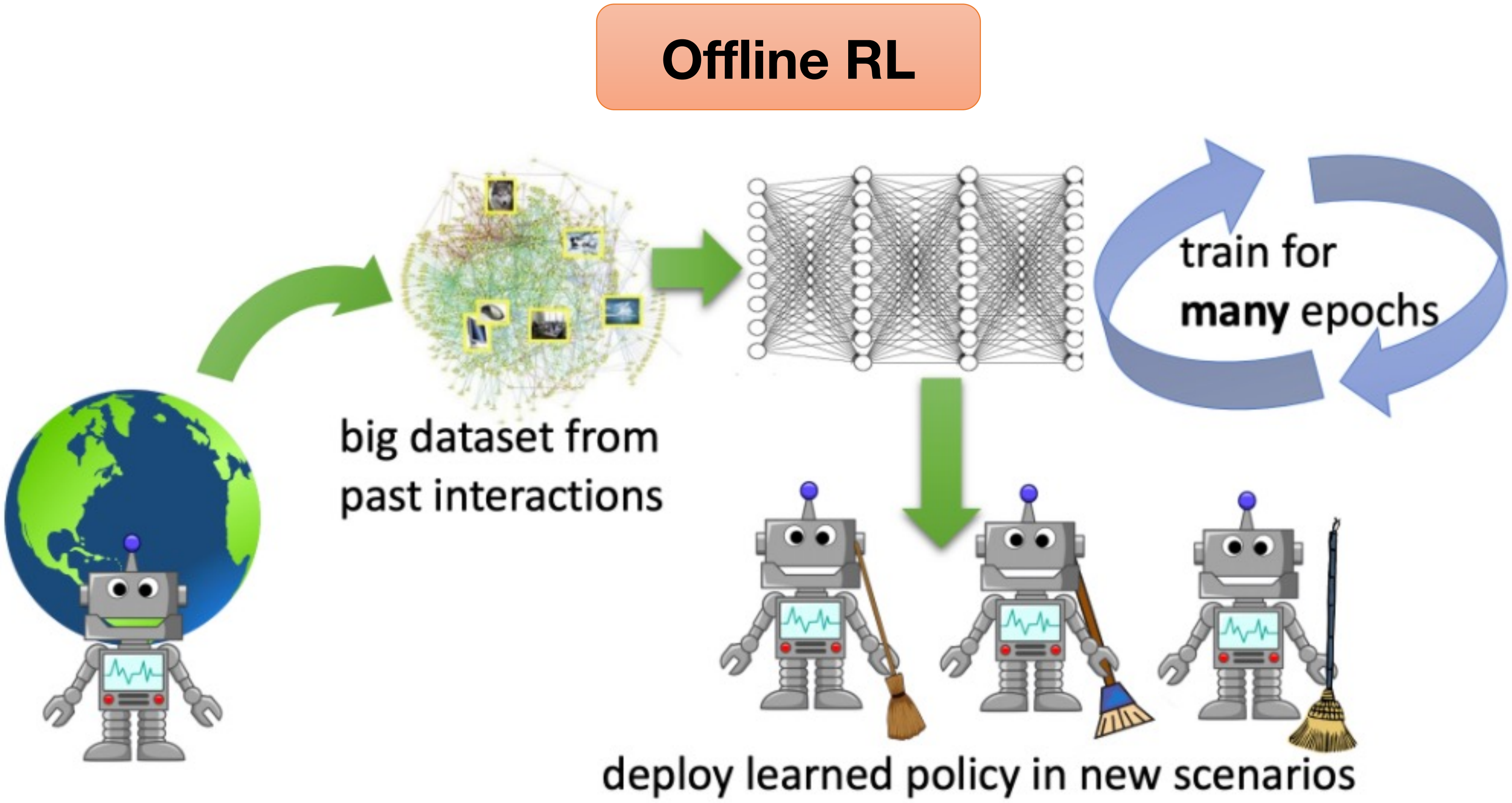
Offline Q-Learning on Diverse Multi-Task Data Both Scales and Generalizes

Aviral Kumar, Rishabh Agarwal, Young Geng,
George Tucker*, Sergey Levine*

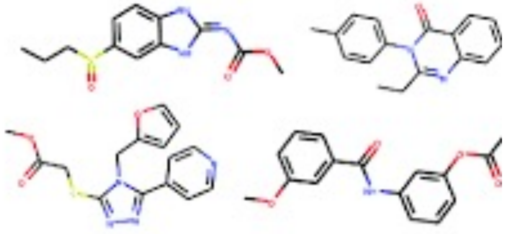
Berkeley
UNIVERSITY OF CALIFORNIA

Offline Reinforcement Learning

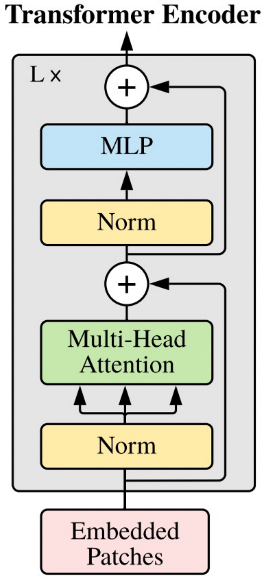


The Generalization Promise of Offline RL

Training on **large, pre-collected** datasets to attain **broad generalization**

Large datasets



Expressive function approximators

Broadly generalizing policies

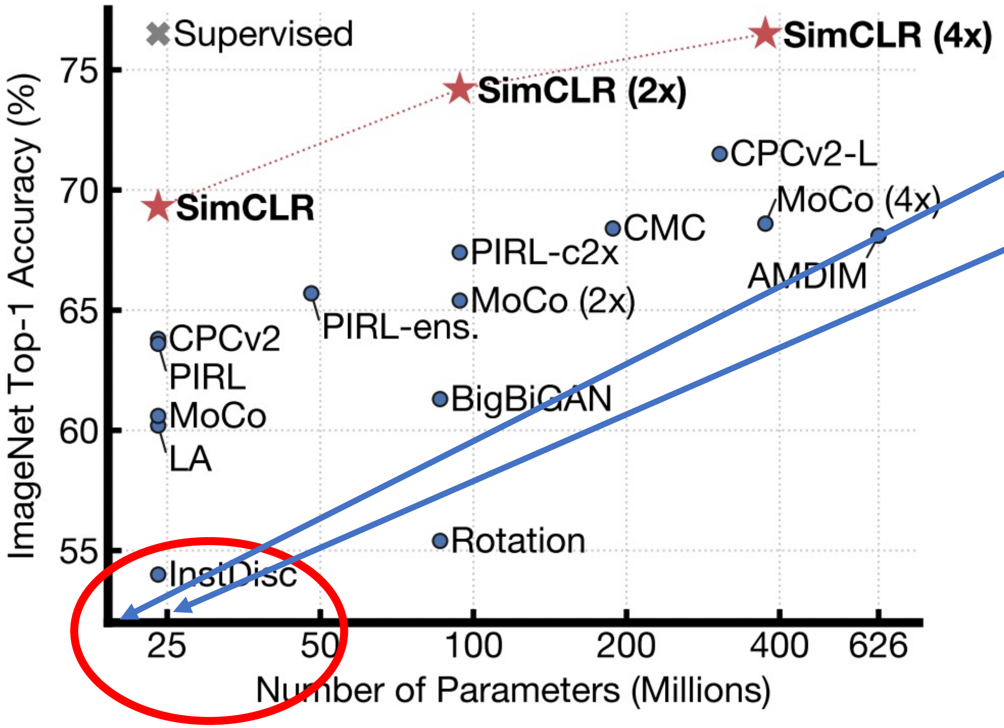
This Talk

What does it take to make
offline RL scale and generalize?

Where Are We At?

Disclaimer: This **definitely** misses some works (sorry!) but reflects the general trend.

Supervised learning



Reinforcement Learning

Gym / D4RL (3/4-layer feed-forward)	~1M	Offline RL
Mu Zero Reanalyze (16 res. blocks)	~25M	Offline RL
IMPALA architecture	~30M	Online RL
AlphaGo Zero (20 res. blocks)		High-profile successes
AlphaStar (~55M)		High-profile successes

Generally, much smaller models compared to supervised learning

Picture taken from the SimCLR paper

Large-Scale Study: Single Policy to Play Atari Games

Train a single policy on 40 Atari games

Evaluate on training games

Fine-tune to new games

Why Atari? Why is this problem challenging?

First large-scale test-bed to evaluate generalization and scaling

Requires **large** networks; offline Q-learning never worked

2 billion transitions, 40 games, **sub-optimal** data

Lee et al. **Multi-Game Decision Transformers.** NeurIPS 2022.

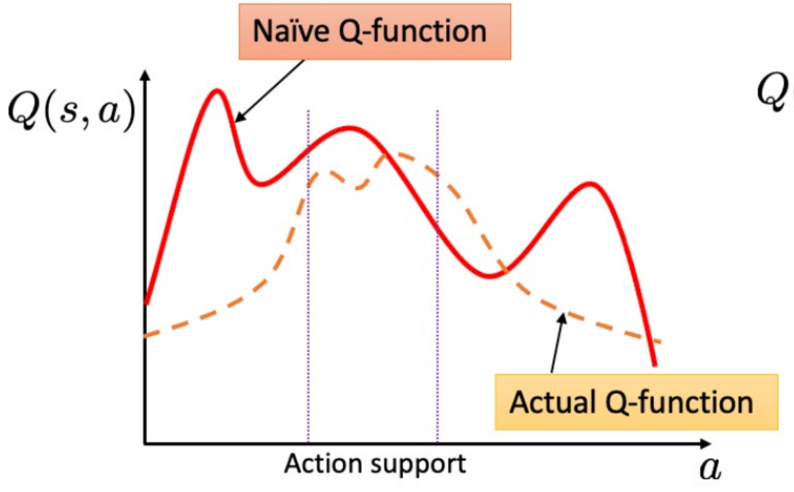
Three key ingredients
in our recipe…..

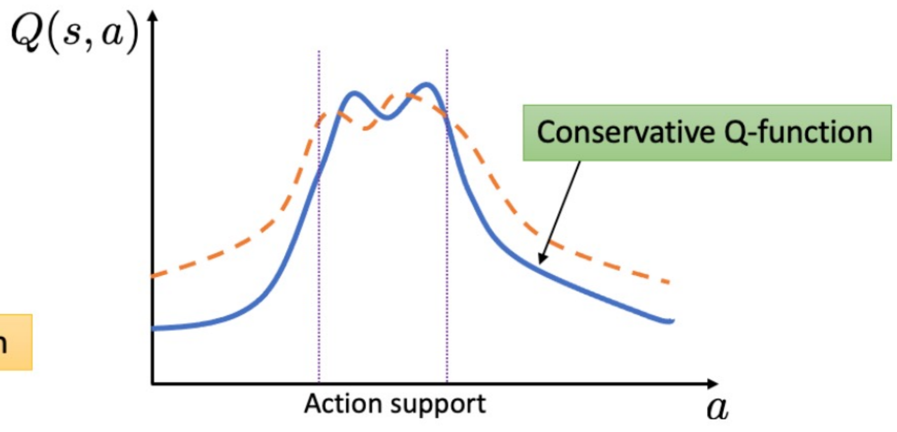
Ingredient 1: An offline RL Algorithm

Conservative Q-Learning (CQL)

$$\min_{\theta} \underbrace{\alpha \left(\mathbb{E}_{\mathbf{s}\sim\mathcal{D}} \left[\log \left(\sum_{\mathbf{a}'} \exp(Q_\theta(\mathbf{s}, \mathbf{a}')) \right) \right] - \mathbb{E}_{\mathbf{s},\mathbf{a}\sim\mathcal{D}} \left[Q_\theta(\mathbf{s}, \mathbf{a}) \right] \right)} + \underline{\text{TDError}(\theta; \mathcal{D})}$$

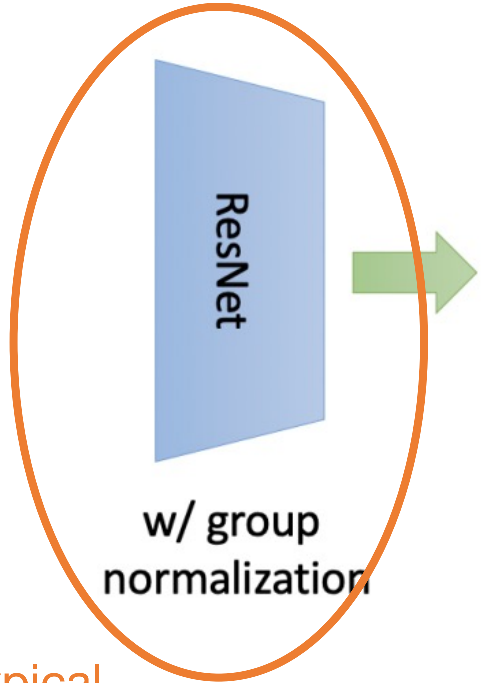
Encourages the Q-function to not over-estimate





K., Zhou, Tucker, Levine. **Conservative Q-Learning for Offline Reinforcement Learning.** NeurIPS 2020.

Ingredient 2: Large Networks

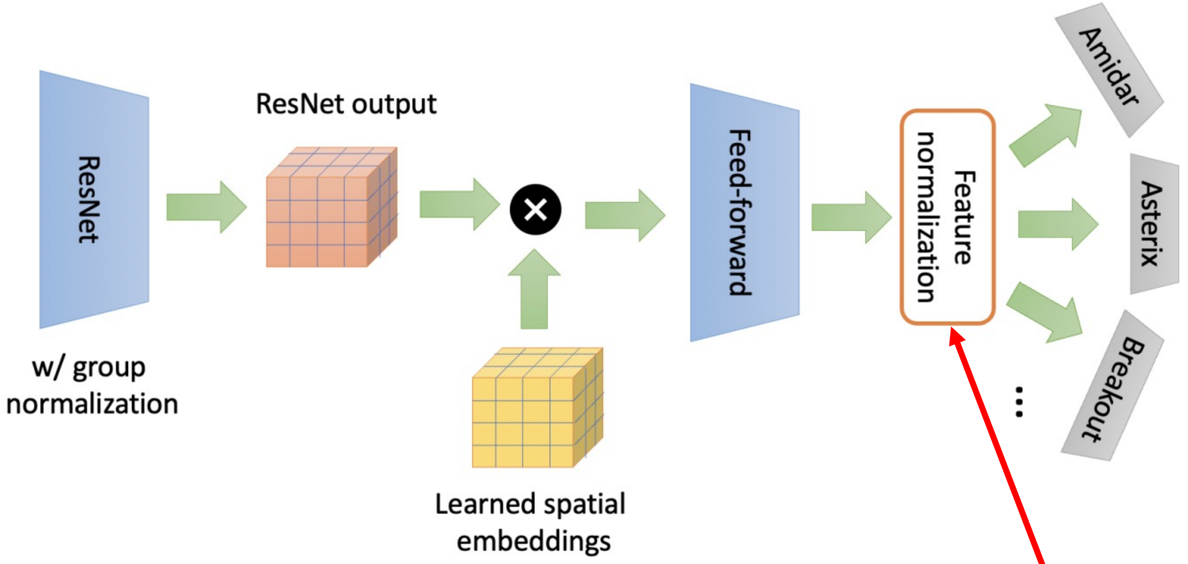


Not the typical
IMPALA architecture!

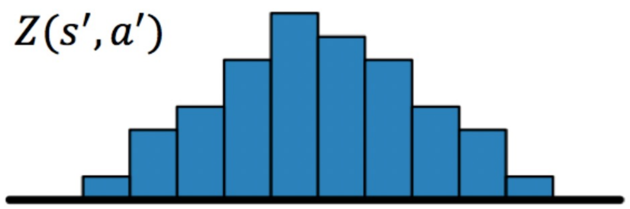
Keep track of
spatial information in the image!

Multi-headed Q-function

Ingredient 3: Methods to Effectively Use Capacity



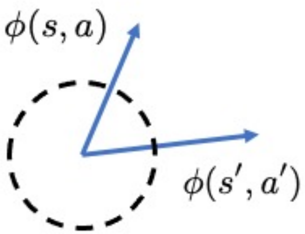
C51 (Q-value = categorical distribution)



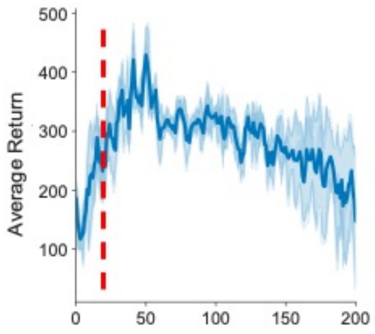
Can train via cross-entropy loss

DR3 normalization: Normalize features to have norm = 1

$$\text{Normalize}(\phi(\mathbf{s}, \mathbf{a})) = \frac{\phi(\mathbf{s}, \mathbf{a})}{||\phi(\mathbf{s}, \mathbf{a})||_2}$$



Iteration 20



K., Agarwal, Ma, Courville, Tucker, Levine. **DR3: Value-Based Deep RL Requires Explicit Regularization.** ICLR 2022

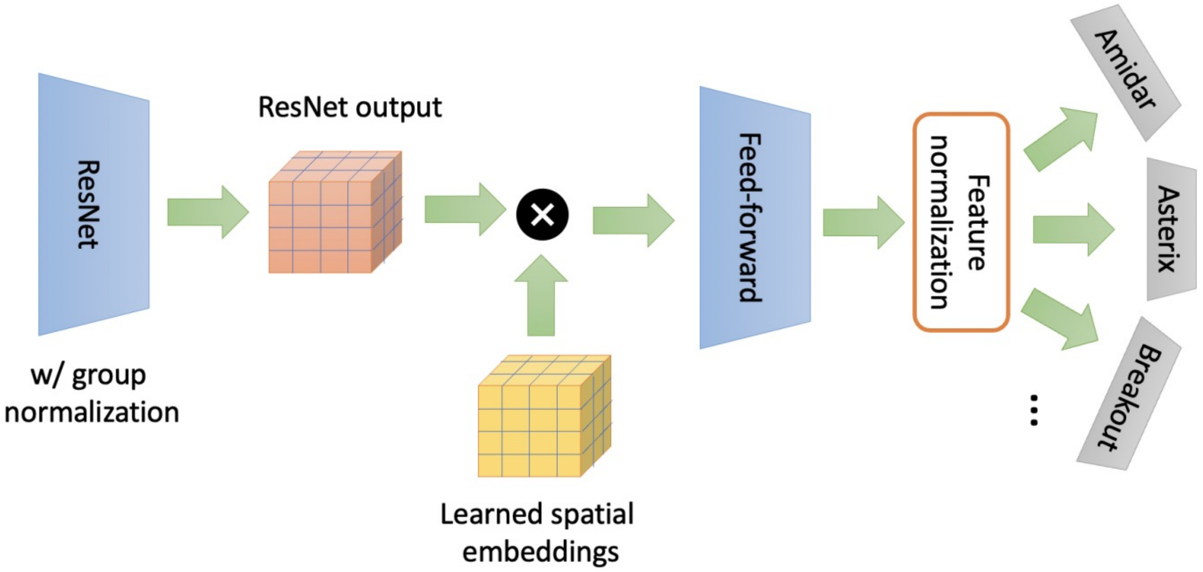
Dabney et al. **The Value Improvement Path: Towards Better Representations for Reinforcement Learning.** AAAI 2021.

Summary: "Scaled Q-Learning"

An offline RL algorithm

Conservative Q-learning (CQL)

A large neural network

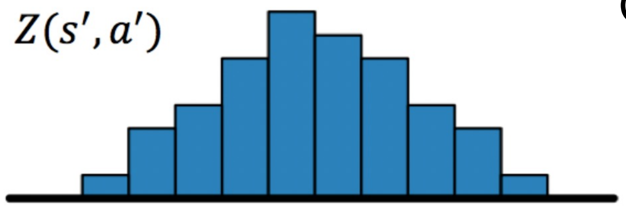


A way to effectively use capacity

DR3 normalization: Normalize features to have norm = 1

$$\text{Normalize}(\phi(\mathbf{s}, \mathbf{a})) = \frac{\phi(\mathbf{s}, \mathbf{a})}{||\phi(\mathbf{s}, \mathbf{a})||_2}$$

Q-value discretization (distributional RL)

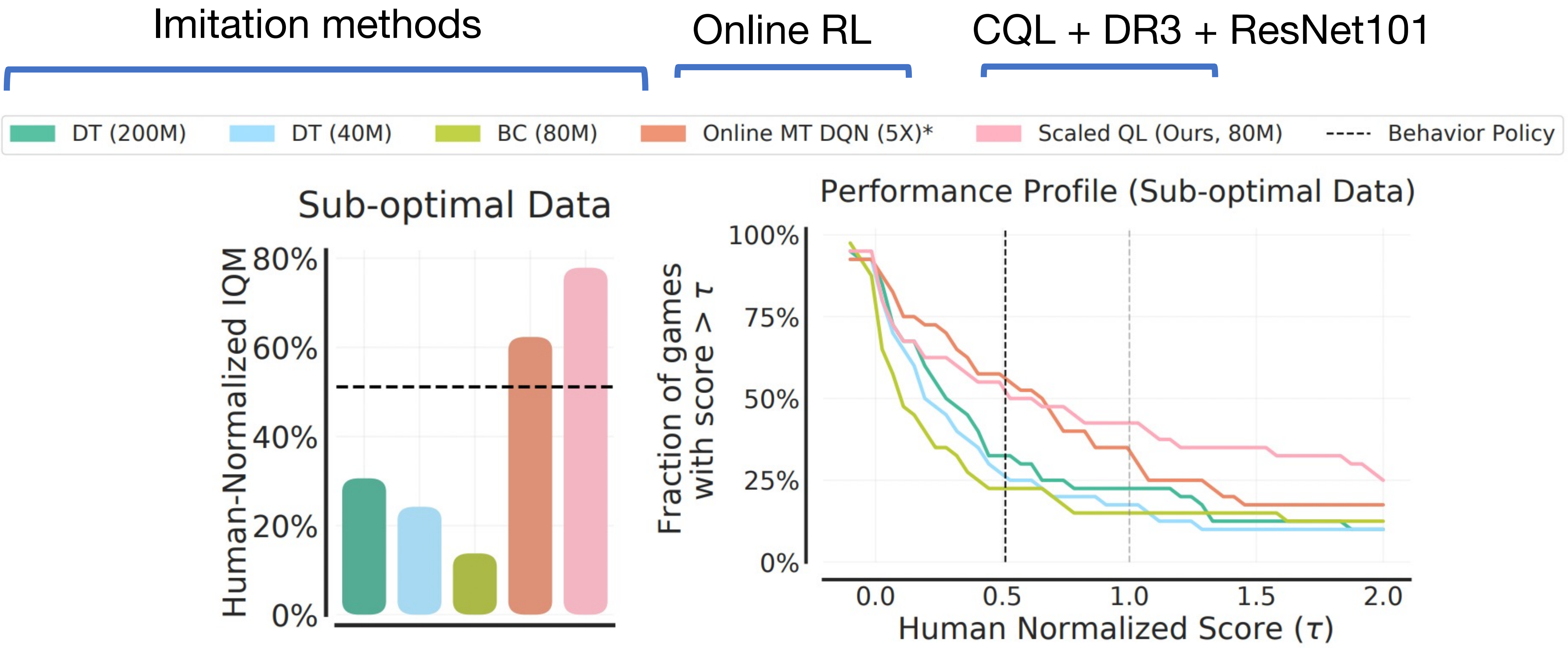


Q-value = categorical distribution

Cross-entropy loss

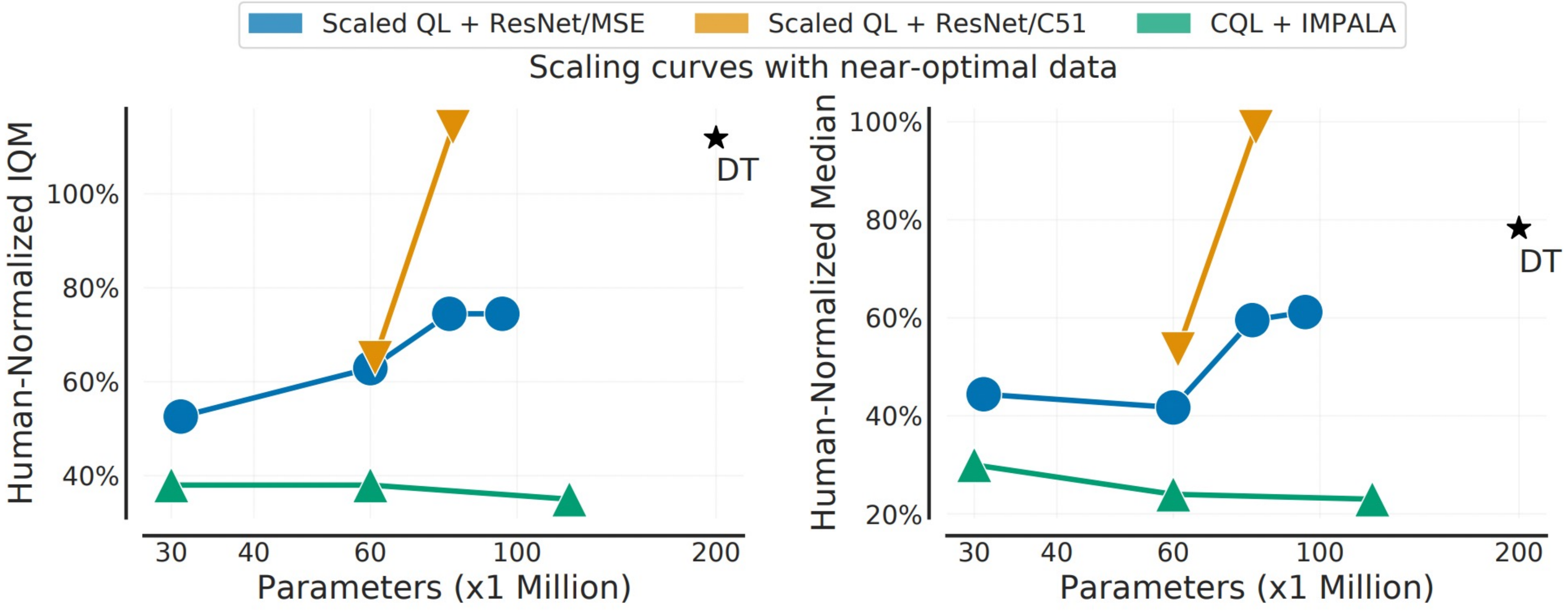
Performance of Scaled Q-Learning

We find: ~80M ResNet + sub-optimal data => better than online RL or decision transformers



Scaling Trends for Scaled Q-Learning

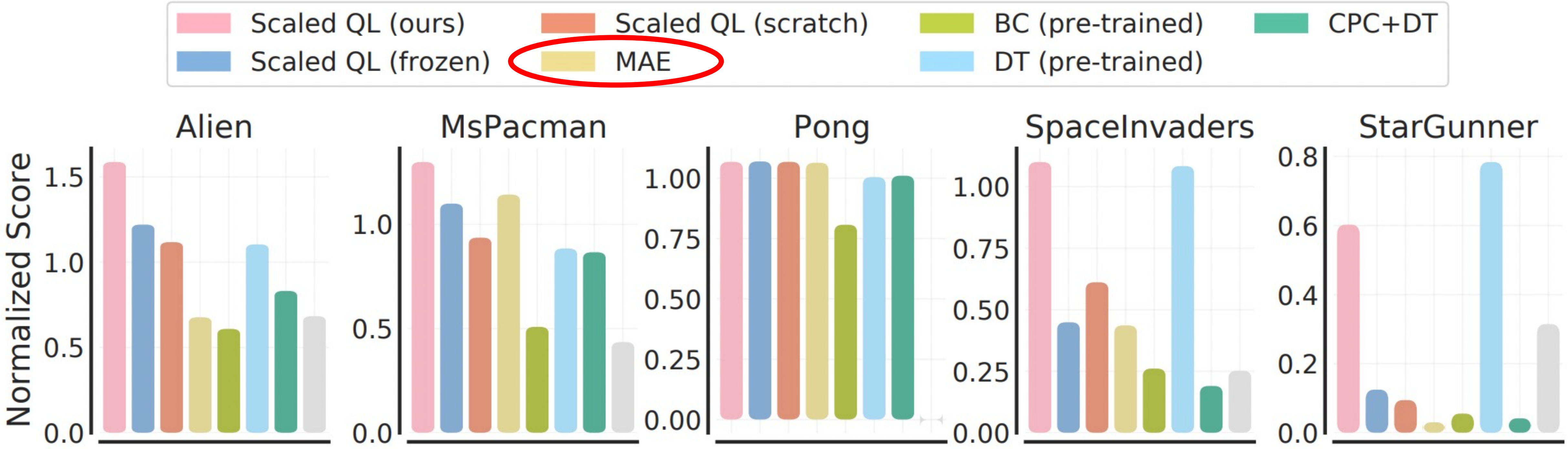
Scaled Q-learning does scale favorably, while other naïve choices (IMPALA) do not



Generalization: Offline Fine-Tuning to New Games

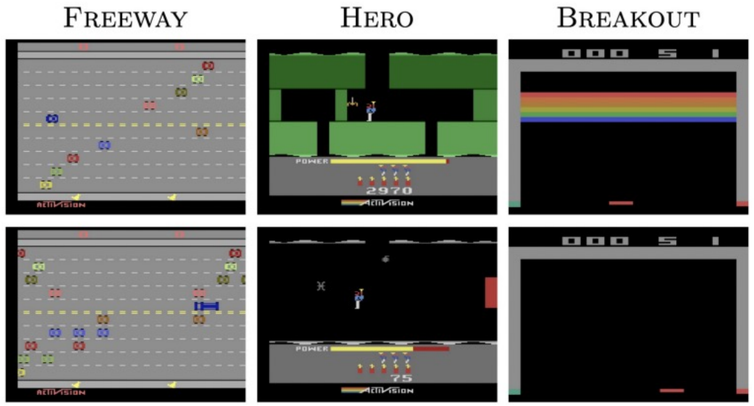
Limited offline data for a new game + pre-trained model on the training games

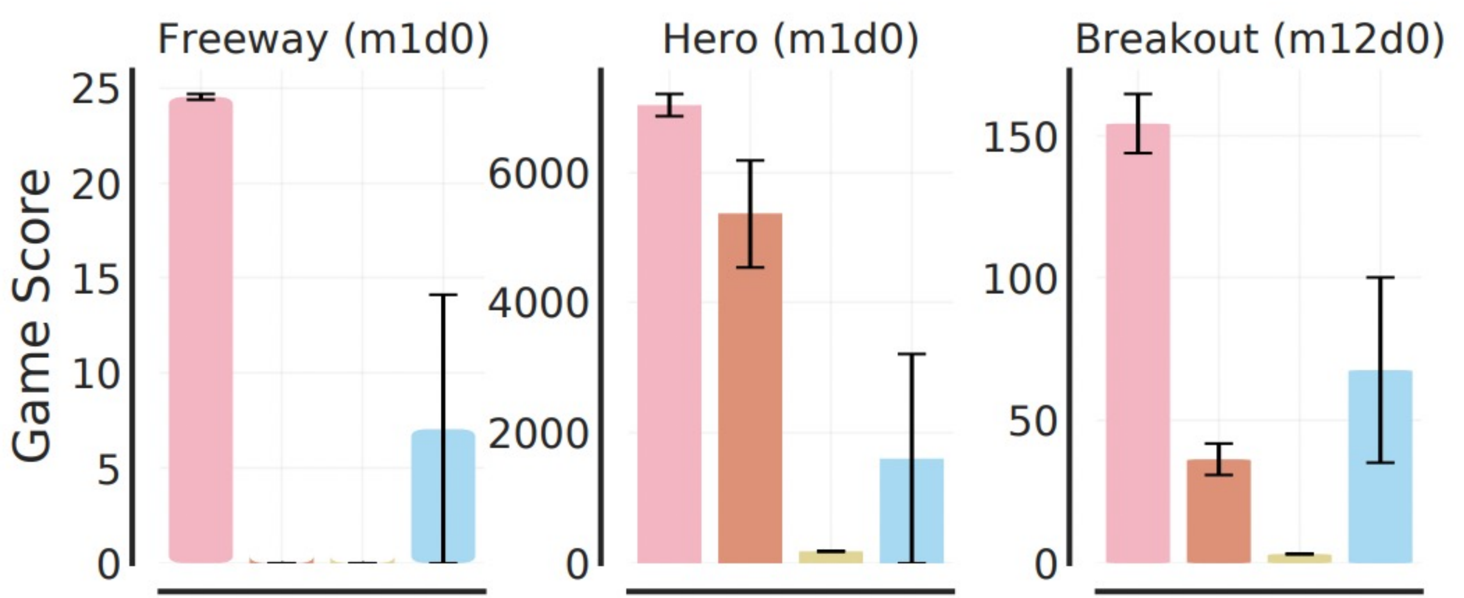
82% improvement



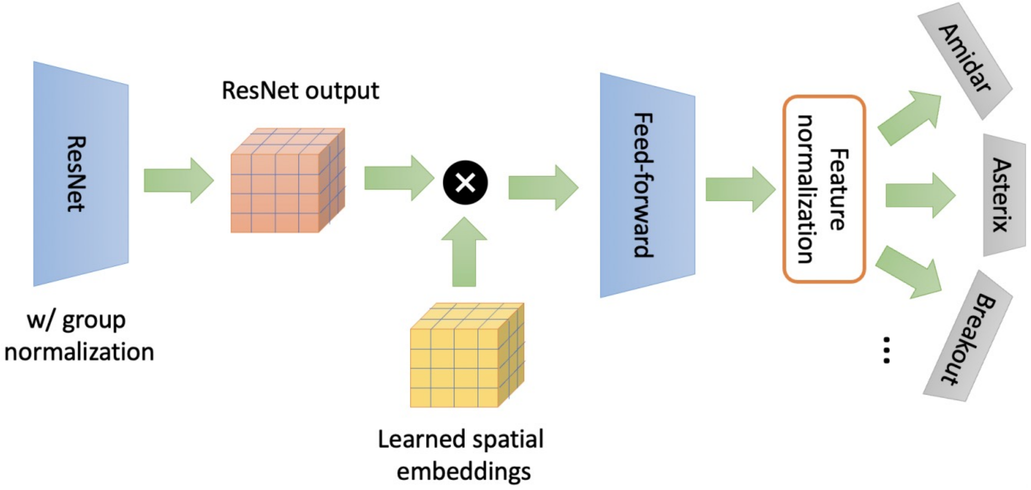
Generalization: Online Fine-Tuning to New Modes

Scaled Q-Learning learns representations useful under changes to the environment





DR3 Enables Effective Use of Capacity



	Scaled QL (ResNet 34)	Scaled QL (ResNet 50)	Scaled QL (ResNet 101)
without feature normalization	50.9%	73.9%	80.4%
with feature normalization	78.0% (+28.9%)	83.5% (+9.6%)	98.0% (+17.6%)

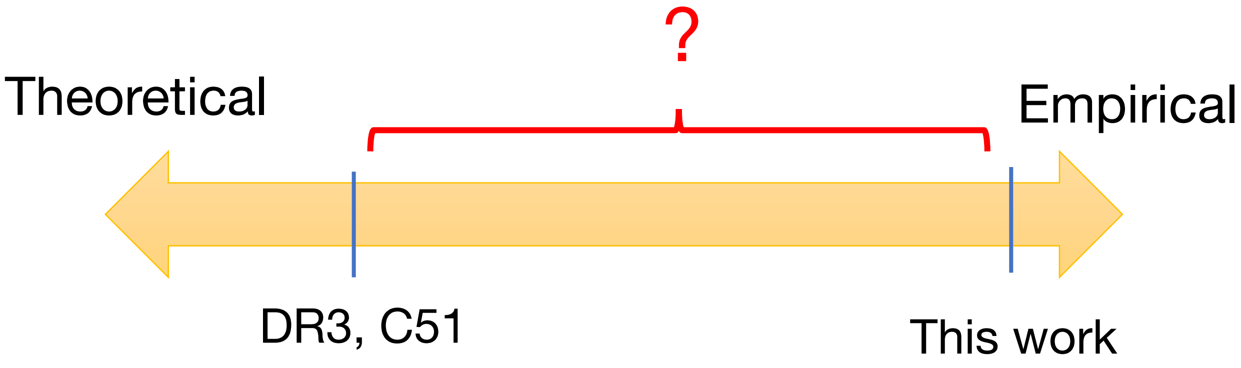
DR3 normalization

DR3 improves consistently along the way

Enables the use of higher capacity more effectively

Summary and Takeaways

- We present a simple way to **scale** Q-learning to large datasets + large models
- Models pre-trained via offline Q-learning learn **generalizing** representations
- Effectively leveraging capacity of large networks seem critical!
- **Preliminary Code:** https://tinyurl.com/scaled-ql-code



Thank You!

Research at Google