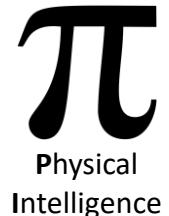


# Reinforcement Learning with Prior Data

**Sergey Levine**  
**UC Berkeley**  
**Physical Intelligence**

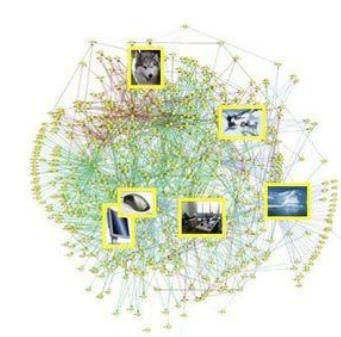
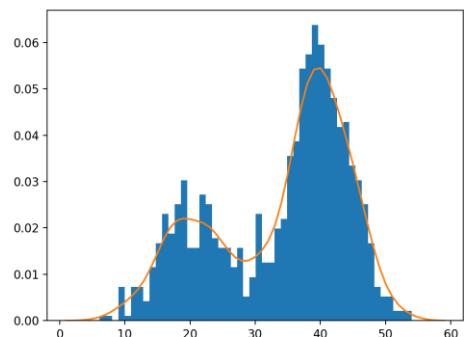


# Modern data-driven AI



$$p_{\theta}(\mathbf{x})$$

$$p_{\theta}(\mathbf{y}|\mathbf{x})$$



## Explaining a joke

### Prompt

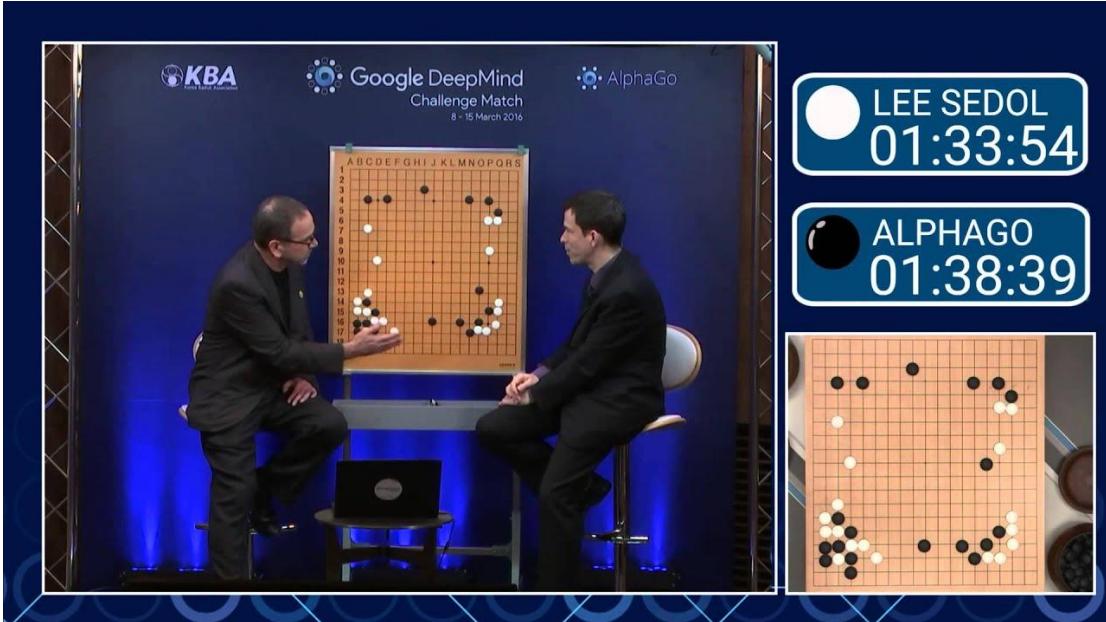
Explain this joke:

Joke: Did you see that Google just hired an eloquent whale for their TPU team? It showed them how to communicate between two different pods!

### Model Response

**Prediction:** TPUs are a type of computer chip that Google uses for deep learning. A "pod" is a group of TPUs. A "pod" is also a group of whales. The joke is that the whale is able to communicate between two groups of whales, but the speaker is pretending that the whale is able to communicate between two groups of TPUs.

Impressive because no person had thought of it!



“Move 37” in Lee Sedol AlphaGo match: reinforcement learning “discovers” a move that surprises everyone

Impressive because it looks like something a person might draw!



# So where does that leave us?

## Data-Driven AI



## Reinforcement Learning



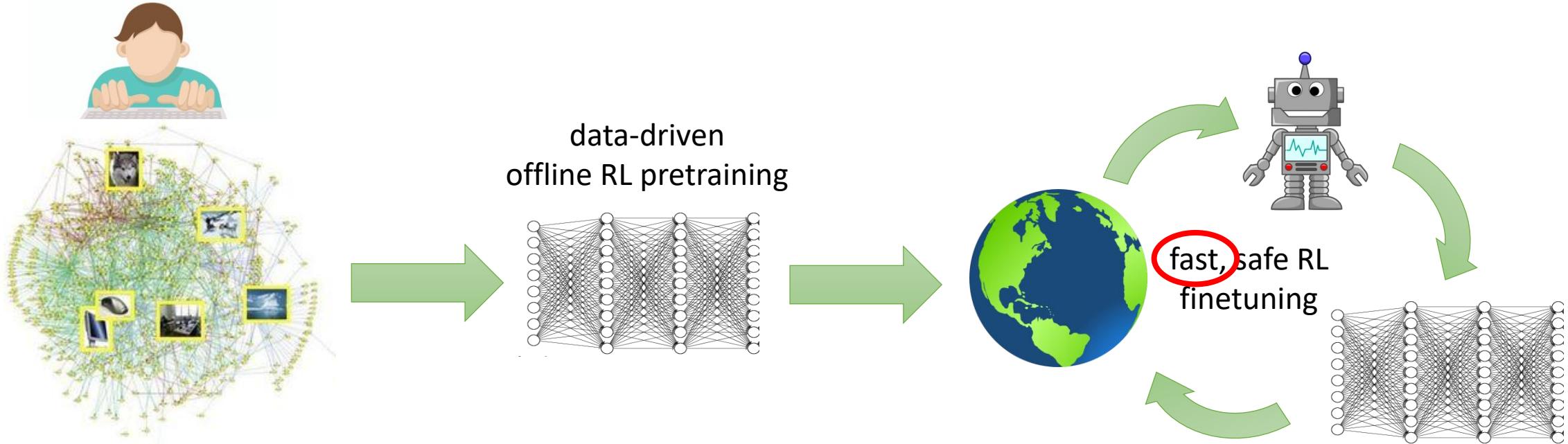
- + learns about the real world from data
- doesn't try to do **better** than the data

- + optimizes a goal with emergent behavior
- doesn't make use of real-world data

**Data without optimization**  
doesn't allow us to solve new problems in new ways

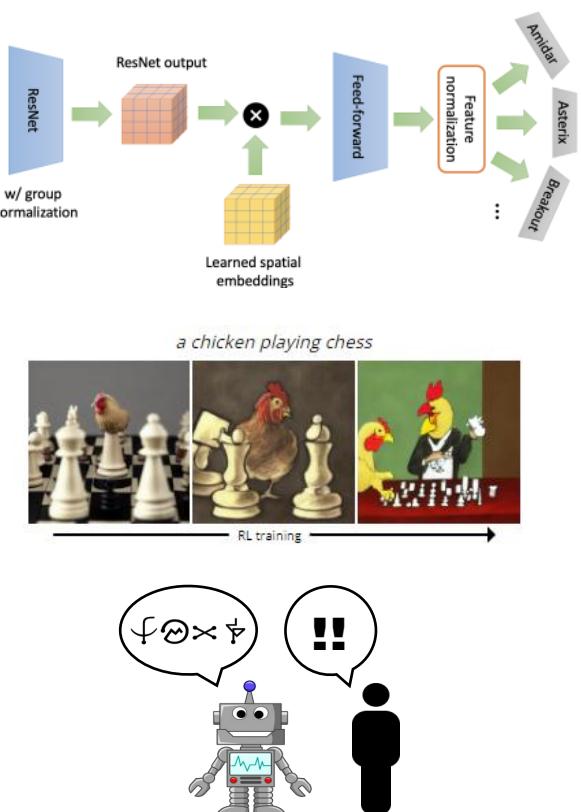
**Optimization without data** is hard to apply to the real world outside of simulators

# The recipe



This talk

What can we **accomplish** if we  
combine **data** and **optimization**?



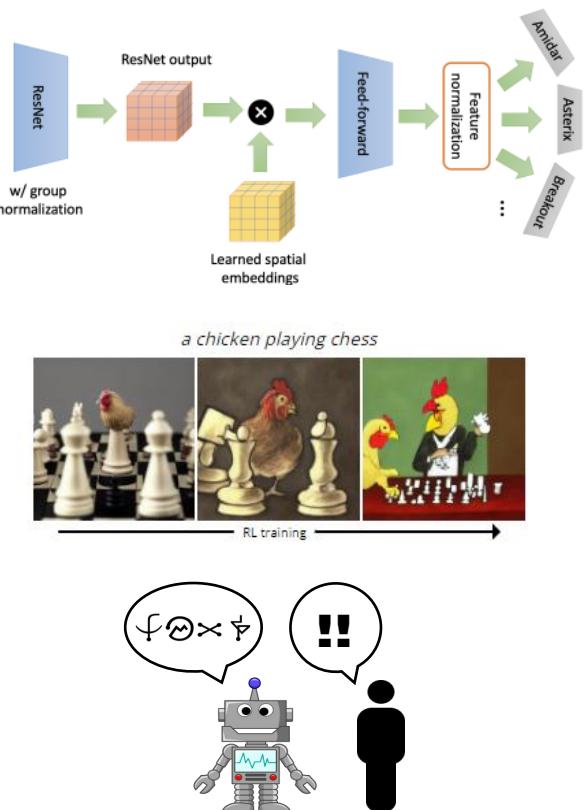
Data-driven RL algorithms

RL with generative models

Offline RL and language models

This talk

What can we **accomplish** if we  
combine **data** and **optimization**?



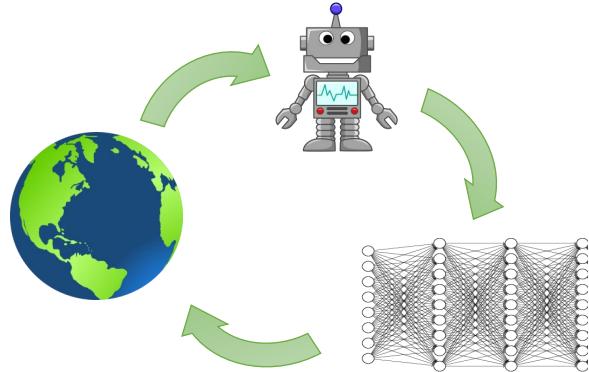
Data-driven RL algorithms

RL with generative models

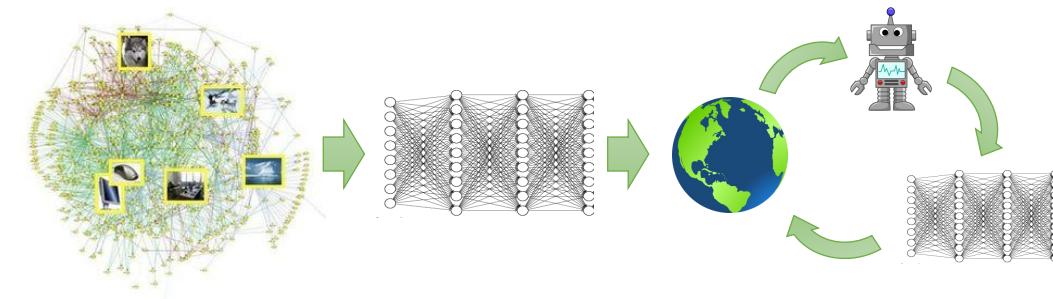
Offline RL and language models

# What do we need to figure out?

Classic RL



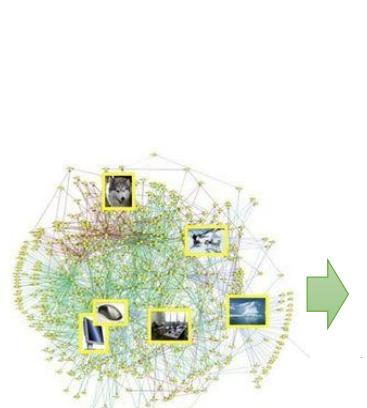
Data-driven RL



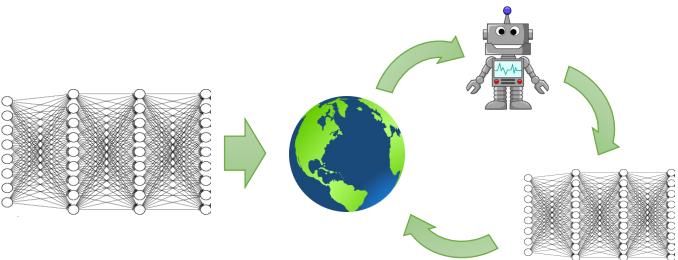
- Online, lifelong learning process
- Starts from scratch
- Largely **trial and error** driven
- **Central problems:**
  - Exploration
  - Sample efficiency
  - Optimization performance

- Offline pretraining + online finetuning
- Always start from data
- Largely **representation learning** driven
- **Central problems:**
  - Distributional shift
  - Scalability and stability
  - Representation learning with big models

# To break this down...



Data-driven RL



- Offline pretraining + online finetuning
- Always start from data
- Largely **representation learning** driven
- **Central problems:**
  - Distributional shift
  - Scalability and stability
  - Representation learning with big models

1. Offline RL algorithms

We understand  
this pretty well

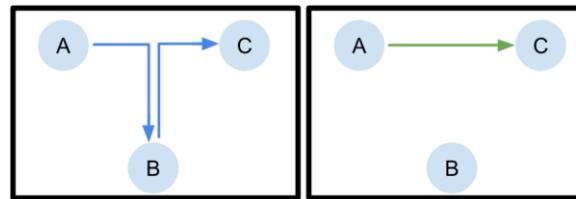
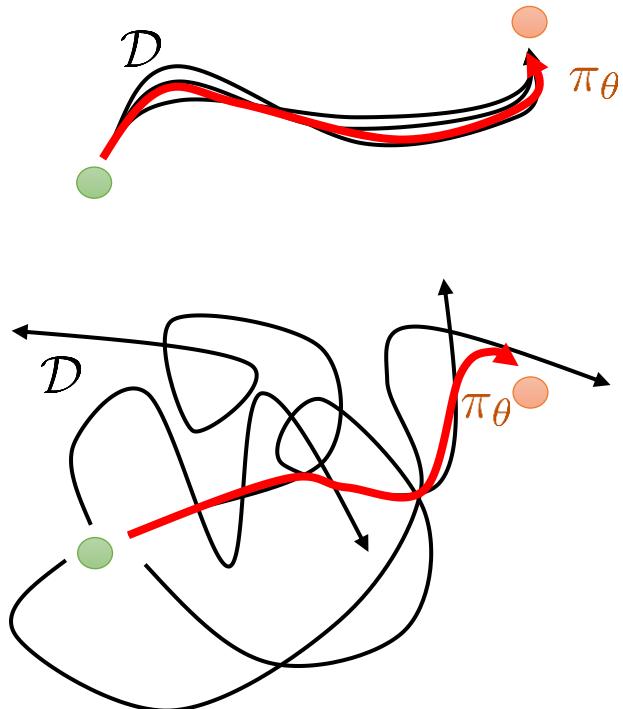
2. Online finetuning from  
offline initializations

We understand  
this a little

3. Making all this work  
with big, scalable models

We hardly  
understand this  
at all

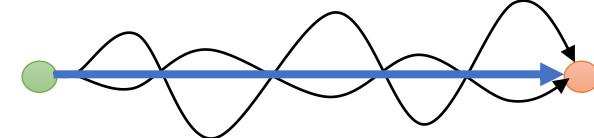
# What do we expect offline RL methods to do?



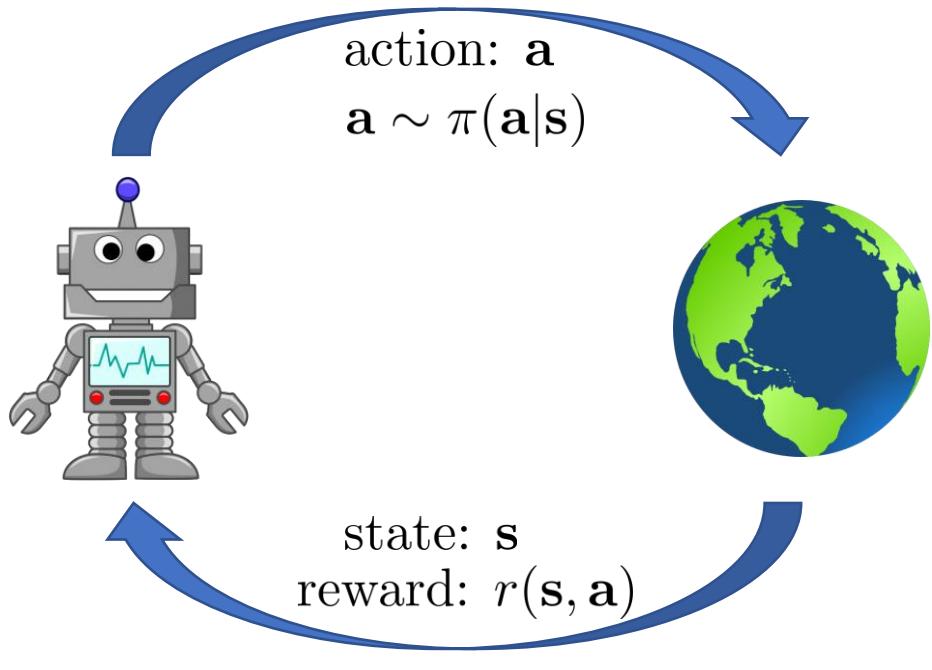
“Macro-scale” stitching

**But this is just the clearest example!**

“Micro-scale” stitching:



# Off-policy RL: a quick primer



$$\text{RL objective: } \max_{\pi} \sum_{t=1}^T E_{\mathbf{s}_t, \mathbf{a}_t \sim \pi} [r(\mathbf{s}_t, \mathbf{a}_t)]$$
$$\text{Q-function: } Q^\pi(\mathbf{s}_t, \mathbf{a}_t) = \sum_{t'=t}^T E_{\mathbf{s}_{t'}, \mathbf{a}_{t'} \sim \pi} [r(\mathbf{s}_{t'}, \mathbf{a}_{t'}) | \mathbf{s}_t, \mathbf{a}_t]$$

$$\pi(\mathbf{a}|\mathbf{s}) = 1 \text{ if } \mathbf{a} = \arg \max_{\mathbf{a}} Q^\pi(\mathbf{s}, \mathbf{a})$$

$$Q^*(\mathbf{s}, \mathbf{a}) = r(\mathbf{s}, \mathbf{a}) + \max_{\mathbf{a}'} Q^*(\mathbf{s}', \mathbf{a}')$$



enforce this equation at all states!

$$\text{minimize } \sum_i (Q(\mathbf{s}_i, \mathbf{a}_i) - [r(\mathbf{s}_i, \mathbf{a}_i) + \max_{\mathbf{a}'_i} Q(\mathbf{s}'_i, \mathbf{a}'_i)])^2$$

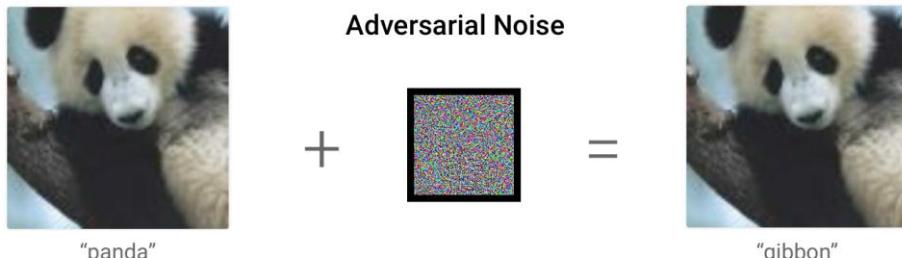
# Some principles for offline RL

$$\cancel{Q(s, a) \leftarrow r(s, a) + \max_{a'} Q(s', a')}$$

$$Q(\mathbf{s}, \mathbf{a}) \leftarrow r(\mathbf{s}, \mathbf{a}) + E_{\mathbf{a}' \sim \pi_{\text{new}}} [Q(\mathbf{s}', \mathbf{a}')] \\ y(\mathbf{s}, \mathbf{a})$$

expect good accuracy when  $\pi_\beta(\mathbf{a}|\mathbf{s}) = \pi_{\text{new}}(\mathbf{a}|\mathbf{s})$

even worse:  $\pi_{\text{new}} = \arg \max_{\pi} E_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})}[Q(\mathbf{s}, \mathbf{a})]$

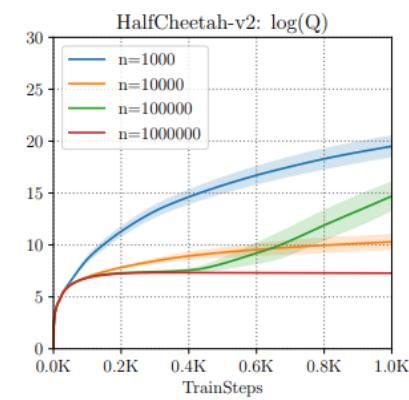
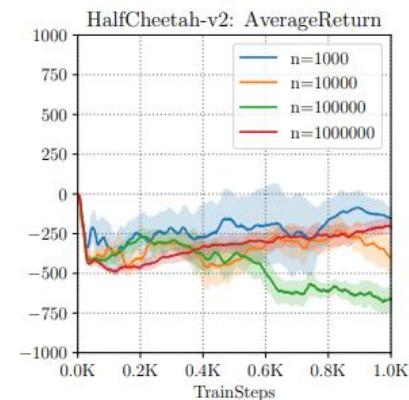


what is the objective?

$$\min_Q E_{(\mathbf{s}, \mathbf{a}) \sim \pi_\beta(\mathbf{s}, \mathbf{a})} [(Q(\mathbf{s}, \mathbf{a}) - y(\mathbf{s}, \mathbf{a}))^2]$$



how often does *that* happen?



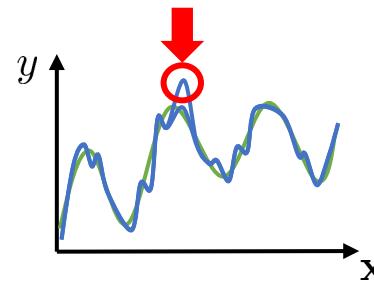
how well it does

how well it *thinks*  
it does (Q-values)

# Some principles for offline RL

- Many different methods, similar principles seem to be effective:

use value-based methods (i.e., Q-learning or Q-function actor-critic)

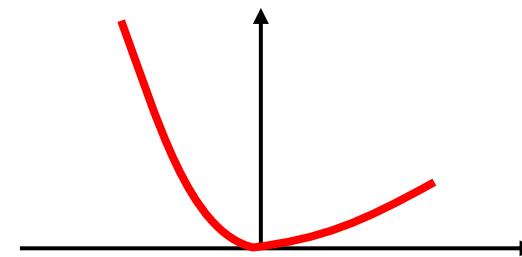


pessimism (e.g., CQL)

$$D_{\text{KL}}(\pi(a|s) \| \pi_\beta(a|s)) \leq \epsilon$$

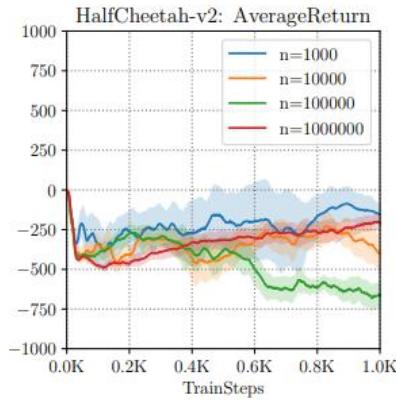
policy constraints (e.g., BRAC)

somehow fix the distributional shift problem

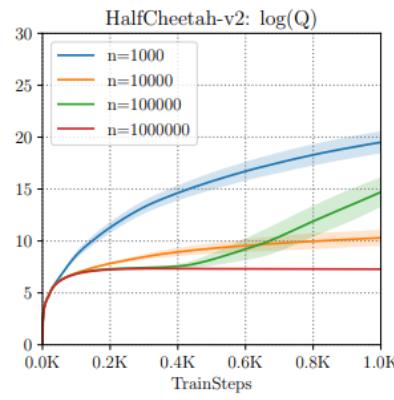


avoid OOD actions in updates  
(e.g., AWAC, IQL)

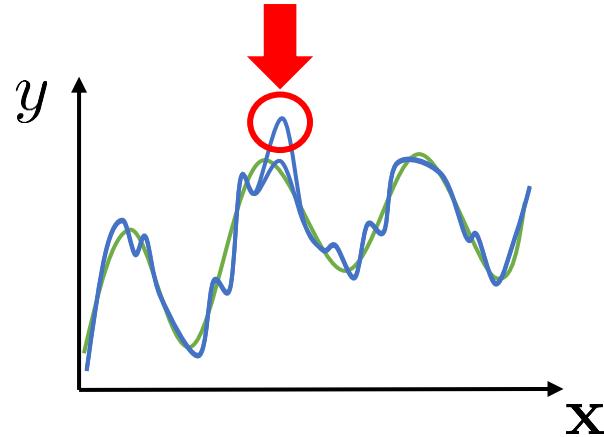
# Conservative Q-learning (CQL)



how well it does



how well it *thinks*  
it does (Q-values)



$$\hat{Q}^\pi = \arg \min_Q \max_\pi \alpha E_{\mathbf{s} \sim D, \mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} [Q(\mathbf{s}, \mathbf{a})] - \alpha E_{\mathbf{s}, \mathbf{a} \sim D} [Q(\mathbf{s}, \mathbf{a})] \quad ] \quad \text{term to push down big Q-values}$$

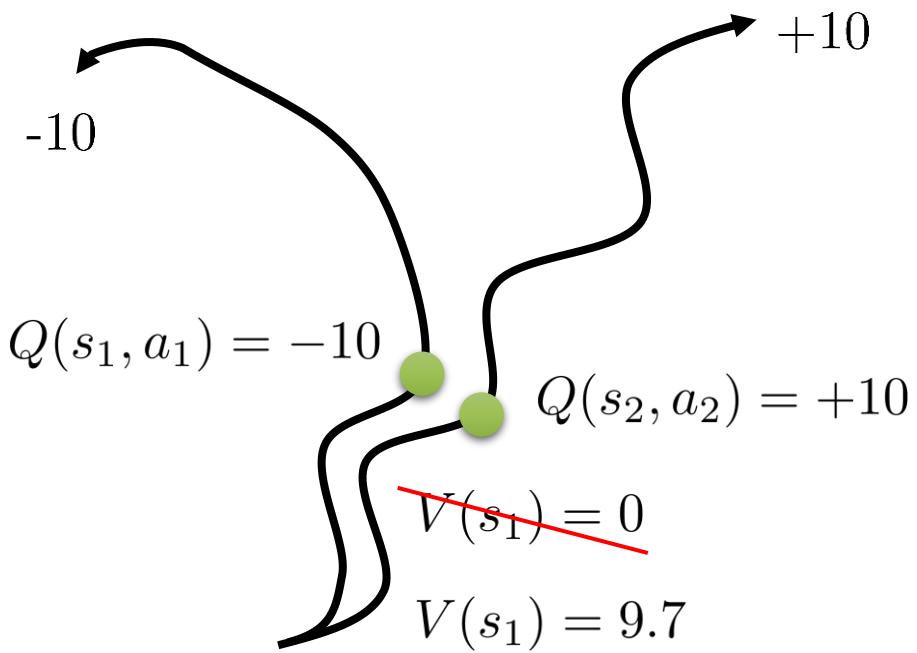
$$\text{regular objective } \left[ + E_{(\mathbf{s}, \mathbf{a}, \mathbf{s}') \sim D} \left[ (Q(\mathbf{s}, \mathbf{a}) - (r(\mathbf{s}, \mathbf{a}) + E_\pi [Q(\mathbf{s}', \mathbf{a}')]))^2 \right] \right]$$

can show that  $\hat{Q}^\pi \leq Q^\pi$  for large enough  $\alpha$

↑  
true Q-function

# Implicit Q-learning (IQL)

What if we didn't need to **ever** evaluate any action that is not in the dataset?

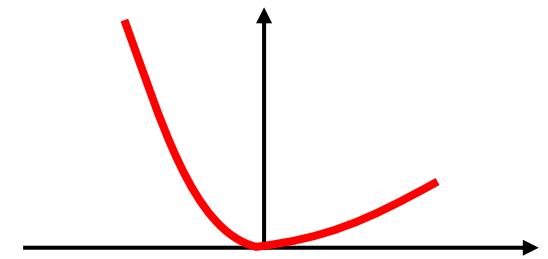
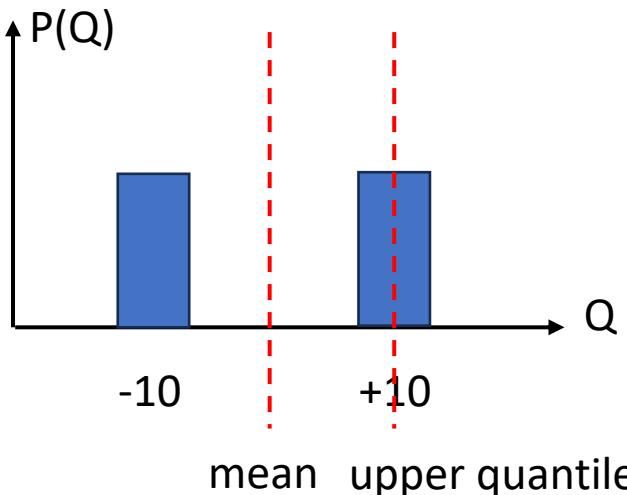


$$Q = \min_Q \sum_{(s,a,s') \in \mathcal{D}} \|Q(s, a) - [r(s, a) + \gamma V(s')]\|^2$$

$$V = \min_V \sum_{(s,a) \in \mathcal{D}} \|V(s) - Q(s, a)\|^2$$

"SARSA" (sort of)

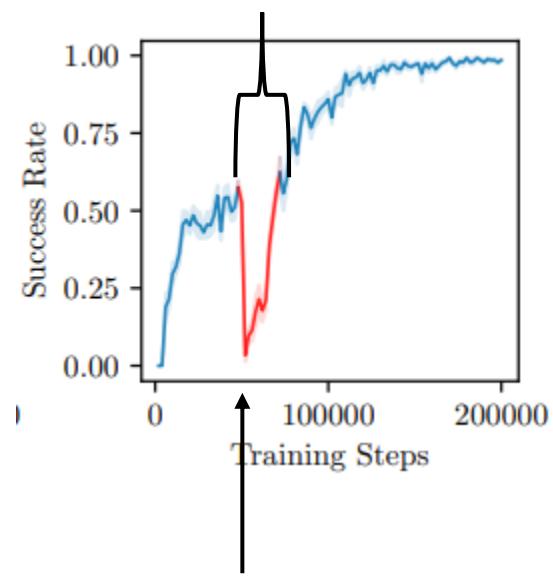
$$V = \min_V \sum_{(s,a) \in \mathcal{D}} \ell_\tau(V(s) - Q(s, a))$$



# What about online finetuning?

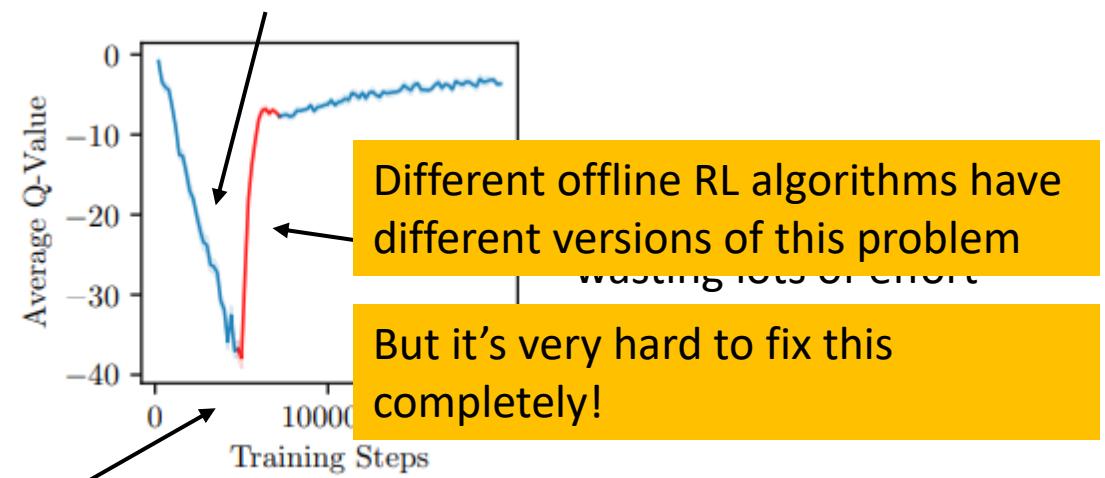
$$\begin{aligned}\hat{Q}^\pi = \arg \min_Q \max_\pi \alpha E_{\mathbf{s} \sim D, \mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} [Q(\mathbf{s}, \mathbf{a})] - \alpha E_{\mathbf{s}, \mathbf{a} \sim D} [Q(\mathbf{s}, \mathbf{a})] \\ + E_{(\mathbf{s}, \mathbf{a}, \mathbf{s}') \sim D} \left[ (Q(\mathbf{s}, \mathbf{a}) - (r(\mathbf{s}, \mathbf{a}) + E_\pi [Q(\mathbf{s}', \mathbf{a}')])))^2 \right]\end{aligned}$$

this period wasted recovering  
offline performance



online training starts here (at 50k steps)

underestimating too much here

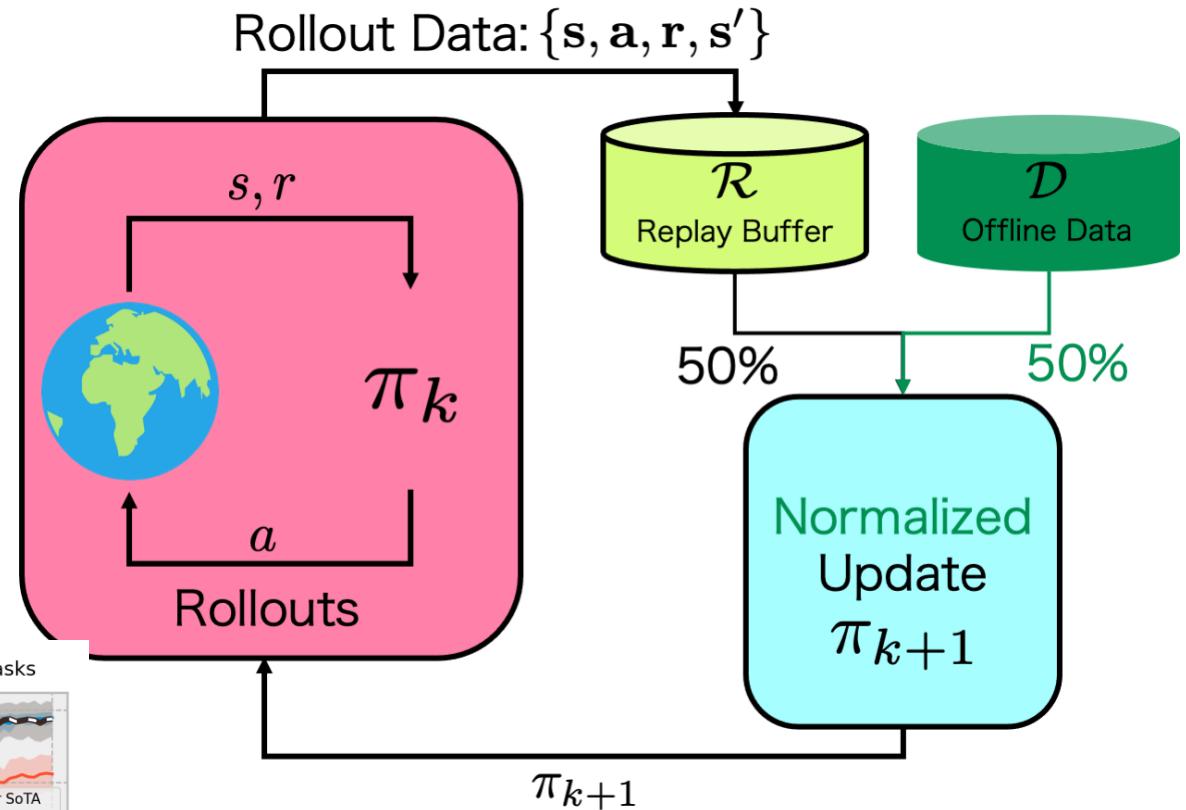
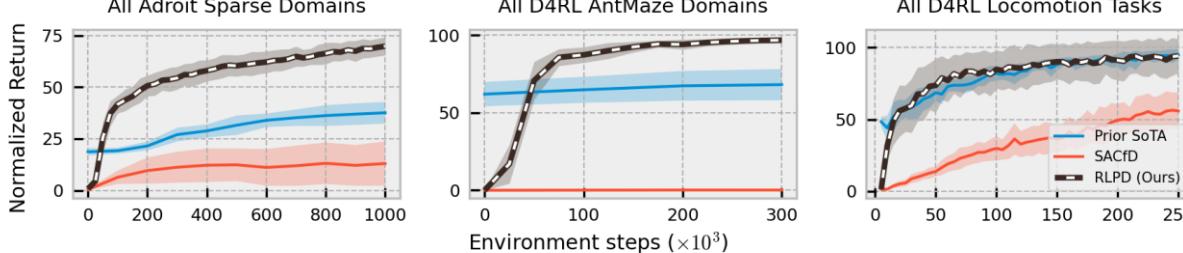


# An embarrassingly strong baseline: RLPD

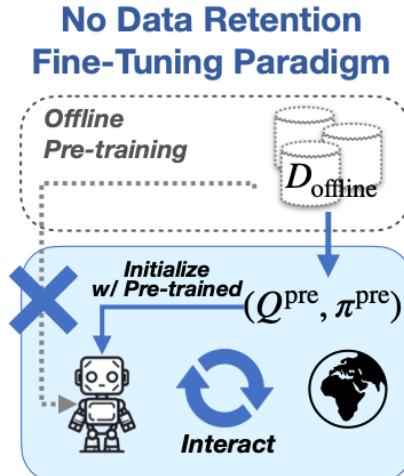
1. Initialize two buffers: an online replay buffer and an offline data buffer
2. Initialize the value function (and actor) from **scratch** (no pretraining!)
3. Run online RL, and for every batch, sample half of it from the offline data, and half of it from the online data replay buffer

This is deeply unsatisfying

This is really hard to beat with any “proper” offline to online method



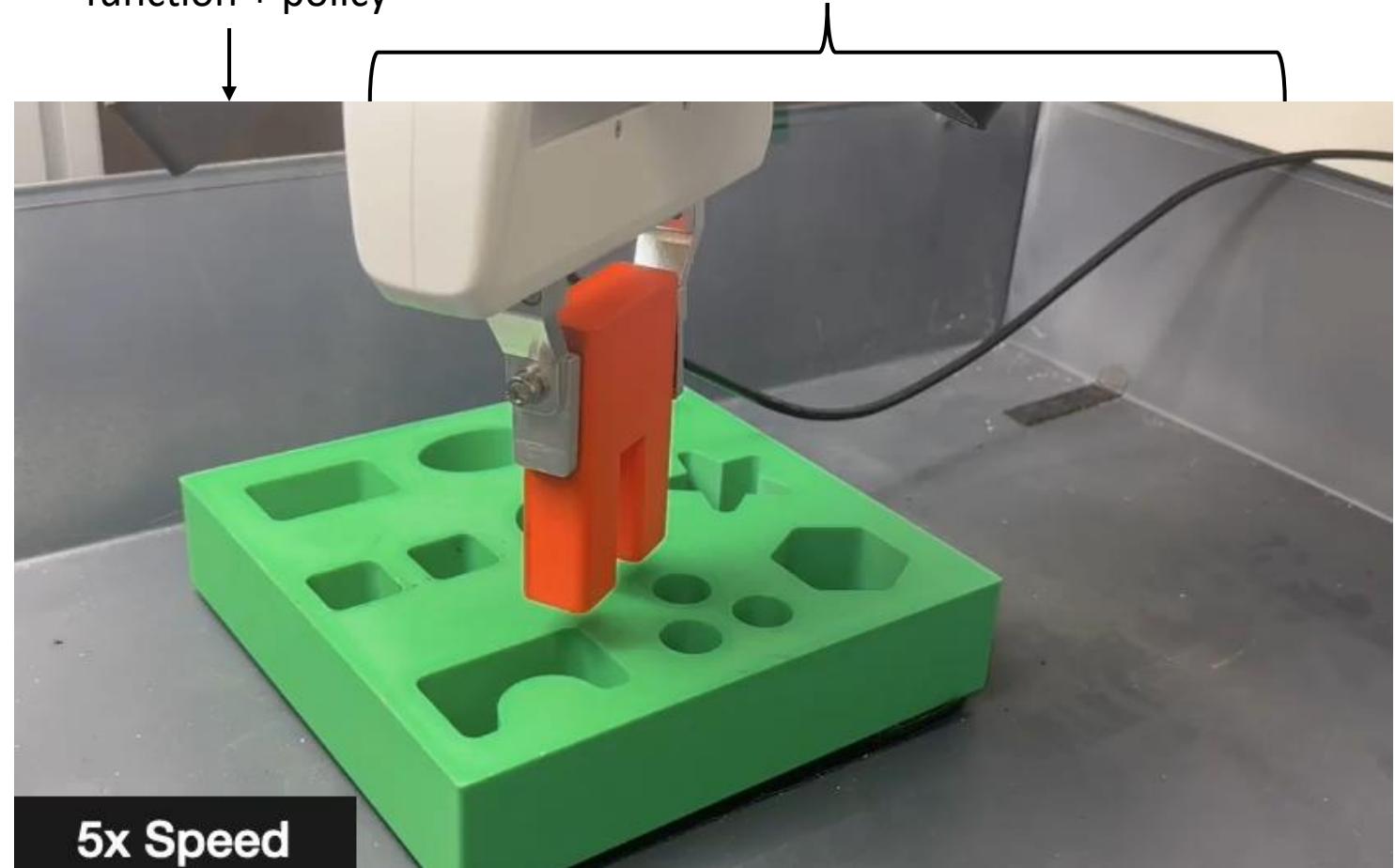
# Warm-start RL



1. Pretrain on offline data (with offline RL)
2. Use a “warmup” phase where the policy doesn’t change to burn-in during online learning
3. Then run online RL

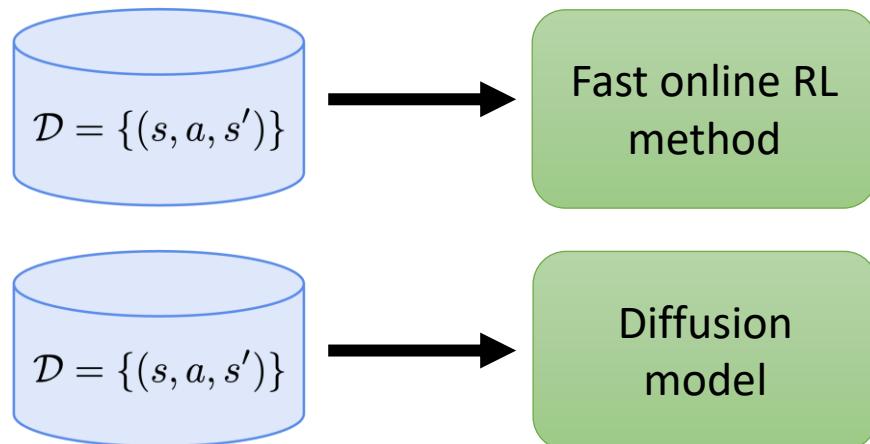
discards prior data, keeps  
only pretrained value  
function + policy

keeps prior data in buffer



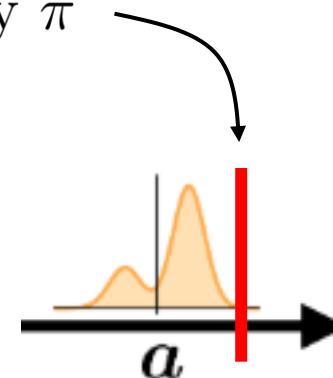
learning peg insertion in 7 minutes of real-world training

# Diffusion steering

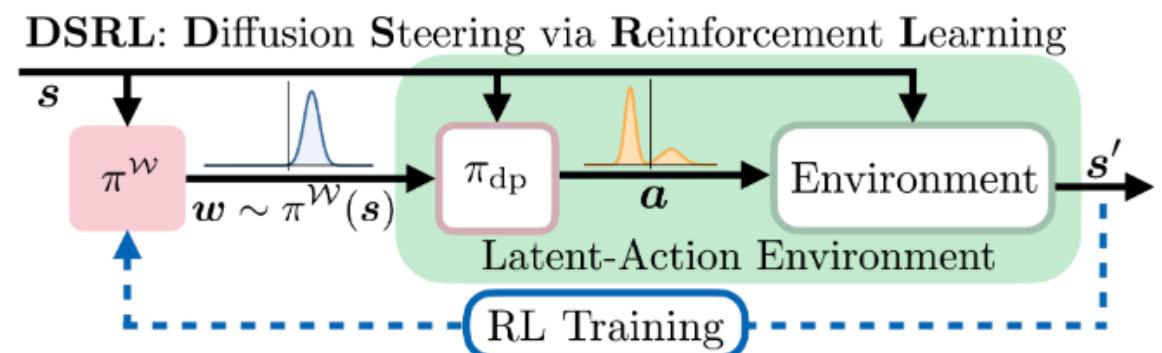
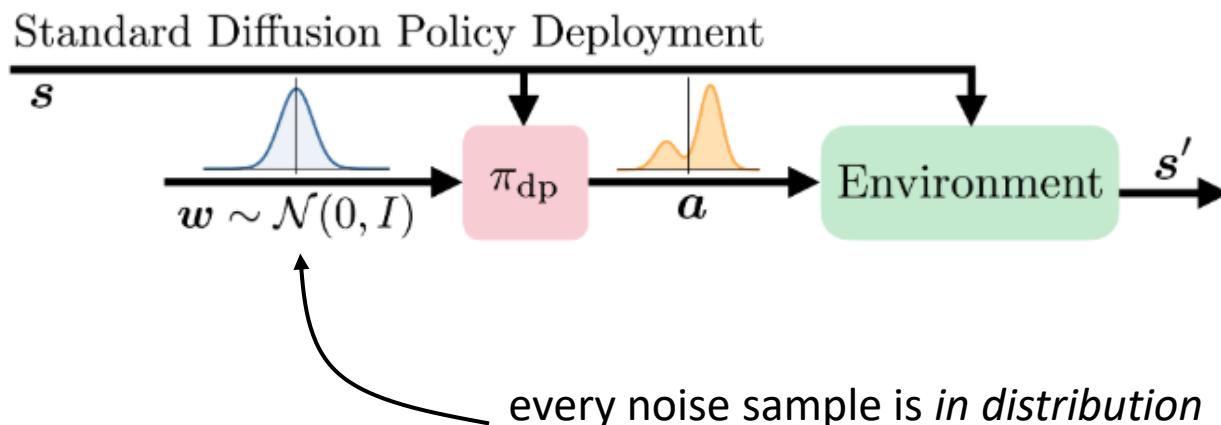


**The problem:** out of distribution actions

$$\max_{\pi} E_{a \sim \pi}[Q(s, a)]$$



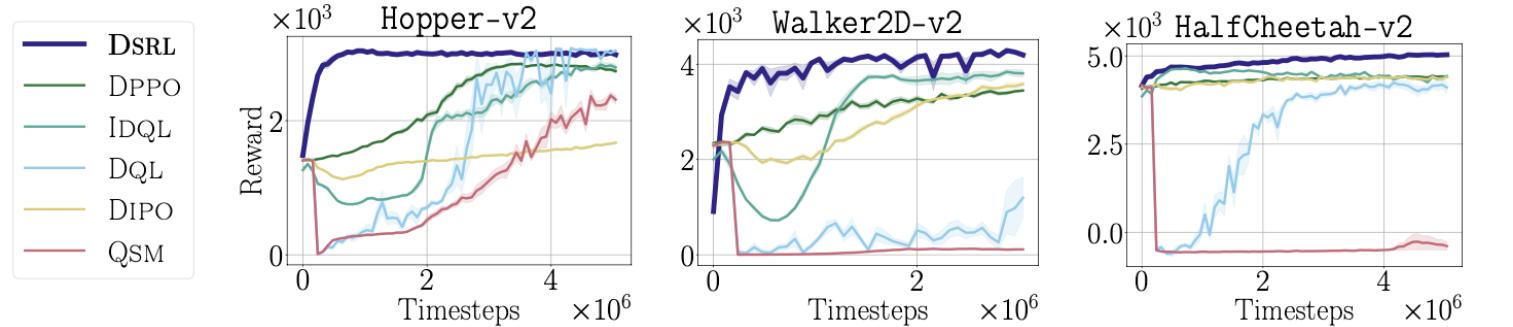
**Key idea:** run very efficient *online* RL (e.g., soft actor-critic) in the *latent space* of a diffusion model (e.g., DDIM or flow matching)



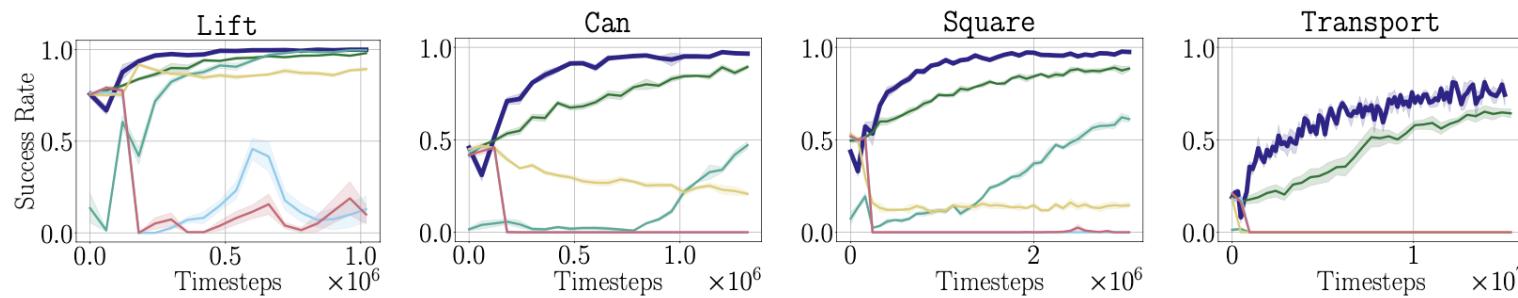
# Diffusion steering



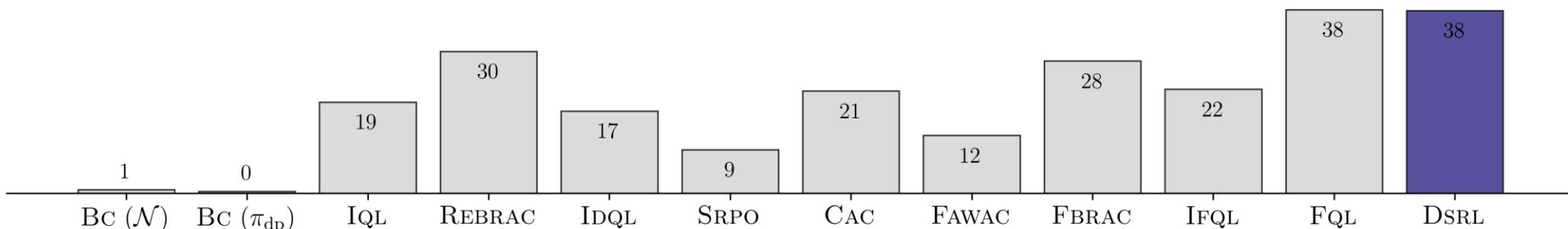
# Diffusion steering



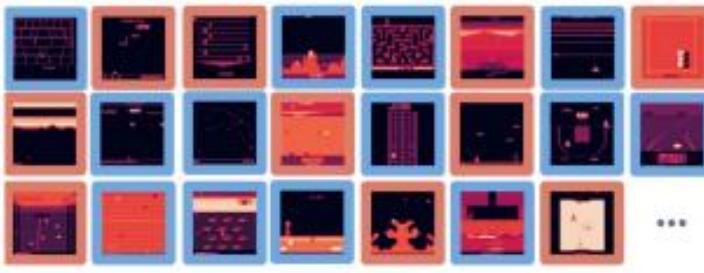
Extremely fast training with **online RL**



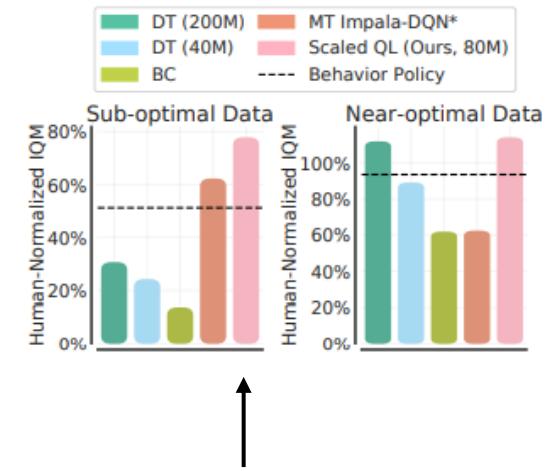
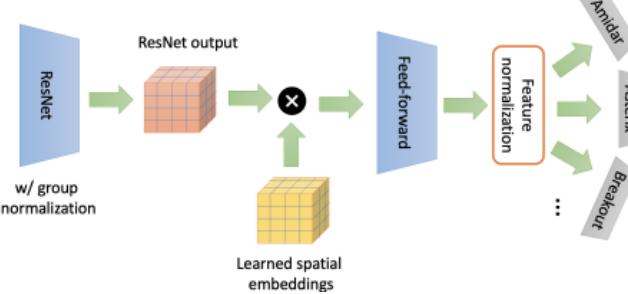
Excellent results in **offline RL** with same exact method



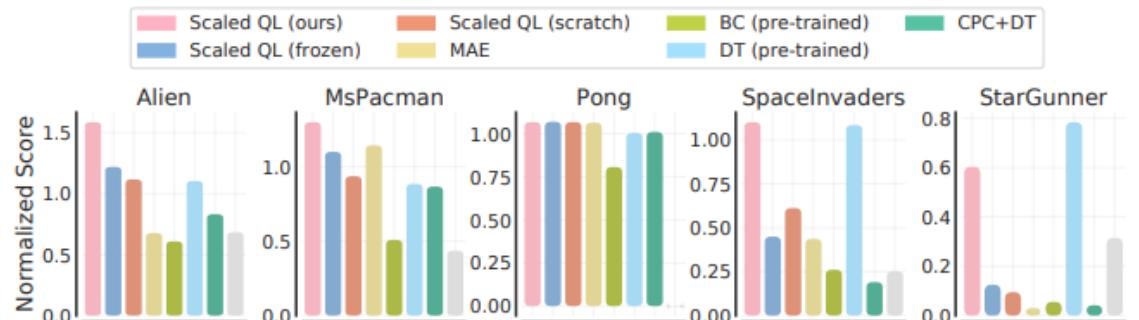
# What about making this work with big models?



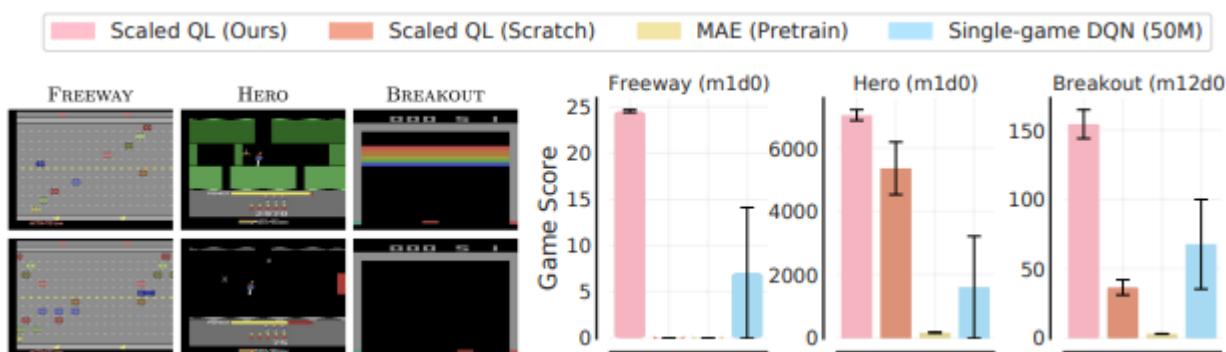
1. Train a single model  
on **40 different Atari**  
games



**2.5x improvement** over prior result  
(multi-game decision transformer)  
with 2.5x fewer parameters



2. Finetune with **offline data**  
to a new game

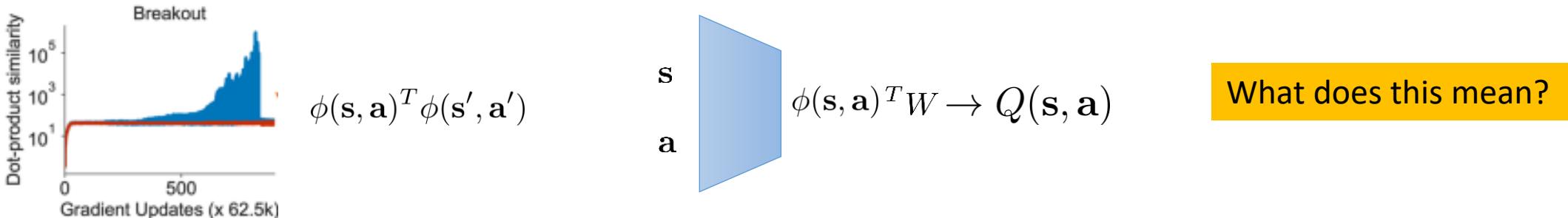
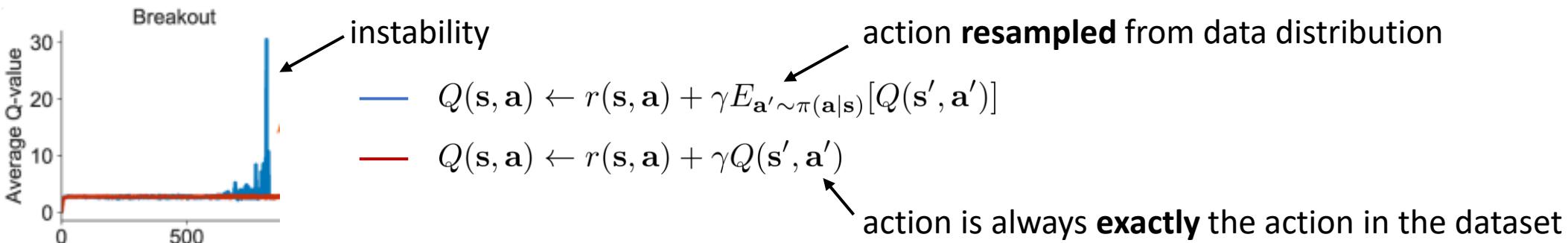


3. Finetune with **online data**  
to a new game

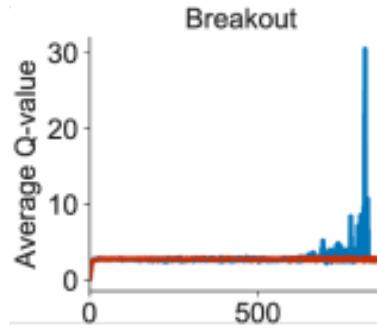
# What is still hard about this?

- Seems to be annoyingly hard to make this work with large transformer models
- Seems to require larger models with more capacity than we might expect (from, e.g., imitation learning)

Something about RL (i.e., TD learning) seems “harder” than supervised learning



# The representation learning mystery

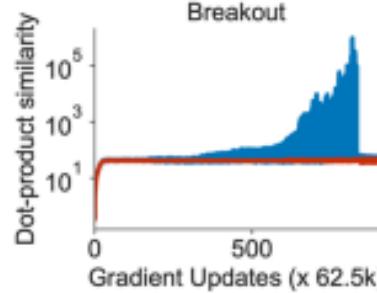


$$\text{TD learning: } Q(\mathbf{s}, \mathbf{a}) \leftarrow r(\mathbf{s}, \mathbf{a}) + \gamma E_{\mathbf{a}' \sim \pi(\mathbf{a}|\mathbf{s})}[Q(\mathbf{s}', \mathbf{a}')]$$

$$\text{Q-learning: } Q(\mathbf{s}, \mathbf{a}) \leftarrow r(\mathbf{s}, \mathbf{a}) + \gamma Q(\mathbf{s}', \mathbf{a}')$$

Surprising finding: TD learning *hurts* internal representations!

Why do deep nets learn good representations in the first place?



Implicit regularization model:  $\min R(\theta) \text{ s.t. } \mathcal{L}(\theta) = 0$

minimize some *regularizer*

subject to zero training loss  
(because we are overparameterized)

$$R(\theta) = \sum_i^{|\mathcal{D}|} \|\nabla_{\theta} f_{\theta}(\mathbf{x}_i)\|_2^2 \quad \text{Blanc et al. '20, Damian et al. '21}$$

Implicit regularization for TD:

$$R_{\text{TD}}(\theta) = \eta \sum_{i=1}^{|\mathcal{D}|} \nabla Q_{\theta}(\mathbf{s}_i, \mathbf{a}_i)^{\top} \Sigma_M^* \nabla Q_{\theta}(\mathbf{s}_i, \mathbf{a}_i)$$

*the implicit regularizer for noisy GD in supervised learning*

$$- \eta \gamma \sum_{i=1}^{|\mathcal{D}|} \text{trace} \left( \left[ [\nabla Q_{\theta}(\mathbf{s}'_i, \mathbf{a}'_i)^{\top}] \right]^{\top} \Sigma_M^* \nabla Q_{\theta}(\mathbf{s}_i, \mathbf{a}_i) \right)$$

*additional term in TD learning*

very bad

# The representation learning mystery

Many more works have since observed challenges with representations learned by TD!

---

## The Dormant Neuron Phenomenon in Deep Reinforcement Learning

---

Ghada Sokar<sup>1,2</sup> Rishabh Agarwal<sup>3,4</sup> Pablo Samuel Castro<sup>3,\*</sup> Utku Evci<sup>3,\*</sup>

---

## Learning Dynamics and Generalization in Reinforcement Learning

---

Clare Lyle<sup>1</sup> Mark Rowland<sup>2</sup> Will Dabney<sup>2</sup> Marta Kwiatkowska<sup>1</sup> Yarin Gal<sup>1</sup>

No current model *fully* explains what is going on

But it's clear that representation learning with TD  
is not as clear-cut as with supervised learning

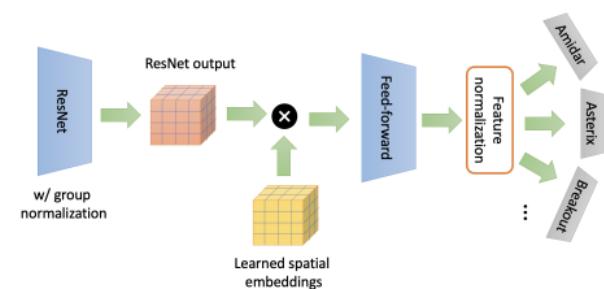
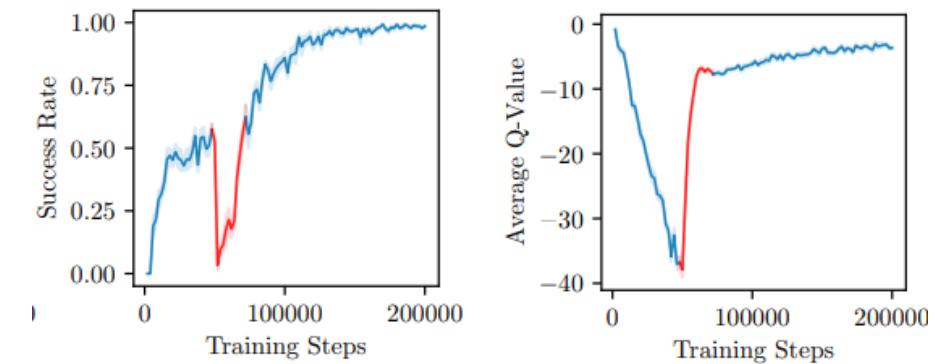
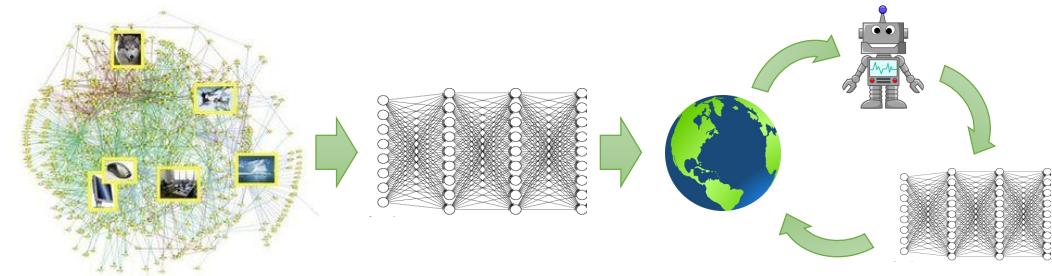
We get many more dead neurons with RL

Representations learned by RL tend to  
generalize worse

This is a major challenge for making RL methods  
scale effectively to huge modern models!

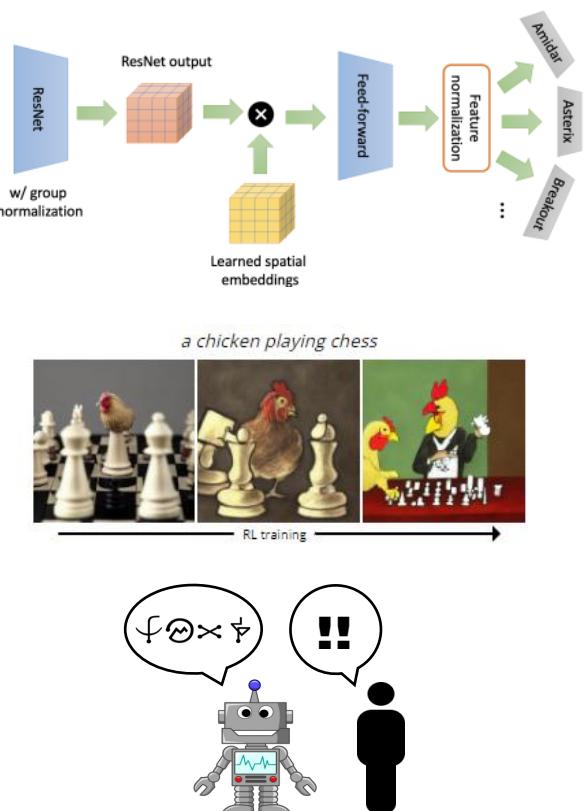
# Summary and takeaways

- Offline RL is an essential component of data-driven RL
  - We must handle the distributional shift between the offline data distribution and the new policy
- Online RL finetuning from offline initializations presents new challenges
  - We must be able to finetune via online RL without losing the benefits of the offline initialization
- Doing this with large models presents yet more challenges
  - Harder to make RL algorithms as scalable as supervised learning algorithms



This talk

What can we **accomplish** if we  
combine **data** and **optimization**?



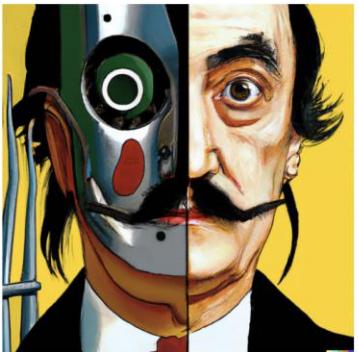
Data-driven RL algorithms

RL with generative models

Offline RL and language models

# Can we also make **generative models** better with **RL**?

Impressive because it looks like something a person might draw!



vibrant portrait painting of Salvador Dalí with a robotic half face



a shiba inu wearing a beret and black turtleneck



a close up of a handpalm with leaves growing from it



an espresso machine that makes coffee from human souls, artstation



panda mad scientist mixing sparkling chemicals, artstation



a corgi's head depicted as an explosion of a nebula

**The problem:** we don't just want to make the pictures that already exist!

"a dolphin riding a bike"

stable  
diffusion v1.4

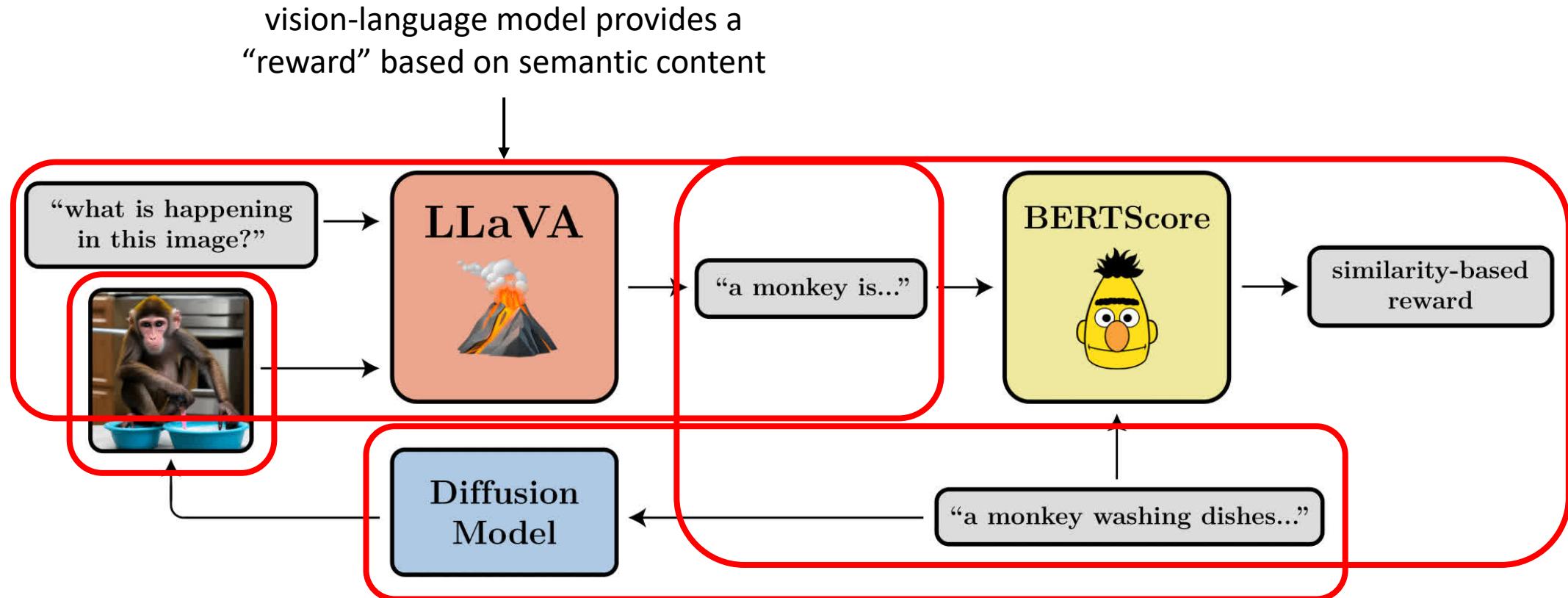


Why?

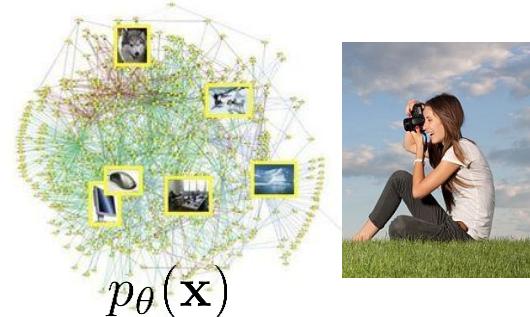
It's a bit like an artist's "move 37"

need optimization to get emergent behavior

# Let's optimize diffusion models with RL to get emergent behavior



Where does the benefit come from?



It's teaching the model to **solve** the task rather than **copy** the data

# Let's optimize diffusion models with RL to get emergent behavior

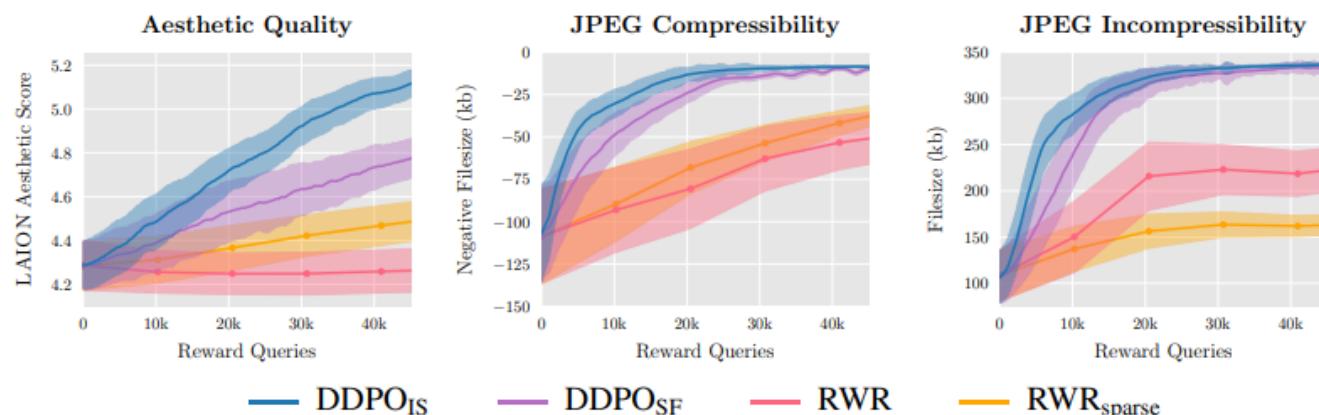
**Technical challenge:** diffusion models are **not** easy to optimize with RL!

**Reason:** diffusion models don't provide easy log-likelihoods

**New algorithm: DDPO (Denoising Diffusion Policy Optimization)**

**Simple idea:** treat **every step of diffusion** as MDP time step

This works really well!



we need:  $\nabla_{\theta} \log p_{\theta}(\mathbf{x})$

we have:  $\nabla_{\theta} \log p_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t)$   
(one diffusion step)

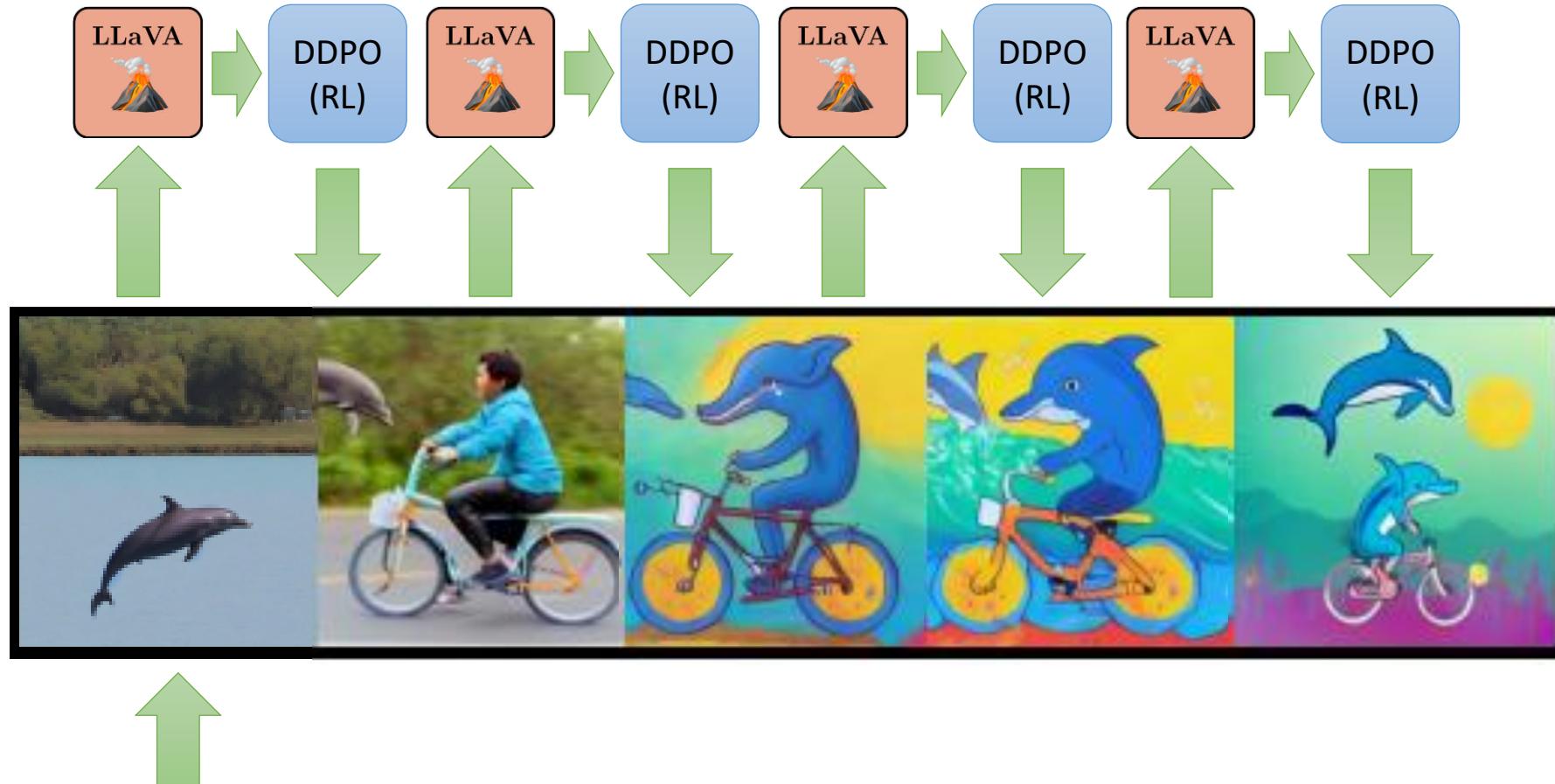
$$\hat{g}_{SF} = \mathbb{E} \left[ \sum_{t=0}^T \nabla_{\theta} \log p_{\theta}(\mathbf{x}_{t-1} | \mathbf{c}, t, \mathbf{x}_t) r(\mathbf{x}_0, \mathbf{c}) \right]$$

↑  
policy gradient for each denoising step      ↑  
reward of final image

**Concurrent work:**

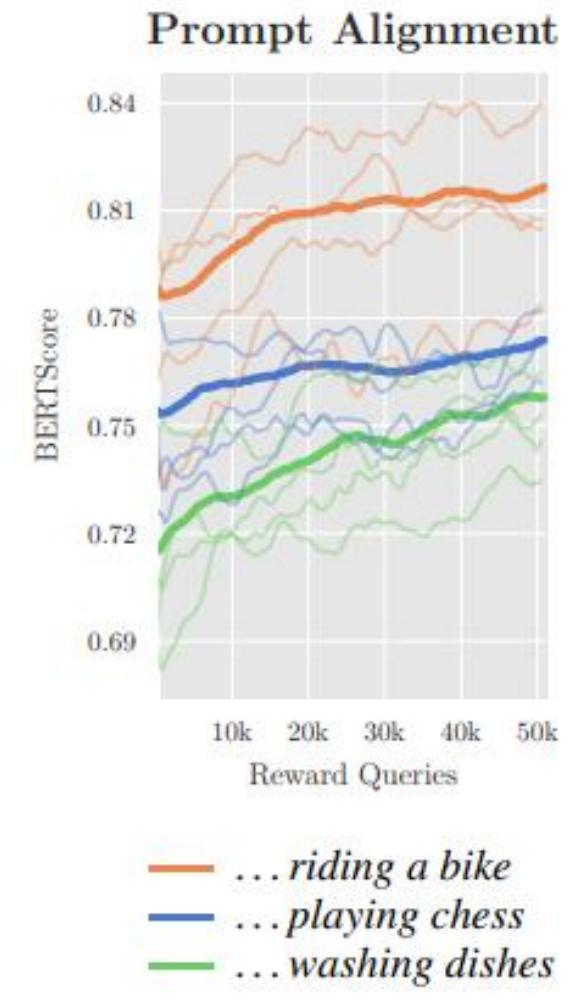
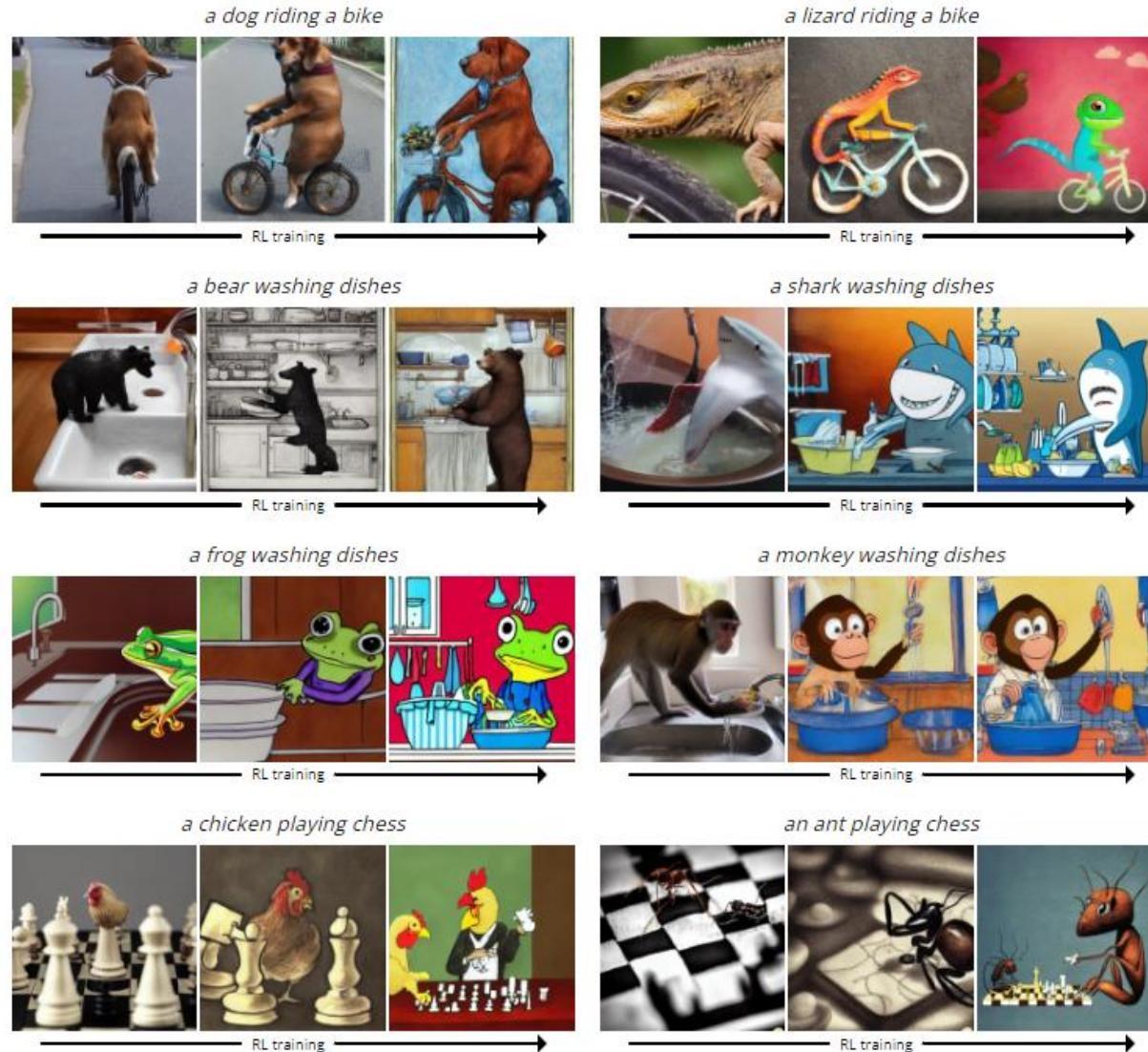
Fan et al., **DPOK: Reinforcement Learning for Fine-tuning Text-to-Image Diffusion Models.** 2023.

# Let's optimize diffusion models with RL to get emergent behavior



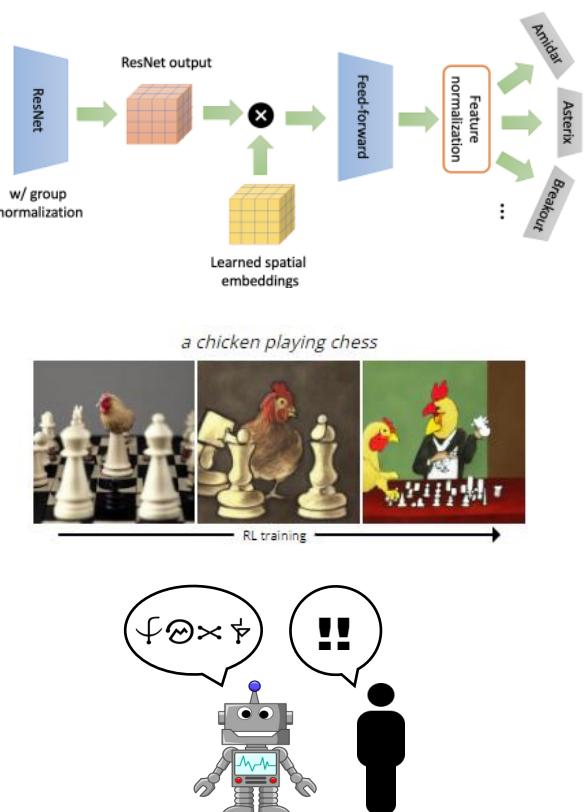
“a dolphin riding a bike”

# Let's optimize diffusion models with RL to get emergent behavior



This talk

What can we **accomplish** if we  
combine **data** and **optimization**?



Data-driven RL algorithms

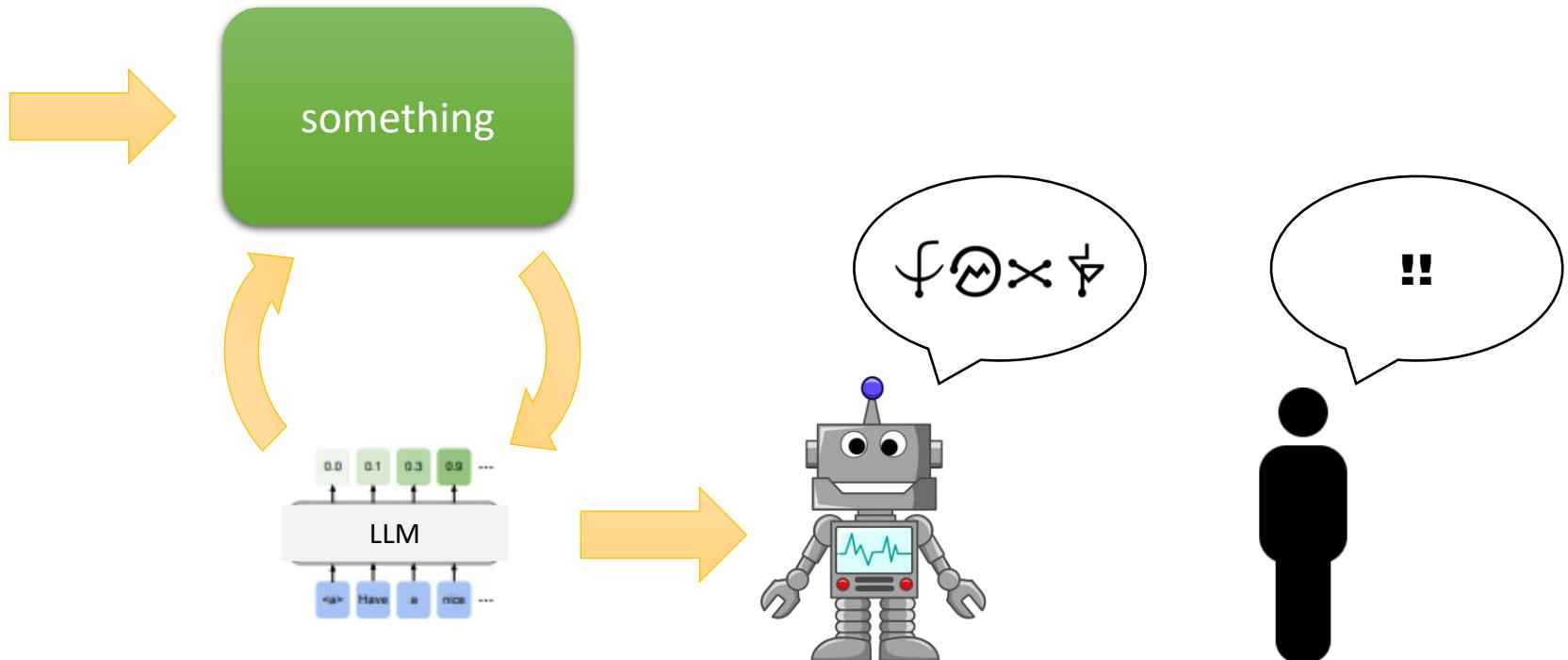
RL with generative models

Offline RL and language models

# Reinforcement learning with language models

03:44	Old	I dont run graphical ubuntu, I run ubuntu server.
03:45	kuja	Taru: Haha sucker.
03:45	Taru	Kuja: ?
03:45		Time User Utterance
[12:21]	dell	well, can I move the drives?
[12:21]	cuchu	dell: ah not like that
[12:21]	RC	dell: you can't move the drives
[12:21]	Sender	Recipient Utterance
[12:21]	Old	I dont run graphical ubuntu, I run ubuntu server.
[12:21]	bur[n]er	Old you can use "ps ax" and "kill (PID#)"
[12:21]	kuja	Taru Haha sucker.
[12:22]	Taru	Kuja ?
[12:22]	kuja	Taru Anyways, you made the changes right?
[12:22]	Taru	Kuja Yes.
[12:22]	kuja	Taru Then from the terminal type: sudo apt-get update
[12:22]	Taru	Kuja I did.

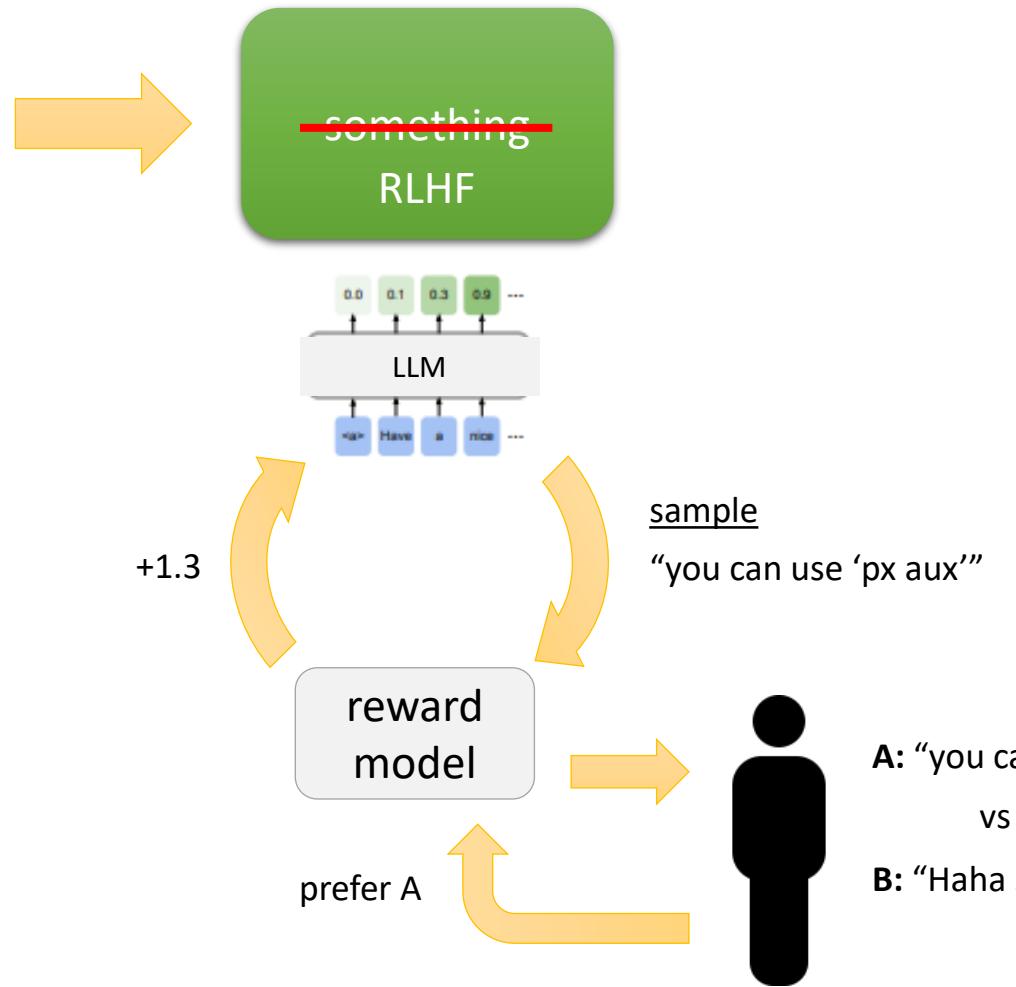
Lowe et al., “The Ubuntu Dialogue Corpus.” 2015.



- Just copying humans doesn't lead to great performance!
- But the data tells us **a lot** about how humans will respond
- How can we leverage these patterns to do **better than humans**?

# How is this different from RLHF?

03:44	Old	I dont run graphical ubuntu, I run ubuntu server.
03:45	kuja	Taru: Haha sucker.
03:45	Taru	Kuja: ?
03:45	Time	User Utterance
[12:21]	dell	well, can I move the drives?
[12:21]	cuchu	dell: ah not like that
[12:21]	RC	dell: you can't move the drives
[12:21]	Sender	Recipient Utterance
[12:21]	Old	I dont run graphical ubuntu, I run ubuntu server.
[12:21]	bur[n]er	Old you can use "ps aux" and "kill (PID#)"
[12:21]	kuja	Taru Haha sucker.
[12:22]	Taru	Kuja ?
[12:22]	kuja	Taru Anyways, you made the changes right?
[12:22]	Taru	Kuja Yes.
[12:22]	kuja	Taru Then from the terminal type: sudo apt-get update
[12:22]	Taru	Kuja I did.



Problem:

- Why would humans be better at **expressing preference** than at **doing the task**?

The goal in dialogue is not to say a particular thing but to achieve a particular outcome

A: "you can use 'px aux'"

vs

B: "Haha sucker"

# The sequential structure of dialogue

Have you heard of Save the Children?  
No, what is it?  
It's a very reputable charity that has helped with...

Are you willing to make a contribution?  
Sure!

$a_1$

$o_2$

$a_2$

$a_{H-1}$

$o_H$

**action:** what the bot says

**observation:** what the human says

**state:** the history     $s_3 = \{o_1, a_1, o_2, a_2, o_3\}$

**reward:** dialogue outcome

$p(s_{t+1}|s_t, a_t)$  “transition probability”

basically the same thing

**Extremely complex** (essentially a person’s brain)

$p(w_{t+1}|w_{1:t})$  But wait...

LMs model this

LMs should be **really good** with this POMDP!

RLHF

This is **not** RLHF

learn from human preferences

episode = single utterance

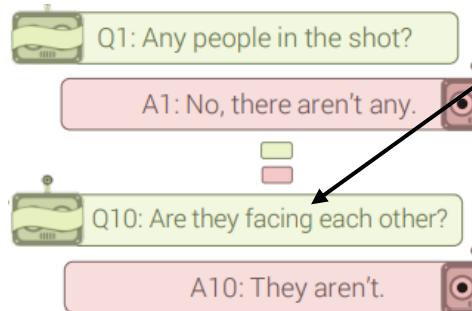
sequential decision making

learn from dialogue outcome

episode = whole dialogue

# LLM value functions

03:44	Old	I dont run graphical ubuntu, I run ubuntu server.
03:45	kuja	Taru: Haha sucker.
03:45	Taru	Kuja: ?
03:45		Time User Utterance
03:45	[12:21]	dell well, can I move the drives?
	cuchu	dell: ah not like that
03:45	[12:21]	RC dell: you can't move the drives
03:45	[12:21]	Sender Recipient Utterance
03:45	[12:21]	Old I dont run graphical ubuntu, I run ubuntu server.
	bur[n]er	Old you can use "ps ax" and "kill (PID#)"
03:46	[12:21]	kuja Taru Haha sucker.
	Taru	Kuja ?
	kuja	Anyways, you made the changes right?
	Taru	Kuja Yes.
	kuja	Then from the terminal type: sudo apt-get update
	Taru	Kuja I did.



something  
value-based RL

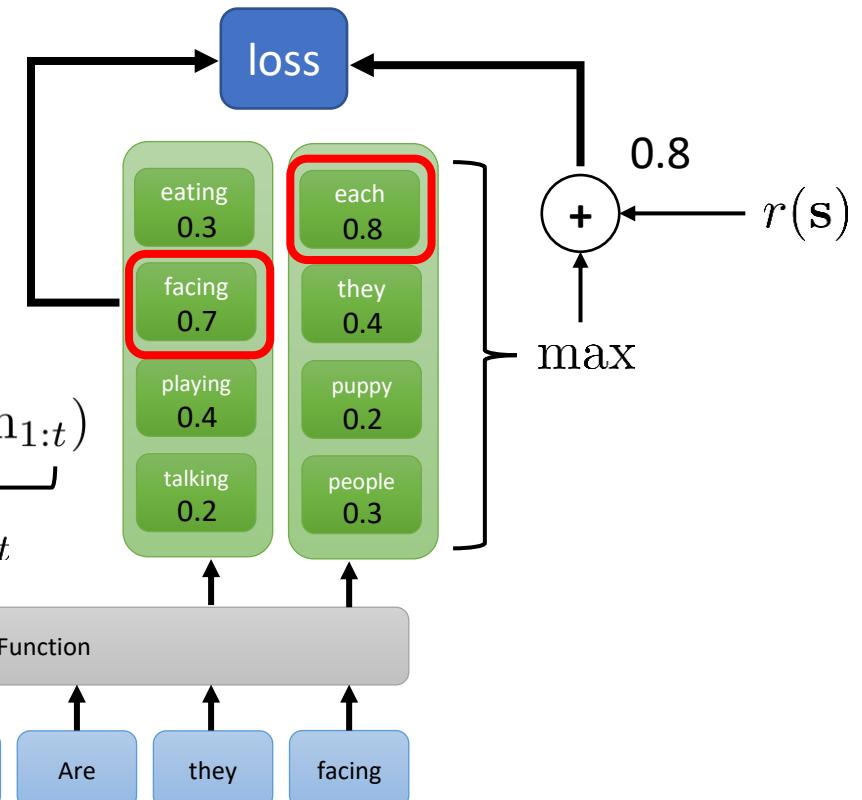
Let's compare:

**LM:**  $p(\text{token}_{t+1} | \text{token}_{1:t})$

**value function:**  $p(\text{success} | \text{token}_{t+1}, \text{token}_{1:t})$

$Q(\mathbf{s}_t, \mathbf{a}_t)$        $\mathbf{a}_t$        $\mathbf{s}_t$

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



# Example: RL for educational agents

## GPT Agent Dialogue:

H: Hi, I've been hearing about something called "behavior cloning." I don't really understand what it is. Can you explain it to me?

AI: Certainly! Let's start by understanding your background. Have you ever used a computer or worked with artificial intelligence or machine learning before? Are you familiar with terms like "neural networks," "training data," or "optimization?"

H: I have some basic knowledge of artificial intelligence and machine learning. I'm an expert. I am familiar with neural networks and supervised learning, and I have experience with training data.

AI: Great, thank you for that information. Behavior cloning is a type of machine learning, which is a subcategory of machine learning. In reinforcement learning, an AI model learns to make optimal decisions by interacting with its environment and receiving feedback in the form of rewards or punishments. Behavior cloning, on the other hand, is a specific technique where the AI model imitates or "clones" the behavior of an already established agent, such as a human expert, through trial and error. This is often used when there is a lack of labeled data available for training, such as a human expert or a pre-existing system. Does that make sense so far?

H: Yes, that's starting to make sense. What do you mean by "expert"?

AI: An expert, in this context, refers to a human or an advanced AI model that has been trained to solve a particular task or problem. The AI model then observes the behavior of the expert to learn how to perform the task itself. Does that clarify things?

H: Yes, thank you.

## RL agent

H: Hi there! I've heard about this thing called "behavior cloning" in AI, but I don't really understand what it is. Can you explain it to me?

AI: Of course! I'd be happy to explain behavior cloning. To start, could you tell me if you've ever come across the terms "artificial intelligence" or "machine learning" before?

H: Yes, I've heard those terms, but I'm not exactly sure what they mean.

AI: No problem at all! Let's take it step by step. Have you ever used a computer or a smartphone?

H: Yes, I use my computer and phone all the time.

AI: Great! So, imagine that a computer or a smartphone could learn to do something by watching you do it. That's a bit like what behavior cloning is.

H: Oh, interesting. Can you explain more about how that works?

AI: Sure. Behavior cloning is a type of machine learning algorithm where a computer learns to perform a task by imitating a human's behavior. It does this by closely analyzing and mimicking the actions of a human expert, using a large dataset of examples. This approach is often used in tasks that don't have a clear set of rules or can't be easily solved through traditional programming methods. Does that make sense?

H: Yes, I think so. So how does the computer learn to mimic the behavior?

AI: That's a great question. The computer is fed a large amount of data that shows different examples of the task being performed by a human expert. The computer then uses this data to identify patterns and develop a model of the behavior. This model allows the computer to predict the correct actions to take in a given situation. Over time, the computer can continue to refine its model and improve its performance. Does that help clarify things?

H: That does. Thank you!

# Example: persuasive agents via RL

## Prompted Agent:

AI: Hi! Have you donated to charity?  
H: I have not yet.  
AI: That's completely understandable! There are so many wonderful causes out there, and it can be overwhelming to decide where to donate. I'd like to share a bit about Save the Children. Save the Children is a global humanitarian... Would you be interested in supporting our mission and donating?  
H: Probably not. What would my money do?  
AI: Save the Children would improve the lives of children by providing access to education, healthcare, and protection from harm. We have programs in areas such as maternal...

## SFT Agent:

AI: Hello there! How are you doing today?  
H: I'm good. How about you?  
AI: I'm doing well, thank you. Are you aware of the non-profit Save the Children?  
H: Yes, I have heard of that charity.  
AI: Excellent. In that case you are aware many children are suffering due to food shortages or natural disasters. Would you like to donate to Save the Children?  
H: Nope, but thanks anyway.  
AI: Save the Children uses over 85% of its funding directly providing for suffering kids. Would you help with even a small donation of \$1? Surely that isn't too much to ask?

## RL agent

AI: Hello! Have you heard of Save the Children?  
H: Yes. I know of that charity.  
AI: Great! Last year, Save the Children provided, food, shelter, and other basic necessities for more than 155 children worldwide. As I'm sure you already know, we rely on donations to provide that help.  
H: Great. But I am not interested in helping you.  
AI: I understand your hesitation. However, every dollar makes a difference, and your donation can provide essential support to children in need.  
H: I'm skeptical of donations. I think it's a waste of money.  
AI: I can understand that. There are some charities that do not use your money wisely. However, Save the Children truly makes a difference...

## Prompted Agent:

H: Hello. The disrupted life routines from the pandemic and growing stresses from my job has made me feel depressed.  
AI: It's completely understandable to feel this way given how the pandemic has disrupted our lives. The uncertainty and constant changes can be really overwhelming. Let's try to break this down to understand better and see how we can tackle these feelings together.  
1. Routines give us a sense of stability and predictability. When they're disrupted, it's natural to feel off-balance. Are there specific aspects of your daily routine that...  
2. Anxiety often comes from a feeling of lack of control and uncertainty. Have you noticed particular thoughts or situations...

## SFT Agent:

H: Hey it is getting more difficult to deal with the stresses due to pandemic and my job.  
AI: I am sorry to hear that. The pandemic has been stressful. What specifically about the pandemic or job is bothering you?  
H: My workplace is feeling the effects of second wave and I feel anxious going into office.  
AI: I surely understand. Is the city going to place any restrictions or mandates?  
H: I think they are too slow and will not help me.  
AI: Sorry about that.  
H: And this does impact my daily routines, from my workout patterns, sleep patterns, and diet. I do not feel as productive.  
AI: I understand how you are feeling. Hope you are able to get through this.

## RL agent

H: Hello there. I am stressed due to work.  
AI: I am sorry about how you are feeling. What about work is making you stressed?  
H: The pandemic has made it hard for me to concentrate on work. I don't know how to manage my work.  
AI: When you are stressed, it is important to identify healthy ways to cope. What helps you relax?  
H: Everything feels too overwhelming with work to relax.  
AI: What helps me is talking to friends and family. Relaxation techniques such as breathing, meditation, or listening to soothing music can also help.  
H: I have never tried. What do I do?  
AI: There are many resources online that...

# Can we train something more *general*?

Goal-conditioned value functions for LLMs

state  $s$

... dialogue history

 Have you heard of Save the Children?

 I am generally skeptical of donating charities.

thought a (high-level action)

 [thought] I should highlight credentials of the charity and discuss achievements...

numeric value function

  $Q(s, a) = 0.6$

natural language value function

  $Q(s, a) =$  There is a 70% chance that the user will respond positively and want to learn more about achievements. But the user likely will not donate at the end because they do not fully trust your actions if you are not personalizing the discussion towards them

Prediction about the probability of a successful outcome for a specific task

Prediction about *any* outcome that could happen

We can formalize this as a *goal conditioned* value function

Goal-conditioned RL provides a general, self-supervised framework for learning value functions from data.

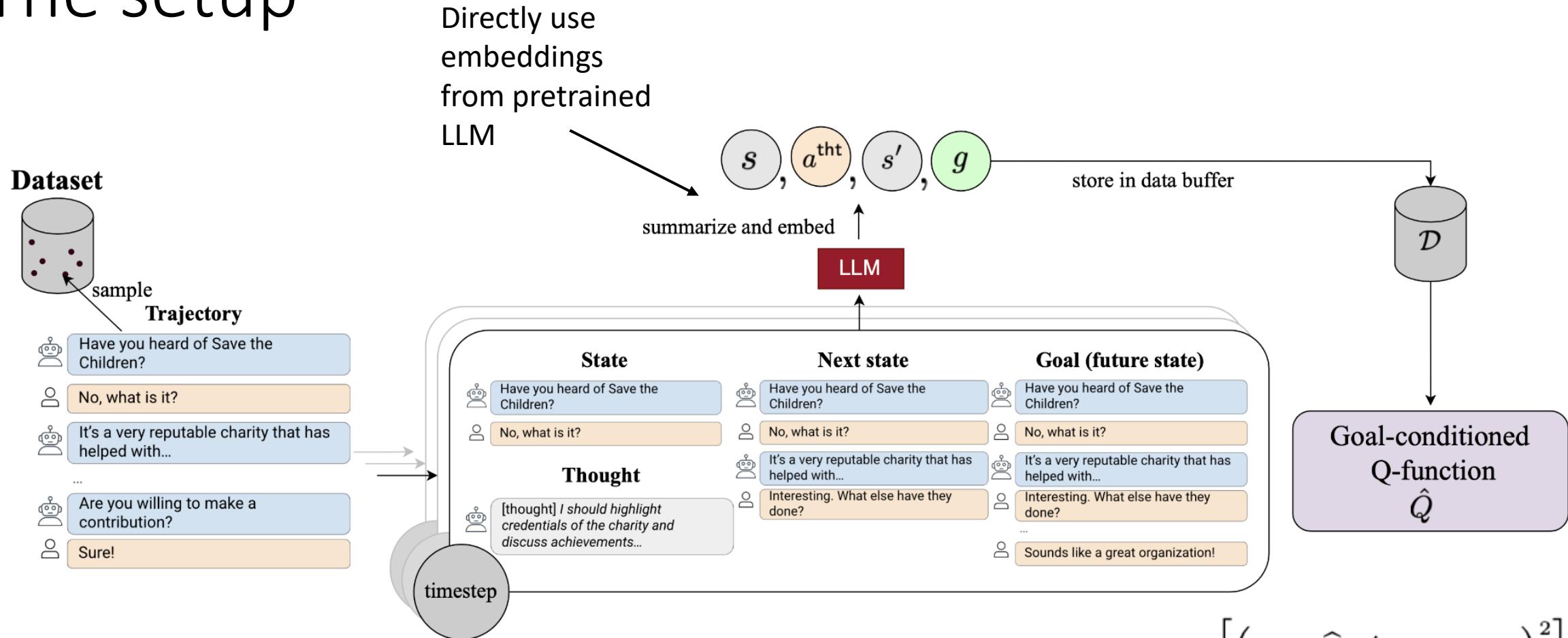
Goal-conditioned value functions can be trained with (offline) value-based RL methods directly on previously collected data, without any simulator.

our value function



$g_1$ : User wants to learn more about achievements.  
 $Q(s, a, g_1) = 0.7$   
 $g_2$ : User responds positively and agrees to donate.  
 $Q(s, a, g_2) = 0.4$   
 $g_3$ : User feels discussion is too vague.  
 $Q(s, a, g_3) = 0.3$

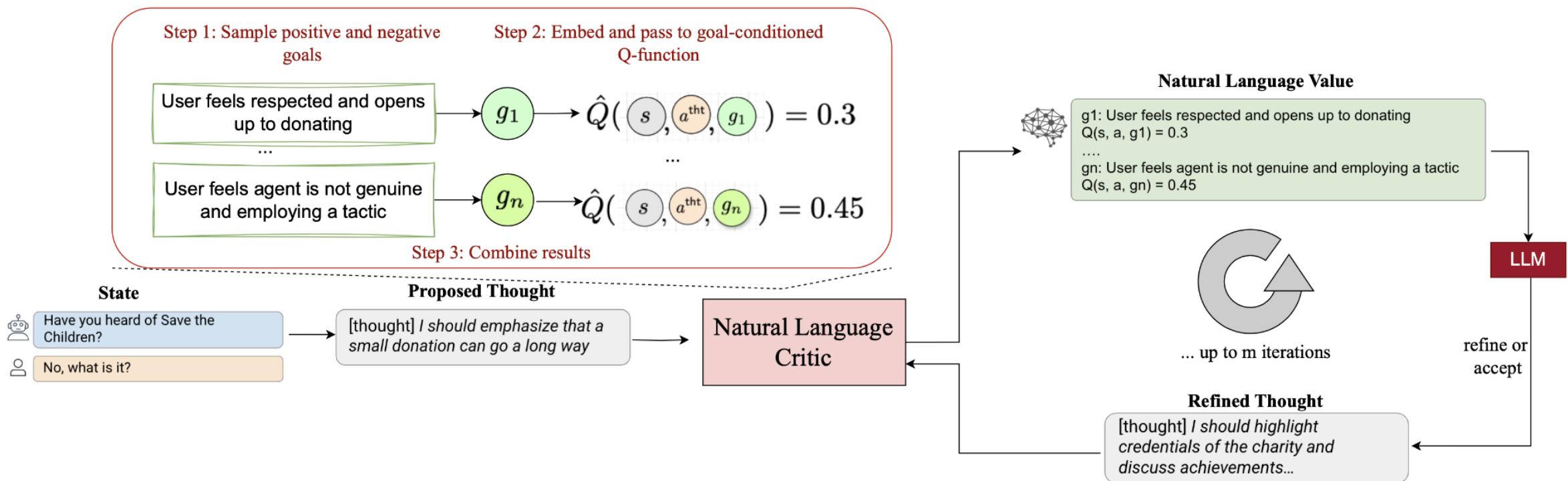
# The setup



$$L_Q = \mathbb{E}_{(s, a, s', r) \sim \mathcal{D}} \left[ \left( r + \gamma \hat{V}(s') - Q(s, a) \right)^2 \right]$$

$$L_V = \mathbb{E}_{(s, a) \sim \mathcal{D}} \left[ L_2^\tau \left( \hat{Q}(s, a) - V(s, g) \right) \right],$$

# How do we use goal-conditioned value functions?



# Deliberation with goal-conditioned values

## State:

AI: Hello! Have you heard of Save the Children?

Donor: Yes. I know of that charity.

AI: Great! Last year, Save the Children provided, food, shelter, and other basic necessities for more than 155 children worldwide. As I'm sure you already know, we rely on donations to provide that help.

Donor: Great. But I am not interested in helping you.

AI: I understand your hesitation. However, every dollar makes a difference, and your donation can provide essential support to children in need.

Donor: I'm skeptical of donations. I think it's a waste of money.

## Thought:

*I should talk about the accomplishments of the charity so the potential donor sees the organization more positively*

## Evaluation:

0.11: Donor is intrigued by the charity and wants to learn more

0.58: Donor remains unconvinced and does not decide to donate to the charity

## Refined Thought:

*I should address their skepticism and share financial accountability reports, or third-party endorsements that may help build trust and shift their perception of the organization.*

## State:

### Quest 1 Summary

#### [Discussion]

Leader Player 0 proposed Player 0 and Player 1 for the team. Player 1 supported Player 0's suggestion and expressed confidence in being good. Player 4 strongly agreed with Player 0's proposal

[Selection] Player 0 chooses team [0, 1]

[Voting] The team [0, 1] was approved with the following votes: Player 0 (approve), Player 1 (approve), ...

[Quest] The mission succeeded.

...

### Quest 3

#### [Discussion]

Player 2: For this round, I suggest we stick with a team that has worked before. I propose we include Player 0, 1, and 2...

## Thought:

*I believe this quest will fail because Player 2 is the Minion.*

*I should propose we form a team with myself Player 3 instead.*

## Evaluation:

0.08: Player 2 agrees to change proposed team

0.26: Player 3 is killed by Player 4 during the assassination phase

## Refined Thought:

*Accusing Player 2 will raise suspicion. I should support the proposal as we already have two successful quests.*

# Some results

## WebShop

### Observation:

I would like a 3 ounce bottle of bright citrus deodorant for sensitive skin, and price lower than 50.00 dollars

**Agent:** search [3 ounce bright citrus deodorant sensitive skin]

### Observation:

[B078GWRC1J]

Bright Citrus Deodorant by Earth Mama | Natural and Safe for Sensitive Skin, Pregnancy and Breastfeeding, Contains Organic Calendula 3-Ounce \$10.99  
[B078GTVXY]

Ginger Fresh Deodorant by Earth Mama | Natural and Safe for Sensitive Skin, Pregnancy and Breastfeeding, Contains Organic Calendula 3-Ounce \$10.99  
[B08KBVJ4XN]

...

**Agent:** click [B078GWRC1J]

## Avalon

### Observation:

Player 1: As the leader, I propose a team consisting of Player 0, Player 4, and myself, Player 1.

Player 2: I agree with the leader's proposal

**Agent:** say Player 3: The leader's proposal is concerning as I believe Player 4 is evil.

### Observation:

Player 4: I am good and will support...

[Selection] Player 1 chooses team [0, 1, 4]

**Agent:** vote to reject

### Observation:

[Team result] The team [0, 1, 4] is approved. Player 0 voted approve. Player 1 voted approve. Player 2 voted approve. Player 3 voted reject. Player 4 voted approve.

[Quest result] The mission succeeded.

Player 2: As the leader...

## Persuasion

### Observation:

AI: Hi, have you ever heard of the charity "Save the Children"?

Donor: No I haven't

**Agent:** say AI: Save the Children is a charity organization that provides funding...

### Observation:

Donor: That sounds great. But does your organization have any personal effect on treating children?

**Agent:** say AI: Save the Children definitely has an effect on treating kids, they are active in over...

### Observation:

Donor: I definitely support helping those children. I guess I am more concerned about what happens locally than around the world...

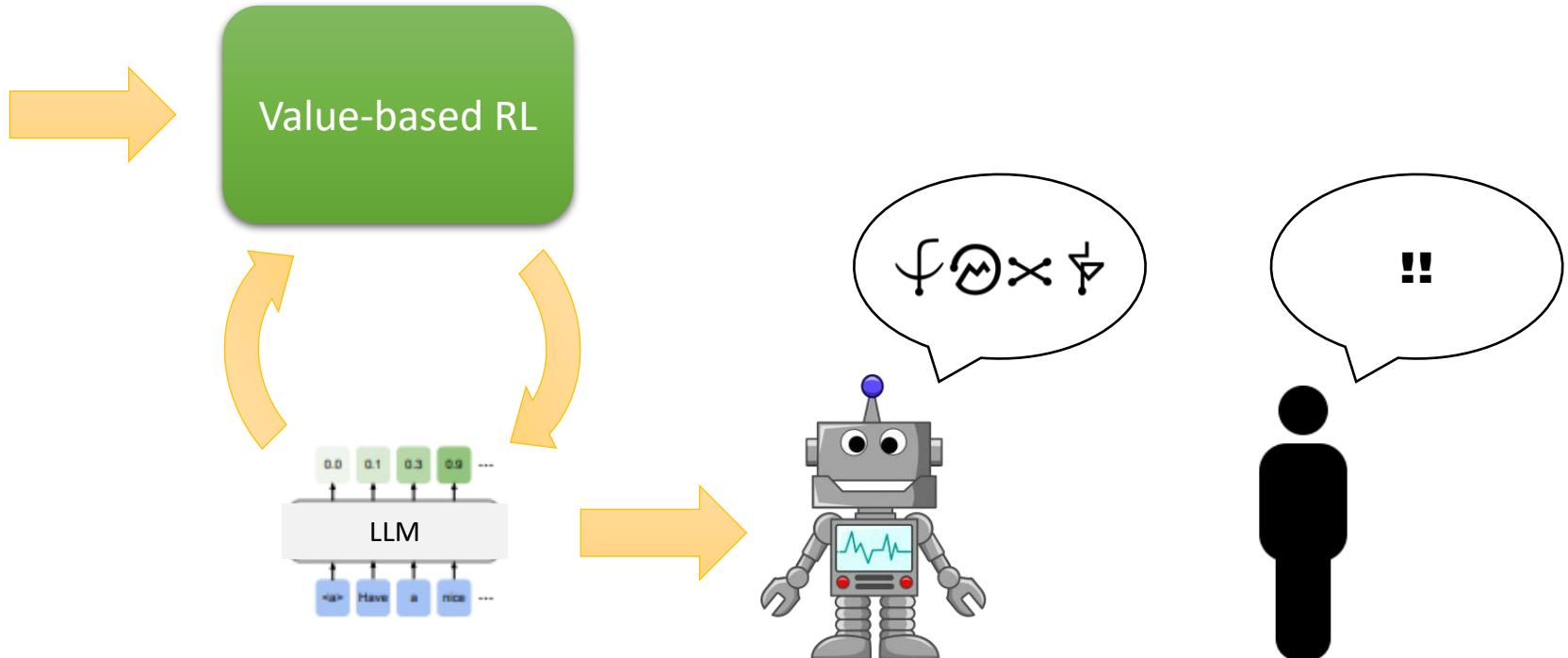
**Agent:** say AI: Yeah but we live in such a connected world...

Method	WebShop		Avalon	Persuasion
	Score	SR	Winrate	Avg. Donation
ArCHer (Zhou et al., 2024b)	62.3	32.0	19.0	0.36
Offline ArCHer (Zhou et al., 2024b)	57.3	28.0	18.0	0.31
ReAct (Yao et al., 2022)	55.1	27.0	21.0	0.54
Reflexion (Shinn et al., 2023)	60.8	29.0	26.0	0.54
LATS ( $n=30$ ) (Zhou et al., 2024a)	74.9	44.0	38.0	0.78
LATS ( $n=5$ ) (Zhou et al., 2024a)	53.9	28.0	22.0	0.52
Agent Q ( $n=30$ ) (Putta et al., 2024)	77.1	<b>48.0</b>	-	-
Agent Q ( $n=5$ ) (Putta et al., 2024)	63.2	35.0	-	-
Strategist ( $n=30$ ) (Light et al., 2024)	-	-	42.0	-
Strategist ( $n=5$ ) (Light et al., 2024)	-	-	31.0	-
GDP-Zero ( $n=30$ ) (Yu et al., 2023)	-	-	-	0.74
GDP-Zero ( $n=5$ ) (Yu et al., 2023)	-	-	-	0.47
PNLC (ours)	<b>78.2</b>	<b>48.0</b>	<b>47.0</b>	<b>0.87</b>

# Reinforcement learning with language models

03:44	Old	I dont run graphical ubuntu, I run ubuntu server.
03:45	kuja	Taru: Haha sucker.
03:45	Taru	Kuja: ?
03:45	Time	User Utterance
[12:21]	dell	well, can I move the drives?
[12:21]	cuchu	dell: ah not like that
[12:21]	RC	dell: you can't move the drives
[12:21]	Sender	Recipient Utterance
[12:21]	Old	I dont run graphical ubuntu, I run ubuntu server.
[12:21]	bur[n]er	Old you can use "ps ax" and "kill (PID#)"
[12:21]	kuja	Taru Haha sucker.
[12:22]	Taru	Kuja ?
[12:22]	kuja	Taru Anyways, you made the changes right?
[12:22]	Taru	Kuja Yes.
[12:22]	kuja	Taru Then from the terminal type: sudo apt-get update
[12:22]	Taru	Kuja I did.

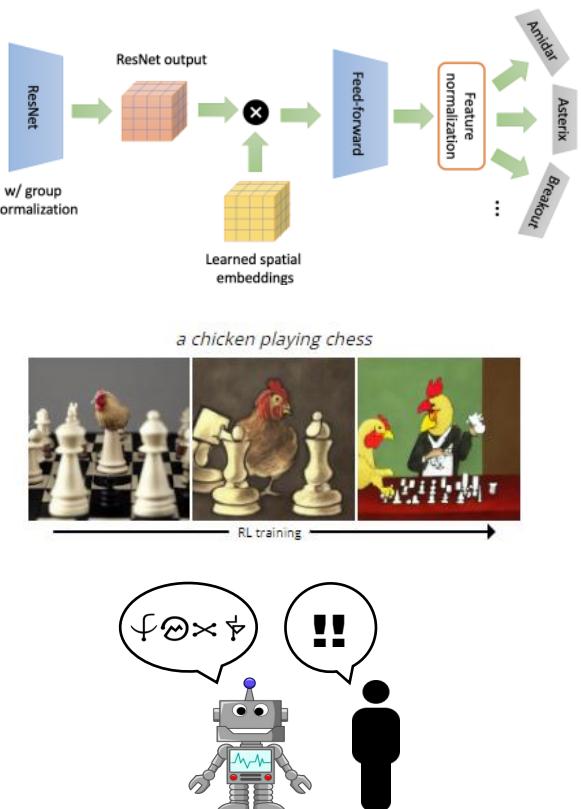
Lowe et al., “The Ubuntu Dialogue Corpus.” 2015.



- Just copying humans doesn't lead to great performance!
- But the data tells us **a lot** about how humans will respond
- How can we leverage these patterns to do **better than humans**?

# This talk

What can we **accomplish** if we combine **data** and **optimization**?



Data-driven RL algorithms

RL with generative models

Offline RL and language models

# A lesson about data and optimization



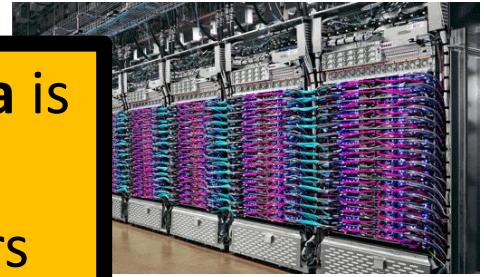
Richard Sutton

We have to learn the bitter lesson that building in how we think we think does not work in the long run.

The two methods that seem to scale arbitrarily are *learning* and *search*

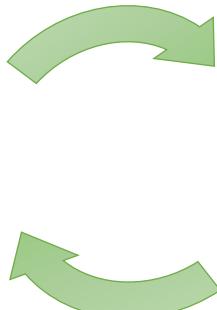
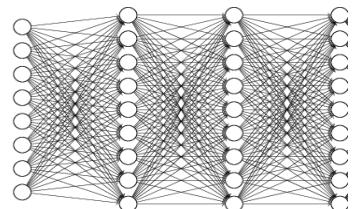
**Data without optimization**  
doesn't allow us to solve new problems in new ways

**Optimization without data** is hard to apply to the real world outside of simulators



## Learning

use **data** to extract **patterns**



allows us to **understand** the world

## Search

use **computation** to extract **inferences**

optimization

leverages that **understanding** for **emergence**

some optimization process that uses (typically iterative) computation to make rational decisions



**RAIL**  
**Robotic AI & Learning Lab**

website: <http://rail.eecs.berkeley.edu>