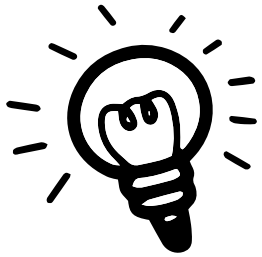
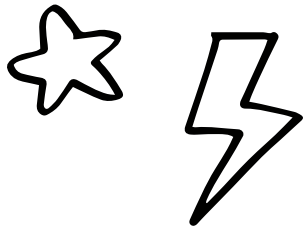
Вклад в будущее
СБЕР

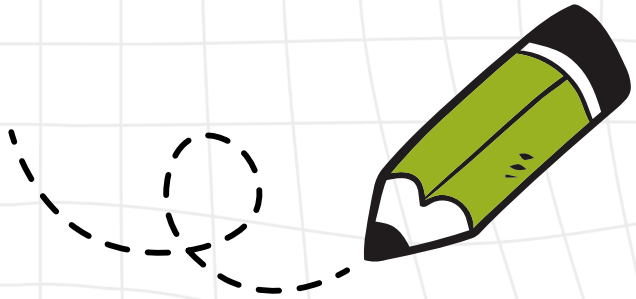
АКАДЕМИЯ
искусственного интеллекта
для школьников

СТРУКТУРИРОВАНИЕ И АНАЛИЗ ДАННЫХ





Урок 2



ДАННЫЕ СТРУКТУРИРОВАННЫЕ (КОЛИЧЕСТВЕННЫЕ ДАННЫЕ)

Данные [data] – это сведения, факты, показатели, выраженные как в числовой, так и любой другой форме.

Структурированные – отражают отдельные факты предметной области.

Записываются в стандартизованном формате (в основном в виде таблиц), что позволяет применять к ним различные методы автоматизированной обработки.

ID Клиента	Фамилия	Имя	Отчество	Пол
1	Иванов	Иван	Иванович	М
2	Иванова	Людмила	Андреевна	Ж
3	Сидоров	Андрей	Анатольевич	М
4	Сидорова	Юлия	Ивановна	Ж
5	Петров	Аркадий	Алексеевич	М
6	Петрова	Анна	Александровна	Ж

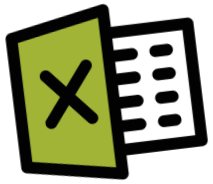
ДАННЫЕ СТРУКТУРИРОВАННЫЕ (КОЛИЧЕСТВЕННЫЕ ДАННЫЕ)

Большинство алгоритмов машинного обучения, статистического и интеллектуального анализа данных работают только со структурированными данными.

Структурированные данные хранятся в специальных хранилищах. Это компактные хранилища или репозитории с определенной структурой, которую сложно изменить.

СТРУКТУРИРОВАННЫЕ ДАННЫЕ:

ПРИМЕРЫ



- ФАЙЛЫ EXCEL
- БАЗЫ ДАННЫХ SQL
- РЕЗУЛЬТАТЫ ЗАПОЛНЕНИЯ ВЕБ-ФОРМЫ
- ТЕГИ ОПТИМИЗАЦИИ ДЛЯ ПОИСКОВЫХ СИСТЕМ (SEO)
- КАТАЛОГИ ПРОДУКТОВ
- СИСТЕМЫ БРОНИРОВАНИЯ

ДАННЫЕ НЕСТРУКТУРИРОВАННЫЕ (КАЧЕСТВЕННЫЕ ДАННЫЕ)

Неструктурированные – не имеют заранее определенной структуры и представляются в любом виде – текстовом, графическом, звуковом, видео.

Традиционные методы и инструменты не могут быть использованы для их анализа и обработки.

Данные хранят в репозиториях хранения – озерах данных. **Озеро данных** – это хранилище или система, предназначенная для хранения огромных объемов данных в естественном / необработанном формате.

НЕСТРУКТУРИРОВАННЫЕ ДАННЫЕ:

ПРИМЕРЫ

ЭЛЕКТРОННАЯ ПОЧТА

ТЕКСТОВЫЕ ФАЙЛЫ

СООБЩЕНИЯ В СОЦИАЛЬНЫХ СЕТЯХ

ВИДЕО

ИЗОБРАЖЕНИЯ

АУДИО

ДАННЫЕ ДАТЧИКОВ

СТРУКТУРИРОВАННЫЕ ДАННЫЕ: ПРЕИМУЩЕСТВА

Доступность использования

Формат организации данных (таблица) понятен любому пользователю. Организовать доступ к данным просто.

Масштабируемость

Структурированные данные масштабируются алгоритмически. Для хранения данных добавляются новые хранилища.

Аналитика

Используется язык структурированных запросов (SQL) для создания отчетов, а также для изменения и обслуживания данных. Алгоритмы машинного обучения могут анализировать структурированные данные и выявлять общие закономерности.

СТРУКТУРИРОВАННЫЕ ДАННЫЕ: ПРОБЛЕМЫ

Ограниченное использование

Предопределенная структура является преимуществом, но может быть и проблемой. Структурированные данные можно использовать только по назначению.

Отсутствие гибкости

Изменение схемы структурированных данных по мере изменения обстоятельств и появления новых отношений или требований может быть дорогостоящим и ресурсоемким.

ПЕРСПЕКТИВЫ РАЗВИТИЯ СТРУКТУРИРОВАННЫХ ДАННЫХ

По прогнозам IBM ожидается, что в 2025 году глобальный объем данных достигнет 175 зеттабайт.

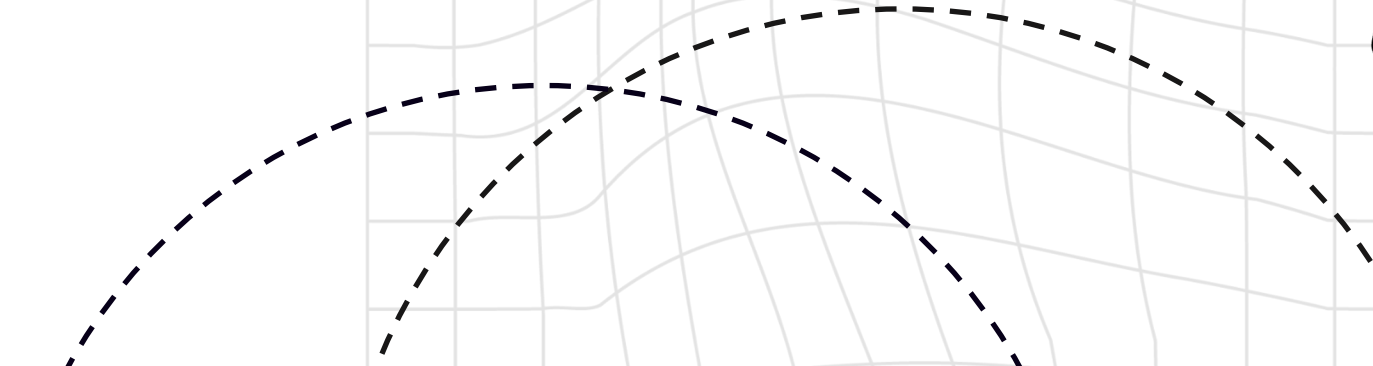
35ZB содержит примерно 1 триллион часов фильмов. Чтобы посмотреть все эти фильмы, потребуется 115 миллионов лет.

Сейчас большая часть данных (около 80%), неструктурированная.

Только 20% всей генерируемой информации происходит на базе структурированных данных.

АНАЛИЗ ДАННЫХ

Это процесс исследования, фильтрации, преобразования и моделирования данных с целью извлечения полезной информации и принятия решений.

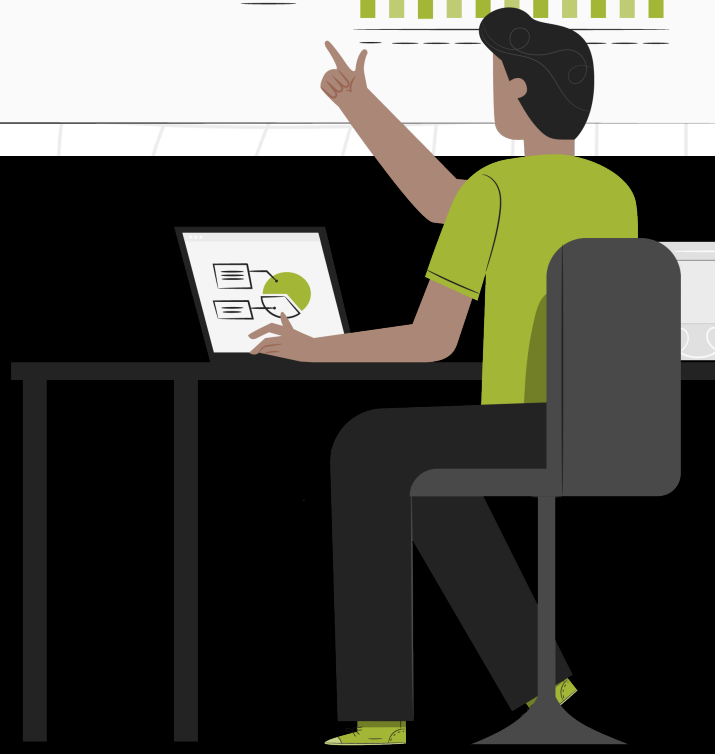


ЦЕЛИ И ЗАДАЧИ АНАЛИЗА ДАННЫХ

Цель анализа данных — комплексное понимание исследуемой ситуации для принятия решений (выявление тенденций, посторенние прогнозов, выработка рекомендаций).

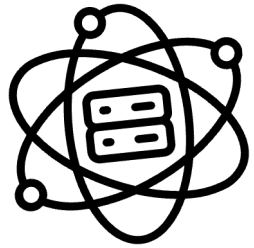
Для достижения цели ставятся следующие задачи:

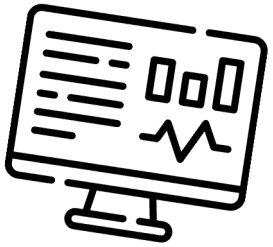
- сбор данных
- структурирование данных
- выявление закономерностей
- прогнозирование и выработка рекомендаций



ПРОГРАММНЫЕ СРЕДСТВА ПО АНАЛИЗУ ДАННЫХ

1. Аналитика Плюс.
2. Loginom — аналитическая low-code платформа, которая позволяет проводить анализ данных любого уровня сложности без программирования.
3. АСУ-аналитика.
4. IBM SPSS Statistics.
5. N3.Аналитика.



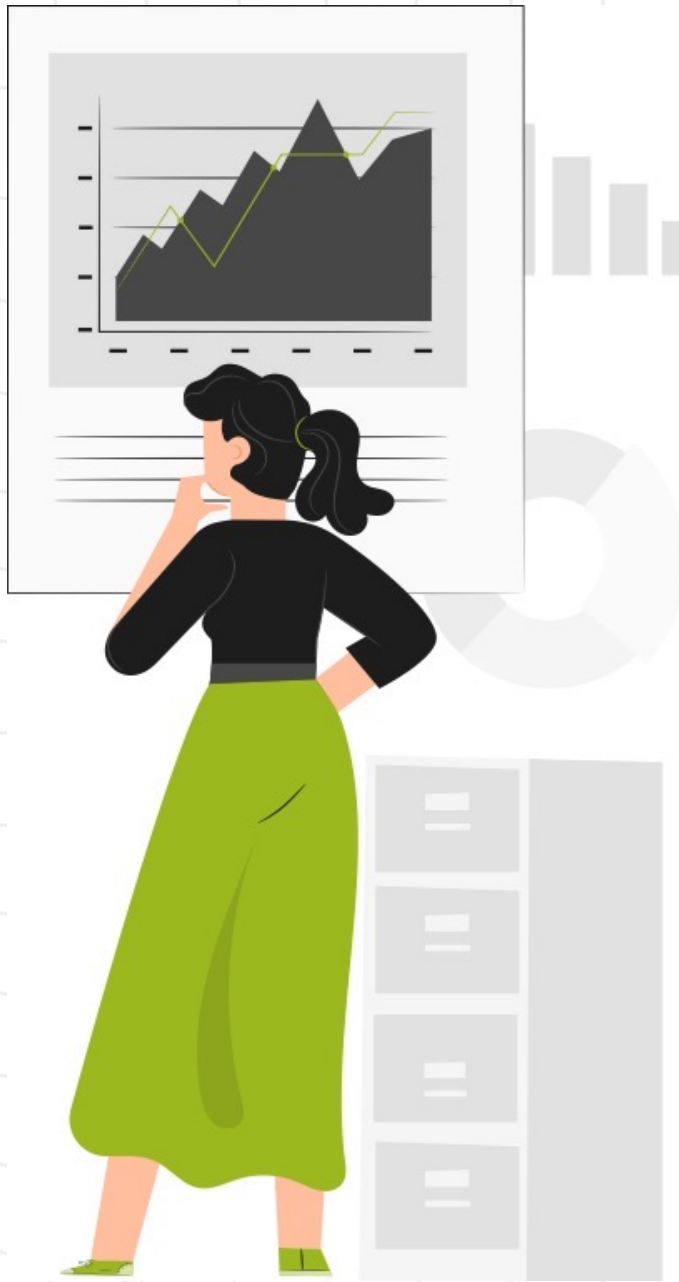


ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ

Интеллектуальный анализ данных преобразует необработанные данные в практические знания.

Компании используют знания для решения проблем, анализа будущего влияния бизнес-решений и повышения прибыли.

ИИ способен анализировать неструктурированные данные.

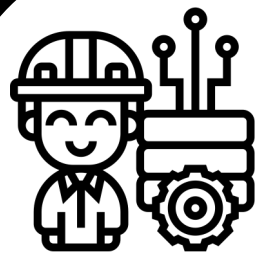


СПЕЦИАЛИСТЫ ПО РАБОТЕ С ДАННЫМИ

Аналитик данных (Data Analyst или дата-аналитик) — это специалист по анализу больших данных: сбор, обработка, выводы. На основании его отчетов в компаниях принимают важные решения. Профессия аналитика данных находится на стыке IT, менеджмента и математики.

Дата-сайентист (data scientist) – это программист, который создаёт модели, предсказывающие результат. Он ищет в массивах данных связи и закономерности, на основе которых и строит модель. Разница между дата-сайентистом и дата-аналитиком в том, что аналитик не строит модели, а занимается анализом данных.

СПЕЦИАЛИСТЫ ПО РАБОТЕ С ДАННЫМИ



Архитектор баз данных (Database Architect) — это ИТ-специалист экспертного уровня, который занимается выбором технологии для хранения данных, составляет план разработки БД, может выполнять проектирование и оптимизацию БД, следит за ее безопасностью.

Разработчик баз данных (Database Developer или Database Programmer) создает, настраивает, оптимизирует, модернизирует и обслуживает базы данных (БД), которые входят в информационные системы.

ВЫВОДЫ ПО ТЕМЕ. СИНКВЕЙН

1-я строка – одно ключевое слово, определяющее содержание синквейна;

2-я строка – два прилагательных, характеризующих данное понятие;

3-я строка – три глагола, обозначающих действие в рамках заданной темы;

4-я строка – короткое предложение, раскрывающее суть темы или отношение к ней;

5-я строка – синоним ключевого слова (существительное).

Например:

- Данные
- Сложные, структурированные
- Собираем, храним, анализируем
- Данные нужны для принятия решений
- Информация