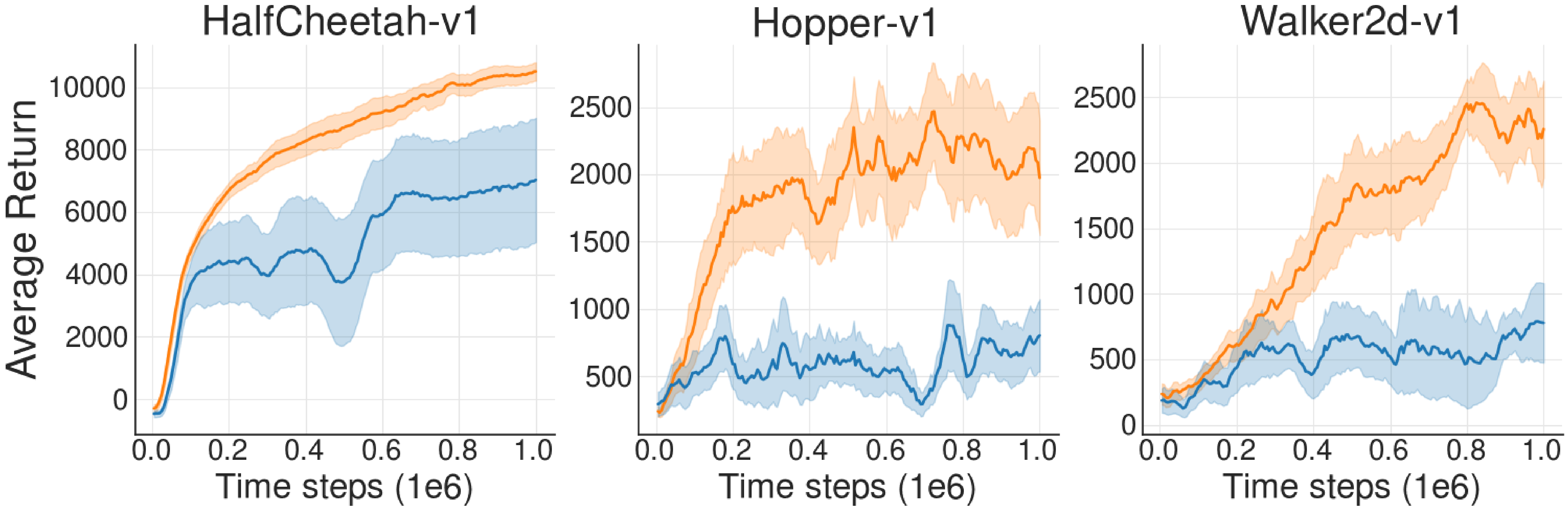
Off-Policy Deep Reinforcement Learning without Exploration

Scott Fujimoto, David Meger, Doina Precup

Mila, McGill University


HalfCheetah-v1
Hopper-v1
Walker2d-v1
Average Return
Time steps (1e6)
Time steps (1e6)
Time steps (1e6)
10000
8000
6000
4000
2000
0
2500
2000
1500
1000
500
0.0
0.2
0.4
0.6
0.8
1.0

Surprise!

Agent orange and agent blue are trained with…

1. The **same off-policy algorithm** (DDPG).

2. The **same dataset.**

The Difference?

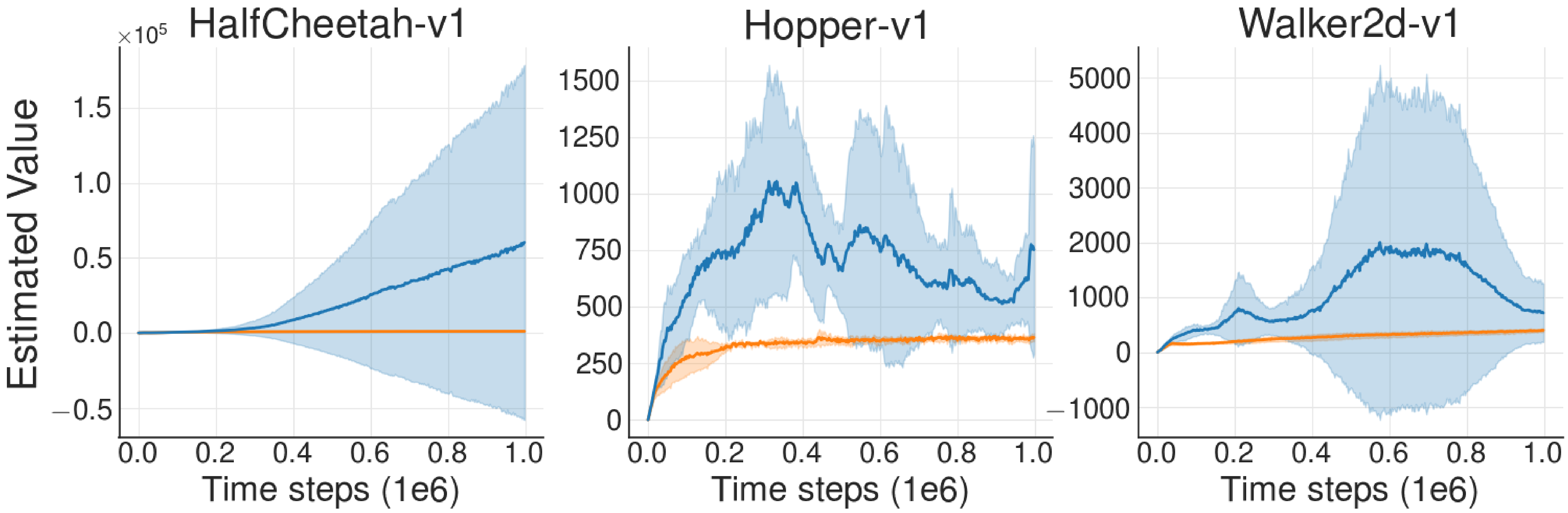
1. Agent orange: Interacted with the environment.
    - Standard RL loop.
    - Collect data, store data in buffer, train, repeat.

2. Agent blue: Never interacted with the environment.
    - Trained with data collected by agent orange concurrently.

1. Trained with the same off-policy algorithm.
2. Trained with the same dataset.
3. One interacts with the environment. One doesn’t.

Off-policy deep RL fails when **truly off-policy**.

Value Predictions



Extrapolation Error

$$Q(s, a) \leftarrow r + \gamma Q(s', a')$$

Extrapolation Error

$$Q(s, a) \leftarrow r + \gamma Q(s', a')$$

GIVEN

GENERATED

Extrapolation Error

$$Q(s, a) \leftarrow r + \gamma Q(s', a')$$

1. $(s, a, r, s') \sim Dataset$
2. $a' \sim \pi(s')$

Extrapolation Error

$$Q(s,a) \leftarrow r + \gamma Q(s',a')$$

$$(s',a') \notin Dataset \rightarrow Q(s',a') = \mathbf{bad}$$

$$\rightarrow Q(s,a) \quad = \mathbf{bad}$$

Extrapolation Error

$$Q(s,a) \leftarrow r + \gamma \boxed{Q(s',a')}$$

$$(s',a') \notin Dataset \rightarrow \boxed{Q(s',a') = \mathbf{bad}}$$

$$\rightarrow Q(s,a) \quad = \mathbf{bad}$$

Extrapolation Error

$$\boxed{Q(s,a)} \leftarrow r + \gamma Q(s',a')$$

$$(s',a') \notin Dataset \rightarrow Q(s',a') = \textbf{bad}$$

$$\rightarrow \boxed{Q(s,a) \quad = \textbf{bad}}$$

Extrapolation Error

Attempting to evaluate π without (sufficient) access to the (s, a) pairs π visits.

Batch-Constrained Reinforcement Learning

Only choose π such that we have access to the (s, a) pairs π visits.

Batch-Constrained Reinforcement Learning

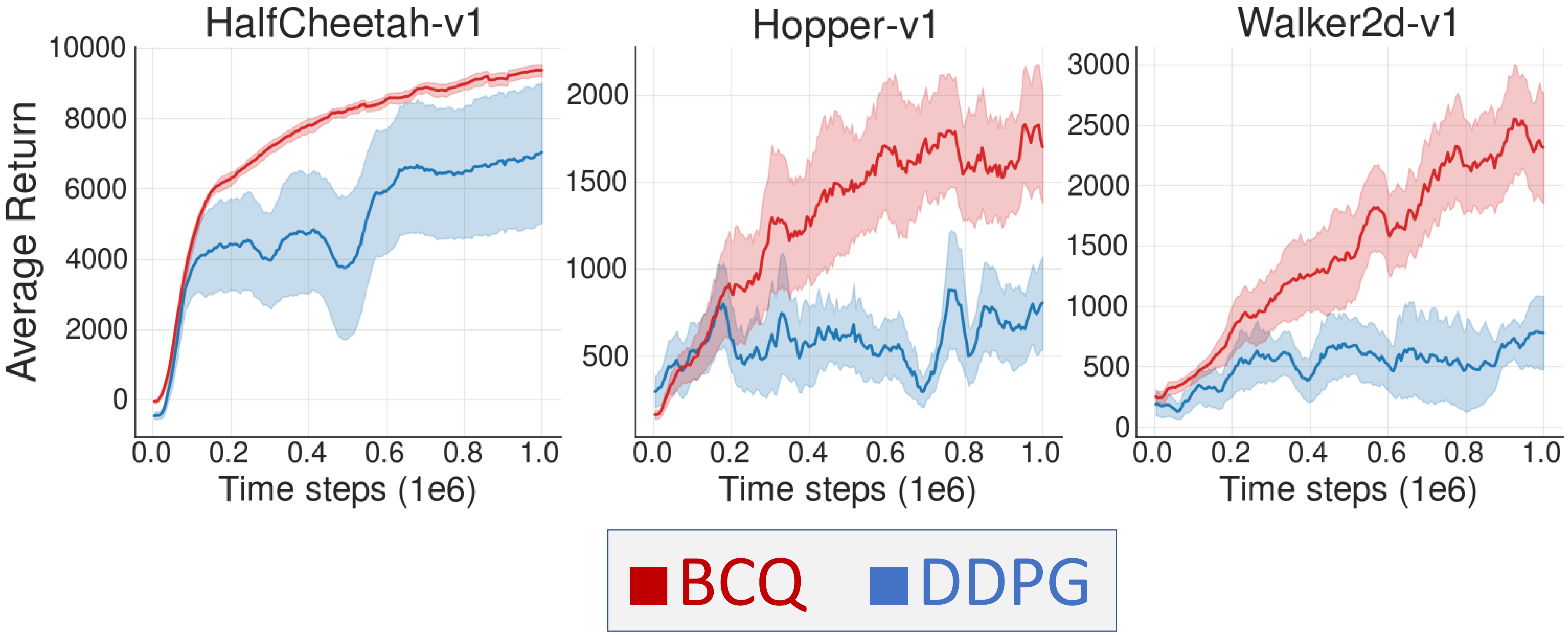
1. a$\sim\pi(s)$ such that $(s, a) \in Dataset$.
2. a$\sim\pi(s)$ such that $\left(s', \pi(s')\right) \in Dataset$.
3. a$\sim\pi(s)$ such that $Q(s, a)$ is maxed.

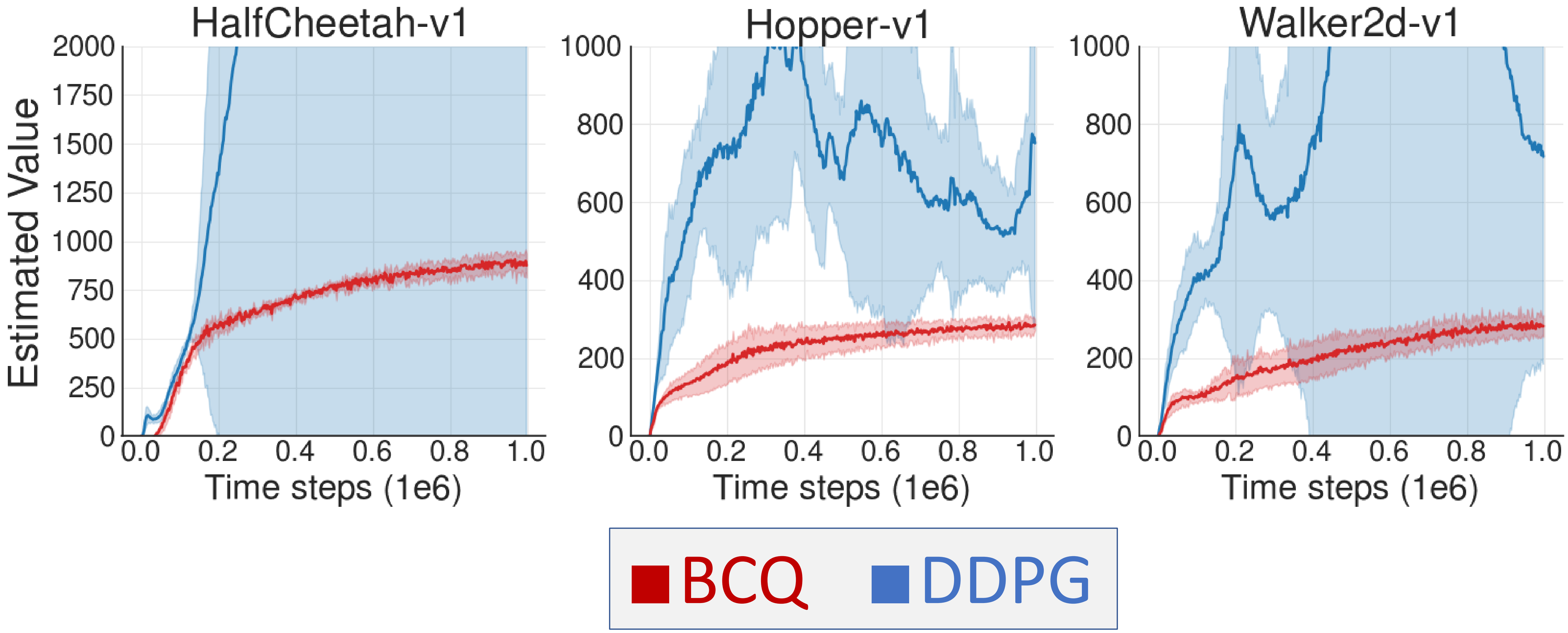
Batch-Constrained Deep Q-Learning (BCQ)

First imitate dataset via generative model:
$G(a|s) \approx P_{Dataset}(a|s)$.

$\pi(s) = \operatorname{argmax}_{a_i} Q\,(s, a_i)$, where $a_i \sim G$
(I.e. select the best action that is likely under the dataset)

(+ some additional deep RL magic)

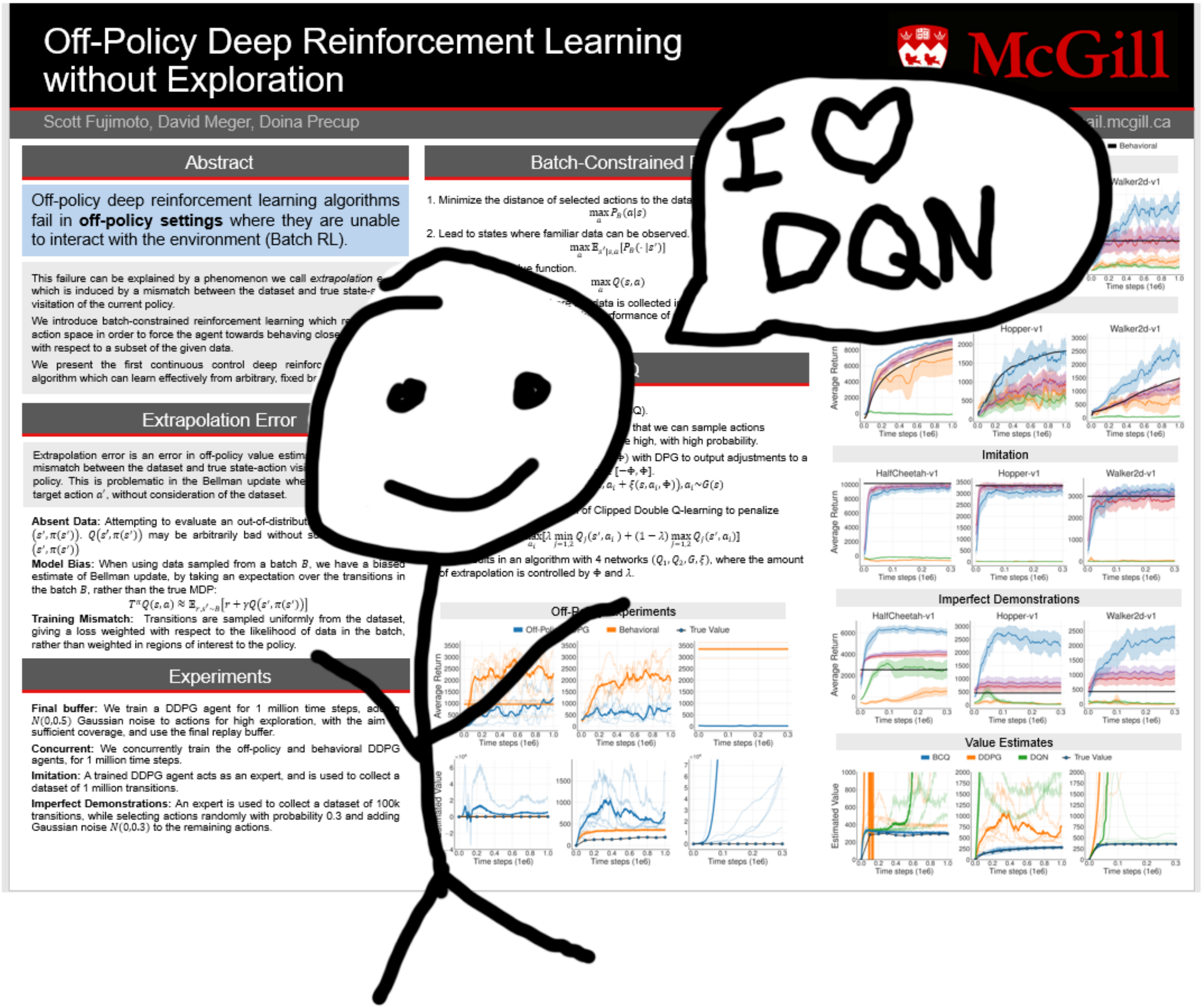

HalfCheetah-v1
Hopper-v1
Walker2d-v1
Average Return
10000
8000
6000
4000
2000
0
2000
1500
1000
500
3000
2500
2000
1500
1000
500
0
0.0
0.2
0.4
0.6
0.8
1.0
Time steps (1e6)
BCQ
DDPG


HalfCheetah-v1
Hopper-v1
Walker2d-v1
Estimated Value
2000
1750
1500
1250
1000
750
500
250
0
800
600
400
200
0.0
0.2
0.4
0.6
0.8
1.0
Time steps (1e6)
BCQ
DDPG

Come say Hi @ Pacific Ballroom #38 (6:30 Tonight)

---

https://github.com/sfujim/BCQ



(Artist's rendition of poster session)