

Deep Reinforcement Learning with Real-World Data

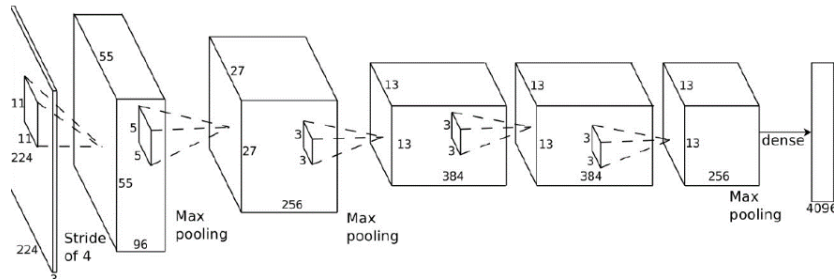
Sergey Levine

UC Berkeley

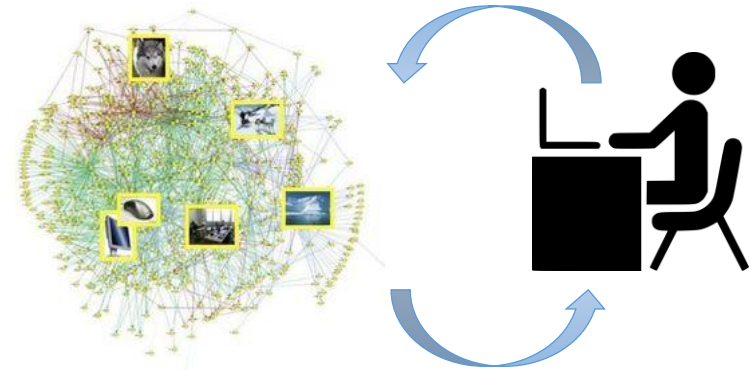


Large-scale machine learning

Why does deep learning work?

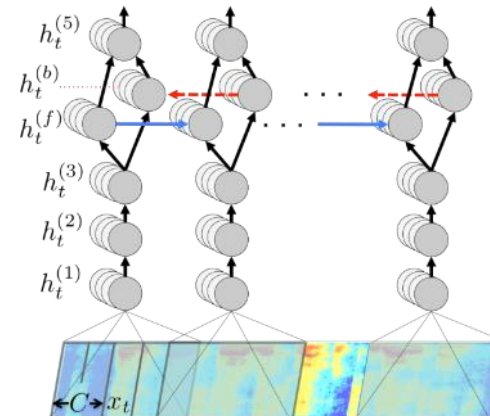
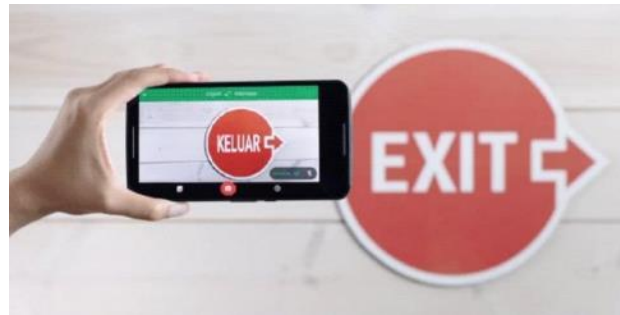


big models

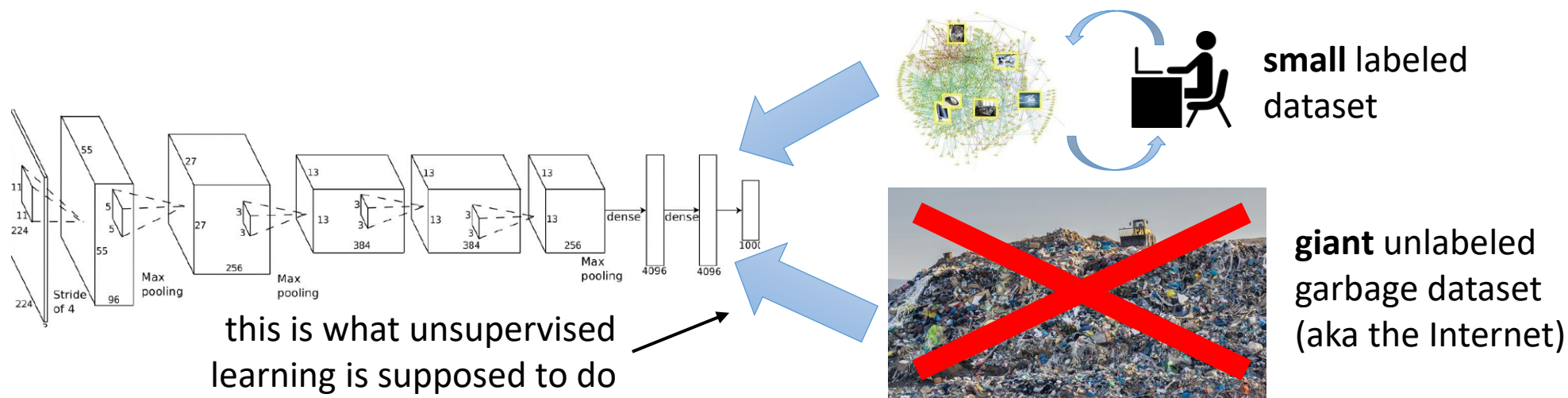


big datasets

labeled



Reducing the supervision burden



But then where does the knowledge come from?

“Classic” unsupervised learning: $p_{\theta}(\mathbf{x})$
(this is what, for example, large language models do)



Aside: perhaps this is why “prompting” large language models is such an art!

Stepping back a bit...

Why do we need **machine learning** anyway?

A postulate:

We need machine learning for one reason and one reason only – that's **to produce adaptable and complex decisions.**



Decision: how do I move my joints?



Decision: how do I steer the car?



What is the decision? The image label?

What happens with that label **afterwards**?

Is it used to tag a user's photo?

Detect an endangered animal in a camera trap?

these are **decisions**

they lead to **consequences**

Aside: why do we need **brains** anyway?

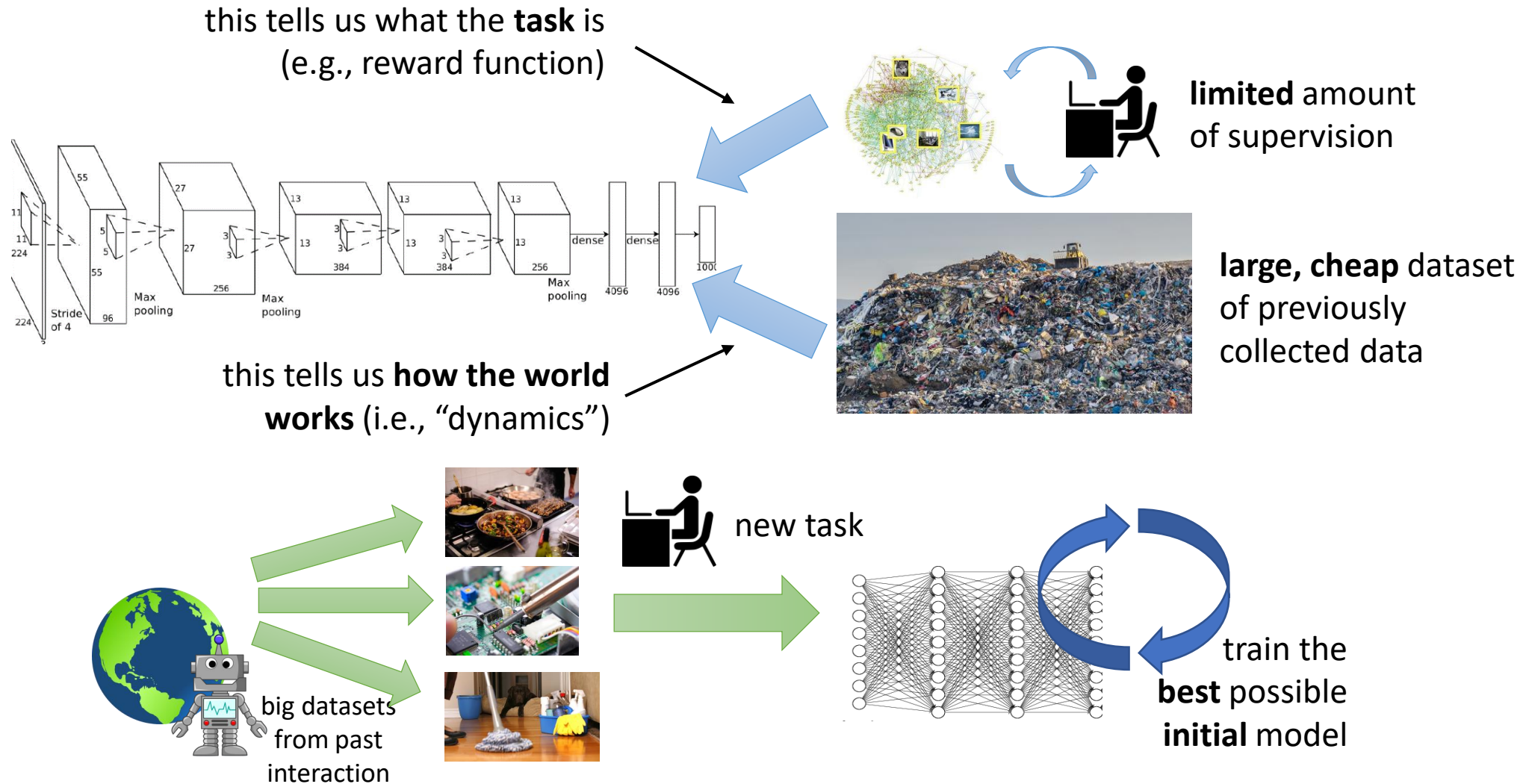


Daniel Wolpert
(knows quite a lot
about brains)



"We have a brain for one reason and one reason only – that's **to produce adaptable and complex movements**. Movement is the only way we have affecting the world around us... I believe that to understand movement is to understand the whole brain."

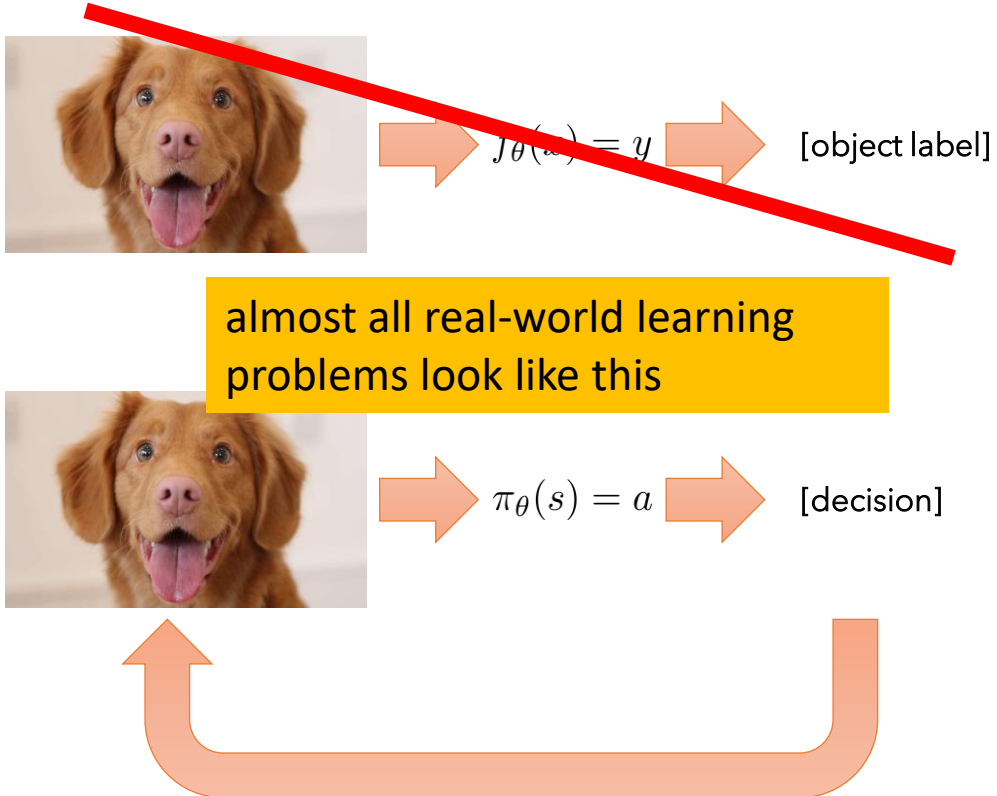
Reinforcement learning as a way to use “cheap” (previously collected) data



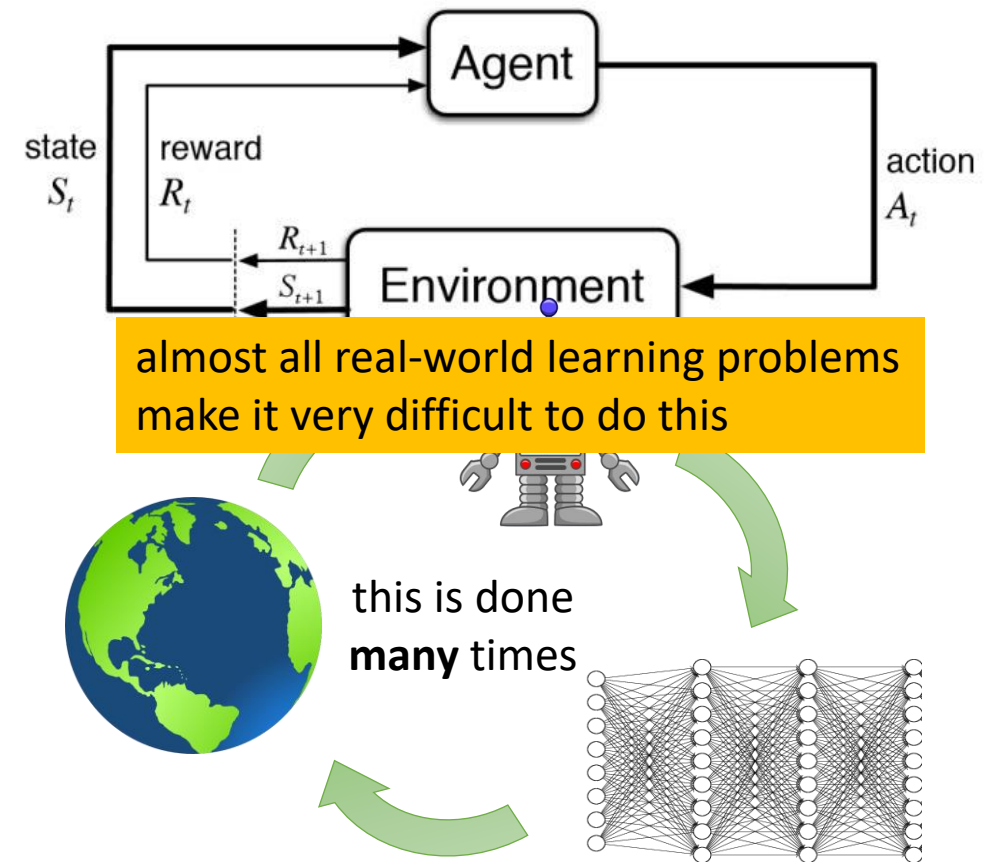
So what's the problem?

Reinforcement learning is two different things:

1. Framework for learning-based **decision making**

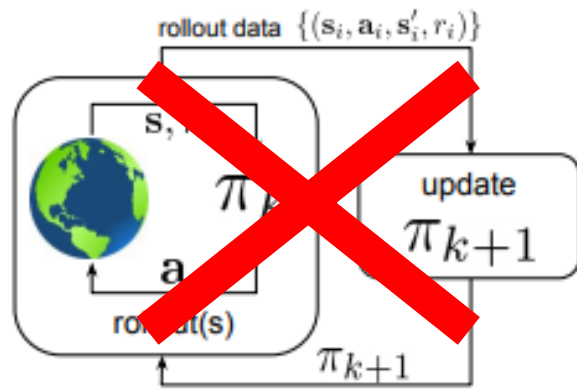


2. **Active, online** learning algorithms for control

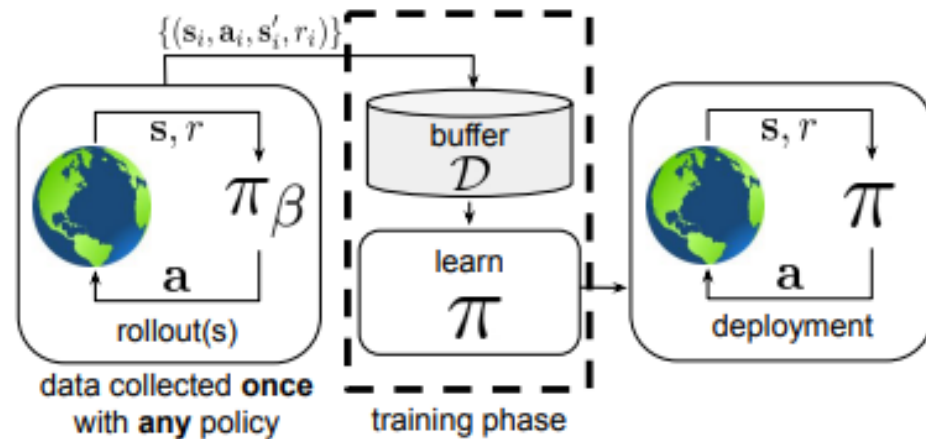


Making RL look more like data-driven learning

on-policy RL

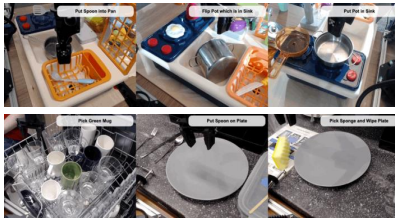


offline reinforcement learning

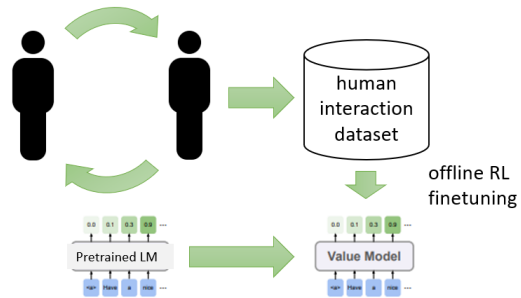




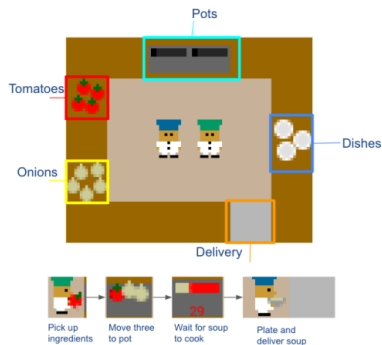
Offline RL fundamentals



Offline RL pretraining in robotics



Offline RL for large language models



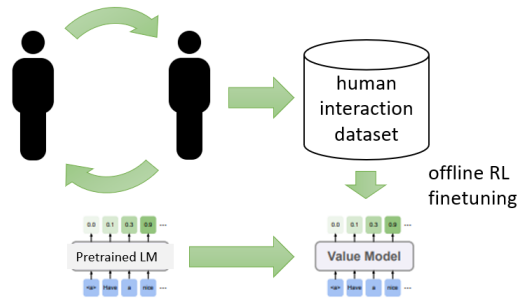
Offline RL for human influence



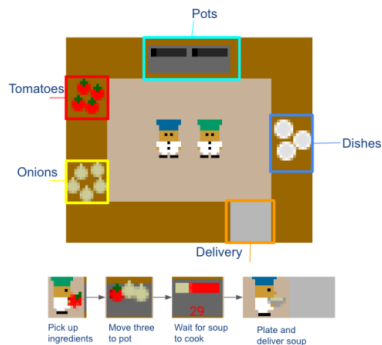
Offline RL fundamentals



Offline RL pretraining in robotics



Offline RL for large language models



Offline RL for human influence

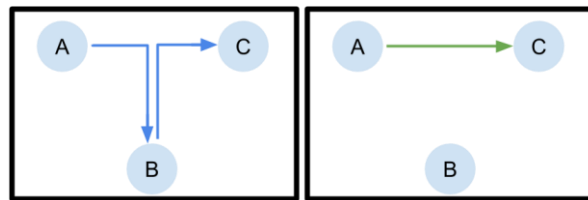
What do we expect offline RL methods to do?

Bad intuition: it's like imitation learning

Though it can be shown to be **provably** better than imitation learning even with optimal data, under some structural assumptions!

See: Kumar, Hong, Singh, Levine. Should I Run Offline Reinforcement Learning or Behavioral Cloning?

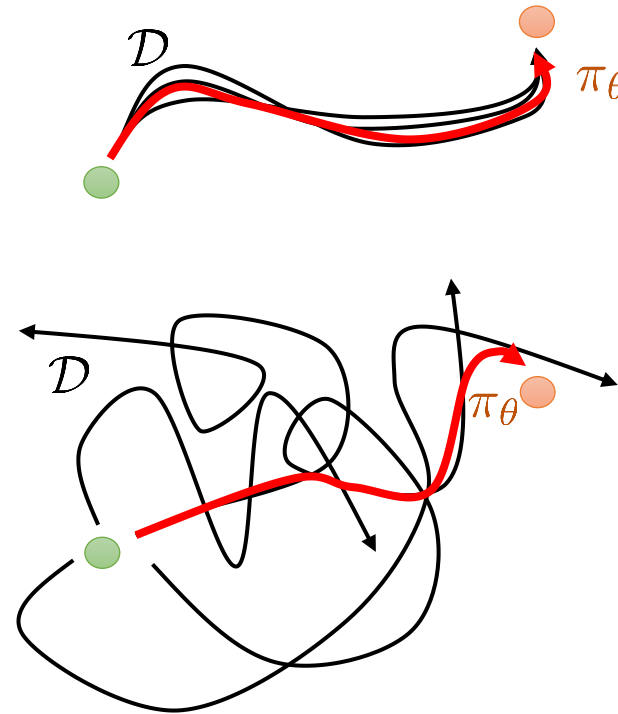
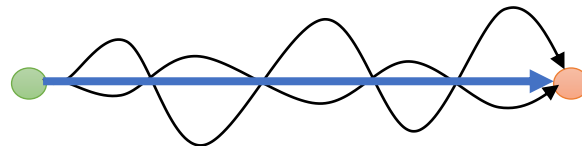
Better intuition: get order from chaos



“Macro-scale” stitching

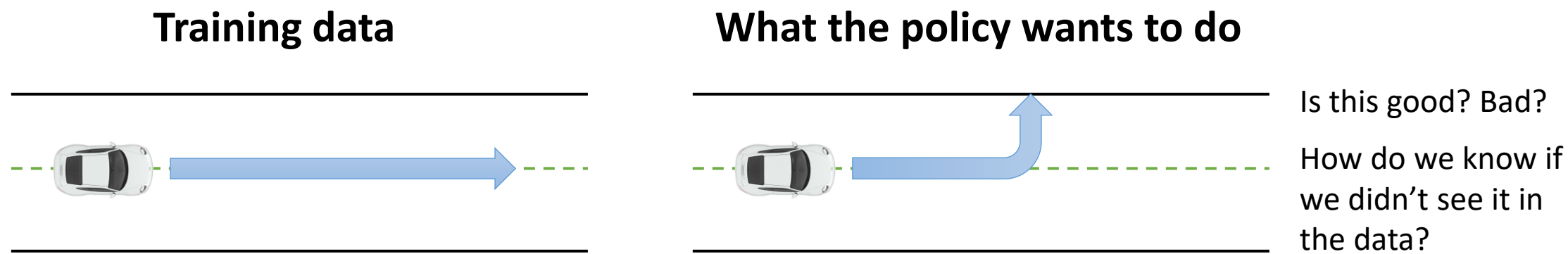
But this is just the clearest example!

“Micro-scale” stitching:



What is **hard** about offline RL?

Fundamental problem: counterfactual queries

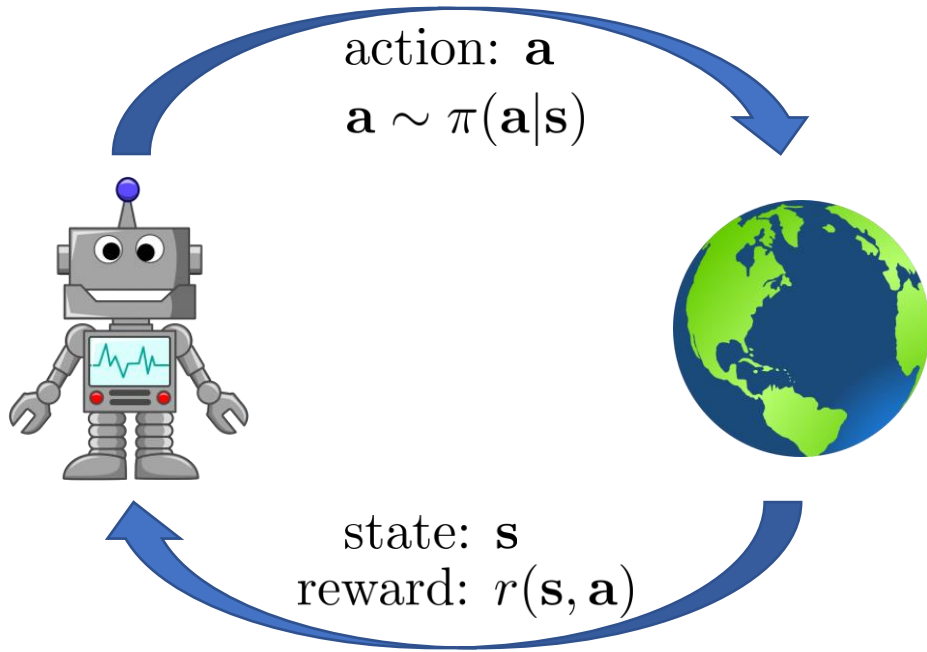


Online RL algorithms don't have to handle this, because they can simply **try** this action and see what happens

Offline RL methods must somehow account for these unseen ("out-of-distribution") actions, ideally in a safe way

...while still making use of generalization to come up with behaviors that are better than the best thing seen in the data!

Off-policy RL: a quick primer



RL objective: $\max_{\pi} \sum_{t=1}^T E_{\mathbf{s}_t, \mathbf{a}_t \sim \pi} [r(\mathbf{s}_t, \mathbf{a}_t)]$

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

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

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

enforce this equation at all states!

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

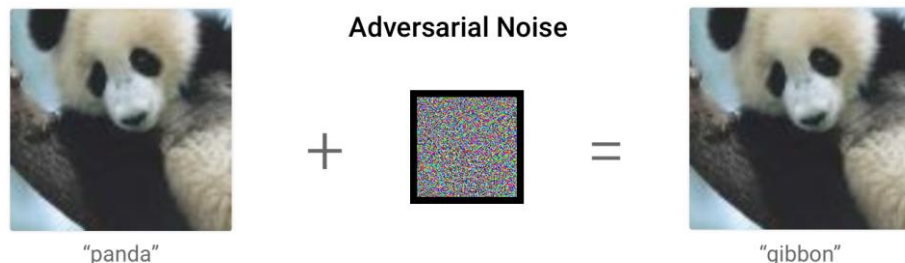
Where do we suffer from distribution shift?

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

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

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

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

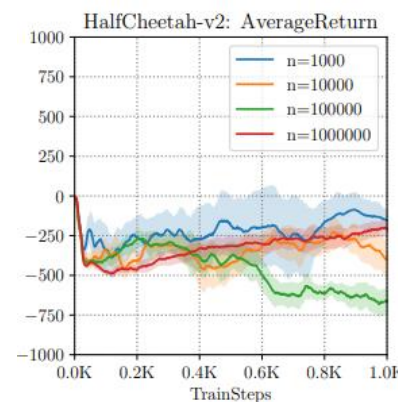


what is the objective?

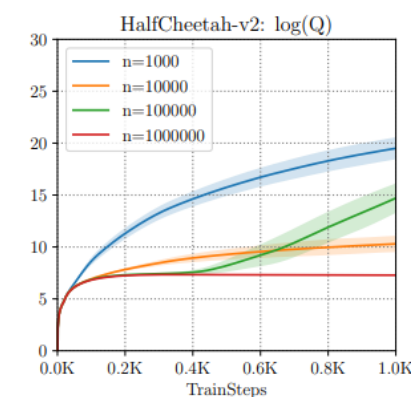
$$\min_Q E_{(\mathbf{s}, \mathbf{a}) \sim \pi_{\beta}(\mathbf{s}, \mathbf{a})} [(Q(\mathbf{s}, \mathbf{a}) - y(\mathbf{s}, \mathbf{a}))^2]$$

behavior policy target value

how often does *that* happen?

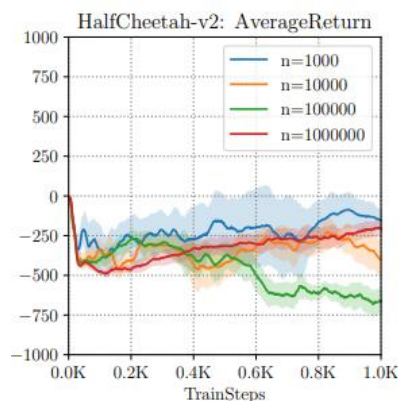


how well it does

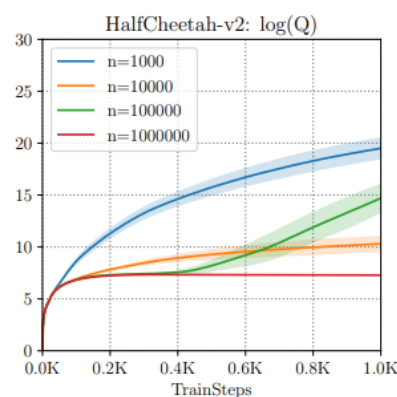


how well it *thinks*
it does (Q-values)

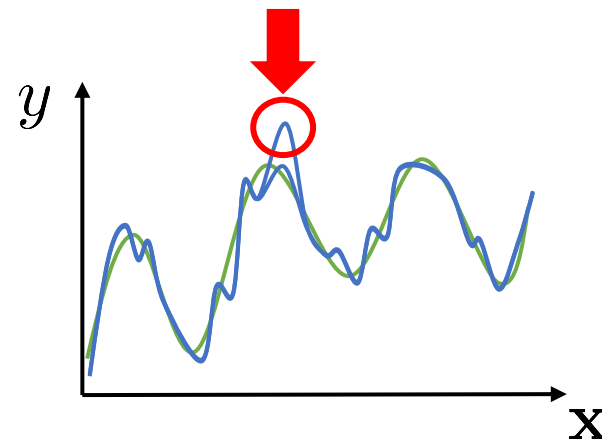
Conservative Q-learning



how well it does



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



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

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

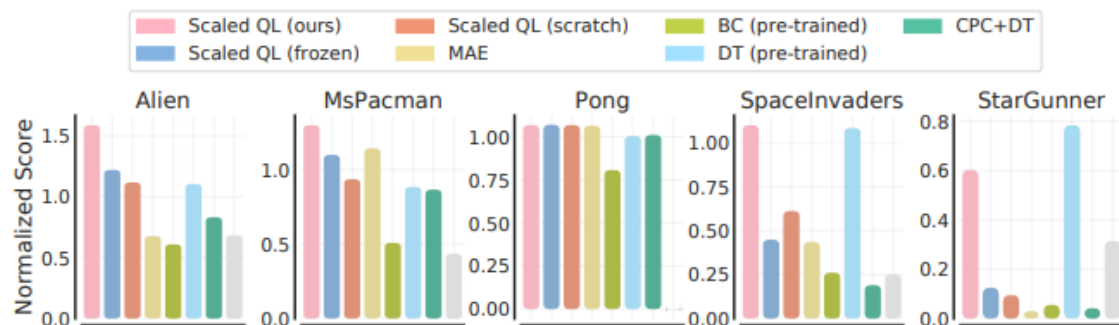
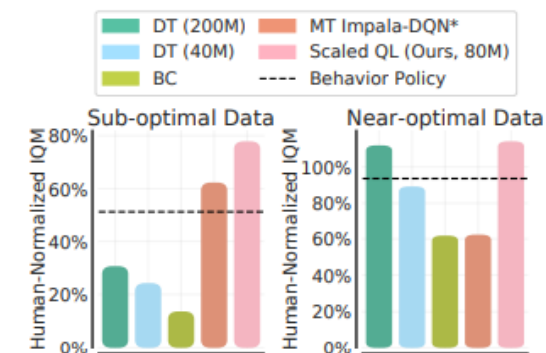
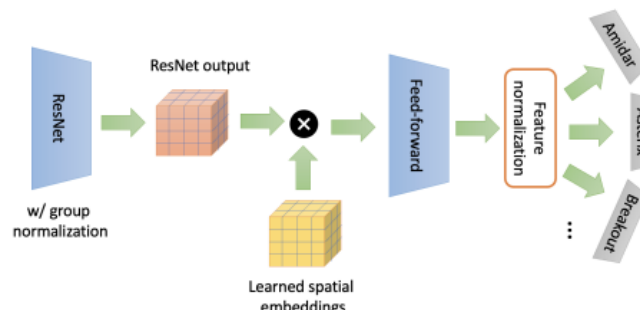
can show that $\hat{Q}^\pi \leq Q^\pi$ for large enough α

true Q-function

Application: multi-game Atari policies

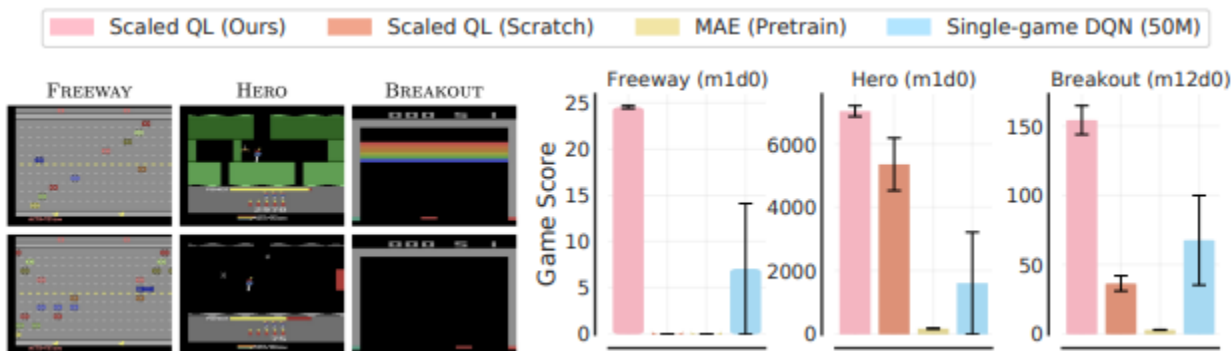


1. Train a single model on **40 different Atari games**



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

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



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

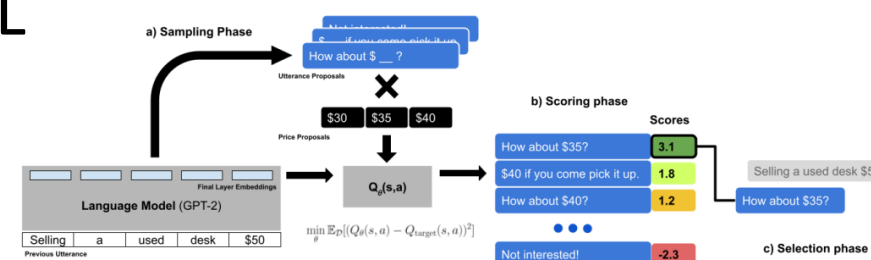
Offline RL & conservative Q-learning: applications and extensions

Robotics: offline training from diverse data



Dialogue systems: finetuning LLMs with offline RL

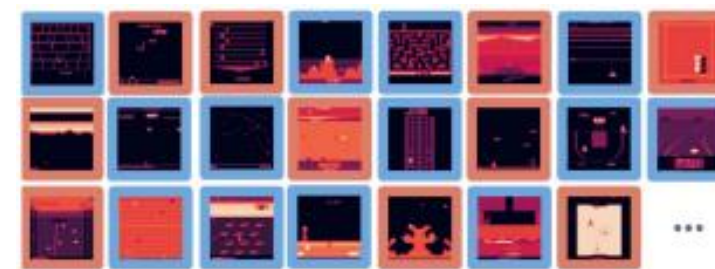
Atari games: state-of-the-art results learning policies to play many Atari games



Mobile notifications: optimizing notification policies at LinkedIn



Prabhakar et al. Multi-objective Optimization of Notifications Using Offline Reinforcement Learning.



Digital management systems (DMS): optimizing agreement conversion



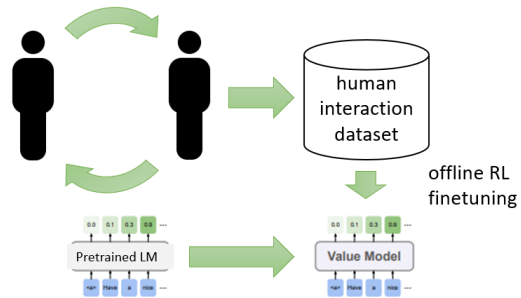
Soares et al. PulseRL: Enabling Offline Reinforcement Learning for Digital Marketing Systems via Conservative Q-Learning.



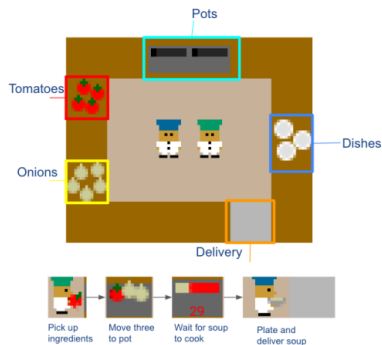
Offline RL fundamentals



Offline RL pretraining in robotics

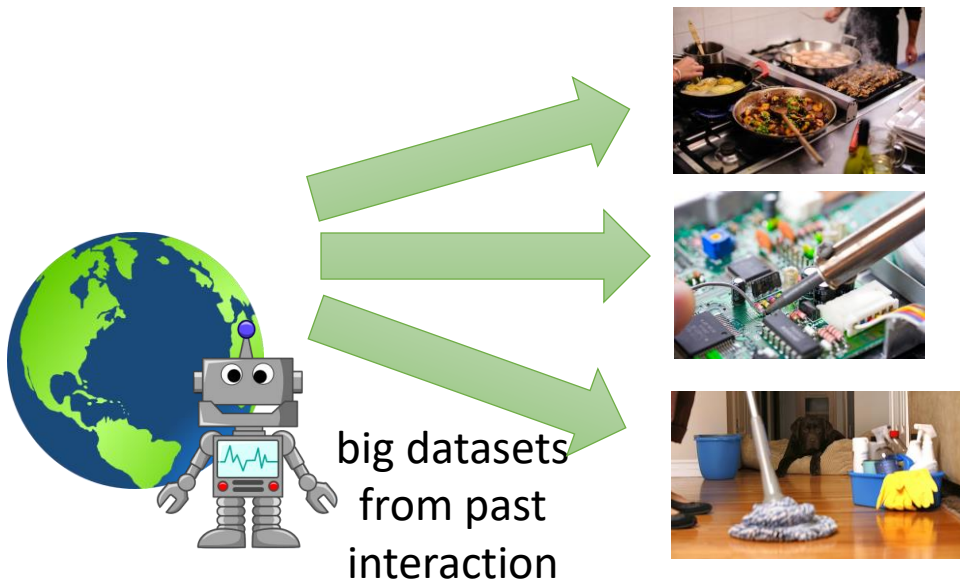


Offline RL for large language models

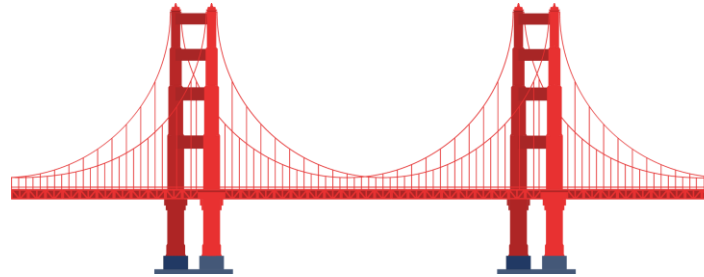


Offline RL for human influence

Offline data for robotic manipulation



- 12,000+ demonstrations
- 10 environments
- 70 tasks
- Designed to be reusable by other researchers in **new domains** and for **new tasks**



Pretraining on bridge data with offline RL (PTR)

Offline RL Pretraining on Bridge Dataset



Put Sushi in Pot



Put Eggplant on Plate



Put Eggplant on Plate



10 domains
100 tasks
12k demos

Offline RL Fine-tuning on Target Data + Bridge Data



put corn in bowl



Target dataset



batch-mixing bridge and target data



1. Pre-train on bridge data

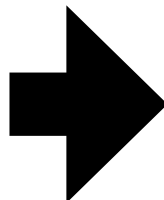


2. Fine-Tune on Mix of Bridge Data and Target Data

PTR (Pre-Training for Robots) Results

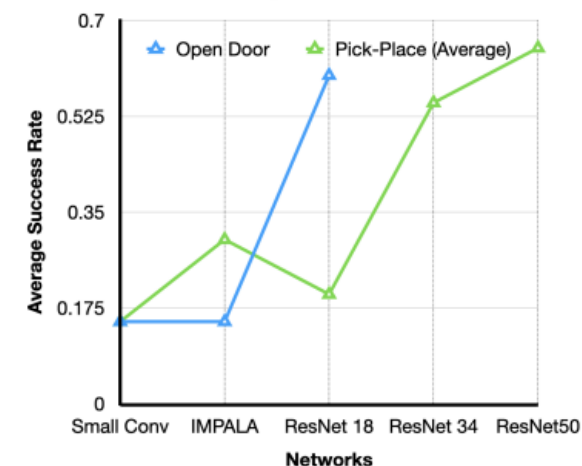


learning **entirely new** skills after pretraining on the bridge dataset



the data was **not** collected specifically for offline RL!

Scaling Trend for PTR



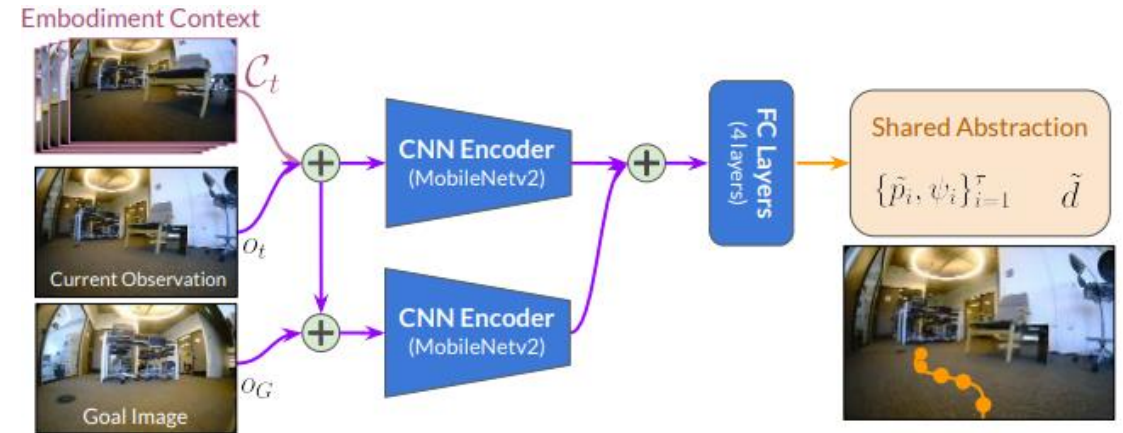
Task	BC finetuning				Joint training		Target data only		Pre-train. rep. + BC finetune	
	PTR (Ours)	BC (fine.)	Autoreg. BC	BeT	COG	BC	CQL	BC	R3M	MAE
Take croissant from metal bowl	7/10	3/10	5/10	1/10	4/10	4/10	0/10	1/10	1/10	3/10
Put sweet potato on plate	7/20	1/20	1/20	0/20	0/20	0/20	0/20	0/20	0/20	1/20
Place knife in pot	4/10	2/10	2/10	0/10	1/10	3/10	3/10	0/10	0/10	0/10
Put cucumber in pot	5/10	0/10	1/10	0/10	2/10	1/10	0/10	0/10	0/10	0/10

Pretraining for navigation

the question

Can we create a dataset and model that can generalize in **zero shot** to control entirely new robots?

the setup



RC-Car
(Kahn et al. 2018)



TurtleBot
(Hirose et al. 2019)



Jackal
(Shah et al. 2021, 2022)



Spot
(Karman et al. 2022)



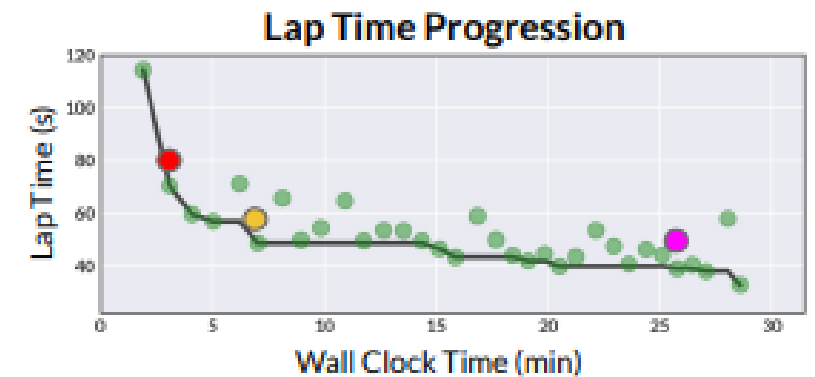
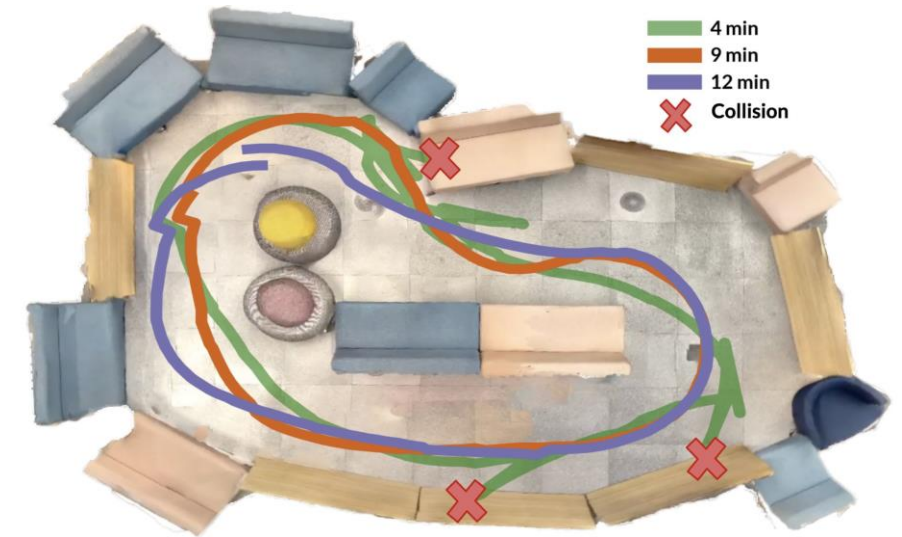
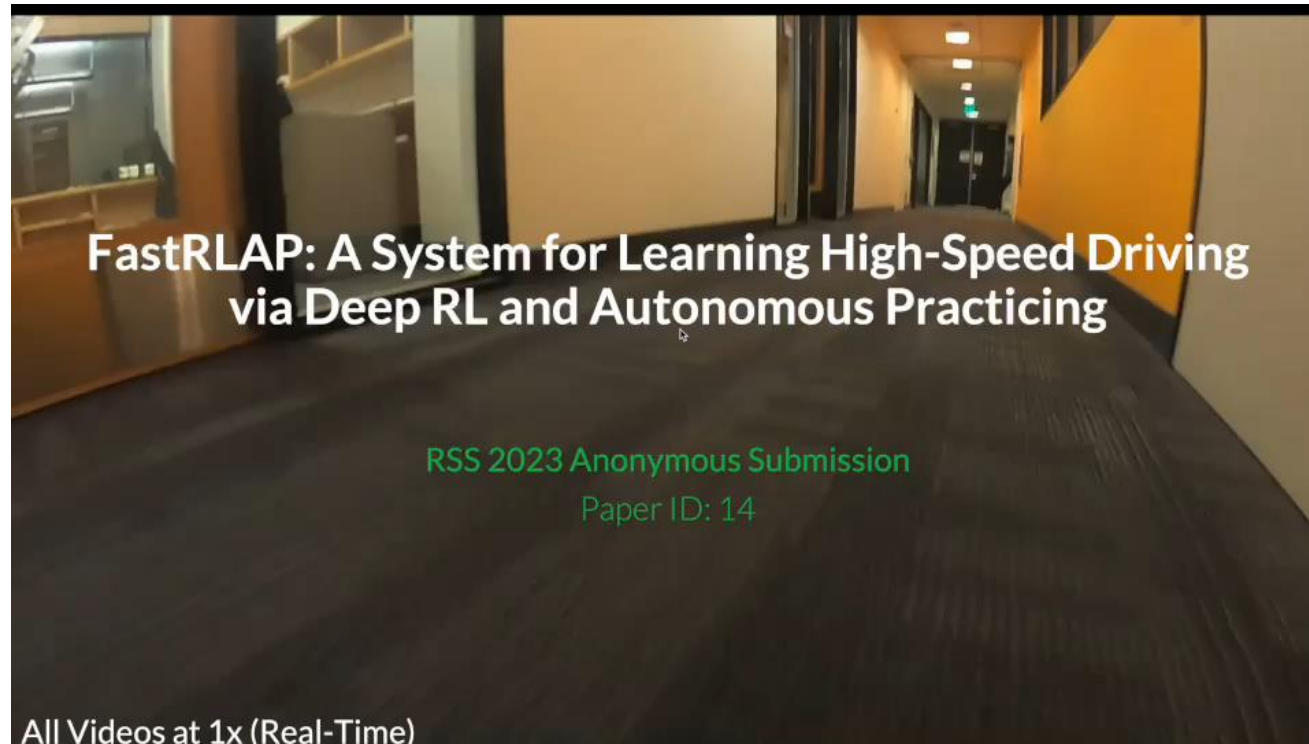
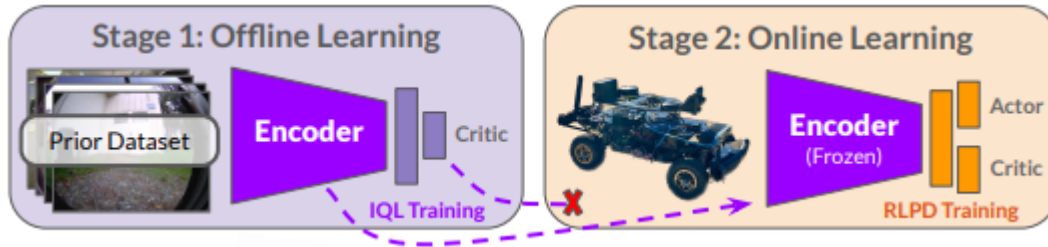
Warthog
(Shaban et al. 2021)



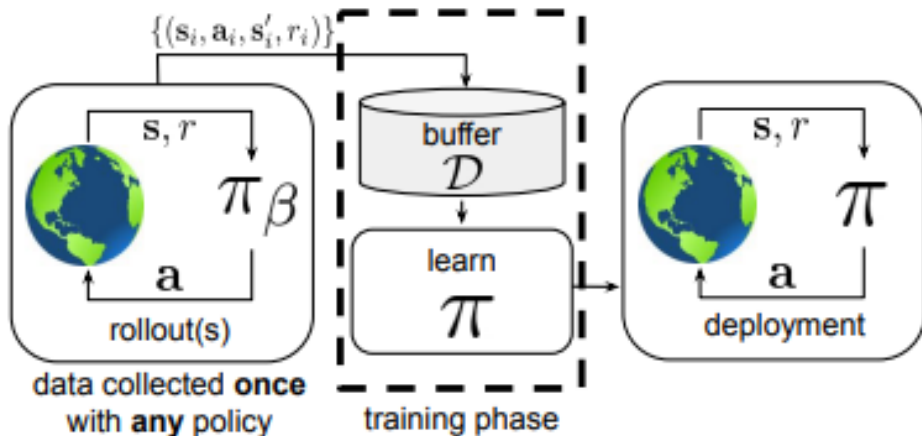
ATV
(Triest et al. 2022)



Now make it go fast!



What have we done, and what remains open?



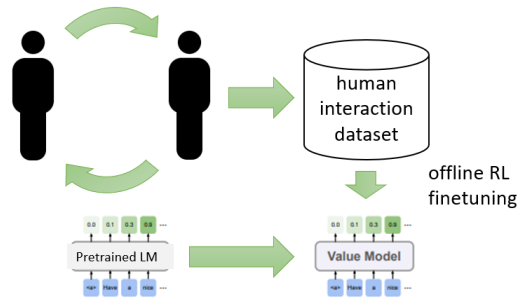
- Large quantities of robot data can give us initializations for downstream skills
- Offline RL serves as a powerful framework for **both** pretraining and finetuning
- Pretraining with offline RL (vs. “traditional” representation learning) really helps
- The skills are still simple
- The data is still narrow
- A lot more to be done in the future!



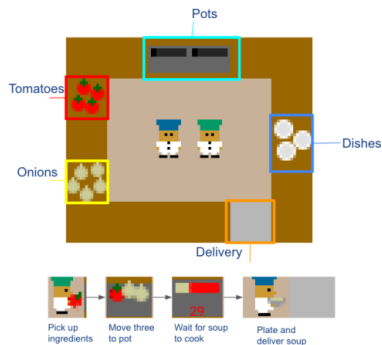
Offline RL fundamentals



Offline RL pretraining in robotics



Offline RL for large language models

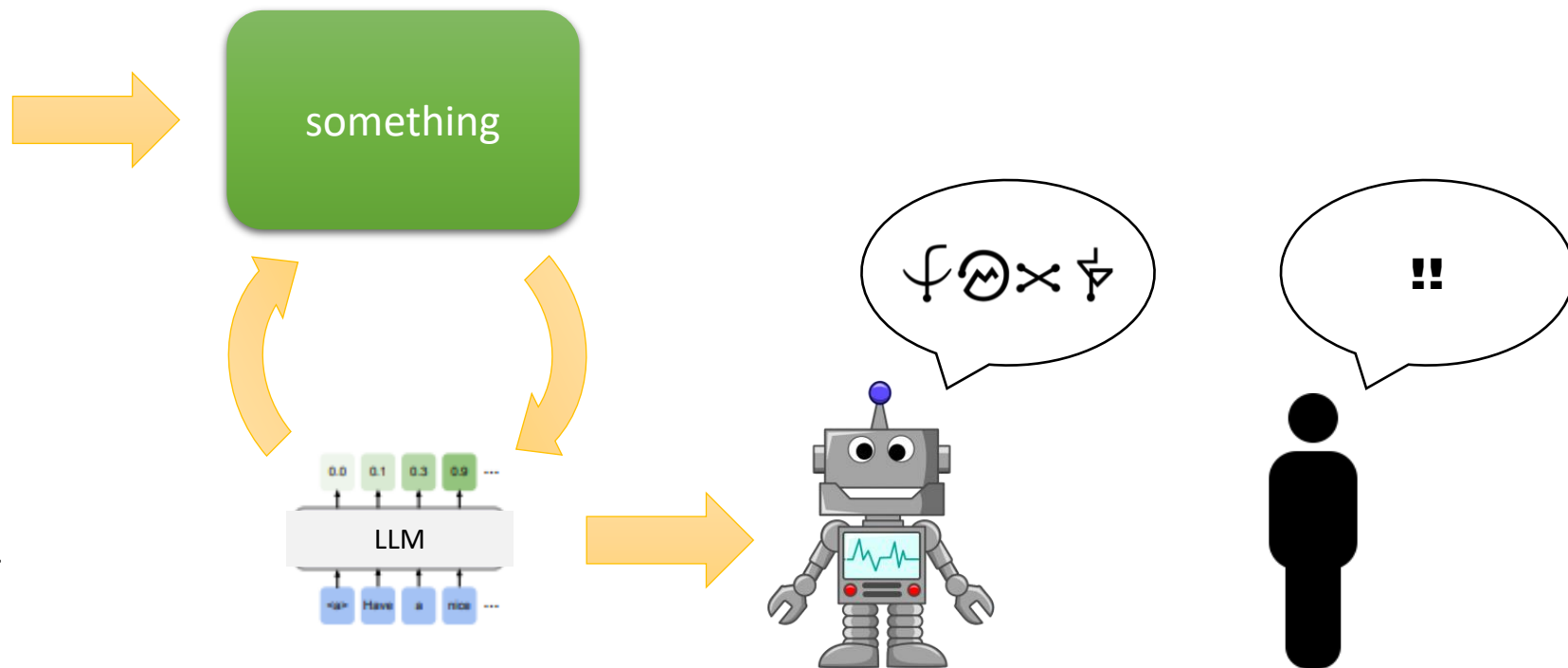


Offline RL for human influence

Learning to interact with humans

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?
	[12:21]	cucho	dell: ah not like that
03:45	[12:21]	RC	dell: you can't move the drives
03:45	[12:21]	Sender	Recipient
03:45	[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#)"
03:46	[12:21]	kuja	Taru Haha sucker.
	[12:22]	Taru	Kuja ?
		kuja	Taru Anyways, you made the changes right?
03:46	[12:22]	Taru	Kuja Yes.
		kuja	Taru Then from the terminal type: sudo apt-get update
		Taru	Kuja I did.

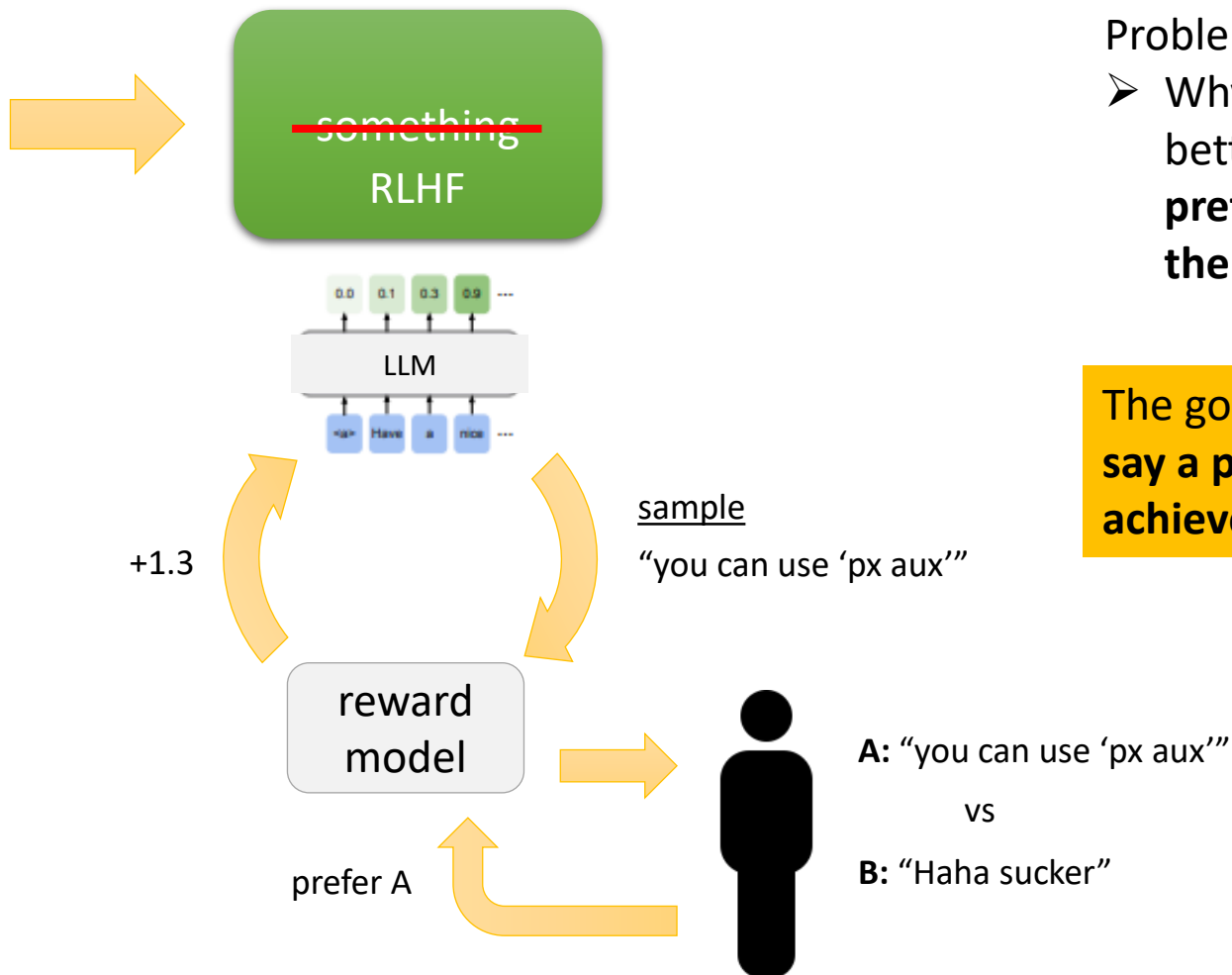
Lowe et al., "The Ubuntu Dialogue Corpus." 2015.



- 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**?

Learning to interact with humans via RL

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?
	[12:21]	cucho	dell: ah not like that
03:45	[12:21]	RC	dell: you can't move the drives
03:45	[12:21]	Sender	Recipient
03:45	[12:21]	Old	I dont run graphical ubuntu, I run ubuntu server.
	[12:21]	bur[n]er	Old
03:46	[12:21]		you can use "ps ax" and "kill (PID#)"
	[12:21]		
	[12:22]	kuja	Taru
		Taru	Kuja
		kuja	Taru
03:46			Anyways, you made the changes right?
	[12:22]	Taru	Kuja
		kuja	Taru
		Taru	Kuja
			Then from the terminal type: sudo apt-get update
			I did.

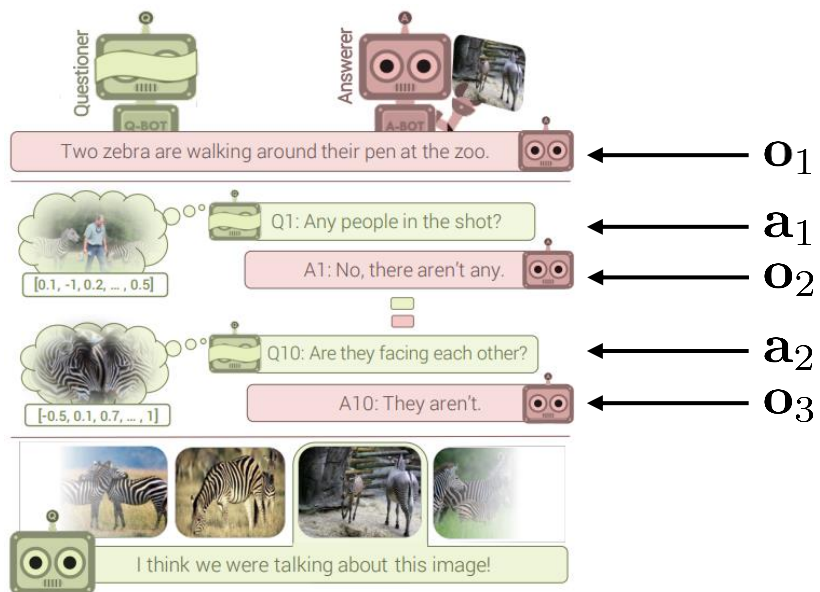


Problem:

- 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**

The sequential structure of dialogue



Das et al. Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning. 2017.

LMs model this

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



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

Extremely complex (essentially a person's brain)

But wait...

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

RLHF

learn from human preferences

episode = single utterance

This is **not** RLHF

sequential decision making

learn from **dialogue outcome**

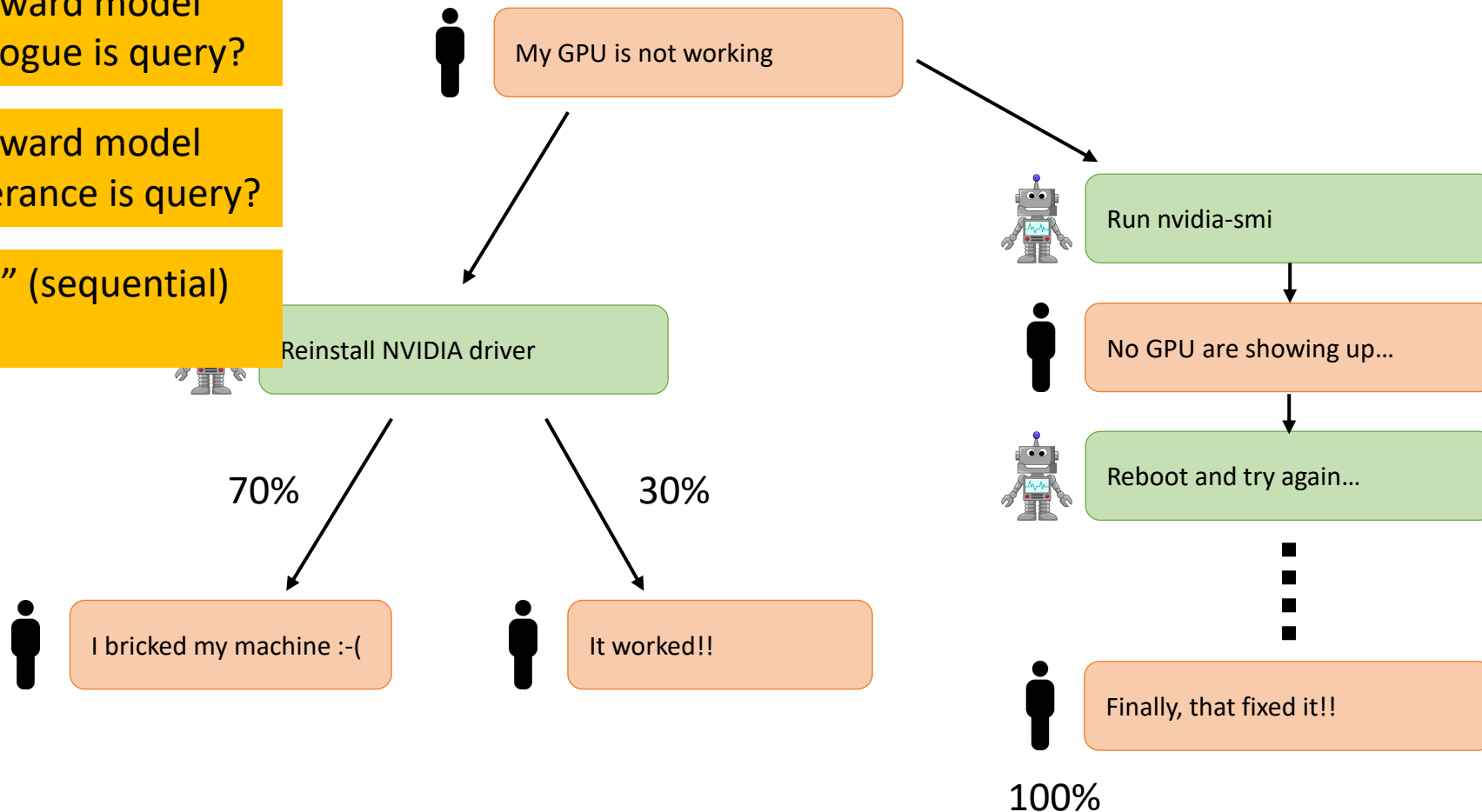
episode = whole dialogue

Understanding cause and effect

What would a reward model pick if entire dialogue is query?

What would a reward model pick if single utterance is query?

What about “full” (sequential) RL?



Learning to interact with humans via offline RL

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?
	[12:21]	cucho	dell: ah not like that
03:45	[12:21]	RC	dell: you can't move the drives
03:45	[12:21]	Sender	Utterance
03:45	[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#)"
03:46	[12:21]		
	[12:21]	kuja	Taru Haha sucker.
	[12:22]	Taru	Kuja ?
03:46		kuja	Taru Anyways, you made the changes right?
	[12:22]	Taru	Kuja Yes.
		kuja	Taru Then from the terminal type: sudo apt-get update
		Taru	Kuja I did.



~~something~~
offline RL

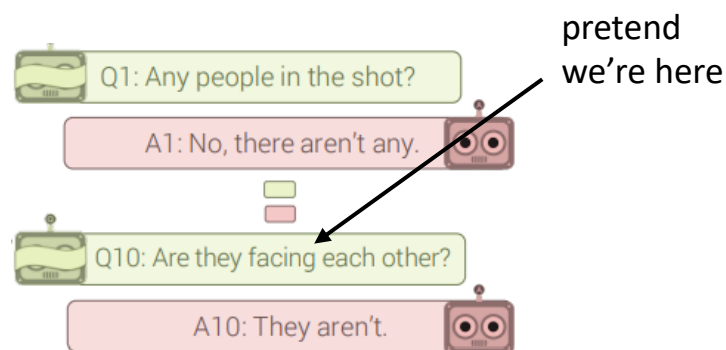
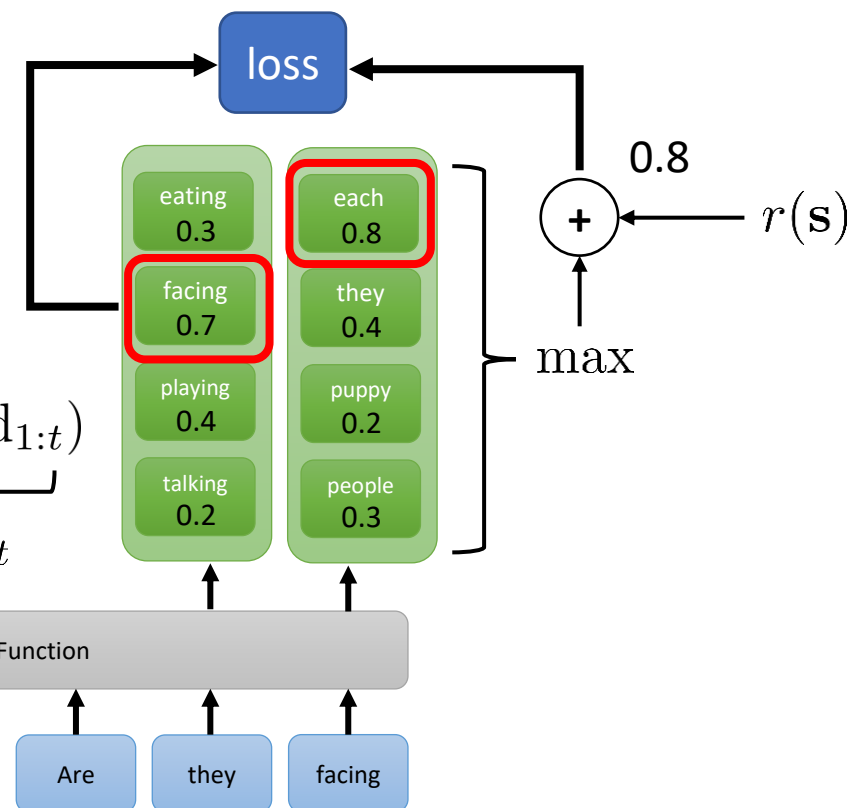
$$Q(s, a) \leftarrow r(s) + \gamma \max_{a'} Q(s', a')$$

Let's compare:

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

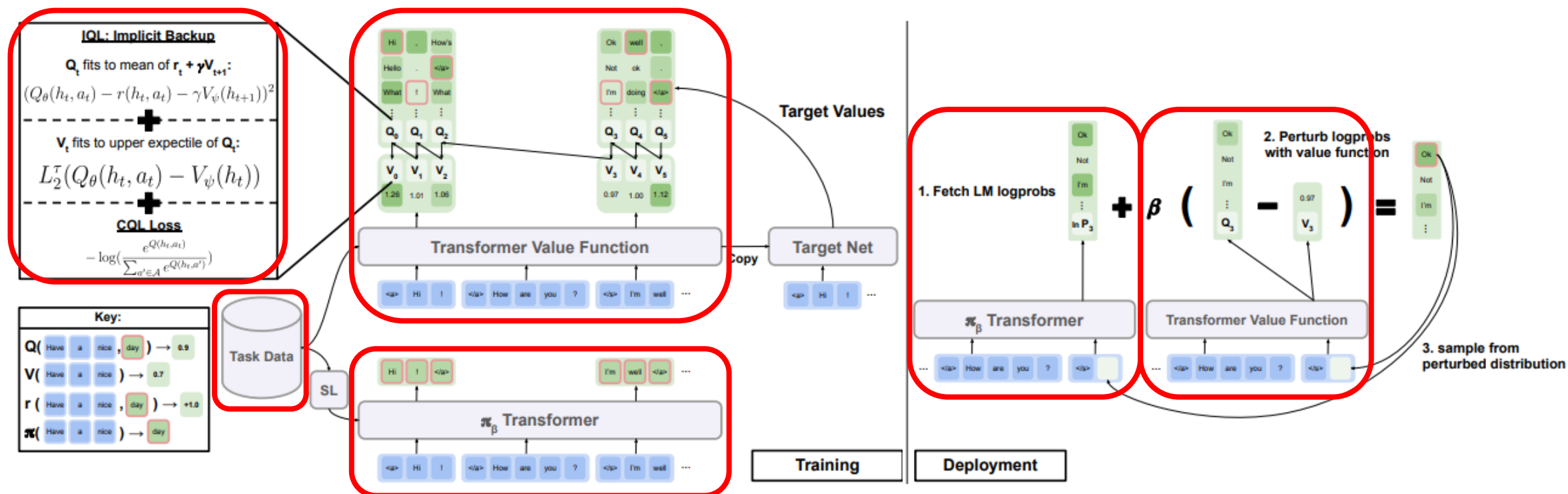
value

$$\text{function: } p(\underbrace{\text{success}}_{Q(s_t, a_t)} | \underbrace{\text{word}_{t+1}}_{a_t}, \underbrace{\text{word}_{1:t}}_{s_t})$$



Implicit language Q-learning (ILQL)

Basic idea: combination of IQL (implicit Q-learning) and CQL (conservative Q-learning) for training LLMs



Questioner: how many people? (everything starts with pretrained models)

Answerer: 2

Questioner: what gender are people?

Answerer: 1 is male and 1 is female

Questioner: what are they wearing?

Evaluation on visual dialogue

Standard Reward Function:

Image Caption: A man, with other people in the background, is sitting at a picnic table with a cell phone in front of him while he uses a fork and knife to cut a dessert on a plate.

Questioner: is man young?

Answerer: yes

Questioner: is he wearing glasses?

Answerer: no

Questioner: is there food on table?

Answerer: yes

Questioner: is there food on table?

Answerer: yes

Yes/No Penalty:

Image Caption: A person flying a kite over a sand dune.

Questioner: what color is kite?

Answerer: it's black and white

Questioner: is person male or female?

Answerer: can't tell

Questioner: what is person wearing?

Answerer: i can't tell

Questioner: what color is sand?

Answerer: dark brown

Questioner: is person flying kite flying in air?

Answerer: yes

Questioner: what color is kite?

Answerer: black

Conservative Yes/No Penalty:

Image Caption: Tour buses are lined up on the street waiting for people.

Questioner: how many buses?

Answerer: 2

Questioner: what color are buses?

Answerer: white and red

Questioner: how many people?

Answerer: 2

Questioner: what gender are people?

Answerer: 1 is male and 1 is female

Questioner: what are they wearing?

Answerer: 1 is wearing shorts and other is wearing shorts and shirt

Questioner: what color is their hair?

Answerer: dark brown

method	standard	y/n	conservative y/n
ILQL	-5.22 $\pm$ 0.13	-5.69 $\pm$ 0.13	-6.57 $\pm$ 0.18
SARSA	-5.14 $\pm$ 0.13	-6.19 $\pm$ 0.15	-7.77 $\pm$ 0.20
%BC	-5.07 $\pm$ 0.13	-7.48 $\pm$ 0.21	-9.13 $\pm$ 0.22
BC	-5.25 $\pm$ 0.13	-10.85 $\pm$ 0.27	-15.16 $\pm$ 0.35

Evaluation on comment toxicity

Context: previous comment on reddit

Goal: generate a new comment that...

avoids “toxic” behavior

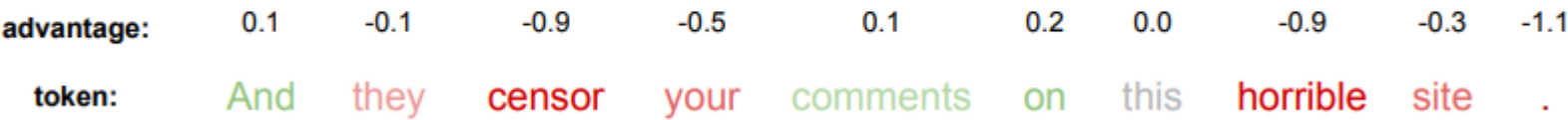
is likely to receive upvotes

highly stochastic reward

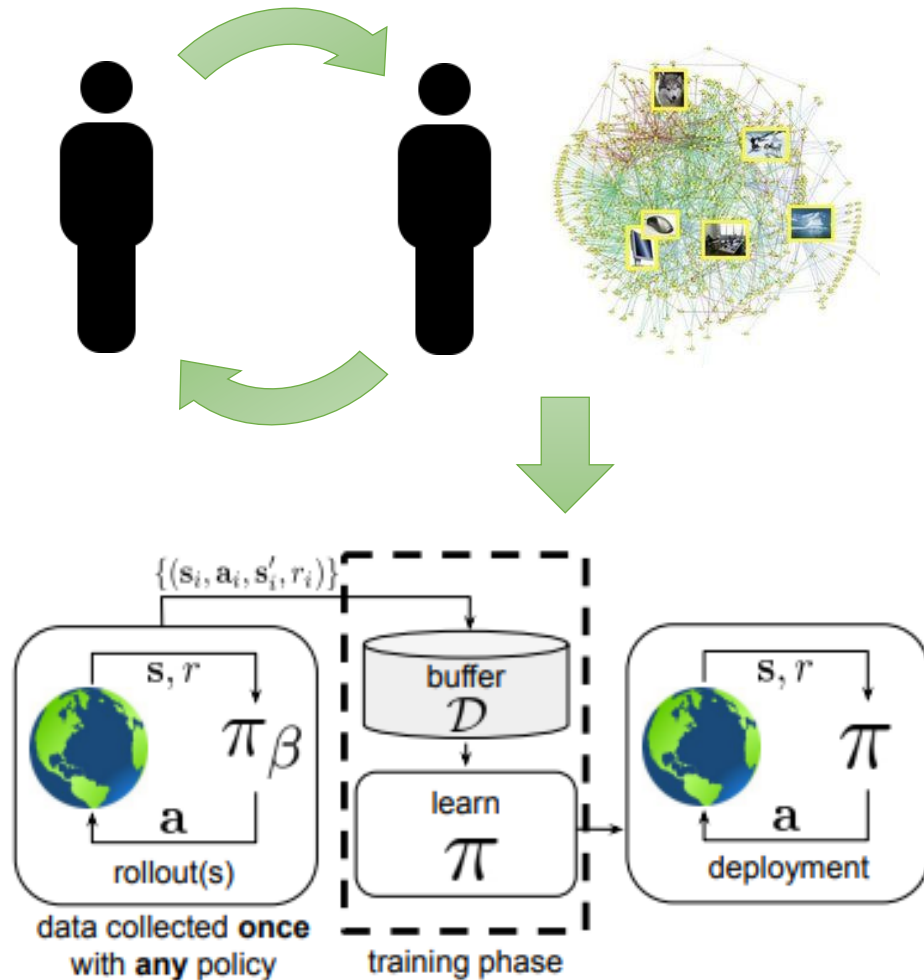


method	toxicity	upvotes real	upvotes model
ILQL	0.0 ±0.0	9.83 ±0.04	10.0 ±0.0
SARSA	0.0 ±0.0	6.23±0.15	10.0 ±0.0
%BC	-0.74±0.07	7.06±0.14	7.86±0.13
BC	-3.51±0.13	4.87±0.16	4.87±0.16

ILQL per-token advantages for toxic comments generated by filtered finetuning model



What have we done, and what remains open?



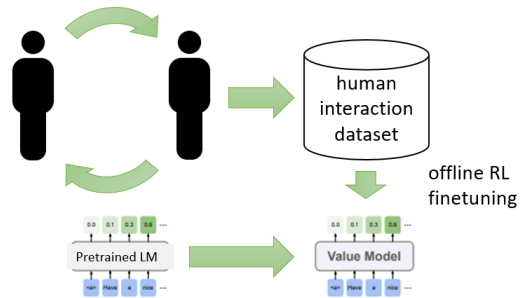
- Offline RL can learn from humans interacting with other humans
- This can be used to train large language models to achieve goals
- Evaluations are still at a small scale – the big benefit of large language models and other models that leverage human-human data is that they can pick up on subtle patterns in behavior
- Remains to be seen in future work how far LLM training with offline RL can go!



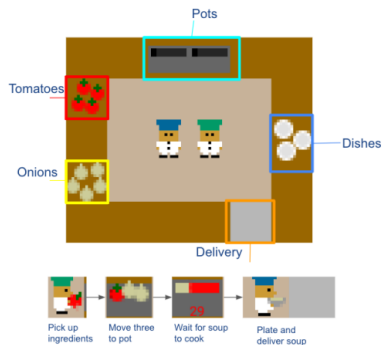
Offline RL fundamentals



Offline RL pretraining in robotics

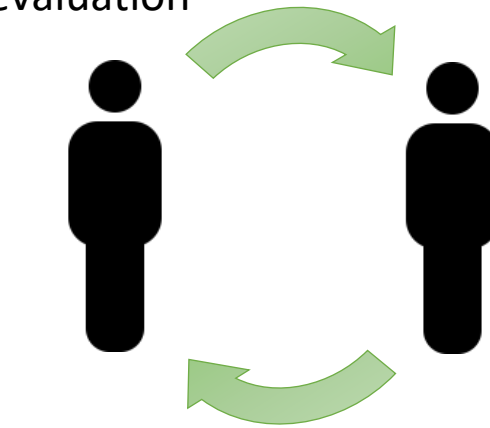
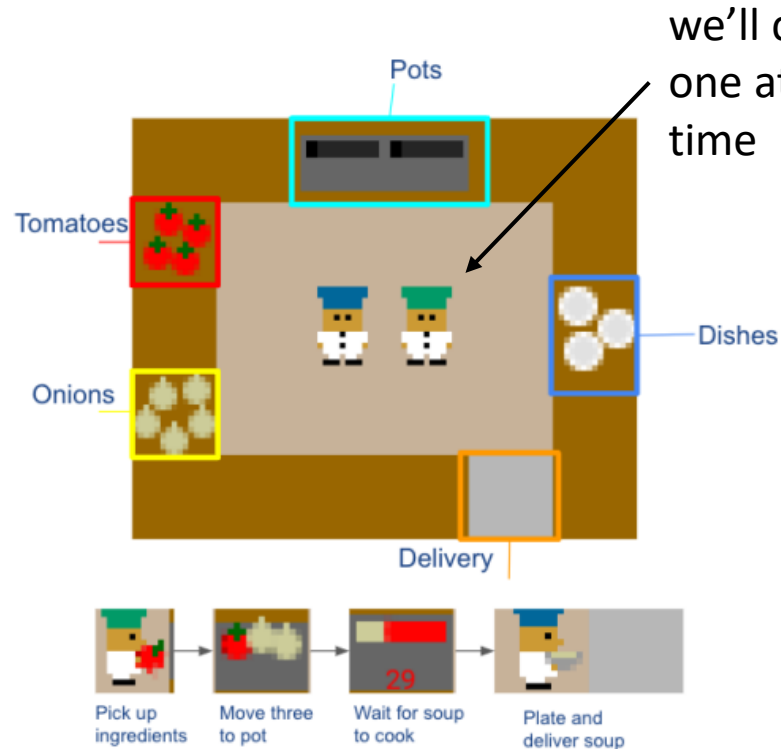


Offline RL for large language models



Offline RL for human influence

Influencing humans to be better cooks



The catch: tomato soup is worth more than onion soup

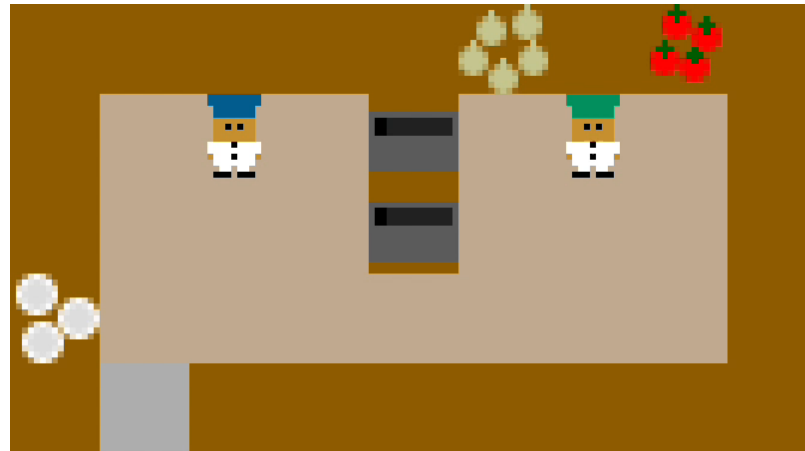
- Humans don't know this
- But the reward function tells our algorithm this

- Humans “accidentally” influence each other during normal play
- They sometimes make one type of soup, sometimes the other
- By analyzing lots of games, offline RL can figure out what causes a player to do one thing vs. another

Can offline RL influence humans to get them to play better?

Influencing humans to be better cooks

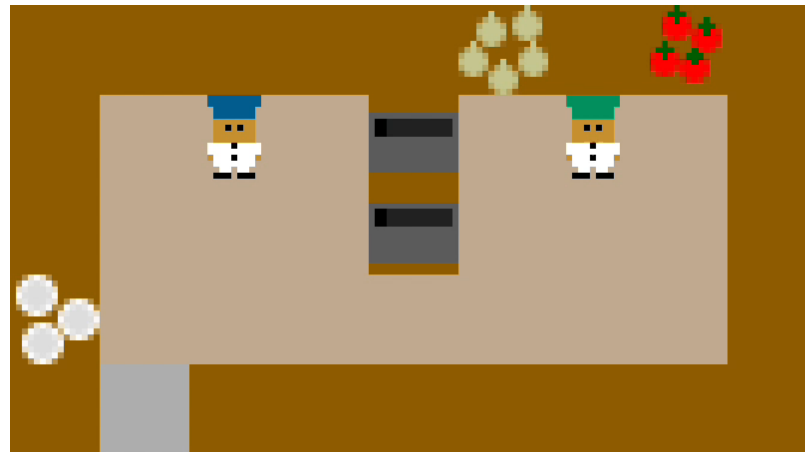
Scenario 1: direct physical influence



Scenario 2: make sure they get the hint



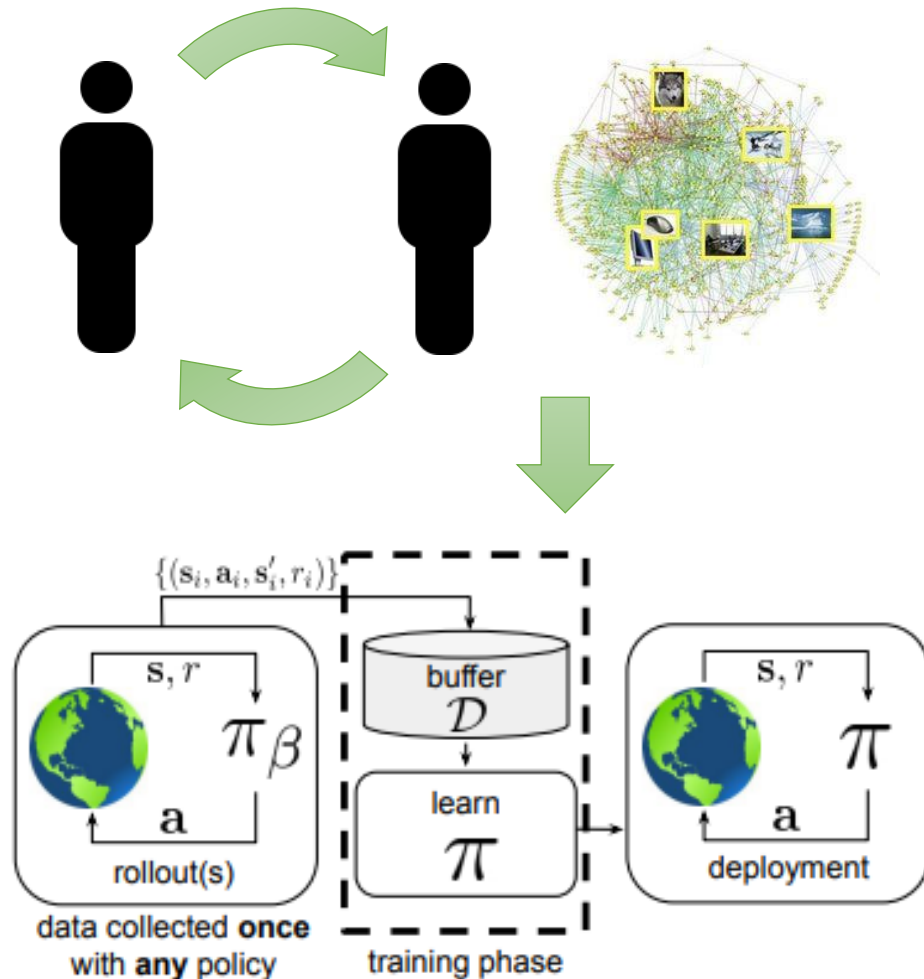
Naïve baseline
(imitation learning)



Offline RL
(same data)



Human influence and its implications



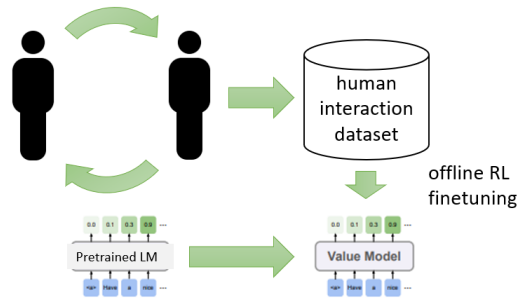
- Offline RL can pick up on how humans **accidentally** influence each other, and figure out how bots can **intentionally** influence humans
- We showed that this works even when the dataset doesn't contain **overt** and **intentional** influence
- We showed that bots can influence humans to perform **new** tasks that differ (a little) from the data
- Larger datasets will let us discern more subtle and interesting patterns
- Obviously this is ripe for abuse, and we should think about how to develop reward functions that promote **honesty** and other desirable traits



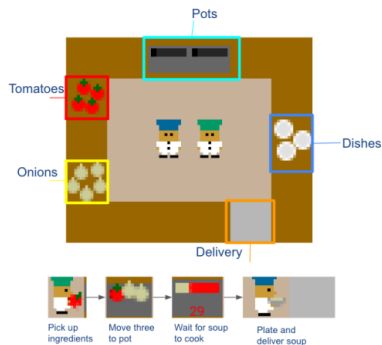
Offline RL fundamentals



Offline RL pretraining in robotics



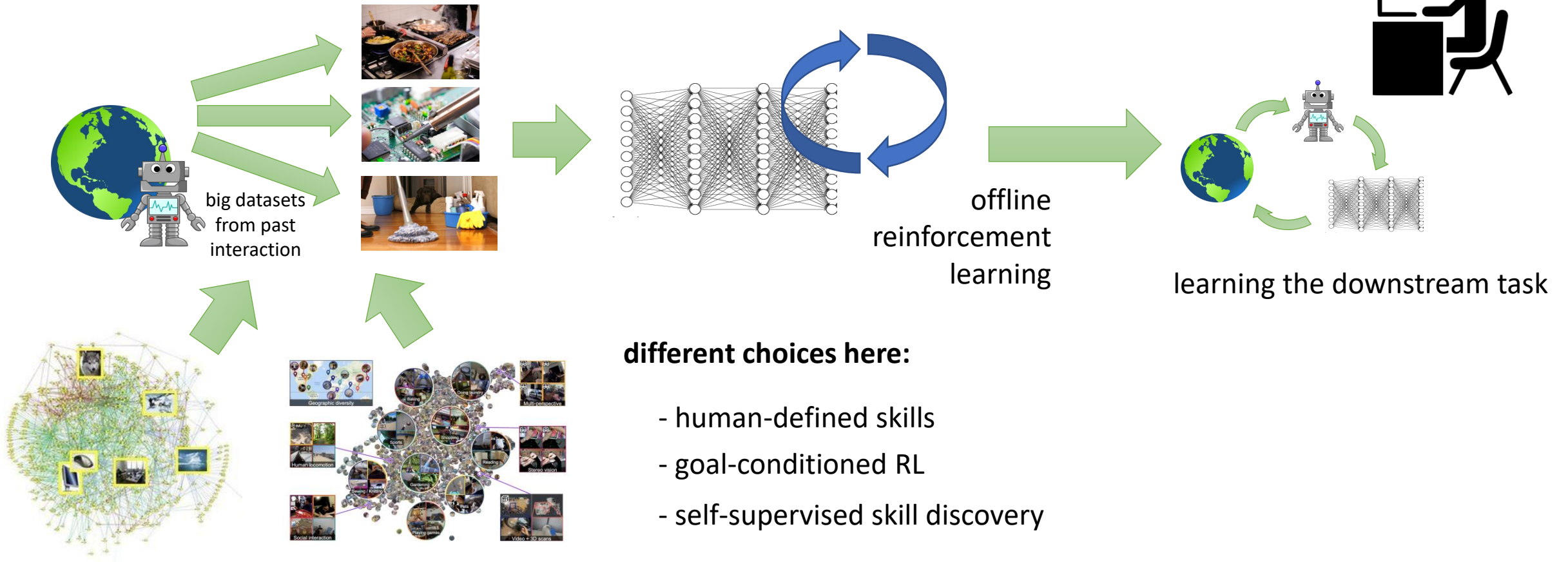
Offline RL for large language models



Offline RL for human influence

A hypothetical offline RL pipeline?

We need machine learning for one reason and one reason only – that's to produce adaptable and complex decisions.





RAIL
Robotic AI & Learning Lab

website: <http://rail.eecs.berkeley.edu>
source code: <http://rail.eecs.berkeley.edu/code.html>