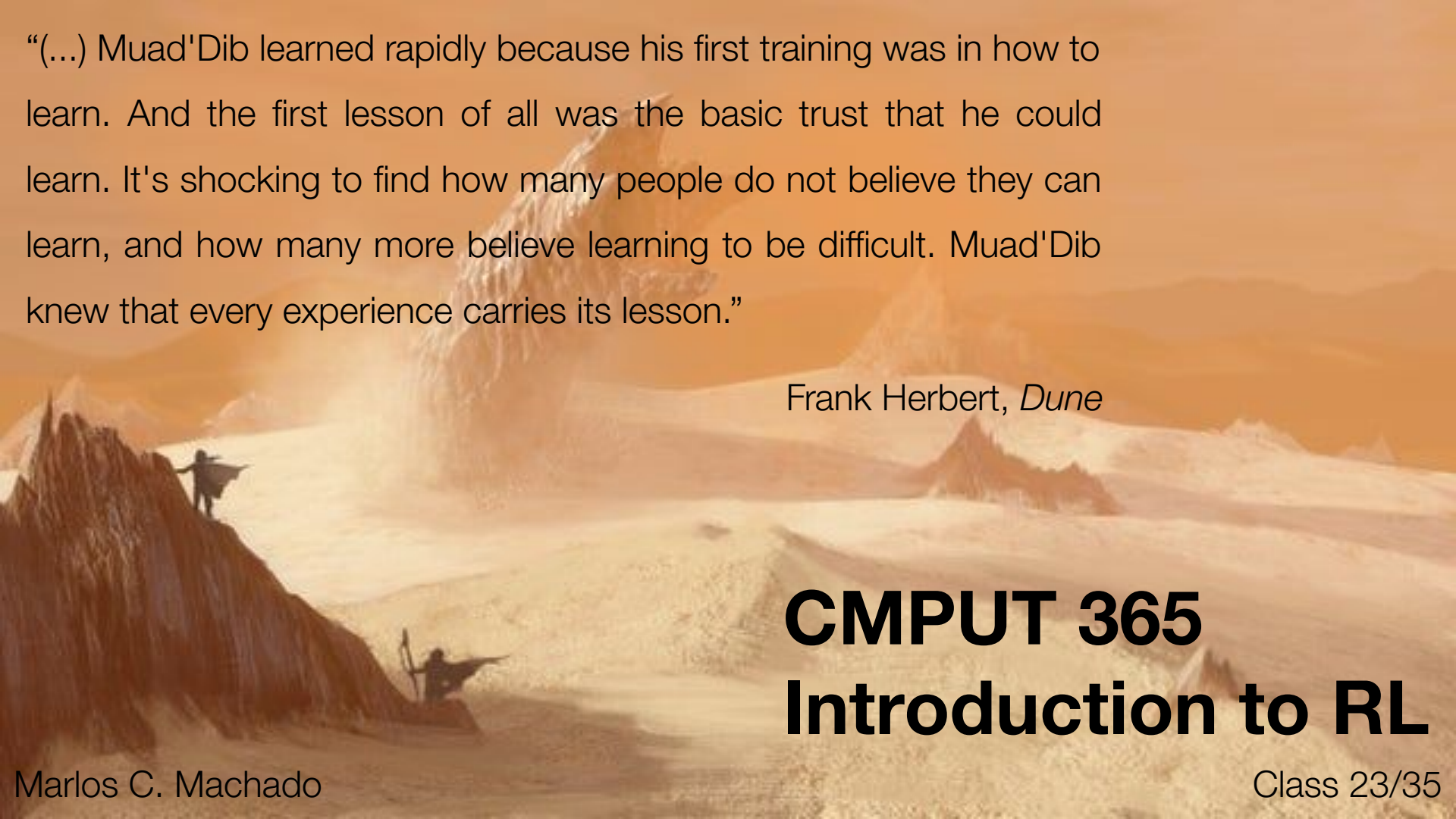

“(...) Muad'Dib learned rapidly because his first training was in how to learn. And the first lesson of all was the basic trust that he could learn. It's shocking to find how many people do not believe they can learn, and how many more believe learning to be difficult. Muad'Dib knew that every experience carries its lesson.”
Frank Herbert, *Dune*
CMPUT 365
Introduction to RL
Marlos C. Machado
Class 23/35

Reminder I

You **should be enrolled in the private session** we created in Coursera for CMPUT 365.

I **cannot** use marks from the public repository for your course marks. You **need** to **check**, **every time**, if you are in the private session and if you are submitting quizzes and assignments to the private section.

At the end of the term, I **will not port grades** from the public session in Coursera.

If you have any questions or concerns, **talk with the TAs** or email us `cmput365@ualberta.ca`.

Plan / Reminder II

- What I plan to do today:
  - Wrap up TD Learning for Control (Second half of Chapter 6 of the textbook)
  - Maybe some exercises
- Starting Friday: Sample-based learning methods: Planning, learning, & acting
  - Practice Quiz is due on Friday

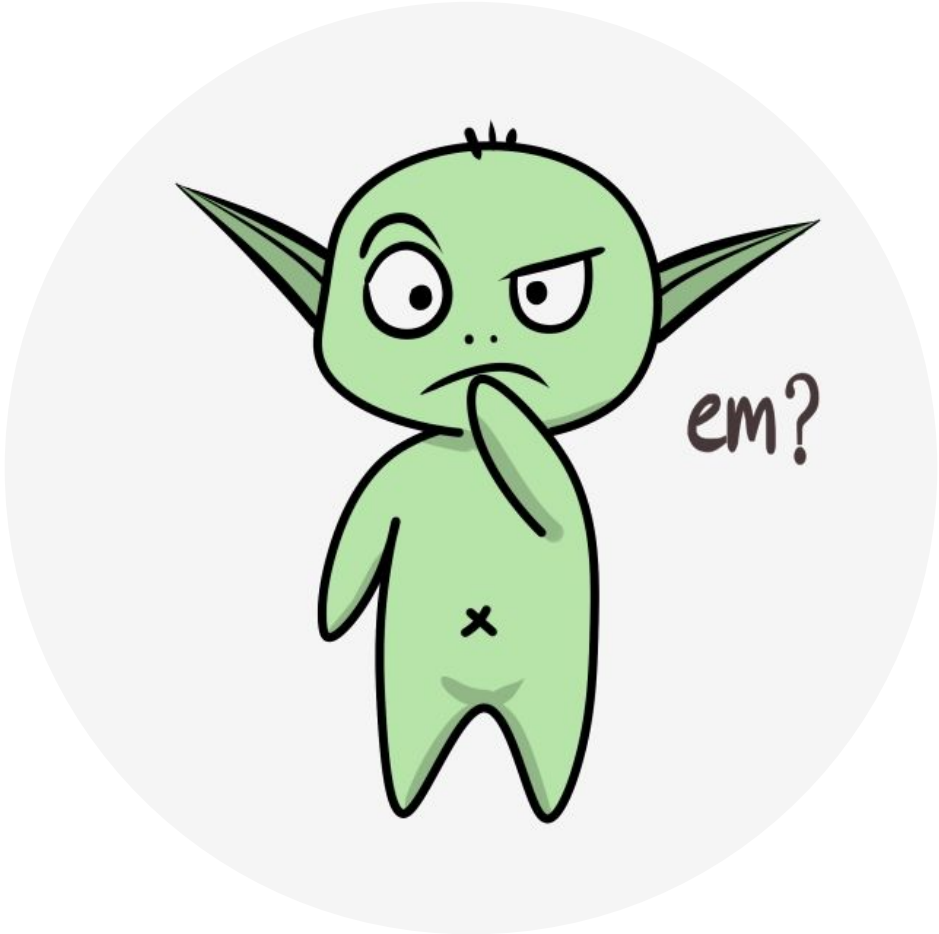
Please, interrupt me at any time!

https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

Last Class: Expected Sarsa and Double Learning

$$
\begin{aligned}
Q(S_t, A_t) &\leftarrow Q(S_t, A_t) + \alpha \Big[R_{t+1} + \gamma \mathbb{E}_\pi [Q(S_{t+1}, A_{t+1}) \mid S_{t+1}] - Q(S_t, A_t) \Big] \\
&= Q(S_t, A_t) + \alpha \Big[R_{t+1} + \gamma \sum_a \pi(a|S_{t+1}) Q(S_{t+1}, a) - Q(S_t, A_t) \Big],
\end{aligned}
$$

$$
Q_1(S_t, A_t) \leftarrow Q_1(S_t, A_t) + \alpha \Big[R_{t+1} + \gamma Q_2 \big(S_{t+1}, \arg\max_a Q_1(S_{t+1}, a) \big) - Q_1(S_t, A_t) \Big]
$$



https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

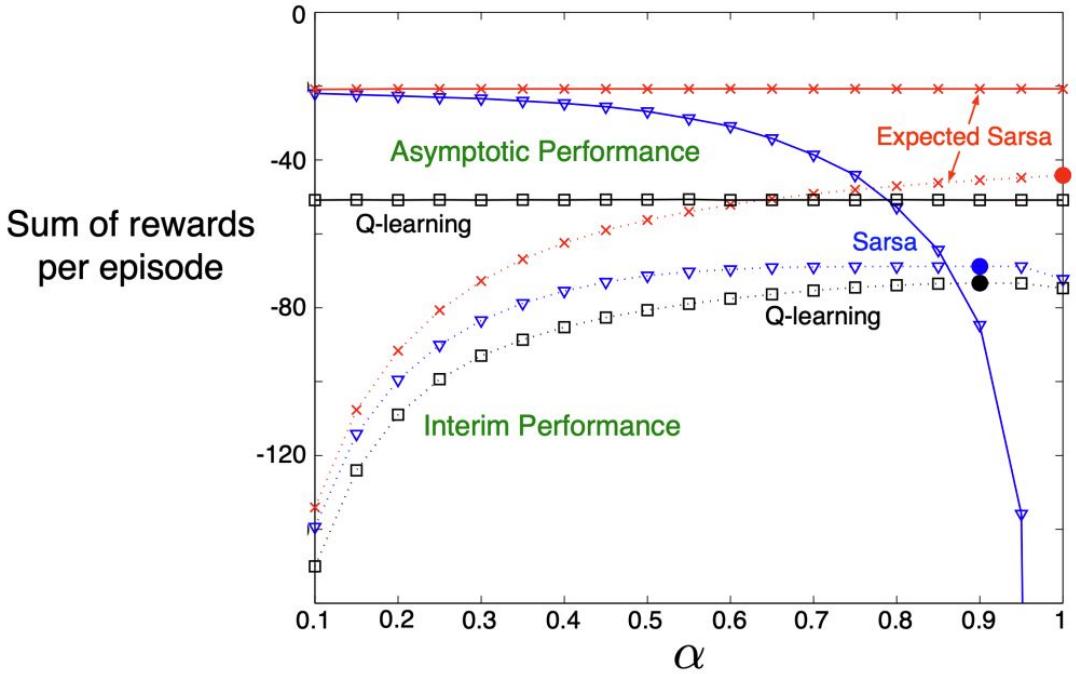
Choosing to Choose your RL Algorithm?

- What's the nature of your task, episodic or continuing?
- Do you have access to the model, p(s', r | s, a)?
- Is changing the policy during an episode important for good performance?
- Will my initial estimate of the value function, Q_0, be really terrible?
- Do I need the optimal policy or near-optimal is good enough?
- *Are we measuring performance online or offline?*

The Devil is in the Details

- Let’s say we decide to go with Q-Learning.
- How should we:
  - Decide on the target policy?
  - Decide on the behaviour policy?
  - How should we initialize Q_0?
  - Set ε? Set α? Should they change over time?
- What do we care about? The area under the curve or just the final policy?
- How long will you run it for?

The Devil is in the Details

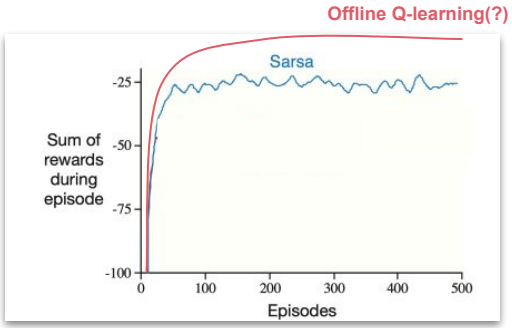


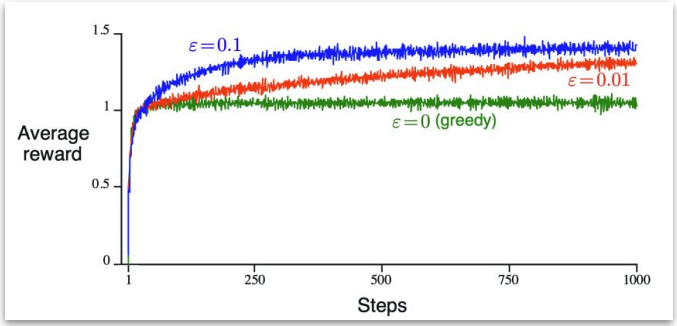
Each point on this plot represents a different choice of:

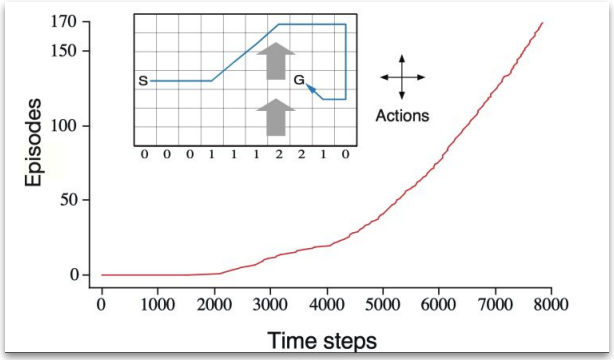
1. **Target policy.**
2. **α.**
3. **How long it was run for.**

Online vs Offline Performance

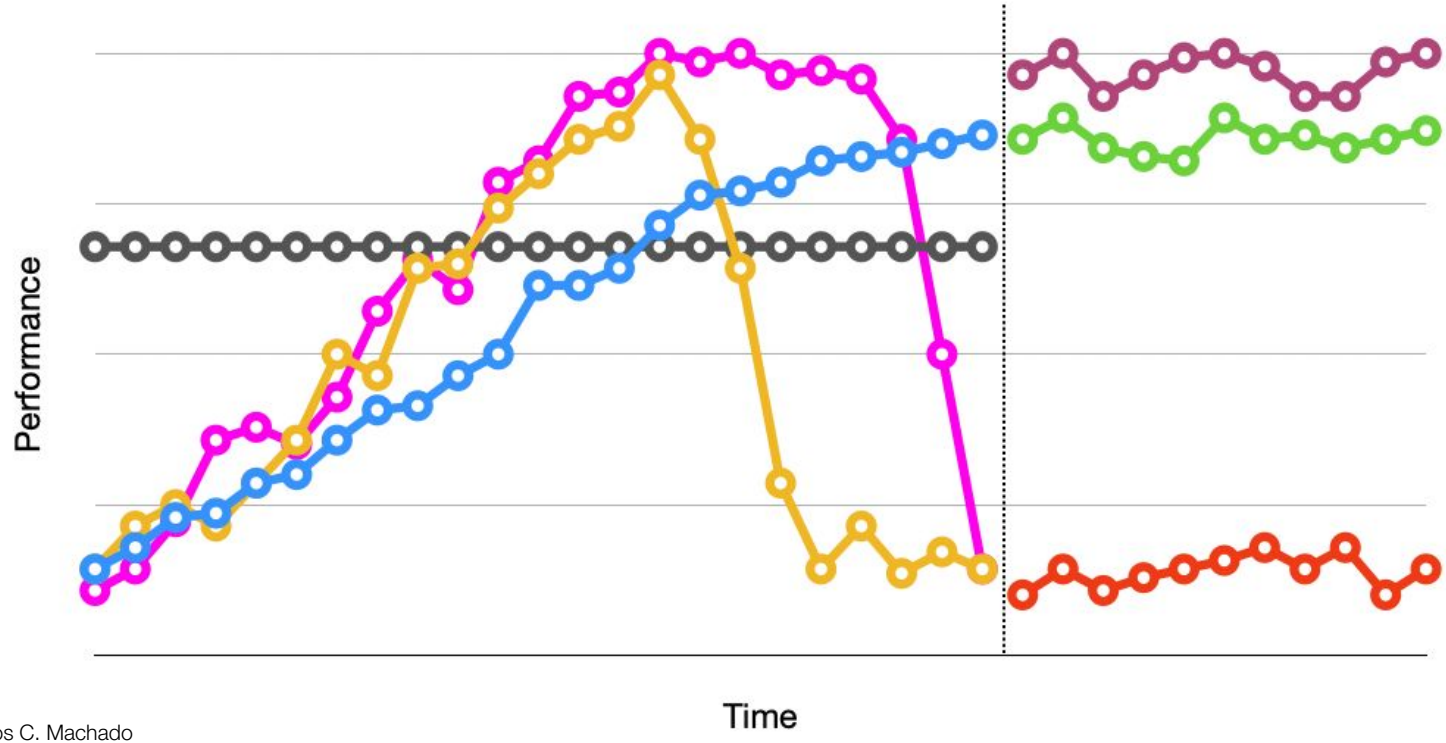
- Online is what we have been doing all along.
- In offline evaluation we have two phases:
  - **Learning** (updating the value function) and taking actions, with no performance evaluation.
  - **Testing**, when learning is disabled and we evaluate the current policy π.



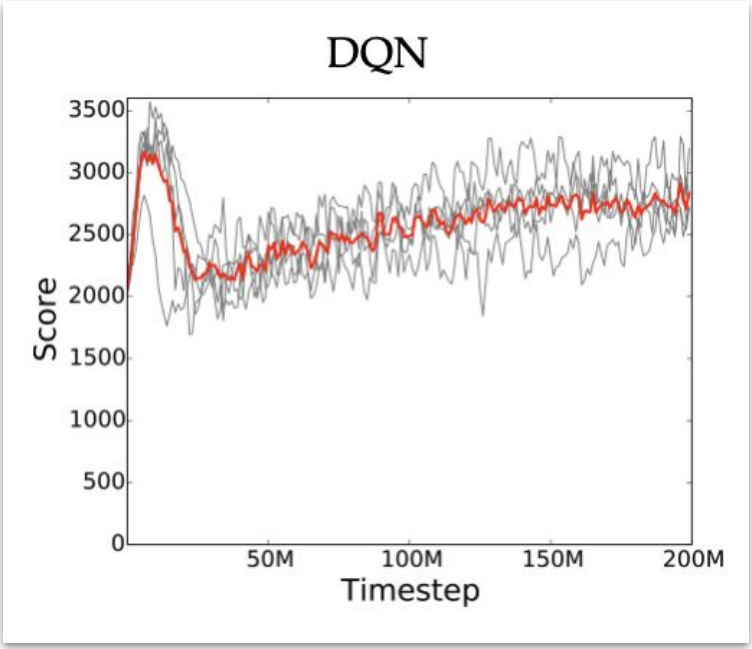




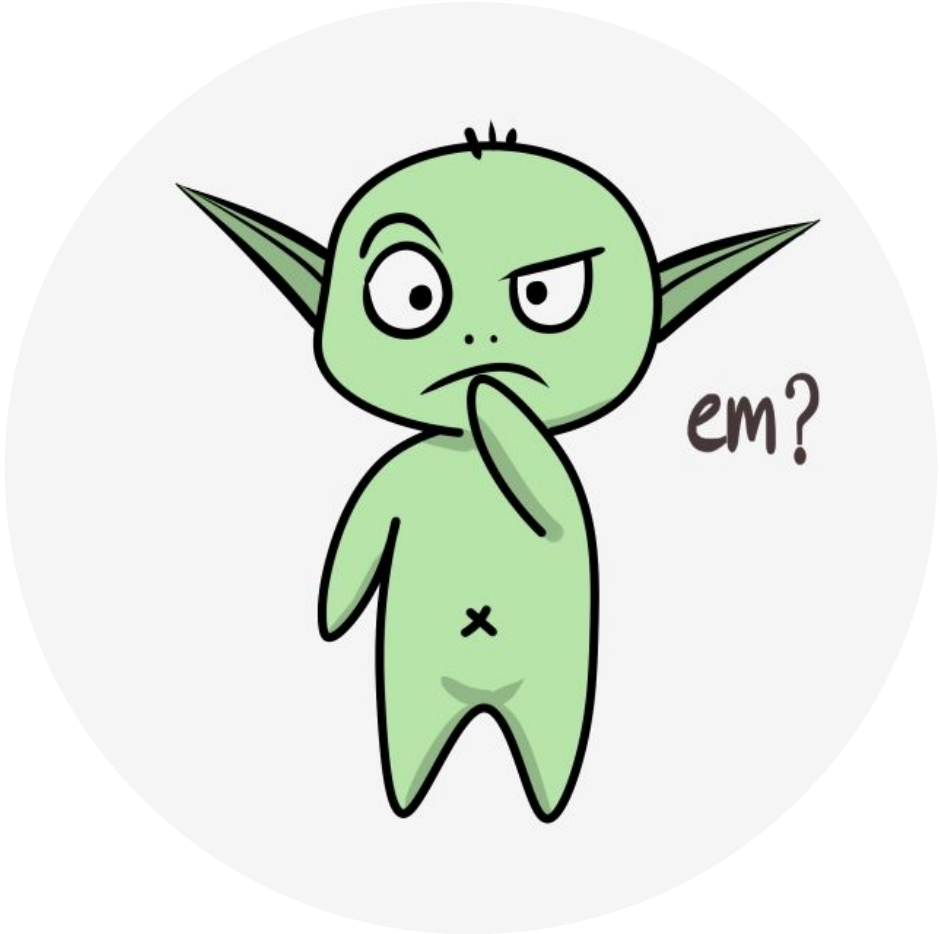
Which one do you prefer?



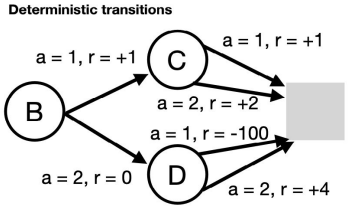
It does happen!



[Machado et al., 2018]



https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht



Exercise

Consider the following MDP, with three states B, C, and D ($\mathcal{S}$ = {B, C, D}), and 2 actions ($\mathcal{A}$ = {1, 2}), with γ = 1.0. Assume the action values are initialized Q(s, a) = 0 $\forall$ s $\in \mathcal{S}$. The agent takes actions according to an ε-greedy policy with ε = 0.1.

1. What is the optimal policy for this MDP and what are the action-values corresponding to the optimal policy: q*(s, a)?
2. Imagine the agent experienced a single episode, and the following experience: S_0 = B, A_0 = 2, R_1 = 0, S_1 = D, A_1 = 2, R_2 = 4. What are the Sarsa updates during this episode, assuming α = 0.1? Start with state B, and perform the Sarsa update, then update the value of state D.
3. Using the sample episode above, compute the updates Q-learning would make, with α = 0.1? Again, start with state B, and then state D.
4. Let's consider one more episode: S_0 = B, A_0 = 2, R_1 = 0, S_1 = D, A_1 = 1, R_2 = -100. What would the Sarsa updates be? And what would the Q-learning updates be?
5. Assume you see one more episode, and it's the same on as in 4. Once more update the action values, for Sarsa and Q-learning. What do you notice?