Learning Best Practices

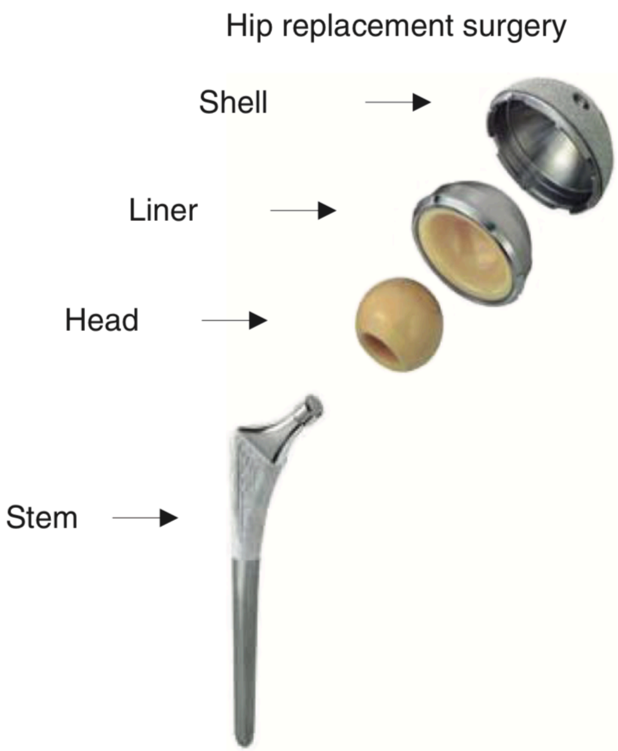
Can Machine Learning Substitute for Experiences?

INFORMS 2019

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

Learning Takes Time

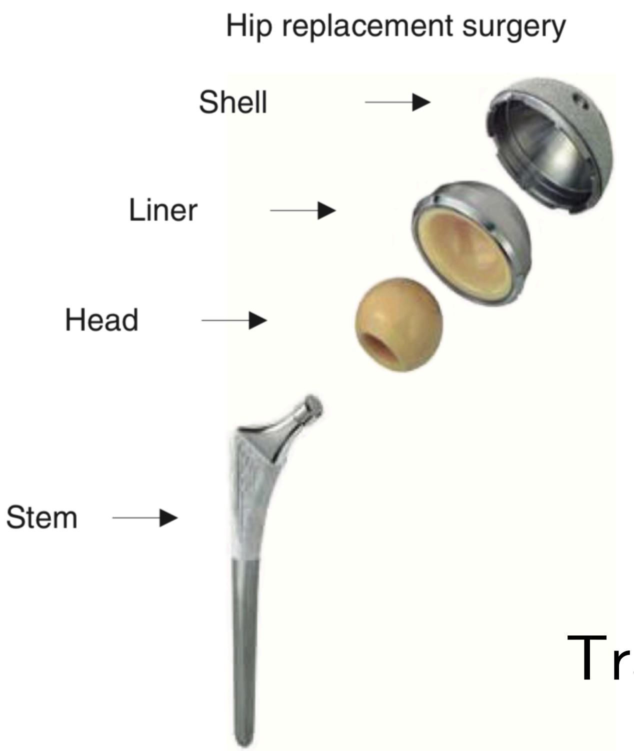
Learning Takes Time



"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

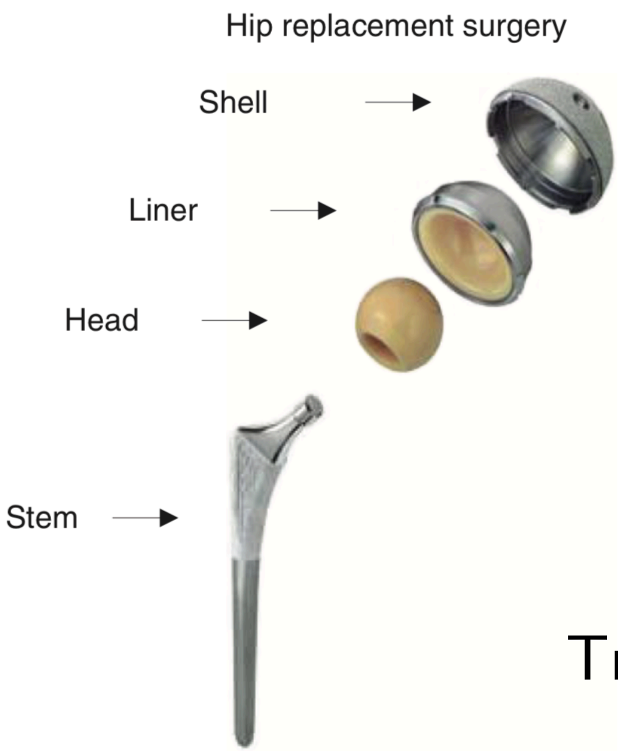


"The first use of certain device versions can result in at least a **32.4% increase in surgery duration**, hurting quality and productivity."

- 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."

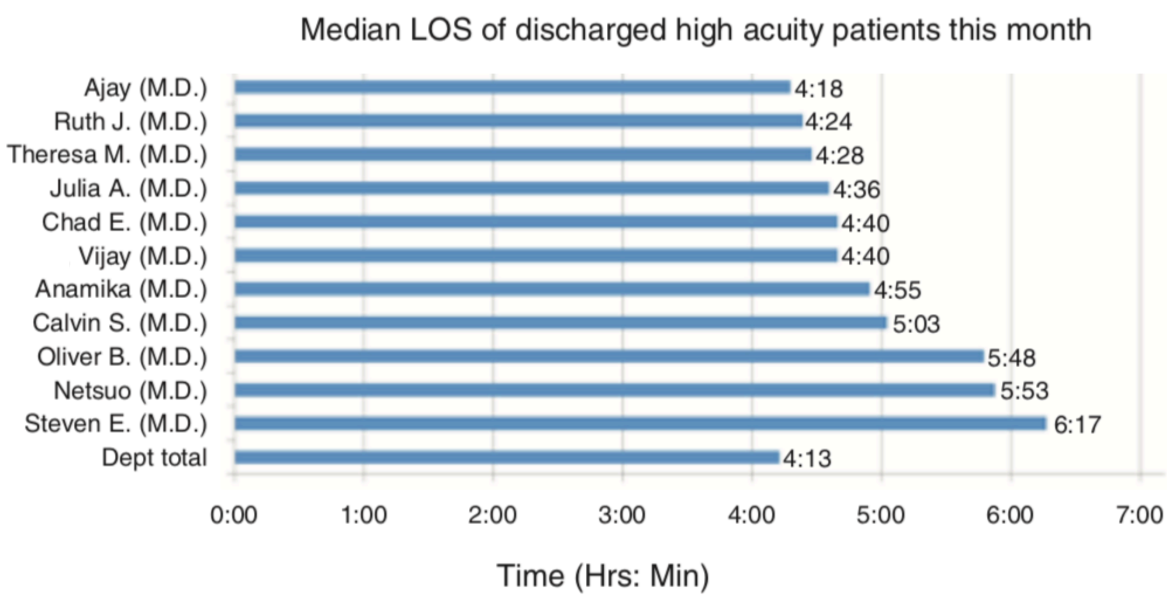
- Ramdas et al. 2018

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

Simple Tips?

Learning from Experts

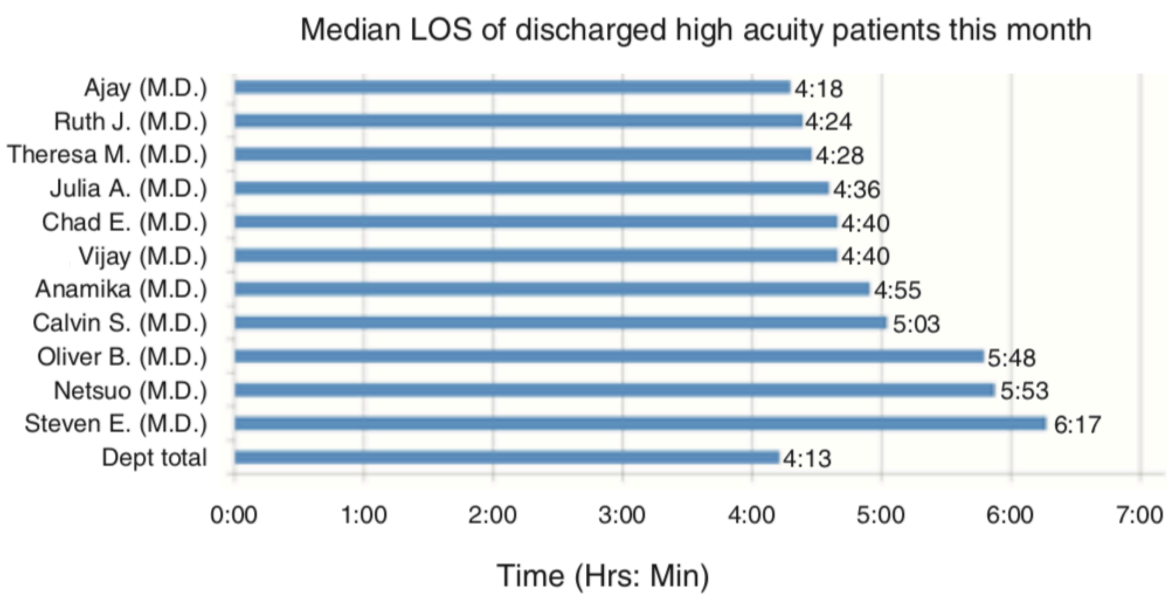
Learning from Experts



"The public disclosure of RPF allowed workers to **identify their top-performing coworkers**, which in turn **enabled the identification and validation of best practices within the work group.**"

- Song et al. 2018

Learning from Experts

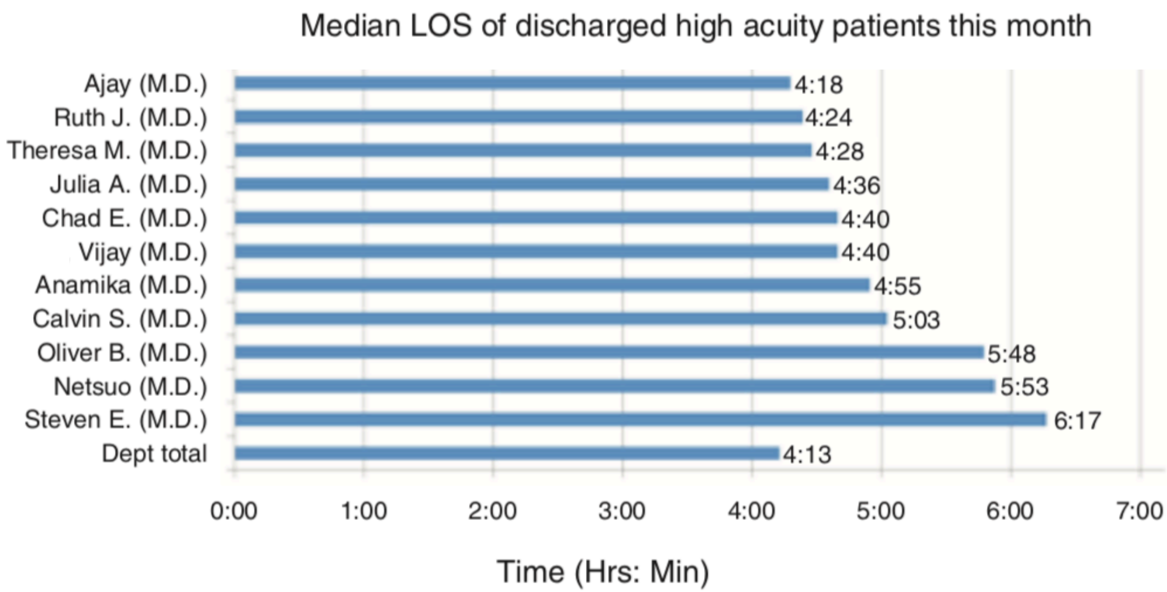


"The public disclosure of RPF allowed workers to **identify their top-performing coworkers,** which in turn **enabled the identification and validation of best practices within the work group.**"

- Song et al. 2018

But It's Still Hard

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 within the work group.**"

- Song et al. 2018

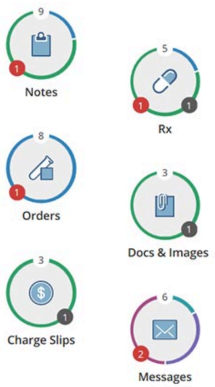
But It's Still Hard

Data!

Trace Data is Everywhere

Trace Data is Everywhere

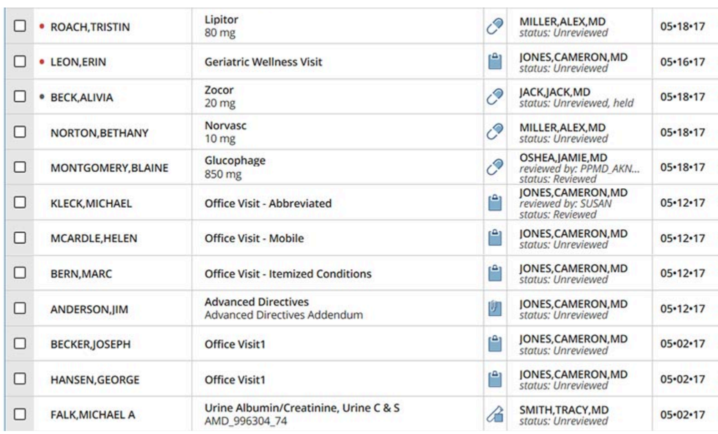
Physicians



ROACH,TRISTIN	Lipitor 80 mg		MILLER,ALEX,MD *status: Unreviewed*	05•18•17
LEON,ERIN	Geriatric Wellness Visit		JONES,CAMERON,MD *status: Unreviewed*	05•16•17
BECK,ALIVIA	Zocor 20 mg		JACK,JACK,MD *status: Unreviewed, held*	05•18•17
NORTON,BETHANY	Norvasc 10 mg		MILLER,ALEX,MD *status: Unreviewed*	05•18•17
MONTGOMERY,BLAINE	Glucophage 850 mg		OSHEA,JAMIE,MD *reviewed by: PPMD_AKN...* *status: Reviewed*	05•18•17
KLECK,MICHAEL	Office Visit - Abbreviated		JONES,CAMERON,MD *reviewed by: SUSAN* *status: Reviewed*	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,JIM	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

Trace Data is Everywhere

Physicians



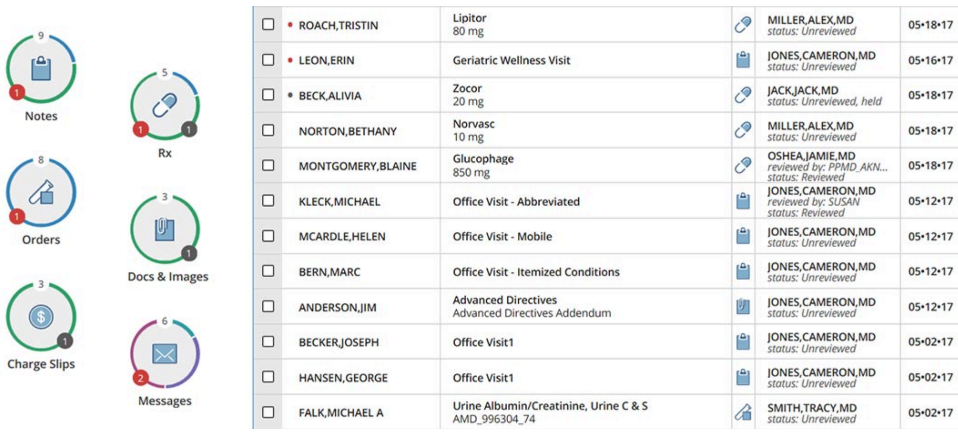
ROACH,TRISTIN	Lipitor 80 mg		MILLER,ALEX,MD status: Unreviewed	05•18•17
LEON,ERIN	Geriatric Wellness Visit		JONES,CAMERON,MD status: Unreviewed	05•16•17
BECK,ALIVIA	Zocor 20 mg		JACK,JACK,MD status: Unreviewed, held	05•18•17
NORTON,BETHANY	Norvasc 10 mg		MILLER,ALEX,MD status: Unreviewed	05•18•17
MONTGOMERY,BLAINE	Glucophage 850 mg		OSHEA,JAMIE,MD reviewed by: PPMD_AKN... status: Reviewed	05•18•17
KLECK,MICHAEL	Office Visit - Abbreviated		JONES,CAMERON,MD reviewed by: SUSAN status: Reviewed	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,JIM	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

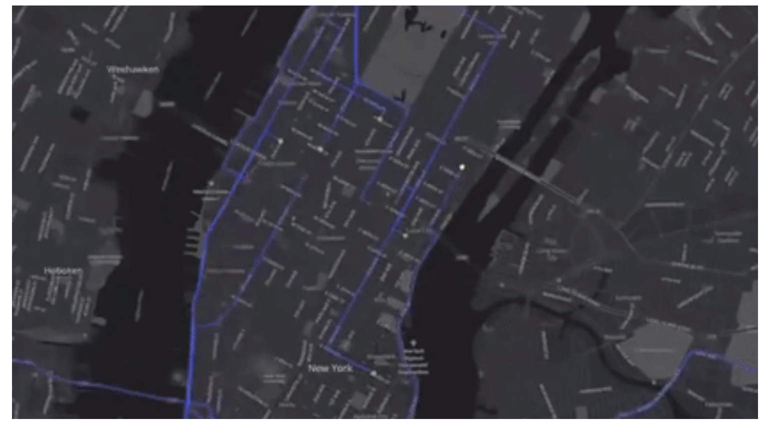
Trace Data is Everywhere

Physicians



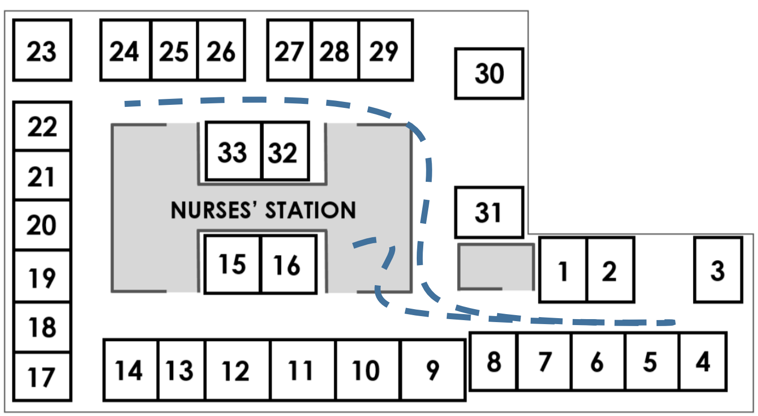
ROACH,TRISTIN	Lipitor 80 mg	MILLER,ALEX,MD status: Unreviewed	05•18•17
LEON,ERIN	Geriatric Wellness Visit	JONES,CAMERON,MD status: Unreviewed	05•16•17
BECK,ALIVIA	Zocor 20 mg	JACK,JACK,MD status: Unreviewed, held	05•18•17
NORTON,BETHANY	Norvasc 10 mg	MILLER,ALEX,MD status: Unreviewed	05•18•17
MONTGOMERY,BLAINE	Glucophage 850 mg	OSHEA,JAMIE,MD reviewed by: PPMD_AKN... status: Reviewed	05•18•17
KLECK,MICHAEL	Office Visit - Abbreviated	JONES,CAMERON,MD reviewed by: SUSAN status: Reviewed	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,JIM	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

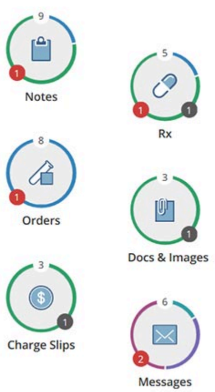
Nurses

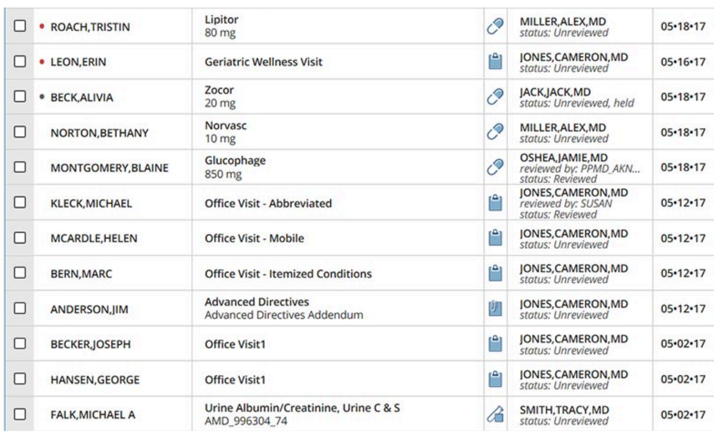


Meng et al. 2018

Trace Data is Everywhere

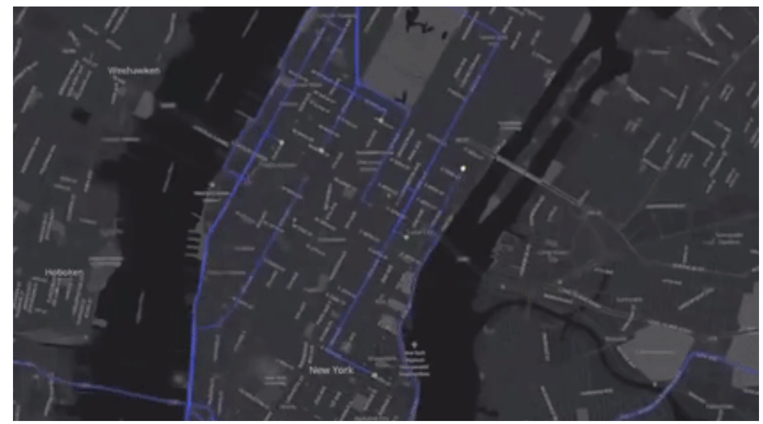
Physicians





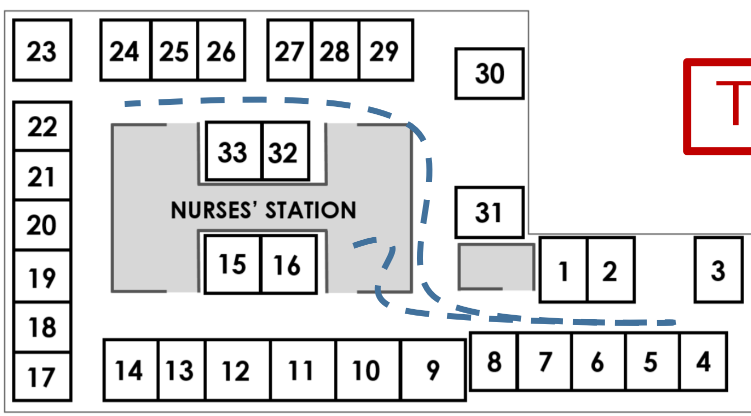
ROACH,TRISTIN	Lipitor 80 mg	MILLER,ALEX,MD status: Unreviewed	05•18•17
LEON,ERIN	Geriatric Wellness Visit	JONES,CAMERON,MD status: Unreviewed	05•16•17
BECK,ALIVIA	Zocor 20 mg	JACK,JACK,MD status: Unreviewed, held	05•18•17
NORTON,BETHANY	Norvasc 10 mg	MILLER,ALEX,MD status: Unreviewed	05•18•17
MONTGOMERY,BLAINE	Glucophage 850 mg	OSHEA,JAMIE,MD reviewed by: PPMD_AKN... status: Reviewed	05•18•17
KLECK,MICHAEL	Office Visit - Abbreviated	JONES,CAMERON,MD reviewed by: SUSAN status: Reviewed	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,JIM	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



Meng et al. 2018

Trace data

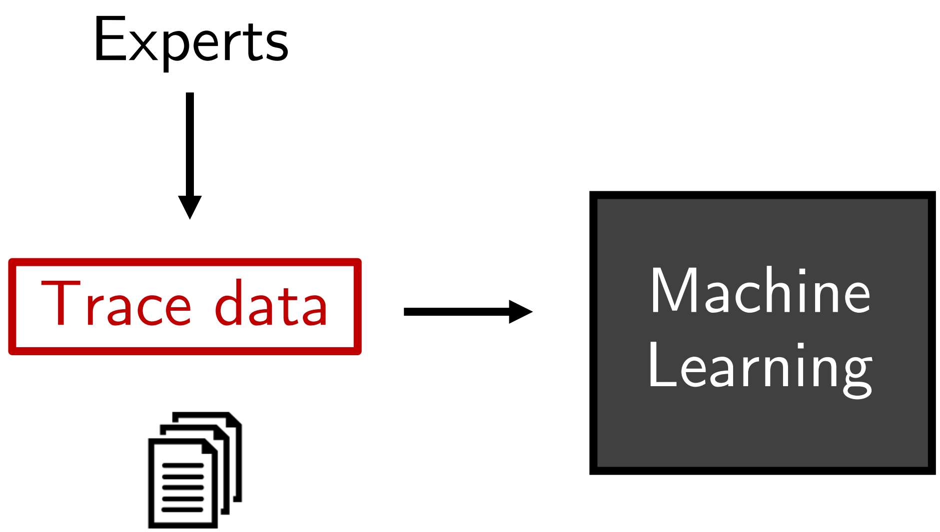
Tips

Our Paper

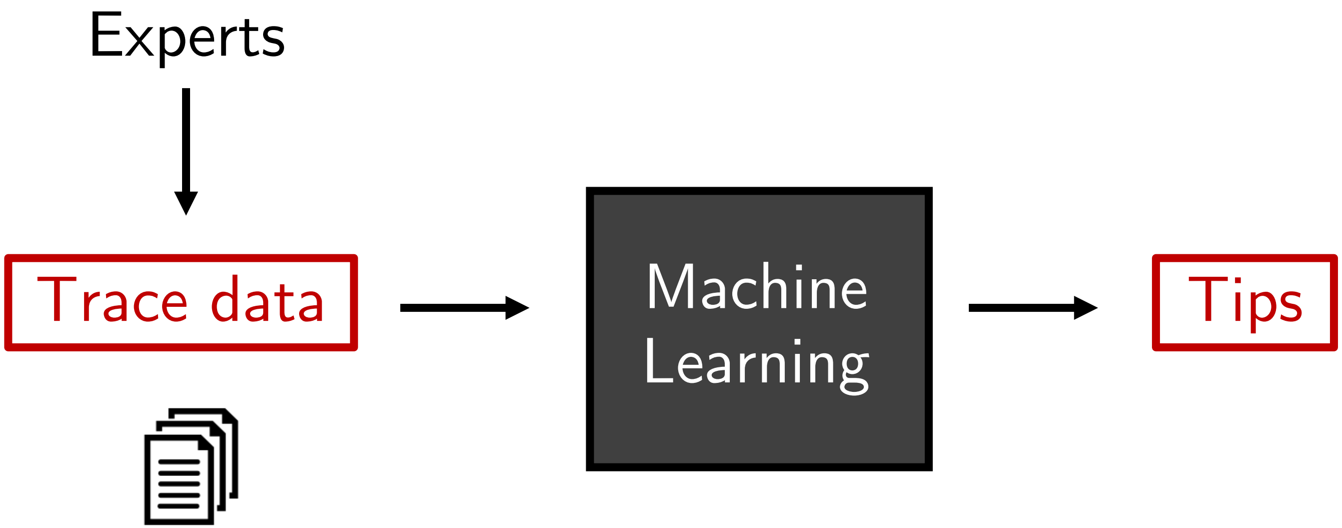
Experts

Trace data

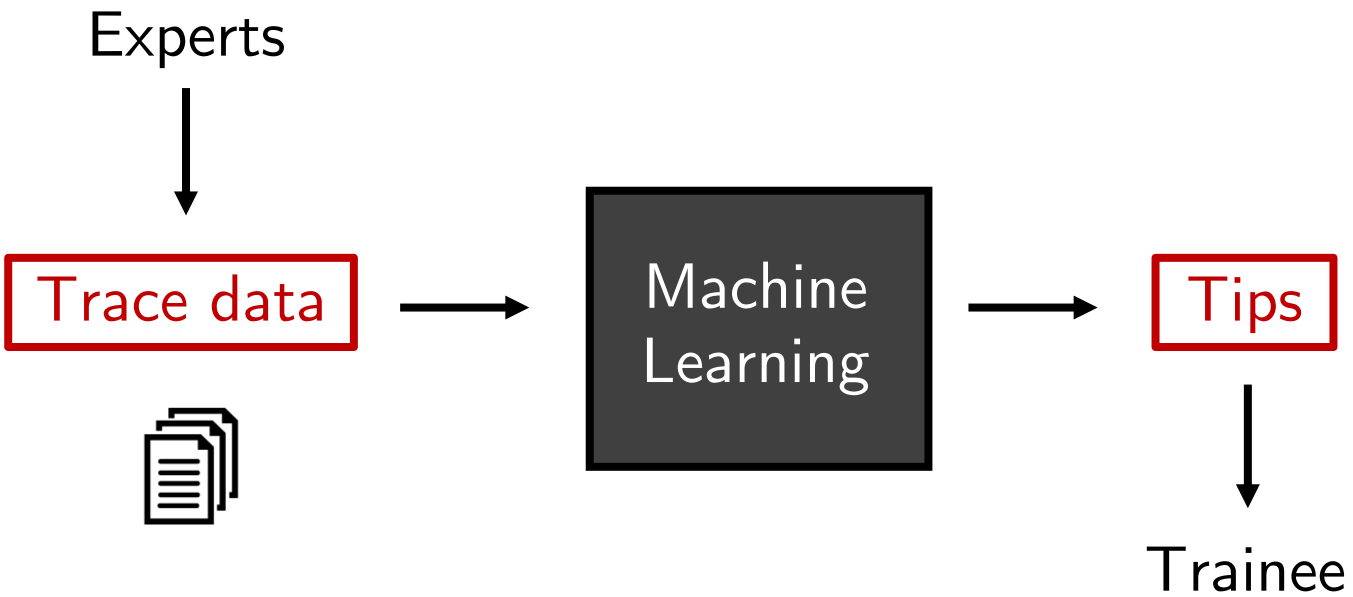
Our Paper



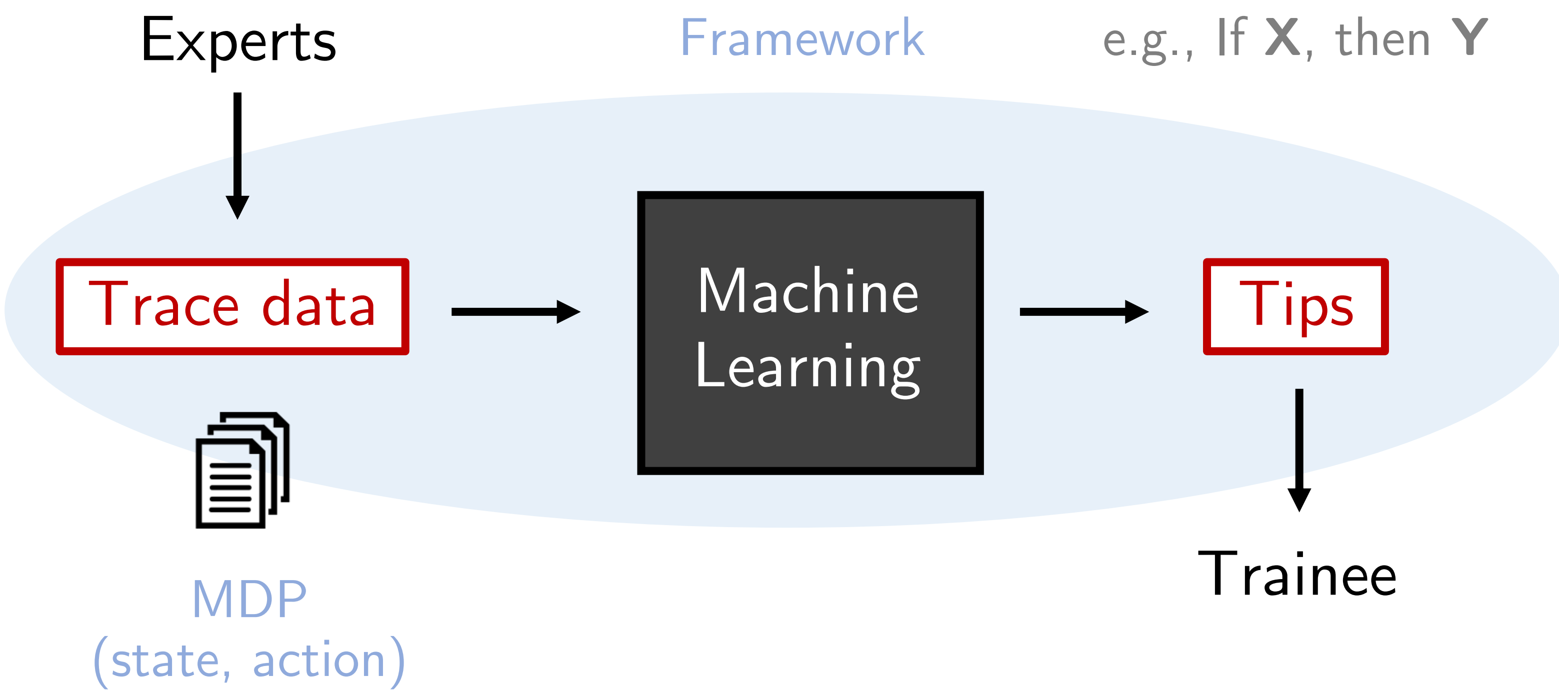
Our Paper



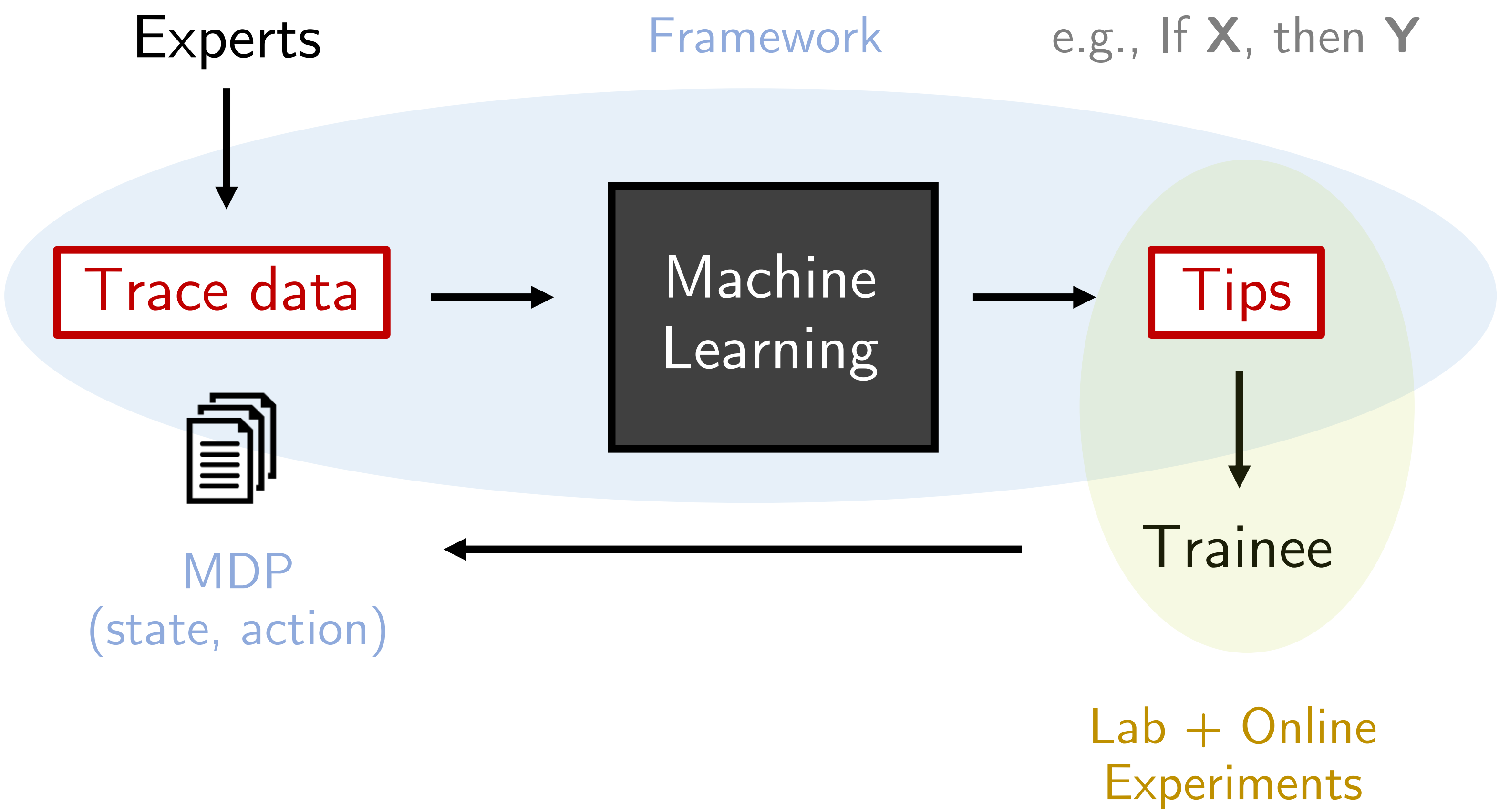
Our Paper



Our Paper


Experts
Framework
e.g., If X, then Y
Trace data
Machine Learning
Tips
MDP
(state, action)
Trainee

Our Paper



Experts
Framework
e.g., If **X**, then **Y**
Trace data
Machine Learning
Tips
Trainee
MDP
(state, action)
Lab + Online
Experiments

Our Framework

Policy = the probability distribution of actions given a state

Our Framework

Trace Data

1

Trainee Policy

π^{h}

Imitation Learning

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

Policy = the probability distribution of actions given a state

Our Framework

Trace Data

1 Trainee Policy

π^{h}

Imitation Learning

Trace Data

2 Expert Policy

π^{*}

Imitation Learning

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

Our Framework

Trace Data

1 Trainee Policy

π^h

Imitation Learning

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

Trace Data

2 Expert Policy

π^*

Imitation Learning

Deep RL $\leftarrow$ Known MDP

Our Framework

Trace Data

Trace Data

1 Trainee Policy

2 Expert Policy

$$\pi^{\text{tip}} = \Delta\left(\pi^{h}, \pi^{*}\right)$$

Imitation Learning

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

Imitation Learning

Deep RL

Our Framework

Trace Data

Trace Data

1 Trainee Policy

2 Expert Policy

$$\pi^{\text{tip}} = \Delta\left(\boxed{\pi^{h}}, \boxed{\pi^{*}} \right)$$

Imitation Learning

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

Imitation Learning

Deep RL

3 Tip Mining

Interpretable ML

Letham et al 2015

Learning Trainee Policy

- Supervised Learning
  - Random Forests/
    Gradient Boosting Machine

State $\longrightarrow$ Action

1

Trainee Policy

π^h

Learning Trainee Policy

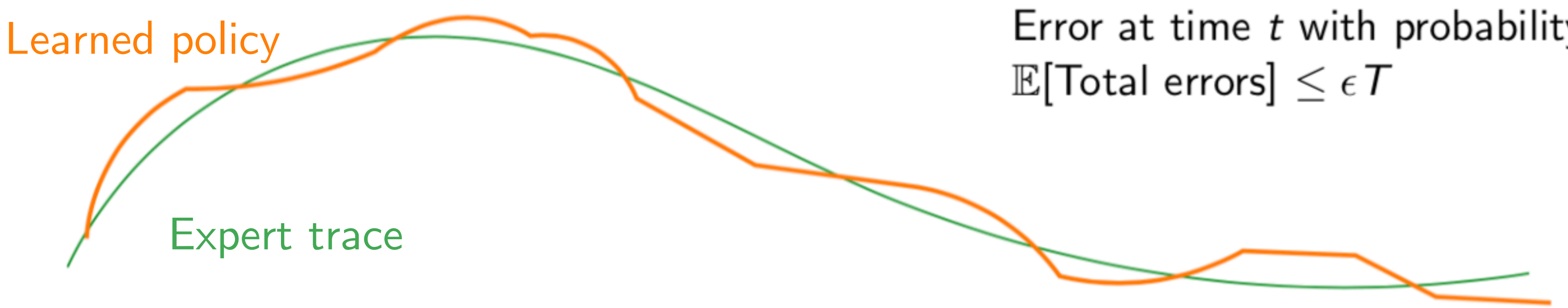
- Supervised Learning
  - Random Forests/ Gradient Boosting Machine

1

Trainee Policy

π^h

State → Action



Error at time t with probability ϵ

$\mathbb{E}[\text{Total errors}] \leq \epsilon T$

Learning Trainee Policy

- Supervised Learning
  - Random Forests/ Gradient Boosting Machine

1

Trainee Policy

π^h

State → Action

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

Learning Trainee Policy

- Supervised Learning
  - Random Forests/
    Gradient Boosting Machine

1

Trainee Policy

π^h

State	→	Action
		0
		0
		0
		0
		0
		1
		0
		0

→ Predict 0

Learning Trainee Policy

- Supervised Learning
  - Random Forests/
    Gradient Boosting Machine

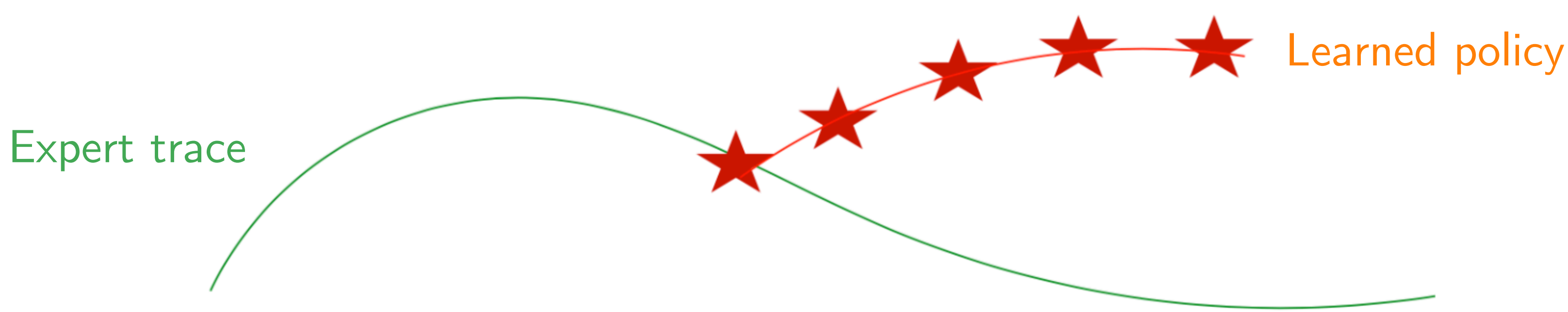
1

Trainee Policy

π^h

State $\longrightarrow$ Action

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



Learning Trainee Policy

- Supervised Learning
  - Random Forests/
    Gradient Boosting Machine

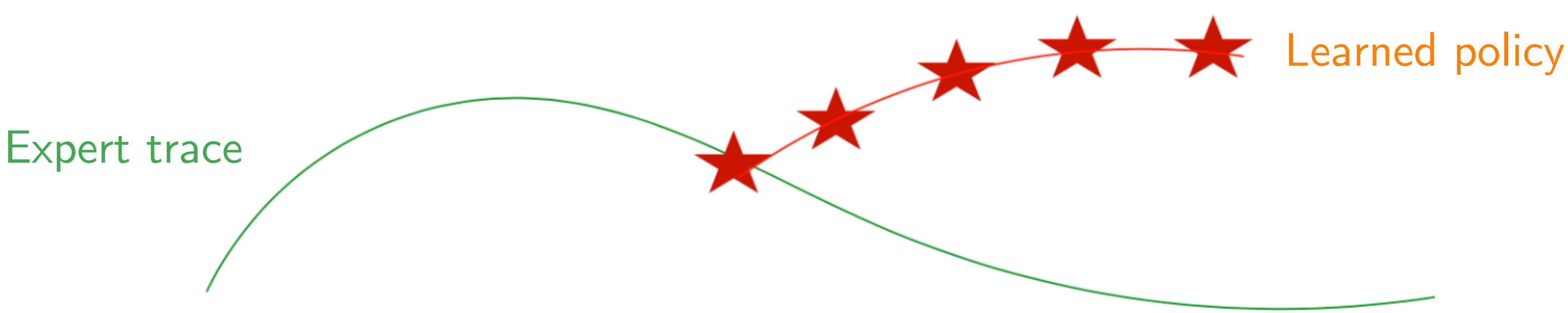
1

Trainee Policy

π^{h}

State $\longrightarrow$ Action

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



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

- 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.

1

Trainee Policy

π^{h}

Imitation Learning

- Generative Adversarial IL
  - Learn the policy directly

1

Trainee Policy

$\boldsymbol{\pi}^{\mathrm{h}}$

- Ho & Ermon 2016

Imitation Learning

- Generative Adversarial IL
  - Learn the policy directly
  - Two neural networks:
    - Train a policy $\boldsymbol{\pi}^{\boldsymbol{\theta}}$ (gradient descent)
    - Discriminator (trained on trace data)

1

Trainee Policy

$\boldsymbol{\pi}^{\mathrm{h}}$

- Ho & Ermon 2016

Imitation Learning

- Generative Adversarial IL
  - Learn the policy directly
  - Two neural networks:
    - Train a policy $\boldsymbol{\pi}^{\boldsymbol{\theta}}$ (gradient descent)
    - Discriminator (trained on trace data)
    - Find $\boldsymbol{\pi}^{\boldsymbol{\theta}}$ such that discriminator cannot distinguish between states from $\boldsymbol{\pi}^{\boldsymbol{\theta}}$ as opposed to those from $\boldsymbol{\pi}^{\mathbf{h}}$

1

Trainee Policy

$\boldsymbol{\pi}^{\mathrm{h}}$

- Ho & Ermon 2016

Imitation Learning

- Generative Adversarial IL
  - Learn the policy directly
  - Two neural networks:
    - Train a policy $\boldsymbol{\pi}^{\boldsymbol{\theta}}$ (gradient descent)
    - Discriminator (trained on trace data)
    - Find $\boldsymbol{\pi}^{\boldsymbol{\theta}}$ such that discriminator cannot distinguish between states from $\boldsymbol{\pi}^{\boldsymbol{\theta}}$ as opposed to those from $\boldsymbol{\pi}^{\mathbf{h}}$
  - **Our setting:** Finite state space, don't need discriminator!
  - Minimize $D_{KL}(d_{\boldsymbol{\pi\theta}}, d_{\boldsymbol{\pi}\mathbf{h}})$

1

Trainee Policy

$\boldsymbol{\pi}^{h}$

- Ho & Ermon 2016

Learning Expert Policy

Reward function unknown

- Imitation learning from experts

2

Expert Policy

$$\pi^*$$

Learning Expert Policy

2

Expert Policy

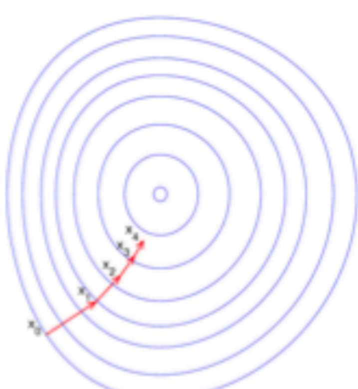
π^*

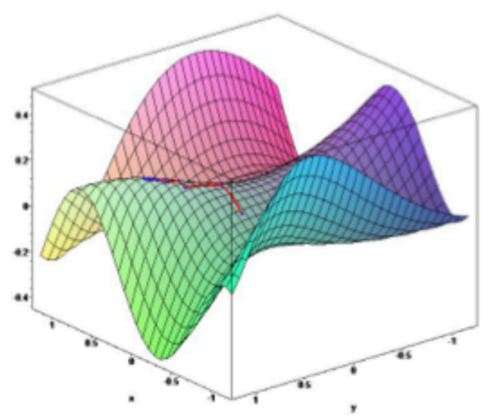
Reward function unknown

- Imitation learning from experts

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]**

$$
\begin{aligned}
&\textbf{if} \quad x \text{ obeys } a_1 \textbf{ then } y \sim \text{Binomial}(\theta_1),\ \theta_1 \sim \text{Beta}(\boldsymbol{\alpha} + \mathbf{N}_1)\\
&\textbf{else if } x \text{ obeys } a_2 \textbf{ then } y \sim \text{Binomial}(\theta_2),\ \theta_2 \sim \text{Beta}(\boldsymbol{\alpha} + \mathbf{N}_2)\\
&\vdots\\
&\textbf{else if } x \text{ obeys } a_m \textbf{ then } y \sim \text{Binomial}(\theta_m),\ \theta_m \sim \text{Beta}(\boldsymbol{\alpha} + \mathbf{N}_m)\\
&\textbf{else } y \sim \text{Binomial}(\theta_0),\ \theta_0 \sim \text{Beta}(\boldsymbol{\alpha} + \mathbf{N}_0).
\end{aligned}
$$

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

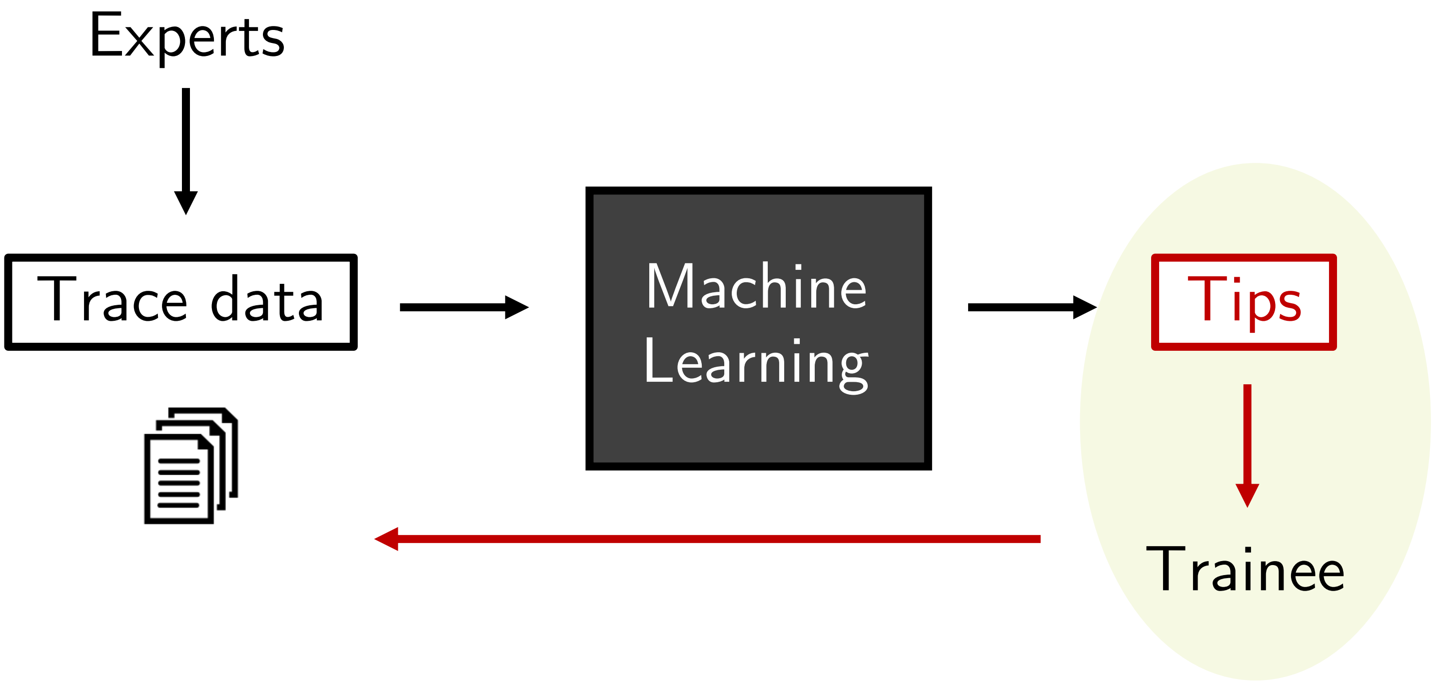
Learning Tips

3 Tip Mining

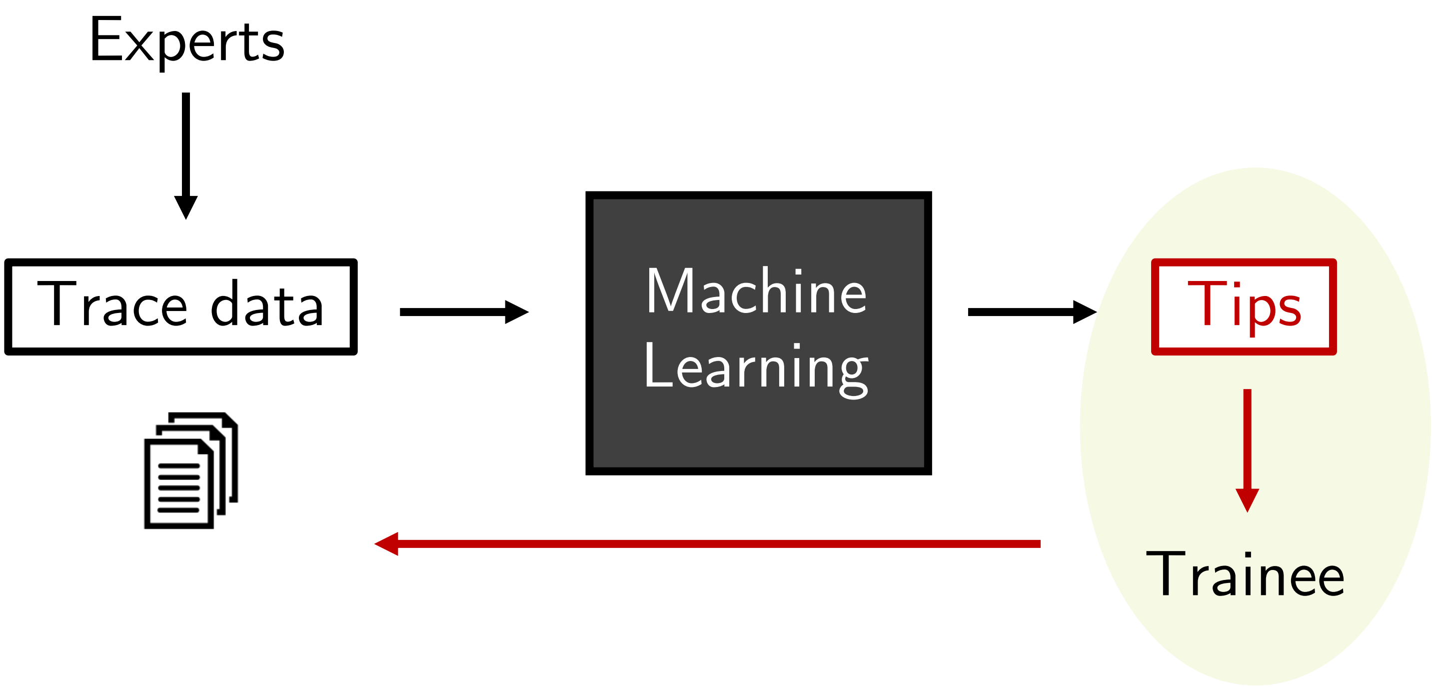
- Interpretable ML: Bayesian Rule Lists
  - Adapt from Letham et al 2015
  - If **[feature x],** then **[label y]**
  - e.g., if high blood pressure, then stroke
  - Pre-mined frequent patterns into a decision list using Bayesian statistics

- Our context
  - If **[state s]**, then take **[action a]**

Our Paper



Our Paper



Queueing Game

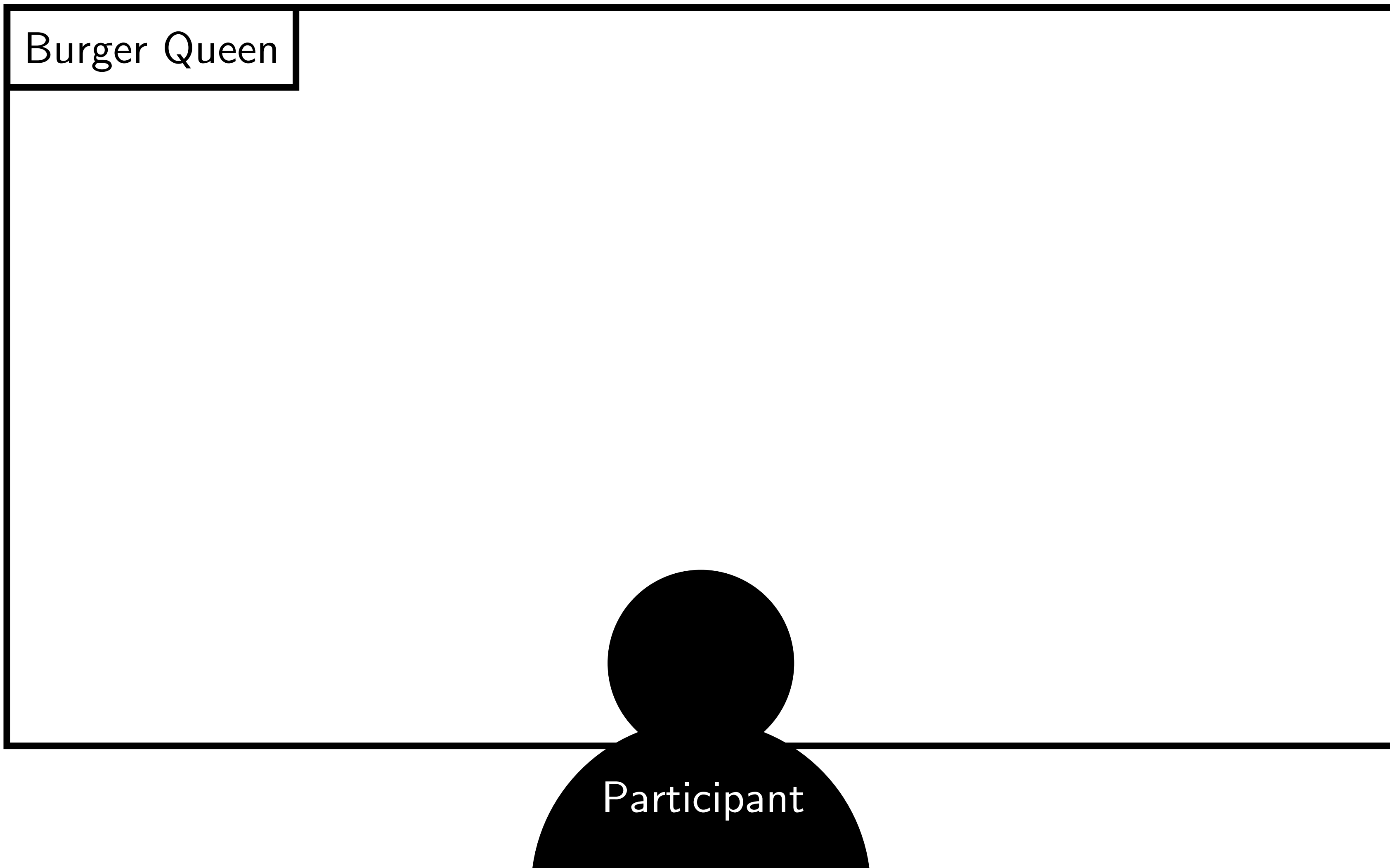
Amazon Mechanical Turk

$N = 146$

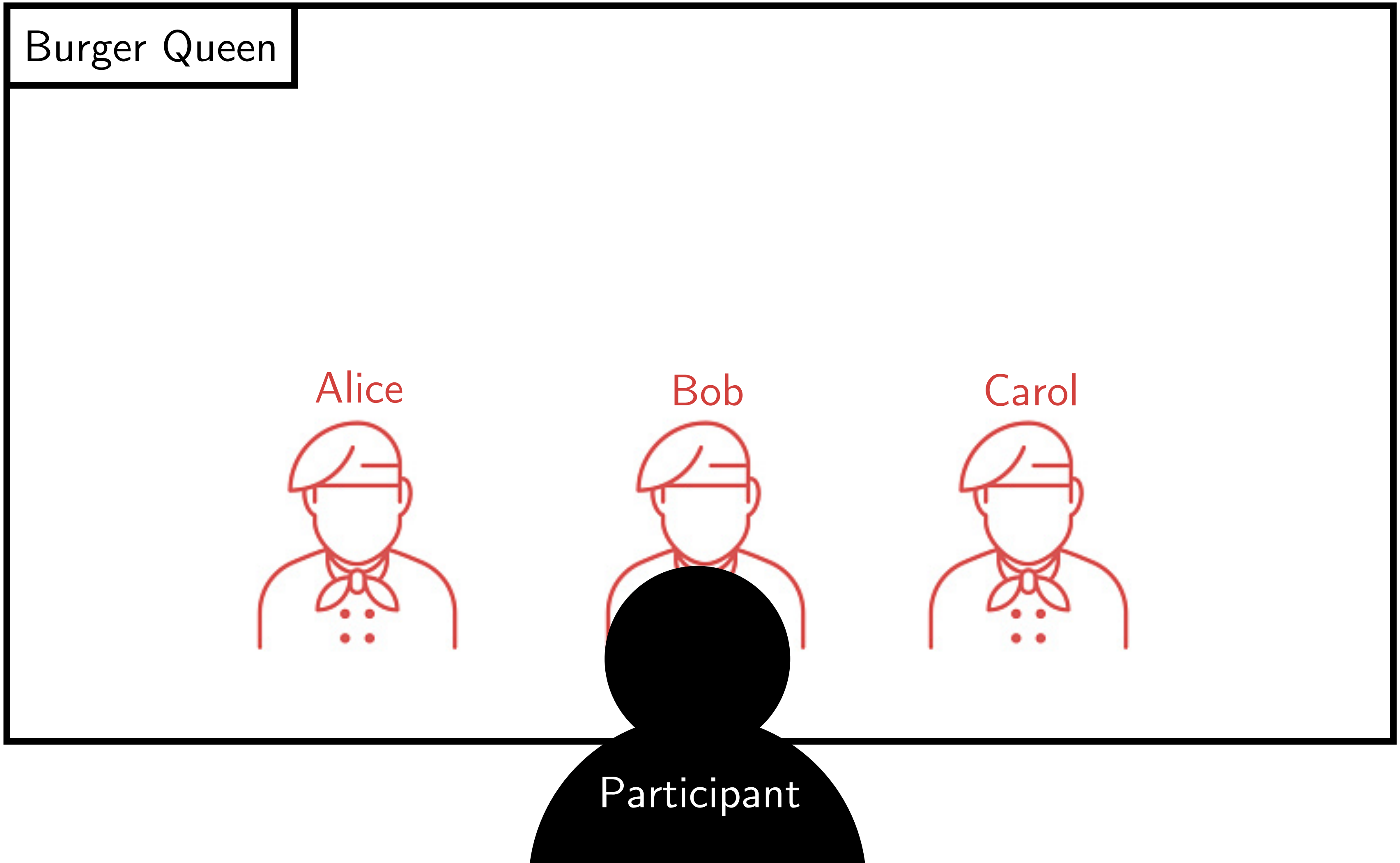
Wharton Behavioral Lab

$N = 203$

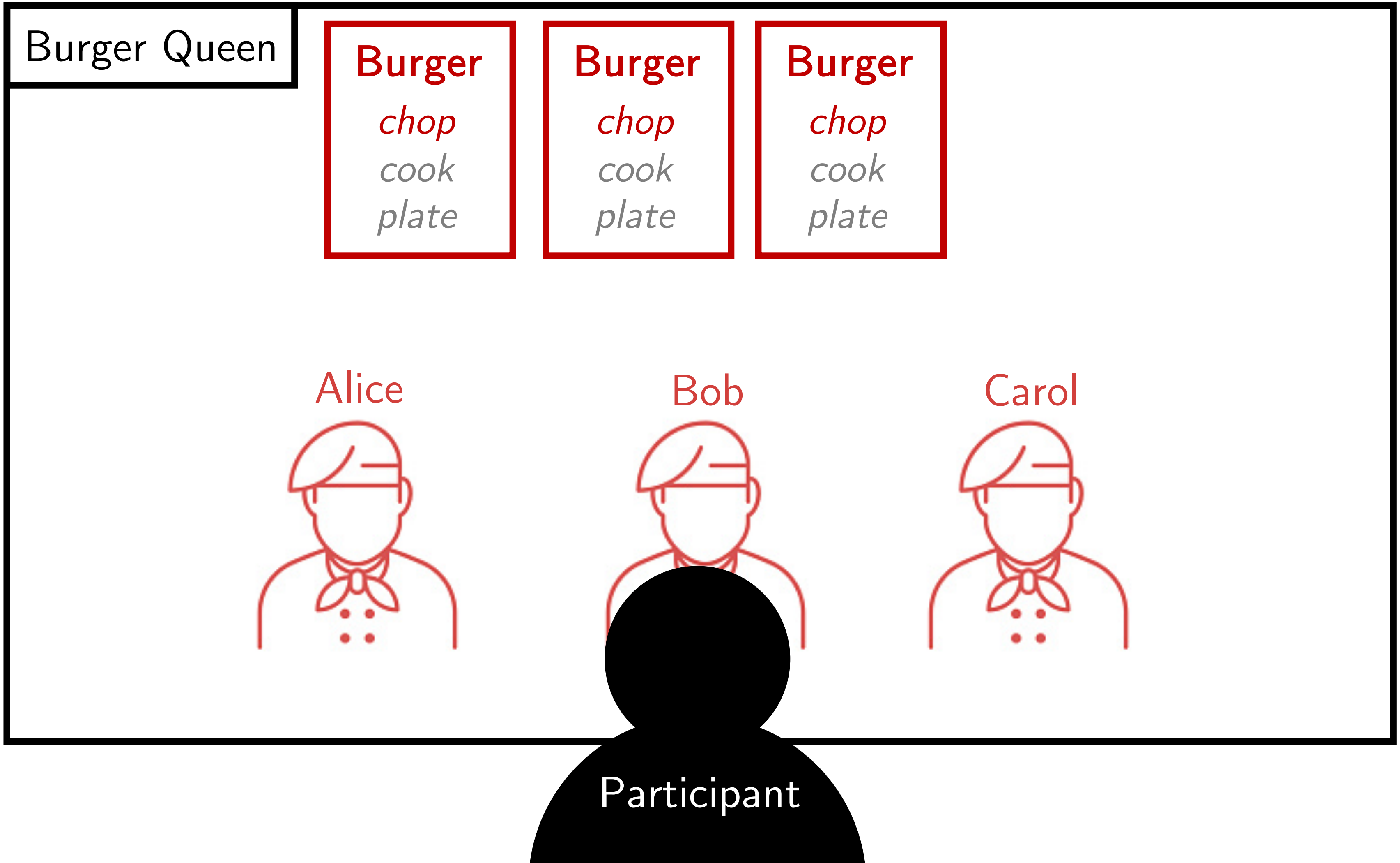
Queueing Game



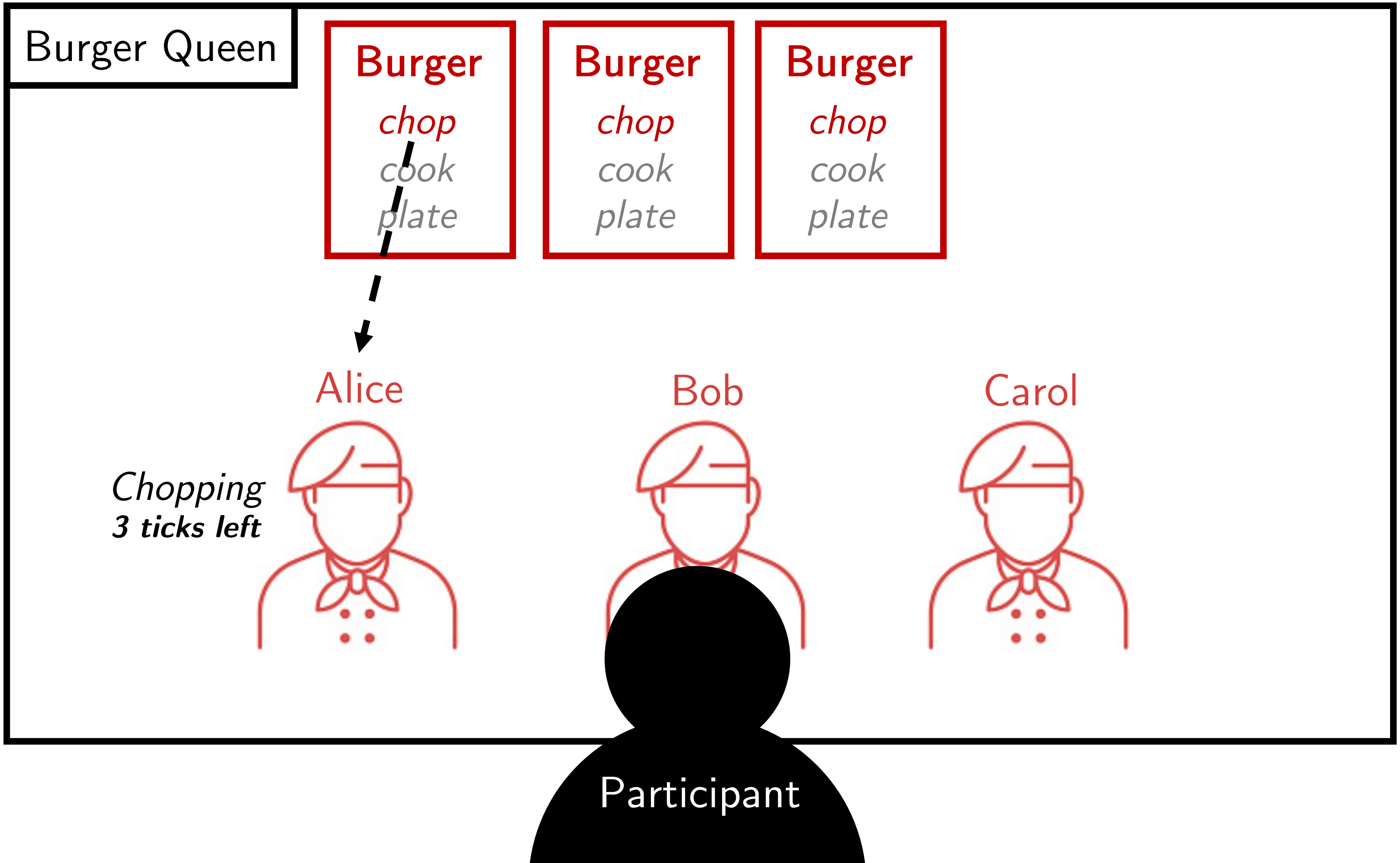
Queueing Game



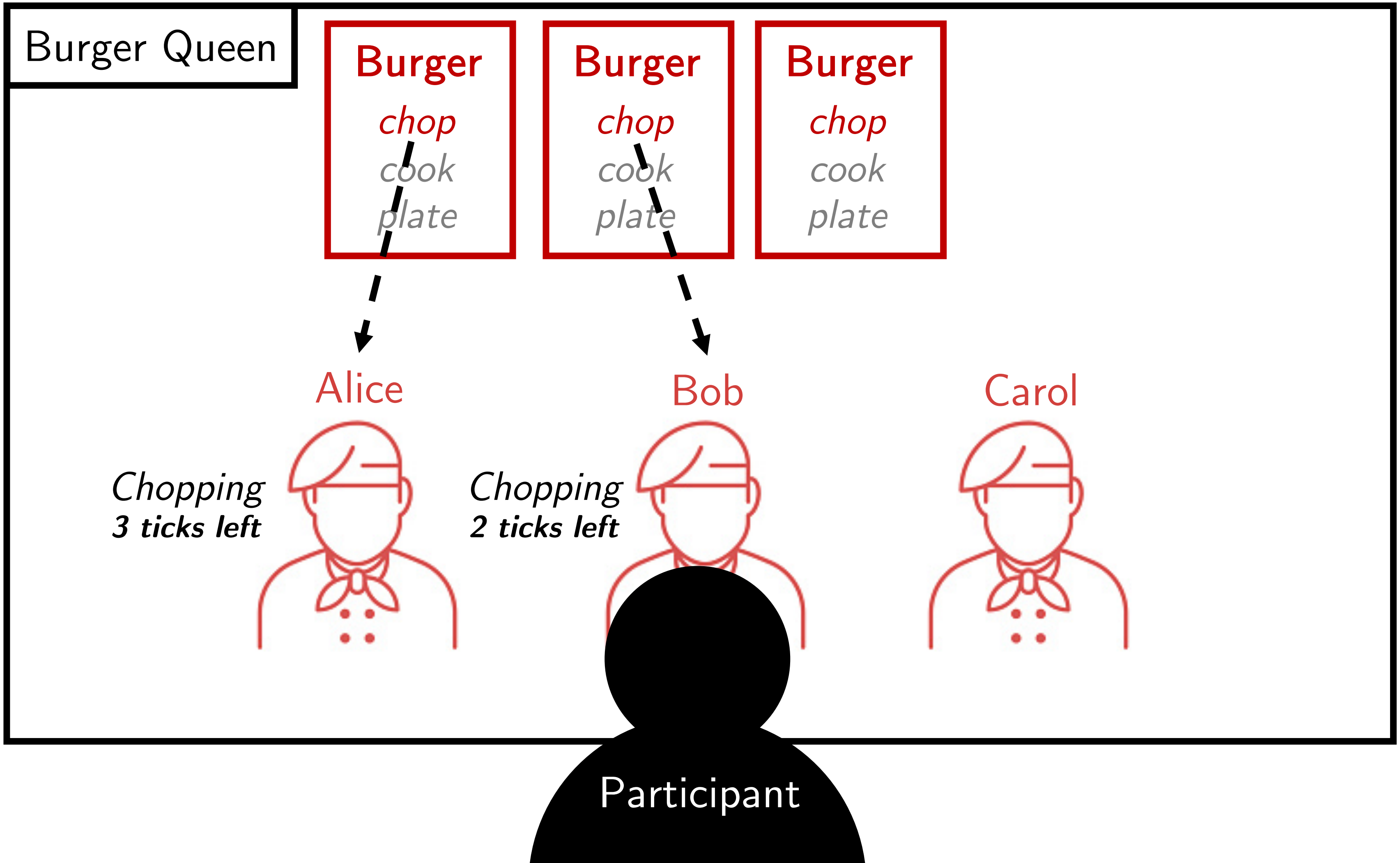
Queueing Game



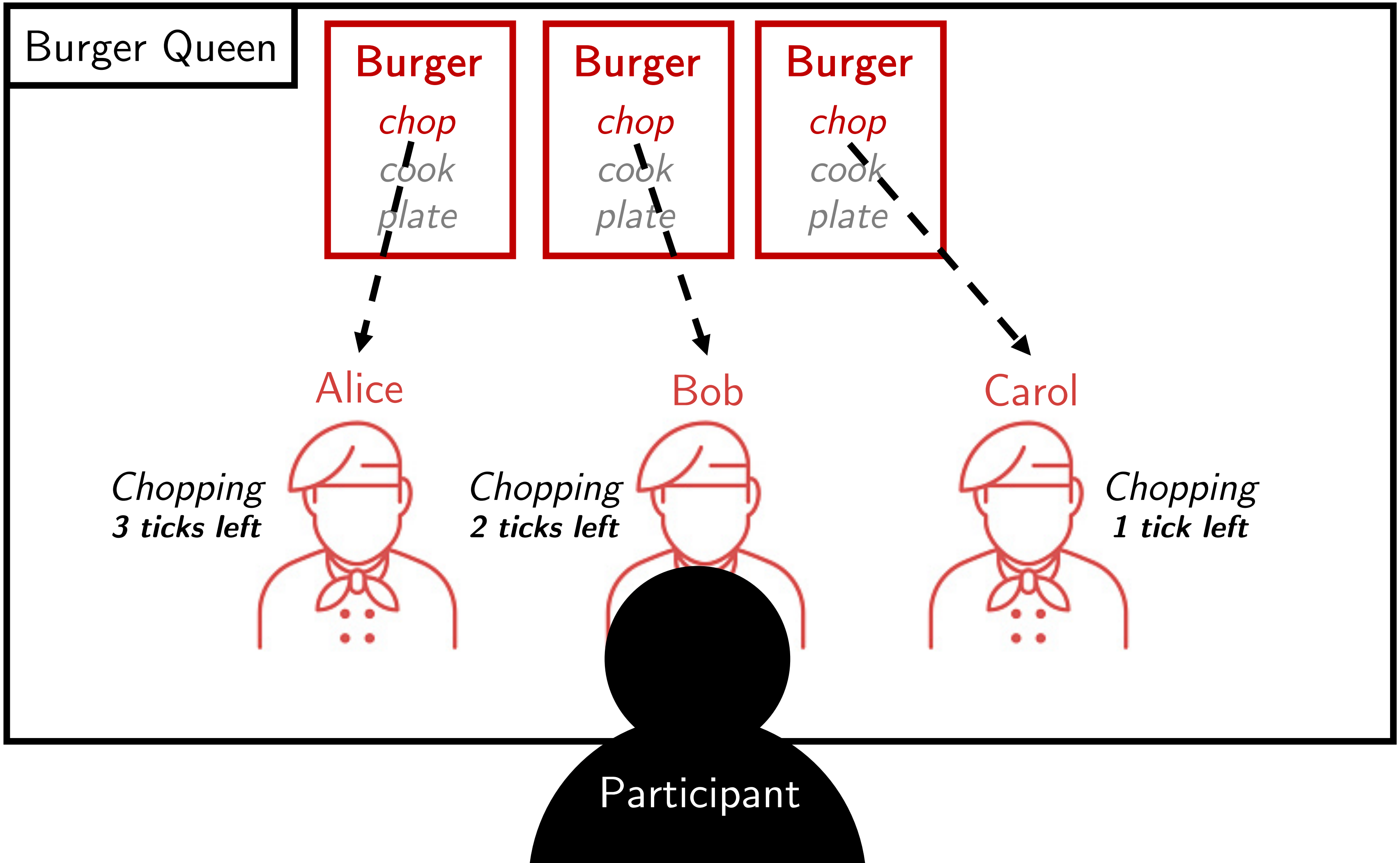
Queueing Game



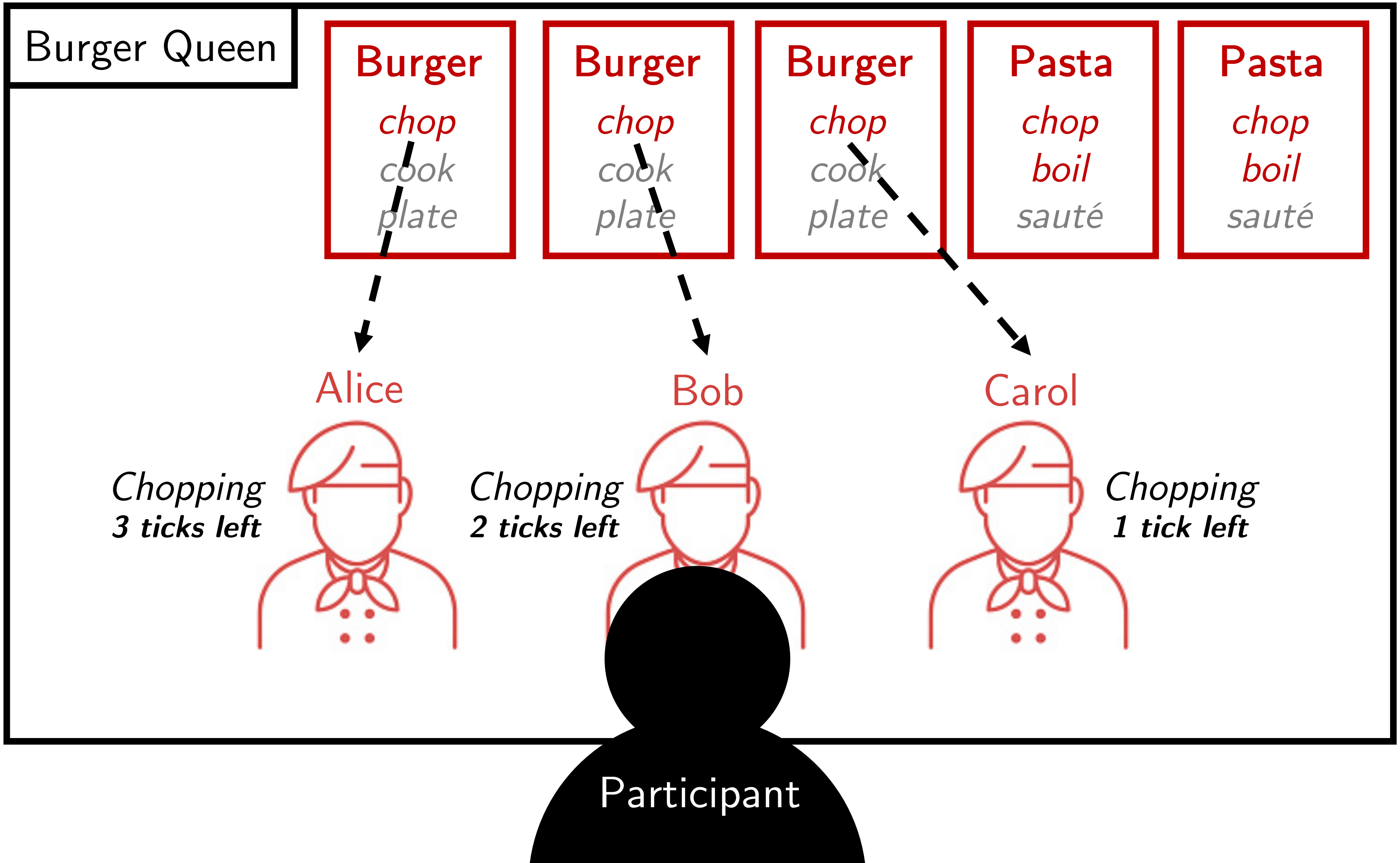
Queueing Game

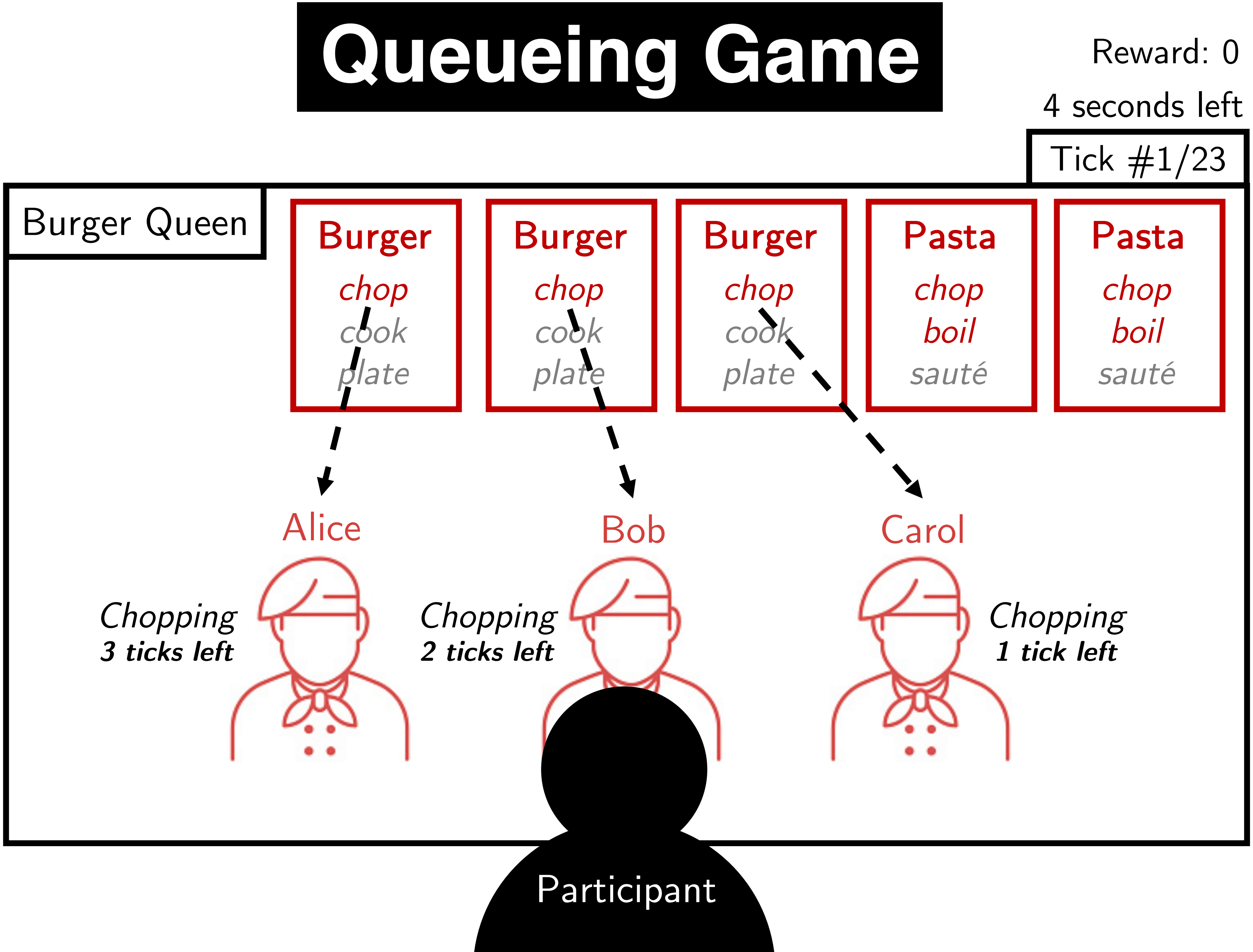


Queueing Game



Queueing Game



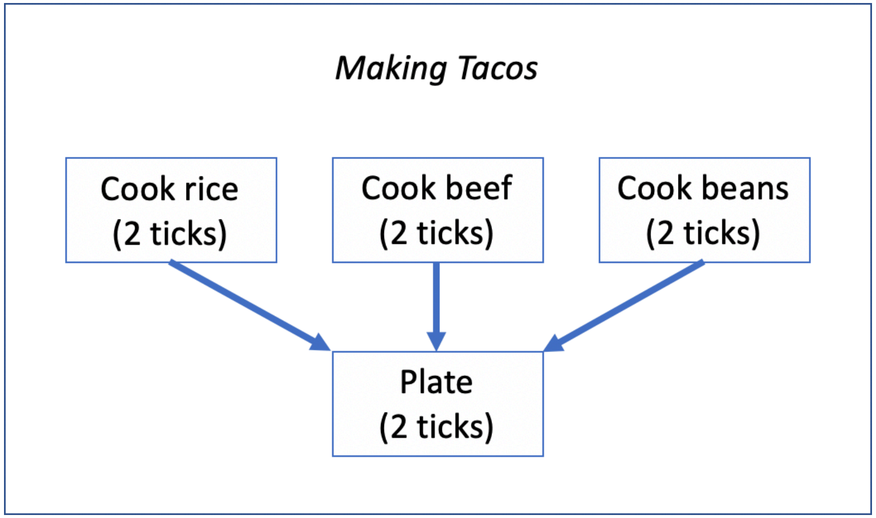

Queueing Game
Reward: 0
4 seconds left
Tick #1/23
Burger Queen
Burger
chop
cook
plate
Burger
chop
cook
plate
Burger
chop
cook
plate
Pasta
chop
boil
sauté
Pasta
chop
boil
sauté
Alice
Bob
Carol
Chopping
3 ticks left
Chopping
2 ticks left
Chopping
1 tick left
Participant

Experiment Learning Workers' Skills

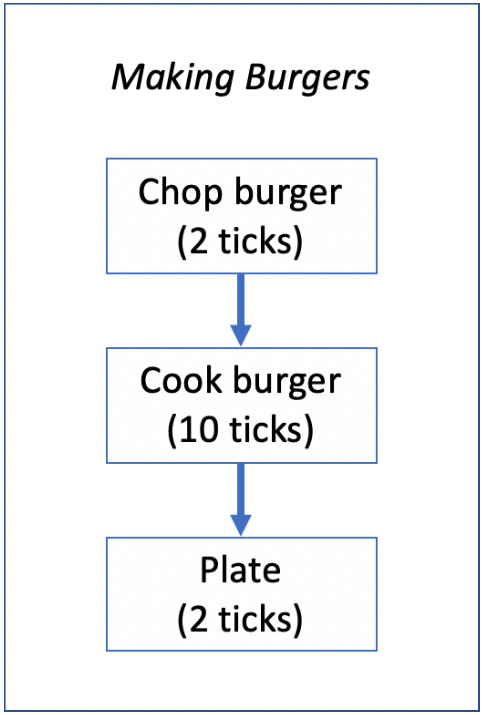
	Alice	Bob	Carol
Role:	Server	Sous-Chef	Chef
Chopping:	Slow	Fast	Fast
Cooking:	Slow	Medium	Fast
Plating:	Fast	Medium	Slow

Unknown to participants

Experiment Learning Tasks



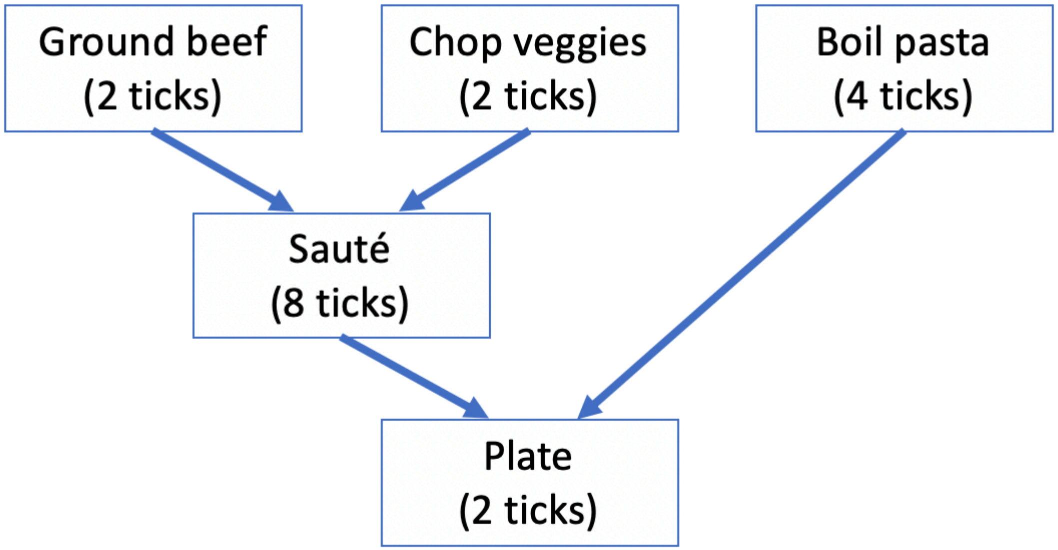
Parallel Tasks



Sequential Tasks

Experiment Learning Tasks

Making Pasta



Hybrid Tasks (Both Parallel and Sequential)

Experiment Handcrafted Tips

Control

No tips.

Tips

Experiment Handcrafted Tips

Control

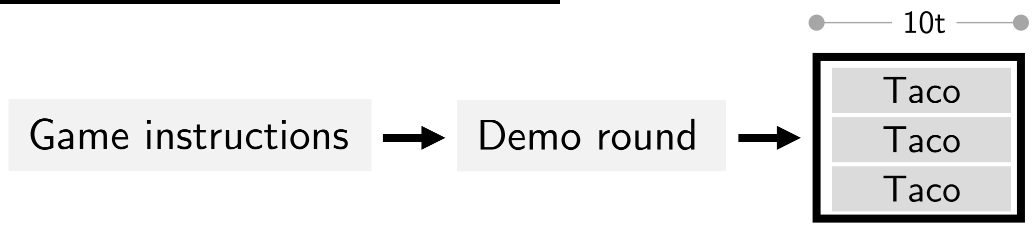
No tips.

Tips

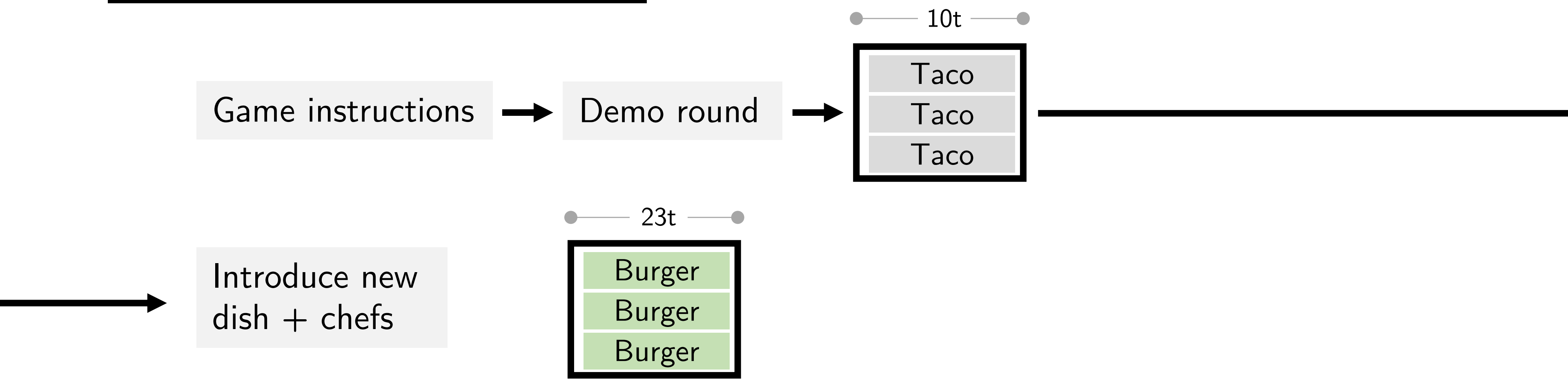
“Prioritize cooking and chopping tasks for Bob and Carol, and plating tasks for Alice.”

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

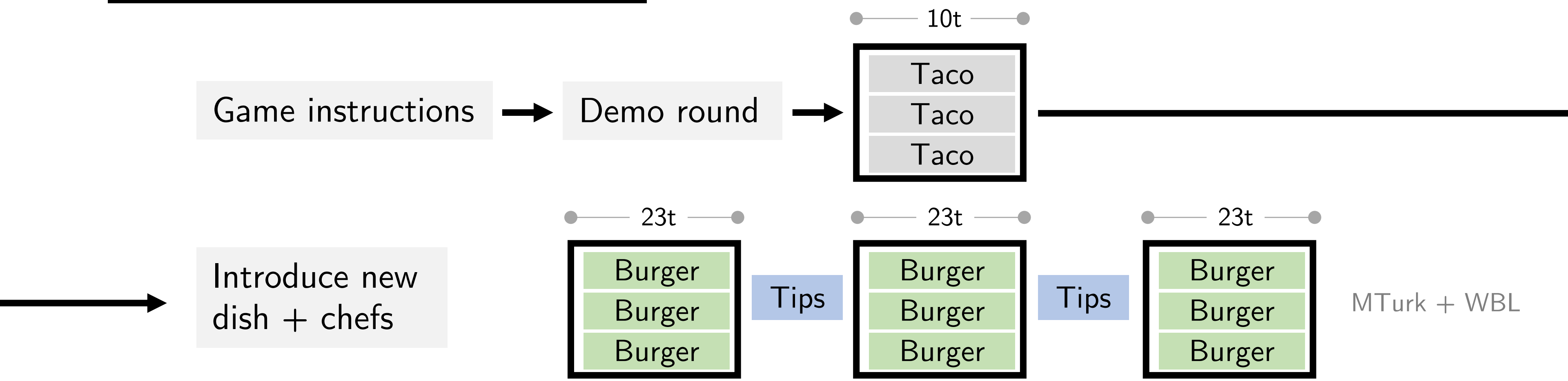
Experiment Flow



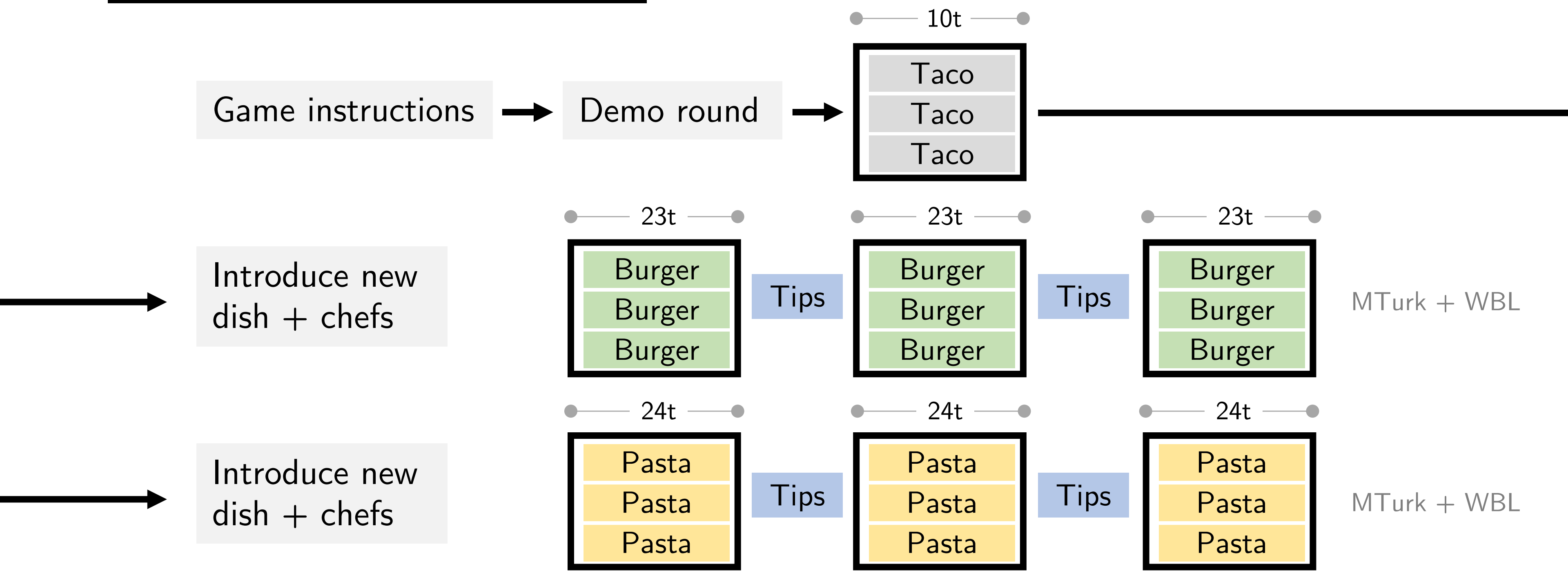
Experiment Flow



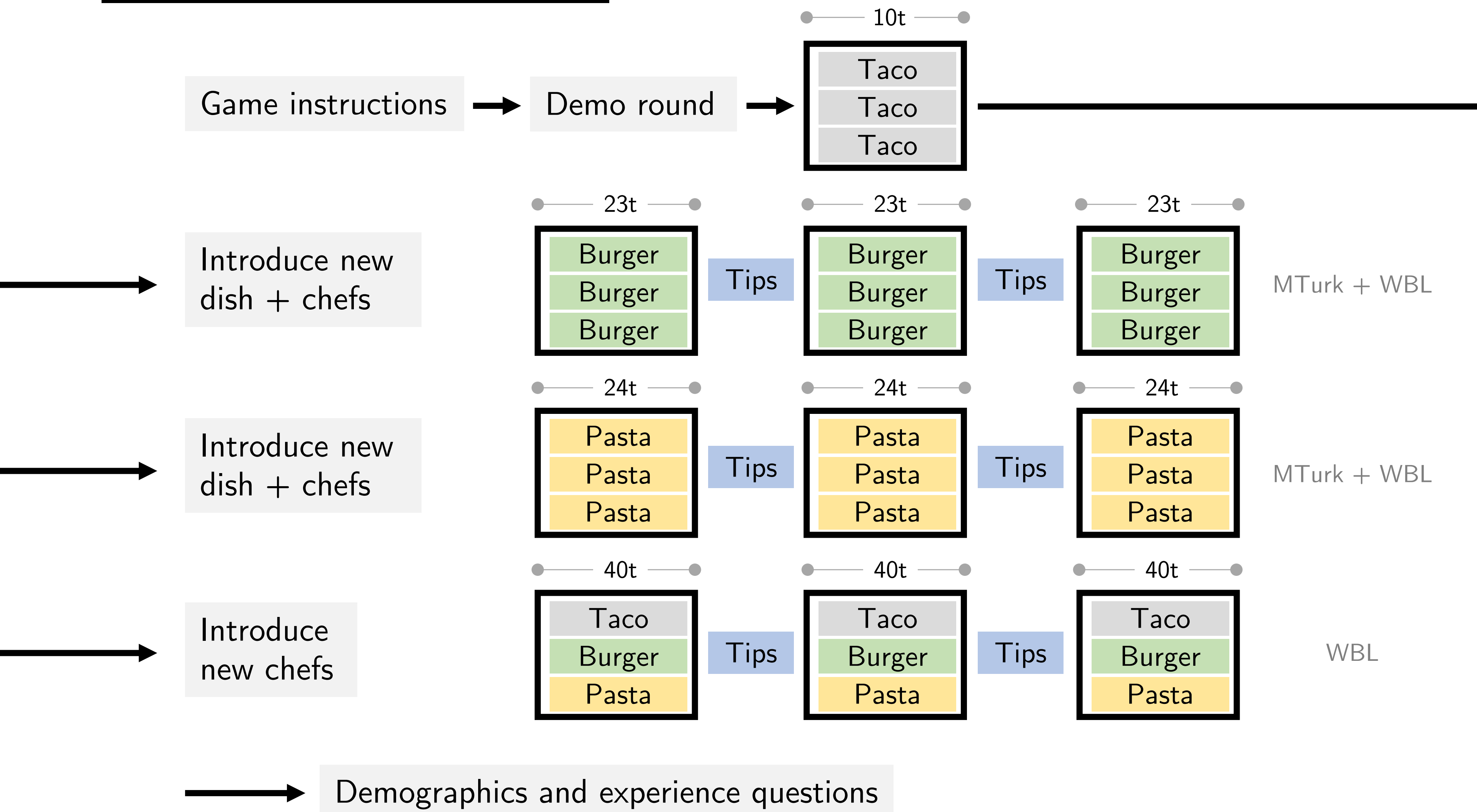
Experiment Flow



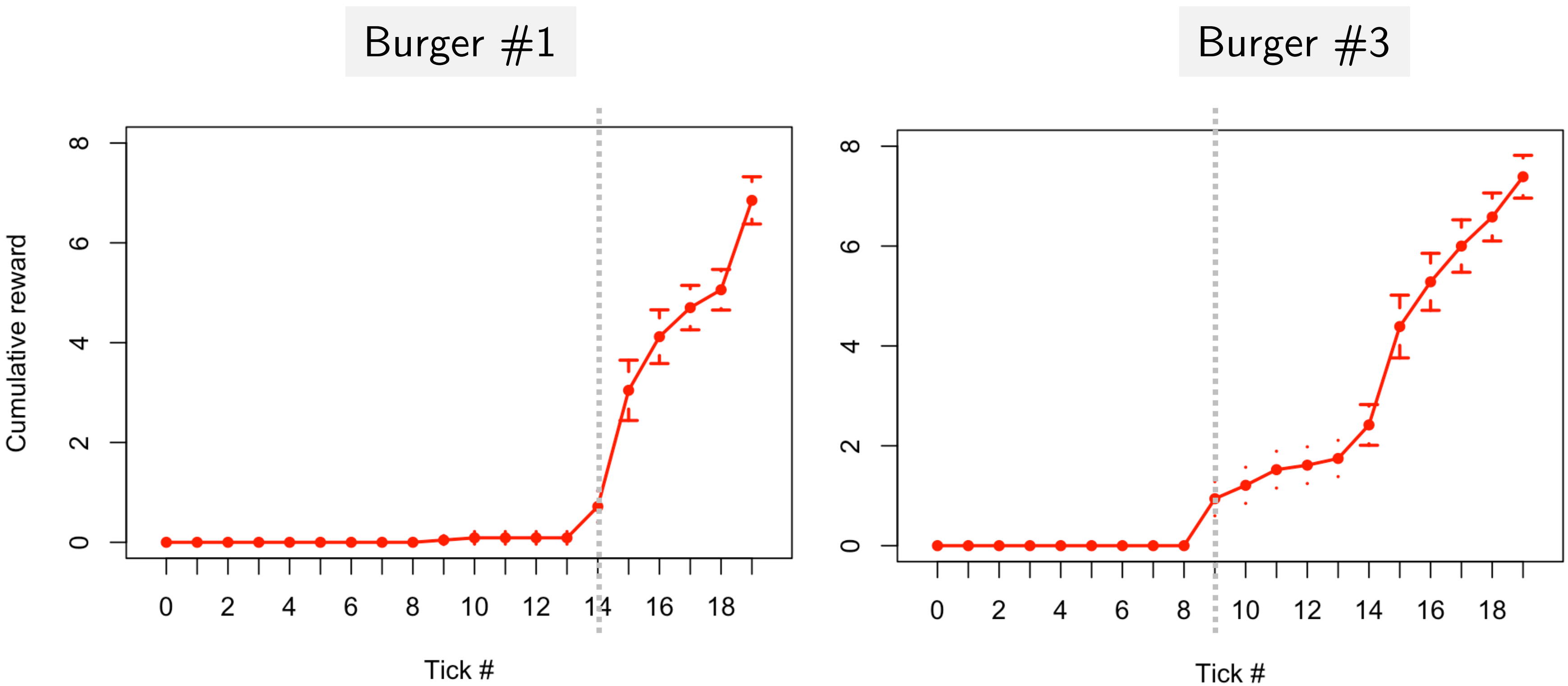
Experiment Flow



Experiment Flow



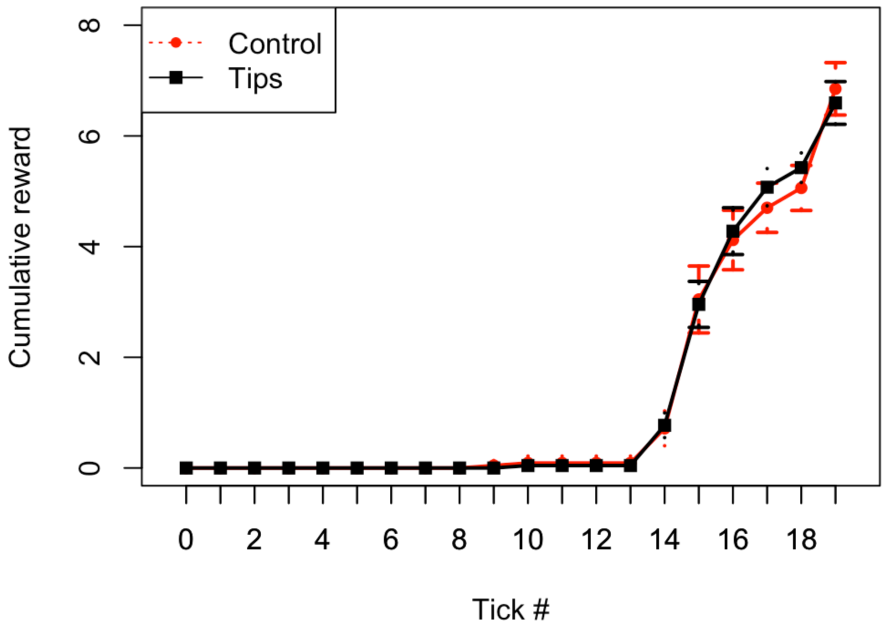
Results Performance



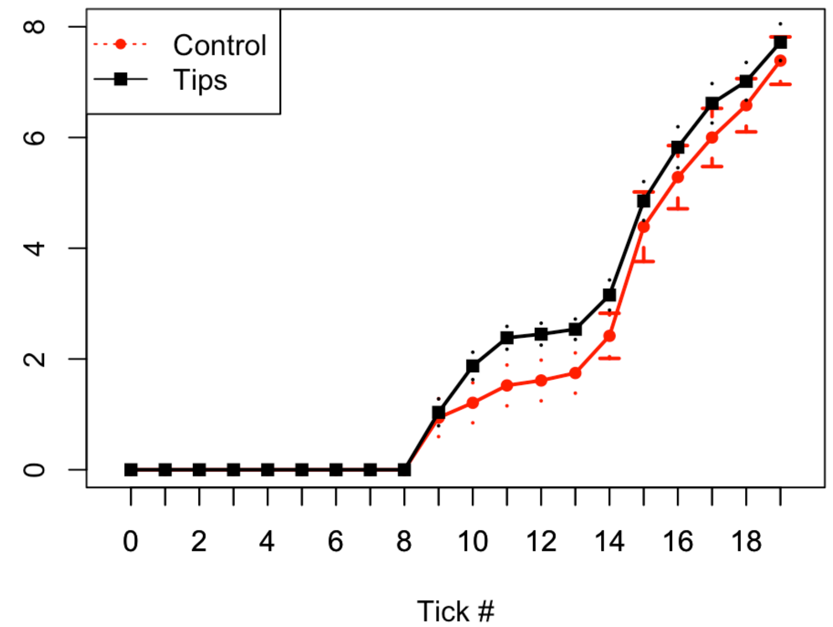
Control: Over time, people improve.

Results Tips Help

Burger #1



Burger #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	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 ✓

Obtain trace data to estimate policies ✓

Automatically extract tips ✓

Results Extracting Policies

All Participants
Trace Data

1

Trainee Policy

π^h

Imitation Learning

High Performing-
Participants
Trace Data

2

Expert Policy

π^*

Imitation Learning

Policy Gradient ← MDP

Results Extracting Policies

All Participants
Trace Data

1

Trainee Policy

$$\pi^{h}$$

High Performing-
Participants
Trace Data

2

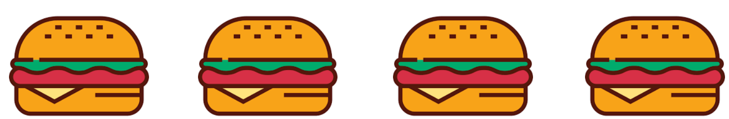
Expert Policy

$$\pi^{*}$$

26 ticks — Imitation Learning

Imitation Learning — 25 ticks

Policy Gradient — 19 ticks



Results Extracting Tips

3 **Tip Mining**

- Balance best improvement and confidence

Available Workers	Open Tasks	Action	Best Perf/Freq

Results Extracting Tips

3 Tip Mining

- Balance best improvement and confidence

	Available Workers	Open Tasks		Action	Best Perf/Freq
1.	Everyone	Cooking	→	Leave server idle	**19** / 19%

Results Extracting Tips

3 Tip Mining

- Balance best improvement and confidence

	Available Workers	Open Tasks		Action	Best Perf/Freq
1.	Everyone	Cooking	→	Leave server idle	**19** / 19%
2.	Chef Server	Plating	→	Leave chef idle	**19** / 4%

Results Extracting Tips

3 Tip Mining

- Balance best improvement and confidence

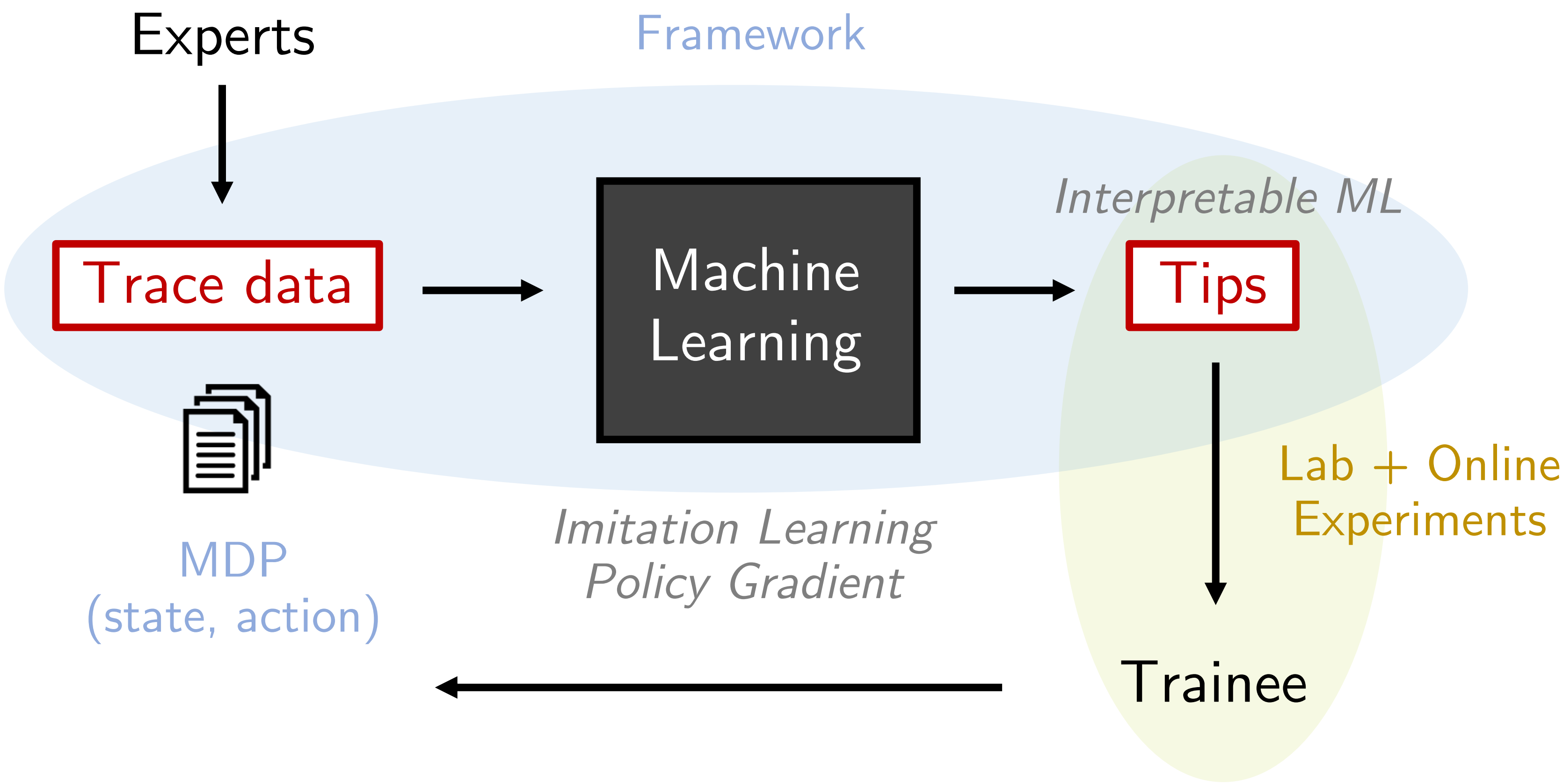
	Available Workers	Open Tasks	Action	Best Perf/Freq
1.	Everyone	Cooking	Leave server idle	**19** / 19%
2.	Chef, Server	Plating	Leave chef idle	**19** / 4%
3.	Sous Chef	Cooking	Assign to sous chef	**22** / 23%

On-Going Work

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

Summary

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



Hamsa Bastani, Osbert Bastani, **Park Sinchaisri** (swich@wharton.upenn.edu | @wichinpong)

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 $\pi^* : S \to A$ 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 ny 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.*,

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

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.