Машинное обучение

Лекция 1
Введение в машинное обучение
(на основе лекций Евгения Соколова)

Сергей Корпачев
korpachev.ss@phystech.edu
sskorpachev@gmail.com

НИУ ВШЭ, 2026

Как перевести часы в минуты?

Как перевести часы в минуты?

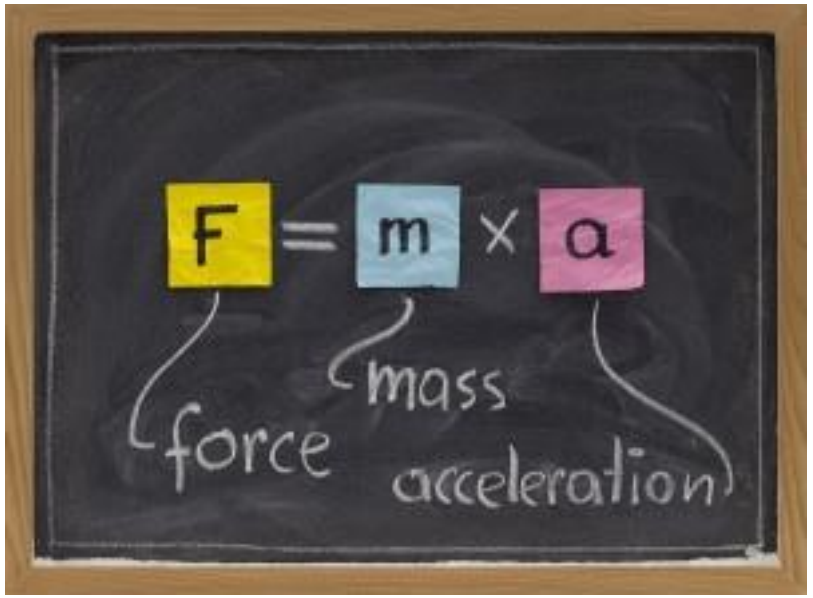
- x — часы
- $f(x) = 60x$ — преобразование в минуты, функция

Какая сила приложена к телу?

- Известны масса тела m и его ускорение a
- Чему равна сила F?

Какая сила приложена к телу?

- Известны масса тела m и его ускорение a
- Чему равна сила F?
- Второй закон Ньютона: $F = ma$



Как предсказать погоду?

Уравнения Навье-Стокса

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} + w\frac{\partial u}{\partial z} = -\frac{\partial P}{\partial x} + Re\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}\right),$$

$$\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} + w\frac{\partial v}{\partial z} = -\frac{\partial P}{\partial y} + Re\left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} + \frac{\partial^2 v}{\partial z^2}\right),$$

$$\frac{\partial w}{\partial t} + u\frac{\partial w}{\partial x} + v\frac{\partial w}{\partial y} + w\frac{\partial w}{\partial z} = -\frac{\partial P}{\partial z} + Re\left(\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} + \frac{\partial^2 w}{\partial z^2}\right),$$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0.$$

Уравнения Навье-Стокса

Дифференциальные уравнения

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} + w\frac{\partial u}{\partial z} = -\frac{\partial p}{\partial x} + Re\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}\right),$$

$$\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} +$$

Позволяют найти скорость воздуха и давление в любой точке

$$\frac{\partial w}{\partial t} + u\frac{\partial w}{\partial x} + v\frac{\partial w}{\partial y} +$$

Очень тяжело решать

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0.$$

Анализ тональности текста

- Какой эмоциональный окрас имеет текст?
- Варианты: позитивный, нейтральный, негативный
- Применение: автоматический анализ отзывов от пользователей

Анализ тональности текста

«*Большое спасибо! Сюдя по всему, это как раз то, чего не хватает всем зарубежным курсам по Machine Learning и Knowledge Discovery. Это теория, математика, объяснение того, как оно устроено “в кишках”.*»

Какой окрас?

Анализ тональности текста

«Я вижу очень большой минус, что курс будет на готовой библиотеке sci-kit. Курс от Andrew лучше тем, что ученик сам пишет алгоритм и видит изнутри, как он работает.»

Какой окрас?

Анализ тональности текста

- x — текст на русском языке
- $f(x)$ — его окрас (принимает значения -1, 0, 1)
- Можно ли выписать формулу для $f(x)$?

- На входе — вовсе не числа
- Точная зависимость может не существовать

Больше сложных задач!

- Какой будет спрос на товар в следующем месяце?
- Сколько денег заработает магазин за год?
- Вернет ли клиент кредит?

- Что изображено на картинке?
- Как перевести запись речи в текст?
- **Какое слово будет следующим в тексте?**

Больше сложных задач!

- Везде — очень сложные неявные зависимости
- Нельзя выразить их формулой
- Но есть некоторое число примеров
  - Тексты с известным окрасом
- Будем приближать зависимости, используя примеры

Машинное обучение

— это про то, как восстановить сложные зависимости

по конечному числу примеров.

- **Машинное обучение** – набор способов воспроизведения связей между событиями и результатом.

- **Машинное обучение** – обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться.

- **Machine learning** – the field of study that gives computers the ability to learn without being explicitly programmed.

Организационное

Команда курса:

- Лекции и семинары: Сергей Корпачев
- Ассистент: Дмитрий Денисенко

Про курс

- Планируется 16 лекций и 16 семинаров по “классическому” машинному обучению.

Что будем в курсе:

Популярные задачи машинного обученения

- Регрессия
- Классификация
- Ранжирование
- Кластеризация
- Снижение размерностей

Про оценку

$$О_{\text{итоговая}} = 0.4 * ДЗ + 0.1 * ПР + 0.2 * КР + 0.3 * Э$$

- Домашние задания (4 шт.)
- Проверочные работы (в начале лекций по теор. вопросам, 3 шт.)
- Контрольная работа (после 3 модуля)
- Письменный экзамен (ближе к концу 4 модуля)

Про курс: критерии автомата

- $\text{О}_{\text{итоговая}} = \text{О}_{\text{накопленная}} = \frac{10}{7}(0.4 * \text{ДЗ} + 0.1 * \text{ПР} + 0.2 * \text{КР})$

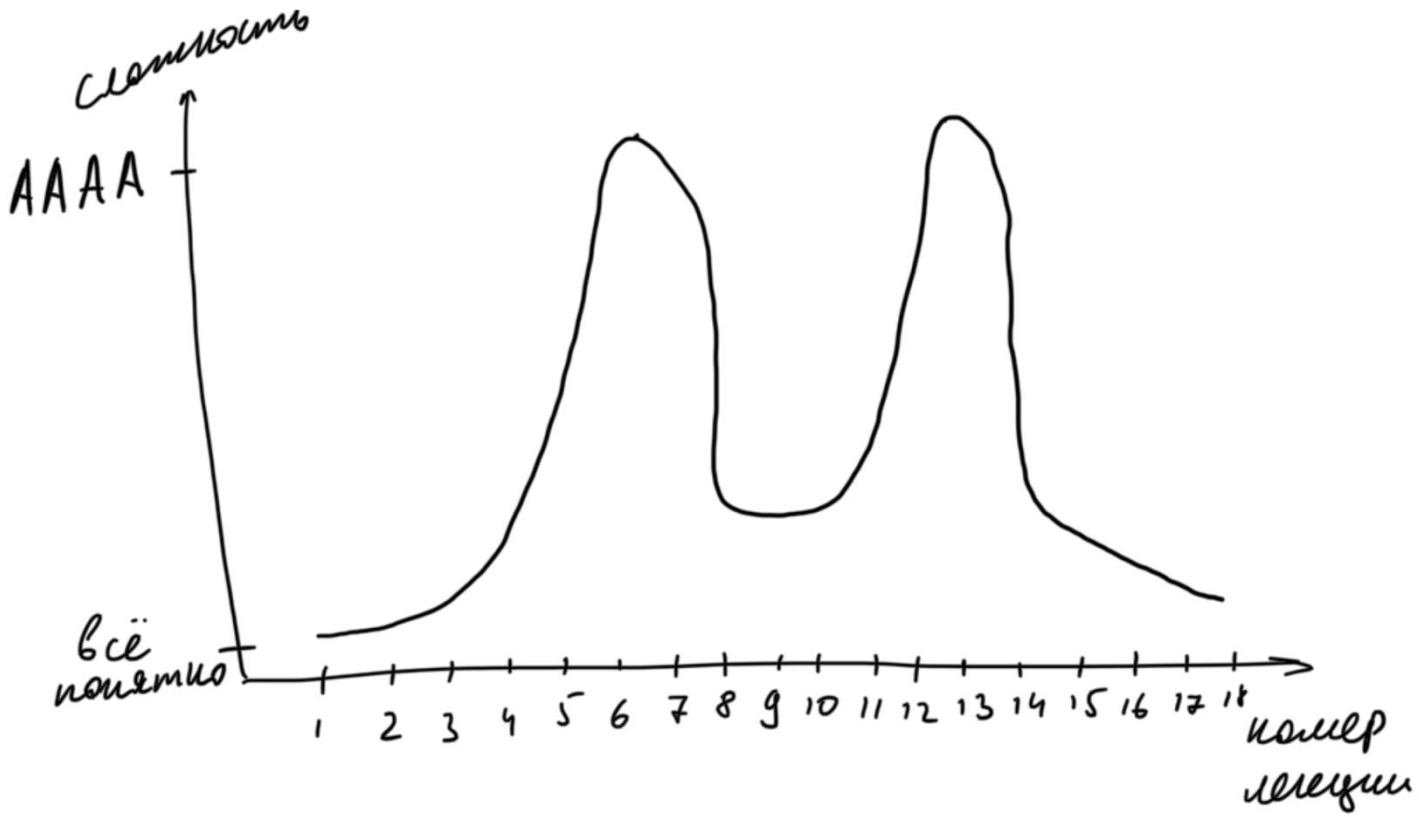
Можно получить автомат **при выполнении следующих условий:**

1. Накопленная оценка должна быть не ниже 5.5
2. Оценка за контрольную работу должна быть не ниже 5.5
3. Не был установлен факт плагиата ни одной домашней, проверочной и контрольной работ.

Про план курса

- Введение
- Метод k ближайших соседей
- Линейные методы
- Решающие деревья и случайные леса
- Градиентный бустинг
- Кластеризация
- …

Про план курса



Про литературу

- Luis Pedro Coelho and Willi Richert. Building Machine Learning Systems with Python.
- Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. An Introduction to Statistical Learning.
- Mohammed J. Zaki, Wagner Meira Jr. Data Mining and Analysis. Fundamental Concepts and Algorithms.

Про литературу

- Курсы ПМИ ФКН:
  - http://wiki.cs.hse.ru/Машинное_обучение_1
  - http://wiki.cs.hse.ru/Машинное_обучение_2
- Онлайн-курсы:
  - https://www.coursera.org/learn/**machine-learning**
  - https://openedu.ru/course/hse/INTRML/

Что нам пригодится?

Математический анализ

- Производные
- Частные производные
- Градиент

Что нам пригодится?

Линейная алгебра

- Векторы и матрицы
- Нормы, метрики, скалярное произведение
- Умножение матриц
- Обращение матриц
- Собственные числа и собственные векторы

Что нам пригодится?

Теория вероятностей и статистика

- Можно и обойтись

Но если не лень разбираться:

- Основные дискретные и непрерывные распределения
- Математическое ожидание, дисперсия, моменты
- Ковариация и корреляция

Что нам пригодится?

Писать код на Python, помнить про ООП

- Это всегда больно, нужны время и практика, чтобы привыкнуть
- Семинарист и ассистент помогут!

Что будет потом?

- Основы глубинного обучения
  - Общие принципы работы и обучения нейронных сетей
  - Свёрточные нейронные сети
  - Задачи компьютерного зрения
  - Нейронные сети для последовательностей
- Прикладные задачи анализа данных
  - Задачи NLP
  - Работа со звуком
  - Генеративные модели
  - Рекомендательные системы
  - Временные ряды
  - Основы DevOps

О чём ещё помнить?

- Мы проверяем домашние задания на плагиат
- Важно наладить контакт с ассистентами по поводу проверки!
- **Не забывайте про дедлайны**
- 9 и 10 — это очень высокие оценки
- Делать не всё из домашних заданий — нормально
- Задавать любые вопросы — не стыдно
- Тратить много времени и немножко мучиться — нормально
  - Но это окупится!

Основные термины

Пример задачи

- Сеть ресторанов
- Хотим открыть еще один
- Несколько вариантов размещения
- Какой из вариантов принесет максимальную прибыль?

* см. kaggle.com, TFI Restaurant Revenue Prediction

Обозначения

- x — объект — для чего хотим делать предсказания
  - Конкретное расположение ресторана
- $\mathbb{X}$ — пространство всех возможных объектов
  - Все возможные расположения ресторанов
- y — ответ, целевая переменная, target — что предсказываем
  - Прибыль в течение первого года работы
- $\mathbb{Y}$ — пространство ответов — все возможные значения ответа
  - Все вещественные числа

Обучающая выборка

- Мы ничего не понимаем в экономике
- Зато имеем много объектов с известными ответами
- $X = (x_i, y_i)_{i=1}^{\ell}$ — обучающая выборка
- ℓ — размер выборки

Признаки

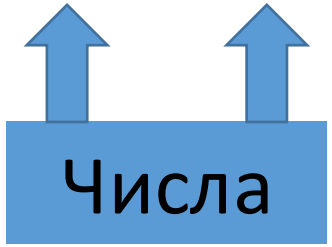
- Объекты — абстрактные сущности
- Компьютеры работают только с числами
- Признаки, факторы, features — числовые характеристики объектов
- d — количество признаков
- $x = (x_1, \dots, x_d)$ — признаковое описание

Признаки

- Объекты — абстрактные сущности
- Компьютеры работают только с числами
- Признаки, факторы, features — числовые характеристики объектов
- d — количество признаков
- $x = (x_1, \dots, x_d)$ — признаковое описание

Признаки

- Объекты — абстрактные сущности
- Компьютеры работают только с числами
- Признаки, факторы, features — числовые характеристики объектов
- d — количество признаков
- $x = (x_1, \dots, x_d)$ — признаковое описание



Признаки

- Про демографию:
  - Средний возраст жителей ближайших кварталов
  - Динамика количества жителей
- Про недвижимость:
  - Средняя стоимость квадратного метра жилья поблизости
  - Количество школ, банков, магазинов, заправок
  - Расстояние до ближайшего конкурента
- Про дороги:
  - Среднее количество машин, проезжающих мимо за день

Алгоритм

- $a(x)$ — алгоритм, модель — функция, предсказывающая ответ для любого объекта
- Отображает $\mathbb{X}$ в $\mathbb{Y}$
- Линейная модель: $a(x) = w_0 + w_1 x_1 + \cdots + w_d x_d$
- Например:

$$a(x) = 1.000.000 + 100.000 * (\text{расстояние до конкурента}) \\ -100.000 * (\text{расстояние до метро})$$

Функция потерь

- Не все алгоритмы полезны
- $a(x) = 0$ — не принесет никакой выгоды
- Функция потерь — мера корректности ответа алгоритма
- Предсказали $10000 прибыли, на самом деле $5000 — хорошо или плохо?
- Квадратичное отклонение: $(a(x) - y)^2$

Функционал ошибки

- Функционал ошибки, метрика качества — мера качества работы алгоритма на выборке
- Среднеквадратичная ошибка (Mean Squared Error, MSE):

$$\frac{1}{\ell}\sum_{i=1}^{\ell}(a(x_i) - y_i)^2$$

- Чем меньше, тем лучше

Функционал ошибки

- Должен соответствовать бизнес-требованиям
- Одна из самых важных составляющих анализа данных

Обучение алгоритма

- Есть обучающая выборка и функционал ошибки
- Семейство алгоритмов $\mathcal{A}$
  - Из чего выбираем алгоритм
  - Пример: все линейные модели
  - $\mathcal{A} = \{w_0 + w_1 x_1 + \cdots + w_d x_d \mid w_0, w_1, \ldots, w_d \in \mathbb{R}\}$
- Обучение: поиск оптимального алгоритма с точки зрения функционала ошибки

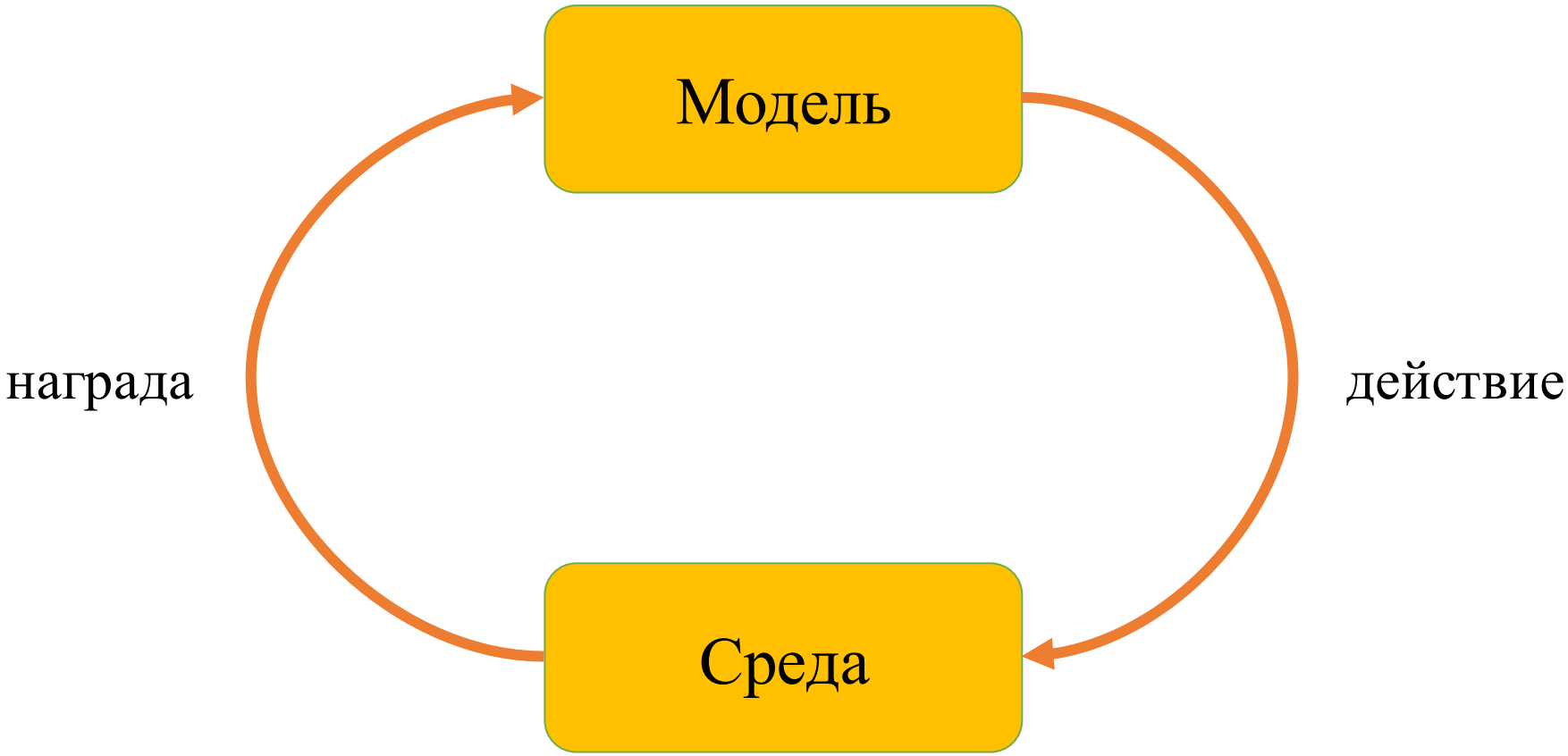
$$a(x) = \arg\min_{a \in \mathcal{A}} Q(a, X)$$

Машинное обучение

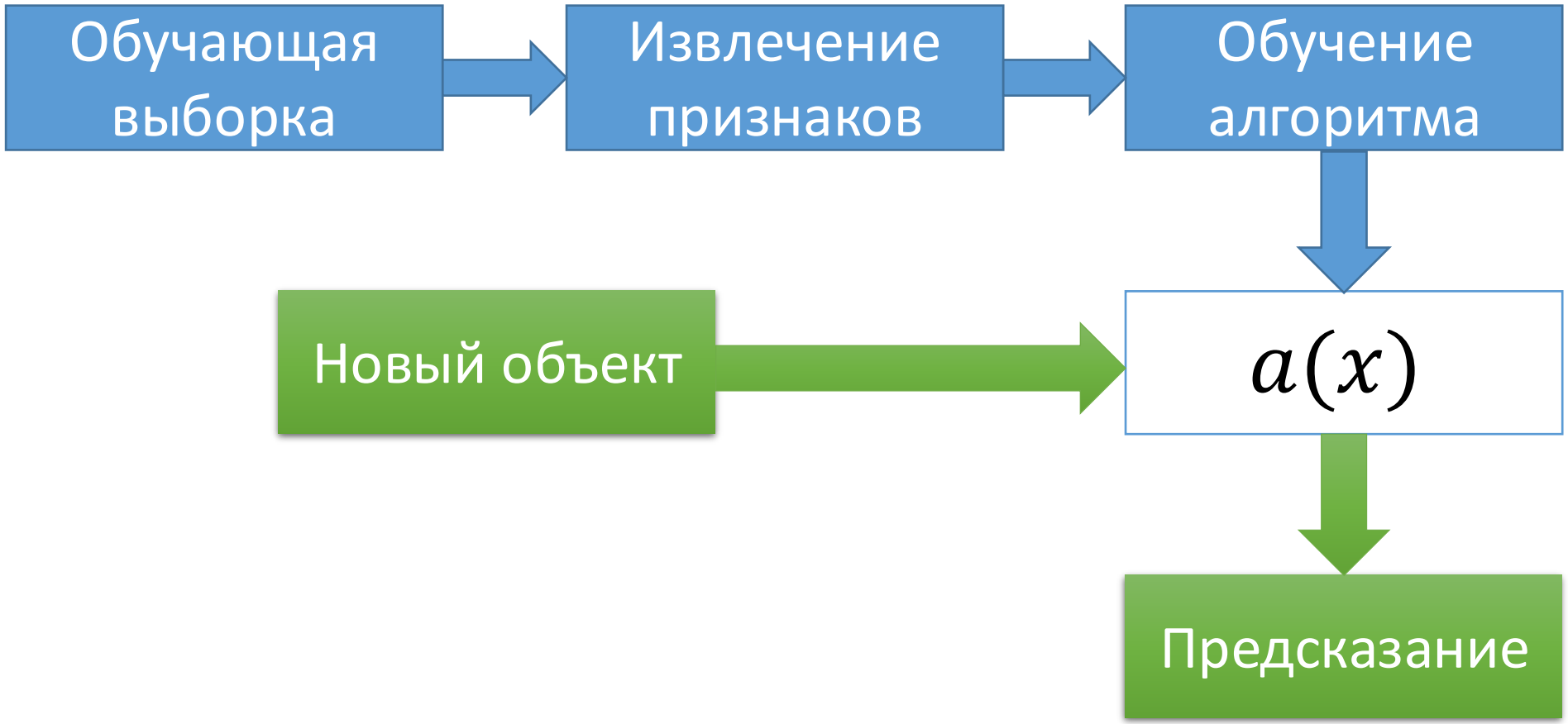
- Не все задачи имеют такую формулировку!

- Обучение без учителя
- Обучение с подкреплением
- И т.д.

Обучение с подкреплением



Машинное обучение



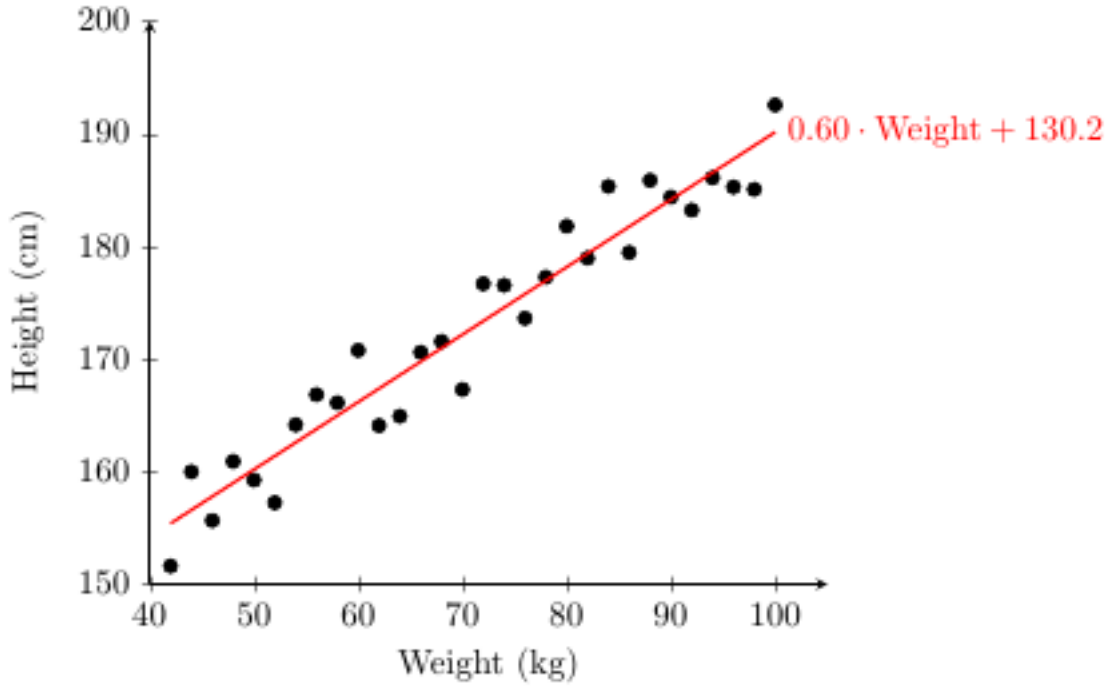
Что нужно знать

1. Как сформулировать задачу?
2. Какие признаки использовать?
3. Откуда взять обучающую выборку?
4. Как подготовить обучающую выборке?
5. Как выбрать метрику качества?
6. Как обучить алгоритм?
7. Как оценить качество алгоритма?
8. Как потом внедрить алгоритм и поддерживать его?

Типы ответов

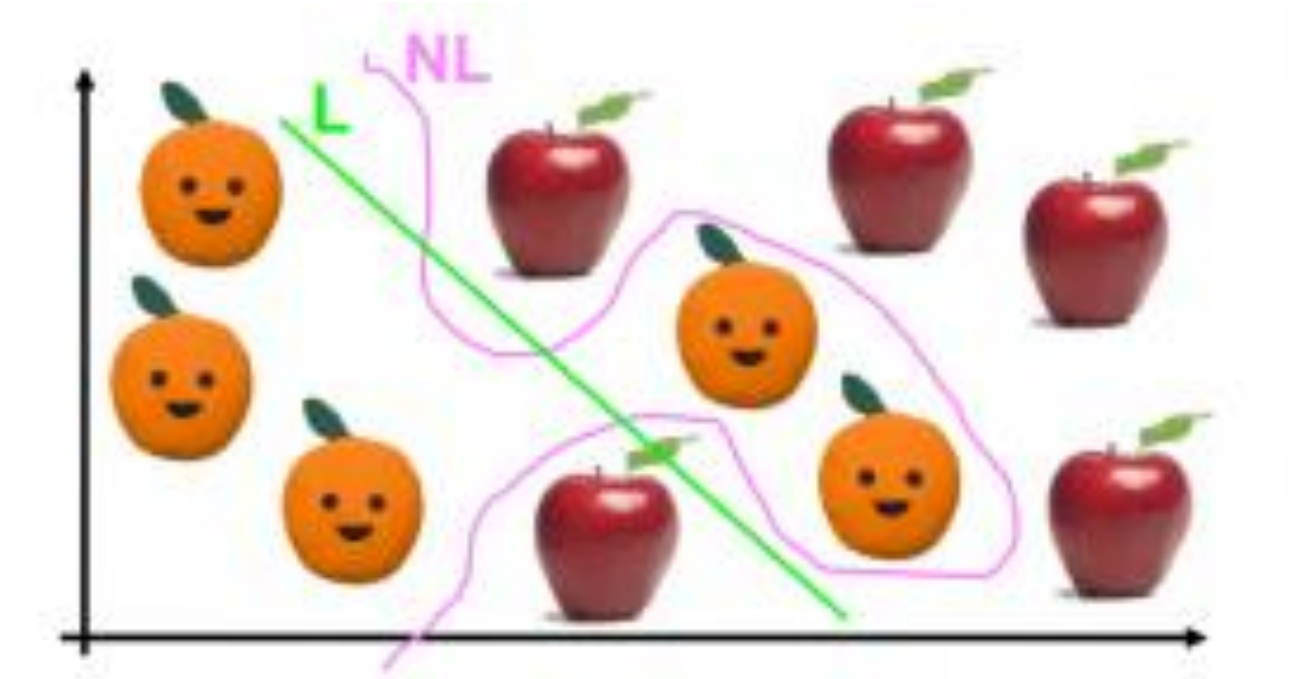
Регрессия

- Вещественные ответы: $\mathbb{Y} = \mathbb{R}$
- (вещественные числа — числа с любой дробной частью)
- Пример: предсказание роста по весу



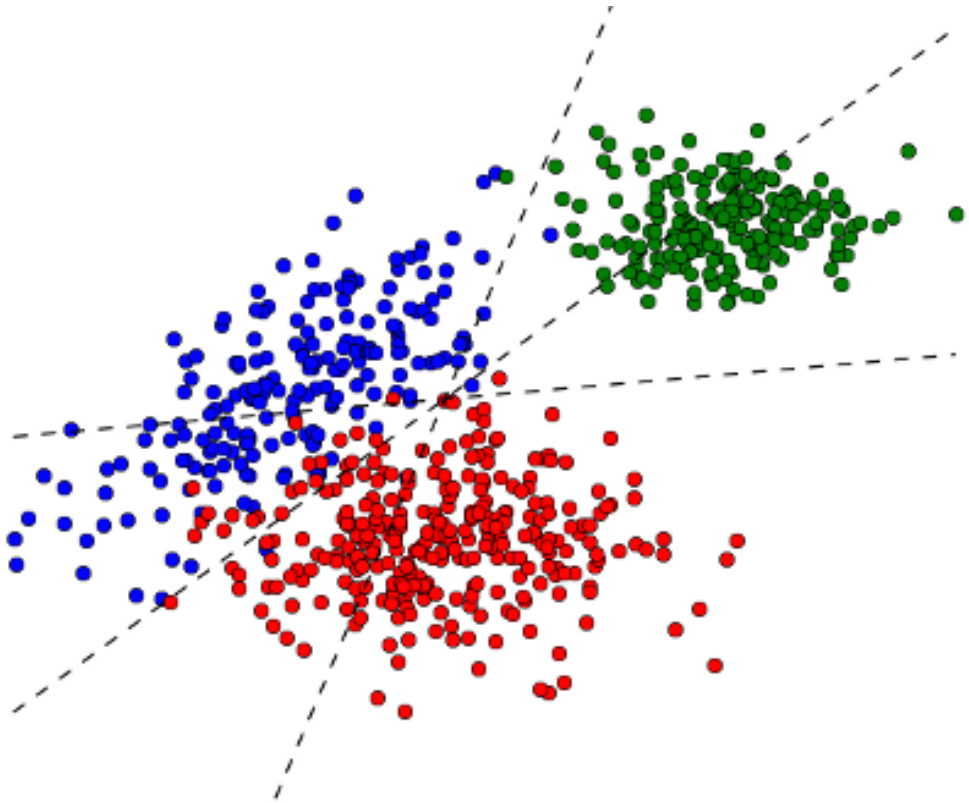
Классификация

- Конечное число ответов: $|\mathbb{Y}| < \infty$
- Бинарная классификация: $\mathbb{Y} = \{-1, +1\}$



Классификация

- Многоклассовая классификация: $\mathbb{Y} = \{1, 2, \dots, K\}$



Классификация

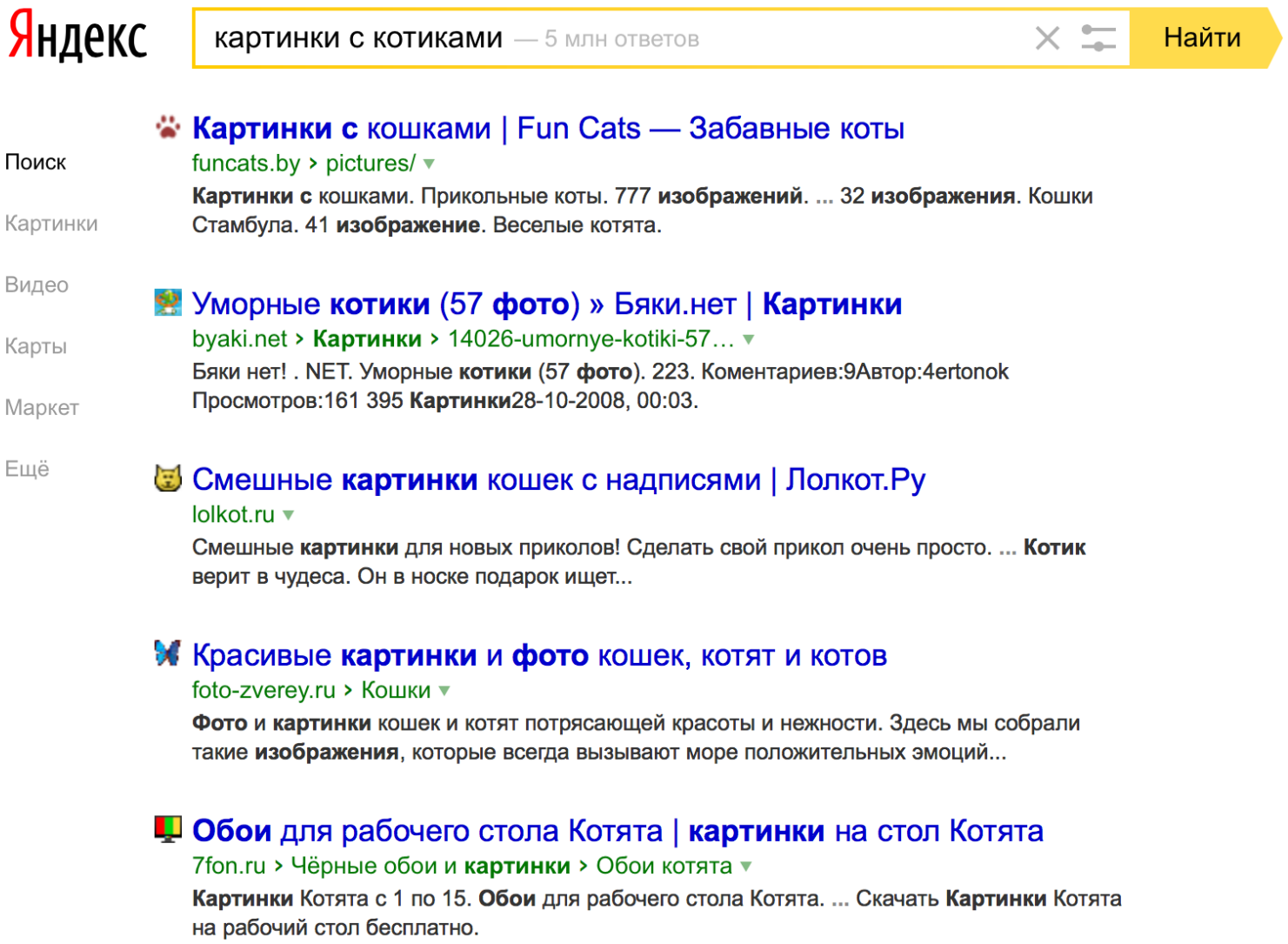
- Классификация с пересекающимися классами: $\mathbb{Y} = \{0, 1\}^K$
  - (multi-label classification)
- Ответ — набор из K нулей и единиц
- i-й элемент ответа — принадлежит ли объект i-му классу

- Какие темы присутствуют в статье?
- (математика, биология, экономика)

Ранжирование

- Набор документов $d_1, \ldots, d_n$
- Запрос q
- Задача: отсортировать документы по *релевантности* запросу
- $a(q, d)$ — оценка релевантности

Ранжирование



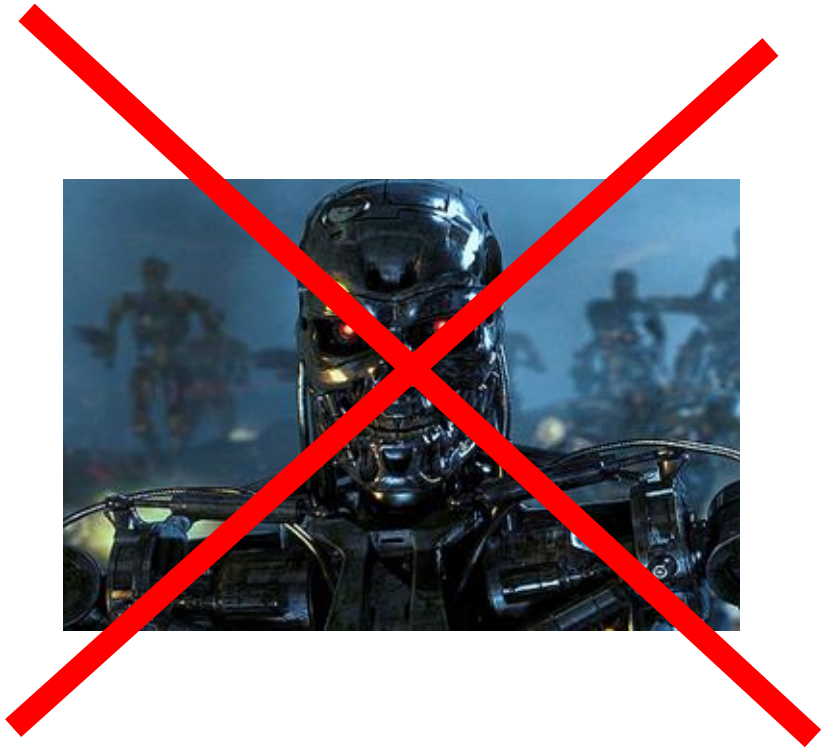
Кластеризация

- $\mathbb{Y}$ — отсутствует
- Нужно найти группы похожих объектов
- Сколько таких групп?
- Как измерить качество?

- Пример: сегментация пользователей мобильного оператора

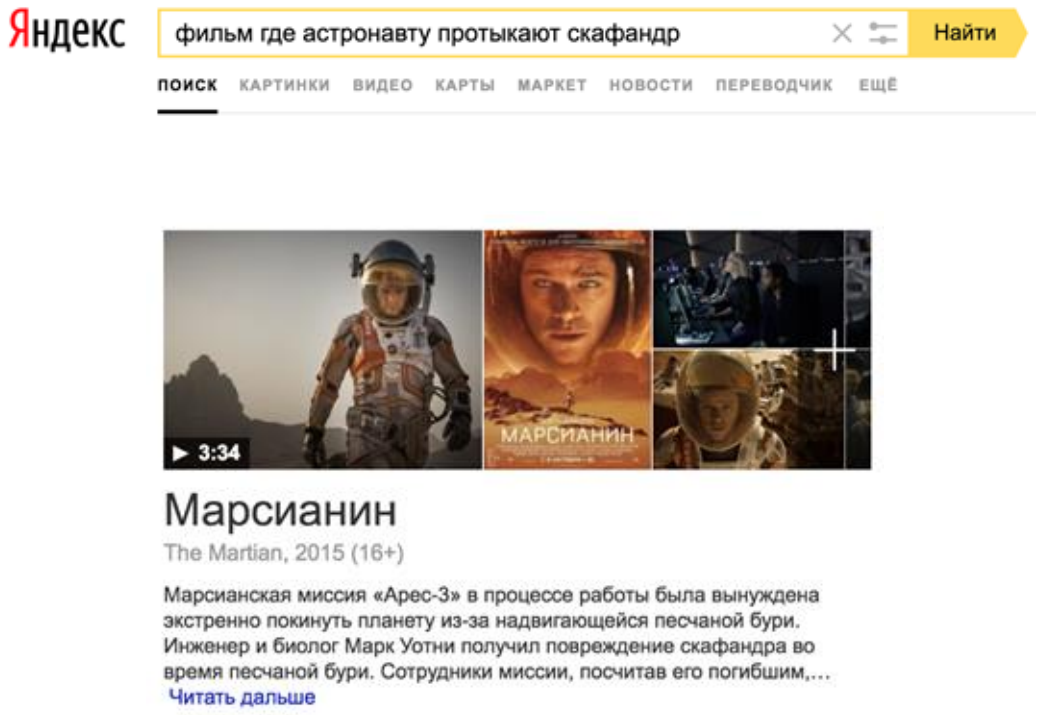
Зачем это нужно?

Искусственный интеллект



Сильный ИИ

через 20-100 лет



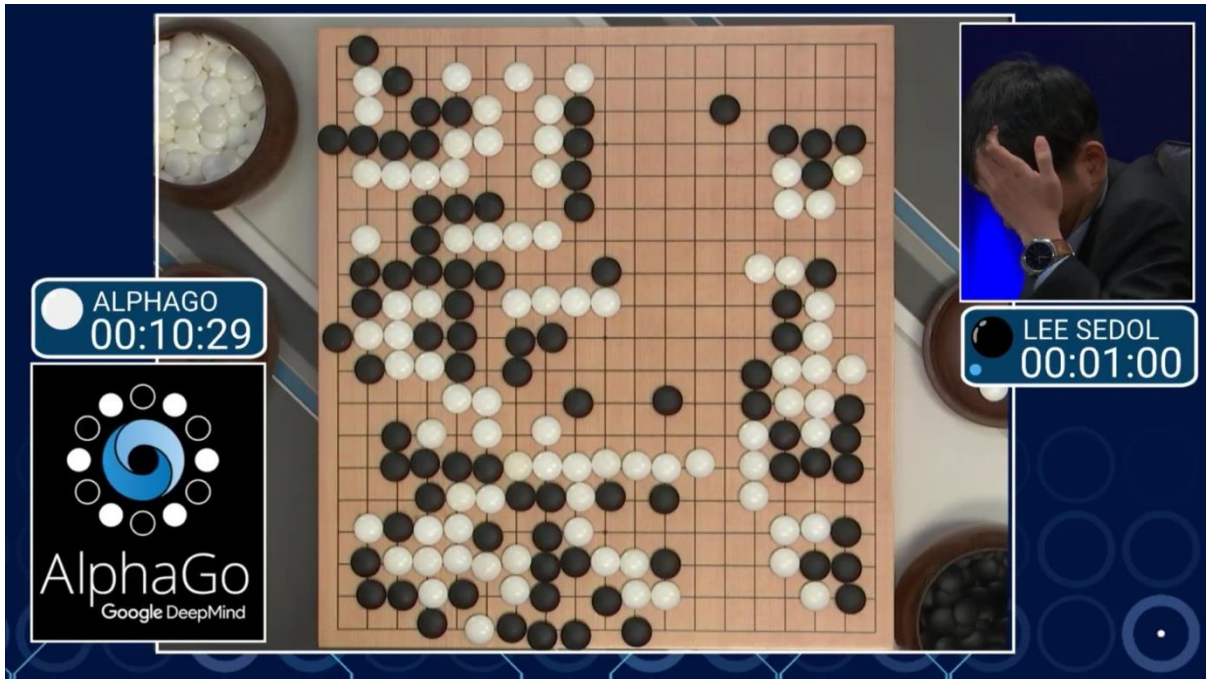
Специализированный ИИ

уже сейчас

AlphaGo

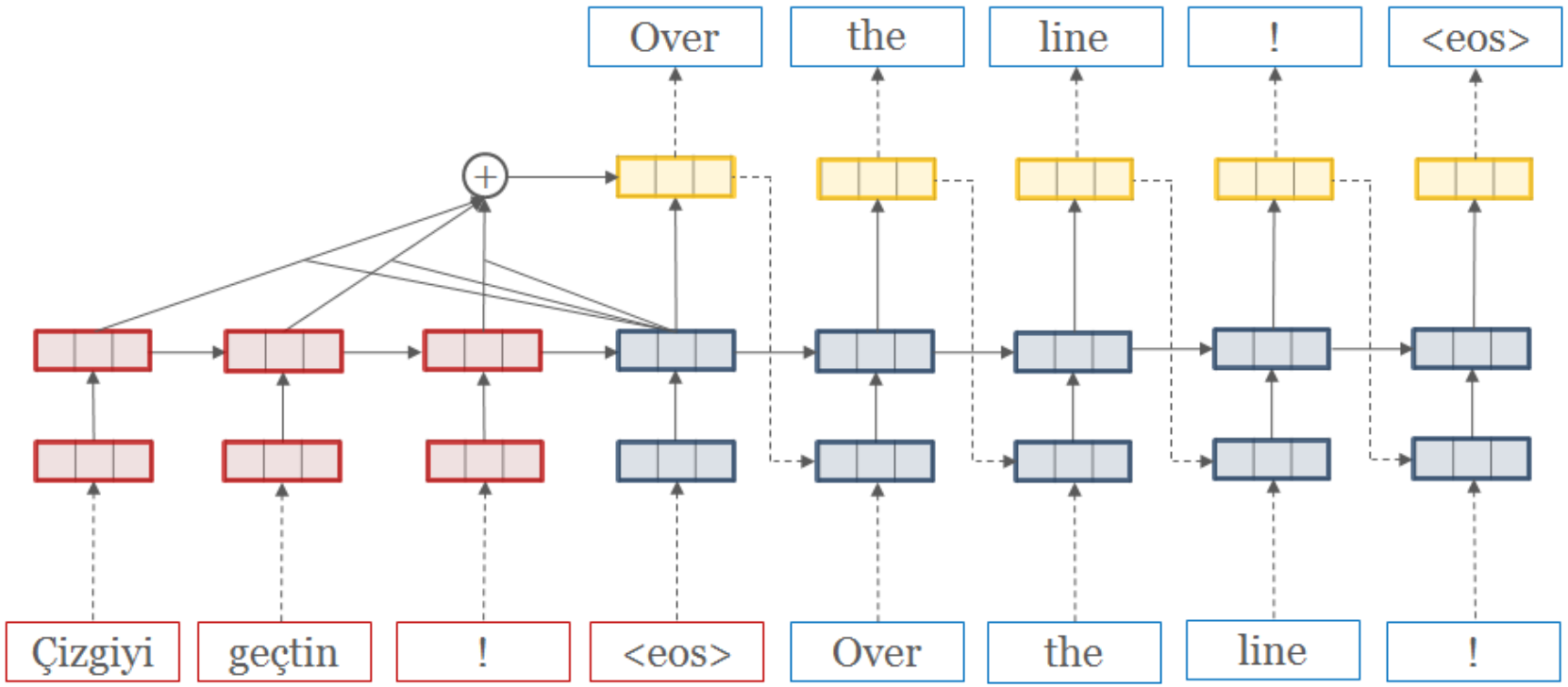
- Модель для игры в Го
- Оценивает успешность хода
- Обучалась путём игры с собой
- Победила чемпиона мира в 2016 году

- Долгое время игра в Го считалась невозможной задачей для компьютера



Перенос стиля

Машинный перевод

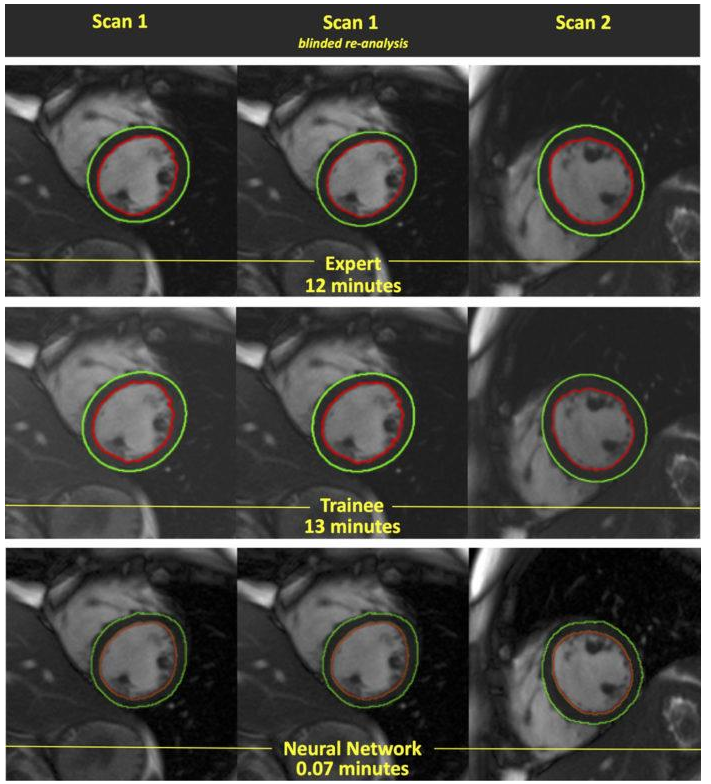


Генерация текста

- GPT-3 от OpenAI
- https://arxiv.org/abs/2005.14165
- https://talktotransformer.com

Биоинформатика и медицина

- Поиск связей между ДНК и заболеваниями (23andme и другие)
- Таргетные лекарства
- Анализ медицинских снимков



Сельское хозяйство

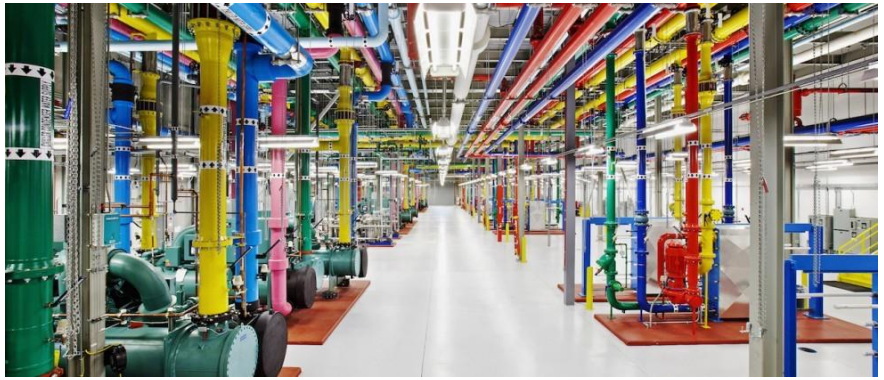
- Робототехника
- Мониторинг посевов и почвы
- Прогнозирование болезней и урожайности

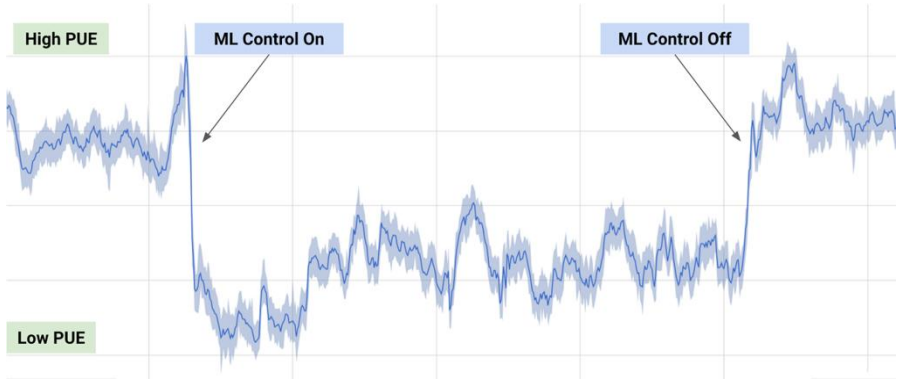
Машинное обучение в HR

- Поиск кандидатов и предсказание исхода собеседования
- Помощь при ротации
- Предсказание ухода сотрудника
- Анализ внутренних форумов, выделение жалоб

Автоматизация системы охлаждения

- Одна из главных компонент дата-центра — система охлаждения
- Результат работы системы сложным образом зависит от её параметров
- Необходимо быстро адаптироваться под изменение условий (нагрузка на серверы, погода)
- Все дата-центры разные — эвристические правила одного центра не работают в другом
- Машинное обучение позволило сократить затраты электричества на охлаждение на 40%





Рекомендательные системы

- Полки рекомендаций на Amazon генерируют 35% от всех покупок
- Рекомендации на основе машинного обучения и анализа больших объёмов данных

Зачем это нужно?

- Это круто
  - Сложные задачи
  - Движение к искусственному интеллекту
- Это полезно
  - Извлечение прибыли из данных
  - Data-driven companies

Как можно заниматься анализом данных?

- Data scientist
  - Работа с данными
  - Знание инструментов и методов
  - Опыт решения задач
- Менеджер
  - Понимание, как работает машинное обучение
  - Понимание узких мест, оценивание сроков
- Заказчик
  - Метрики качества
  - Требования к данным
  - Ограничения современных подходов

На следующей лекции

- Типы задач в машинном обучении
- Типы признаков
- Примеры задач
- Метод k ближайших соседей