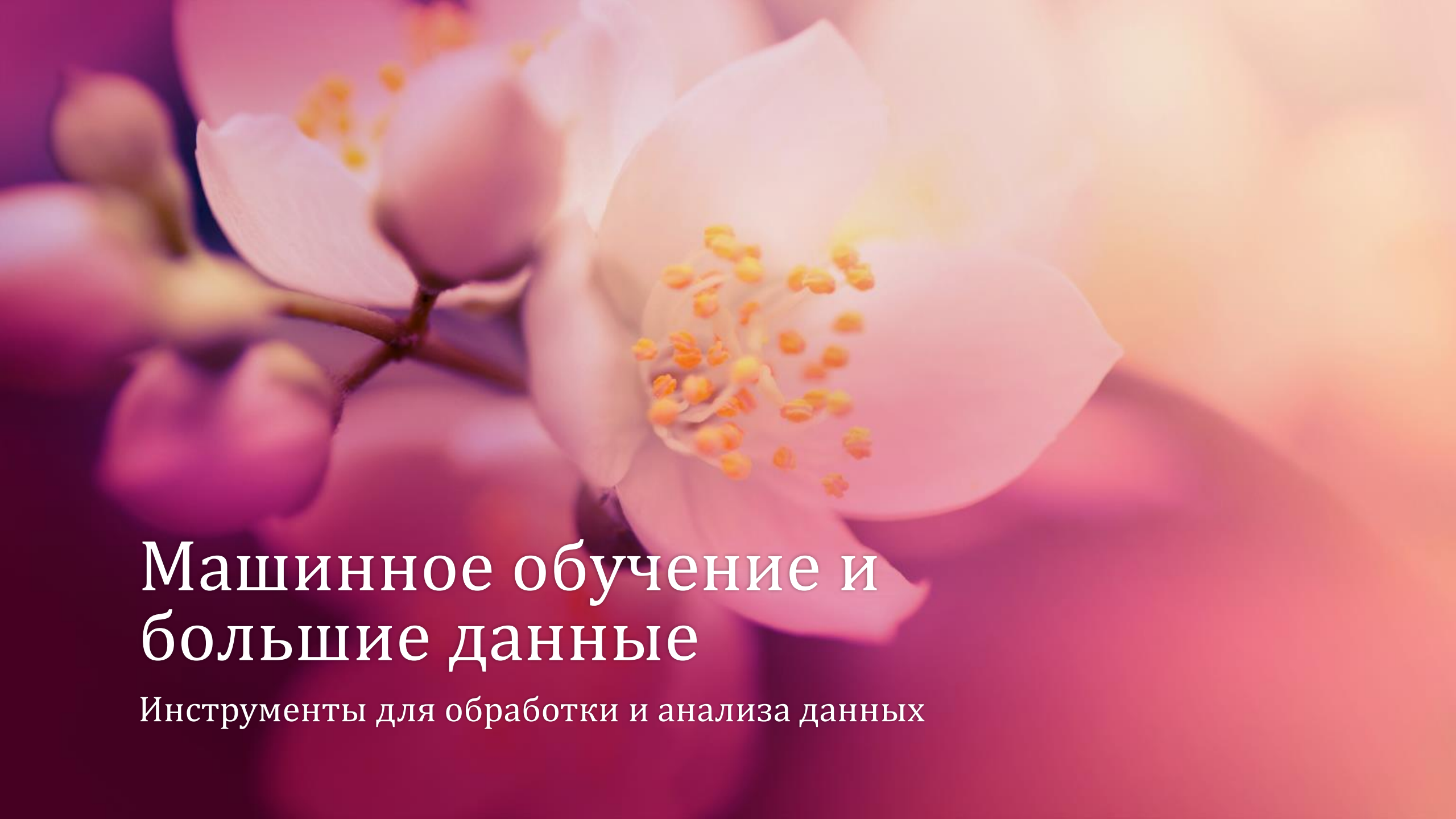


Машинное обучение и большие данные

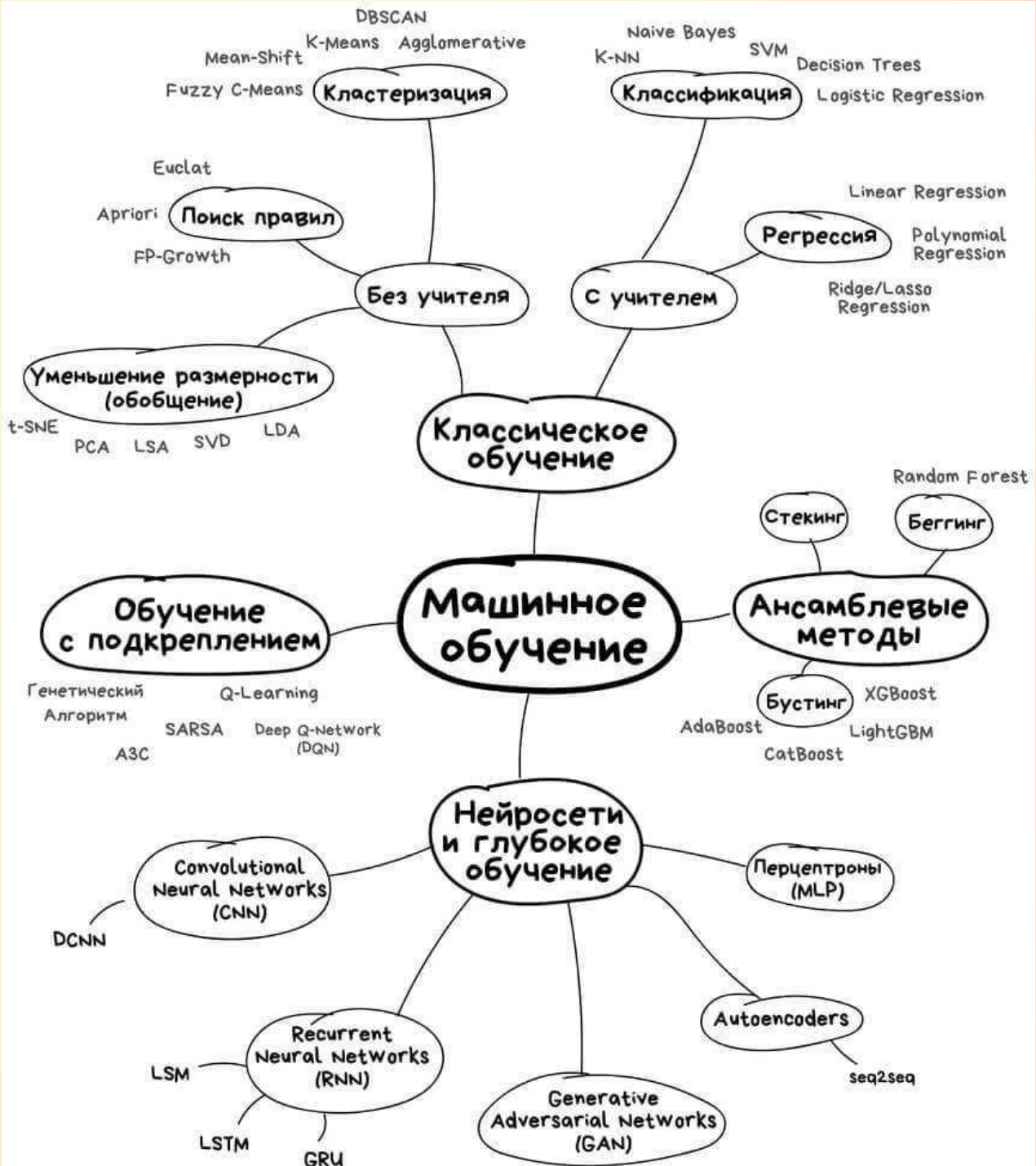
Инструменты для обработки и анализа данных

Основные понятия

- **Машинное обучение (Machine Learning) —** обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться.

- Построение систем машинного обучения является на сегодняшний день одной из самых популярных, актуальных и современных областей человеческой деятельности на стыке информационных технологий, математического анализа и статистики.

Основные стандартные типы задач



Основные стандартные типы задач

Основные типы задач

- **Задача регрессии** – прогноз на основе выборки объектов с различными признаками. На выходе должно получиться вещественное число, к примеру цена квартиры, стоимость ценной бумаги по прошествии полугода, ожидаемый доход магазина на следующий месяц, качество вина при слепом тестировании.
- **Задача классификации** – получение категориального ответа на основе набора признаков. Имеет конечное количество ответов (как правило, в формате «да» или «нет»): есть ли на фотографии кот, является ли изображение человеческим лицом, болен ли пациент раком.
- **Задача кластеризации** – распределение данных на группы: разделение всех клиентов мобильного оператора по уровню платёжеспособности, отнесение космических объектов к той или иной категории (планета, звёзда, чёрная дыра и т. п.).
- **Задача уменьшения размерности** – сведение большого числа признаков к меньшему (обычно 2–3) для удобства их последующей визуализации (например, сжатие данных).
- **Задача выявления аномалий** – отделение аномалий от стандартных случаев. На первый взгляд она совпадает с задачей классификации, но есть одно существенное отличие: аномалии – явление редкое, и обучающих примеров, на которых можно обучить машинно обучающуюся модель на выявление таких объектов, либо исчезающе мало, либо просто нет, поэтому методы классификации здесь не работают. На практике такой задачей является, например, выявление мошеннических действий с банковскими картами.

Области применения машинного обучения

- Автономные машины
- Анализ эмоциональной окраски (например, классификация рецензий на фильмы на отрицательные, положительные и нейтральные)
- Выявление аномалий
- Выявление закономерностей в данных
- Выявление попыток мошенничества с кредитными картами
- Выявление попыток страхового мошенничества
- Глубокий анализ данных в социальных сетях (Facebook, Twitter, LinkedIn)
- Диагностическая медицина
- Исследование данных
- Классификация новостей: спорт, финансы, политика и т. д.
- Классификация электронной почты и выделение спама
- Маркетинг: деление клиентов на группы
- Обнаружение вторжений в компьютерные системы
- Обнаружение объектов в сценах
- Перевод естественных языков (с английского на испанский, с французского на японский и т. д.)
- Прогнозирование временных рядов — например, предсказание будущих котировок акций и прогнозы погоды
- Прогнозирование нарушений выплат ипотечных кредитов
- Прогнозирование оттока клиентов
- Распознавание голоса
- Распознавание лиц
- Распознавание образов и классификация изображений
- Распознавание рукописного текста
- Рекомендательные системы («тем, кто купил этот продукт, также понравились...»)
- Сжатие данных
- Фильтрация спама
- Чат-боты

Примеры задач из современной реальной жизни. Диагностика заболеваний

- Пациенты в данном случае являются объектами, а признаками – все наблюдающиеся у них симптомы, анамнез, результаты анализов, уже предпринятые лечебные меры (фактически вся история болезни, формализованная и разбитая на отдельные критерии).

- Некоторые признаки – пол, наличие или отсутствие головной боли, кашля, сыпи и иные – рассматриваются как бинарные. Оценка тяжести состояния (крайне тяжёлое, средней тяжести и др.) является порядковым признаком, а многие другие – количественными: объём лекарственного препарата, уровень гемоглобина в крови, показатели артериального давления и пульса, возраст, вес.

- Собрав информацию о состоянии пациента, содержащую много таких признаков, можно загрузить её в компьютер и с помощью программы, способной к машинному обучению, решить следующие задачи:

– провести дифференциальную диагностику (определение вида заболевания);

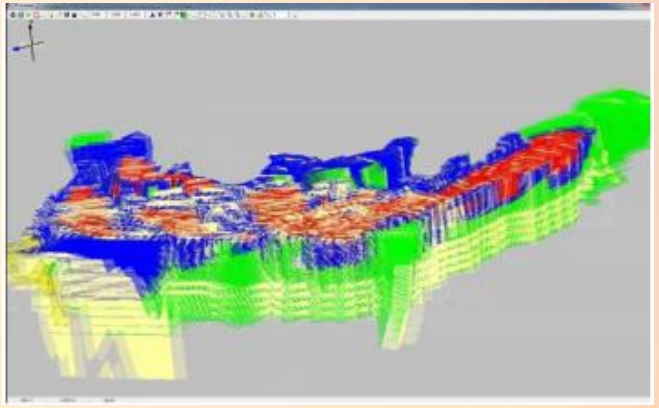
– выбрать наиболее оптимальную стратегию лечения;

– спрогнозировать развитие болезни, её длительность и исход;

– просчитать риск возможных осложнений;

– выявить синдромы – наборы симптомов, сопутствующие данному заболеванию или нарушению.

Примеры задач из современной реальной жизни. Поиск мест залегания полезных ископаемых

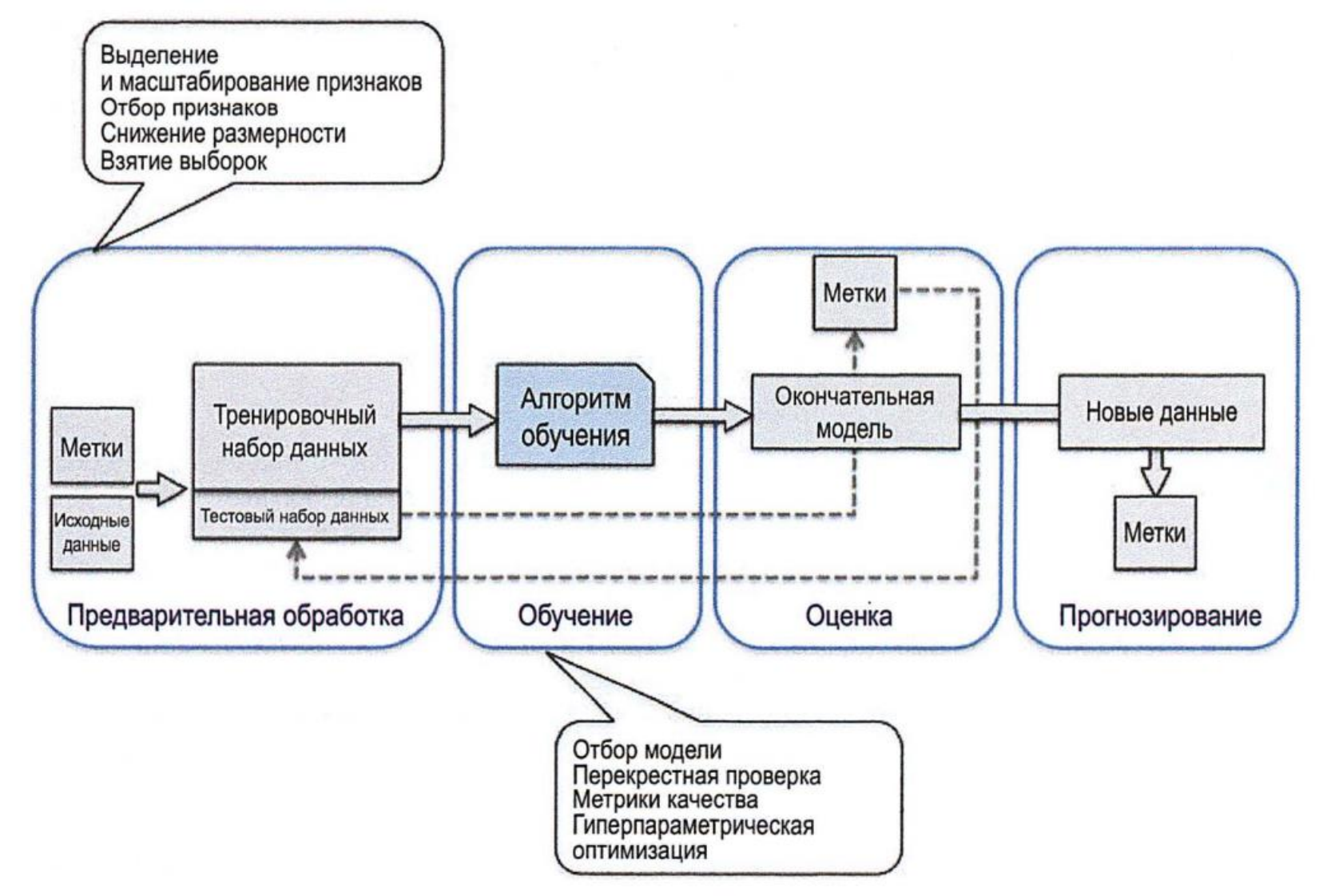


- В роли признаков здесь выступают сведения, добытые при помощи геологической разведки: наличие на территории местности каких-либо пород (и это будет признаком бинарного типа), их физические и химические свойства (которые раскладываются на ряд количественных и качественных признаков).
- Для обучающей выборки берутся 2 вида прецедентов: районы, где точно присутствуют месторождения полезных ископаемых, и районы с похожими характеристиками, где эти ископаемые не были обнаружены. Но добыча редких полезных ископаемых имеет свою специфику: во многих случаях количество признаков значительно превышает число объектов, и методы традиционной статистики плохо подходят для таких ситуаций. Поэтому при машинном обучении акцент делается на обнаружение закономерностей в уже собранном массиве данных. Для этого определяются небольшие и наиболее информативные совокупности признаков, которые максимально показательны для ответа на вопрос исследования – есть в указанной местности то или иное ископаемое или нет. Можно провести аналогию с медициной: у месторождений тоже можно выявить свои синдромы.

Примеры задач из современной реальной жизни. Оценка надёжности и платёжеспособности кандидатов получение кредитов

- Лица, запрашивающие у банка заём, – это объекты, а вот признаки будут отличаться в зависимости от того, физическое это лицо или юридическое. Признаковое описание частного лица, претендующего на кредит, формируется на основе данных анкеты, которую оно заполняет. Затем анкета дополняется некоторыми другими сведениями о потенциальном клиенте, которые банк получает по своим каналам. Часть из них относятся к бинарным признакам (пол, наличие телефонного номера), другие – к порядковым (образование, должность), большинство же являются количественными (величина займа, общая сумма задолженностей по другим банкам, возраст, количество членов семьи, доход, трудовой стаж) или номинальными (имя, название фирмы-работодателя, профессия, адрес).
- Для машинного обучения составляется выборка, в которую входят кредитополучатели, чья кредитная история известна. Все заёмщики делятся на классы, в простейшем случае их 2 – «хорошие» заёмщики и «плохие», и положительное решение о выдаче кредита принимается только в пользу «хороших».

Дорожная карта для построения систем машинного обучения

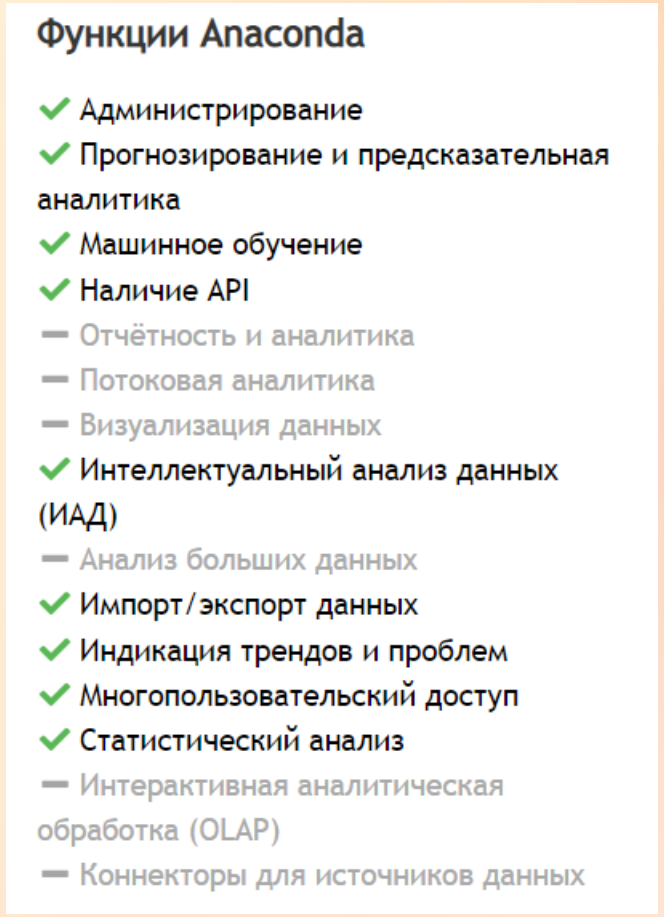


Инструменты для анализа данных

- **Python** (в русском языке встречаются названия питóн или пáйтон) — высокоуровневый язык программирования общего назначения с динамической строгой типизацией и автоматическим управлением памятью, ориентированный на повышение производительности разработчика, читаемости кода и его качества, а также на обеспечение переносимости написанных на нём программ.
- Язык является полностью объектно-ориентированным в том плане, что всё является объектами.
- Необычной особенностью языка является выделение блоков кода пробельными отступами. Синтаксис ядра языка минималистичен, за счёт чего на практике редко возникает необходимость обращаться к документации. Сам же язык известен как интерпретируемый и используется в том числе для написания скриптов.
- Недостатками языка являются зачастую более низкая скорость работы и более высокое потребление памяти написанных на нём программ по сравнению с аналогичным кодом, написанным на компилируемых языках, таких как C или C++

Инструменты для анализа данных

- **Anaconda** – это платформа управления пакетами приложений анализа данных (для языков Python и R) с открытым исходным кодом. Система позволяет специалистам по обработке данных быстро разворачивать проекты машинного обучения, предоставляя необходимую информацию для лиц, принимающих решения.
- Программный продукт Anaconda (рус. Анаконда) от одноимённой компании предназначен для управления приложениями анализа данных, основанных на моделях машинного обучения, современных алгоритмах интеллектуального анализа данных и статистических методах анализа. Платформа подходит для работы как студентов, так и опытных специалистов (имеются разные варианты инсталляции). Продукт подходит также для предприятий, которым требуется обеспечить полный жизненный цикл машинного обучения и платформу для поддержки искусственного интеллекта.



Инструменты для анализа данных

- **Jupyter Notebook** — это веб-приложение с открытым исходным кодом, которое позволяет создавать и обмениваться документами, которые содержат живой код и результат его выполнения, уравнения, визуализации и текст с пояснениями по вычислениям. Часто используется для: очистки и преобразования данных, числового моделирования, статистического моделирования, визуализации данных, машинного обучения и др.

Инструменты для анализа данных

- **PyCharm** — интегрированная среда разработки для языка программирования Python. Предоставляет средства для анализа кода, графический отладчик, инструмент для запуска юнит-тестов и поддерживает веб-разработку на Django. PyCharm разработана компанией JetBrains на основе IntelliJ IDEA.
- PyCharm — это кроссплатформенная среда разработки, которая совместима с Windows, macOS, Linux. PyCharm Community Edition (бесплатная версия) находится под лицензией Apache License, а PyCharm Professional Edition (платная версия) является проприетарным ПО.
- PyCharm предоставляет умную проверку кода, быстрое выявление ошибок и оперативное исправление, вкупе с автоматическим рефакторингом кода, и богатыми возможностями в навигации.

Пакеты Python для обработки и анализа данных

Библиотеки для научных вычислений (scientific computing libraries)

- **Pandas** — это библиотека для работы с данными, а именно для анализа данных, трансформации данных (обработки данных), загрузки данных из различных источников и сохранения данных в разных форматах как в файловую систему, так и в базу данных. В качестве структуры данных используется Pandas DataFrame или Pandas Series.

- **Numpy** — это библиотека языка Python, которая позволяет работать с многомерными массивами и матрицами, в том числе внутри библиотеки есть большой выбор математических функций для выполнения операций над массивами и матрицами.

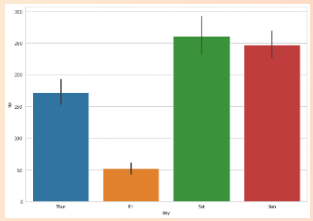
- **SciPy** — это пакет прикладных математических процедур (или научных инструментов), основанный на расширении Numpy Python. Содержит модули для оптимизации, интегрирования, специальных функций, обработки сигналов, обработки изображений, генетических алгоритмов, решения обыкновенных дифференциальных уравнений и других задач, обычно решаемых в науке и при инженерной разработке.

Пакеты Python для обработки и анализа данных

Библиотеки для визуализации (visualization libraries)

- **Matplotlib —** библиотека на языке программирования Python для визуализации данных двумерной и трёхмерной графикой. Получаемые изображения могут быть использованы в качестве иллюстраций в публикациях.

- **Seaborn —** это библиотека для создания статистических графиков на Python. Она основывается на matplotlib и тесно взаимодействует со структурами данных pandas.



Пакеты Python для обработки и анализа данных

Библиотеки алгоритмы (algorithmic libraries)

- **Scikit-Learn —** это инструмент для обработки изображений и имитации искусственного интеллекта. Библиотека построена на SciPy (Scientific Python). Библиотека Scikit-Learn использует в своей реализации NumPy массивы. Она предоставляет широкий выбор алгоритмов обучения с учителем и без учителя. В этой библиотеке находится большое количество алгоритмов для задач, связанных с классификацией и машинным обучением в целом.

- **Statsmodels —** это модуль Python, который предоставляет классы и функции для оценки множества различных статистических моделей, а также для проведения статистических тестов и исследования статистических данных.