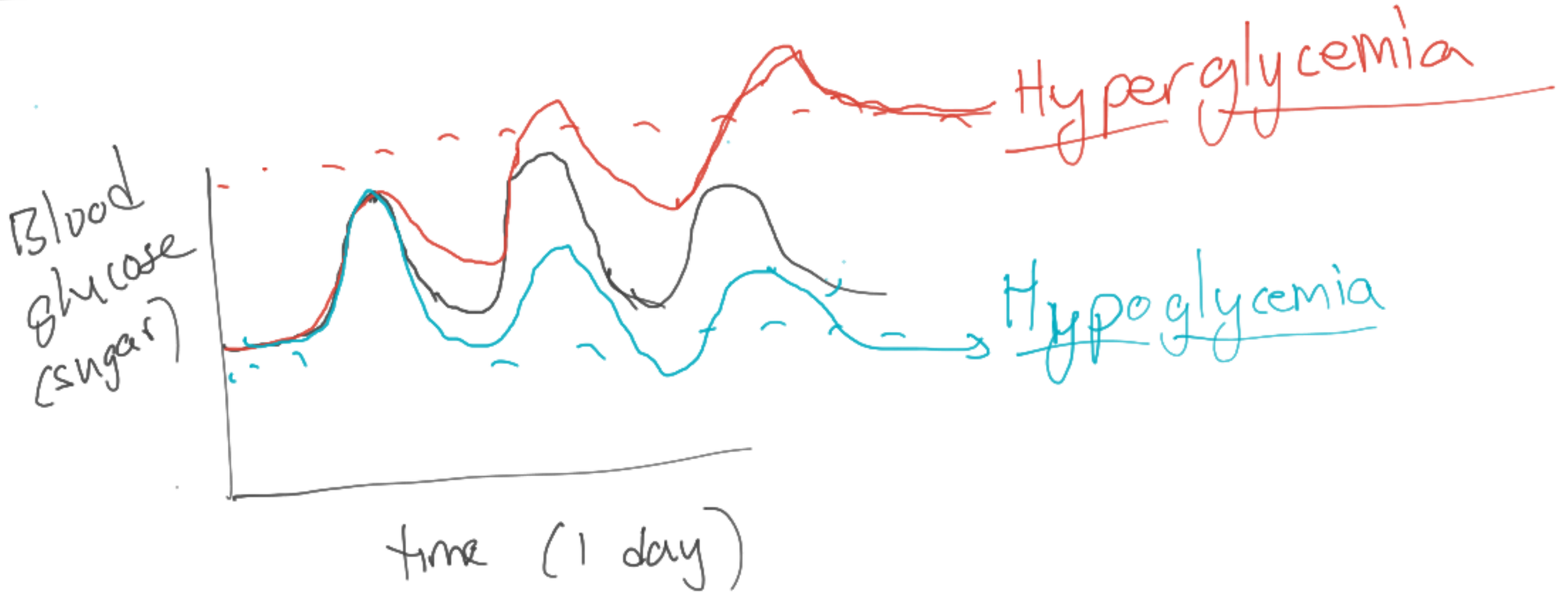
Agent - Environment Diagram





Agent: Child, dog, robot, program, ITS, diabetes treatment.

Environment: World, lab, software environment



Neuroscience: How do animals learn?
A specific agent or set of agents.

- The study of some examples of learning and intelligence.

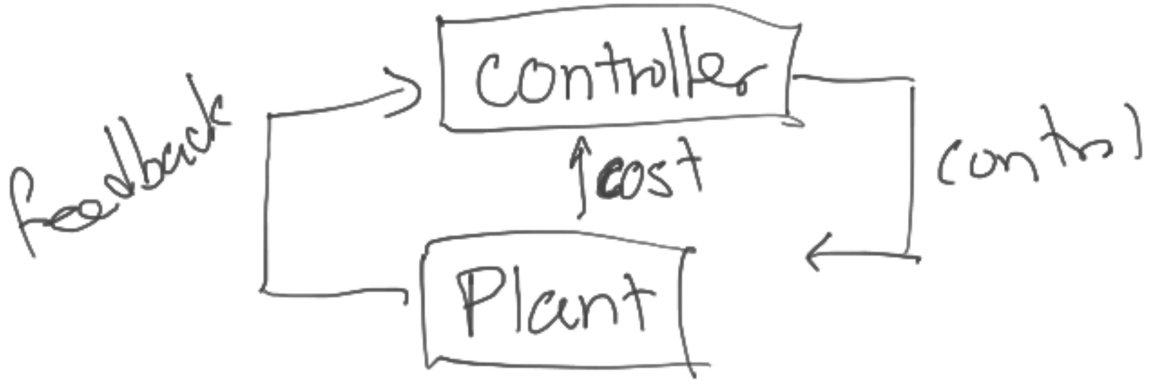
(ML)
RL : How can we make an agent that learns?

- The study of learning & intelligence. (in general)
(animal or not),

Dopamine ≈ Temporal difference error.

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.

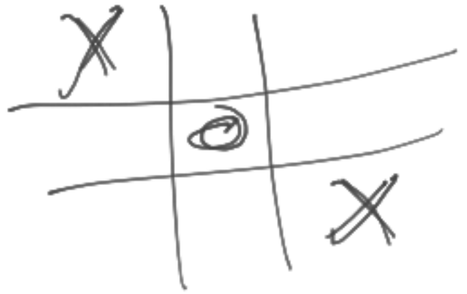
When should you use RL?

→ As a last resort.

-

Key properties:

Next time!

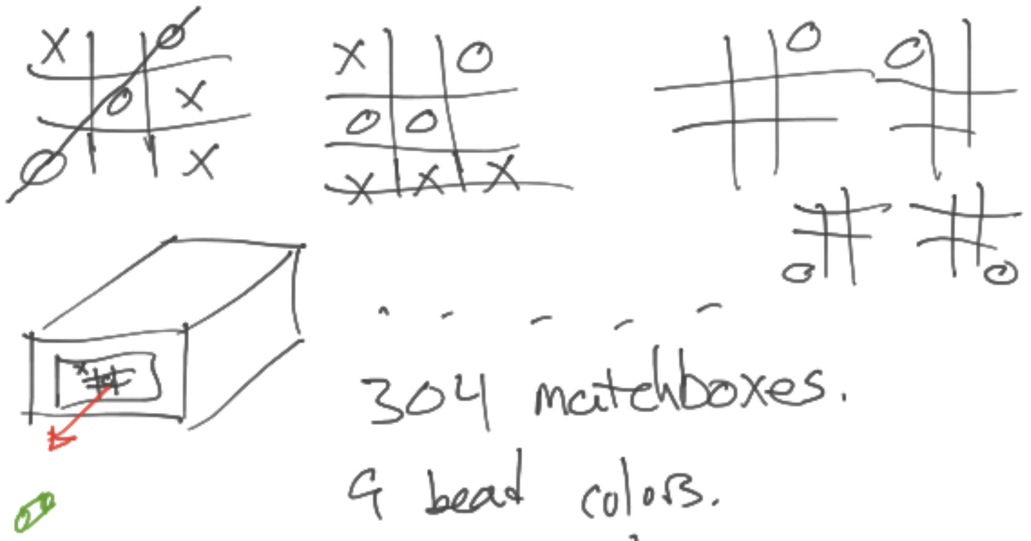


needs a colors.

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 Mitchie, 1961)



304 matchboxes.

4 bead colors.

6 possible moves.

Win: Put back bead, +3 more of same color.

Lose: Remove beads.

Tie: (Draw) Put back bead, +1 more of same color.

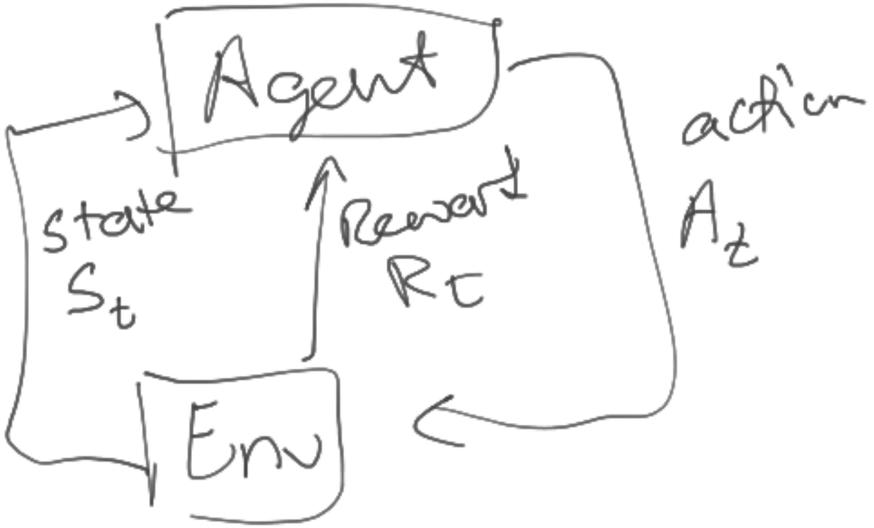
Idea:

- Act at least slightly random.
- Good outcome
  - to take chosen actions more often
- Bad outcome
  - to take chosen actions less often.

CMPSCI 687

Act randomly ???

- "Necessary" for learning!
- "exploration" vs "exploitation".
  - ↳ taking actions that think are suboptimal
  - ↳ choosing actions think are best.



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

S_t: State at time t.
(Agent's observation of state)

A_t: Action at time t.

R_t: Reward at time t.
$R_t \in \mathbb{R}$

A policy π is a way of selecting actions.
$\pi(s,a) = Pr(A_t = a \mid S_t = s)$

~~$A_t = \pi(S_t)$~~

- 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: $J(\pi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t ; \pi\right]$

Means "assuming R_t generated using π"

- Could be infinite! → $J(\pi)$
- Doesn't discount based on time

YAHP
↳ yet another hyperparameter.

$\gamma \in [0,1]$

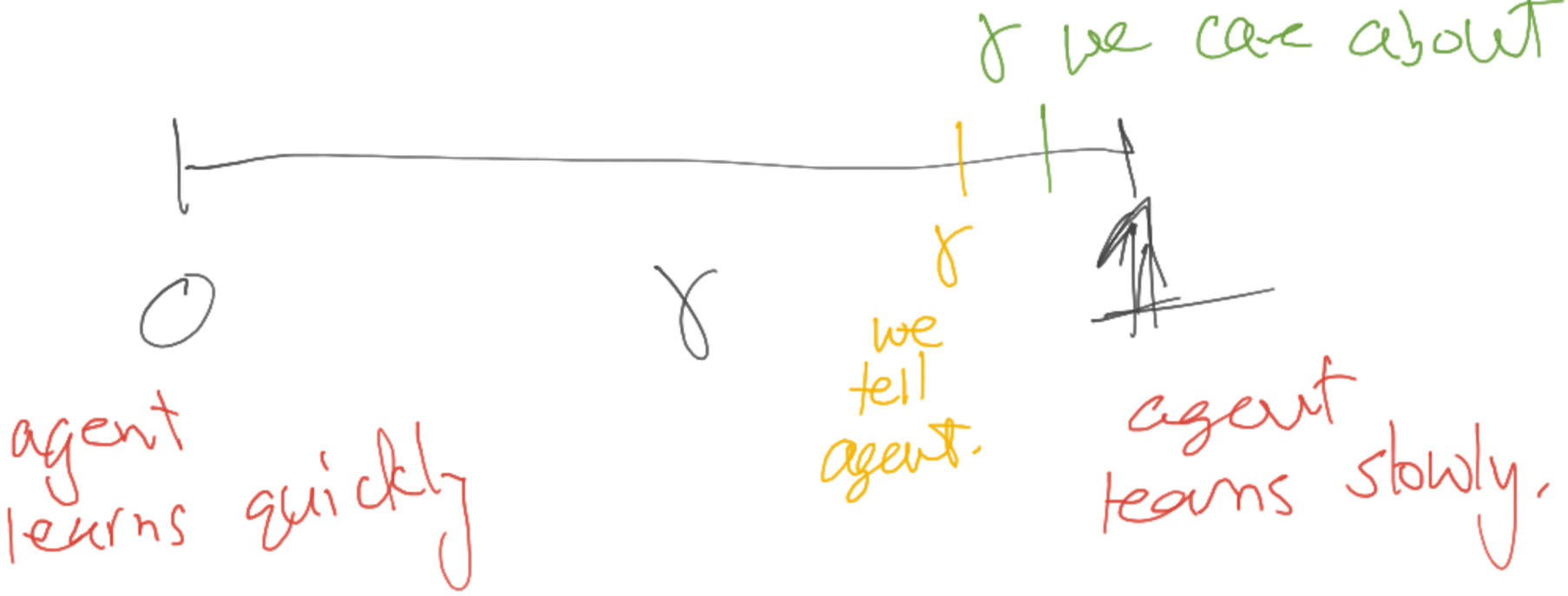
0 → care about short term

1 → care about long term.

Typically $\gamma \approx .9 \leftrightarrow 1$

$\gamma^0 R_0 + \gamma^1 R_1 + \gamma^2 R_t \cdots$

$1 R_0 + .5 R_1 + .25 R_t + \cdots$



Agent's goal is to find or approximate

$$\pi^* \in \arg\max_{\pi} J(\pi)$$

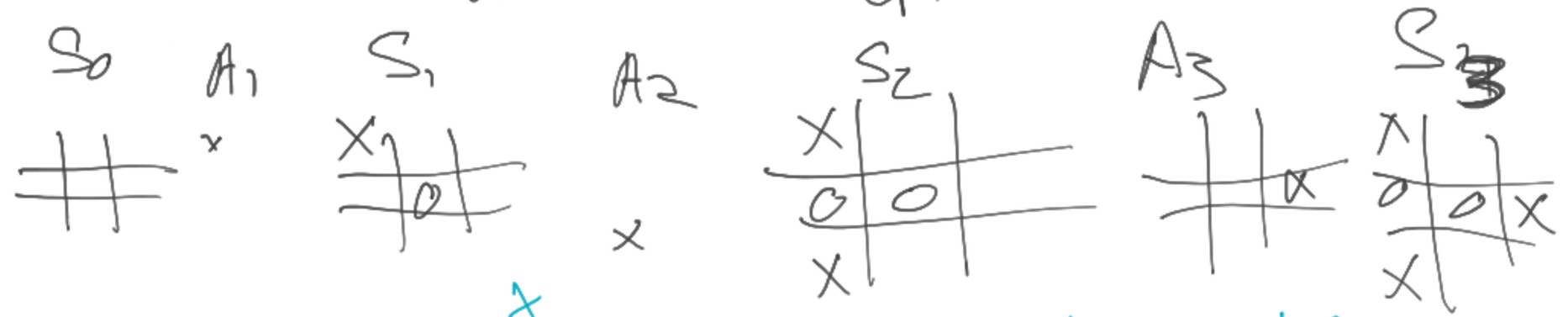
π^* is called an optimal policy.

- Midterm format.
  - In class (video on!)
  - T/F
  - Short answer
  - Math derivations.
  - Open notes.
    - closed internet.
    - individual

(No RL / No generalization) Lecture 13
-More RL!

"Independent" trials: each starting from $t=0$, are called "episodes."

one episode: $S_0, A_1, S_1, A_2, S_2, A_3, S_3$



next episode.

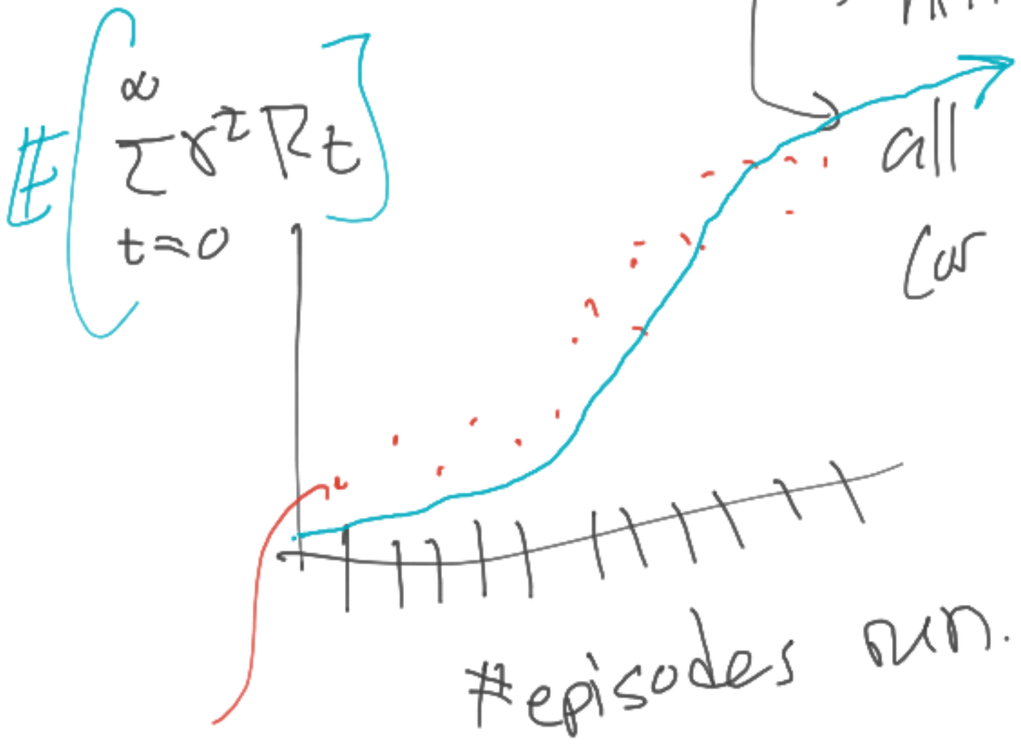
The dynamics of env do not change across trials, e.g. actions in one episode do not impact environment in later episodes.

- Way agent acts (agent's policy) can change across episodes.

$(S_t, A_t, R_t)_{t=0}^{\infty}$ independent rand variables across episodes.

Episodes end when the environment reaches a state called a "terminal state"

(↳ hill top in Mountain Car, all cells filled Car someone won) in Tic-Tac-Toe



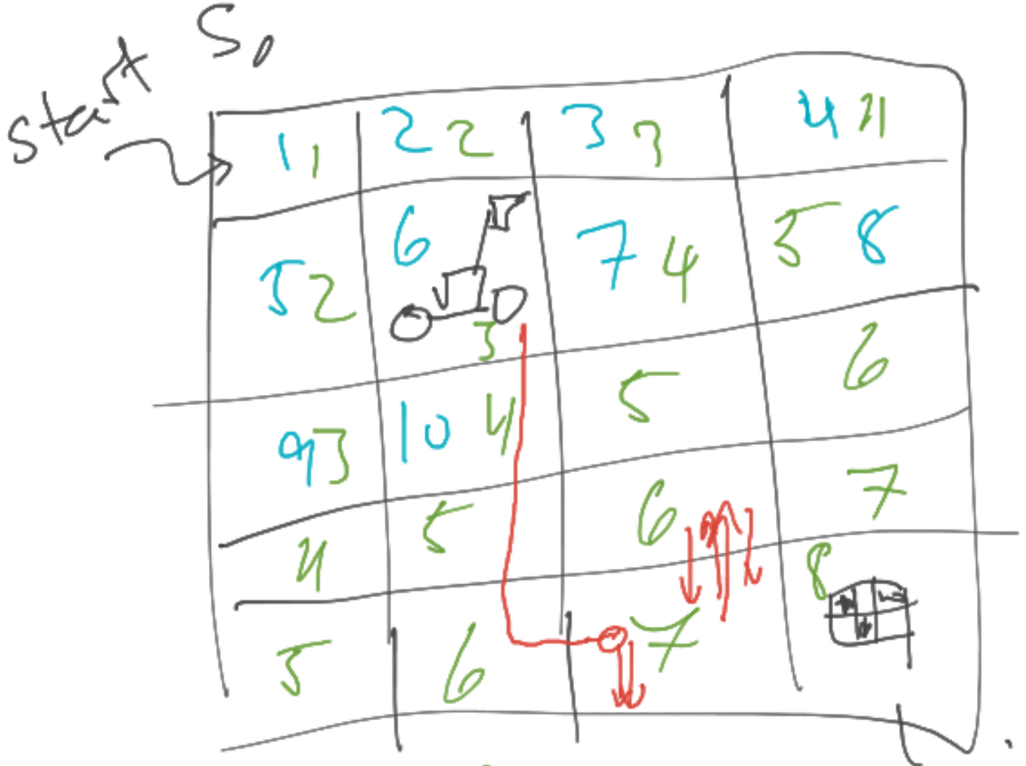
Flow:

$(S_0, A_0, R_0, S_1, A_1, R_1, \ldots)$

→ For all times after a terminal state, $R_t = 0$.

Gridworld:

$A_t \in \{\text{up}, \text{down}, \text{left}, \text{right}\}$



$S_t = (2,4)$

$S_t = 6$

$A_2 = \text{right}$

$R_t =$

$S_{t+1} = (3,4)$

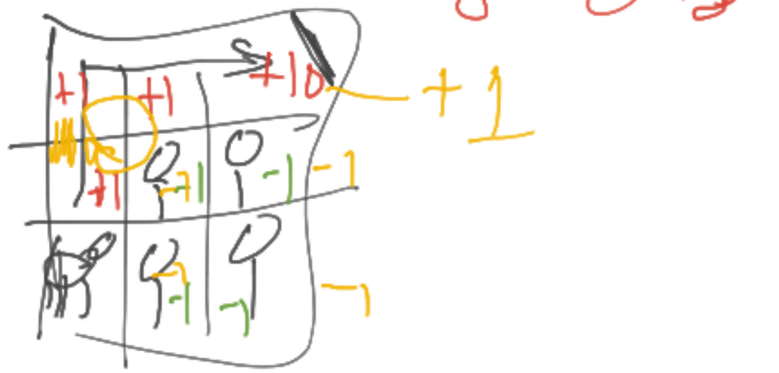
$R_t = 4$

$t = 10$
$t = 20$

R_t design flower bed

How to define R_t is yet another hyperparameter.

$\gamma = 0.5$

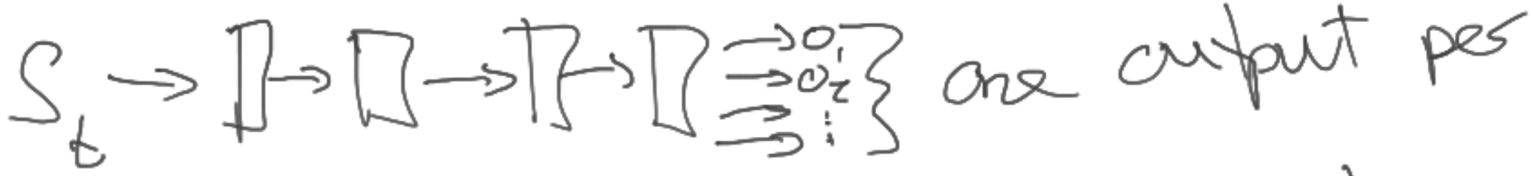


→ Give rewards for the outcomes you want, not for how you think the agent should achieve the outcome.

"reward shaping" Andrew Ng.

How represent π?

- Simplifying assumption: A_t has a small finite number of possible values.



one output per possible action.

Softmax to select the action:

$$\pi(s,a) = Pr(A_t = a \mid S_t = s) = \frac{e^{o_a}}{\sum_i e^{o_i}}$$

Supervised learning

RL

model → policy

w → θ

"model parameters" → "policy parameters"

f_w → π_θ

↳ stochastic, like classification (uses softmax)

↓

deterministic for regression.

ℓ → $-J$