



**МИНИСТЕРСТВО СЕЛЬСКОГО ХОЗЯЙСТВА РОССИЙСКОЙ ФЕДЕРАЦИИ**  
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ  
**«РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ АГРАРНЫЙ УНИВЕРСИТЕТ**  
**– МСХА имени К.А. ТИМИРЯЗЕВА»**  
**(ФГБОУ ВО РГАУ - МСХА имени К.А. Тимирязева)**

---

## **Открытый Мир. Старт в науку**

Тема работы: «Визуализация результатов анализа данных»

Выполнил  
студент 1 курса колледжа  
группы ТК-И-11-11  
Сары Максим Валерьевич

Научный руководитель:  
ассистент кафедры  
статистики и кибернетики РГАУ-  
МСХА имени К.А. Тимирязева  
Ульянкин Александр Евгеньевич

Москва,  
2023

## Оглавление

|   |    |
|---|----|
| Введение.....   | 3  |
| Глава 1 Теоретические аспекты методов визуализации больших данных ..... | 4  |
| Глава 2 Основные методы визуализации больших данных в python .....      | 7  |
| Заключение .....  | 14 |
| Список использованной литературы.....                                   | 15 |

## Введение

Визуализация (от лат. visualis, "зрительный") – общее название приёмов представления числовой информации или физического явления в виде, удобном для зрительного наблюдения и анализа.

В компьютерной графике визуализацией называют процесс получения изображения по модели.

Визуализация данных – это наглядное представление массивов различной информации.

Визуальная информация лучше воспринимается и позволяет быстро и эффективно донести до зрителя собственные мысли и идеи. Физиологически, восприятие визуальной информации является основной для человека. Есть многочисленные исследования, подтверждающие, что:

- 90% информации человек воспринимает через зрение
- 70% сенсорных рецепторов находятся в глазах
- около половины нейронов головного мозга человека задействованы в обработке визуальной информации
- на 19% меньше при работе с визуальными данными используется когнитивная функция мозга, отвечающая за обработку и анализ информации
- на 17% выше производительность человека, работающего с визуальной информацией
- на 4,5% лучше вспоминаются подробные детали визуальной информации.

# **Глава 1 Теоретические аспекты методов визуализации больших данных**

Визуализация данных относится к областям как научной, так и информационной визуализации. В первом случае данные возникают в результате сложного компьютерного моделирования различных объектов и процессов. Во втором - имеет место визуальное описание и представление абстрактной информации, получаемой в результате процесса сбора и обработки многокатегориальных данных, для анализа которых необходимо применение нескольких количественных и качественных мер оценки.

Работы, описывающие результаты по визуализации больших данных, появились сравнительно недавно. Среди них можно выделить «Белую книгу» компании Intel, посвященную визуализации результатов «большого счета», “установочную” публикацию известного специалиста в области компьютерной визуализации и человеко-компьютерного взаимодействия Б. Шнейдермана и введение к спецвыпуску, посвященному визуализации больших данных.

Среди задач визуализации данных рассматриваются следующие:

- визуализация потоков данных;
- визуальный интеллектуальный анализ данных (Visual data mining);
- визуальный поиск и рекомендации (Visual search and recommendation);
- описание ситуаций на основе больших данных с использованием визуализации (Big data storytelling using visualization);
- масштабируемые методы параллельной визуализации;
- современные аппаратные средства и архитектуры для анализа и визуализации данных;
- человеко-компьютерный интерфейс и визуализация больших данных;

- приложения визуализации больших данных, включая кибер разведку и контрразведку, бизнес-анализ (бизнес разведку), электронную коммерцию, анализ научной информации, образование и т.д.

Выбирая наиболее подходящий вид графика для визуализации данных, следует, прежде всего, определить **цель анализа** и/или представления информации, например:

- сравнить разные показатели;
- продемонстрировать распределение данных – какие значения встречаются чаще или реже других;
- показать состав и структуру;
- выявить взаимосвязи между переменными.

Для этих целей используется более 20 видов различных диаграмм, от линейных графиков до корреляционных матриц. Выбор конкретной диаграммы для визуализации данных также зависит от числа анализируемых переменных и временных периодов (рис. 1).

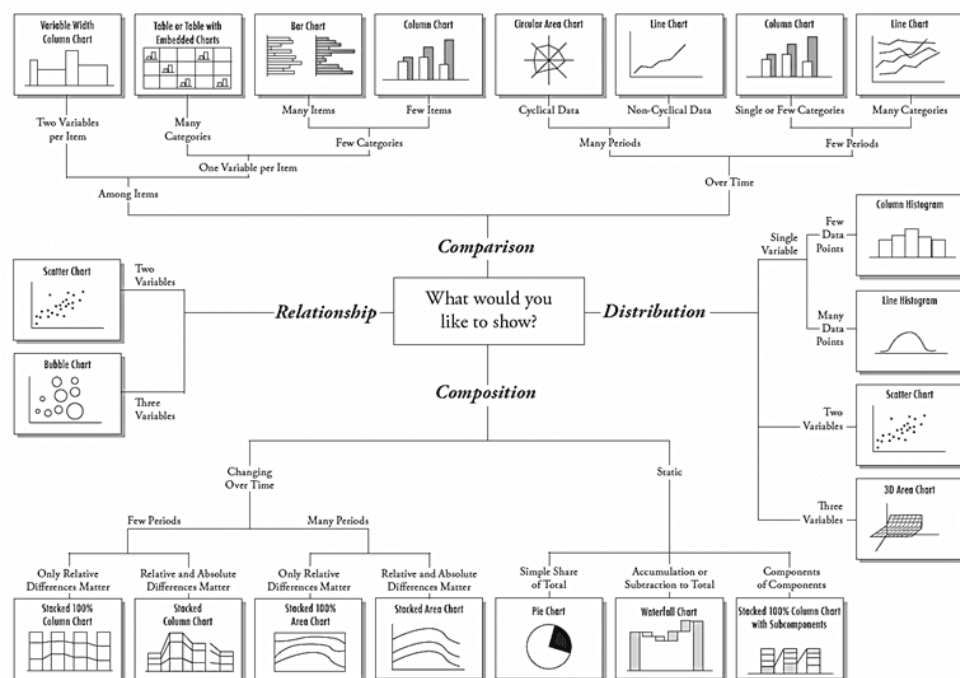


Рисунок 1 – Многообразие диаграмм для визуализации данных

Наглядное представление результатов анализа больших данных имеет принципиальное значение для их интерпретации. Не секрет, что восприятие человека ограничено, и ученые продолжают вести исследования в области

совершенствования современных методов представления данных в виде изображений, диаграмм или анимаций.

На практике в Data Science наиболее часто используются следующие виды графиков и диаграмм:

**Гистограмма** для визуализации распределения данных в рамках непрерывного интервала или ограниченного периода времени, определения концентрации значений, а также выявления предельных показателей, пропусков или отклонений;

**Диаграмма рассеяния** для выявления корреляции между двумя переменными;

**Диаграмма размаха** (ящик с усами) для отображения групп числовых данных через квартили, что удобно при сравнении распределений между большим количеством датасетов;

**Тепловая матрица** для многомерного анализа данных и выявления корреляций;

**Пузырьковая диаграмма** для сравнения и отображения взаимосвязей между разными переменными с помощью их местоположения и пропорций – часто используется для анализа паттернов и поиска корреляций;

Правильно выбранный вид диаграммы для визуализации данных соответствует следующим критериям:

- **краткость** – возможность одновременно отобразить много разнотипных данных;

- **относительность и близость** – способность демонстрировать кластеры, относительные размеры групп, их схожесть и различие, выпадающие значения;

- **концентрацию и контекст** – возможность легко и оперативно взаимодействовать с выбранным объектом путем его интерактивного просмотра (отображение структуры и связей);

- **масштабируемость** – возможность легко и быстро изменять размеры представления;

•**удобство пользователя** за счет максимальной наглядности предоставления и поддержка интуитивных действий по выявлению закономерностей.

## Глава 2 Основные методы визуализации больших данных в python

Наверное, самый привычный для нас вид визуализации данных. Именно графики мы видим в учебниках в школе, с ними же первым делом знакомимся, когда начинаем осваивать Excel.

Точечные и линейные графики являются наиболее распространенными формами графического представления данных, обеспечивая визуальное представление функции  $y = f(x)$ , определенной набором точек  $(x, y)$ . Точечные графики просто показывают точки данных, в то время как линейные графики соединяют их или интерполируют для определения непрерывной функции  $f(x)$ .

Для наглядной демонстрации возможностей языка программирования Python для визуализации данных был использован датасет о марках автомобилей, включающей в себя следующую информацию:

model – модель автомобиля

year – год регистрации

price – цена в фунтах стерлингов

transmission – тип коробки передач

mileage – пробег

fuelType – моторное топливо

tax – дорожный налог

mpg – миля на галлон

engineSize – размер двигателя в литрах

**Линейные графики** используются для отображения количественных показателей за непрерывный интервал или определенный период времени.

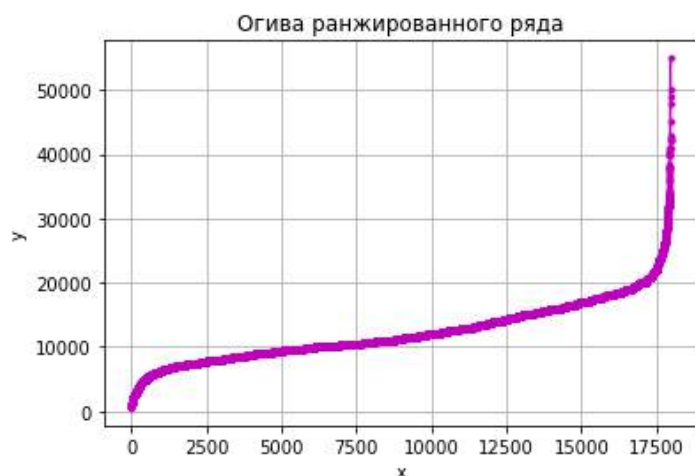


Рисунок 2 – Огива ранжированной цены

По данному графику видно, что наименьшая цена на автомобиль марки ford составила примерно 495 фунтов стерлингов, а максимальная примерно 54 995 фунтов стерлингов.

**Точечный график** передает порядок ранжирования предметов. А поскольку он выровнен вдоль горизонтальной оси, можно визуальнo оценить, как далеко точки находятся друг от друга.



Рисунок 3 – Точечный график

На графике видно, что Mustang стоит дороже всего и очень сильно разнится с ценой Edge.

**Диаграммы размаха** («ящик с усами») (Box and Whisker Plot или Box Plot) – это удобный способ визуального представления групп числовых данных через квантили.



Виды наблюдений, которые можно сделать на основе ящика с усами:

- Каковы ключевые значения, например: средний показатель, медиана 25го перцентиля и так далее.
- Существуют ли выбросы и каковы их значения.
- Симметричны ли данные.
- Насколько плотно сгруппированы данные.
- Смещены ли данные и, если да, то в каком направлении.

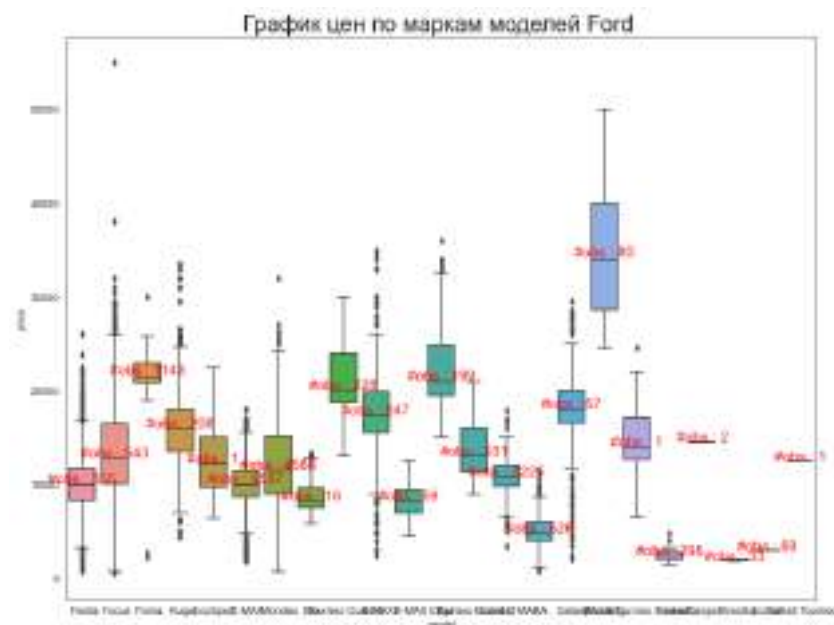


Рисунок 4 – Box Plot

Прямые линии, исходящие из ящика, называются «усами» и используются для обозначения степени разброса (дисперсии) за пределами верхнего и нижнего квартилей. Выбросы иногда отображаются в виде отдельных точек, находящихся на одной линии с усами.

На *диаграммах рассеяния* ряд точек, размещенных в декартовой системе координат, отображает значения по двум переменным. Присвоив каждой оси переменную, можно определить, существуют ли отношения или корреляция между этими двумя переменными.



Рисунок 5 – Диаграмма рассеяния

На данной диаграмме рассеяния видна зависимость цены от пробега, связь отрицательна. Цвет точек различен и обозначает модели марки Ford.

**Диаграмма корреляции** используется для визуального просмотра метрики корреляции между всеми возможными парами числовых переменных в данном наборе данных (или двумерном массиве).

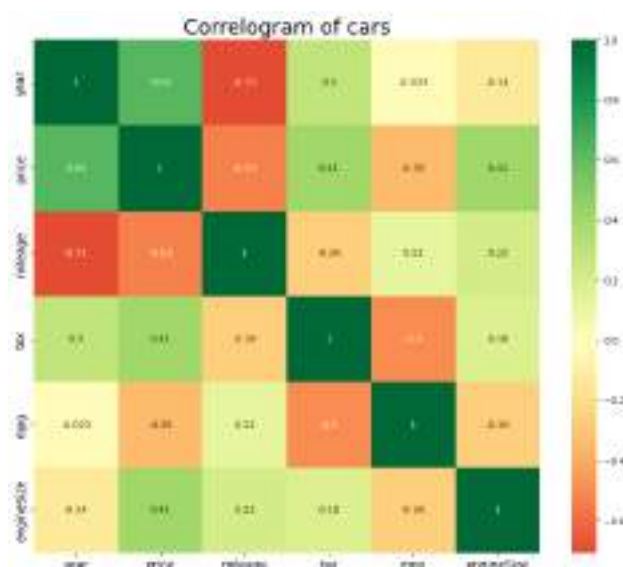


Рисунок 6 – Коррелограмм

На данном рисунке показана взаимосвязь между показателями. Так видно, что связь между ценой и пробегом обратная умеренная.

**Пузырьковая диаграмма** – это многомерный график, который находится на пересечении диаграммы рассеяния и **Диаграмма с пропорциональными областями**. Пузырьковые диаграммы, как правило, используются для

сравнения и отображения взаимосвязей между отмеченными/классифицированными окружностями с помощью определения их местоположения и пропорций.

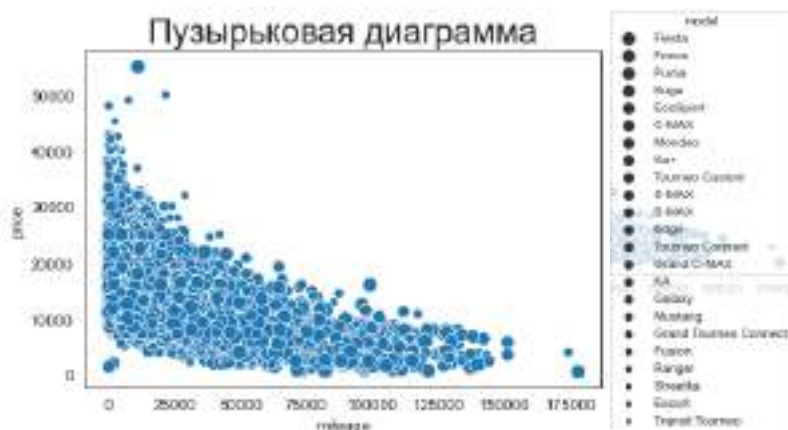


Рисунок 7 – Пузырьковая диаграмма

На данной диаграмме отображена связь цены и пробега. Размер пузырьков зависит от модели.

**Столбчатая диаграмма** эффективно передает порядок ранжирования элементов. Но, добавив значение показателя над диаграммой, пользователь получает точную информацию от самой диаграммы.

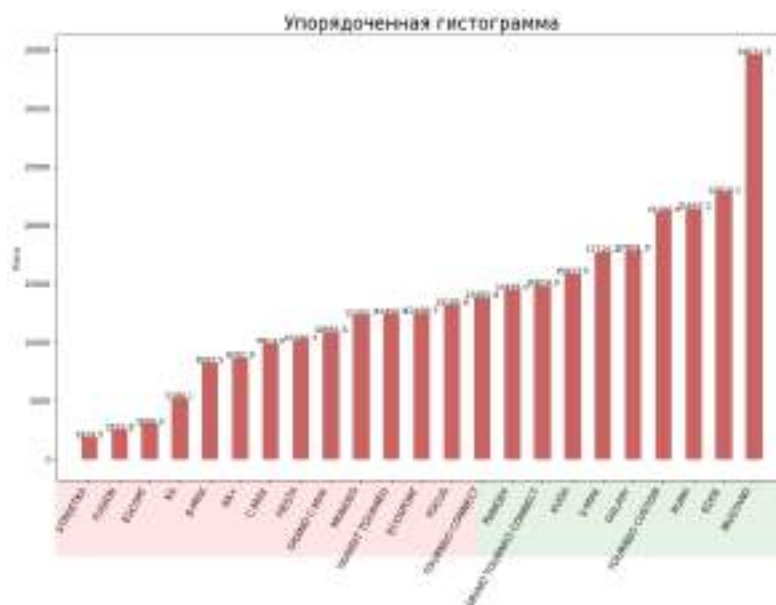


Рисунок 8 – столбчатая диаграмма

На гистограмме видно распределение моделей марки Ford по возрастанию цены. Красным отмечены модели ниже средней цены.

**Гистограмма** визуализирует распределение данных в рамках непрерывного интервала или ограниченного периода времени. Гистограммы помогают определить концентрацию значений, предельные значения и наличие пробелов или отклонений.

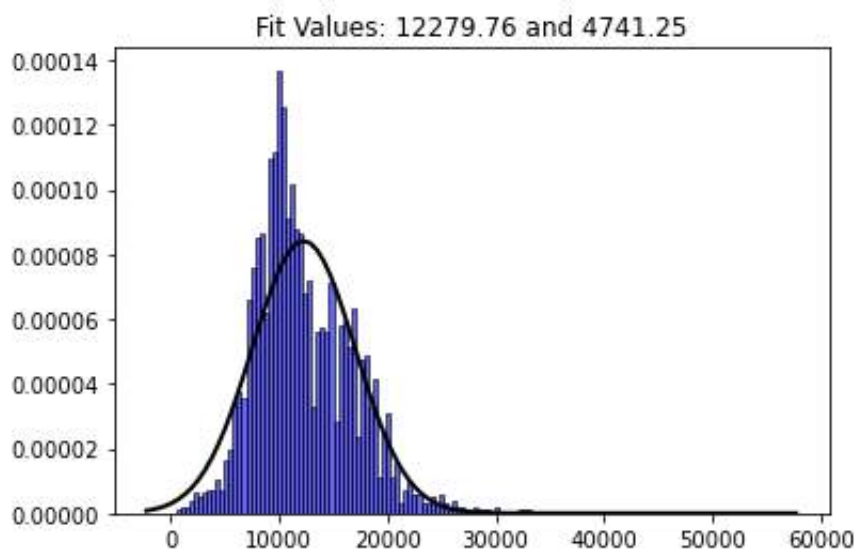


Рисунок 9 – Гистограмма с диаграммой нормального распределения

Максимальное значение в одном интервале – 993. Также по диаграммам видно, что имеются нулевые интервалы. Распределение не соответствует нормальному закону распределения.

**Круговые диаграммы** широко используются в презентациях и офисной документации. Они позволяют показать пропорциональное и процентное соотношение между категориями за счет деления круга на пропорциональные сегменты. Длина каждой дуги представляет собой пропорциональную долю каждой категории, в то время как круг целиком представляет общую сумму всех данных, равную 100%.

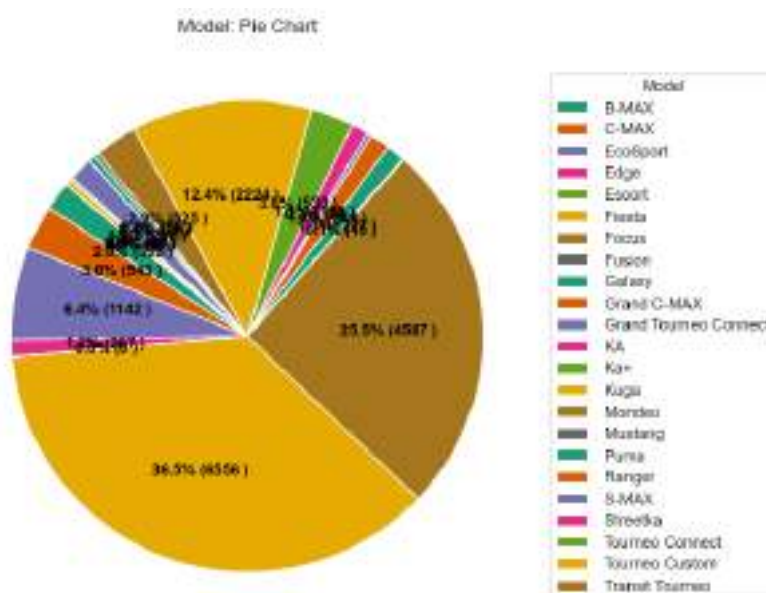


Рисунок 10 – Круговая диаграмма

Больше всего машин модели Fiesta около 36,5% всей совокупности. Из-за большого количества моделей визуально плохо видно кластеры и их подписи.

*Древовидная карта* похожа на круговую диаграмму и работает лучше, не вводя в заблуждение долю каждой группы.

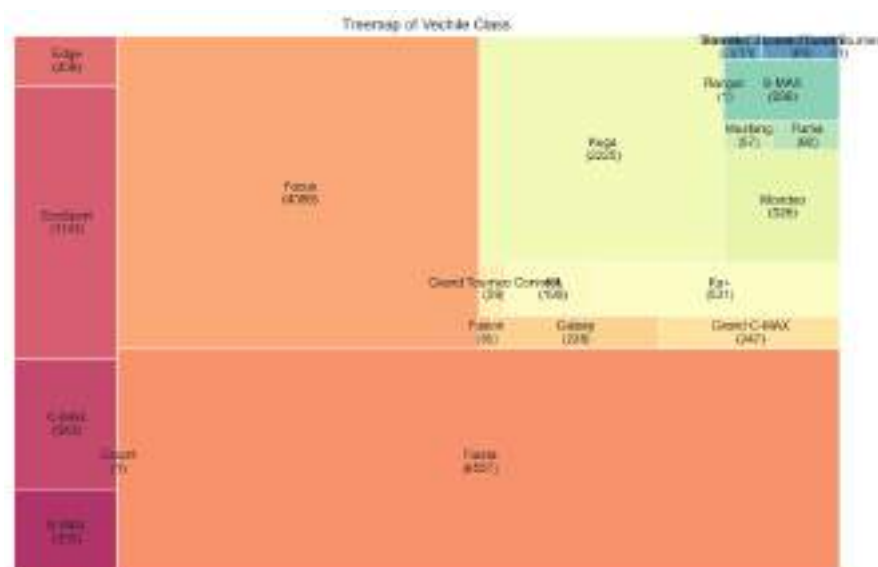


Figure 2 – Древовидная карта

Древовидная карта отобразила тоже само, что и круговая диаграмма, однако визуально она лучше отображает большое количество моделей маши.

## **Заключение**

Анализ и интерпретация результатов анализа требуют использования новых технологий компьютерной графики, сред виртуальной и расширенной реальности. Отсюда следует необходимость анализа влияния «человеческого фактора», проведения комплексных исследований не только с точки зрения компьютерных наук и математики, но и когнитивной психологии. Поэтому среди серьезных проблем нового направления можно указать на проблему «больших картинок» (big pictures), отображающих большие данные. Основные трудности здесь связаны с восприятием и интерпретацией сверхбольших объемов графической информации.

Проблемы визуализации данных порождают предельные на данный момент случаи, что требует комплексного решения проблем компьютерной графики и визуальной аналитики. Анализ зарубежных работ по визуализации больших данных за последние годы показывает, что они включают в себя целый ряд направлений компьютерной визуализации, среди которых научная и информационная визуализация, визуализация программного обеспечения, визуальный анализ данных, верификация и валидация визуализации, изучение восприятия и когнитивной составляющей визуализации при использовании «больших экранов» и сред виртуальной реальности. Необходим учет этого опыта при разворачивании аналогичных исследований в нашей стране.

## Список использованной литературы

1. Бахтерев М.О., Васёв П.А., Казанцев А.Ю., Манаков Д.В. Система удалённой визуализации для инженерных и суперкомпьютерных вычислений // Вестник ЮжУрГУ, N 17 (150), 2009, серия «Математическое моделирование и программирование», Выпуск 3. Стр. 4-11.
2. Васёв П.А. Среда поддержки интерактивной визуализации для суперкомпьютерных вычислений // Вопросы атомной науки и техники. Серия: Математическое моделирование физических процессов. 2009. Выпуск 4. Стр. 67- 77.
3. Zubov M.V., Pustygina A.N., Starceva E.V. Получение типов данных в языках с динамической типизацией для статического анализа исходного кода с помощью универсального классового представления // Вестн. Астрахан. гос. техн. ун-та. Сер. управление, вычисл. техн. информ., 2013, N 2, Стр. 66–74.
4. Михайлов И.О., Авербух В.Л. Современные методы визуализации больших и сверхбольших объёмных данных // XV Международная конференция «Супервычисления и Математическое Моделирование». Тезисы. ФГУП «РФЯЦ ВНИИЭФ». Саров. 2014, стр. 97-98.
5. <http://www.intel.com/content/dam/www/public/us/en/documents/white-papers/big-data-visualization-turning-big-data-into-big-insights.pdf>