Human Influence for Reinforcement Learning

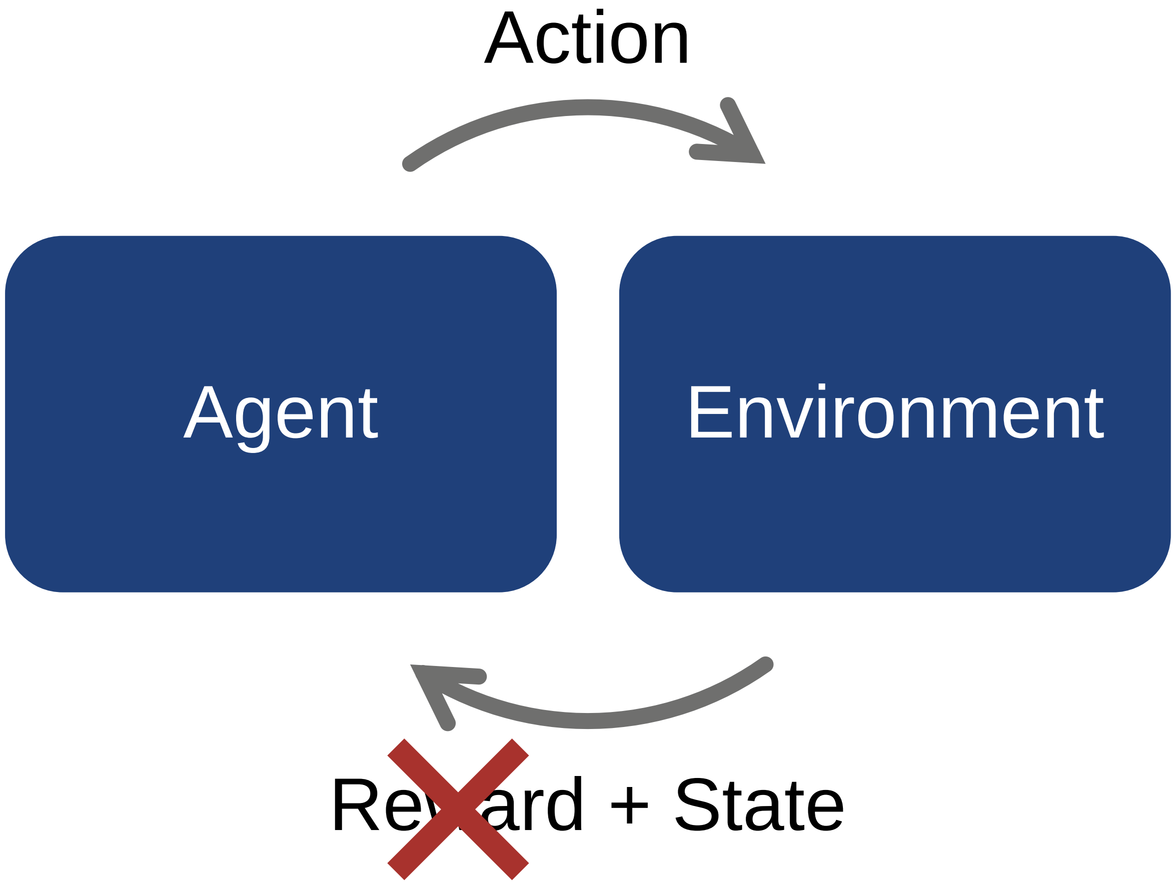
Deep Q-learning from Demonstrations

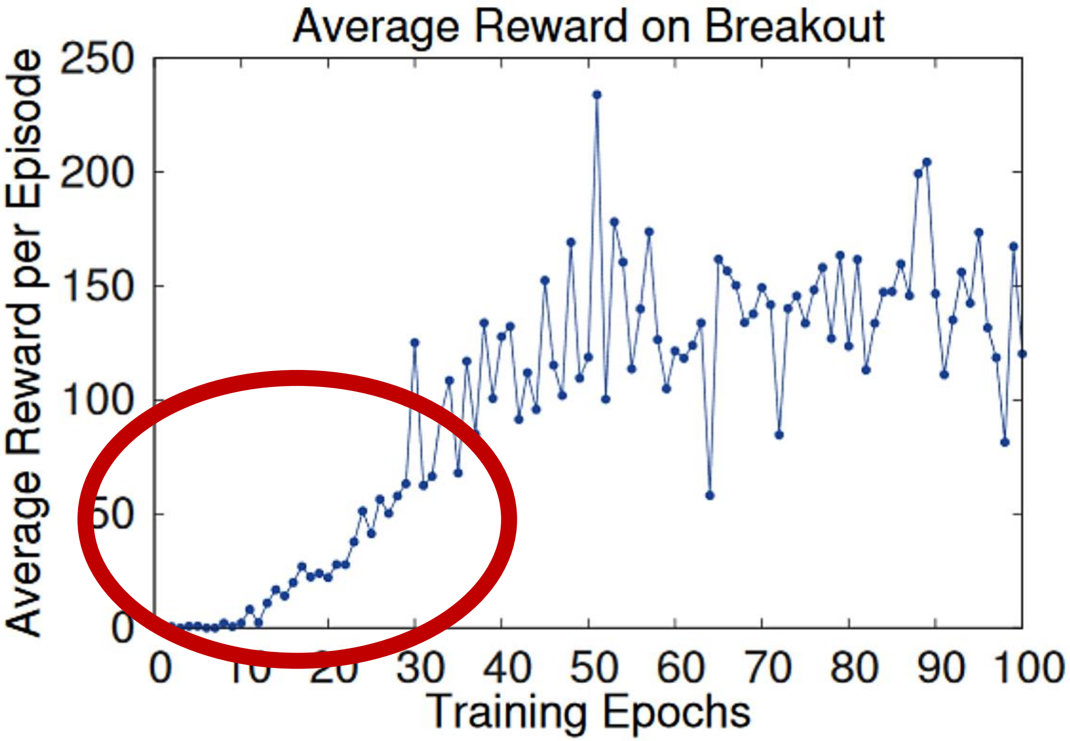
T. Hester et al.

Deep Reinforcement Learning from Human Preferences

P. Christiano et al.

Conventional Reinforcement Learning





[1] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing atari with deep reinforcement learning.

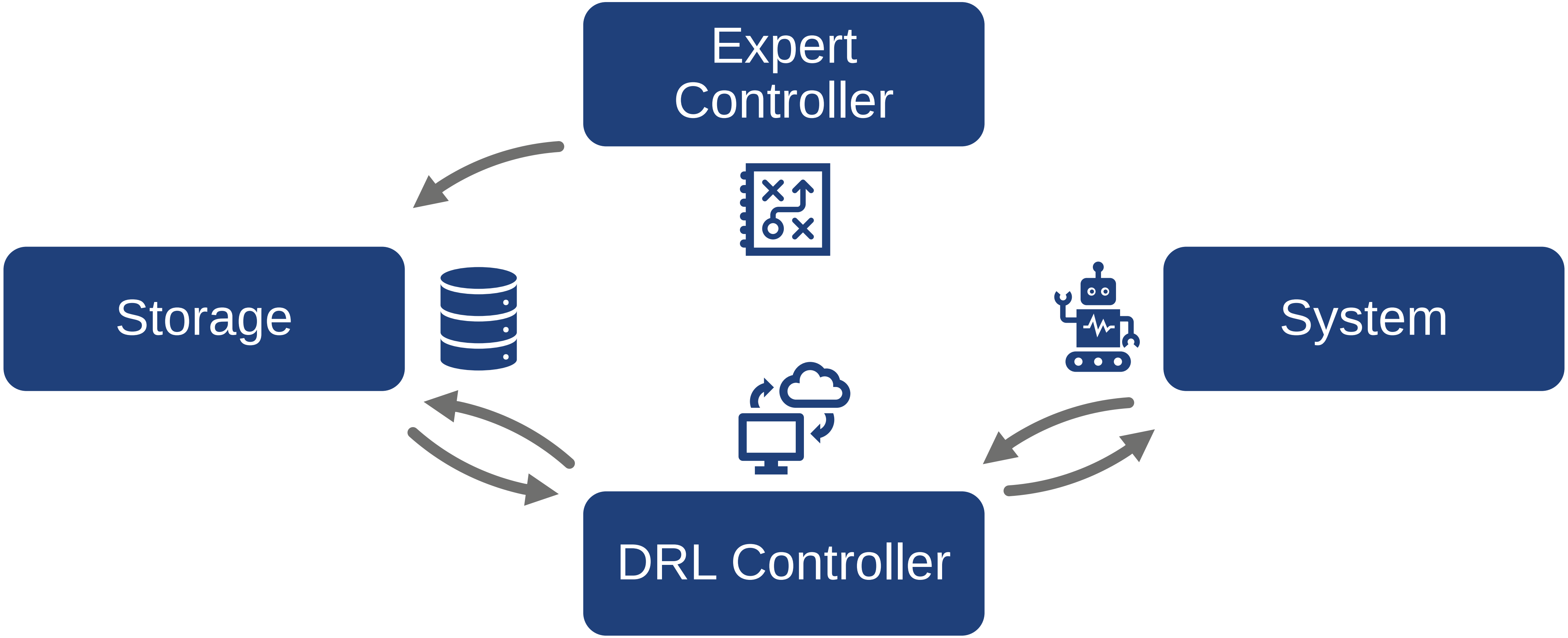
Learning from Demonstrations

- Imitation learning:
  - By design never outperform human experts
  - Only exploit narrow area of state-action space
- Combined reinforcement and imitation learning:
  - Reward / policy shaping
- Teacher / apprentticeship agents:
  - Learning from trained agents

Deep Q-learning from Demonstrations

Deep Q-learning from Demonstrations

Deep Q-learning from Demonstrations



Base Network

- Double DQN with prioritized experience replay [1,2]
  - Double DQN: reduced reward overestimation
  - Prioritized experience replay: increased number of hard tasks

$$J_{DQ}(Q) = \left(R(s,a) + \gamma Q(s_{t+1}, a_{t+1}^{\max}; \theta') - Q(s,a;\theta)\right)^2$$

[1] Van Hasselt, H., Guez, A., & Silver, D. (2016, March). Deep reinforcement learning with double q-learning. In *Thirtieth AAAI Conference on Artificial Intelligence*.

[2] Schaul, T., Quan, J., Antonoglou, I., & Silver, D. (2015). Prioritized experience replay. *arXiv preprint arXiv:1511.05952*.

Two Phase Learning

- Pre-training (offline)
  - Replay buffer:
    - Controller data
  - Loss:
    - 1-step double Q-learning loss
    - n-step double Q-learning loss (n=10)
    - Supervised large margin classification loss
    - L2 regularization loss

- Online learning
  - Replay buffer:
    - Controller data (not overwritten + prioritized)
    - Self-generated data
  - Loss:
    - 1-step double Q-learning loss
    - N-step double Q-learning loss
    - (Supervised large margin classification loss) for controller data
    - L2 regularization loss

Loss Function

- Supervised large margin classification loss [1]
  - Limits value of unseen actions

$$J_E(Q) = \max_{a \in A}[Q(s,a) + l(a_E, a)] - Q(s, a_E), \quad l(a_E, a) = \begin{cases} 0 & a = a_E \\ c > 0 & a \neq a_E \end{cases}$$

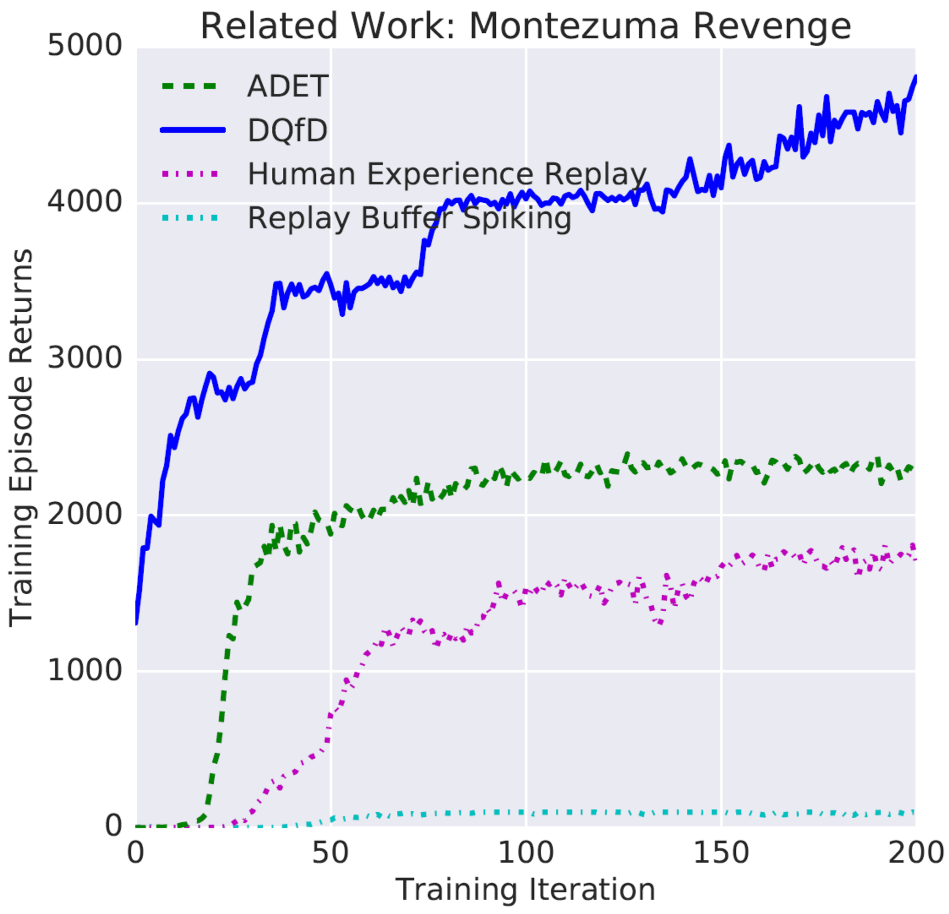
- 1-step + N-step double Q-learning loss
  - Guarantee Bellman equation
- L2 regularization loss
  - Network weight + bias regularization

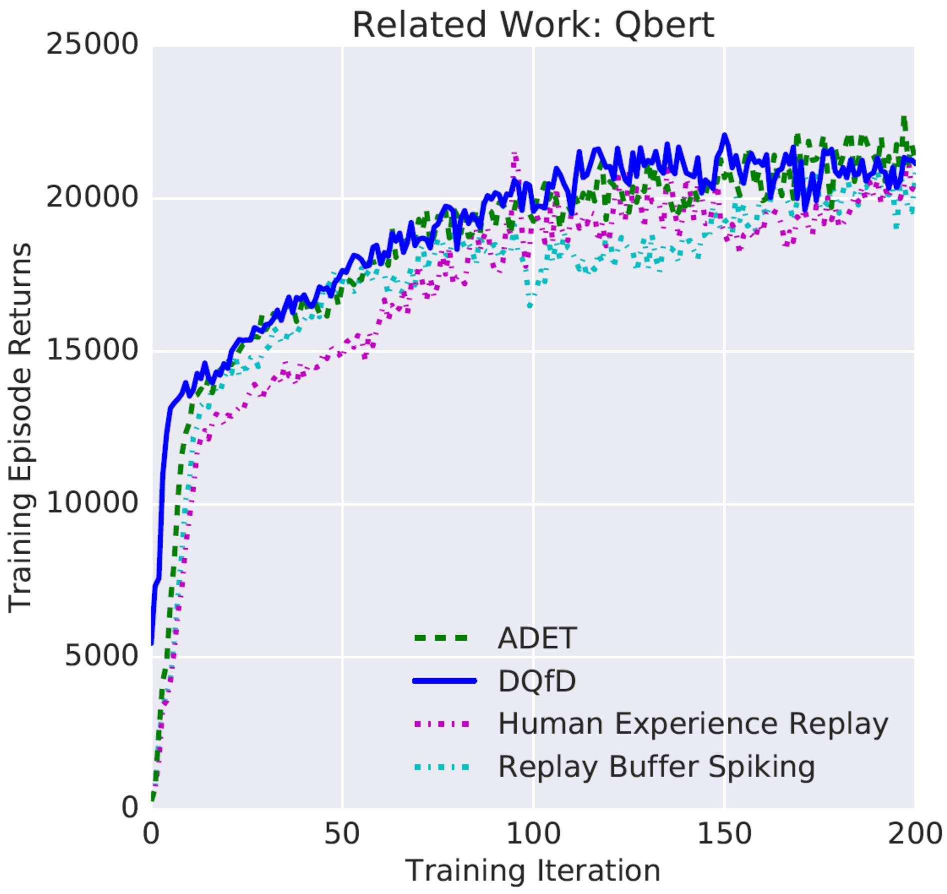
[1] Piot, B., Geist, M., & Pietquin, O. (2014, September). Boosted bellman residual minimization handling expert demonstrations. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 549-564). Springer, Berlin, Heidelberg.

Experiments

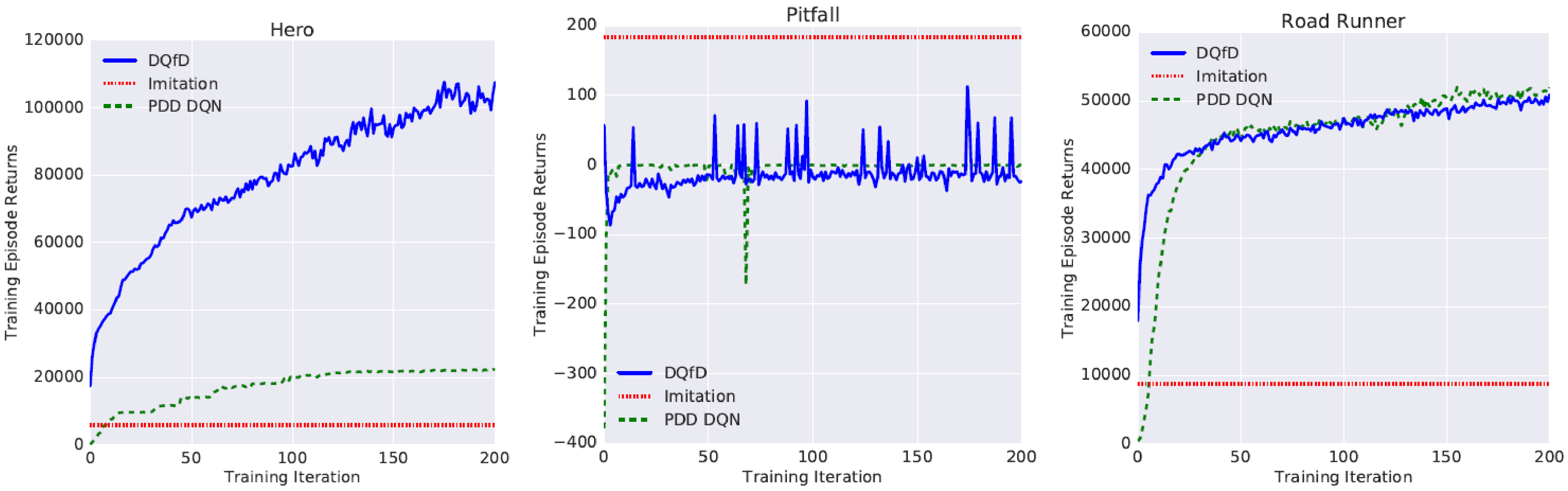
- 42 Atari games played 3-12 times → 5,574 to 75,472 transitions/game
  - Outperforms worst demonstration in 29 games
  - Outperforms best demonstration in 14 games

Results


Related Work: Montezuma Revenge
ADET
DQfD
Human Experience Replay
Replay Buffer Spiking
Training Episode Returns
0
1000
2000
3000
4000
5000
0
50
100
150
200
Training Iteration

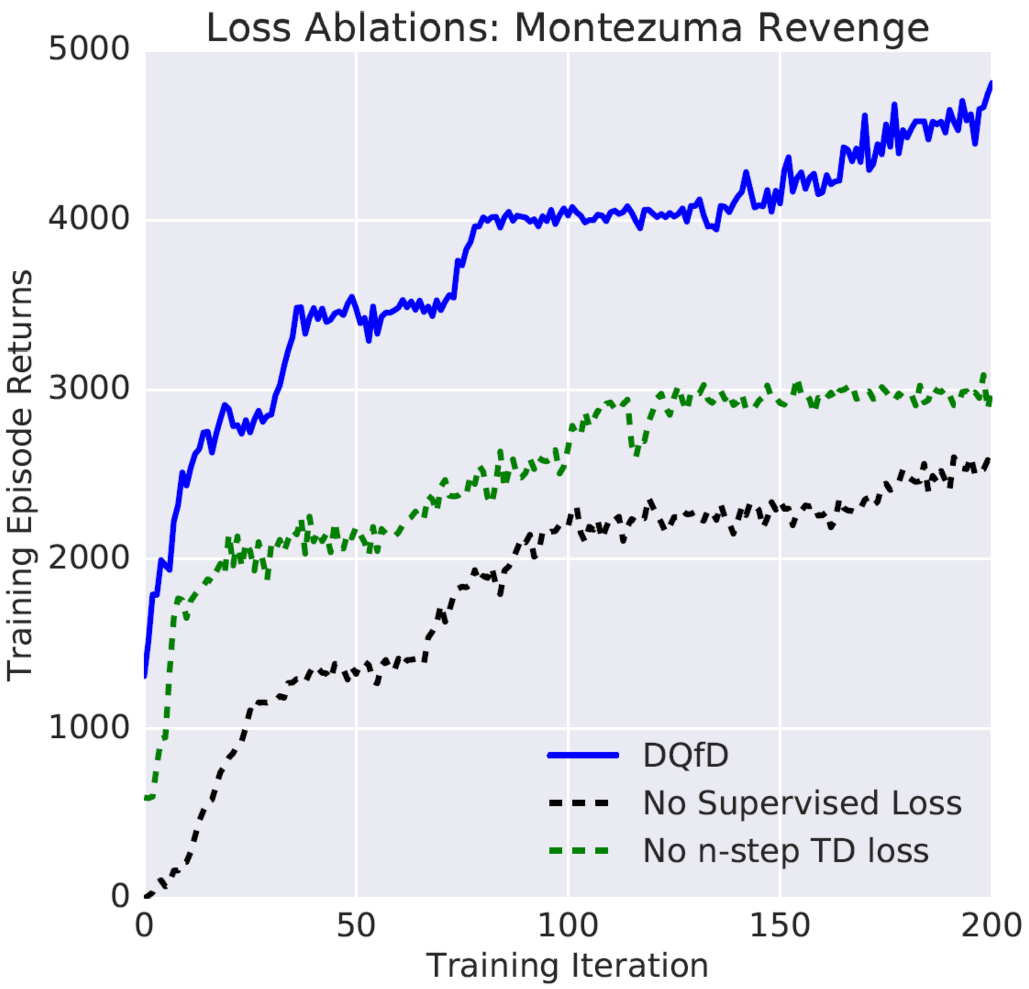

Related Work: Qbert
ADET
DQfD
Human Experience Replay
Replay Buffer Spiking
Training Episode Returns
0
5000
10000
15000
20000
25000
0
50
100
150
200
Training Iteration

Results

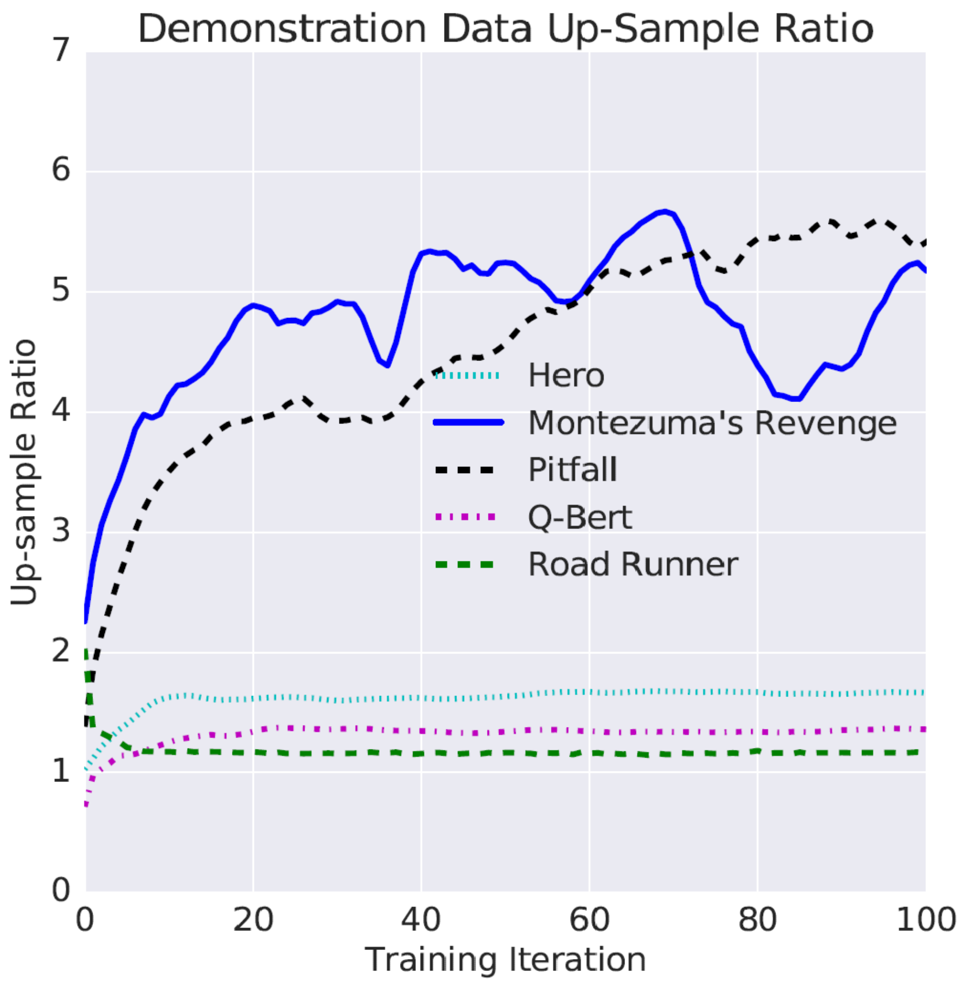


[1] Hester, T., Vecerik, M., Pietquin, O., Lanctot, M., Schaul, T., Piot, B., ... & Dulac-Arnold, G. (2018, April). Deep q-learning from demonstrations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

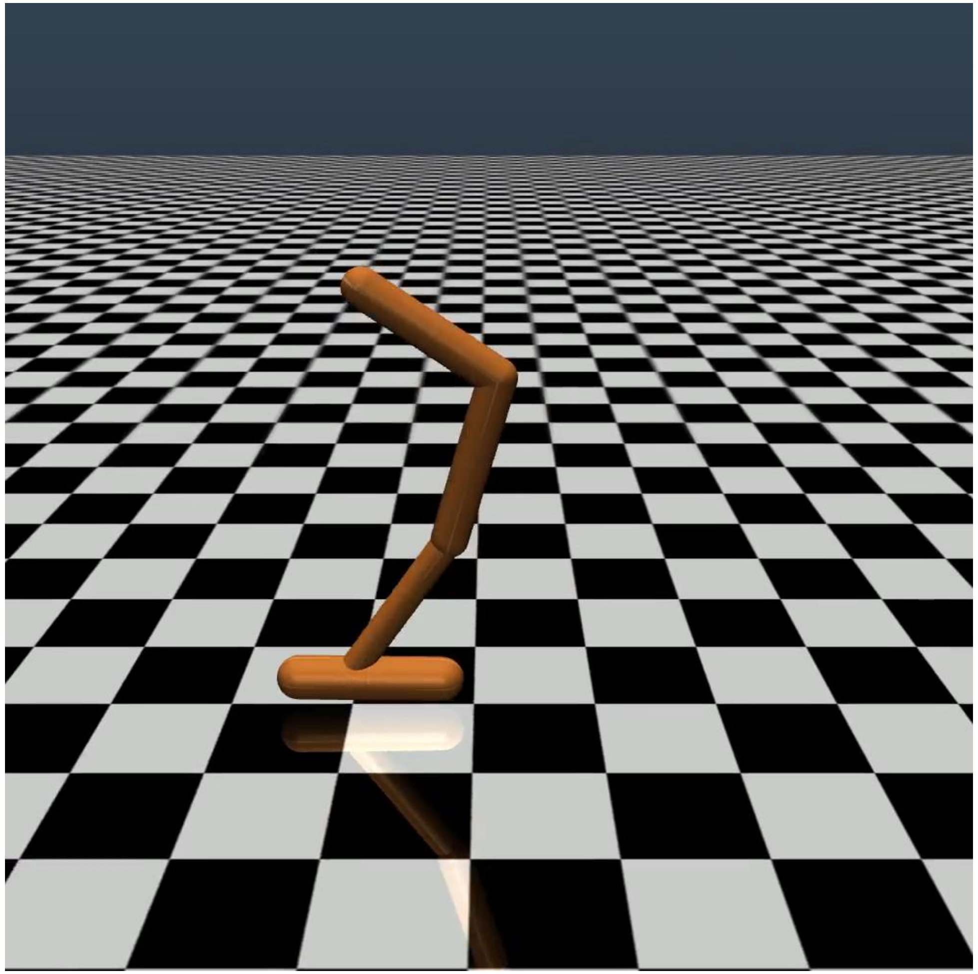
Ablation Study



Demonstration Up-Sample Ratio



We can intuitively define complex reward functions!

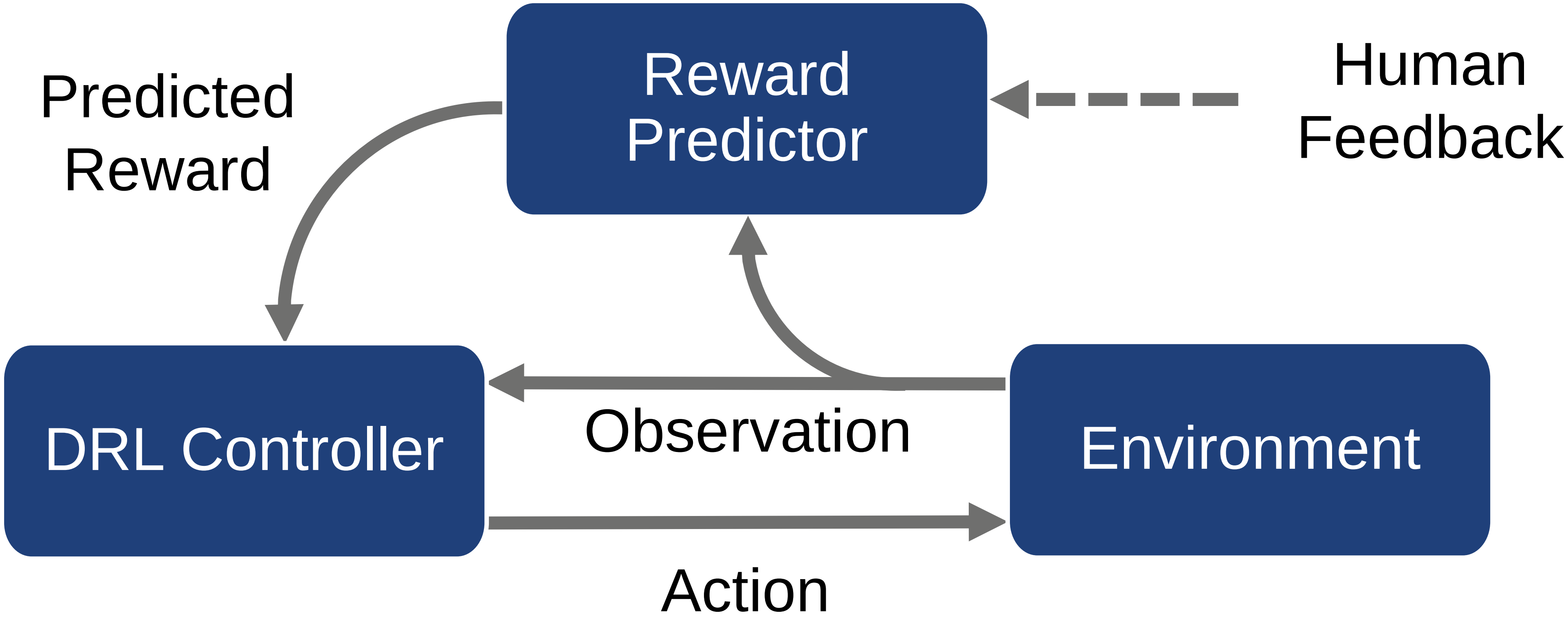


[1] Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In Advances in Neural Information Processing Systems (pp. 4299-4307).

Deep Reinforcement Learning from Human Preferences

- Solve DRL tasks without observing the true reward
- Comparison of video sequences ➔ intuitive evaluation
- Not contingent on human performing task
- Potential to outperform conventional DRL

Deep Reinforcement Learning from Human Preferences



Deep Reinforcement Learning from Human Preferences

- Conventional DRL step:

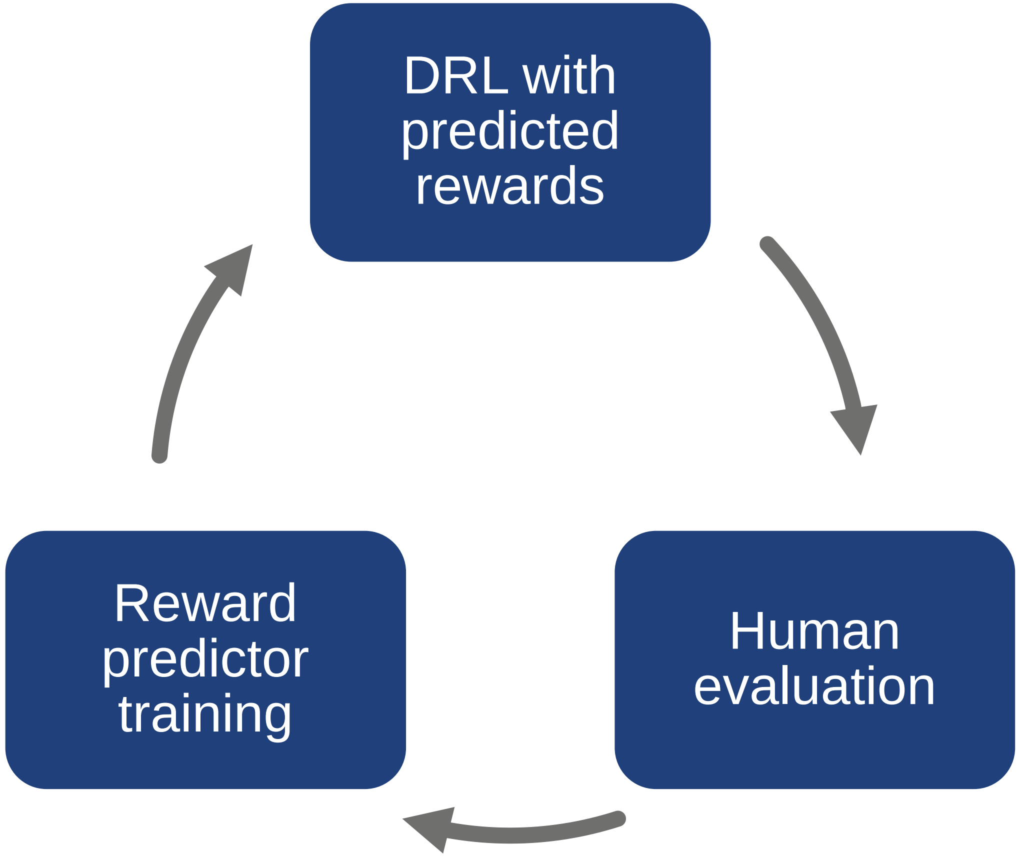
  $(o_i, a_i) \rightarrow r_i$

- Trajectory segment:

  $\sigma = ((o_0, a_0), (o_1, a_1), ..., (o_k - 1, a_k - 1)) \rightarrow r_{k-1}$

- Human can rate order of trajectory segments:
  - Goal in human language
  - Present video segments of agent's attempts
  - Rate videos $\sigma^1 \succ \sigma^2$

Training pipeline



- DRL with predicted rewards:
  - Interaction with environment
  - Trajectory generation
- Human evaluation:
  - Trajectory comparison
- Reward predictor training:
  - Optimization of reward predictor

DRL with Predicted Rewards

- Tasks:
  - Interaction with environment
  - Generation of trajectories
- Methods:
  - Conventional DRL with non-stationary reward function
  - Atari: advantage actor critic (A2C) [1]
  - Robots: trust region policy optimization (TRPO) [2]

[1] Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., ... & Kavukcuoglu, K. (2016, June). Asynchronous methods for deep reinforcement learning. In *International conference on machine learning* (pp. 1928-1937).
[2] Schulman, J., Levine, S., Abbeel, P., Jordan, M., & Moritz, P. (2015, June). Trust region policy optimization. In International Conference on Machine Learning (pp. 1889-1897).

Human Evaluation

- 1s – 2s segments are evaluated
- Database $\mathcal{D}$ of triples $(\sigma^1, \sigma^2, \mu)$
- Querries based on prediction variance ➔ approximates value of information

Reward Predictor Training

- Preference predictor: latent factor of human judgement

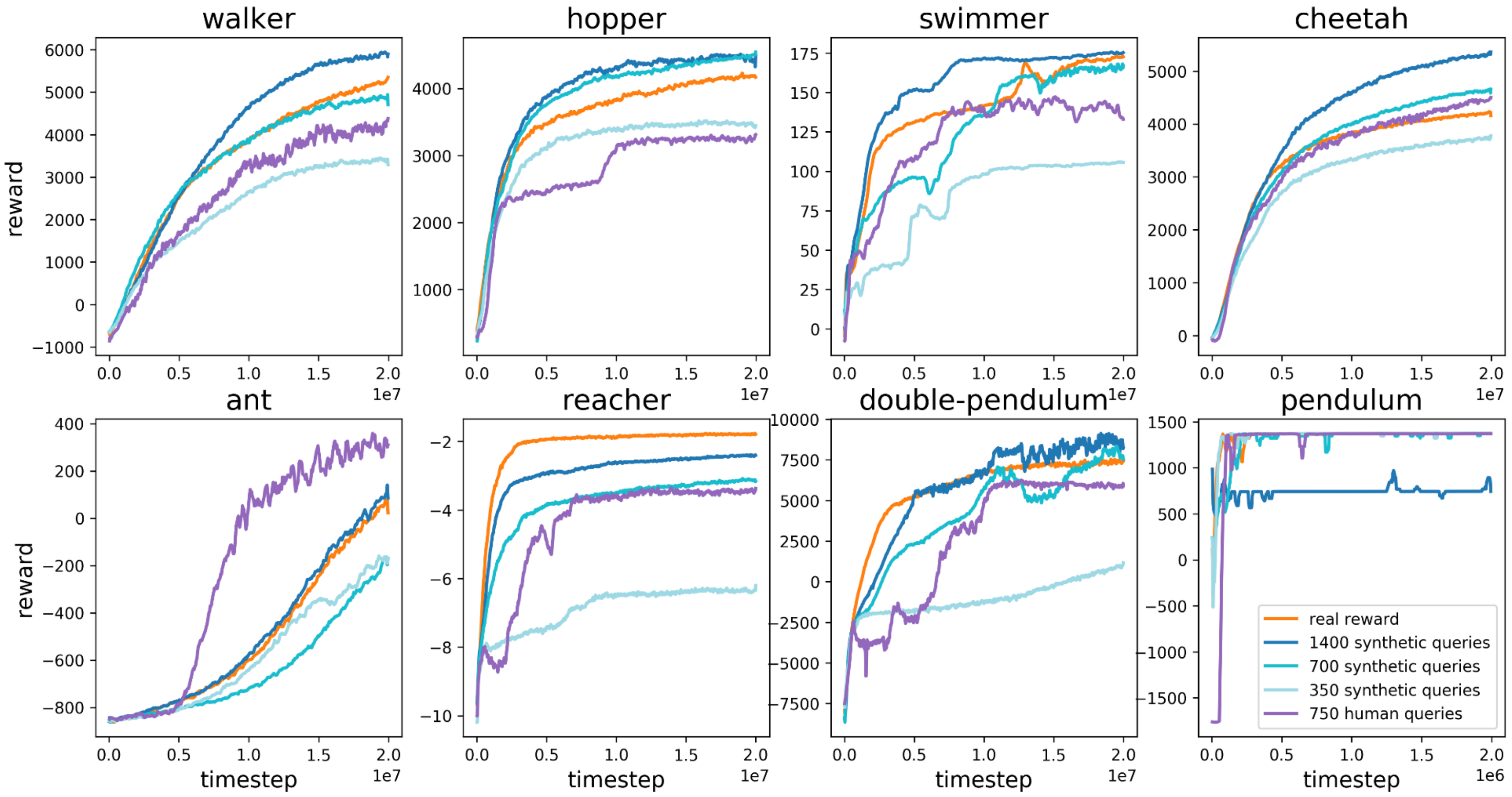
$$\hat{P}[\sigma^1 \succ \sigma^2] = \frac{\exp \sum \hat{r}(o_t^1, a_t^1)}{\exp \sum \hat{r}(o_t^1, a_t^1) + \exp \sum \hat{r}(o_t^2, a_t^2)}$$

- Training with cross entropy loss

$$\operatorname{loss}(\hat{r}) = - \sum_{(\sigma^1, \sigma^2, \mu) \in \mathcal{D}} \mu(1) \log \hat{P}[\sigma^1 \succ \sigma^2] + \mu(2) \log \hat{P}[\sigma^2 \succ \sigma^1]$$

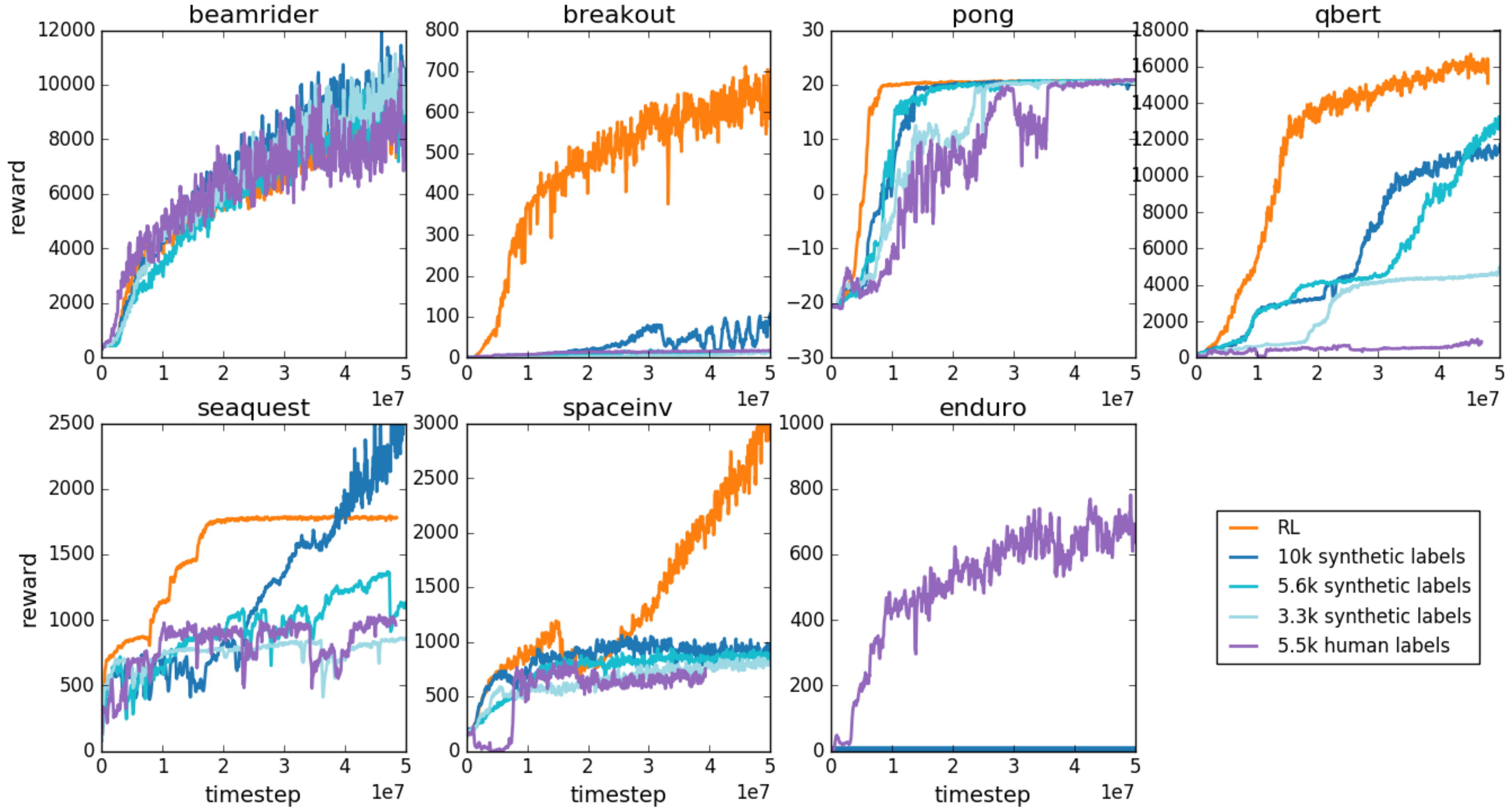
- Implementation details:
  - Ensemble of predictors
  - L2 regularization optimized on validation set
  - Assumption: Human choice 10% at random

Results Simulated Robotics



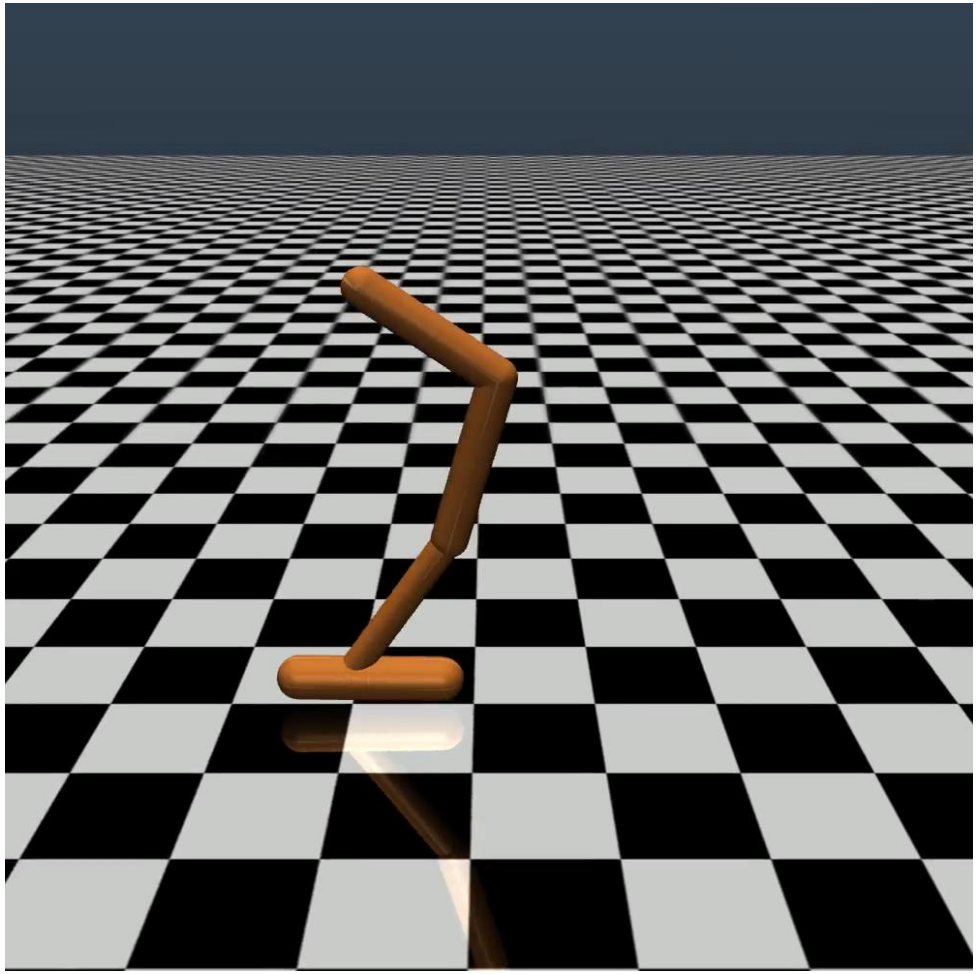
[1] Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In Advances in Neural Information Processing Systems (pp. 4299-4307).

Results Atari



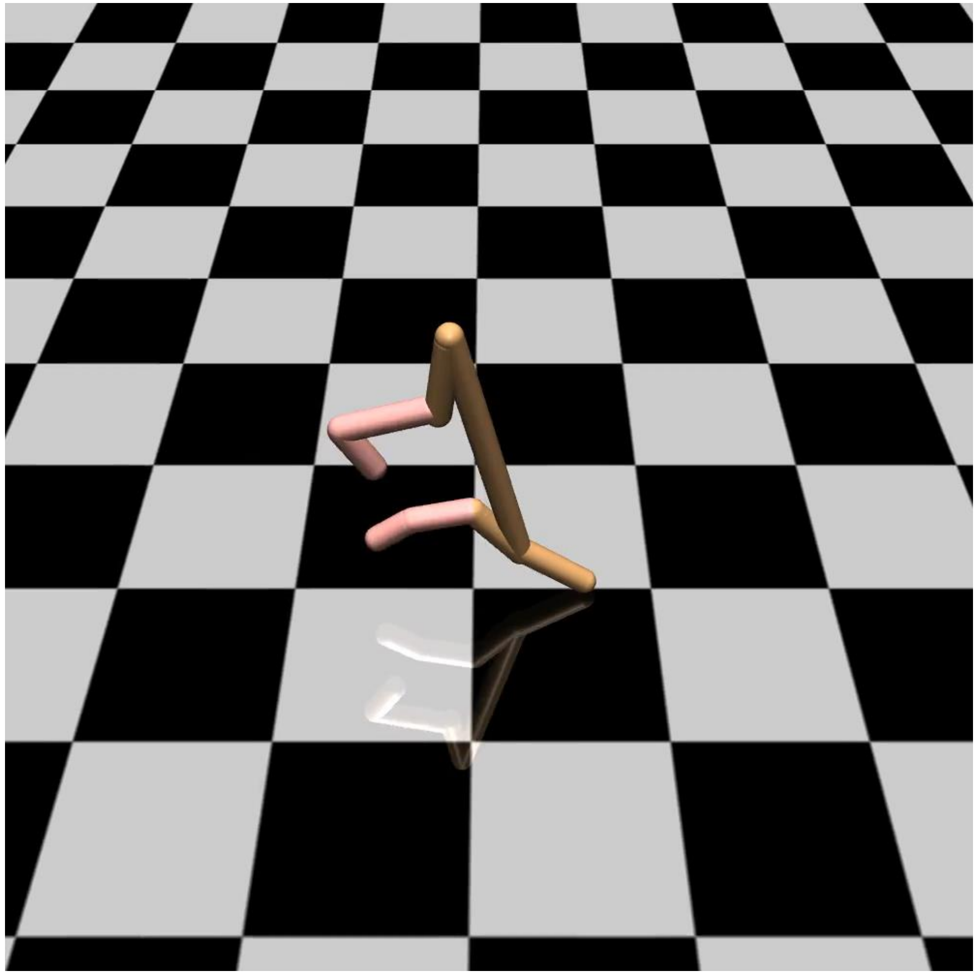
[1] Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In Advances in Neural Information Processing Systems (pp. 4299-4307).

Complex Task: Hopper Backflip



[1] Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In Advances in Neural Information Processing Systems (pp. 4299-4307).

Complex Task: Half-Cheetah Handstand

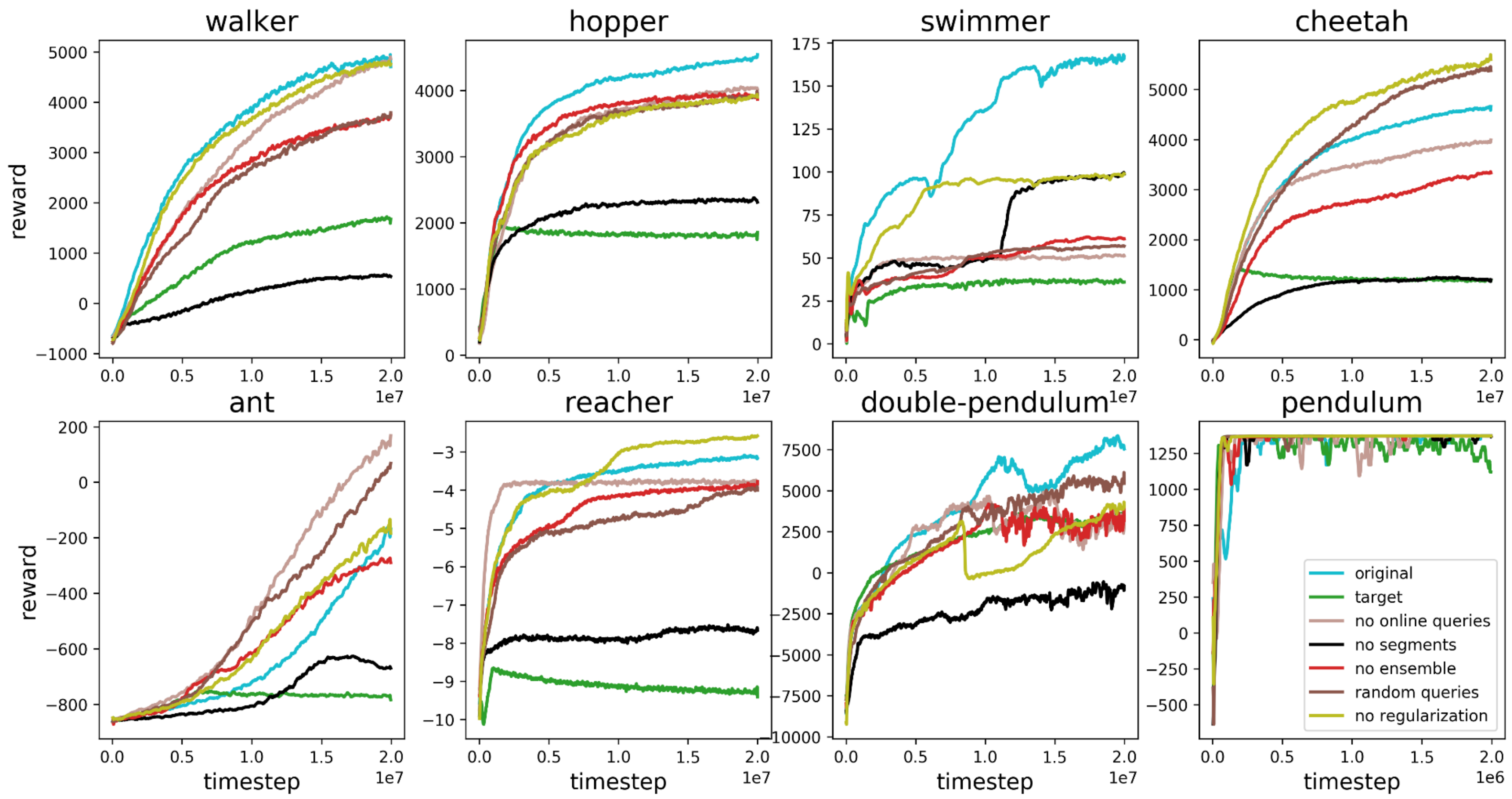


[1] Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In Advances in Neural Information Processing Systems (pp. 4299-4307).

Complex Task: Enduro keep alongside cars

[1] Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In Advances in Neural Information Processing Systems (pp. 4299-4307).

Ablation Simulated Robotics



[1] Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In Advances in Neural Information Processing Systems (pp. 4299-4307).

Ablation Results

- Offline reward predictor training results in strange behavior
- Querying comparisons is more helpful than absolute scores
- Sequences are more helpful than single frames

Summary

- DRL for hard tasks can profit from human intuition
- Boost initial performance with demonstrations
- Behavioral ratings for not directly solvable tasks