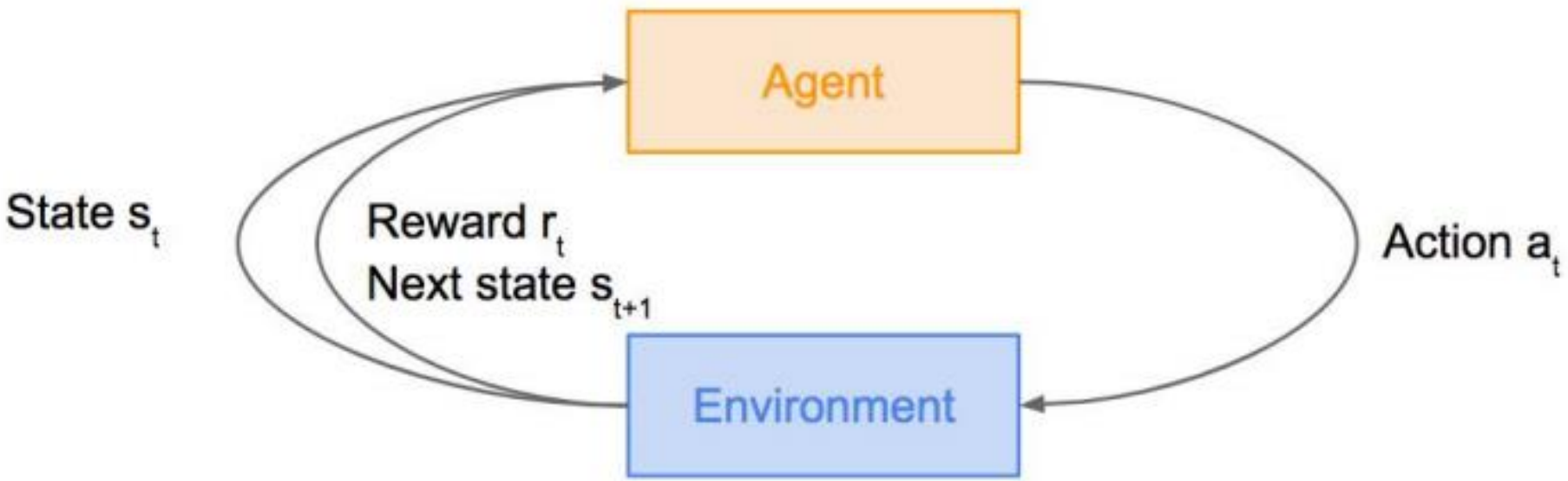
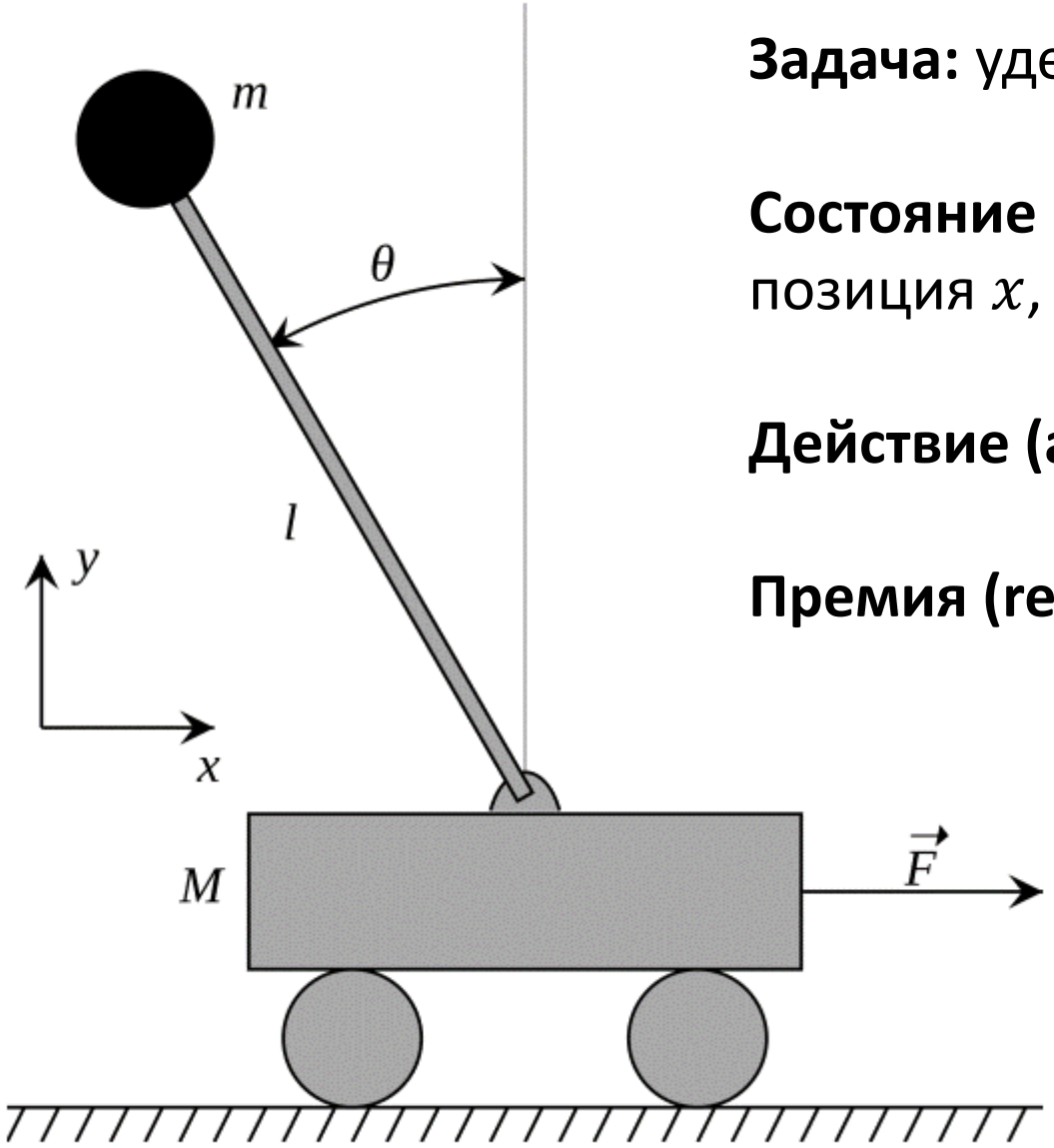
Reinforcement Learning

(обучение с подкреплением, обучение систем управления)



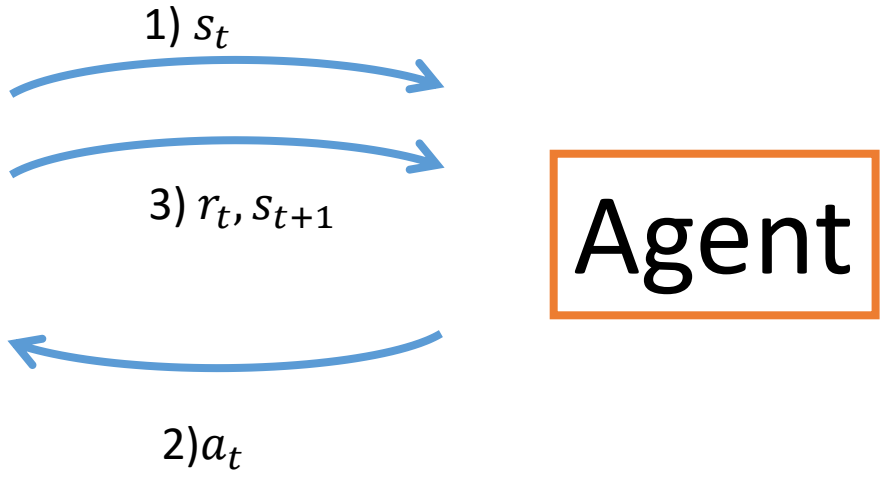


Задача: удержать маятник.

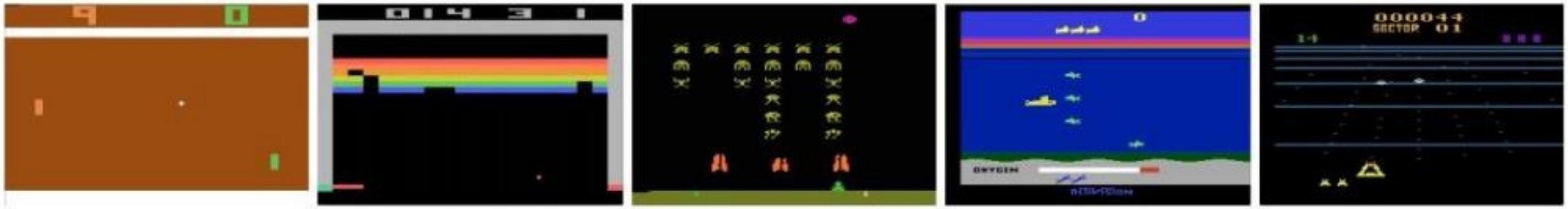
Состояние (state): угол θ, угловая скорость, позиция x, линейная скорость.

Действие (action): приложить силу F вправо или влево.

Премия (reward): +1 на каждом шаге при $|\theta| < 30°$



Atari Games



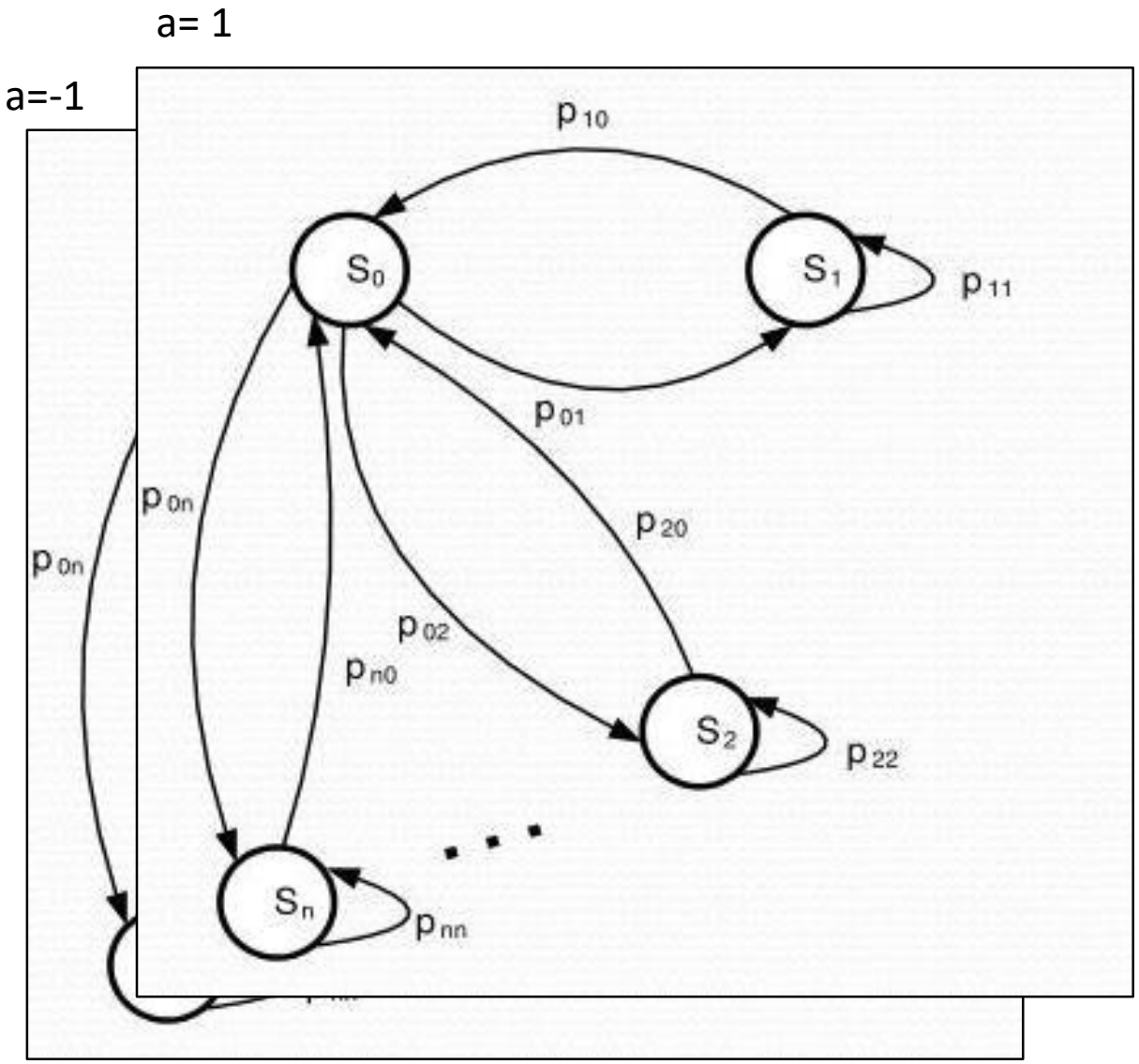
Задача: набрать максимальное количество очков.

Состояние (state): изображение на экране $W \times H \times C$.

Действие (action): кнопки **Up, Down, Left, Right, Fire**.

Премия (reward): +N очков на каждом шаге.

Марковский процесс – вероятности перехода зависят только от текущего состояния **s** и действия **a**



$\mathbb{P}$ - распределение вероятностей перехода в состояние s_{t+1} для пары (s, a)

a = - 1

	s_0	s_1	s_n
s_0	0	0.9	0.1
s_1	0.2	0.2	0.6
s_n	0.6	0.1	0.3

a = 1

	s_0	s_1	s_1
s_0	0.1	0.2	0.7
s_1	0.6	0.1	0.3
s_n	0.2	0.2	0.6

$$(\mathcal{S}, \mathcal{A}, \pi^*, \mathcal{R}, \mathbb{P}, \gamma)$$

$\mathcal{S}$ - множество состояний s

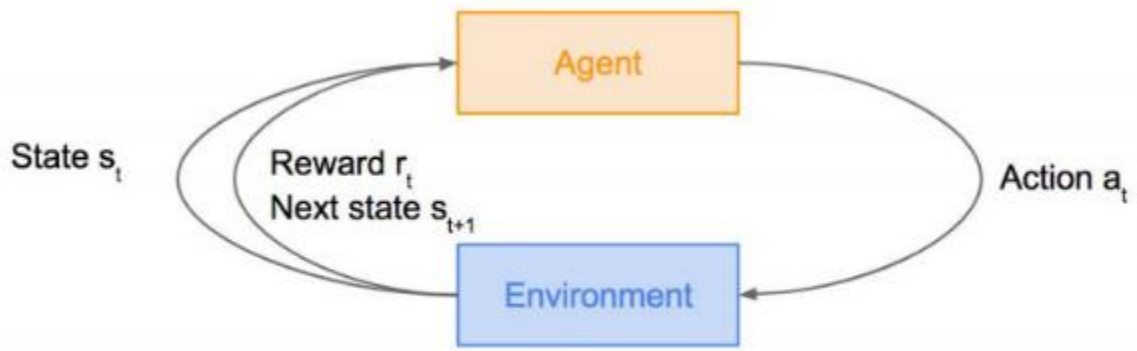
$\mathcal{A}$ - множество действий *a*

π^*- политика агента при выборе действия a в зависимости от состояния s

$\mathbb{P}$ - распределение вероятностей перехода среды в состояние s_{t+1} для пары (s, a)

$\mathcal{R}$ - распределение премий (reward) за пару (s, a)

γ – дисконт премии на каждом шаге.



Алгоритм управления

В момент t_0 среда находится в состоянии $s_0 \sim p(s_0)$

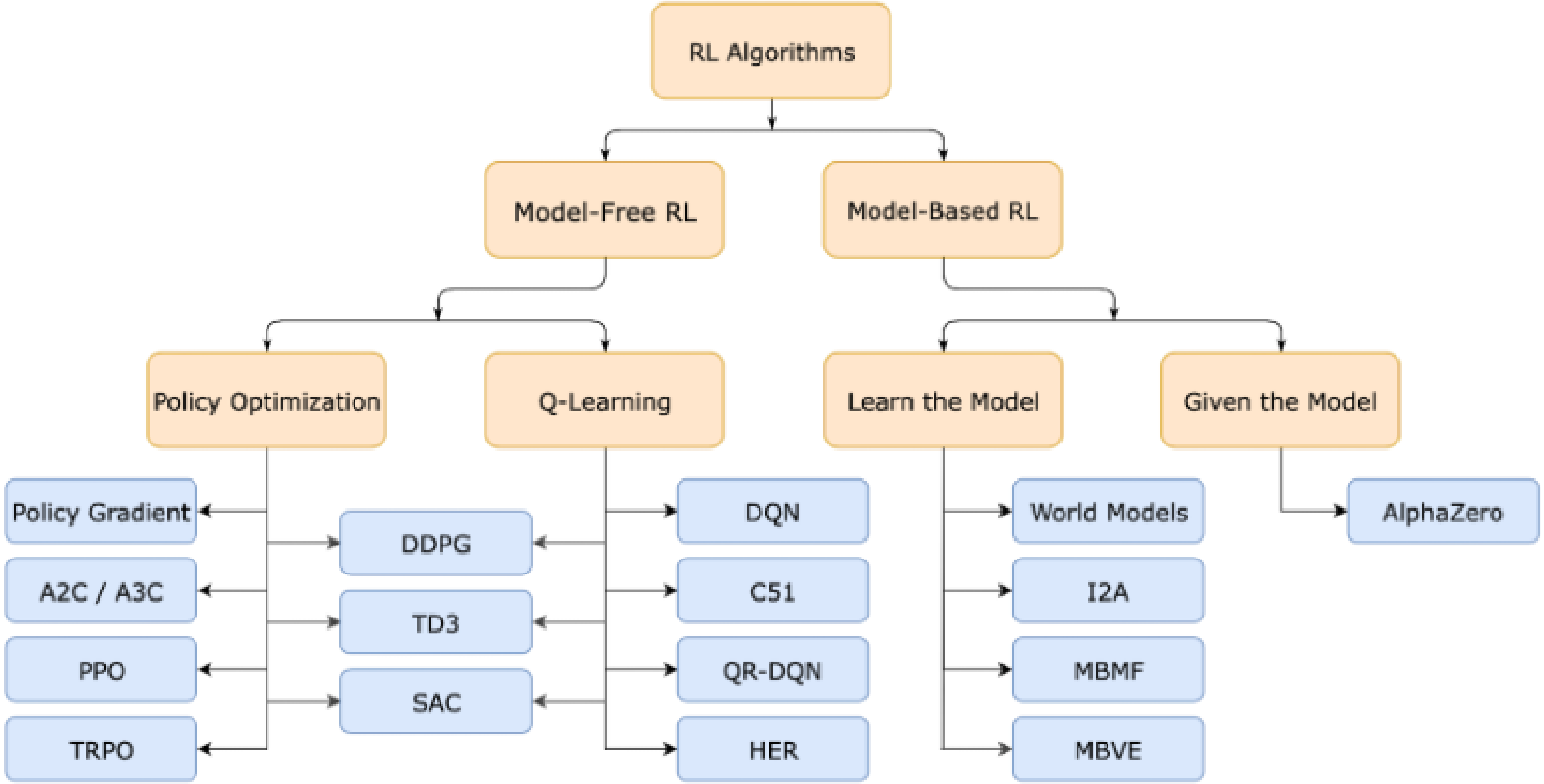
Цикл :

- агент совершает действие a_t в соответствии с политикой $\pi^*(s_t)$
- среда выдает премию $r_t \sim R(.|s_t, a_t)$
- среда переходит в состояние $s_{t+1} \sim P(.|s_t, a_t)$
- агент получает премию r_t и следующее состояние s_{t+1}

Политика π - это функция $S \rightarrow A$ которая аппроксимируется нейросетью.

Оптимальная политика π^*- это такая функция, которая позволяет получить максимальную сумму премий с дисконтом:

$$\sum_{t>0} \gamma^t r_t \rightarrow max$$


RL Algorithms
Model-Free RL
Model-Based RL
Policy Optimization
Q-Learning
Learn the Model
Given the Model
Policy Gradient
A2C / A3C
PPO
TRPO
DDPG
TD3
SAC
DQN
C51
QR-DQN
HER
World Models
I2A
MBMF
MBVE
AlphaZero