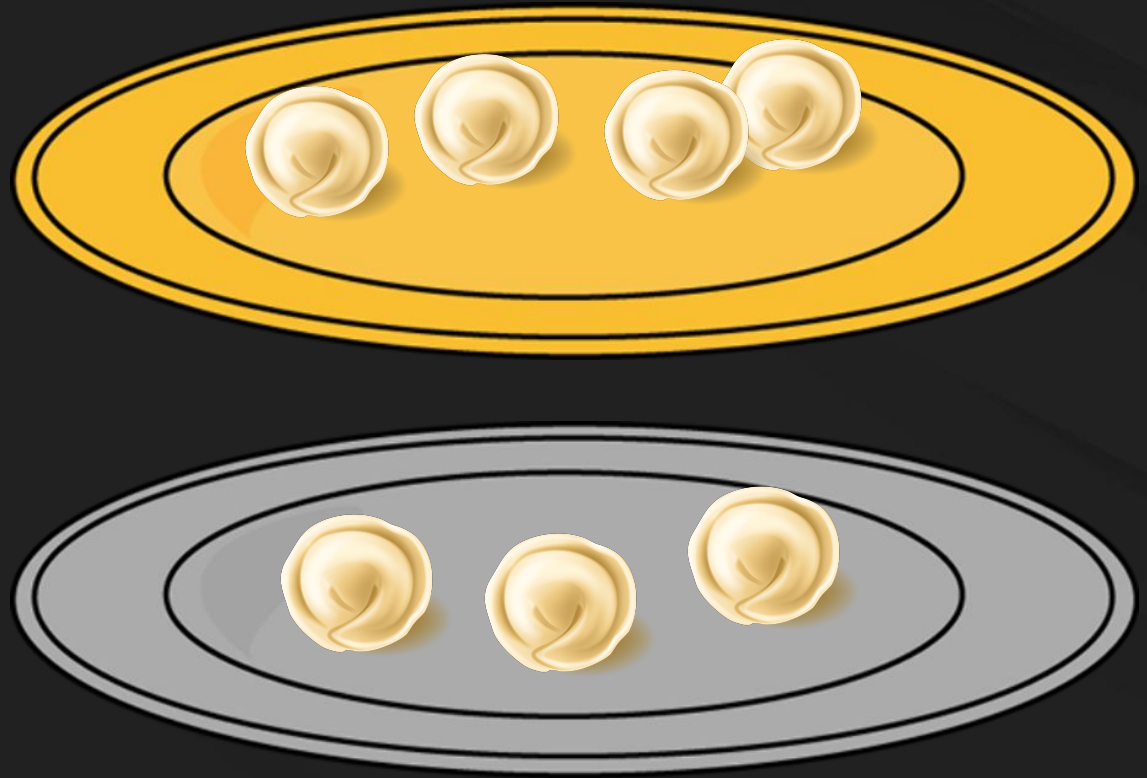
RL–02

Ключевые понятия

Начнем в 20:01

otus.ru



Тема вебинара

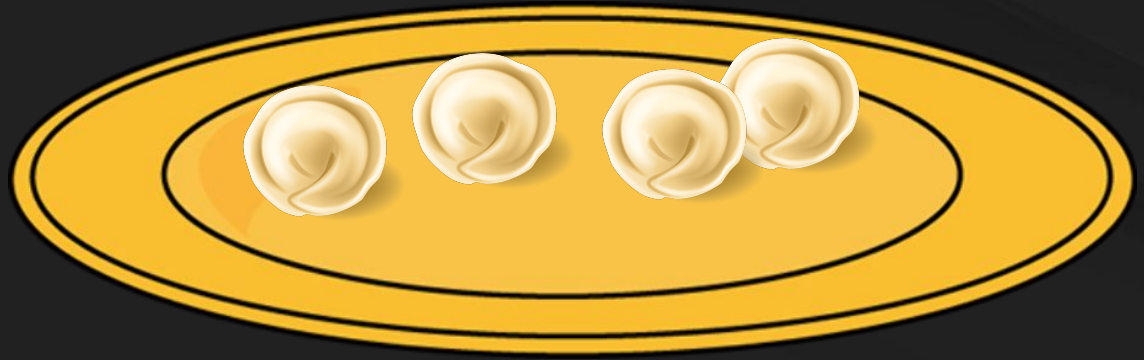
Базовые понятия
reinforcement learning

Катя

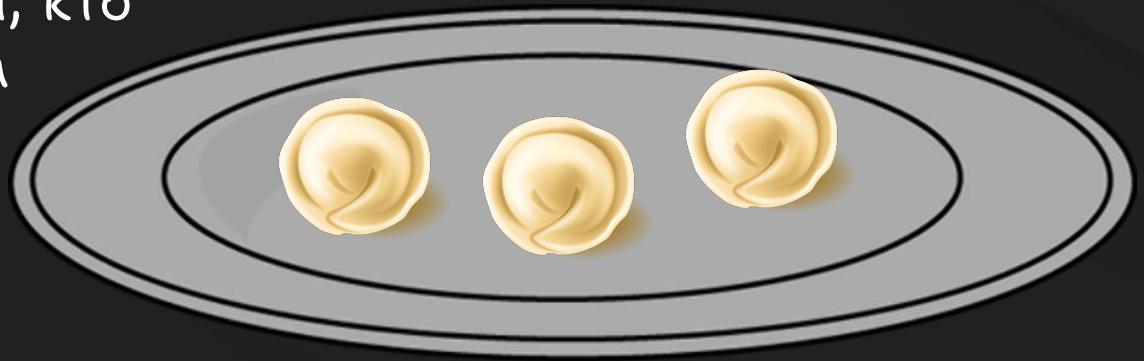
Екатерина Дмитриева
Telegram: @dmi3eva

За раз можно съесть любое число пельменей

... но только из одной тарелки

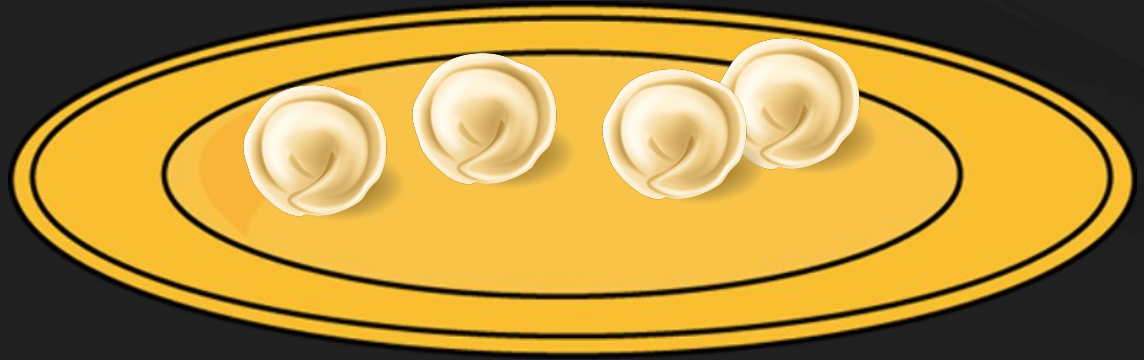


Выигрывает тот, кто съест последний пельмень

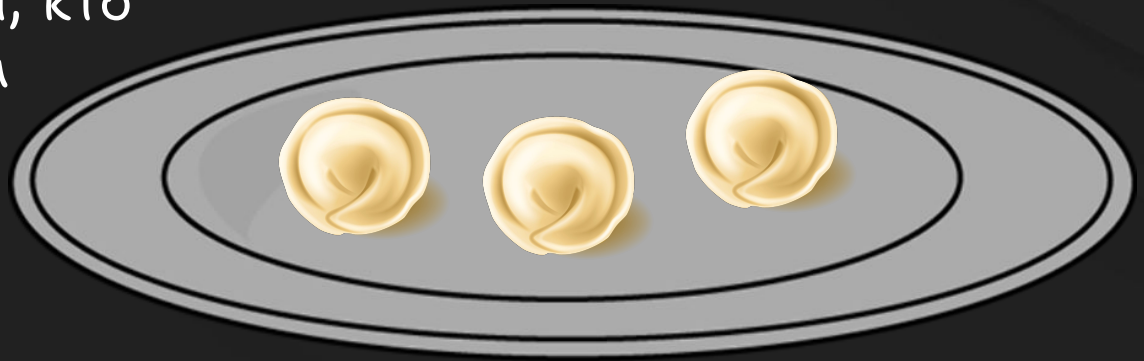


За раз можно съесть любое число пельменей

... но только из одной тарелки

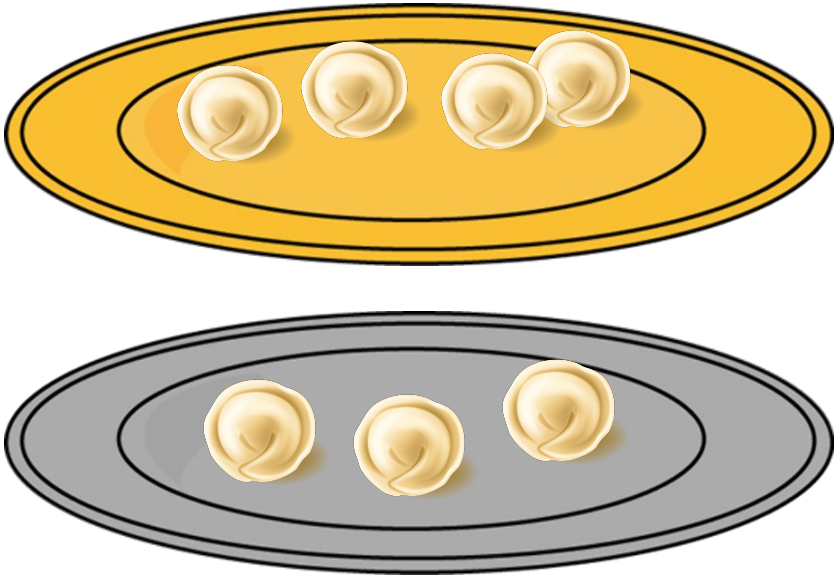


Выигрывает тот, кто съест последний пельмень



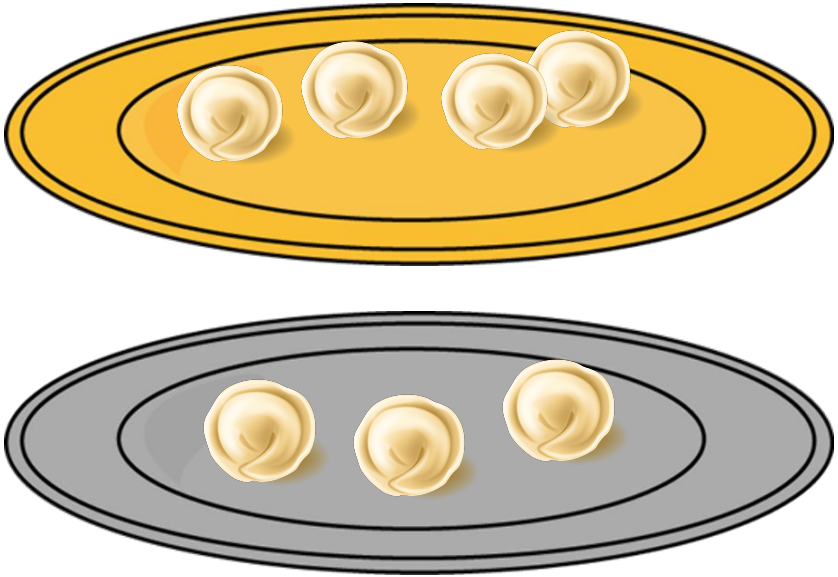
Выигрышная стратегия

- Выигрышная стратегия есть у первого игрока:
  - **Ход 1**: Съесть 1 оранжевый
  - **Ход 2:** Повторять ходы второго игрока в другой тарелке



Выигрышная стратегия

- Выигрышная стратегия есть у первого игрока:
  - **Ход 1**: Съесть 1 оранжевый
  - **Ход 2:** Повторять ходы второго игрока в другой тарелке



В чем суть метода RL

- **Дискретное программирование:** аналитически найти оптимальную стратегию и реализовать

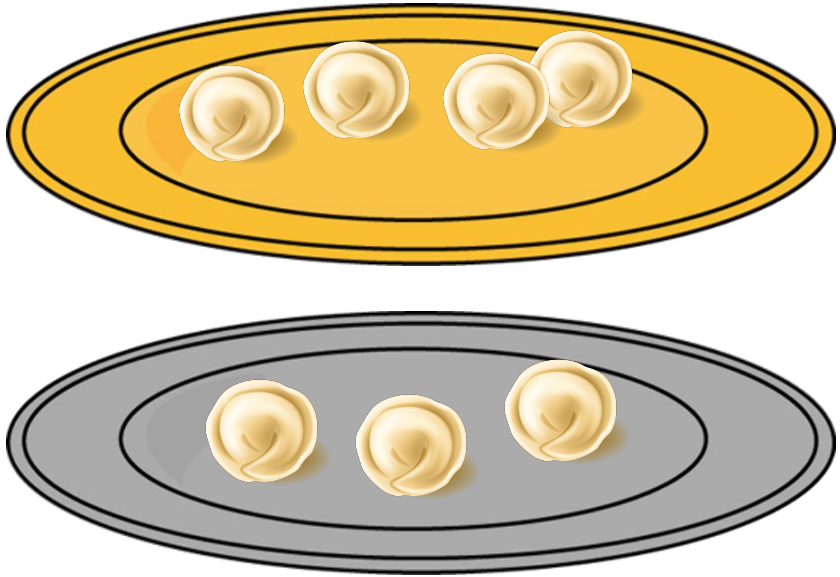
В чем суть RL?

- **Дискретное программирование:** аналитически найти оптимальную стратегию и реализовать
- **RL:** Будем находить стратегию методом проб и ошибок

Постановка задачи “почти” RL

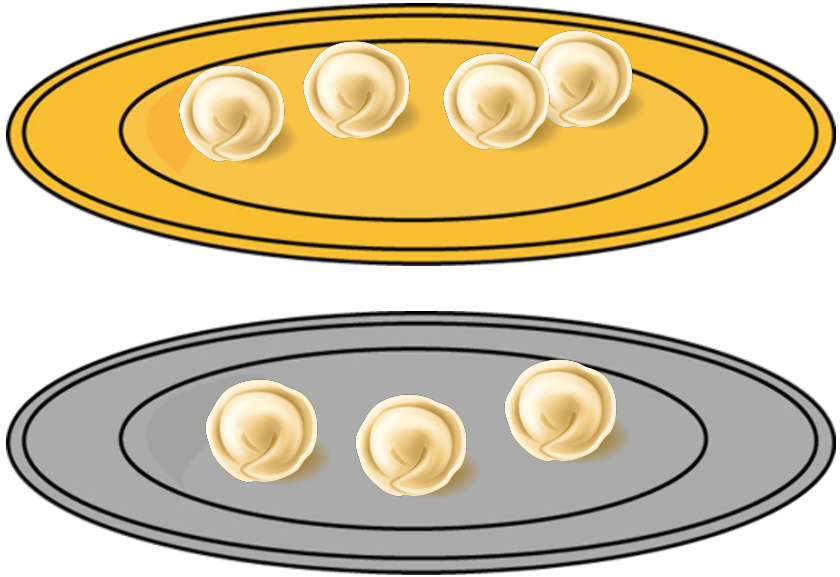
Цель: Найти выигрышную стратегию

… но метод не RL



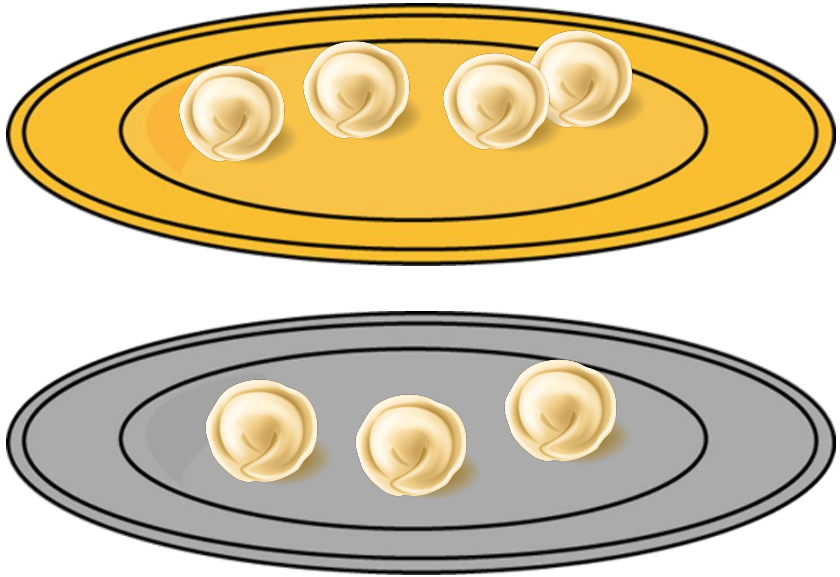
Постановка задачи RL

1. **Environment** (среда, мир, окружение)
2. **Agent** (агент)



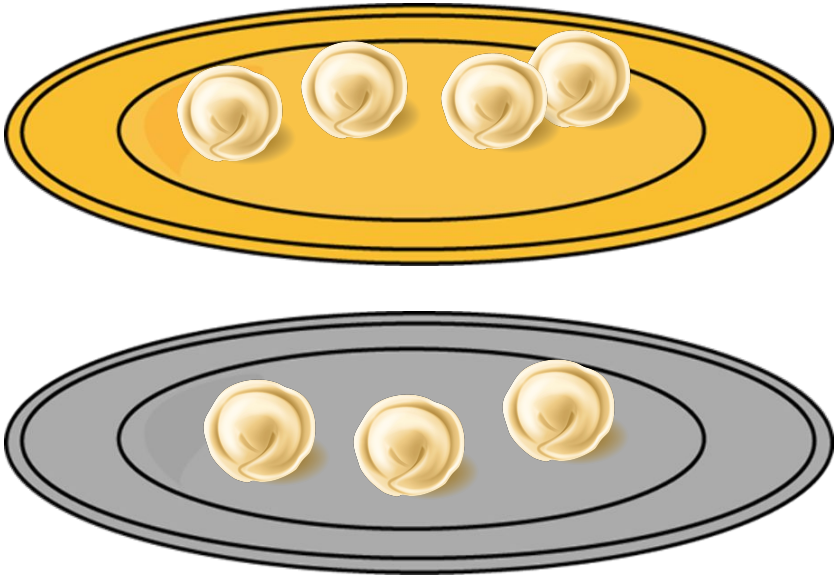
Постановка задачи RL

1. **Environment** (среда, мир, окружение)
2. **Agent** (агент) = 👄



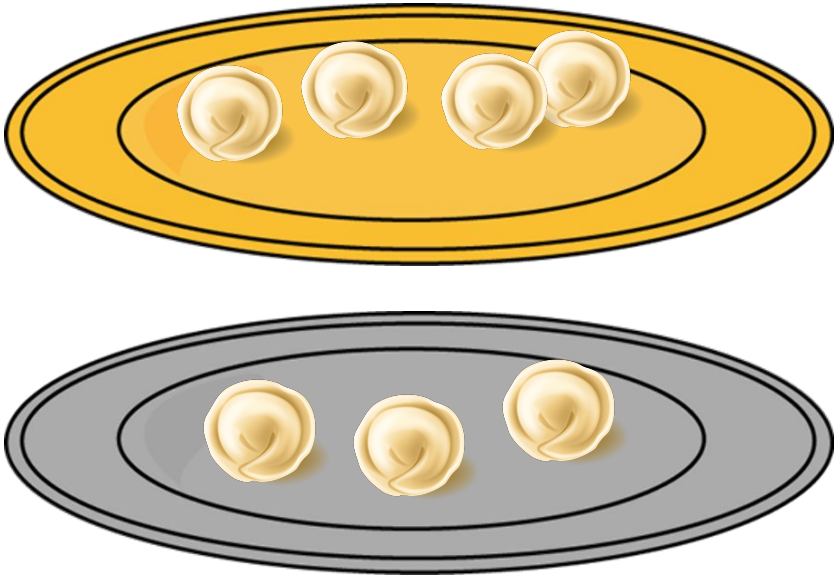
Постановка задачи RL

1. **Environment** (среда, мир, окружение)
2. **Agent** (агент) = 👄



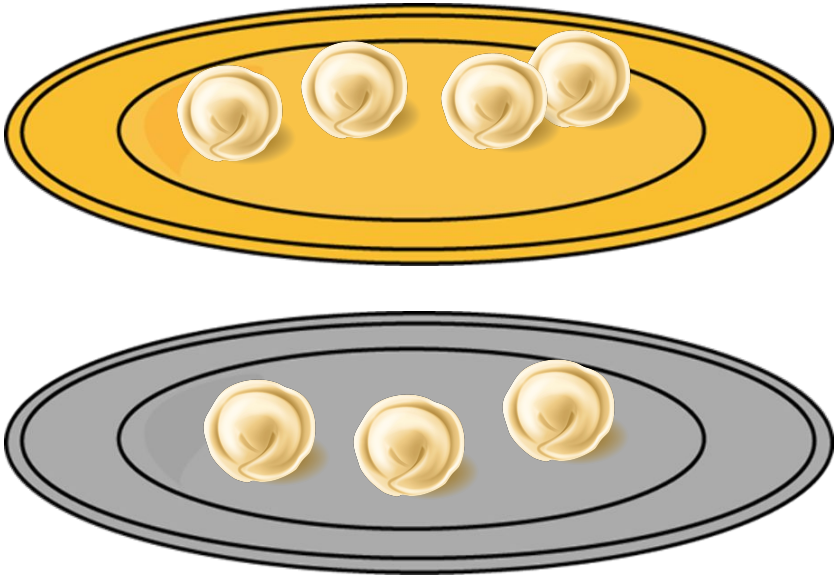
Постановка задачи RL

1. **Environment** (среда, мир, окружение)
   - **S**
   - **A**
   - **P**



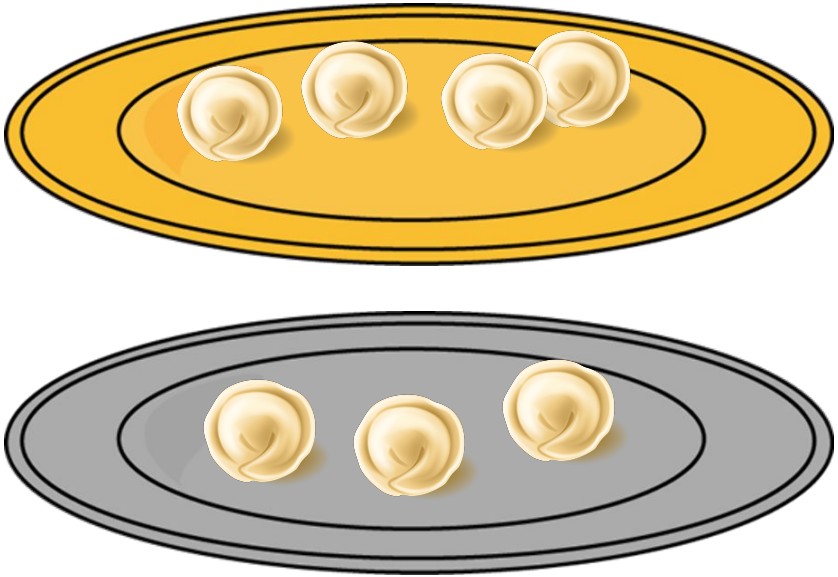
Постановка задачи RL

1. **Environment** (среда, мир, окружение)
    - **S** — пространство состояний (state space)
    - **A**
    - **P**



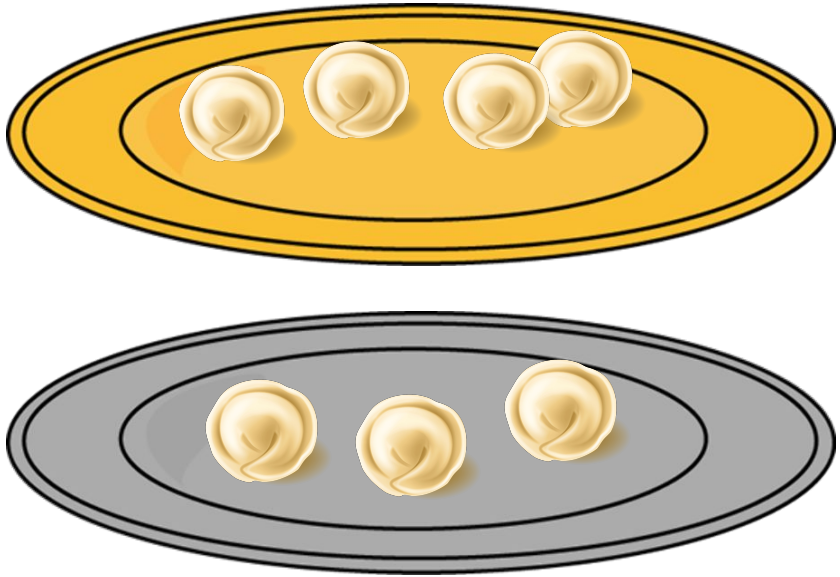
Постановка задачи RL

1. **Environment** (среда, мир, окружение)
   - **S** — пространство состояний (state space)
   - **A** — пространство действий (action space)
   - **P**



Постановка задачи RL

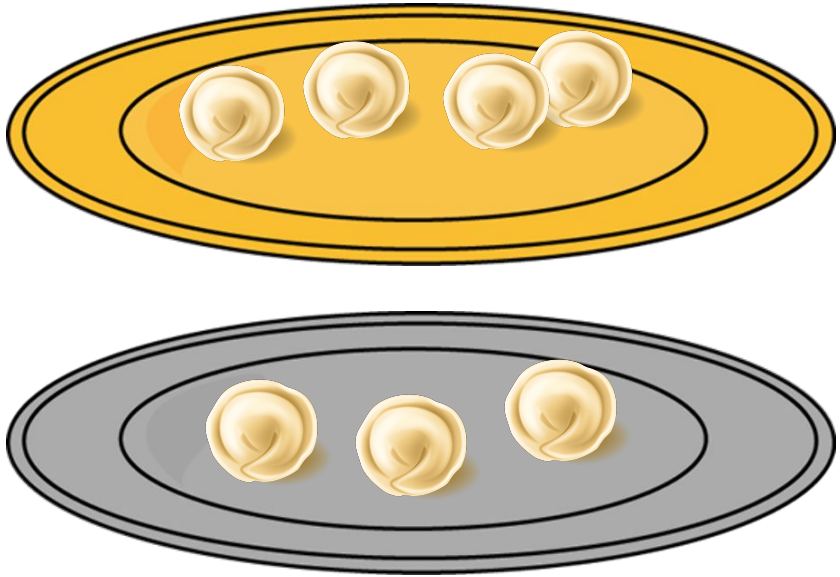
1. **Environment** (среда, мир, окружение)
   - **S** — пространство состояний (state space)
   - **A** — пространство действий (action space)
   - **P** — функция переходов (transition function)
     или динамика среды (world dynamics)



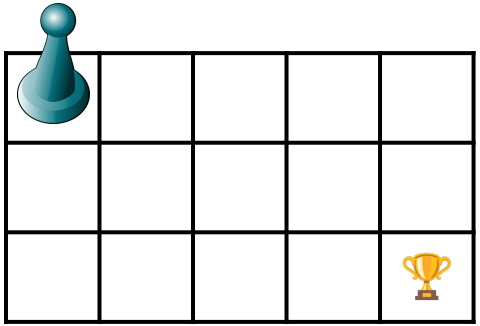
Практика

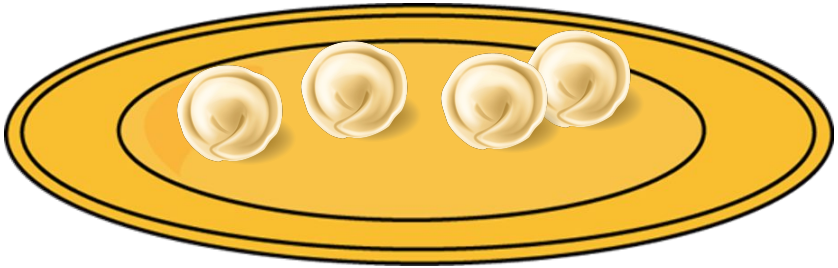
Постановка задачи RL

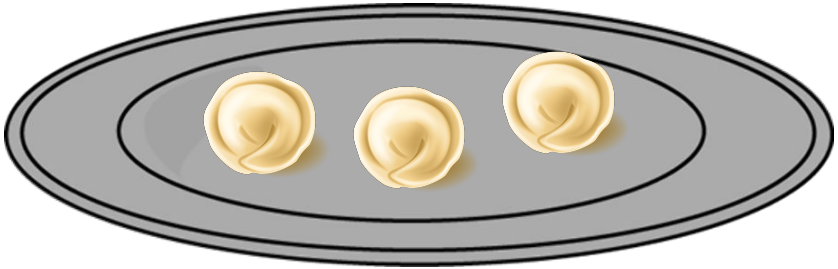
1. **Environment:**
   State, Actions, Dynamics
2. **Agent**



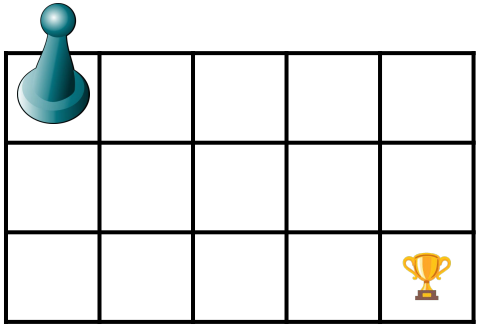
Вопрос на понимание



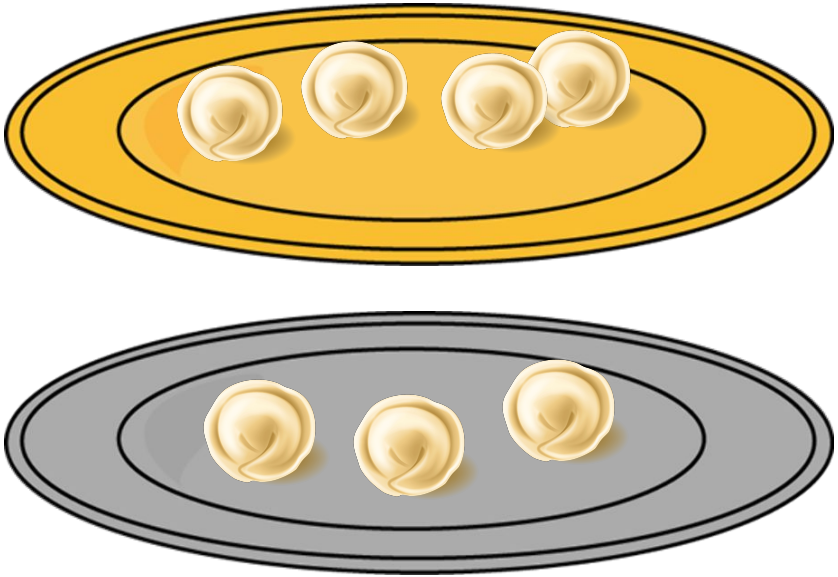




Вопрос на понимание

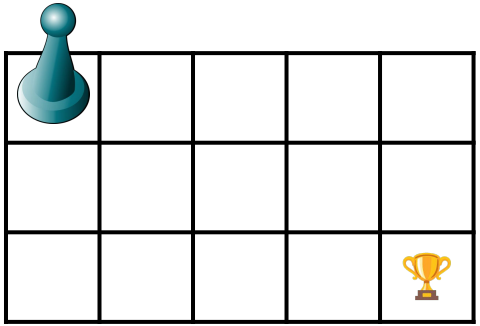


Ход: Сдвинуться на любое число клеток по горизонтали или вертикали



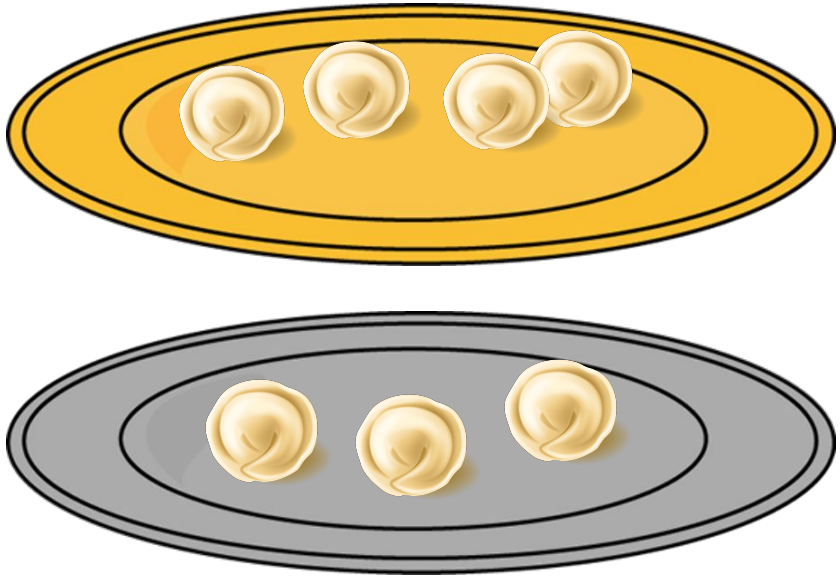
Ход: Съесть любое число пельменей из одной тарелки

Вопрос на понимание



Ход: Сдвинуться на любое число клеток по горизонтали или вертикали

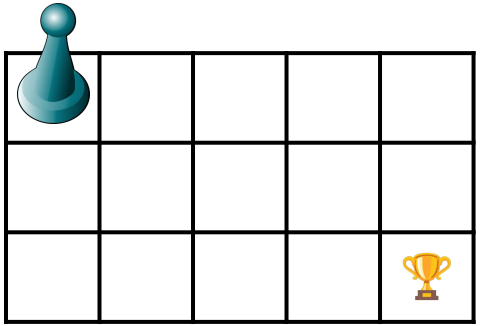
Победа: (0, 0)



Ход: Съесть любое число пельменей из одной тарелки

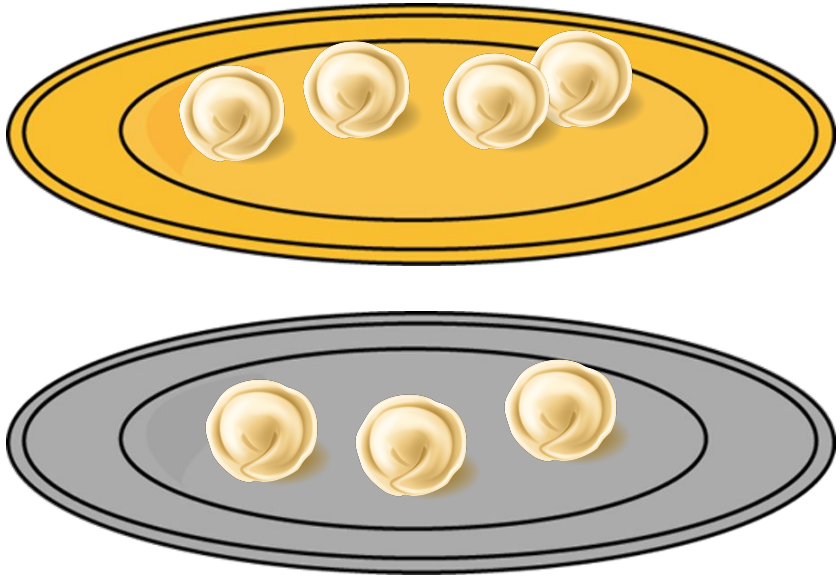
Победа: (0, 0)

Среда [environment] точно такая же!



Ход: Сдвинуться на любое число клеток по горизонтали или вертикали

Победа: (0, 0)



Ход: Съесть любое число пельменей из одной тарелки

Победа: (0, 0)

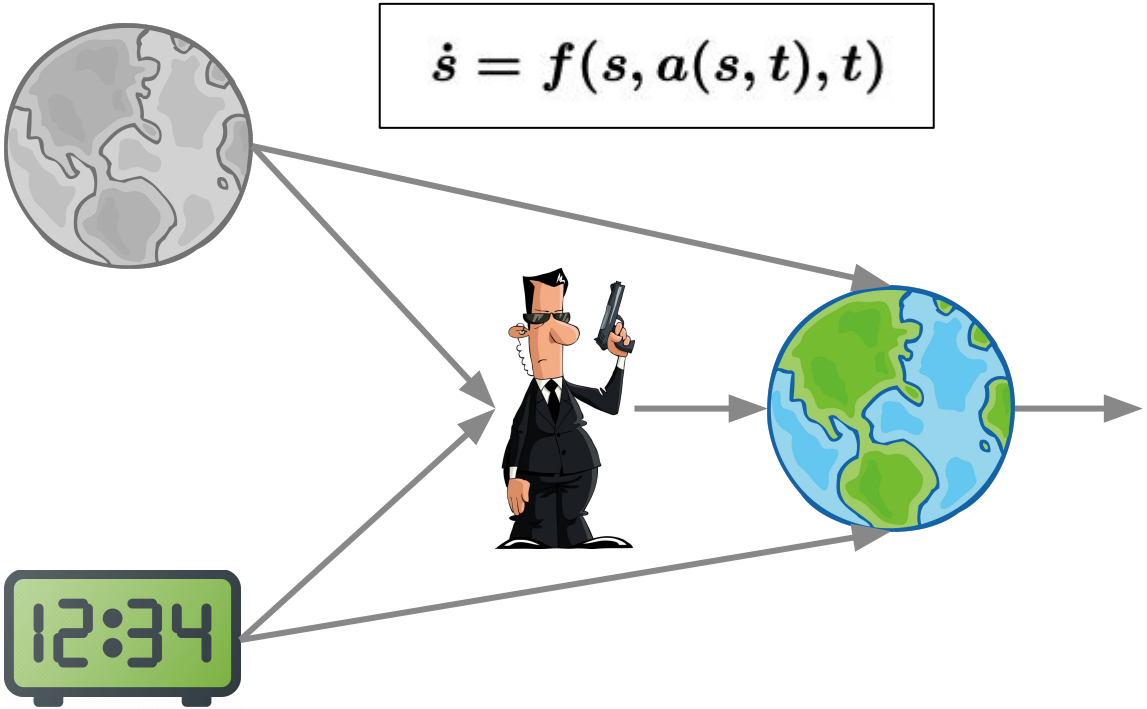
Модель мира

1. **Environment:**
   State, Actions, Dynamics
2. **Agent**

$$\dot{s} = f(s, a(s,t), t)$$

Модель мира

1. **Environment:**
   State, Actions, Dynamics
2. **Agent**

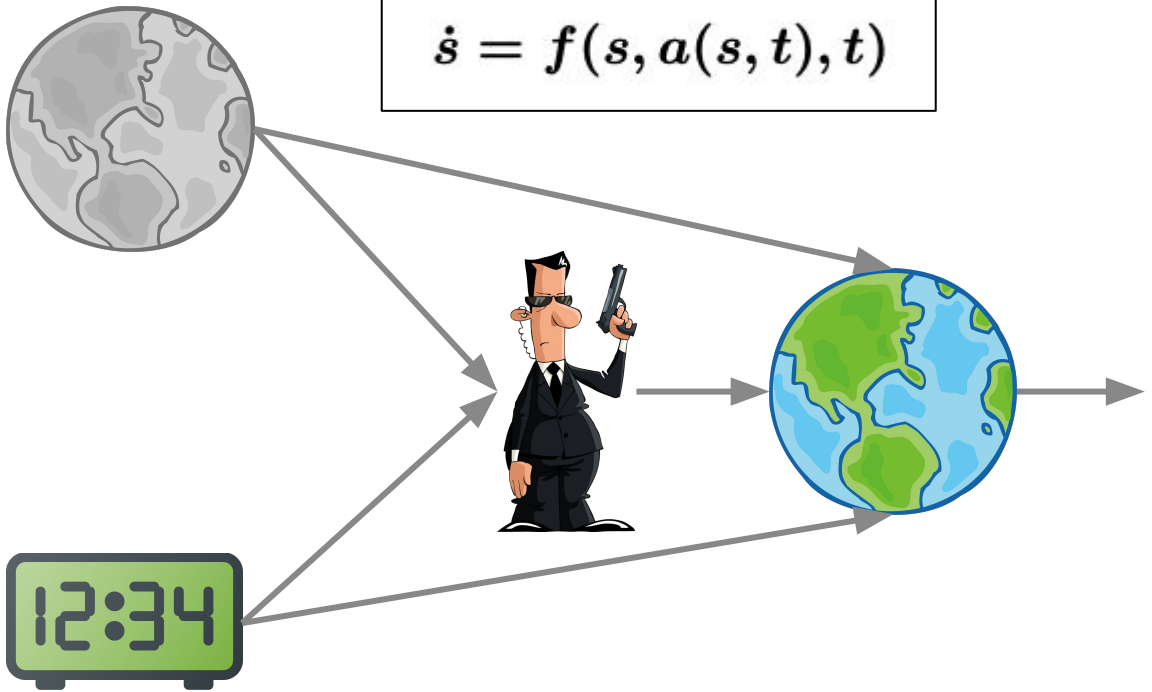


Модель мира

1. **Environment:**
   State, Actions, Dynamics
2. **Agent**

$a(s, t) = ?$

$$\dot{s} = f(s, a(s, t), t)$$

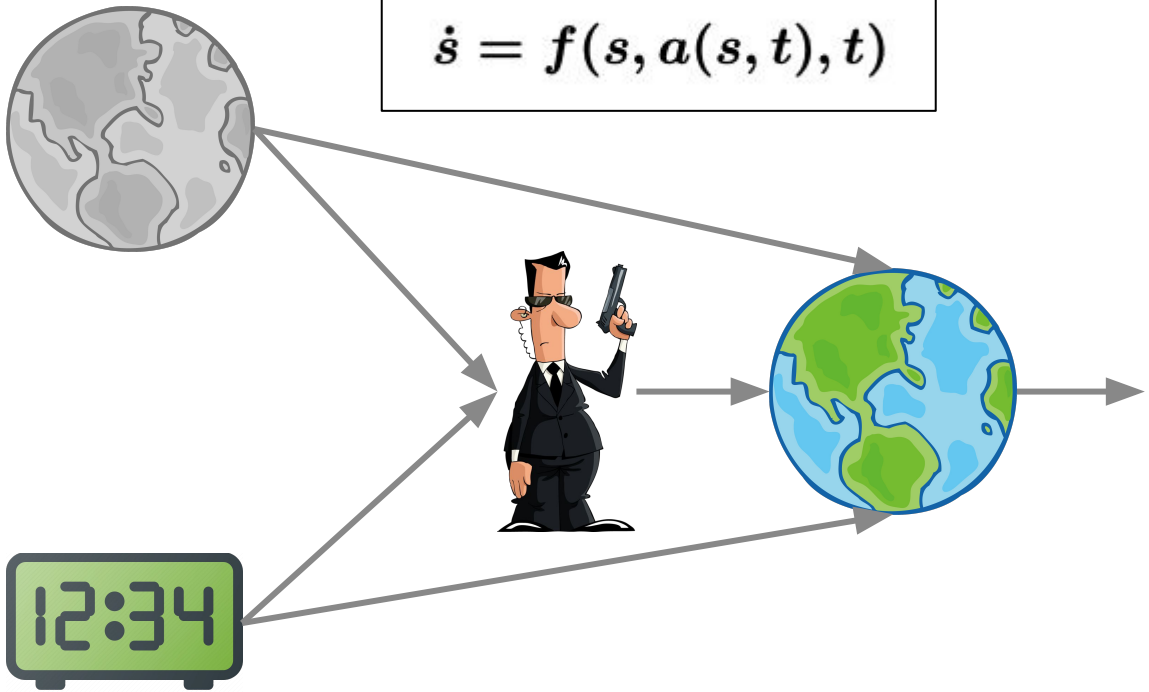


Модель мира

1. **Environment:**
   State, Actions, Dynamics
2. **Agent**

a(s, t) = ?
Какую именно **a** найти?

$$\dot{s} = f(s, a(s, t), t)$$

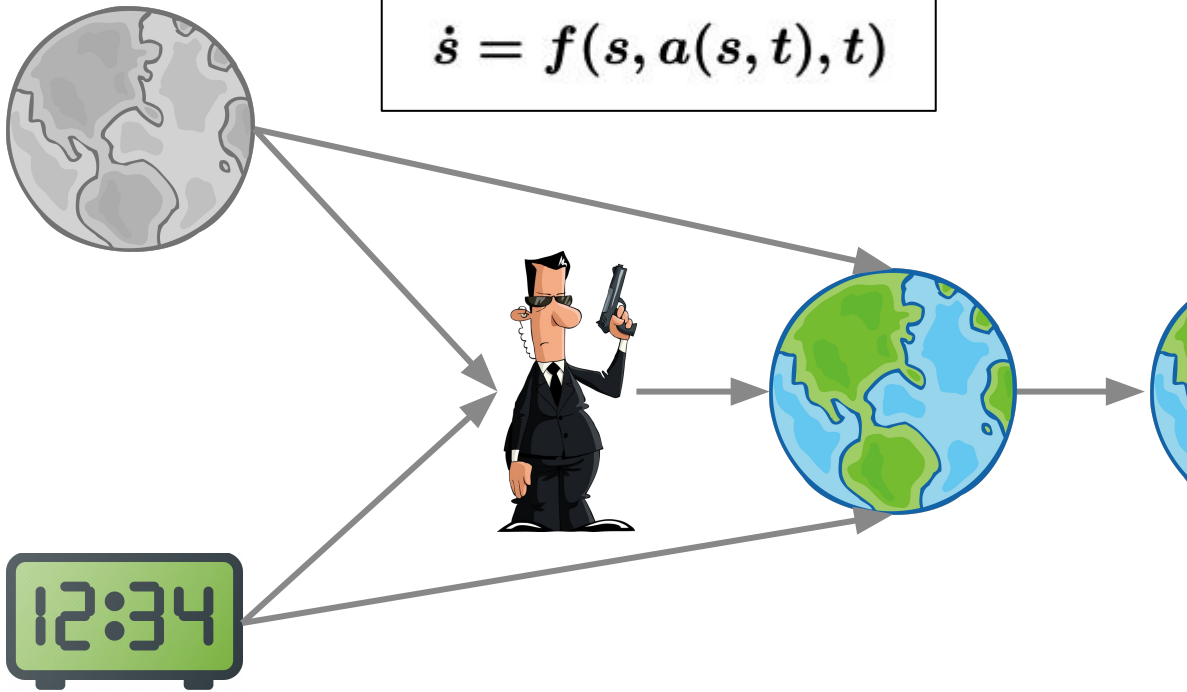


Модель мира

1. **Environment:**
   State, Actions, Dynamics
2. **Agent**

a(s, t) = ?
Какую именно **a** найти?

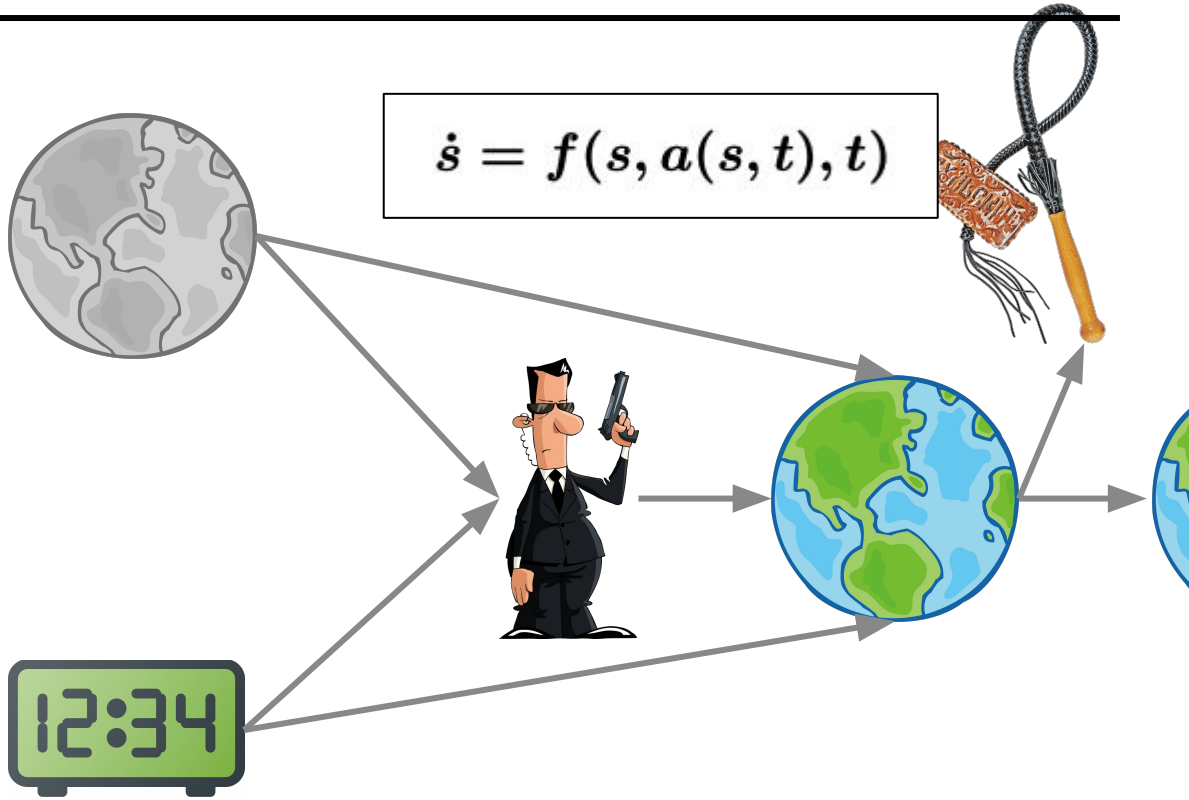
$$\dot{s} = f(s, a(s,t), t)$$



Модель мира

1. **Environment:**
   State, Actions, Dynamics
2. **Agent**

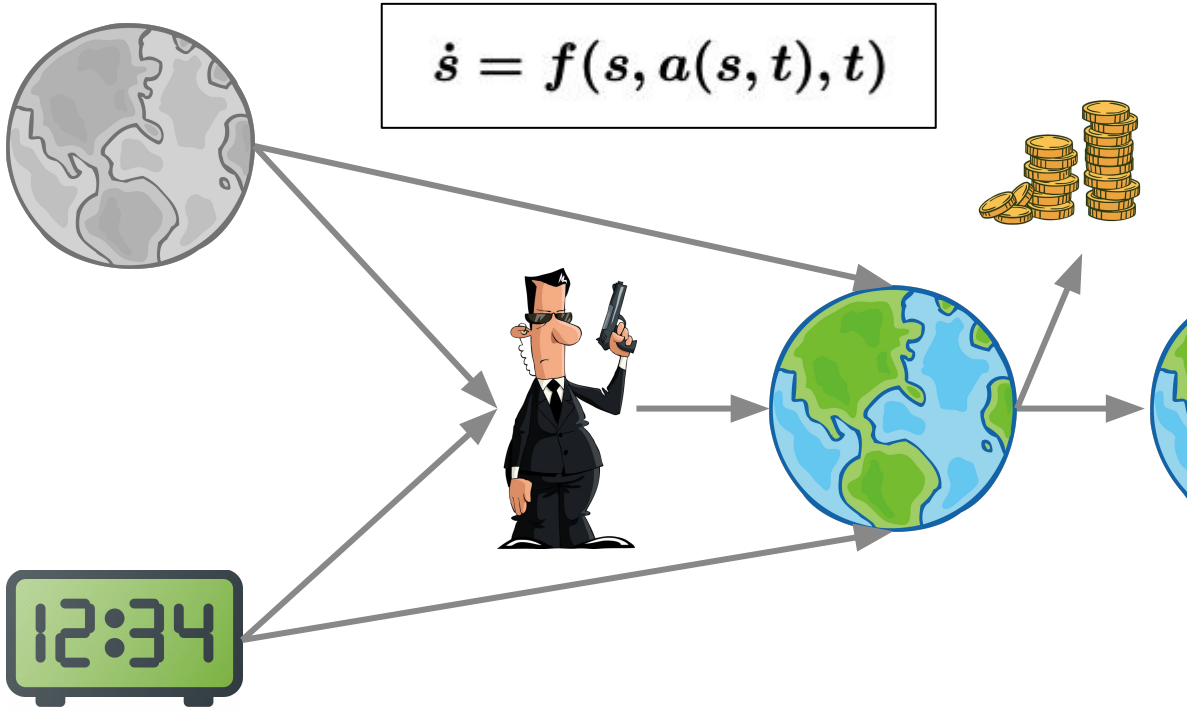
a(s, t) = ?
Какую именно **a** найти?



Модель мира

1. **Environment:**
   State, Actions, Dynamics
2. **Agent**

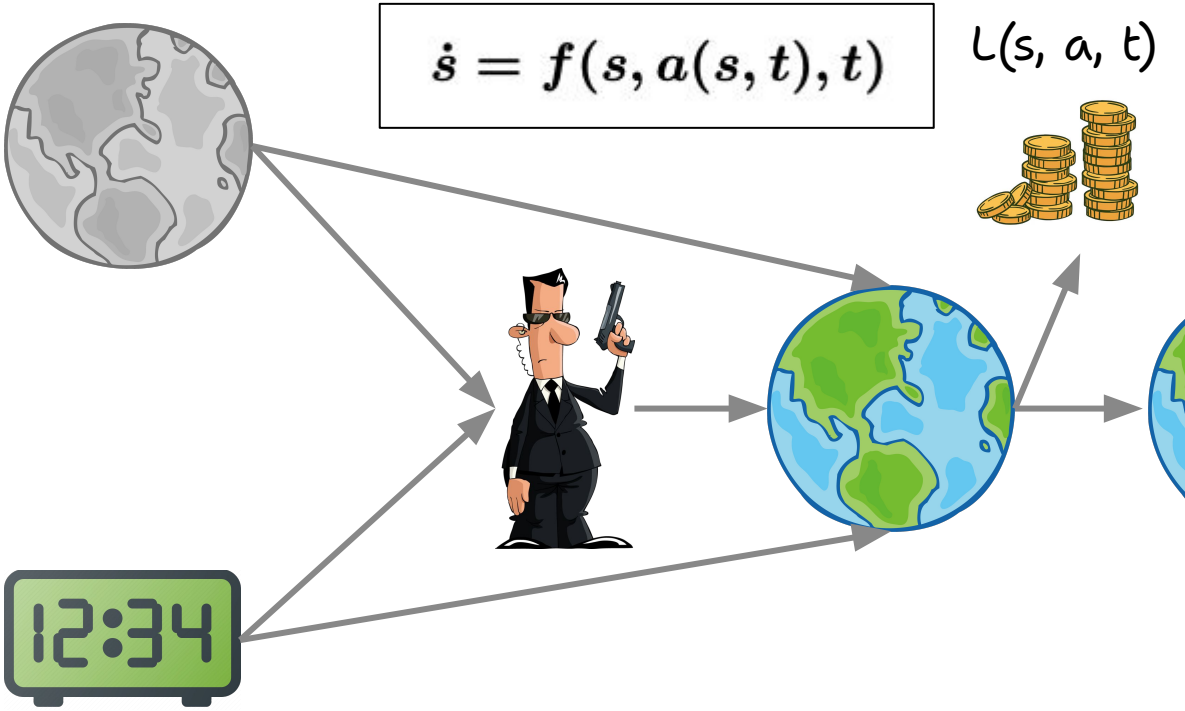
a(s, t) = ?
Какую именно **a** найти?



Модель мира

1. **Environment:**
   State, Actions, Dynamics
2. **Agent**

a(s, t) = ?
Какую именно **a** найти?



Нужны награды!

1. **Environment:**
   State, Actions, Dynamics
2. **Agent**

$$\begin{cases} \dot{s} = f(s, a(s,t), t) \\ -\int L(s, a(s,t), t)\,\mathrm{d}t \to \max\limits_{a(s,t)} \end{cases}$$

a(s, t) = ?

Какую именно **a** найти?

Нужны награды!

Точное аналитическое решение:

- Теория принятия решений
- Теория оптимального управления

$$\dot{s} = f(s, a(s,t), t)$$

$$-\int L(s, a(s,t), t)\,\mathrm{d}t \to \max_{a(s,t)}$$

$a(s, t) = ?$

Приближенное:

- Reinforcement Learning

Нужны награды!

Точное аналитическое решение:

- Теория принятия решений
- Теория оптимального управления

$$\dot{s} = f(s, a(s,t), t)$$

$$-\int L(s, a(s,t), t)\,\mathrm{d}t \to \max_{a(s,t)}$$

$a(s, t) = ?$

1. Время дискретно
2. Среда стохастична
3. Среда стационарна
4. Модель мира может быть до конца неизвестна

Приближенное:

- Reinforcement Learning

MDP = Марковский Процесс Принятия Решений

1. **Agent**
2. **Environment:**
   State, Actions, Dynamics
3. **Rewards**

} MDP

Что из себя представляет решение?

1. **Agent**
2. **Environment:**
   State, Actions, Dynamics

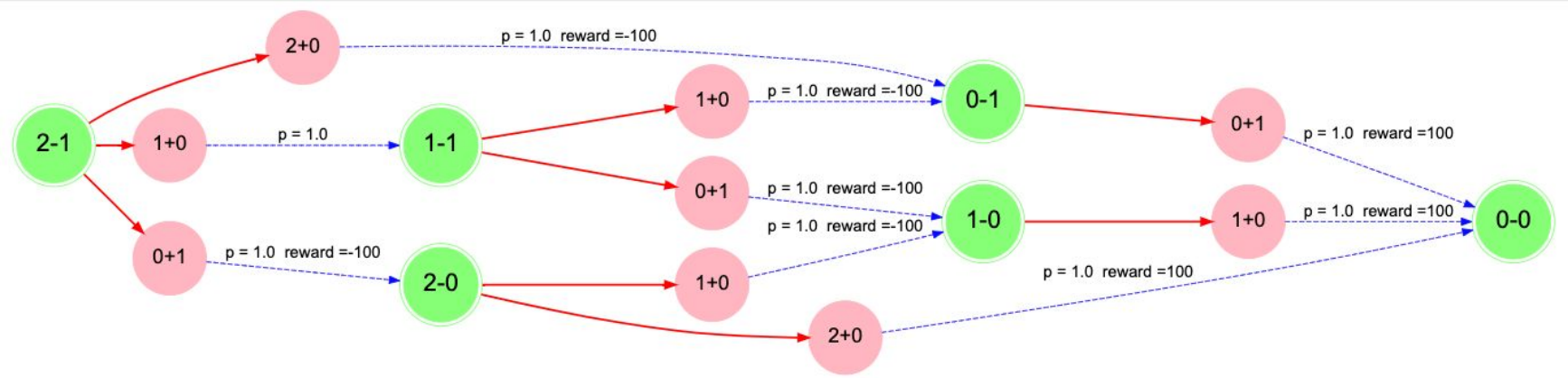
Дано

3. Rewards
4. Алгоритм

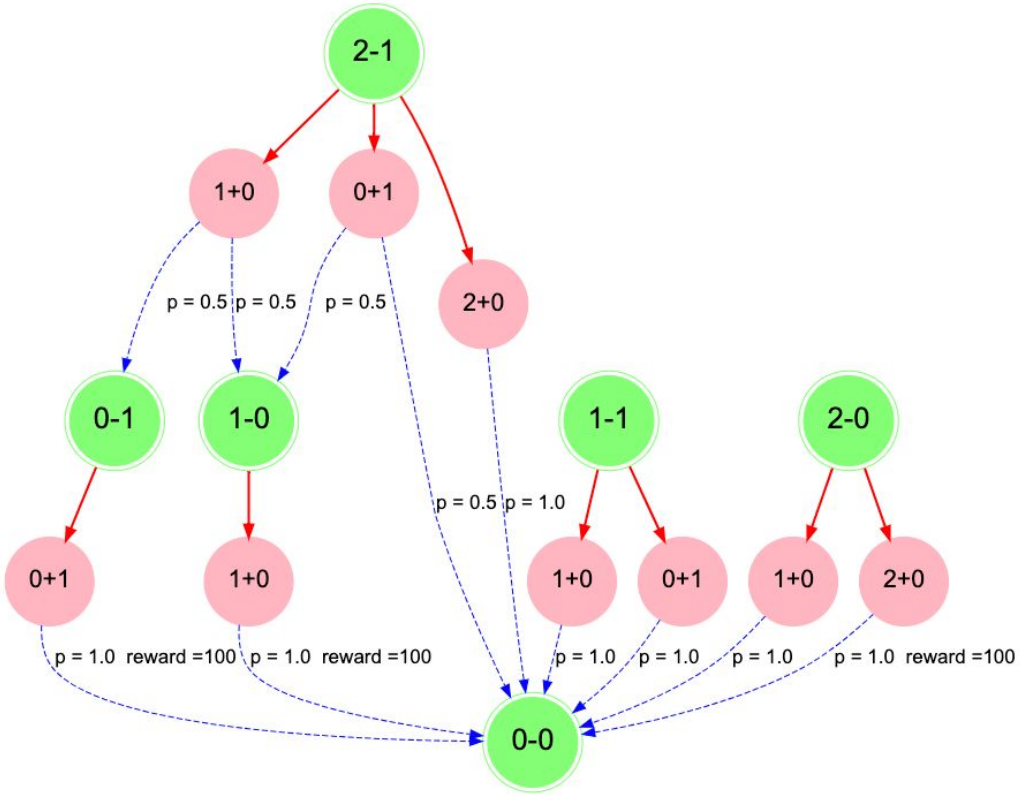
Найти

Практика

Представлене в виде графа без соперника

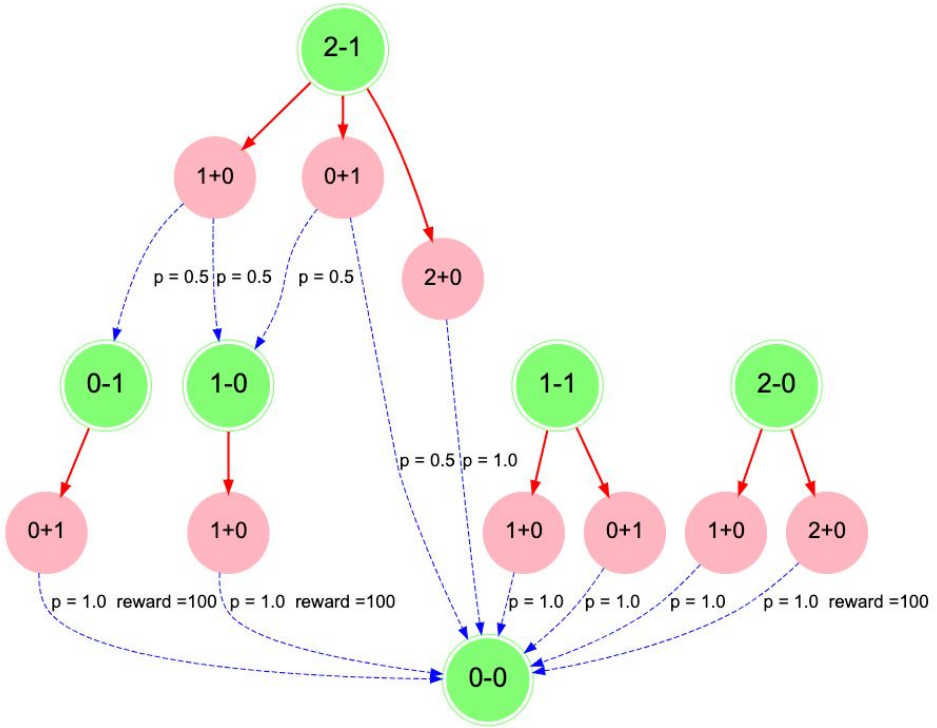


Представлене в виде графа с соперником



Эпизод

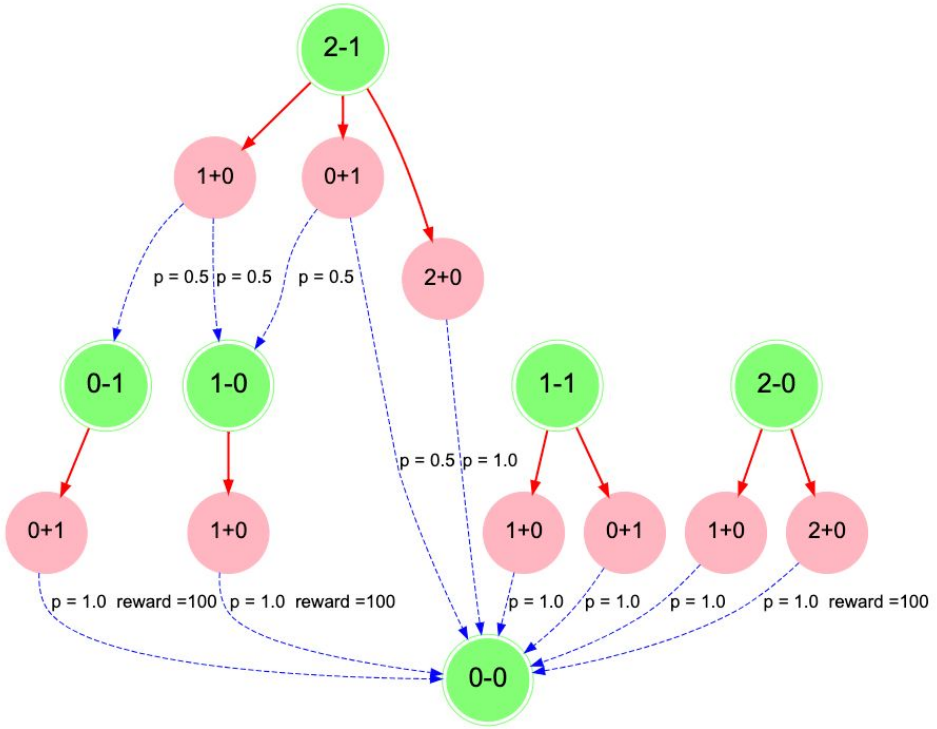
- **Терминальное состояние** = агент не может покинуть его



$\forall a \in A: P(s' = s \mid s, a) = 1, r(s, a) = 0$

Эпизод

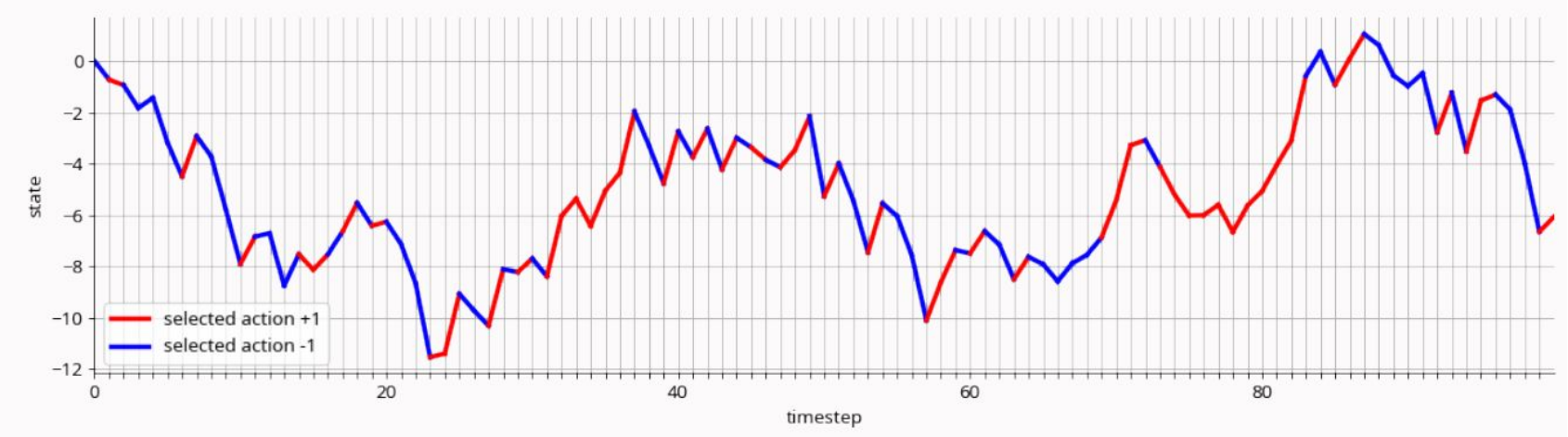
- **Терминальное состояние** = агент не может покинуть его
- **Стартовое состояние** = состояние, с которого начинаем
- **Эпизод** = один цикл процесса от стартового состояния до терминального называется



$\forall a \in A: P(s' = s \mid s, a) = 1, r(s, a) = 0$

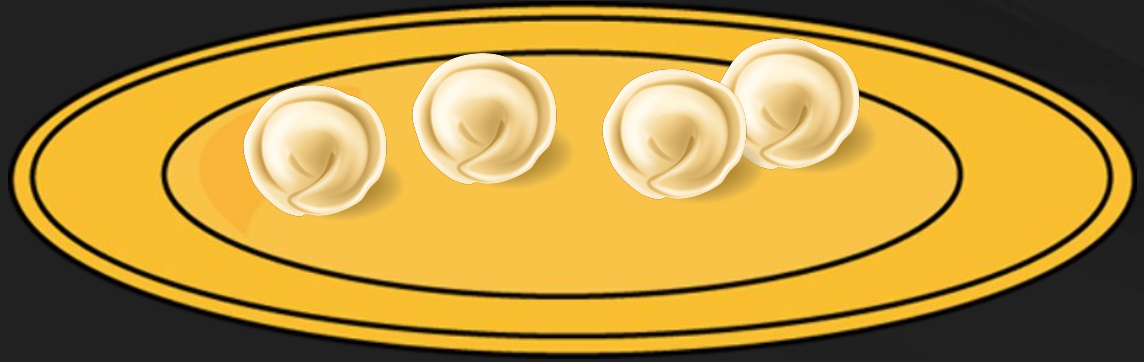
Траектория

Траектория: (s0, a0, s1, a1, s2, a2, s3, a3 . . .)

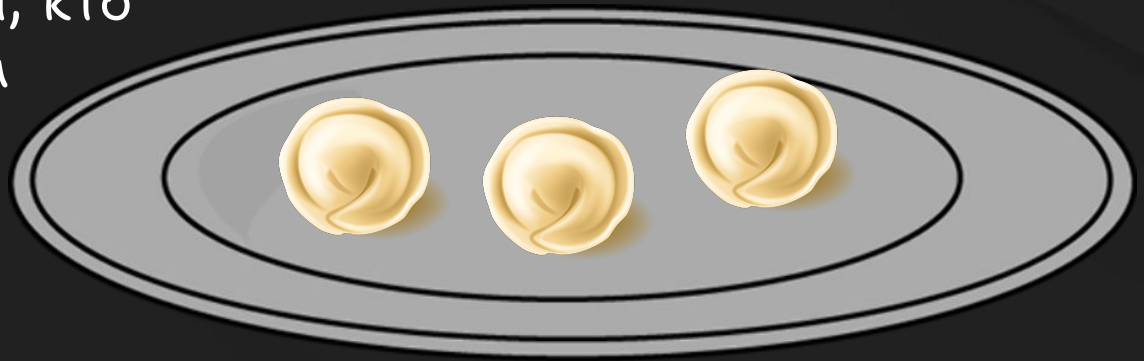


За раз можно съесть любое число пельменей

... но только из одной тарелки

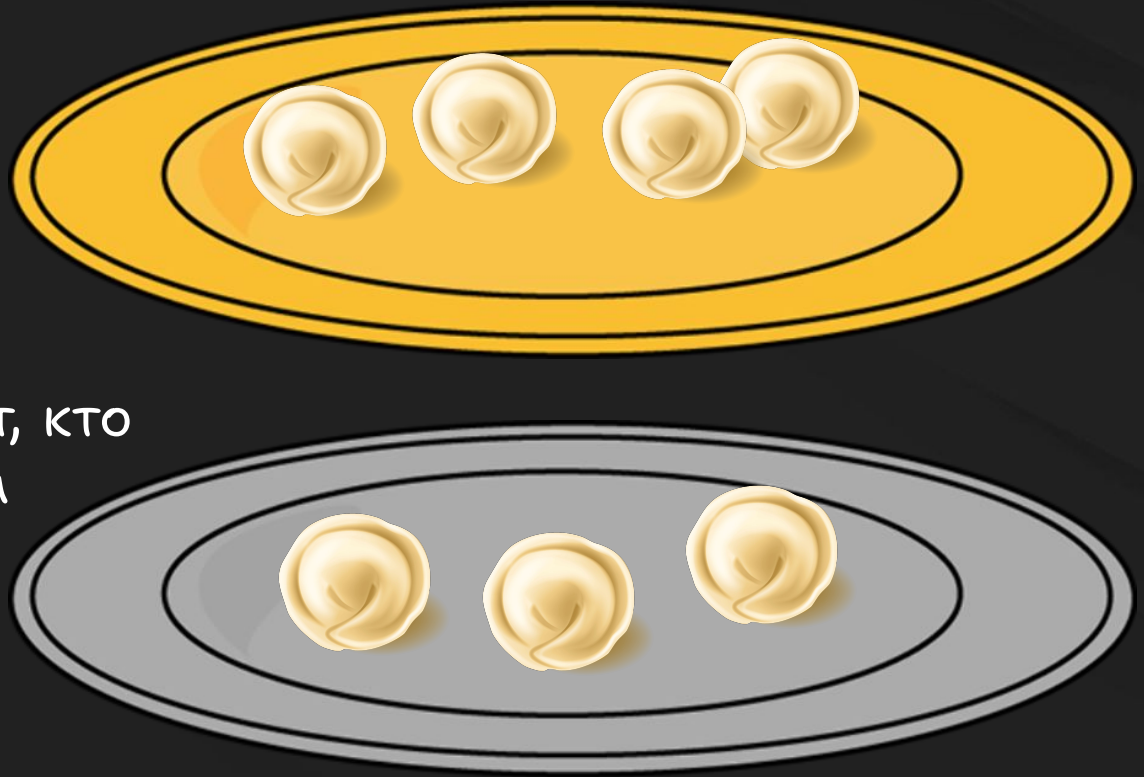


Выигрывает тот, кто съест последний пельмень



За раз можно съесть любое число пельменей

... но только из одной тарелки



Проигрывает

~~Выигрывает~~ тот, кто съест последний пельмень

Pearls Before Swine

https://grandgame.net/game/?flash=pearls_before_swine

Практика

Что из себя представляет решение?

1. **Agent**
2. **Environment:**
   State, Actions, Dynamics

Дано

3. Rewards
4. Алгоритм

Найти

5. Политика

Результат

Какое из утверждений истинно?

А) RL является частным случаем обучения с учителем

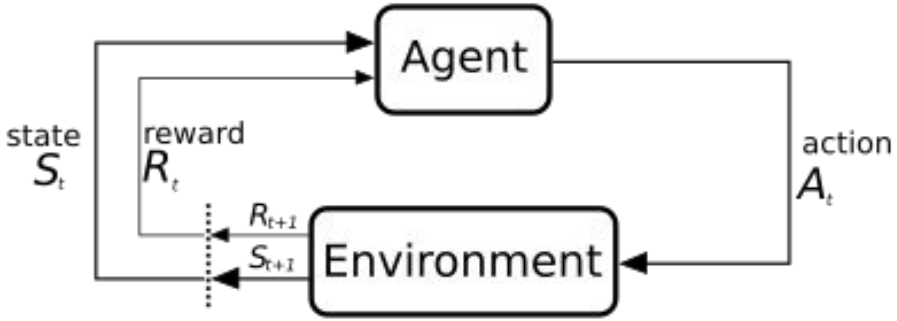
Б) Обучение с учителем – частный случай RL

В) Оба утверждения ложные

Какое из утверждений истинно?

Обучение с учителем:

Имя	Пол	Возраст	ЗП	Дать кредит?
Иван	М	30	50 000	1
Олег	М	36	20 000	1
Анна	Ж	18	500 000	1
Алла	Ж	36	500 000	0
Петр	М	20	10 000	?



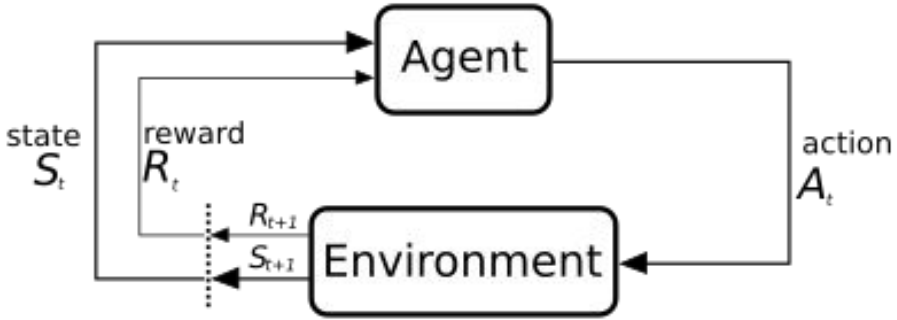
Утверждение:

Задача обучения с учителем является частным случаем задачи RL

В нашем случае?

Обучение с учителем:

Траектория	Награда
`5-4 -> 1+0 -> 4-4 -> 0-4 -> 0+4 -> 0-0`	100
`5-4 -> 1+0 -> 4-4 -> 4-0 -> 4+0 -> 0-0`	100
`5-4 -> 5+0 -> 0-4 -> 0-0`	-100



Утверждение:

Задача обучения с учителем является частным случаем задачи RL

Что из себя представляет решение?

1. **Agent**
2. **Environment:**
   State, Actions, Dynamics

Дано

3. Rewards
4. Алгоритм

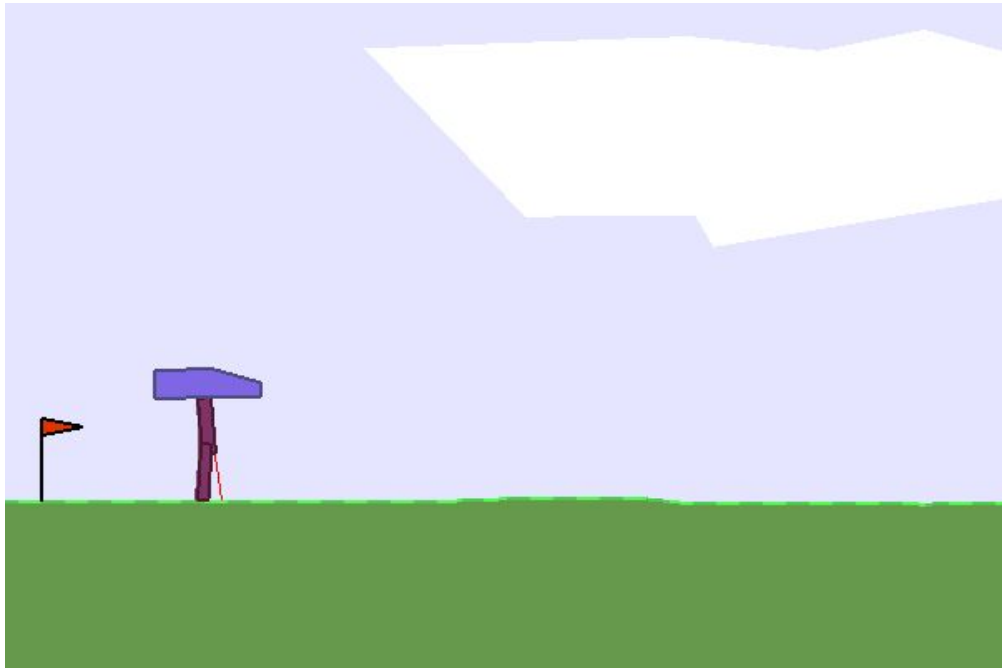
Найти

5. Политика

Результат

BiPedal Walker

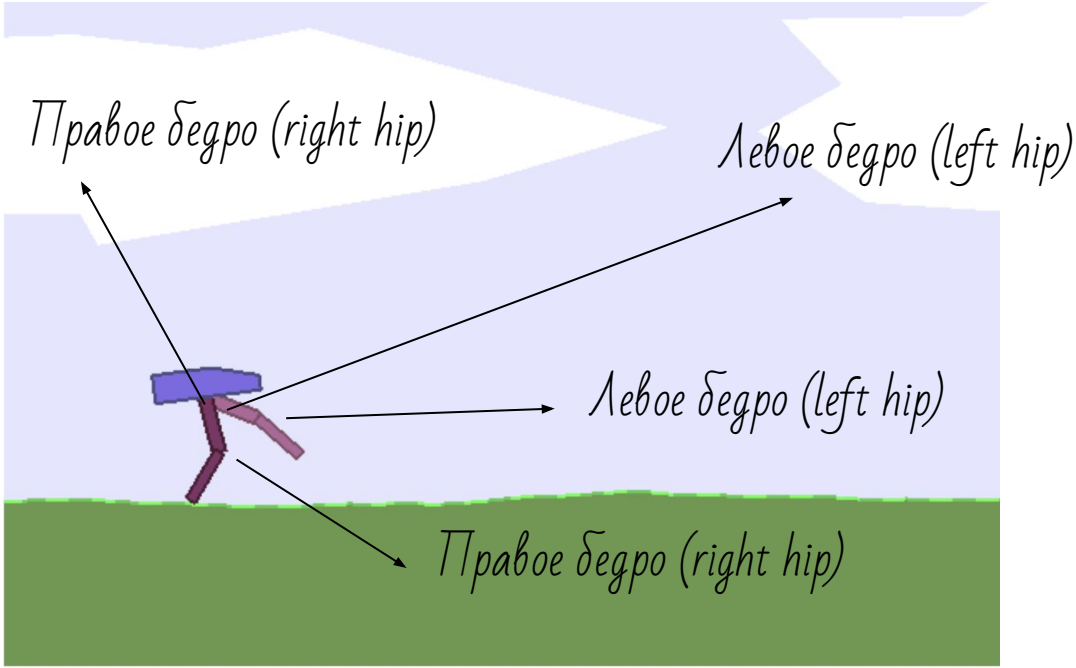
Про среду



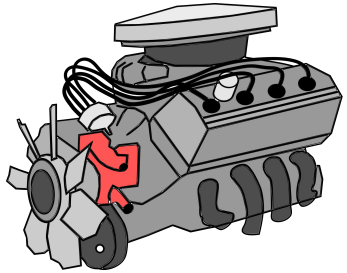
Цель: 300 очков за 1600 шагов

- Двигаться быстро
- Не падать
- Израсходовать меньше энергии

Действия



Задаем скорость на четырех моторах: [-1, 1]

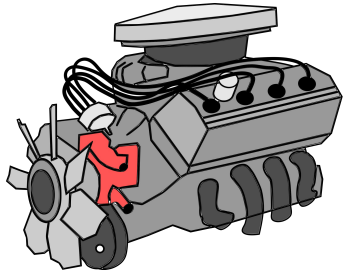


torque = крутящий момент

Действия

Num	Name	Min	Max
0	Hip_1 (Torque / Velocity)	-1	+1
1	Knee_1 (Torque / Velocity)	-1	+1
2	Hip_2 (Torque / Velocity)	-1	+1
3	Knee_2 (Torque / Velocity)	-1	+1

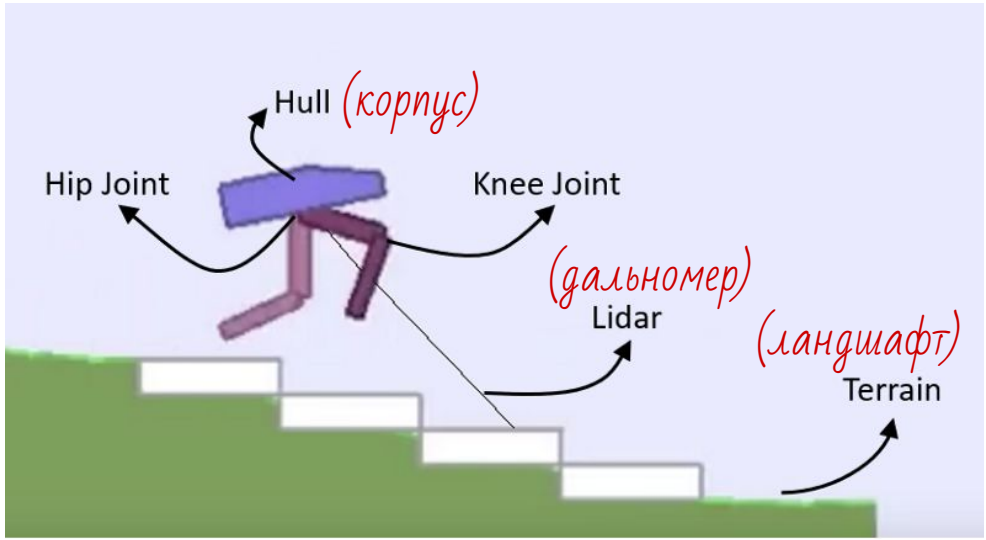
Задаем скорость на четырех моторах: [-1, 1]



torque = крутящий момент

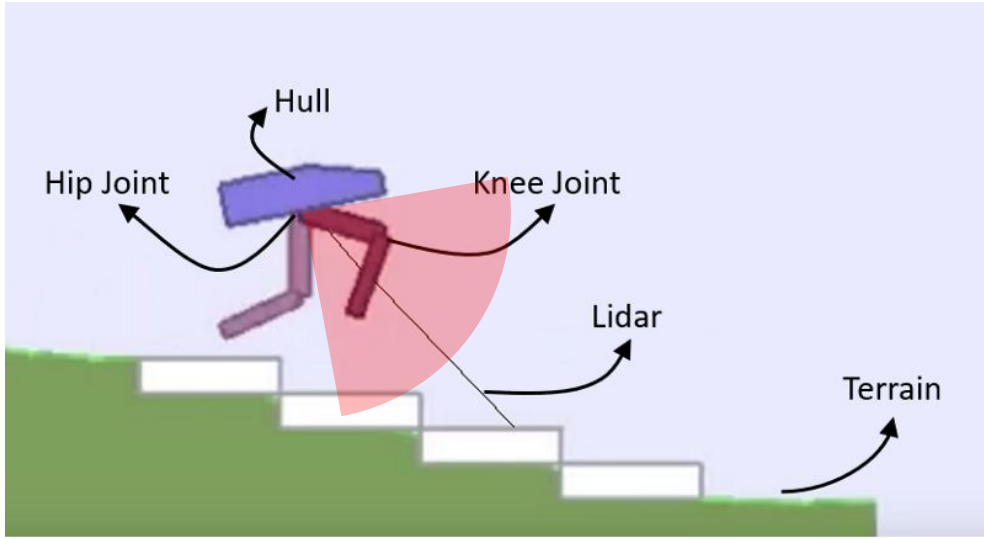
Состояния среды

Num	Observation	Min	Max	Mean
0	hull_angle	0	2*pi	0.5
1	hull_angularVelocity	-inf	+inf	-
2	vel_x	-1	+1	-
3	vel_y	-1	+1	-
4	hip_joint_1_angle	-inf	+inf	-
5	hip_joint_1_speed	-inf	+inf	-
6	knee_joint_1_angle	-inf	+inf	-
7	knee_joint_1_speed	-inf	+inf	-
8	leg_1_ground_contact_flag	0	1	-
9	hip_joint_2_angle	-inf	+inf	-
10	hip_joint_2_speed	-inf	+inf	-
11	knee_joint_2_angle	-inf	+inf	-
12	knee_joint_2_speed	-inf	+inf	-
13	leg_2_ground_contact_flag	0	1	-
14-23	10 lidar readings	-inf	+inf	-

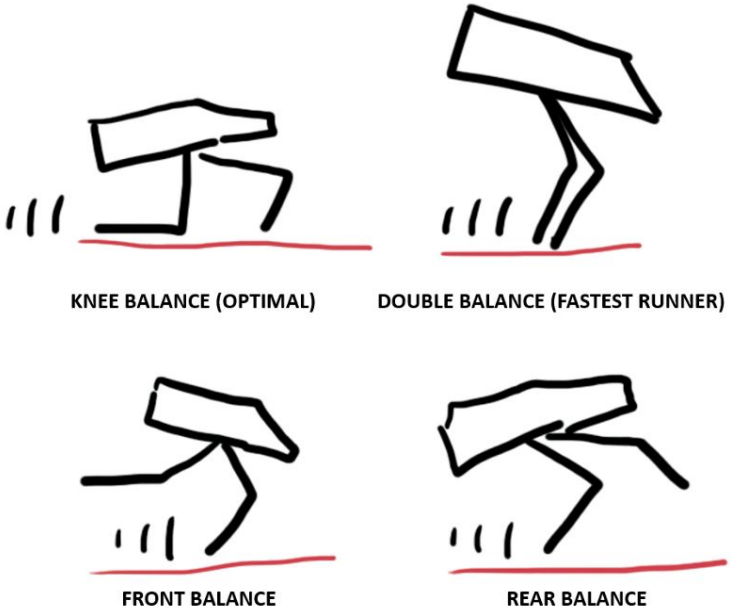


Про лидар

- 10 равномерных измерений
- По дуге в 90^0
- Перпендикулярно корпусу



Варианты стратегий



Резюме

Итого

1. **Agent**
2. **Environment:**
   State, Actions, Dynamics

Дано

MDP

3. Rewards
4. Алгоритм

Найти

5. Политика

Результат

Опрос в конце: https://otus.ru/polls/141246/

Дисклеймер: В презентации использованы личные материалы **@dmi3eva**.

Образовательная площадка **Otus** не несет за них ответственность.