

*Тыщенко Данил Эдуардович*

*ФГБОУ ВО Уральский государственный университет путей  
сообщения (УрГУПС)*

*г. Екатеринбург, Россия*

*Хорожев Глеб Олегович*

*ФГБОУ ВО Уральский государственный университет путей  
сообщения (УрГУПС)*

*г. Екатеринбург, Россия*

*Научный руководитель: Грибанова Александра Валерьевна*

*Ассистент кафедры «Экономика транспорта»*

*ФГБОУ ВО Уральский государственный университет путей  
сообщения (УрГУПС)*

*г. Екатеринбург, Россия*

## **АНАЛИЗ БОЛЬШИХ ДАННЫХ: СТАТИСТИЧЕСКИЕ ПОДХОДЫ К ОБРАБОТКЕ И АНАЛИЗУ БОЛЬШИХ ДАННЫХ**

*Аннотация.* Статья посвящена применению статистических методов при анализе больших данных. Рассматриваются различные этапы обработки и анализа больших объемов информации, начиная от сбора данных до построения прогнозных моделей. Основное внимание уделено таким аспектам, как сбор и подготовку данных, визуализации данных, регрессивному анализу, временным рядам и прогнозированию, ограничениям и перспективам развития. Описаны современные инструменты и алгоритмы, используемые для решения задач анализа больших данных, а также приведены примеры практических приложений этих методов.

**Ключевые слова:** Большие данные, сбор и подготовка данных, визуализация данных, регрессивный анализ, временные ряды и прогнозирование, ограничения и перспективы развития.

**Annotation.** The article is devoted to the application of statistical methods in the analysis of big data. Various stages of processing and analysing large amounts of information are considered, ranging from data collection to the construction of predictive models. The focus is on aspects such as data collection and preparation, data visualization, regression analysis, time series and forecasting, constraints and development prospects. Modern tools and algorithms used to solve big data analysis problems are described, as well as examples of practical applications of these methods.

**Keywords:** Big data, data collection and preparation, data visualization, regression analysis, time series and forecasting, constraints and development prospects.

## Введение

Выражение «большие данные» всё чаще встречается в нашей жизни, но далеко не все понимают что это такое на самом деле. Итак, под большими данными понимаются разнообразные данные, поступающие с высокой скоростью, объем которых постоянно растет. Таким образом, три ключевые характеристики больших данных - это разнообразие, быстрое поступление и большой объем.

Проще говоря, большие данные - это большие и более сложные наборы данных, особенно из новых источников. Размер этих наборов данных настолько велик, что традиционные программы обработки не могут с ними справиться. Однако эти большие данные могут быть использованы для решения ранее неразрешимых задач.

Как было сказано ранее большие данные помогают решить ранее неразрешимые задачи в разных областях жизни, так и в статистике они

используются для статистического подхода в анализе данных и нахождения ответов на следующие вопросы:

- Какие признаки являются наиболее важными?
- Каким требованиям должен отвечать эксперимент, чтобы на его основе можно было сформировать стратегию продукта?
- Каковы ключевые показатели рентабельности?
- Каковы наиболее распространённые результаты применения того или иного подхода? Оправдывает ли он ожидания?
- Как можно установить достоверность данных?

Большие данные могут использоваться в самых разных областях деятельности, от взаимодействия с клиентами до аналитики. Вот несколько практических примеров использования:

- Разработка продуктов - Такие компании, как Netflix и Procter & Gamble, используют большие данные для прогнозирования потребительского спроса. Классифицируя ключевые характеристики существующих и устаревших продуктов и услуг, создаётся взаимосвязь между этими характеристиками и коммерческим успехом продукта, чтобы прогнозировать новые продукты и услуги.
- Взаимодействие с заказчиками - Получить точные данные о клиентском опыте стало проще, чем когда-либо прежде. Большие данные позволяют извлекать полезные сведения из социальных сетей, информации о посещении веб-сайтов и других источников, чтобы повысить качество взаимодействия с клиентами и предоставить максимально полезный сервис.
- Машинное обучение - Машинное обучение одна из самых популярных тем. Большие данные, являются одной из причин этой популярности. Мы можем обучать машины, а не программировать их.

## Сбор и подготовка данных

Сбор данных является основным и наиболее важным этапом при работе с большими данными. В него входит получение данных и их предварительная обработка для дальнейшего анализа.

К этапам сбора данных относятся:

1. Определение источника данных.
  - 1.1. Данные компании.
  - 1.2. Различные сенсоры.
  - 1.3. Данные из открытых источников (Веб-сайты, соц-сети).
2. Сбор данных.
  - 2.1. Извлечение, трансформация и загрузка данных во внутренние хранилища.
3. Предварительная обработка.
  - 3.1. Очистка от неверных и дублирующихся данных.
  - 3.2. Необходимое форматирование.
  - 3.3. Удаление избытков информации.
4. Хранение данных.

Для облегчения и автоматизации процесса сбора данных используются специально разработанные инструменты и платформы. Ниже приведён пример нескольких из них:

- Apache Nifi - инструмент для управления потоками данных, позволяющий создавать сложные процессы переноса данных из нескольких систем в одну (ETL-процессы).
- Talend - платформа для интеграции данных, включающая функции ETL-процессов и одновременной подготовки данных.
- StreamSets – решение для управления и мониторинга процесса непрерывного сбора данных по мере их создания и перемещения между хранилищами.

## Визуализация данных

Человек не в силах анализировать большие данные, так как они представляют собой гигантский массив цифр, в этом ему способна помочь наука по названию «Big Data». Она предоставляет инструменты и методы для анализа огромных Больших данных. Одним из важнейших разделов этой науки является визуализация результатов анализа.

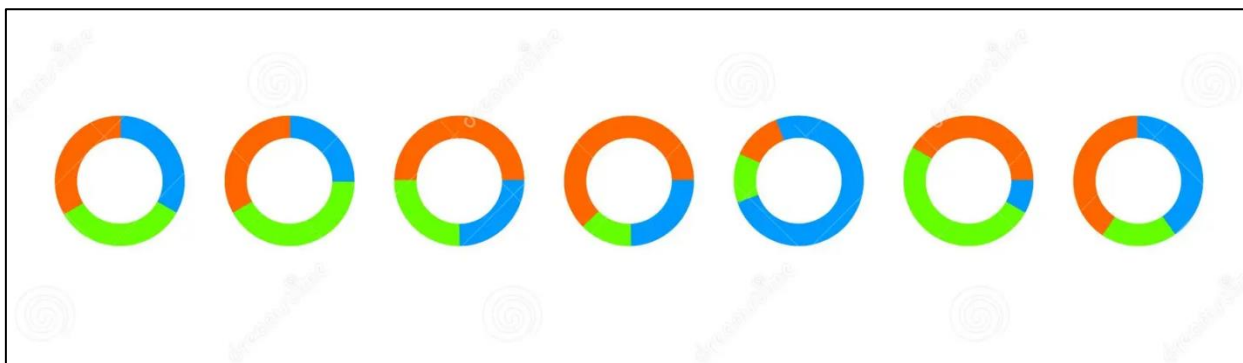
Визуализация больших данных - это представление информации в графическом виде, доступном для анализа и интерпретации. Проще говоря, вместо массивов цифр мы получаем графики, карты и диаграммы.

Графическая визуализация данных помогает:

- Проще находить закономерности и взаимосвязи. На графиках проще выделить общие паттерны в особенно если рассматривать несколько типов визуализации, применяемых к общей базе данных.
- Обнаружить проблемные и аномальные зоны. Провалы и неравномерности четко видны на графиках, устраняя их, можно повышать объективность анализа.
- Принимать обоснованные решения. Чем нагляднее представлены данные, тем проще их анализировать, не упуская из вида все нюансы.

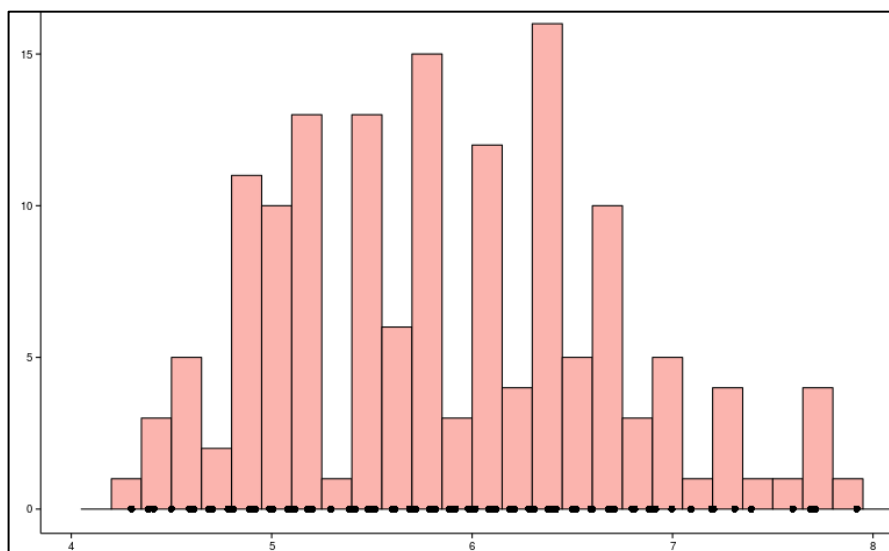
На данный момент в аналитике используется более 60 типов визуализации данных. Рассмотрим наиболее удобные и популярные на сегодняшний день.

- Круговая диаграмма. Это круг, разделенный на несколько секторов. Она полезна для визуальной оценки пропорций и процентных соотношений. В круговой диаграмме полезно учитывать несколько элементов, особенно если разница между ними велика. Если значения почти равны (42 % и 58 %), заметить разницу становится сложнее.



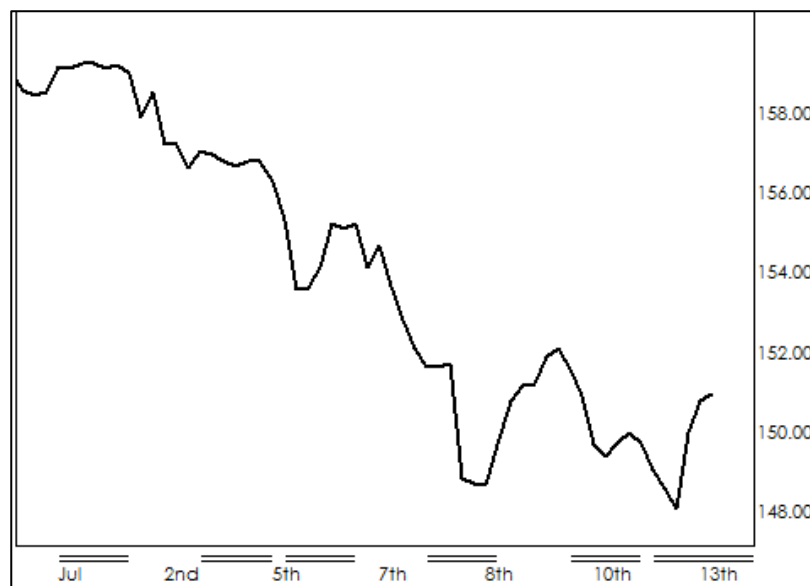
Изображение 1. Круговая диаграмма

- Гистограмма. Данные представлены прямоугольниками, высота которых пропорциональна значению. В анализе она помогает наглядно представить, как часто встречается то или иное значение в данном наборе данных.



Изображение 2. Гистограмма

- Линейный график. Наиболее простой тип графика, позволяющий отслеживать движение одного или нескольких показателей.



Изображение 3. Линейный график

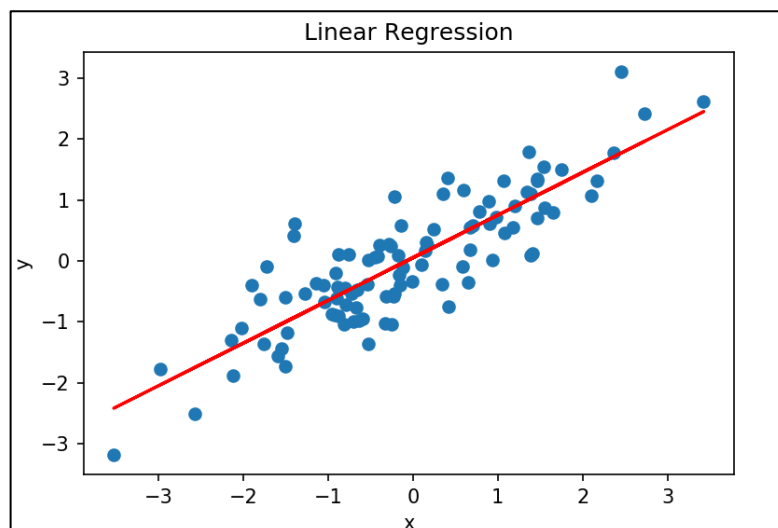
Простейшим примером выявления закономерности при работе с визуализацией может послужить гистограмма, с её помощью определить основную возрастную категорию клиентов.

### Регрессивный анализ

Регрессионный анализ - это статистический метод, используемый для исследования взаимосвязи между переменными. Основная цель регрессионного анализа - найти математическую модель, описывающую зависимость одной переменной от другой. В зависимости от типа зависимой переменной и характера связи различают несколько типов регрессии:

- Линейная регрессия

Линейная регрессия - один из самых простых и часто используемых методов регрессионного анализа. Он предполагает, что связь между независимой переменной  $x$  и зависимой переменной  $y$  может быть выражена в виде линейной функции. Пример линейной регрессии:

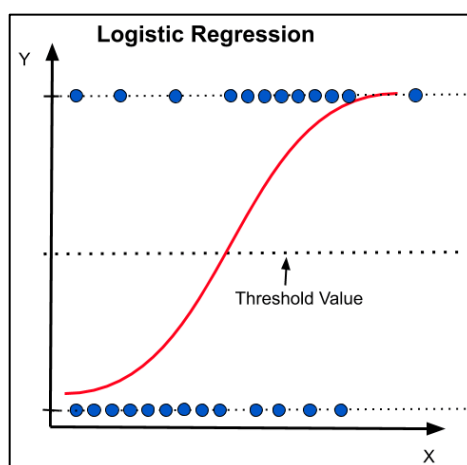


Изображение 4. Линейная регрессия

Линейная регрессия часто используется для прогнозирования значения зависимой переменной на основе известных независимых переменных. Например, ее можно использовать для прогнозирования продаж на основе расходов на рекламу или для прогнозирования стоимости недвижимости на основе площади.

- Логистическая регрессия

Логистическая регрессия используется, когда зависимая переменная принимает два значения (например, да/нет, успех/неудача). В отличие от линейной регрессии, где зависимая переменная является непрерывной, в логистической регрессии используется функция *logit*, которая преобразует вероятность события в диапазон от 0 до 1. Примеры логистической регрессии:

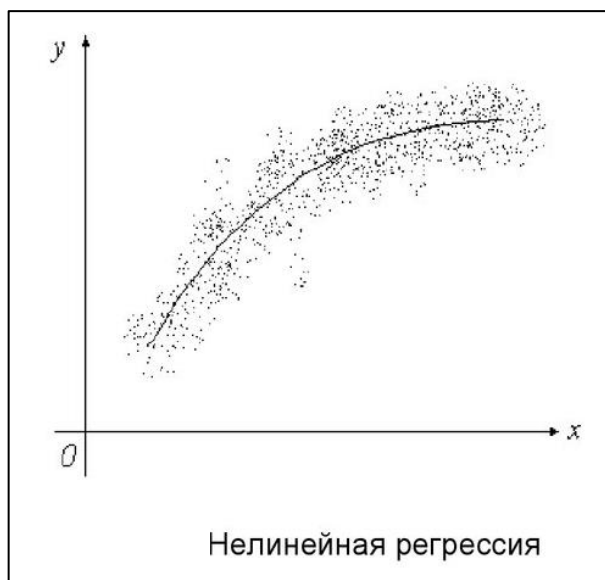


Изображение 5. Логистическая регрессия

Логистическая регрессия широко используется в медицине, маркетинге и финансах для оценки вероятности наступления того или иного события. Например, логистическая регрессия может быть использована для определения вероятности дефолта заемщика на основе его кредитной истории.

- **Нелинейные модели регрессии**

В реальной жизни многие процессы трудно описать с помощью линейных моделей. В таких случаях используются методы нелинейной регрессии. Примерами нелинейных моделей являются полиномиальная регрессия (например, квадратичная или кубическая), экспоненциальные модели и логистические кривые. Пример:



Изображение 6. Нелинейная регрессия

Эти модели позволяют объяснить сложные нелинейные зависимости между переменными и полезны в различных областях науки и техники.

- **Практическое применение регрессионного анализа в больших данных**

В эпоху больших данных регрессионный анализ играет важную роль в анализе и обработке огромных объемов информации. В этом разделе представлены некоторые применения регрессионных моделей в больших данных:

1. Прогнозирование спроса: предприятия используют регрессионные модели для прогнозирования спроса на товары и услуги на основе исторических данных о продажах, ценах, погоде и других факторах.

2. финансовый анализ: банки и финансовые учреждения используют регрессионный анализ для оценки кредитного риска клиентов, прогнозирования доходности инвестиций и управления портфелями активов

3. медицина: врачи и исследователи используют регрессионные модели для выявления факторов риска развития заболеваний, прогнозирования результатов лечения и разработки новых лекарств.

4. маркетинг: маркетологи используют регрессию для анализа эффективности рекламных кампаний, сегментирования рынков и оптимизации маркетинговых стратегий.

5. Интернет вещей (IoT): регрессионные модели могут использоваться для анализа данных с датчиков IoT, мониторинга состояния оборудования, прогнозирования отказов и оптимизации производственных процессов.

### **Временные ряды и прогнозирование**

Анализ и прогнозирование временных рядов - аспект статистики и машинного обучения, используемый для предсказания будущих значений на основе прошлых данных. Временные ряды характеризуются тем, что они включают наблюдения, проводимые непрерывно в течение времени, и обычно включают такие элементы, как тенденции, сезонность и цикличность.

- **Основы анализа временных рядов**

Временной ряд - это серия наблюдений, проводимых через регулярные промежутки времени. Временной ряд состоит из нескольких элементов:

1. тенденция - направление долгосрочных изменений в ряду, отражающее общую тенденцию к увеличению или уменьшению значений.

2. сезонность - периодическое повторение закономерностей через определенные временные интервалы (например, ежедневные, еженедельные, ежемесячные, ежегодные колебания).

3. период - нерегулярный элемент, связанный с экономическим или природным циклом, продолжительность которого превышает один год.

4. шум - случайные колебания, которые не могут быть объяснены другими факторами.

Чтобы проанализировать временной ряд, выполняются следующие действия:

- Графическое исследование - визуализация временного ряда для выявления тенденций, сезонности и циклов.
- Проверка стационарности - проверка постоянства среднего и дисперсии временного ряда.
- Автокорреляция и частичная автокорреляция - анализ зависимости текущих значений от предыдущих.
- Спектральный анализ - исследование частотной характеристики временного ряда.

- Популярные методы прогнозирования

Модель ARIMA (Auto Regressive Integrated Moving Average) - это статистическая модель, используемая для анализа и прогнозирования временных рядов. Она состоит из трёх компонентов: AR (авторегрессионная средняя), I (интегральная средняя) и MA (скользящая средняя). AR отражает зависимость текущего значения ряда от предыдущего значения, I - это дифференциальная операция для устранения тренда и достижения стационарности, а MA моделирует зависимость от случайной ошибки предыдущего периода. Параметры модели ARIMA выражаются в виде  $(p,d,q)$ , где  $p$  - порядок авторегрессии,  $d$  - порядок производной, а  $q$  - порядок скользящих средних. Модели ARIMA эффективно используются для

прогнозирования того, когда данные показывают тенденции или сезонные колебания, но они требуют предварительной обработки, чтобы привести их в устойчивый вид.

Существует также метод Холта-Уинтерса, метод прогнозирования временных рядов, используемый для объяснения тенденций и сезонности данных. Он содержит три компонента: уровень, тренд и сезонность, и позволяет делать прогнозы для временных рядов с явной сезонной компонентой. Суть метода заключается в обновлении этих компонентов на основе наблюдаемых данных с использованием средневзвешенного значения.

Есть два варианта этого метода: аддитивный и мультипликативный. В аддитивной модели амплитуда сезонных колебаний постоянна, в то время как в мультипликативной модели амплитуда сезонных колебаний изменяется пропорционально уровню ряда. Метод Холта-Уинтерса использует три основных параметра —  $\alpha$  (гладкость уровня),  $\beta$  (гладкость тренда) и  $\gamma$  (гладкость сезонности), которые настраиваются для минимизации ошибки прогноза. Это позволяет методу гибко реагировать на изменения в данных, обеспечивая точные прогнозы для временных рядов с трендами и сезонными колебаниями.

### **Ограничения и перспективы развития**

- Проблемы, возникающие при использовании статистических методов в анализе больших данных
- Масштабируемость. Статистические методы традиционно разрабатывались для работы с относительно небольшими наборами данных. Когда объемы данных становятся очень большими (петабайты и терабайты), многие классические методы (например, линейная регрессия, анализ временных рядов) становятся трудными или невозможными для применения без модификаций.

- Проблемы с качеством данных. В больших наборах данных часто присутствуют ошибки, пропуски, выбросы и другие аномалии, что снижает точность статистических моделей. К тому же данные могут быть негармонизированными, собранными с разных источников, что создаёт дополнительные проблемы.
- Необходимость в вычислительных ресурсах. Обработка и анализ больших данных требуют значительных вычислительных мощностей. Для выполнения статистических анализов, таких как многократные прогонные модели или обучение на огромных наборах данных, необходимо использование мощных серверов, кластеров или распределенных систем, что делает такие подходы дорогостоящими.
- Интерпретируемость моделей. Сложные статистические модели, такие как нейронные сети или случайные леса, могут дать хорошие прогнозы, но зачастую сложно интерпретировать их результаты. В реальных приложениях важно не только получить предсказание, но и объяснить, почему модель пришла к определённому выводу.
- Возможные пути решения этих проблем

Использование распределенных и параллельных вычислительных платформ (например, Hadoop, Spark) позволяет значительно ускорить обработку больших объемов данных. Статистические методы и алгоритмы должны быть адаптированы для работы в таких средах. Например, были разработаны распределенные версии алгоритмов кластеризации, регрессии и других методов. Для работы с высокоразмерными данными были разработаны методы уменьшения размерности, такие как анализ главных компонент (PCA) и t-SNE.

Также активно развиваются методы машинного обучения, такие как регуляризация (например, Lasso и Ridge), которые позволяют обрабатывать большое количество признаков и снижают риск перебора. Для динамических потоковых данных были разработаны алгоритмы онлайн-обучения, которые

обновляют модели по мере поступления новых данных, не перерабатывая весь набор данных. Это делает статистические методы более адаптивными и эффективными в условиях изменения данных во времени.

Одной из актуальных задач является разработка интерпретируемых моделей в машинном обучении. Методы объяснимого ИИ (Explainable AI, XAI) помогают понимать, почему модель приняла те или иные решения. Используются такие подходы, как локальные интерпретируемые модели (LIME), Shapley values и другие.

- Перспективы дальнейшего развития статистических методов в условиях роста объёмов данных

#### 1. Автоматизация анализа данных.

Одна из главных перспектив - развитие технологий автоматизации анализа данных. К ним относится автоматическое обучение моделей (AutoML), при котором система автоматически выбирает оптимальный алгоритм и настраивает параметры, снижая потребность в экспертах и упрощая применение статистических методов.

#### 2. Гибридные подходы с использованием статистики и машинного обучения.

В будущем классические статистические методы будут интегрированы с новыми подходами машинного обучения. Это позволит создавать более мощные и адаптивные модели, способные эффективно работать как с малыми, так и с большими массивами данных. По мере того как ИИ и статистические методы находят все большее применение в различных областях, все большее значение приобретают интерпретируемость и этичность моделей. Будут разработаны новые подходы для объяснения алгоритмических решений и обеспечения соблюдения этических и правовых требований, особенно в таких областях, как медицина, финансы и право.

#### 3. Интеграция с реальным временем и новые источники данных.

Прогнозирование и анализ потоковых данных в реальном времени будут приобретать все большее значение. Статистические методы будут развиваться,

чтобы обрабатывать непрерывно поступающие данные, открывая новые возможности для мониторинга, прогнозирования и принятия решений в режиме реального времени. В будущем будут развиваться методы для работы с новыми источниками данных, такими как сенсоры IoT, генетические данные, видео- и аудиофайлы, что откроет новые горизонты для применения статистики в таких областях, как умные города, здравоохранение, автономные системы и другие.

### **Заключение**

Анализ больших данных с помощью статистических методов - главный инструмент для извлечения ценной информации из огромных массивов данных, с которыми сталкиваются современные организации и исследовательские проекты.

Особое внимание уделяется важности хорошей визуализации данных, которая полезна не только для интерпретации результатов, но и для эффективной коммуникации с заинтересованными сторонами. Статистические методы, такие как регрессионный анализ, методы прогнозирования и анализ временных рядов, доказали свою эффективность в решении широкого круга задач - от предсказания будущих тенденций до улучшения процессов принятия решений в режиме реального времени.

Стоит отметить, что, несмотря на значительные достижения в области обработки больших данных, перед исследователями и практиками по-прежнему стоят серьезные задачи. Среди них - проблемы качества данных, ограниченность вычислительных ресурсов и разработка новых алгоритмов, способных эффективно обрабатывать еще более крупные и сложные данные.

Перспективы развития аналитики больших данных связаны с дальнейшим совершенствованием существующих инструментов и алгоритмов, развитием технологий искусственного интеллекта и машинного обучения, а также ростом возможностей обработки данных в режиме реального времени. В условиях стремительного технологического прогресса важно продолжать

поиск и адаптацию новых методов для максимального извлечения знаний из данных и обеспечения устойчивого развития в экономической, научной и технологической сферах.

### Список литературы

1. big data большие данные / [Электронный ресурс] // : [сайт]. — URL: <https://dspace.www1.vlsu.ru/bitstream/123456789/9407/1/02292.pdf>
2. К вопросу статистического анализа больших данных / [Электронный ресурс] // : [сайт]. — URL: <https://cyberleninka.ru/article/n/k-voprosu-statisticheskogo-analiza-bolshih-dannyh>
3. Методы и техники анализа больших данных / [Электронный ресурс] // : [сайт]. — URL: <https://proglib.io/p/big-data-metody-i-tehniki-analiza-bolshih-dannyh-2021-08-31>
4. BigData: анализ больших данных сегодня / А. В. Веретенников. — Текст : непосредственный // URL: <https://moluch.ru/archive/166/45354>
5. Что такое аналитика больших данных / [Электронный ресурс] // : [сайт]. — URL: <https://azure.microsoft.com/ru-ru/resources/cloud-computing-dictionary/what-is-big-data-analytics>