

# CMSC 671 Principles of Artificial Intelligence: (Guest Lecture)

## Markov Decision Processes I

October 12, 2023

Cassandra Kent, PhD

Robot gridworld example adapted from Brian Hrolenok

---

By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# Guest lecturer bio

- My name is Cassandra Kent, my pronouns are she/her
- I have a PhD in robotics from Georgia Tech, and recently completed a postdoc at UPenn
- As a teacher, I'm a certified [CIRTL associate](#) through GT's Tech to Teaching program

---

By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# Planning Review

Problem: Find a sequence of **actions** that **transition** the **initial state** into a state which passes the **goal test**

Solution: a **plan** - sequence of actions

$$[a_0, a_1, a_2, \dots, a_n]$$

For environments that are:

- Fully observable, deterministic, sequential, static, discrete/continuous

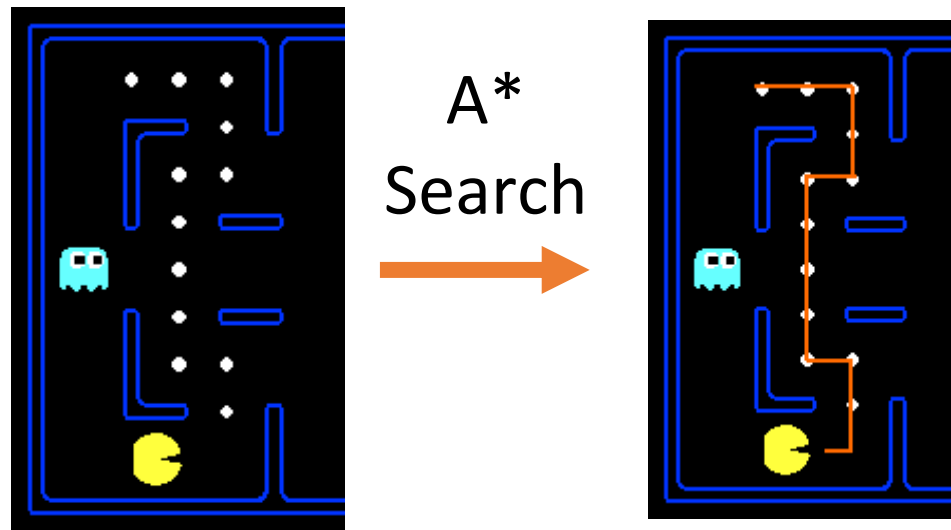
---

By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# How stochasticity affects plans

What happens when we allow for **stochastic** environments instead of **deterministic** environments?



By the end of class today, you will be able to:

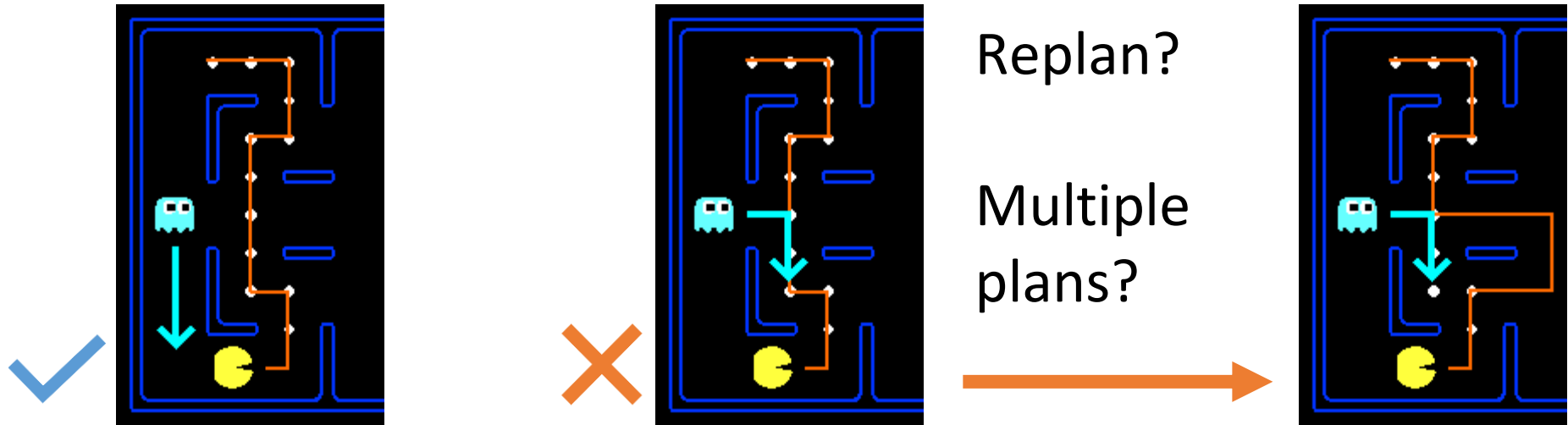
1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# How stochasticity affects plans

What happens when we allow for **stochastic** environments instead of **deterministic** environments?

Pacman example: When we move one grid cell, the ghost will also move one **random** grid cell.

Our plan may be fine, or it may not work!



By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# How stochasticity affects plans

- Re-planning after every action is expensive
- Contingency plans become complicated in highly stochastic environments
- Instead of a **plan** ( $[a_0, a_1, a_2, \dots, a_n]$ ), we need to compute a **policy**

**Policy:** the best action to take, for *every* state in the state space

---

By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# Roadmap for this lecture

We're introducing a lot of new and interrelated concepts:

1. Episodic decision making
2. Sequential decision making
  1. Formalizing the sequential decision making problem: Markov Decision Processes
  2. Deeper look at stochastic actions
  3. Rewards and their effect on policies
  4. Utility for sequential decision making
  5. How to solve Markov Decision Processes (next lecture!)

# Roadmap for this lecture

We're introducing a lot of new and interrelated concepts:

1. Episodic decision making
2. Sequential decision making
  1. Formalizing the sequential decision making problem: Markov Decision Processes
  2. Deeper look at stochastic actions
  3. Rewards and their effect on policies
  4. Utility for sequential decision making
  5. How to solve Markov Decision Processes (next lecture!)



# What would you choose?

Let's consider **episodic** decision making before we look at sequential decision making.

Given the following options, which deal would you take?

1. Receive \$10 right now.
2. Flip a coin. If you call it correctly, you win \$25. Otherwise, you get nothing.



By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# What would you choose?

Let's consider **episodic** decision making before we look at sequential decision making.

Given the following options, which deal would you take?

1. Don't play, receive no prize.
2. Flip a coin. If you call it correctly, you win \$1000. Otherwise, you pay me \$900.



By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# What would you choose?

Let's consider **episodic** decision making before we look at sequential decision making.

Given the following options, which deal would you take?

1. Receive \$100 right now
2. Flip a coin. If you call it correctly, you win \$200. Otherwise, you get nothing.
3. Roll a 6-sided die. If you get a 6, receive \$600. Otherwise, you get nothing.



By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# How should a rational agent choose?

Prompts:

1. Can you explain your process for decision making?
2. What does it mean to be “rational”?

*Potential ideas from the class:*

- *Compute the probabilities, pick an option with favorable odds*
- *Conservative and take options that are safest*
- *Goals matter*
- *Depends on your definition of rational*
- *Have an understanding of laws, rules, penalties, rewards*

---

By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# Episodic decision making

How should a rational agent make decisions?

For each action, calculate the **expected utility**:

$$EU(a | s) = \sum_{s'} P(\text{Result}(a) = s' | s, a) U(s')$$

Expected utility of Action  $a$

Given we are in State  $s$

Average utility over all possible result states  $s'$

Weighted by the probability of reaching each result state

By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# Episodic decision making

**Maximum Expected Utility (MEU):** A rational agent should choose the action that maximizes the agent's **expected utility**

$$action(s) = \underset{a}{\operatorname{argmax}} EU(a | s)$$

---

By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# What would a rational agent choose?

Given the following options, which deal would you take?

1. Receive \$10 right now.
2. **Flip a coin. If you call it correctly, you win \$25. Otherwise, you get nothing.**

$$EU(\textit{option 1}) = (1)(\$10) = \$10$$

$$EU(\textit{option 2}) = (0.5)(\$25) + (0.5)(\$0) = \$12.50$$

---

By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# What would a rational agent choose?

Given the following options, which deal would you take?

1. Don't play, receive no prize.
2. **Flip a coin. If you call it correctly, you win \$1000. Otherwise, you pay me \$900.**

$$EU(\textit{option 1}) = (1)(\$0) = \$0$$

$$EU(\textit{option 2}) = (0.5)(\$1000) + (0.5)(-\$900) = \$50$$

---

By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments



# What would a rational agent choose?

Given the following options, which deal would you take?

1. **Receive \$100 right now**
2. **Flip a coin. If you call it correctly, you win \$200. Otherwise, you get nothing.**
3. **Roll a 6-sided die. If you get a 6, receive \$600. Otherwise, you get nothing.**

$$EU(\text{option 1}) = (1)(\$100) = \$100$$

$$EU(\text{option 2}) = (0.5)(\$200) + (0.5)(\$0) = \$100$$

$$EU(\text{option 3}) = (1/6)(\$600) + (5/6)(\$0) = \$100$$

---

By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

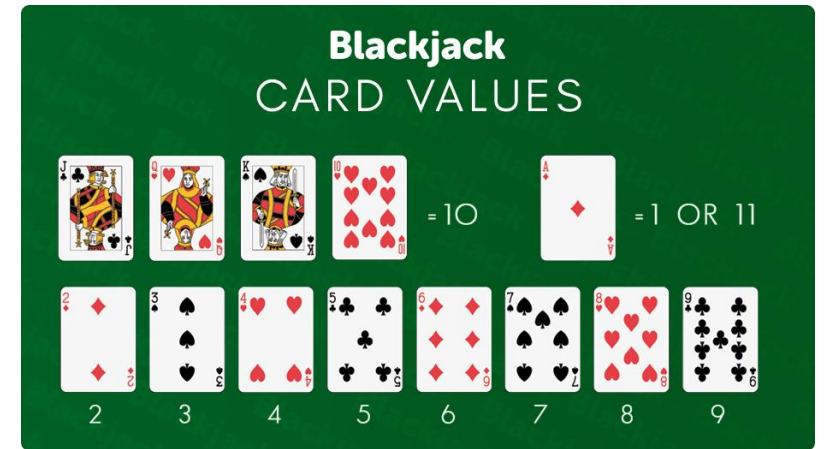
# A More Complex Example: Blackjack

State: sum of values of your cards

Actions:

- Hit – receive a random card to add to your value
- Stay – stop receiving new cards

Goal: Reach a value of 21 without going over

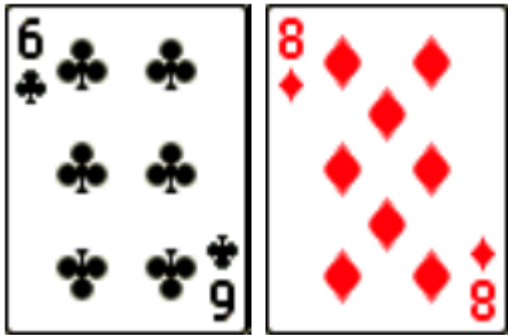


By the end of class today, you will be able to:

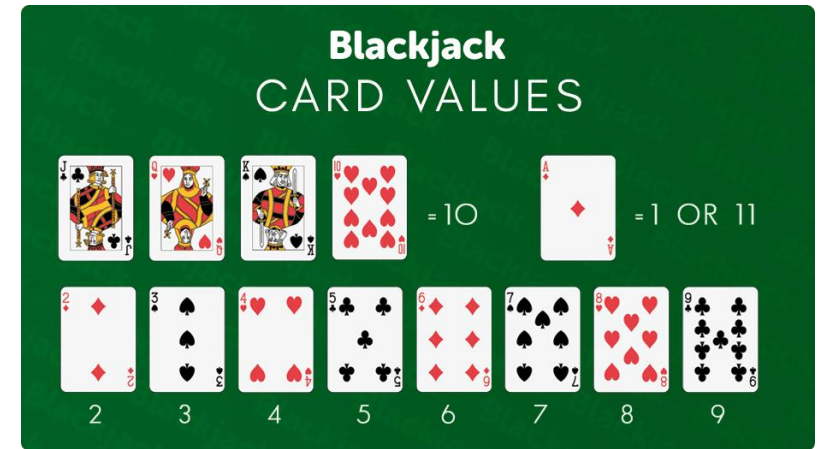
1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# A More Complex Example: Blackjack

Your current hand:



Hit or stay?



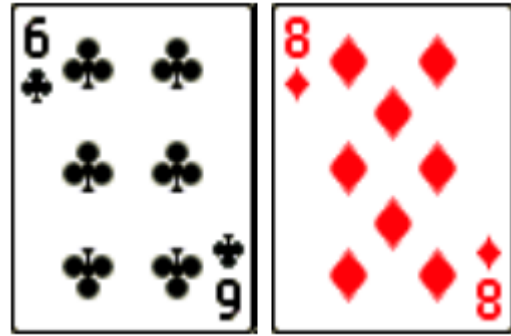
$$\begin{array}{ll} U(< 12) = 0 & U(17) = 7 \\ U(12) = 1 & U(18) = 9 \\ U(13) = 2 & U(19) = 12 \\ U(14) = 3 & U(20) = 16 \\ U(15) = 4 & U(21) = 25 \\ U(16) = 5 & U(> 21) = -100 \end{array}$$

By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# A More Complex Example: Blackjack

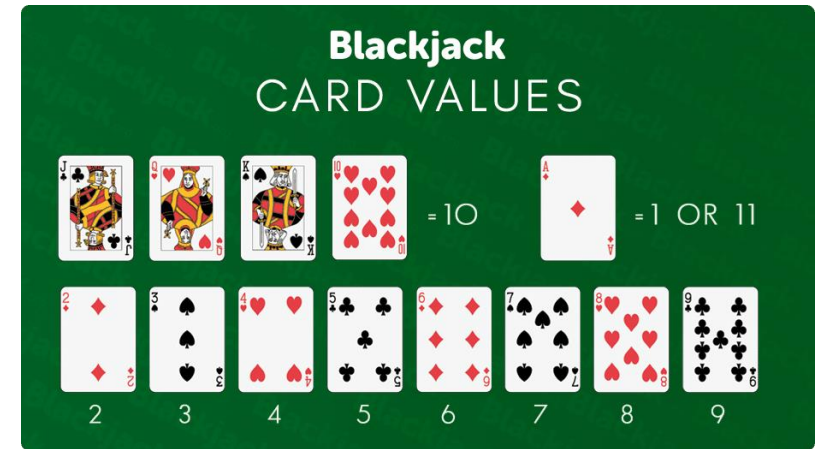
Your current hand:



Hit or stay?

$$U(\textit{stay}) = (1)(U(14)) = 3$$

$$\begin{aligned}
 U(\textit{hit}) &= \frac{4}{50} U(15) + \frac{4}{50} U(16) + \frac{4}{50} U(17) \\
 &+ \frac{4}{50} U(18) + \frac{4}{50} U(19) + \frac{3}{50} U(20) + \frac{4}{50} U(21) \\
 &+ \frac{23}{50} U(> 21) = 5.92
 \end{aligned}$$

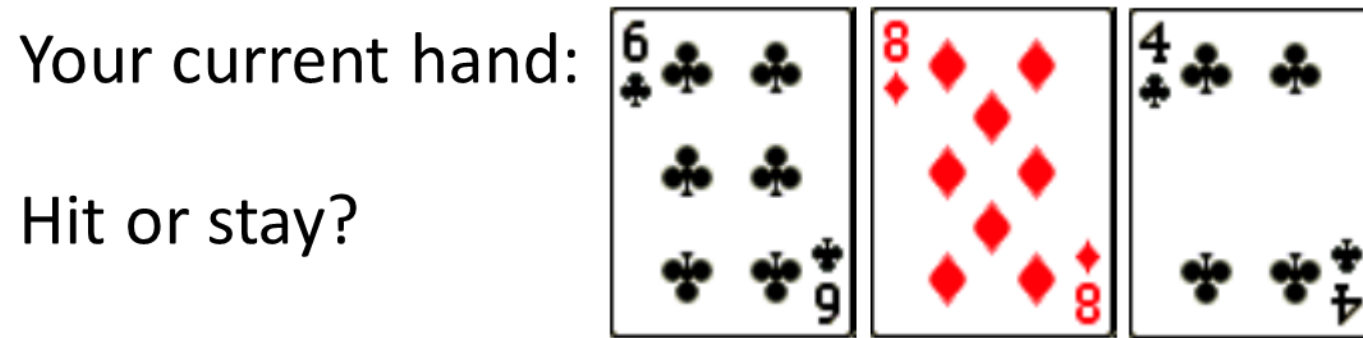


$U(< 12) = 0$	$U(17) = 7$
$U(12) = 1$	$U(18) = 9$
$U(13) = 2$	$U(19) = 12$
$U(14) = 3$	$U(20) = 16$
$U(15) = 4$	$U(21) = 25$
$U(16) = 5$	$U(> 21) = 0$

By the end of class today, you will be able to:

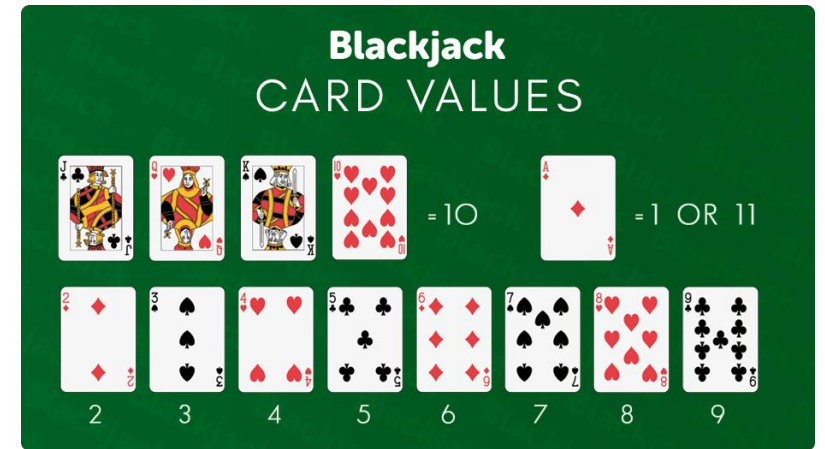
1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# A More Complex Example: Blackjack



$$U(\textit{stay}) = (1)(U(18)) = 9$$

$$U(\textit{hit}) = \frac{4}{49}U(19) + \frac{4}{49}U(20) + \frac{4}{49}U(21) + \frac{37}{49}U(> 21) = 4.33$$



$U(< 12) = 0$	$U(17) = 7$
$U(12) = 1$	$U(18) = 9$
$U(13) = 2$	$U(19) = 12$
$U(14) = 3$	$U(20) = 16$
$U(15) = 4$	$U(21) = 25$
$U(16) = 5$	$U(> 21) = 0$

By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# Applying MEU to Episodic Decisions

- Buy a lottery ticket: costs \$2, 1/292,201,338 chance to win \$75,000,000

$$EU = \frac{1}{292201338} (75000000 - 2) + \frac{292201337}{292201338} (-2) = -1.74$$

- Buy a scratchoff ticket: costs \$1, 1/10 chance to win \$5

$$EU = \frac{1}{10} (5 - 1) + \frac{9}{10} (-1) = -0.50$$

(Note: Example taken from real odds and amounts from various US lottery games in 2020)

- Keep your money: costs \$0, guaranteed chance to win nothing

$$EU = 1(0) = 0.00$$

By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# Applying MEU to Episodic Decisions

- Cook dinner: Takes time and effort (-10 utility), 80% chance food will be good (+50 utility), 20% chance food will be bad (+10 utility)

$$EU = 0.8(50 - 10) + 0.2(10 - 10) = 32$$

- Order sushi: Costs money (-5 utility), 95% chance food will be good (+50 utility), 5% chance food will be bad (-10 utility)

$$EU = 0.95(50 - 5) + 0.05(-10 - 5) = 42$$

By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# Roadmap for this lecture

We're introducing a lot of new and interrelated concepts:

1. Episodic decision making
2. Sequential decision making
  1. Formalizing the sequential decision making problem: Markov Decision Processes
  2. Deeper look at stochastic actions
  3. Rewards and their effect on policies
  4. Utility for sequential decision making
  5. How to solve Markov Decision Processes (next lecture!)



# MEU in Sequential Decision Making

How would we extend the principle of Maximum Expected Utility to **sequential** environments?

- (Reminder: Our decisions now will affect our decisions in the future)
- How do we define utility in this case? Depends on:
  - Which state we're currently in
  - Which states our actions can lead to
  - How much we value individual states
  - The future actions we'll take
  - How many actions we'll be taking in the future
  - ...
- Less straightforward than the **episodic** case. We need to formalize a way to approach this problem!

---

By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# Roadmap for this lecture

We're introducing a lot of new and interrelated concepts:

1. Episodic decision making
2. Sequential decision making
  1. Formalizing the sequential decision making problem: Markov Decision Processes
  2. Deeper look at stochastic actions
  3. Rewards and their effect on policies
  4. Utility for sequential decision making
  5. How to solve Markov Decision Processes (next lecture!)

# Markov Decision Processes (MDPs)

A **problem description** for **sequential** decision problems in **fully observable, stochastic** environments with **Markovian transition functions**

- (More on this later)
- 

Solution is a **policy** that *maximizes expected utility*

---

By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# Problem Definition: MDP

Components of an MDP:  $(S, A(s), T(s, a, s'), R(s), \gamma)$

- **State space** ( $S$ ): all states the agent can be in, with **initial state**  $s_0$
- **Actions** ( $A(s)$ ): set of applicable actions agent can execute in each state  $s$
- **Transition model** ( $T(s, a, s')$ ): function describing how actions change states; modeled as probability distribution  $P(s' | s, a)$ 
  - What states *can* be reached when taking action  $a$  in state  $s$ ?
- **Reward function** ( $R(s)$ ): the value of being in state  $s$ 
  - Often defined by designer's intuition
- **Discount factor** ( $\gamma$ ): how much to prioritize current rewards over future rewards,  $0 \leq \gamma \leq 1$

---

By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# Problem Definition: MDP

Compare to Search  
problem definition:

Same definition

Adapted for  
stochastic actions

Replaces goal test  
and step cost

Components of an MDP:  $(S, A(s), T(s, a, s'), R(s), \gamma)$

- **State space** ( $S$ ): all states the agent can be in, with **initial state**  $s_0$
- **Actions** ( $A(s)$ ): set of applicable actions agent can execute in each state  $s$
- **Transition model** ( $T(s, a, s')$ ): function describing how actions change states; modeled as probability distribution  $P(s' | s, a)$ 
  - What states *can* be reached when taking action  $a$  in state  $s$ ?
- **Reward function** ( $R(s)$ ): the value of being in state  $s$ 
  - Often defined by designer's intuition
- **Discount factor** ( $\gamma$ ): how much to prioritize current rewards over future rewards

---

By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# Problem Definition: MDP

Solution to an MDP: a *policy*

According to Maximum Expected Utility!

**Policy:** the **best action** to take, for *every* state in the state space;  
a mapping from **states** to **actions**

$$\pi: S \rightarrow A, \quad \pi(s) = a$$

---

By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

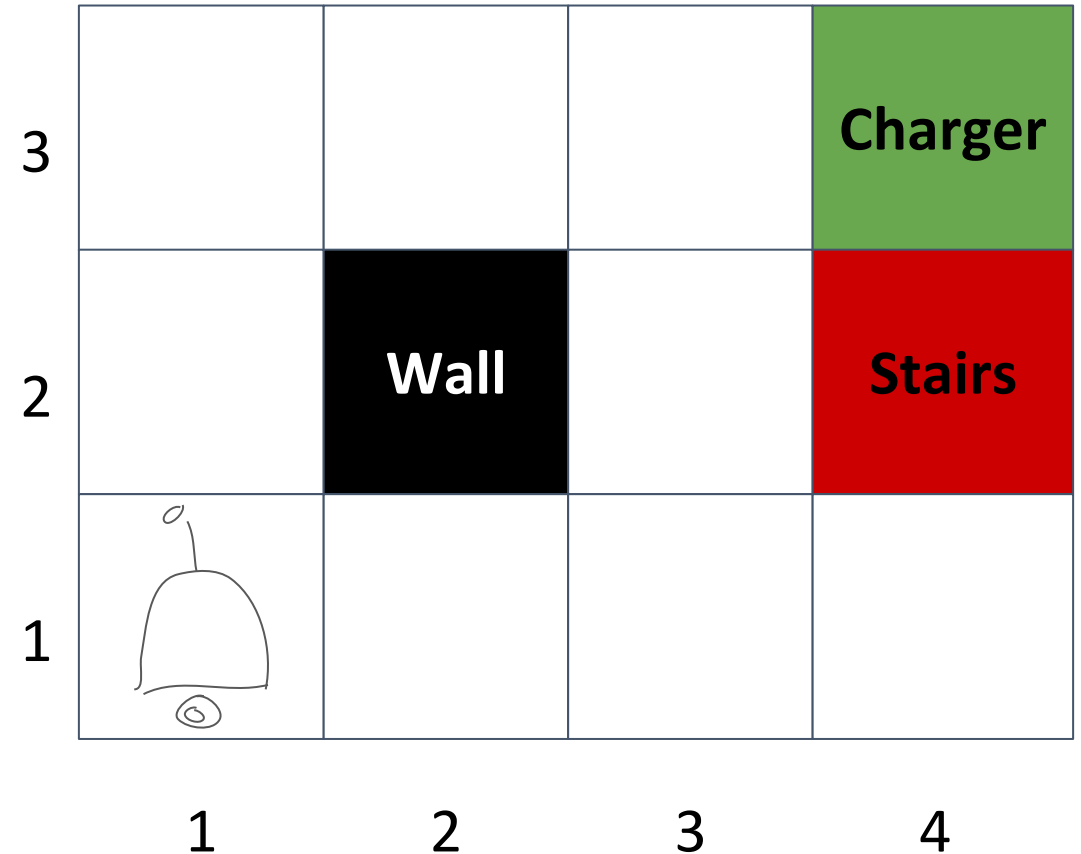
# Example Environment: Robot Gridworld

**State:** Robot can be in any of the non-wall cells  $(x,y)$

**Actions:** Up, Down, Left, Right. (moving into wall or out of bounds  $\rightarrow$  stays put)

**Transition model:** Actions have a chance of moving the wrong way

**Robot's goal:** Low battery, get to the charger  $(4,3)$ ! Avoid falling down the stairs  $(4,2)$ !



By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# Roadmap for this lecture

We're introducing a lot of new and interrelated concepts:

1. Episodic decision making
2. **Sequential decision making**
  1. Formalizing the sequential decision making problem: Markov Decision Processes
  2. **Deeper look at stochastic actions**
  3. Rewards and their effect on policies
  4. Utility for sequential decision making
  5. How to solve Markov Decision Processes (next lecture!)



# Describing the Environment: Determinism

## Deterministic

- Environment changes in exactly one way as a result of an agent's action

## Stochastic

- Randomness
- Action uncertainty
- Partial-observability

---

By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# Describing the Environment: Determinism

## Deterministic

- Environment changes in exactly one way as a result of an agent's action

## Stochastic

- **Randomness**
- **Action uncertainty**
- Partial-observability

---

By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# Robot Gridworld Transition Model

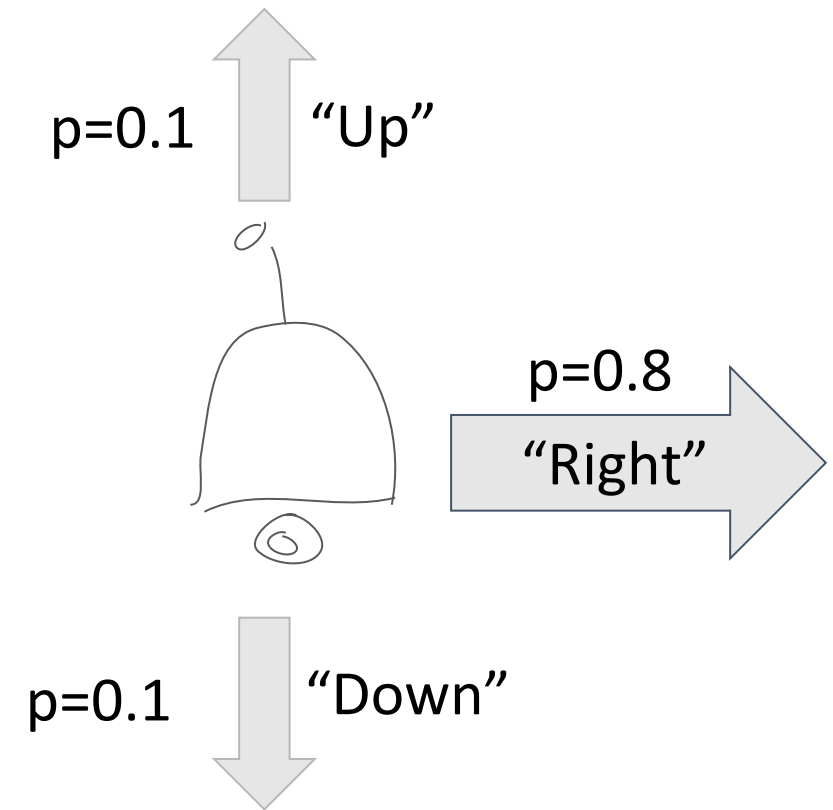
Let's be precise about "chance of moving the wrong way"

There is an 80% chance of moving in the intended direction. The remaining 20% is split evenly between the two orthogonal directions.

## Notation

Probability of transitioning from  $s$  to  $s'$  with action  $a$

$$P(s' | s, a)$$



By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# Robot Gridworld Transition Model

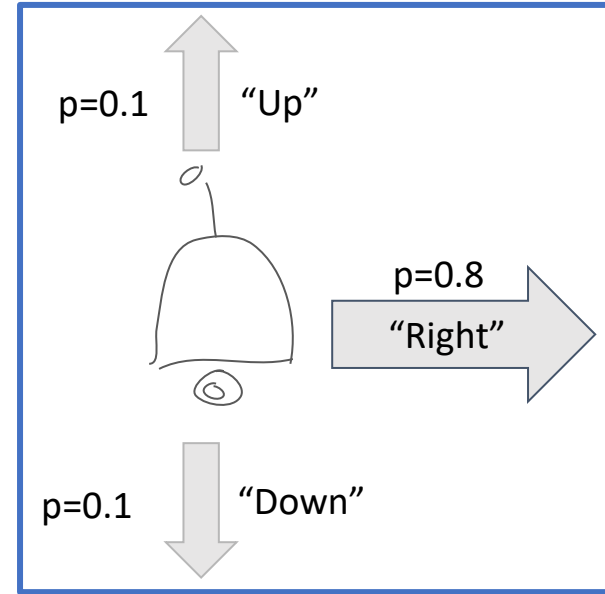
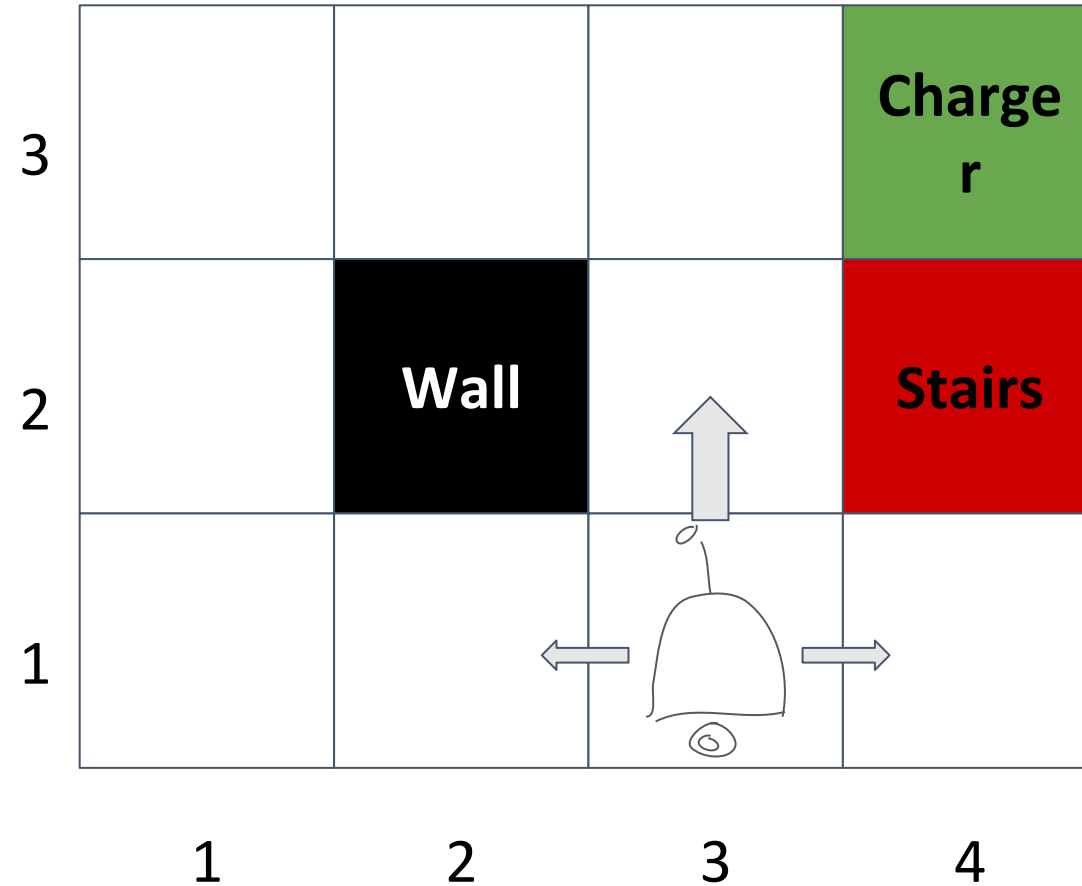
“Up” in (3,1) →  
 (3,2) w/  $p=0.8$   
 (2,1) w/  $p=0.1$   
 (4,1) w/  $p=0.1$

$T(s, a, s')$

$$T((3, 1), \text{Up}, (3, 2)) = 0.8$$

$$T((3, 1), \text{Up}, (2, 1)) = 0.1$$

$$T((3, 1), \text{Up}, (4, 1)) = 0.1$$



By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

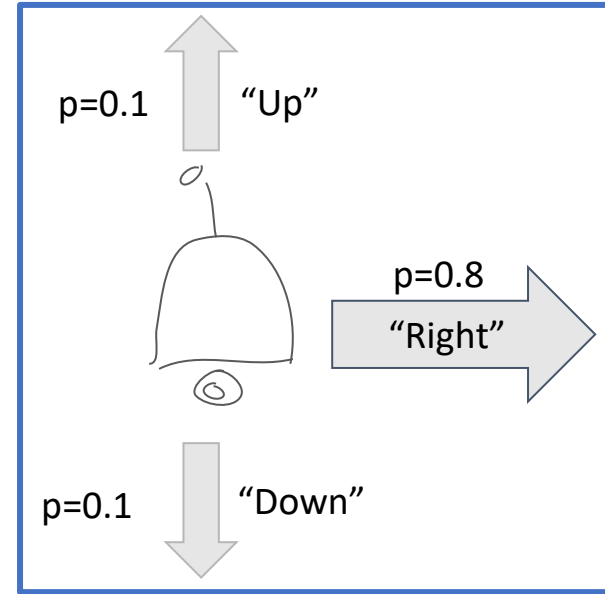
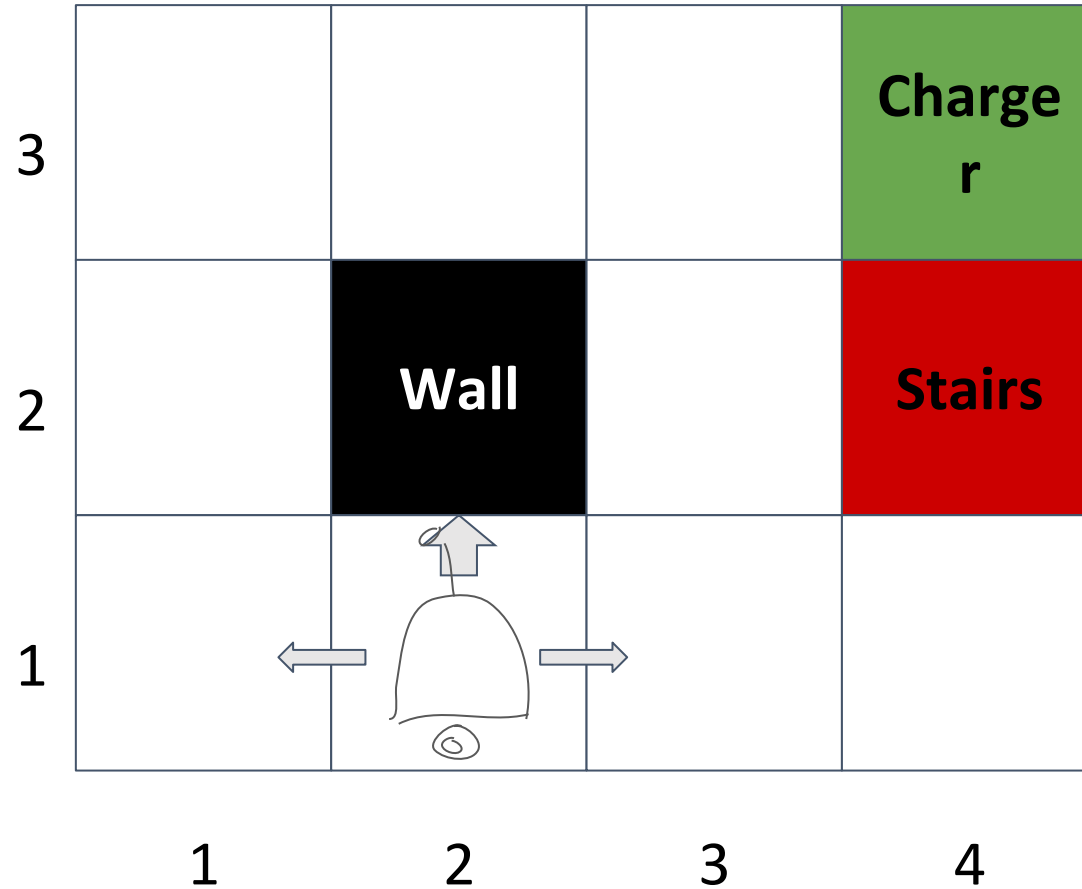
# Robot Gridworld Transition Model

“Up” in (2,1) →  
 (2,1) w/  $p=0.8$   
 (1,1) w/  $p=0.1$   
 (3,1) w/  $p=0.1$

$$T((2, 1), U_p, (2, 1)) = 0.8$$

$$T((2, 1), U_p, (1, 1)) = 0.1$$

$$T((2, 1), U_p, (3, 1)) = 0.1$$



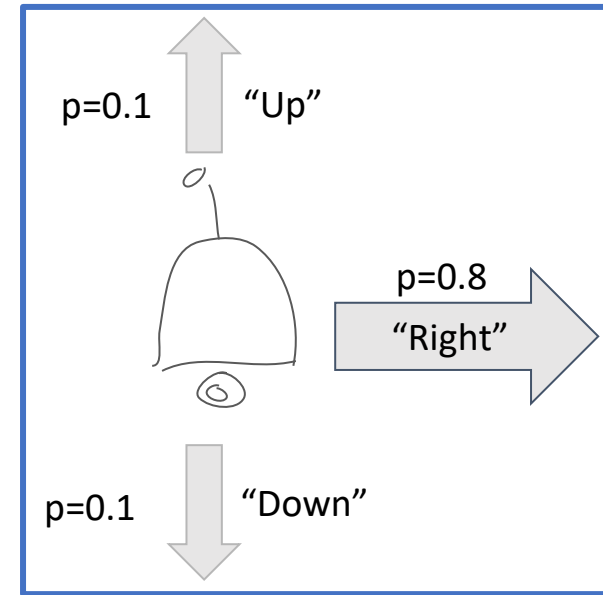
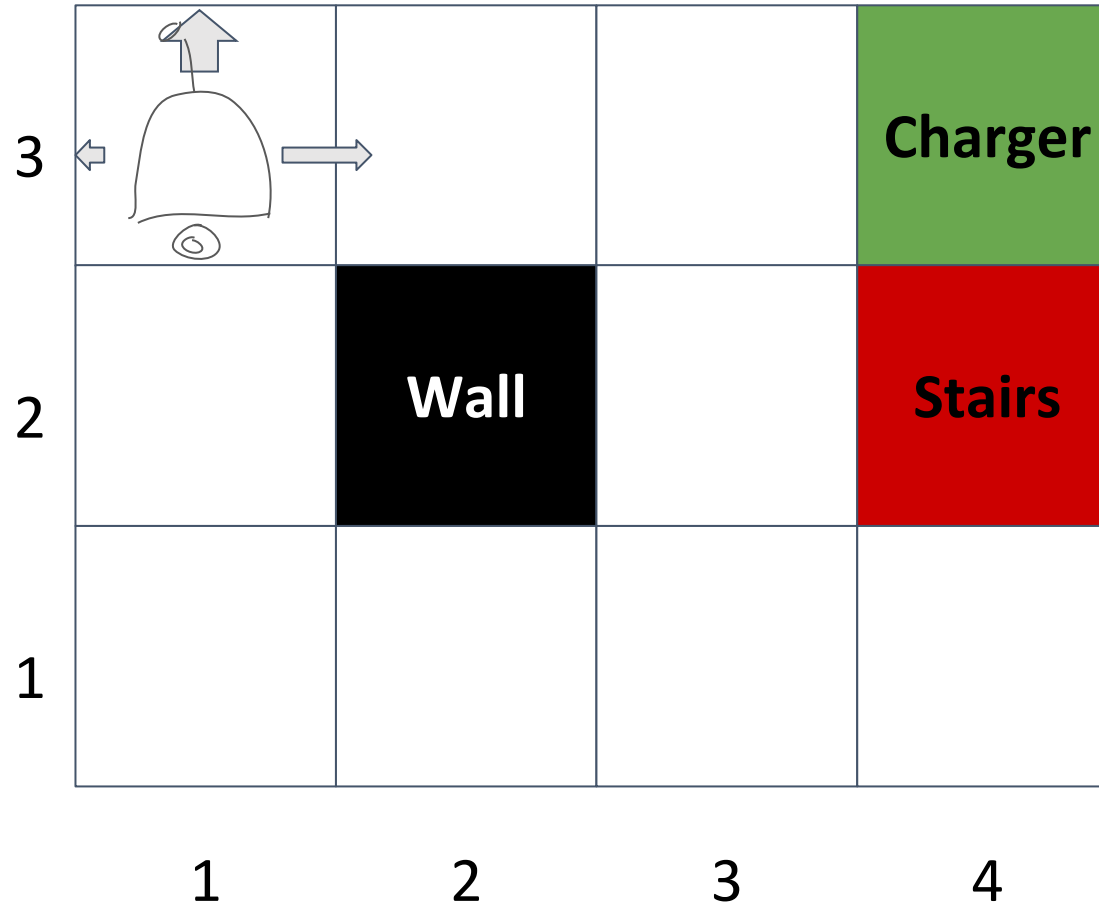
By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# Robot Gridworld Transition Model

“Up” in (1,3) →  
(1,3) w/  $p=0.9$   
(2,3) w/  $p=0.1$

$$T((1, 3), U_p, (1, 3)) = 0.9$$
$$T((1, 3), U_p, (2, 3)) = 0.1$$



By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# Robot Gridworld Planning Solution

What's the planning solution for getting to the goal?

**Deterministic version:**

[U,U,R,R,R] or [R,R,U,U,R]

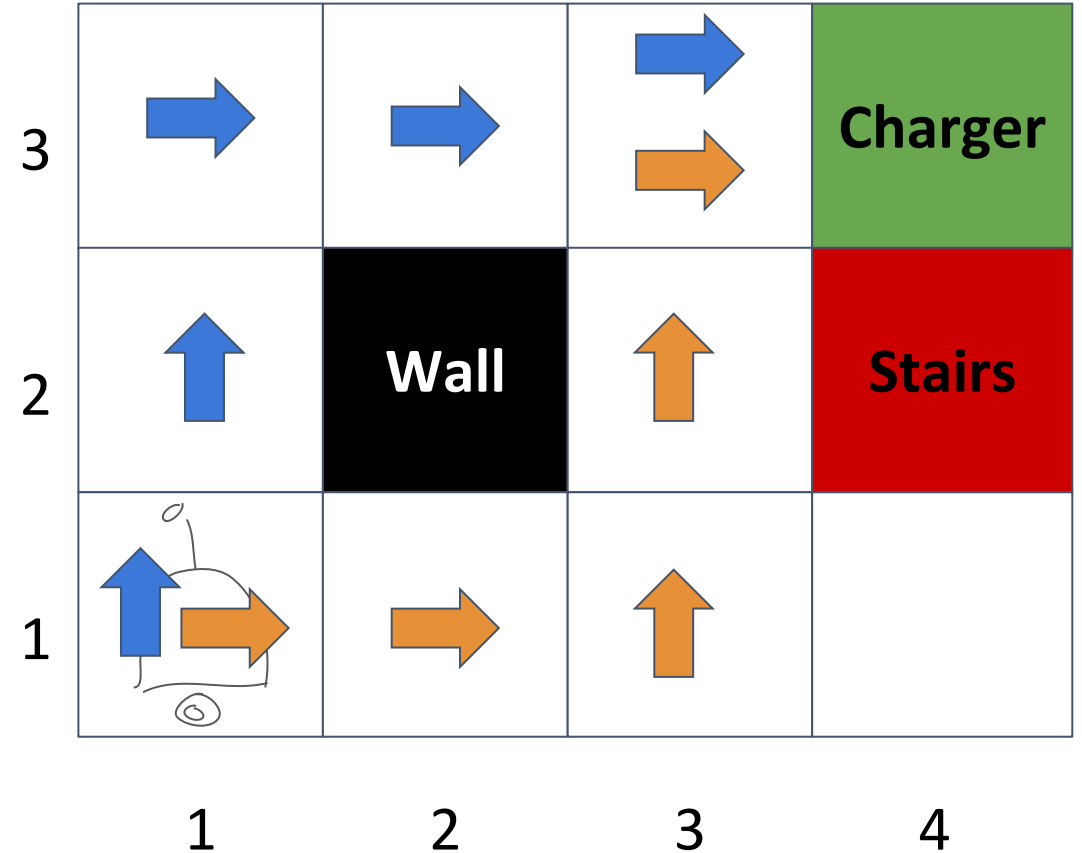
Probability of success?

$$(0.8) * (0.8) * (0.8) * (0.8) * (0.8) = 32\%$$

Probability of success (by accident)?

$$(0.1) * (0.1) * (0.1) * (0.1) * (0.8) = 0.008\%$$

Small chance of success!



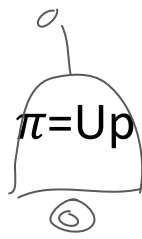
By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# Robot Gridworld Policy Solution

What's an effective **policy** for getting to the charger?

The “best” policy is going to depend on our **performance measure** which we encode as a **reward function**

3	$\pi$ =Right	$\pi$ =Right	$\pi$ =Right	Charger
2	$\pi$ =Up	Wall	$\pi$ =Up	Stairs
1	 $\pi$ =Up	$\pi$ =Right	$\pi$ =Up	$\pi$ =Left
	1	2	3	4

By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments



# Roadmap for this lecture

We're introducing a lot of new and interrelated concepts:

1. Episodic decision making
2. **Sequential decision making**
  1. Formalizing the sequential decision making problem: Markov Decision Processes
  2. Deeper look at stochastic actions
  3. **Rewards and their effect on policies**
  4. Utility for sequential decision making
  5. How to solve Markov Decision Processes (next lecture!)

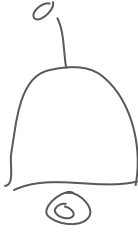
# Robot Gridworld Reward Example

Reward  $R(s)$  is a mapping from states to real numbers

$$R(s): S \rightarrow \mathbb{R}$$

Often defined by designer's intuition, we'll base rewards off of **robot's goal**

- $R(\text{charger}) = +1$
- $R(\text{stairs}) = -1$
- $R(\text{other states}) = 0$ 
  - To get a faster robot (i.e. one with less starting battery), set the reward at other states to a negative value

3	R=0	R=0	R=0	R=+1
2	R=0	Wall	R=0	R=-1
1		R=0	R=0	R=0
	1	2	3	4

By the end of class today, you will be able to:

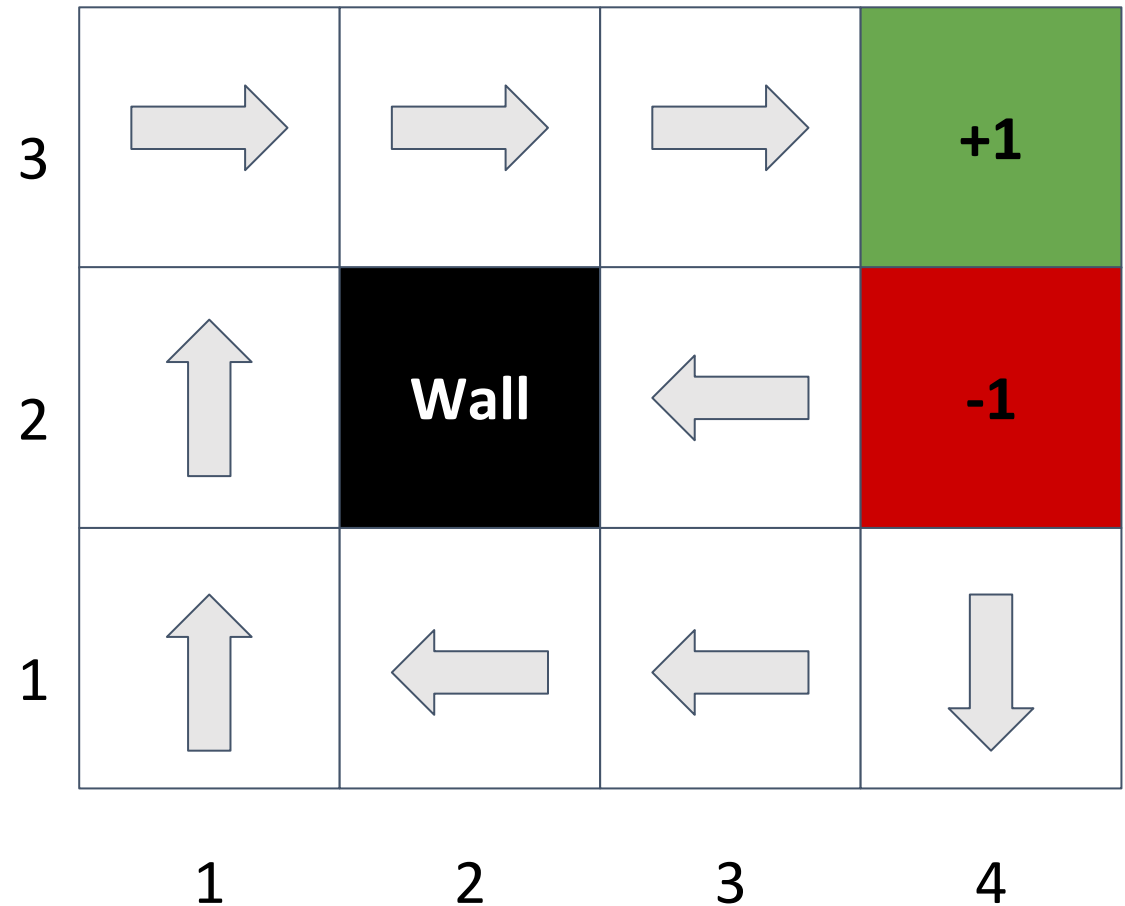
1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# Robot Gridworld Reward Example

Our reward choice matters!  
Different rewards give different  
“best” policies:

Conservative version:

- Keep  $R(s)$  of red and green fixed
- For every other state:
  - $-0.0221 < R(s) < 0$



By the end of class today, you will be able to:

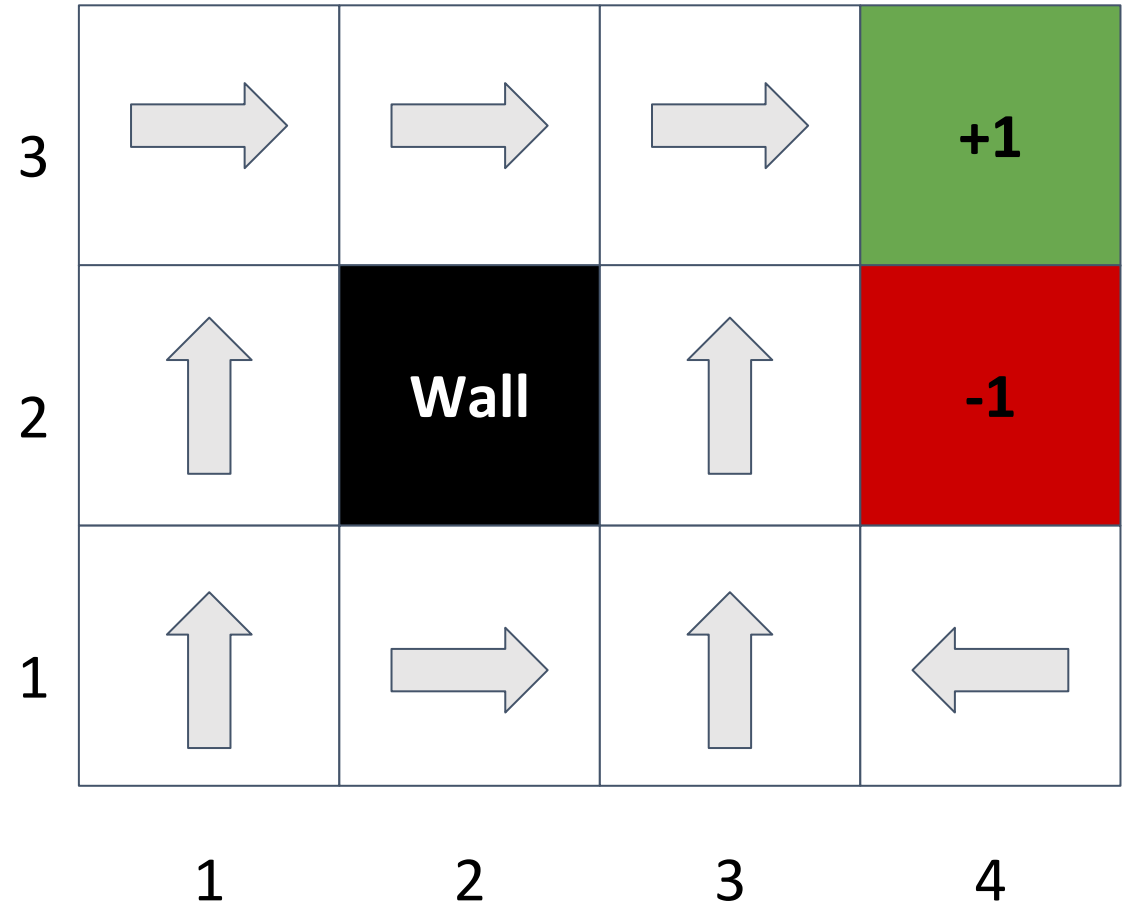
1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# Robot Gridworld Reward Example

Our reward choice matters!  
Different rewards give different  
“best” policies:

Speedy (risky) version:

- Keep  $R(s)$  of red and green fixed
- For every other state:
  - $-0.4278 < R(s) < -0.085$



By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# Finding the Best Policy

How do we determine what the best policy is?

Returning to the principle of maximum expected utility, define the optimal policy,  $\pi^*(s)$ , as:

$$\pi^*(s) = \operatorname{argmax}_a EU(a | s) \quad \leftarrow \text{MEU (general)}$$

$$\pi^*(s) = \operatorname{argmax}_a \sum_{s'} T(s, a, s') U(s') \quad \leftarrow \text{MEU for an MDP}$$

**We need to define utility for sequential environments!**

By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# Roadmap for this lecture

We're introducing a lot of new and interrelated concepts:

1. Episodic decision making
2. **Sequential decision making**
  1. Formalizing the sequential decision making problem: Markov Decision Processes
  2. Deeper look at stochastic actions
  3. Rewards and their effect on policies
  4. **Utility for sequential decision making**
  5. How to solve Markov Decision Processes (next lecture!)

# Utility for Sequential Environments

**Utility** (sequential): indication of how good a state is with regard to future possible states

We can calculate this in two ways: Additive utility or discounted utility

---

By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# Utility for Sequential Environments

**Utility** (sequential): indication of how good a state is with regard to future possible states


We can calculate this over a *state sequence* one of two ways:

- Additive Rewards:

$$U([s_0, s_1, s_2, \dots]) = R(s_0) + R(s_1) + R(s_2) + \dots$$

- Discounted Rewards:

Discount factor from MDP definition

$$U([s_0, s_1, s_2, \dots]) = R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \dots$$


---

By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments



# Comparing Additive and Discounted Rewards

- Additive Rewards:

$$U([s_0, s_1, s_2, \dots]) = R(s_0) + R(s_1) + R(s_2) + \dots$$

- Prioritize future rewards equally with current reward
- Special case of Discounted Rewards when  $\gamma = 1$
- Can **not** handle infinite state sequences

- Discounted Rewards:

$$U([s_0, s_1, s_2, \dots]) = R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \dots$$

- Discount the value of rewards in the future
- $0 \leq \gamma \leq 1$
- **Can** handle infinite state sequences – infinite sum will converge

By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# Utility for Sequential Environments

Discounted rewards let us handle **infinite horizon** sequential decision problems

**Finite horizon:** agent has a fixed amount of time to take actions, after which no rewards can be accumulated

**Infinite horizon:** agent has an infinite amount of time to accumulate rewards

- Problems can still terminate (e.g. robot gridworld ends when robot reaches the charger or falls down the stairs)
- Starting state  $s_0$  does not change the optimal policy

---

By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# Utility for Sequential Environments

**Utility** (sequential): indication of how good a state is with regard to future possible states

Utility for a state, with discounted rewards:

$$U(s) = E \left[ \sum_{t=0}^{\infty} \gamma^t R(S_t) \right]$$

Diagram illustrating the components of the utility equation:

- $U(s)$ : Expected utility
- $E$ : Expected utility
- $\sum_{t=0}^{\infty}$ : Infinite horizon
- $\gamma^t R(S_t)$ : Discounted rewards
- $R(S_t)$ : Stochastic actions (random variable)

By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# Finding the Best Policy

How do we determine what the best policy is?

Returning to the principle of maximum expected utility, define the optimal policy,  $\pi^*(s)$ , as:

$$\pi^*(s) = \operatorname{argmax}_a EU(a | s)$$

$$\pi^*(s) = \operatorname{argmax}_a \sum_{s'} T(s, a, s') U(s'),$$

$$U(s) = E \left[ \sum_{t=0}^{\infty} \gamma^t R(S_t) \right]$$

---

By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# Utility vs. Reward

**Reward:** indication of how good it is to be in state  $s$ , in the *present*)

**Utility:** indication of how good a state is with regard to *future possible states*

**Reward** and **utility** can be very different in sequential environments!

---

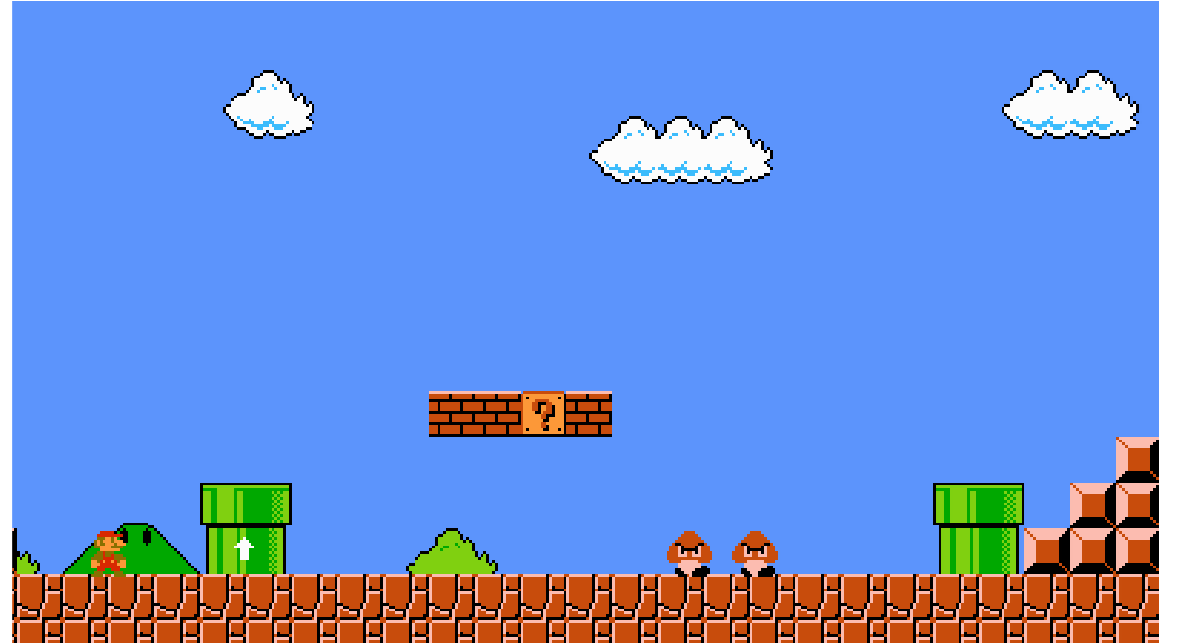
By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# Utility vs. Reward

Example from Mario:

- Hitting [?] blocks from below gives points (+ reward)
- Finishing the level to the right gives points (+ reward)
- Hitting a goomba (little mushroom guys) loses a life (- reward)



By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

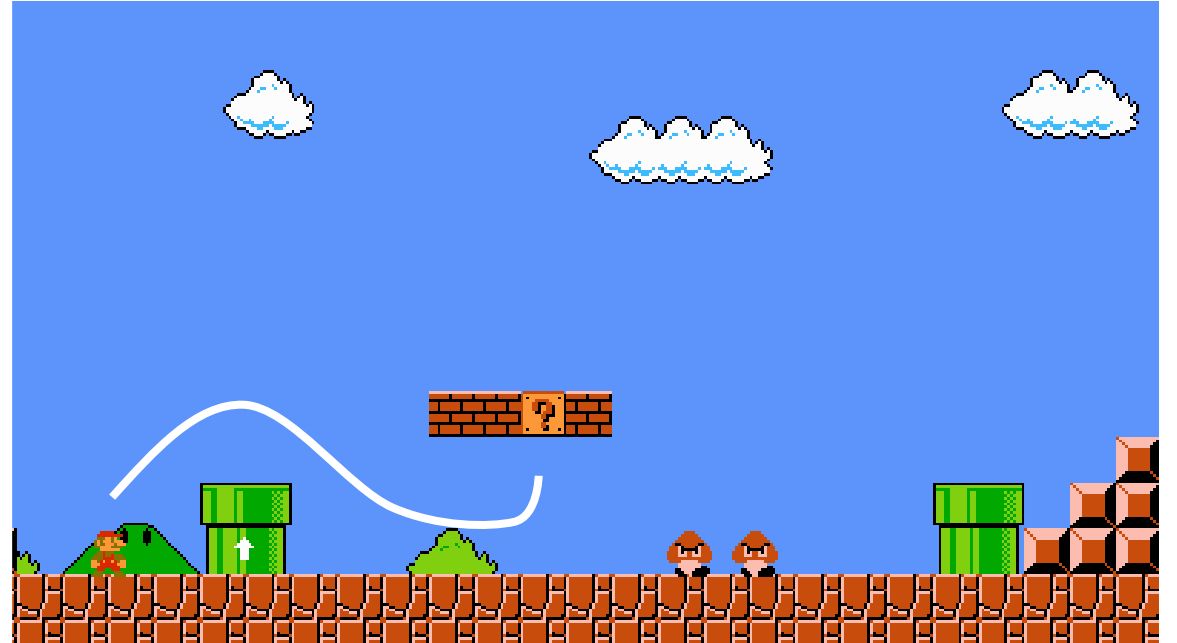
# Utility vs. Reward

Example from Mario:

- Hitting [?] blocks from below gives points (+ **small reward**)
- Finishing the level to the right gives points (+ **large reward**)
- Hitting a goomba (little mushroom guys) loses a life (- **large reward**)

What's the **reward** and **utility** for taking these actions?

**Higher Reward, Lower Utility**



By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

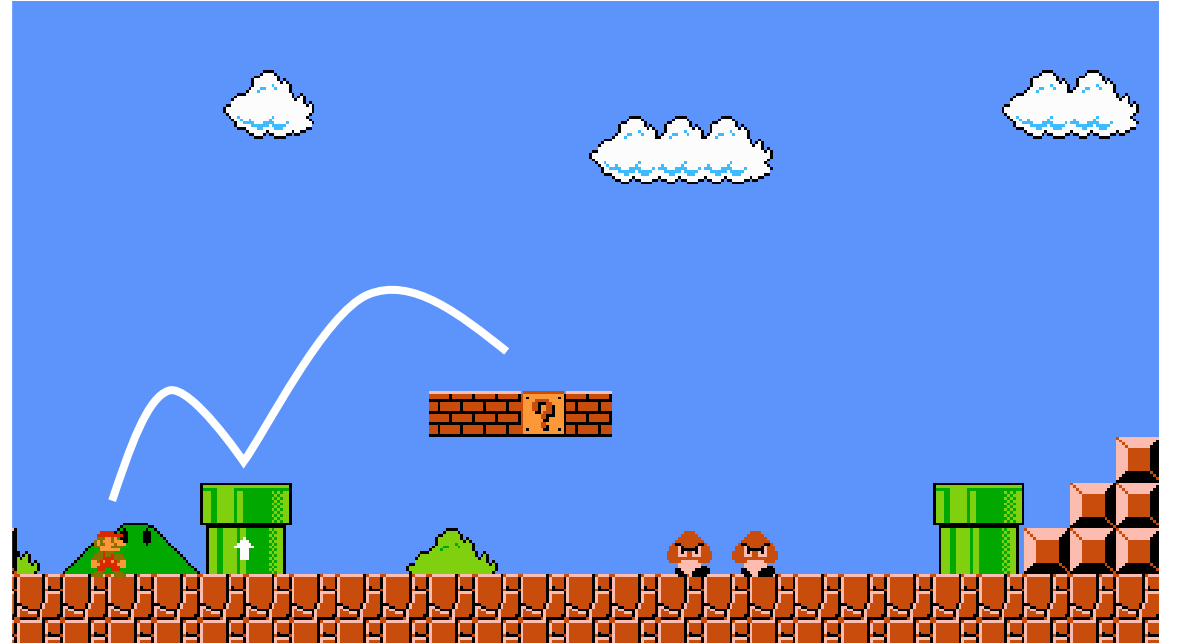
# Utility vs. Reward

Example from Mario:

- Hitting [?] blocks from below gives points (+ **small reward**)
- Finishing the level to the right gives points (+ **large reward**)
- Hitting a goomba (little mushroom guys) loses a life (- **large reward**)

What's the **reward** and **utility** for taking these actions?

**Zero Reward, Higher Utility**

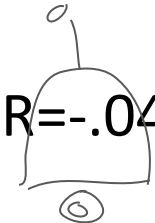


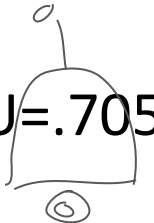
By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments



# Utility vs. Reward

3	R=-.04	R=-.04	R=-.04	R=+1
2	R=-.04	Wall	R=-.04	R=-1
1	 R=-.04	R=-.04	R=-.04	R=-.04
	1	2	3	4

3	U=.812	U=.868	U=.918	U=+1
2	U=.762	Wall	U=.660	U=-1
1	 U=.705	U=.655	U=.611	U=.388
	1	2	3	4

By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# Roadmap for this lecture

We're introducing a lot of new and interrelated concepts:

1. Episodic decision making
2. **Sequential decision making**
  1. Formalizing the sequential decision making problem: Markov Decision Processes
  2. Deeper look at stochastic actions
  3. Rewards and their effect on policies
  4. Utility for sequential decision making
  5. **How to solve Markov Decision Processes (next lecture!)**

# Next Class

How do we actually calculate the optimal policy?

...We'll be using these equations,

$$\pi^*(s) = \operatorname{argmax}_a \sum_{s'} T(s, a, s') U(s'),$$
$$U(s) = E \left[ \sum_{t=0}^{\infty} \gamma^t R(S_t) \right],$$

thanks to the **Markov assumption!**

By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments

# Next Class

Where do reward functions come from? Is there a danger to leaving them to a designer to define based only on their intuition?

We'll discuss the **value alignment** problem.

It looks like you're trying to optimize arbitrary performance criteria. Have you accounted for implicit human values?

- Yes
- No
- Don't show this again*



By the end of class today, you will be able to:

1. Define and compare plans and policies
2. Determine the best action to take using the principle of Maximum Expected Utility
3. Formally define sequential decision making problems using Markov Decision Processes
4. Distinguish between reward and utility in sequential environments