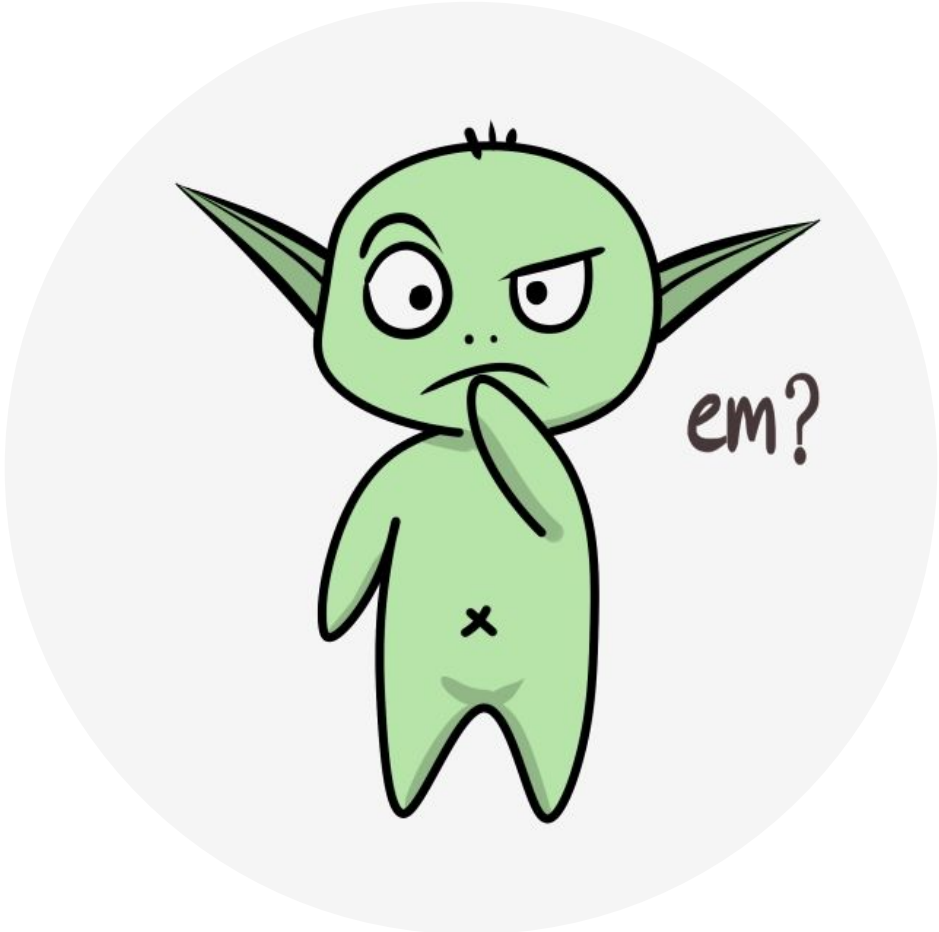

“There is nothing like looking, if you want to find something. You certainly usually find something, if you look, but it is not always quite the something you were after.”
J. R. R. Tolkien, *The Hobbit*
CMPUT 655
RL 1
Marlos C. Machado
Class 9¾/12

Reminders

- **Midterm is today at 15:30. We'll stop at 15:00 so you have a longer break.**
- Marks for the project proposal and Coursera activities are out.
- It is Reading Week next week. We won't have class nor office hours.
- Last set of activities on Coursera are due November 24 at noon.
  - There's a typo in the syllabus.
- We'll have Rich Sutton as guest lecturer in December.



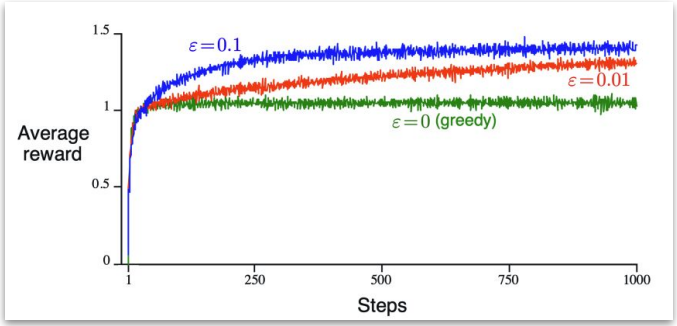
https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

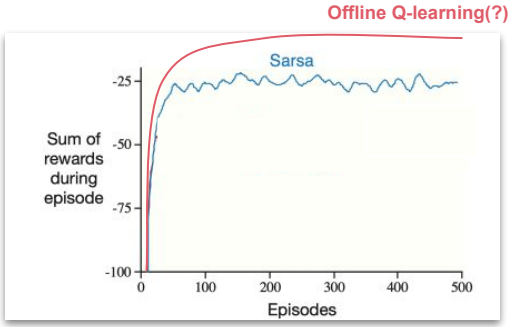
How should one choose an RL Algorithm?

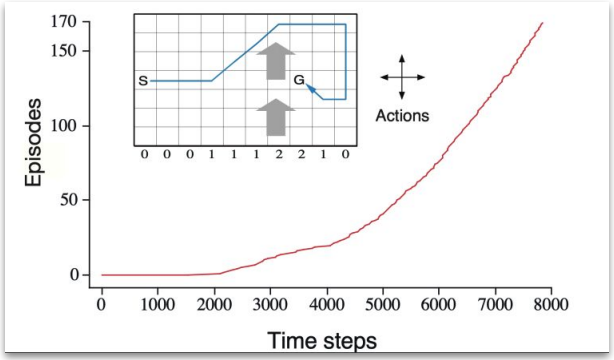
- What's the nature of your task, episodic or continuing?
- How many students and actions does the problem have? Are they discrete?
- Do you have access to the model, p(s', r | s, a)?
- Is changing the policy during an episode important for good performance?
- Will the initial estimate of the value function, Q_0, be really terrible?
- Is the optimal policy needed or is a near-optimal policy good enough?
- How much compute do we have access to?
- What do the rewards look like?
- *Are we measuring performance online or offline?*

Online vs Offline Performance

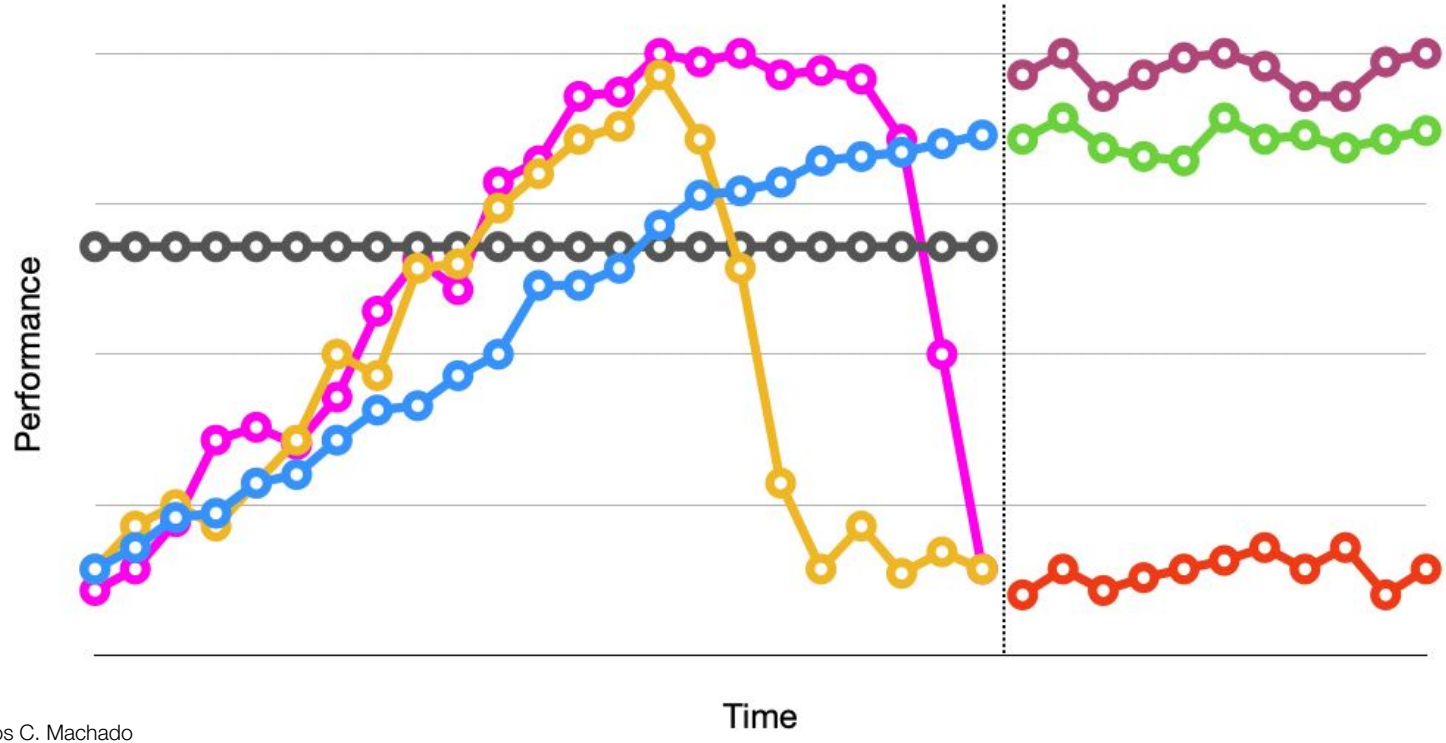
- Online is what we have been doing all along.
- In offline evaluation we have two phases:
  - **Learning** (updating the value function) and taking actions, with no performance evaluation.
  - **Testing**, when learning is disabled and we evaluate the current policy π.



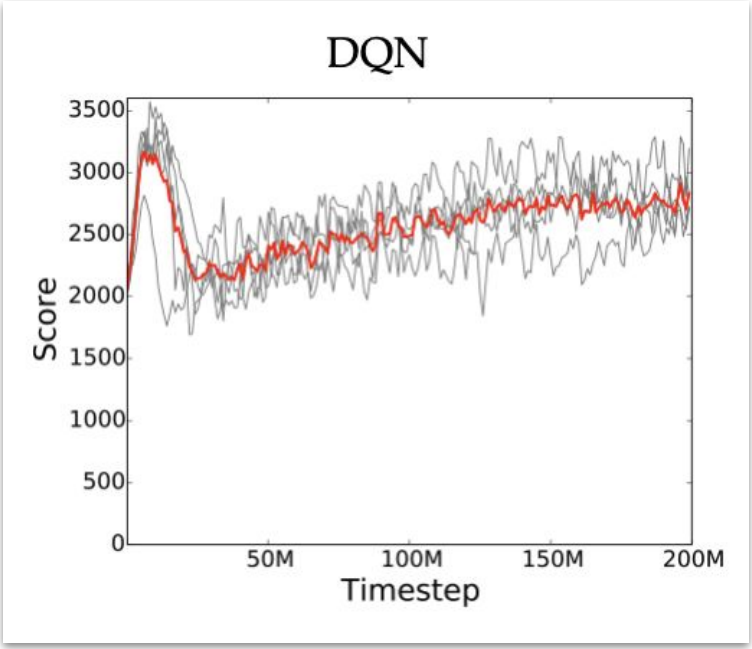




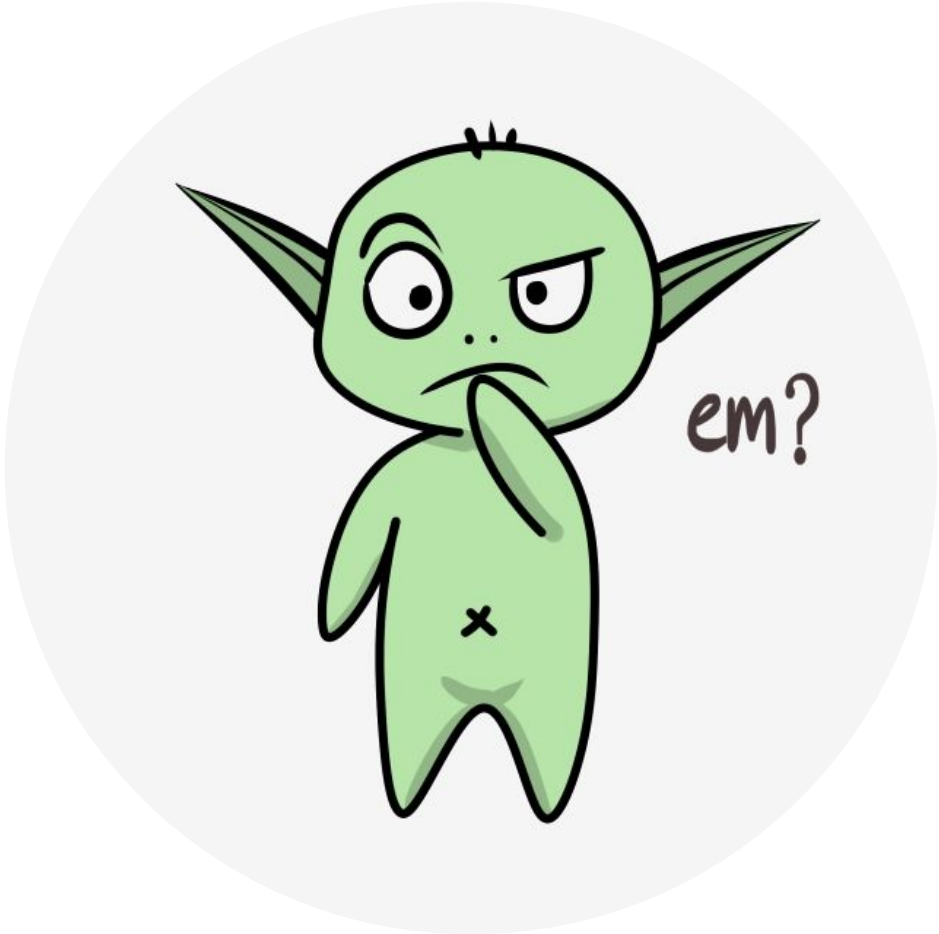
Which one do you prefer?



It does happen!



[Machado et al., 2018]



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

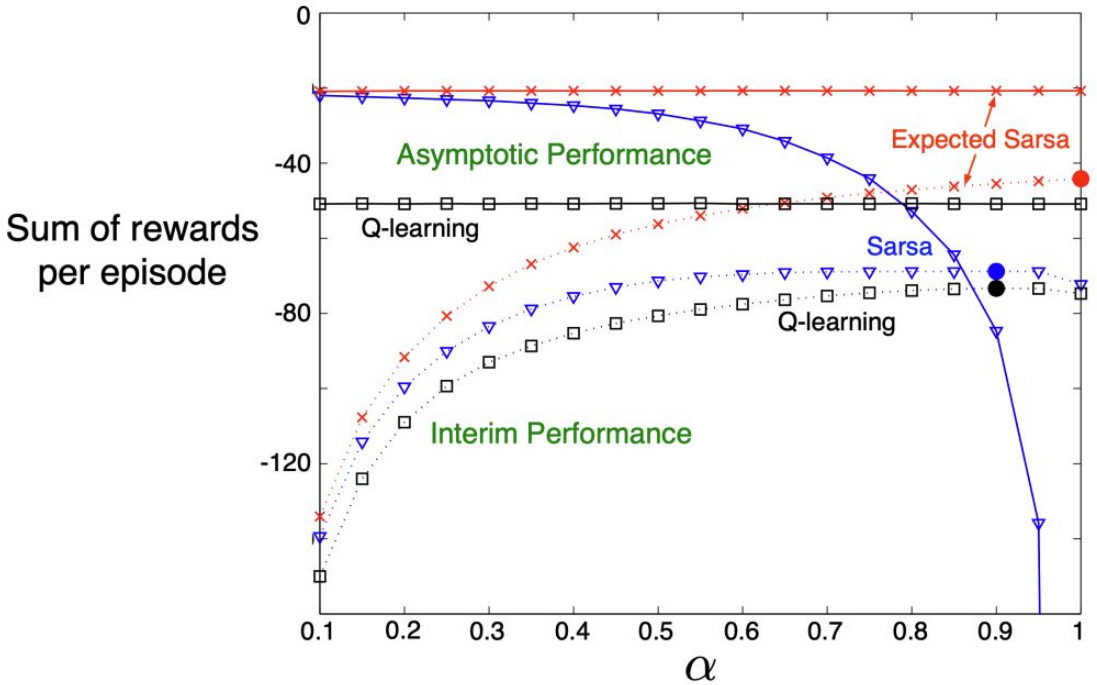
How should one choose an RL Algorithm?

- What's the nature of your task, episodic or continuing?
- How many students and actions does the problem have? Are they discrete?
- Do you have access to the model, p(s', r | s, a)?
- Is changing the policy during an episode important for good performance?
- Will the initial estimate of the value function, Q_0, be really terrible?
- Is the optimal policy needed or is a near-optimal policy good enough?
- How much compute do we have access to?
- What do the rewards look like?
- *Are we measuring performance online or offline?*

The Devil is in the Details

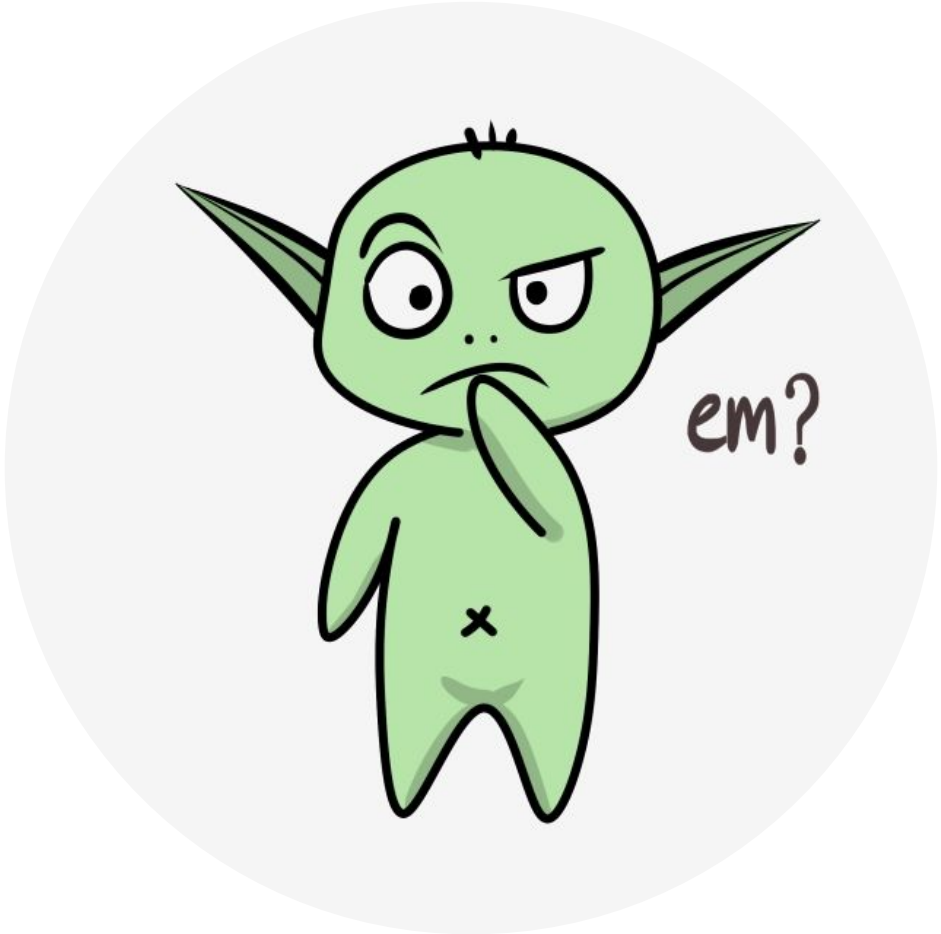
- Let’s say we decide to go with Q-Learning.
- How should we:
  - Decide on the target policy?
  - Decide on the behaviour policy?
  - How should we initialize Q_0?
  - Set ε? Set α? Should they change over time?
  - Do we need function approximation? Which one should we use?
- What do we care about? 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. **Algorithm (thus target policy).**
2. **Step-size, α.**
3. **How long it was run for.**



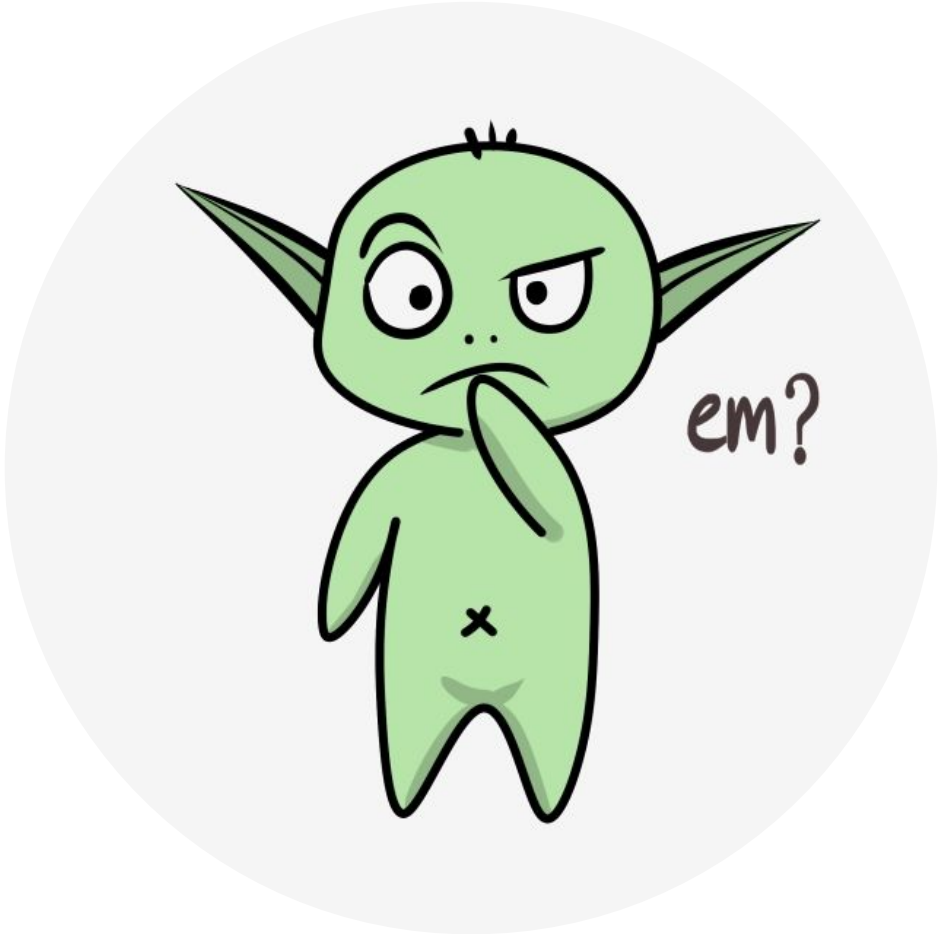
https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

Some additional advice for experimental research

- Use more seeds in our experiments
- Make our source code (and dataset) available
- Be fair to the baselines we are using
- Be thorough on reporting what we actually did
- Use proper statistics
- Don’t report results selectively
- Don’t overclaim

So much has been written about this already

- Protecting against evaluation overfitting in empirical reinforcement learning
  [Whiteson, Tanner, Taylor, & Stone; ADPRL 2011]
- Revisiting the ALE: Evaluation protocols and open problems for general agents
  [Machado, Bellemare, Talvitie, Veness, Hausknecht, & Bowling; JAIR 2018]
- Deep reinforcement learning that matters
  [Henderson, Islam, Bachman, Pineau, Precup, & Meger; AAAI 2018]
- On bonus based exploration methods in the Arcade Learning Environment
  [Taïga, Fedus, Machado, Courville, & Bellemare; ICLR 2020]
- Deep reinforcement learning at the edge of the statistical precipice
  [Agarwal, Schwarzer, Castro, Courville, & Bellemare; NeurIPS 2021]
- Empirical design in reinforcement learning
  [Patterson, Neumann, White, & White; Under review 2023]



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