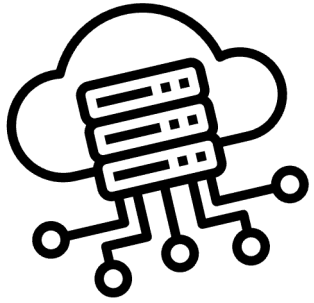
Вклад в будущее
СБЕР

АКАДЕМИЯ
искусственного интеллекта
для школьников

ЧТО ТАКОЕ ДАННЫЕ И КАК С НИМИ РАБОТАТЬ?



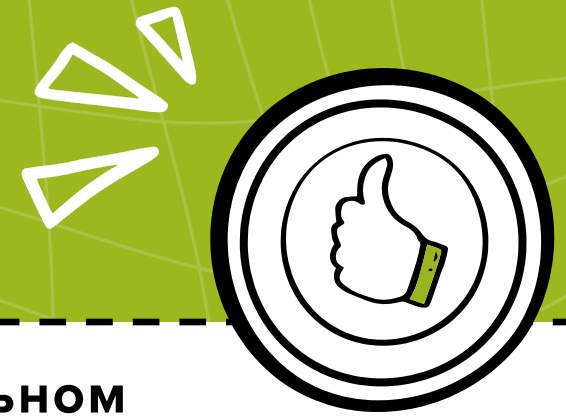
Урок 1

КОМИТЕТ ПО ДАННЫМ ДЛЯ НАУКИ И ТЕХНИКИ

1966 год создан Комитет по данным для науки и техники (англ. Committee on Data for Science and Technology; CODATA), в дальнейшем Комитет по данным Международного совета по науке. https://codata.org/

Цель комитета — сбор, критическая оценка, хранение и предоставление данных для задач науки и техники.

ЧТО ТАКОЕ ДАННЫЕ
СОГЛАСНО СТАНДАРТАМ



Данные (data): Представление информации в формальном виде, пригодном для передачи, интерпретации или обработки. Примечание. Данные могут быть обработаны автоматически или вручную.

ГОСТ Р ИСО/МЭК 20546-2021. Информационные технологии. БОЛЬШИЕ ДАННЫЕ.

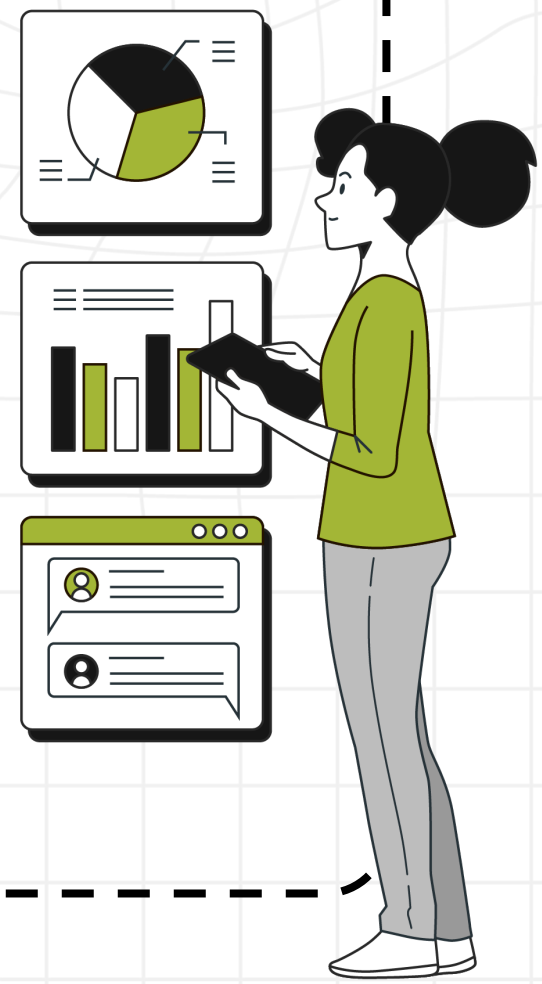
Данные – это факты или информация, которые можно использовать для отчетности, расчетов, планирования или анализа. www.dama.org

Свод знаний по управлению данными DMBOK (Data Management Body of Knowledge), 2009

ЧТО ТАКОЕ ДАННЫЕ

Данные — числовые и нечисловые значения характеристик кого-либо (чего-либо), с которыми выполняет операции человек или вычислительное устройство, в этом случае их называют машиночитаемыми данными.

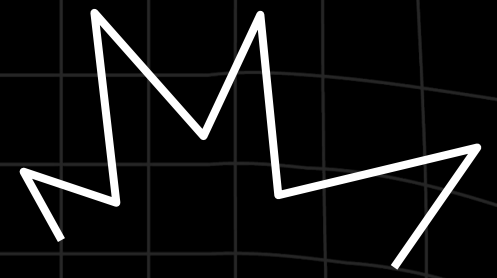
Данные — это зафиксированные и сохраненные человеком или компьютером факты или сведения (значения, показателя, характеристики в любом формате) о событии, явлении, процессе или объекте из окружающего мира.



ДАННЫЕ VS ИНФОРМАЦИЯ

Информация — это обработанные данные.

Информация — форматированные данные, обработанные с определенной целью, которым придан смысл посредством добавления контекста.



Данные — это форма информации,

а **информация** — это форма данных.

ТИПЫ ДАННЫХ ПО НАЗНАЧЕНИЮ

Метаданные — описывают структуру и характеристики данных

Основные данные — данные об объектах

Справочные данные — данные из справочников, международных, общероссийских и отраслевых классификаторов и т. п.

ТИПЫ ДАННЫХ ПО СТРУКТУРЕ

СТРУКТУРИРОВАННЫЕ — данные, имеющие строго определённую структуру, определяемую формальной моделью данных (например, таблицы)

НЕСТРУКТУРИРОВАННЫЕ — данные, произвольные по форме, не имеющие строго определенной структуры и не организованные по определенным правилам (например текст или видео)

Датасет (англ. dataset) — это обработанный и структурированный массив данных, готовый для анализа, исследования и использования в различных вычислительных задачах.

БОЛЬШИЕ ДАННЫЕ (BIG DATA)

Большие данные — это разнообразные данные, поступающие с высокой скоростью и требующие специальных средств их обработки.

Свойства больших данных:

- разнообразие
- высокая скорость поступления
- большой объем

ФОРМАТЫ ДАННЫХ

1. **TSV (tab-separated values)** — это текстовый формат файла, в котором данные в столбцах разделены знаками табуляции, а строки — знаками перевода строки.

2. **CSV (Comma-Separated Values)** — это текстовый формат данных, в котором значения разделены запятыми.

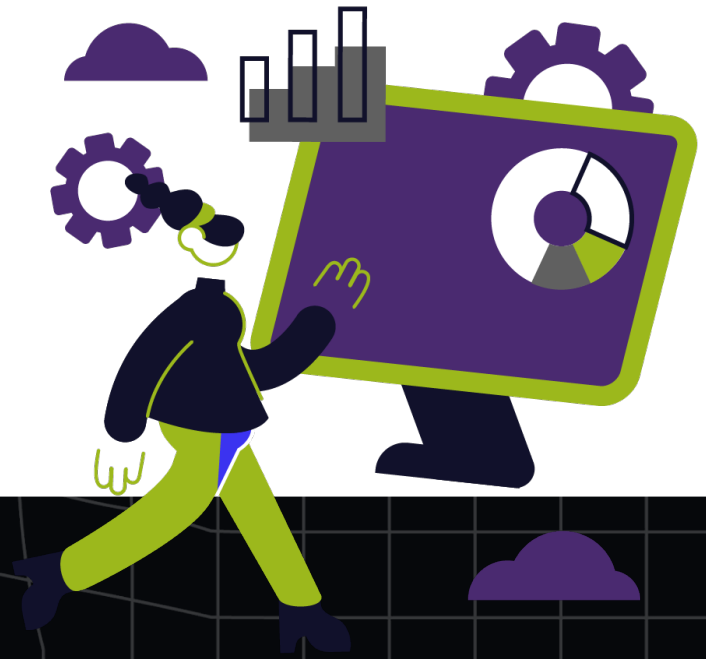
3. **JSON (JavaScript Object Notation)** — это формат данных, основанный на синтаксисе JavaScript, и он используется для представления структурированных данных.

4. **XML (eXtensible Markup Language)** — это язык разметки, используемый для описания структуры данных.

ОТКРЫТЫЕ ДАННЫЕ

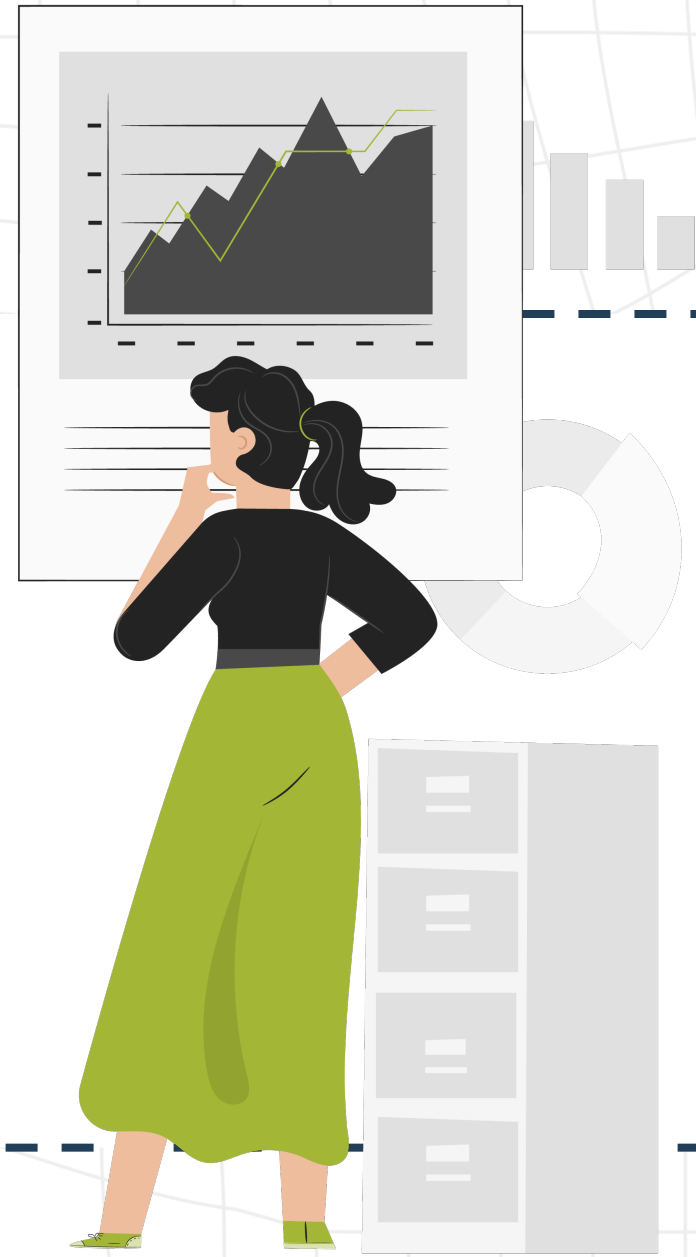
Открытые данные – это информация, размещенная в сети интернет в виде систематизированных данных, организованных в формате, обеспечивающем ее автоматическую обработку без предварительного изменения человеком, в целях неоднократного, свободного и бесплатного использования.

Федеральный Закон «Об обеспечении доступа к информации о деятельности государственных органов и органов местного самоуправления», 2009 год



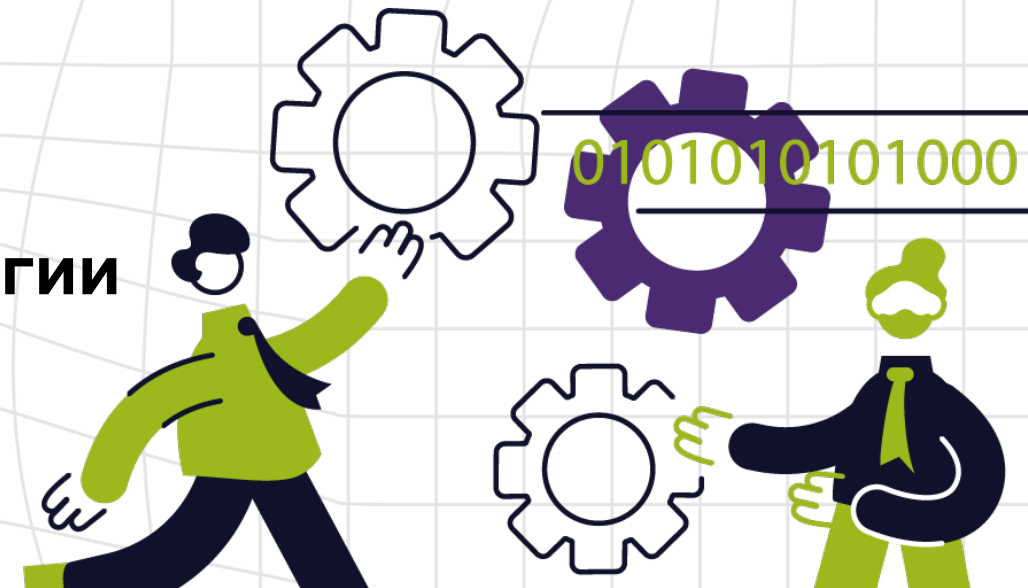
ГДЕ БРАТЬ НАДЕЖНЫЕ ДАННЫЕ?

1. Человек
2. Государственные данные
3. Данные компаний
4. Научные учреждения
5. Независимые профессиональные сообщества и международные организации



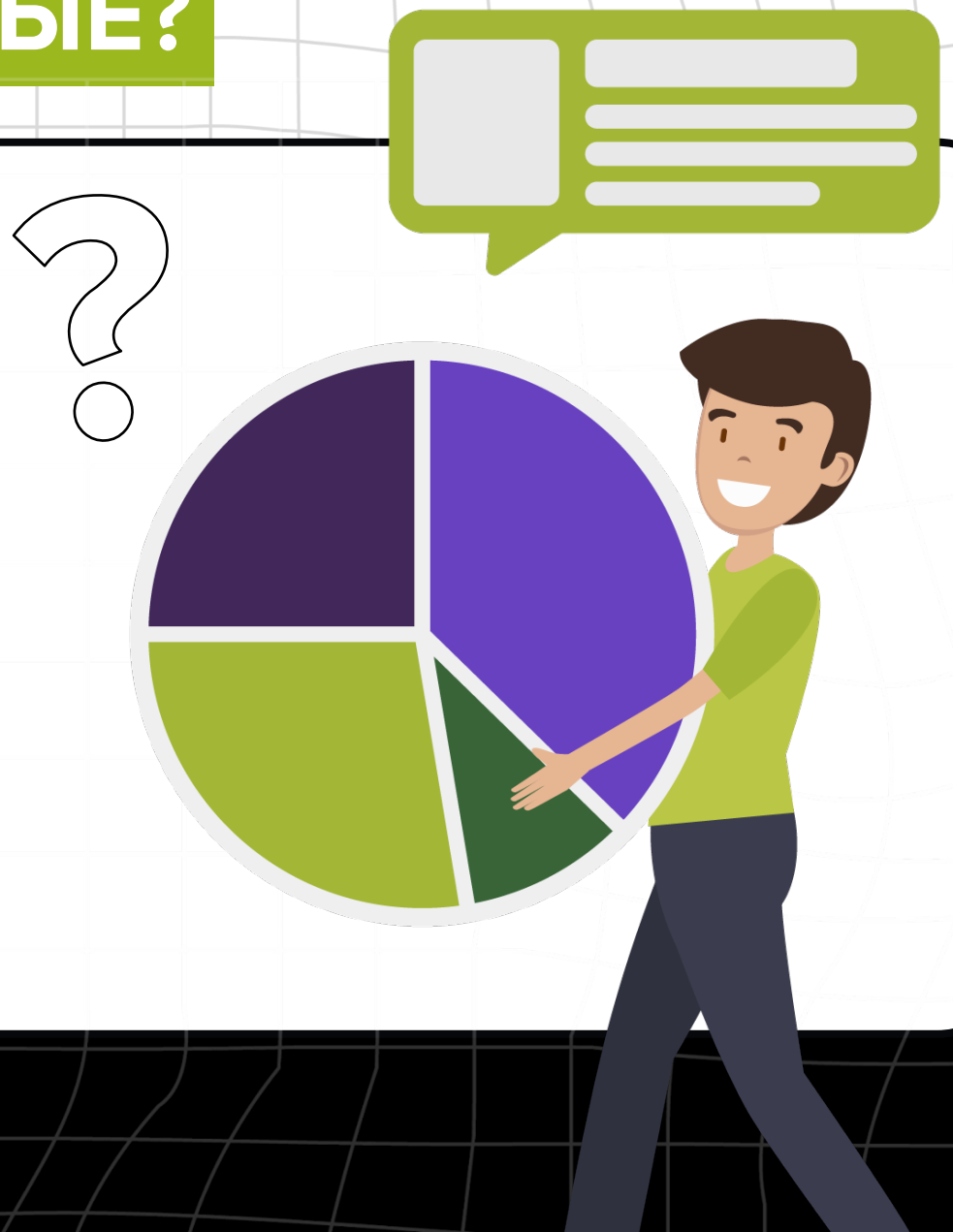
КРИТЕРИИ ВЫБОРА ИСТОЧНИКОВ ОТКРЫТЫХ ДАННЫХ

1. Известный и авторитетный владелец данных
2. Полные реквизиты данных (метаданные)
3. Надежность места хранения и наличие точки входа
4. Доступность данных
5. Наличие информации о методологии и способах сбора данных



КАК СОБИРАЮТ ДАННЫЕ?

1. Ручной сбор
2. Автоматизированный сбор
3. Сбор данных с помощью сенсоров и датчиков
4. Опросы и интервью
5. Статистические данные
6. Веб-скрейпинг или парсинг

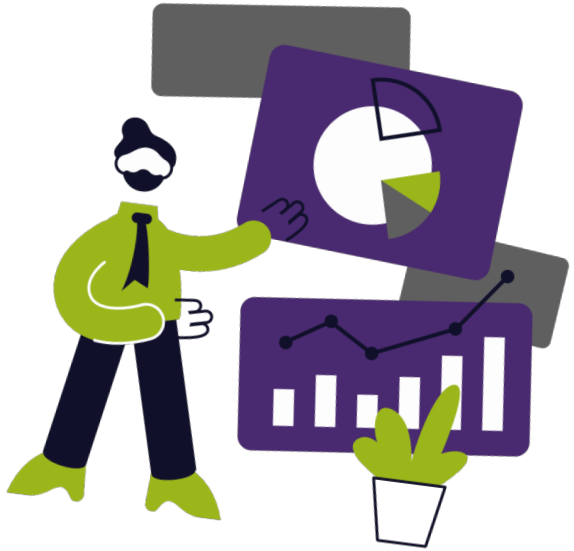


НАУКА О ДАННЫХ

Наука о данных (англ. data science; иногда даталогия — datalogy) — раздел информатики, изучающий проблемы анализа, обработки и представления данных в цифровой форме.

Сферы применения науки о данных :

- **Прогнозирование спроса**
- **Система рекомендаций**
- **Динамическое ценообразование**
- **Поиск аномалий**



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
- Большие, разнообразные
- Ищем, обрабатываем, используем
- Данные – вторая нефть
- Сведения