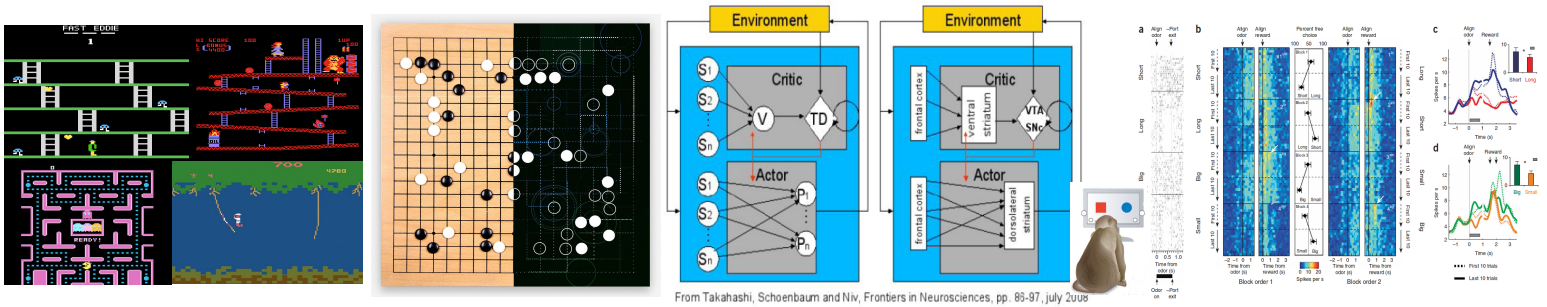
Tabular Reinforcement Learning

1. Introduction

Olivier Sigaud

Sorbonne Université
http://people.isir.upmc.fr/sigaud

Why this class?



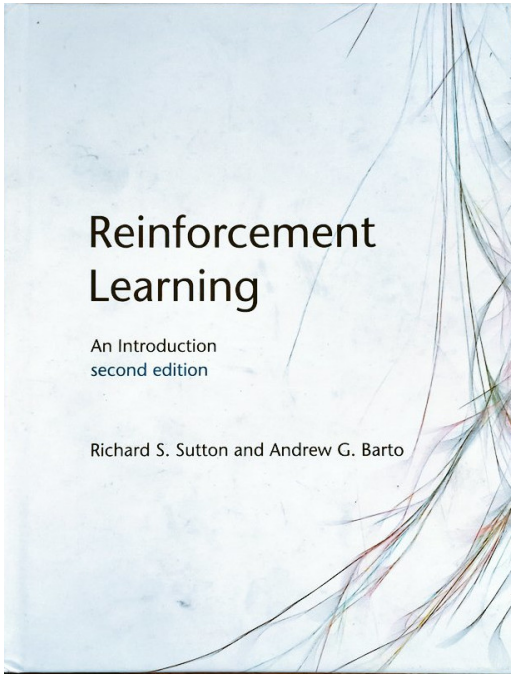
- A lot of buzz about deep reinforcement learning as an engineering tool
- The reinforcement learning framework is also relevant in computational neuroscience and psychology but this aspect will be left out

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015) Human-level control through deep reinforcement learning. *Nature*, 518(7540), 529–533.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017) Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359

Roesch, M. R., Calu, D. J., & Schoenbaum, G. (2007) Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nature Neuroscience*, 10(12):1615–1624

Introductory book: the bible

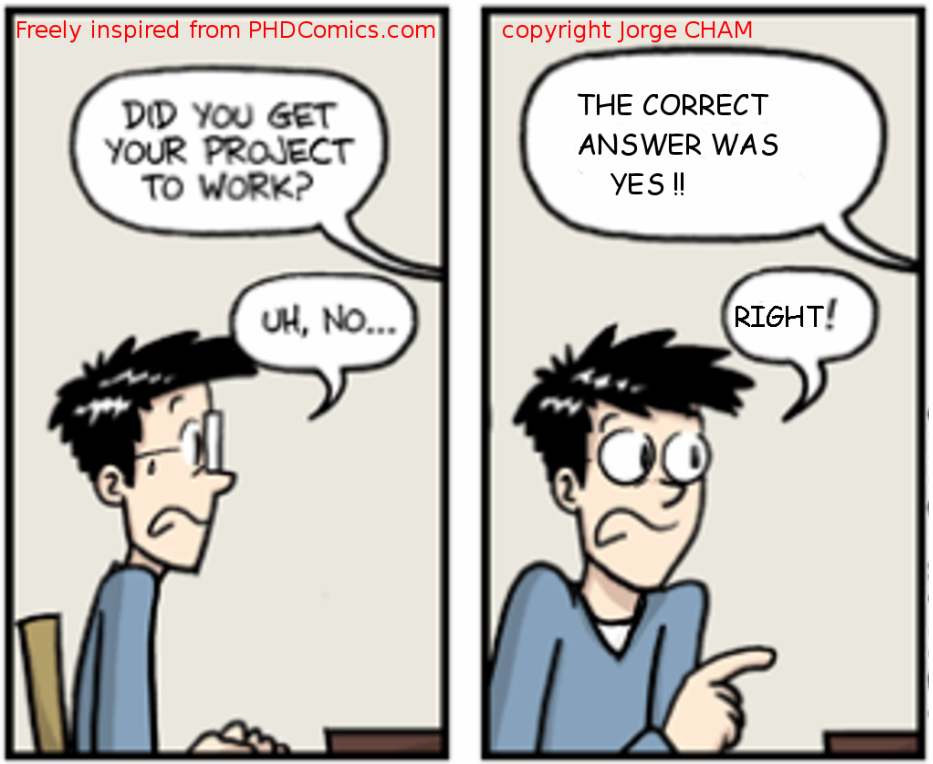


- [Sutton and Barto, 2018]: the ultimate introduction to the field, in the discrete case
- Available online:

  `https://drive.google.com/file/d/1xeUDVGWGUUv1-ccUMAZHJLej2C7aAFWY/view`
- First version in 1998

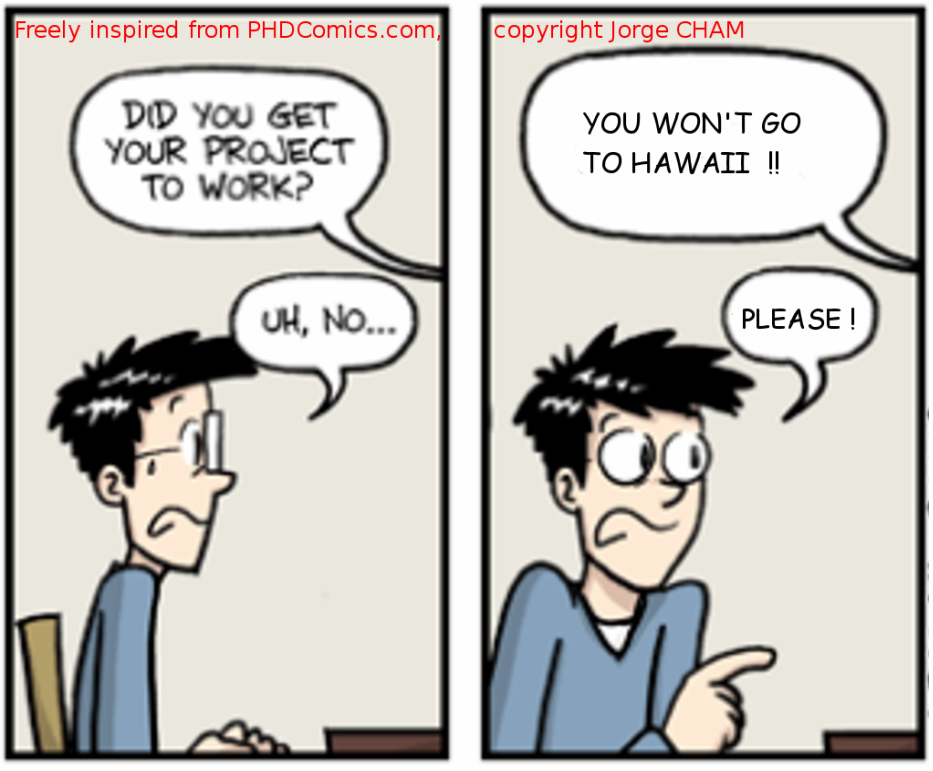
Sutton, R. S. & Barto, A. G. (2018) *Reinforcement Learning: An Introduction*. MIT Press.

Supervised learning



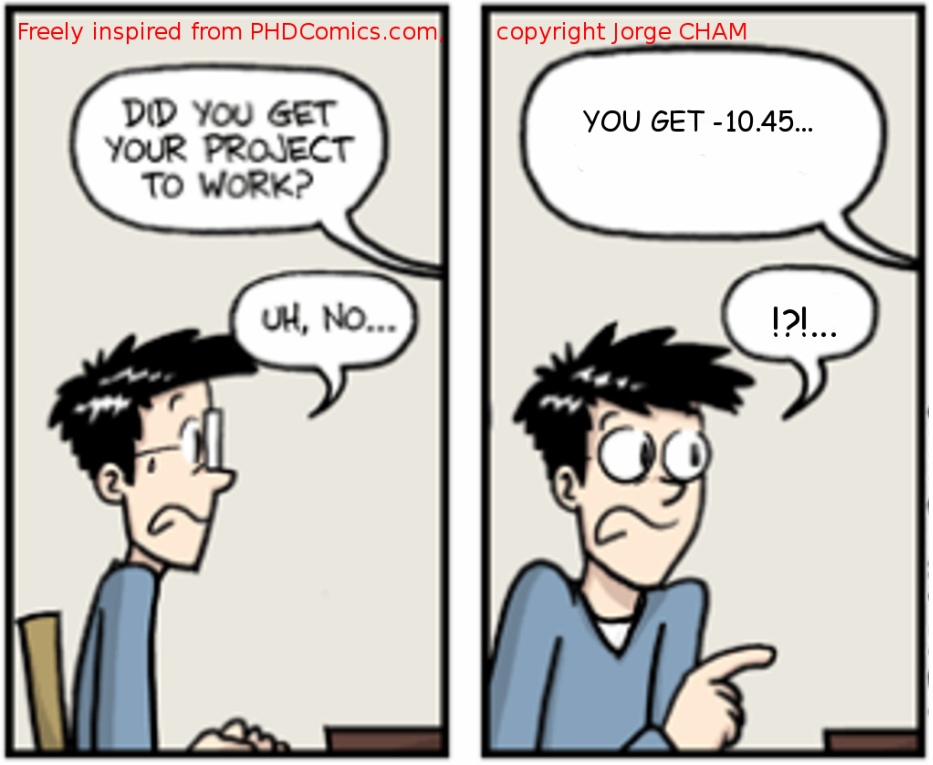
- The supervisor indicates to the agent the expected answer
- The agent corrects a model based on the answer
- Typical mechanism: gradient backpropagation, RLS
- Applications: classification, regression, function approximation...

Cost-Sensitive Learning



- The environment provides the value of action (reward, penalty)
- Application: behaviour optimization

Reinforcement learning



- In RL, the value signal is given as a scalar
- How good is -10.45?
- Necessity of exploration

The exploration/exploitation trade-off

- Exploring can be (very) harmful
- Shall I exploit what I know or look for a better policy?
- Am I optimal? Shall I keep exploring or stop?
- Decrease the rate of exploration along time
- ϵ-*greedy*: take the best action most of the time, and a random action from time to time

Sequentiality: Bandits problems vs Sequential problems

s_0 0.48	s_3 0.53	0.59	0.66	0.73
s_1 0.43	s_2 0.48	0.53		0.81
0.39		0.48		0.9
0.35	0.39	0.43		1

- A multi-armed bandit problem is a one step/state RL problem where each arm is an action (and reward is stochastic for more fun)
- RL problem are sequential decision making problems
- The state at step $t+1$ depends on the action at state t
- This makes exploration/optimization much more difficult

Any question?

Send mail to: Olivier.Sigaud@isir.upmc.fr

ISIR
INSTITUT
DES SYSTÈMES
INTELLIGENTS
ET DE ROBOTIQUE

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015).
Human-level control through deep reinforcement learning.
Nature, 518(7540):529–533.

Roesch, M. R., Calu, D. J., and Schoenbaum, G. (2007).
Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards.
Nature Neuroscience, 10(12):1615–1624.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. (2017).
Mastering chess and shogi by self-play with a general reinforcement learning algorithm.
arXiv preprint arXiv:1712.01815.

Sutton, R. S. and Barto, A. G. (2018).
Reinforcement Learning: An Introduction (Second edition).
MIT Press.