

# Reinforcement Learning & Learning to Promote Learning

Emma Brunskill  
Stanford University

# AI to Automate Humans

Will Robots Replace Human Drivers,  
Doctors and Other Workers?

Robots could take over 38% of U.S.  
jobs within about 15 years

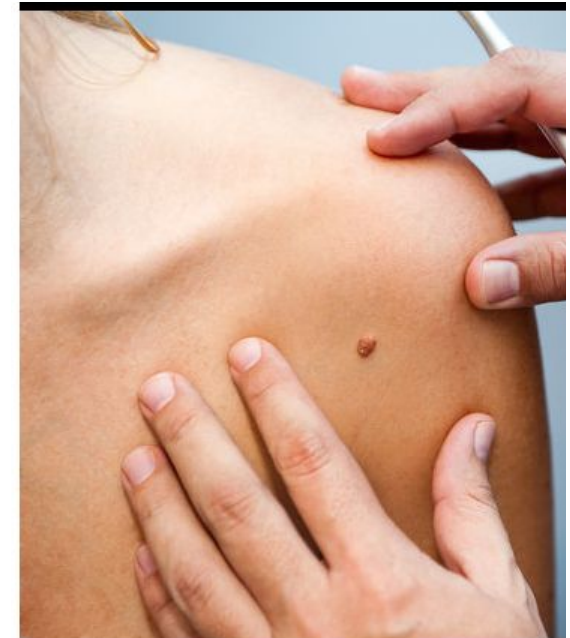
Will robots replace workers by 2030?



[https://www.google.com/url?sa=t&ci=j&q=&esrc=s&source=newssearch&cd=6&ved=0ahUKEwj9oy9tC7VANDSyWMKHRnOAwUQqQIINygAMAU&url=https%3A%2F%2Fwww.wired.com%2Fstory%2Fdriverless-cars-need-e-ars-as-well-as-eyes%2F&usg=AFQjCNHYIWQaBgaSNcJILzHIY\\_i8klygdw](https://www.google.com/url?sa=t&ci=j&q=&esrc=s&source=newssearch&cd=6&ved=0ahUKEwj9oy9tC7VANDSyWMKHRnOAwUQqQIINygAMAU&url=https%3A%2F%2Fwww.wired.com%2Fstory%2Fdriverless-cars-need-e-ars-as-well-as-eyes%2F&usg=AFQjCNHYIWQaBgaSNcJILzHIY_i8klygdw)



<http://money.cnn.com/2017/08/18/news/economy/us-farmers-immigration-automation/index.html>



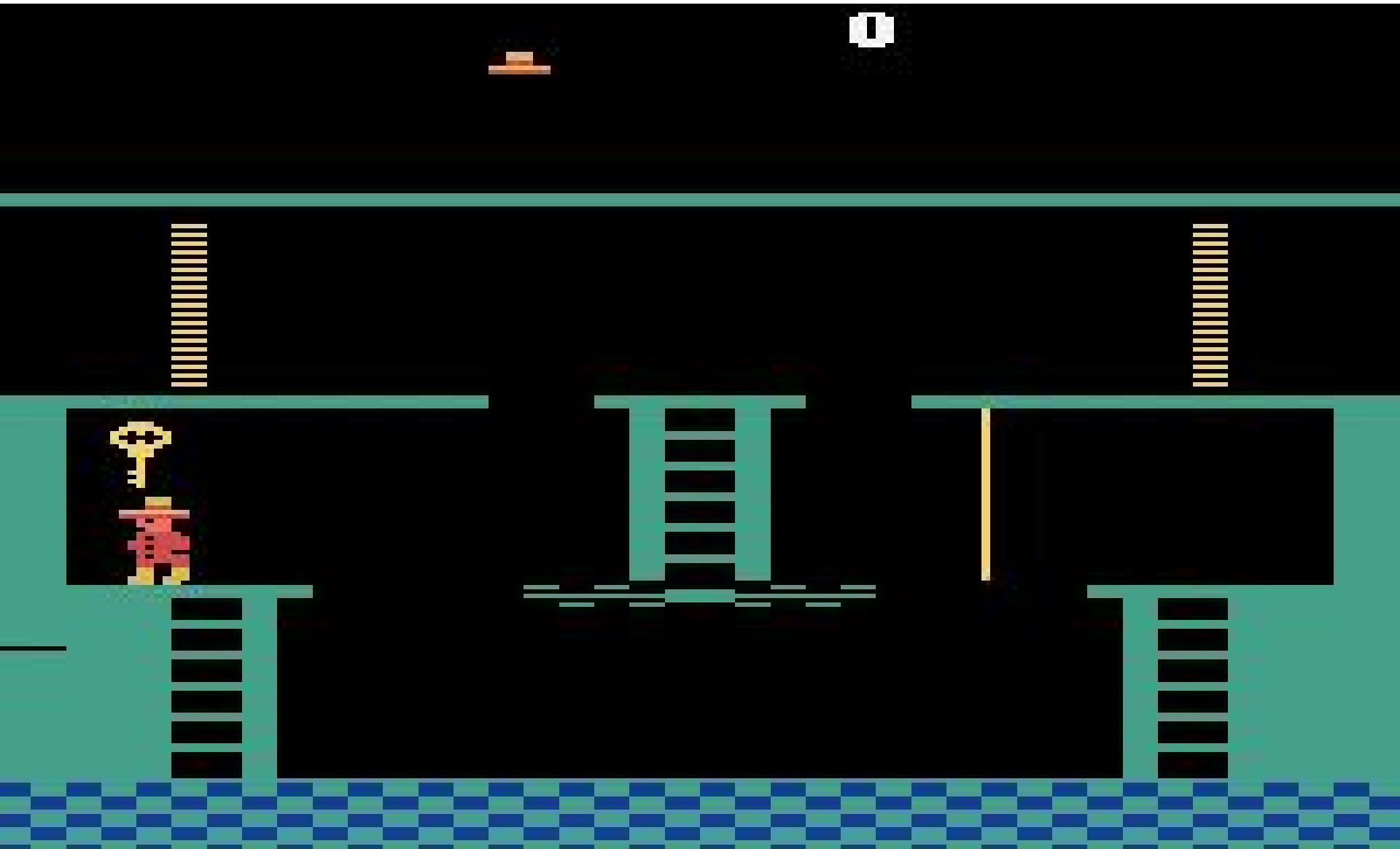
DAMIANGRETKA VIA GETTY IMAGES

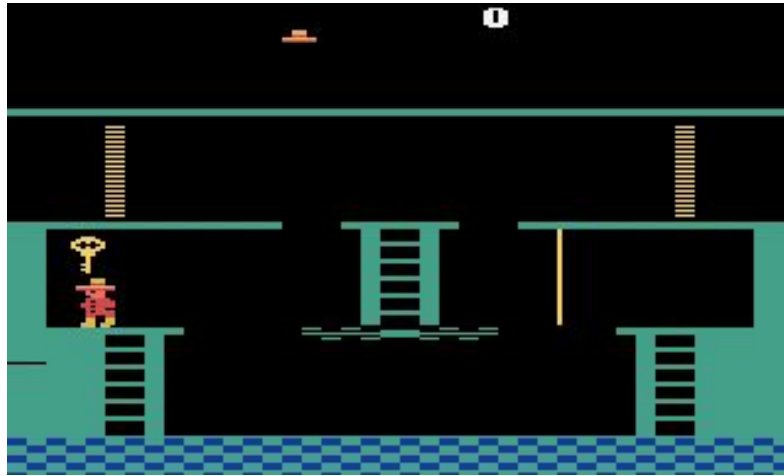
[http://www.huffingtonpost.co.uk/entry/artificial-intelligence-helping-doctors-diagnose-skin-cancer-faster\\_uk\\_599d428be4b0a296083b0778](http://www.huffingtonpost.co.uk/entry/artificial-intelligence-helping-doctors-diagnose-skin-cancer-faster_uk_599d428be4b0a296083b0778)

# Artificial Intelligence to Amplify People

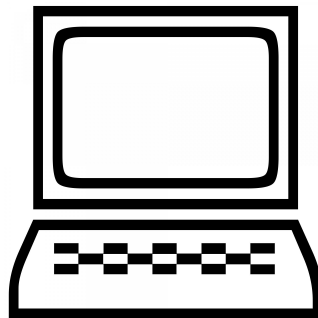


# Reinforcement Learning





Action  
(turn left)



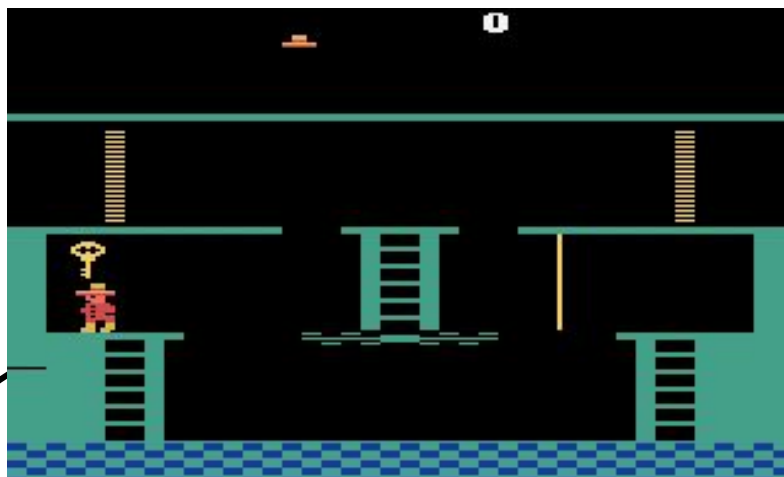
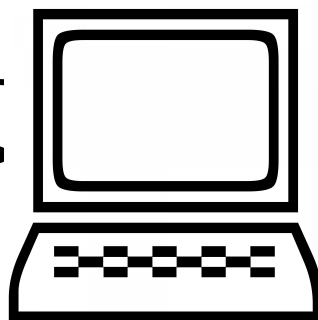


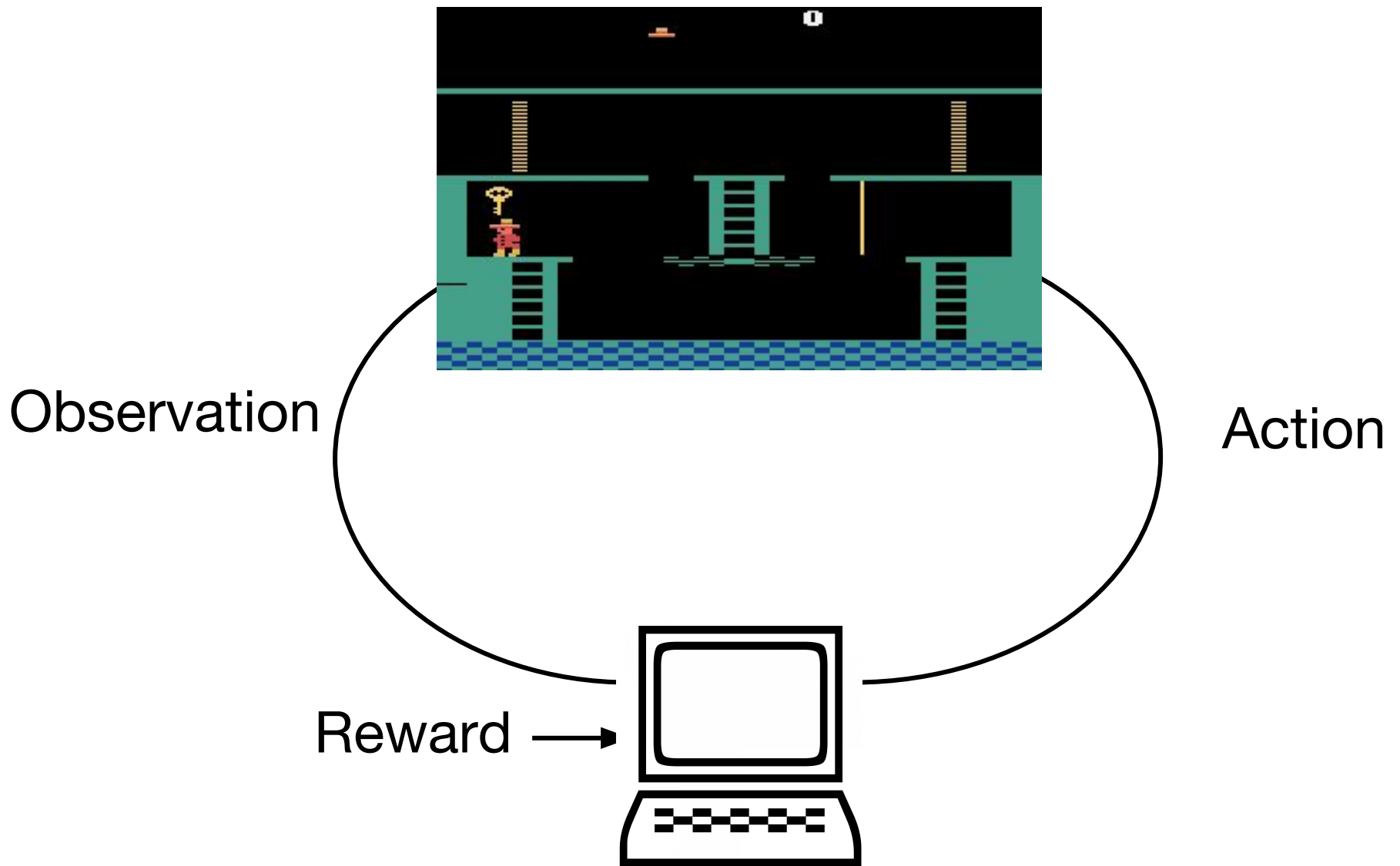
Image  
(pixel colors)

Action  
(turn left)

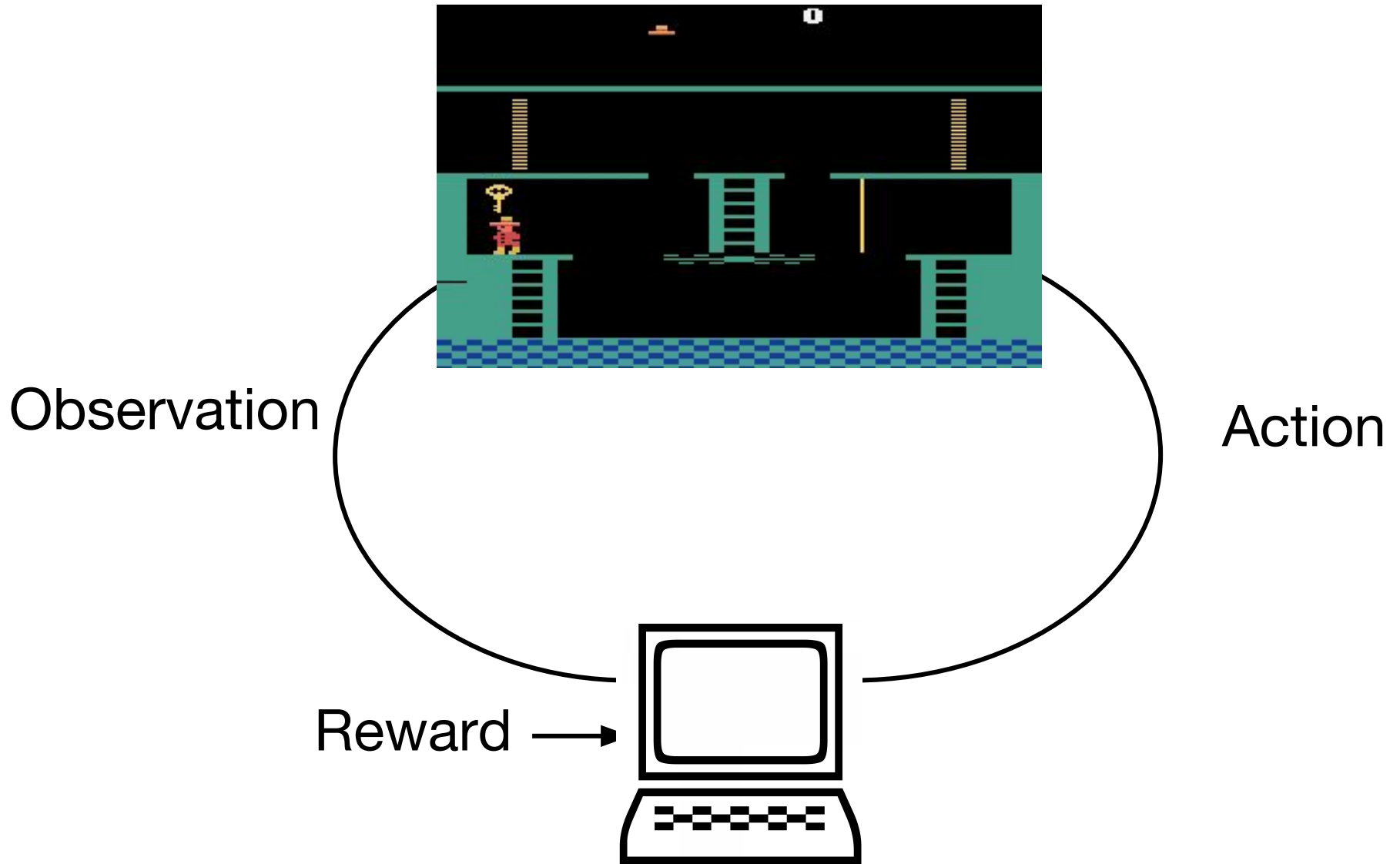
Score



# Reinforcement Learning



# Reinforcement Learning

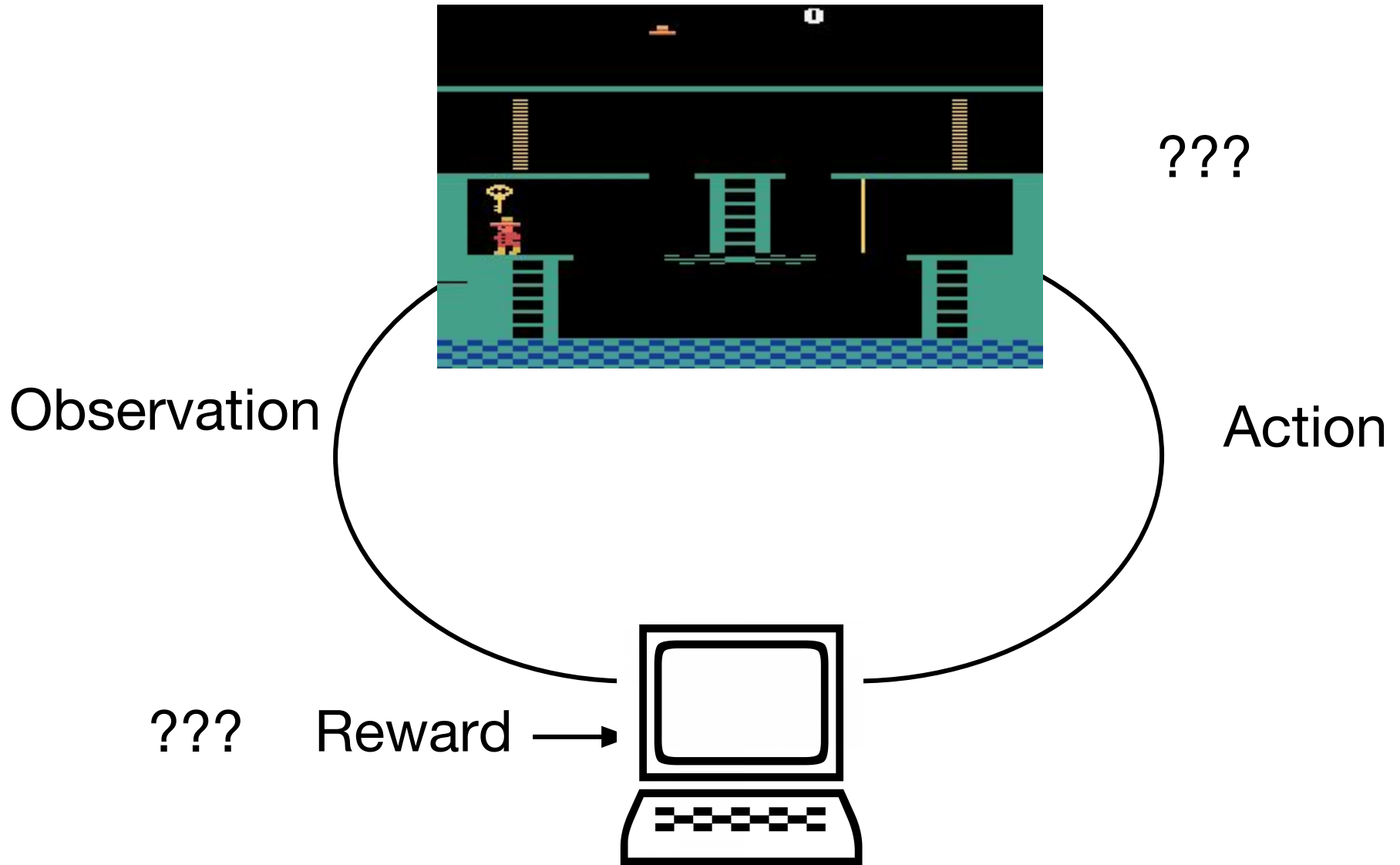


*Policy: Map Observations  $\rightarrow$  Actions*

*Goal: Choose actions to maximize expected rewards*



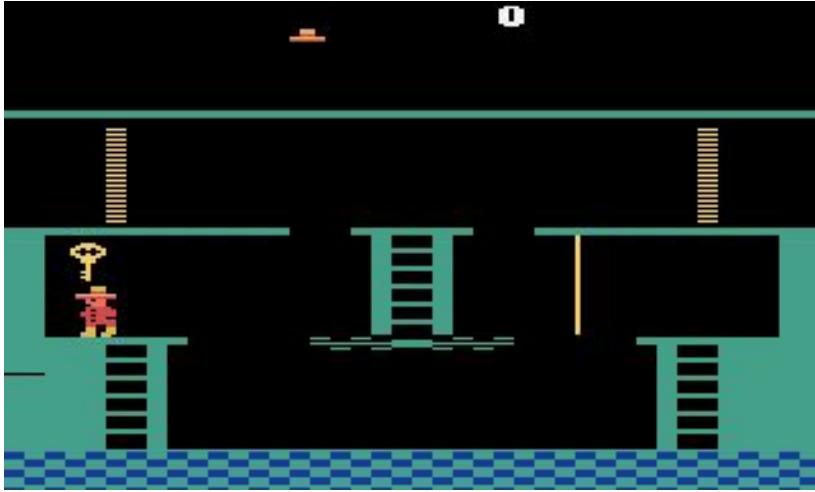
# But Don't Know How World Works!



*Policy: Map Observations  $\rightarrow$  Actions*

*Goal: Choose actions to maximize expected rewards*

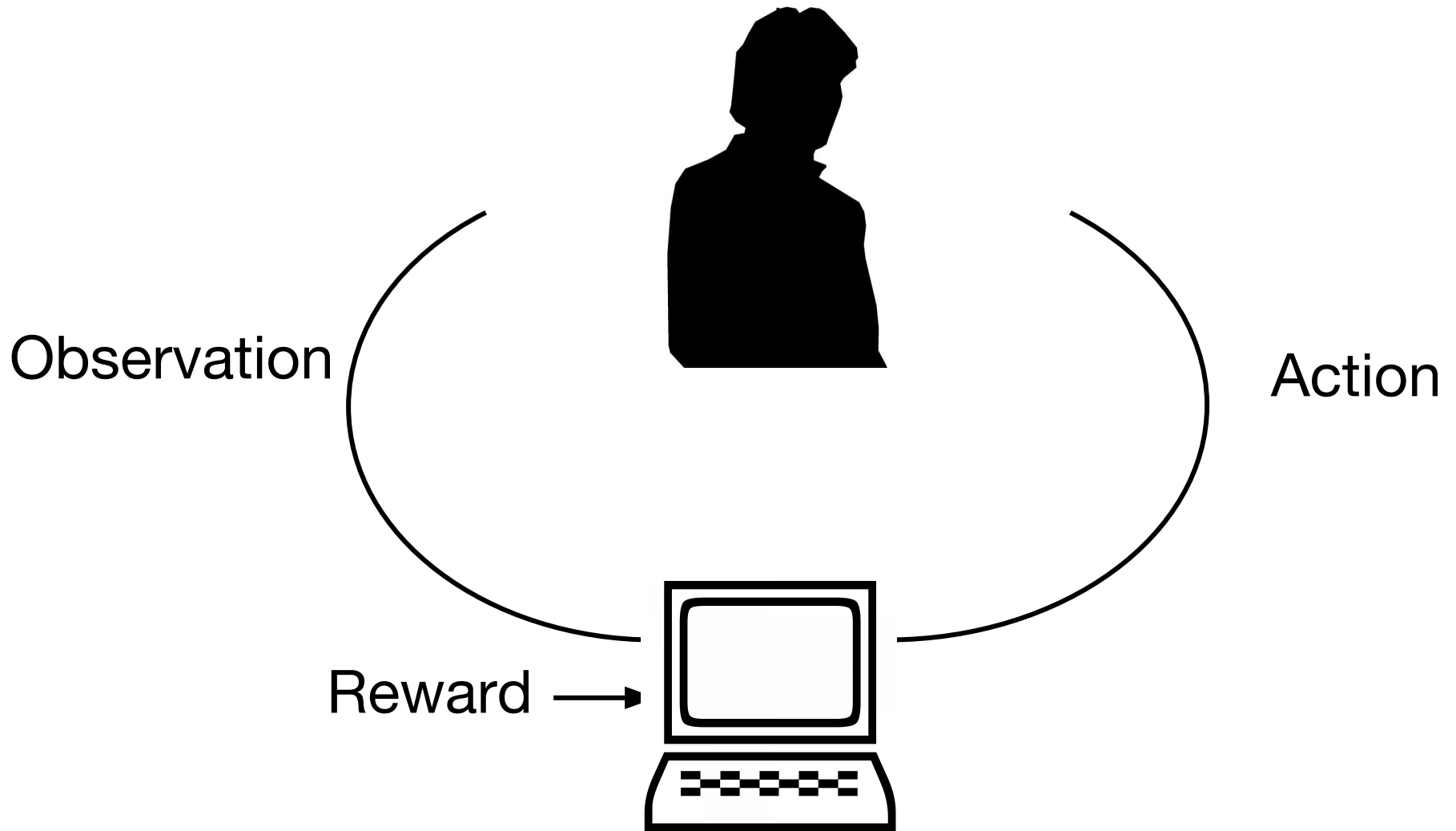
# Reinforcement Learning Progress



# Real Potential: Humans & AI



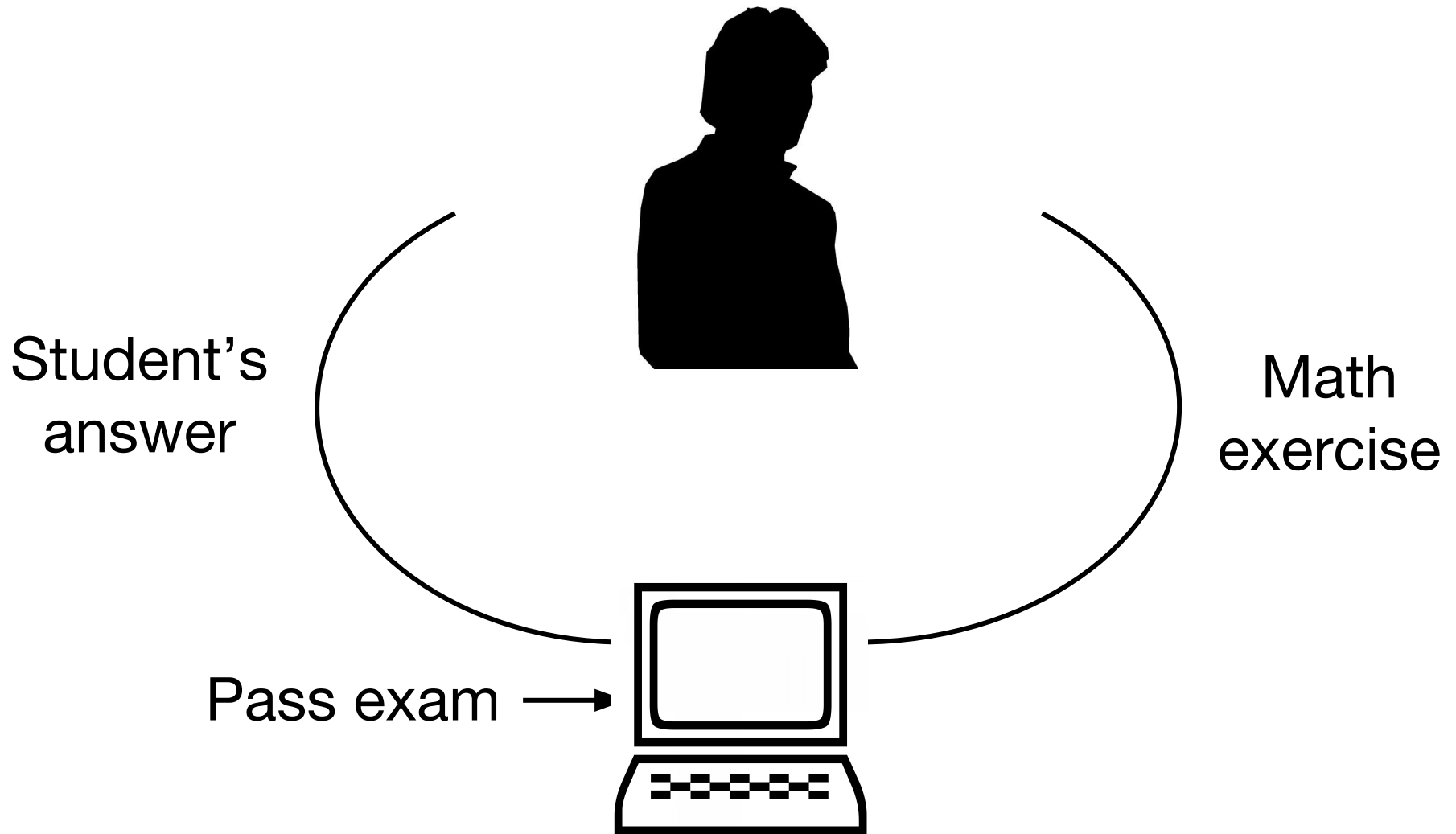
# Reinforcement Learning with and for People



*Policy: Map Observations  $\rightarrow$  Actions*

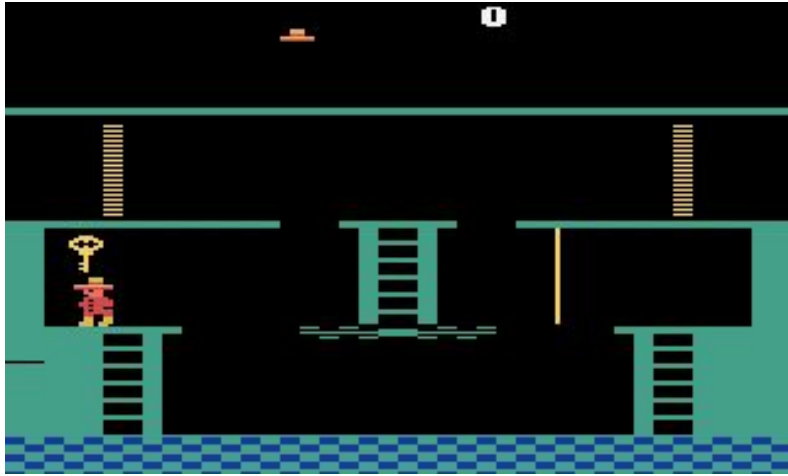
*Goal: Choose actions to maximize expected rewards*

# Reinforcement Learning with and for People

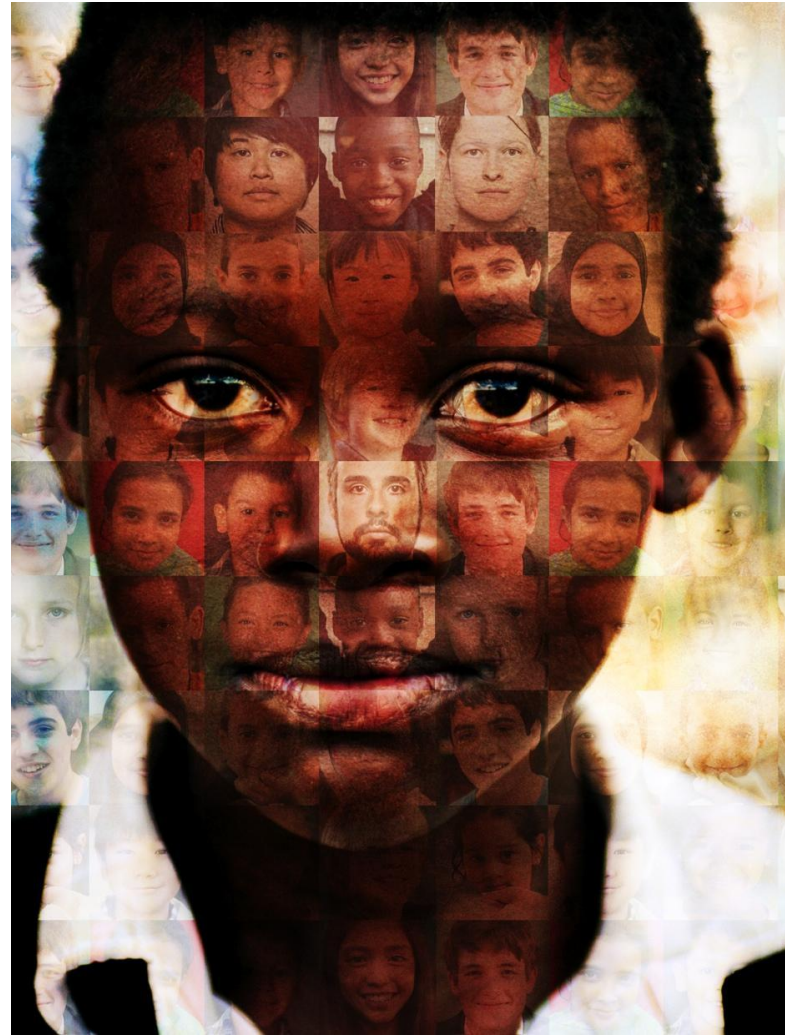


*Policy: Map Observations  $\rightarrow$  Actions*

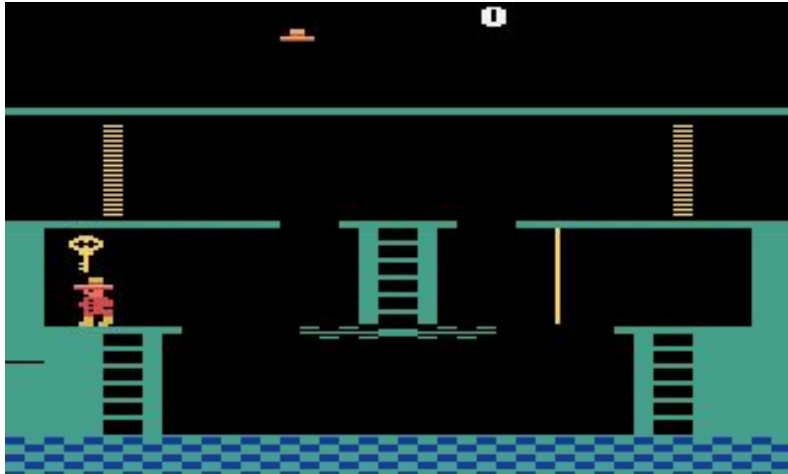
*Goal: Choose actions to maximize student outcomes*



≠







≠



Cheap to try things, or  
Simulate

High stakes  
Hard to model

# Reinforcement Learning & Learning to Promote Learning

Making better decisions by

- 1) Learning from past experience
- 2) Having humans help machines

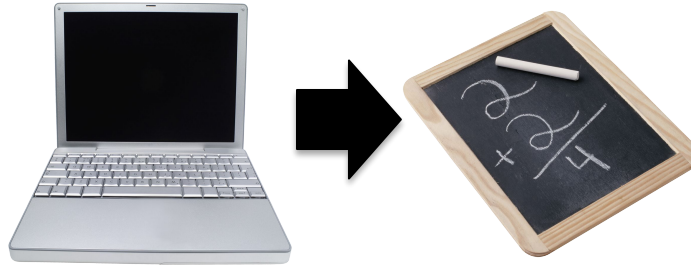


A Classrooms



Avg Score: 95

A Classrooms



Avg Score: 95

B Classrooms



Avg Score: 92

# What should we do for a new student?

A Classrooms



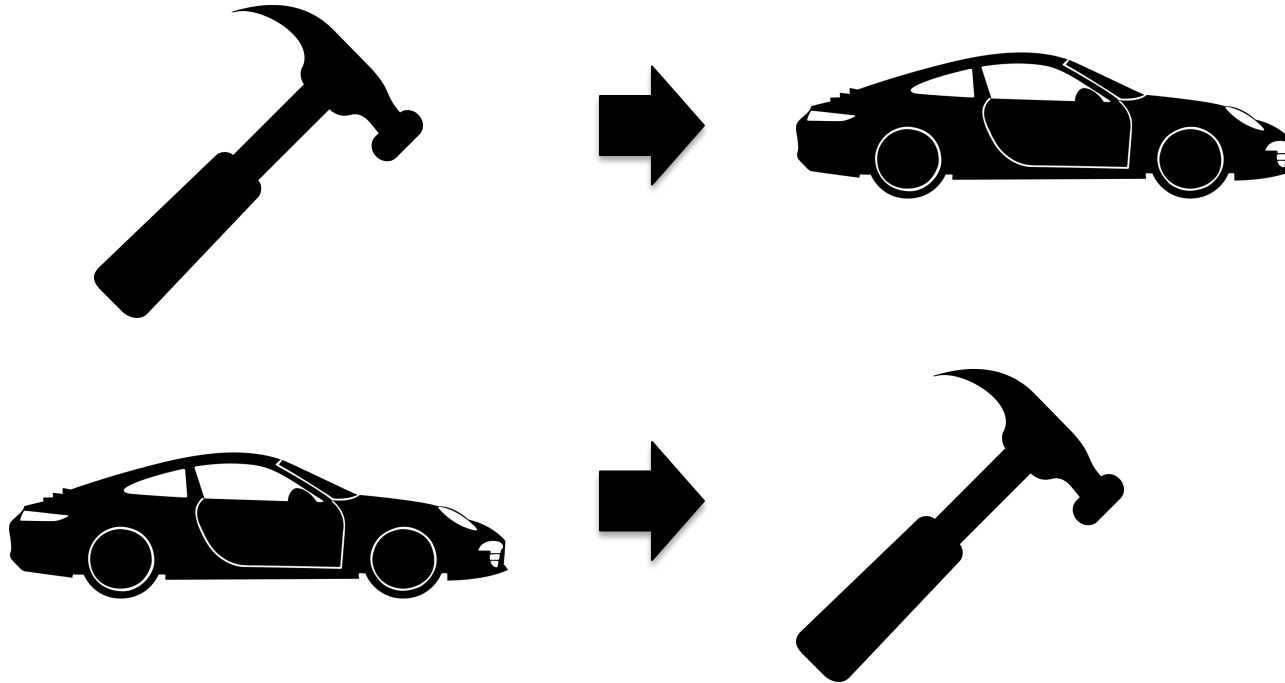
Avg Score: 95

B Classrooms

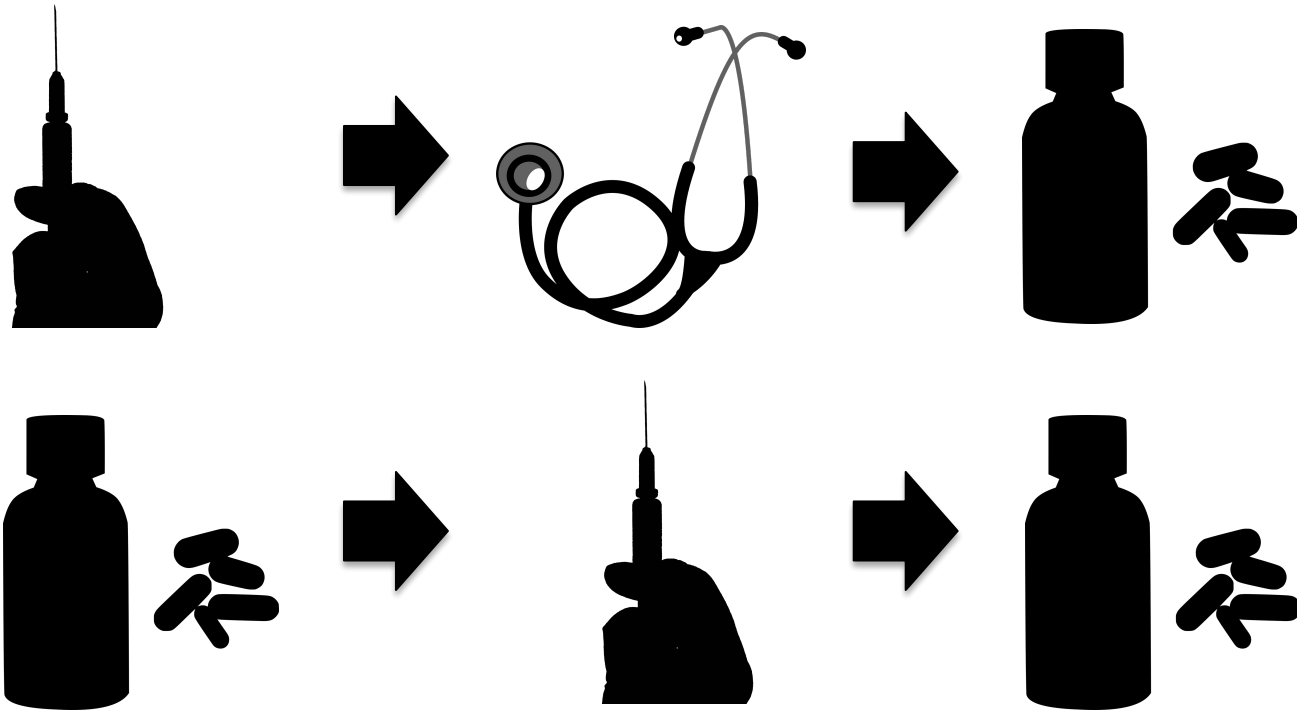


Avg Score: 92

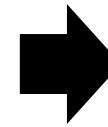
Comes Up in Many Domains: e.g.  
Equipment Maintenance Scheduling



# Comes Up in Many Domains: e.g. Patient Treatment Ordering



# Core Aspect of Intelligent Behavior

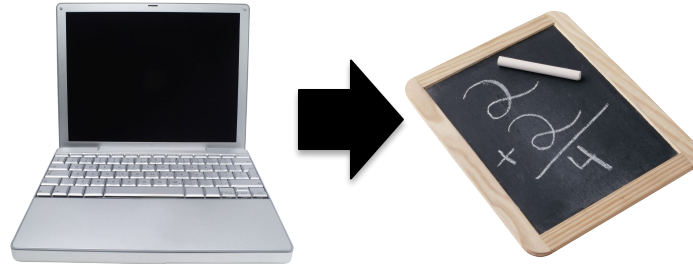


How best to act  
in the future?

Data about past decisions & outcomes

# Challenge: Counterfactual Reasoning

A Classrooms



Avg Score: 95

B Classrooms



Avg Score: 92

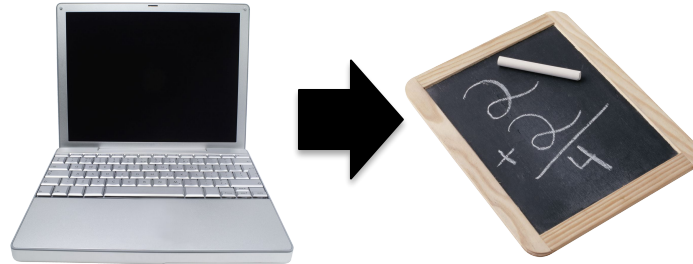
B Classrooms



Avg Score: ????

# Challenge: Generalization to Untried Policies

A Classrooms



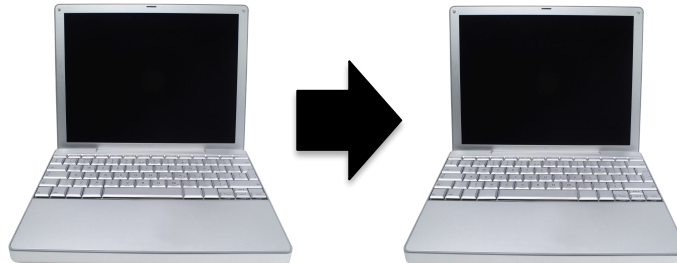
Avg Score: 95

B Classrooms



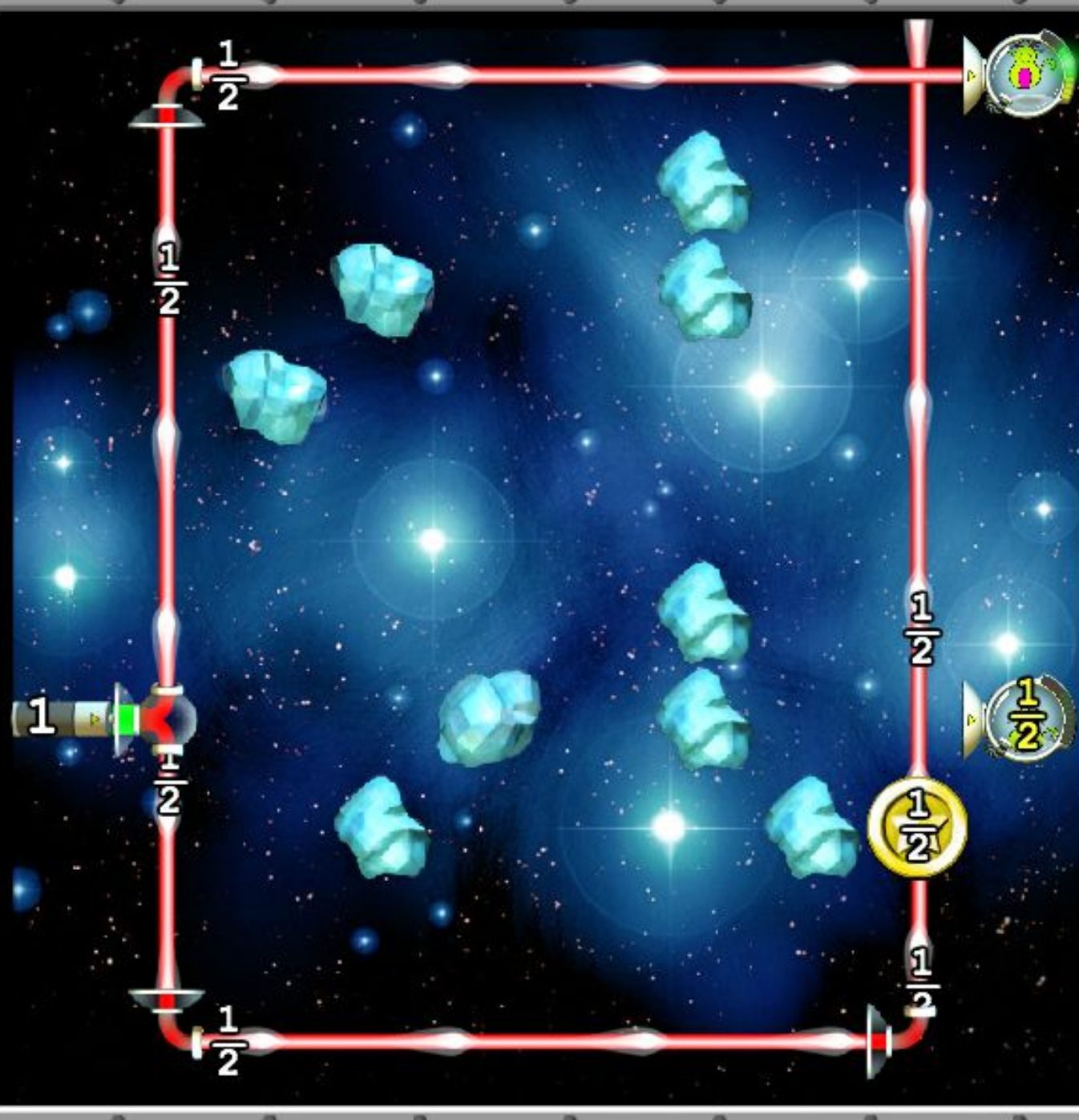
Avg Score: 92

B Classrooms

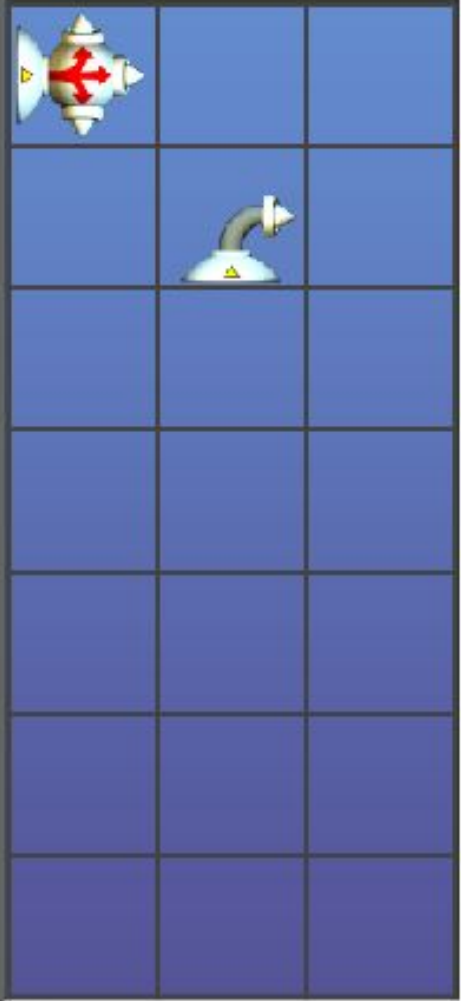


Avg Score: ????





Level 1:8  
Fork



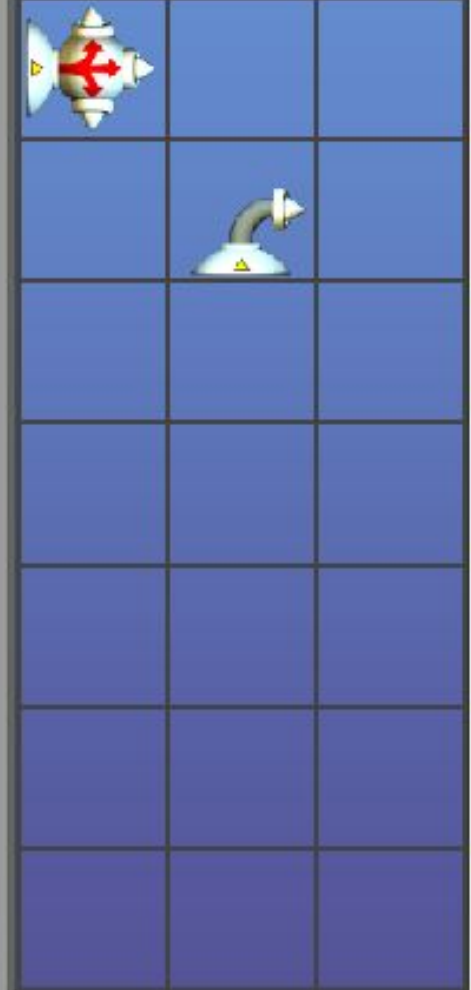
MENU

OPTIONS



Policy: Player state  $\rightarrow$  level  
Goal: Maximize engagement  
Old data: ~11,000 students

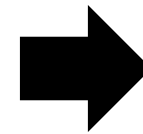
Level 1:8  
Fork



MENU

OPTIONS

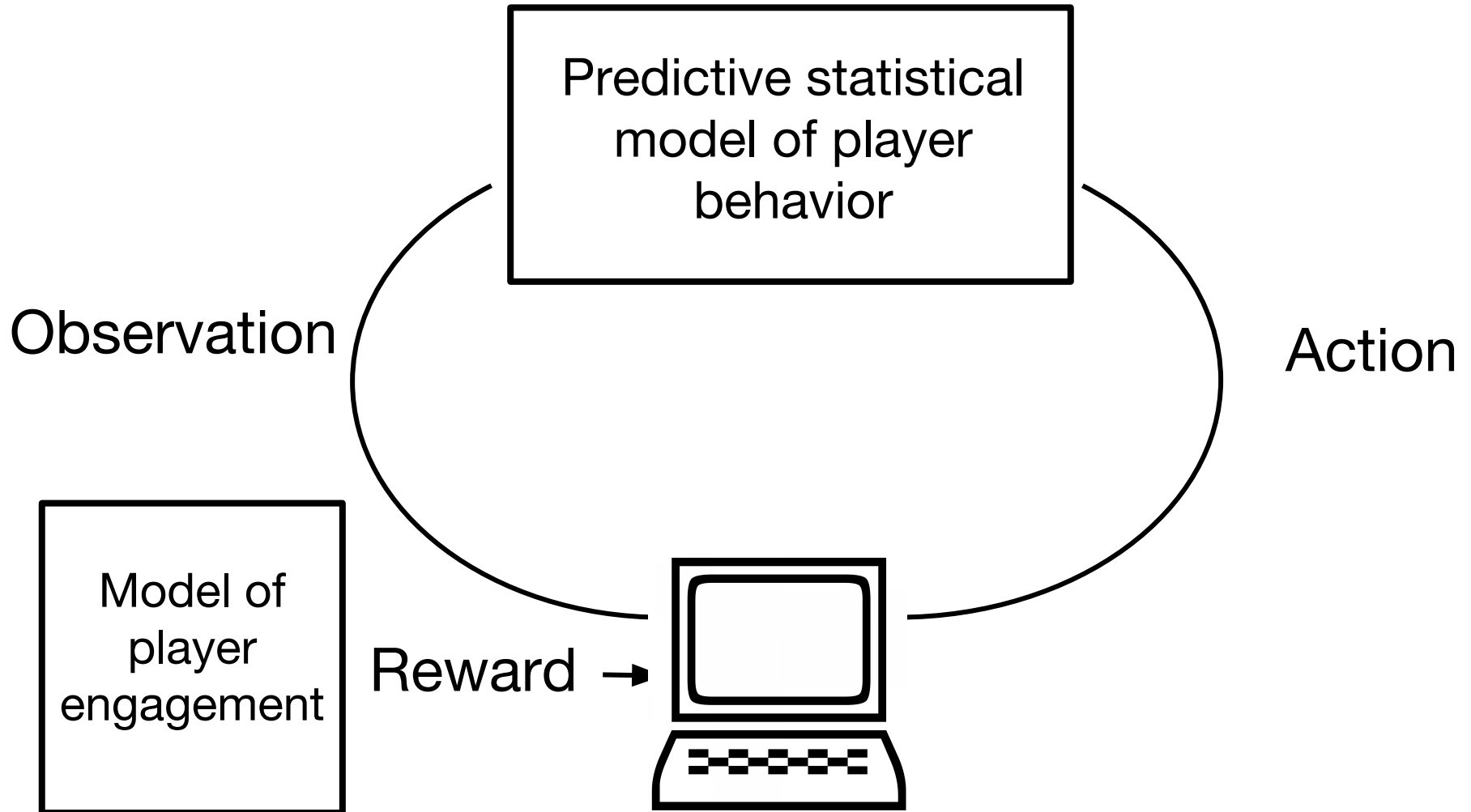




## Statistical Predictive model

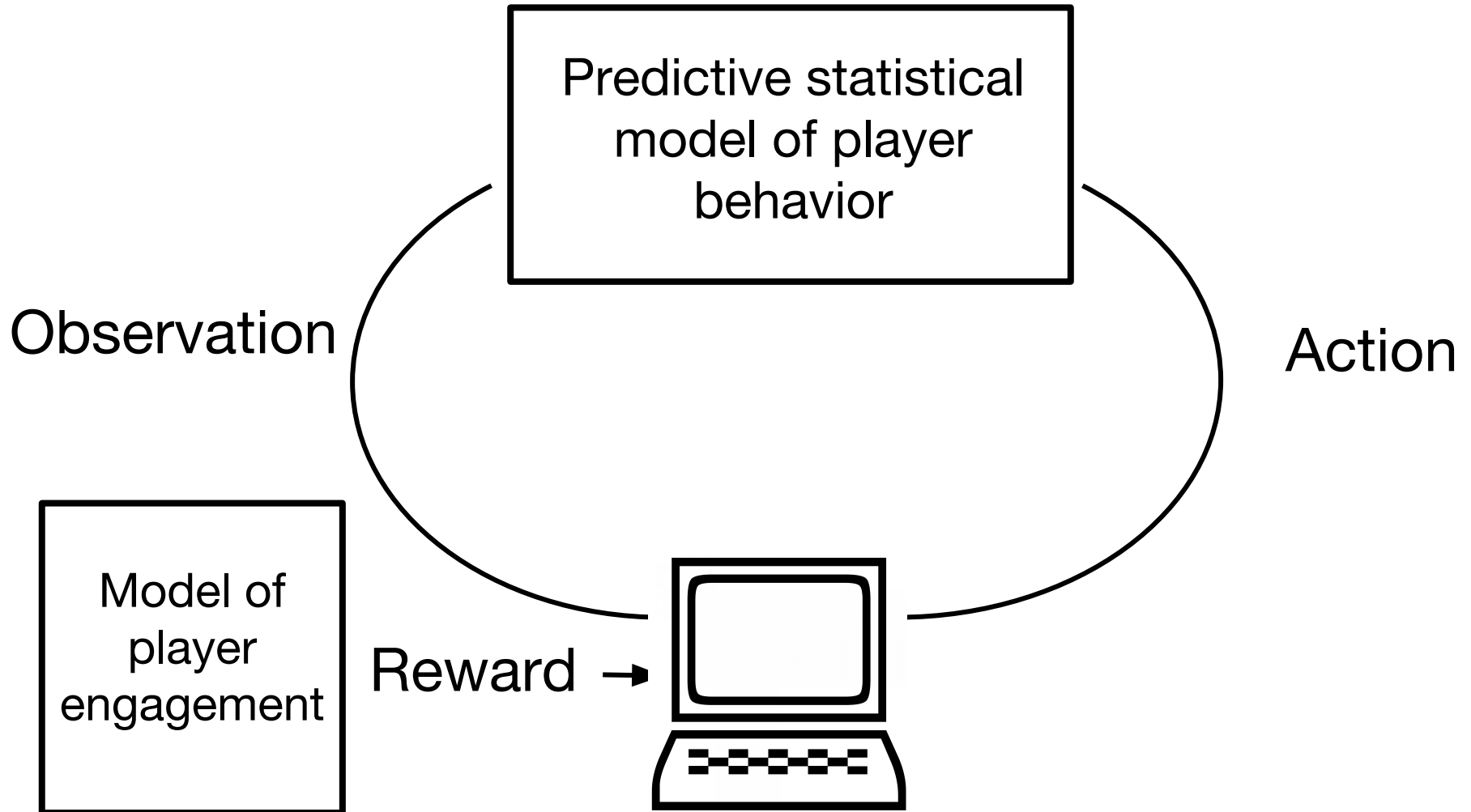
(e.g. Predict if student  
will get next level  
correct)

# Use Models as a Simulator



*Goal: Choose actions to maximize expected rewards*

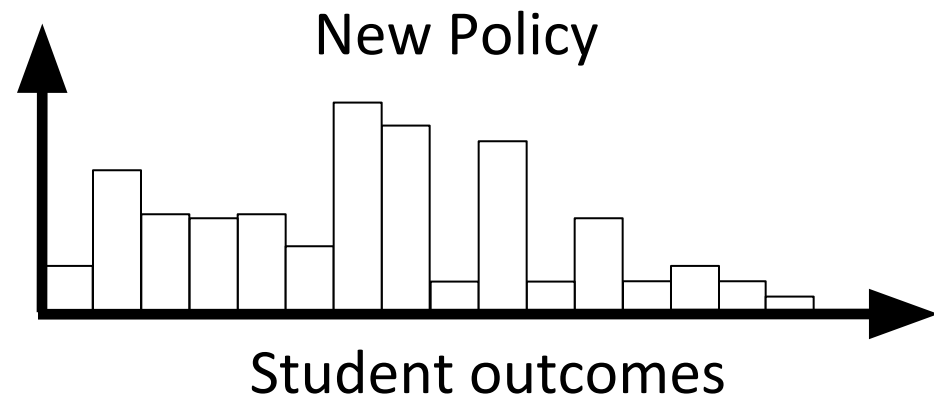
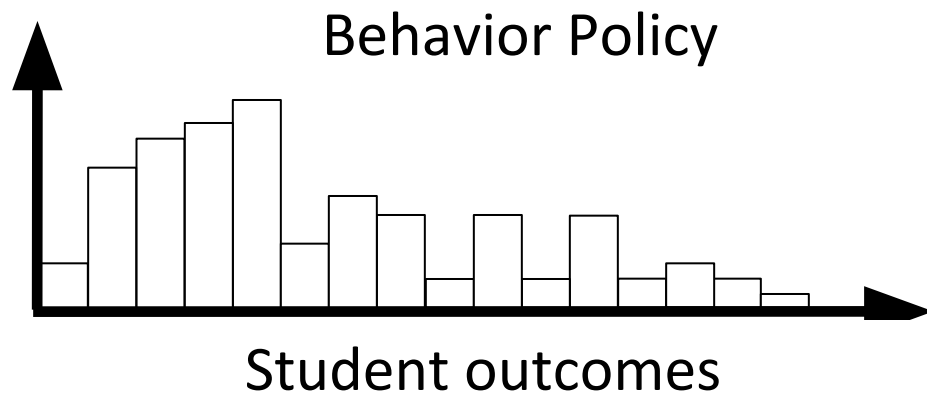
# Problem: When is a Model Good Enough?



*Goal: Choose actions to maximize expected rewards*

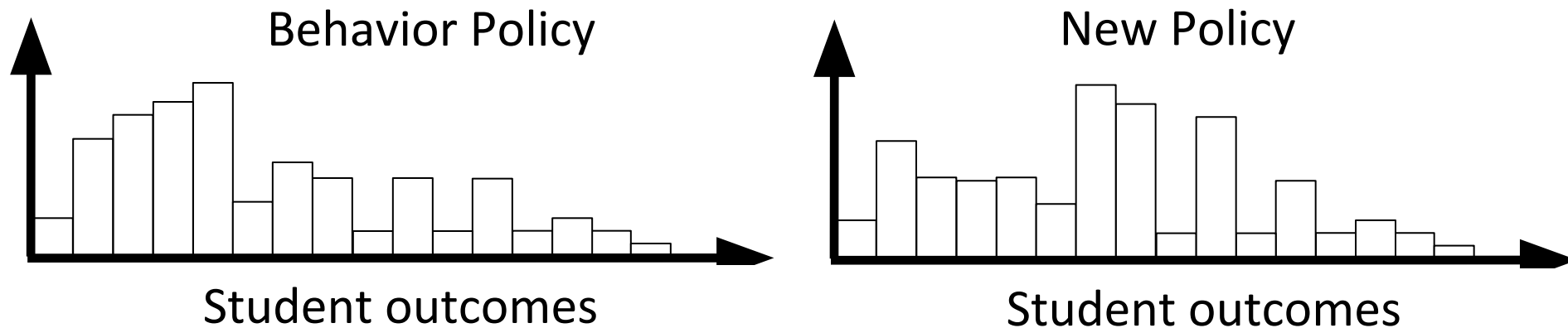
# Alternative: Reweigh Old Experience to Look Like New Policy

- No statistical predictive model assumptions



# Alternative: Reweigh Old Experience to Look Like New Policy

- No statistical predictive model assumptions

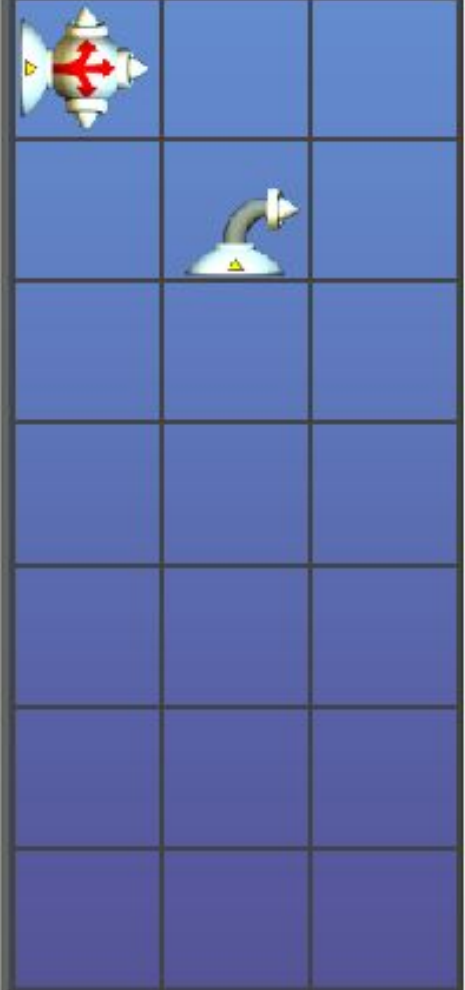


- Unbiased\* estimate of new policy's performance

\*Under mild assumptions

We used to find a policy with  
30% higher engagement  
(Mandel et al. 2014)

Level 1:8  
Fork



MENU

OPTIONS



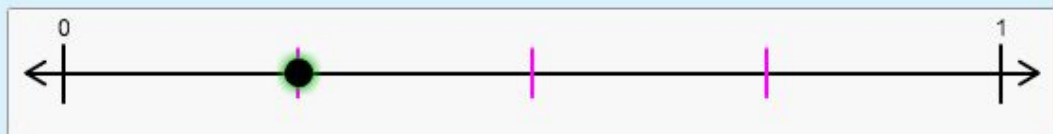
# When Making Many Decisions...

Fractions Tutor: Part 1: ◀ 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 ▶

## Fraction Identification Tutor

A **Let's name fractions  
using number lines!**

- 1 Brittany bought a watermelon to share with three of her friends. Each of the watermelon pieces were equal-sized. Brittany ate  $\frac{1}{4}$  of the watermelon. Use the number line to show how much of the watermelon Brittany ate.



Number of  
sections:

4

?

Hint

← Previous

Next →



Super!

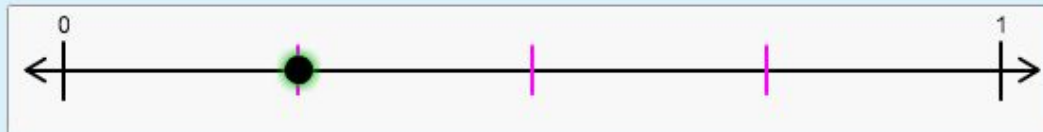
# Towards Better Estimates of New Policies

Fractions Tutor: Part 1: ◀ 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 ▶

## Fraction Identification Tutor

A **Let's name fractions  
using number lines!**

- 1 Brittany bought a watermelon to share with three of her friends. Each of the watermelon pieces were equal-sized. Brittany ate  $\frac{1}{4}$  of the watermelon. Use the number line to show how much of the watermelon Brittany ate.



Number of  
sections:

4

?

Hint

← Previous

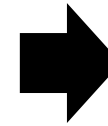
Next →



Super!

- Trade bias and variance
- New methods to combine models & direct evaluation (Guo, Thomas, B 2017; Thomas and B 2016)

# Towards Using Old Data to Confidently Identify Better Policies for Future Use



How best to act  
in the future?

Data about past decisions & outcomes

# Reinforcement Learning & Learning to Promote Learning

- Making better decisions by
  - Learning from past experience
  - Having humans help machines

# Histogram Tutor

**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

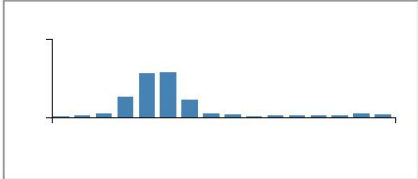
Practice Labeling No Histogram: Voters

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

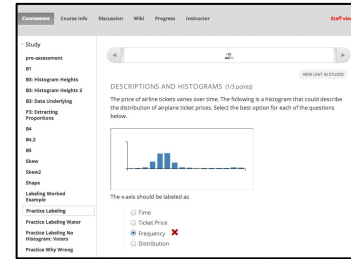
☐ Time

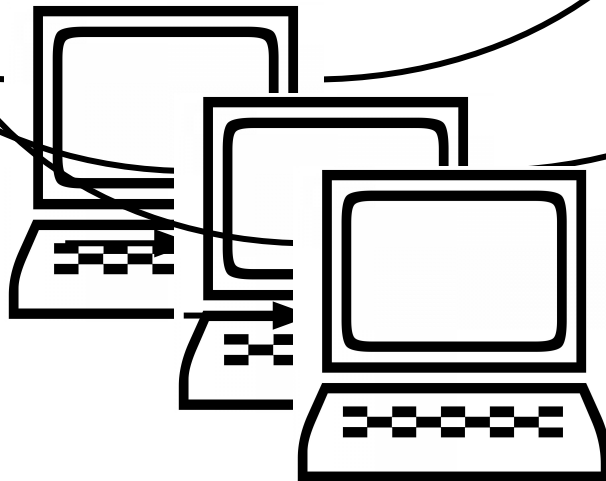
☐ Ticket Price

☒ Frequency **✗**

☐ Distribution

Correct/  
Wrong





# Over Time Tutoring System Stopped Giving Some Problems to Students

**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

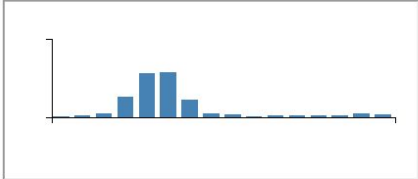
Practice Labeling No Histogram: Voters

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



# System Self-Diagnosed that Problems Weren't Helping Student Learning

**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

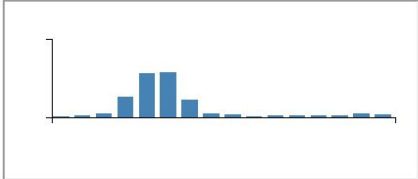
Practice Labeling No Histogram: Voters

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

# Humans are Invention Machines

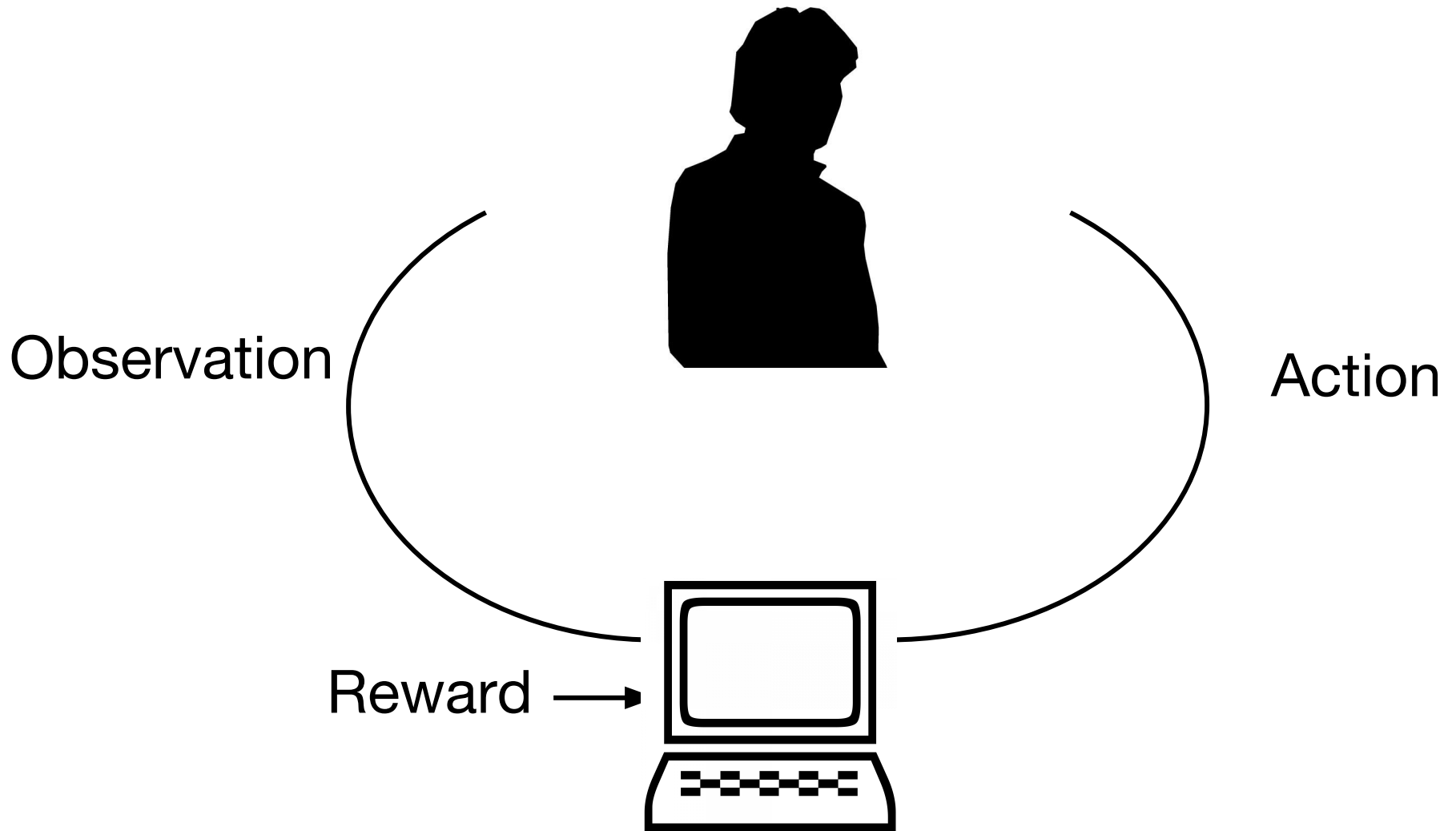


New actions



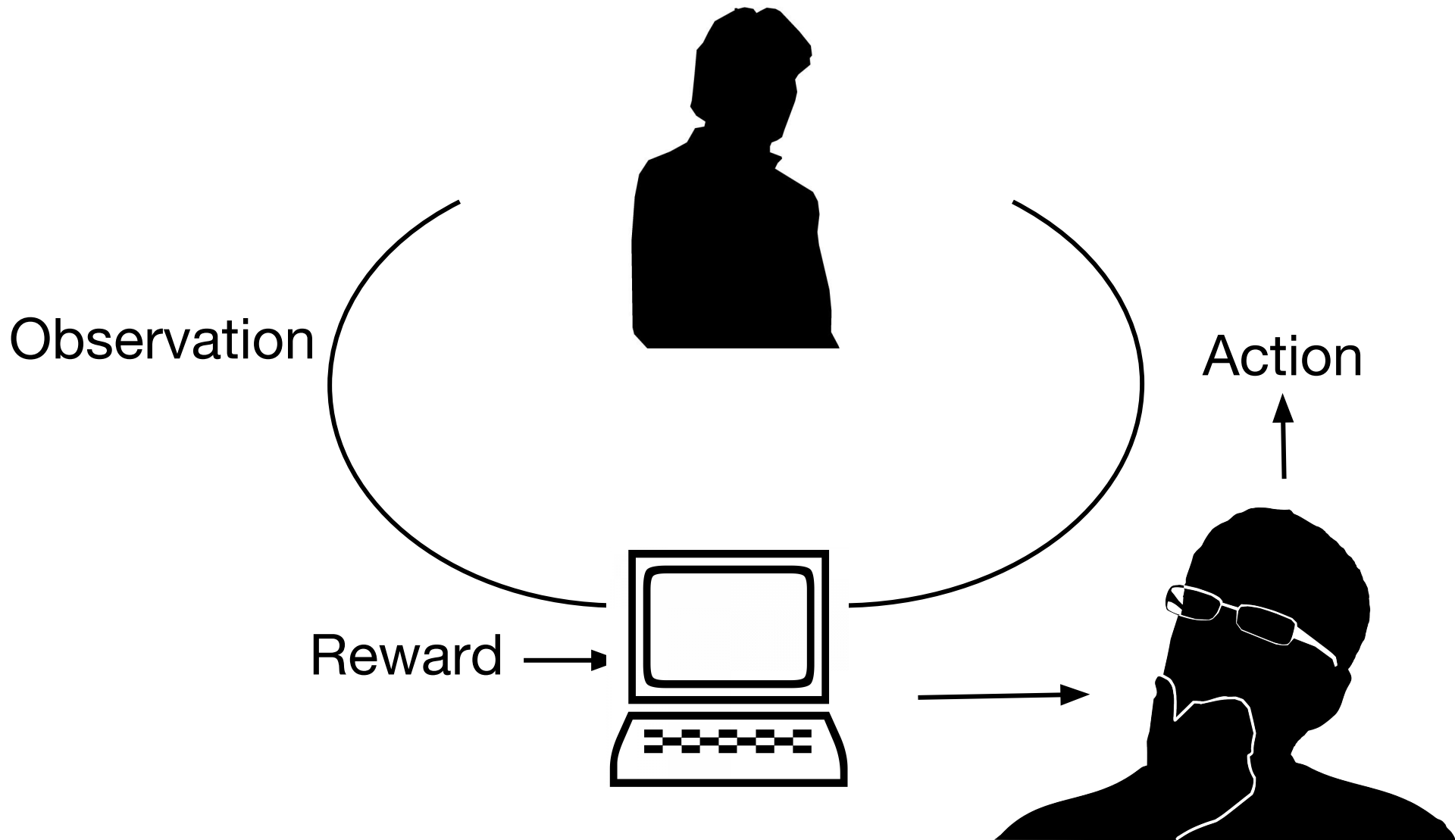
New sensors

# Reinforcement Learning



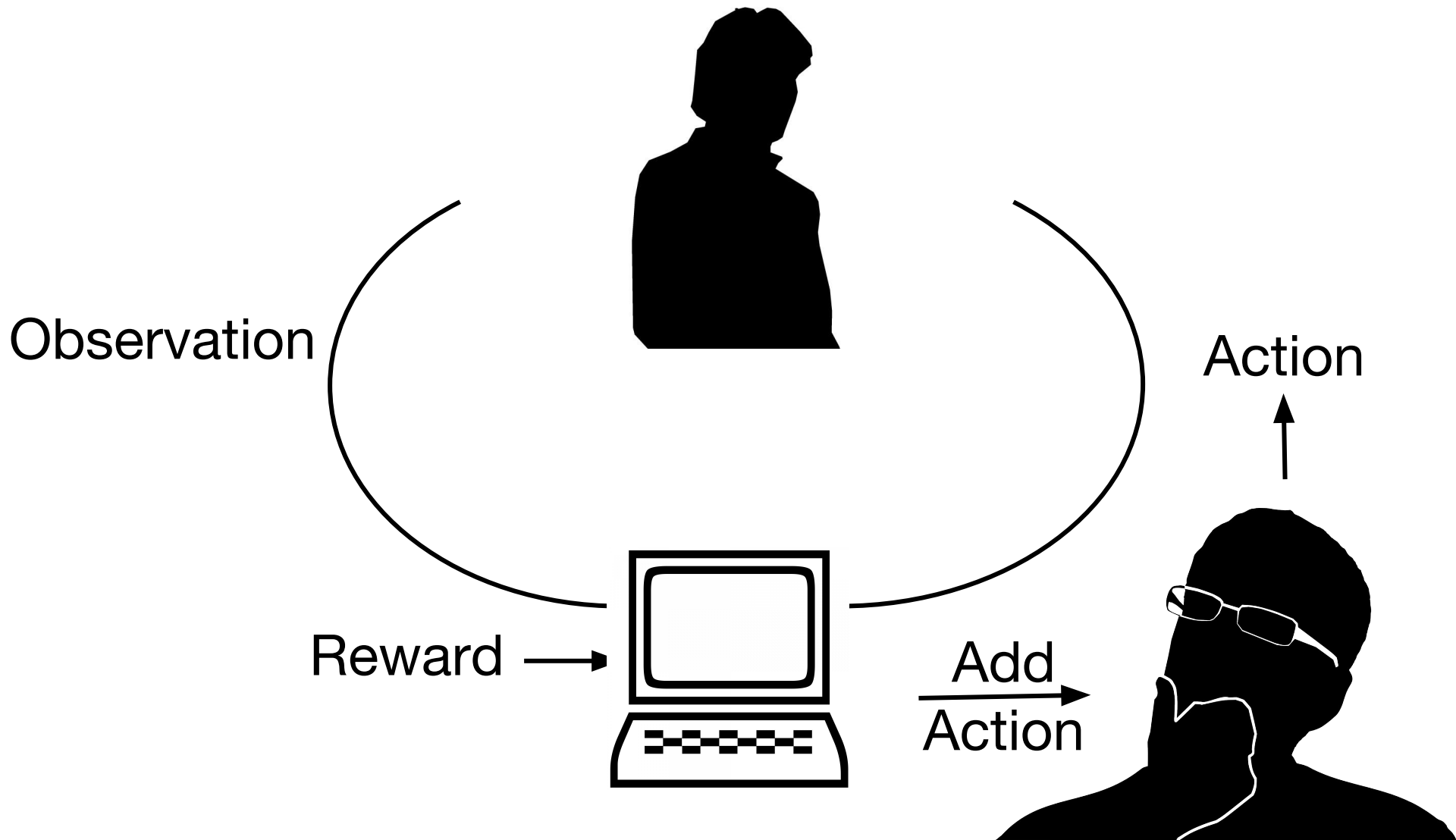
*Goal: Choose actions to maximize expected rewards*

# Human in the Loop Reinforcement Learning



*Goal: Choose actions to maximize expected rewards*

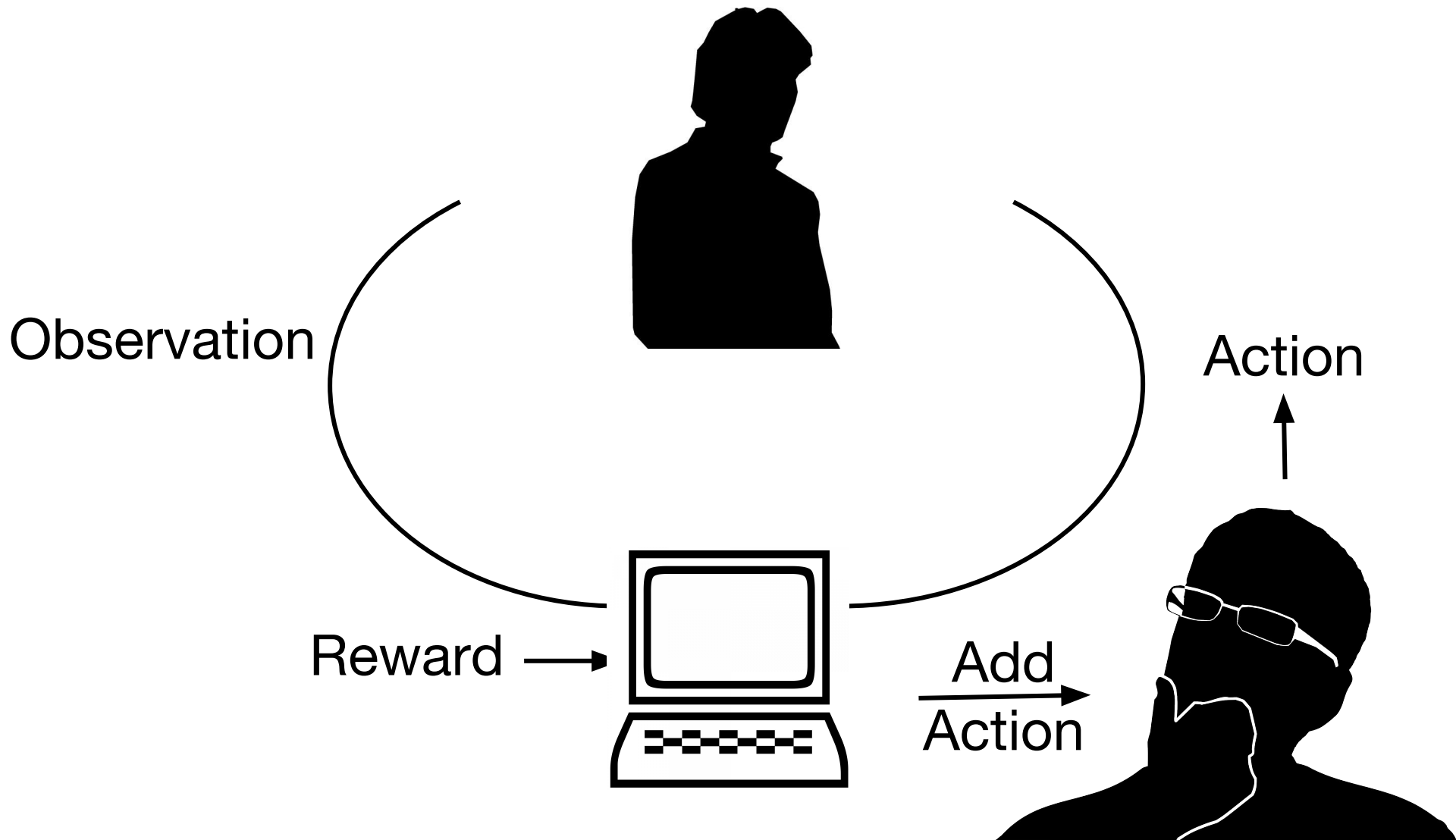
# Human in the Loop Reinforcement Learning



*Goal: Choose actions to maximize expected rewards*

# Where to Ask for New Actions?

Mandel, Liu, Brunskil & Popovic, AAAI 2017



*Goal: Choose actions to maximize expected rewards*

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?



'animals' needs to be the total of all important parts.

8

12

animals





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?



'animals' needs to be the total of all important parts.

8

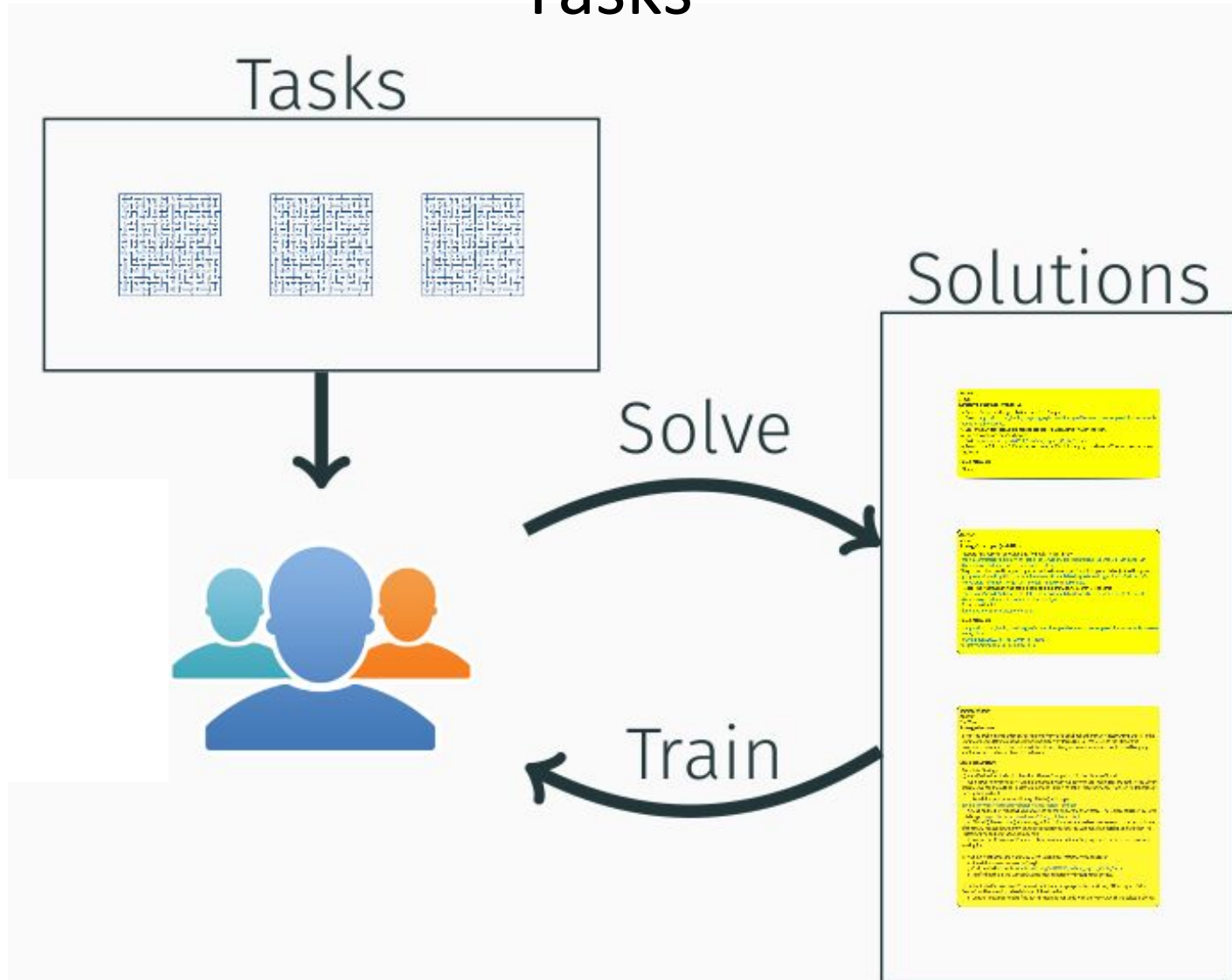
12

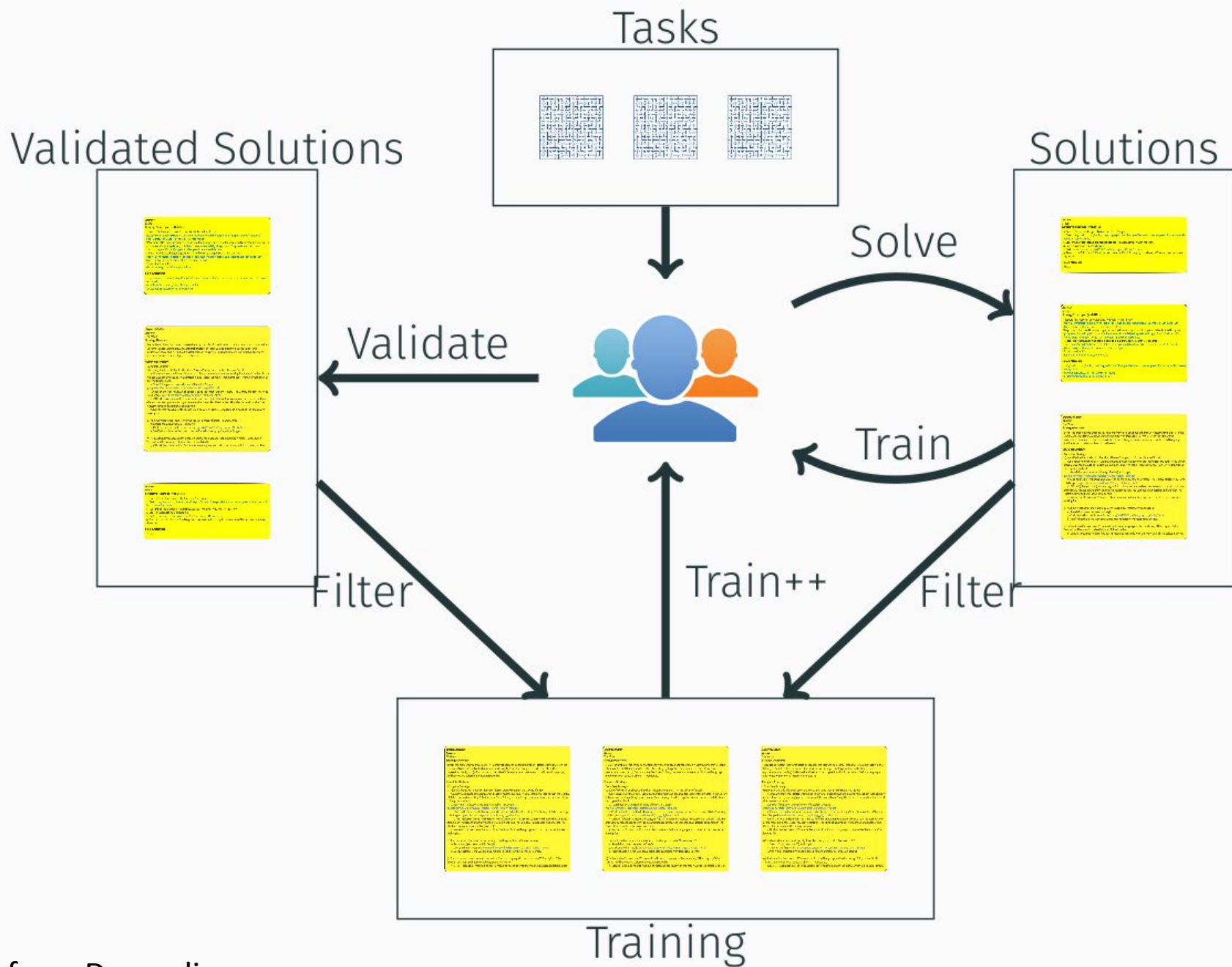
animals

- New actions = new hints
- Learning where to ask for new hints
- **People helping computers to teach people**



# People Helping Computers to Teach People Tasks







# Reinforcement Learning & Learning to Promote Learning

- Making better decisions by
  - Learning from past experience
  - Having humans help machines



# Thanks to



and Karan Goel, Travis Mandel, Yun-En Liu, NSF,  
ONR, Microsoft, Google, Yahoo & IES



# Reinforcement Learning & Learning to Promote Learning

- Making better decisions by
  - Learning from past experience
  - Having humans help machines

