A decorative graphic on the left side of the slide consists of a network of thin, light blue lines. These lines form a complex, branching pattern that resembles a circuit board or a neural network. The lines are connected to small, empty circles at various points, creating a series of nodes and connections. The overall style is clean and modern, with a focus on geometric shapes and a limited color palette of blue and white.

# Занятие 2

## Введение в машинное обучение

Елена Кантонистова

ВШЭ, 2021

# ЧТО ТАКОЕ МАШИННОЕ ОБУЧЕНИЕ

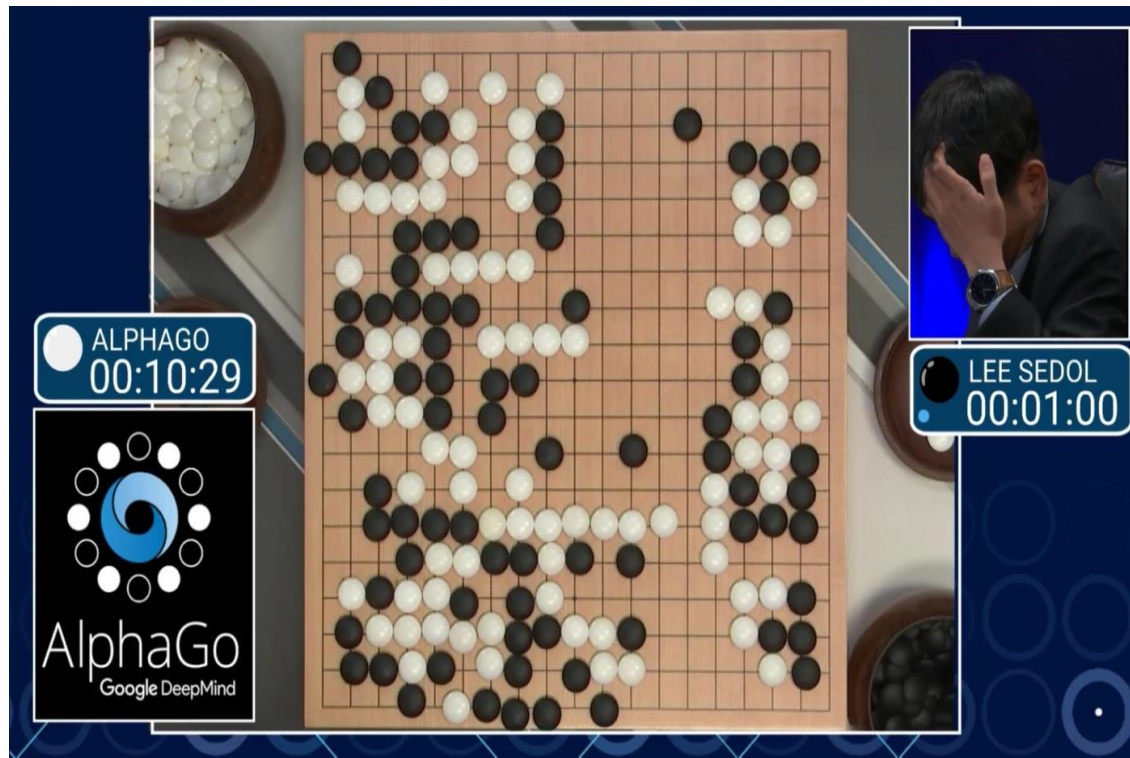
**Машинное обучение** – набор способов воспроизведения связей между событиями и результатом.

**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.

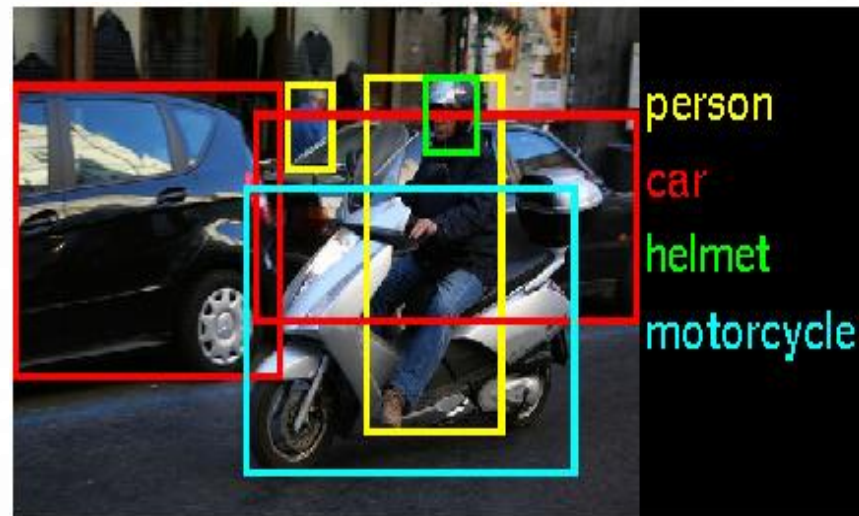
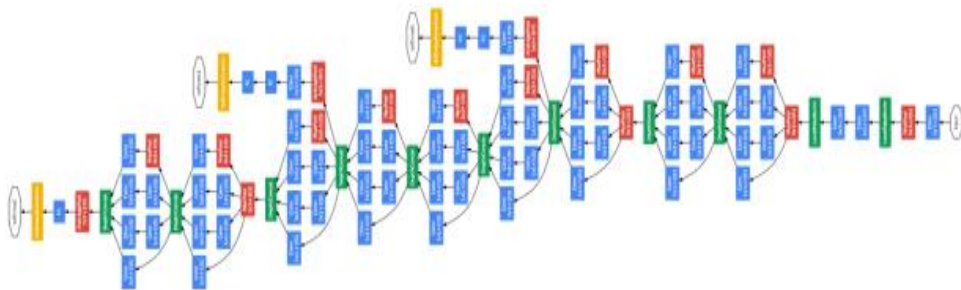
# ПРИМЕРЫ

- **Нейронная сеть, играющая в Го**
- **Март 2016** – победа над мировым чемпионом
- Нейронная сеть обучалась, играя сама с собой для увеличения объёмов входных данных (принцип обучения с подкреплением, reinforcement learning)



# ПРИМЕРЫ

- **ImageNet** — задача распознавания объектов на изображении
- Решается с помощью нейронных сетей с точностью, превышающей точность работы человека



# ПРИМЕРЫ

- Аннотирование изображений



"man in black shirt is playing guitar."



"construction worker in orange safety vest is working on road."



"two young girls are playing with legos toy."



"boy is doing backflip on wakeboard."



"girl in pink dress is jumping in air."



"black and white dog jumps over bar."



"young girl in pink shirt is swinging on swing."



"man in blue wetsuit is surfing on wave."



# ПРИМЕРЫ

- Чтение по губам

*Google Deepmind в **2017** году создали модель, обученную на телевизионном датасете, которая смогла превзойти профессионального lips reader'а с канала BBC.*



# BERT ДЛЯ РЕШЕНИЯ ЗАДАЧ NLP

В октябре **2018** года Google

выпустила модель для работы с текстовыми данными под названием BERT –

*Bidirectional Encoder Representations from Transformers.*



Эта модель даёт state-of-the-art

результаты во многих задачах машинного обучения, связанных с обработкой естественного языка:

- *Определение тональности текста*
- *Перевод с одного языка на другой*
- *Определение связности предложений в тексте и др.*

# ПРИМЕР: BERT ДЛЯ АНАЛИЗА ТОНАЛЬНОСТИ ТЕКСТА

Применение BERT для анализа тональности:

Query	Score
How good is the iPhone 11	0.6 (positive)
How can a person get mental peace	0.4 (negative)
My boyfriend is not talking to me	0.7 (negative)
How to download video from youtube	0.0 (neutral)



# ПРИМЕР: BERT ДЛЯ АНАЛИЗА СУЩНОСТЕЙ

Запрос: “what is the age of Selena Gomez?”

Ответ Google с использованием BERT:

The image is a screenshot of a Google search page. The search bar at the top contains the text "what is the age of selena gomez". Below the search bar, there are tabs for "All", "News", "Images", "Videos", "Shopping", "More", "Settings", and "Tools". The search results show "About 98,200,000 results (1.27 seconds)". The main result is for "Selena Gomez / Age", which displays "27 years" and "July 22, 1992" next to a photo of Selena Gomez. Below this, there is a paragraph of text: "Obvi. Named for the late Tejano singer, Selena Quintanilla-Perez, Selena Maria Gomez was born in Grand Prairie, Texas on **July 22, 1992**. Now 26 years old, the exotic beauty of mixed Italian-Mexican descent was raised by a single mom who was a part-time stage actress. Feb 21, 2019". Below the paragraph is a link titled "How Old Is Selena Gomez and When Did She Start Acting?" with the URL "https://www.cheatsheet.com > entertainment > how-old-is-selena-gomez-and-...". At the bottom, there is a section "People also search for" with three items: "Justin Bieber 25 years", "Ariana Grande 26 years", and "Taylor Swift 29 years". On the right side, there is a sidebar for "Selena Gomez" with the subtitle "American singer". It lists "Available on" with icons for YouTube, Spotify, and Apple Music, and a link to "More music services". Below this, there is a paragraph of text: "Selena Marie Gomez is an American singer, songwriter, actress, and television producer. After appearing on the children's series Barney & Friends, she received wider recognition for her portrayal of ... Wikipedia". At the bottom of the sidebar, there are fields for "Born: July 22, 1992 (age 27 years), Grand Prairie, TX", "Height: 5' 5\"", "Net worth: US \$50 million (September 2018)", and "Parents: Mandy Teefey, Ricardo Joel Gomez".

Google

what is the age of selena gomez

All News Images Videos Shopping More Settings Tools

About 98,200,000 results (1.27 seconds)

Selena Gomez / Age

**27 years**  
July 22, 1992

Obvi. Named for the late Tejano singer, Selena Quintanilla-Perez, Selena Maria Gomez was born in Grand Prairie, Texas on **July 22, 1992**. Now 26 years old, the exotic beauty of mixed Italian-Mexican descent was raised by a single mom who was a part-time stage actress. Feb 21, 2019

**How Old Is Selena Gomez and When Did She Start Acting?**  
<https://www.cheatsheet.com > entertainment > how-old-is-selena-gomez-and-...>

People also search for

Justin Bieber 25 years Ariana Grande 26 years Taylor Swift 29 years

**Selena Gomez**  
American singer

Available on

YouTube Spotify Apple Music

More music services

Selena Marie Gomez is an American singer, songwriter, actress, and television producer. After appearing on the children's series Barney & Friends, she received wider recognition for her portrayal of ...  
[Wikipedia](#)

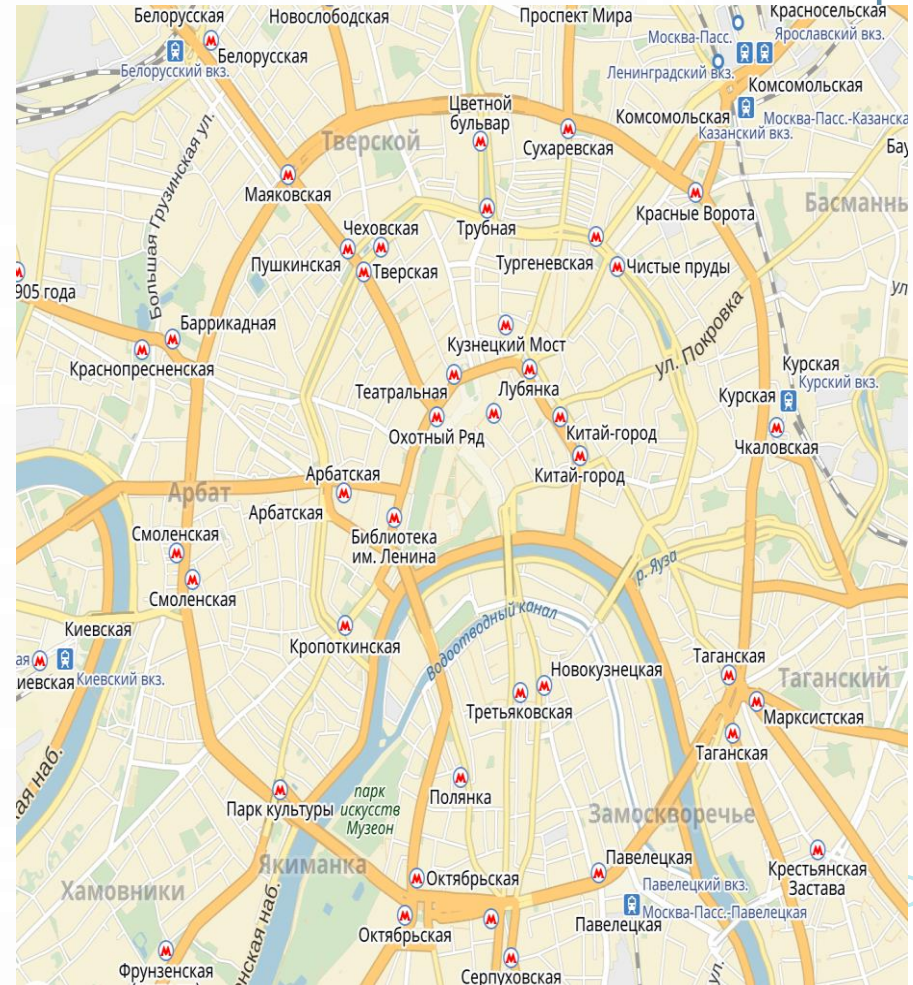
**Born:** July 22, 1992 (age 27 years), Grand Prairie, TX  
**Height:** 5' 5"  
**Net worth:** US \$50 million (September 2018)  
**Parents:** Mandy Teefey, Ricardo Joel Gomez

The image features a light gray background with a subtle pattern of concentric circles. In the four corners, there are decorative elements resembling circuit board traces or neural network connections, consisting of thin blue lines and small circles.

# **ОСНОВНЫЕ ПОНЯТИЯ МАШИННОГО ОБУЧЕНИЯ**

# ПЕРВЫЙ ПРИМЕР: ЗАДАЧА О РЕСТОРАНАХ

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



# ФОРМАЛИЗАЦИЯ

$X$  – множество объектов

$Y$  – множество ответов

$a: X \rightarrow Y$  – неизвестная зависимость

**Дано:**

$\{x_1, \dots, x_n\} \subset X$  – обучающая выборка

$\{y_1, \dots, y_n\}, y_i = y(x_i)$  - известные ответы

**Найти:**

$a: X \rightarrow Y$  – алгоритм (решающую функцию),  
приближающую  $y$  на всем множестве  $X$

# ПРИЗНАКОВОЕ ОПИСАНИЕ ОБЪЕКТОВ

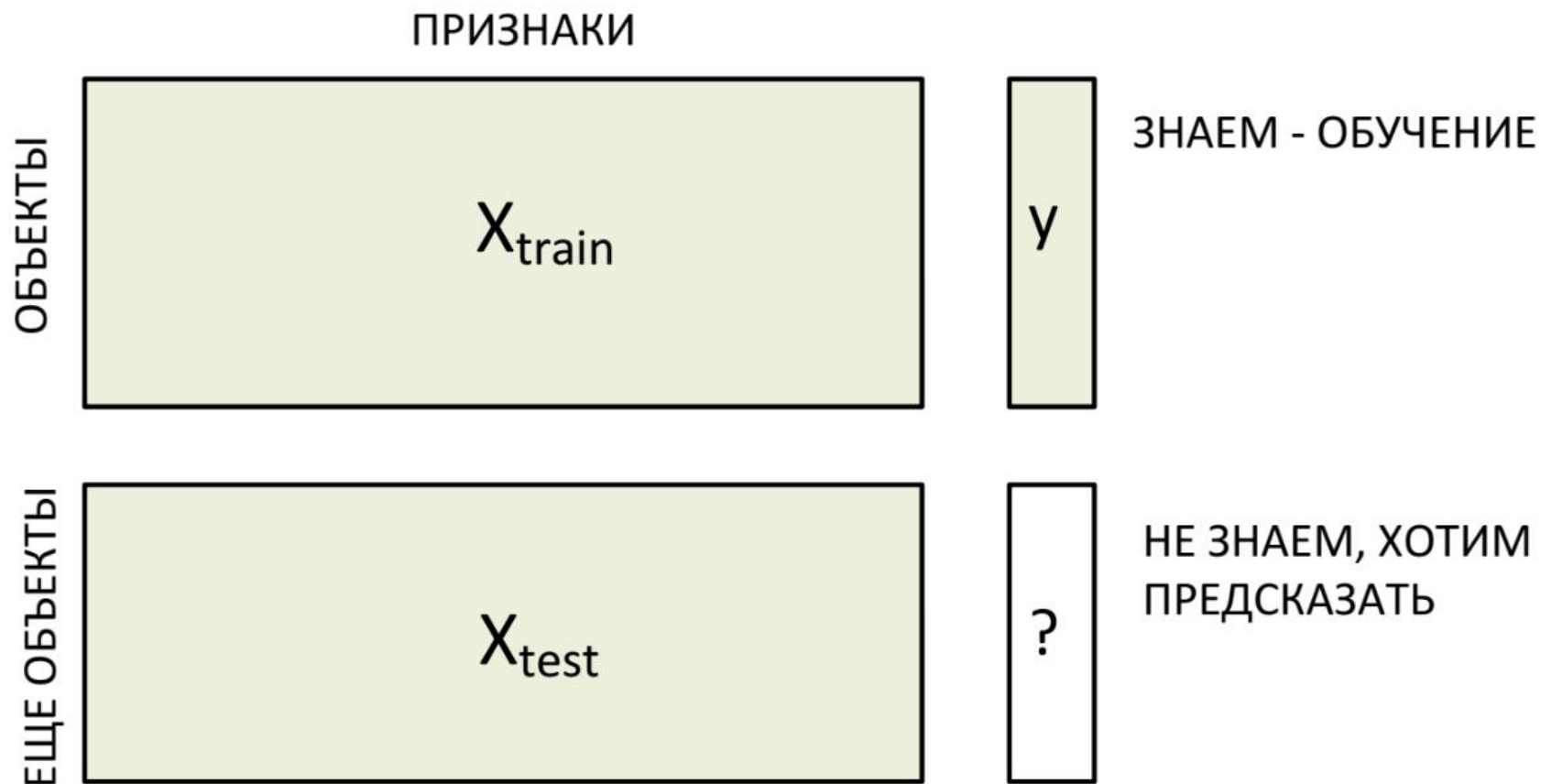
Признаки объекта  $x$  можно записать в виде вектора  
 $(f_1(x), \dots, f_n(x))$

Матрица “объекты-признаки”:

$$\begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_\ell) & \dots & f_n(x_\ell) \end{pmatrix}$$



# СТАНДАРТНАЯ ПОСТАНОВКА ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ



# СХЕМА ПОЛУЧЕНИЯ ПРЕДСКАЗАНИЯ

В задачах обучения с известными классами (обучение по прецедентам) всегда есть два этапа:

- Этап обучения (training):  
по выборке  $X = \{(x_i, y_i)\}$  строим алгоритм  $a$
- Этап применения (testing):  
алгоритм  $a$  для новых объектов  $x$  выдает ответы  $a(x)$

# ОПРЕДЕЛЕНИЯ

- **Признаки, факторы (features)** — количественные характеристики объекта
- **Обучающая выборка (training set)** — конечный набор объектов, для которых известны значения целевой переменной

**Пример:** набор ресторанов, открытых более года назад, для которых известна их прибыль за первый год

- **Объекты** — абстрактные сущности (но компьютеры работают только с числами)
- **Признаки** описывают объекты с помощью чисел

Специалист по анализу данных не является экспертом в предметной области — вся необходимая информация содержится в обучающей выборке. Эксперты нужны при формировании признаков.

# ВИДЫ ПРИЗНАКОВ

- Числовые
- Бинарные (0/1)
- Категориальные (название города, марка машины)
- Признаки со сложной внутренней структурой (изображение)

# ТИПЫ ЗАДАЧ В ЗАВИСИМОСТИ ОТ ЦЕЛЕВОЙ ПЕРЕМЕННОЙ

## Классификация

- $Y = \{0, 1\}$  – классификация на 2 класса
- $Y = \{1, \dots, M\}$  – классификация на  $M$  непересекающихся классов
- $Y = \{0, 1\}^M$  - классификация на  $M$  классов, которые могут пересекаться



# ПРИМЕРЫ ЗАДАЧ КЛАССИФИКАЦИИ

- Задачи медицинской диагностики (пациент здоров или болен)
- Задачи кредитного скоринга (выдаст банк кредит данному клиенту или нет)
- Задача предсказания оттока клиентов (уйдет клиент в следующем месяце или нет)
- Предсказание поведения пользователя (кликнет пользователь по данному баннеру или нет)
- Классификация изображений (на изображении кошка или собака)

# ПРИМЕРЫ ЗАДАЧ КЛАССИФИКАЦИИ

## Мультиклассовая классификация

- Определение типа объекта на изображении



Pedestrian



Car



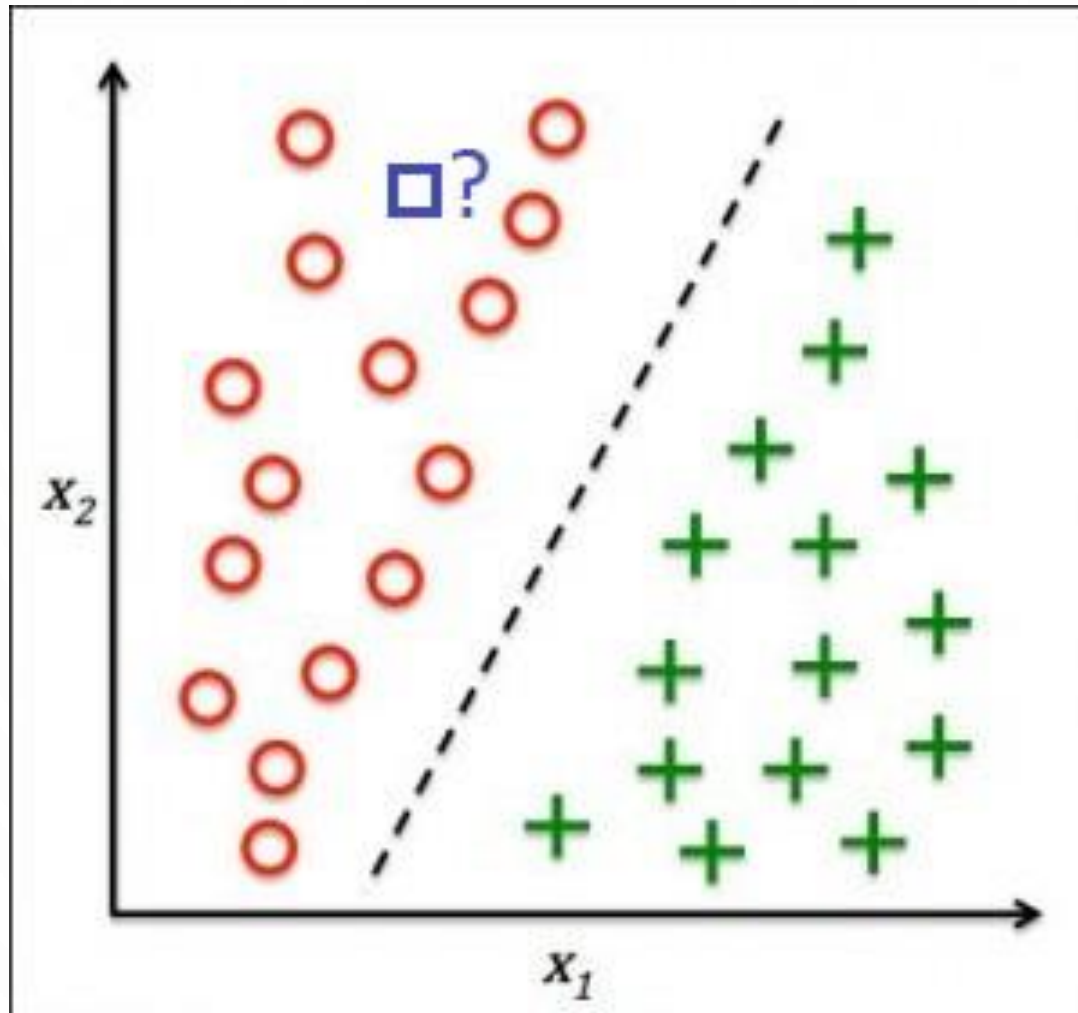
Motorcycle



Truck

- Определение наиболее подходящей профессии для данного кандидата

# ЗАДАЧА КЛАССИФИКАЦИИ



# ТИПЫ ЗАДАЧ В ЗАВИСИМОСТИ ОТ ЦЕЛЕВОЙ ПЕРЕМЕННОЙ

## Классификация

- $Y = \{0, 1\}$  – классификация на 2 класса
- $Y = \{1, \dots, M\}$  – классификация на  $M$  непересекающихся классов
- $Y = \{0, 1\}^M$  - классификация на  $M$  классов, которые могут пересекаться

## Регрессия

- $Y = R$  или  $Y = R^n$

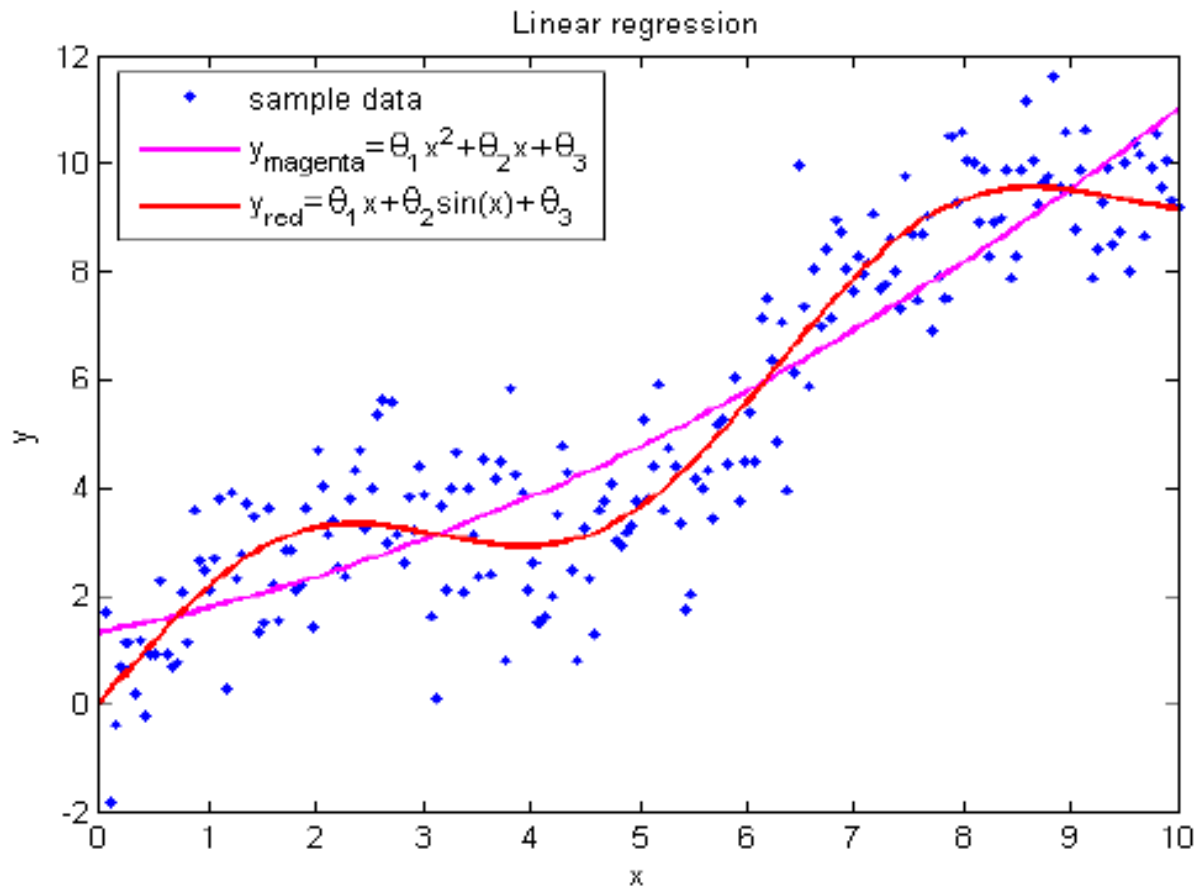
# ПРИМЕРЫ ЗАДАЧ РЕГРЕССИИ

- Предсказание стоимости недвижимости (стоимость квартиры в Москве)
- Предсказание прибыли ресторана
- Предсказание поведения временного ряда в будущем (стоимость акций)
- Предсказание зарплаты выпускника вуза по его оценкам



# ЗАДАЧА РЕГРЕССИИ

$X = Y = \mathbb{R}$ ,  $\ell = 200$ ,  $n = 3$  признака:  $\{x, x^2, 1\}$  или  $\{x, \sin x, 1\}$



# ТИПЫ ЗАДАЧ В ЗАВИСИМОСТИ ОТ ЦЕЛЕВОЙ ПЕРЕМЕННОЙ

## Классификация

- $Y = \{0, 1\}$  – классификация на 2 класса
- $Y = \{1, \dots, M\}$  – классификация на  $M$  непересекающихся классов
- $Y = \{0, 1\}^M$  - классификация на  $M$  классов, которые могут пересекаться

## Регрессия

- $Y = R$  или  $Y = R^n$

## Ранжирование

- $Y$  – конечное упорядоченное множество

# ЗАДАЧИ, В КОТОРЫХ НЕТ ЦЕЛЕВОЙ ПЕРЕМЕННОЙ

- **Кластеризация** – задача разделения объектов на группы, при этом где целевые переменные для объектов неизвестны (или не существуют). Разделение происходит только на основе признаков описаний объектов.
- **Понижение размерности** – задача генерации новых признаков (их число меньше, чем число старых), так, что с их помощью задача решается не хуже, чем с исходными.
- **Оценивание плотности** – задача приближения распределения объектов.
- **Визуализация** – задача изображения многомерных объектов в 2х или 3х мерном пространстве с сохранением зависимостей между ними.

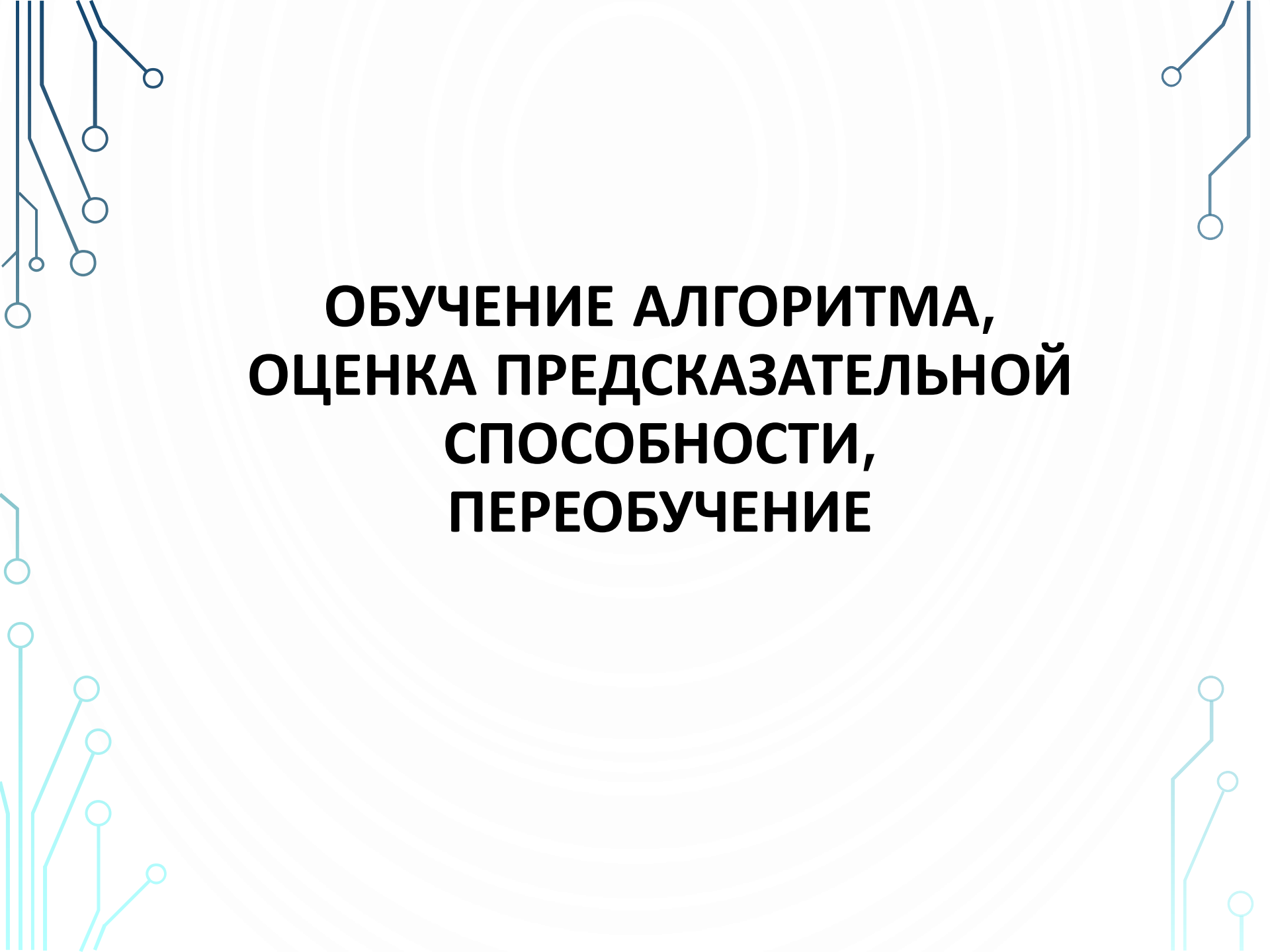
# ТИПЫ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

- Если нам известны значения целевой переменной, то есть алгоритм обучается так, чтобы правильно предсказывать целевую переменную – это **обучение с учителем**. Сюда относят классификацию, регрессию и ранжирование.

# ТИПЫ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

- Если нам известны значения целевой переменной, то есть алгоритм обучается так, чтобы правильно предсказывать целевую переменную – это **обучение с учителем**. Сюда относят классификацию, регрессию и ранжирование.
- Если нам неизвестны значения целевой переменной или целевая переменная вообще отсутствует, то есть алгоритм обучается только по признакам объектов, то это **обучение без учителя**. Примерами обучения с учителем являются кластеризация, понижение размерности и др.



The image features a light blue background with a subtle pattern of concentric circles. In the four corners, there are decorative elements resembling circuit board traces or neural network connections, consisting of thin blue lines and small circles.

# **ОБУЧЕНИЕ АЛГОРИТМА, ОЦЕНКА ПРЕДСКАЗАТЕЛЬНОЙ СПОСОБНОСТИ, ПЕРЕОБУЧЕНИЕ**

# ПОСТРОЕНИЕ МОДЕЛИ

- На этапе *обучения* происходит настройка параметров алгоритма  $a$ , который для каждого объекта  $x$  в нашей задаче выдает предсказание:  $a(x)$ .
- Например, если наш алгоритм имеет вид

$$a(x) = w_0 + w_1 x,$$

то в процессе обучения определяются значения параметров  $w_0, w_1$ .

# МЕТРИКИ КАЧЕСТВА

В задачах машинного обучения для оценки качества моделей и сравнения различных алгоритмов используются *метрики качества*.

# МЕТРИКИ КАЧЕСТВА

В задачах машинного обучения для оценки качества моделей и сравнения различных алгоритмов используются ***метрики качества***.

## Примеры:

- Среднеквадратичная ошибка – для регрессии

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

# МЕТРИКИ КАЧЕСТВА

В задачах машинного обучения для оценки качества моделей и сравнения различных алгоритмов используются ***метрики качества***.

## Примеры:

- Среднеквадратичная ошибка – для регрессии
- **Доля правильных ответов** – для классификации

$$\text{accuracy}(a, X) = \frac{1}{l} \sum_{i=1}^l [a(x_i) = y_i]$$

# ОЦЕНКА ПРЕДСКАЗАТЕЛЬНОЙ СПОСОБНОСТИ АЛГОРИТМА

- Перед началом обучения отложим часть обучающих объектов и не будем использовать их для построения модели (отложенная выборка).

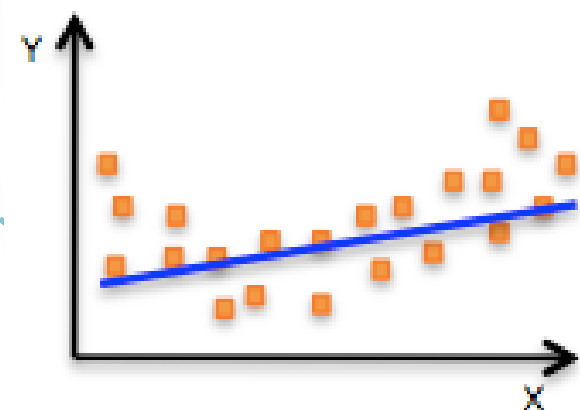


# ОТЛОЖЕННАЯ ВЫБОРКА

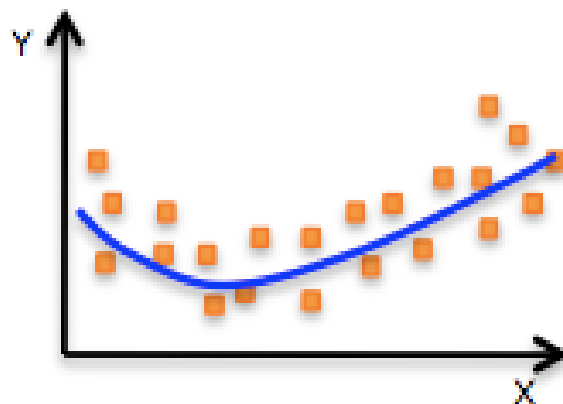
- Перед началом обучения отложим часть обучающих объектов и не будем использовать их для построения модели (отложенная выборка).
- Тогда можно измерить качество построенной модели на отложенной выборке и оценить ее предсказательную силу.



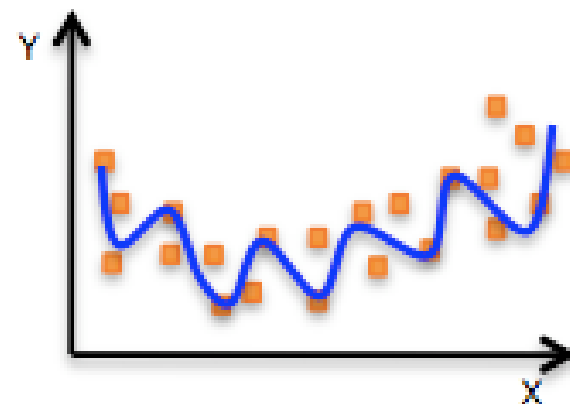
# ПЕРЕОБУЧЕНИЕ И НЕДООБУЧЕНИЕ



**Underfitting**

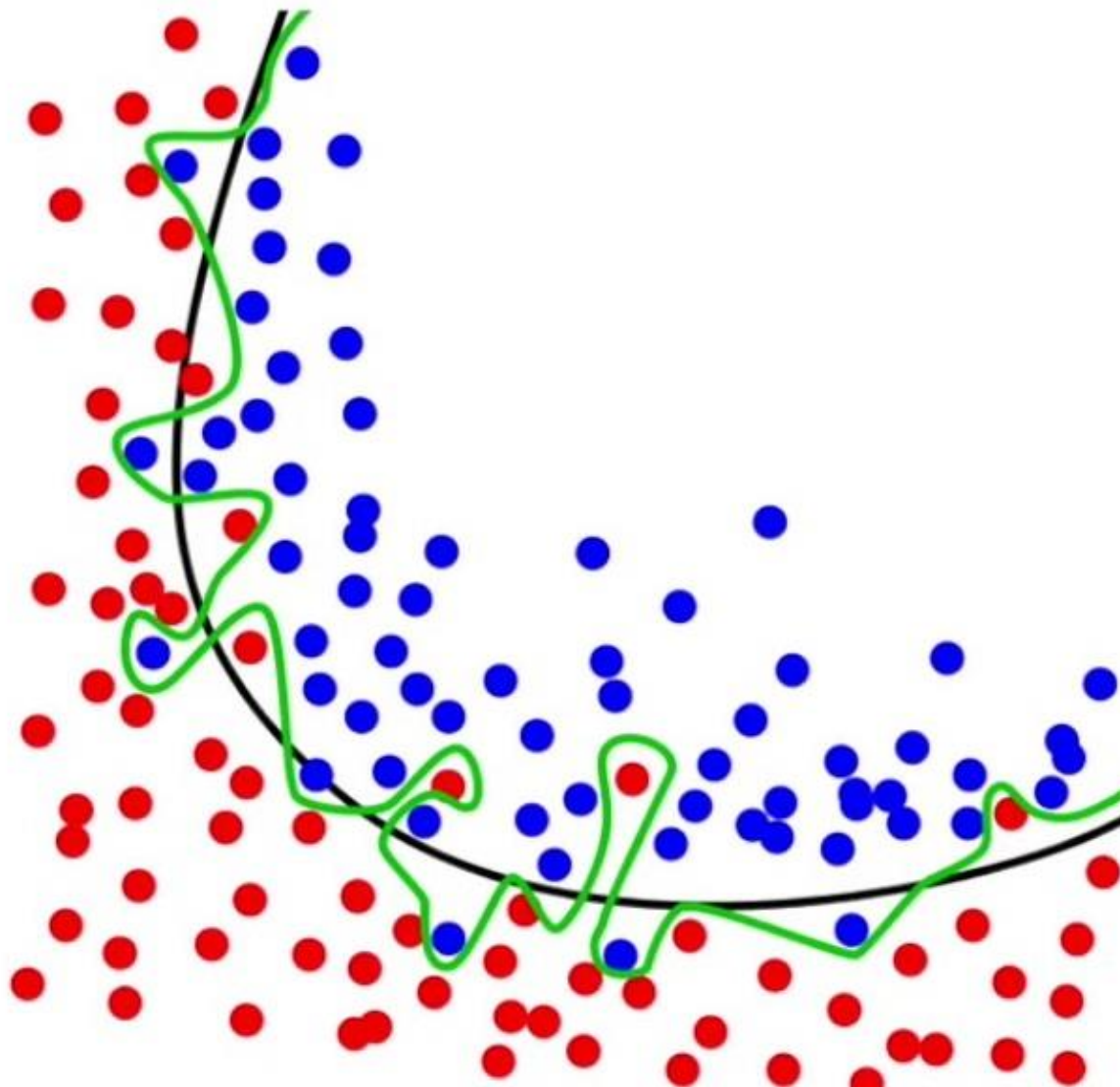


**Just right!**



**overfitting**

# ПРИМЕР ПЕРЕОБУЧЕНИЯ В ЗАДАЧЕ КЛАССИФИКАЦИИ


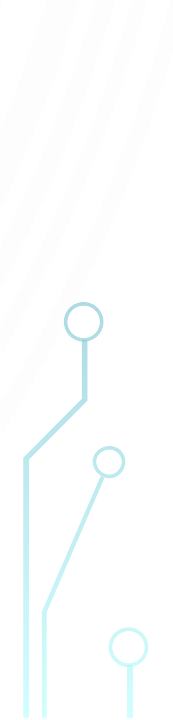


# ПРИЗНАК ПЕРЕОБУЧЕНИЯ

- *Если качество на отложенной выборке сильно ниже качества на обучающих данных, то происходит переобучение*



# АЛГОРИТМ РЕШЕНИЯ ЗАДАЧИ АНАЛИЗА ДАННЫХ

1. Постановка задачи
  2. Сбор данных
  3. Выделение признаков
  4. Выбор функции потерь и метрики качества
  5. Предобработка данных
  6. Построение модели
  7. Оценивание качества модели
- 
- 

# ЗАДАЧА О ДВИЖЕНИЯХ ЧЕЛОВЕКА

- В *открытую папку* вам необходимо загрузить по 4 трека каждого типа движений. В итоге мы соберем набор данных для построения алгоритма.
- Данные из *закрытой папки* (1 трек каждого типа от каждого студента) будут загружены на платформу kaggle, ответы для треков из этой папки вам не будут даны. Качество вашего алгоритма будем измерять как долю верно классифицированных треков из закрытой папки.

# ЗАДАНИЕ К СЛЕДУЮЩЕМУ ЗАНЯТИЮ

- Загрузить треки в гугл-папки
- Написать код, выполняющий следующее:
  - 1) Считывание всех треков из папки
  - 2) Определение по названию трека типа движения и сохранение названия в отдельную переменную
  - 3) Визуализация графиков g-force по данным из папки. Подпись графика – тип движения.
  - 4) Поизучать глазами/в python треки и попытаться выявить различия между типами движений. Записать свои идеи куда-нибудь – будем обсуждать.