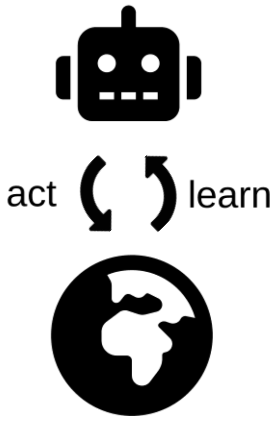
A Dataset Perspective on Offline Reinforcement Learning

Kajetan Schweighofer*, Andreas Radler*, Marius-Constantin Dinu*,
Markus Hofmarcher, Vihang Patil, Angela Bitto-Nemling,
Hamid Eghbal-zadeh, Sepp Hochreiter

* Equal contribution

Offline Reinforcement Learning

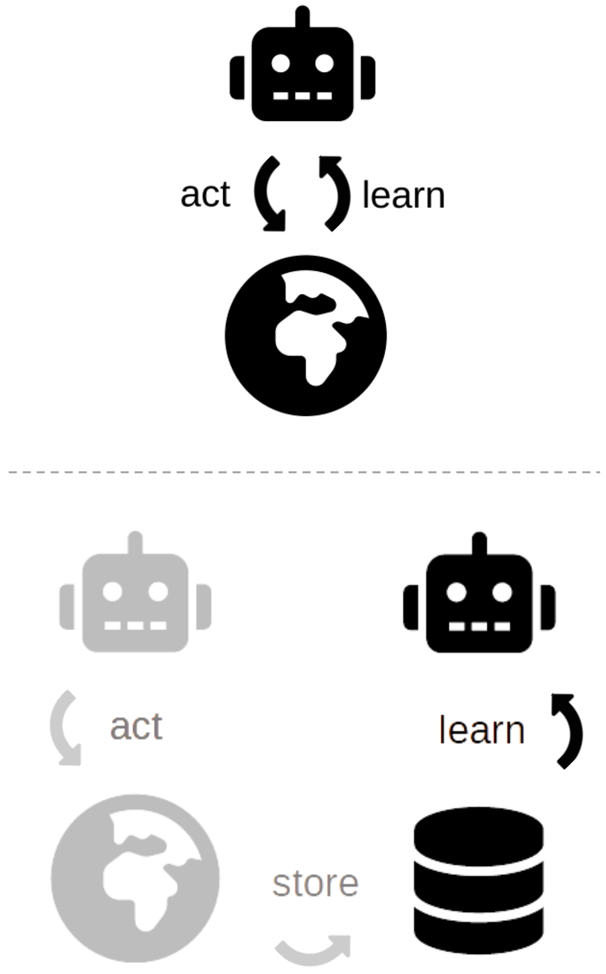
- Reinforcement Learning (RL)
  - Active interaction with environment
  - Corrective feedback
  - Interaction is potentially dangerous & slow



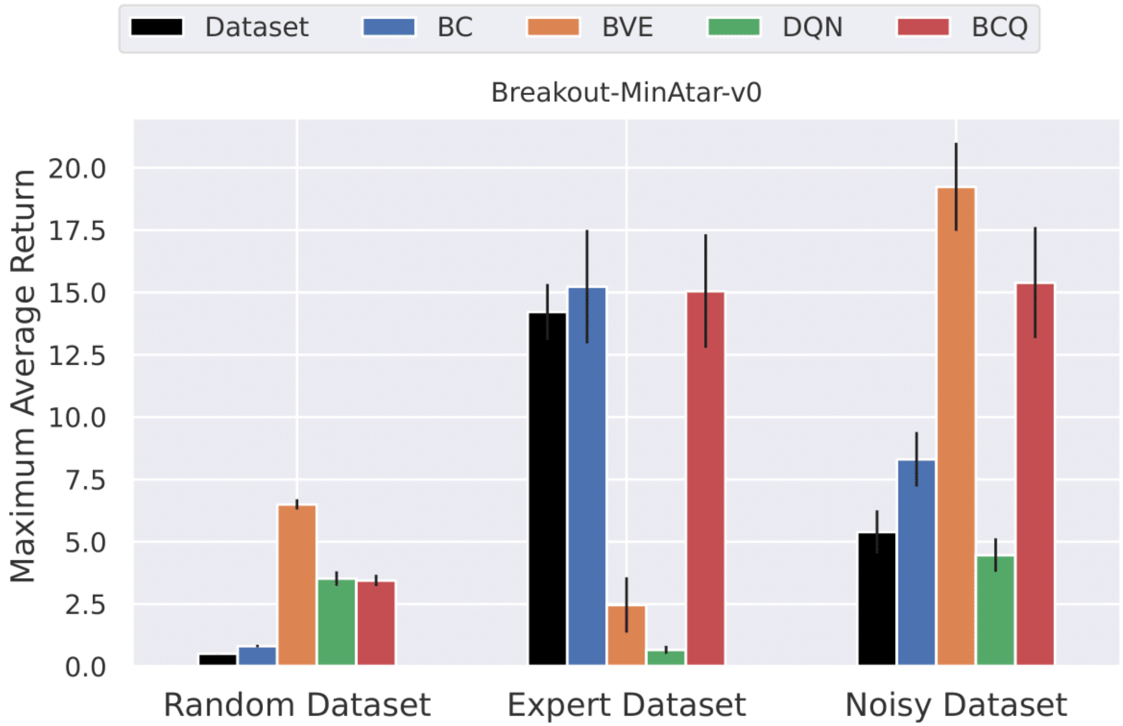
Offline Reinforcement Learning

- Reinforcement Learning (RL)
  - Active interaction with environment
  - Corrective feedback
  - Interaction is potentially dangerous & slow

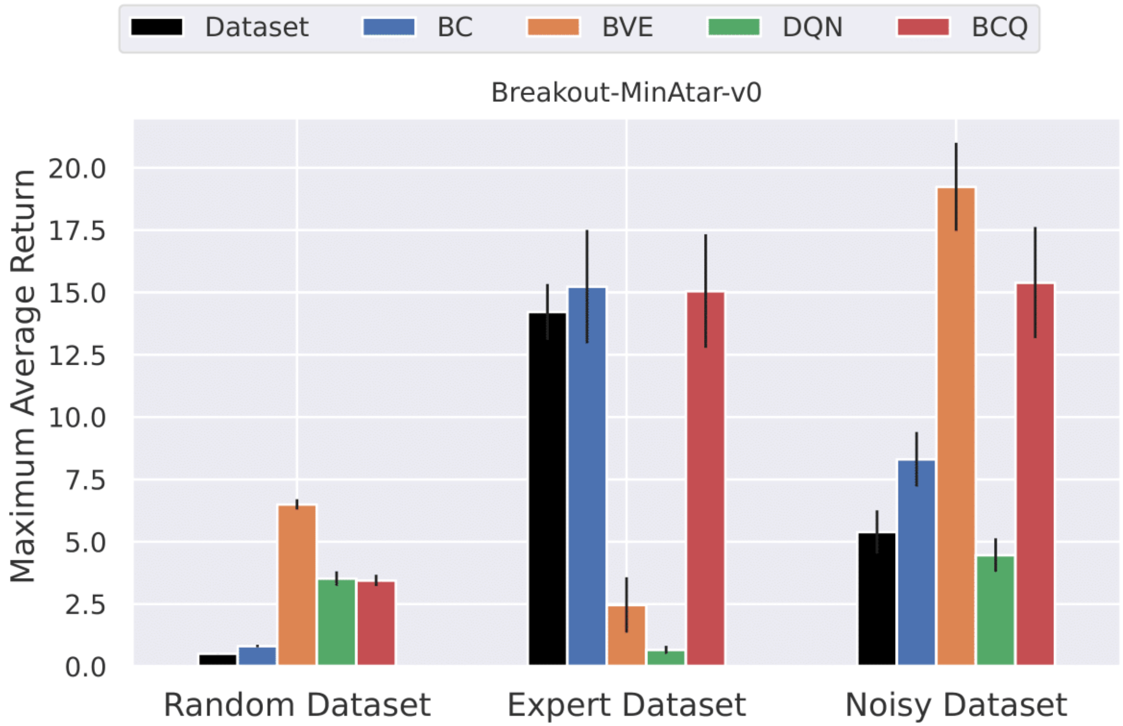
- Offline RL
  - No interaction with environment
  - Leverage safely collected transitions
  - No corrective feedback
  - Distribution shifts & iterative error amplification



Effects of different Datasets



Effects of different Datasets

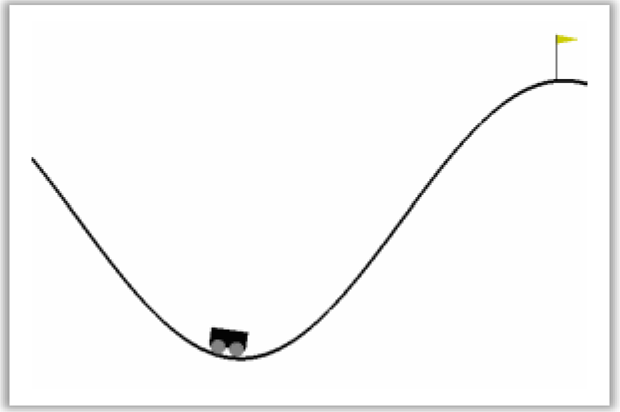


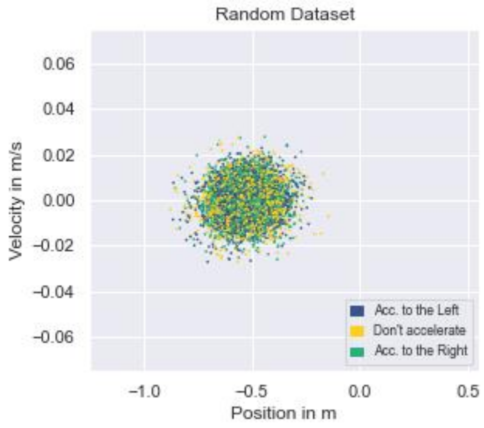
How do dataset characteristics influence algorithms in Offline RL?

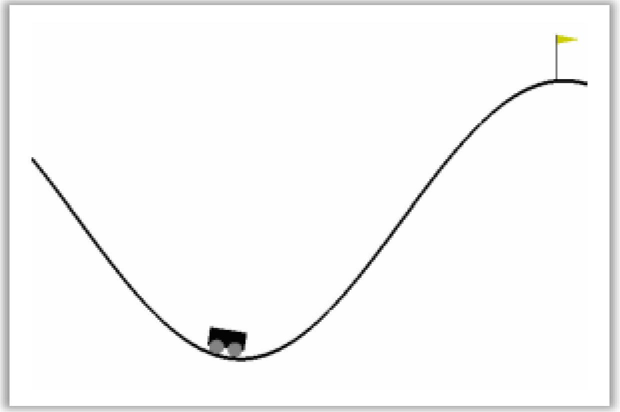
Characterizing RL Datasets

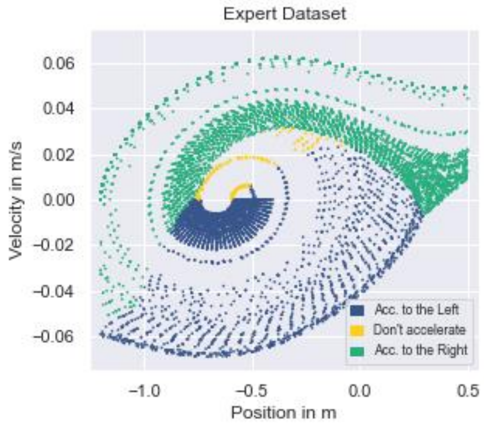
- How to quantify different behavior?
    - Inspection through UI
    - Visualizations
    - Measures

- Characterize attributes of the behavioral policy
    - Exploitation
    - Exploration









Exploitation Measure

- **Theoretical:** Expected Return

$$g_\pi = \mathbb{E}_\pi\left[\sum_{t=0}^{T}\gamma^t R_{t+1}\right]$$

- **Empirical:** Average Return

$$\bar{g}(\mathcal{D}) = \frac{1}{B}\sum_{b=0}^{B}\sum_{t=0}^{T_b}\gamma^t r_{b,t}$$

- **Normalized:** Trajectories Quality (TQ)

$$TQ(\mathcal{D}) := \frac{\bar{g}(\mathcal{D}) - \bar{g}(\mathcal{D}_{\text{min}})}{\bar{g}(\mathcal{D}_{\text{expert}}) - \bar{g}(\mathcal{D}_{\text{min}})}$$

Exploration Measure

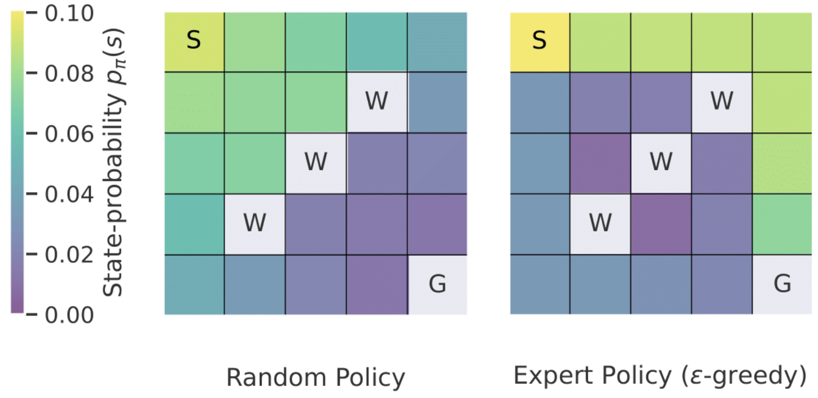
- **Theoretical:** Transition-Entropy

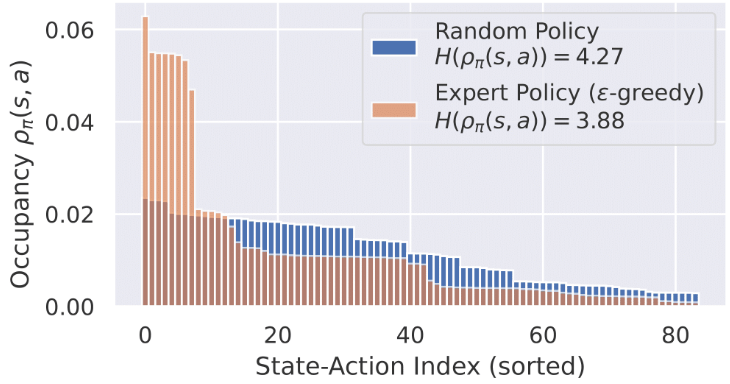
$$
\begin{aligned}
&H(p_\pi(s,a,r,s')) \\
&\quad = -\sum_{s,a,r,s'} p_\pi(s,a,r,s') \log\left(p_\pi(s,a,r,s')\right) \\
&\quad = \sum_{s,a} \rho_\pi(s,a)\ H(p(r,s' \mid s,a)) + H(\rho_\pi(s,a))
\end{aligned}
$$

- Deterministic MDPs: Occupancy-Entropy

$$H(\rho_\pi(s,a)) = -\sum_{s,a} \rho_\pi(s,a) \log\left(\rho_\pi(s,a)\right)$$

$$p_\pi(s) = \sum_a \rho_\pi(s,a)$$





JKU dynatrace

Exploration Measure

- **Empirical:** Estimator for occupancy-entropy $\hat{H}(\mathcal{D})$

- Upper bound: Unique state-action pairs

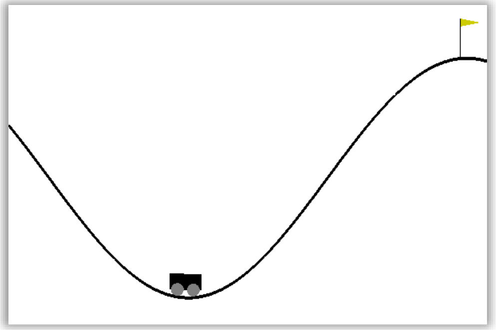
$$\hat{H}(\mathcal{D}) \leq \log(u_{s,a}(\mathcal{D}))$$

$$e^{\hat{H}(\mathcal{D})} \leq u_{s,a}(\mathcal{D})$$

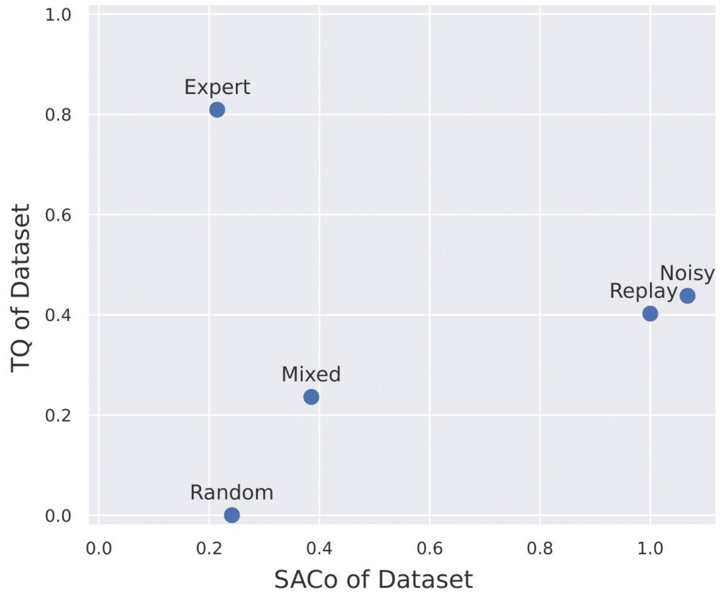
- **Normalized:** State-Action Coverage (SACo) $SACo(\mathcal{D}) := \frac{u_{s,a}(\mathcal{D})}{u_{s,a}(\mathcal{D}_{\text{ref}})}$

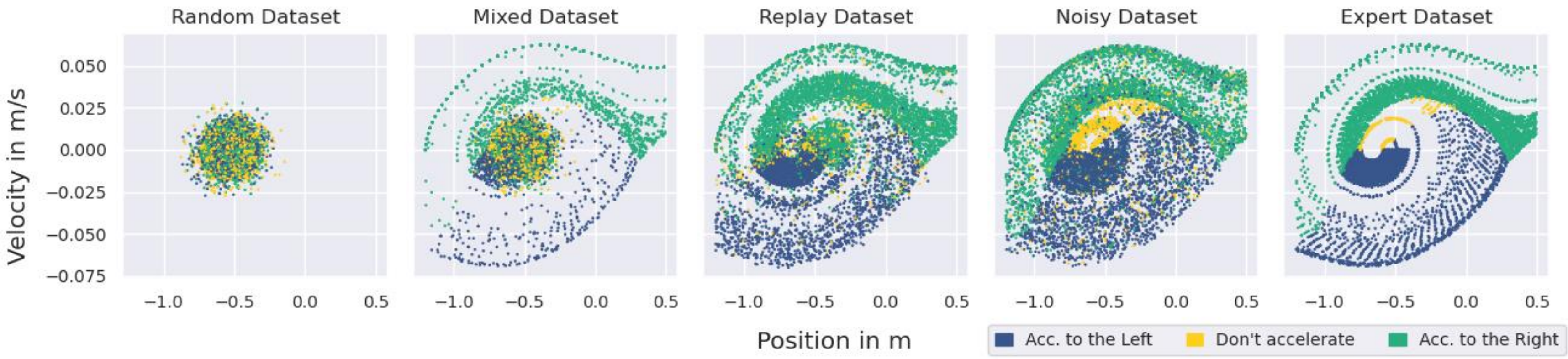
Dataset Generation

- Random
- Mixed
- Replay
- Noisy
- Expert



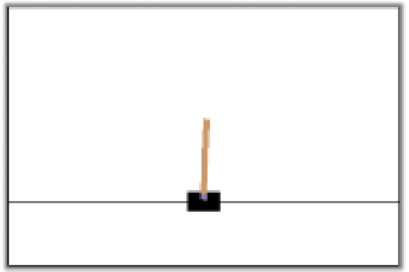
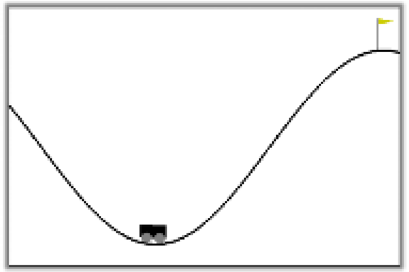
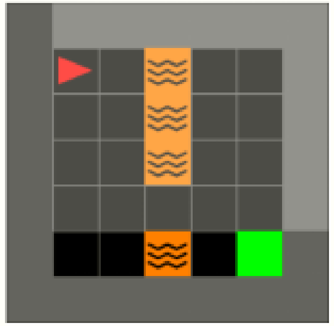
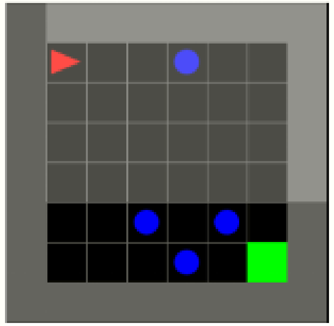
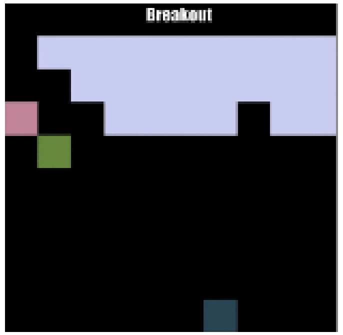
MountainCar

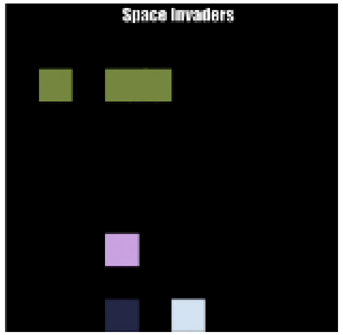




Environments

- Classic Control
  - CartPole
  - MountainCar
- MiniGrid
  - LavaGap
  - DynamicObstacles
- MinAtar
  - Breakout
  - SpaceInvaders

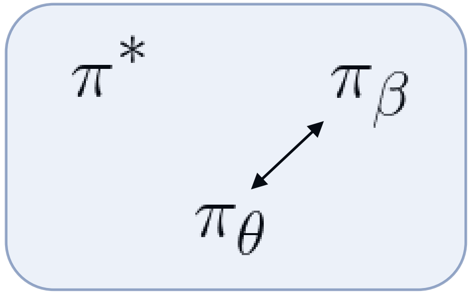








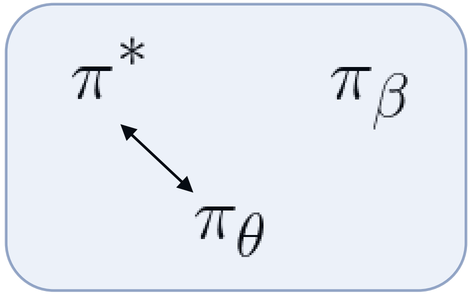
Images created using / taken from
Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. (2016). Openai gym. ArXiv:1606.01540.
Chevalier-Boisvert, M., Willems, L., and Pal, S. (2018). Minimalistic gridworld environment for openai gym. https://github.com/maximecb/gym-minigrid .
Young, K. and Tian, T. (2019). Minatar: An atari-inspired testbed for thorough and reproducible reinforcement learning experiments. arXiv, 1903.03176.

Algorithms

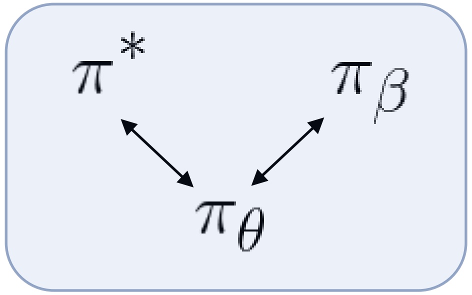
- Baselines
  - Behavior Cloning (**BC**)
  - Behavior Value Estimation (**BVE**)
  - Monte Carlo Estimation (**MCE**)



- Unconstrained off-policy algorithms
  - Deep Q-Network (**DQN**)
  - Quantile Regression DQN (**QRDQN**)
  - Random Ensemble Mixture (**REM**)



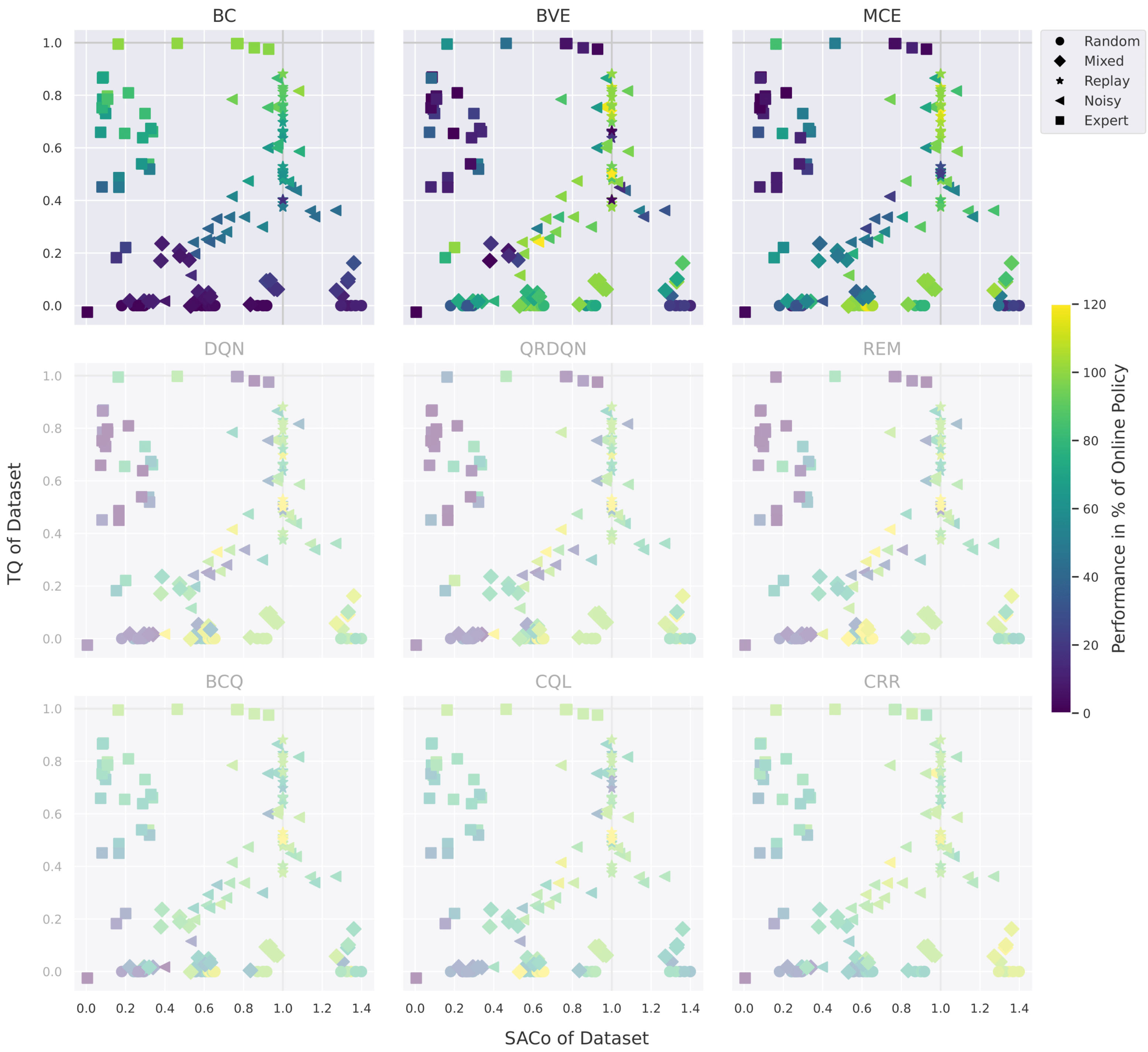
- Dataset-constrained off-policy algorithms
  - Batch-Constrained Q-learning (**BCQ**)
  - Conservative Q-learning (**CQL**)
  - Critic Regularized Regression (**CRR**)



Results

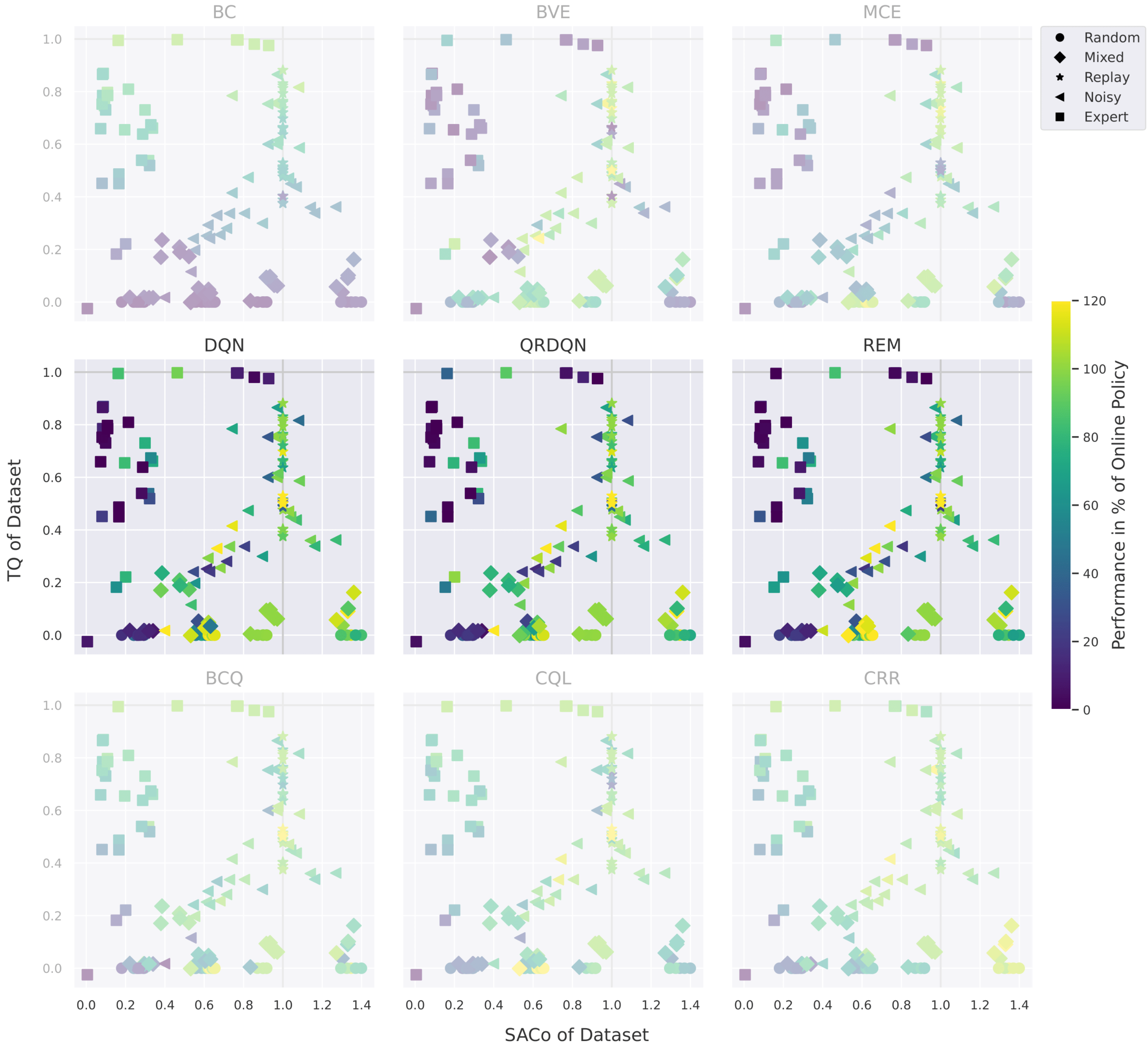
- Baseline Algorithms

- Best performance
  - BC: high TQ
  - BVE: moderate SACo
  - MCE: moderate SACo



Results

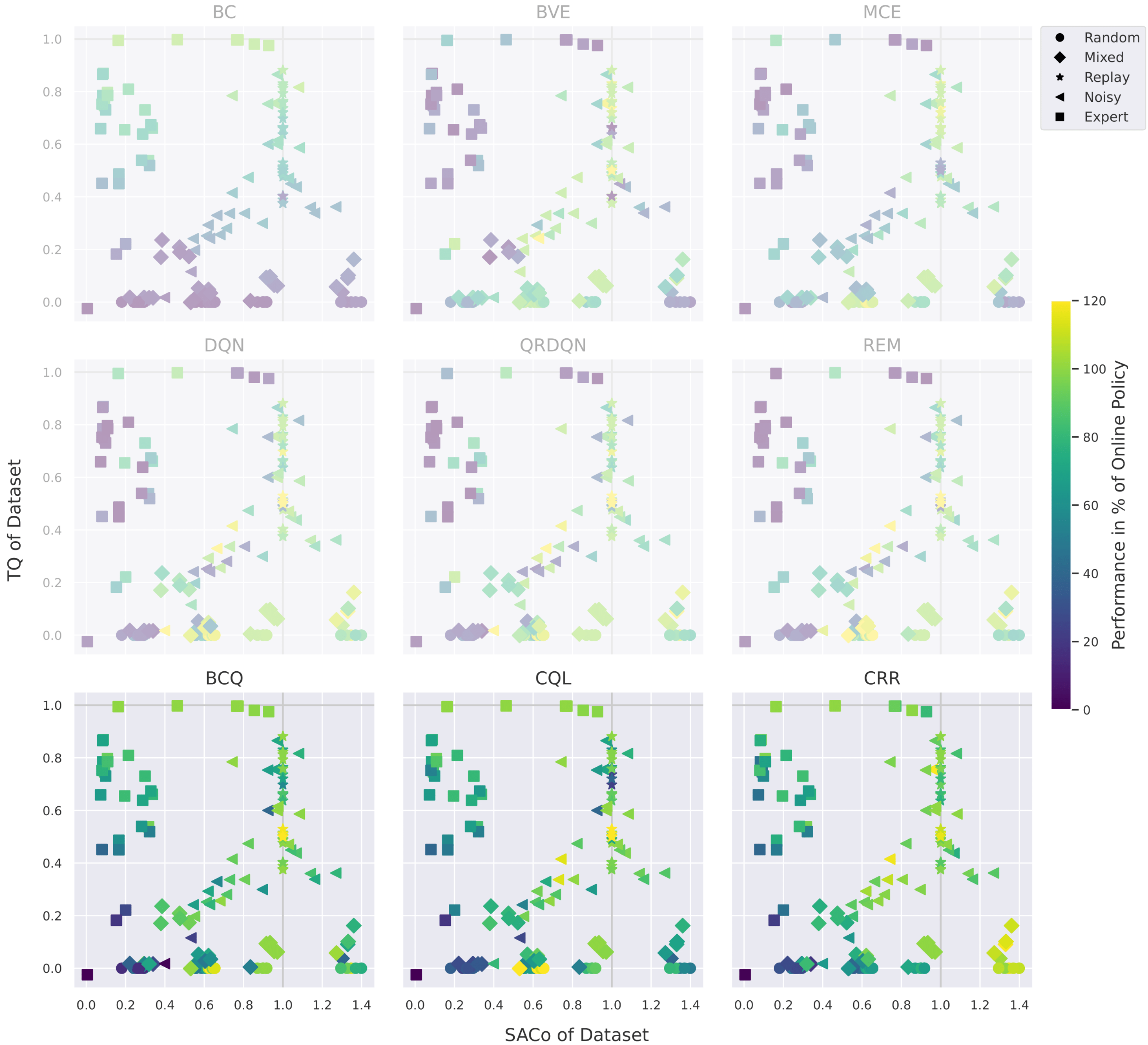
- Unconstrained off-policy Algorithms
- Best performance
  - High SACo
- Very similar results



Results

- Dataset-constrained off-policy Algorithms

- Best performance on
  - High TQ
  - High SACo
  - Moderate TQ & SACo



JKU dynatrace

Future Work

- Extend initial results on continuous state-action spaces
  - Continuous action-spaces require different set of algorithms
- Effects on model-based algorithms
- Different definitions of exploration
- Monitor replay buffer in off-policy algorithms

Summary

- Dataset composition matters a lot in Offline RL
- Comparing algorithms using average performance over multiple datasets might not be sufficient
- Capturing RL dataset characteristics through TQ and SACo

- Paper: https://arxiv.org/abs/2111.04714
- Code: https://github.com/ml-jku/OfflineRL
- Contact us: schweighofer@ml.jku.at