

# Машинное обучение

Занятие 1.  
Введение в машинное обучение.  
Подготовка данных.

# Обо мне



- Тимлид DS-команды СВМ ТСХ
- Старшая за стажеров Департамента анализа данных



- Эксперт Центра непрерывного образования
- Преподаватель бакалавриата и ЦНО
- PhD студент, магистр'18 и бакалавр'15



# Связь и общение

- @elentevanyan
- [elentevanyan.edu@gmail.com](mailto:elentevanyan.edu@gmail.com)

Домашки с нужными темами:

- [elentevanyan.edu@gmail.com](mailto:elentevanyan.edu@gmail.com)

# Материалы курса

- Слайды, конспекты, ссылки, ноутбуки:

[https://github.com/elentevanyan/rdp\\_ml\\_course](https://github.com/elentevanyan/rdp_ml_course)

# ПО

- Jupyter Notebook
- Ставим Anaconda и наслаждаемся:  
<https://anaconda.cloud/installers>

# Расписание и программа курса

<b>Дата</b>	<b>День недели</b>	<b>Тема</b>
7 декабря	Вторник	Введение в машинное обучение
10 декабря	Пятница	Функционалы и метрики качества
14 декабря	Вторник	Линейные алгоритмы
17 декабря	Пятница	Решающие деревья
22 декабря	Среда	Бэггинг и бустинг
24 декабря	Пятница	Кластеризация
14 января	Пятница	Поиск аномалий и Feature Extraction
18 января	Вторник	Рекомендательные системы
21 января	Пятница	Нейронные сети
25 января	Вторник	Введение в статистику
28 января	Пятница	Планирование экспериментов

# Контроль обучения

- 4 обязательных рецензируемых домашних задания
- 1 inclass competition
- После каждого занятия опциональные упражнения для тренировки и закрепления

# Для нашего комфорта

- Стартуем в 10.30
- В каждую встречу работаем по 2 часа с перерывом в 15 минут
- Перерыв в 11.40-11.55
- Дедлайны по работам ставим вместе, чтили их тоже вместе ☺

Слайд для организационных  
вопросов,

о которых я не подумала

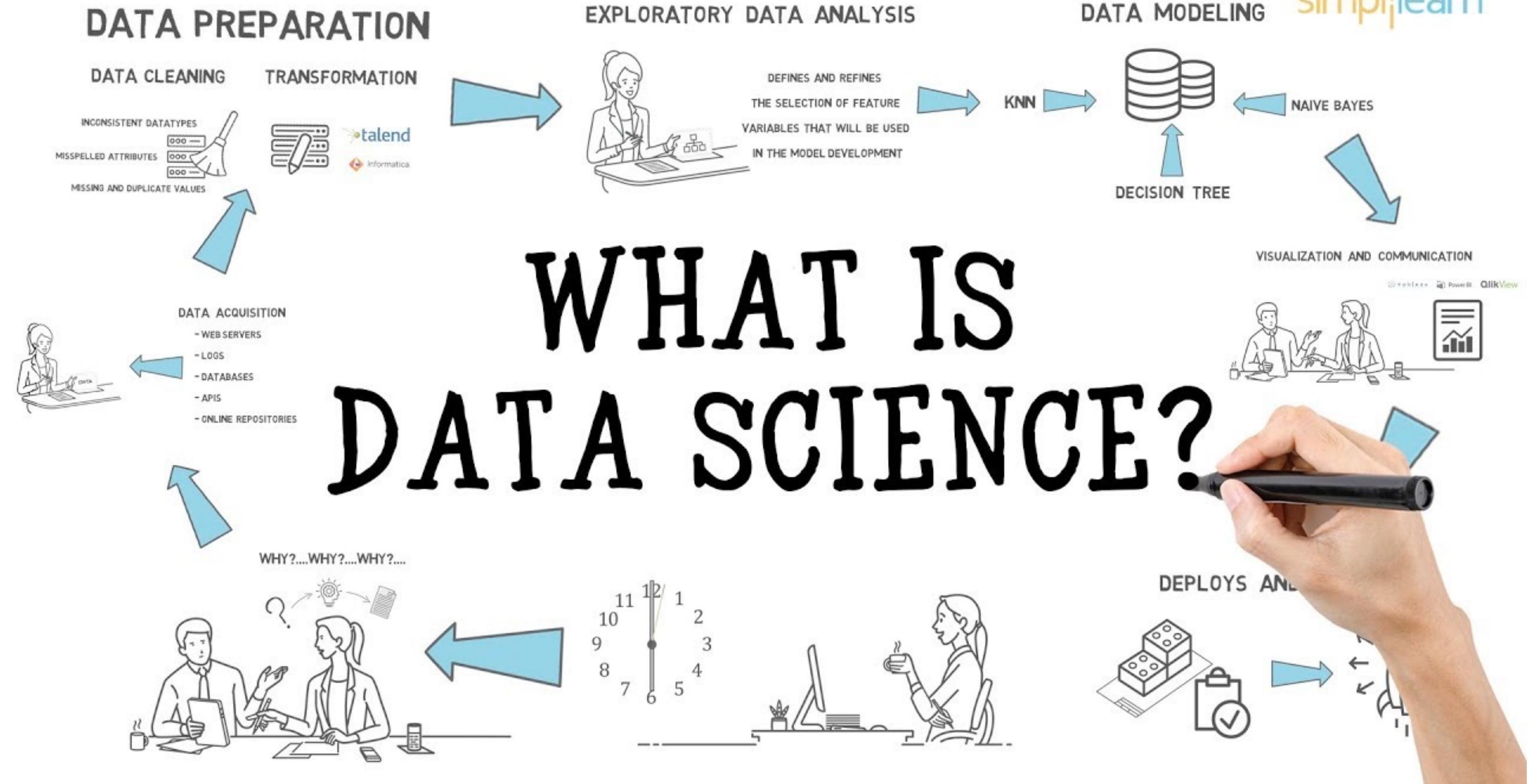
Краткое введение в область

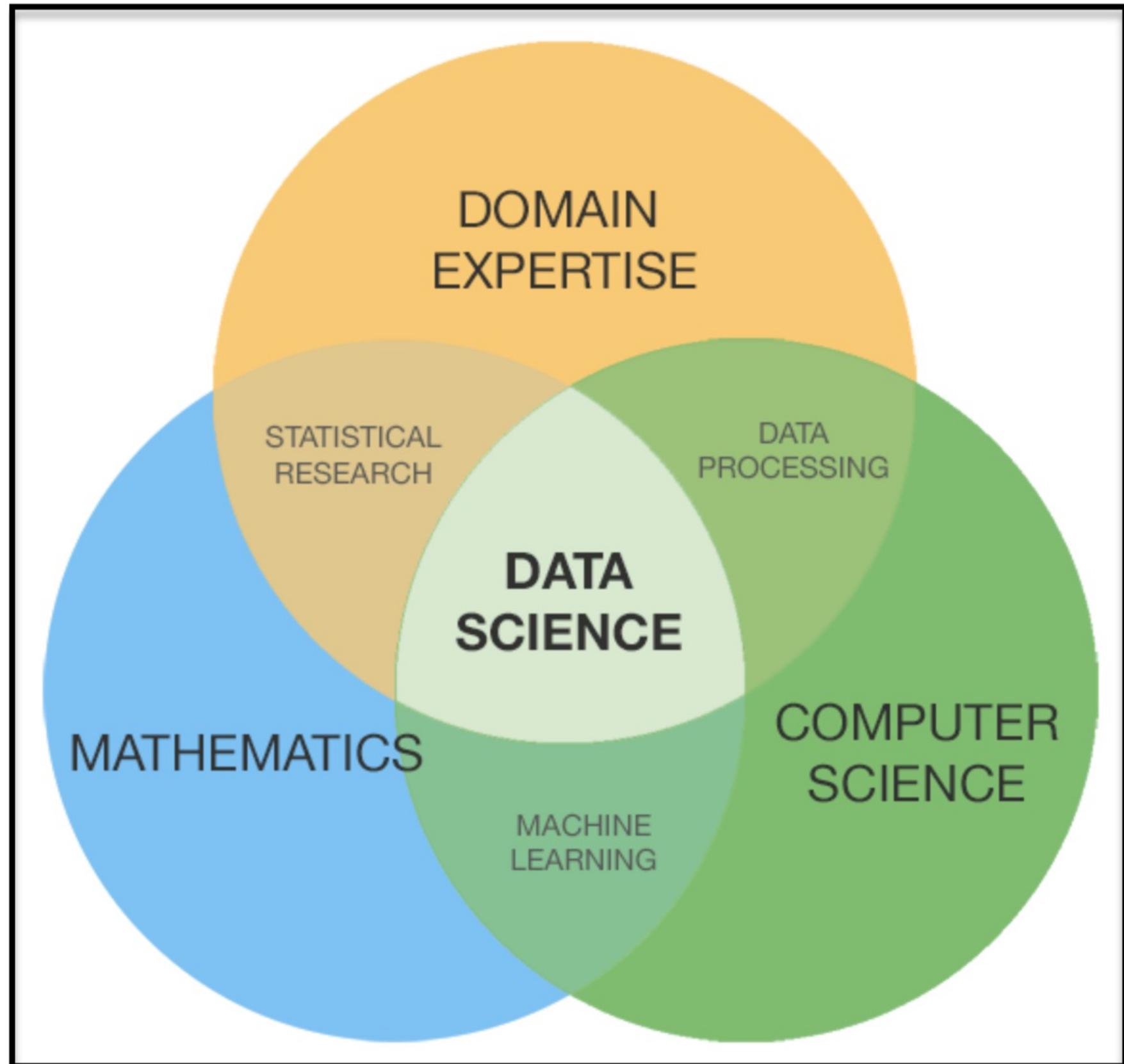
DATA

# Data Scientist: The Sexiest Job of the 21st Century

by [Thomas H. Davenport](#) and [D.J. Patil](#)

FROM THE OCTOBER 2012 ISSUE







**что делать если скучно**

**что делать если скучно за компом**

**что делать если скучно дома**

**что делать если скучно на уроке**

**что делать если скучно трум трум**

**что делать если скучно в симс 4**

**что делать если скучно дома мне 12 лет девочке**

**что делать если скучно дома мне 10 лет девочке**

**что делать если скучно за компом ссылки на прикольные сайты**

**что делать если скучно с подругой дома**

наука о данных (англ. *data science*, иногда датасайенс — *datascience*) — раздел информатики, изучающий проблемы анализа, обработки и представления данных в цифровой форме. Объединяет методы по обработке данных в условиях больших объёмов и высокого уровня параллелизма, статистические методы, методы интеллектуального анализа данных и приложения искусственного интеллекта для работы с данными, а также методы проектирования и разработки баз данных. Скрыть

## 🕒 Что такое **data science** и как это работает? | Ru...

[rb.ru](#) > Авторские колонки > [что-такое-dt](#) ▾

Термин **data science** на русский переводят как «наука о данных», а в профессиональной среде часто просто транслитерируют – «дата сайенс». Формально это набор некоторых взаимосвязанных дисциплин и методов из... Читать ещё >



## >Data Science Skills / Хабр

[habr.com](#) > [post/271085/](#) ▾

Data Science также немного пересекается с такими областями деятельности как ... Data Science – это новая область деятельности, поэтому требования к Data Scientists еще не до конца сформированы. Читать ещё >

## 귤 Дорога в **Data Science** глазами новичка

[pikabu.ru](#) > [story/doroga\\_v\\_data\\_science\\_glaz...](#) ▾

Что такое **Data Science**? В 21 веке информация повсюду. Вы буквально не можете жить, не оставляя вокруг себя информационный след. Читать ещё >



[Поиск](#) [Картинки](#) [Видео](#) [Карты](#) [Маркет](#) [Новости](#) [Эфир](#) [Коллекции](#) [Знатоки](#) [Услуги](#) [Ещё](#)

1200x1200

700x700

600x600

694x694

## ● Купить Наушники в интернет-магазине М.Виде...

[Наушники Bluetooth](#) [Наушники-вкладыши](#)[mvideo.ru > naushniki/naushniki-3967](#)

Наушники в интернет-магазине «М.Видео» представлены широким ассортиментом устройств. Цены варьируются от 390 до 109990 рублей. Читать ещё >



## W Наушники — Википедия

[ru.wikipedia.org > Наушники](#)

Стереофонические наушники (наушники) — два телефона с оголовьем, предназначенные для подключения к бытовым радиоэлектронным аппаратам. Читать ещё >



## ● Купить наушники в Москве, низкие цены на на...

[svyaznoy.ru > catalog/audiovideo/1558](#)

В интернет магазине Связной представлен широкий выбор наушников с онлайн-подбором совместимых брендов и моделей. В нашем каталоге Вы можете заказать... Читать ещё >



## ● Наушники купить наушники и гарнитуры... - Мо...

[doctorhead.ru > Наушники](#)

Лучшие наушники в интернет магазине Doctorhead.ru, широкий ассортимент, проводные и беспроводные наушники, для плеера и профессиональные.



## OZON.ru | Количество излучателей в каждом на...

[OZON.ru > catalog/1196662/](#)

Наушники и гарнитуры. 2364 товара. ... Наушники TWS Наушники Беспроводные с микрофоном TWS I9S, TWS-I9S, белый. Читать ещё >

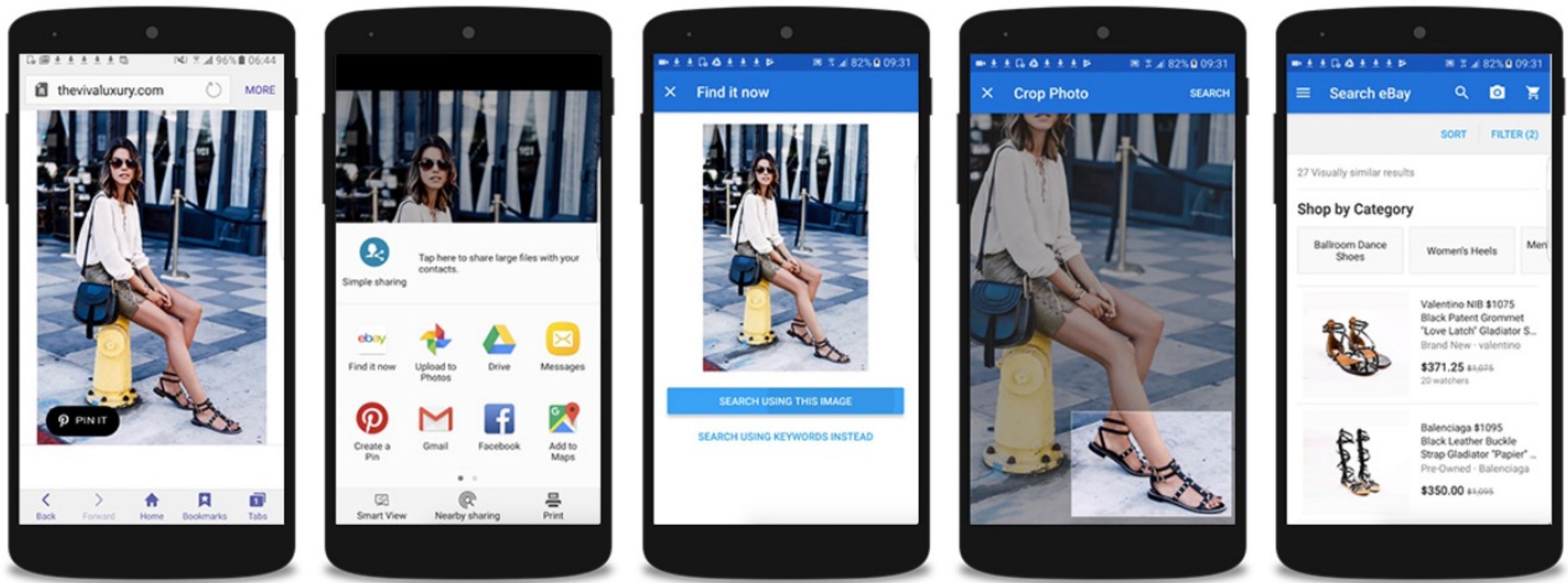


## ● Наушники и Bluetooth-гарнитуры на Маркете

[Яндекс.Маркет > Наушники и Bluetooth-гарнитуры](#)

Изображение: РИА Новости

© 2014



ebay

**CREDIT SCORE**

FAIR

GOOD

EXCELLENT



## Watch It Again



## Because you watched Outlander



## Critically-acclaimed TV Dramas >

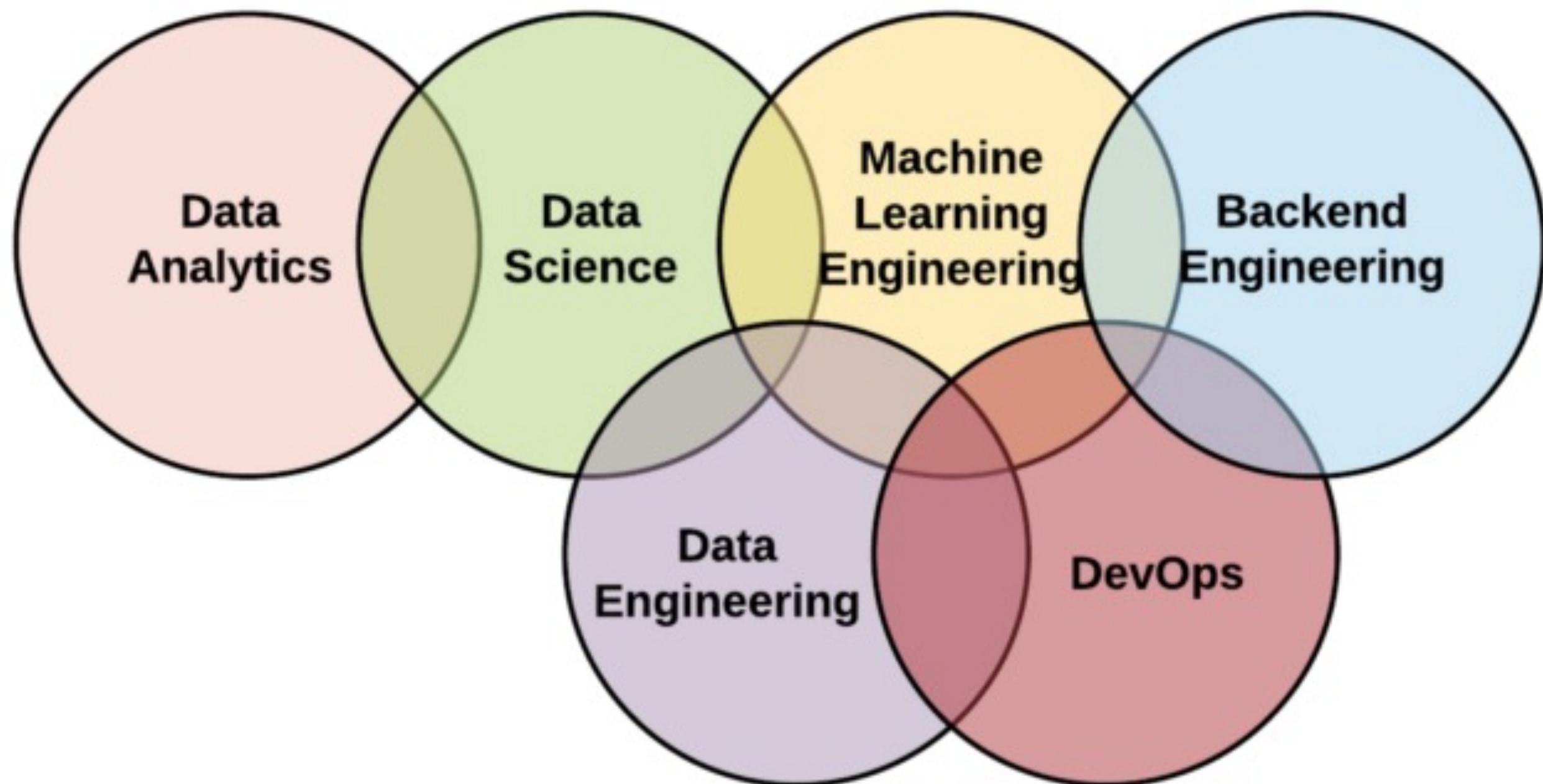


## Romantic TV Dramas





Data профессии



# Основная тройка

## Data Analyst



### Что делают:

Помогают анализировать данные с помощью отчетов и графиков

### Общаются с бизнес-линией

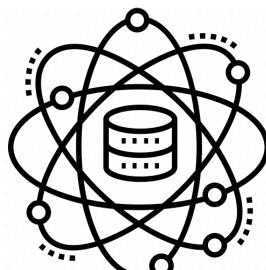
#### Навыки:

Статистика,  
Экспертиза в индустрии

#### Технологический стек:

Excel, SQL, Tableau, R/Python

## Data Scientist



### Что делают:

Ищут паттерны и закономерности в данных, строят предиктивные модели

### Общаются с бизнес-линией

**Навыки:** Статистика, Computer Science, Programming

#### Технологический стек:

SQL, Python, Hive, Spark

## Data Engineer



### Что делают:

Поддерживают всю инфраструктуру обработки данных, выводят решения DS в «бой»

### Общаются с ИТ и с DS

#### Навыки:

Programming, Computer Science

#### Технологический стек:

SQL/NoSQL, Python, C++, Hadoop, Hive, Spark,

Два мифа

# BIG DATA

- Это вообще все данные
- Это данные, превышающие определенный объем (больше 100ГБ, 500ГБ, 1 ТБ, 10 ТБ и т.д.)
- Это данные, которые невозможно обработать на одном компьютере
- Термин не существует, придуман маркетологами

# BIG DATA

- Это подходы, методы и инструменты для хранения и обработки структурированных и неструктурированных данных

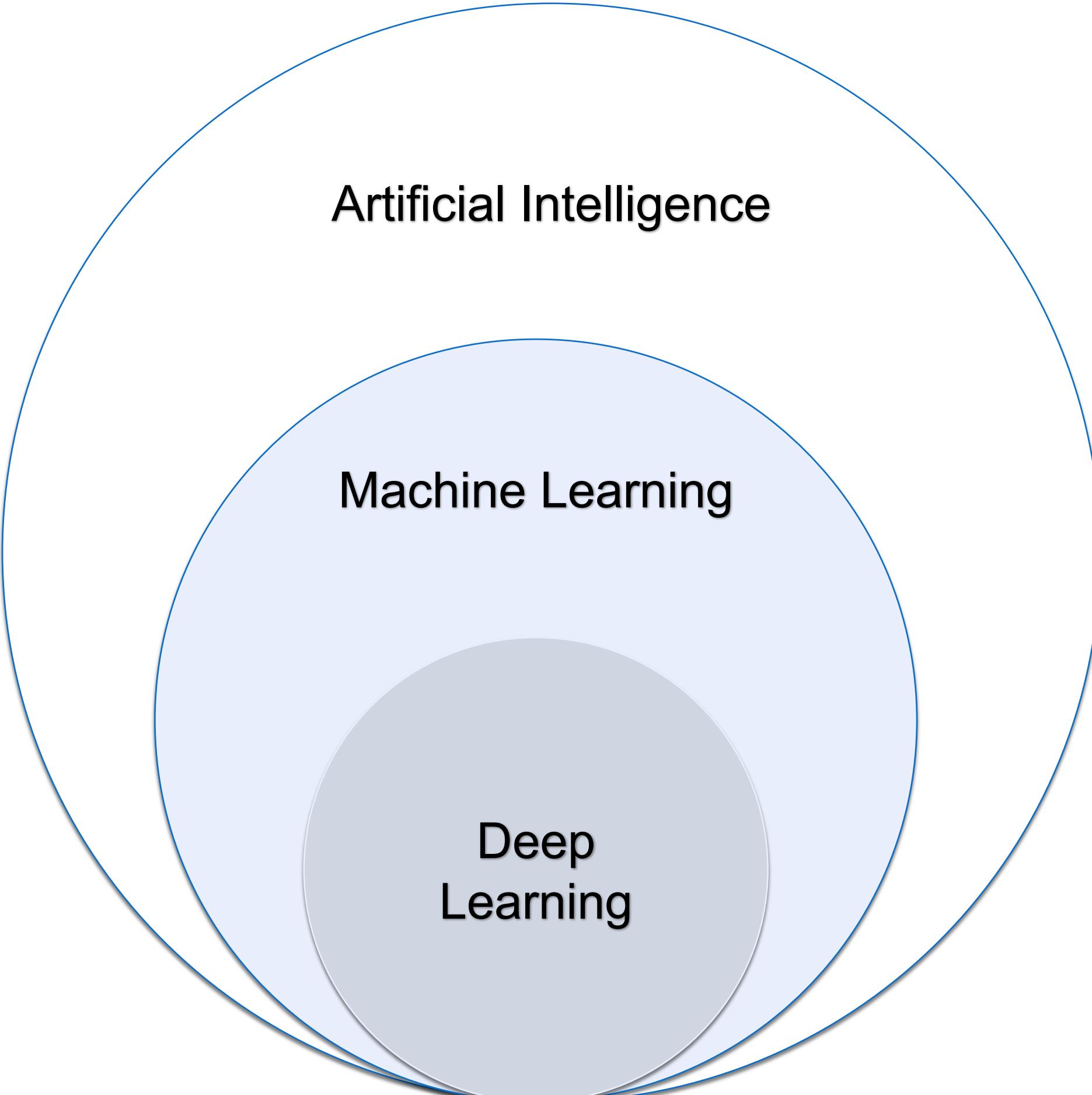
# АНОНС

- Наш курс не по бигдате, а по машинному обучению

**Artificial  
Intelligence**  
(Искусственный  
интеллект)

**Machine  
Learning**  
(Машинное  
обучение)

**Deep Learning**  
(Глубокое  
обучение)



**Artificial Intelligence**

**Machine Learning**

**Deep  
Learning**

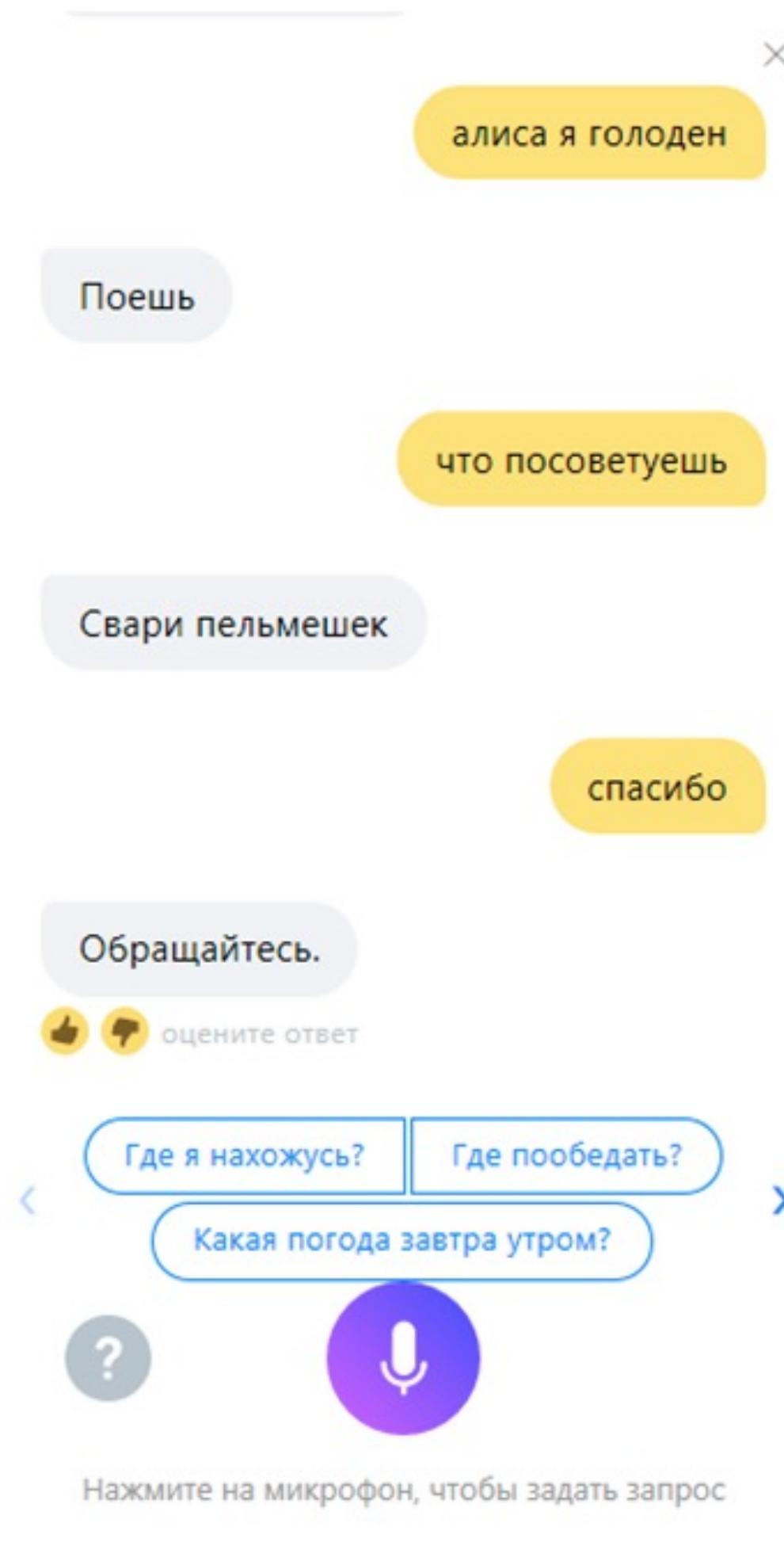
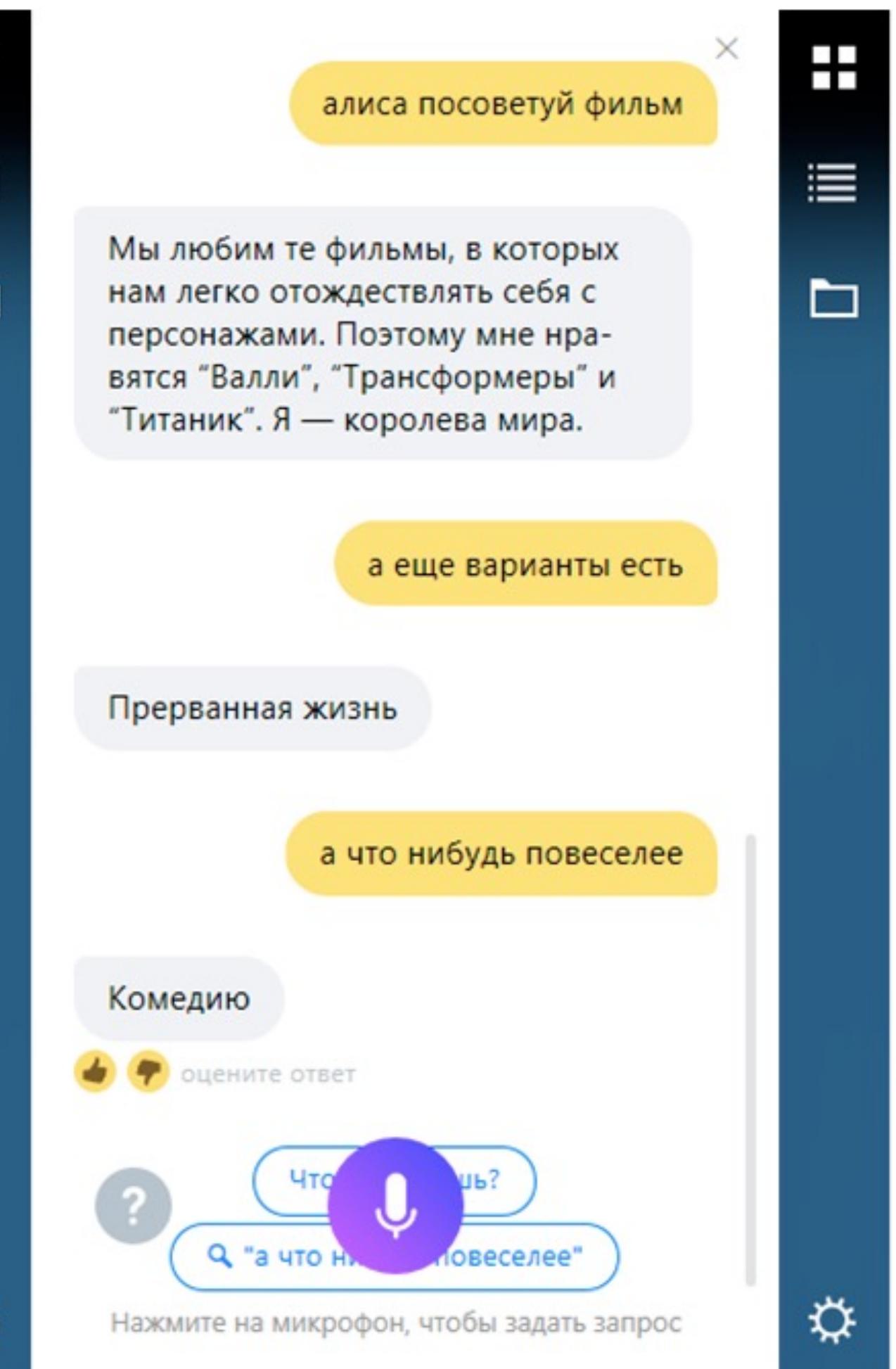
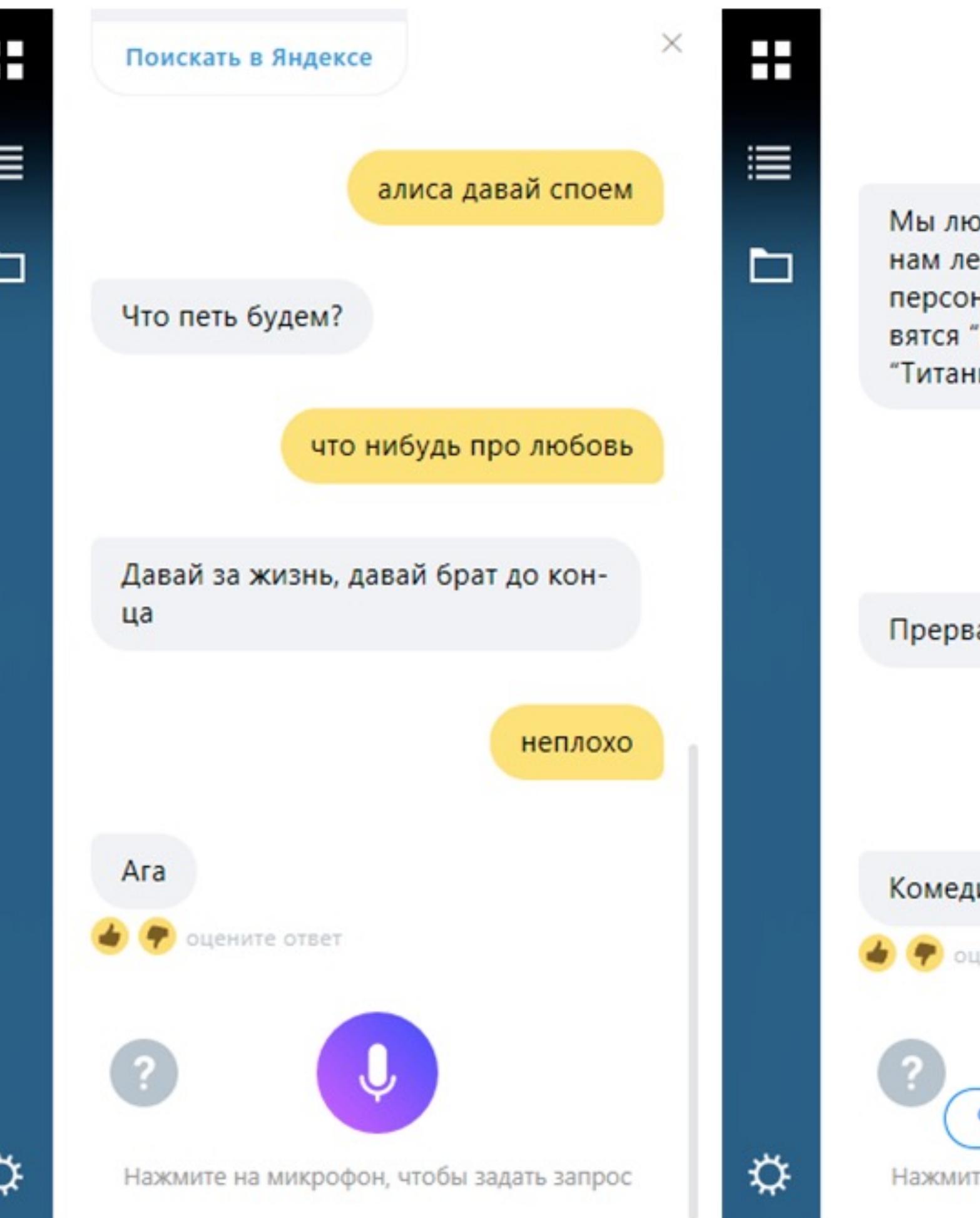
# Основные определения

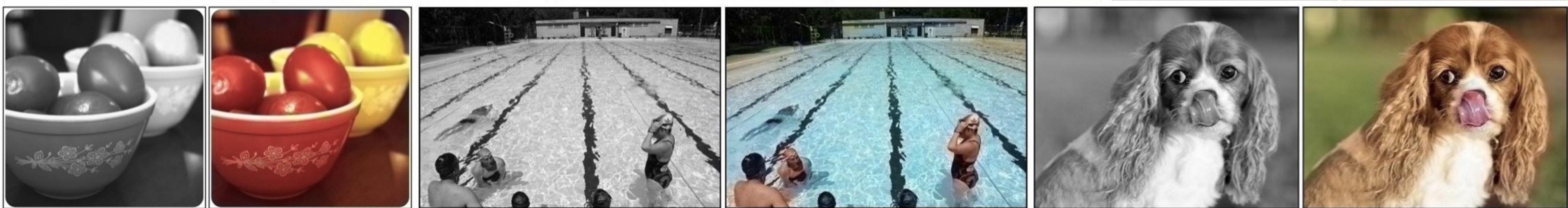
- Искусственный интеллект – широкая область, в которой изучают процесс принятия решений.
- Машинное обучение – подобласть искусственного интеллекта, в которой на основании данных машины учатся принимать решения без прямого, явного программирования по сценариям.
- Глубокое обучение – подобласть машинного обучения, сфокусированная на нейронных сетях.



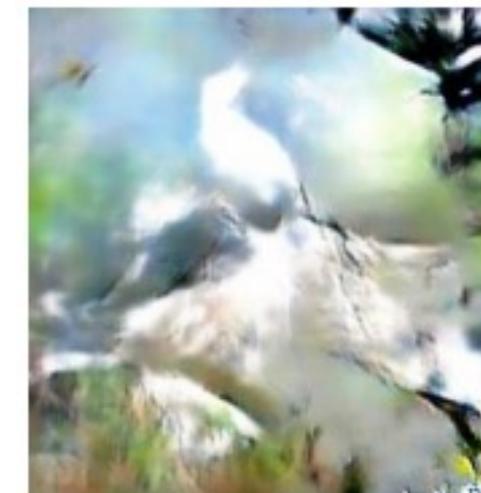
Yandex Taxi

SELF-DRIVING CAR  
prototype

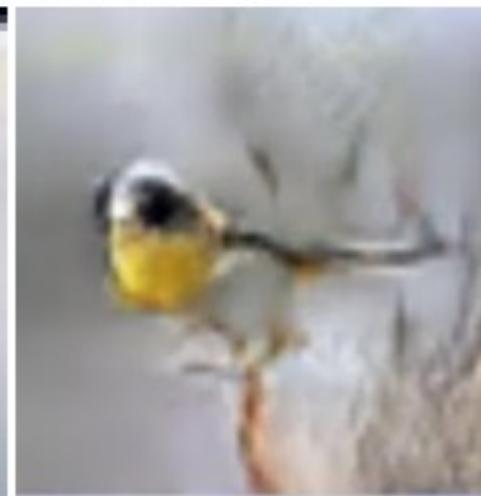




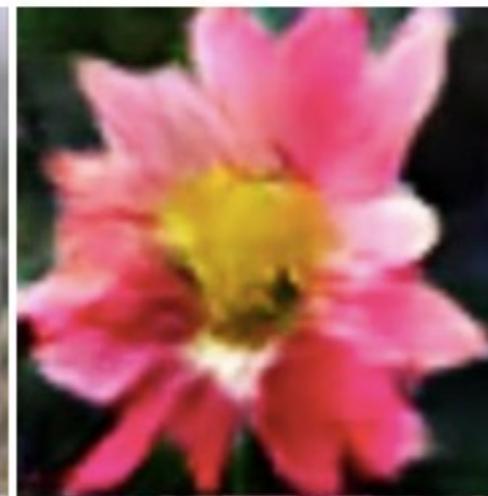
This bird is white with some black on its head and wings, and has a long orange beak



This bird has a yellow belly and tarsus, grey back, wings, and brown throat, nape with a black face



This flower has overlapping pink pointed petals surrounding a ring of short yellow filaments



# Пример задачи

- Для улучшения эффективности диспетчерских служб такси важно знать, когда водитель закончит один заказ и будет готов принять следующий.
- Оценка длительности текущей поездки – один из факторов эффективного распределения заказов.
- Как оценить длительность поездки?

# Терминология

- $x$  (**sample**) – объект, для которой хотим делать предсказания
- Поездки
- $y$  (**target**) – ответ, целевая переменная, т.е. То, что хотим предсказать
- Длительность поездки
- $(x_i, y_i)_{i=1}^{\ell}$  – обучающая выборка, прецеденты, т.е. все объекты, для которых известны значения целевого признака
- $\ell$  – размер выборки.

# Признаки

- Компьютер умеет работать с числовой информацией
- Объекты характеризуются числовой информацией – признаками, факторами, «фичами» (от англ. features)
- $m$  – число признаков
- $x = (x^1, \dots, x^m)$

# Признаки

- Временные
  - Дата и время посадки
  - Дата и время высадки
- Географические
  - Ширина и долгота места посадки
  - Ширина и долгота места высадки
- Погодные
  - Осадки: дождь, снег, шторм
  - Сила осадков
- Маршруты
  - Наиболее быстрые маршруты
  - Скорость по маршрутам
- Пассажиры
  - Число пассажиров

# Признаки

- Временные
  - Дата и время посадки
  - Дата и время высадки
- Географические
  - Ширина и долгота места посадки
  - Ширина и долгота места высадки
- Погодные
  - Осадки: дождь, снег, шторм
  - Сила осадков
- Маршруты
  - Наиболее быстрые маршруты
  - Скорость по маршрутам
- Пассажиры
  - Число пассажиров

# Обучение

- Задача прогнозирования длительности поездки на такси
- Есть обучающая выборка  $X$  и соответствующие наблюдениям  $Y$
- Задача - восстановить функцию
- $a: X \rightarrow Y$
- $a(X)$  - алгоритм, или модель

# Обучение

- $a(x)$  – алгоритм/модель
- Это функция, предсказывающая ответ для любого объекта
- Алгоритм предсказал 100 минут, а поездка длилась 83 минут. Хорошее ли предсказание или плохое?
- Функционал качества – мера корректности алгоритма
- Для нашей задачи можно использовать среднеквадратическую ошибку Mean Square Error:

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

# Обучение

- Есть прецеденты
- Определен функционал качества
- Фиксируется параметризованное семейство алгоритмов:
- «Если время после  $t$  часов, то длительность заказа сокращается на 10%»
- Обучение – поиск оптимальных алгоритмов с точки зрения функционала качества.

# Виды целевой переменной

- Регрессия:  $Y = \mathbb{R}$
- Бинарная классификация:  $Y = \{0; 1\}$
- Многоклассовая классификация:  $Y = \{0, 1, \dots, K\}$
- Многоклассовая классификация  
с пересечением классов  $Y = \{0; 1\}^K$

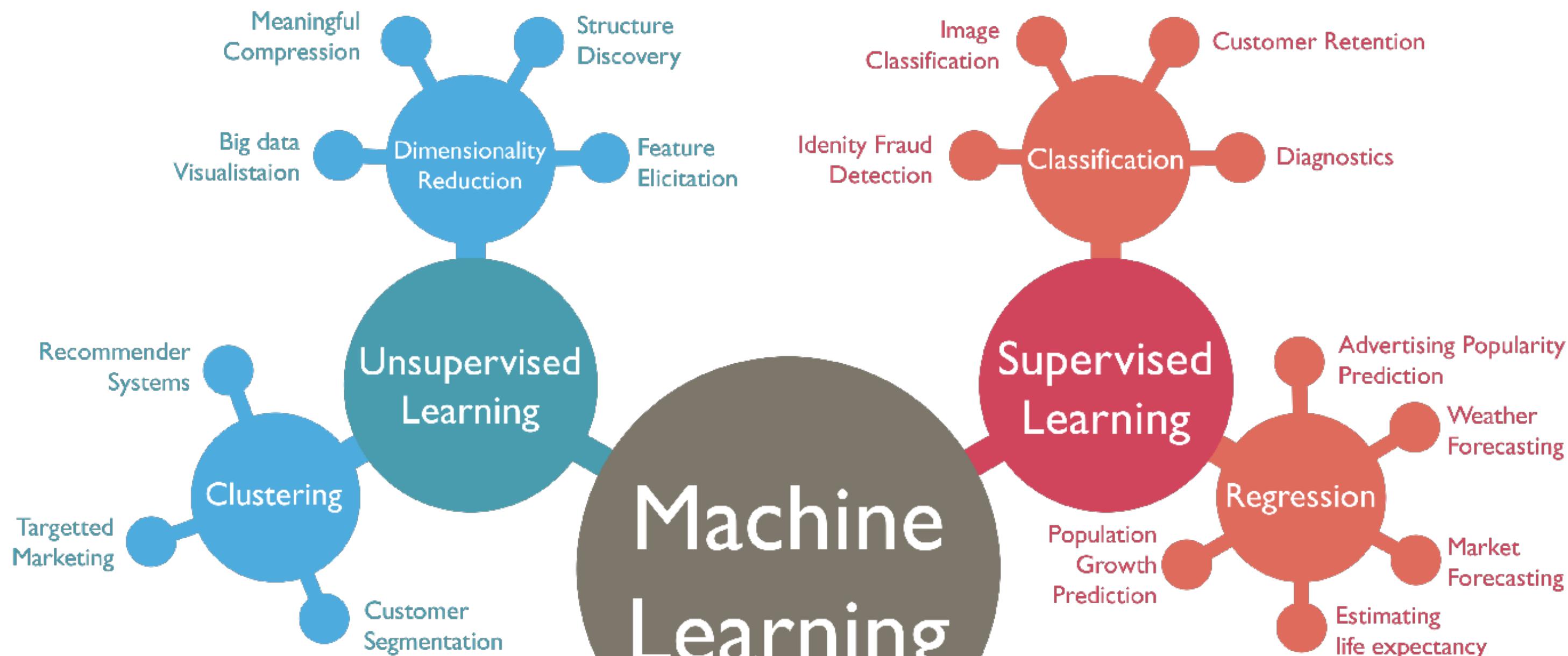
# Примеры классификации

- Предсказание пола для неизвестного пользователя
- Определение типа документа
- Определение языка документа
- Определение эмоционального окраса отзыва
- Вероятность ухода сотрудника/клиента
- Предсказание типов писем: спам/не спам
- Определение объектов на фотографии
- Оценка состояния человека по ЭЭГ

# Примеры регрессионных задач

- Прогнозирование цены дома
- Прогнозирование заработной платы по описанию вакансии
- Прогнозирование спроса на товар в ближайшую неделю
- Прогнозирование уровня экспрессии гена
- Прогнозирование температуры воздуха
- Прогнозирование суммы компенсаций по страховке
- Прогнозирование объема потребления электроэнергии

# Machine Learning



# Разница в двух классах задач

## Обучение с учителем

- $x$  (**sample**) – объект, для которой хотим делать предсказания
- $y$  (**target**) – ответ, целевая переменная, т.е. То, что хотим предсказать

## Обучение без учителя

- $x$  (**sample**) – объект, для которой хотим делать предсказания

# Примерный порядок действий

- Постановка задачи
- Выделение признаков
- Формирование выборки
- Выбор метрики качества
- Предобработка данных
- Обучение модели
- Оценка качества модели

# Резюме:

- Бигдатой заниматься не будем, но ИИ - точно
- Термины МЛ:
  - объект, таргет, фичи
  - функция потерь
- Основные задачи МЛ
- Виды задач:
  - Обучение с учителем и без учителя