

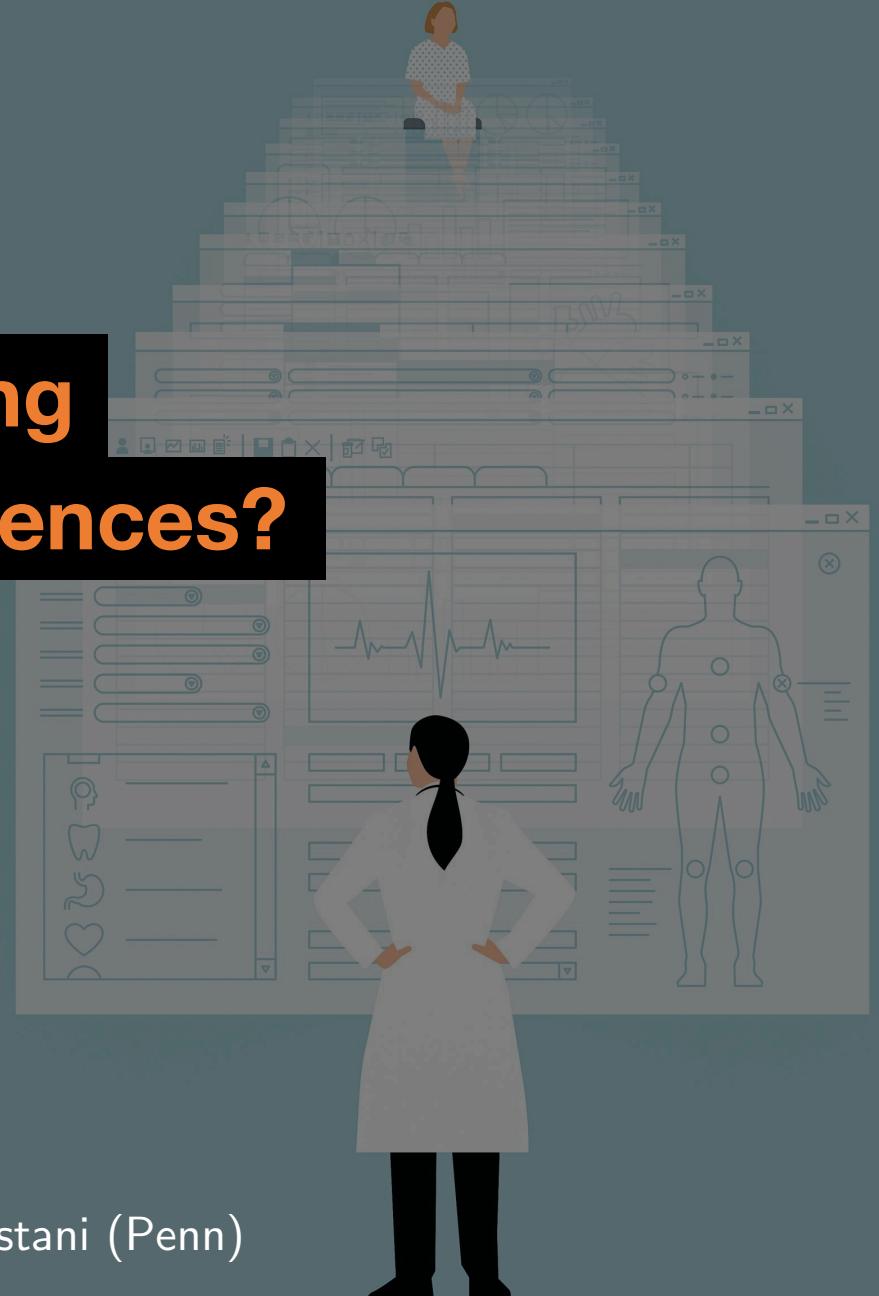
Learning Best Practices

Can Machine Learning
Substitute for Experiences?

INFORMS 2019

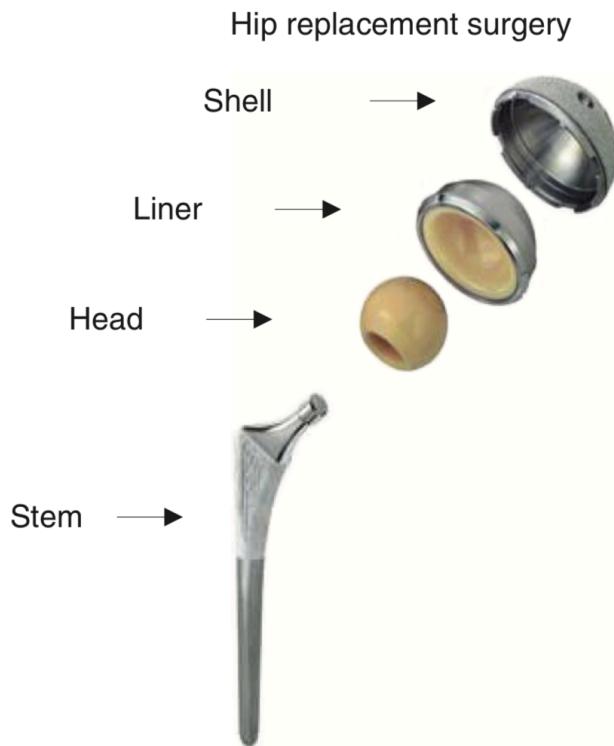


Park Sinchaisri (Wharton)
with Hamsa Bastani (Wharton), Osbert Bastani (Penn)



Learning Takes Time

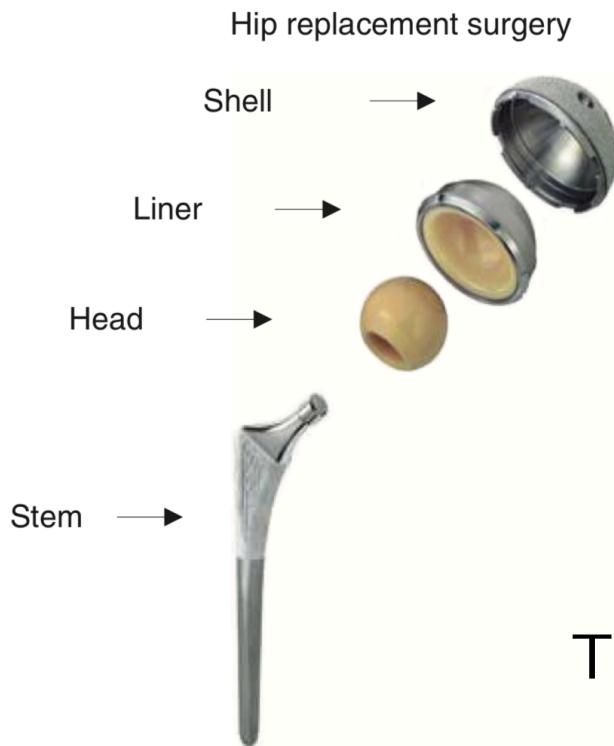
Learning Takes Time



“The first use of certain device versions can result in at least a **32.4% increase in surgery duration**, hurting quality and productivity.”

- Ramdas et al. 2018

Learning Takes Time

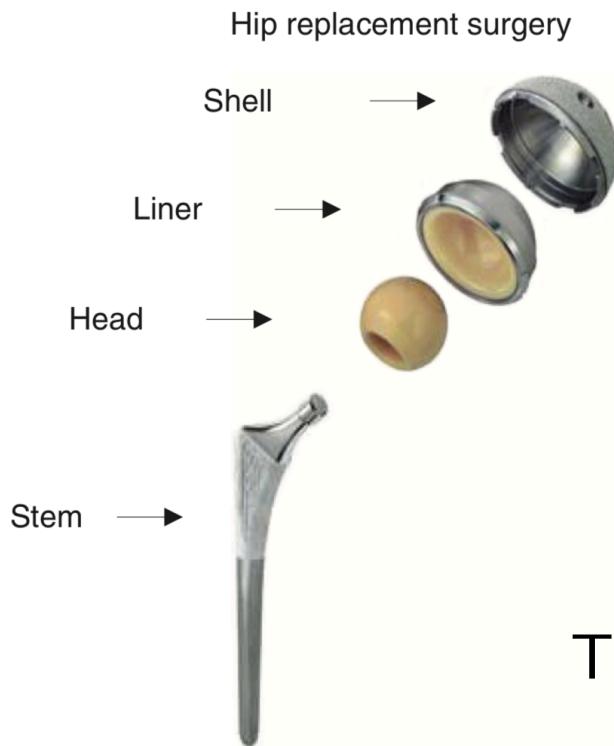


“The first use of certain device versions can result in at least a **32.4% increase in surgery duration**, hurting quality and productivity.”

- Ramdas et al. 2018

Training workers is really important, but there is a big cost upfront.

Learning Takes Time



“The first use of certain device versions can result in at least a **32.4% increase in surgery duration**, hurting quality and productivity.”

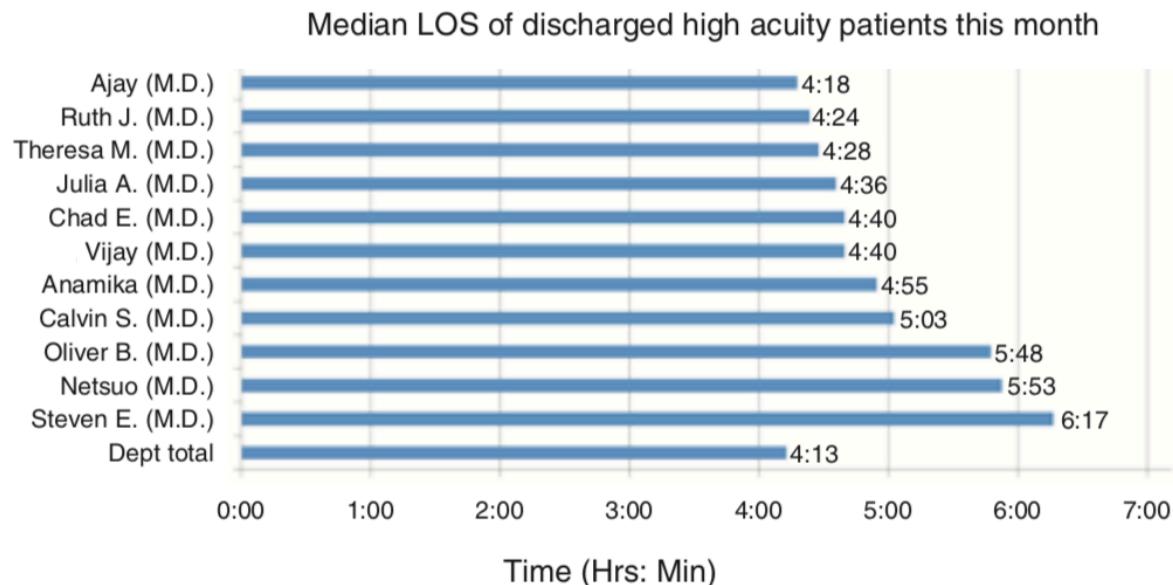
- Ramdas et al. 2018

Training workers is really important, but there is a big cost upfront.

Simple Tips?

Learning from Experts

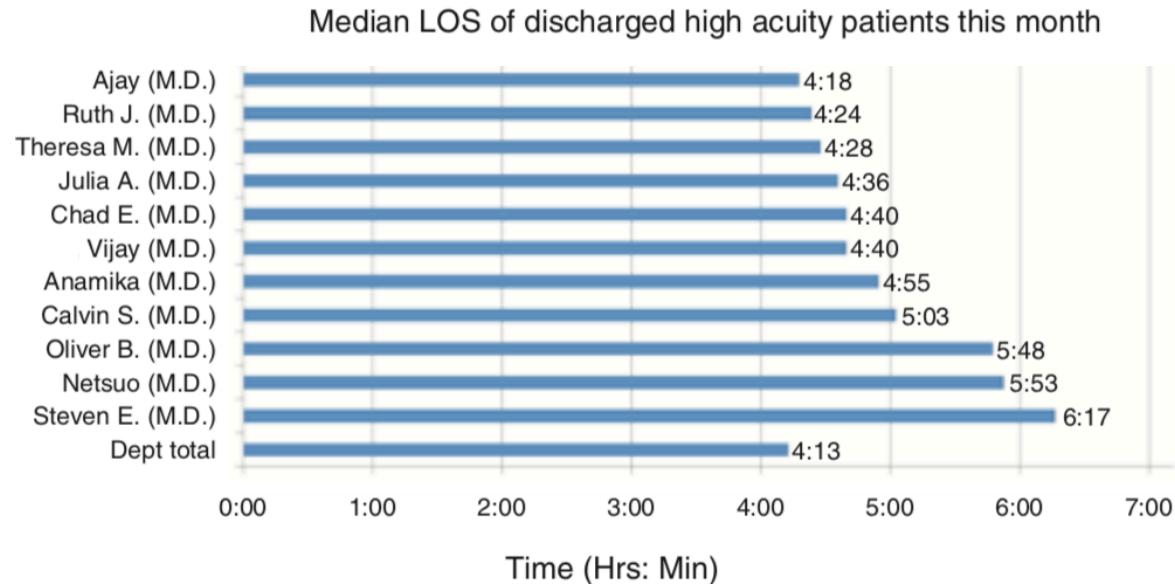
Learning from Experts



“The public disclosure of RPF allowed workers to identify their top-performing coworkers, which in turn enabled the identification and validation of best practices...”

- Song et al. 2018

Learning from Experts



“The public disclosure of RPF allowed workers to identify their top-performing coworkers, which in turn enabled the identification and validation of best practices...”

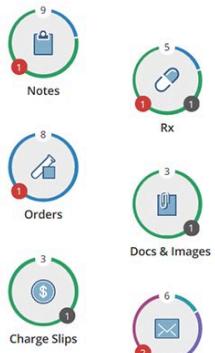
- Song et al. 2018

But It's Still Hard

Trace Data is Everywhere

Trace Data is Everywhere

Physicians

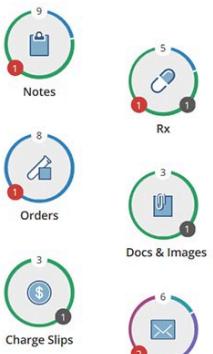


<input type="checkbox"/> • ROACH,TRISTIN	Lipitor 80 mg		MILLER,ALEX,MD status: Unreviewed	05-18-17
<input type="checkbox"/> • LEON,ERIN	Geriatric Wellness Visit		JONES,CAMERON,MD status: Unreviewed	05-16-17
<input type="checkbox"/> • BECK,ALIVIA	Zocor 20 mg		JACK,JACK,MD status: Unreviewed, held	05-18-17
<input type="checkbox"/> NORTON,BETHANY	Norvasc 10 mg		MILLER,ALEX,MD status: Unreviewed	05-18-17
<input type="checkbox"/> MONTGOMERY,BLAINE	Glucophage 850 mg		OSHEAJAMIE,MD reviewed by: PPMD_AKN... status: Reviewed	05-18-17
<input type="checkbox"/> KLECK,MICHAEL	Office Visit - Abbreviated		JONES,CAMERON,MD status: Reviewed by: SUSAN	05-12-17
<input type="checkbox"/> MCARDLE,HELEN	Office Visit - Mobile		JONES,CAMERON,MD status: Unreviewed	05-12-17
<input type="checkbox"/> BERN,MARC	Office Visit - Itemized Conditions		JONES,CAMERON,MD status: Unreviewed	05-12-17
<input type="checkbox"/> ANDERSON,JIM	Advanced Directives Advanced Directives Addendum		JONES,CAMERON,MD status: Unreviewed	05-12-17
<input type="checkbox"/> BECKER,JOSEPH	Office Visit1		JONES,CAMERON,MD status: Unreviewed	05-02-17
<input type="checkbox"/> HANSEN,GEORGE	Office Visit1		JONES,CAMERON,MD status: Unreviewed	05-02-17
<input type="checkbox"/> FALK,MICHAEL A	Urine Albumin/Creatinine, Urine C & S AMD_996304_74		SMITH,TRACY,MD status: Unreviewed	05-02-17

Trace Data is Everywhere

Physicians

<input type="checkbox"/> ROACH,TRISTIN	Lipitor 80 mg	<input type="checkbox"/> MILLER,ALEX,MD status: Unreviewed	05-18-17
<input type="checkbox"/> LEON,ERIN	Geriatric Wellness Visit	<input type="checkbox"/> JONES,CAMERON,MD status: Unreviewed	05-16-17
<input type="checkbox"/> BECK,ALIVIA	Zocor 20 mg	<input type="checkbox"/> JACK,JACK,MD status: Unreviewed, held	05-18-17
<input type="checkbox"/> NORTON,BETHANY	Norvasc 10 mg	<input type="checkbox"/> MILLER,ALEX,MD status: Unreviewed	05-18-17
<input type="checkbox"/> MONTGOMERY,BLAINE	Glucophage 850 mg	<input type="checkbox"/> OSHEA,JAMIE,MD reviewed by: PPMD_AKN... status: Reviewed	05-18-17
<input type="checkbox"/> KLECK,MICHAEL	Office Visit - Abbreviated	<input type="checkbox"/> JONES,CAMERON,MD status: Unreviewed	05-12-17
<input type="checkbox"/> MCARDLE,HELEN	Office Visit - Mobile	<input type="checkbox"/> JONES,CAMERON,MD status: Unreviewed	05-12-17
<input type="checkbox"/> BERN,MARC	Office Visit - Itemized Conditions	<input type="checkbox"/> JONES,CAMERON,MD status: Unreviewed	05-12-17
<input type="checkbox"/> ANDERSON,JIIM	Advanced Directives Advanced Directives Addendum	<input type="checkbox"/> JONES,CAMERON,MD status: Unreviewed	05-12-17
<input type="checkbox"/> BECKER,JOSEPH	Office Visit1	<input type="checkbox"/> JONES,CAMERON,MD status: Unreviewed	05-02-17
<input type="checkbox"/> HANSEN,GEORGE	Office Visit1	<input type="checkbox"/> JONES,CAMERON,MD status: Unreviewed	05-02-17
<input type="checkbox"/> FALK,MICHAEL A	Urine Albumin/Creatinine, Urine C & S AMD_996304_74	<input type="checkbox"/> SMITH,TRACY,MD status: Unreviewed	05-02-17



Uber Drivers

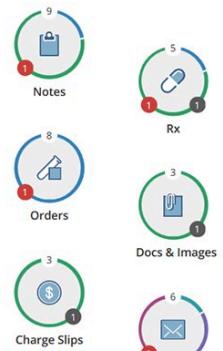


Whong 2014

Trace Data is Everywhere

Physicians

<input type="checkbox"/> • ROACH,TRISTIN	Lipitor 80 mg	<input type="checkbox"/> MILLER,ALEX,MD status: Unreviewed	05-18-17
<input type="checkbox"/> • LEON,ERIN	Geriatric Wellness Visit	<input type="checkbox"/> JONES,CAMERON,MD status: Unreviewed	05-16-17
<input type="checkbox"/> • BECK,ALIVIA	Zocor 20 mg	<input type="checkbox"/> JACK,JACK,MD status: Unreviewed, held	05-18-17
<input type="checkbox"/> NORTON,BETHANY	Norvasc 10 mg	<input type="checkbox"/> MILLER,ALEX,MD status: Unreviewed	05-18-17
<input type="checkbox"/> MONTGOMERY,BLAINE	Glucophage 850 mg	<input type="checkbox"/> OSHEA,JAMIE,MD reviewed by: PPMD_AKN... status: Reviewed	05-18-17
<input type="checkbox"/> KLECK,MICHAEL	Office Visit - Abbreviated	<input type="checkbox"/> JONES,CAMERON,MD status: Reviewed by: SUSAN status: Reviewed	05-12-17
<input type="checkbox"/> MCARDLE,HELEN	Office Visit - Mobile	<input type="checkbox"/> JONES,CAMERON,MD status: Unreviewed	05-12-17
<input type="checkbox"/> BERN,MARC	Office Visit - Itemized Conditions	<input type="checkbox"/> JONES,CAMERON,MD status: Unreviewed	05-12-17
<input type="checkbox"/> ANDERSON,JIIM	Advanced Directives Advanced Directives Addendum	<input type="checkbox"/> JONES,CAMERON,MD status: Unreviewed	05-12-17
<input type="checkbox"/> BECKER,JOSEPH	Office Visit1	<input type="checkbox"/> JONES,CAMERON,MD status: Unreviewed	05-02-17
<input type="checkbox"/> HANSEN,GEORGE	Office Visit1	<input type="checkbox"/> JONES,CAMERON,MD status: Unreviewed	05-02-17
<input type="checkbox"/> FALK,MICHAEL A	Urine Albumin/Creatinine, Urine C & S AMD_996304_74	<input type="checkbox"/> SMITH,TRACY,MD status: Unreviewed	05-02-17

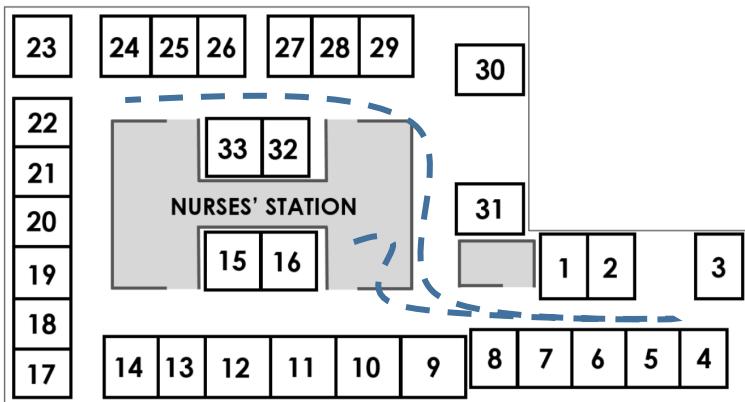


Uber Drivers



Whong 2014

Nurses



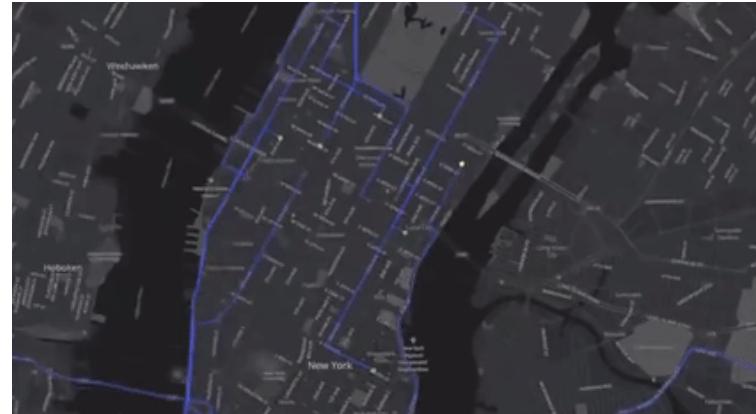
Meng et al. 2018

Trace Data is Everywhere

Physicians

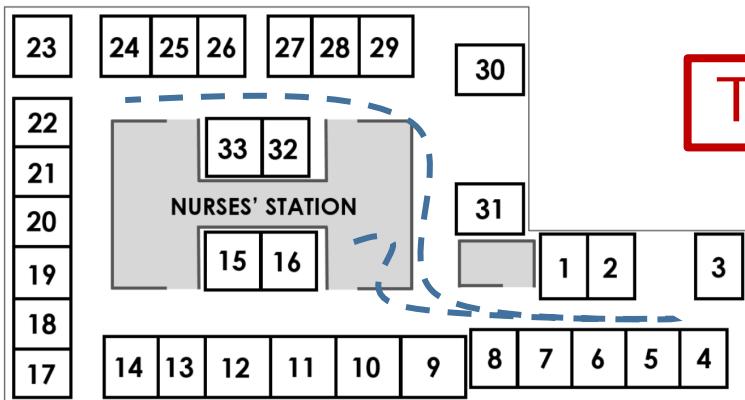
 1 Notes	Lipitor 80 mg	 MILLER,ALEX,MD status: Unreviewed	05-18-17
 5 Rx	• LEON,ERIN Geriatric Wellness Visit	 JONES,CAMERON,MD status: Unreviewed	05-16-17
 8 Orders	• BECK,ALIVIA Zocor 20 mg	 JACK,JACK,MD status: Unreviewed, held	05-18-17
 3 Docs & Images	NORTON,BETHANY Norvasc 10 mg	 MILLER,ALEX,MD status: Unreviewed	05-18-17
 3 Charge Slips	MONTGOMERY,BLAINE Glucophage 850 mg	 OSHEA,JAMIE,MD reviewed by: PPMD_AKN... status: Reviewed	05-18-17
 2 Messages	KLECK,MICHAEL Office Visit - Abbreviated	 JONES,CAMERON,MD status: Unreviewed	05-12-17
	MCARDLE,HELEN Office Visit - Mobile	 JONES,CAMERON,MD status: Unreviewed	05-12-17
	BERN,MARC Office Visit - Itemized Conditions	 JONES,CAMERON,MD status: Unreviewed	05-12-17
	ANDERSON,JIIM Advanced Directives Advanced Directives Addendum	 JONES,CAMERON,MD status: Unreviewed	05-12-17
	BECKER,JOSEPH Office Visit1	 JONES,CAMERON,MD status: Unreviewed	05-02-17
	HANSEN,GEORGE Office Visit1	 JONES,CAMERON,MD status: Unreviewed	05-02-17
	FALK,MICHAEL A Urine Albumin/Creatinine, Urine C & S AMD_996304_74	 SMITH,TRACY,MD status: Unreviewed	05-02-17

Uber Drivers



Whong 2014

Nurses



Trace data

Meng et al. 2018

Tips

Our Paper

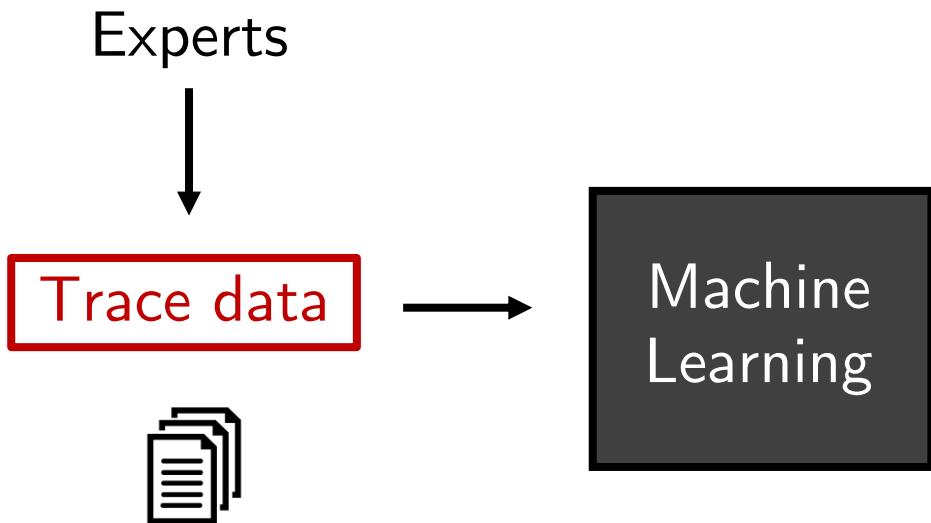
Experts



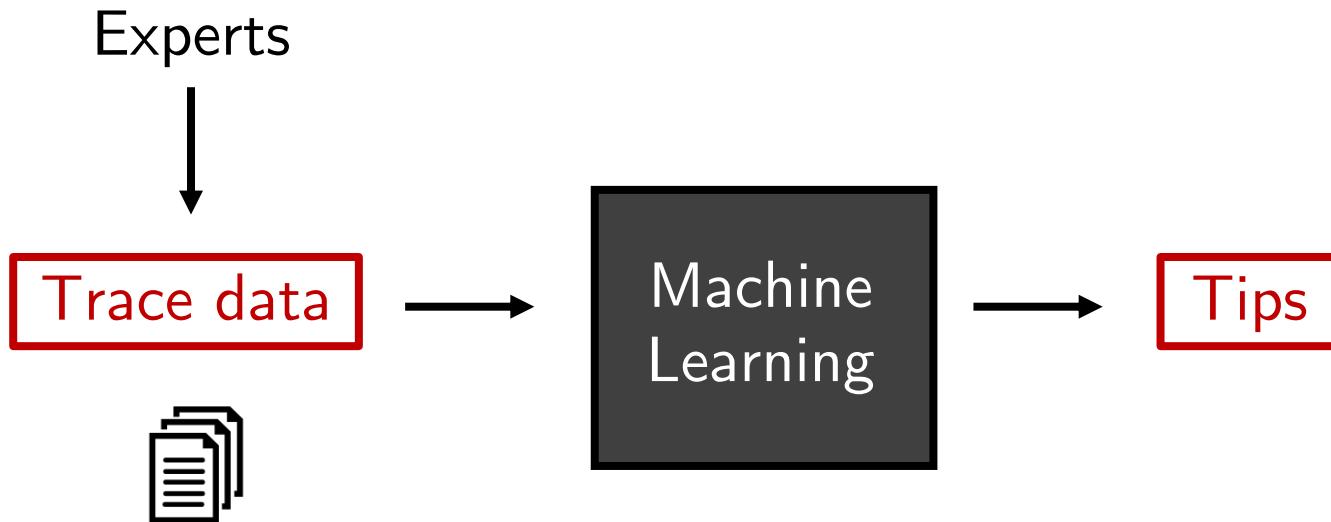
Trace data



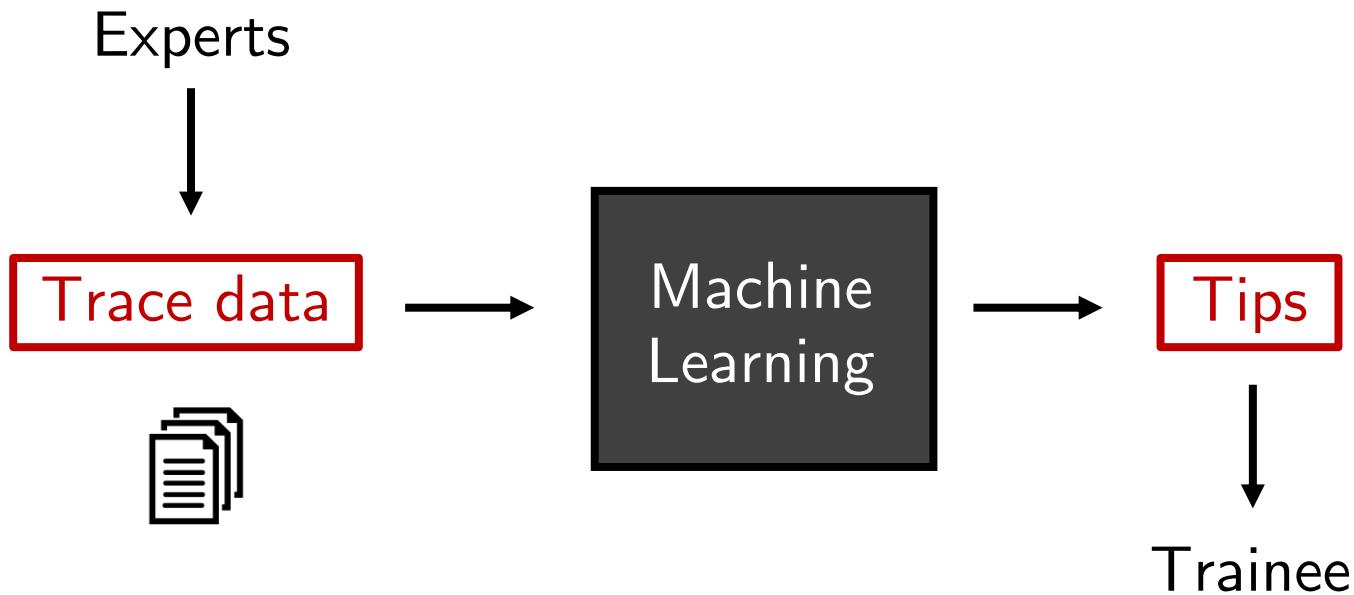
Our Paper



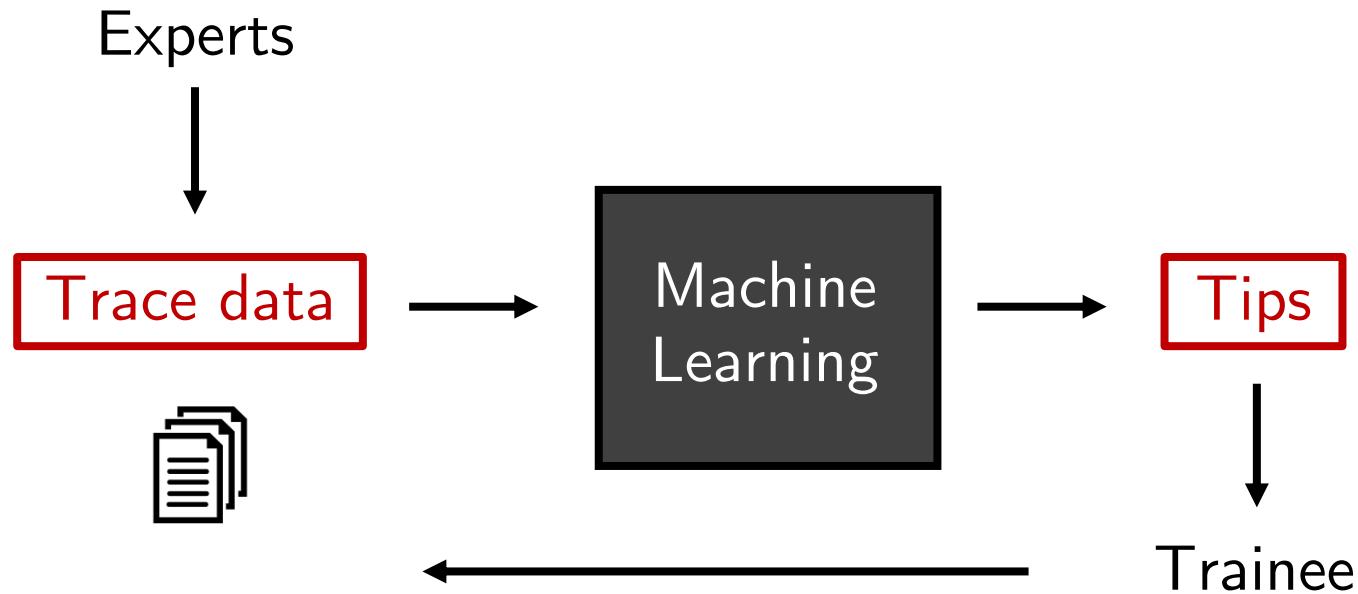
Our Paper



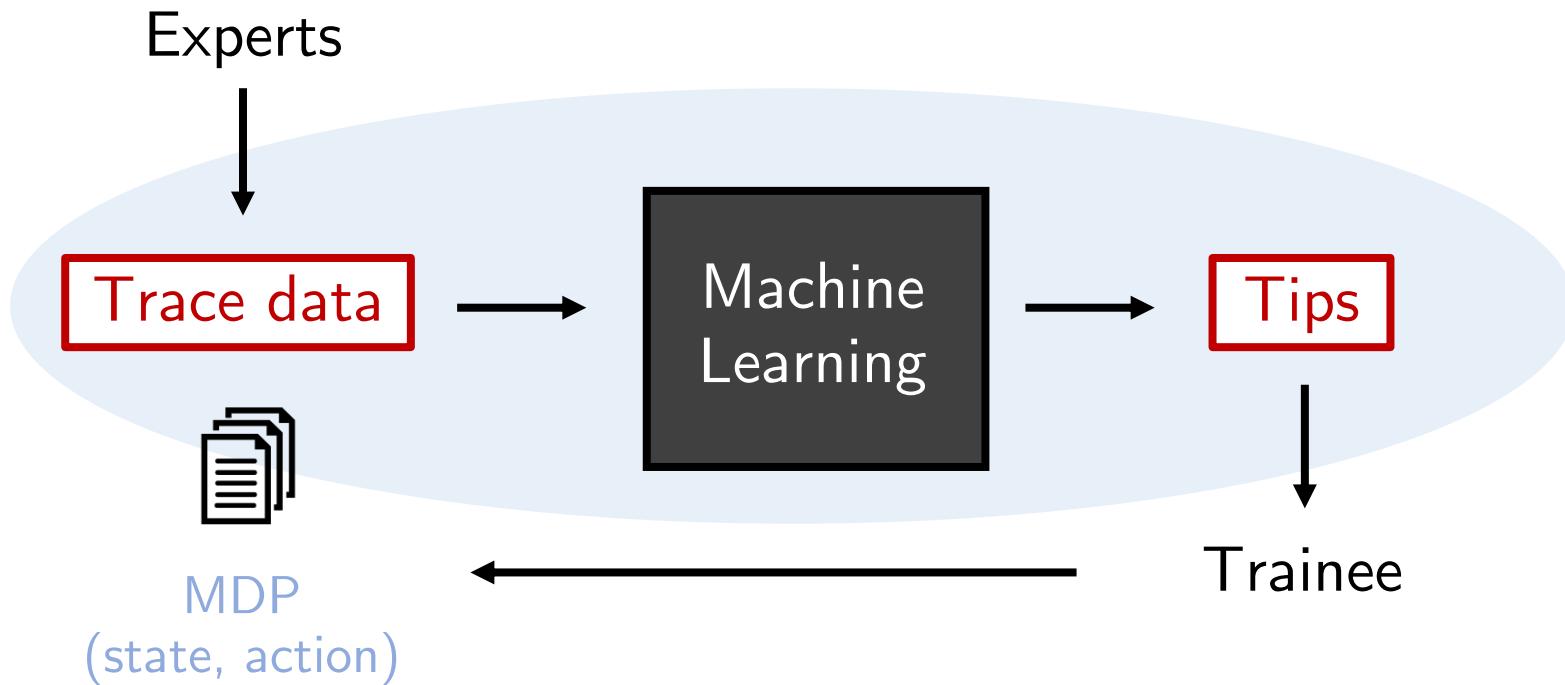
Our Paper



Our Paper

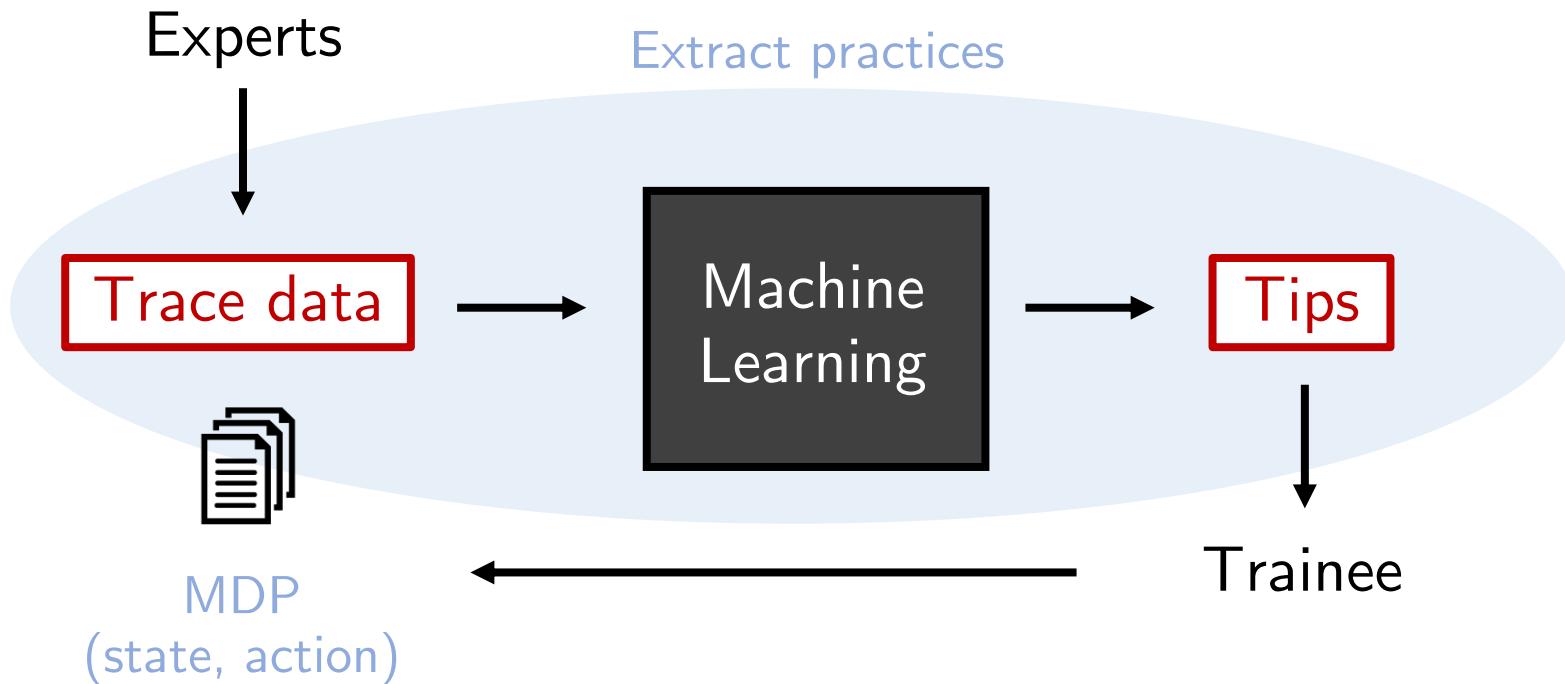


Our Paper



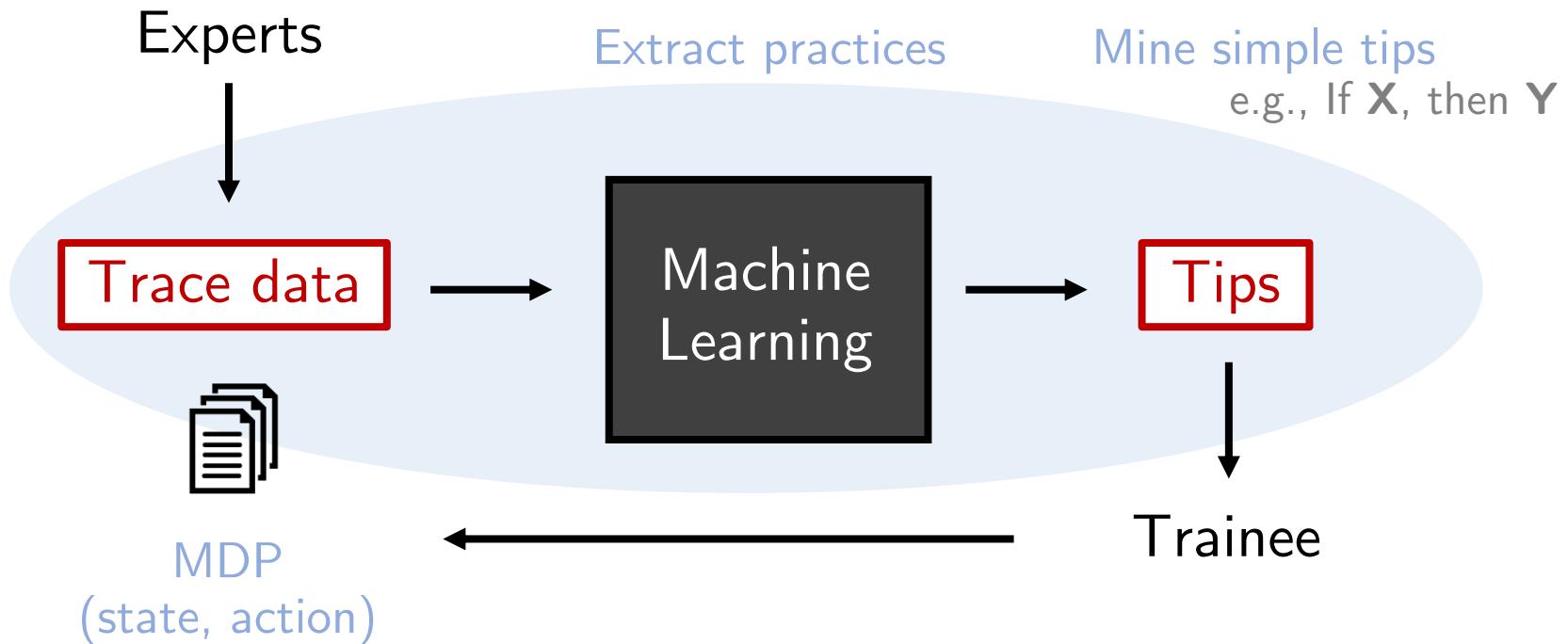
Today: Framework

Our Paper



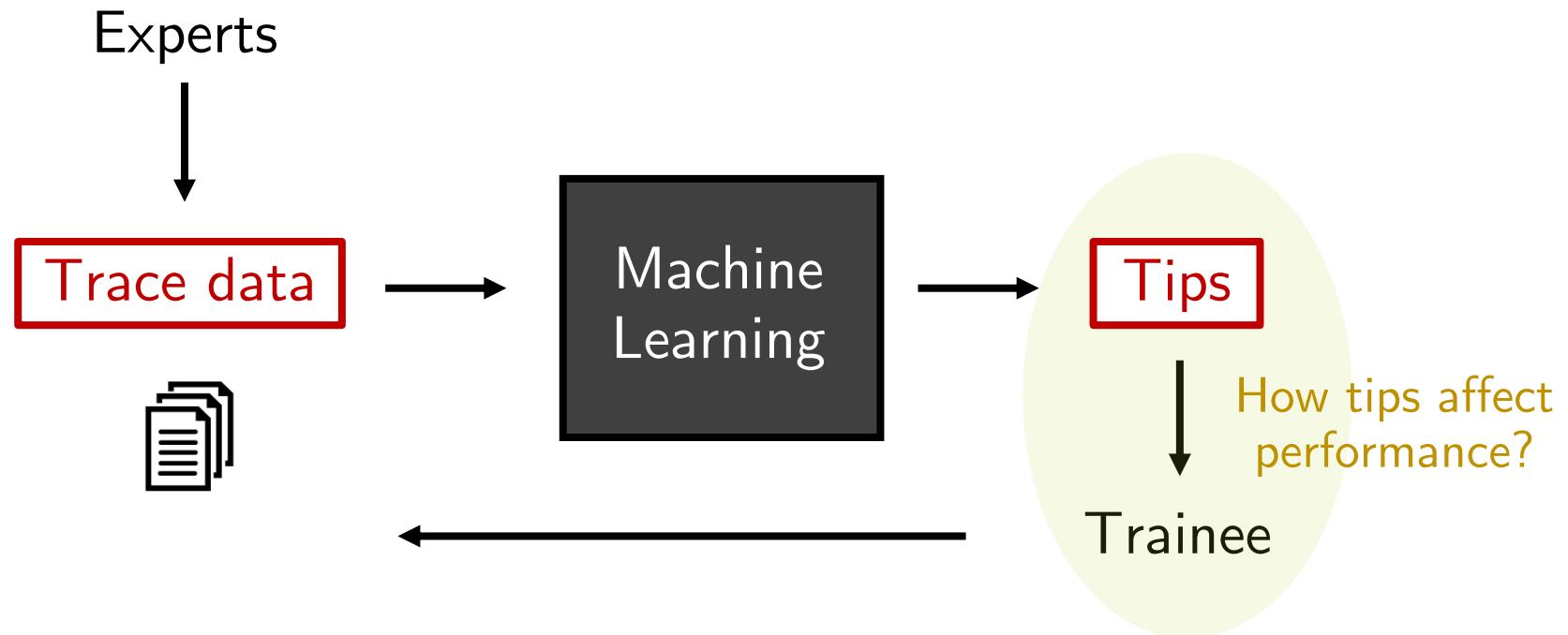
Today: Framework

Our Paper



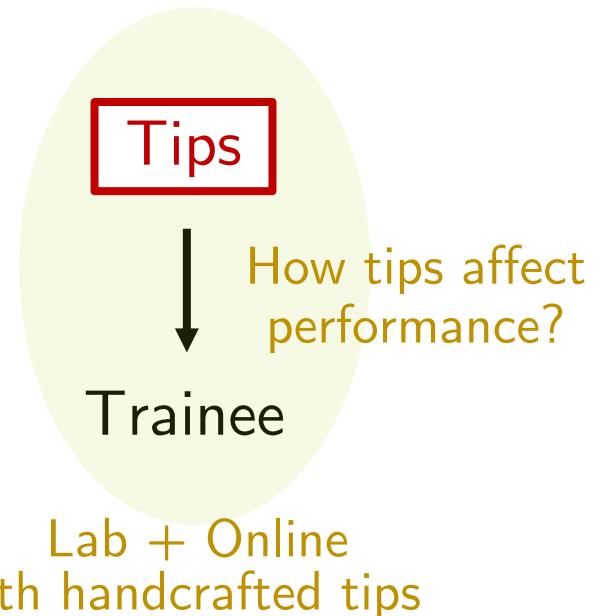
Today: Framework

Our Paper



Today: Framework

Our Paper

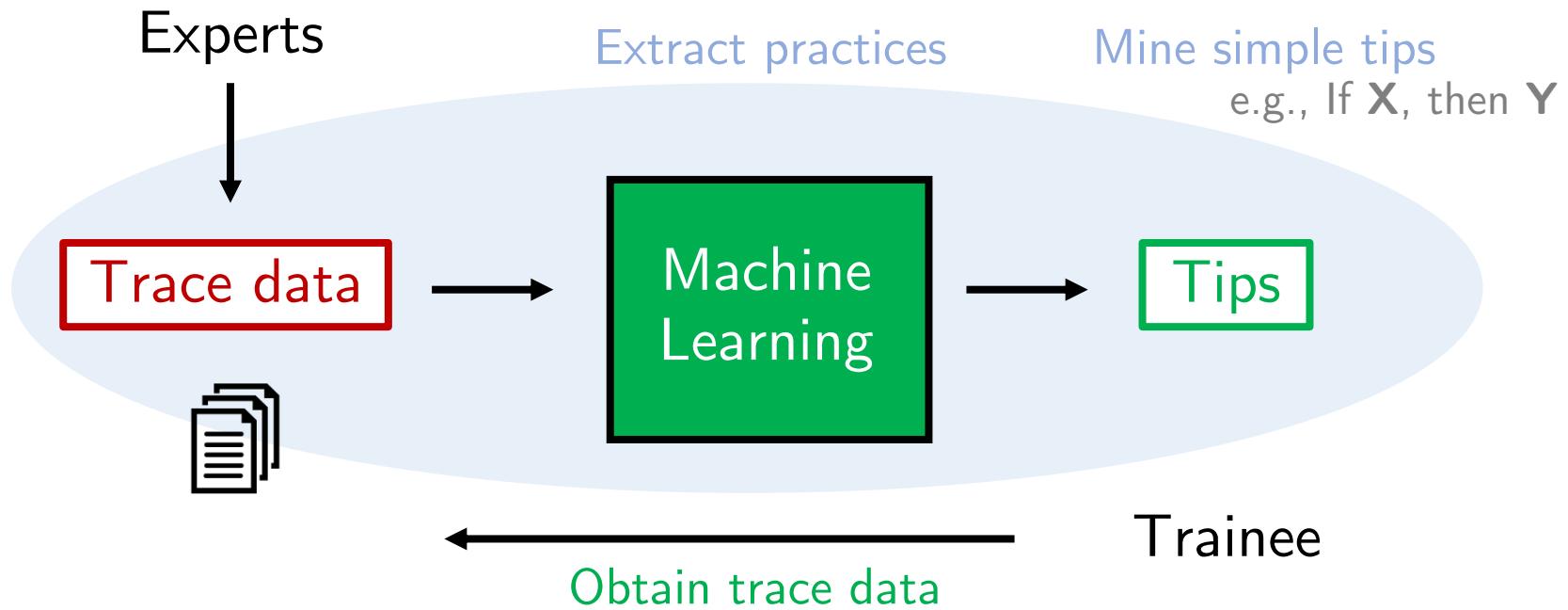


Today:

Framework

Pilot Experiments

Our Paper



Today:

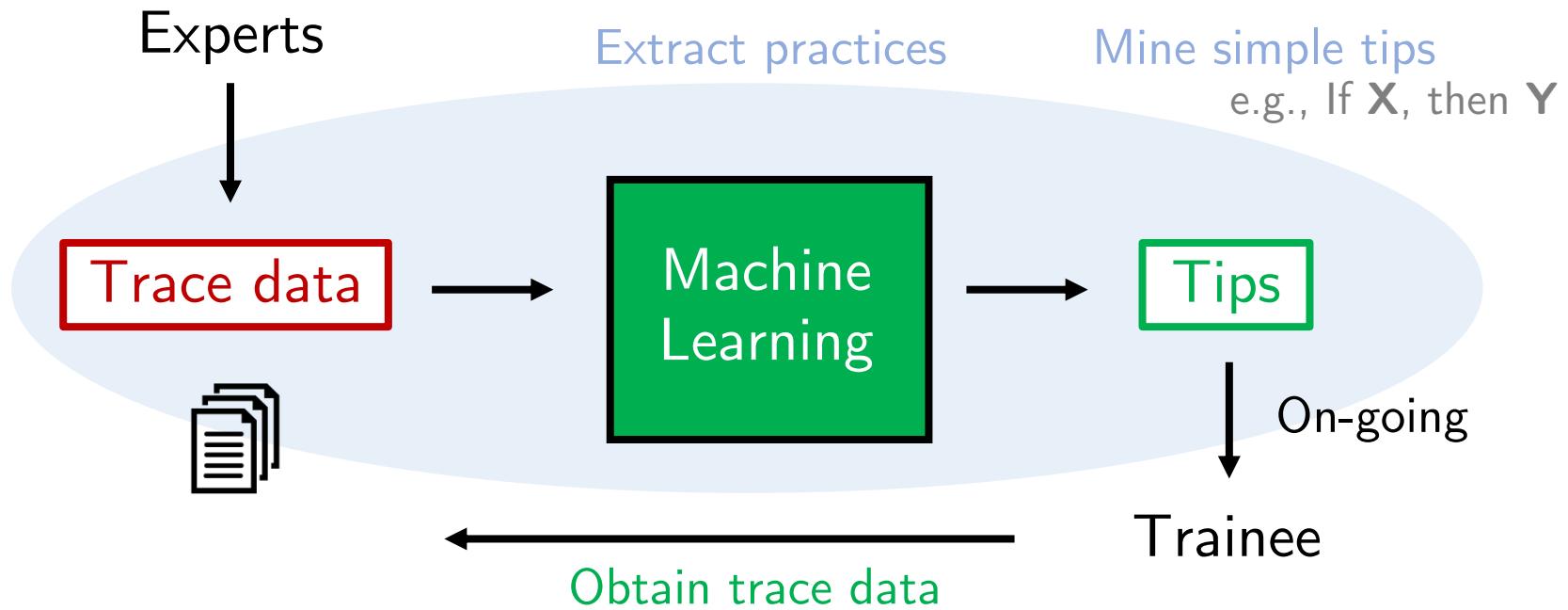
Framework

Pilot Experiments

Results

Extracted practices + tips

Our Paper



Today:

Framework

Pilot Experiments

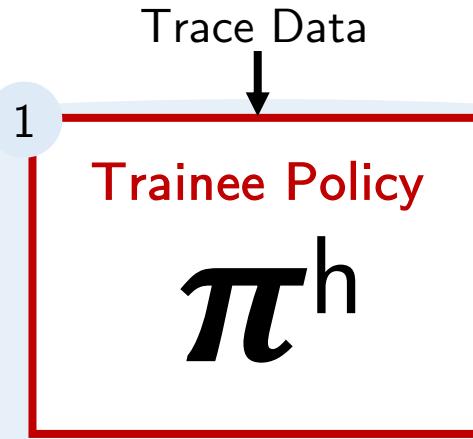
Results

Extracted practices + tips

Our Framework

Policy = the probability distribution of actions given a state

Our Framework

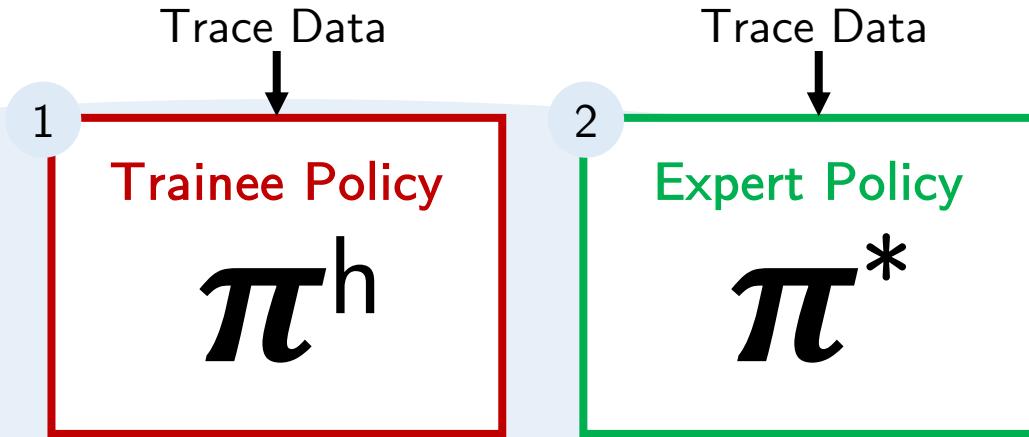


Policy = the probability distribution of actions given a state

Imitation Learning

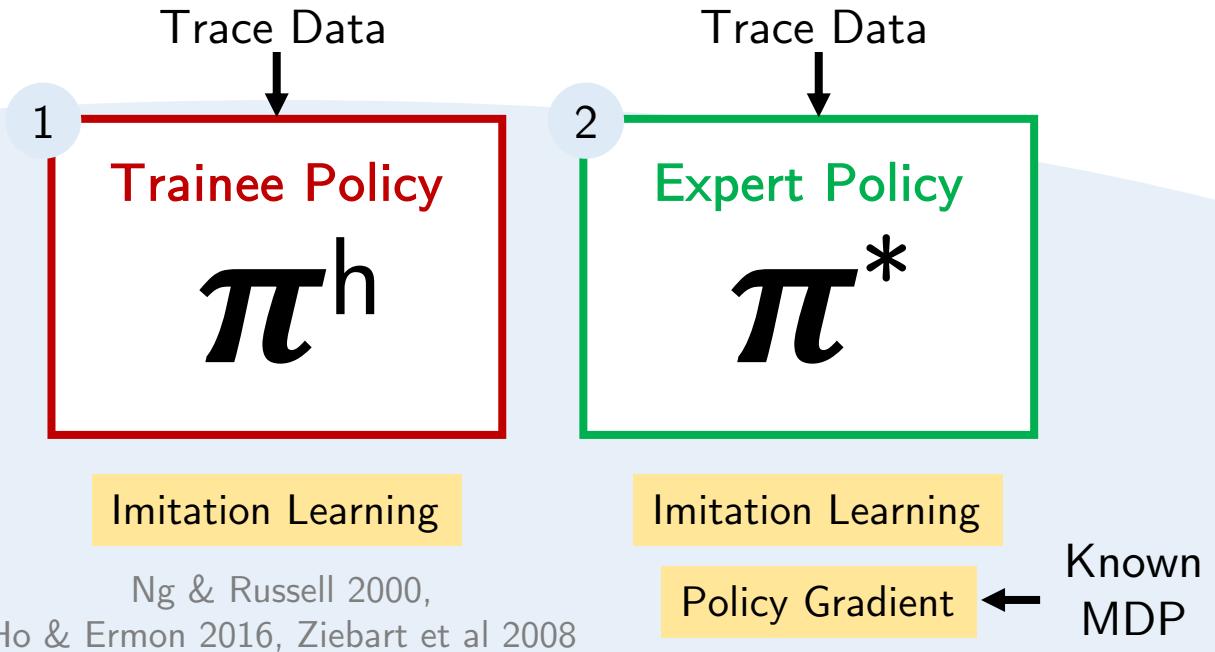
Ng & Russell 2000,
Ho & Ermon 2016, Ziebart et al 2008

Our Framework

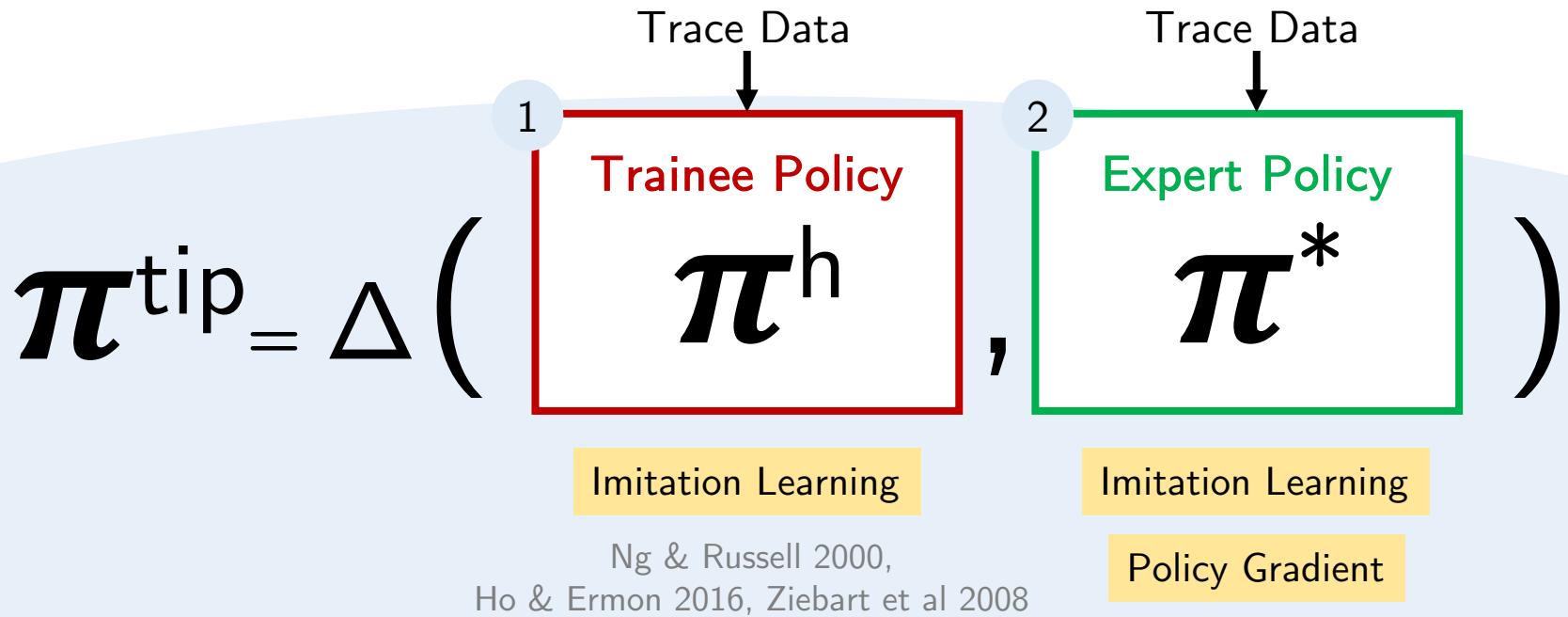


Ng & Russell 2000,
Ho & Ermon 2016, Ziebart et al 2008

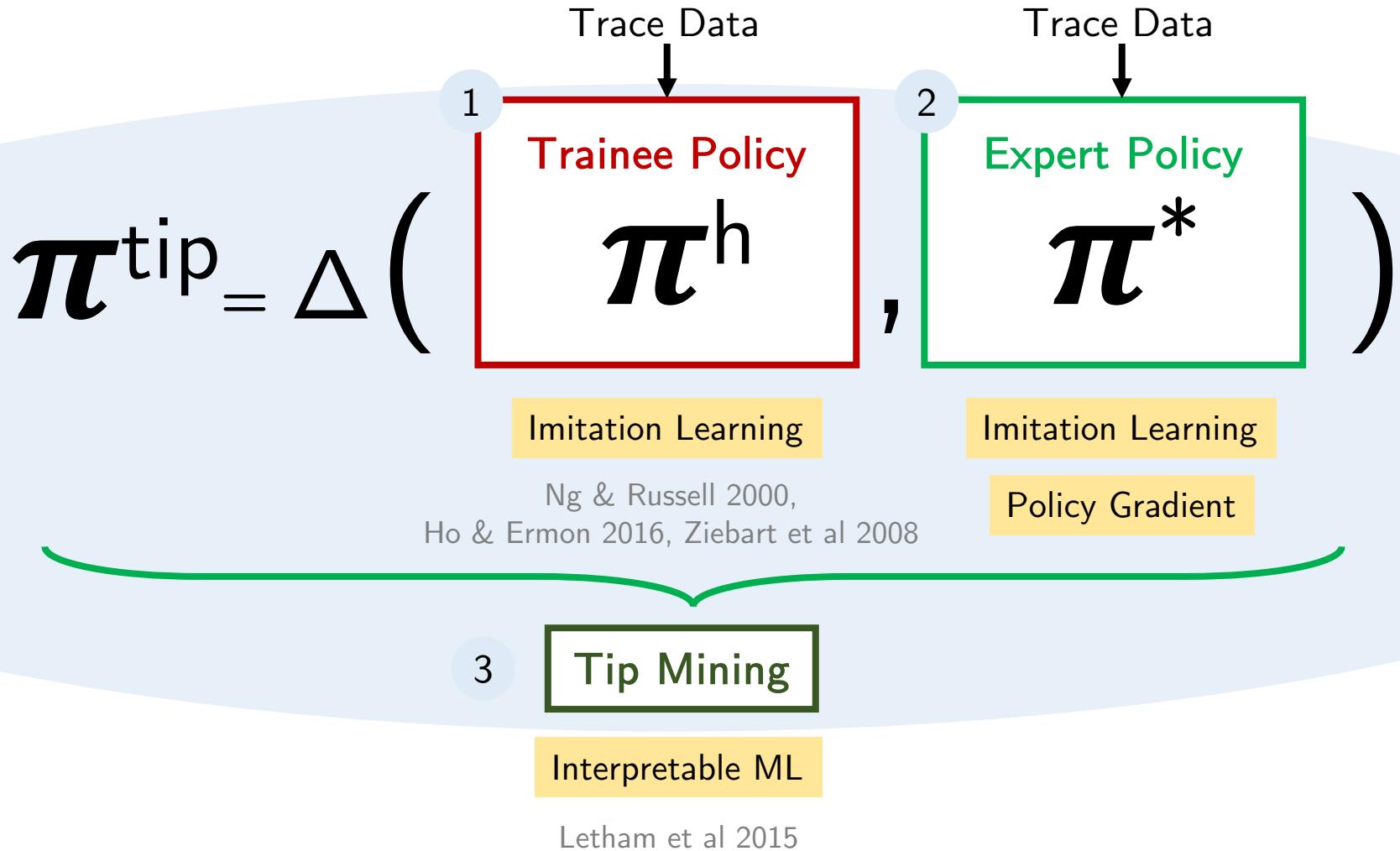
Our Framework



Our Framework

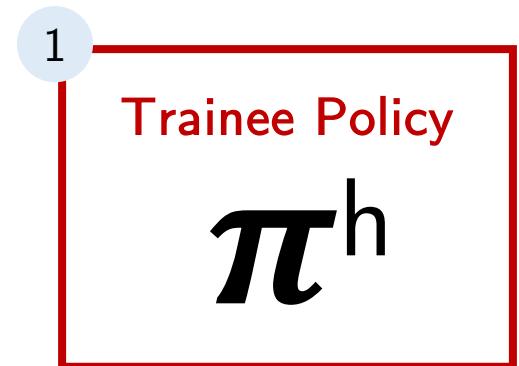


Our Framework



Learning Trainee Policy

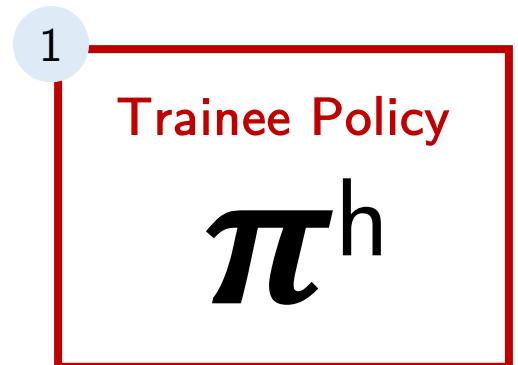
- Supervised Learning
 - Random Forests/
Gradient Boosting Machine



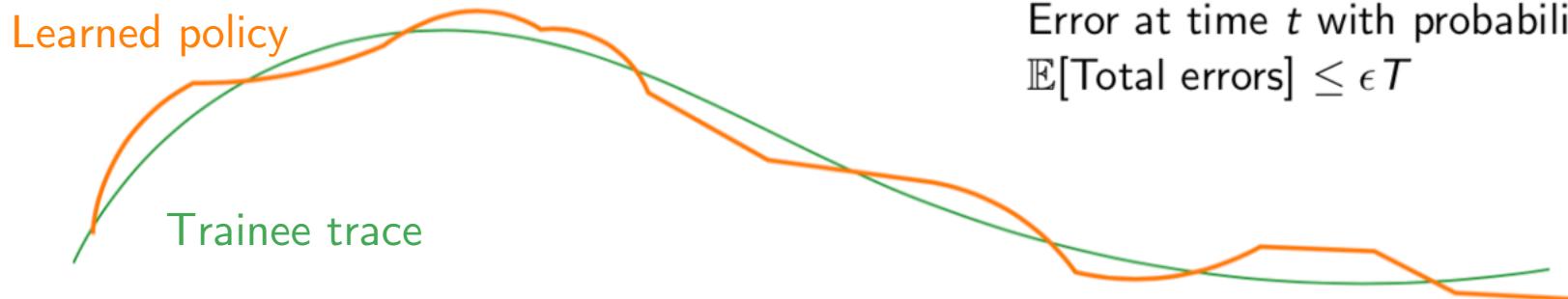
State \longrightarrow Action

Learning Trainee Policy

- Supervised Learning
 - Random Forests/
Gradient Boosting Machine

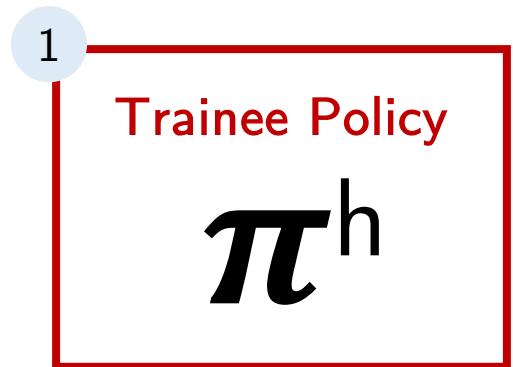


State \longrightarrow Action



Learning Trainee Policy

- Supervised Learning
 - Random Forests/
Gradient Boosting Machine

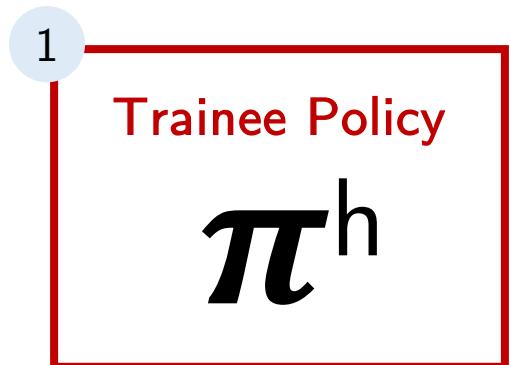


State \longrightarrow Action

i.i.d. (state, action) pairs, ignores temporal structure

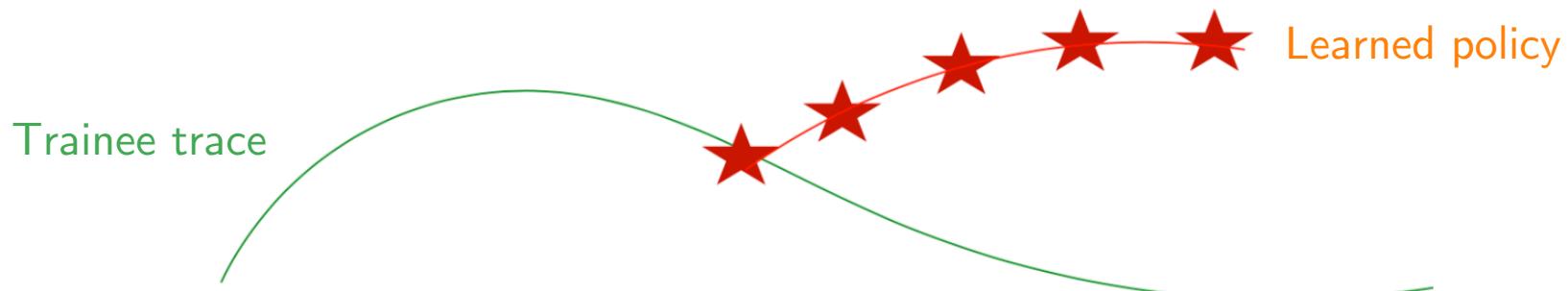
Learning Trainee Policy

- Supervised Learning
 - Random Forests/
Gradient Boosting Machine



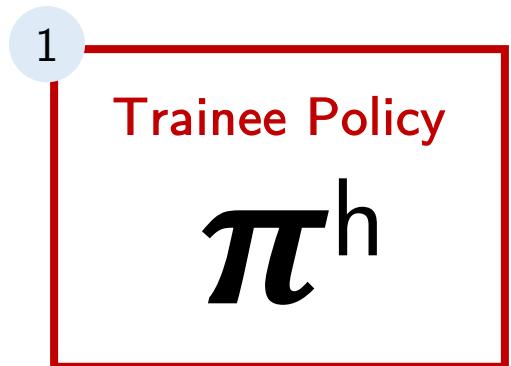
State → Action

i.i.d. (state, action) pairs, ignores temporal structure



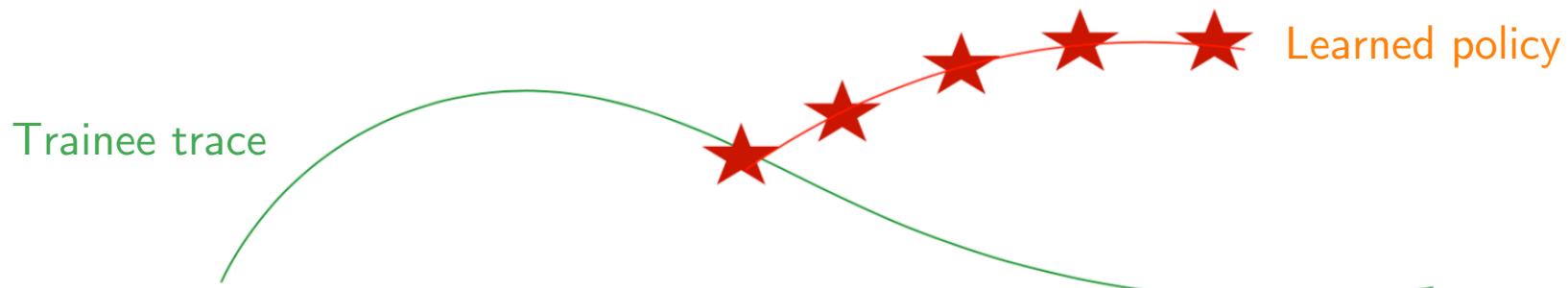
Learning Trainee Policy

- Supervised Learning
 - Random Forests/
Gradient Boosting Machine



State \longrightarrow Action

i.i.d. (state, action) pairs, ignores temporal structure

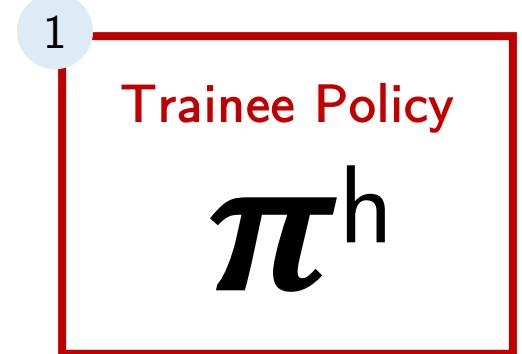


$$\mathbb{E}[\text{Total errors}] \leq \epsilon(T + (T - 1) + (T - 2) \dots + 1) \propto \epsilon T^2$$

- Ross et al 2011

Imitation Learning

- Preserve (s, a) distribution
- Need human experts
 - Useful when it is easier for experts to demonstrate the desired behavior rather than:
 - Specifying a reward that would generate such behavior
 - Specifying the desired policy directly.



Imitation Learning

- Preserve (s,a) distribution
- Need human experts
 - Useful when it is easier for experts to demonstrate the desired behavior rather than:
 - Specifying a reward that would generate such behavior
 - Specifying the desired policy directly.
- **Expensive!**

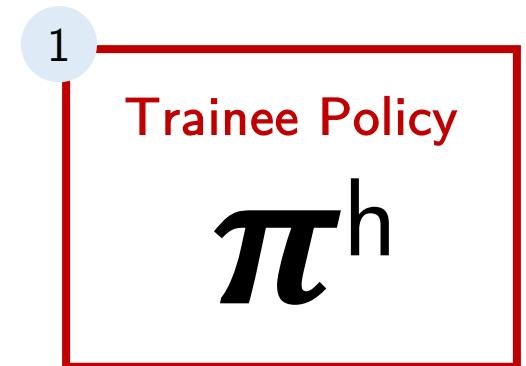
1

Trainee Policy

π^h

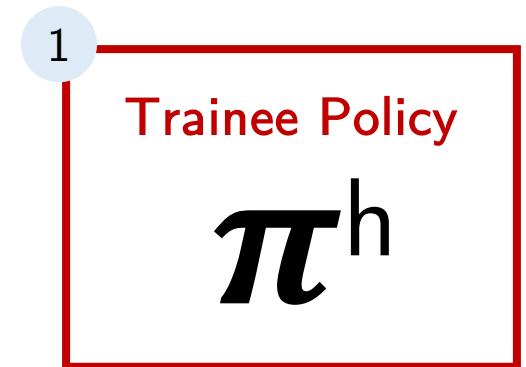
Imitation Learning

- Generative Adversarial IL
 - Distribution from trace data: \mathbf{d}_{π^h}



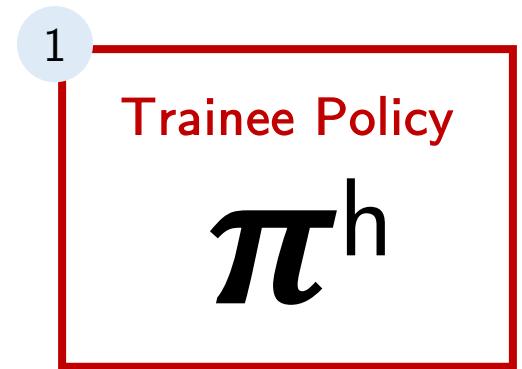
Imitation Learning

- Generative Adversarial IL
 - Distribution from trace data: \mathbf{d}_{π^h}
 - Train a policy π^θ (gradient descent)
 - Obtain its (s, a) distribution: \mathbf{d}_{π^θ}



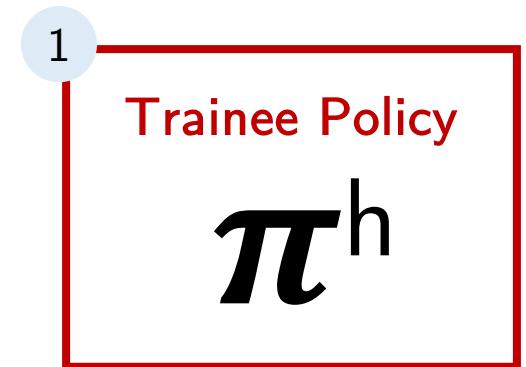
Imitation Learning

- Generative Adversarial IL
 - Distribution from trace data: \mathbf{d}_{π_h}
 - Train a policy π^θ (gradient descent)
 - Obtain its (s, a) distribution: \mathbf{d}_{π^θ}
 - Find π^θ that minimizes $D_{KL}(\mathbf{d}_{\pi^\theta}, \mathbf{d}_{\pi_h})$



Imitation Learning

- Generative Adversarial IL
 - Distribution from trace data: \mathbf{d}_{π^h}
 - Train a policy π^θ (gradient descent)
 - Obtain its (s, a) distribution: \mathbf{d}_{π^θ}
 - Find π^θ that minimizes $D_{KL}(\mathbf{d}_{\pi^\theta}, \mathbf{d}_{\pi^h})$

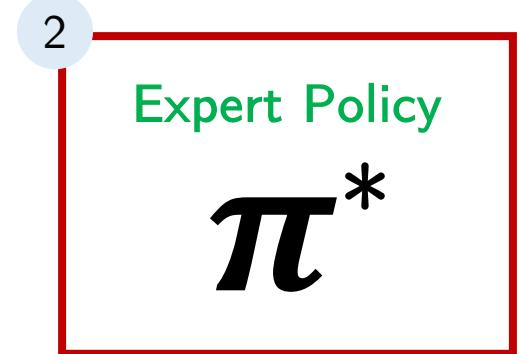


= States visited and actions taken in those states under π^θ to be the same as those from π^h

Learning Expert Policy

Reward function unknown

- Imitation learning from experts



Learning Expert Policy

Reward function unknown

- Imitation learning from experts

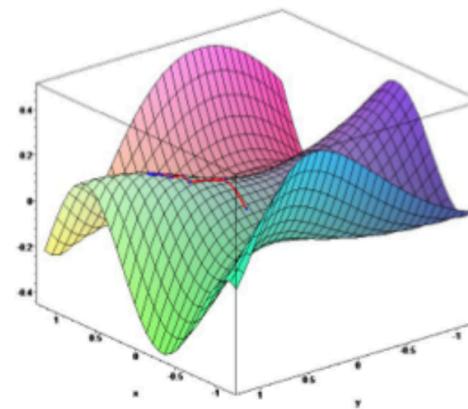
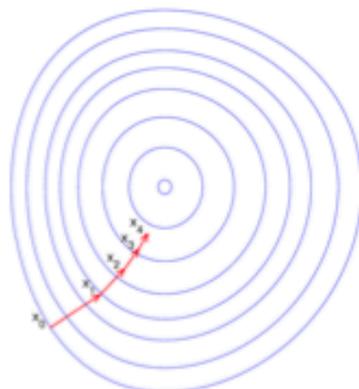
2

Expert Policy

π^*

Inferred reward function

- Our experiment: we know MDP.
- Solve for optimal policy with policy gradient algorithm



Learning Tips

3

Tip Mining

- Interpretable ML: Bayesian Rule Lists
 - Adapt from Letham et al 2015
 - If **[feature x]**, then **[label y]**

Learning Tips

3

Tip Mining

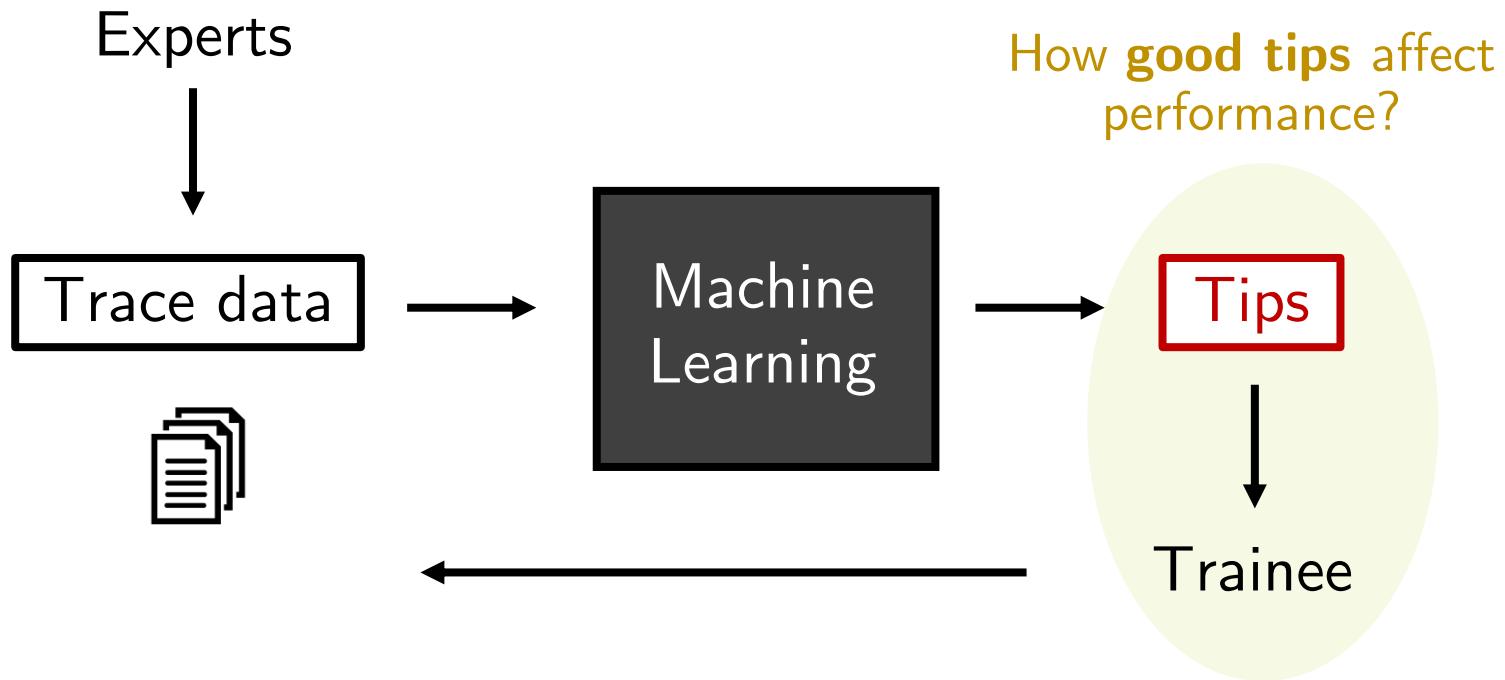
- Interpretable ML: Bayesian Rule Lists
 - Adapt from Letham et al 2015
 - If **[feature x]**, then **[label y]**
 - e.g., if **high blood pressure**, then **stroke**
 - Pre-mined frequent patterns into a decision list using Bayesian statistics

Learning Tips

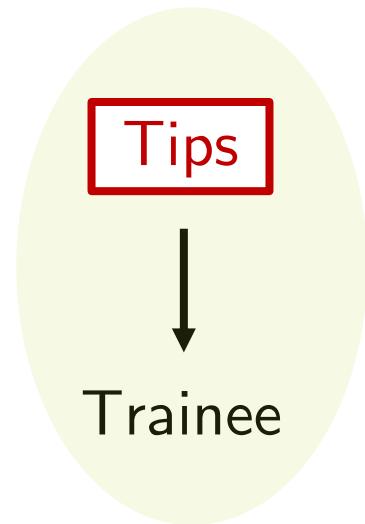
3

Tip Mining

- Interpretable ML: Bayesian Rule Lists
 - Adapt from Letham et al 2015
 - If **[feature x]**, then **[label y]**
 - e.g., if **high blood pressure**, then **stroke**
 - Pre-mined frequent patterns into a decision list using Bayesian statistics
- Our context
 - If **[state s]**, then take **[action a]**



How **good tips** affect performance?

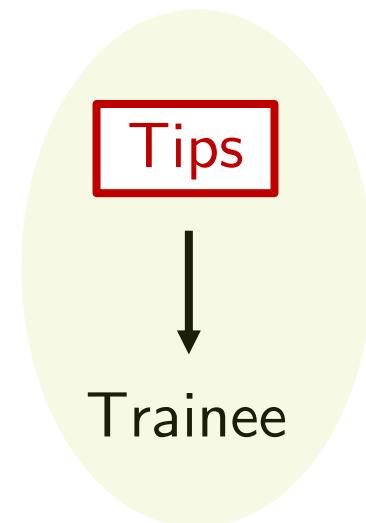


Pilot Experiments

Queueing Game

How good tips affect performance?

Start with
handcrafted tips



Amazon Mechanical Turk

N = 146

Wharton Behavioral Lab

N = 203

Queueing Game

Burger Queen

Participant

Queueing Game

Burger Queen

Alice



Bob



Carol



Participant

Queueing Game

Reward: 0

4 seconds left

Tick #1/23

Burger Queen

Burger

*chop
cook
plate*

Burger

*chop
cook
plate*

Burger

*chop
cook
plate*

Alice



Bob



Carol



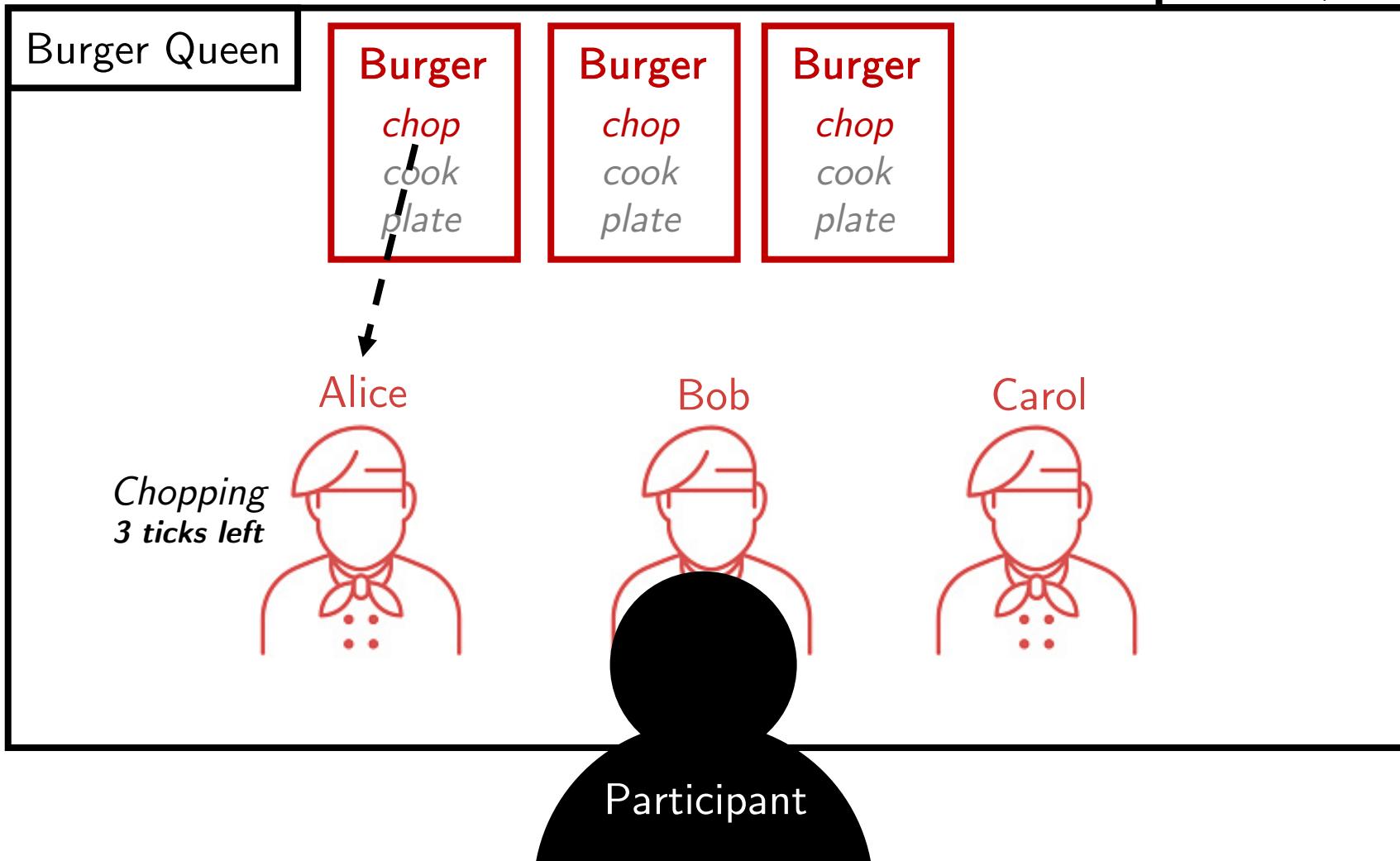
Participant

Queueing Game

Reward: 0

4 seconds left

Tick #1/23

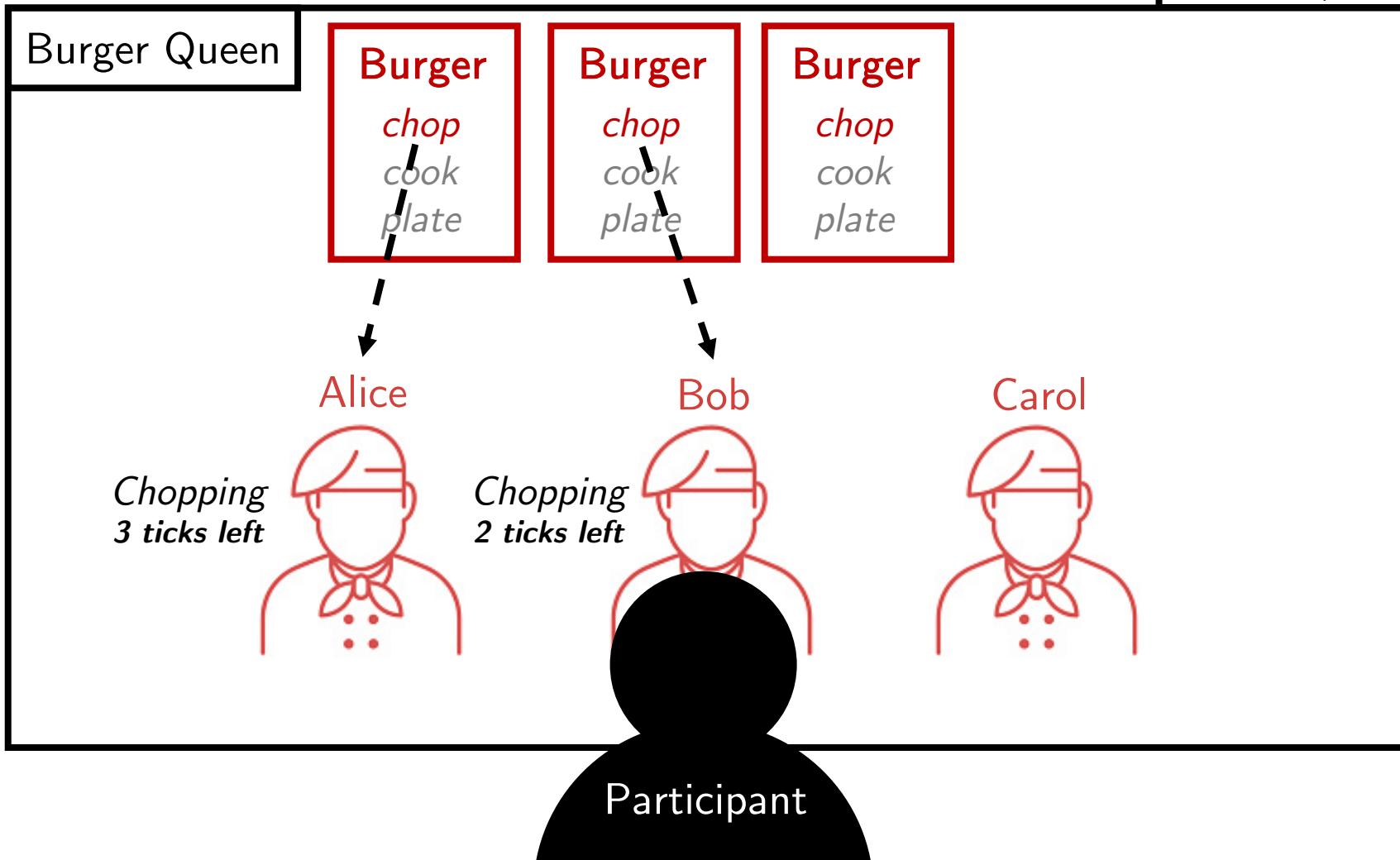


Queueing Game

Reward: 0

4 seconds left

Tick #1/23

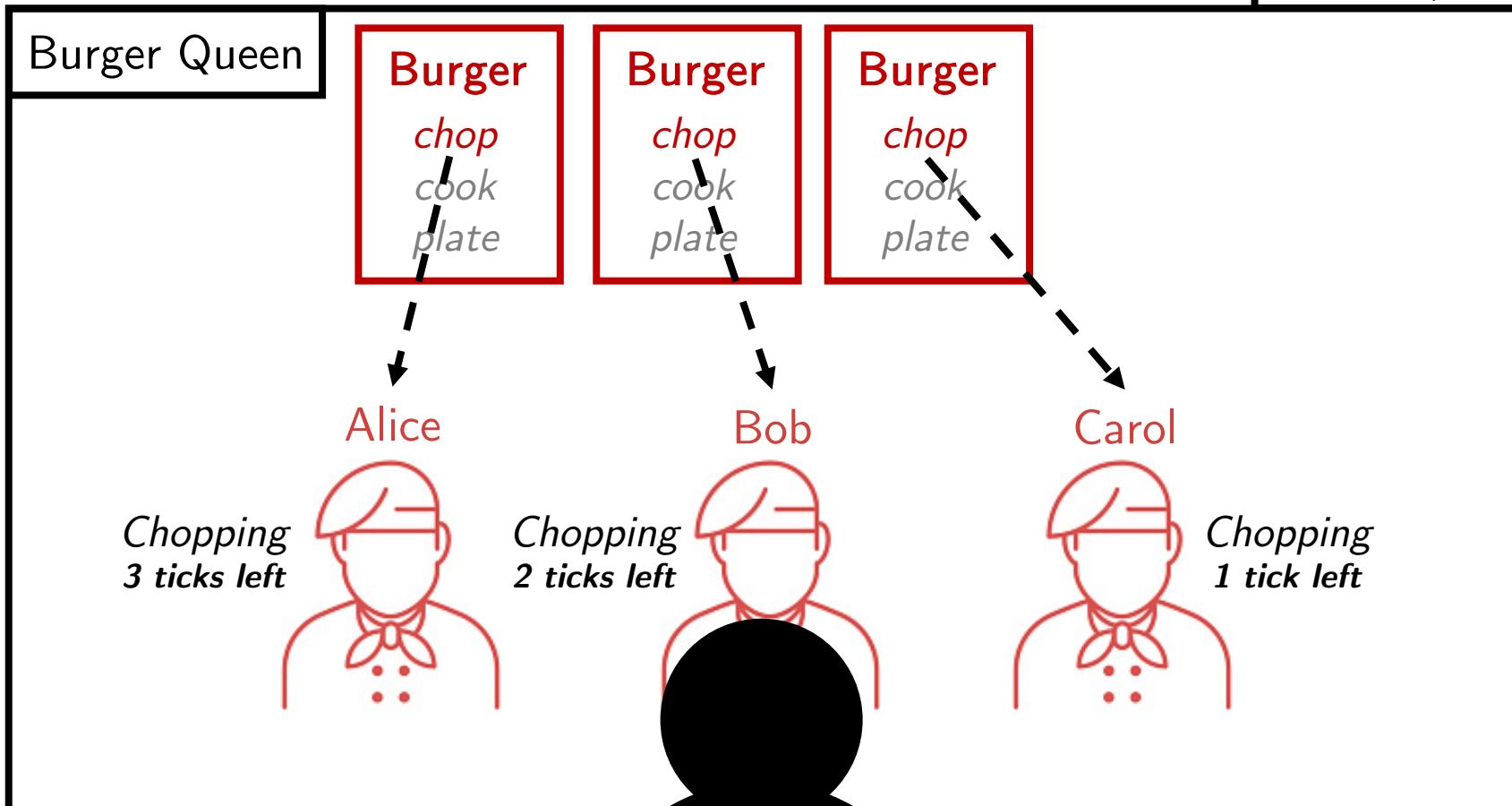


Queueing Game

Reward: 0

4 seconds left

Tick #1/23

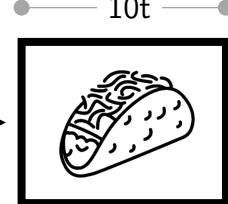


Experiment Flow

Game instructions



Demo: tacos

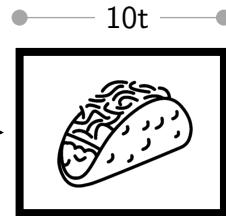


Experiment Flow

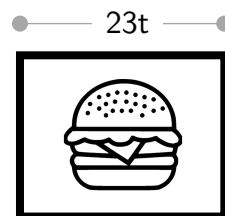
Game instructions



Demo: tacos



Introduce burgers
+ new workers

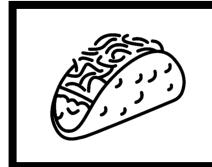


Experiment Flow

Game instructions

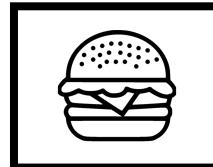
→ Demo: tacos

● — 10t — ●

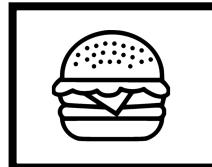


→ Introduce burgers
+ new workers

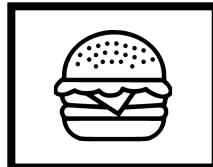
● — 23t — ●



● — 23t — ●



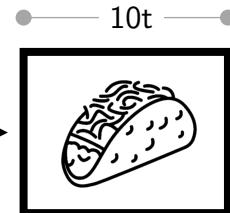
● — 23t — ●



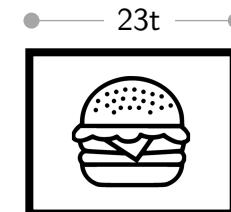
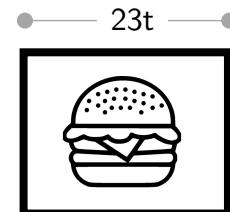
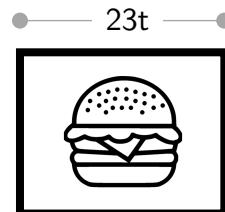
Experiment Flow

Game instructions

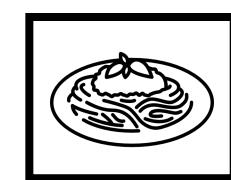
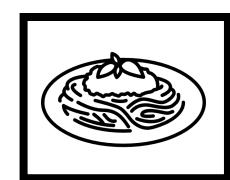
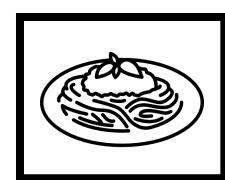
Demo: tacos



Introduce burgers
+ new workers



Introduce pasta
+ new workers

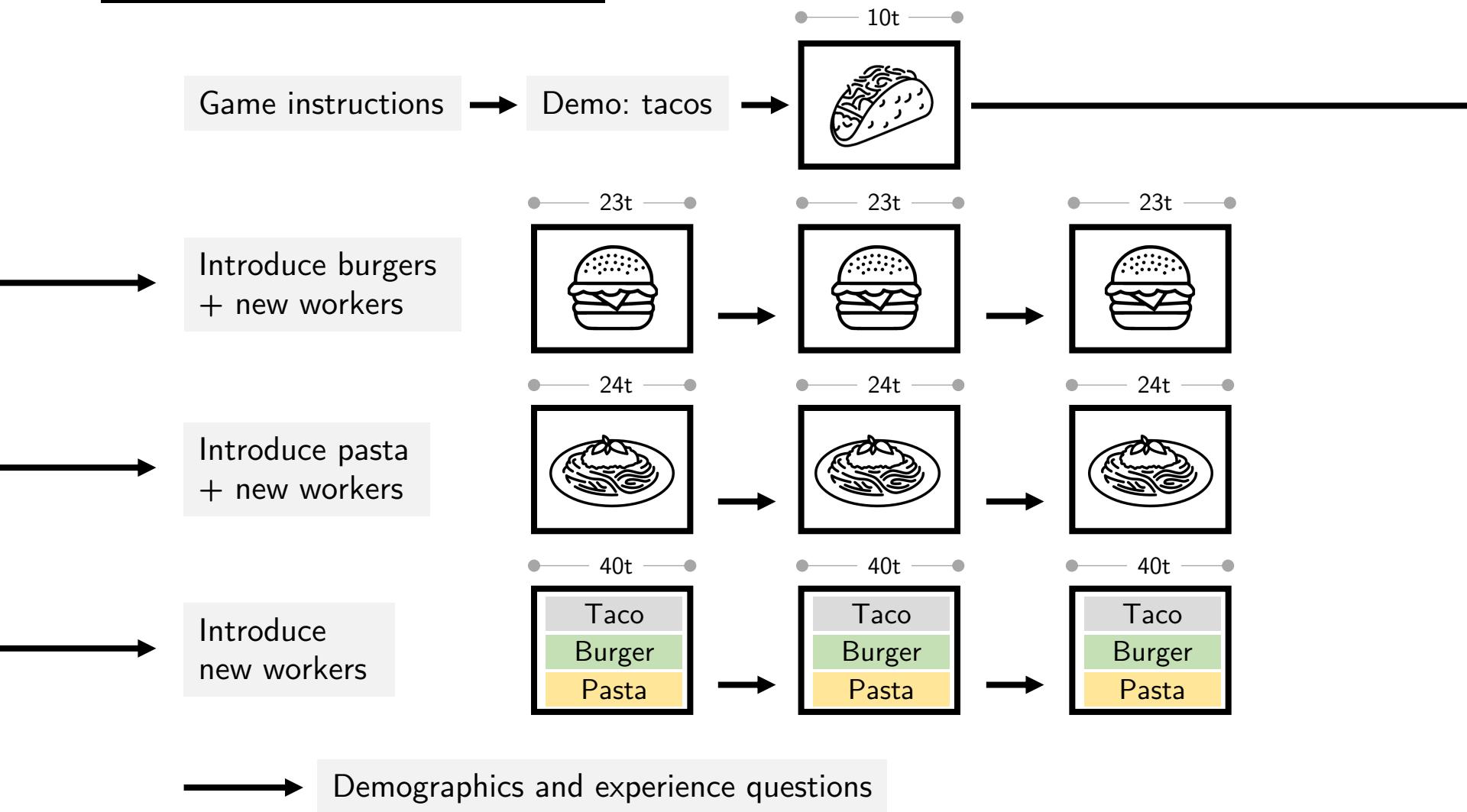


10t

23t

24t

Experiment Flow



Experiment

Learning Workers' Skills



Experiment Learning Workers' Skills



Role:

Server

Sous-Chef

Chef

Unknown to participants

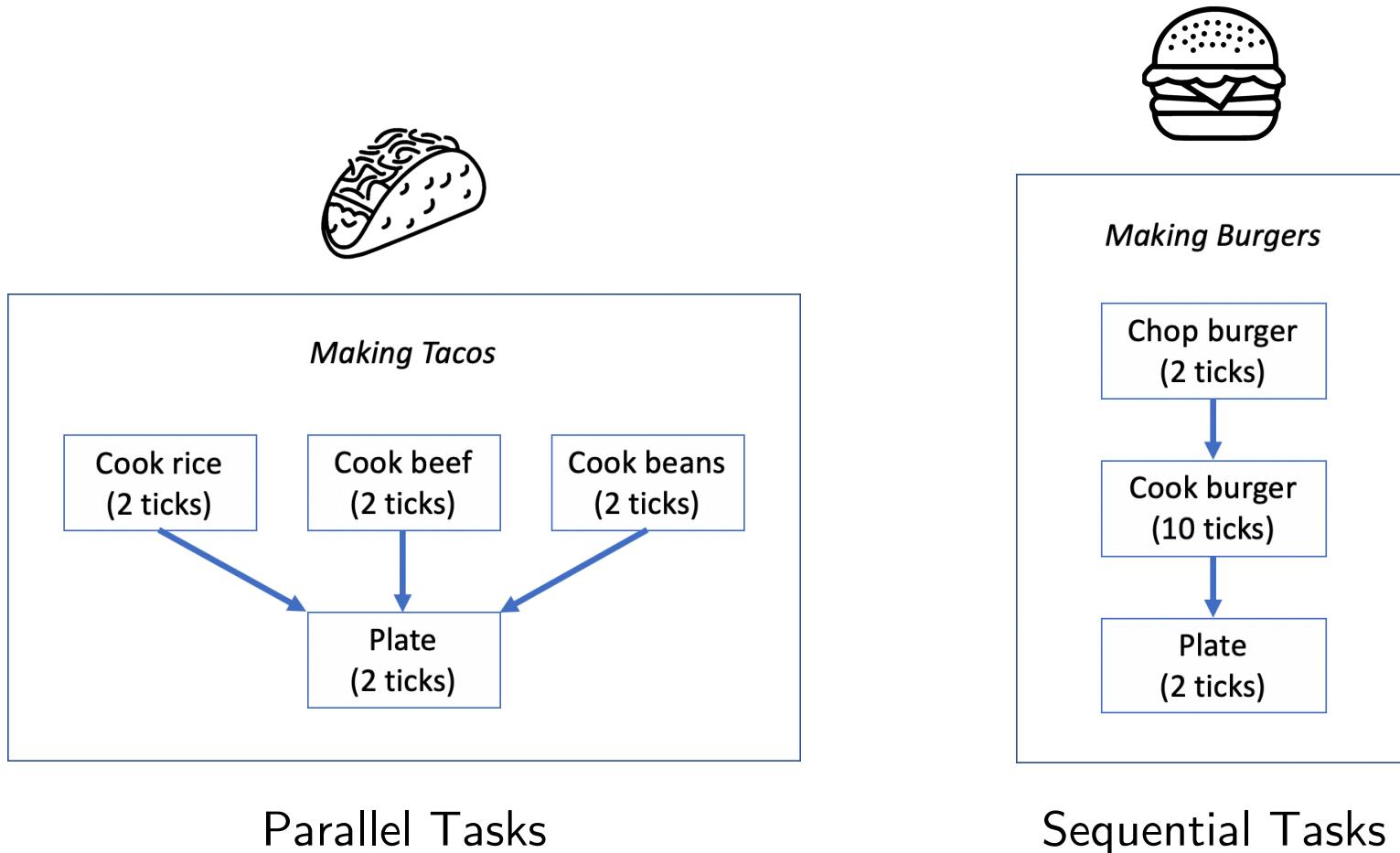
Experiment Learning Workers' Skills



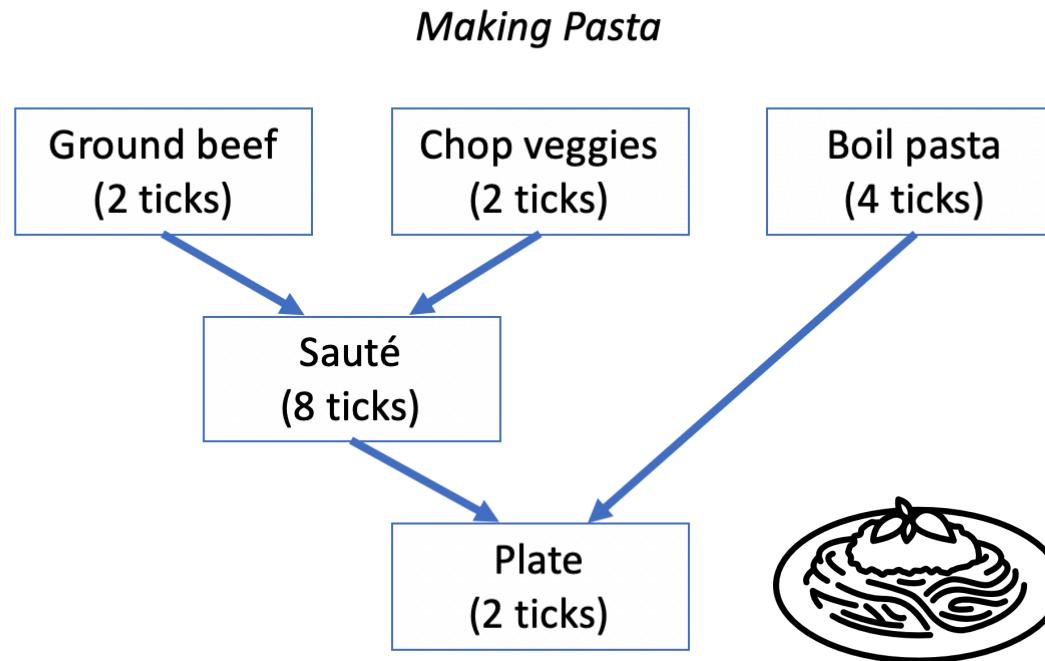
<i>Role:</i>	Server	Sous-Chef	Chef
<i>Chopping:</i>	Slow	Fast	Fast
<i>Cooking:</i>	Slow	Medium	Fast
<i>Plating:</i>	Fast	Medium	Slow

Unknown to participants

Experiment Learning Tasks



Experiment Learning Tasks



Hybrid Tasks (Both Parallel and Sequential)

Experiment

Handcrafted Tips

Control

No tips.

Tips

Experiment Handcrafted Tips

Control

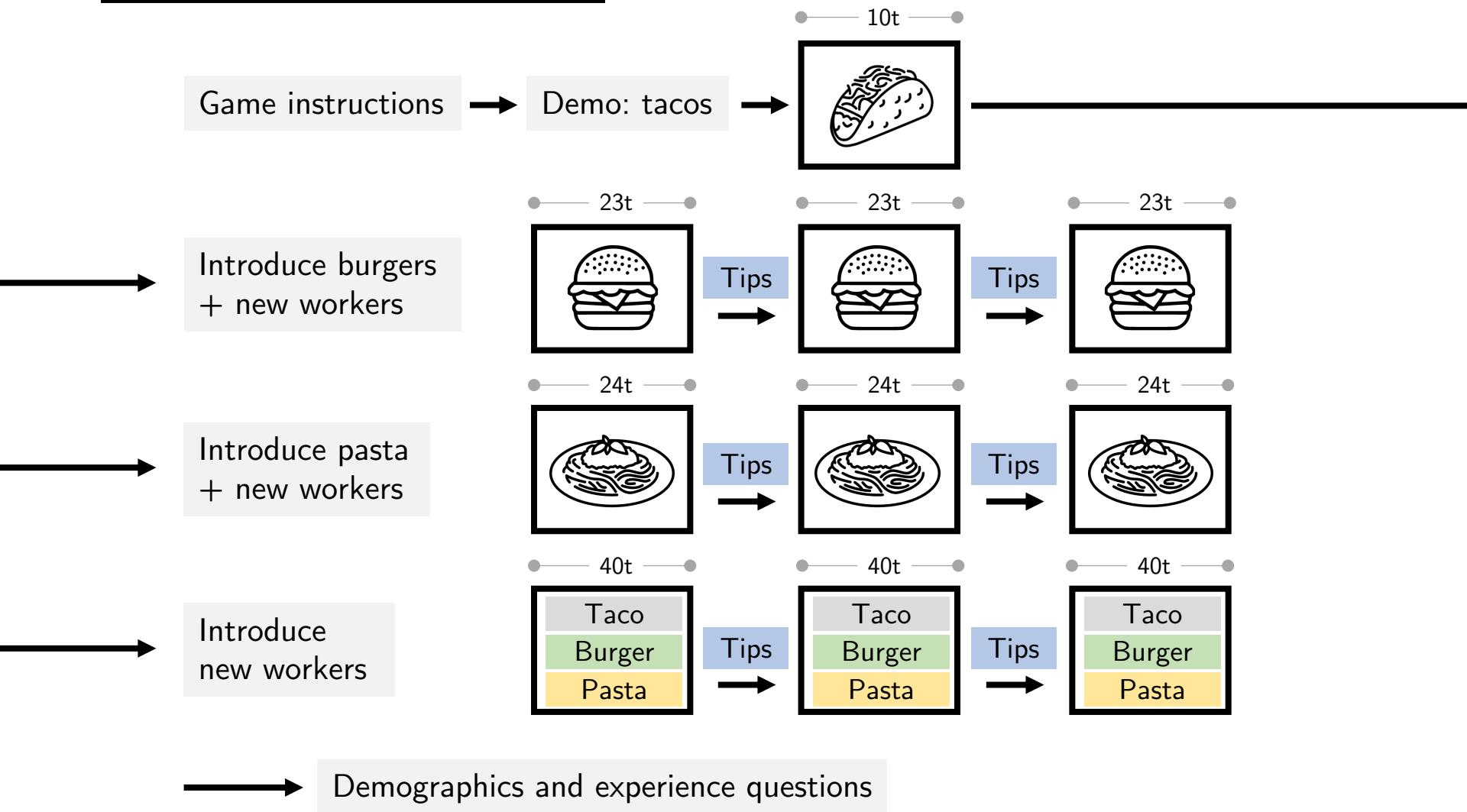
No tips.

Tips

“Prioritize cooking and chopping tasks for Bob and Carol, and plating tasks for Alice.”

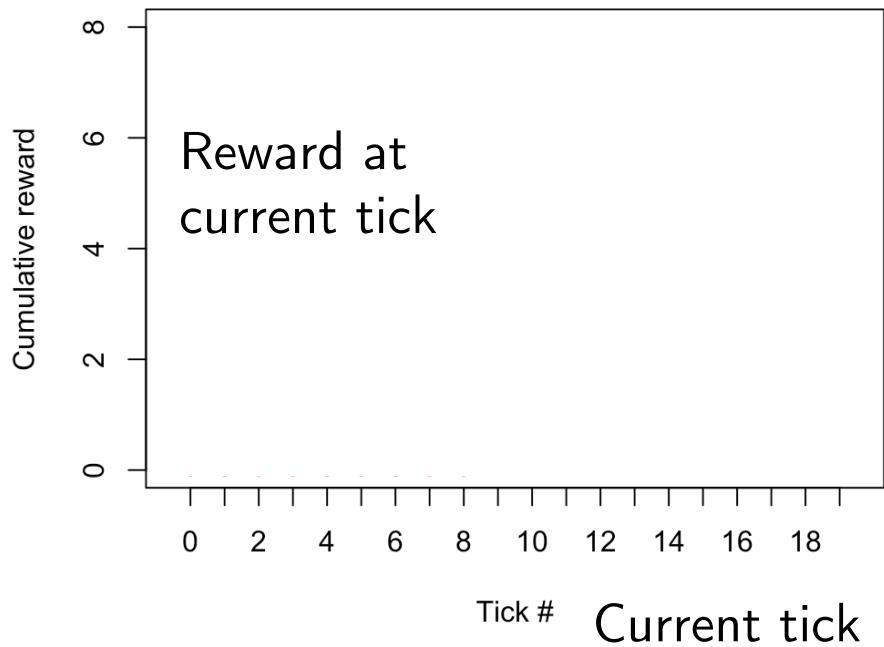
“...Carol should NEVER plate, and Alice should NEVER cook (it’s better to leave them idle than to give them these tasks.)”

Experiment Flow

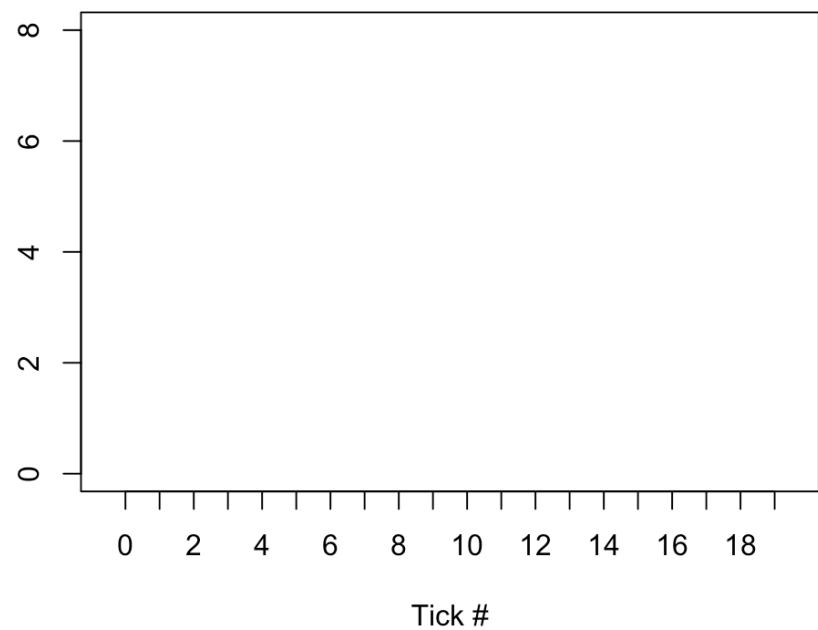


Results Performance

Burger Round 1

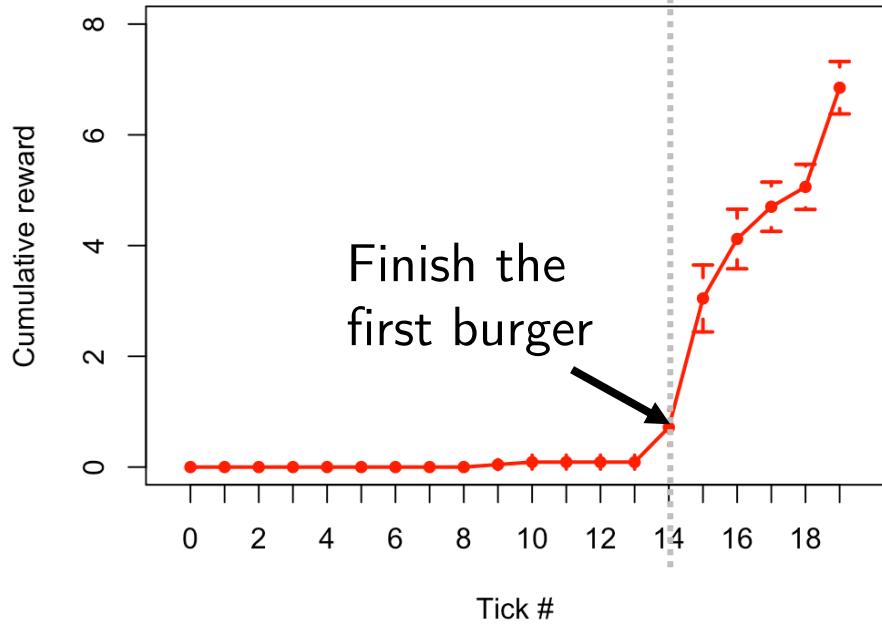


Burger Round 3

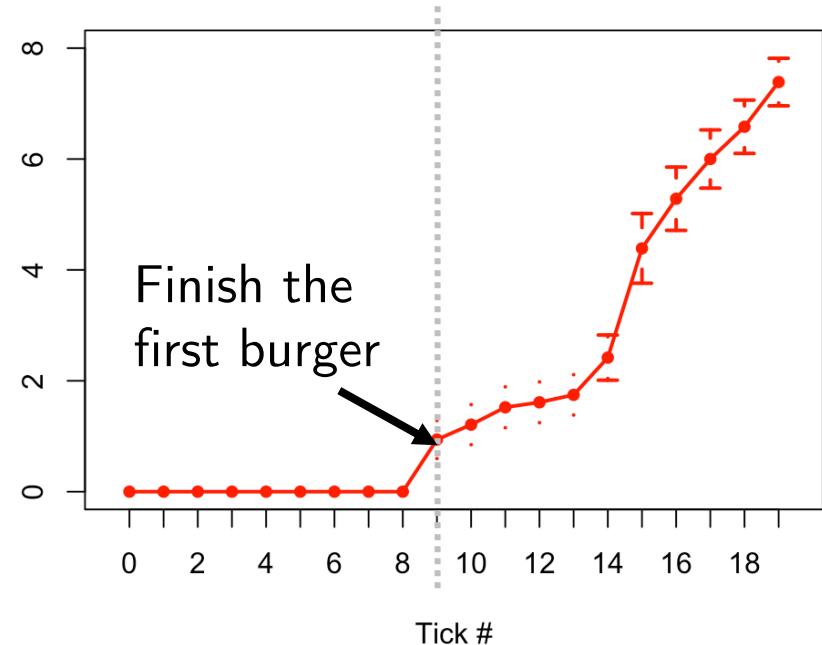


Results Performance

Burger Round 1



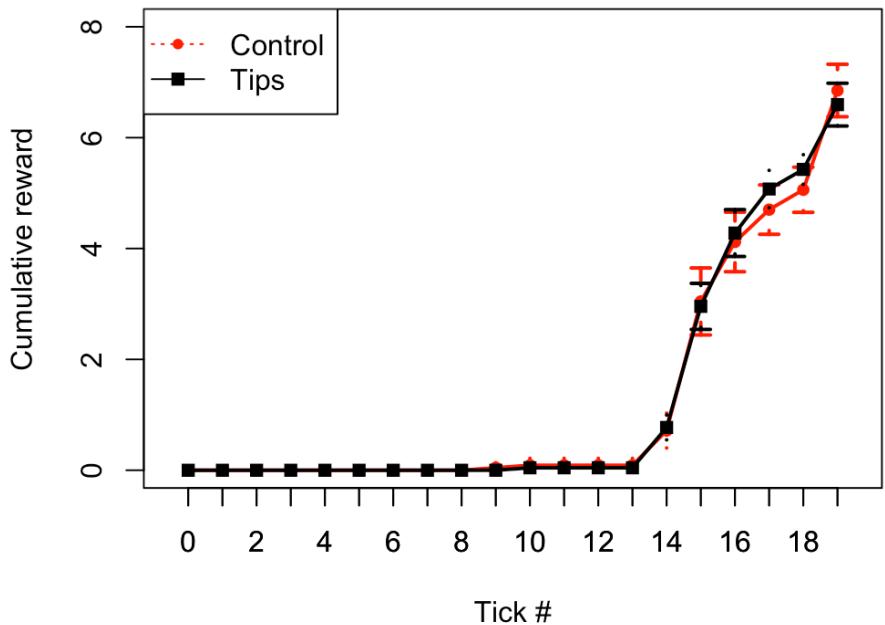
Burger Round 3



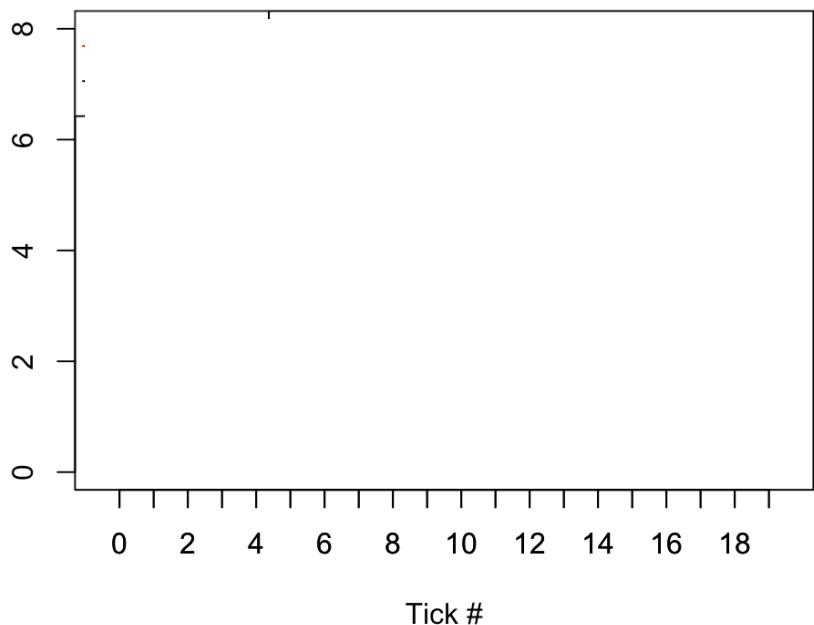
Control: Over time, people improve.

Results

Burger Round 1



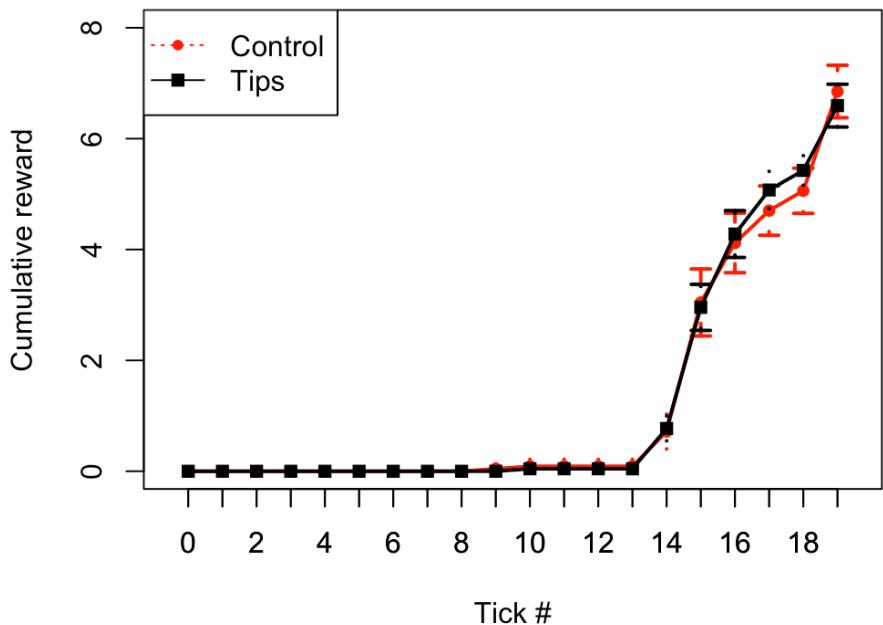
Burger Round 3



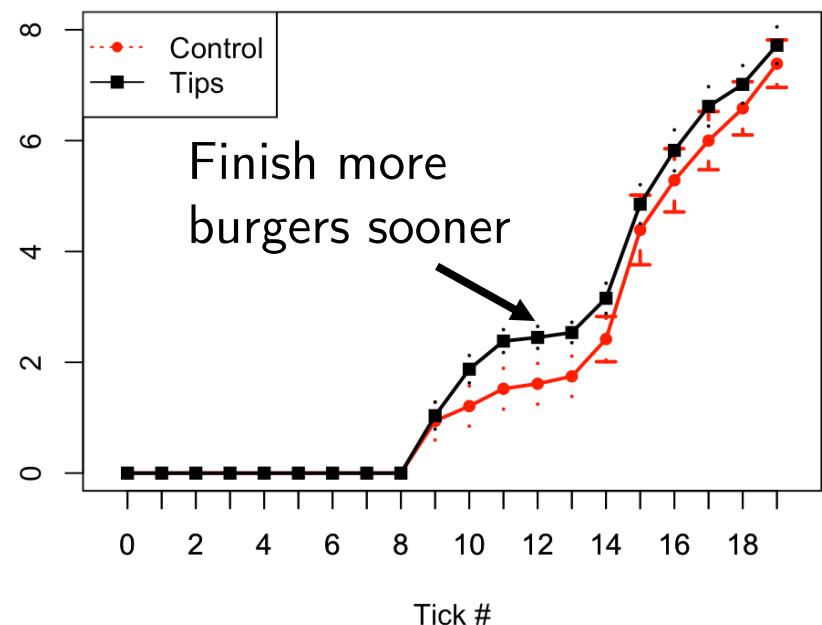
Results

Tips Help

Burger Round 1



Burger Round 3



With Tips: Higher score, improve faster

Results

Tips Help

Suboptimal Decisions

e.g., assigning high-skilled chef to plate a dish

%	Burger #1	#2	#3
Control	5.49	3.88	3.31
Tips	2.78	1.53	1.30

With Tips: Higher score, improve faster, make fewer suboptimal decisions

Results

Tips Help

Suboptimal Decisions e.g., assigning high-skilled chef to plate a dish

%	Burger #1	#2	#3	Pasta #1	#2	#3
Control	5.49	3.88	3.31	3.26	2.81	2.54
Tips	2.78	1.53	1.30	1.38	1.16	0.84

With Tips: Higher score, improve faster, make fewer suboptimal decisions

Results

Tips Help

Suboptimal Decisions e.g., assigning high-skilled chef to plate a dish

%	Burger #1	#2	#3	Pasta #1	#2	#3
Control	5.49	3.88	3.31	3.26	2.81	2.54
Tips	2.78	1.53	1.30	1.38	1.16	0.84

With Tips: Higher score, improve faster, make fewer suboptimal decisions

In our setting, people can potentially improve ✓

From experiments
Experts + Trainee

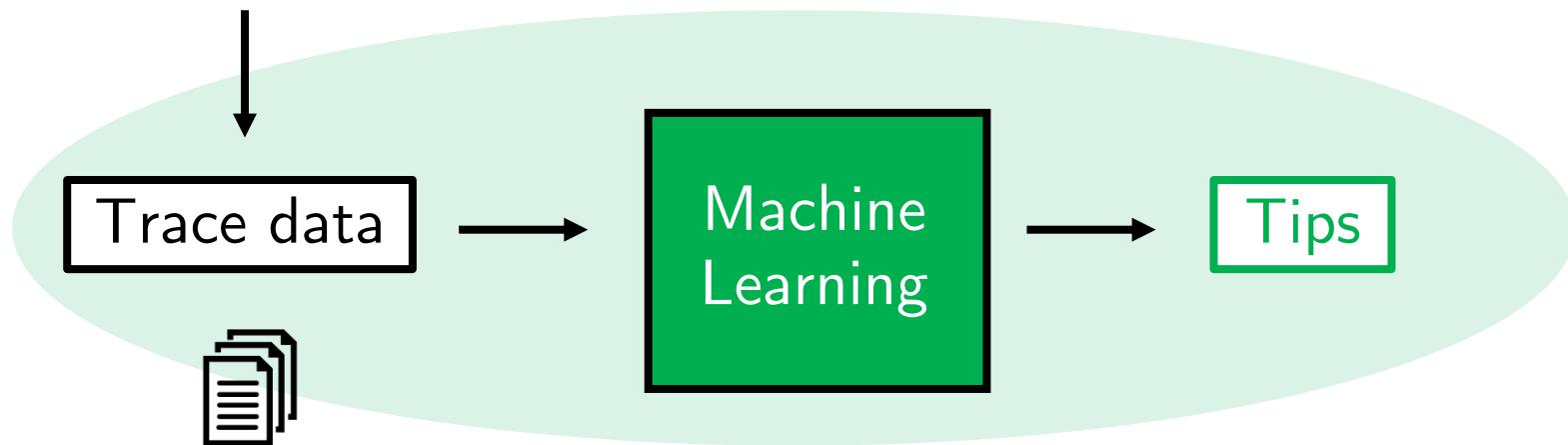


Trace data



Testing Our Framework

From experiments
Experts + Trainee



Testing Our Framework

From experiments
Experts + Trainee

Extract policies

Trace data

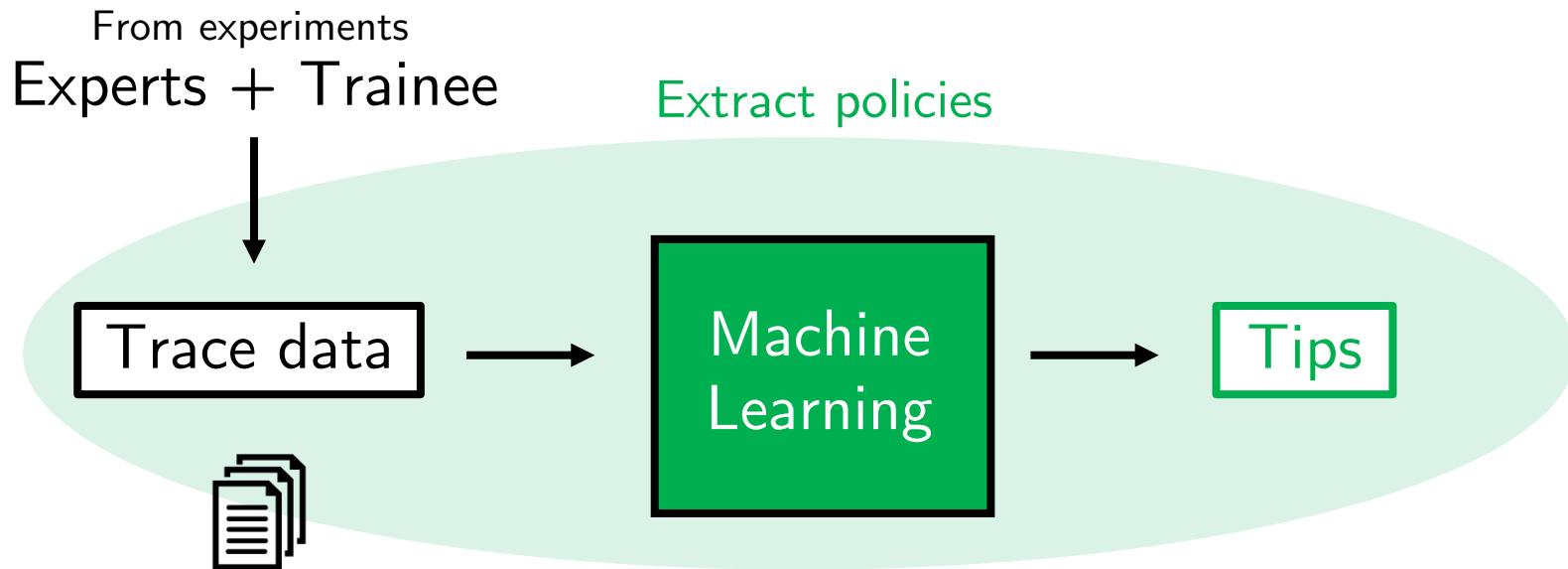
Machine Learning

Tips

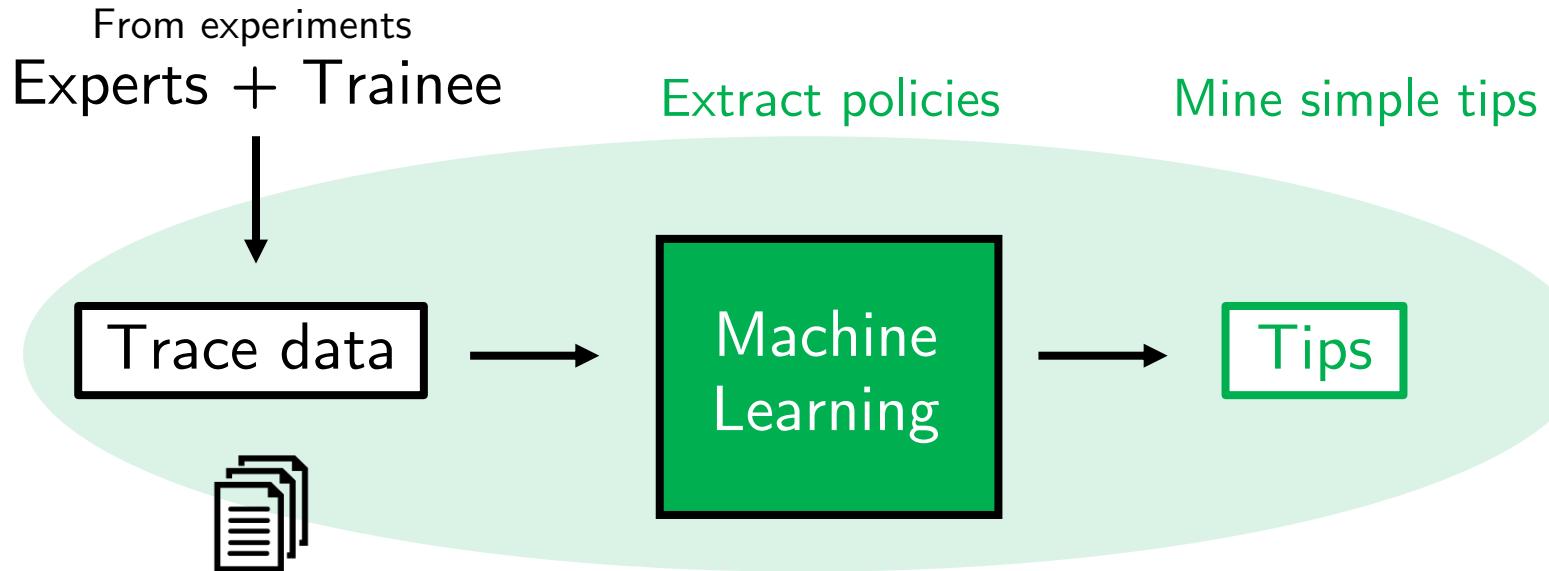


turns needed to finish 4 burgers
(smaller is better)

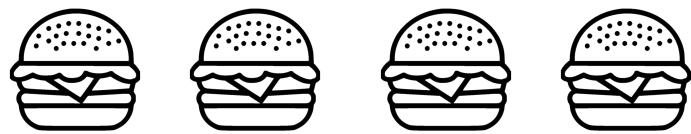
Testing Our Framework



Testing Our Framework



Extracting Good Tips



	Trainee	Experts	Optimal
	26	25	19

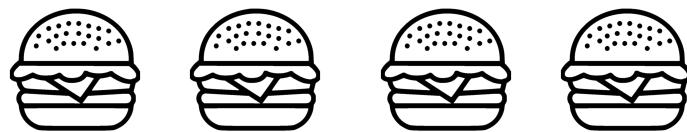
Available Workers

Open Tasks



Action

Extracting Good Tips



	Trainee	Experts	Optimal
	26	25	19

Available Workers

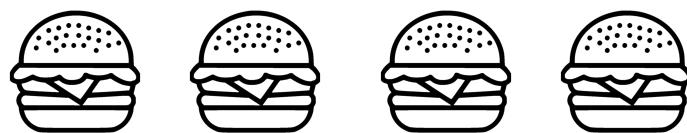
Open Tasks



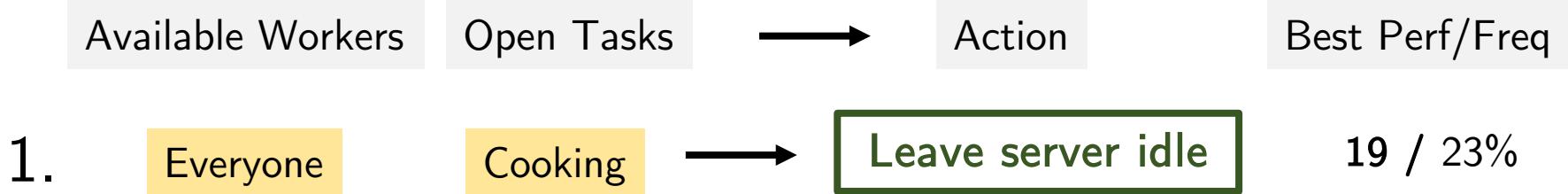
Action

Best Perf/Freq

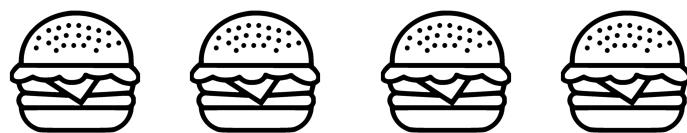
Extracting Good Tips



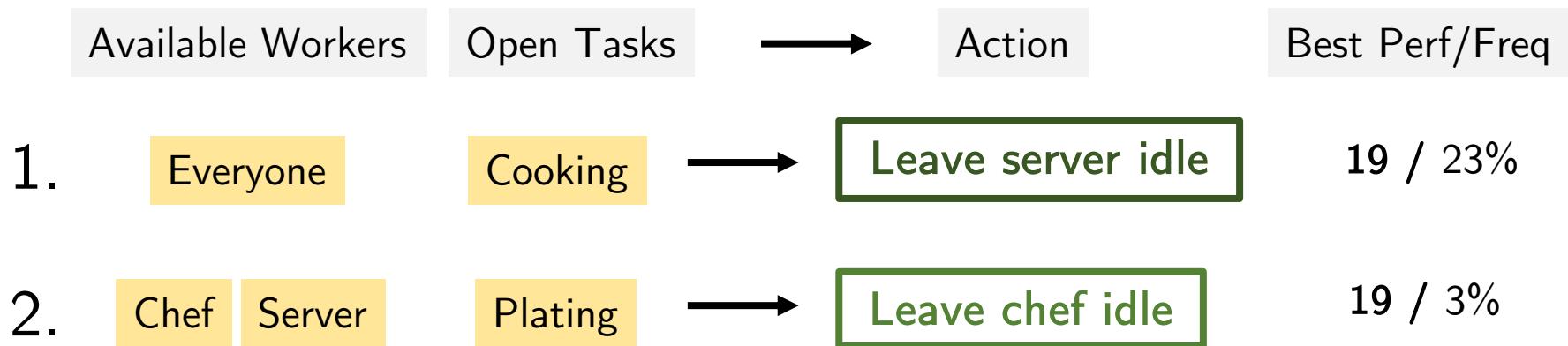
	Trainee	Experts	Optimal
	26	25	19



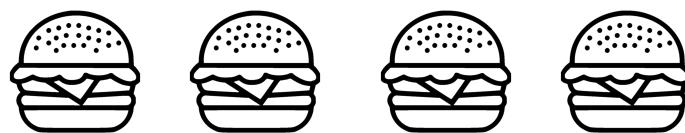
Extracting Good Tips



	Trainee	Experts	Optimal
	26	25	19



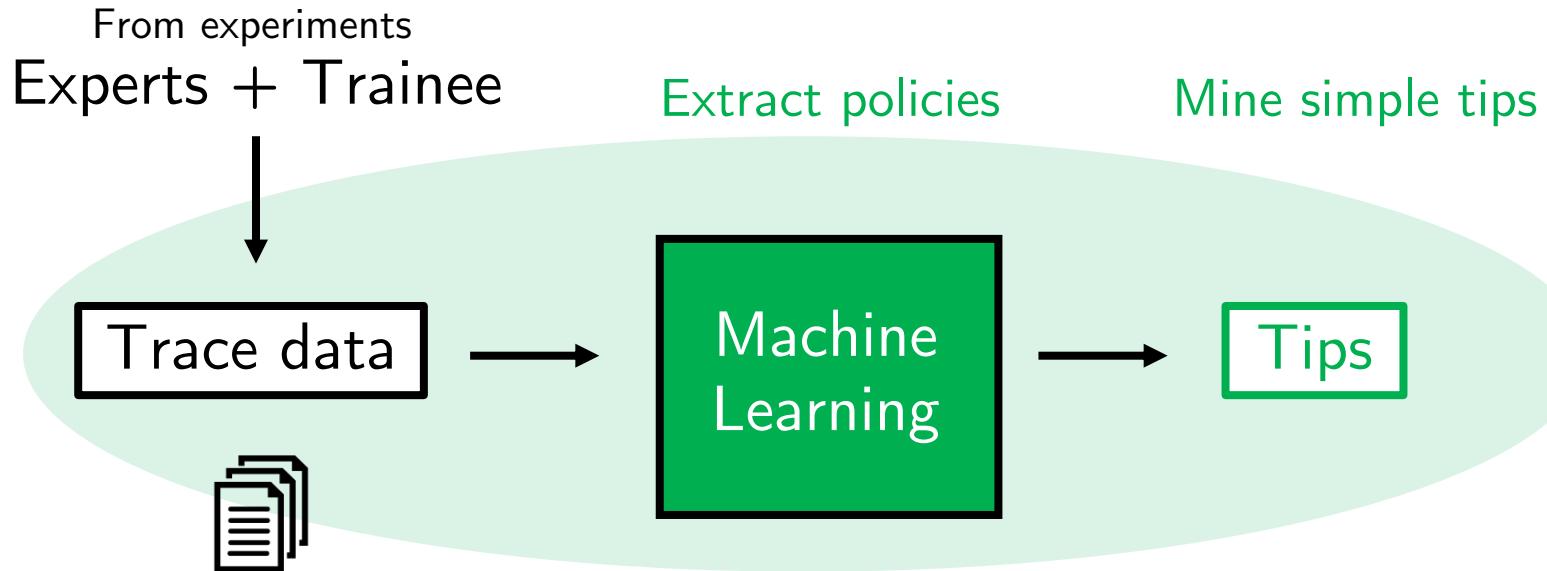
Extracting Good Tips



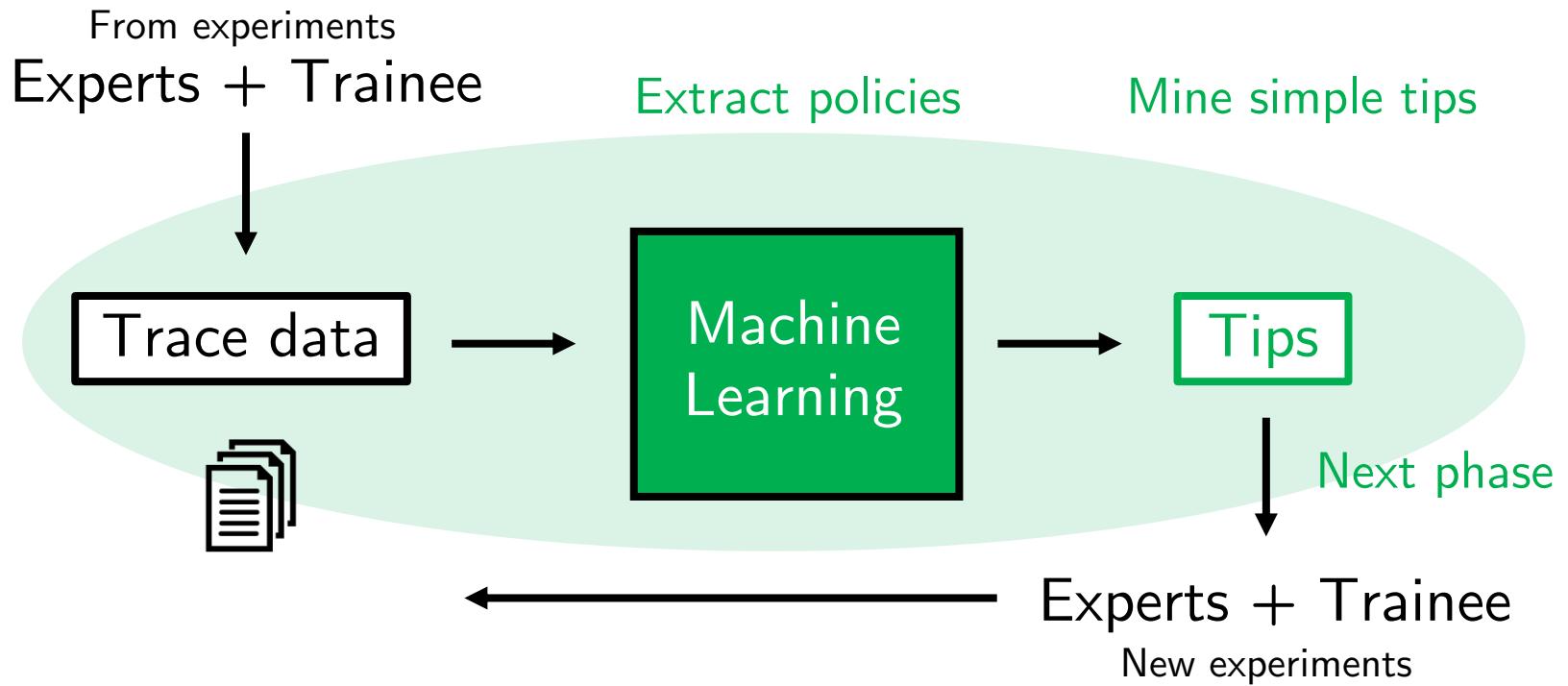
	Trainee	Experts	Optimal
	26	25	19

	Available Workers	Open Tasks	Action	Best Perf/Freq
1.	Everyone	Cooking	Leave server idle	19 / 23%
2.	Chef Server	Plating	Leave chef idle	19 / 3%
3.	Sous Chef	Cooking	Assign to sous chef	19 / 2%

Testing Our Framework



Testing Our Framework



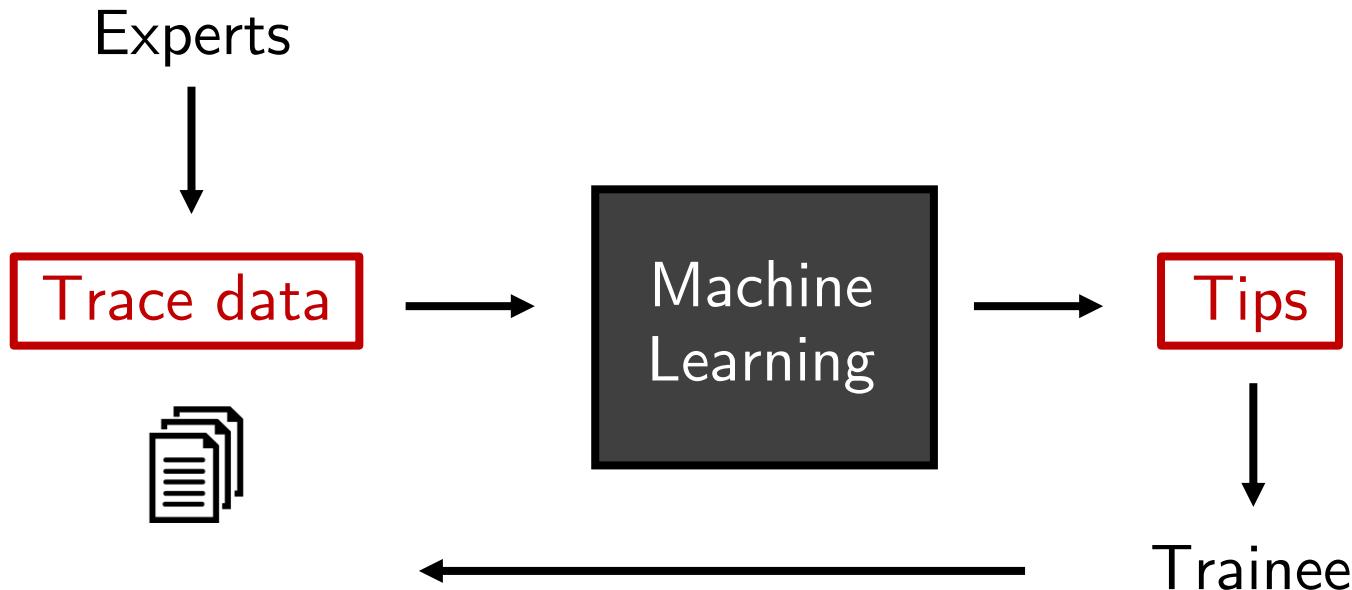
On-Going Work

- Implement personalized tips in the next phase of experiments, dynamically choosing tips
- Improve tips learning algorithms
 - Can use imitation learning as well
 - Fit a model predicting when $(s, \pi^*(a^*))$ differ $(s, \pi^*(a))$
- Extend to team collaboration



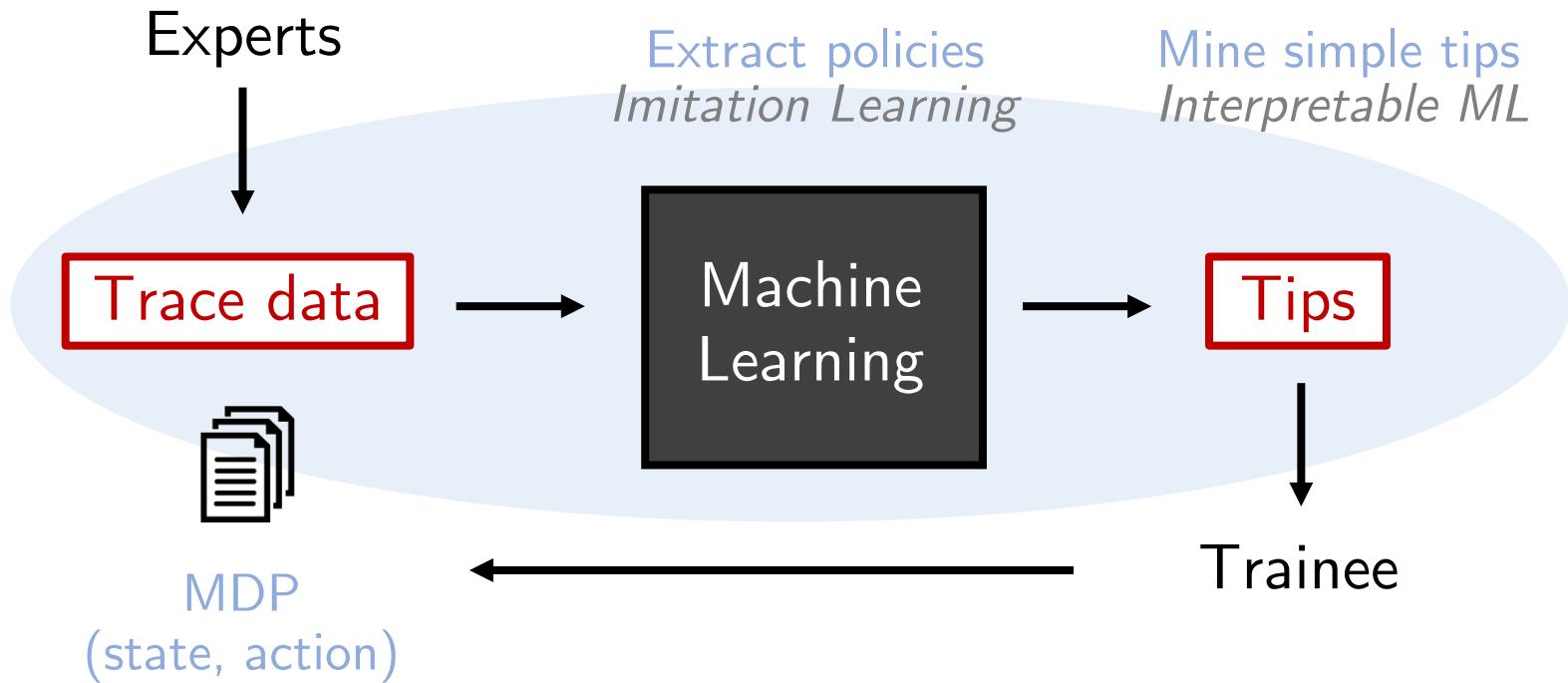
Summary

ML to automatically help people improve in a personalized and dynamic way



Summary

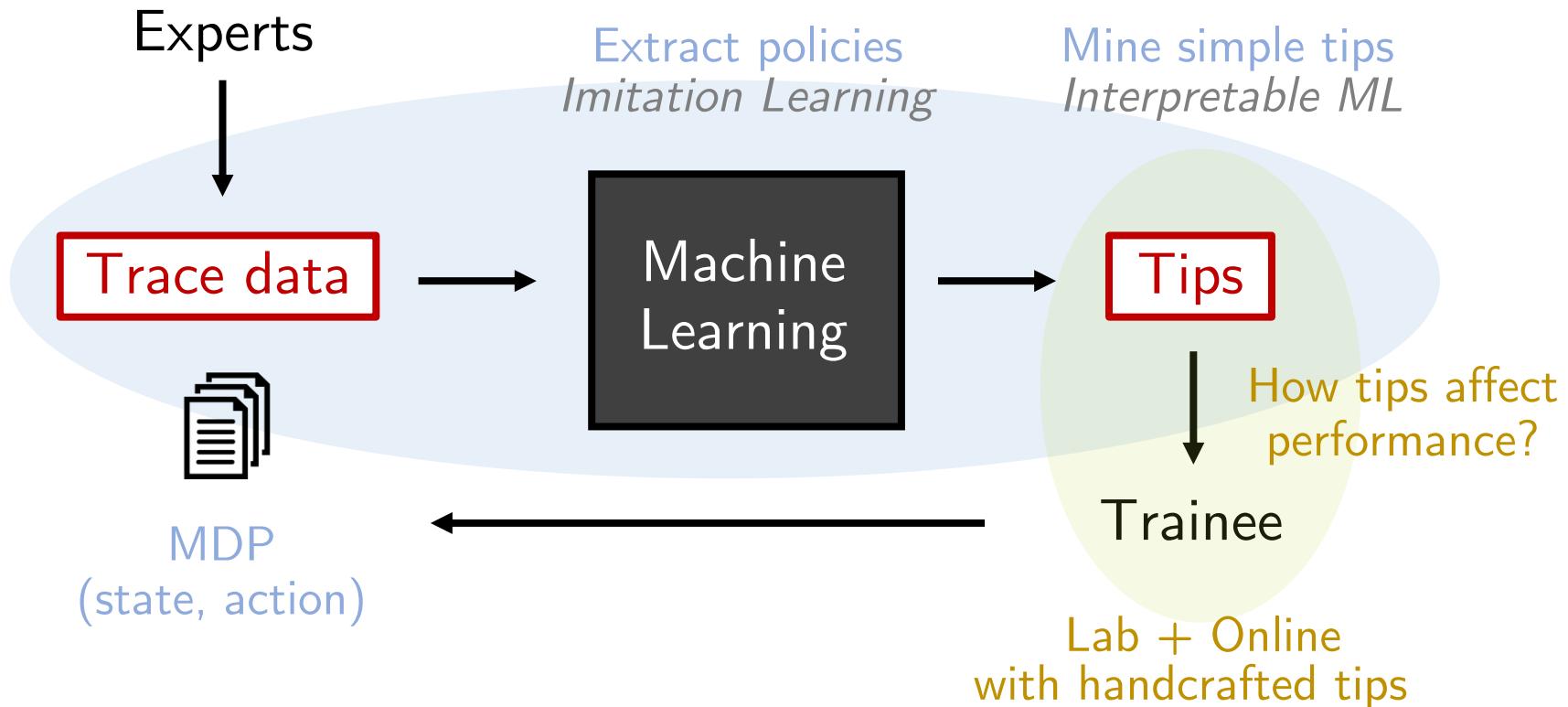
ML to automatically help people improve
in a personalized and dynamic way



Today: Framework

Summary

ML to automatically help people improve in a personalized and dynamic way



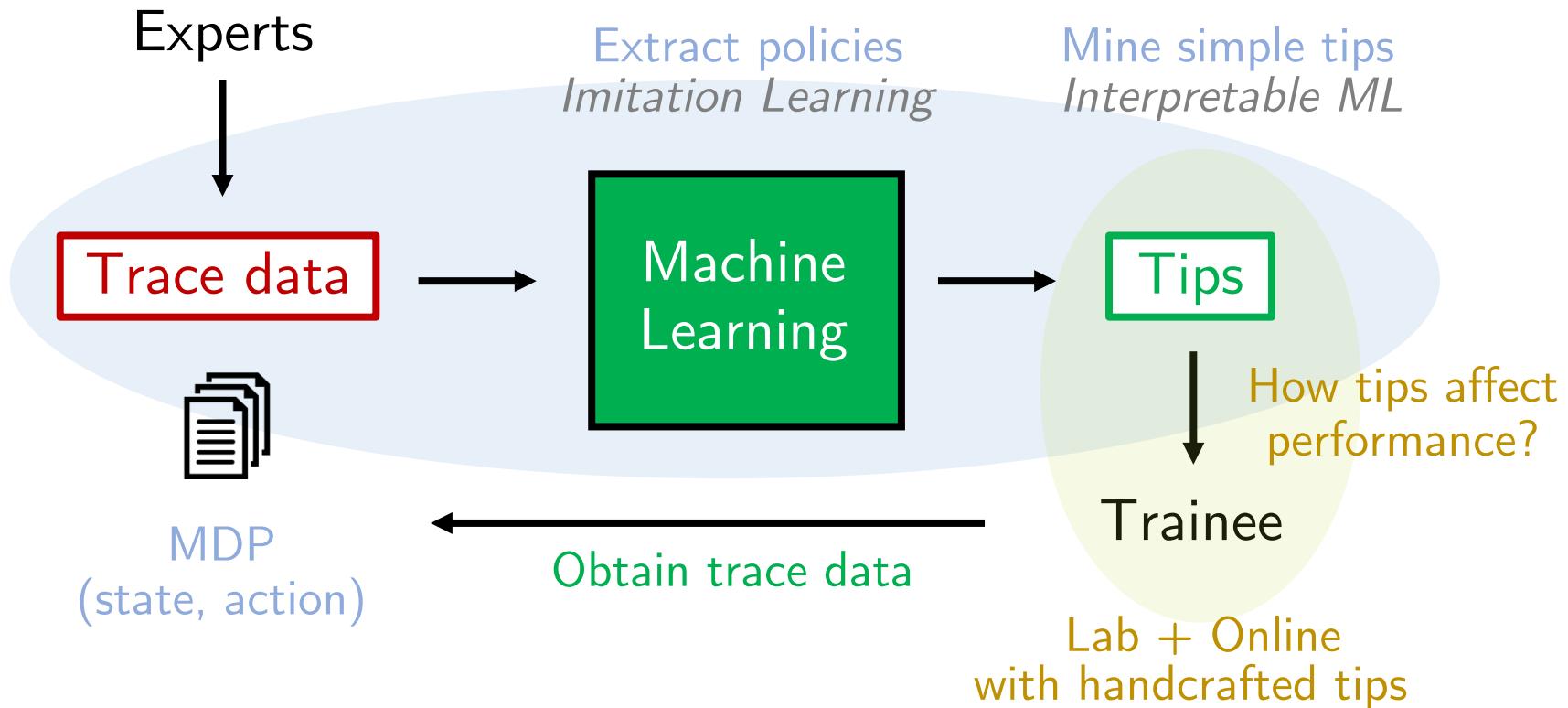
Today:

Framework

Pilot Experiments

Summary

ML to automatically help people improve in a personalized and dynamic way



Today:

Framework

Pilot Experiments

Policies + Tips

Policy Gradient

Recall that we are given a Markov Decision Process (S, A, P, R) denoting states, actions, transition probabilities, and rewards respectively. We seek an optimal policy $\pi^* : S \rightarrow A$ over a horizon of T time steps and a discount factor γ . In order to perform gradient descent, we will featurize all quantities of interest, so that we can parametrize π by some θ .

We start by defining the value function:

$$V^{(\theta)}(s) = R(s) + \gamma \sum_{s' \in S} \sum_{a \in A} \pi(a|s) \cdot P(s'|s, a) \cdot V^{(\theta)}(s') .$$

The cost-to-go function is then

$$J(\theta) = \mathbb{E}_{\substack{s \sim D^{(\pi)} \\ a \sim \pi(\cdot|s)}} V^{(\theta)}(s) .$$

Sometimes it is easier to work with the Q -function:

$$Q^{(\theta)}(s, a) = R(s) + \gamma \sum_{s' \in S} P(s'|s, a) \cdot V^{(\theta)}(s') .$$

We wish to optimize $J(\theta)$ as a function of our policy using gradient descent techniques. Therefore, we require the *policy gradient*:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\substack{s \sim D^{(\pi)} \\ a \sim \pi(\cdot|s)}} \left[Q^{(\theta)}(s, a) \cdot \nabla_\theta \log \pi_\theta(a|s) \right]$$

Policy Gradient

However, we clearly do not know the Q -function, so we cannot directly compute this gradient. Instead, we will estimate it using the following simple procedure:

1. Sample N rollouts using π_θ , *i.e.*,

$$\zeta = (s_0, a_0, r_0, s_1, a_1, r_1, \dots, s_{T-1}, a_{T-1}, r_{T-1}).$$

2. Estimate the Q -function as

$$\hat{Q}_t = \sum_{t'=t}^{T-1} \gamma^{t'-t} \cdot \tilde{r}_{t'},$$

where \tilde{r}_t are normalized rewards, *i.e.*, $\tilde{r}_t = \frac{r_t - \mu}{\sigma}$, where μ and σ are the average and standard deviation of the rewards from the most recent N rollouts

3. Estimate the policy gradient

$$\nabla_\theta J(\theta) \approx \sum_{t=0}^{T-1} \hat{Q}_t \cdot \nabla_\theta \log \pi_\theta(a_t | s_t).$$

The above is for a single rollout. It is recommended to average this quantity over all N rollouts.

Feeding a randomly chosen initial policy as well as the gradient computation procedure above to an optimizer like ADAM in Pytorch should yield an estimated optimal policy.