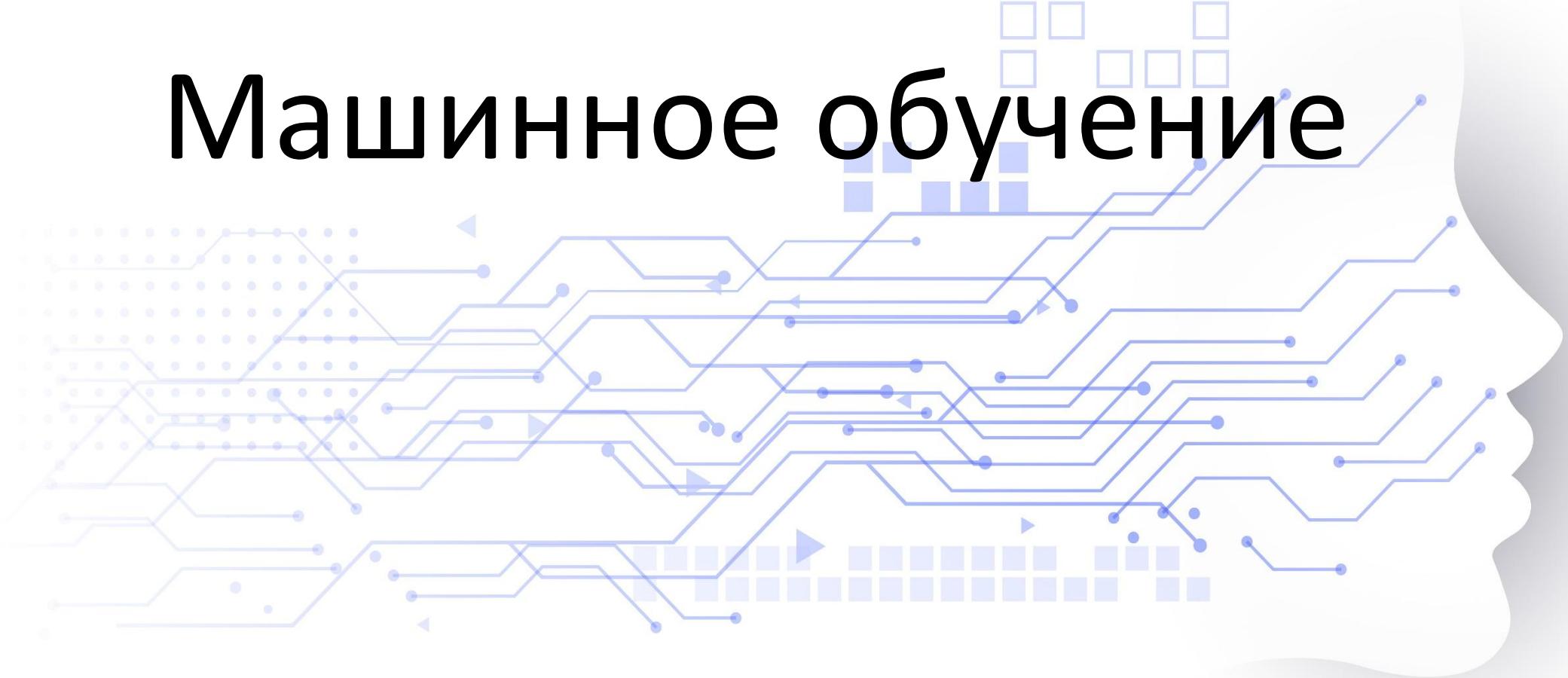


# Машинное обучение





**Резаиан Наим**

E-mail: [rezaian-n@rudn.ru](mailto:rezaian-n@rudn.ru)

Telegram: [@NaeimRezaeian](https://t.me/NaeimRezaeian)

1. Заведующий лабораторией искусственного интеллекта
2. Руководитель направления разработок Центра развития цифровых технологий в образовательных процессах
3. Старший преподаватель факультета искусственного интеллекта

# Вы должны знать

## ✓ Разработка алгоритмов

Временная сложность алгоритма и вычислительная сложность

## ✓ Линейная алгебра

Матрицы и операции над матрицами, Векторы, Система линейных алгебраических уравнений

Обратная матрица, Собственный вектор, Невырожденная матрица, Сингулярное разложение

## ✓ Анализ функций многих переменных

Производная функции, Интеграл, Касательное пространство

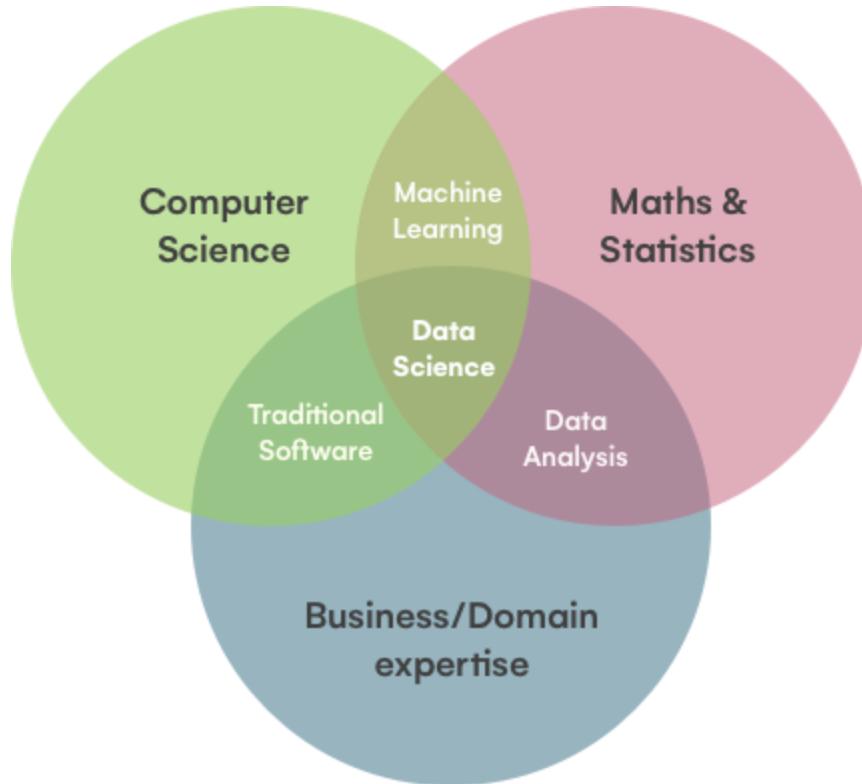
## ✓ Теория вероятностей

Случайная величина, Математическое ожидание , Дисперсия случайной величины, ...

## ✓ Программирование

Python

# Что такое машинное обучение?



- **Домашнее задание – 50 баллов**
- **Промежуточная аттестация – 20 баллов**
- **Итоговая аттестация – 20 баллов**
- **Активность на занятиях – 10 баллов**

# История развития искусственного интеллекта

❖ Идея — 1943 год

❖ Активное развитие — с 2022 года

1943

Первая математическая модель искусственной сети

1950

Алан Тьюринг публикует статью в журнале *Mind*, где предлагает тест на «интеллектуальность машины»

1956

Проведён Дартмутский семинар, заложивший теоретические основы искусственного интеллекта

1980-е

Разработаны первые свёрточные нейронные сети

1991

Опубликована первая версия языка программирования Python

1997

Deep Blue побеждает Гарри Каспарова в шахматном турнире

2005

Джеффри Хинтон и Йошуа Бенджи начинают обучать первые глубокие нейронные сети

2011

IBM Watson побеждает в *Jeopardy!*

2012

Свёрточная нейросеть AlexNet выигрывает конкурс по распознаванию изображений

# История развития искусственного интеллекта

❖ Идея — 1943 год

❖ Активное развитие — с 2022 года

2013

Создана исследовательская лаборатория искусственного интеллекта Facebook\*, которая позже превратится в Meta\* AI, создающую ИИ для метавселенной

2015

Илон Маск и Сэм Альтман основывают НКО Open AI

2015

DeepMind создаёт AlphaGo — нейросеть, способную победить человека в игре го

2016

AlphaGo побеждает Ли Седоля

2017

Выход знакового препарата, заложившего архитектуру «Attention is all you need»

2018

DeepMind создаёт AlphaFold

2020

DeepMind выпускает AlphaFold 2

2022

Запуск ChatGPT

2022

Опубликована «белковая» языковая модель ESM-2

2023

Разработана генеративная модель для белкового дизайна RFdiffusion

Мы ежедневно взаимодействуем с ИИ

(и не всегда знаем об этом)

текущее положение дел: миром правит слабый ИИ



кинопоиск



Яндекс Go Такси



Яндекс Музыка

СБЕР БАНК

маруся

NETFLIX



иви



Яндекс Директ



# Почему сейчас?



Российский университет  
дружбы народов

## **1. Большие данные**

## **2. Вычислительные ресурсы**

Графические ускорители вычислений (GPU)

Специализированные процессоры (Tensor Processing Unit, TPU)

Облачные вычисления

## **3. ML engineer**

Модная профессия, за это деньги платят

# Иерархия в искусственном интеллекте

Искусственный интеллект

Машинное обучение  
сотни других методов обучения

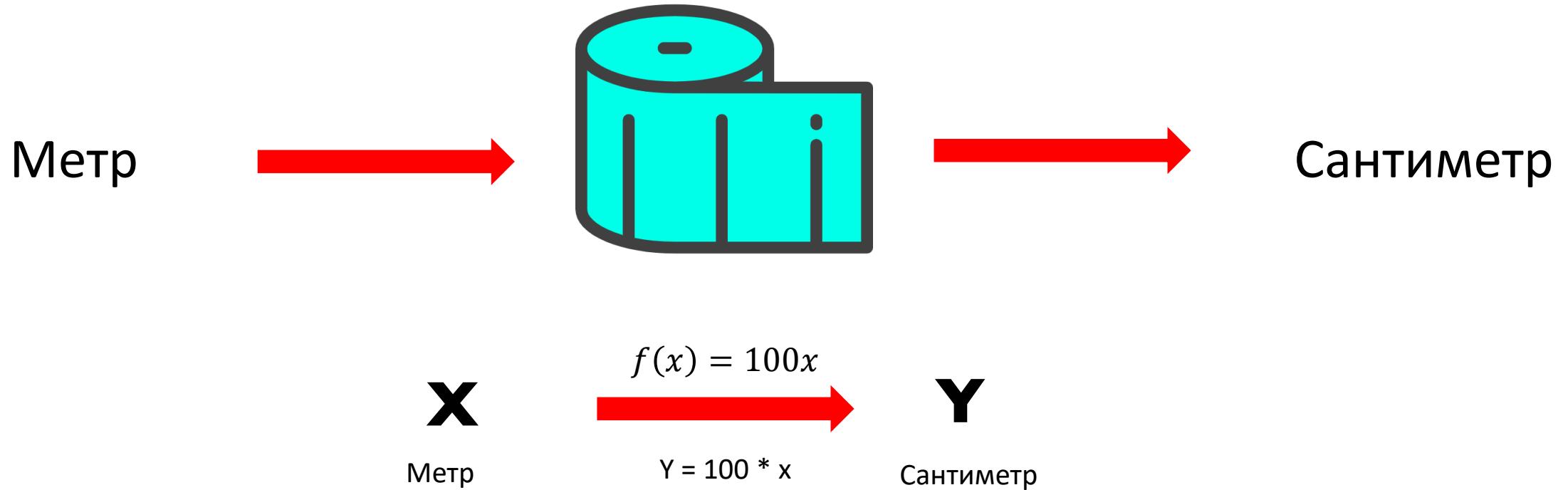
Нейросети

Глубокое обучение

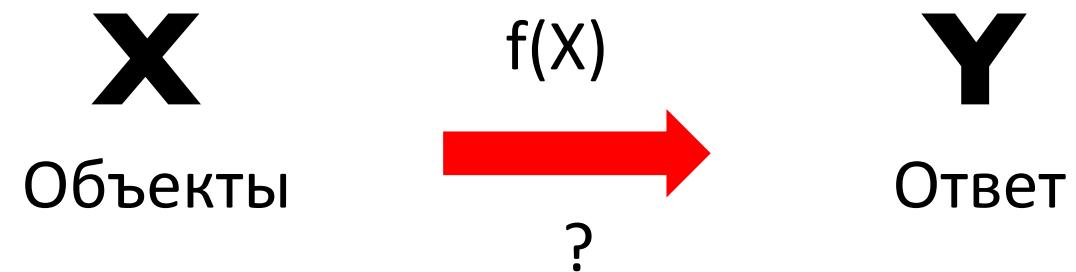
Три главные составляющие:

- ⌚ **Данные** – исходные результаты измерений
- ⌚ **Признаки (features)** – выделение характерных особенностей в данных
- ⌚ **Алгоритм** – программа, которая способна подстраиваться (обучаться) под входные данные и выдавать требуемый результат

Как перевести метр в сантиметр?



## Больше сложных зависимостей



$$Y \approx f(X)$$

# Моделирования прогноза погоды



Уравнение Навье - Стокса

$$\frac{\partial \vec{v}}{\partial t} = -(\vec{v} \cdot \nabla) \vec{v} + \nu \Delta \vec{v} - \frac{1}{\rho} \nabla p + \vec{f}$$

## Прогнозирование цены акций на рынке



Авторегрессии проинтегрированного скользящего среднего

$$X_t - \alpha_1 X_{t-1} - \cdots - \alpha_{p'} X_{t-p'} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q},$$

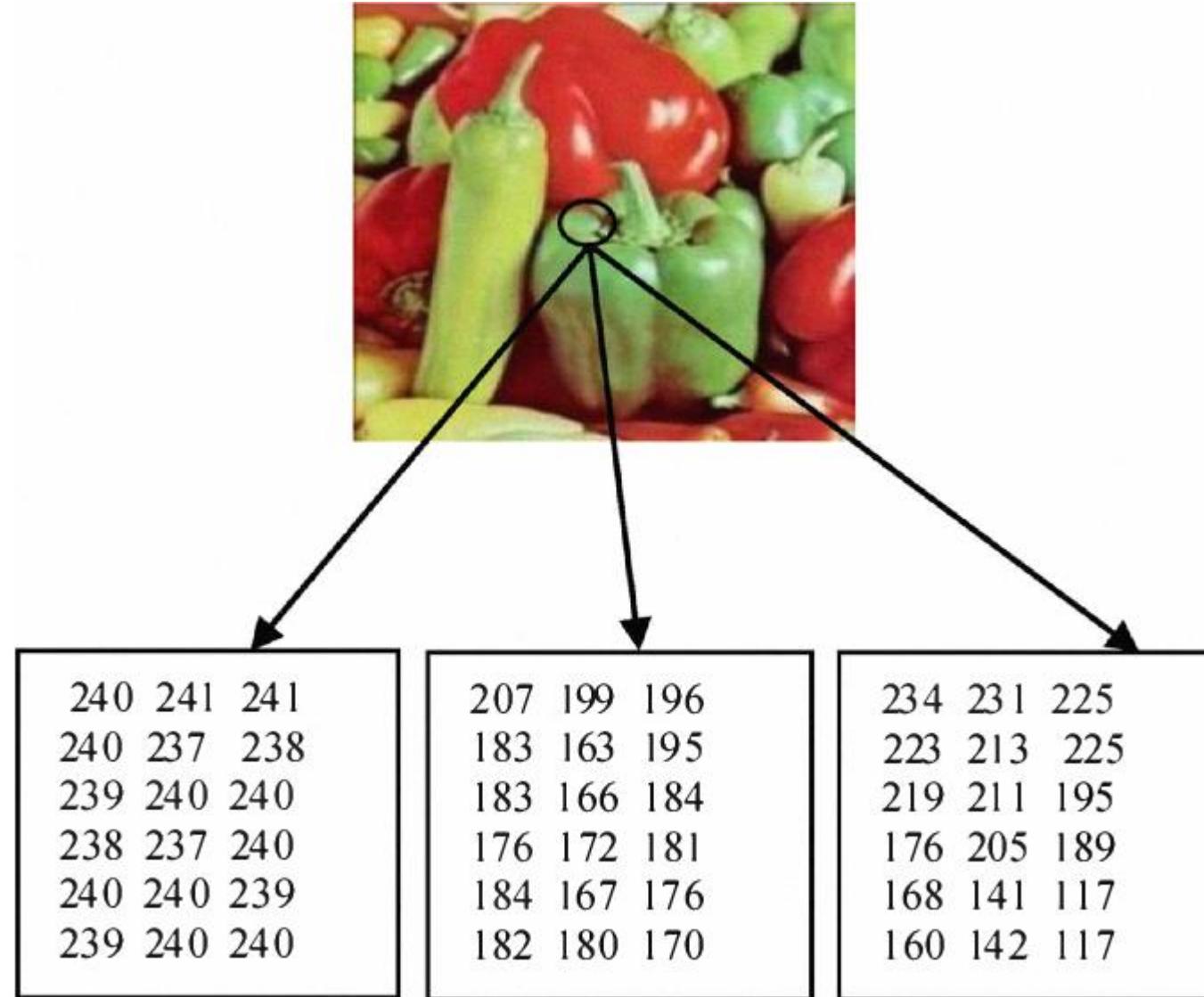
Как отличить кота от собаки?



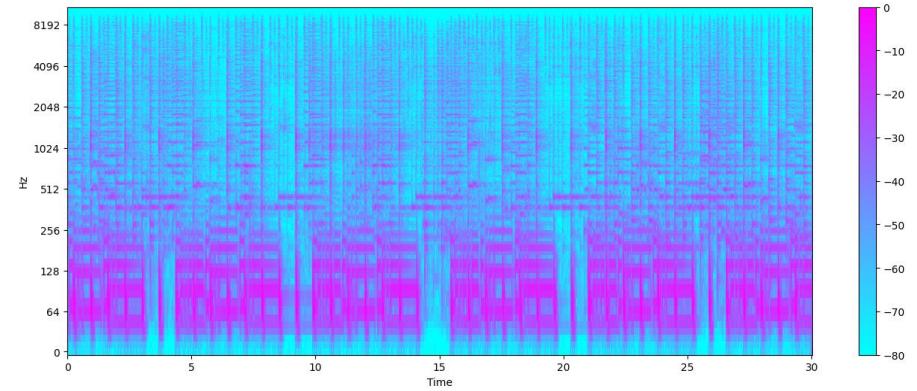
Или



# Что такое изображение для компьютера?



# Как определить жанр музыки?



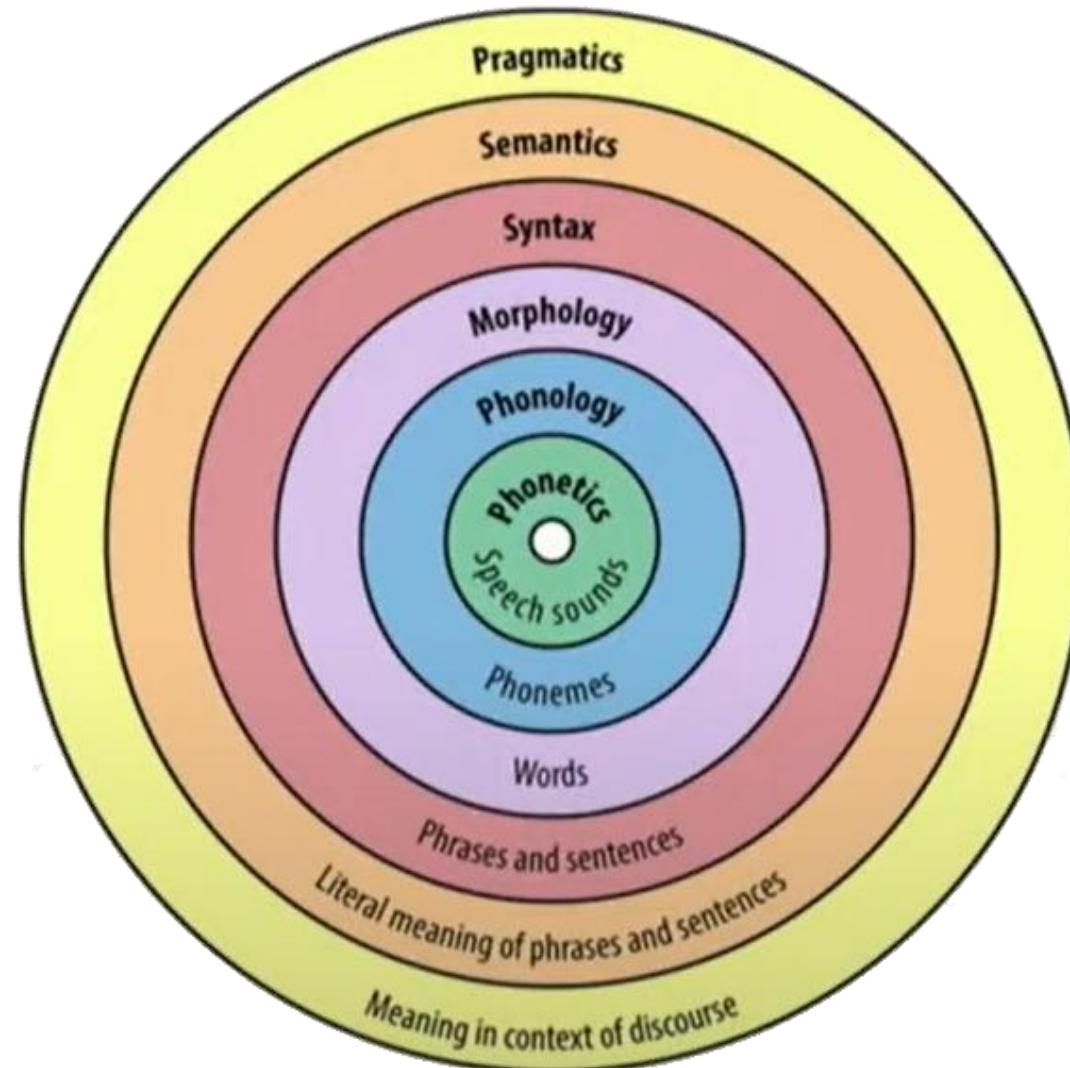
{Хип-хоп, Рок, Шансон, ... }

## Мама мыла раму

\u041c\u0430\u043c\u0430  
\u043c\u044b\u043b\u0430  
\u0440\u0430\u043c\u0443



# Язык

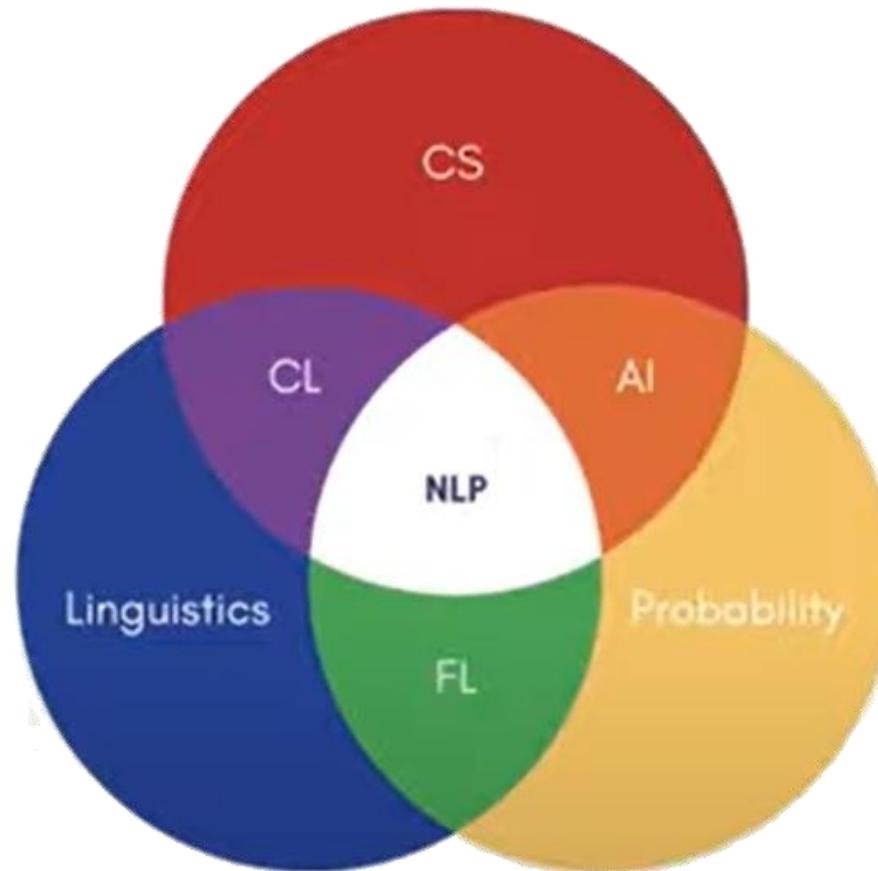




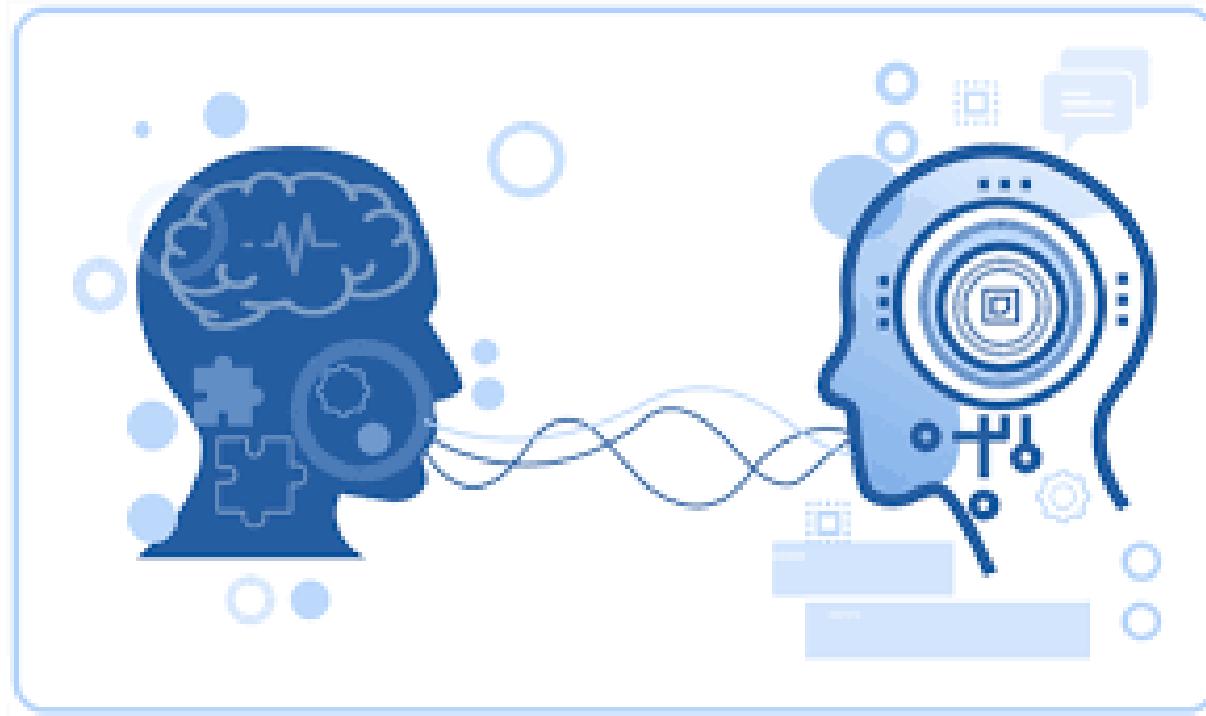
# Что такое NLP

## Natural Language Processing

### обработка естественного языка



## Цель NLP



Научить компьютер понимать человеческий язык так, чтобы выполнять практические задачи.

# Как определить тональность текста?

Крайне разочарован покупкой. Товар не соответствует описанию на сайте, качество оставляет желать лучшего. Доставка заняла гораздо больше времени, чем было обещано, и при этом товар пришел с повреждениями. Обслуживание клиентов также оставляет желать лучшего — на мои вопросы и жалобы отвечали медленно и неохотно. Опыт был неприятный, и я бы не рекомендовал этот сайт для покупок.



# Как определить тональность текста?

Прекрасный магазин! Доставка заняла всего месяц, что позволяет с нетерпением ждать каждую посылку. Качество товара удивило – редко удается найти что-то настолько уникальное. Не каждый день встретишь товар с таким количеством сюрпризов! В общем, если любите неожиданные открытия, этот магазин точно для вас!



# Как определить тональность текста?

X – текст на русском языке

$$f(X) = \{ -1, 1 \}$$

На входе нечисловые данные и вряд ли есть точная формула для  $f(x)$



# Неоднозначность

## **Ключ**

Металлический стержень для открывания и закрывания замков

Инструмент для отвинчивания и завинчивания гаек

Родник, источник воды

## **Некоторые типы стали есть в нашем цехе**

Нехорошие люди (типы) едят прямо в цехе

В цехе имеется несколько различных видов сплава железа с углеродом

## **Речь**

Я и крыс ем еще

Я икры съем еще

# Особенности русского языка

## **Флективность**

Форма слова изменяется (приставки, суффиксы, окончания)

Существительные: падеж, число

Прилагательные: падеж, число, род

Глагол: время, вид

Английский: cat, cats

Русский: кошка, кошки, кошке, кошку, кошкою, кошке, кошек,  
кошкам, кошками, кошках.

# Особенности русского языка

## **Свободный порядок слов**

Я люблю обработку естественного языка

Люблю я обработку естественного языка

Обработку естественного языка люблю я



# Особенности русского языка

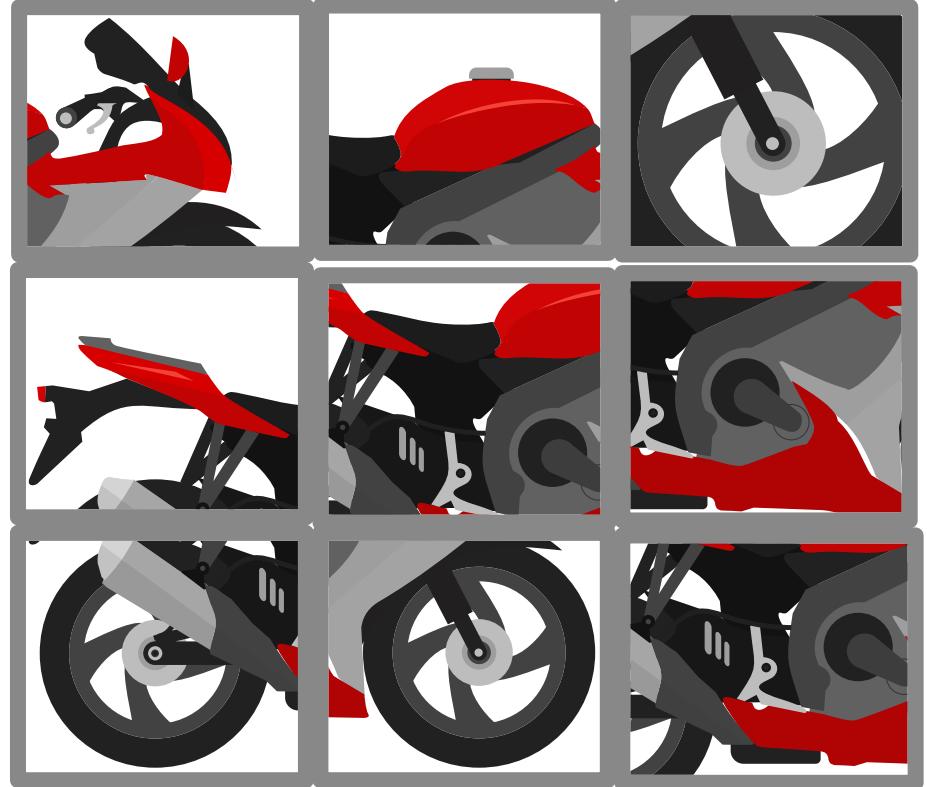
## Анафора

Купил телефон, но он мне не понравился и я сдал его назад.

Ранее главный редактор газеты опроверг свое высказывание, где он выступил с сенсационной новостью, которую ему сообщил достоверный источник, который ввел его в заблуждение.

## Feature extraction

Извлечение признаков



# Основные виды методов машинного обучения



## Обучение с учителем (Supervised Learning)

Модель обучается на основе размеченных данных, где для каждого примера входных данных имеется соответствующая метка.

Целью этого обучения является разработка модели, которая способна предсказывать целевую переменную для новых, ранее не виденных данных.



## Обучение без учителя (Unsupervised Learning)

Модель обучается на неразмеченных данных, то есть данных, для которых нет предварительно заданных меток классов или целевых переменных.

Вместо этого, задачей в обучении без учителя является выявление скрытых структур, паттернов, группировок или зависимостей в данных.



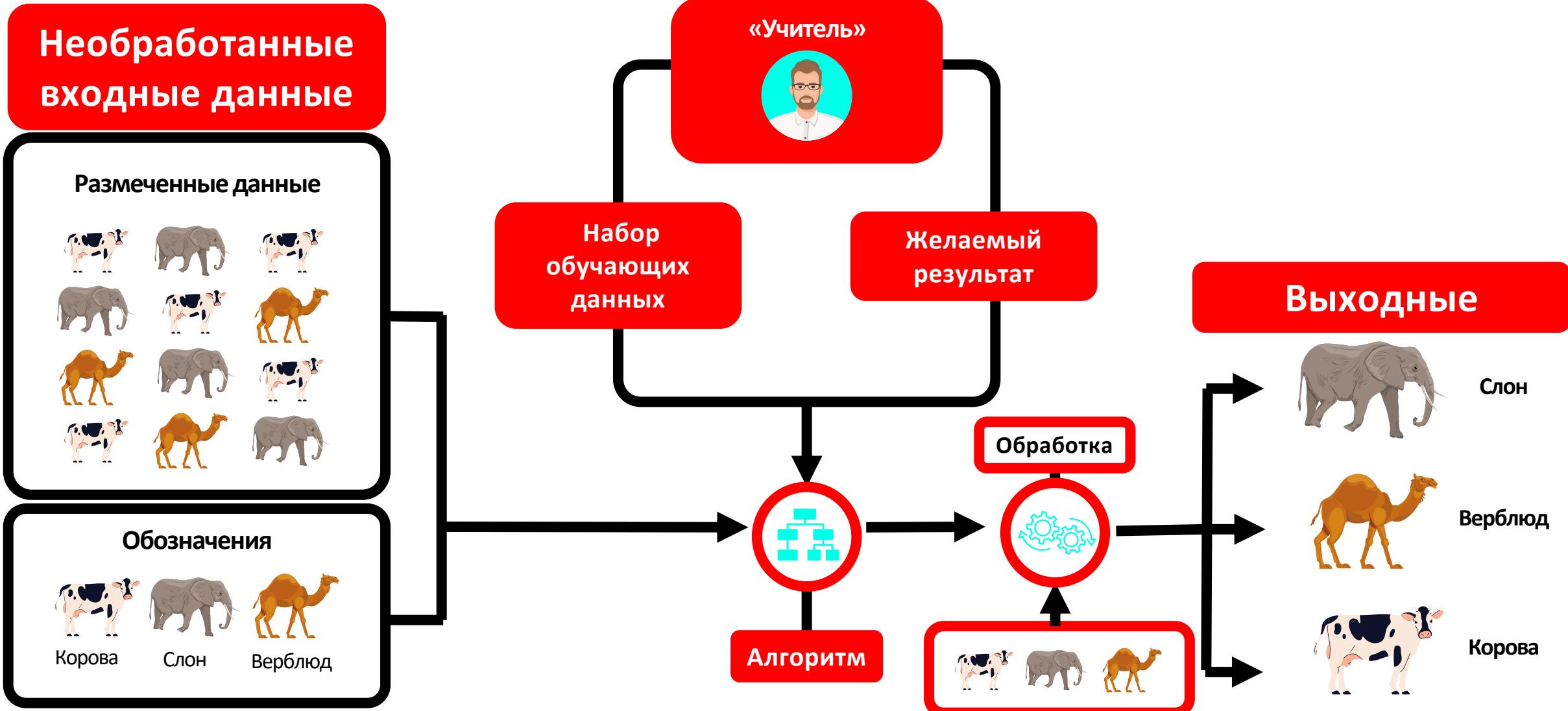
## Обучение с подкреплением (Reinforcement Learning):

Это вид машинного обучения, в котором агент (искусственный интеллект) взаимодействует с окружающей средой и принимает решения с целью максимизации награды или определенного критерия успеха. В RL нет предварительных пар входных данных и целевых переменных, как в обучении с учителем.

Вместо этого агент учится, испытывая различные действия в среде и анализируя их последствия.

# Обучение с учителем

## Supervised Learning



# Обучение с учителем

## Supervised Learning



### Набор данных:

Обучающий набор, в котором для каждой записи предоставлен **правильный ответ**



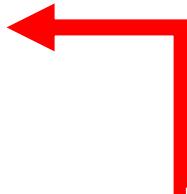
### Цель:

Поиск функции  $f$ , которая наилучшим образом приближает зависимость между  $x$  и  $y$

$$f : X \rightarrow Y$$

Входные  
данные

$$\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})\}$$



Целевая переменная



# Обучение с учителем (Supervised Learning)



## Классификация (Classification):

Модель обучается присваивать входным данным одну из заранее определенных меток классов.

Примеры включают задачи, такие как определение категории электронного письма (спам или не спам) или распознавание изображений (классификация на категории).



## Регрессия (Regression):

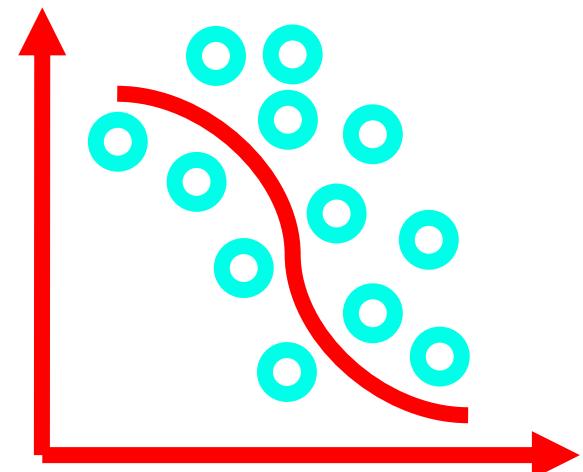
Модель предсказывает числовое значение на основе входных данных.

Примеры включают задачи, такие как прогнозирование цен на недвижимость или предсказание временных рядов, таких как температура и курс валюты.



## С учителем:

- Классификация
- Регрессия
  - Какого цвета конфета?
  - Сколько будет стоить квартира?
  - Какая будет цена на нефть в 2025 году?



# Идентификация спама



**Вход:** письмо



**Выход:** спам / не спам



Dear Sir. First, I must solicit your confidence in this transaction, this is by virtue of its nature as being utterly confidential and top secret. ...



To be removed from future mailings, simply reply to this message and put "remove" in the subject. in the subject. 99 million email addresses for only \$99



Ok, I know this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.



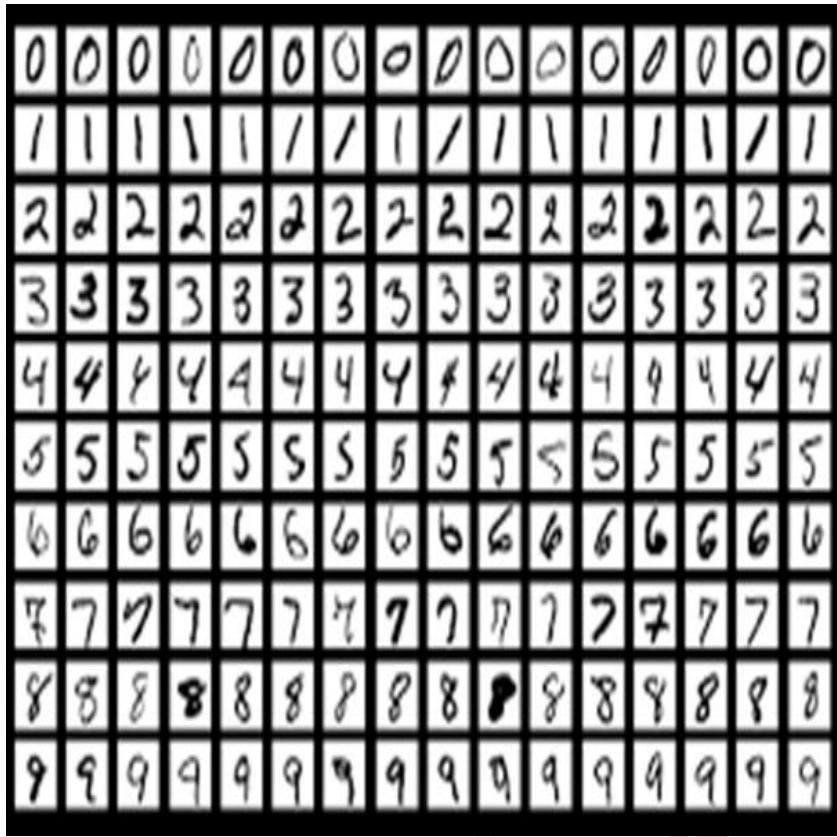
# Распознавание рукописных номеров



Вход: Изображение рукописного номера

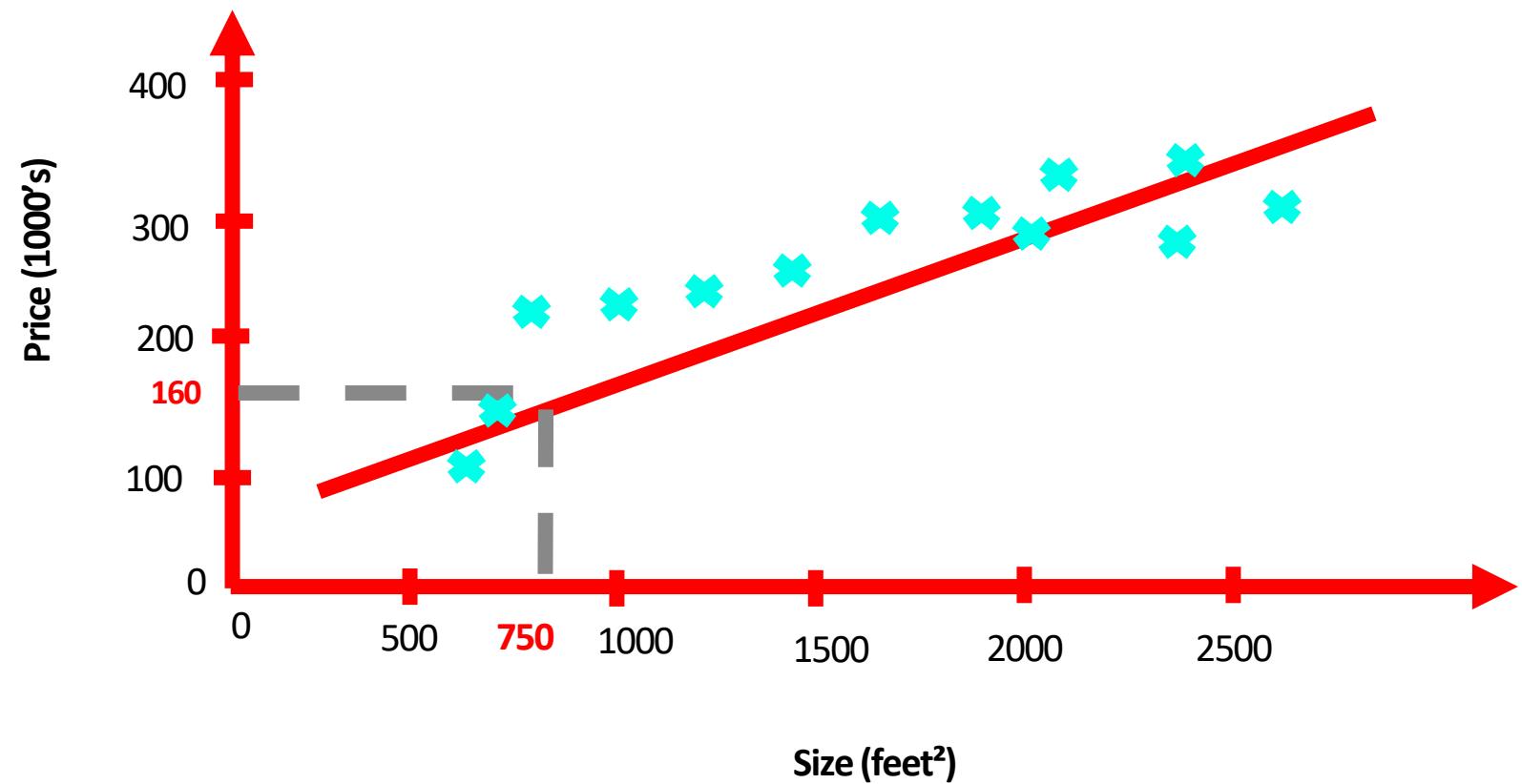


Выход: Число



# Стоимость жилого недвижимого имущества

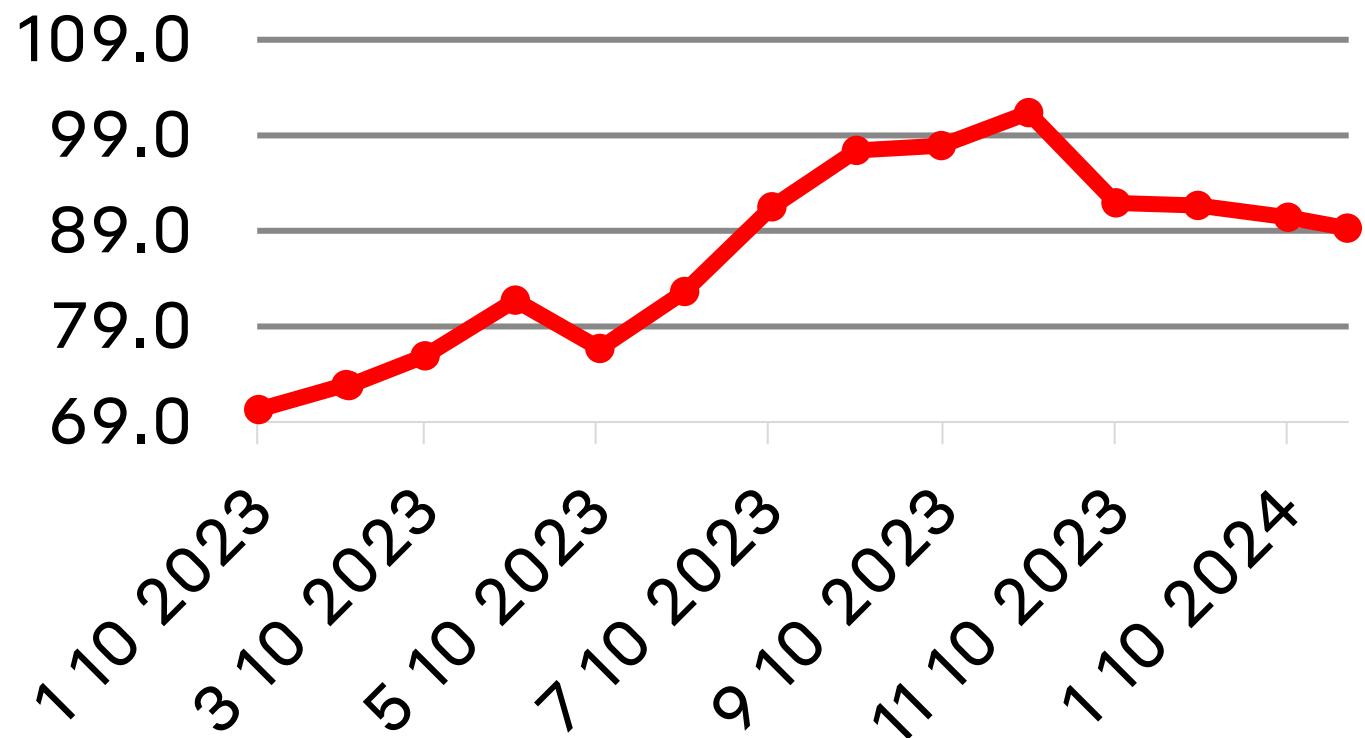
- ➡ Вход: площадь квадрата
- ✖ Выход: стоимость квартиры



## Курс валют

⌚ Вход: дата

💸 Выход: стоимость курса в рублях



# Задача кредитного скоринга

❖ Набор содержит  
13 функций

1	Кредит	Уникальный идентификатор
2	Пол	Пол заявителя Мужчина/женщина
3	Женат	Семейное положение заявителя, значения будут равны «Да»/«Нет»
4	Зависимые	Он показывает, есть ли у заявителя какие-либо иждивенцы или нет
5	Образование	Это покажет нам, получил ли заявитель высшее образование или нет
6	Self_Employed	Это определяет, что заявитель является самозанятым, т.е. «Да»/«Нет»
7	Доход соискателя	Доход кандидата
8	Coapplicantincome	Доход соавтора заявки
9	Количество кредитов	Сумма кредита (в тысячах)
10	Loan_Amount_Term	Условия кредита (в месяцах)
11	Credit_History	Кредитная история погашения физическим лицом своих долгов
12	Property_Area	Площадь собственности, т.е. Сельская/Городская/Полугородская
13	Loan_Status	Статус одобренного кредита: Y – да, N – нет

# Обучение без учителя

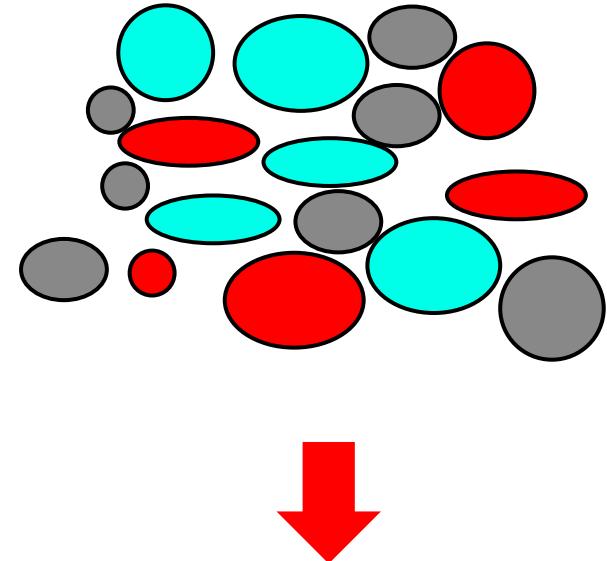
## Unsupervised Learning

❖ Модель обучается на неразмеченных данных, то есть данных, для которых нет предварительно заданных меток классов или целевых переменных. Вместо этого, задачей в обучении без учителя является выявление скрытых структур, паттернов, группировок или зависимостей в данных.

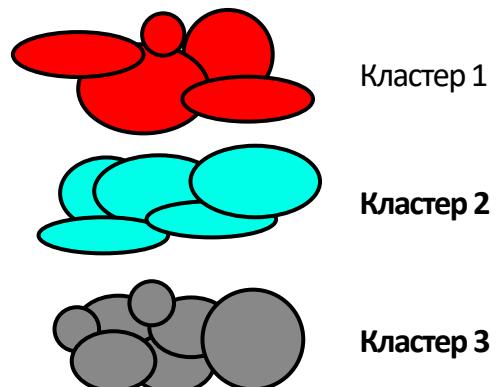
### 3 Кластеризация

- Машинное обучение – финансы – игры
- Разложить похожие вещи по кучкам

### До кластеризации



### После кластеризации



# Обучение без учителя

## Unsupervised Learning



# Кластеризация

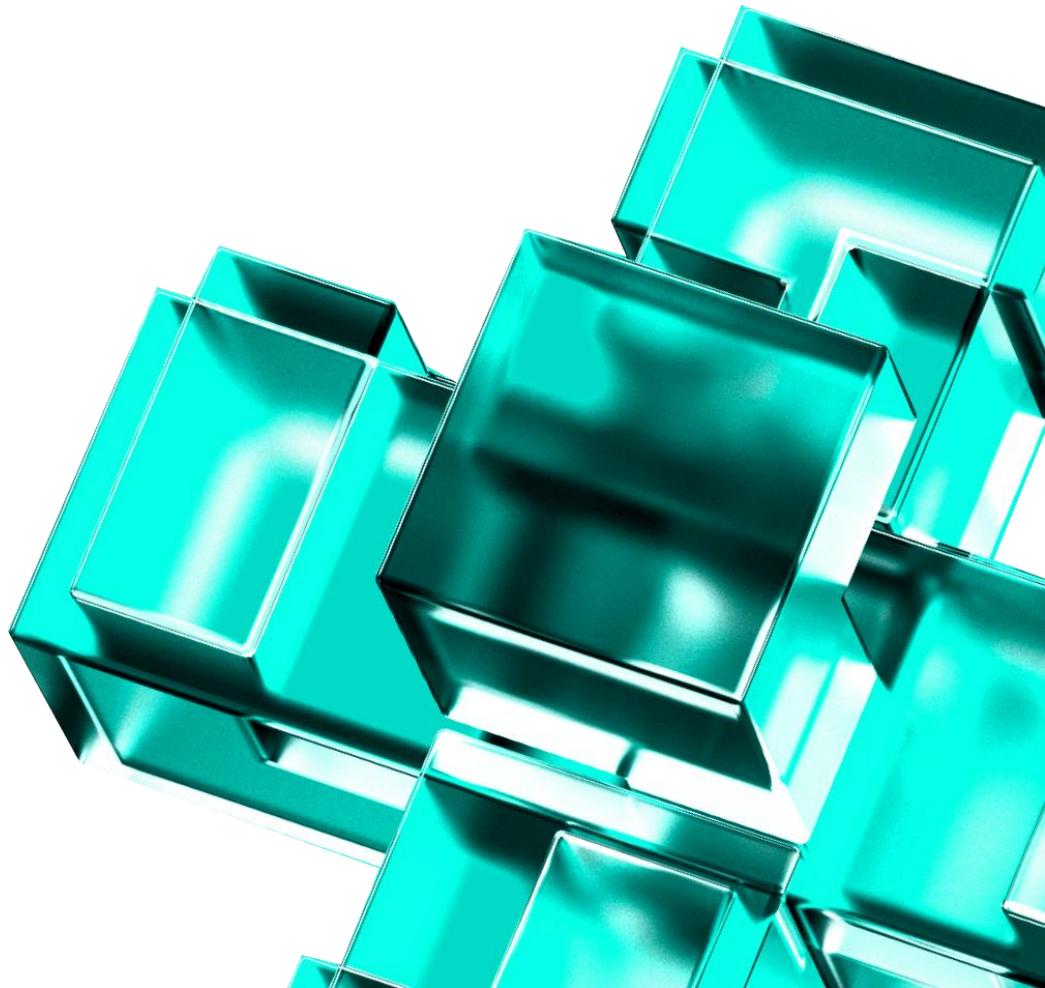


Найти похожих людей

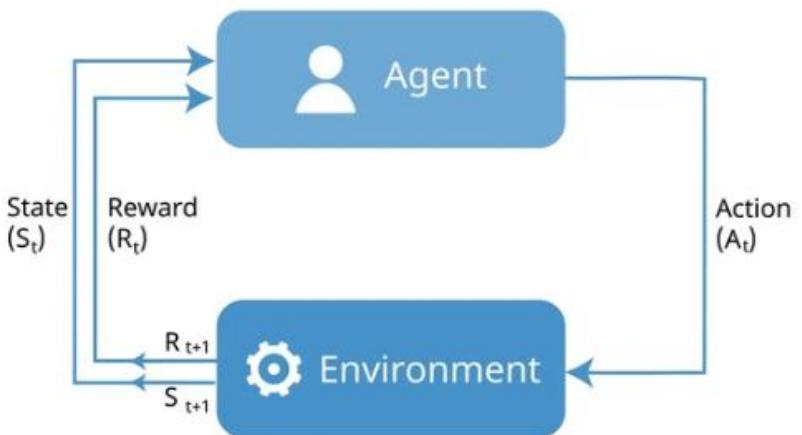


## Сфера применения

- ⌚ Анализ поведения отдельных клиентских групп
- ⌚ Исследование рынка конкурентов
- ⌚ Изучение статистики выздоровления
- ⌚ Анализ мнений при опросах в разных группах людей
- ⌚ SEO-ключи для формирования тематик страниц сайта
- ⌚ Группировка документов по тематике



# Обучение с подкреплением (Reinforcement Learning)

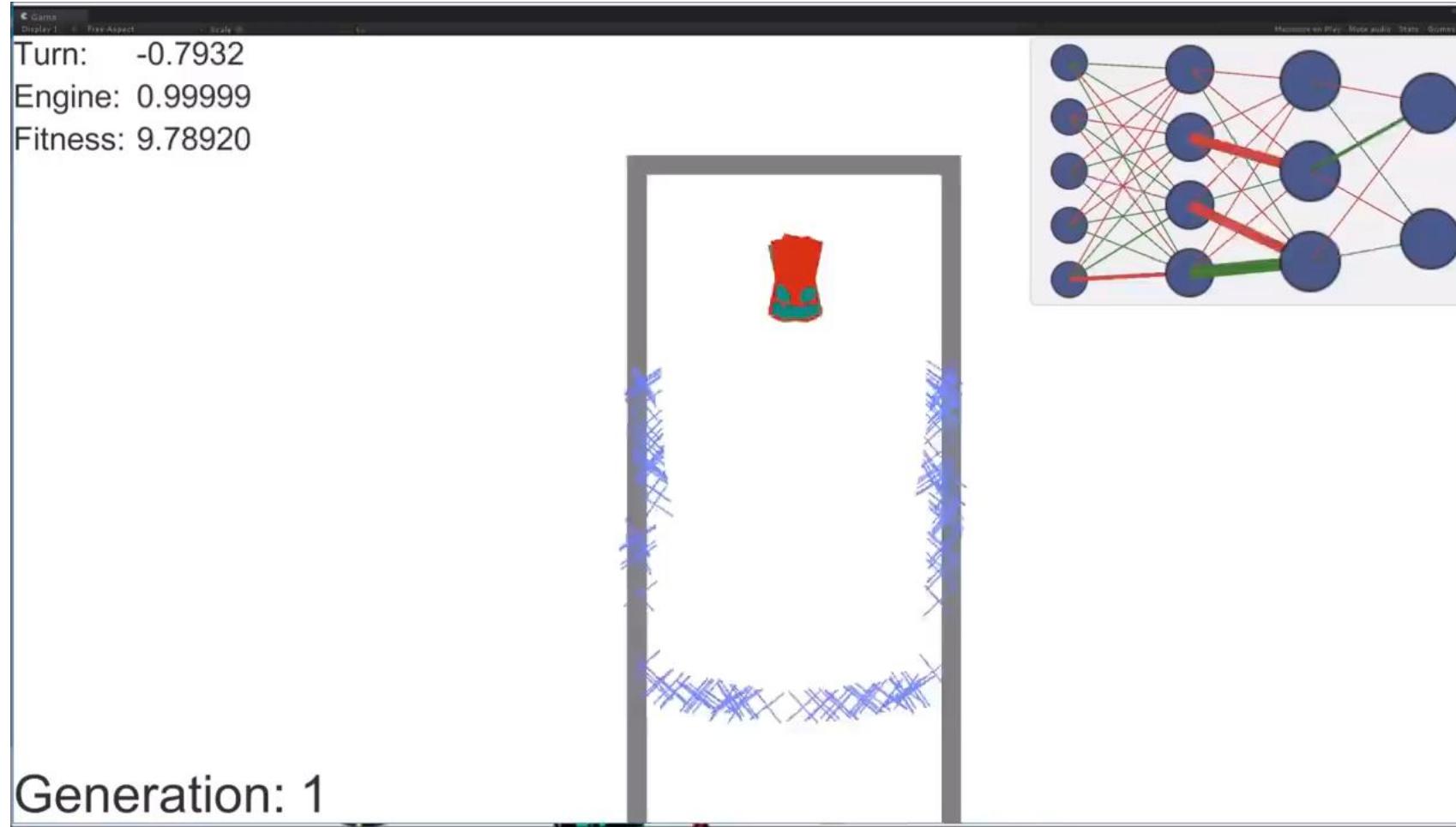




## Обучение с подкреплением (Reinforcement Learning)

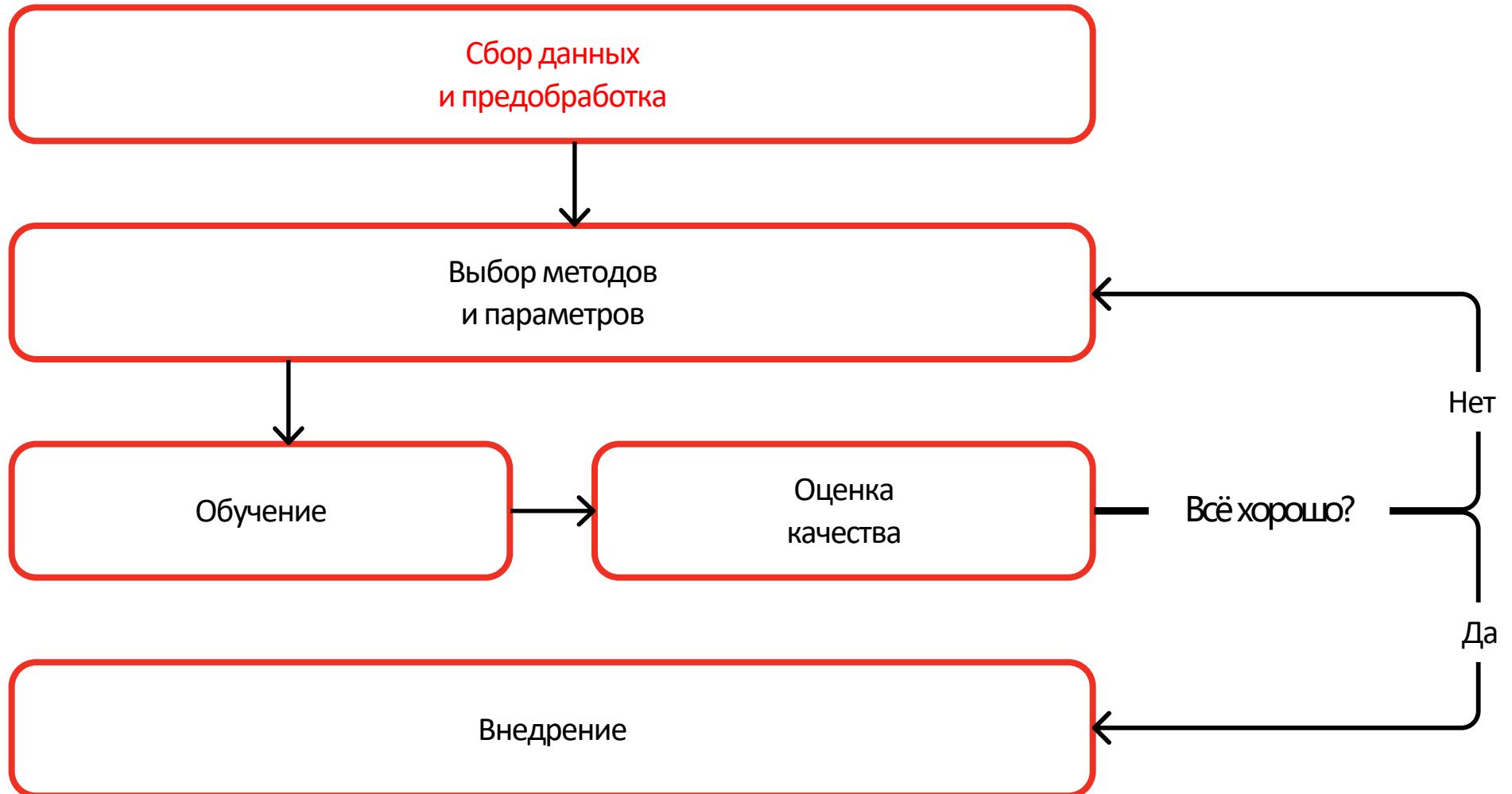


# Обучение с подкреплением (Reinforcement Learning)



Этапы работы алгоритмов машинного обучения

# Ход работы



# Данные и информация

Данные — это набор фиксированных сведений



Данные ≠ Информация



Обработка

**DIKW**  
data, information, knowledge, wisdom



# Данные и информация

## Данные

Название	Требования	Заработка плата
Программист	Знание C#, работа в Jira, ...	150 000 рублей
Библиотекарь	Высшее образование, ...	30 000 рублей
Дизайнер	Внимательность, Photoshop, ...	80 000 рублей
Учитель математики	Знание детской психологии, ...	20 000 рублей
Продавец	Умение работать с кассой, ...	40 000 рублей

## Информация

**Средняя заработная плата:**  
64 000 рублей

# Типы данных (по структуре)

## Структурированные данные

- Имеют четкую организацию, хранятся в таблицах с фиксированными полями
- Легко обрабатываются программно
- Хорошо подходят для аналитики

Пол	Возраст	Должность	Место работы
Женщины	33	Оператор	Банк
Мужчины	36	Преподаватель	Университет
Женщины	28	Тренер	Фитнес-зал
Мужчины	18	Разнорабочий	Маркетплейс

К примеру:

- База данных опроса
- Данные о клиентах

Формат хранения: csv, xlsx

# Типы данных (по структуре)

## Неструктурированные данные

- Нет четкой структуры
- Требуют сложной обработки
- Занимают много места



## К примеру:

- Мультимедиа (фото, видео, аудио)
- Тексты (электронные письма, сообщения в соцсетях, книги, статьи)
- Логи с серверов (неформатированный текст)

# Типы данных (по структуре)

## Полуструктурированные данные

- Имеют частичную организацию, но не требуют строгой схемы
- Гибкость(можно добавлять поля)
- Требуют парсинга, но проще, чем неструктурированные

```
{  
    "productId": 1,  
    "productName": "A green  
door",  
    "price": 12.50,  
    "tags": [ "home", "green" ]  
}
```

• **productId**: идентификатор продукта  
• **productName**: название продукта  
• **price**: затраты для потребителя  
• **tags**: необязательный массив идентифицирующих тегов

## К примеру:

- JSON (API-ответы, настройки приложений)
- XML (конфиги, веб-данные)Логи с серверов (неформатированный текст)
- NoSQL-базы (MongoDB, Elasticsearch)

# Основная терминология

## Датасет (dataset)

это набор данных (структурированных, неструктурированных или полуструктурированных), собранных и организованных для конкретной задачи (аналитики, машинного обучения и др.).



В качестве данных могут выступать:

- Успеваемость студентов
- Отзывы на товары и организации
- Посты в группах в социальных сетях
- Научные статьи и их метаданные
- Заболевания с указанием симптомов

## Столбцы(поля) –

переменные, признаки, характеристики объекта

Возврат кредита в срок	Возраст	Оценка кредитного потенциала
Да	40	1
Нет	50	4
Да	60	3
Да	48	2
Нет	65	5

## Строки(записи) –

наблюдения, объекты



Вид датасета - простая запись

Значение признаков соответствуют конкретным объектам. Например, данные о клиентах

# Как выглядят реальные данные?

Пример датасета – посты социальной сети ВКонтакте

group_id	region_title	id	from_id	owner_id	date
11726883	Свердловская о	7045	-11726883	-11726883	1680144696
11726883	Свердловская о	6998	-11726883	-11726883	1677406916
11726883	Свердловская о	6985	-11726883	-11726883	1677222002
11726883	Свердловская о	6973	-11726883	-11726883	1677071880
1,65E+08	Томская област	651	-164704678	-164704678	1677317400
1,65E+08	Томская област	649	-164704678	-164704678	1677313607
1,65E+08	Томская област	639	-164704678	-164704678	1673957465
1,65E+08	Томская област	647	-164704678	-164704678	1675996213
15511351	Белгородская о	6973	-15511351	-15511351	1685531640
15511351	Белгородская о	6731	-15511351	-15511351	1675160280
1,29E+08	Вологодская об	4705	-129154265	-129154265	1676527922
1,77E+08	Башкортостан	17028	-176863991	-176863991	1681748760
93052045	Белгородская о	56478	-93052045	-93052045	1685698565

comments	likes	reposts	views
0	2	0	748
0	1	0	328
0	0	0	250
0	0	1	459
0	2	0	15
0	2	0	15
0	2	0	10
0	2	0	18
0	0	0	54
0	0	0	177
0	4	6	632
0	5	1	1075
2	18	24	589

# Как выглядят реальные данные?

 schools

followe...	relatives	relation
8		
3		
130		
11		
2		
109		
67	[{"id": 538447394, "type": "parent"}]	{"relation": "4"}
9		
1		
2		
3	[{"id": 579783679, "type": "sibling"}]	
1		
86		{"relation": "2", "relation_pa
634		
114		
47	[{"id": 509365783, "type": "sibling"}]	{"relation": "5"}
66	[{"id": 519197715, "type": "child"}, {"id": -463151722, "name": "Hatar"}]	{"relation": "7", "relation_pa
272		
4		

# Какие данные нам нужны?

Определите **цель** и **гипотезы**

Выберите **тип данных** (таблицы, текст,

изображения) Найдите **источники**

(внутренние/внешние)

Выделите **целевую переменную и признаки**

Проверьте **качество и достаточность** данных

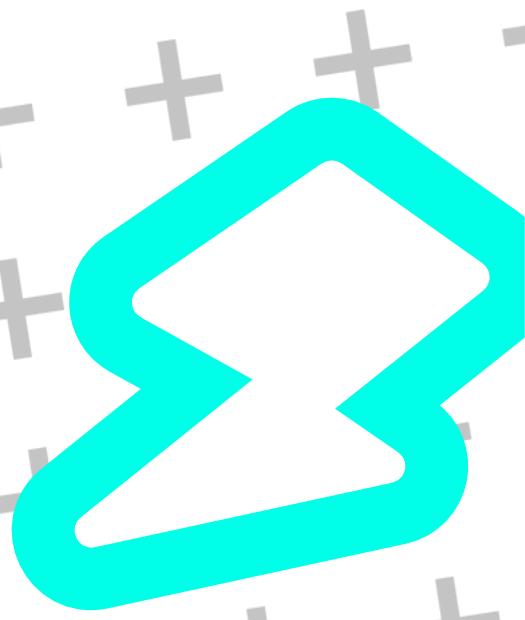
Предобработайте (очистка, трансформация)

Основа любого  
аналитического проекта  
**- ДАННЫЕ**



# Классификация данных по источникам

- **Собственные** — данные, которые у нас уже есть
- **Сторонние** — данные, которые мы можем взять из других источников
- **«Потенциальные»** — необходимо организовать сбор этих данных



# Способы получения данных

- Сбор/получение первичной исходной информации
- Получение данных из вторичных (сторонних) источников → [Дата-брокеры](#)
- Получение данных с использованием API
- Парсинг — автоматизированный сбор информации с сайта, ее анализ, преобразование и выдача в структурированном виде
- Ручной способ формирования датасета

**API** – интерфейс, который предоставляет определенный набор методов, позволяющих собирать нужные данные из базы путем отправки http-запросов к специальному серверу.



- Пользователи (профили, подписки)
- Сообщества (подписчики, стены, профили сообществ)

**VK API:**  
<https://dev.vk.com/reference>



- Публикации
- Комментарии и реакции к публикациям
- Описание Telegram-каналов

**Telegram API:**  
<https://core.telegram.org/methods>

**Парсинг** — это автоматизированный сбор информации с сайта, ее анализ, преобразование и выдача в структурированном виде, чаще всего в виде таблицы с набором данных.

### Если:

- не нарушает законы об интеллектуальной собственности,
- не взламывает защитные системы,
- не извлекает персональные данные пользователя,
- не мешает работе сайта, который подвергается парсингу,
- не нарушает условия использования сайта

# Ручной способ формирования датасета

Когда необходим так способ сбора данных?

- У вас есть специфическая задача (готовых данных нет)
- Неопределенное количество источников
- Источники отличаются по структуре и названиям признаков

- + Вы сами задаете признаковое пространство (контент)
- + Валидация данных на этапе сбора
- + Минимальные технические навыки
- Временные затраты
- Сложности объединение данных из разных источников

# Популярные источники данных

- Блоги и форумы
- Веб-аналитика  
(данные о посещаемости сайтов,  
поведение пользователей)
- Вакансии  
(сервис РосНавык)
- Социальные сети  
(Вконтакте, Telegram)
- Контент тематических сайтов
- Интернет-СМИ
- CRM- системы
- Всевозможные базы данных  
(государственная  
статистика)
- Википедия
- Поисковые запросы  
(Yandex)
- Собственные данные  
организаций
- Научные публикации  
(наукометрическая база  
OpenAlex)

# Открытые данные

## Открытые данные (англ. open data)

— концепция о том, что определенные данные должны быть свободно доступны для машиночитаемого использования и дальнейшей републикации без ограничений авторского права, патентов и других механизмов контроля

[https://ru.wikipedia.org/wiki/Открытые\\_данные](https://ru.wikipedia.org/wiki/Открытые_данные)



### При использовании открытых данных пользователь обязан

- использовать открытые данные только в законных целях
- не искажать открытые данные при их использовании
- сохранять ссылку на источник информации при использовании открытых данных



### Кто публикует открытые данные?

- государственные органы
- научный сектор (исследователи)
- компании

## Примеры источников открытых данных

### 1. ЕГРЮЛ

### 2. Порталы открытых данных

- dateno.io/about
- hubofdata.ru/dataset
- datasets-isc.ru
- datacatalogs.ru
- ruarhive.org/kb/intro
- ngodata.ru/
- data.gov.kg/
- tochno.st/

### 4. Контракты государственных закупок

- (zakupki.gov.ru)

### 5. Данные государственных сайтов

- minobrnauki.gov.ru/opendata
- cikrf.ru/opendata
- trudvsem.ru/opendata

### 6. Научные данные

- (статьи, заявки на гранты и т.д.)
  - ruscorpora.ru

### 7. Общественные открытые данные

- (Kaggle, Huggingface, Github, Mendeley)

# Открытые данные

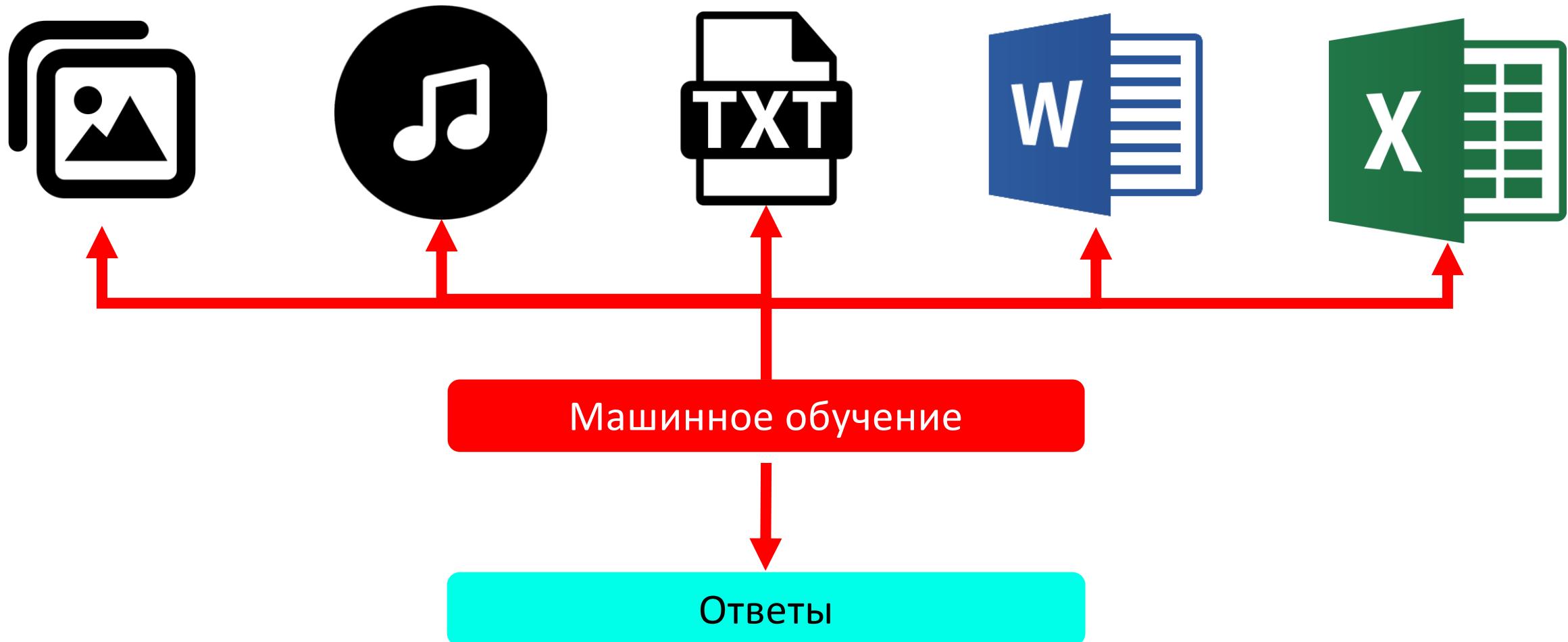
Росстат <https://rosstat.gov.ru/>

The screenshot shows the official website of the Federal Statistical Service of Russia (Rosstat). At the top, there is a navigation bar with links to 'О Росстате' (About Rosstat), 'Статистика' (Statistics), 'Публикации' (Publications), 'Респондентам' (For Respondents), 'Пресс-служба' (Press Office), and 'Контакты' (Contacts). Below the navigation bar, there are four main categories: 'Официальная статистика' (Official statistics), 'Переписи населения' (Censuses), 'BI-система' (BI system), 'Новости статистики' (Statistics news); 'Методология и нормативно-справочная информация' (Methodology and normative reference information), 'Сельскохозяйственные переписи' (Agricultural censuses), 'Инфографика' (Infographics), 'Анонсы' (Announcements); and 'Статистические обследования' (Statistical surveys), 'Понятная статистика' (Accessible statistics). A large banner in the center features an illustration of two people in a cityscape and the text 'Росстат осуществил вторую оценку ВВП за 2024 год' (Rosstat conducted the second estimate of GDP for 2024). Below the banner, there is a 'Новости Росстата' (Rosstat news) section with several news items. To the right, there are three blue buttons for 'Календарь публикаций' (Publication calendar), 'Календарь отчетности' (Report calendar), and 'BI-система'. At the bottom, there is a 'Каталог' (Catalog) section with a search bar and a button to 'Зарегистрироваться' (Register).

НИИД <https://data.rcsi.science/>

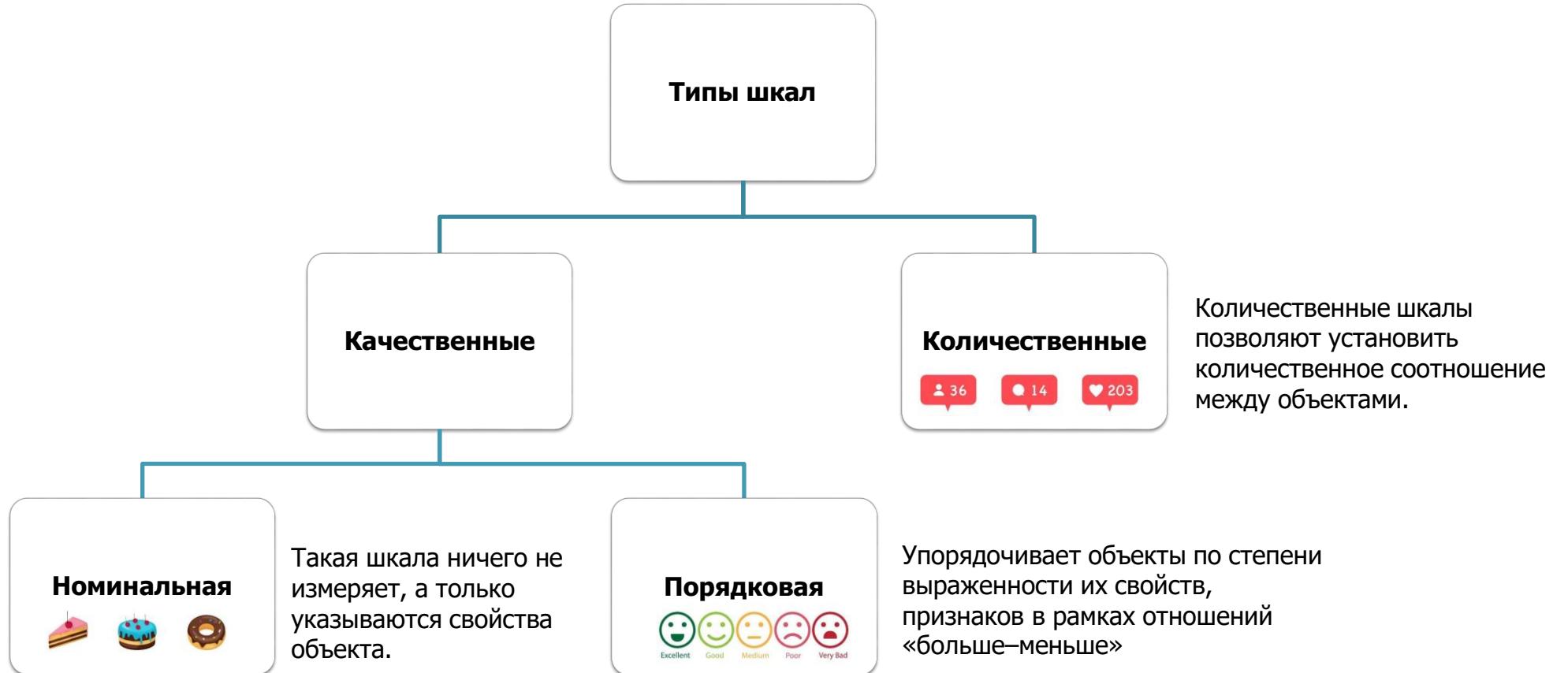
The screenshot shows the catalog page of the INID (Institute for the Study of Scientific Data) website. On the left, there is a sidebar with 'Категории' (Categories) including 'Государственные финансы', 'Образование и наука', 'Доходы и неравенство', 'Здравоохранение', 'Экономика', and 'Статистика'. Below that is a 'Фильтр' (Filter) section. The main content area shows a dataset titled 'Реестр субъектов малого и среднего предпринимательства за 2016–2024 гг. с доходами, расходами, численностью работников и географическими координатами' (Register of small and medium-sized enterprises for 2016–2024 with income, expenses, number of employees and geographical coordinates). The dataset is described as being available to all users and was updated on 06.11.2024. There are filters at the bottom for 'По дате обновления' (By update date), 'По скачиваниям' (By downloads), 'По просмотрам' (By views), and 'Показывать по:' (Show by) with a value of 10.

## Типы данных



## Данные и виды признаков

# Как измеряются признаки?



## Обучающая выборка

Площадь квартиры ( $x_1$ )	Этаж квартиры ( $x_2$ )	Площадь кухни ( $x_3$ )	Количество комнат ( $x_4$ )	Стоимость квартиры ( $\mathbb{Y}$ )
460	2	15	6	195
230	7	9	4	130
315	1	20	3	140
178	3	25	4	80
...	...	...	...	....

## Обучающая выборка

Площадь квартиры ( $x_1$ )	Этаж квартиры ( $x_2$ )	Площадь кухни ( $x_3$ )	Количество комнат ( $x_4$ )	Стоимость квартиры ( $\mathbb{Y}$ )
460	2	15	6	195
230	7	9	4	130
315	1	20	3	140
178	3	25	4	80
...	...	...	...	....

$x$  – объект

## Обучающая выборка

Площадь квартиры ( $x_1$ )	Этаж квартиры ( $x_2$ )	Площадь кухни ( $x_3$ )	Количество комнат ( $x_4$ )	Стоимость квартиры ( $\mathbb{Y}$ )
460	2	15	6	195
230	7	9	4	130
315	1	20	3	140
178	3	25	4	80
...	...	...	...	....

$x$  – объект

$\mathbb{X}$  – пространство объектов

# Обучающая выборка

Площадь квартиры ( $x_1$ )	Этаж квартиры ( $x_2$ )	Площадь кухни ( $x_3$ )	Количество комнат ( $x_4$ )	Стоимость квартиры ( $\mathbb{Y}$ )
460	2	15	6	195
230	7	9	4	130
315	1	20	3	140
178	3	25	4	80
...	...	...	...	....

$x$  – объект

$\mathbb{X}$  – пространство объектов

$y$  – целевая переменная

# Обучающая выборка

Площадь квартиры ( $x_1$ )	Этаж квартиры ( $x_2$ )	Площадь кухни ( $x_3$ )	Количество комнат ( $x_4$ )	Стоимость квартиры ( $\mathbb{Y}$ )
460	2	15	6	195
230	7	9	4	130
315	1	20	3	140
178	3	25	4	80
...	...	...	...	....

$x$  – объект

$\mathbb{X}$  – пространство объектов

$y$  – целевая переменная

$\mathbb{Y}$  – пространство ответов

## Обучающая выборка

Площадь квартиры ( $x_1$ )	Этаж квартиры ( $x_2$ )	Площадь кухни ( $x_3$ )	Количество комнат ( $x_4$ )	Стоимость квартиры ( $\mathbb{Y}$ )
460	2	15	6	195
230	7	9	4	130
315	1	20	3	140
178	3	25	4	80
...	...	...	...	....

$x$  – объект

$\mathbb{X}$  – пространство объектов

$y$  – целевая переменная

$\mathbb{Y}$  – пространство ответов

$(\mathbb{X}, \mathbb{Y})$  – Обучающая выборка

## Обучающая выборка

Площадь квартиры ( $x_1$ )	Этаж квартиры ( $x_2$ )	Площадь кухни ( $x_3$ )	Количество комнат ( $x_4$ )	Стоимость квартиры ( $\mathbb{Y}$ )
460	2	15	6	195
230	7	9	4	130
315	1	20	3	140
178	3	25	4	80
...	...	...	...	....

$x = (x_1, x_2, \dots, x_d)$  характеристики объектов, каждый объект характеризуется набором признаков

$d$  – число признаков

$\ell$  – число объектов или выборки

# Типы признаков объекта

## Вещественные

$Y \in \{0,1\}$  бинарный признак

$Y \in \mathbb{R}$  количественный признак

Число лайков от пользователей

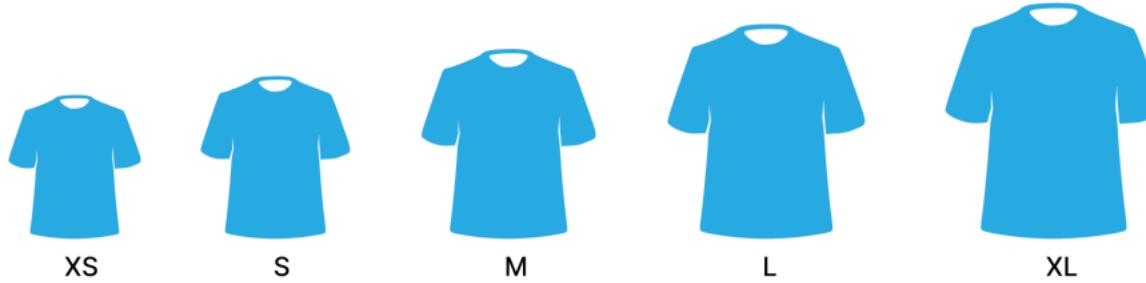


Температура



## Типы признаков объекта (Категориальные $\in \{c_1, c_2, \dots, c_n\}$ )

- Упорядоченные (ординальные) – для каждой пары возможных категорий можем сказать, какая больше, а какая меньше. Например, размер одежды



## Типы признаков объекта (Категориальные)

- Неупорядоченные (номинальные) – категории между собой несравнимы.



Солнечно



Облачно



Снежно



Дождливо



Ветрено



Морозно

## Матрица объектов-признаков

- Матрица признаков  $\mathbb{X}$  это матрица размером  $\ell \times d$ , в которой каждый элемент  $x_{ij}$  представляет значение  $j$  – го признака для  $i$  – го объекта

$$\mathbb{X} = \begin{pmatrix} x_{11} & \cdots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{\ell 1} & \cdots & x_{\ell d} \end{pmatrix}$$

Матрица целевой переменной  $\mathbb{Y}$  это столбец размером  $\ell \times 1$ , в котором каждый элемент  $y_i$  представляет целевое значение для  $i$  – го объекта

$$\mathbb{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_\ell \end{pmatrix}$$

# Задачи медицинской диагностики

Объект - пациент

Классы : {-1,1} где

$$\begin{cases} -1 \text{ у пациента нет рака} \\ 1 \text{ у пациента есть рак} \end{cases}$$

**Примеры признаков:**

Бинарные: пол, головная боль, слабость, тошнота и т. д.

Упорядоченные: тяжесть состояния и т. д.

Количественные: возраст, пульс, артериальное давление и т. д.

# Задачи прогнозирования стоимости недвижимости

Объект – квартира

Числовое значение: стоимость квартиры

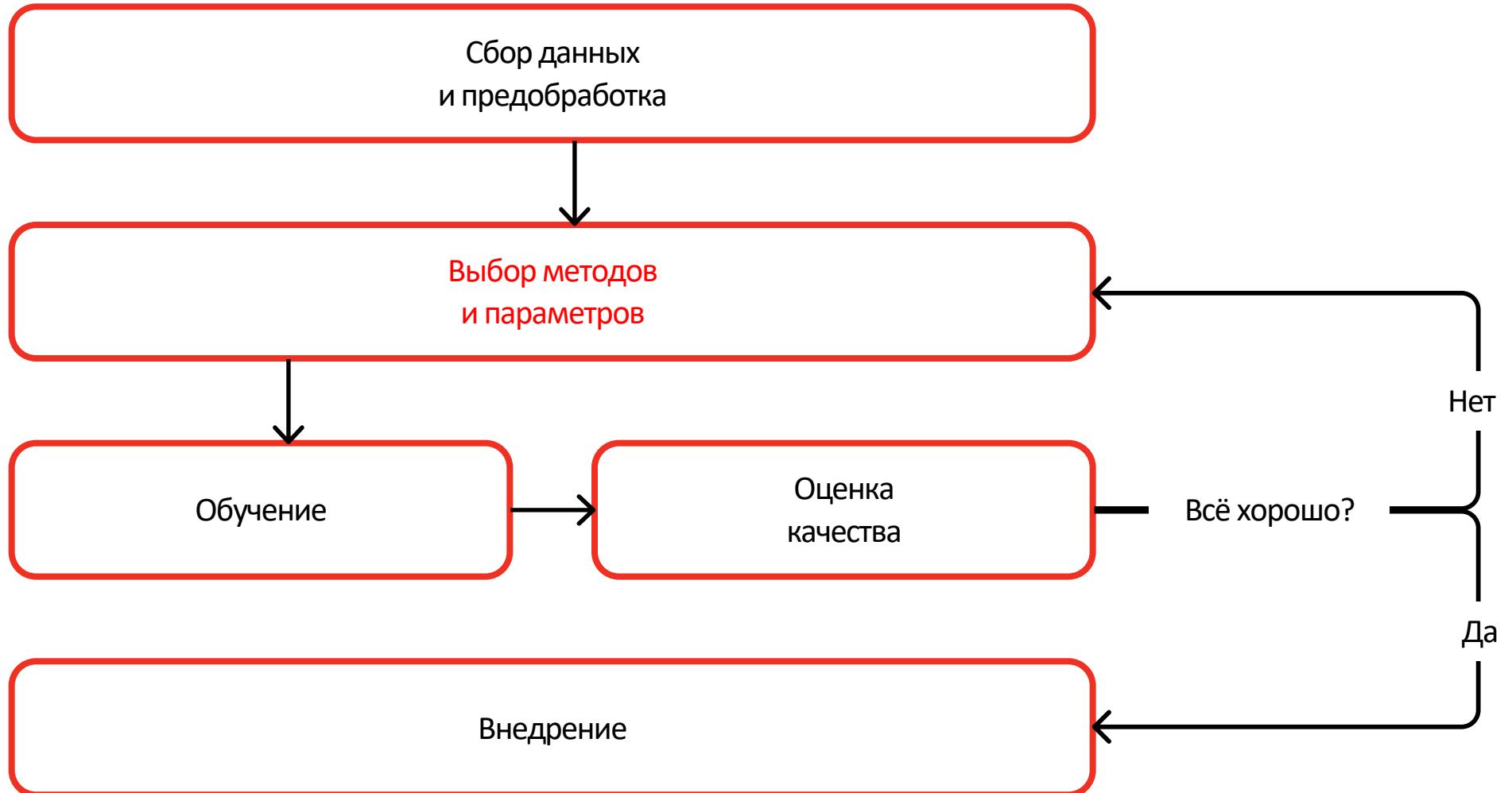
## **Примеры признаков:**

Бинарные: наличие балкона, лифта, охраны и т. д.

Номинальные: район города, тип дома и т. д.

Количественные: жилая площадь, число комнат, расстояние до метро, возраст дома и т. д.

# Ход работы

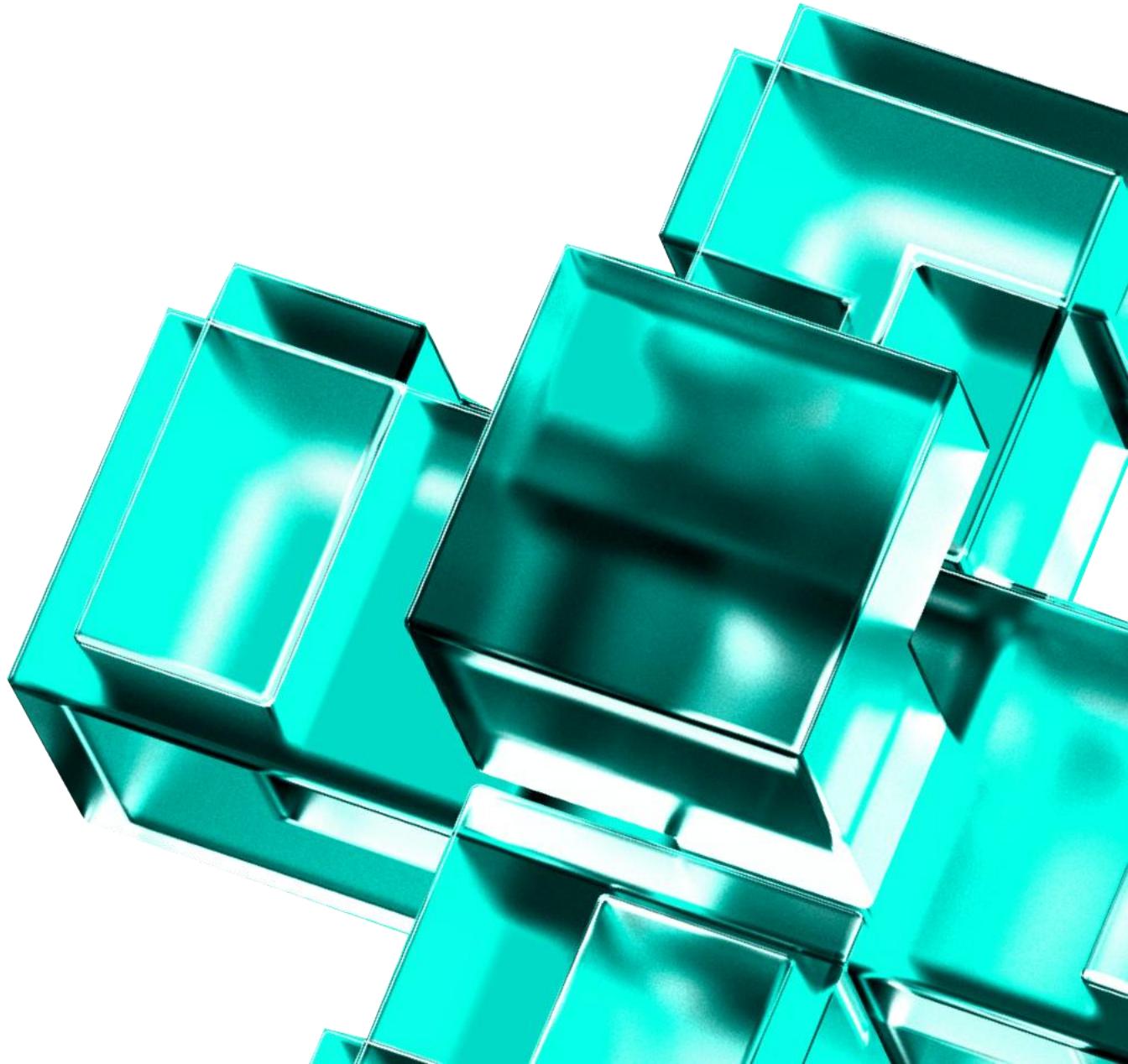


# Какой алгоритм выбрать?



## «Базовый» список моделей машинного обучения

- Наивный байесовский классификатор
- Дерево решений
- Метод опорных векторов
- Регрессия
- Логистическая регрессия
- Принцип главных компонент
- Другие

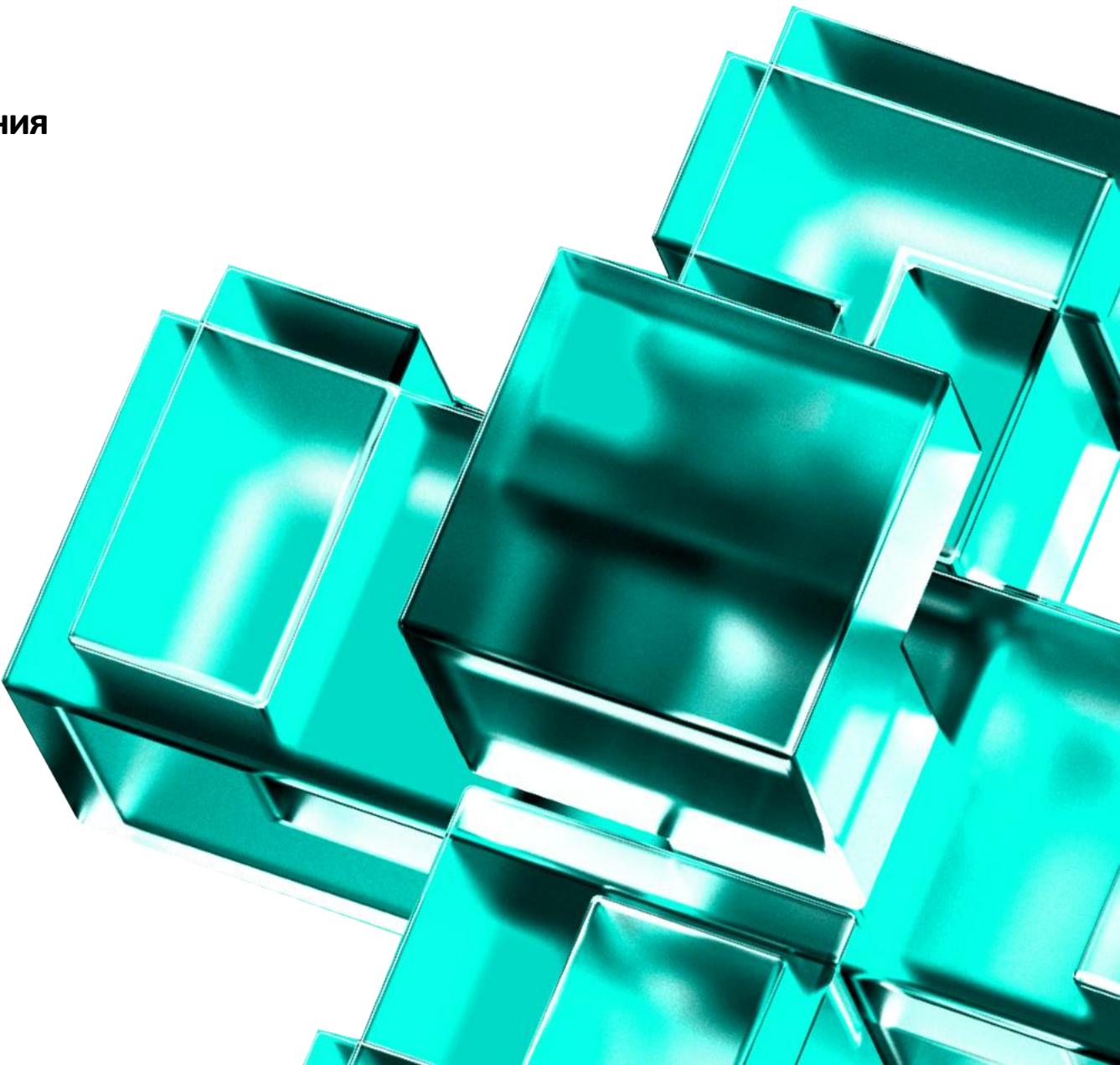


# Какой алгоритм выбрать?

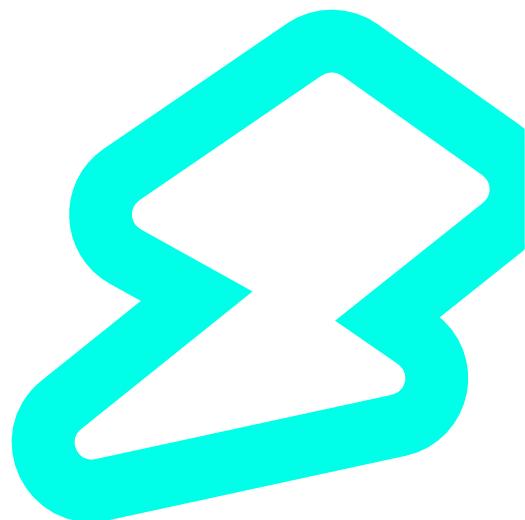
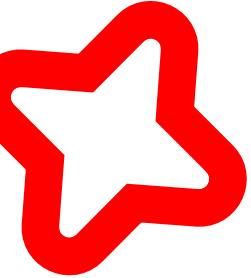
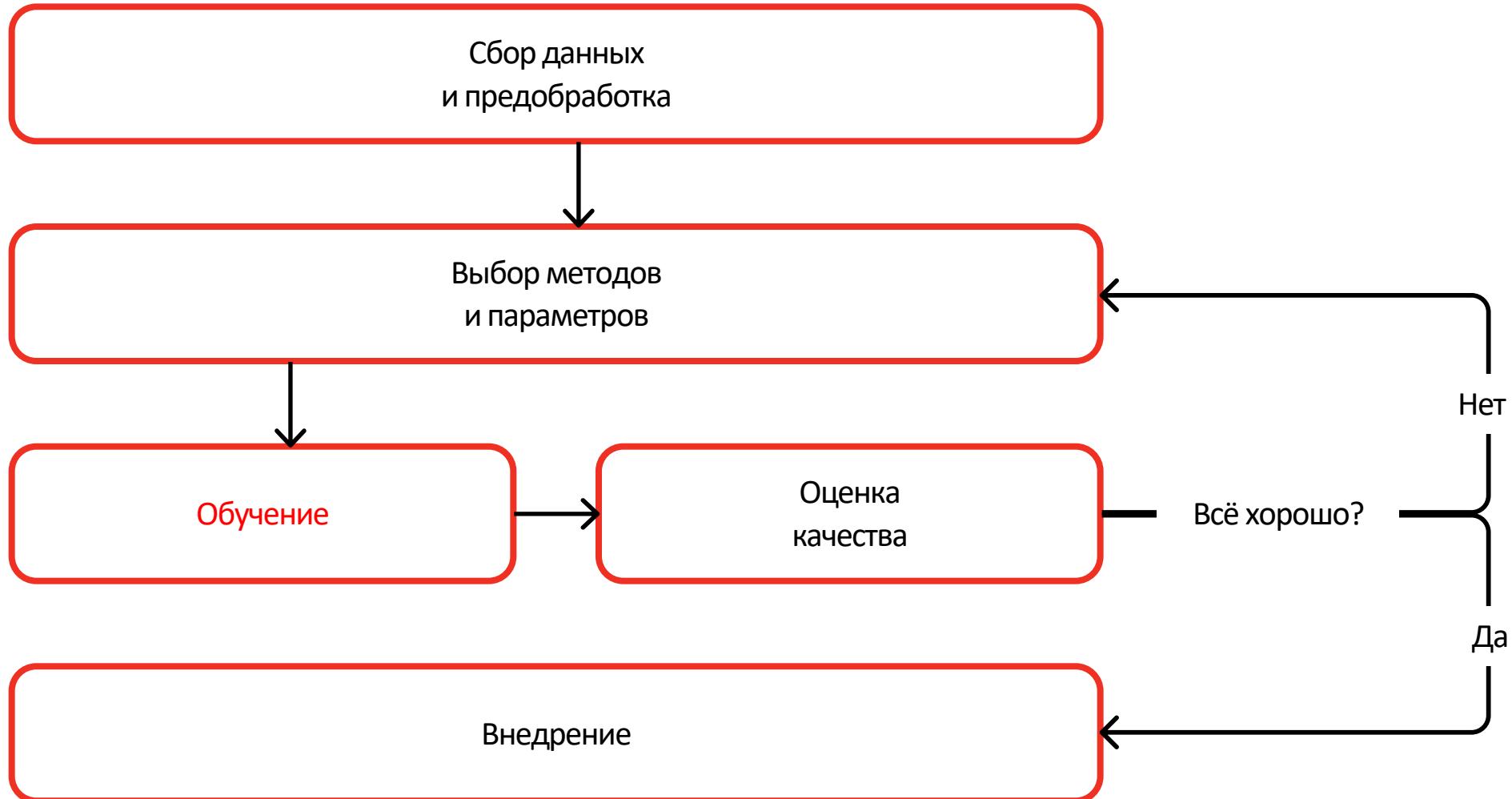
3

## «Расширенный» список моделей машинного обучения

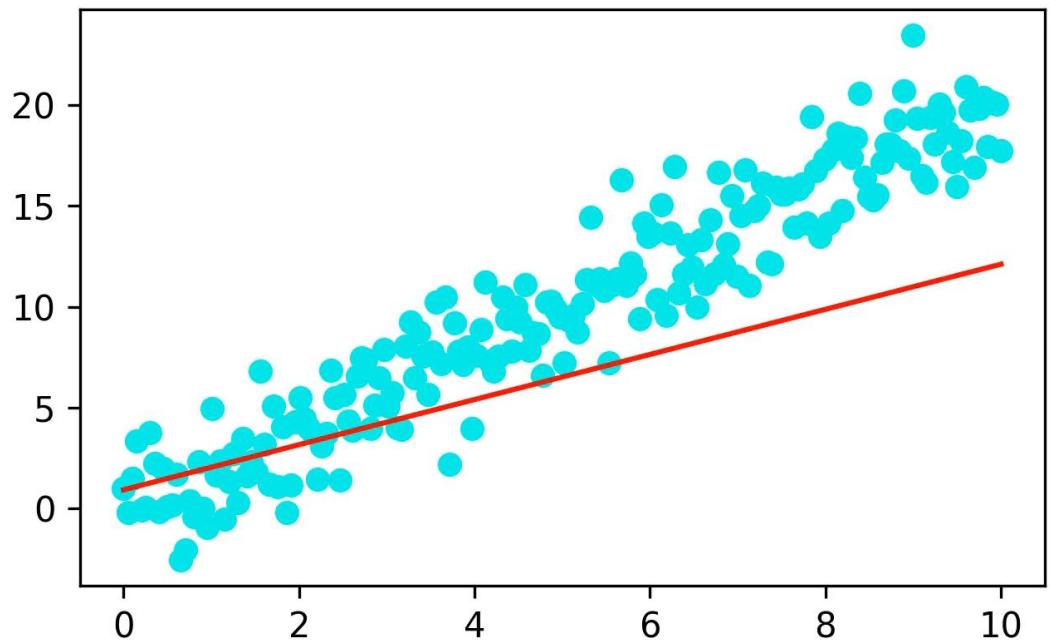
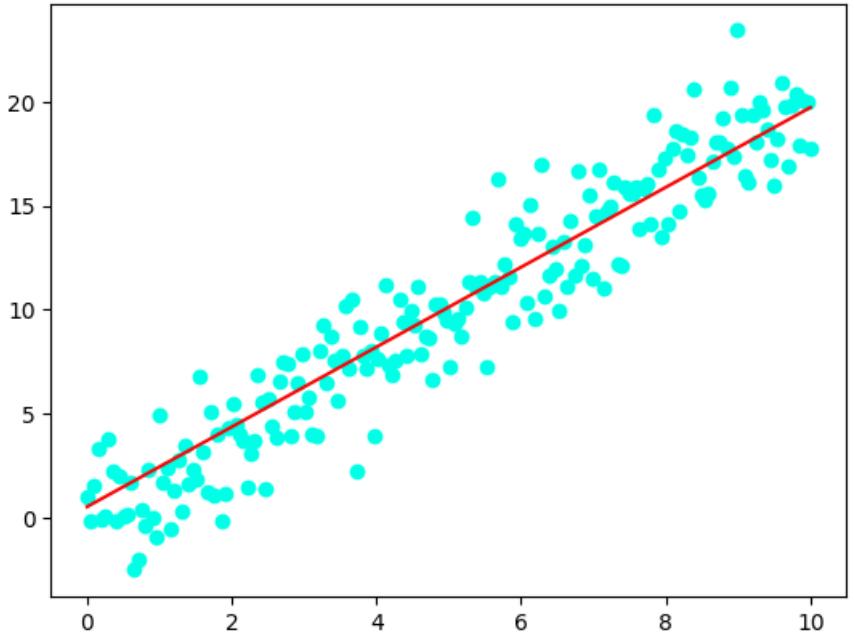
- Нейронная сеть
- Свёрточная нейронная сеть
- К-средних (kNN)
- Случайный лес
- Бустинг над деревьями решений
- Ансамбль моделей
- Другие



# Ход работы

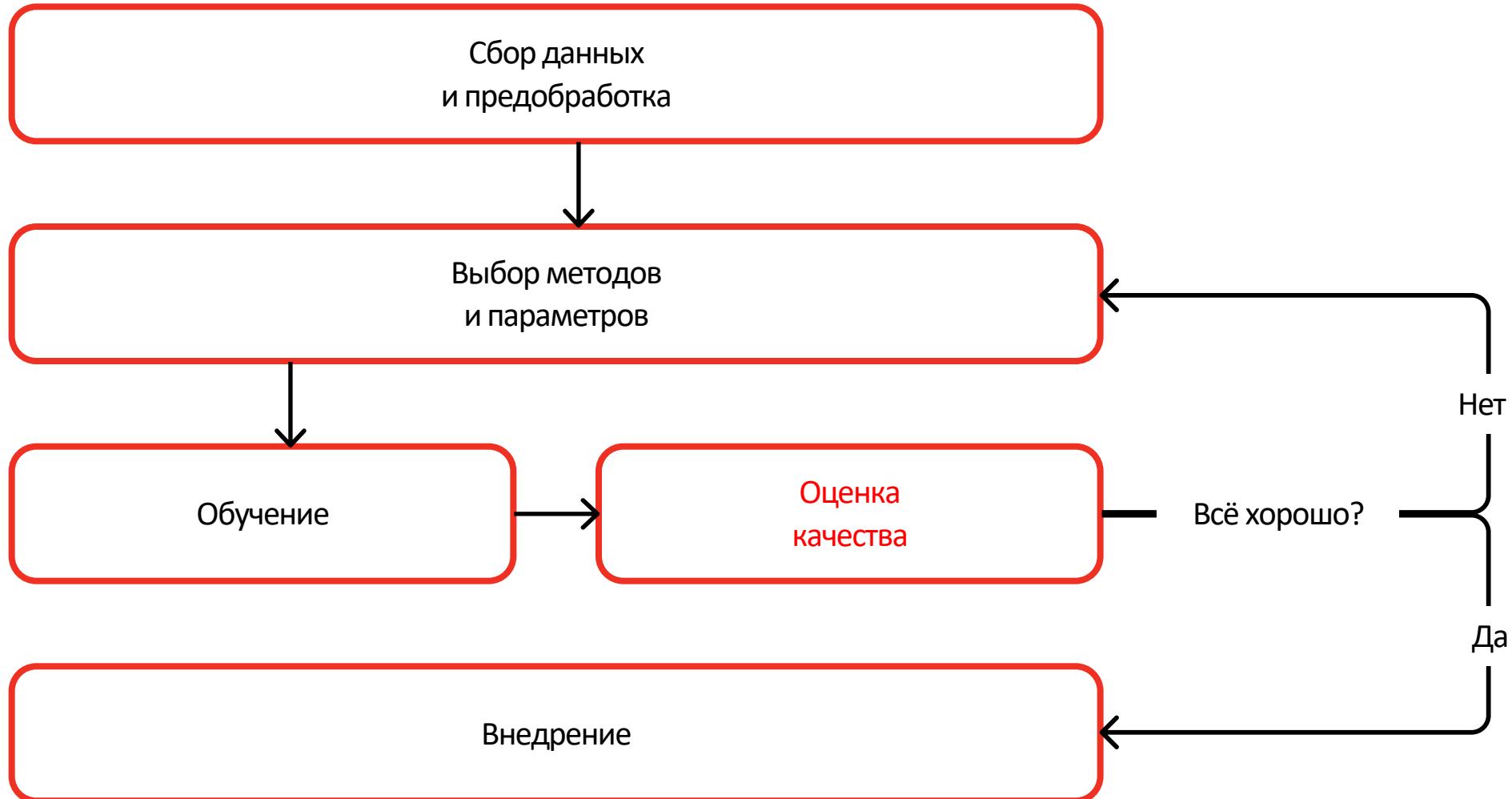


# Обучение



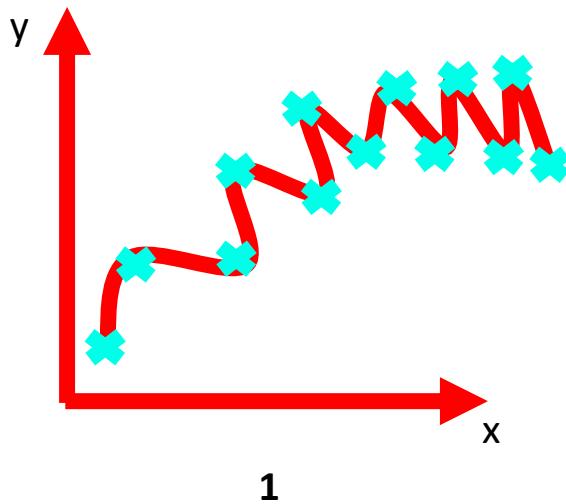
Обучение

# Ход работы

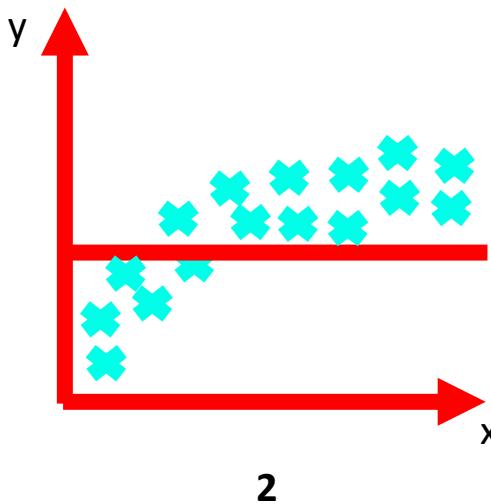


## Интерактивное задание

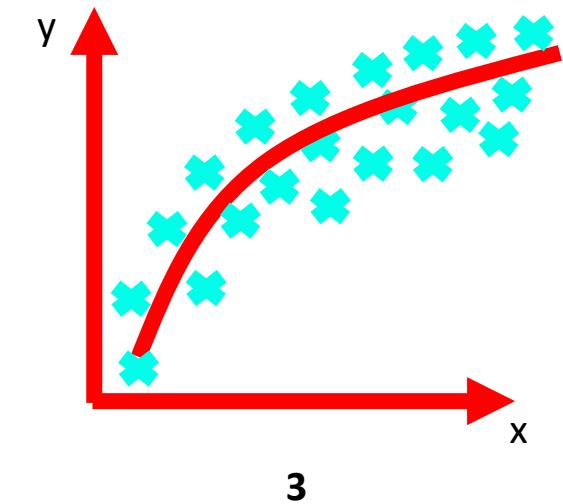
Какой график или модель вы считаете самыми лучшими?



1

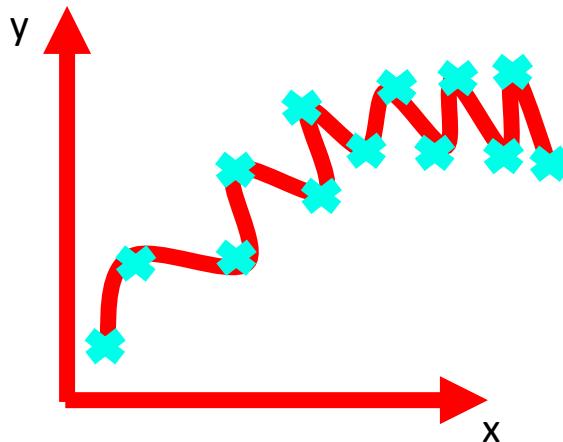


2

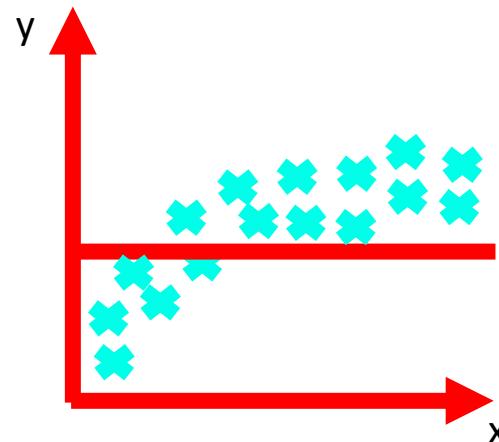


3

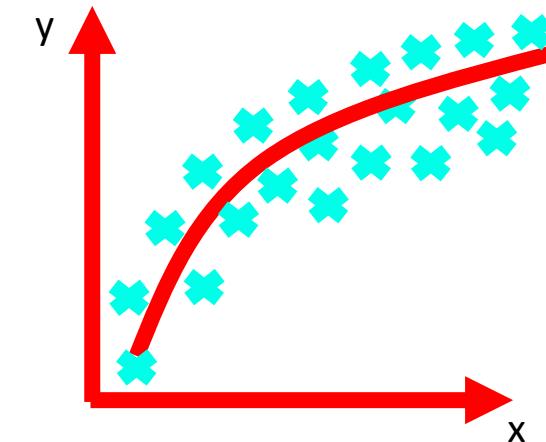
## Оценка качества



Overfitting (High Variance)

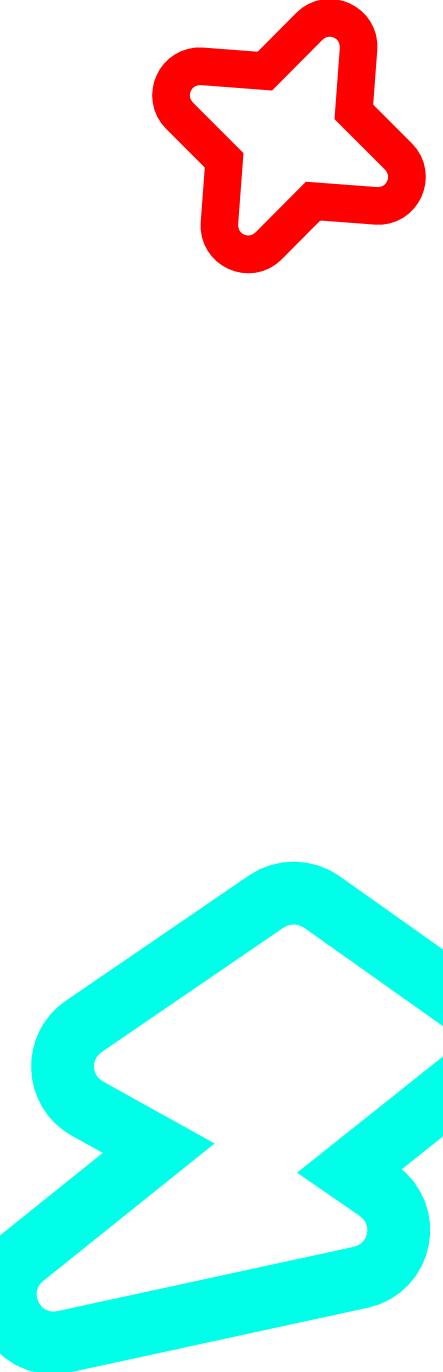
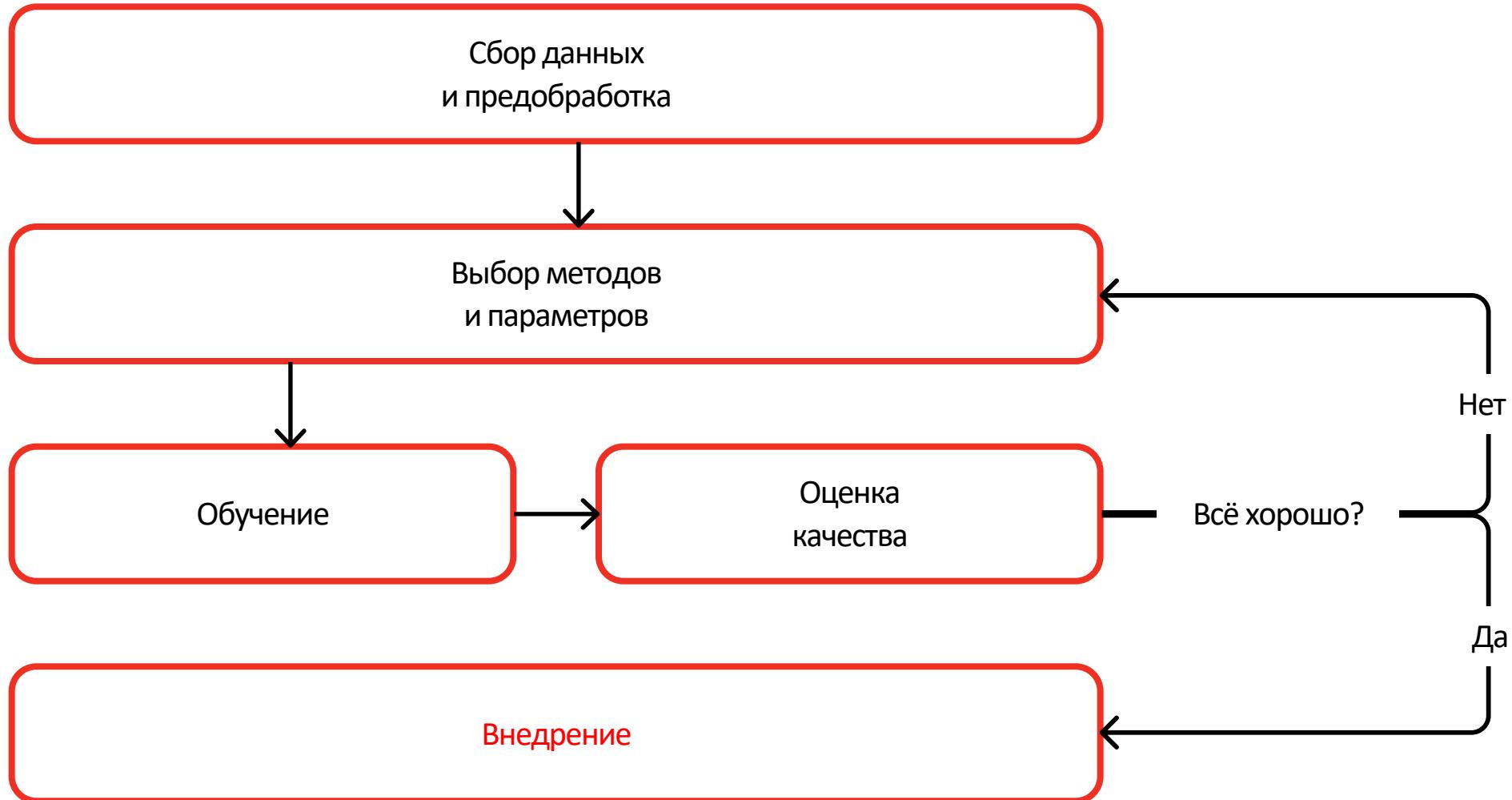


Underfitting (High Bias)



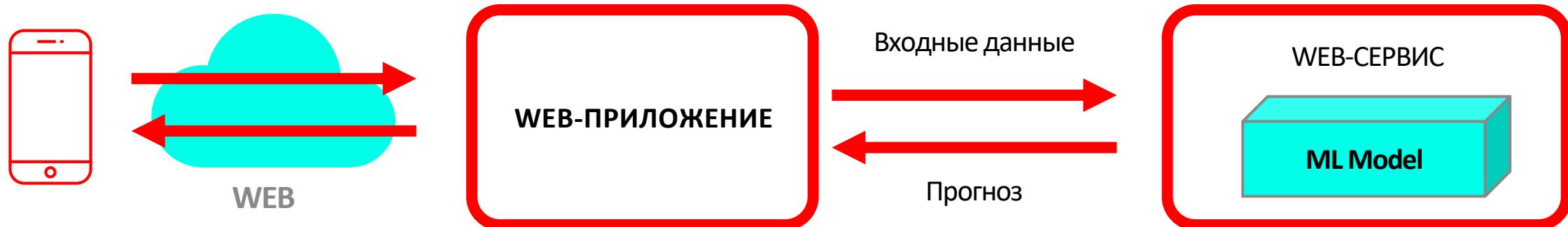
Just Right

# Ход работы

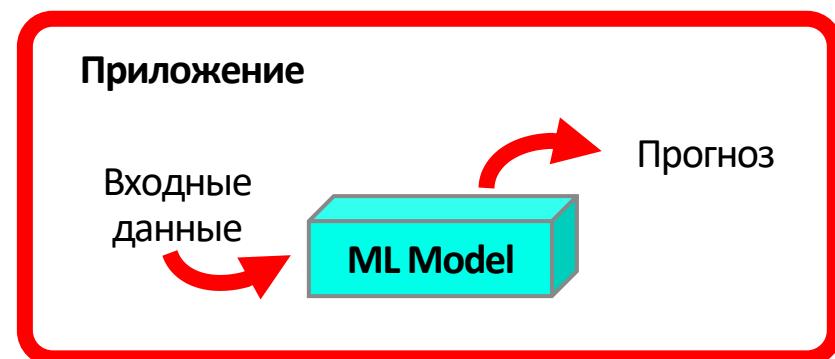


# Внедрение

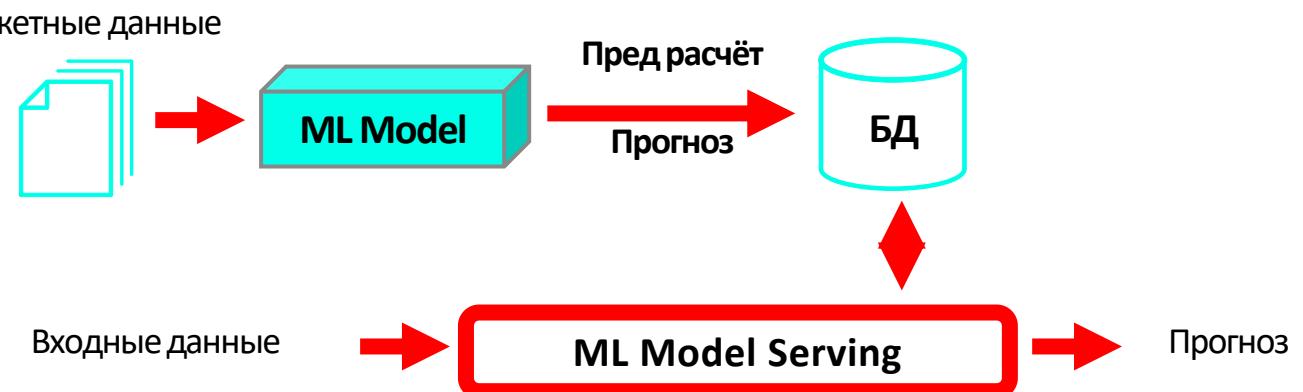
## ⚡ Модель как услуга



## ⚡ Модель как зависимость



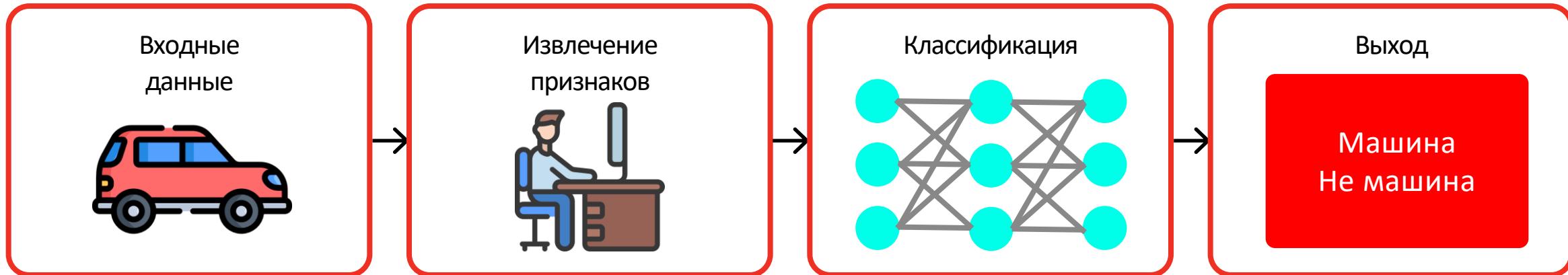
## ⚡ Предварительный расчёт



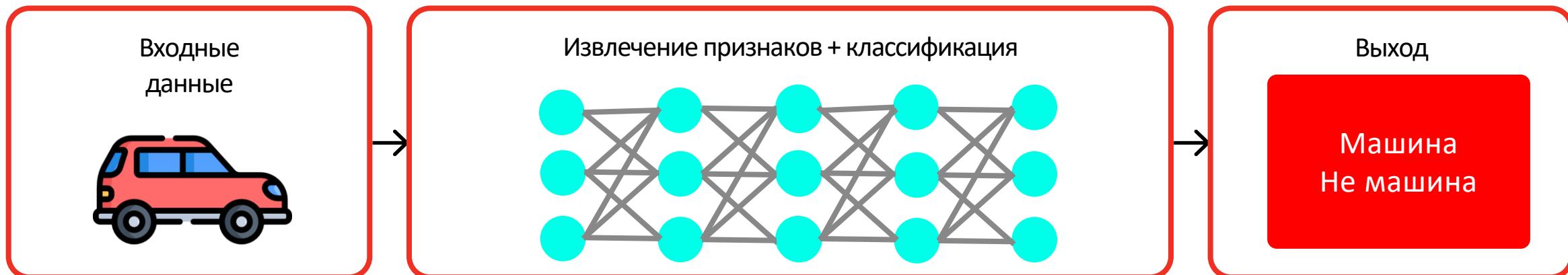
## Feature extraction

### Извлечение признаков

#### Машинное обучение



#### Глубокое обучение

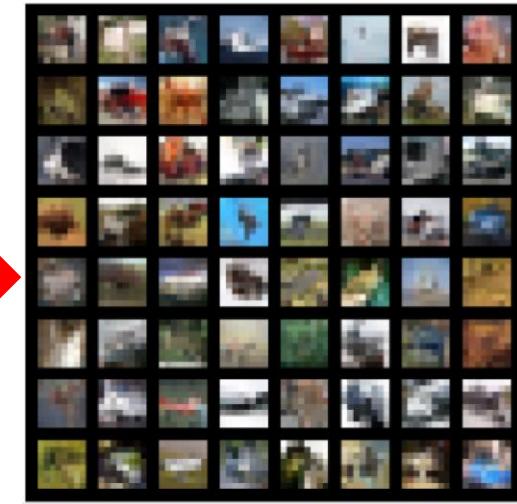
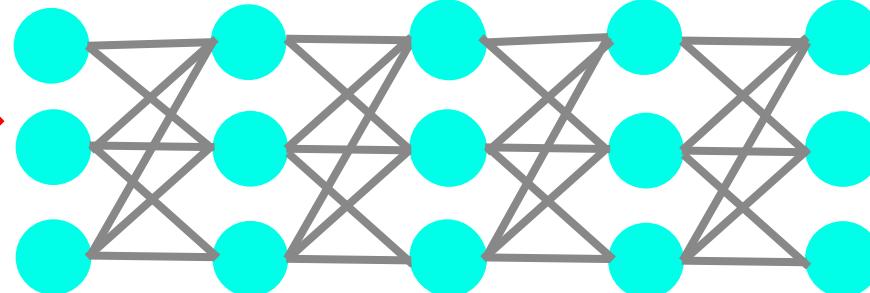


# Feature extraction

## Извлечение признаков



**Глубокое обучение**

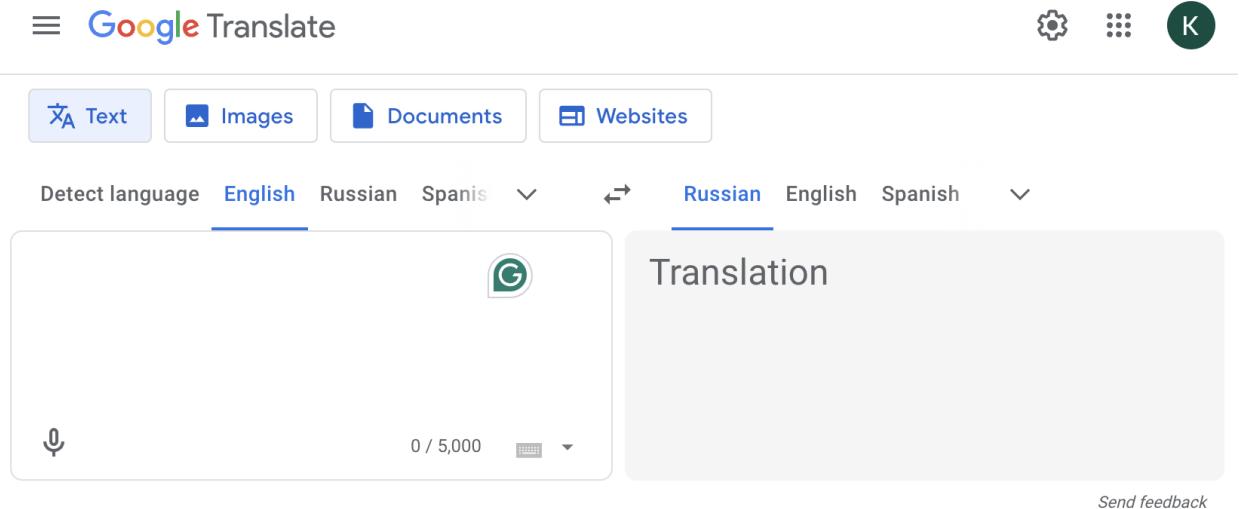
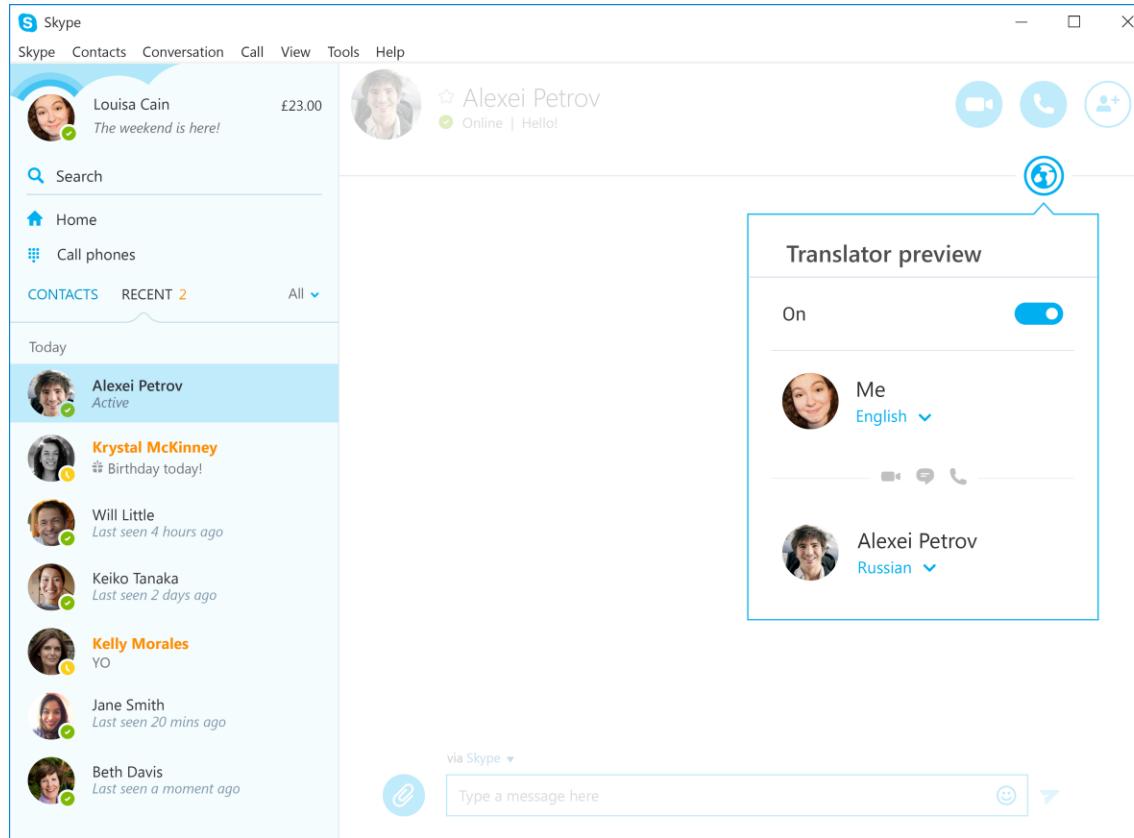


# Применение глубоких нейронных сетей Google's AlphaGo



# Применение глубоких нейронных сетей

## Skype Translator, переводчик Google



# Применение глубоких нейронных сетей

## Беспилотные автомобили и не только



Сингапур



Питтсбург, США



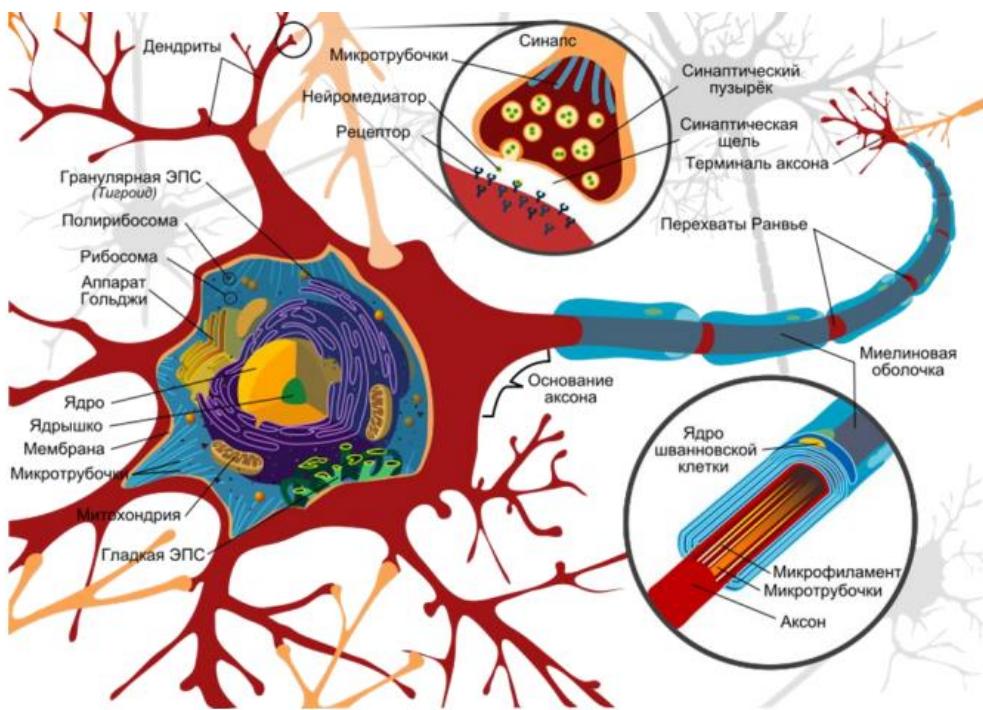
Лондон, Великобритания



Москва,  
Россия

# Нейрон головного мозга

## Нейрон головного мозга



## Искусственный нейрон

