Машинное обучение

Лекция 1
Введение

Логистика Курса

Материалы: https://github.com/WeaselCMC/ml_dpo_2021

Телеграм: По всем вопросам https://t.me/joinchat/u5NGLf6CUoEyNjQy

Логистика Курса : преподаватели

Евгений Егоров

Tg: @iesh00a
Website: https://evgenii-egorov.github.io

Teaching: Skoltech, ВШЭ, ШАД
Индустриальные проекты: Сбербанк, Huawei, Газпром, Азбука Вкуса
Научные публикации: NeurIPS, ICML, AABI

Руслан Костоев

Tg: @mrstrangelove
Website: Coming soon

Teaching: Skoltech, ВШЭ, ШАД
Индустриальные проекты: Газпром, Google, Huawei, Северсталь

Логистика Курса: темы

- Введение и основные задачи
- Линейная регрессия
- Градиентные методы обучения
- Классификация (метрики качества классификации, логистическая регрессия и SVM, Многоклассовая классификация)
- Решающие деревья
- Бэггинг и случайные леса, Градиентный бустинг
- Отбор признаков и снижение размерности
- Кластеризация
- Поиск аномалий
- Рекомендательные системы
- Ранжирование
- Заключение

Логистика Курса: оценки

- Домашние задания (HW)
    - 5-6 самостоятельных заданий
    - У всех домашек одинаковый вес
    - Для зачета нужно получить в среднем 6/10

Логистика Курса: работа в классе

- 1ч 20м лекция
  - Можно перебивать
  - Задавать вопросы
- 10м перерыв
- 1ч 30м практическая работа
  - Задание в ноутбуках
  - Breakout room по 2-3 человека

Логистика Курса

Вопросы?

Что такое машинное обучение ?

“Field of study that gives computers the ability to learn without being explicitly programmed”
- Arthur Samuel

https://en.wikipedia.org/wiki/Arthur_Samuel

Задача: Перевести часы в минуты

Задача: Перевести часы в минуты

x - часы

$f(x) = 60x$

Задача: Как долго будет падать тело?

- Я кинула объект с высоты x
- Через какое время он упадет на Землю?

Задача: Как долго будет падать тело?

- Я кинула объект с высоты x
- Через какое время он упадет на Землю?
- Если я - Галилео Галилей:
  Надо провести эксперименты и замерять время

https://en.wikipedia.org/wiki/Galileo%27s_Leaning_Tower_of_Pisa_experiment

Задача: Как долго будет падать тело?

- Я кинула объект с высоты x
- Через какое время он упадет на Землю?
- Если я - Галилео Галилей:
  Надо провести эксперименты и замерять время
- Я в 21 веке:

$$f(x) = \sqrt{\frac{2x}{g}}$$

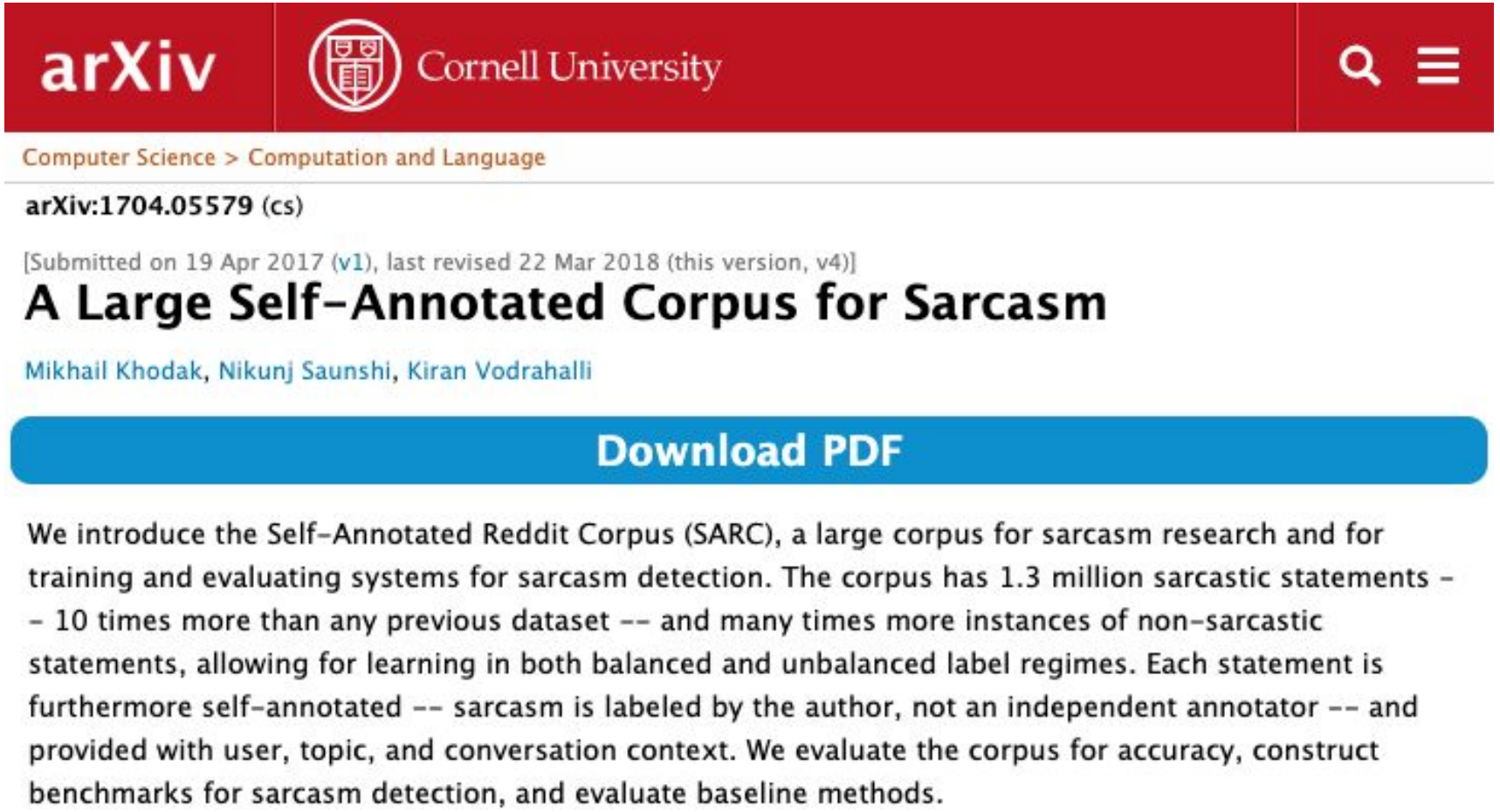
Задача: Кто изображен на фото?

Задача: Анализ эмоциональной окраски текста

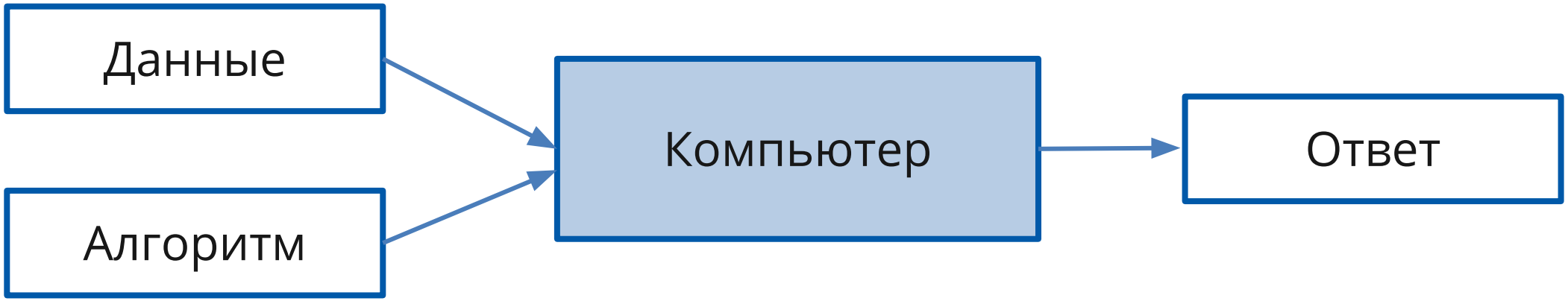
Какой из отзывов положительный?

- “Не могу сказать, что мне понравился суп. Но он лучше чем ничего”
- “Мне не не понравился этот банан”

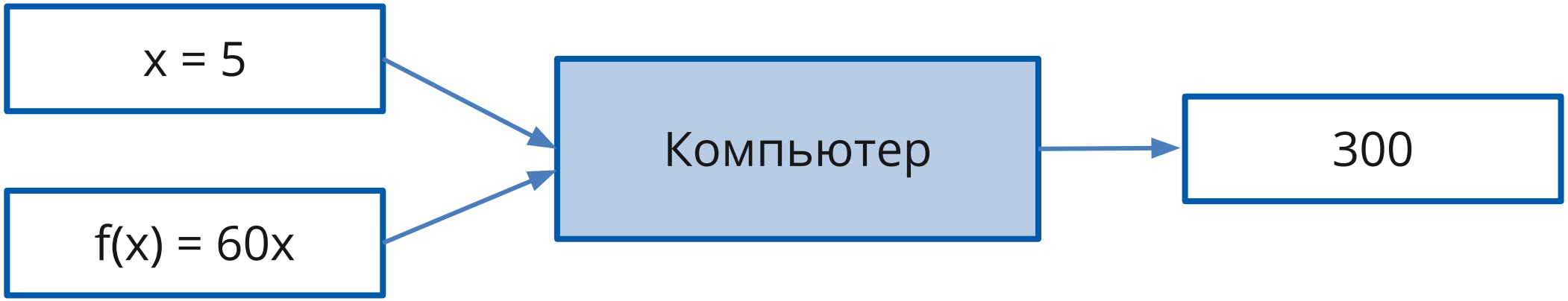
Задача: Анализ эмоциональной окраски текста



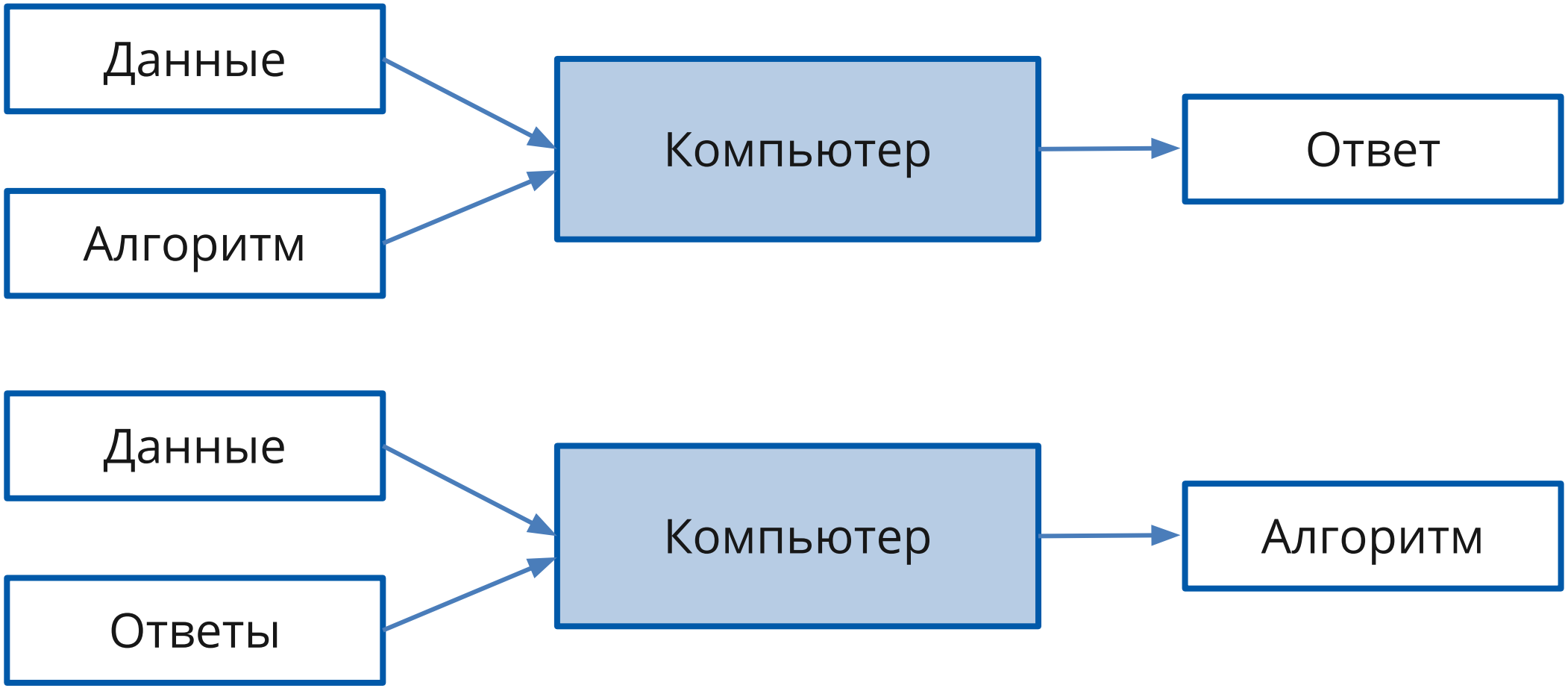
Машинное обучение Классическое Программирование



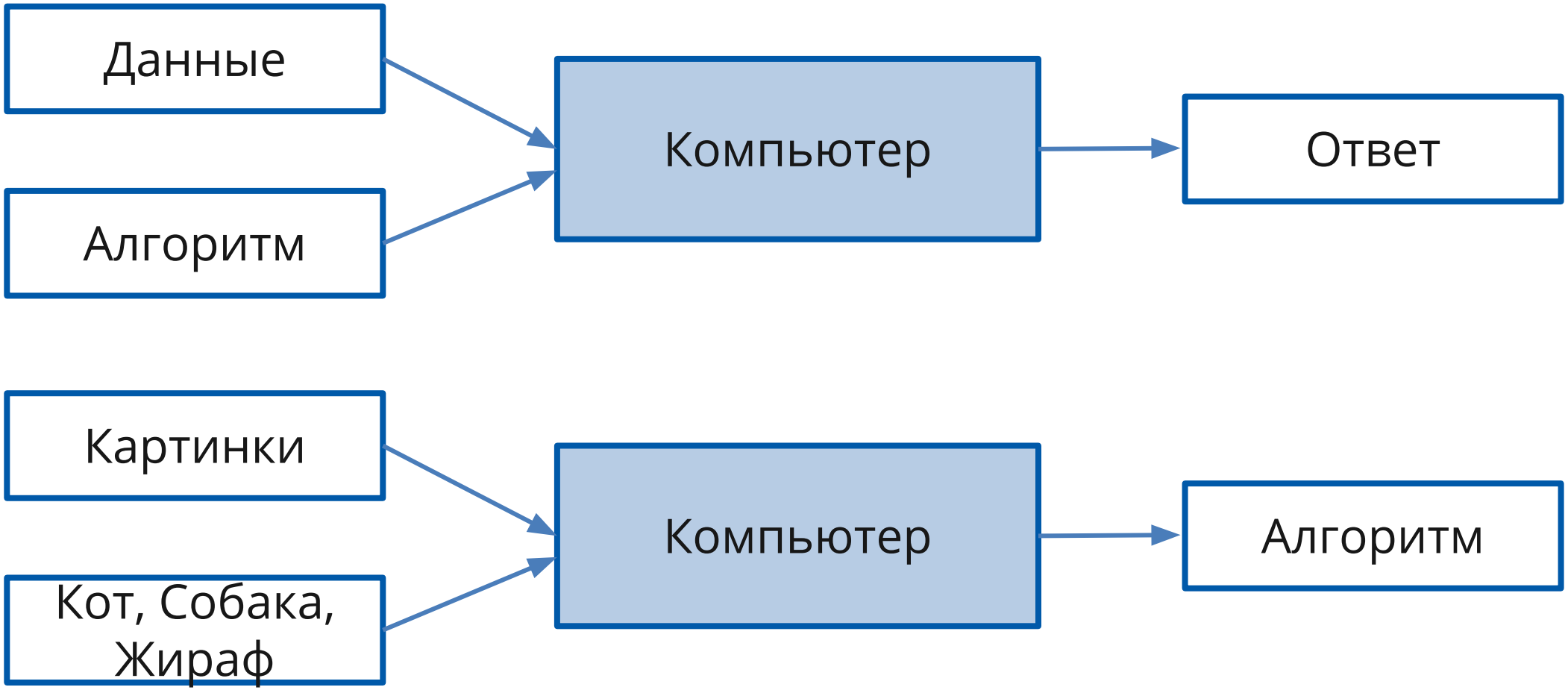
Машинное обучение Классическое Программирование



Машинное обучение Классическое Программирование



Машинное обучение Классическое Программирование



Задача: Шахматы против Го

Deep Blue vs Kasparov (1985)

AlphaGo vs Lee Sedol (2016)

https://deepmind.com/research/case-studies/alphago-the-story-so-far

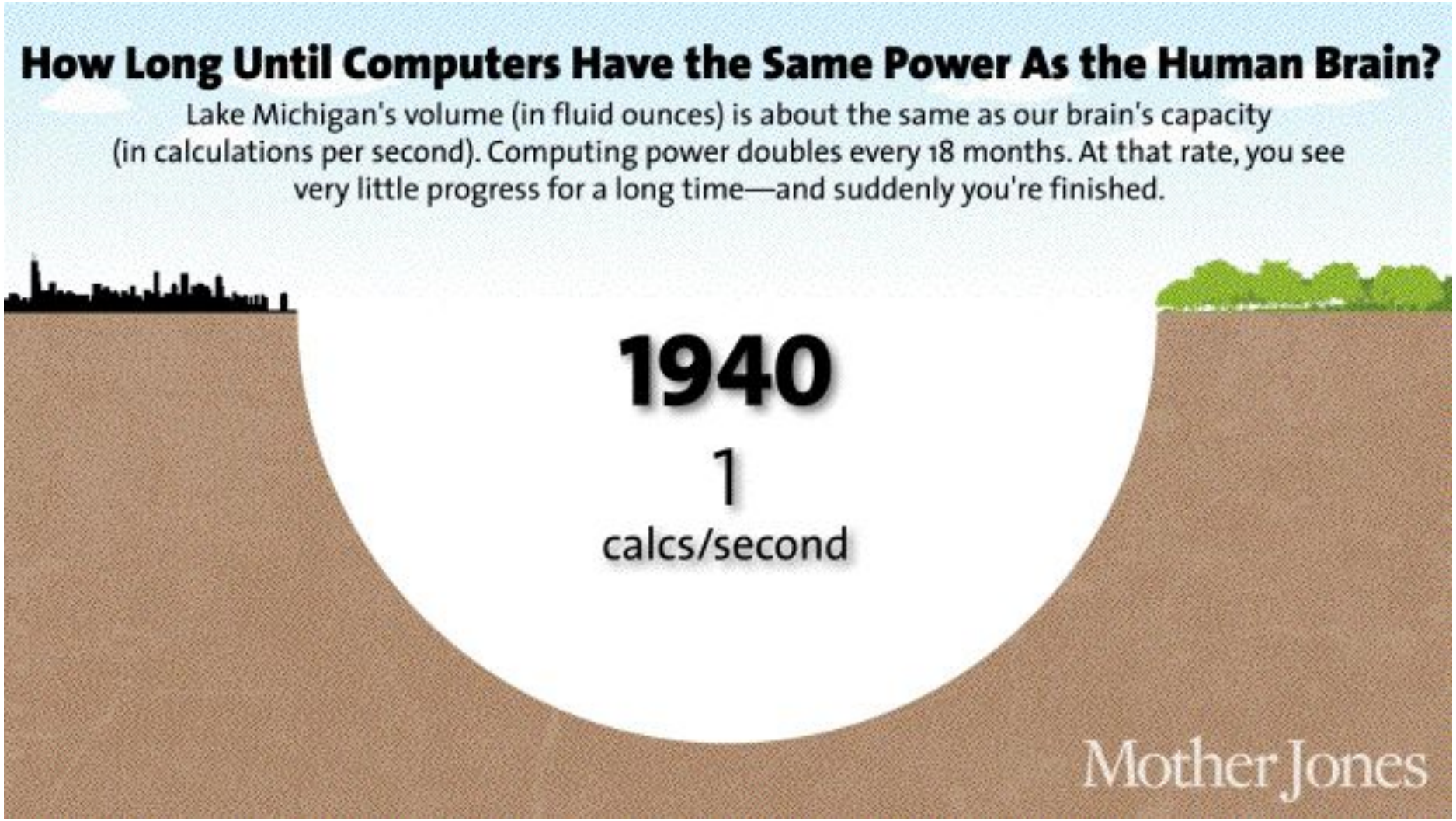
Почему сейчас?

Deep Blue vs Kasparov (1985)

AlphaGo vs Lee Sedol (2016)

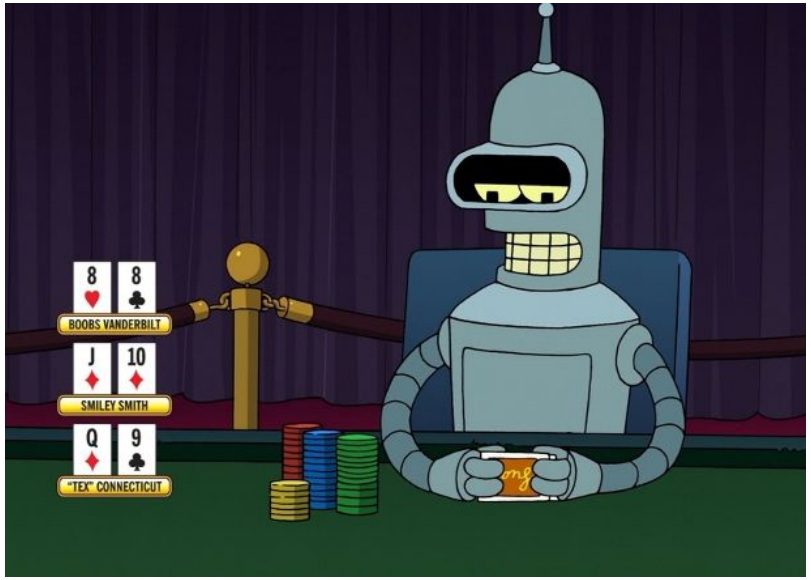
https://deepmind.com/research/case-studies/alphago-the-story-so-far

Почему сейчас?



Еще примеры

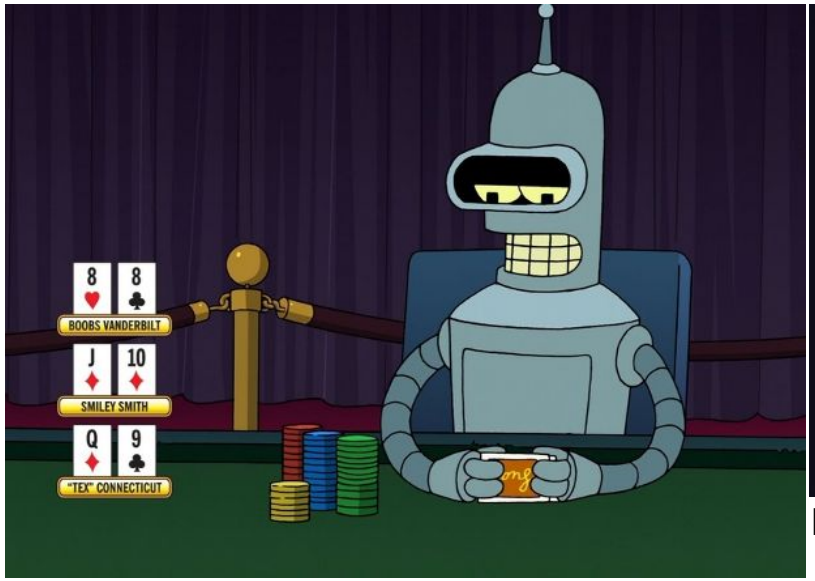
2017



https://www.wired.com/2017/01/mystery-ai-just-crushed-best-human-players-poker/

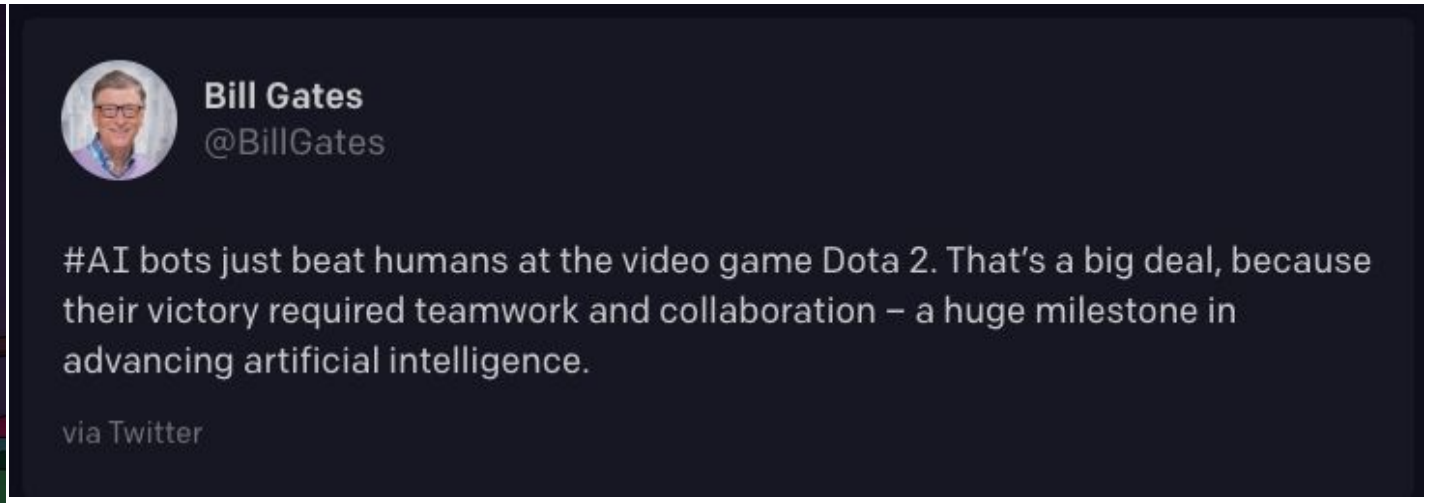
Еще примеры

2017



https://www.wired.com/2017/01/mystery-ai-just-crushed-best-human-players-poker/

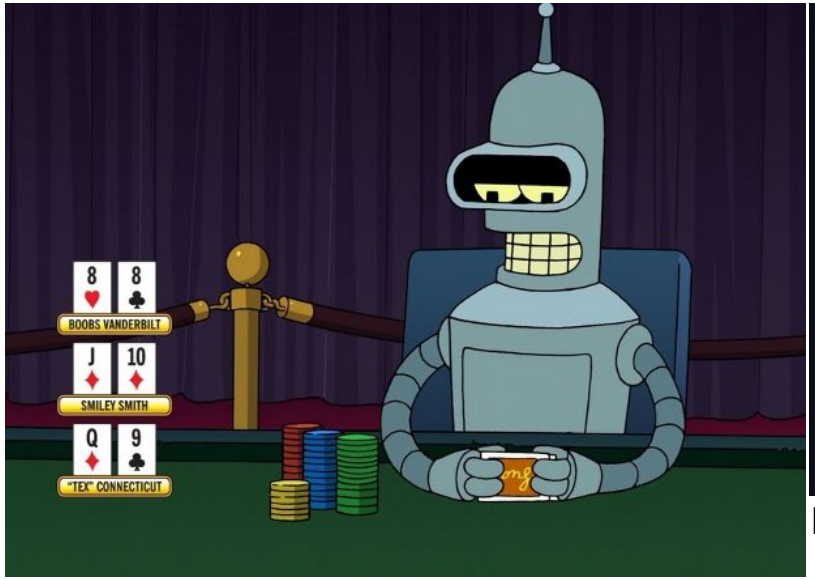
2018



https://openai.com/projects/five/

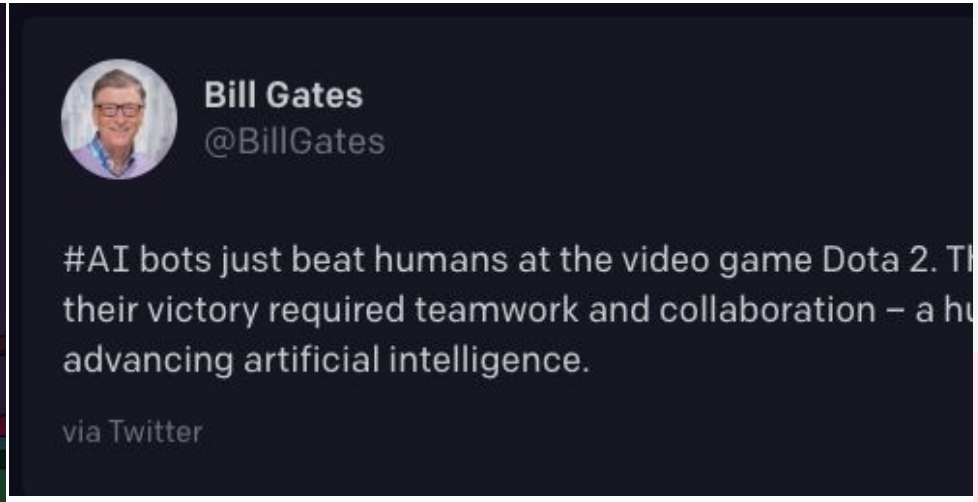
Еще примеры

2017



https://www.wired.com/2017/01/mystery-ai-just-crushed-best-human-players-poker/

2018



https://openai.com/projects/five/

2019

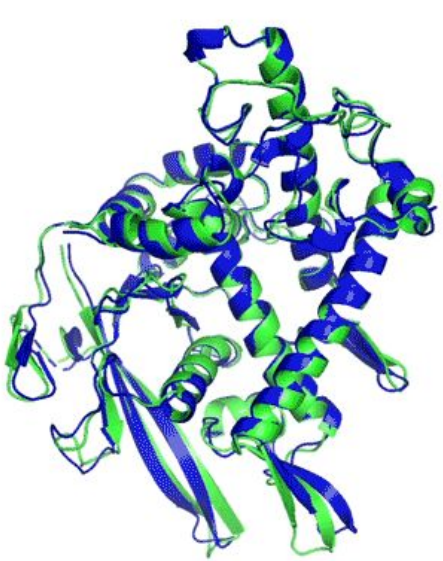
https://deepmind.com/blog/article/alphastar-mastering-real-time-strategy-game-starcraft-ii

Не только игры Что общего у этих людей?

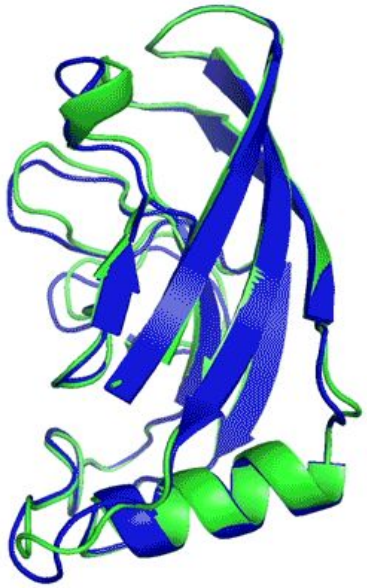
Не только игры Что общего у этих людей?

https://thispersondoesnotexist.com/

Не только игры Фолдинг белка



T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)



T1049 / 6y4f
93.3 GDT
(adhesin tip)

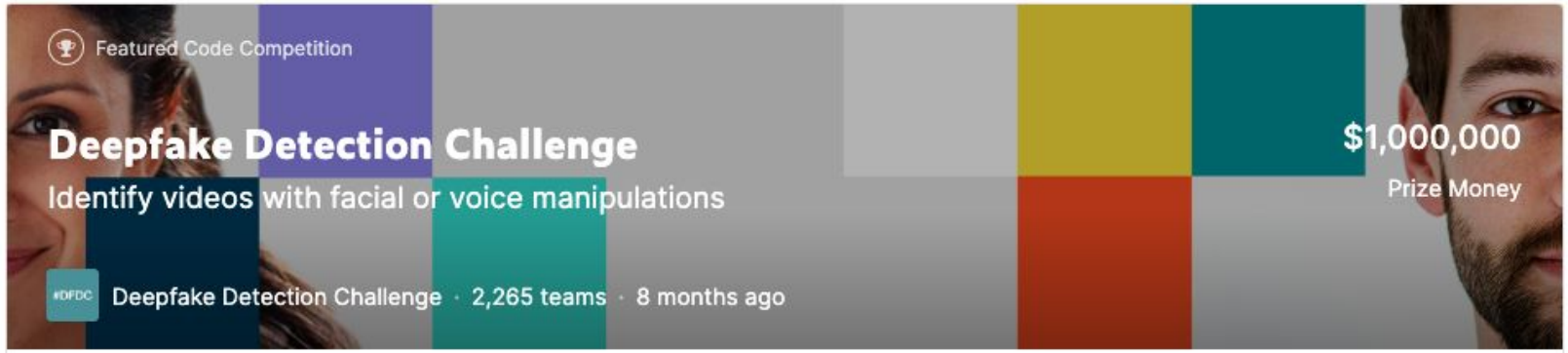
Experimental result
Computational prediction

https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

Не только игры Deep Fakes

https://www.youtube.com/watch?v=2jp4M1cIJ5A

Не только игры Deep Fakes Detection



Overview Data Notebooks Discussion Leaderboard Rules

Public Leaderboard Private Leaderboard

This competition is closed for submissions. The Private Leaderboard was based on a re-run of participants' code by the host on a privately-held test set.
This competition has completed. This leaderboard reflects the final standings.

Refresh

In the money Gold Silver Bronze

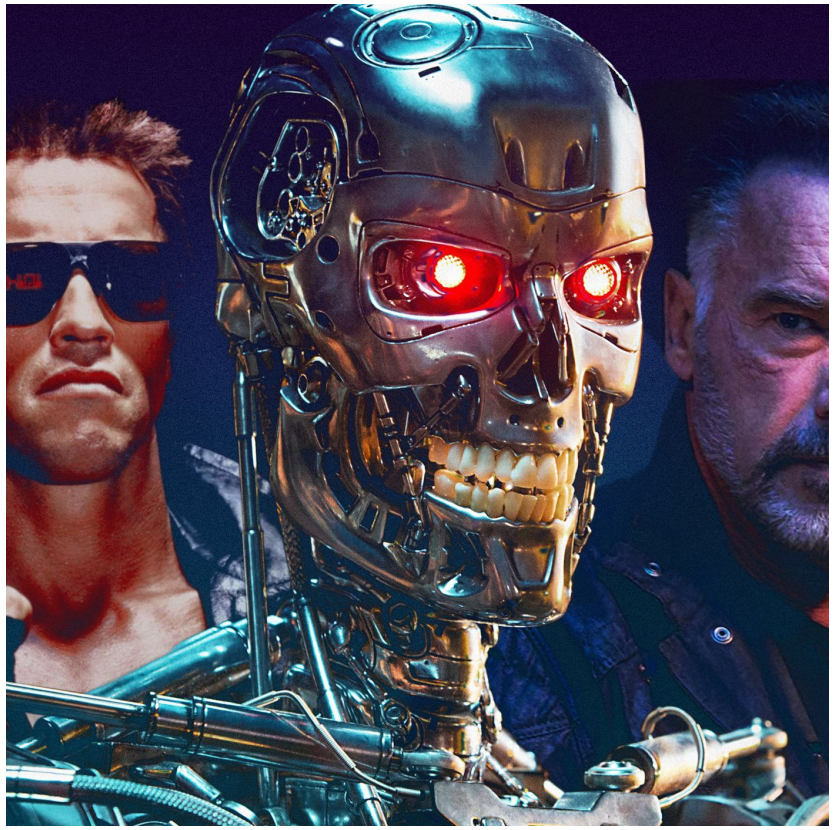
#	△pub	Team Name	Notebook	Team Members	Score	Entries	Last
1	▲3	Selim Seferbekov			0.42798	2	8mo

https://www.kaggle.com/c/deepfake-detection-challenge/leaderboard

Искусственный Интеллект?

- **Слабый** (week)

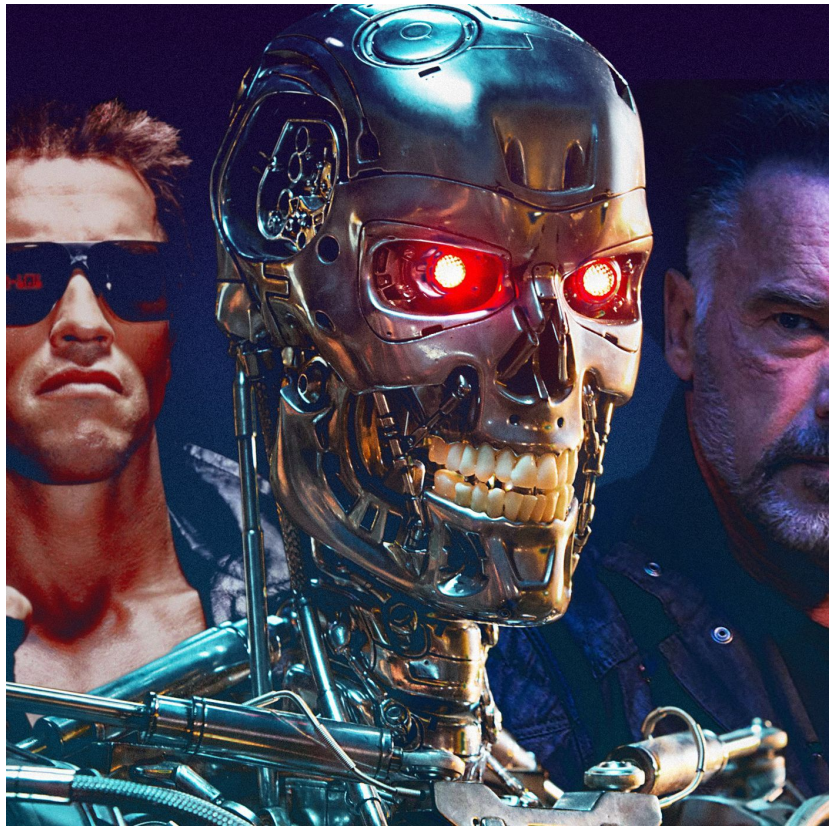
- **Сильный** (general)



Искусственный Интеллект?

- **Слабый** (week)

- ~~**Сильный** (general)~~



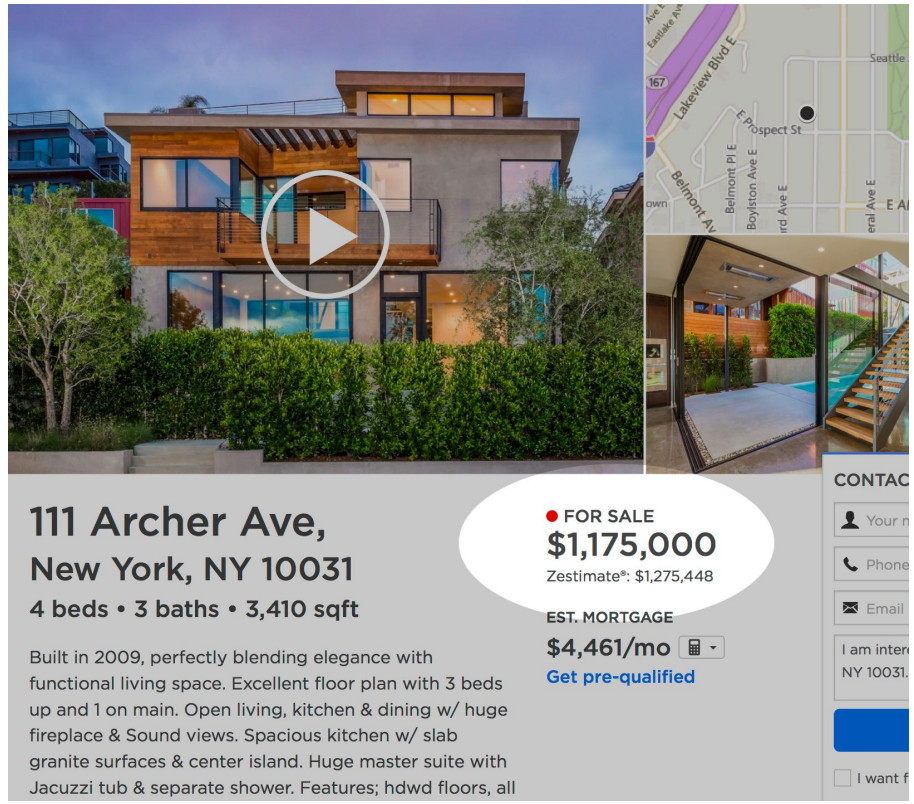
Основные определения

Задача: Сколько стоит дом?

Соревнования на Kaggle.com:

- Zillow’s Home Value Prediction ($25’000)
- Sberbank Russian Housing Market ($12’000)

Хотим научиться предсказывать стоимость дома



https://www.kaggle.com/c/zillow-prize-1/overview/competition-overview

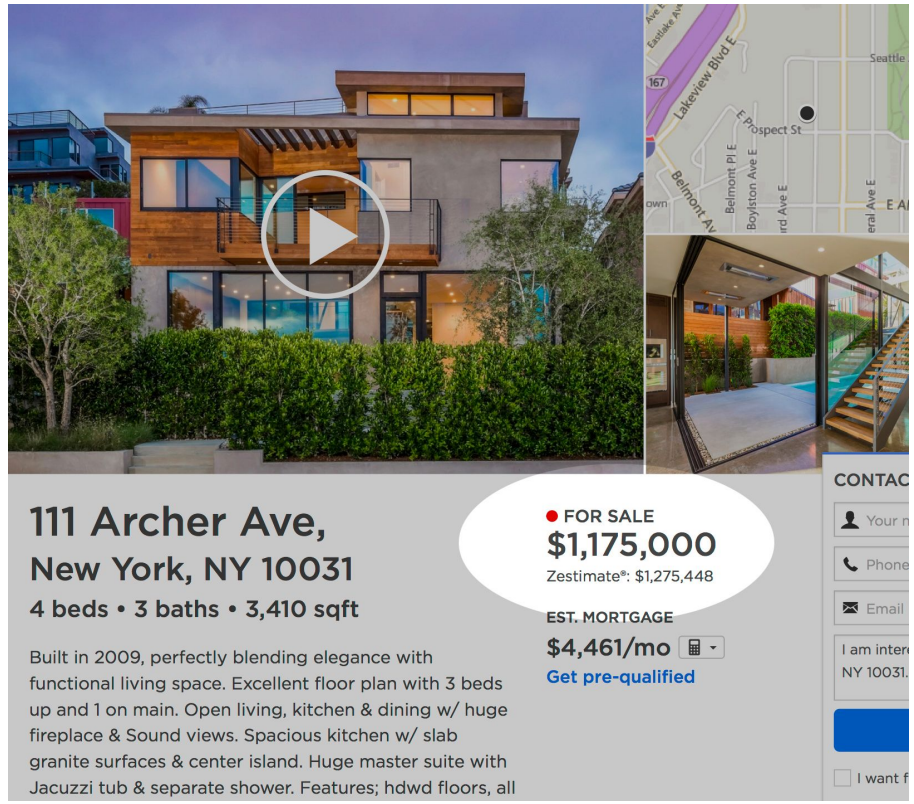
Задача: Сколько стоит дом?

Соревнования на Kaggle.com:

- Zillow’s Home Value Prediction ($25’000)
- Sberbank Russian Housing Market ($12’000)

Хотим научиться предсказывать стоимость дома

... посмотреть на 100500 других домов



https://www.kaggle.com/c/zillow-prize-1/overview/competition-overview

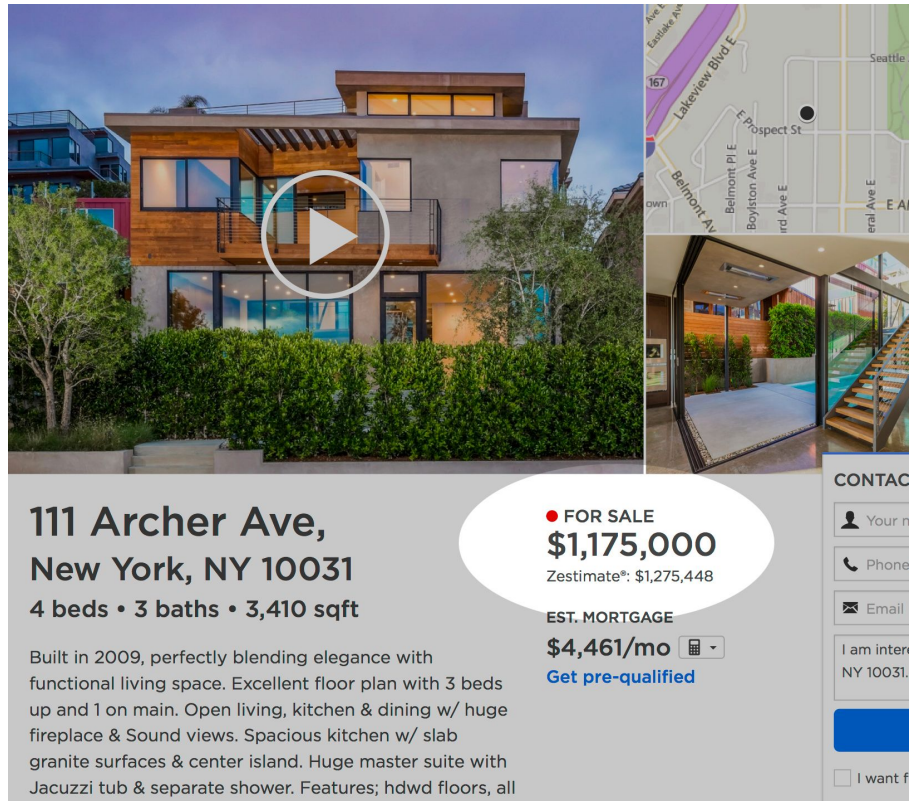
Обозначения

- x - объект (sample)
  - Описание дома

- $\mathbb{X}$ - пространство всех возможных объектов
  - Все возможные описание домов

- y - ответ, целевая переменная (target)
  - Стоимость одного дома

- $\mathbb{Y}$ - пространство всех возможных ответов
  - Все положительные вещественные числа



https://www.kaggle.com/c/zillow-prize-1/overview/competition-overview

Обозначения: признаки

Что такое объект? Компьютеру нужны числа

Признаки (features) - числовые характеристики объектов

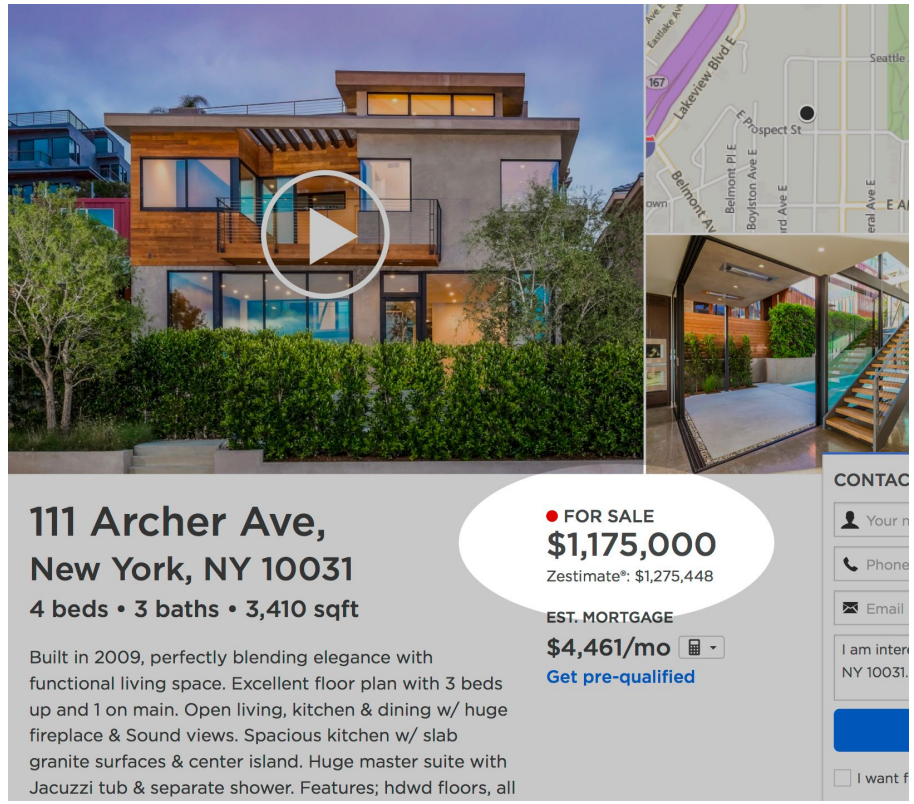
$$x = (x_1, x_2, \dots, x_d)$$

Обозначения: признаки

$$x = (x_1, x_2, \ldots, x_d)$$

Описание дома:

- Этаж
- Количество комнат
- Общая площадь
- Район
- Расстояние до метро
- …



https://www.kaggle.com/c/zillow-prize-1/overview/competition-overview

Обозначения: обучающая выборка

Набор объектов с известными ответами (Training Dataset)

$$(x_i, y_i)_{i=1}^{N}$$

N - размер выборки

Обозначения: алгоритм

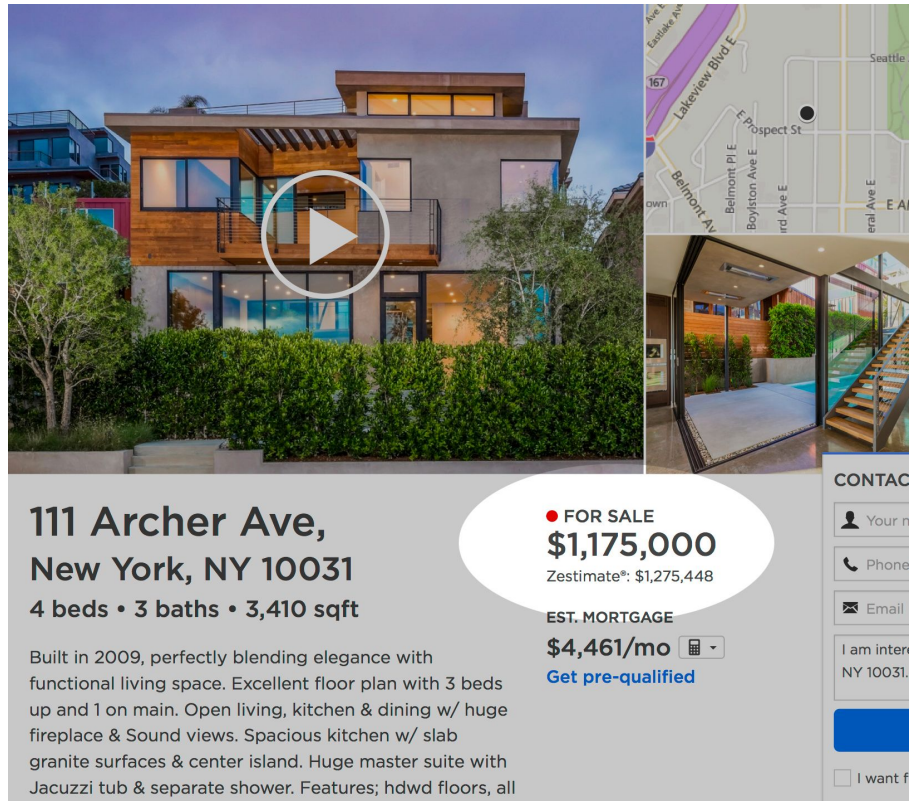
Функция, предсказывающая ответ для любого объекта - алгоритм (модель)

$$a(x): \mathbb{X} \rightarrow \mathbb{Y}$$

Обозначения: алгоритм

$$a(x): \mathbb{X} \rightarrow \mathbb{Y}$$

a(x) = 1’000’000 + 200’000 * (площадь) - 50’000 * (расстояние до метро)



https://www.kaggle.com/c/zillow-prize-1/overview/competition-overview

Обозначения: функция потерь

Функция потерь (Loss function) измеряет как “далеко” наше предсказание от правды.

$$L(a, x) = (a(x) - y)^2$$

Обозначения: функционал качества

Функционал качества, метрика качества — мера качества работы алгоритма на выборке

Среднеквадратичная ошибка (Mean Squared Error, MSE):

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

Обозначения: функционал качества

Функционал качества, метрика качества — мера качества работы алгоритма на выборке

Среднеквадратичная ошибка (Mean Squared Error, MSE):

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

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

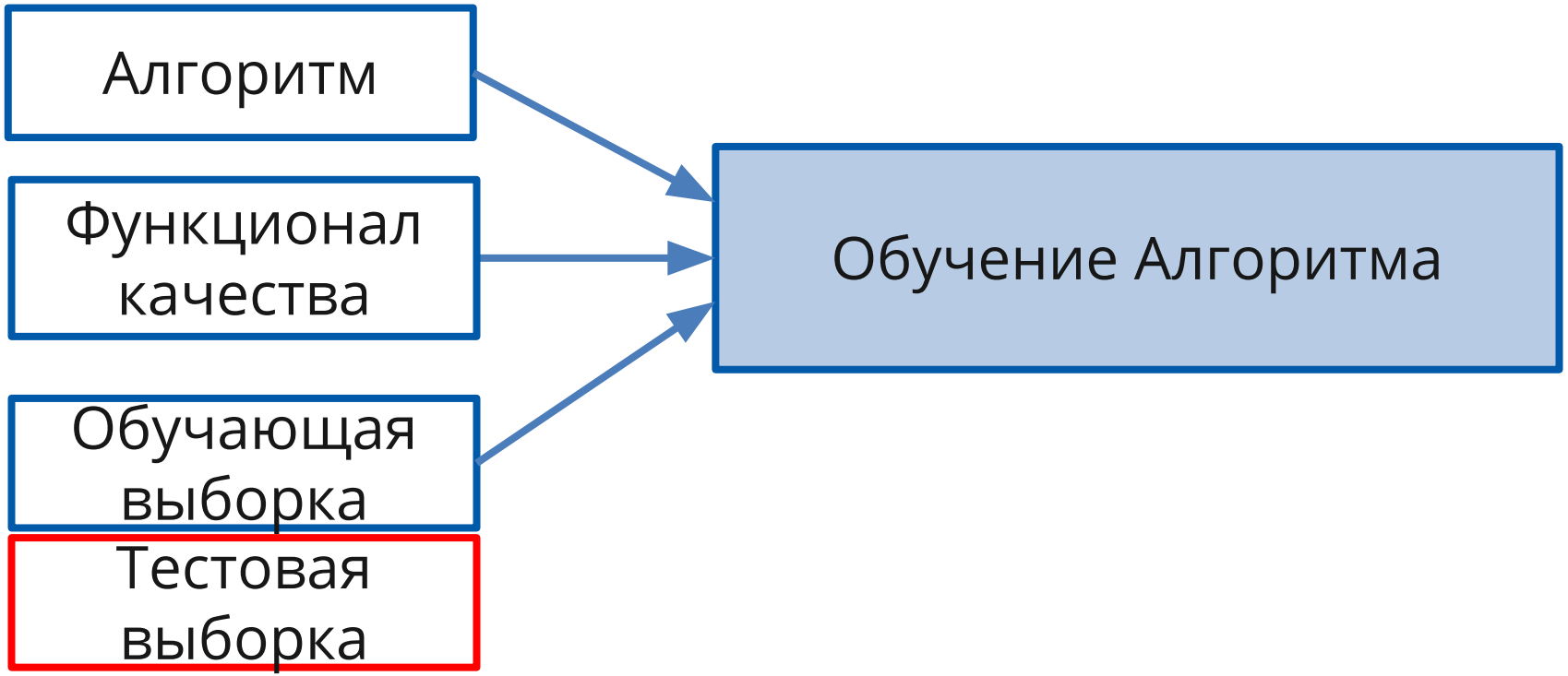
Обозначения: обучение алгоритма

- Есть обучающая выборка и функционал качества

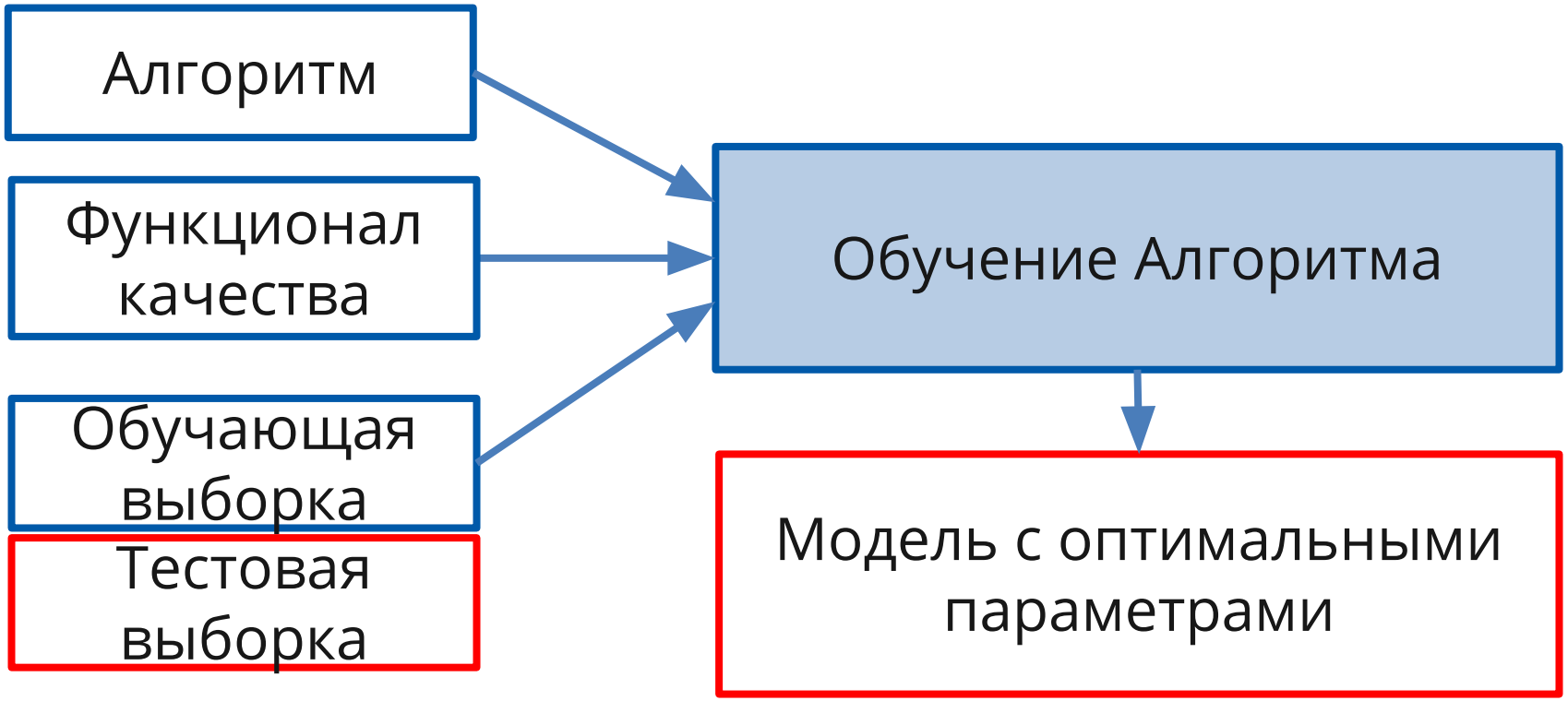
- Семейство алгоритмов $\mathscr{A}$
  - Из чего выбираем алгоритм
  - Пример: все линейные модели
  - $$\mathcal{A} = \{w_0 + w_1 x_1 + \cdots + w_d x_d \mid w_0, w_1, \ldots, w_d \in \mathbb{R}\}$$

- Обучение: поиск оптимального алгоритма с точки зрения функционала качества

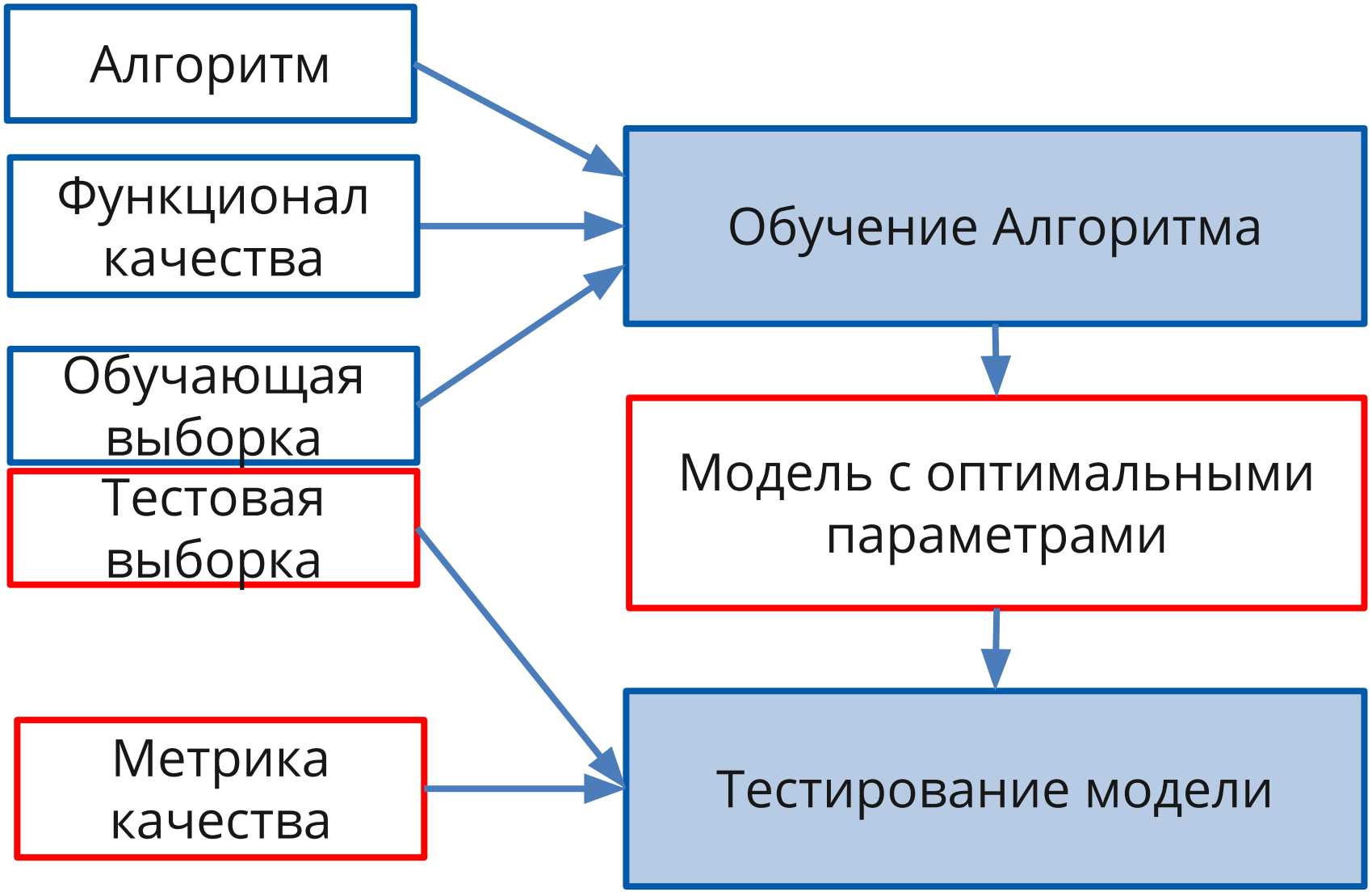
Обучение алгоритма



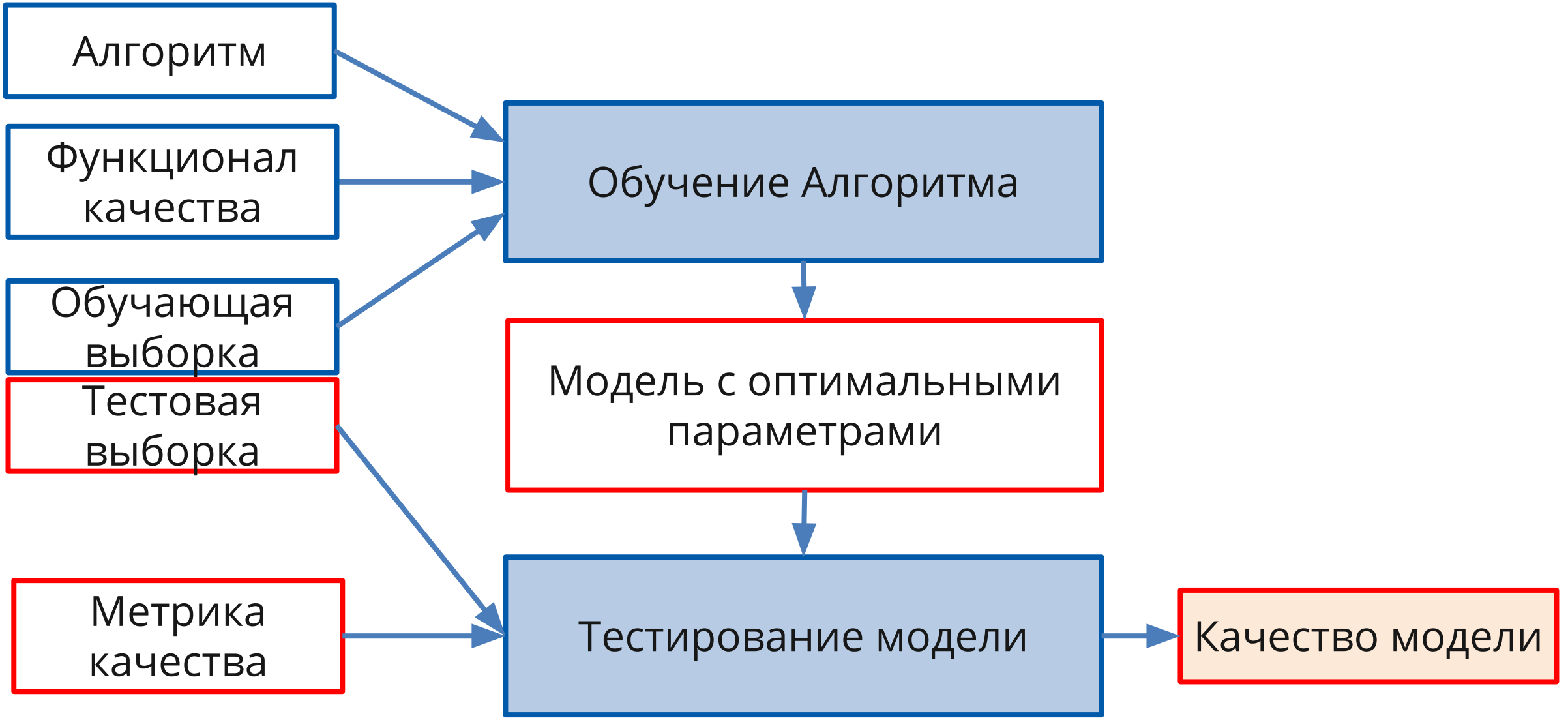
Обучение алгоритма



Обучение алгоритма



Обучение алгоритма

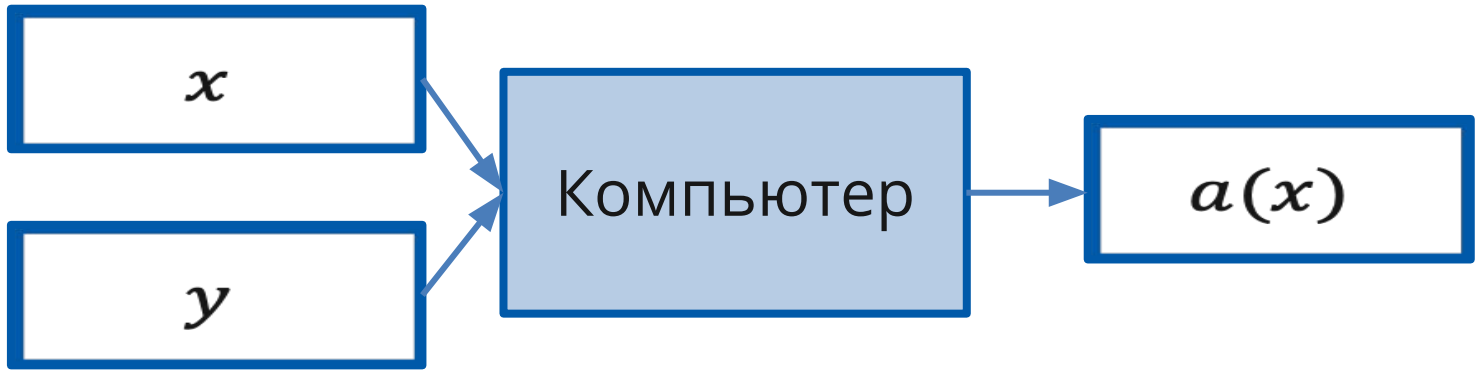


Типы Задач

Обучение с учителем и без

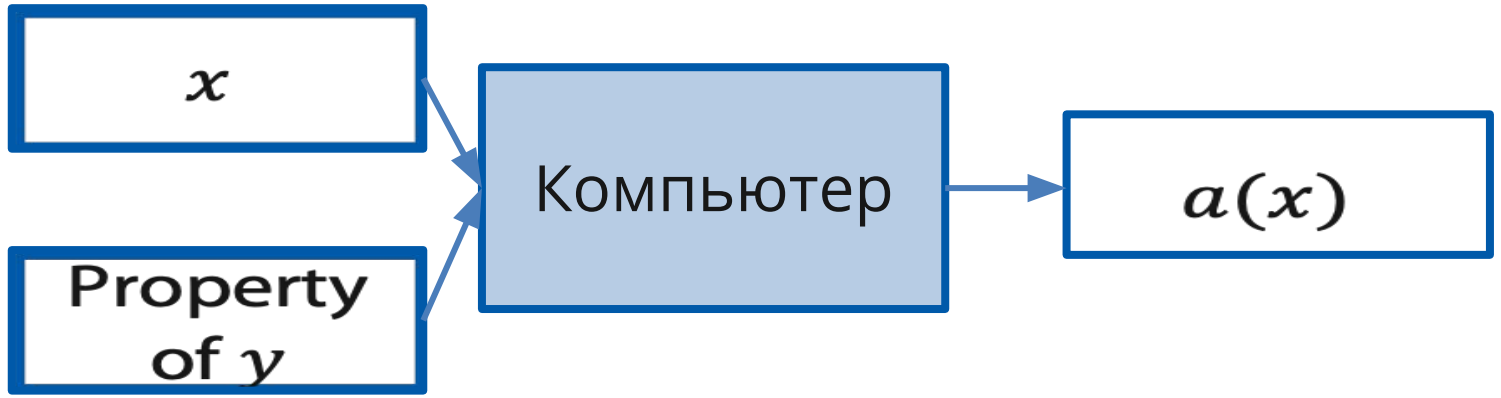
Supervised learning

- У нас есть “правильные ответы”



Unsupervised learning

- У нас есть только объекты



Классификация $\mathbb{Y} = \{1, \dots, K\}$

2006

Computer accuracy = 60% Probability to guess= $0.6^{12} = 0.00217$

2014

Completed • Swag • 215 teams

Dogs vs. Cats

Wed 25 Sep 2013 – Sat 1 Feb 2014 (8 months ago)

Dashboard

Private Leaderboard - Dogs vs. Cats

This competition has completed. This leaderboard reflects the final standings.

See someone

#	Δ1w	Team Name * in the money	Score	Entries	Last Submission UTC (Best – Last
1	—	Pierre Sermanet *	0.98914	5	Sat, 01 Feb 2014 21:43:19 (-
2	↑26	orchid *	0.98309	17	Sat, 01 Feb 2014 23:52:30
3	—	Owen	0.98171	15	Sat, 01 Feb 2014 17:04:40 (-
4	new	Paul Covington	0.98171	3	Sat, 01 Feb 2014 23:05:20

Computer accuracy = 98% Probability to guess = $0.98^{12} = 0.875$

Идеи задач классификации?

http://ml-playground.com/#

Пример 1: Credit scoring

https://www.kaggle.com/c/home-credit-default-risk

Пример 2: Sentiment analysis

https://pypi.org/project/dostoevsky/

https://www.kaggle.com/c/tweet-sentiment-extraction

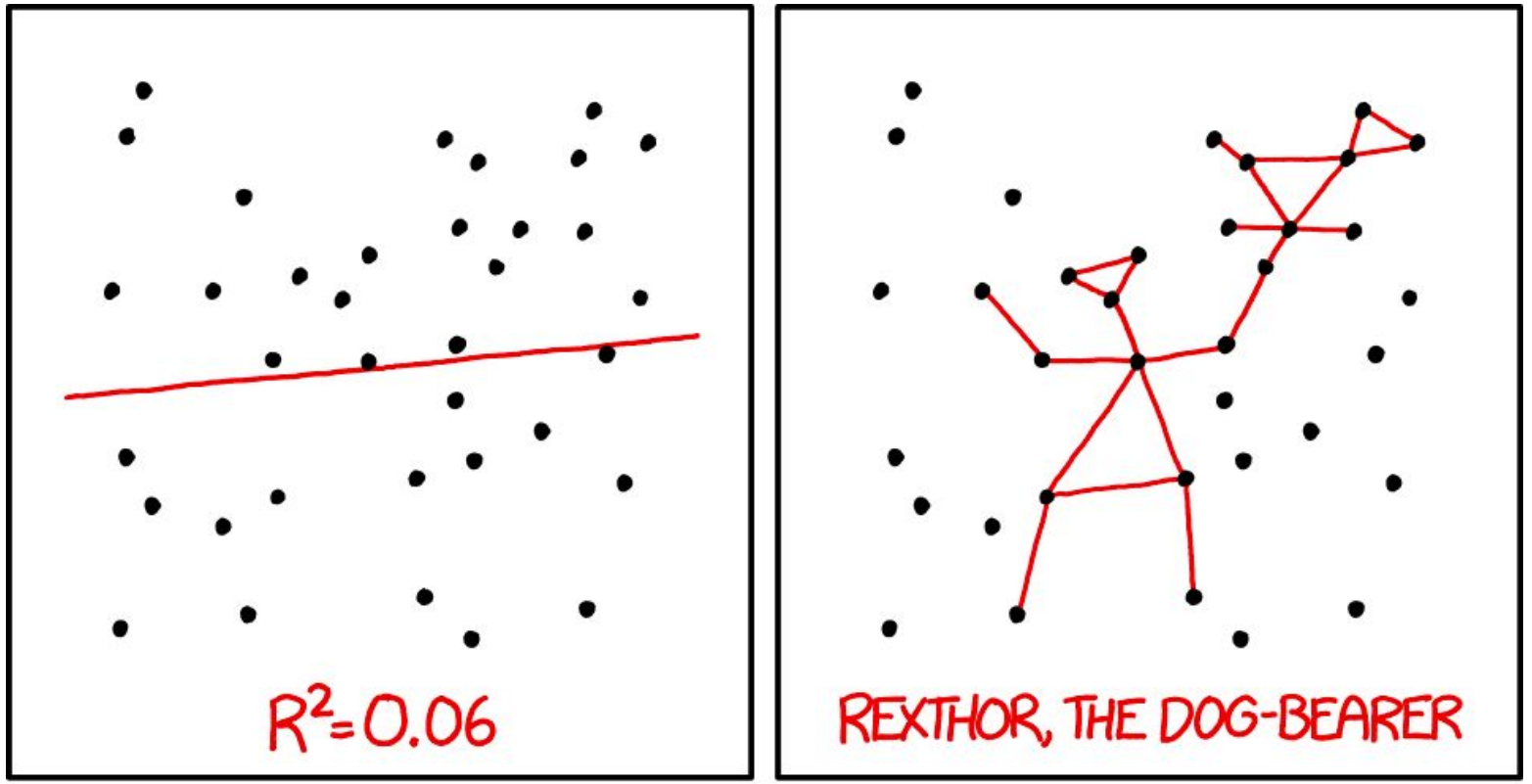
Пример 3: Image Segmentation

https://ai.facebook.com/blog/-detectron2-a-pytorch-based-modular-object-detection-library-/

https://www.med.upenn.edu/cbica/brats2020/tasks.html

Регрессия

$$\mathbb{Y} = \mathbb{R}$$



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

Идеи задач регрессии?

Пример 1: Predicting age

https://www.kaggle.com/c/trends-assessment-prediction
https://www.how-old.net

Пример 2: Detection

https://ai.facebook.com/blog/-detectron2-a-pytorch-based-modular-object-detection-library-/

Кластеризация

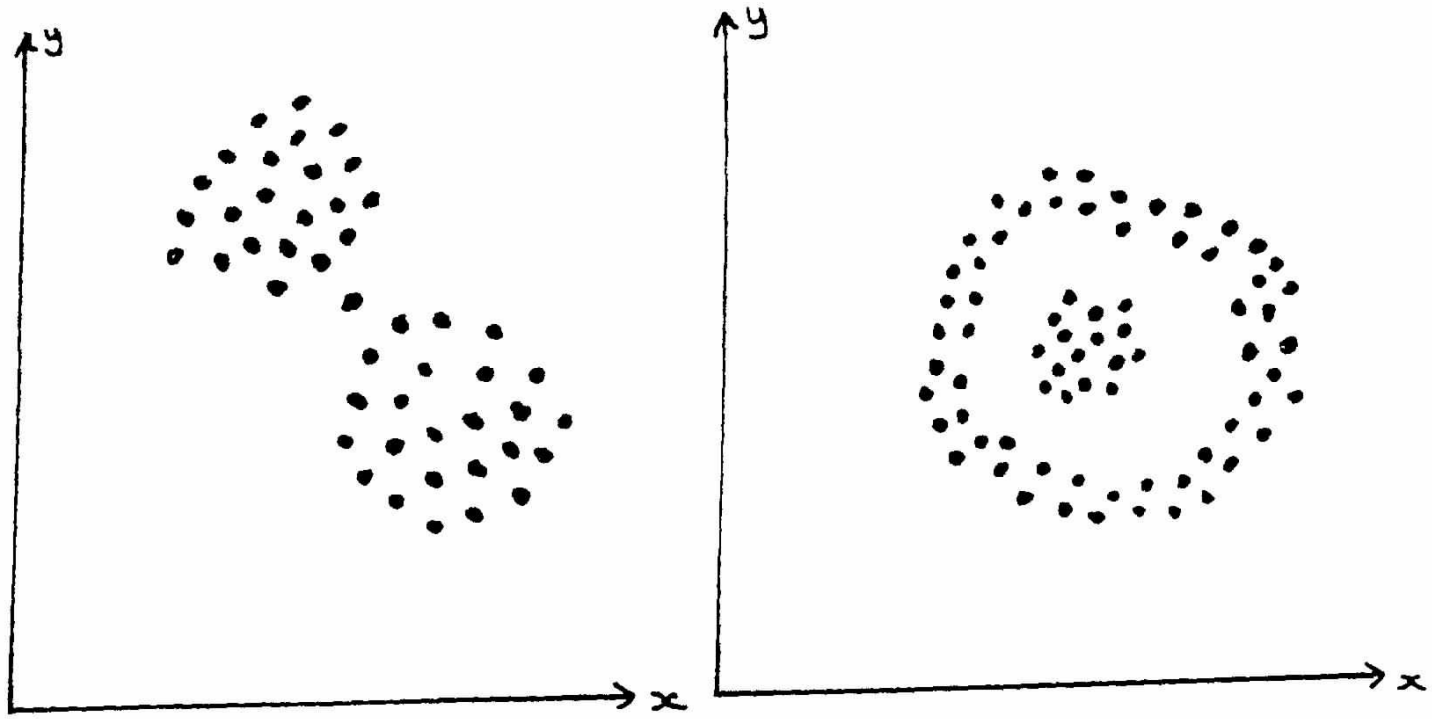
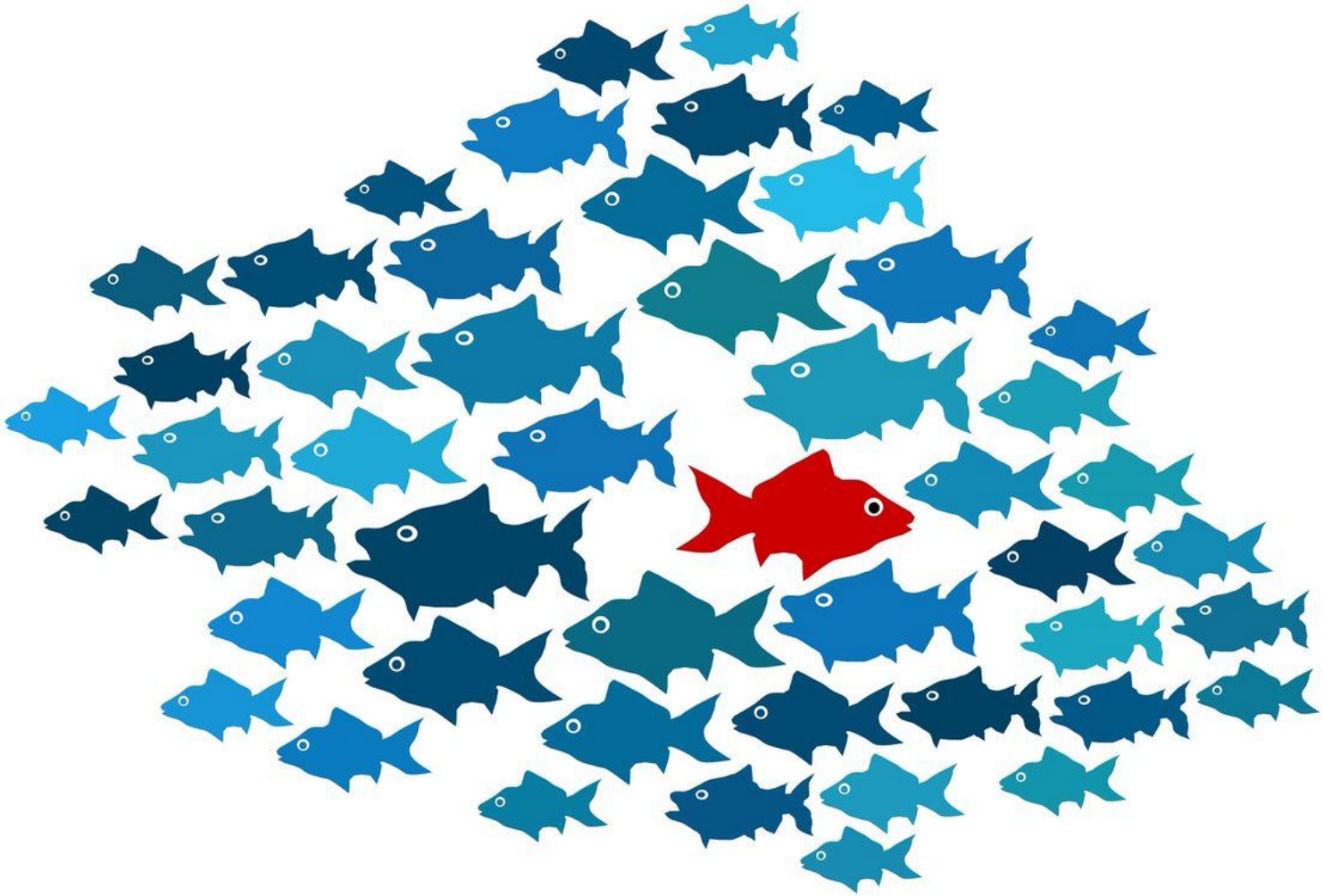
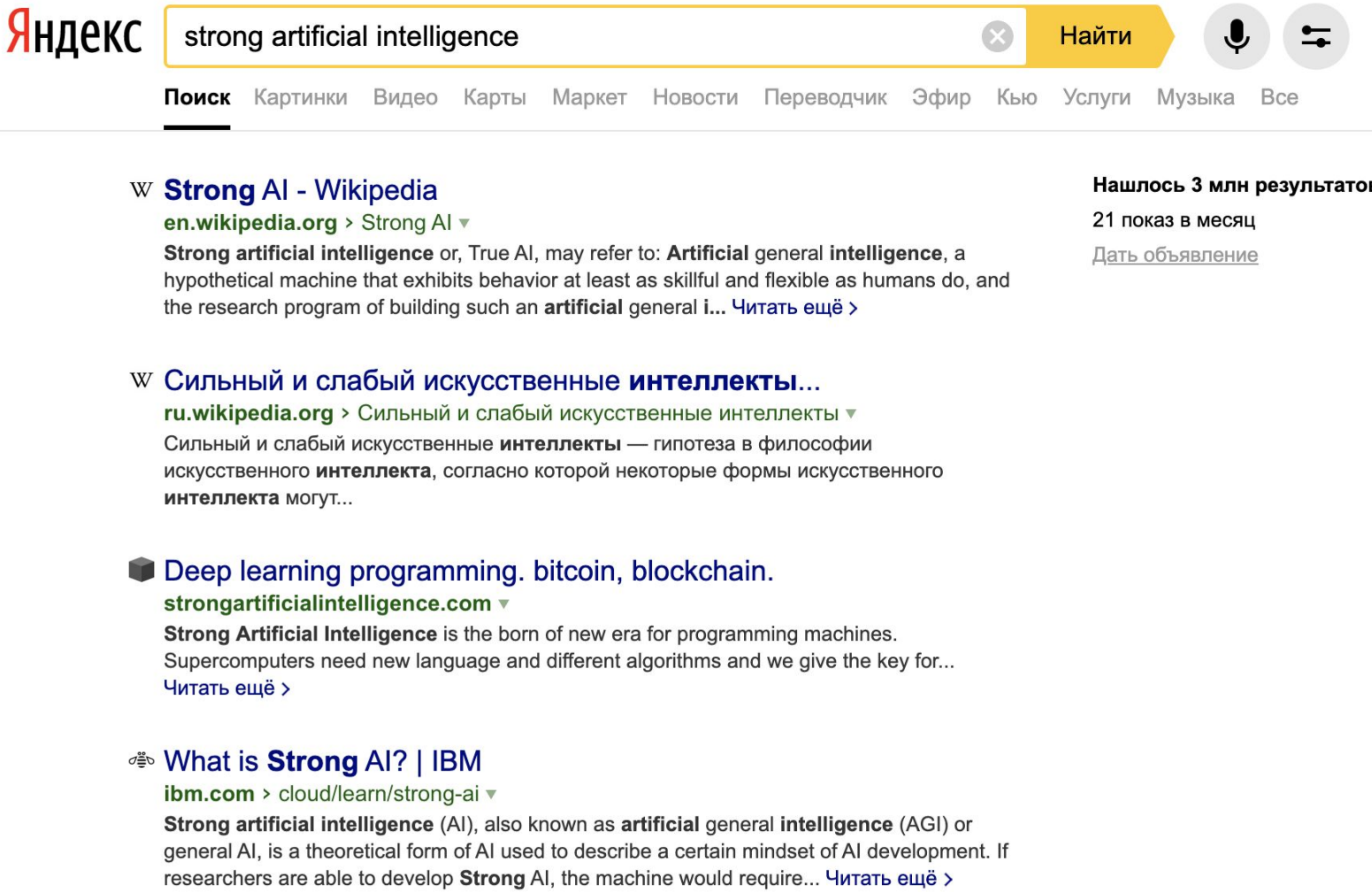


Figure 3.4 Further Examples of Clusters

Поиск Аномалий



Ранжирование



В следующих сериях

- Типы признаков
- Линейная регрессия
- Решение задач оптимизации для обучения алгоритмов МО