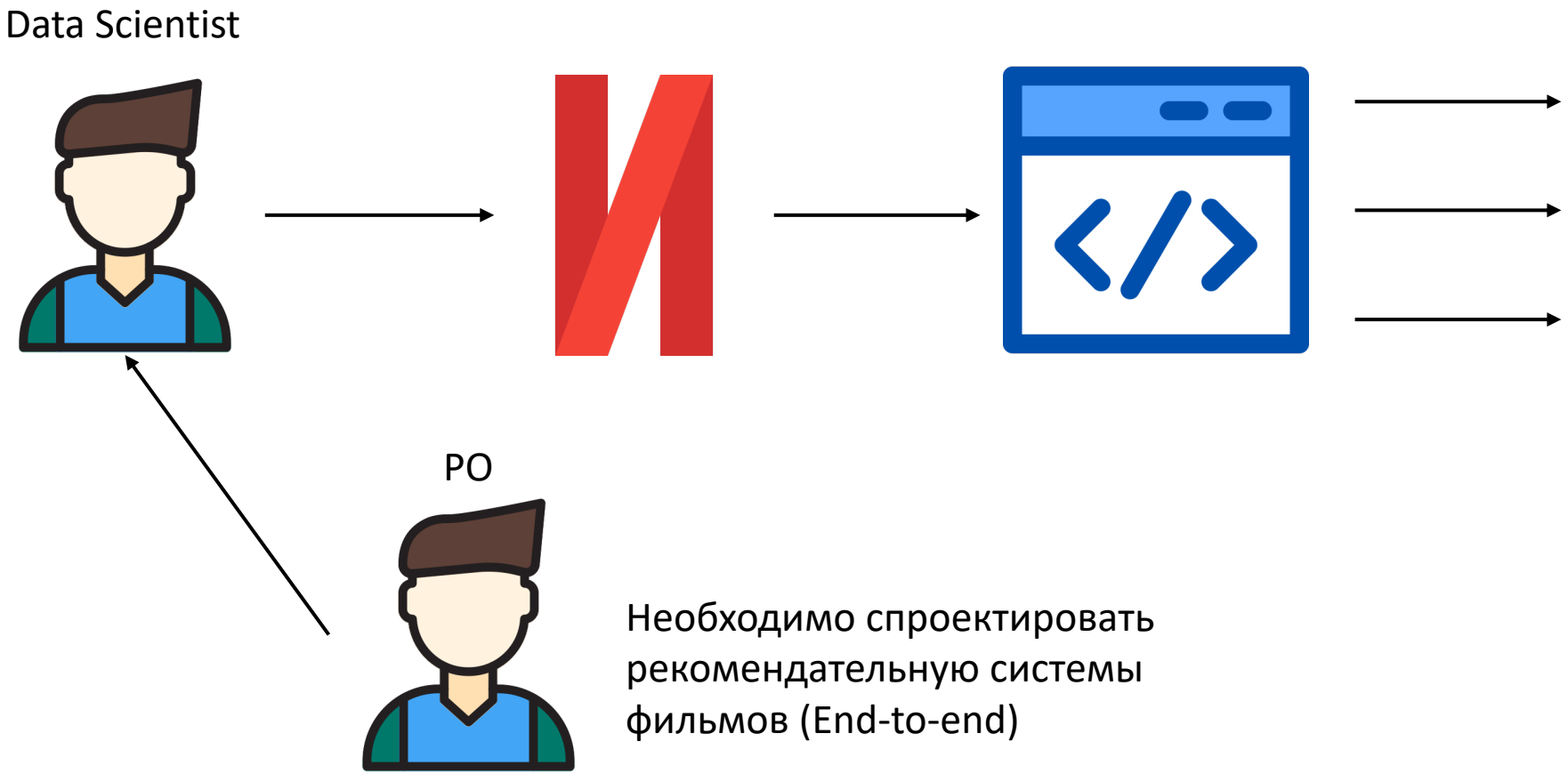
# Введение в машинное обучение и анализ больших данных

*Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed. — Arthur L. Samuel, AI pioneer, 1959*

## Что будет на курсе

- Освоим Python (основы)
- Освоим самые популярные библиотеки Python для анализа данных и ML
- Поговорим о технологиях Big Data (параллельные и распределенные
- вычисления и способы хранения больших данных)
- Изучим основные методы машинного обучения
- Изучим основные методы обработки данных
- Будем писать код и обучать модели
- Будут домашки (3 шт.)
- Будут мини тесты на теорию (каждую пару)
- Поделимся своим опытом работы в ML
- Поговорим о том как стать специалистом в области машинного обучения
- Если останется время поговорим о продвинутых вещах в ML

# Вводный кейс

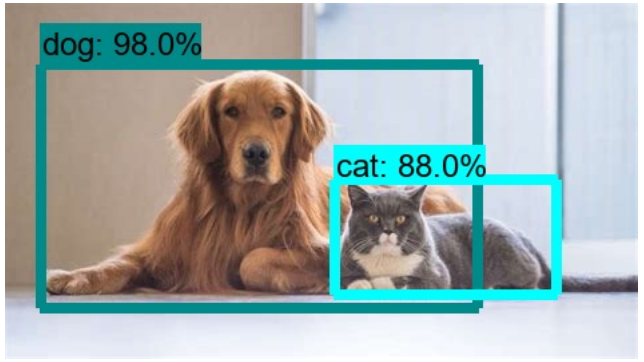


# Пример задач машинного обучения

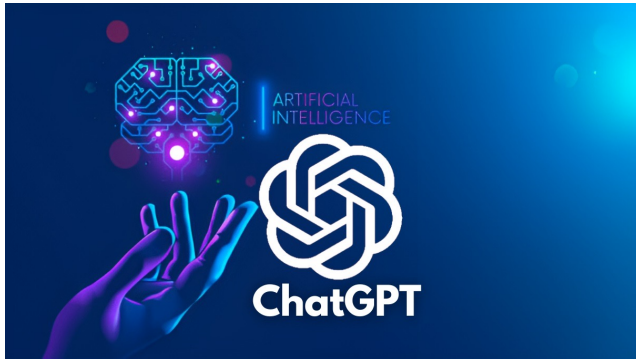
- Email spam detection
- Face detection and matching
- Web search (Yandex, Google)
- Sports predictions
- Post office (e.g., sorting letters by zip codes)
- Credit card fraud
- Stock predictions
- Smart assistants (Apple Siri, Amazon Alexa, . . . )
- Product recommendations (e.g., Walmart, Netflix, Amazon)
- Self-driving cars (e.g., Uber, Tesla)
- Language translation (Google translate)
- Sentiment analysis
- Drug design
- Medical diagnoses
- . . .

# Самые популярные направления

## Машинное зрение




## NLP

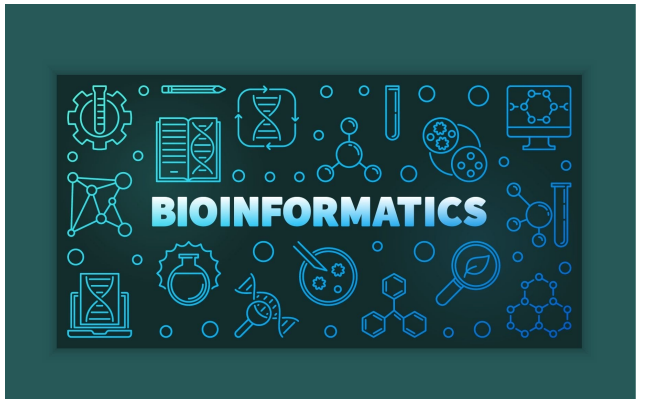



## Решение  бизнес задач


## Обучение с подкреплением


## Беспилотники

## Биология и медицина

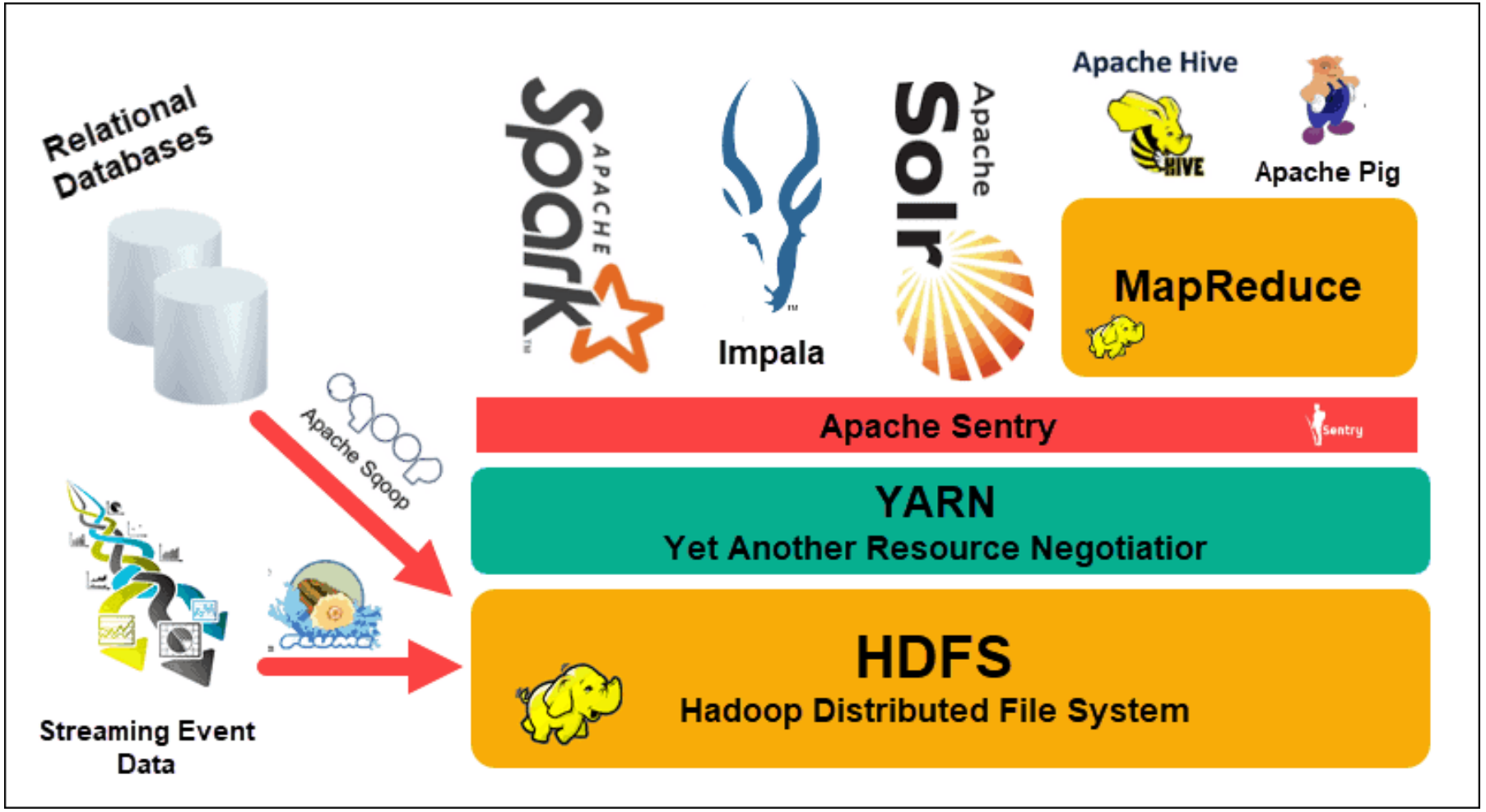



# И многое другое

# Основные ML инструменты


# И многое другое

# Работа с Big Data




# И многое другое

# Основные определения и постановки задач

Машинное обучение — это наука, изучающая способы извлечения закономерностей из ограниченного количества примеров.

В рамках данного курса будут использоваться следующие обозначения: $x$ — объект, $X$ — пространство объектов, $y = y(x)$ — ответ на объекте $x$, $Y$ — пространство ответов.

Объектом называется то, для чего нужно сделать предсказание. В данном примере объектом является пара (пользователь, фильм). Пространство объектов — это множество всех возможных объектов, для которых может потребоваться делать предсказание. В данном примере это множество всех возможных пар (пользователь, фильм). Ответом будет называться то, что нужно предсказать. В данном случае ответ — понравится пользователю фильм или нет. Пространство ответов, то есть множество всех возможных ответов, состоит из двух возможных элементов: -1 (пользователю фильм не понравился) и +1 (понравился). Признаковым описанием объекта называется совокупность всех признаков.

# Выборка, алгоритм обучения

Центральным понятием машинного обучения является обучающая выборка $X = (x_i, y_i)_{i=1}^{l}$
Это те самые примеры, на основе которых будет строиться общая закономерность. Отдельная задача — получение обучающей выборки. В вышеупомянутом случае $y_i$ - это оценка фильма пользователем. Предсказание будет делаться на основе некоторой модели (алгоритма) $a(x)$, которая представляет из себя функцию из пространства $X$ в пространство $Y$. Эта функция должна быть легко реализуема на компьютере, чтобы ее можно было использовать в системах машинного обучения.

Не все алгоритмы подходят для решения задачи. Например константный алгоритм $a(x) = 1$ не подходит. Это довольно бесполезный алгоритм, который вряд ли принесет пользу сайту. Поэтому вводится некоторая характеристика качества работы алгоритма — функционал ошибки. $Q(a, X)$ — ошибка алгоритма а на выборке $X$. Например, функционал ошибки может быть долей неправильных ответов. Следует особо отметить, что $Q$ называется функционалом ошибки, а не функцией. Это связано с тем, что первым его аргументом является функция. Задача обучения состоит в подборе такого алгоритма а, для которого достигается минимум функционала ошибки. Лучший в этом смысле алгоритм выбирается из некоторого семейства $\mathbb{A}$ алгоритмов.

# Обучение на размеченных данных

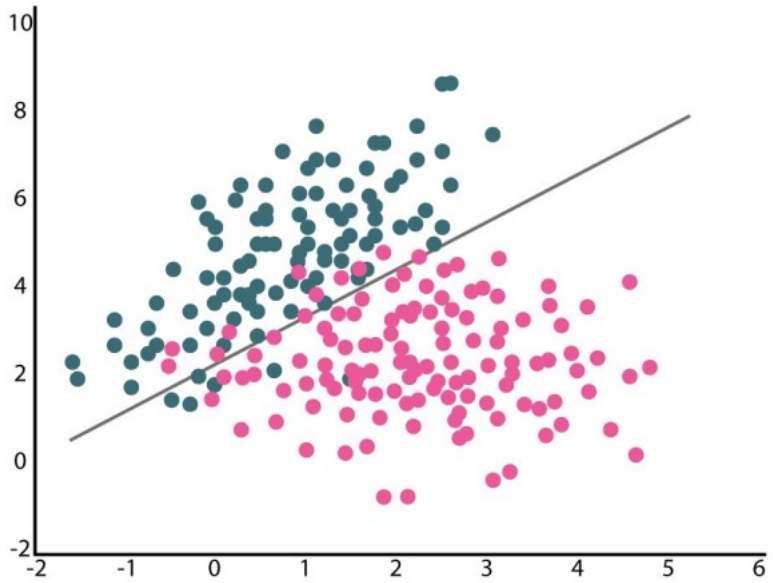
Общая постановка задачи обучения с учителем следующая. Для обучающей выборки $X$= $(x_i, y_i)_{i=1}^{l}$ нужно найти такой алгоритм $a \in A$, на котором будет достигаться минимум функционала ошибки:

$$Q(a, X) \to \min_{a \in \mathbb{A}}.$$

В зависимости от множества возможных ответов $Y$, задачи делятся на несколько типов (их очень много).

# Задача бинарной классификации

В задаче бинарной классификации пространство ответов состоит из двух ответов $Y = \{0, 1\}$. Множество объектов, которые имеют один ответ, называется классом. Говорят, что нужно относить объекты к одному из двух классов, другими словами, классифицировать эти объекты.

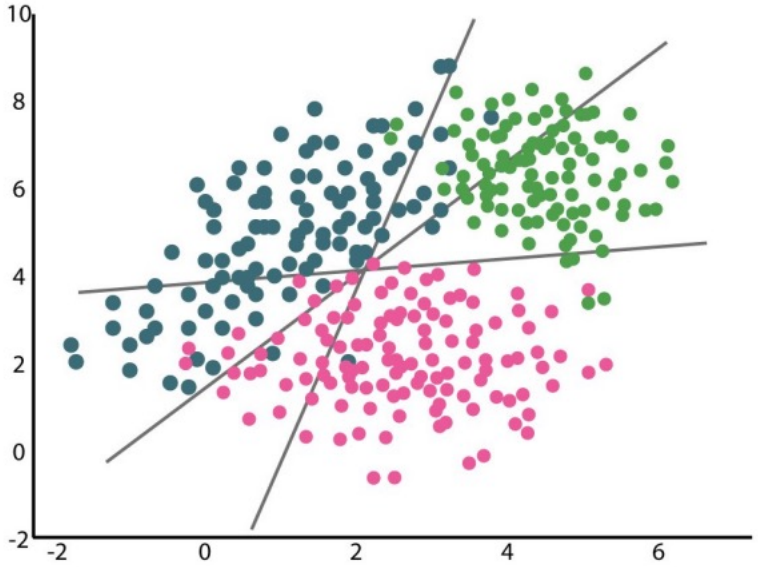



Примеры задач бинарной классификации:

- Понравится ли пользователю фильм?
- Вернет ли клиент кредит?

# Задача многоклассовой классификации

Классов может быть больше, чем два. В таком случае имеет место задача многоклассовой классификации.

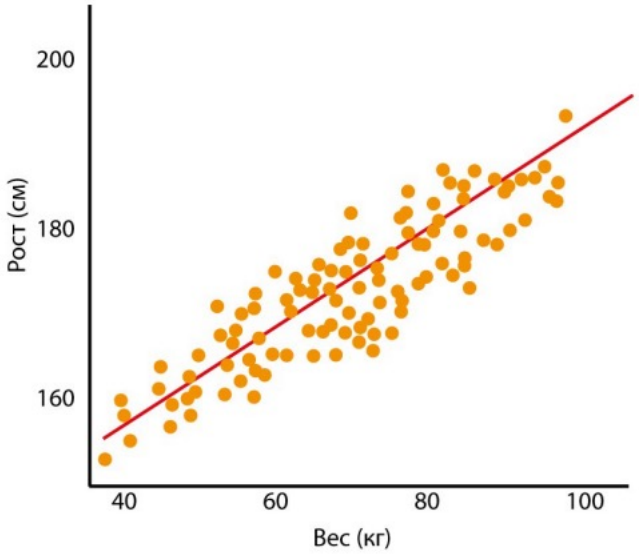



Примеры задач многоклассовой классификации:

- Из какого сорта винограда сделано вино?
- Какая тема статьи?
- Машина какого типа изображена на фотографии: мотоцикл, легковая или грузовая машина?

# Задача регрессии

Когда y является вещественной переменной, говорят о задаче регрессии.

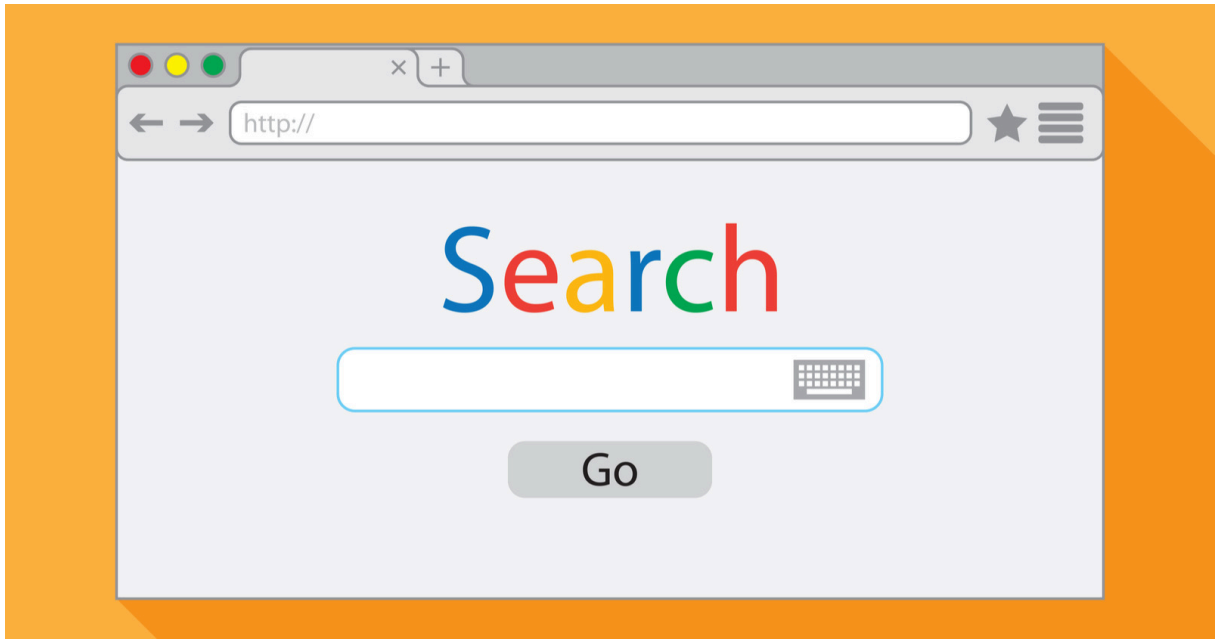



Примеры задач регрессии:

- Предсказание температуры на завтра.
- Прогнозирование выручки магазина за год.
- Оценка возраста человека по его фото.

# Задача ранжирования

Еще одним примером задачи обучения с учителем является задача ранжирования. Эта задача довольно тяжелая, и речь о ней в данном курсе не пойдет, но знать о ней полезно. Мы сталкиваемся с ней каждый день, когда ищем что-либо в интернете. После того, как мы ввели запрос, происходит ранжирование страниц по релевантности их запросу, то есть для каждой страницы оценивается ее релевантность в виде числа, а затем страницы сортируются по убыванию релевантности. Задача состоит в предсказании релевантности для пары (запрос, страница).

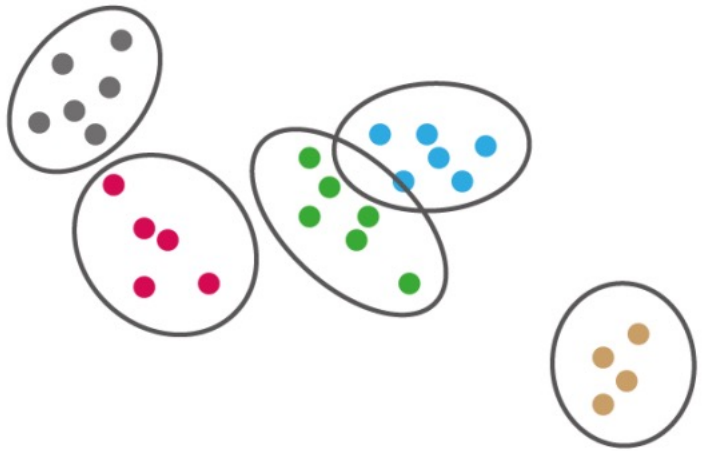


# Обучение без учителя

Обучением с учителем называются такие задачи, в которых есть и объекты, и истинные ответы на них. И нужно по этим парам восстановить общую зависимость. Задача обучения без учителя — это такая задача, в которой есть только объекты, а ответов нет. Также бывают «промежуточные» постановки. В случае частичного обучения есть объекты, некоторые из которых с ответами. В случае активного обучения получение ответа обычно очень дорого, поэтому алгоритм должен сначала решить, для каких объектов нужно узнать ответ, чтобы лучше всего обучиться. Рассмотрим несколько примеров постановки задач без учителя.

# Задача кластеризации

Первый пример — задача кластеризации. Дано множество объектов. Необходимо найти группы похожих объектов. Есть две основные проблемы: не известно количество кластеров и не известны истинные кластеры, которые нужно выделять. Поэтому задача решается очень тяжело — здесь невозможно оценить качество решения. Этим и отличается задача классификации — там тоже нужно делить объекты на группы, но в классификации группы, а точнее классы, фиксированы, и известны примеры объектов из разных групп.



Примеры задач кластеризации:

- Сегментация пользователей (интернет-магазина или оператора связи)
- Поиск схожих пользователей в социальных сетях
- Поиск генов с похожими профилями экспрессии

# Поиск аномалий

Третий пример задачи обучения без учителя — поиск аномалий. Необходимо обнаружить, что данный объект не похож на все остальные, то есть является аномальным. При обучении есть примеры только обычных, не аномальных, объектов. А примеров аномальных объектов либо нет вообще, либо настолько мало, что невозможно воспользоваться классическими методами обучения с учителем (методами бинарной классификации). При этом задача очень важная. Например, к такому типу задач относится:

- Определение поломки в системах самолета (по показателям сотен датчиков)
- Определение поломки интернет—сайта
- Выявление проблем в модели машинного обучения. Все упомянутые задачи не будут обсуждаться в рамках данного курса. Им будет посвящен следующий курс — «Поиск структуры в данных».

# Признаки в машинном обучении

Существует несколько классов, или типов признаков. И у всех свои особенности — их нужно по-разному обрабатывать и по-разному учитывать в алгоритмах машинного обучения. В данном разделе будет обсуждаться используемая терминология, о самих же особенностях речь пойдет в следующих уроках. Признаки описывают объект в доступной и понятной для компьютера форме. Множество значений j-го признака будет обозначаться $D_j$ .

## Бинарные признаки

Первый тип признаков — бинарные признаки. Они принимают два значения: $D_j$ = {0, 1}. К таковым относятся:

- Выше ли доход клиента среднего дохода по городу?
- Цвет фрукта — зеленый?

Если ответ на вопрос да — признак полагается равным 1, если ответ на вопрос нет — то равным 0

# Признаки в машинном обучении

## Вещественные признаки

Более сложный класс признаков — вещественные признаки. В этом случае $D_j = \mathbb{R}$
Примерами таких признаков являются:
- Возраст
- Площадь квартиры
- Количество звонков в call-центр

Множество значений последнего указанного признака, строго говоря, является множеством натуральных чисел, а не, но такие признаки тоже считают вещественными.

## Категориальные признаки

Следующий класс признаков — категориальные признаки. В этом случае $D_j$ — неупорядоченное множество. Отличительная особенность категориальных признаков — невозможность сравнения «больше-меньше» значений признака. К таковым признакам относятся:
- Цвет глаз
- Город
- Образование (В некоторых задачах может быть введен осмысленный порядок)

Категориальные признаки очень трудны в обращении — до сих пор появляются способы учета этих признаков в тех или иных методах машинного обучения.

# Признаки в машинном обучении

## Множествозначные признаки

Множествозначный признак — это такой признак, значением которого на объекте является подмножество некоторого множества. Пример:

- Какие фильмы посмотрел пользователь
- Какие слова входят в текст

## Распределение признака

Далее речь пойдет о проблемах, с которыми можно столкнуться при работе с признаками. Первая из них — существование выбросов. Выбросом называется такой объект, значение признака на котором отличается от значения признака на большинстве объектов.

Наличие выбросов представляет сложность для алгоритмов машинного обучения, которые будут пытаться учесть и их тоже. Поскольку выбросы описываются совершенно другим законом, чем основное множество объектов, выбросы обычно исключают из данных, чтобы не мешать алгоритму машинного обучения искать закономерности в данных. Проблема может быть и в том, как распределен признак. Не всегда признак имеет такое распределение, которое позволяет ответить на требуемый вопрос. Например, может быть слишком мало данных о клиентах из небольшого города, так как собрать достаточную статистику не представлялось возможным.