



ТИНЬКОФФ

Введение в МЛ



О машинном обучении

Человек постоянно пытается упростить себе жизнь

- ▶ Для этого надо перекладывать часть работы на машин



С анализом данных то же самое



Огромный массив данных
машина обработает лучше
человека



Человек нужен, чтобы делать
выводы либо принимать
сложные решения



Для остального можно
приспособить компьютеры

О машинном обучении

- Последние 10 лет все только и говорят что о машинном обучении
- Часто ещё упоминают некий мифический AI (ИИ, искусственный интеллект)
- В чём отличие ML от AI?



О машинном обучении

- ML в сущности является частью AI-методов
- Существует море способов принятия решений на данных



Кто есть в ML



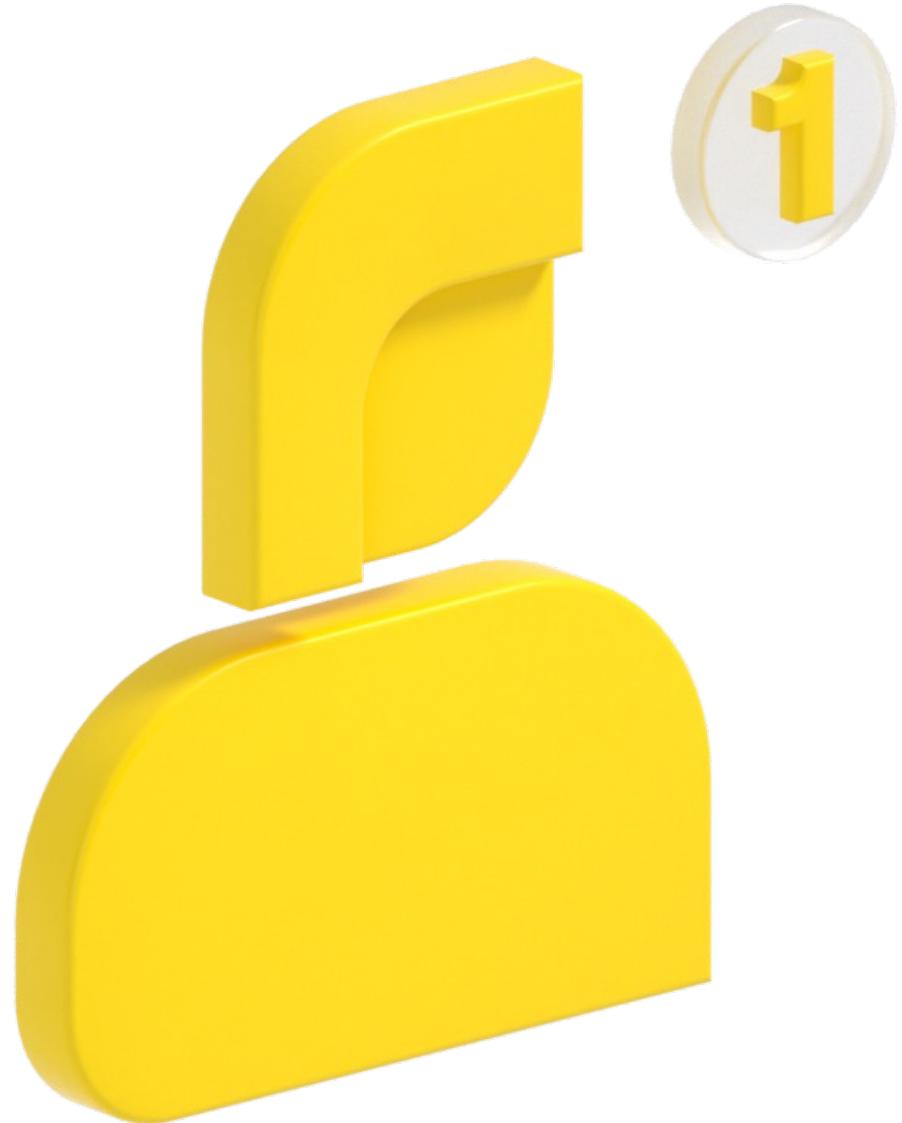
- Аналитики
- ML-инженеры
- ML-ресерчеры
(исследователи)

Кто такой аналитик?



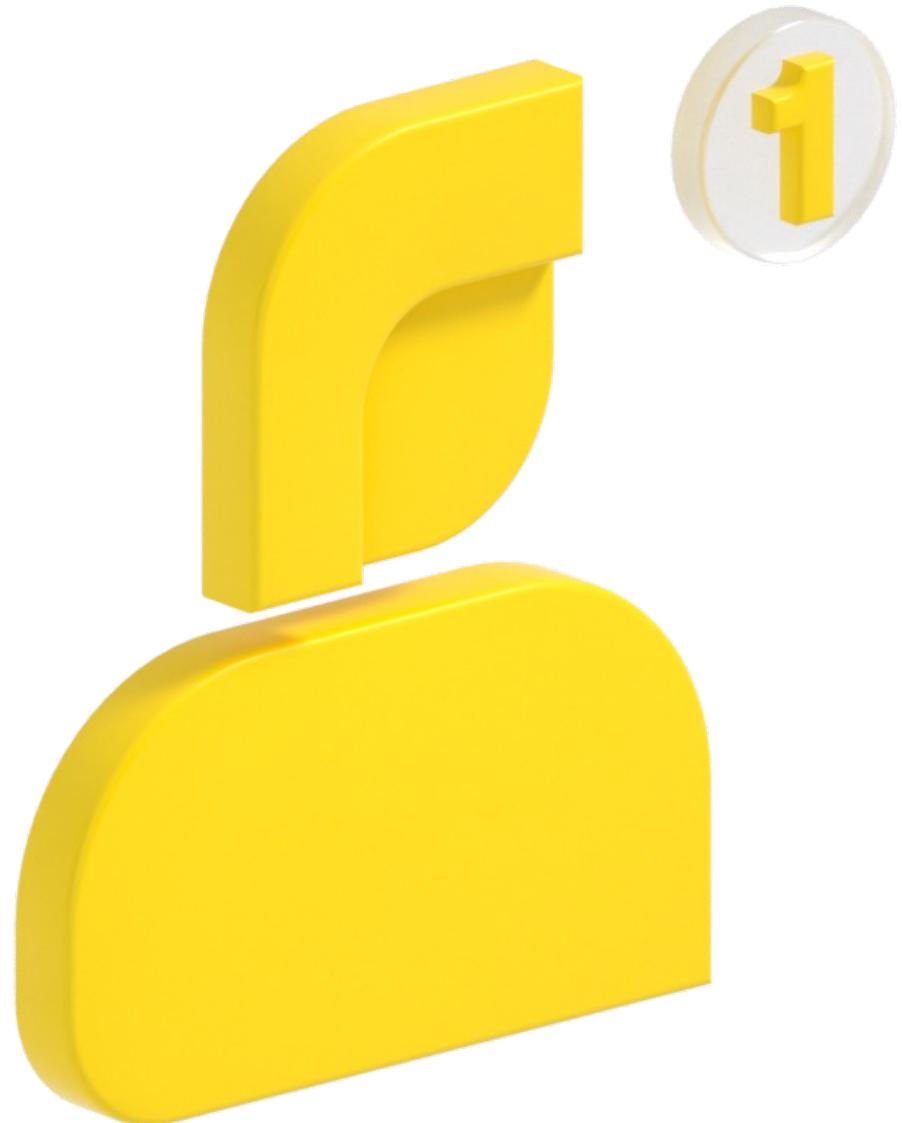
- ➔ Работает с данными, строит по ним отчеты, делает выводы
- ➔ Берёт на себя часть работы инженеров
- ➔ Ищет новые возможности для автоматизации
 - Что хочет пользователь?
 - Как это поможет бизнесу?
- ➔ Проводит тесты новых решений
- ➔ Нужно сильно уметь в математику

Кто такой ML-инженер?



- Работает с данными, пишет код, обучает модели
- Автоматизирует как задачи бизнеса, так и свою рутину
 - ML Ops
- Нужно понимать, как работает математика в ML
- Нужно уметь хорошо писать код
- Важно понимать
 - Какую задачу можно решить эвристиками (решения на правилах, выведенных человеком)
 - Для каких задач без моделей не обойтись

Кто такой исследователь?



- Читает статьи
- Придумывает
 - Новые подходы к обучению
 - Новые способы и трюки для обхода существующих ограничений
- Ставит эксперименты
- Публикует статьи
 - Для публикации нужно ещё больше экспериментов, чтобы показать, что придуманный подход действительно работает лучше существующих
- Важно очень хорошо уметь в математику, алгосы, быть упрямым, но в то же время гибким
- Резюме: ресерчеры – сверхлюди

Какие задачи есть в ML?

01

ML – это почти всегда
про прикладные вещи

02

ML-инженеры и аналитики
решают какие-то бизнес-задачи

03

Ресерчеры также часто
занимаются прикладными
исследованиями

04

Посмотрим на примеры
таких задач

05

В курсе в основном будем
рассматривать задачи ML-
инженеров



Примеры задач в ML



Знаем доход от наших запущенных продуктов

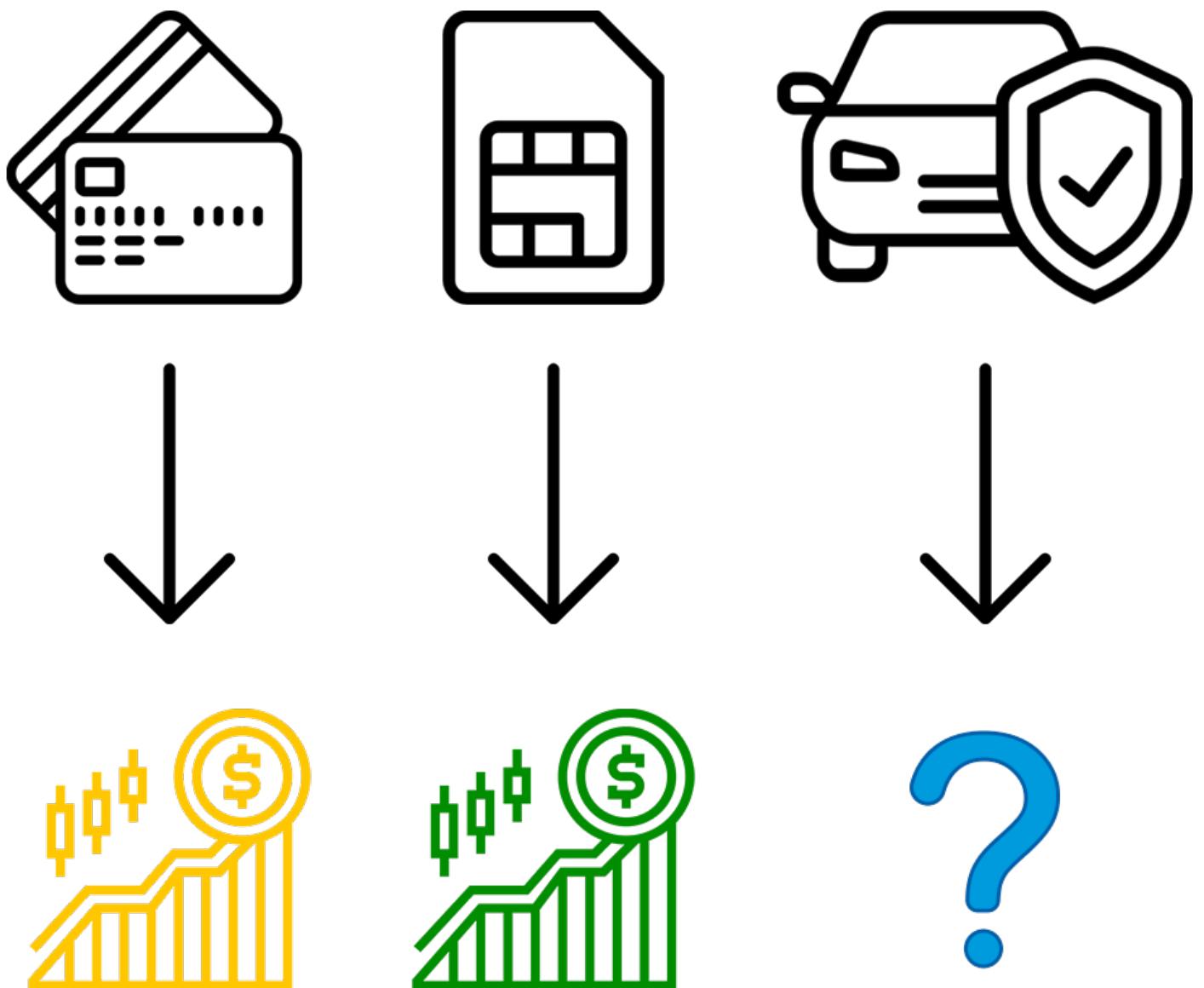


Известны параметры продуктов

- Охват пользователей
- Условия
- Возраст целевой аудитории
- Её средний доход
- И т.д.



Надо предсказать доход от нового продукта

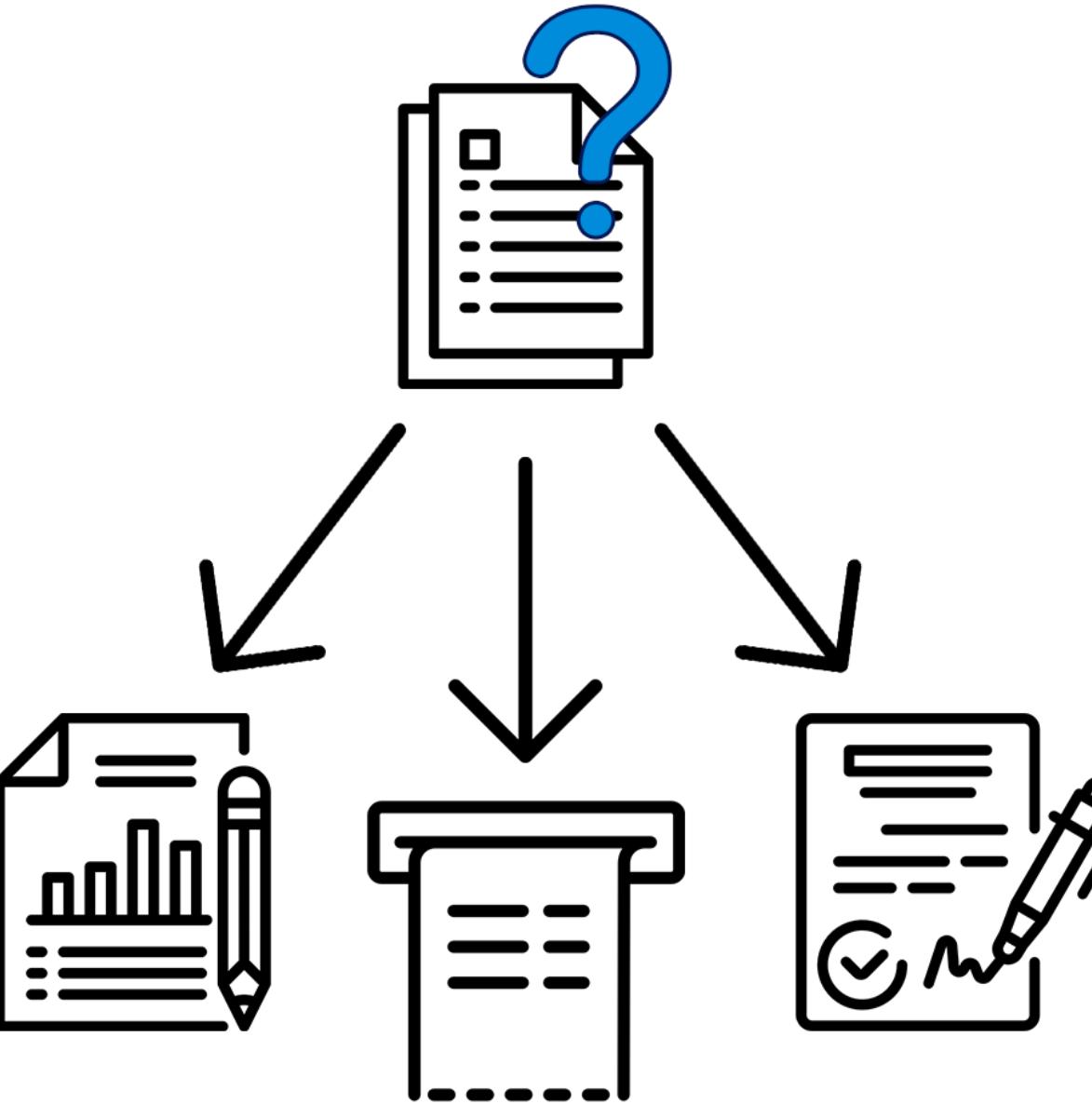


Примеры задач в ML

★ Пользователи присылают нам сканы
своих документов

★ Надо распределять документы
по категориям:

- «Счета»
- «Договоры»
- «Чеки»



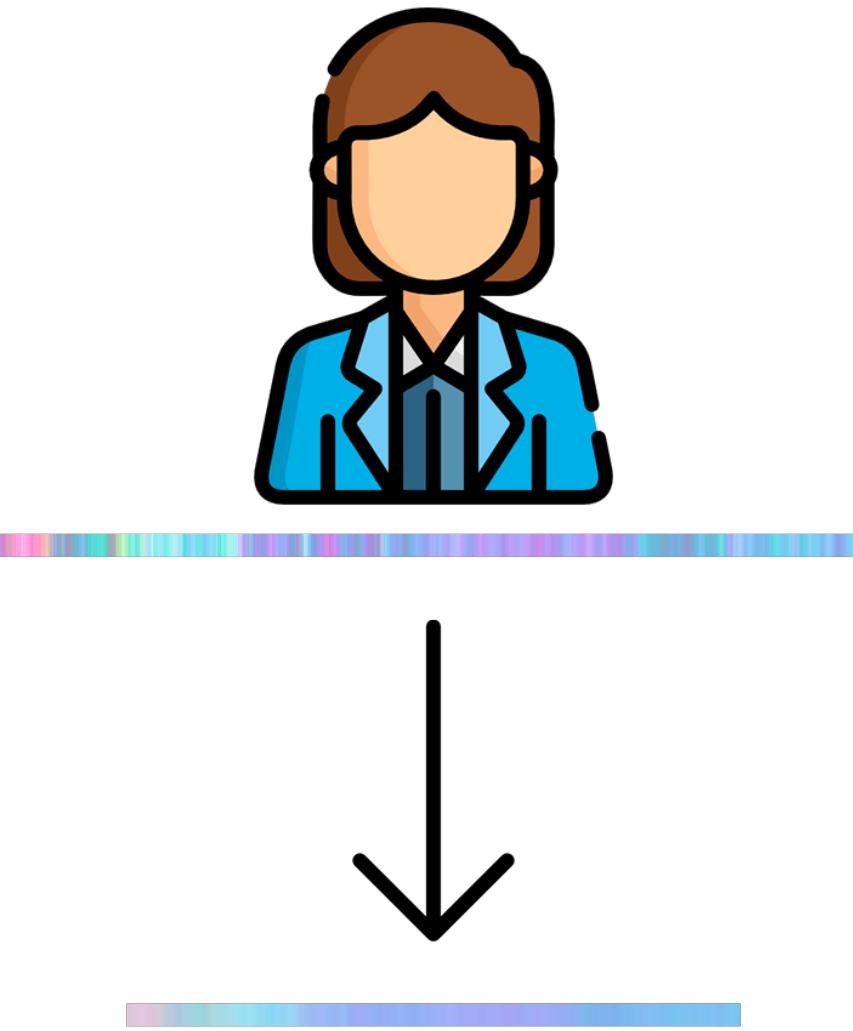
Примеры задач в ML

- ★ Знаем разные параметры пользователей наших продуктов
- ★ Хотим персонализировать будущие предложения
- ★ Надо разбить пользователей на независимые группы



Примеры задач в ML

- ★ Есть большая база данных пользователей
- ★ Они все характеризуются слишком большим числом параметров
- ★ Сгенерировать маленький набор параметров из уже существующих
- ★ Описание пользователей должно быть исчерпывающим
 - Как новых, так и старых



**Можно ли формализовать
возникающие задачи?**



Задача регрессии

- Задачи, в которых по известным признакам объектов надо предсказать вещественное число
- Вещественное число, которое надо предсказать – т.н. метка объекта
- Если объектам сопоставлена метка, говорят, что данные размечены



Примеры задач

- Предсказание цены товара
- Предсказание дохода от нового продукта
- Предсказание котировок на бирже

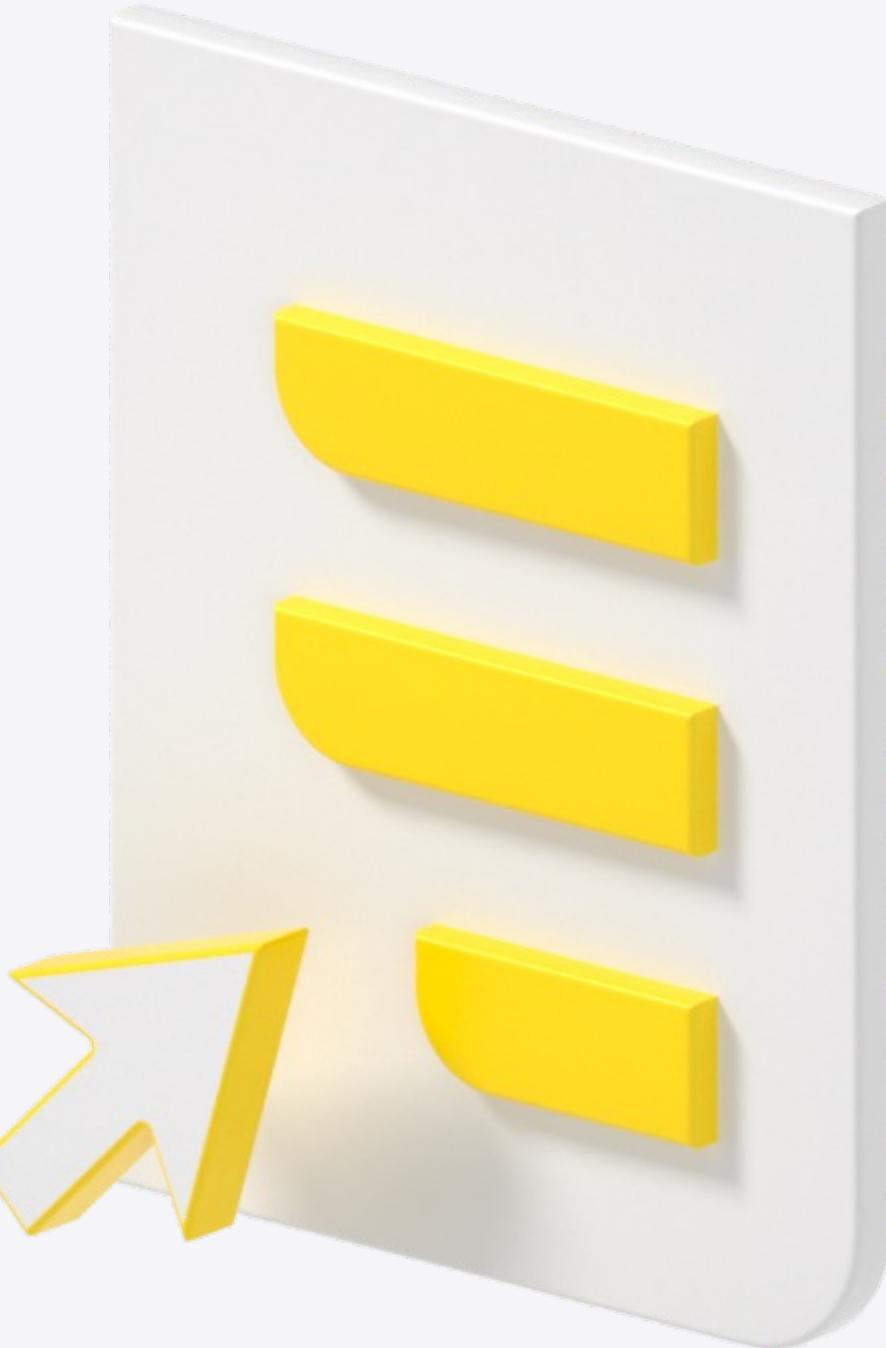
Задача классификации

По заданным признакам объекта предсказать его категорию

Заранее заданное конечное число категорий

- Простейший случай: две категории
- Более сложный случай: много категорий
- Ещё более сложный случай: категории могут быть вложены друг в друга

Категории тоже считаем метками объектов



Примеры задач

▪ Определить мошенническую транзакцию

▪ Определить что хочет пользователь по фразе

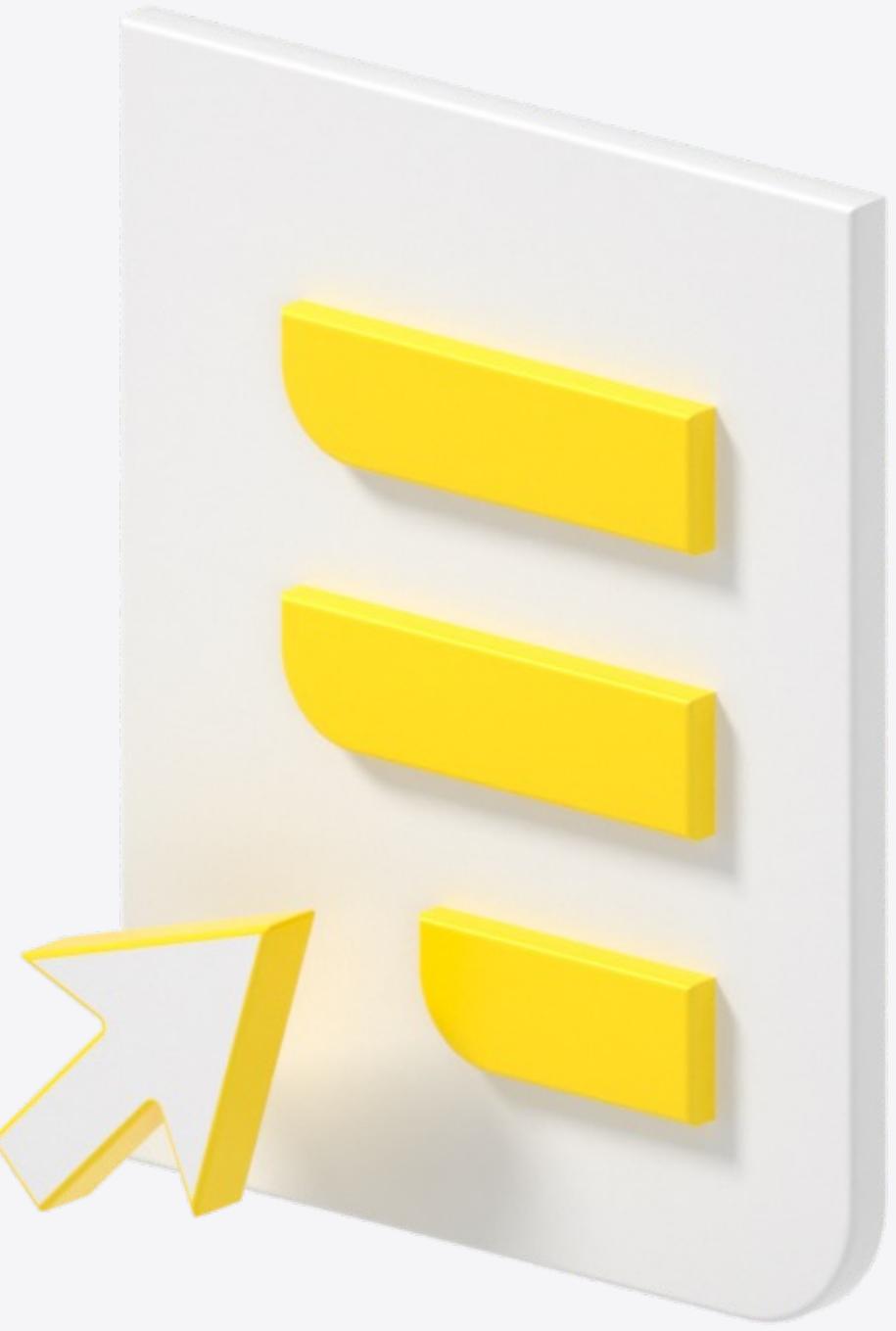
Задача кластеризации



Найти категории в объектах без каких-либо заранее заданных меток

Примеры задач

- Выделение групп пользователей
- Выделение новых категорий (классов) в неразмеченных данных



Задача уменьшения (понижения) размерности



Выбрать наиболее существенные признаки объектов

- Либо сгенерировать новые на основе существующих



Для новых объектов (в той же предметной области) эти признаки также должны быть существенными

Примеры задач

- Уменьшение числа признаков для вычислительной эффективности
- Создание признаков для сложных объектов (текст, звук, картинки)

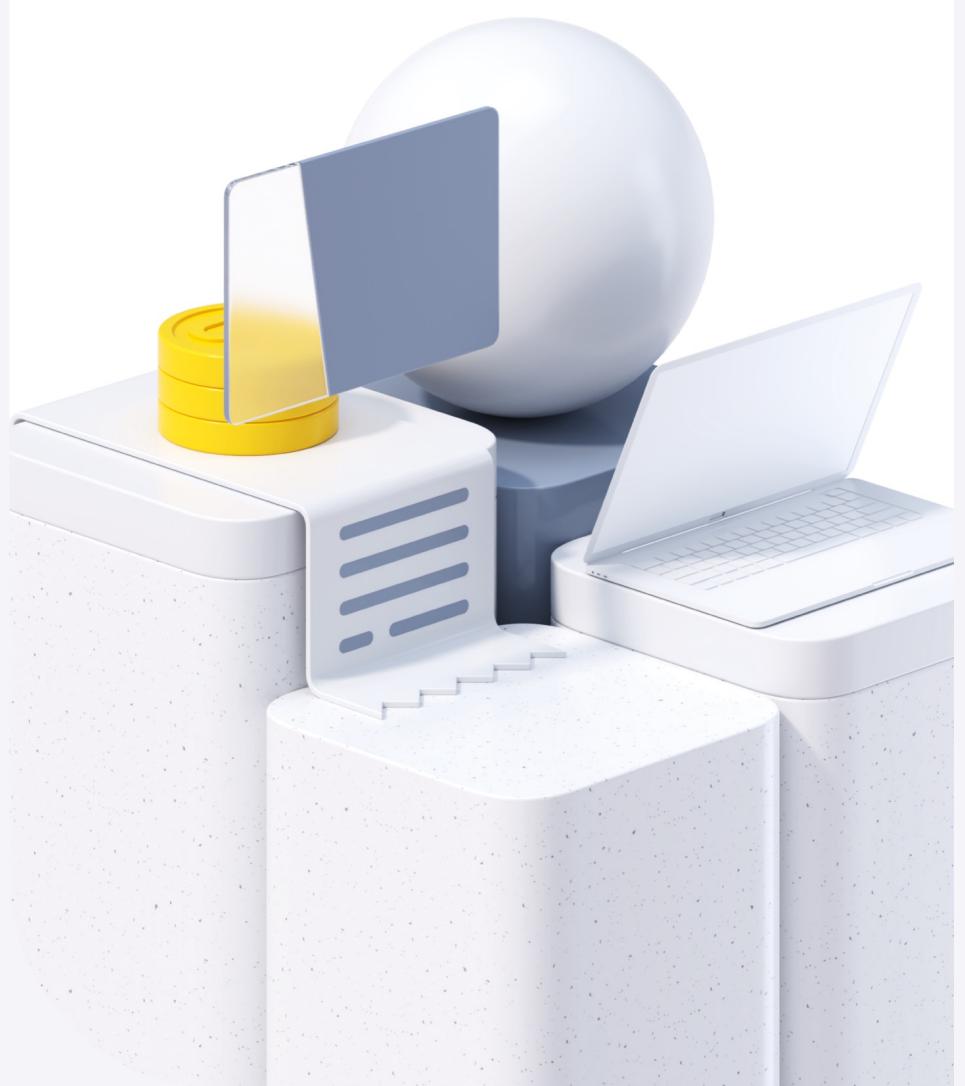


Как ещё сгруппировать задачи?

В ML есть более широкие классы задач

Supervised vs Unsupervised

- Обучение с учителем vs обучение без учителя
- Supervised – есть целевая метка, которую мы хотим предсказывать (регрессия, классификация)
- Unsupervised – у объектов нет целевой метки, работаем только с их признаками (кластеризация, понижение размерности)

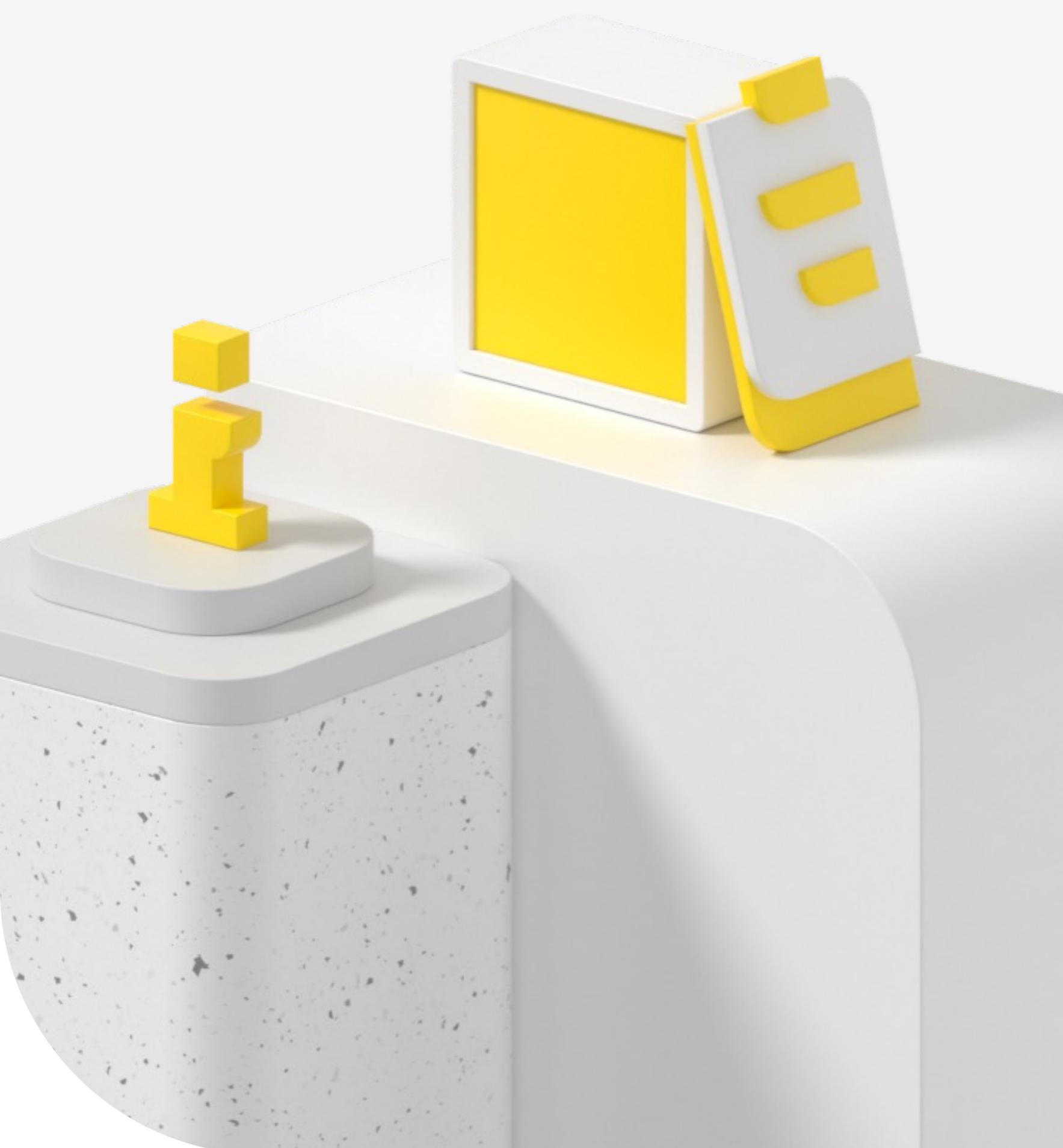


Как ещё сгруппировать задачи?

Discriminative vs Generative

- Более сложное разбиение, более сложные примеры задач
- Дискриминативные задачи: моделируем зависимость меток от признаков
- Генеративные задачи: моделируем сами признаки (генерация новых объектов)





Ближе к практике – Инструментарий

В ML редко пишут что-то с нуля

**Когда может возникнуть потребность
написать с нуля какой-то из алгоритмов?**

- Во время изучения, чтобы разобраться в его тонкостях и подводных камнях
- Если существующие инструменты платные/содержат критичные баги
- Если это новейший подход, придуманный в вашей команде

Обзор

Инструмен- тариЙ

Работа
с математикой

numpy

Работа
с классическими
алгоритмами ML

sklearn

Работа
с табличными
данными

pandas

Не факт,
что вы будете
пользоваться
этим каждый
день

Работа
с график[о/ами]

matplotlib,
seaborn

То, с чего надо
начать

Инструментарий – пипту



- ❖ Ключевая библиотека, делающая ML на питоне быстрым
- ❖ За счет перехода к C++ и AVX-инструкций ускоряет большинство операций
- ❖ Используется в качестве базы в других основных библиотеках
- ❖ Важно научиться писать вместо циклов пипту-код

Инструментарий – pandas



Обертка над питру для работы с таблицами



Естественное представление датасетов



Удобно делать предварительную обработку
данных



Удобно строить всякую статистику, отчетики,
делать аналитику

Инструментарий – matplotlib



- ❖ Строим графики
- ❖ Визуализируем данные
- ❖ Упрощаем восприятие наших выводов
- ❖ Удобно обосновывать свою позицию

Инструментарий – sklearn



- 👉 Реализованы все классические алгоритмы ML
- 👉 Тоже использует внутри себя питону
- 👉 То, чем вы будете пользоваться почти весь семестр



ТИНЬКОФФ

Вопросы

