Машинное обучение

Лекция 1: Введение

Антон Семёнкин

asemenkin@hse.ru | t.me/topshik

НИУ ВШЭ, 2021

Организационные моменты

- Чат в телеграме 
- Материалы занятий 
- Дополнительная информация  

Организационные моменты

- Задания
  - Домашние задания
  - Небольшие квизы
  - Финальный проект
- 10-ти бальная система
- Итоговая оценка:

$$О_{итог} = 0.6 * ДЗ + 0.2 * Квизы + 0.2 * Проект$$

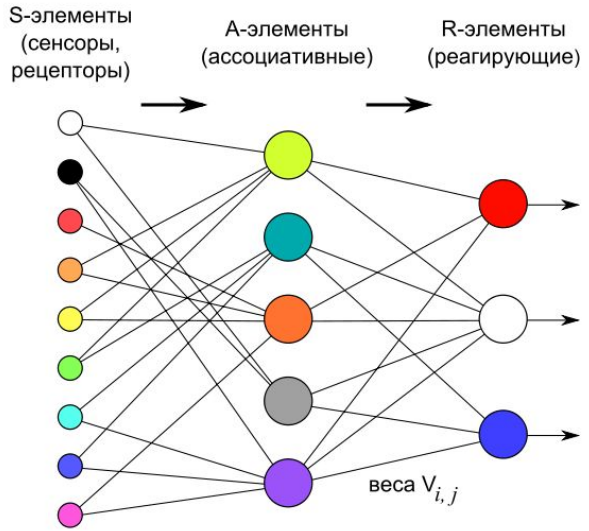
- Для зачёта: 6 и выше

Как всё было

- 1950-ые: первый семинар по проблемам ИИ
  - Задача: моделирование человеческого интеллекта

Как всё было

- 1950-ые: первый семинар по проблемам ИИ
  - Задача: моделирование человеческого интеллекта
- 1960-ые перцептрон Розенблатта



Как всё было

- 1950-ые: первый семинар по проблемам ИИ
  - Задача: моделирование человеческого интеллекта
- 1960-ые перцептрон Розенблатта
- 1980-ые: моделирование работы эксперта

Как всё было

- 1950-ые: первый семинар по проблемам ИИ
  - Задача: моделирование человеческого интеллекта
- 1960-ые перцептрон Розенблатта
- 1980-ые: моделирование работы эксперта
- 1990-ые: нейронные сети, бустинг
- 2000-ые: ядровые методы, обучение без учителя
- 2010-ые: всплеск нейросетевых методов

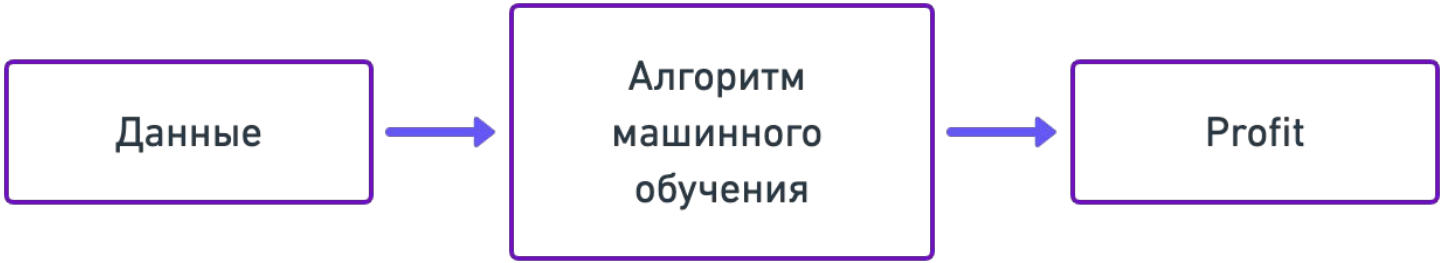
Как всё было

- 1950-ые: первый семинар по проблемам ИИ
  - Задача: моделирование человеческого интеллекта
- 1960-ые перцептрон Розенблатта
- 1980-ые: моделирование работы эксперта
- 1990-ые: нейронные сети, бустинг
- 2000-ые: ядровые методы, обучение без учителя
- 2010-ые: всплеск нейросетевых методов

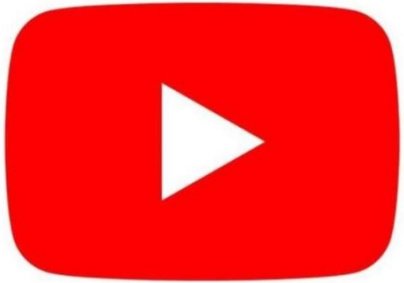
YOU ARE HERE


DATA
DATA EVERYWHERE

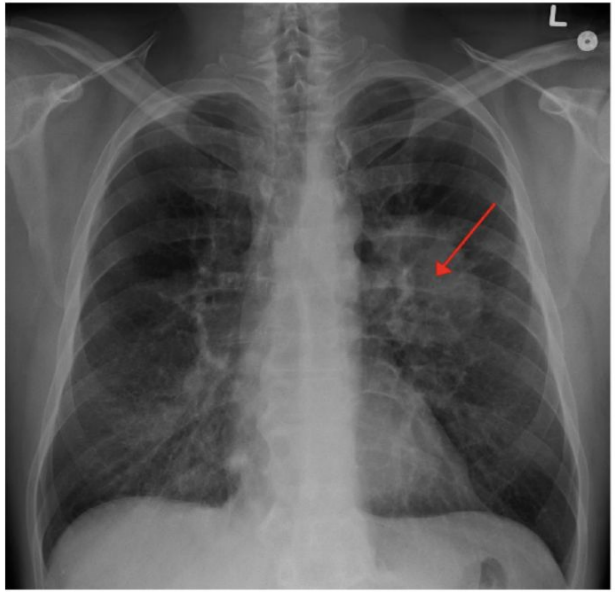
Чем будем заниматься?

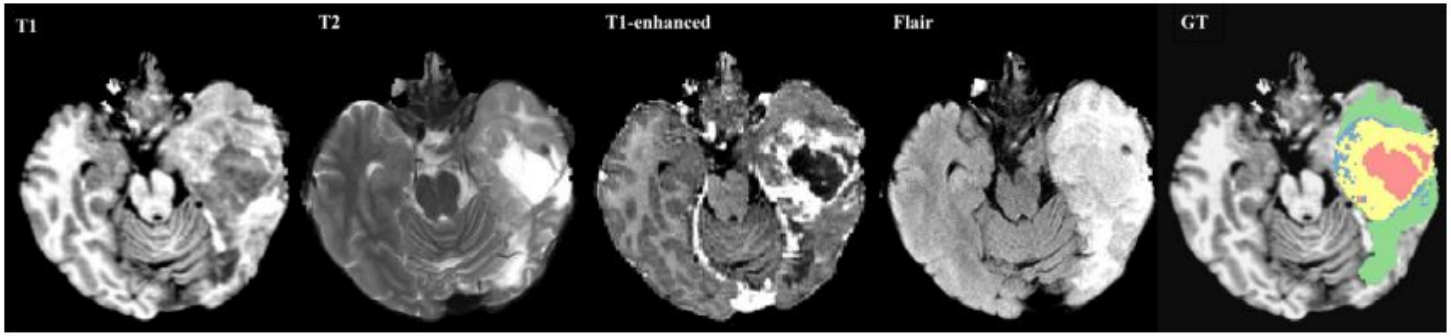



FREE!
SALE!
DEAL!
BUY

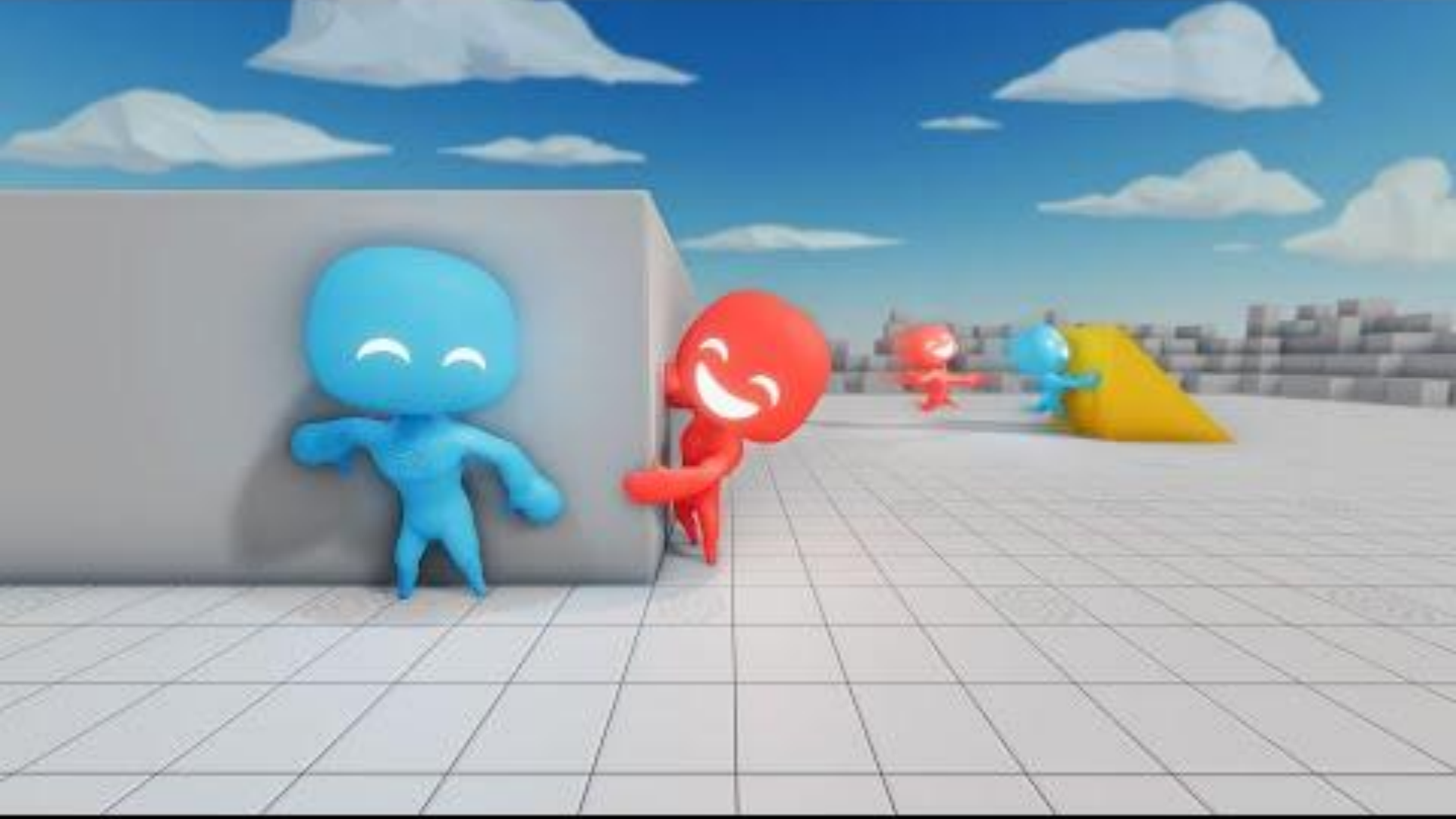


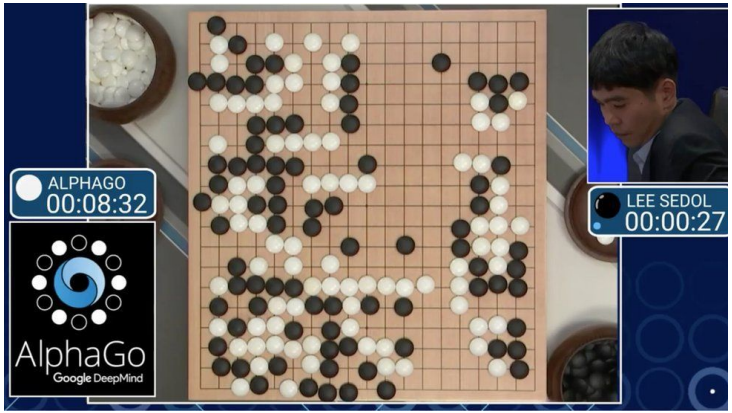

wyscout
N. CASTILLO ('93)
Mainz 05
Overview
Full Matches
Events
Career
Transfer
TSN report
Scouting
Stats
Playlists
Clips
PLAYER INFO
PLAYER STATS
VIDEOS
CURRENTLY ACTIVE COMPETITIONS
PREVIOUS MATCHES



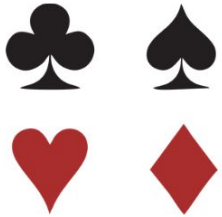


link

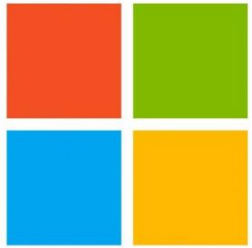



ALPHAGO
00:08:32
LEE SEDOL
00:00:27
AlphaGo
Google DeepMind

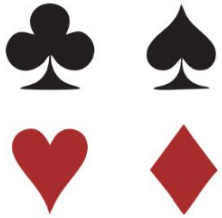

GRZEGORZ 'MANA' KOMINCZ
[PLAYER PERSPECTIVE]
Select Army Units (F2)
BLIZZARD
EXHIBITION GAME
MANA vs ALPHASTAR



https://www.wired.com/story/poker-playing-robot-goes-to-pentagon/



https://www.theverge.com/2019/7/22/20703578/microsoft-openai-investment-partnership-1-billion-azure-artificial-general-intelligence-agi



https://vc.ru/flood/26141-sberbank-holdem

https://www.wired.com/story/facebook-quietly-enters-starcraft-war-for-ai-bots-and-loses/

Что общего у всех этих людей?

Что общего у всех этих людей?

ИХ НЕ СУЩЕСТВУЕТ

thispersondoesnotexist.com

Больше примеров!

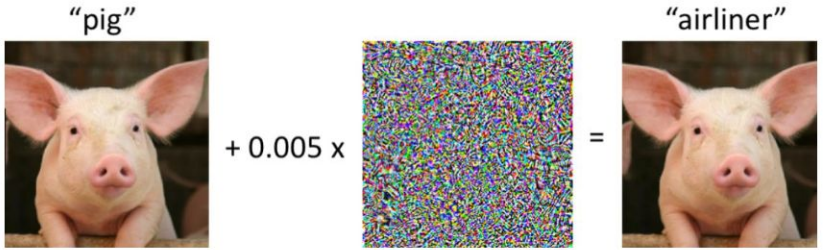
- Deep fake

- Распознавание силуэтов

?=

https://findface.pro/blog/ochertania-budushego-ot-ntechlab-raspoznavanie-siluetov/

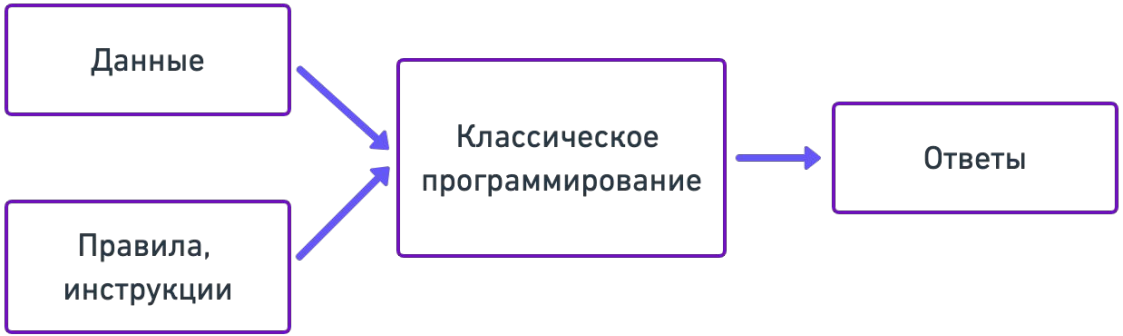
- Adversarial attacks



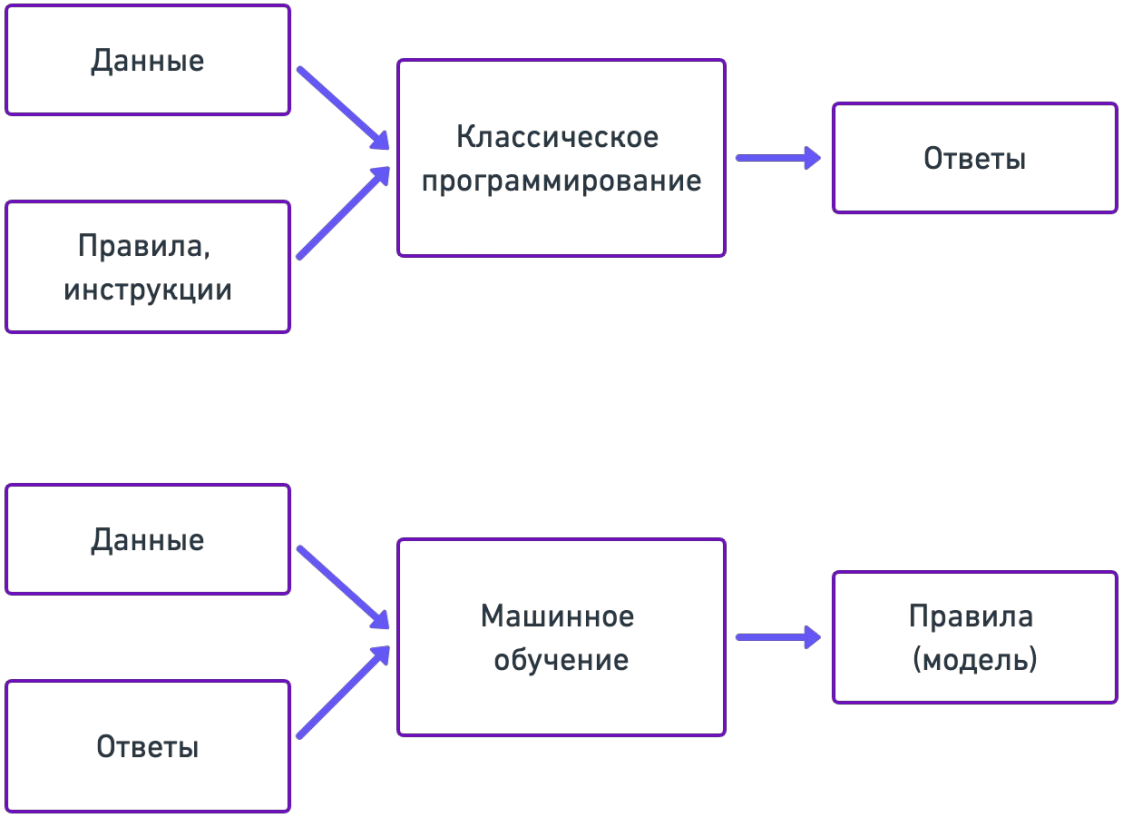
People with no idea about AI, telling me my AI will destroy the world

AI

Алгоритмы VS МО



Алгоритмы VS МО



Типы задач

- Регрессия
- Классификация
- Кластеризация
- Ранжирование

Пример задачи: постановка

- Сеть ресторанов
- Хотим открыть ещё один
- Несколько вариантов размещения
- Какой принесёт наибольшую выгоду?

Пример задачи: обозначения

- x — объект, sample — 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 объекта

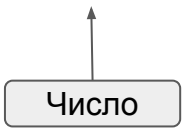
Признаки

- 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 объекта

Вектор

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 объекта



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 $\mathbb{Y}$
- Например, линейная модель: $a(x) = w_0 + w_1 x_1 + \cdots + w_d x_d$

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

Функция потерь

- Не все алгоритмы полезны — нужно как-то оценивать их качество
- $a(x) = 0$ — не принесет никакой выгоды
- Предсказали \$10000 прибыли, а она на самом деле \$5000 — хорошо или плохо?
- Функция потерь $L(a, x)$ — функция, характеризующая величину ошибки алгоритма a на объекте x
- Квадратичное отклонение: $L(a, x) = (a(x) - y)^2$
- Чем меньше, тем лучше

Функционал качества

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

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

Функционал качества

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

Обучение алгоритма

- Есть обучающая выборка и функционал качества
- 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 качества

Машинное обучение

- Не все задачи имеют такую формулировку!

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

Что нужно знать

1. Как сформулировать задачу?
2. Какие признаки использовать?
3. Откуда взять обучающую выборку?
4. Как выбрать метрику качества?
5. Как обучить алгоритм?
6. Как оценить качество алгоритма?