



ОБУЧЕНИЕ С

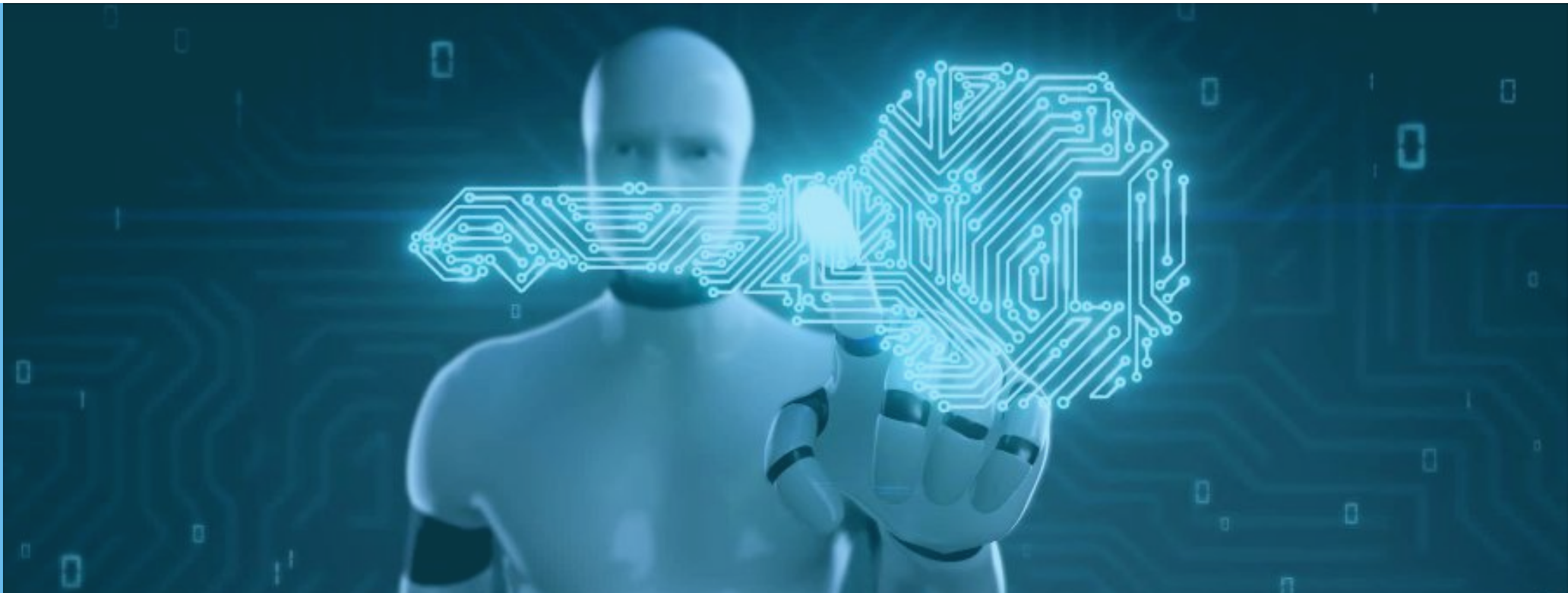
ПОДКРЕПЛЕНИЕМ

ЗАНЯТИЕ #16



ОБУЧЕНИЕ С ПОДКРЕПЛЕНИЕМ В МАШИННОМ ОБУЧЕНИИ

«Брось робота в лабиринт и пусть
ищет себе выход»



Отличие от других подходов



- **Обучение с учителем**
есть база, есть размеченные ответы
- **Обучение без учителя**
есть база, нет размеченных ответов
- **Обучение с подкреплением**
нет базы, нет размеченных ответов



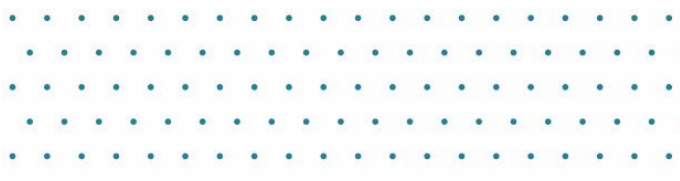
ИДЕЯ

Стремиться
к действиям, ведущим
к награде, избегать
действий, ведущих к
неудачам



ОБЩАЯ ПОСТАНОВКА ЗАДАЧИ

Взаимодействие агента со средой:

- Инициализируется состояние среды, стратегия агента
 - Агент выбирает и совершает действие
 - Среда генерирует награду и своё новое состояние
 - Агент корректирует стратегию
- 

Область применения

- Игры (особенно логические)
- Роботы-манипуляторы
- Навигация машин, роботов
- Боты(трейдинг, чат, игровые)
- И т.д ...



Abstract background featuring binary code (0s and 1s) and mathematical formulas, including $\lim_{h \rightarrow 0} \frac{1}{h} \ln \frac{1}{h} = 0$ and $\lim_{h \rightarrow 0} \frac{1}{h} \ln \frac{1}{h} = 0$.

- 

POLICY GRADIENT

(градиенты политики, стратегии)

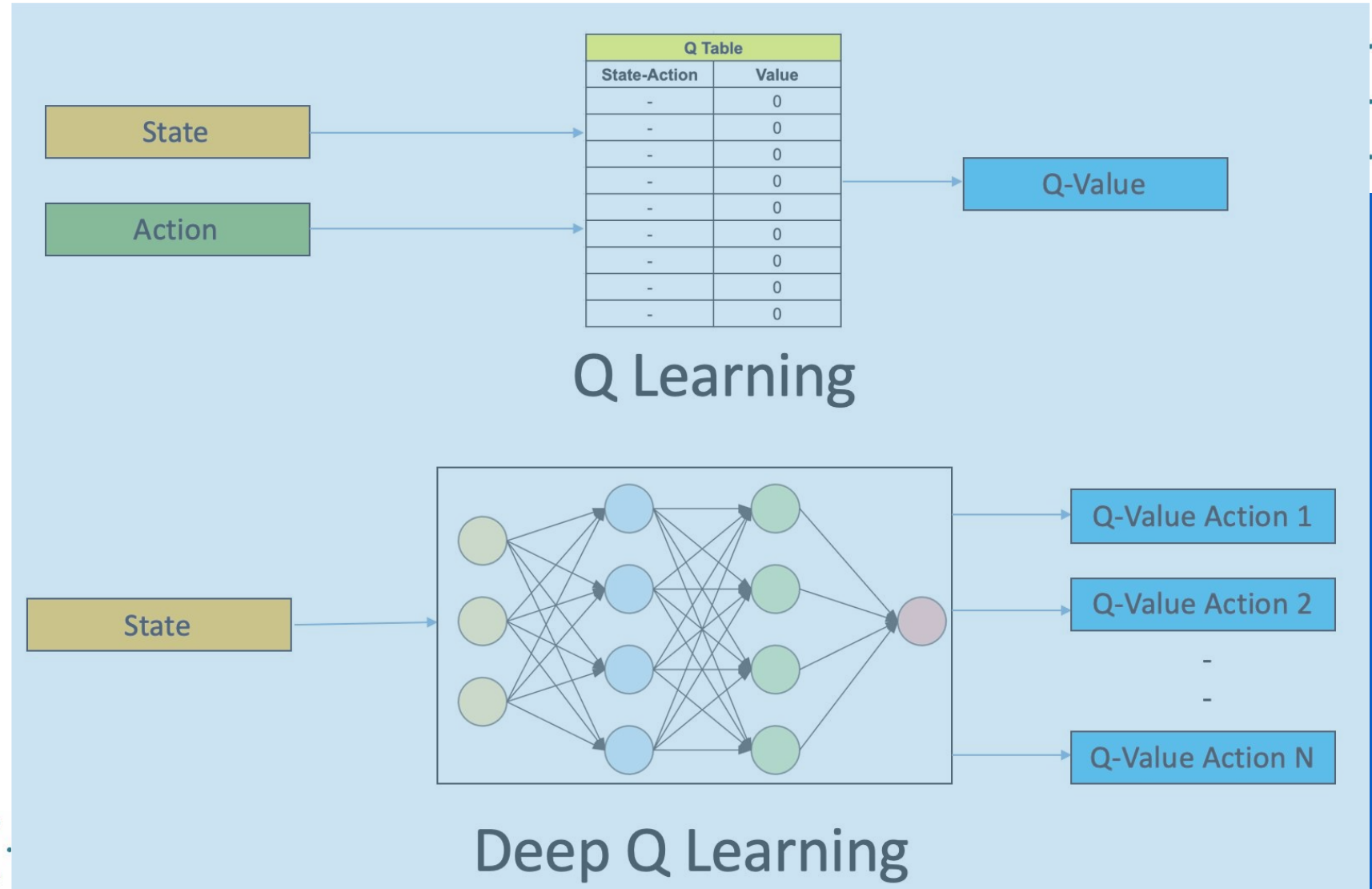
Схожа с обучением с учителем, но:

- Вместо правильных меток - метка выиграл/проиграл
- Если выиграл в эпизоде - все действия в нем получают позитивную метку(и наоб.)
- Функция потерь - кросс-энтропия, в которую вводим вознаграждение (умножаем его на логарифм)
- Делаем тренировку из ряда игровых эпизодов, обновляем веса по этой функции, делаем следующую тренировку



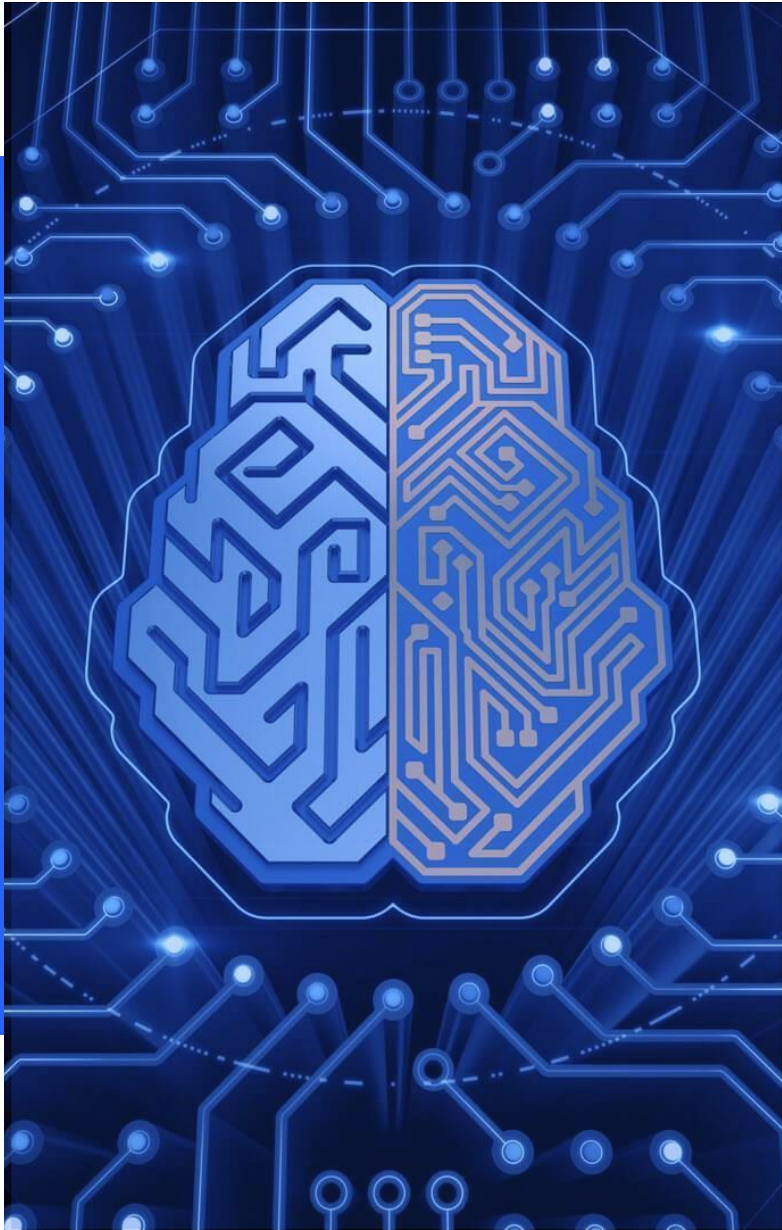
Q-learning

Deep
Q-learning -
DQN





НЕЙРОСЕТЬ УЧИТСЯ ИГРАТЬ В PONG

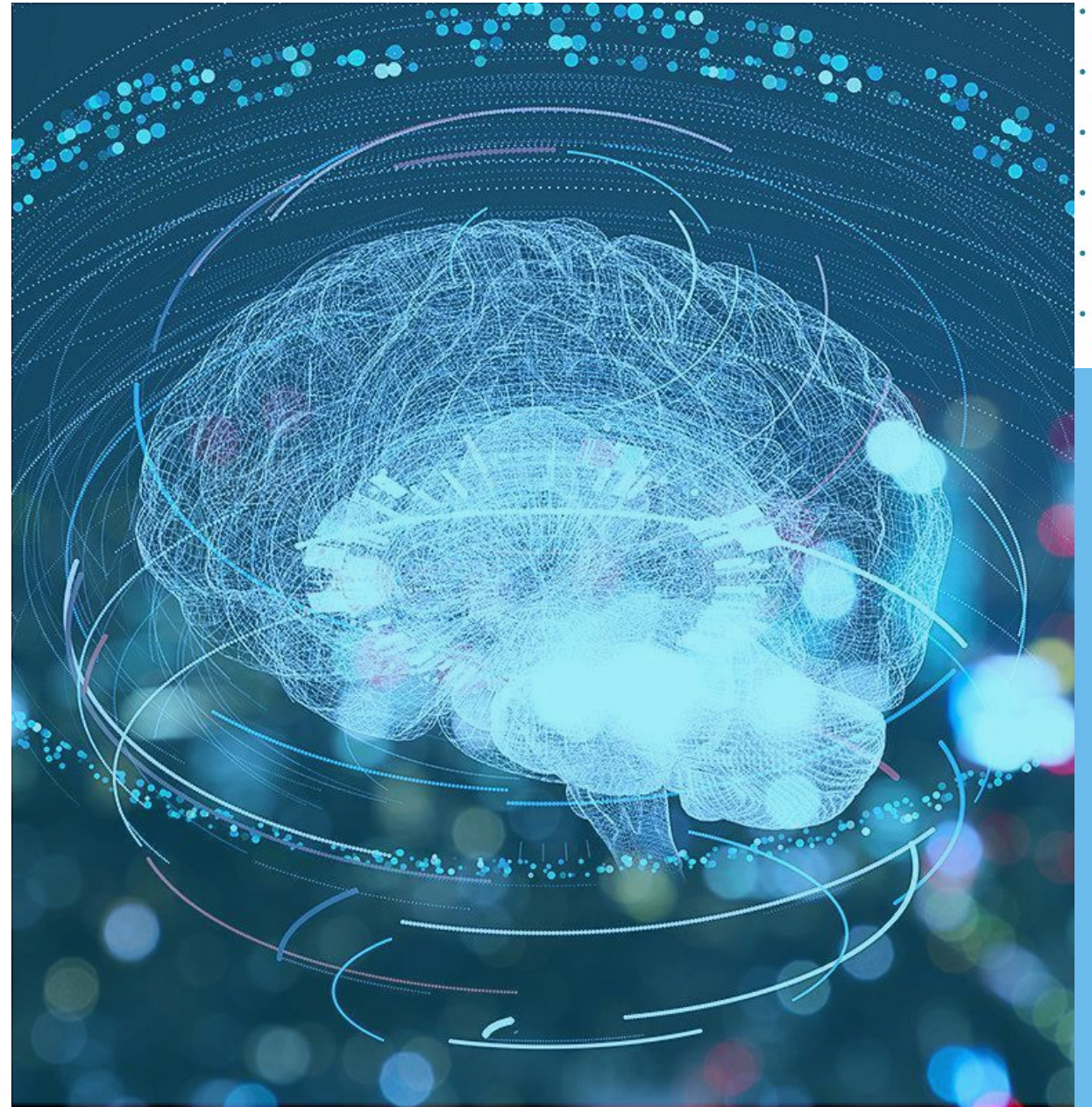


Обучаем нейросеть

- Предобрабатываем данные
- Моделируем нейросеть
- Задаем функцию потерь в соответствии с вознаграждением
- Эффективно определяем вознаграждение
- Тренируем сеть
- Запускаем игру с обученной сетью

Как ускорить обучение с подкреплением

- GPU
- Сверточные слои
- Распараллеливание процессов

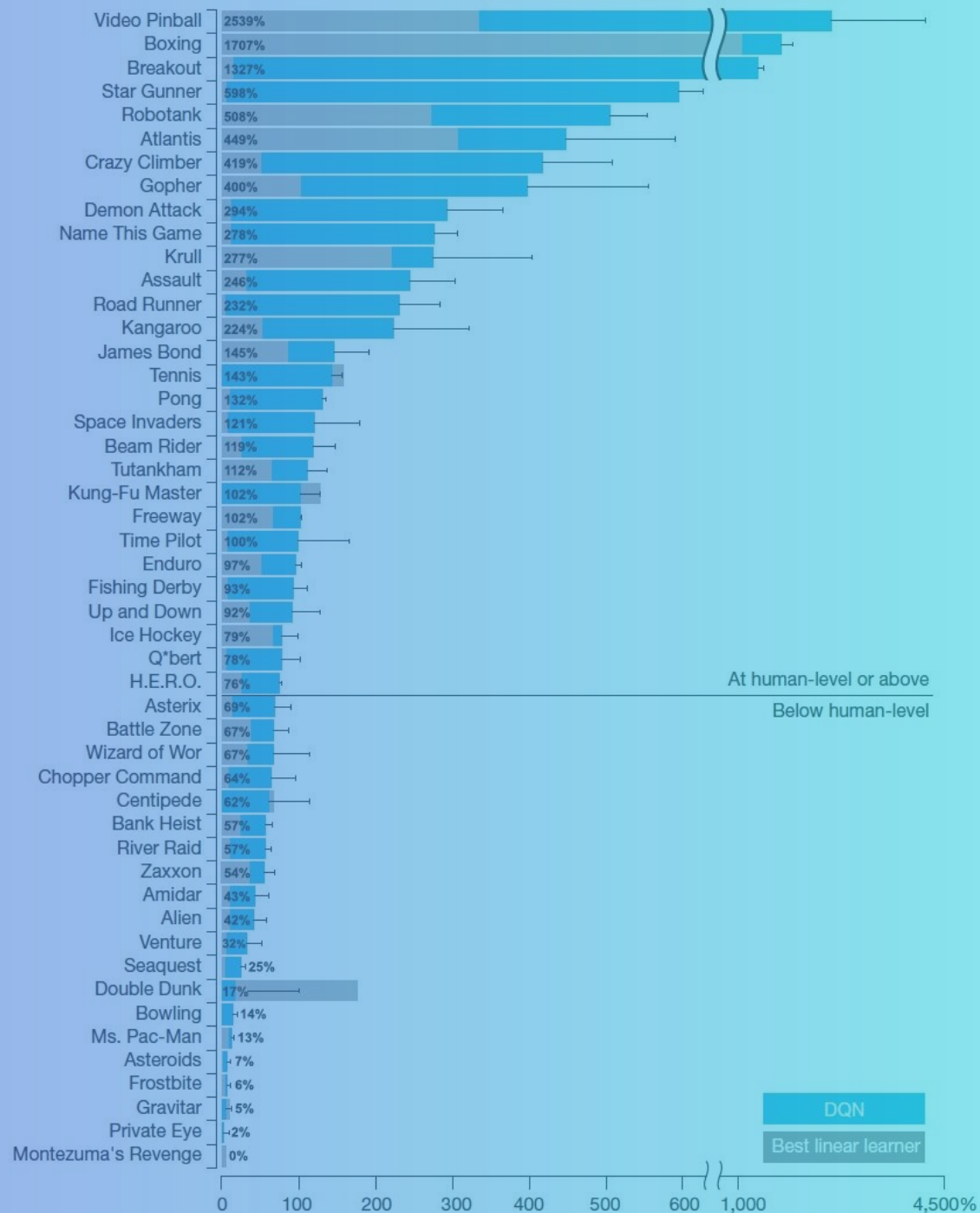


Ограничения и проблемы подхода

- Требуется много данных и запусков
- Долгосрочное и разреженное вознаграждение
- Плохо обучается на очень больших сетках
- Разведка против эксплуатации (зацикливание в локальном оптимуме, доволен текущей наградой)

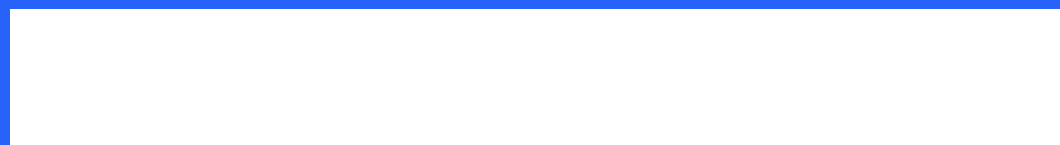


Что сегодня?





Спасибо



За внимание

