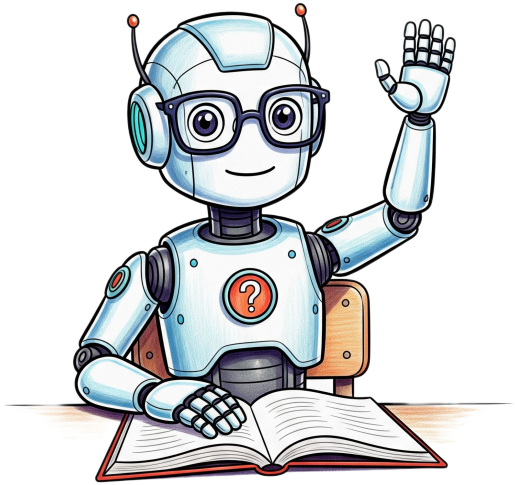


Evolution: RL alternative

28 сентября 2025

О чем сегодня поговорим?

- ➔ Познакомимся с ещё одним методом решения задач последовательного принятия решений
- ➔ Попробуем его на практике

Recap. Monte-Carlo

2) **Обновление:**

   Для каждой пары (s_t, a_t) в эпизоде:

   1. Вычислить возврат(ы) для этого посещения: $G(s_t, a_t) = \sum_{k=0}^{T-t} \gamma^k r_{t+k}$
   2. Вычислить новую оценку: $\hat{Q}(s_t, a_t) \leftarrow \text{Average}[G(s_t, a_t)]$
   3. Плавно обновляем Q: $Q(s_t, a_t) \leftarrow (1 - \alpha) Q(s_t, a_t) + \alpha \hat{Q}(s_t, a_t)$

1) Используем симулятор для сбора данных: $\tau = [(s_0, a_0, r_0), (s_1, a_1, r_1), \ldots, (s_T, a_T, r_T)]$.
2) Обновляем

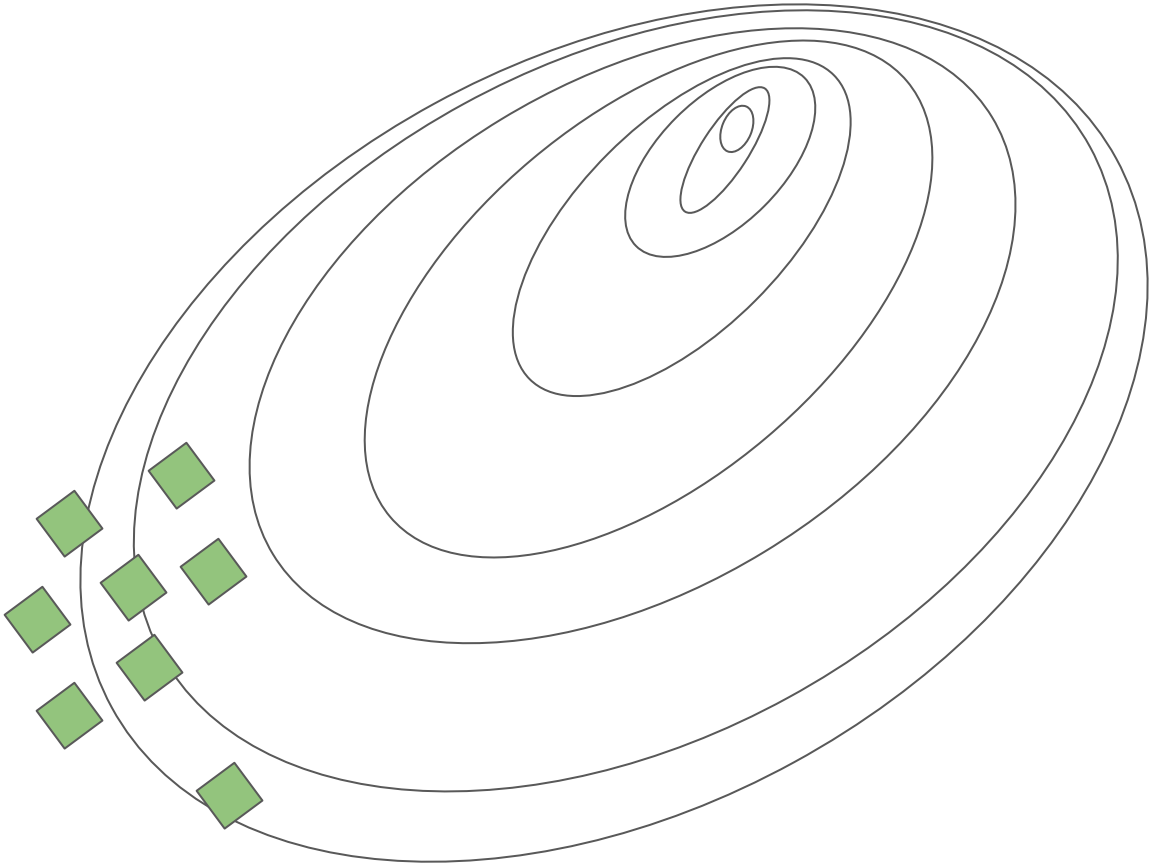
Evolution Strategies

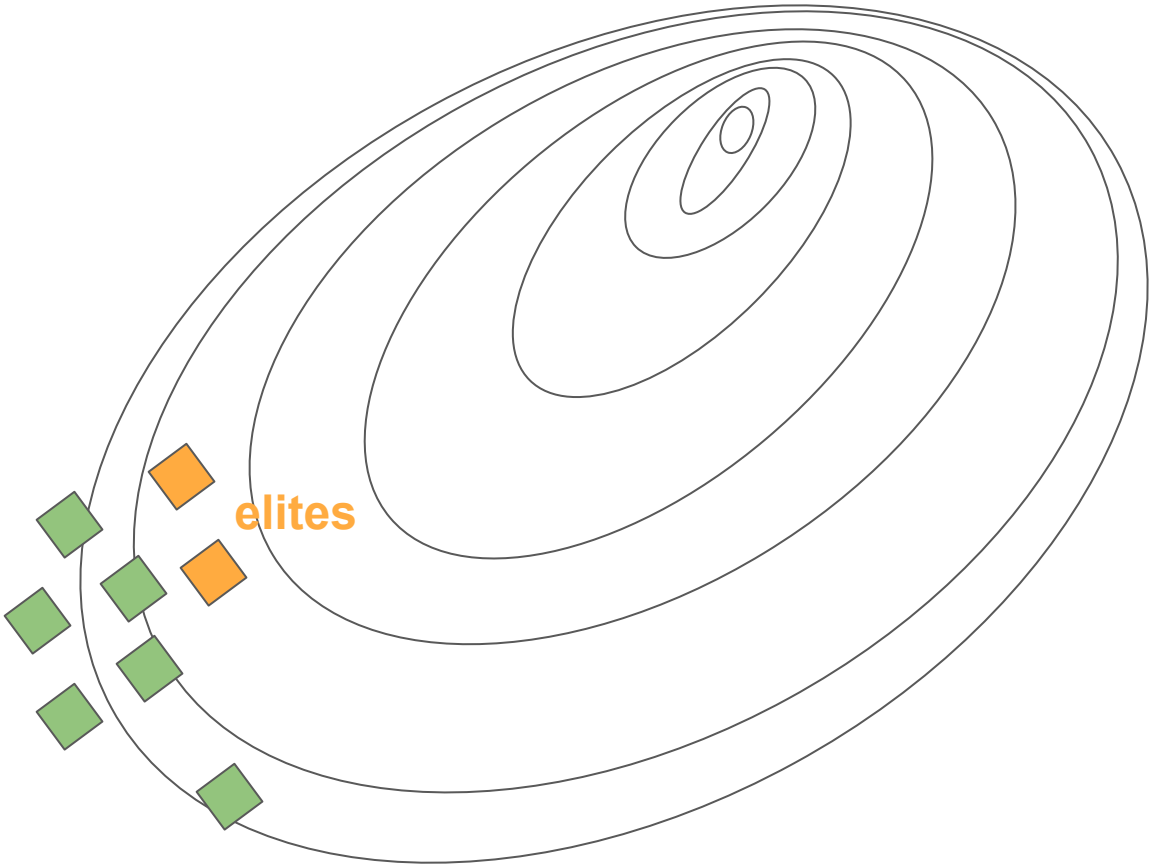
Что такое ES?

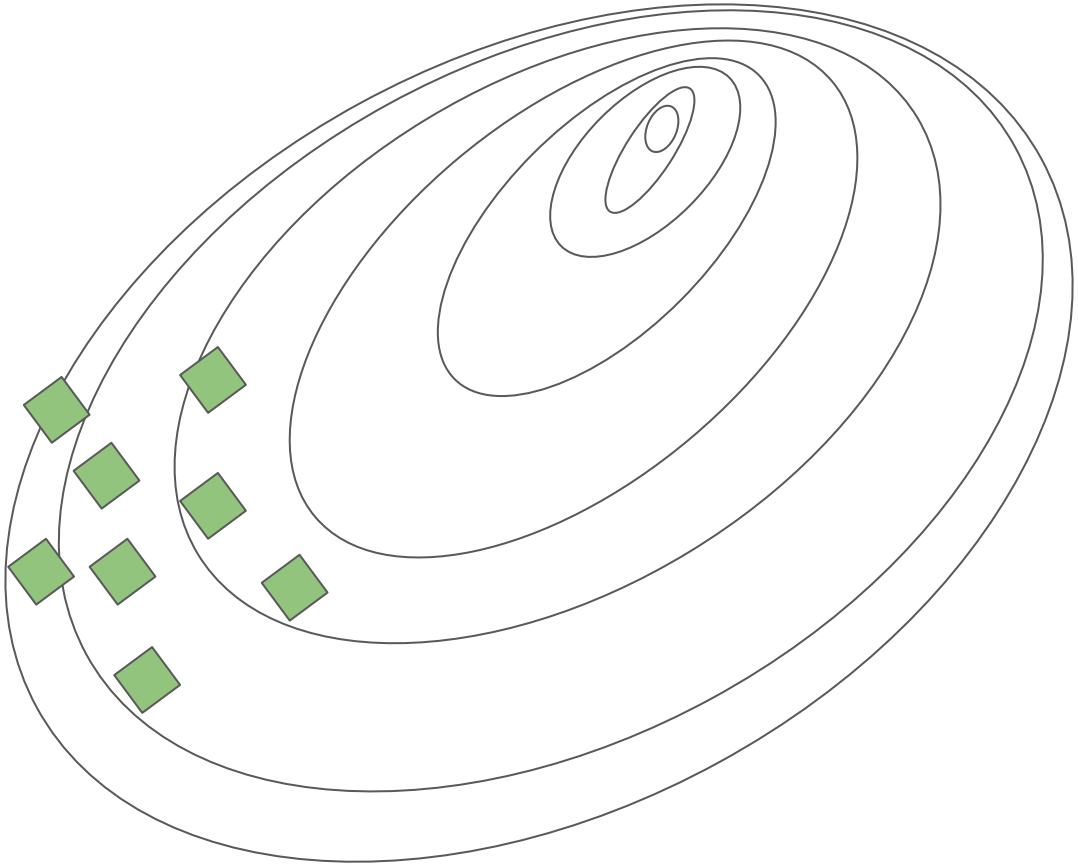
Метод оптимизации, вдохновлённый биологической эволюцией, использующий популяцию решений и их мутации для поиска оптимальных параметров.

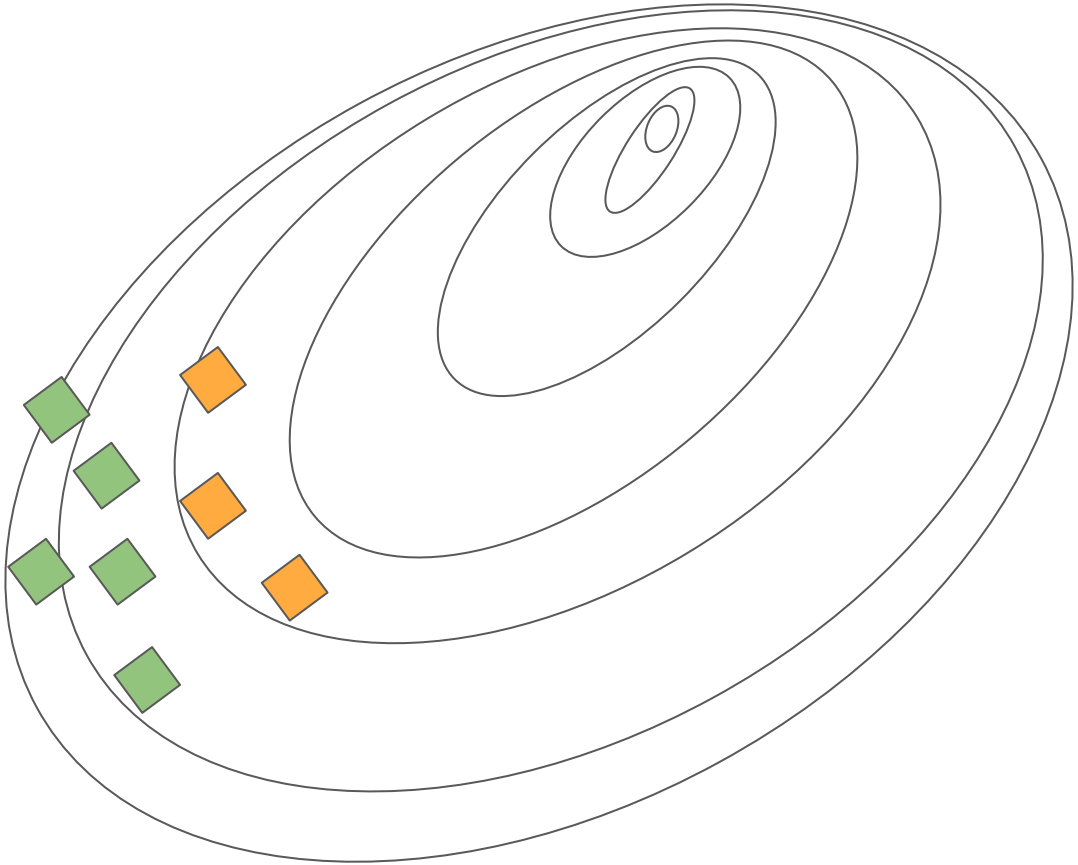
Преимущества ES:

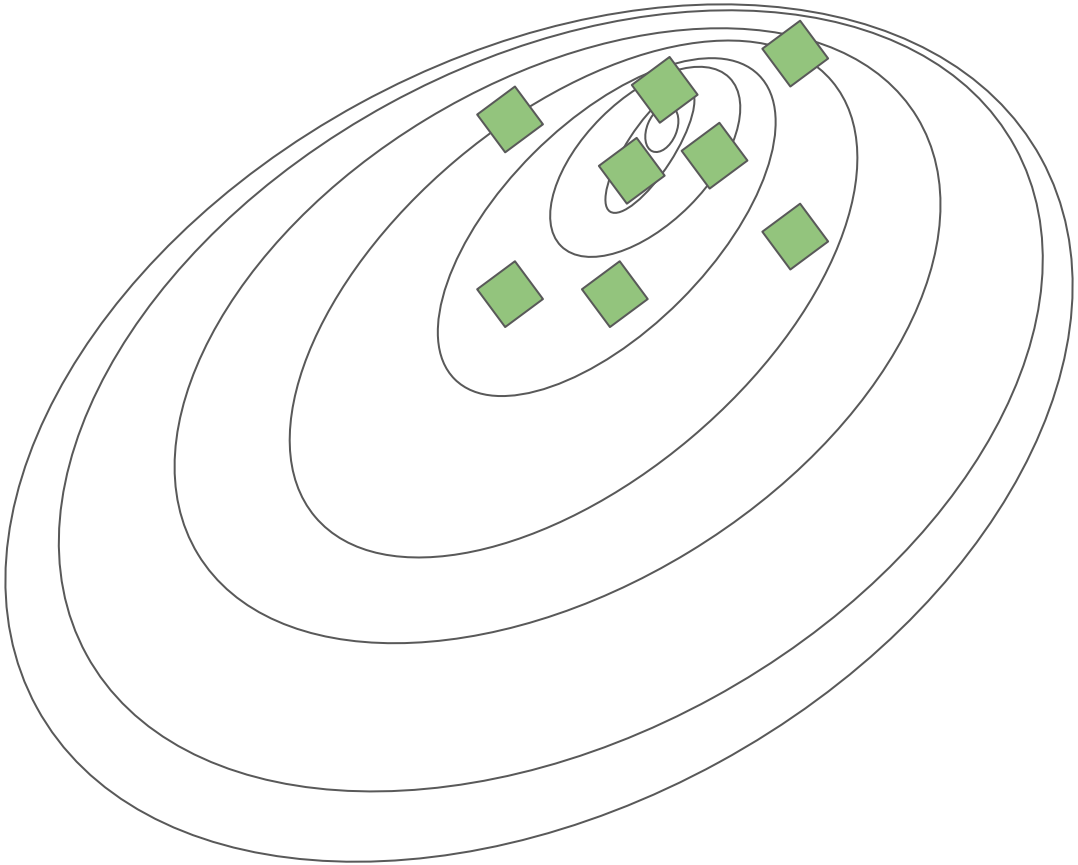
- Не требует обратного распространения ошибки (backpropagation).
- Легко масштабируется в распределённых системах.
- Эффективен при разреженных вознаграждениях.
- Меньше гиперпараметров.

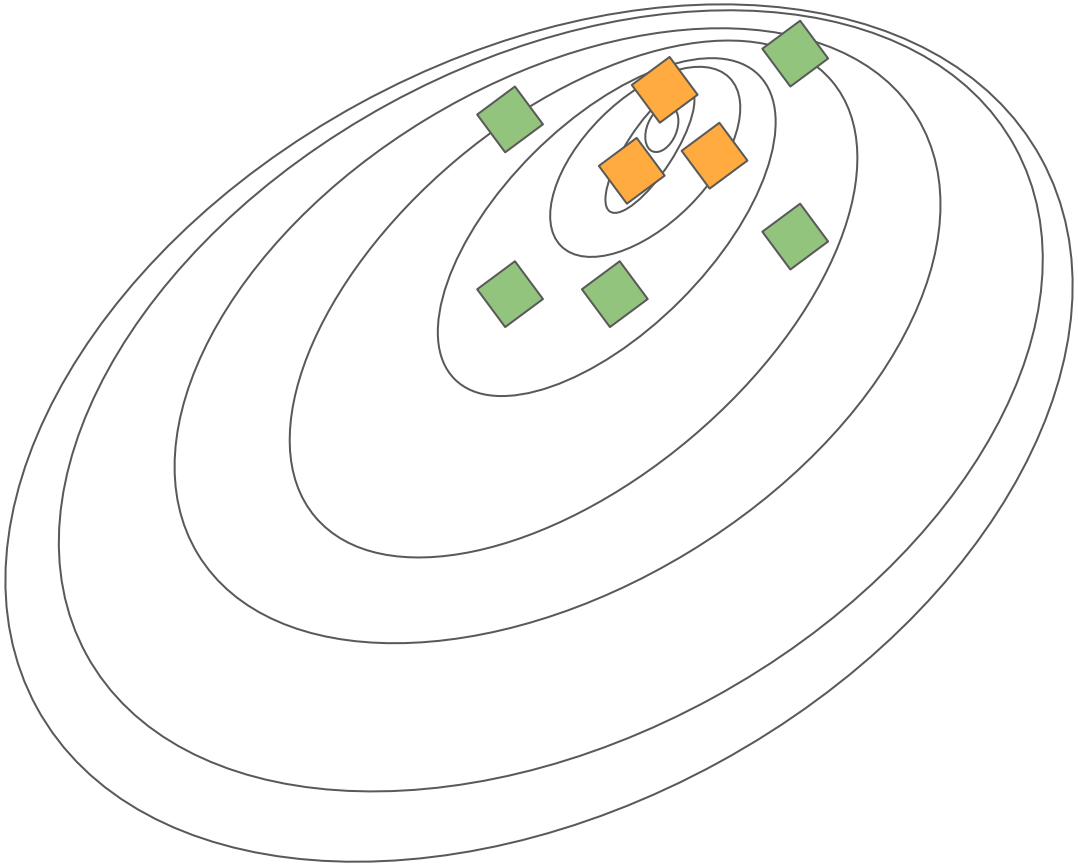


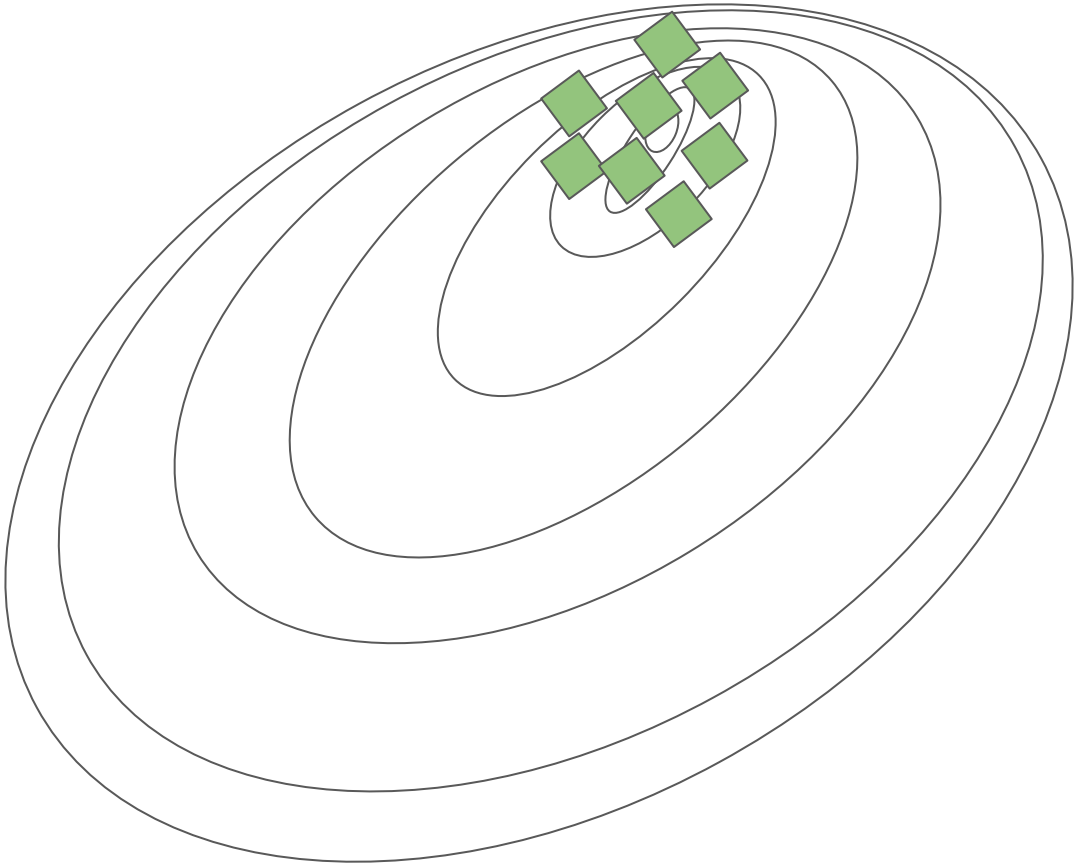

elites



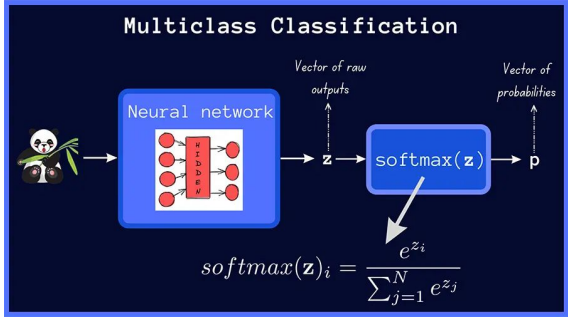








Cross Entropy! (recap)



На основе датасета: [(вопрос: ответ)]

Учимся по вопросу предсказывать ответ

На основе состояния хотим предсказывать действия

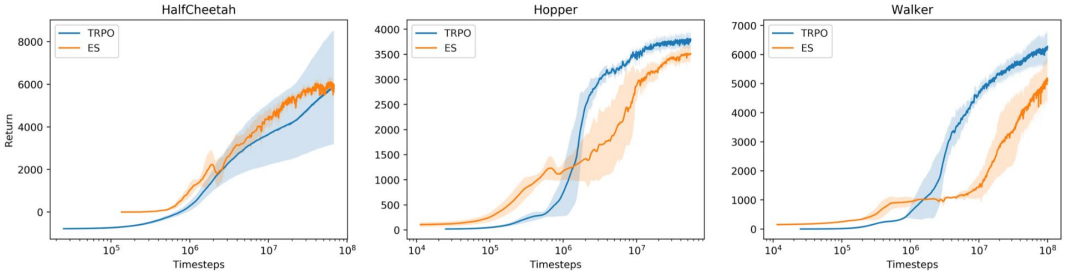
Cross Entropy Method

1) Используем симулятор для сбора траекторий:
   1) Траектория $\tau = [(s_0, a_0), (s_1, a_1), \ldots, (s_T, a_T)]$.
   2) Для каждой τ запоминаем суммарную награду R
2) Обновляем:
   1) Выбираем N наилучших траекторий
   2) Обучаем политику на них

классификатор (s, a) на этих траекториях

https://openai.com/index/evolution-strategies/

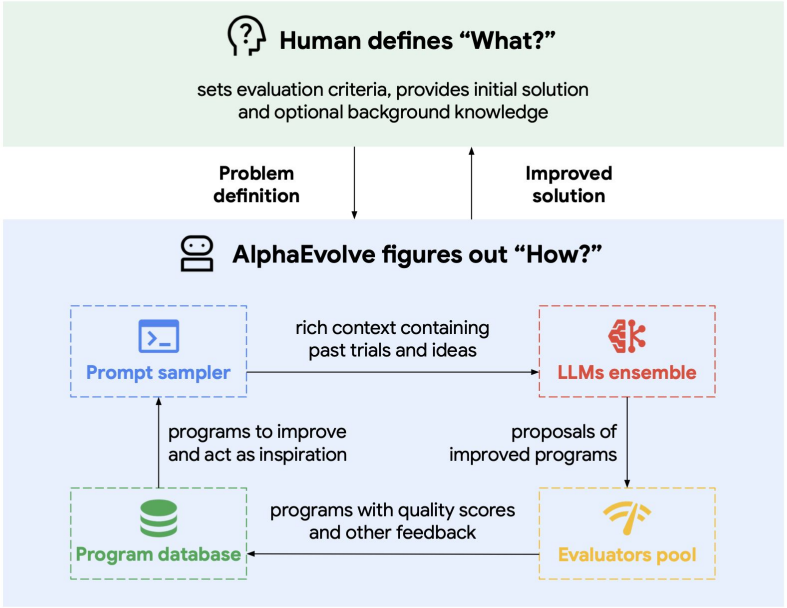
ES vs RL



- **Обучение с подкреплением:**
  - Использует градиентный спуск и обратное распространение ошибки.
  - Требует вычислительных ресурсов и сложных настроек.
- **Стратегия эволюции:**
  - Простота реализации и масштабируемость.
  - Сравнимая производительность с RL на современных бенчмарках (например, Atari/MuJoCo).
- **Результаты:**
  - Обучение 3D-агента в MuJoCo за 10 минут на 1440 CPU-ядрах.
  - Сравнимая производительность с A3C на Atari при сокращении времени обучения с 1 дня до 1 часа.

https://deepmind.google/discover/blog/alphaevolve-a-gemini-powered-coding-agent-for-designing-advanced-algorithms/

AlphaEvolve



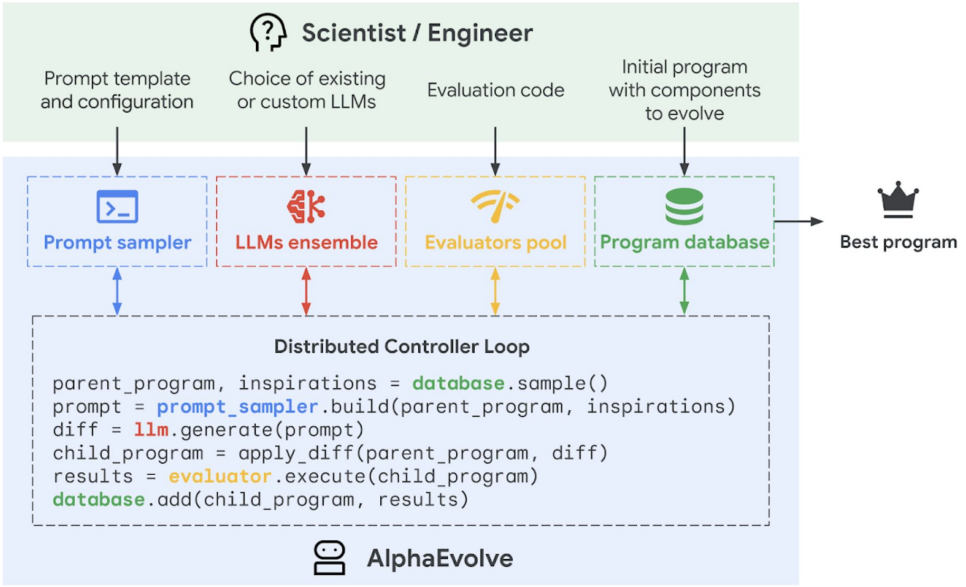


Figure 1 | *AlphaEvolve* high-level overview.

Как работает по шагам:

Инициализация популяции: создаем множество кандидатов (параметров или политик)

1. **Оценка качества**:

   считаем, насколько каждый кандидат «успешен»

2. **Выбор лучших**:

   выбираем топ-кандидатов для формирования нового поколения

3. **Обновление**:

   на основе лучших кандидатов обновляем параметры генерации кандидатов

Повторение цикла: генерируем новые кандидаты и повторяем шаги 2–4 до сходимости.

Итоги

- **Выводы**:
  - ES представляет собой мощный инструмент для оптимизации в задачах обучения с подкреплением.
  - Сравнимая производительность с традиционными методами при меньших вычислительных затратах.
- **Перспективы**:
  - Развитие методов ES для более сложных и высокоразмерных задач.
  - Интеграция ES с другими методами оптимизации и обучения.
- Пора кодить!

Вопросы?