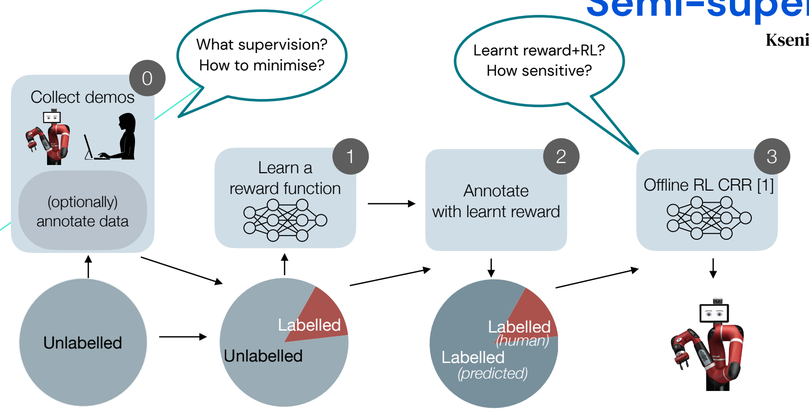
Semi-supervised reward learning for offline reinforcement learning

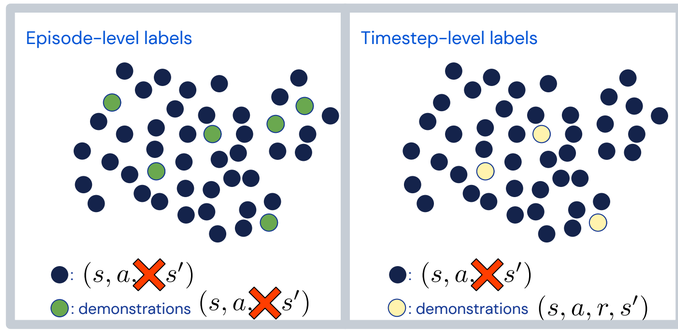
Ksenia Konyushova, Konrad Zolna, Yusuf Aytar, Alexander Novikov, Scott Reed, Serkan Cabi, Nando de Freitas

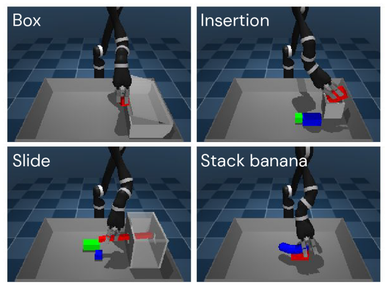
DeepMind



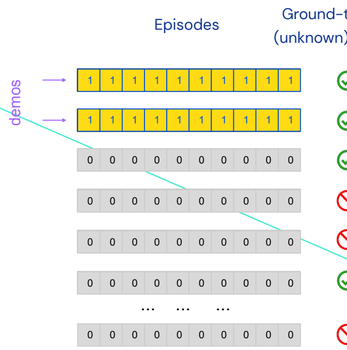
	Problem	Solution
Data	hard to collect • hard to explore • data-hungry	use offline RL
Reward	no reward in practice • hard to engineer • impossible for some tasks	learn reward function

Needs human annotations, try to minimise





Episode-level labels

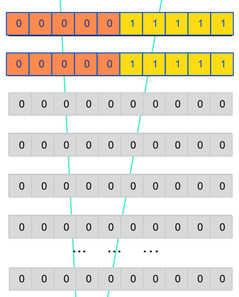


No learning: SQIL [2]

- Assume flat rewards for all episodes
- Set 1 for all timesteps from demos, and 0 otherwise

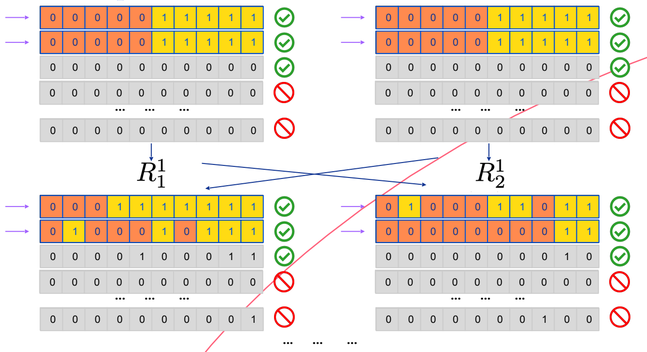
Train a classifier: ORIL [3]

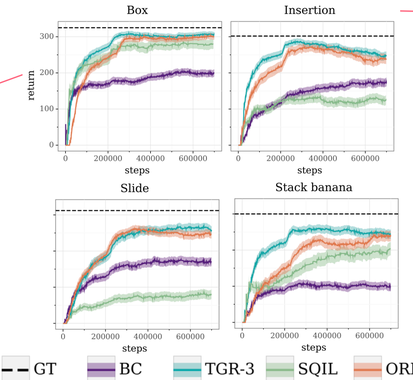
- Train a classifier using these labels
- Additionally, use regularisations:
  - TRAIL,
  - PU-learning,
  - or both.



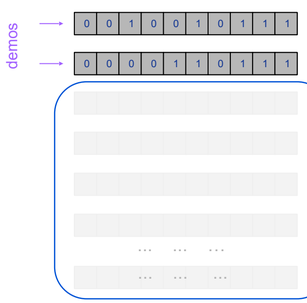
Time-guided reward (TGR)

- The first half of the demo has reward 0, and the second half has reward 1
- Unlabelled data has reward 0
- Train a classifier!



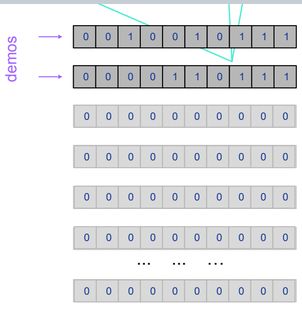


Timestep-level labels



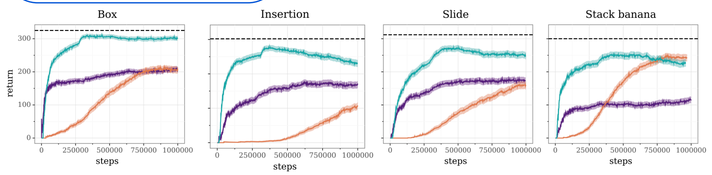
Sup-demo

- Suppose we have only very limited amount of demonstrations
- To compensate for their scarcity, we can decide to annotate them with timestep rewards
- Ignore the rest of the episodes.
- Train a classifier



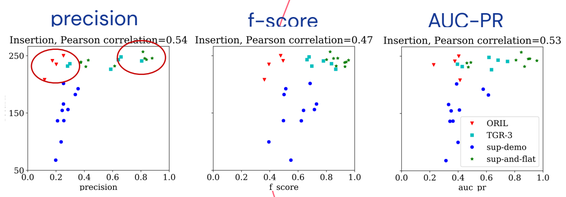
Sup-and-flat

- Take unlabelled data into account
- Assign flat reward
- Train a classifier including these labels

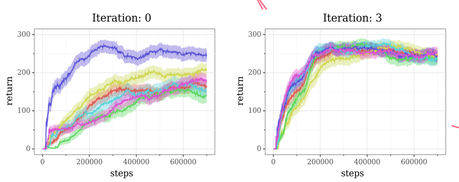


Analysis

How to choose the best reward model?



TGR improvement reduces the variance



Other RL algorithms?

[1] Critic Regularized Regression. Z. Wang, A Novikov, K Zolna, J.T. Springenberg, S. Reed, B. Shahriari, N. Siegel, J. Merel, C. Gulcehre, N. Heess, N. de Freitas, NeurIPS 2020

[2] SQIL: imitation learning via reinforcement learning with sparse rewards. S. Reddy, A. D. Dragan, and S. Levine. In ICLR, 2020.

[3] Offline Learning from Demonstrations and Unlabeled Experience. K. Zolna, A. Novikov, K. Konyushkova, C. Gulcehre, Z. Wang, Y. Aytar, M. Denil, N. de Freitas, S. Reed.