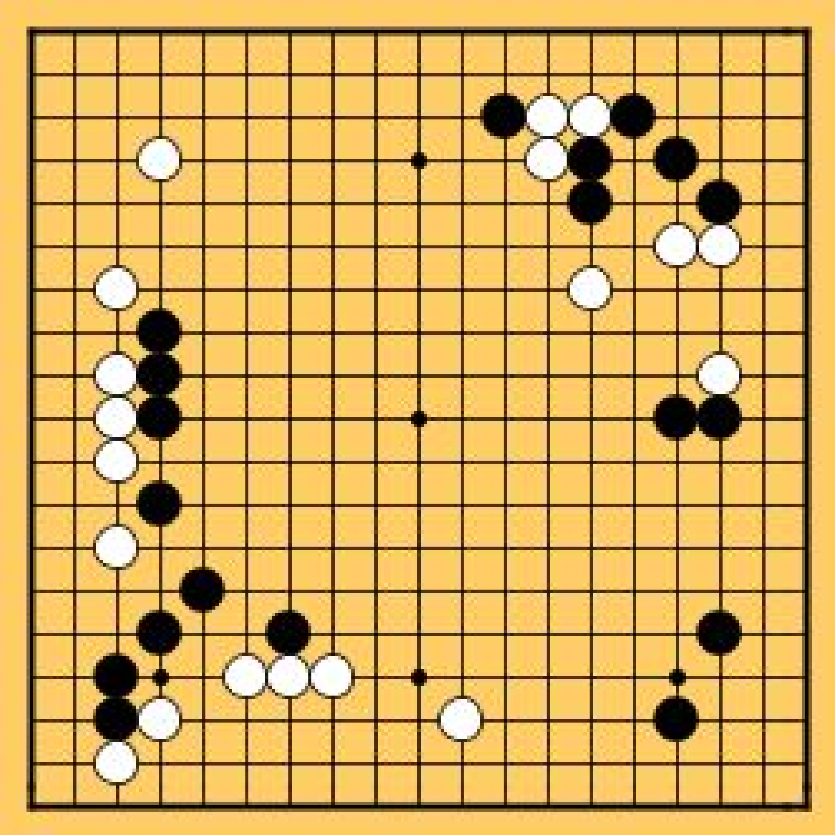
Human-in-the-loop RL

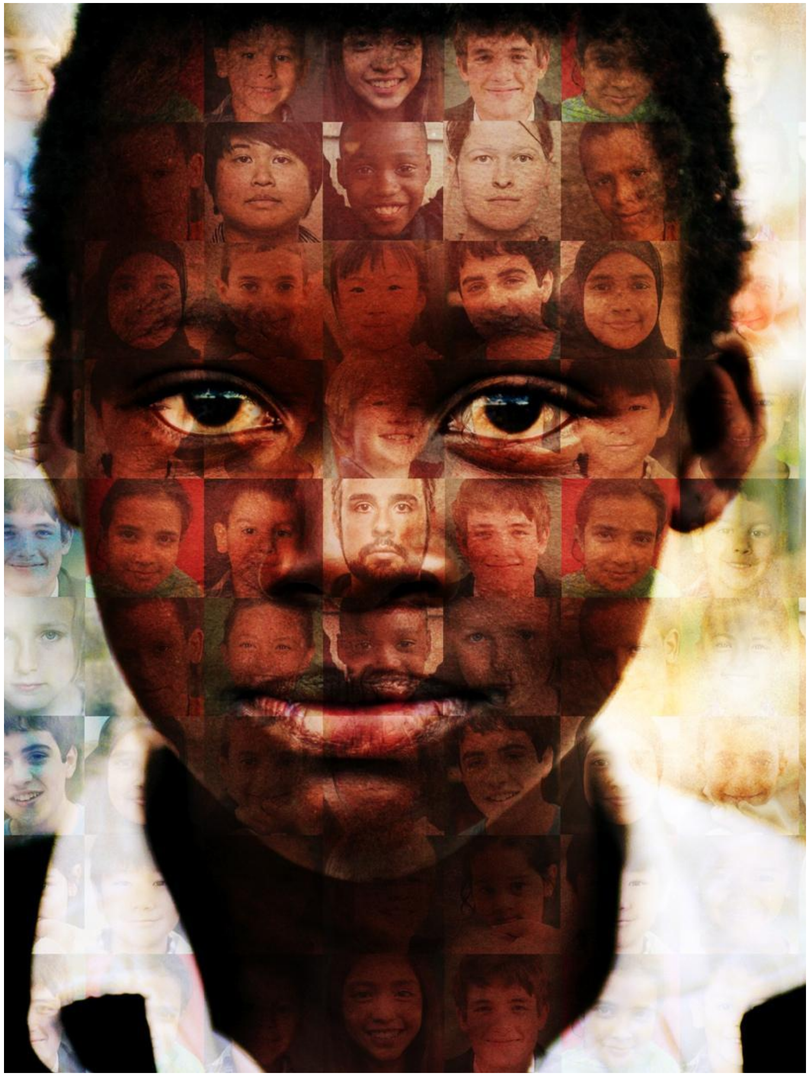
Emma Brunskill

CS234 Spring 2017

From here ....
healthcare…



to education,



Courseware Course Info Discussion Wiki Progress Instructor **Staff view**

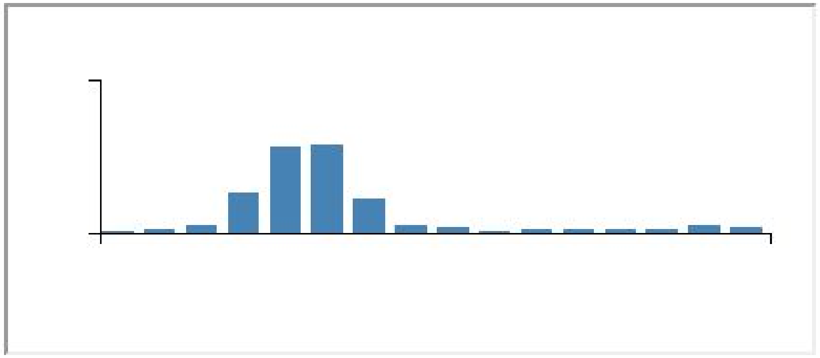
Study

- pre-assessment
- B1
- B3: Histogram Heights
- B3: Histogram Heights 2
- B3: Data Underlying
- P3: Extracting Proportions
- B4
- B4.2
- B5
- Skew
- Skew2
- Shape
- Labeling Worked Example
- Practice Labeling
- Practice Labeling Water
- Practice Labeling No Histogram: Voters
- Practice Why Wrong

VIEW UNIT IN STUDIO

DESCRIPTIONS AND HISTOGRAMS (1/3 points)

The price of airline tickets varies over time. The following is a histogram that could describe the distribution of airplane ticket prices. Select the best option for each of the questions below.



The x-axis should be labeled as

- ○ Time
- ○ Ticket Price
- ● Frequency ✖
- ○ Distribution

w/Karan Goel, Rika Antonova, Joe Runde, Christoph Dann, & Dexter Lee

Setting

- Set of N skills
  - Understand what x-axis represents
  - Estimate the mean value from a histogram
  - ...
- Assume student can learn each skill independently
- Policy is a mapping from the history of prior skill practices & their outcomes to whether or not to give the student another practice problem
  - E.g. (incorrect, incorrect, incorrect) → give another practice
  - (correct,correct) → no more practice
- Use a parameterized policy to characterize the teaching policy for each skill
- Reward is a function of the student's performance on a post test after the policy for each skill says "no more practice" and how much practice gave

Initial Work: Bayesian Optimization Policy Search

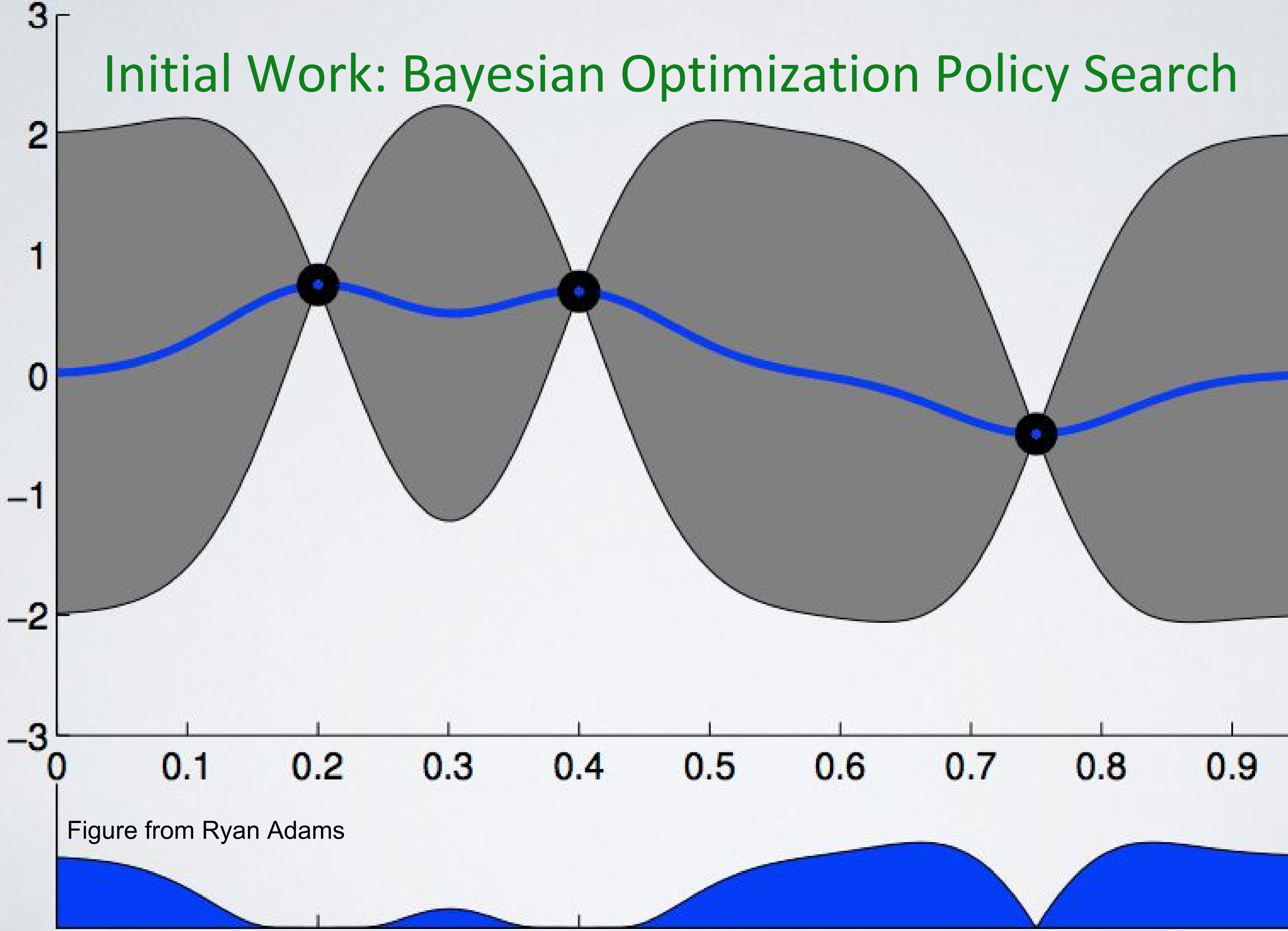
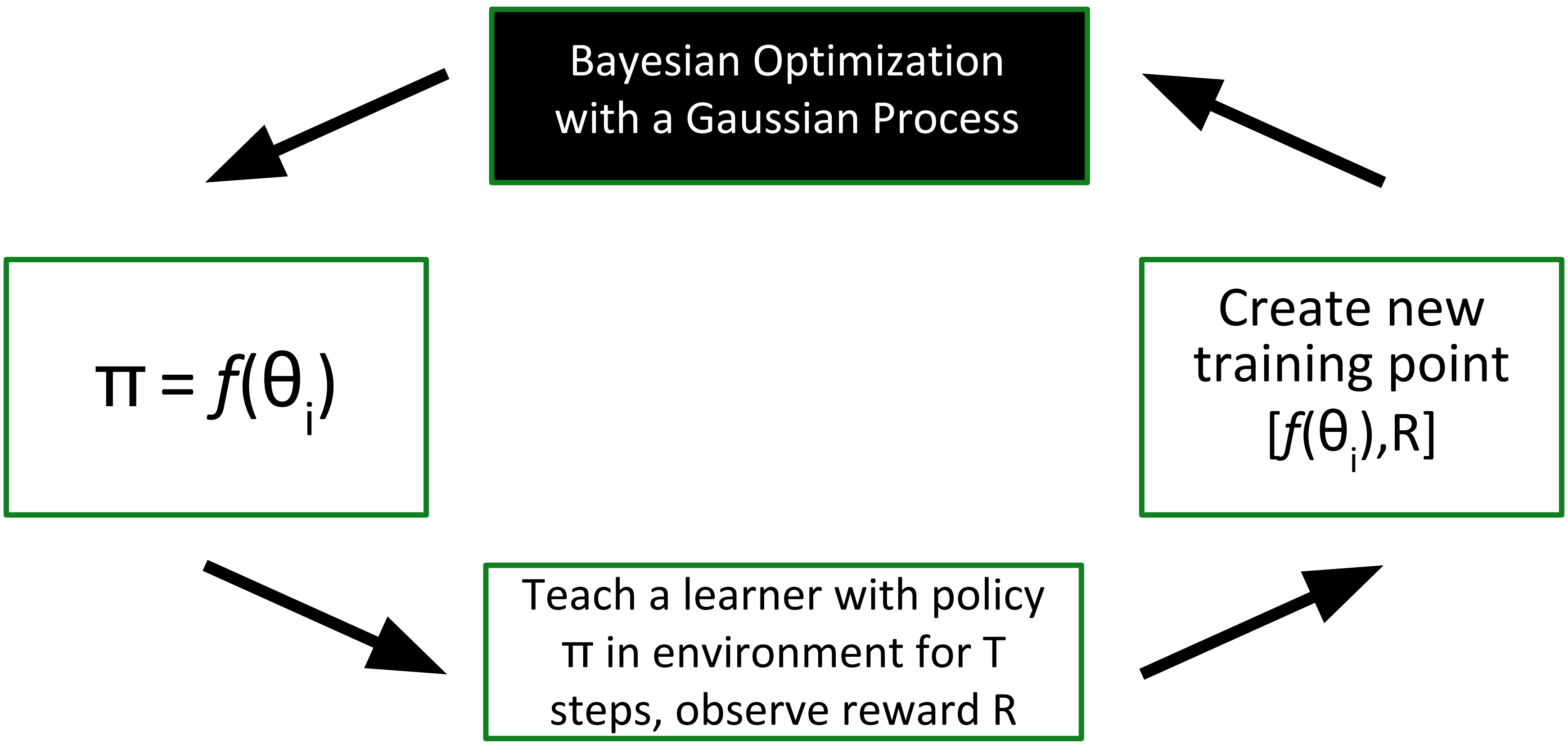


Figure from Ryan Adams

Learning to Teach

Goal: Should Learn Policy That Maximizes Expected Student Outcomes

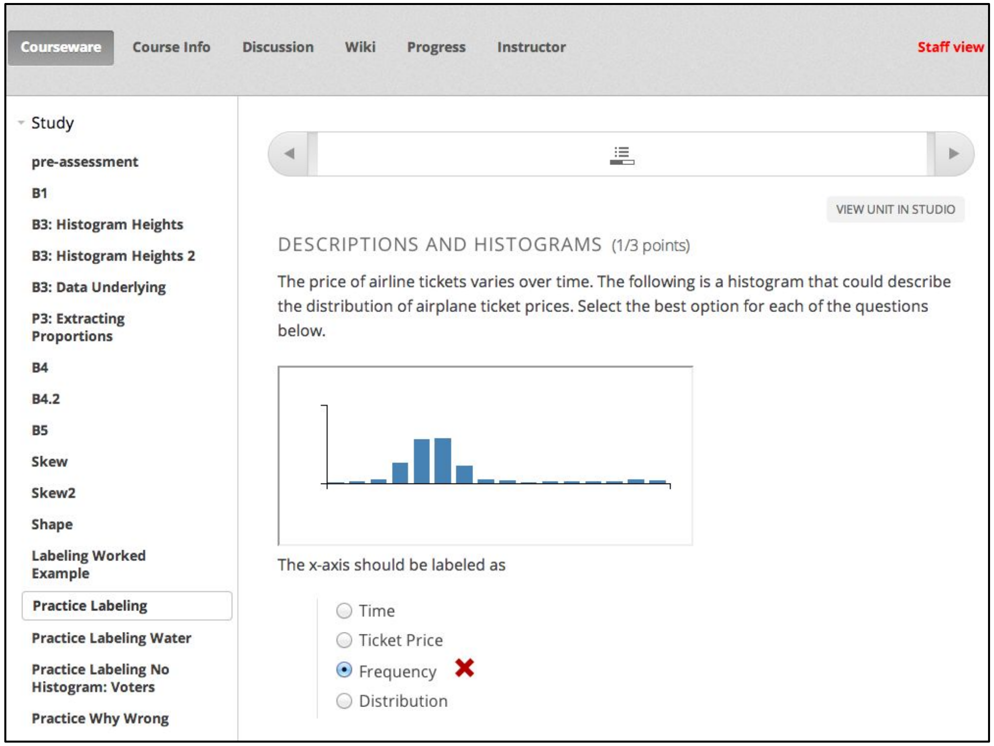


Reward Signal?

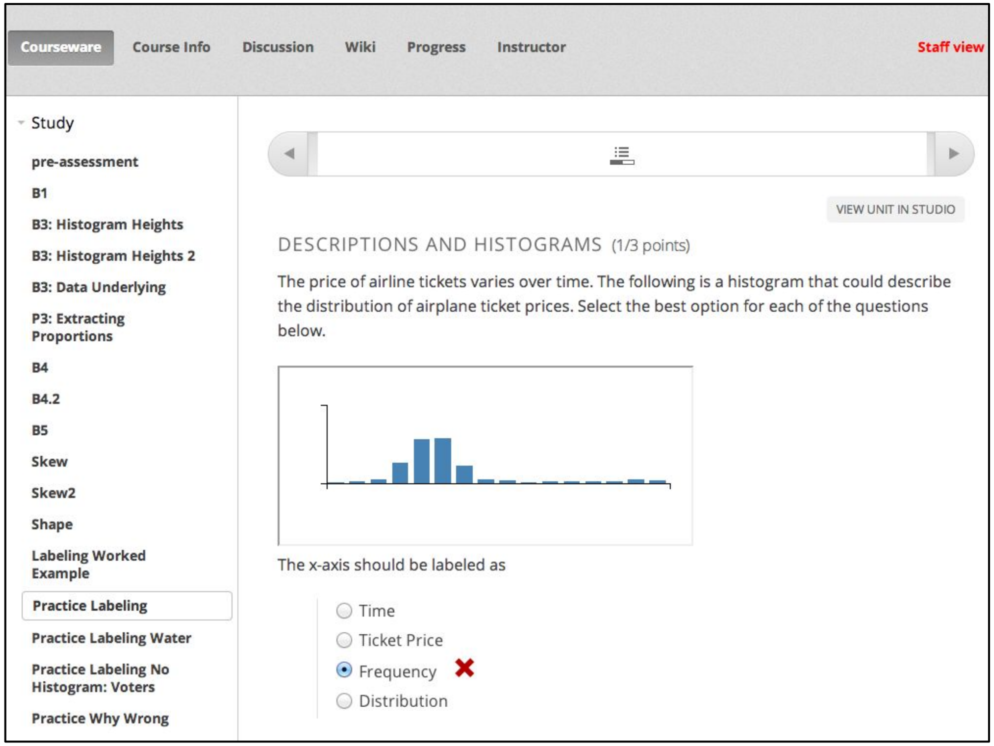
- Balance post test performance with amount of practice needed
- p_s=Performance on skill s,
- p = Post test performance across all skills,
- l = # practices for skill s

$$f(\pi) = \frac{p_s + \mathbb{I}(p > 9)}{\sqrt{l}}$$

During Policy Search Tutoring System Stopped Teaching Some Histogram Skills

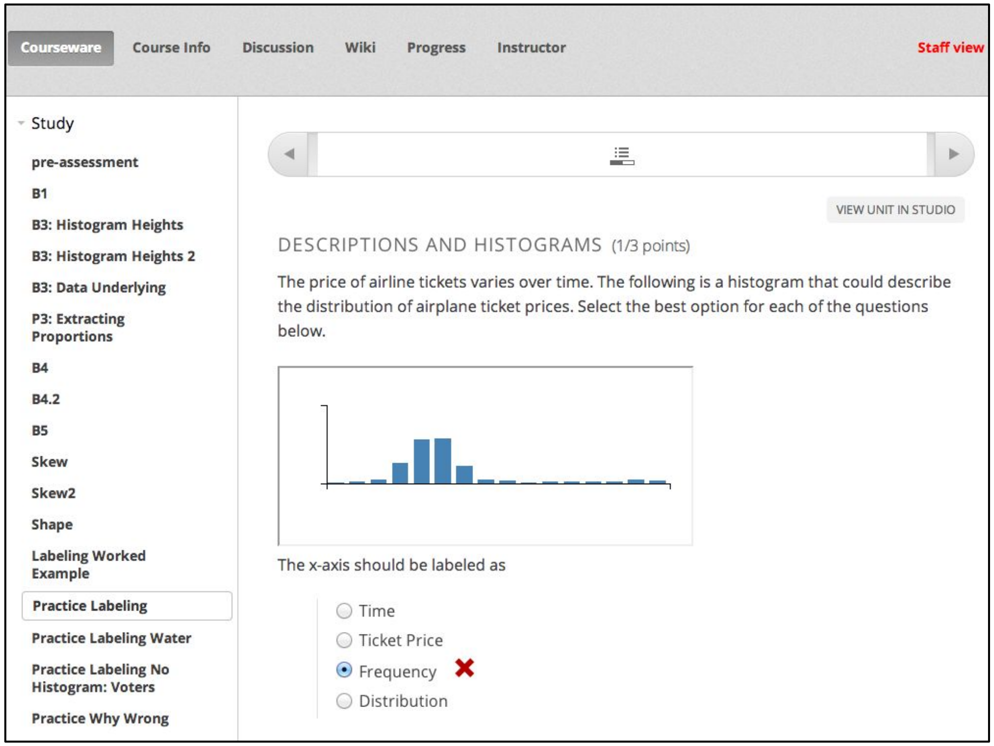


Reward Signal: Post Test / # Problems Given



$$f(\pi) = \frac{p_s + \mathbb{I}(p > 9)}{\sqrt{l}}$$

During Policy Search Tutoring System Stopped Teaching Some Histogram Skills



- No improvement in post test → system had learned that some of our content was inadequate so best thing was to skip it!
- **Content (action space) insufficient to achieve goals**

Humans are Invention Machines

New actions

New sensors

Invention Machines: Creating Systems that Can Evolve Beyond Their Original Capacity To Reach Extraordinary Performance

New actions

New sensors

Problem Formulation

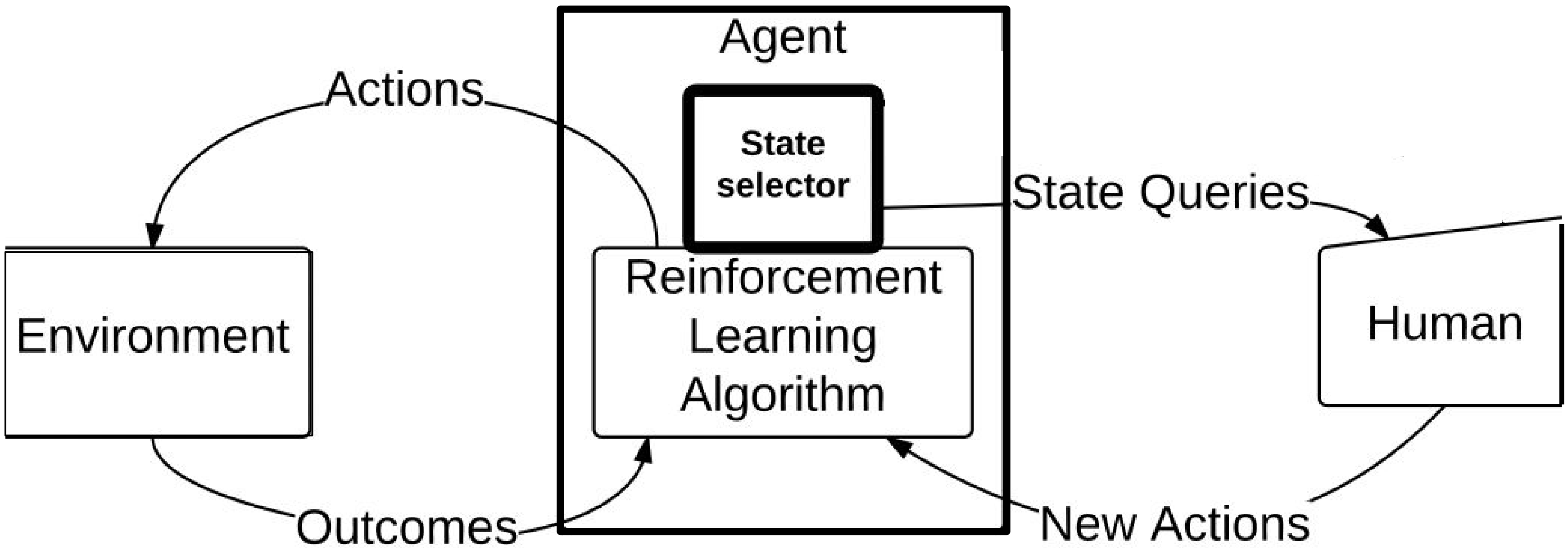
- Maximize expected reward
- Online reinforcement learning
- Directed action invention
  - Where (which states) should we add actions at?

Related Work

- Policy advice / learning from demonstration
- Changing action spaces
  - Almost all work is reactive, not active solicitation

Online reinforcement learning

Active Domain (Action Space) Adaptation



Mandel, Liu, Brunskil & Popovic, AAAI 2017

Requesting New Actions

$$\arg\max_{s} \sum_{s_0 \in S_0} V_{\mathcal{A} \cup a_h}(s_0)\, p(s_0)$$

Current action set

New action

Mandel, Liu, Brunskil & Popovic, AAAI 2017

Expected Local Improvement

$$\arg\max_s \int_a \underbrace{p_s(a_h)}_{\text{Prob. human gives you action } a_h \text{ for state s}} \underbrace{(V_{\mathcal{A}\cup a_h}(s) - V_{\mathcal{A}}(s))}_{\text{Improvement in value at state s if add in action } a_h} da_h$$

Mandel, Liu, Brunskil & Popovic, AAAI 2017

$$
\begin{aligned}
ELI(s) &= \int_a p_s(a_h)(V_{\mathcal{A}\cup a_h}(s) - V_{\mathcal{A}}(s))da_h \\
&\leq \int_{a:V_{\mathcal{A}\cup a_h}(s)>V_{\mathcal{A}}(s)} p_s(a_h)(V_{\mathcal{A}\cup a_h}(s) - V_{\mathcal{A}}(s))da_h \\
&\leq (V_{max} - \underbrace{V_{\mathcal{A}}(s))}_{\substack{\text{V(s) given}\\ \text{current action set}}} \underbrace{\int_{a:V_{\mathcal{A}\cup a_h}(s)>V_{\mathcal{A}}(s)} p_s(a_h)da_h}_{\substack{\text{Probability get a new action}\\ \text{that will increase V(s)}}}
\end{aligned}
$$

Unknown!

Mandel, Liu, Brunskil & Popovic, AAAI 2017

What to Use for $V_{\mathcal{A}}(s)$

$$(V_{max} - V_{\mathcal{A}}(s)) \int_{a:V_{\mathcal{A} \cup a_h}(s) > V_{\mathcal{A}}(s)} p_s(a_h) da_h$$

- Be optimistic (MBIE, Rmax, …)
- Why?
  - Don't need to add in new actions if current action set might yield optimal behavior
  - Avoids focusing on highly unlikely states

Mandel, Liu, Brunskil & Popovic, AAAI 2017

Probability of Getting a Better Action

$$(V_{max} - V_{\mathcal{A}}(s)) \boxed{\int_{a:V_{\mathcal{A}\cup a_h}(s) > V_{\mathcal{A}}(s)} p_s(a_h) da_h}$$

- Don't want to ask for actions at same state forever (maybe no improvement possible)
- Model prob of a better action as $Beta(1, |\mathcal{A}_{s,\ell}| + 1)$
- Chance of better action decays w/ # of actions

$$ELI(s) = \frac{1}{|\mathcal{A}_{s,\ell}| + 2} (V_{max} - V_{\mathcal{A}}(s))$$

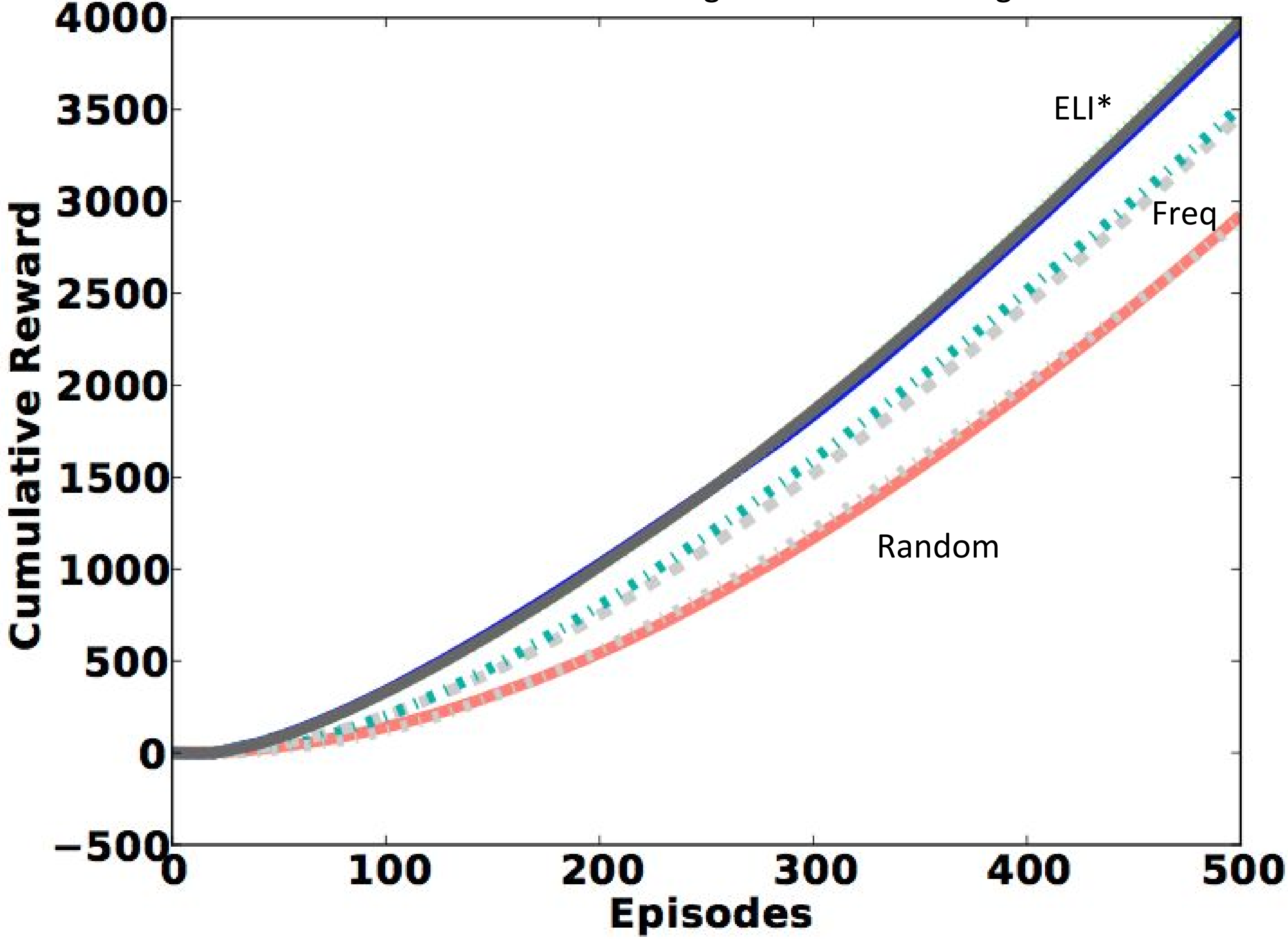
Mandel, Liu, Brunskil & Popovic, AAAI 2017

Simulations

- Large action task* (Sallans & Hinton 2004)
  - 13 states
  - 273 outcomes (next possible states per state)
  - $\mathbf{2^{20}}$ actions per state
- At start each *s* has single *a* (like default π)
- Every 20 steps can request an action
  - Sample action at random from action set for s
  - Compare ELI vs Random s vs High freq s

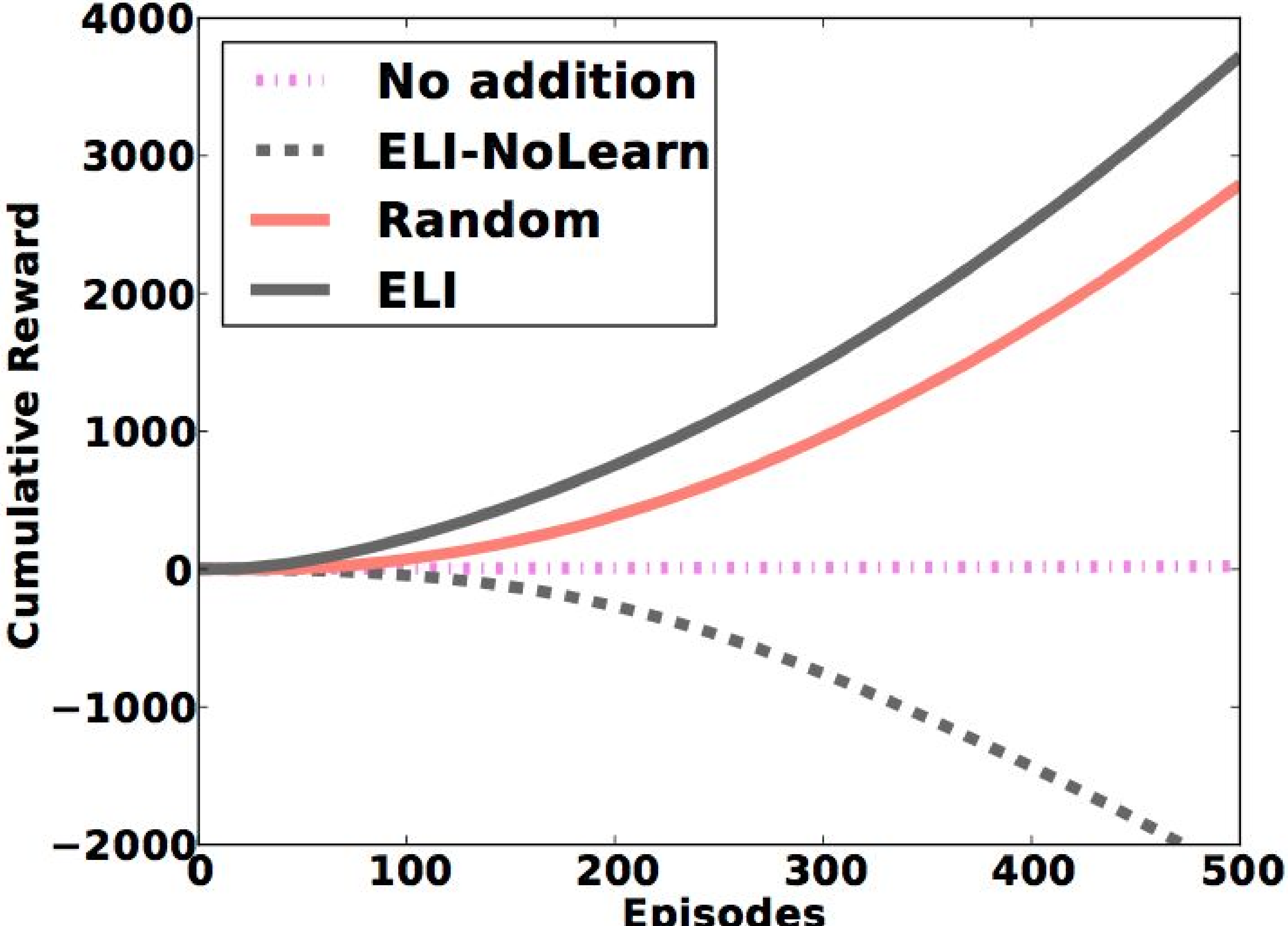
Mandel, Liu, Brunskil & Popovic, AAAI 2017

*With best choice of algorithm for choosing current value



Mandel, Liu, Brunskil & Popovic, AAAI 2017

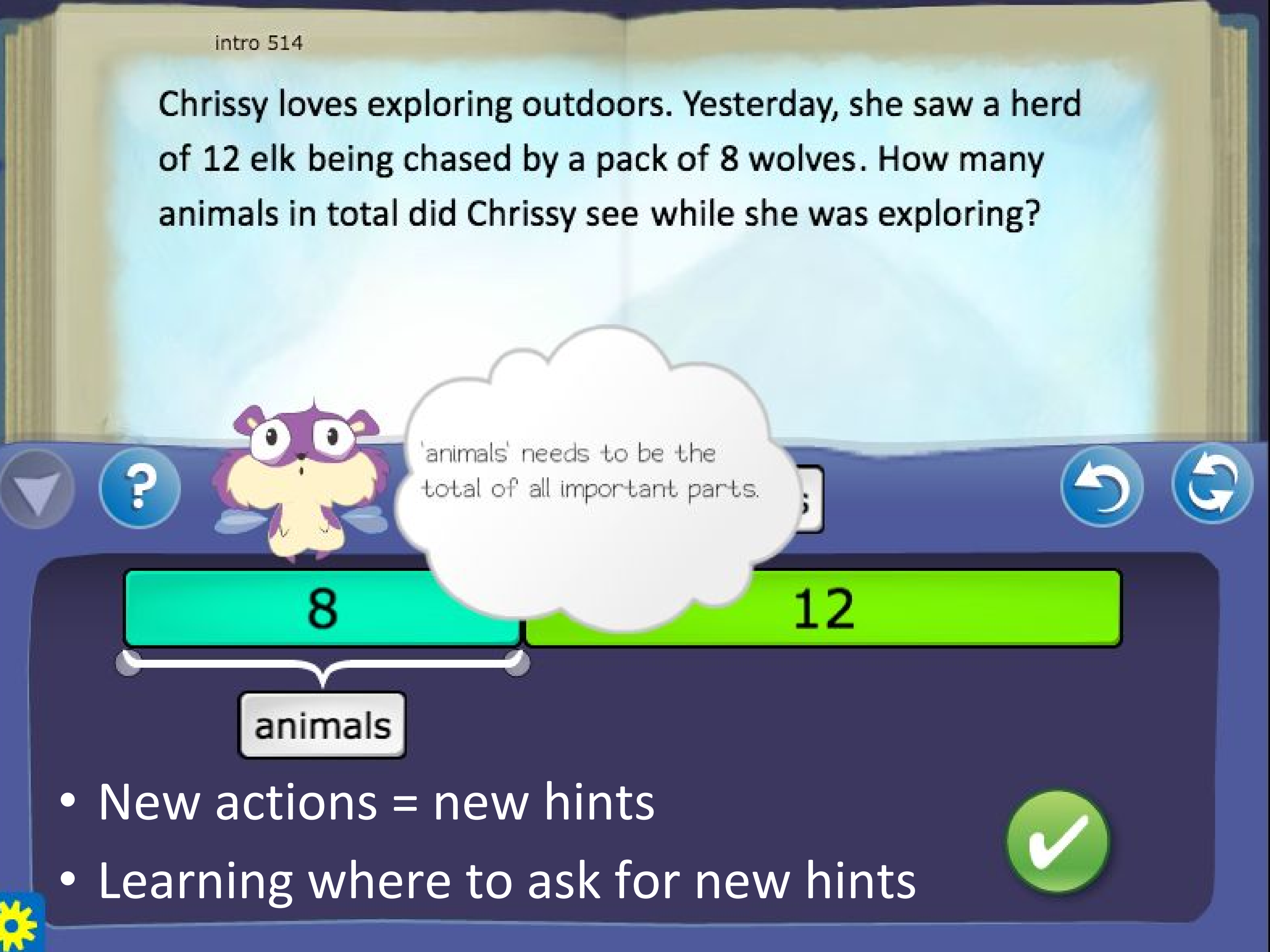
Mostly Bad Human Input



Mandel, Liu, Brunskil & Popovic, AAAI 2017

intro 514

Chrissy loves exploring outdoors. Yesterday, she saw a herd of 12 elk being chased by a pack of 8 wolves. How many animals in total did Chrissy see while she was exploring?



- New actions = new hints
- Learning where to ask for new hints

Summary

- Can use RL towards personalized, automated tutoring
  - More applications next week!
- Can create RL systems that evolve beyond their original specification
  - Not limited by original state/action space
  - Help humans-in-the-loop prioritize effort
  - Towards extraordinary performance