Искусственный Интеллект

Лекция 2: Основы машинного обучения

Мартынюк Полина Антоновна

telegram: @PAMartynyuk

email: pa-martynyuk@yandex.ru

Искусственный интеллект : подходы и методы

Искусственный интеллект, ИИ (Artificial Intelligence, AI)

Искусственный интеллект - это широкая область информатики, которая стремится создать компьютерные системы и программы, способные выполнять задачи, требующие интеллектуальных способностей человека:

- распознавание образов
- понимание естественного языка
- принятие решений
- и другие…

ИИ включает в себя различные подходы и методы, включая машинное обучение, нейронные сети и глубокое обучение.

Машинное обучение

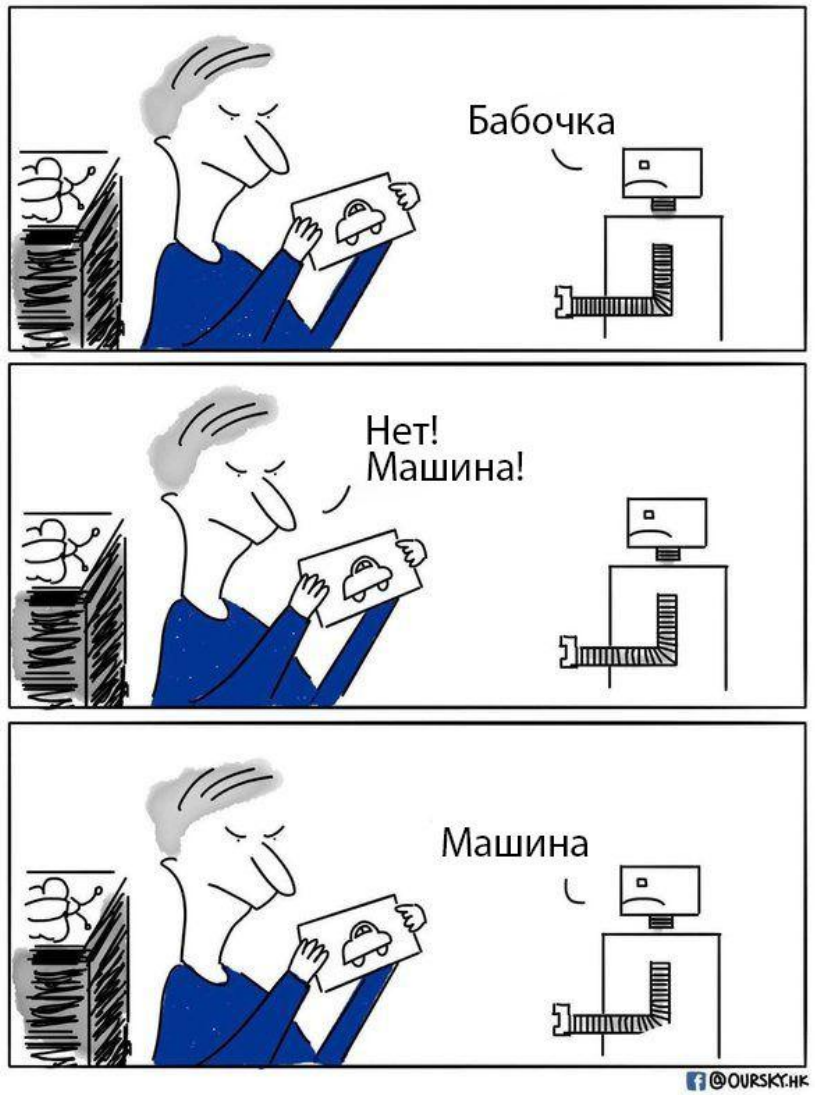
Машинное обучение, МО
(Machine Learning, ML)

Машинное обучение - это конкретная подобласть искусственного интеллекта, которая фокусируется на создании алгоритмов и моделей, которые могут учиться и приспосабливаться к данным без явного программирования.

“Машинное обучение - это область обучения, которая дает компьютерам возможность учиться, не будучи явно запрограммированной.”

- Артур Самюэль

(специалист в области информатики, преподаватель университета, исследователь искусственного интеллекта)



Машинное обучение

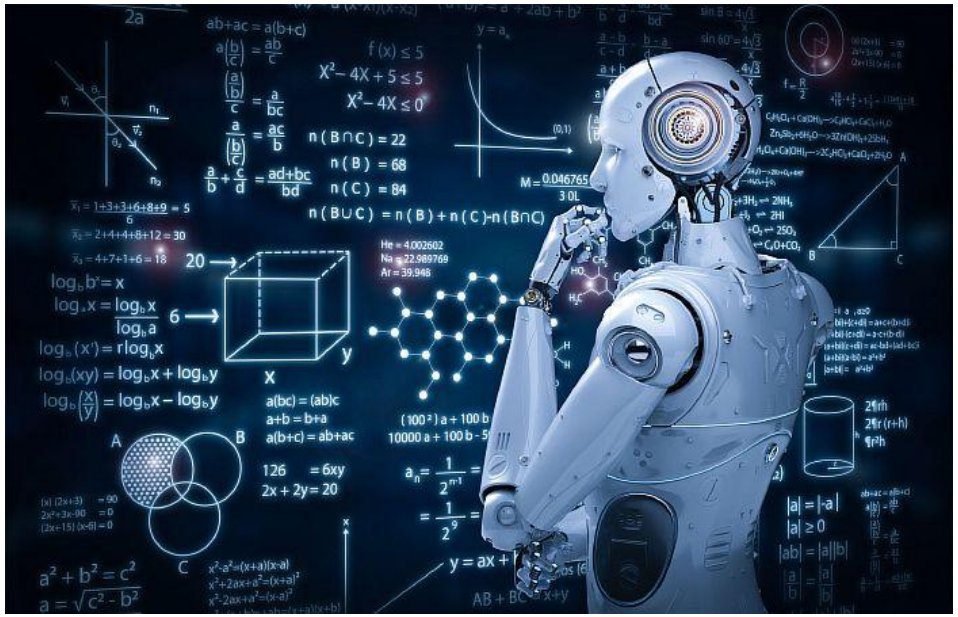
МО используются различные методы, включая нейронные сети и другие статистические и математические техники, чтобы научить компьютеры решать конкретные задачи, например, классификацию изображений или прогнозирование.

Старый подход

```
function register()
{
    if (!empty($_POST)) {
        $msg = '';
        if ($_POST['user_name']) {
            if ($_POST['user_password_new']) {
                if ($_POST['user_password_new'] === $_POST['user_password_repeat']) {
                    if (strlen($_POST['user_password_new']) > 5) {
                        if (strlen($_POST['user_name']) < 65 && strlen($_POST['user_name']) > 1) {
                            if (preg_match('/^[a-z\d]{2,64}$/i', $_POST['user_name'])) {
                                $user = read_user($_POST['user_name']);
                                if (!isset($user['user_name'])) {
                                    if ($_POST['user_email']) {
                                        if (strlen($_POST['user_email']) < 65) {
                                            if (filter_var($_POST['user_email'], FILTER_VALIDATE_EMAIL)) {
                                                create_user();
                                                $_SESSION['msg'] = 'You are now registered so please login';
                                                header('Location: ' . $_SERVER['PHP_SELF']);
                                                exit();
                                            } else $msg = 'You must provide a valid email address';
                                        } else $msg = 'Email must be less than 64 characters';
                                    } else $msg = 'Email cannot be empty';
                                } else $msg = 'Username already exists';
                            } else $msg = 'Username must be only a-z, A-Z, 0-9';
                        } else $msg = 'Username must be between 2 and 64 characters';
                    } else $msg = 'Password must be at least 6 characters';
                } else $msg = 'Passwords do not match';
            } else $msg = 'Empty Password';
        } else $msg = 'Empty Username';
        $_SESSION['msg'] = $msg;
    }
    return register_form();
}
```

«Используем миллион правил, которые учитывают все особенности задачи»

Новый подход



«Зальём обучающийся алгоритм данными и он сам найдёт в них закономерности»

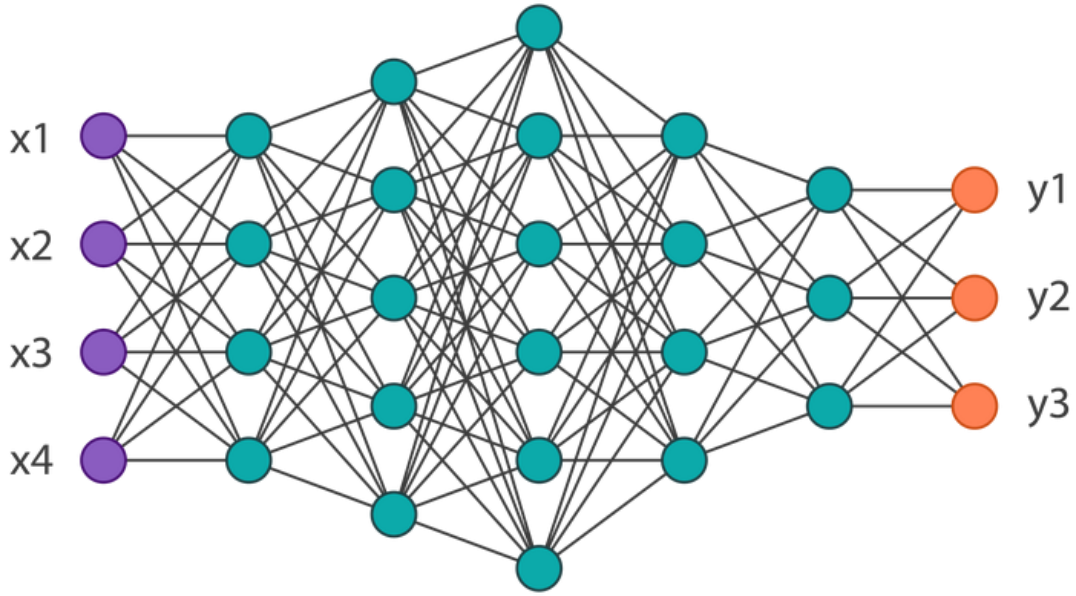
Нейронные сети

Нейронные сети
Artificial Neural networks, ANN

Нейронные сети - это класс моделей машинного обучения, вдохновленных структурой и функцией нейронов в человеческом мозге.

Они состоят из искусственных нейронов, объединенных в слои, и используются для обработки данных и извлечения признаков.

Нейронные сети могут быть использованы в различных задачах, от распознавания образов до обработки текста и много чего еще.



Глубокое обучение

Глубокое обучение
Deep Learning, DL

Глубокое обучение - это подраздел машинного обучения, который сосредотачивается на использовании глубоких нейронных сетей с множеством слоев для решения сложных задач.

Глубокое обучение обычно требует большого количества данных и вычислительных ресурсов, но может достичь выдающихся результатов в различных областях применения.

ChatGPT – модель глубокого обучения

Виды машинного обучения

Виды машинного обучения

По степени привлечения учителя:

- Обучение **с учителем** (supervised learning)
  - **Классификация** (classification)
  - **Регрессия** (regression)
- Обучение **без учителя** (unsupervised learning)
  - **Кластеризация** (clustering)
- Обучение с частичным привлечением учителя (semi-supervised learning)

Основные термины и понятия

Данные (Data)

Данные - информация, используемая для обучения модели.

Она может включать в себя:

- **Признаки** (features, обозначение - **X**)
- **Целевую переменную** (target, обозначение - **Y**) (в задачах обучения с учителем).

Основные термины и понятия

Данные (Data)

Данные - информация, используемая для обучения модели.

Она может включать в себя:

- **Признаки** (features, обозначение - **X**)
- **Целевую переменную** (target, обозначение - **Y**) (в задачах обучения с учителем).

Площадь квартиры (X1, feature)	Количество комнат (X2, feature)
73	2
150	4
34	1
101	3

Данные, содержащие только признаки

Площадь квартиры (X1, feature)	Количество комнат (X2, feature)	Цена квартиры (Y, label, target)
73	2	5.5
150	4	15.8
34	1	4.2
101	3	10.9

Данные, содержащие признаки и целевую переменную

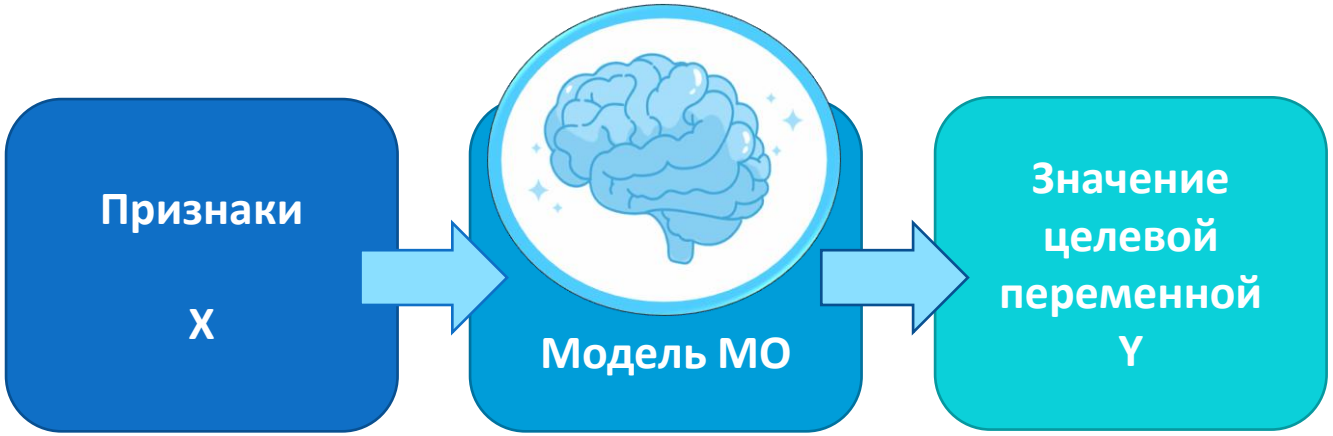
Основные термины и понятия

Данные (Data)

Данные - информация, используемая для обучения модели.

Она может включать в себя:

- **Признаки** (features, обозначение - **X**)
- **Целевую переменную** (target, обозначение - **Y**) (в задачах обучения с учителем).



Основные термины и понятия

Данные (Data)

Данные - информация, используемая для обучения модели.

Она может включать в себя:

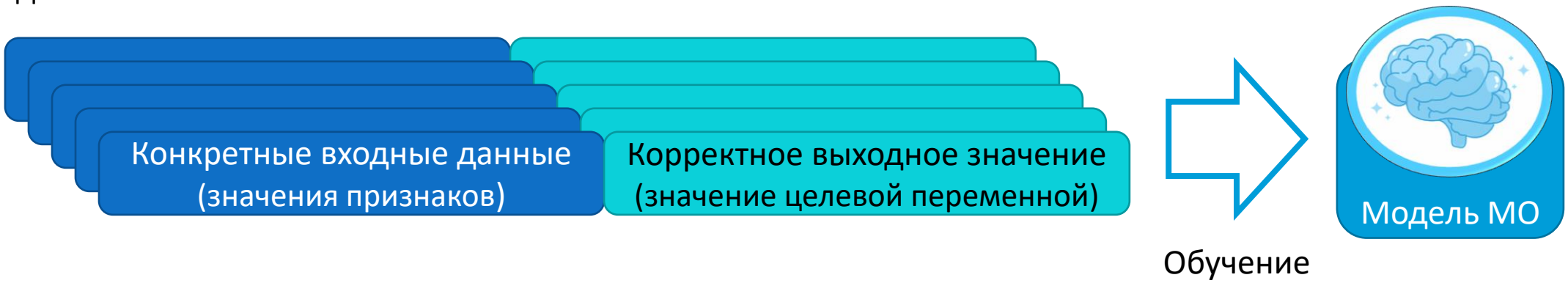
- **Признаки** (features, обозначение - **X**)
- **Целевую переменную** (target, обозначение - **Y**) (в задачах обучения с учителем).

Виды машинного обучения: Обучение с учителем

Обучение с учителем («Делай вот так»)

Модель обучается на основе **размеченных данных (наборе данных, содержащих явно заданное значение целевой переменной)**.

В этом типе задачи каждый обучающий пример представляет собой пару: **входные данные**(признаки) и соответствующее этим данным **значение выходной метки** (целевая переменная).
Задача модели состоит в том, чтобы научиться предсказывать выходные метки на основе входных данных.



Обучение с учителем

Обучение с учителем, основные моменты

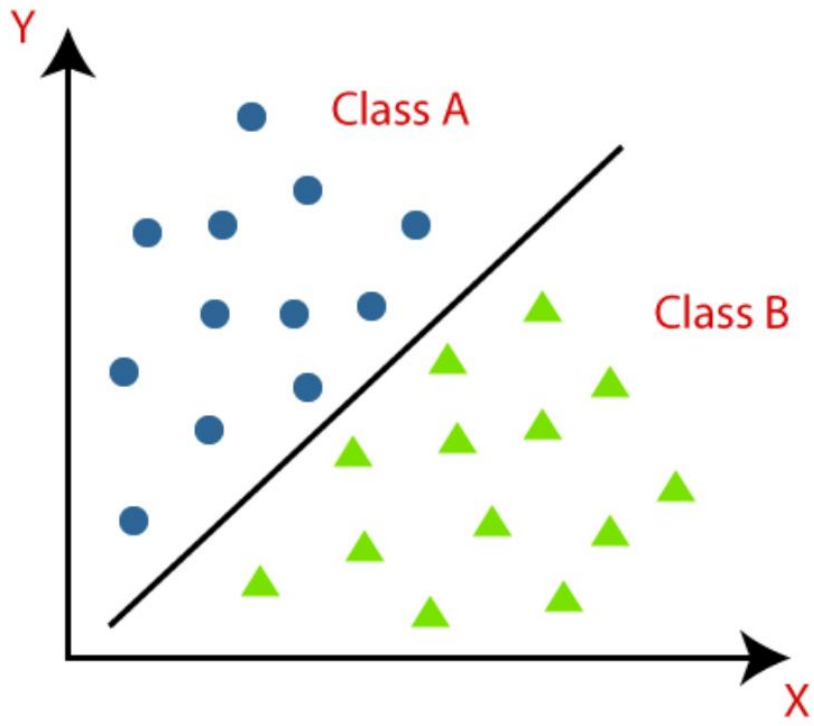
- **Размеченные данные**: Для обучения модели требуются данные, для которых известны правильные ответы или метки.
- **Целевая переменная**: Обучающие примеры включают в себя целевую переменную, которую модель должна предсказать.
- **Обучение на основе примеров**: Модель обучается на основе обучающих примеров, применяя различные алгоритмы и методы для поиска связей и закономерностей между входными данными и целевой переменной.
- **Цель - обобщение**: Основная цель обучения с учителем - научить модель обобщать знания с обучающих данных на новые, ранее не виденные данные. Модель должна способно предсказывать метки для новых примеров.

Задачи обучения с учителем: Классификация

Классификация (Classification)

- **предсказание категории или класса** для входных данных. Выходная величина принимает значение одного из заранее определённых классов.

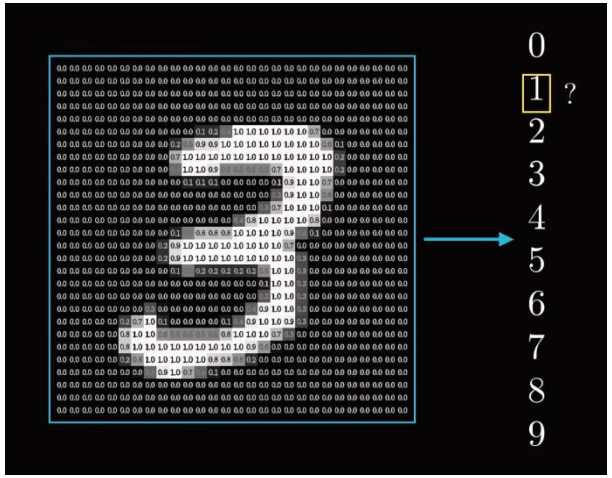
- **Бинарная классификация**: В задачах бинарной классификации метка может принимать одно из двух значений, например, 0 или 1, Да или Нет, Положительный или Отрицательный и т.д.
- **Многоклассовая классификация**: В многоклассовой классификации метка может относиться к одному из нескольких классов или категорий. Например, при классификации изображений метка может указывать на тип объекта (кошка, собака, автомобиль и так далее)

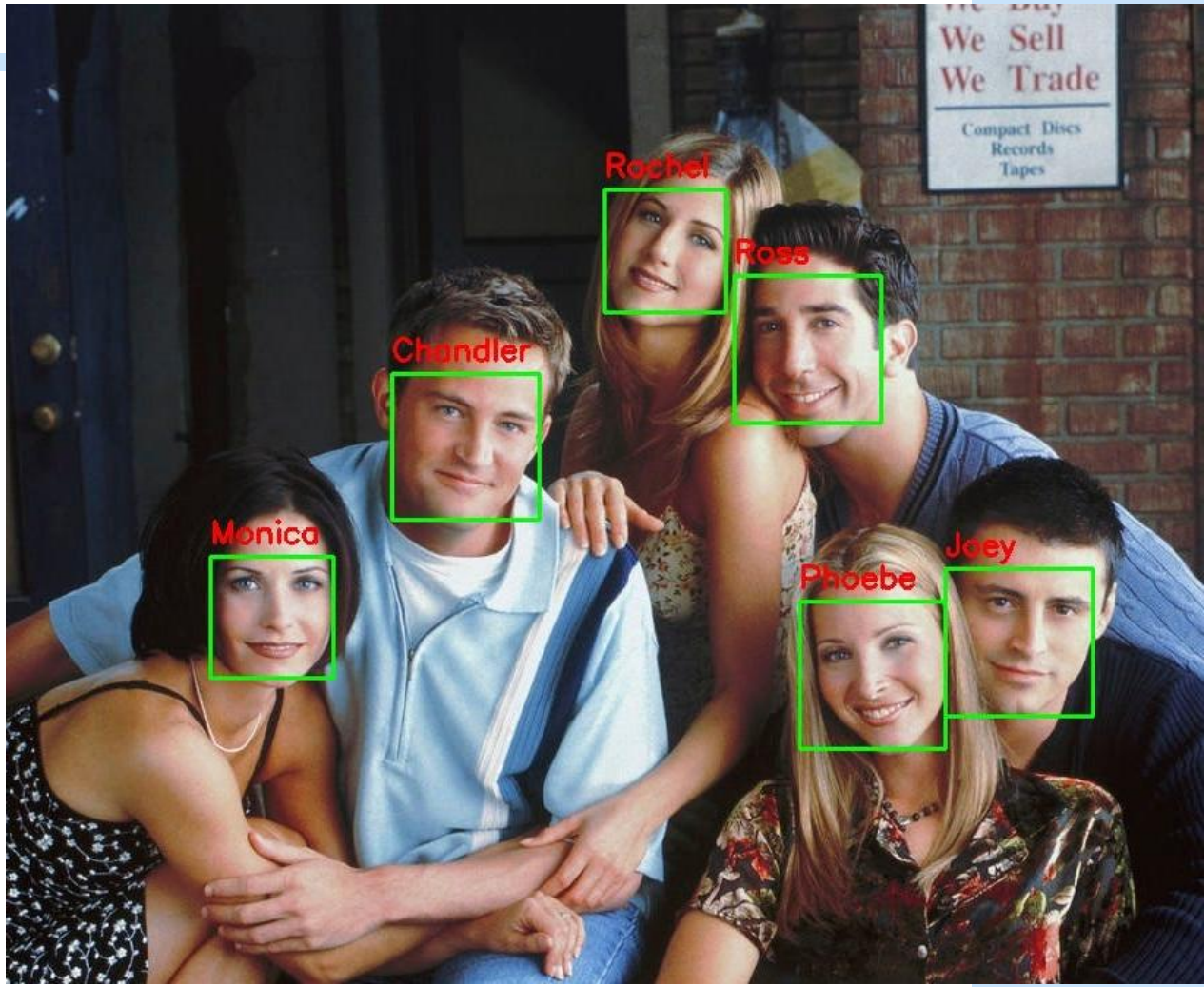


Задачи обучения с учителем: Классификация

Классификация (Classification)

- **Классификация изображений** - распознавание объектов на изображениях. Примеры включают классификацию животных на фотографиях, автомобилей по их маркам и моделям, распознавание лиц и др.

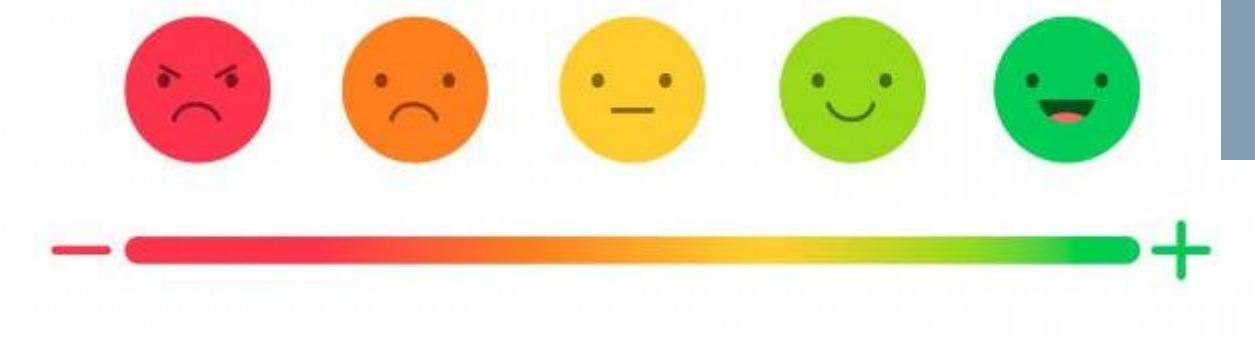


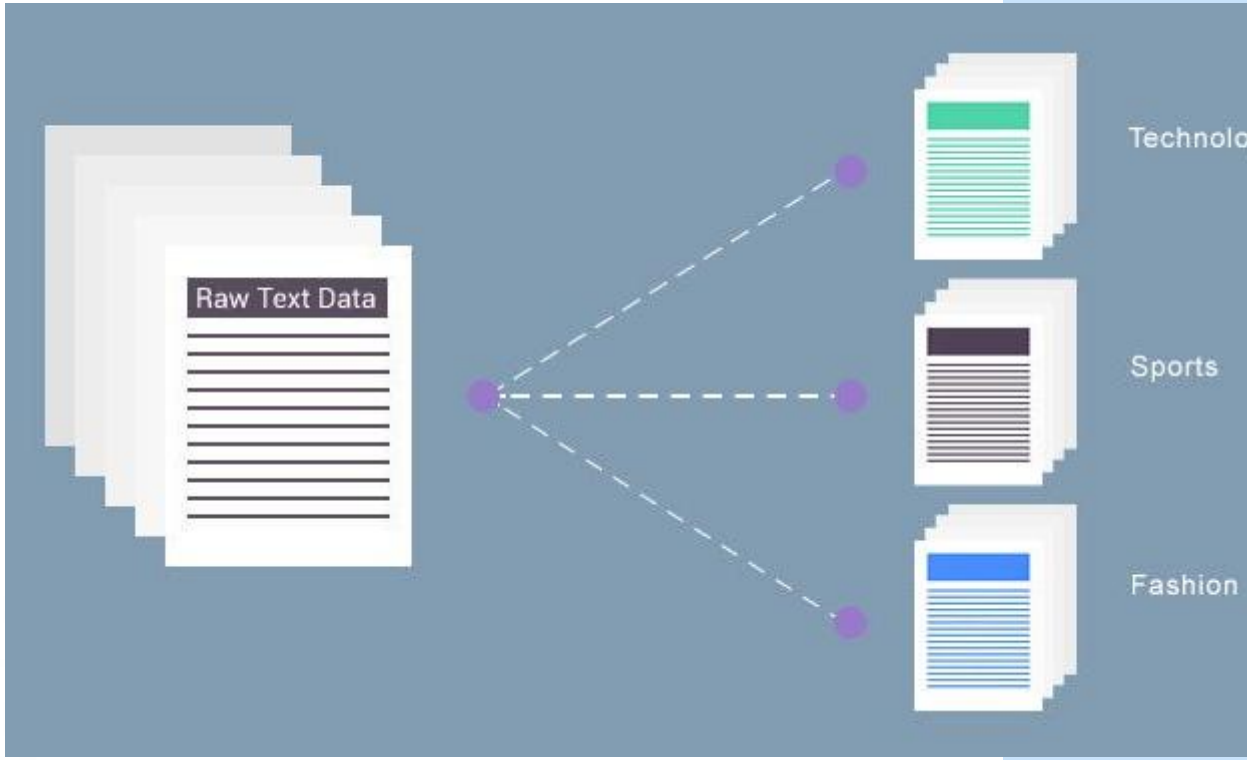


Задачи обучения с учителем: Классификация

Классификация (Classification)

- **Классификация текста** - определение категории или темы текста. Например, классификация новостных статей по темам (политика, спорт, наука), определение тональности текста (позитивная, негативная, нейтральная).





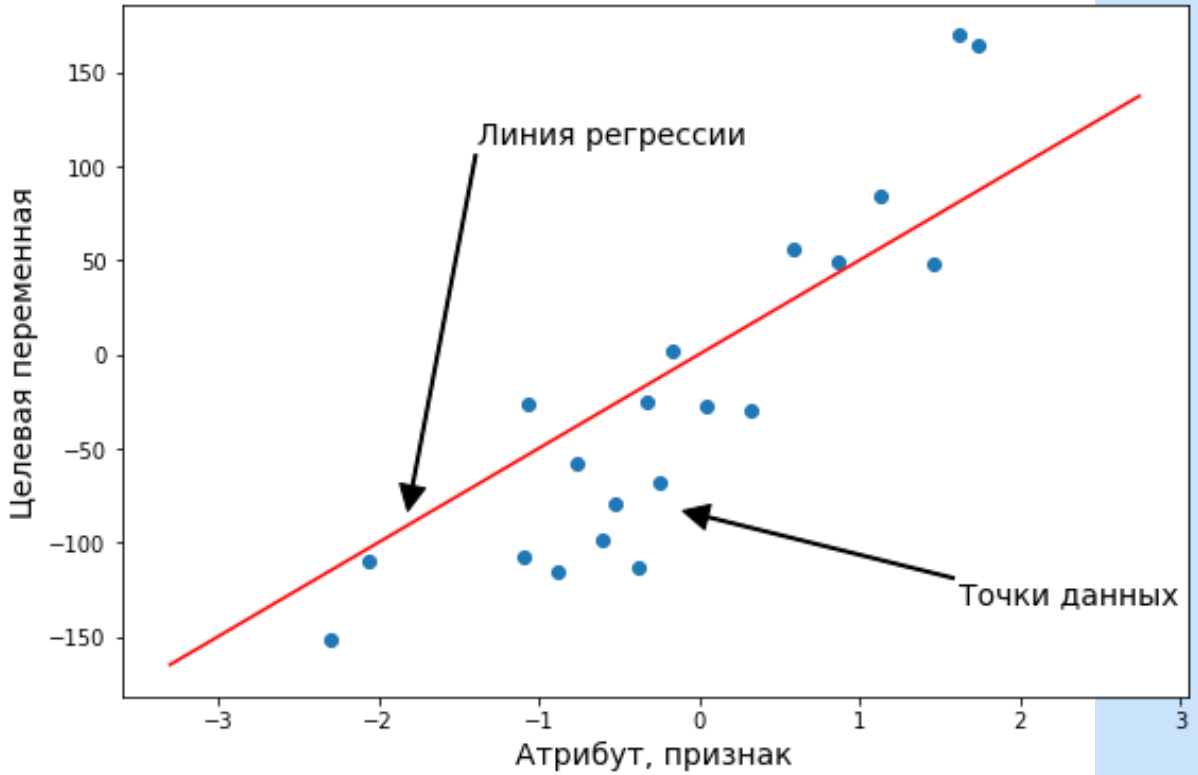
Задачи обучения с учителем: Регрессия

Регрессия (Regression)

- **предсказание числового значения** по входным данным.

В задачах регрессии метка обычно представляет собой числовое значение, которое модель пытается предсказать.

Например, при прогнозировании цены дома метка может быть числом, представляющим стоимость.



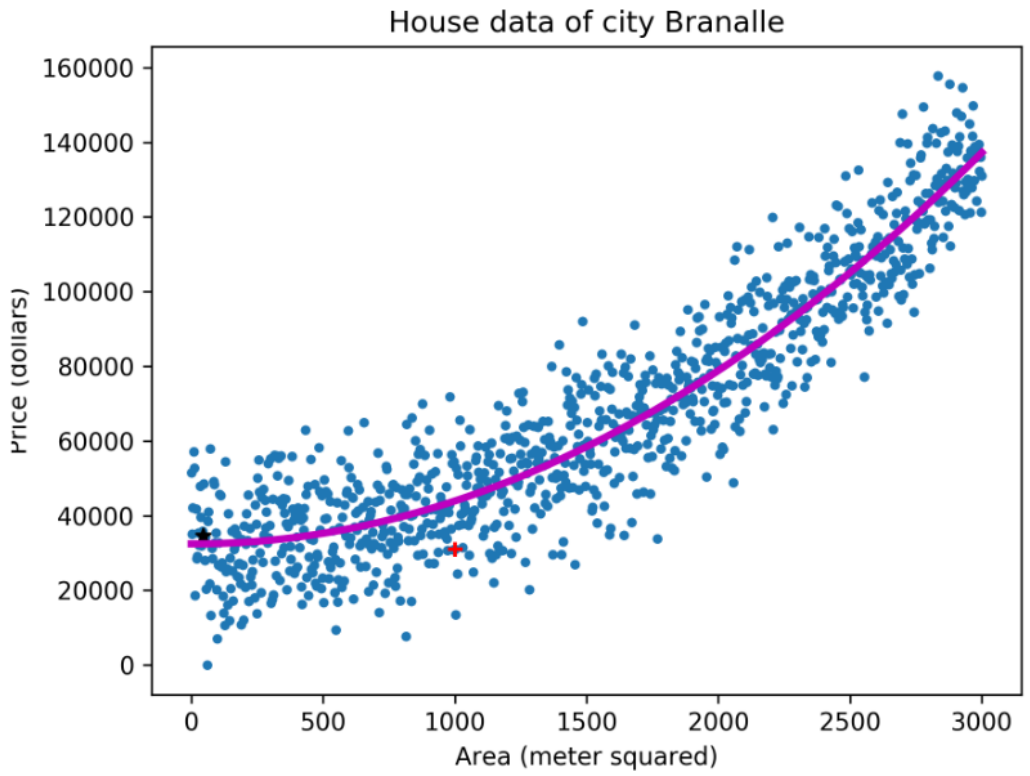
Выходная величина расположена на отрезке значений

Задачи обучения с учителем: Регрессия

Регрессия (Regression)

- **Прогнозирование цены недвижимости** - определение стоимости домов или квартир на основе их характеристик, таких как площадь, количество комнат, район и другие факторы.

Площадь квартиры (X1, feature)	Количество комнат (X2, feature)	Цена квартиры (Y, label, target)
73	2	5.5
150	4	15.8
34	1	4.2
101	3	10.9



Целевое значение

Классификация (Classification)

- **Целевое значение категориальное** – метка дискретных категорий или классов. Например, в задаче классификации изображений целевая переменная может указывать на категорию объекта на изображении (например, "кошка", "собака", "автомобиль").

Регрессия (Regression)

- **Целевое значение численное –** конкретное числовое значение. Модель стремится предсказать точное численное значение.
  Примеры численных целевых переменных: цена товара, температура, доход, количество продаж и так далее.

Виды машинного обучения: Обучение без учителя

Обучение без учителя («Разберись сам»)

Модель обучается на основе **неразмеченных данных (в наборе данных отсутствует значение целевой переменной)**.

Модель пытается извлечь скрытую структуру или закономерности из входных данных **без предварительной информации о правильных ответах.**
Этот тип задач машинного обучения позволяет модели самостоятельно выявлять закономерности, группировать данные и создавать представления о данных.

Обучение без учителя

Обучение без учителя, основные моменты

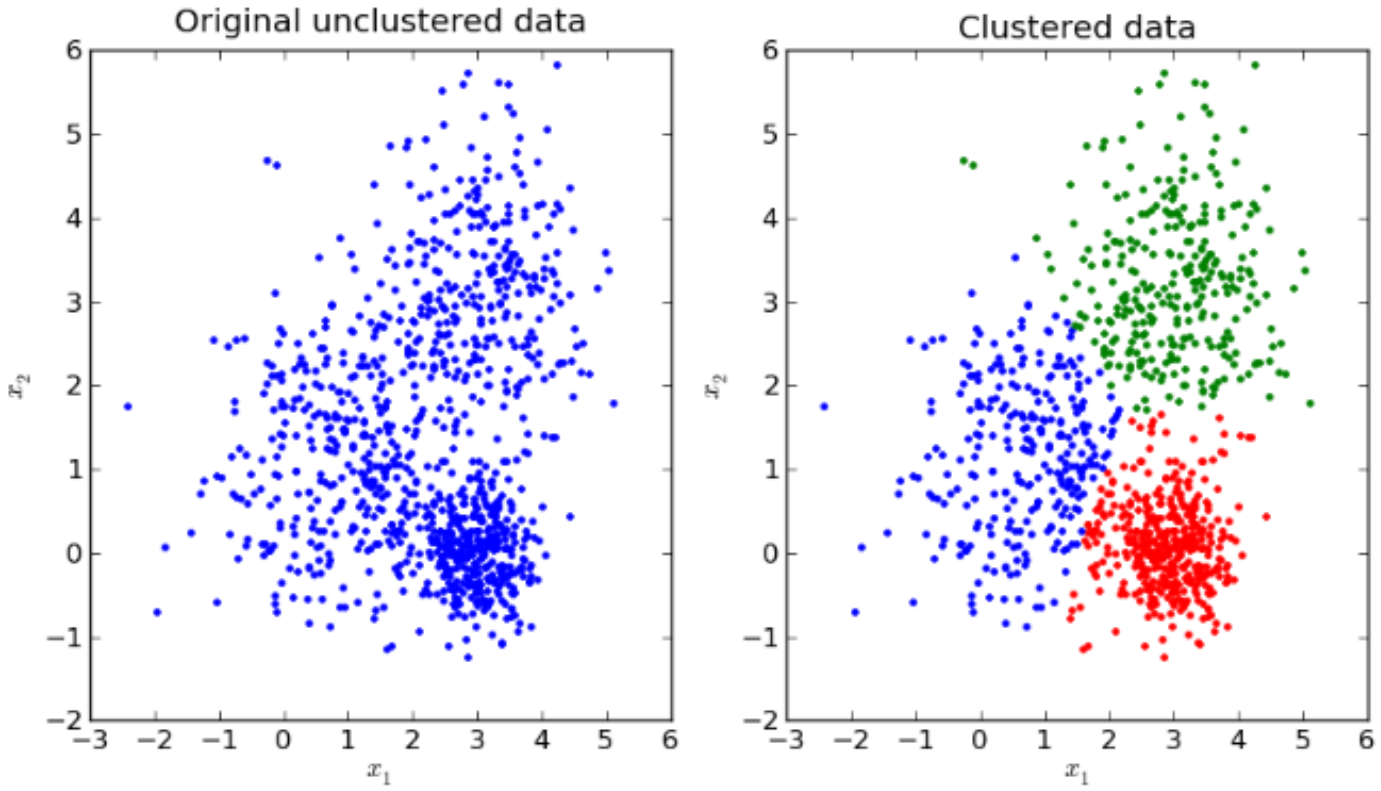
- **Отсутствие целевой переменной** в обучающих данных. В задачах обучения с учителем у нас есть целевая переменная, которую модель пытается предсказать. В обучении без учителя целевая переменная отсутствует.
- **Группировка и структура данных:** Обучение без учителя ориентировано на поиск структуры или закономерностей в данных. Модель стремится выделить группы схожих объектов или другие характеристики данных, такие как кластеры или сниженная размерность.
- **Цель - извлечение информации**: Главной целью обучения без учителя является извлечение полезной информации из данных. Модель создает представление данных, которое может помочь в анализе, визуализации, сегментации или других задачах.
- **Обучение на основе характеристик :** В обучении без учителя модель основывается на характеристиках (признаках - features) данных, чтобы выявить их структуру или закономерности.

Задачи обучения без учителя: Кластеризация

Кластеризация (Clustering)

Группировка объектов в **кластеры** (группы схожих объектов) на основе их сходства.

Кластеризация позволяет выявить внутренние структуры в данных и определить, какие объекты близки по своим характеристикам.



Задачи обучения без учителя: Кластеризация

Кластеризация (Clustering)

Классификация	Кластеризация
Цель: Главной целью классификации является присвоение объектов или данных одной из заранее определенных категорий или классов. В классификации у нас есть явные метки или метки классов, которые мы пытаемся предсказать для новых данных на основе характеристик объектов. **Целевая переменная**: В задачах классификации существует явная целевая переменная, которую модель пытается предсказать для новых данных. Целевая переменная состоит из классов или категорий.	**Цель**: Главной целью кластеризации является группировка схожих объектов или точек данных на основе их характеристик без заранее известных категорий или меток. В кластеризации мы пытаемся найти структуру в данных, выделяя группы или кластеры объектов, которые имеют схожие характеристики или свойства. **Целевая переменная**: В задачах кластеризации нет заданной заранее целевой переменной или меток классов. Модель самостоятельно определяет структуру кластеров в данных.

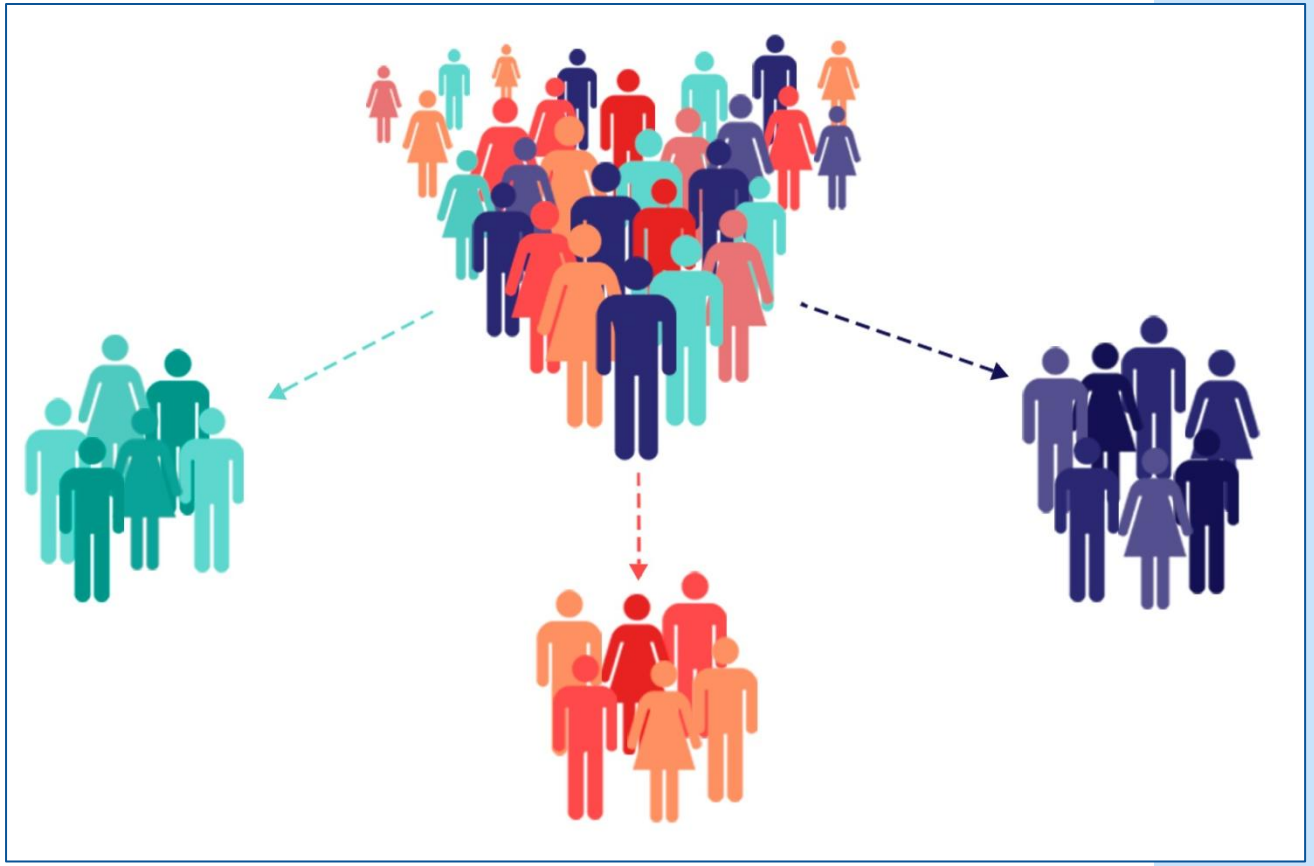
Задачи обучения без учителя: Кластеризация

Кластеризация (Clustering)

Сегментация клиентской базы

В маркетинге можно использовать кластеризацию для разделения клиентов на группы схожих потребителей.

Например, вы можете определить кластеры клиентов, которые предпочитают определенные категории товаров или имеют схожие покупательские привычки.

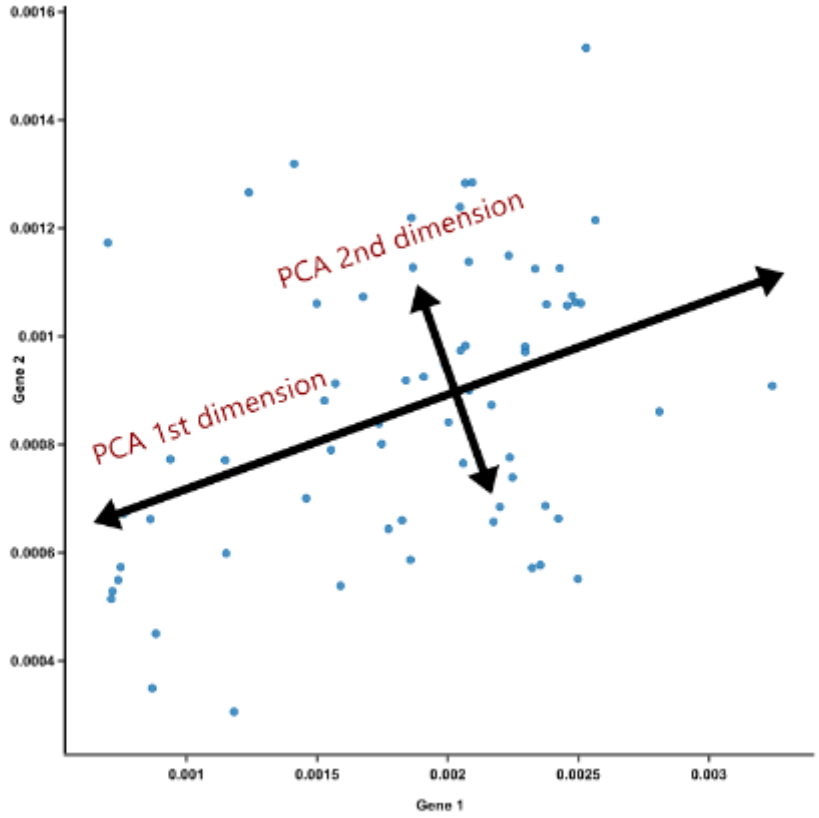


Задачи обучения без учителя: Кластеризация

Снижение размерности (Dimensionality Reduction)

Уменьшение размерности данных путем проекции на более низкоразмерное пространство, при этом стремятся сохранить максимальное количество информации.

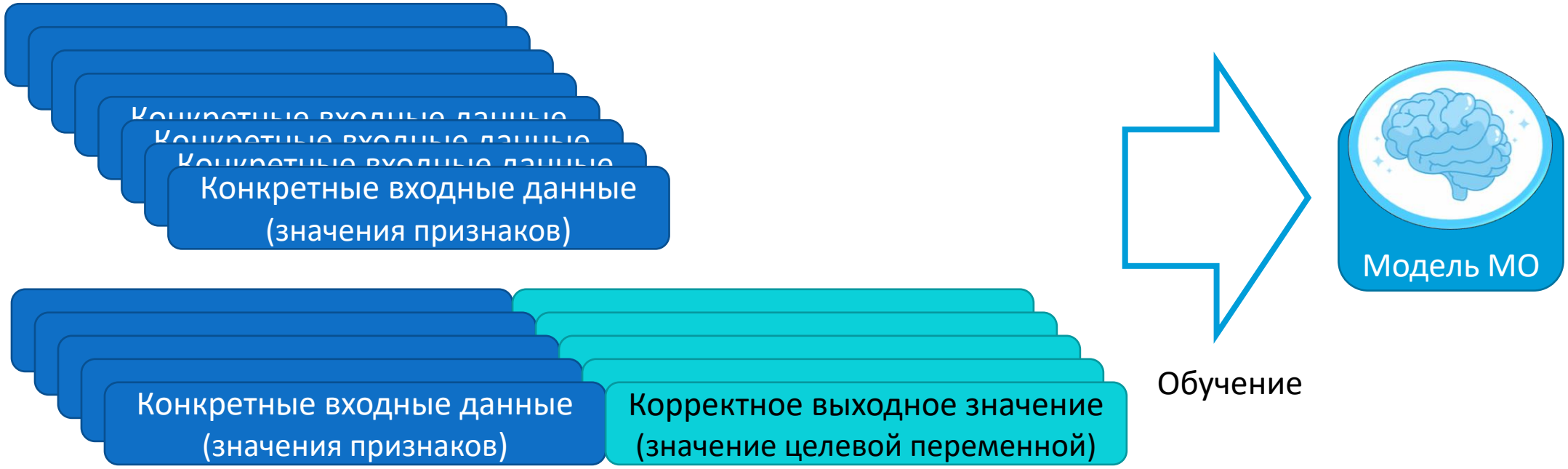
Примером является метод главных компонент (Principal Component Analysis, PCA).



Виды машинного обучения: Обучение с частичным привлечением учителя

Обучение с частичным привлечением учителя

Модель обучается на основе **частично размеченных данных.**



Обучение с частичным привлечением учителя

Обучение с частичным привлечением учителя, основные моменты

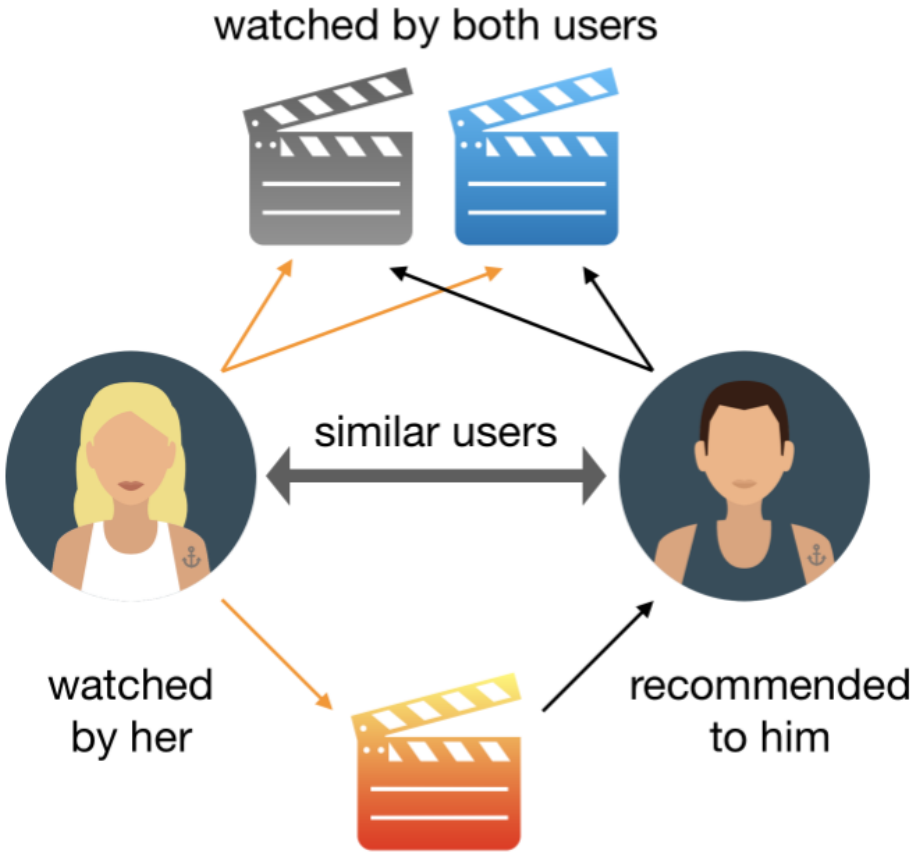
- **Использование размеченных данных**: Модель использует размеченные данные для обучения, чтобы научиться предсказывать целевую переменную.
- **Использование неразмеченных данных**: Дополнительно к размеченным данным, модель также использует немаркированные данные для извлечения структуры данных или для улучшения обобщения. Это помогает модели учить более общие закономерности и снижать риск переобучения.
- **Воссоздание разметки неразмеченных данных**: Существует ряд алгоритмов и методов, разработанных специально для обучения с частичным привлечением учителя. Некоторые из них используют методы активного обучения, которые позволяют модели запрашивать разметку для наиболее неопределенных или важных примеров.
- **Снижение нагрузки на разметку данных**: Этот метод позволяет снизить трудозатратность на разметку данных, поскольку требуется разметить только часть данных, а не всю выборку.

Другие виды машинного обучения: Рекомендательные системы

Рекомендательные системы

Класс программных и алгоритмических инструментов, которые анализируют данные о предпочтениях или поведении пользователей, чтобы предложить им релевантные и персонализированные рекомендации.

Они используются для предсказания, какие товары, услуги, контент или информация могут быть интересны конкретному пользователю.

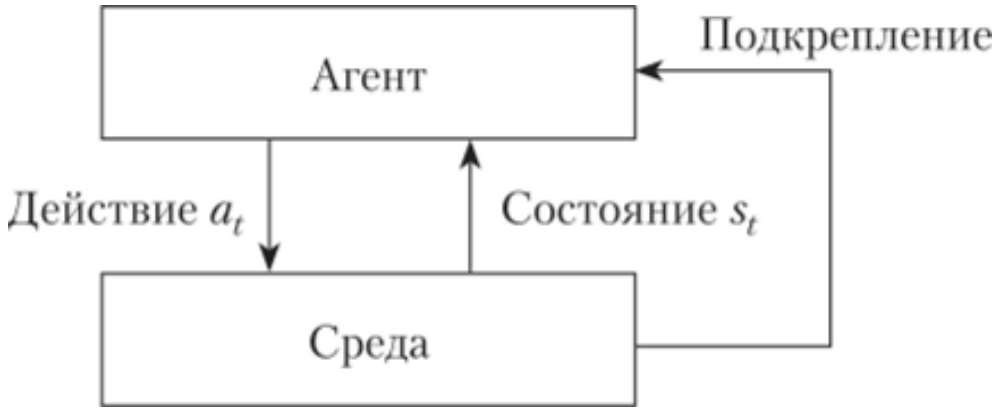


Другие виды машинного обучения: Обучение с подкреплением

Обучение с подкреплением (Reinforcement Learning, RL)

Подход в машинном обучении, в котором агент (или программа) учится принимать последовательность решений, чтобы максимизировать некоторую численную награду (или минимизировать некоторый штраф) в определенной среде.

Этот подход аналогичен обучению, которое происходит у человека или животного, когда они взаимодействуют с окружающей средой, принимают решения и адаптируются на основе получаемого опыта.



Другие виды машинного обучения: Обучение с подкреплением

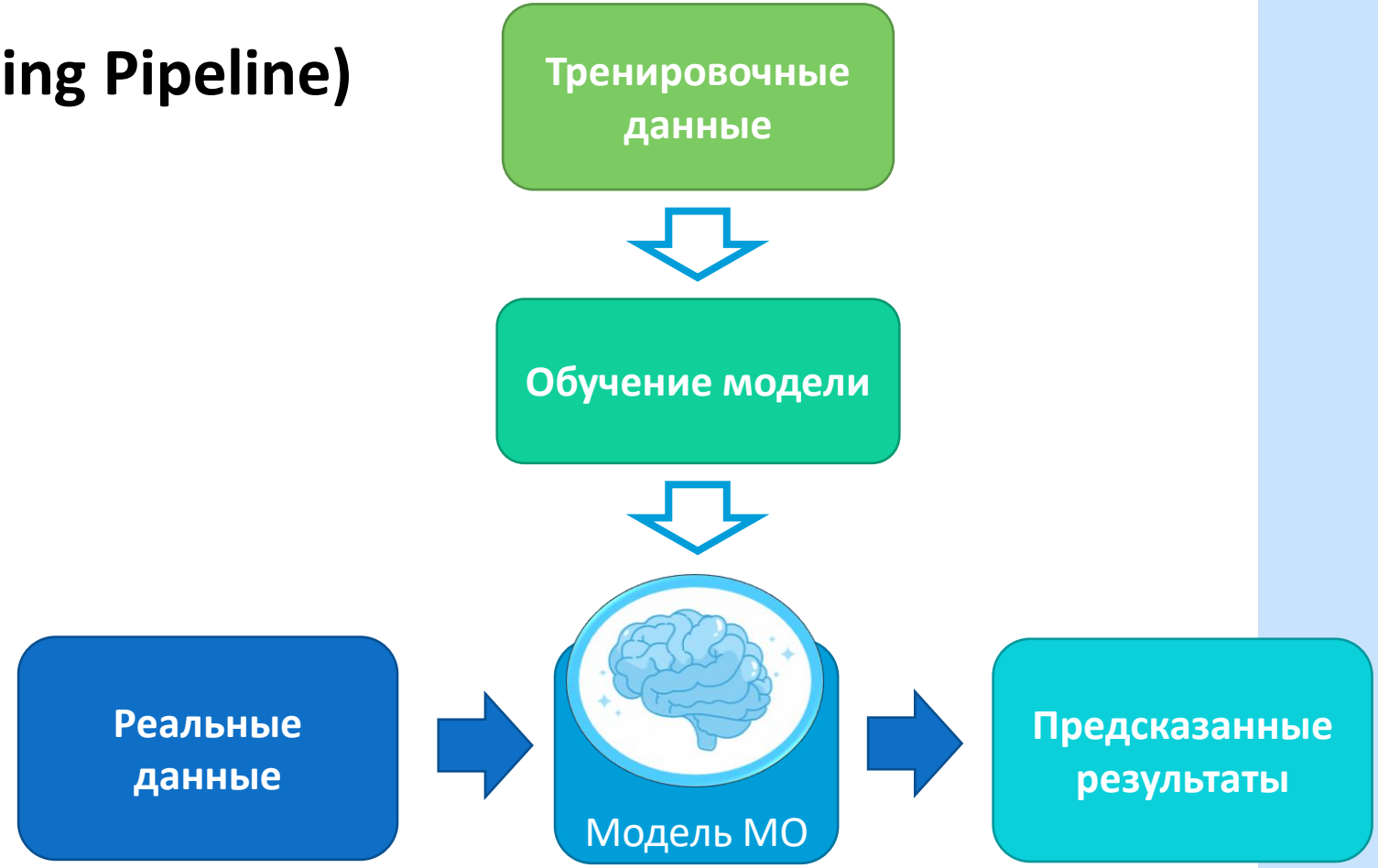
Обучение с подкреплением (Reinforcement Learning, RL)

- **Агент (Agent)**: Это сущность, которая принимает решения в среде.
- **Среда (Environment)**: Это окружающее пространство, в котором действует агент. Среда может быть физической (например, робот в реальном мире) или виртуальной (например, компьютерная игра).
- **Действия (Actions)**: Агент может выполнять определенные действия, которые воздействуют на среду.
- **Состояния (States)**: Состояния представляют текущее состояние среды и содержат информацию, необходимую для принятия решений.
- **Награда (Reward)**: Награда представляет собой численное значение, которое агент получает от среды после выполнения действия. Цель агента - максимизировать награду, то есть получить максимальное накопленное вознаграждение за выполнение последовательности действий.
- **Стратегия (Policy)**: Стратегия определяет, какие действия агент должен выбирать в каждом состоянии, чтобы максимизировать свою награду.
- **Цель обучения с подкреплением** - научить агента находить оптимальную стратегию, которая позволяет ему достигать максимальной награды в заданной среде.

Понятие пайплайна в МО

Пайплайн (Machine Learning Pipeline)

Это последовательность шагов и операций, которые необходимо выполнить для решения конкретной задачи машинного обучения



Понятие пайплайна в МО

Пайплайн (Machine Learning Pipeline)

Данные не всегда сразу могут быть использованы для обучения модели – как правило, требуется их **предобработка.**

Данные могут быть
нормализованы,
масштабированы,
устранены выбросы,
заполнены пропущенные значения и т.д.
Это шаг важен для того, чтобы данные были в правильном формате для обучения модели.

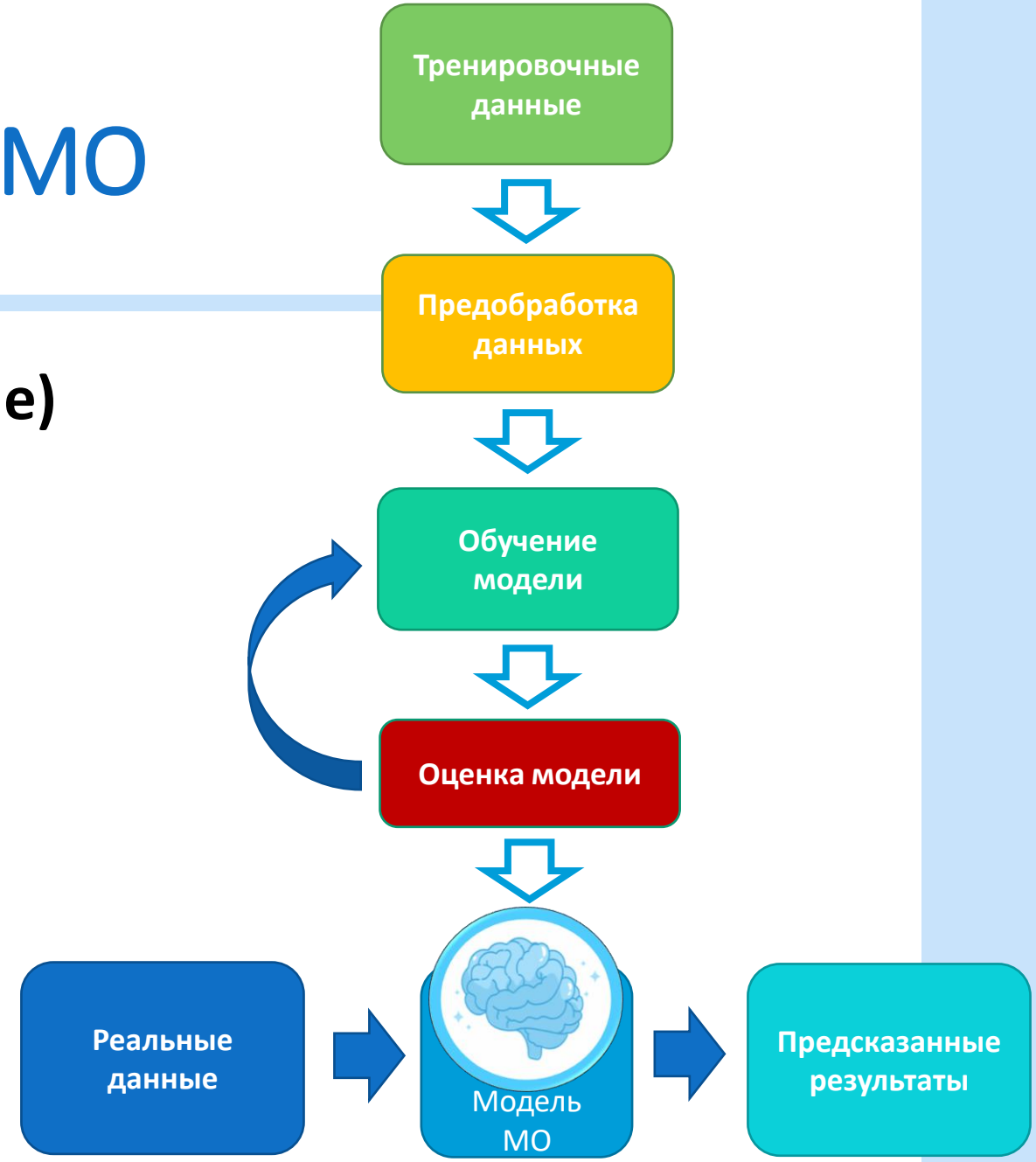
Понятие пайплайна в МО

Пайплайн (Machine Learning Pipeline)

Как понять, хорошо ли обучилась модель?

Модель тестируется на отложенной выборке для оценки ее производительности.

Здесь могут использоваться метрики, такие как точность (accuracy), среднеквадратичная ошибка (mean squared error), F1-мера и другие.

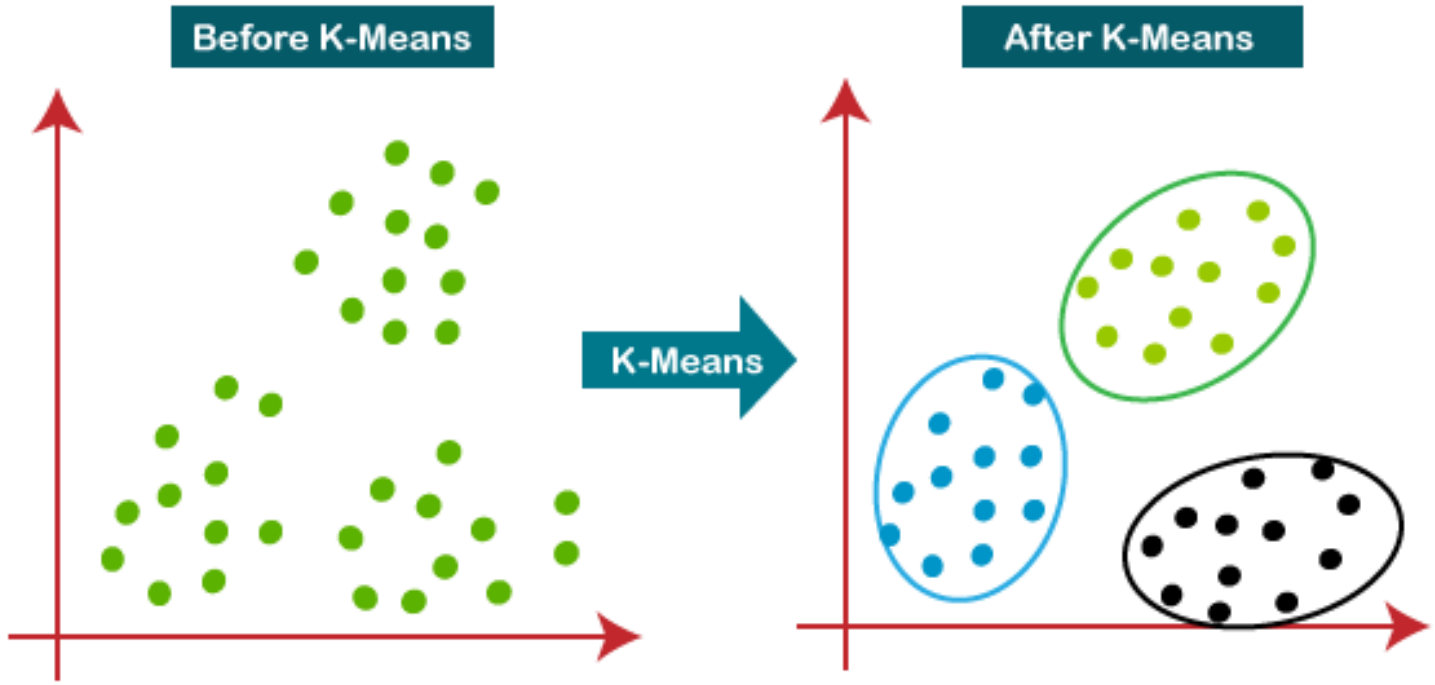


Кластеризация. Метод k-средних (k-means)

Метод k-средних
Простой подход разделение данных на К различных непересекающихся кластеров.

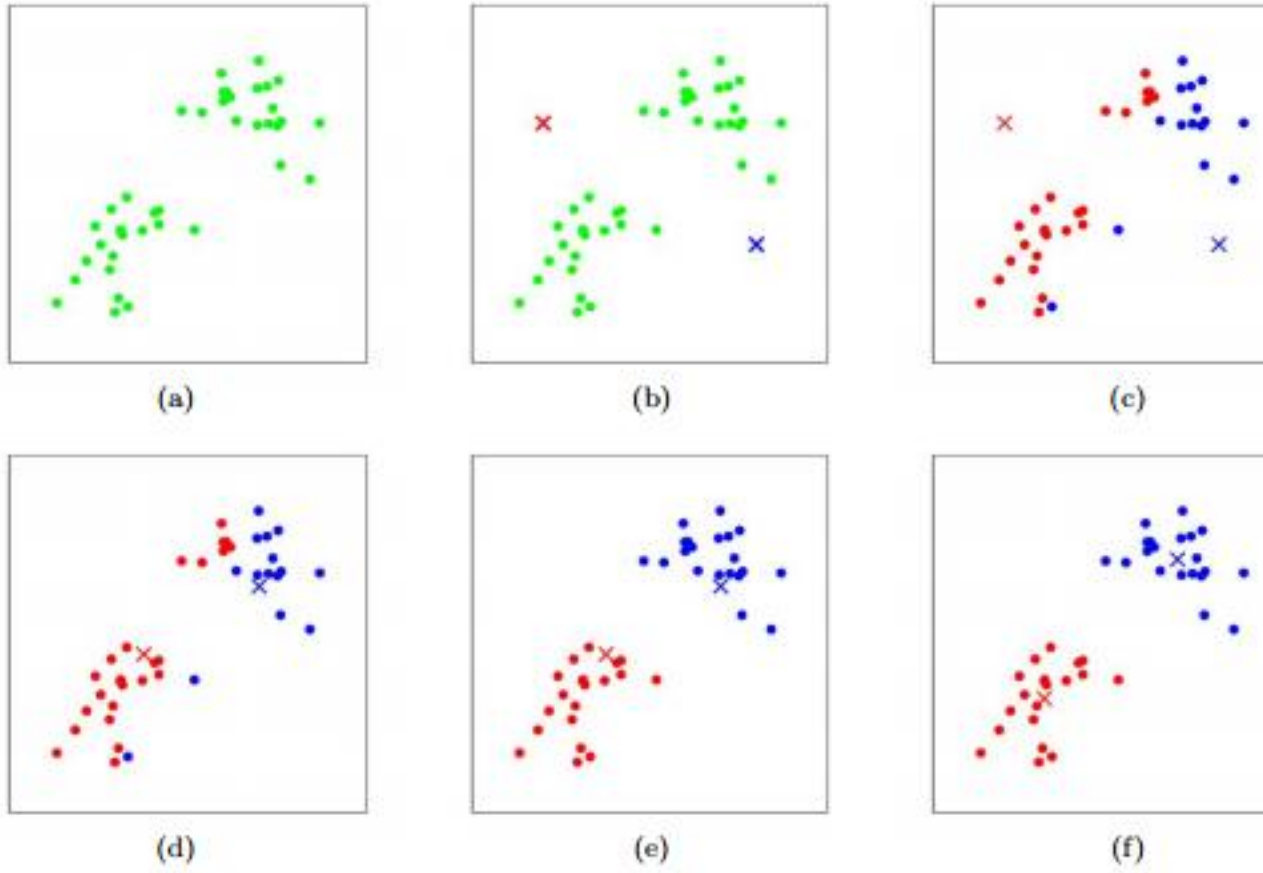
Алгоритм **кластеризации** в машинном обучении, который используется для группировки схожих объектов в заданное количество кластеров (групп).

Пример обучения **без учителя**, поскольку алгоритм самостоятельно находит структуру в данных без использования явных меток классов.



Кластеризация. Метод k-средних (k-means)

1. **Инициализация центроидов**: Сначала выбираются начальные положения центроидов - это будут точки, представляющие центры кластеров. Чаще всего центроиды выбираются случайным образом из набора данных или согласно каким-либо эвристическим правилам.
2. **Назначение объектов к кластерам**: Каждый объект данных назначается к ближайшему центроиду. Расстояние между объектами и центроидами измеряется, обычно, с использованием евклидова расстояния, но также могут использоваться и другие метрики.
3. **Пересчет центроидов**: После назначения объектов к кластерам пересчитываются центроиды для каждого кластера. Это делается путем вычисления среднего значения всех объектов, принадлежащих кластеру.
4. **Повторение шагов 2 и 3**: Эти шаги выполняются итеративно до тех пор, пока центроиды перестают изменяться или до достижения максимального числа итераций.
5. **Завершение алгоритма**: Когда алгоритм завершает свою работу, каждый объект данных находится в одном из кластеров, и центроиды представляют собой средние точки внутри кластеров.

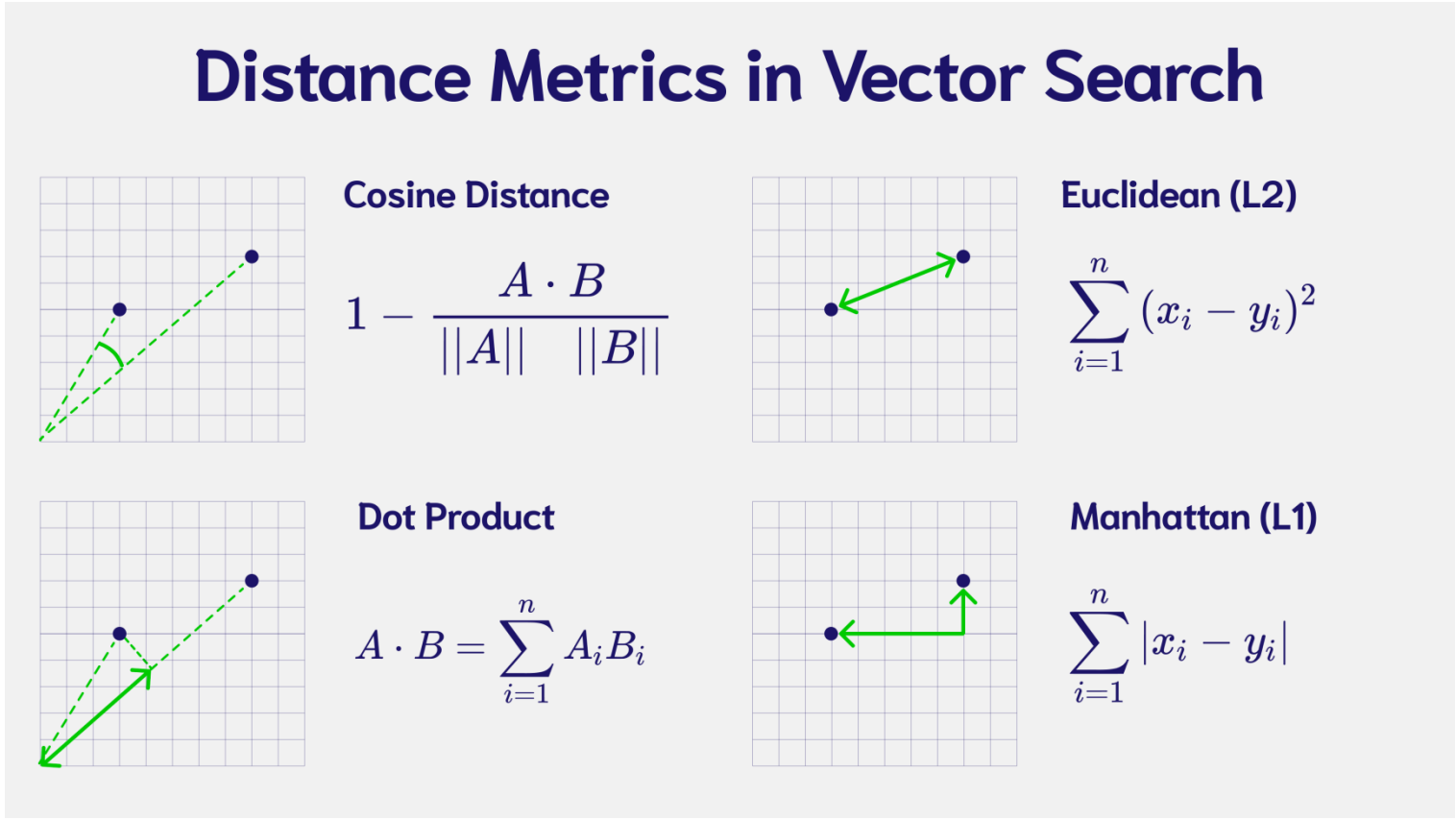


Кластеризация. Метод k-средних (k-means)

$X_1 = (x_{11}, x_{12}, \ldots, x_{1n})$

$X_2 = (x_{21}, x_{22}, \ldots, x_{2n})$

$Dist(X_1, X_2)$ - ?



Кластеризация. Метод k-средних (k-means)

Особенности	Достоинства	Недостатки
• Нужно знать число кластеров • Выбрать способы задать начальные кластеры • Каждое наблюдение только в одном кластере • Нужно несколько раз прогнать алгоритм	• Простота • Сходимость • Применимость к большим наборам данных	• Нужно знать число кластеров • Зависимость от начального задания кластеров • Чувствителен к выбросам • Плохо ведет себя при большой размерности данных • Не учитывает плотность и размеры кластеров

Спасибо за внимание!

Конец Лекции 2