

ВЫЧИСЛИТЕЛЬНЫЙ КЛАСТЕР МГИМО CL120

МГИМО
УНИВЕРСИТЕТ



В проектировании, строительстве и монтаже компьютерного класса в аудитории 120 принимали участие несколько коллективов, десятки людей.

Курировал проект Первый проректор МГИМО
Геннадий Петрович Толстопятенко.

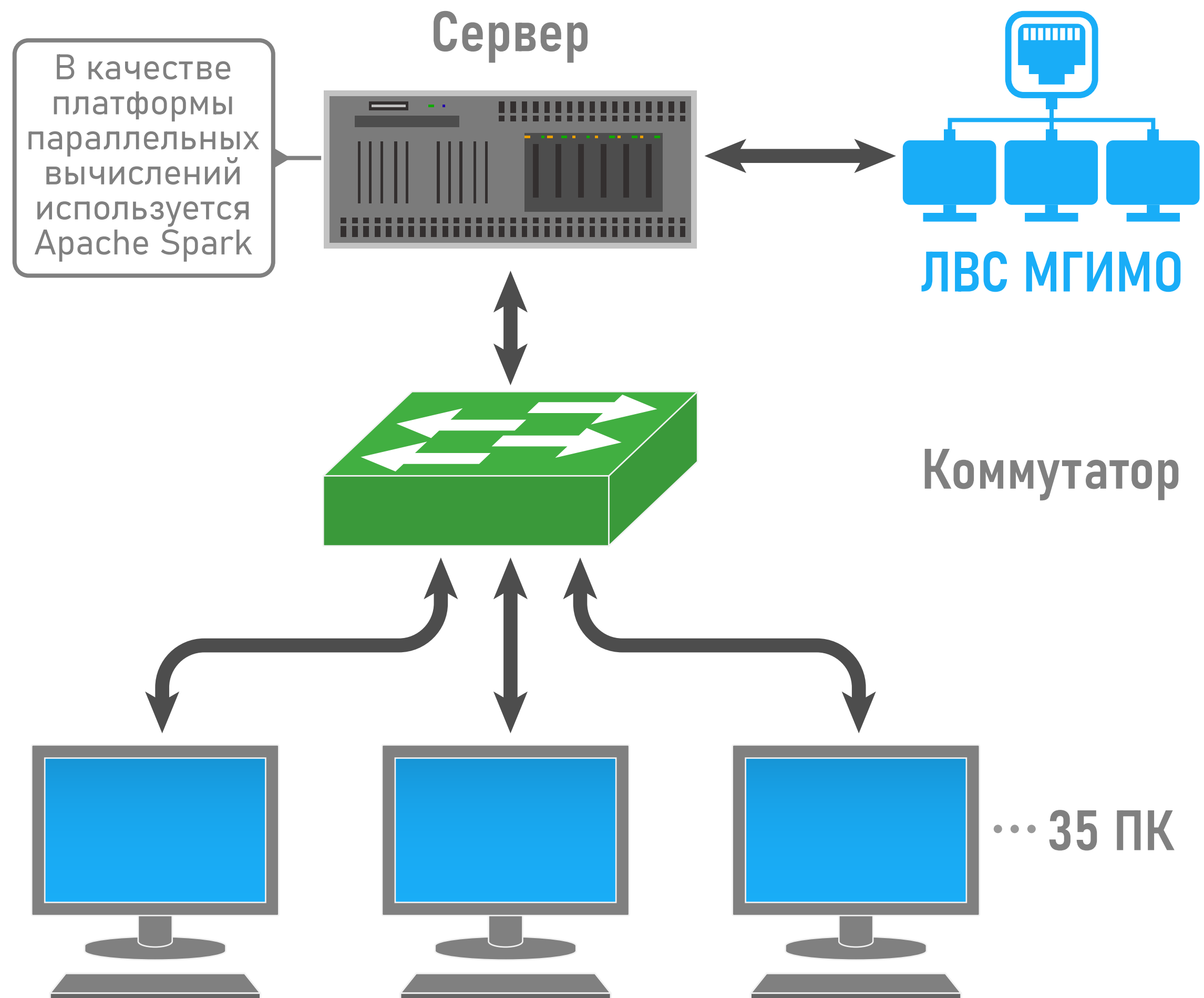
На базе компьютерного класса спроектировали и разработали
Вычислительный кластер в 2023 – 2025 гг.

А.Б. Немченко, М.М. Звягинцев, О.В. Щербаков, А. Слюсарь.

Курировал проект заведующий кафедрой МЭИТ факультета МЭО
Н.В. Артамонов.



КАФЕДРА
МАТЕМАТИКИ
ЭКОНОМЕТРИКИ и
ИНФОРМАЦИОННЫХ
ТЕХНОЛОГИЙ



Apache Spark — это фреймворк для параллельной обработки неструктурированных или слабо-структурированных данных.

Также его называют **фреймворком для кластерных вычислений**, так как для обработки данных используется группа компьютеров (кластер)

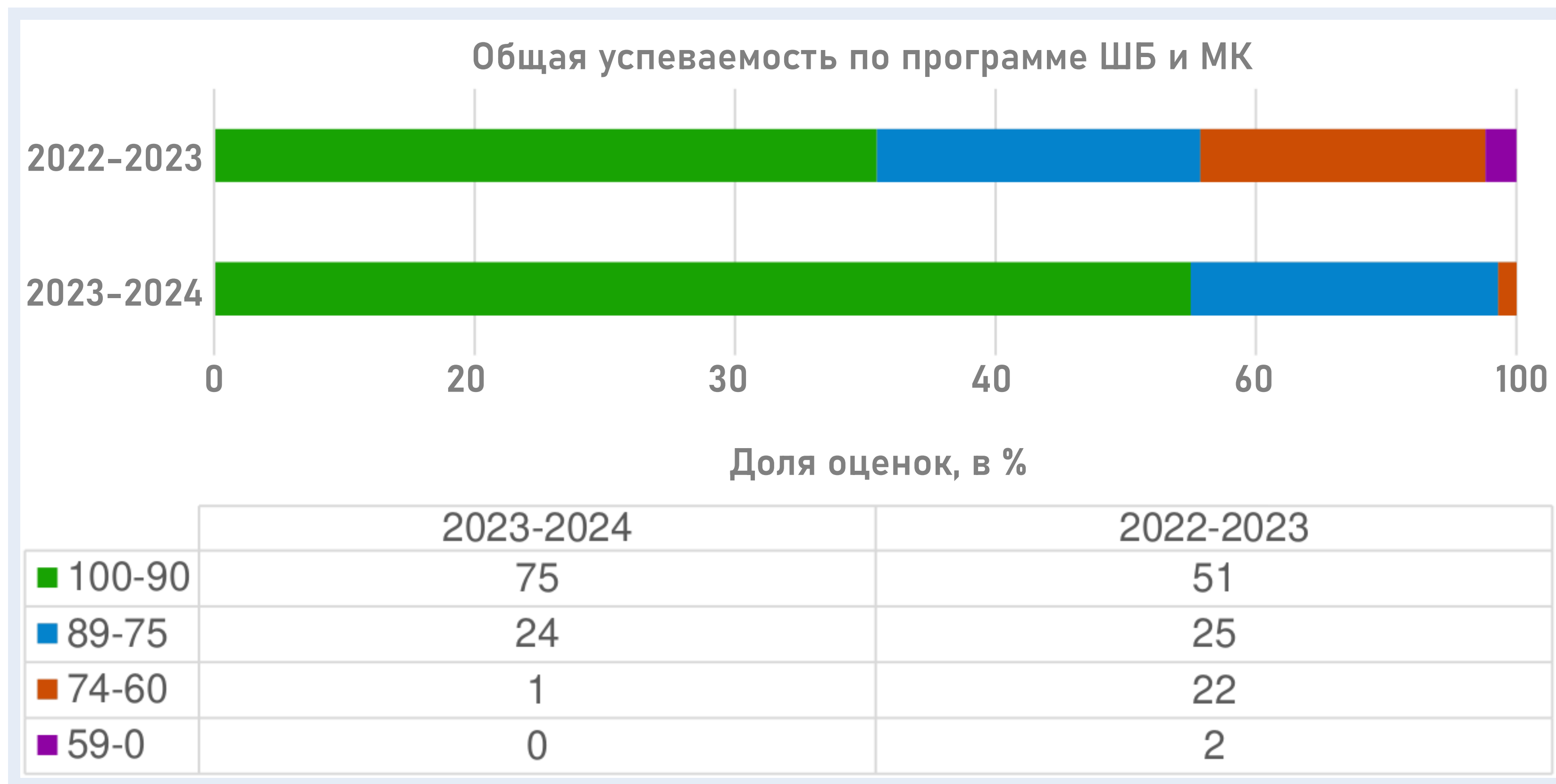


Spark поддерживает языки программирования:

- Scala,
- Java,
- Python,
- R
- SQL.

В структуру Spark входят:

- **Spark Core (ядро)** — движок, который отвечает за хранение данных, управление памятью, распределение и отслеживание задач в кластере;
- **MLlib** — набор библиотек для машинного обучения;
- **SQL** — компонент, который отвечает за запрос данных;
- **GraphX** — модуль для работы с графами (абстрактными представлениями связей между множеством объектов);
- **Streaming** — средство для обработки потоков Big Data в реальном времени.

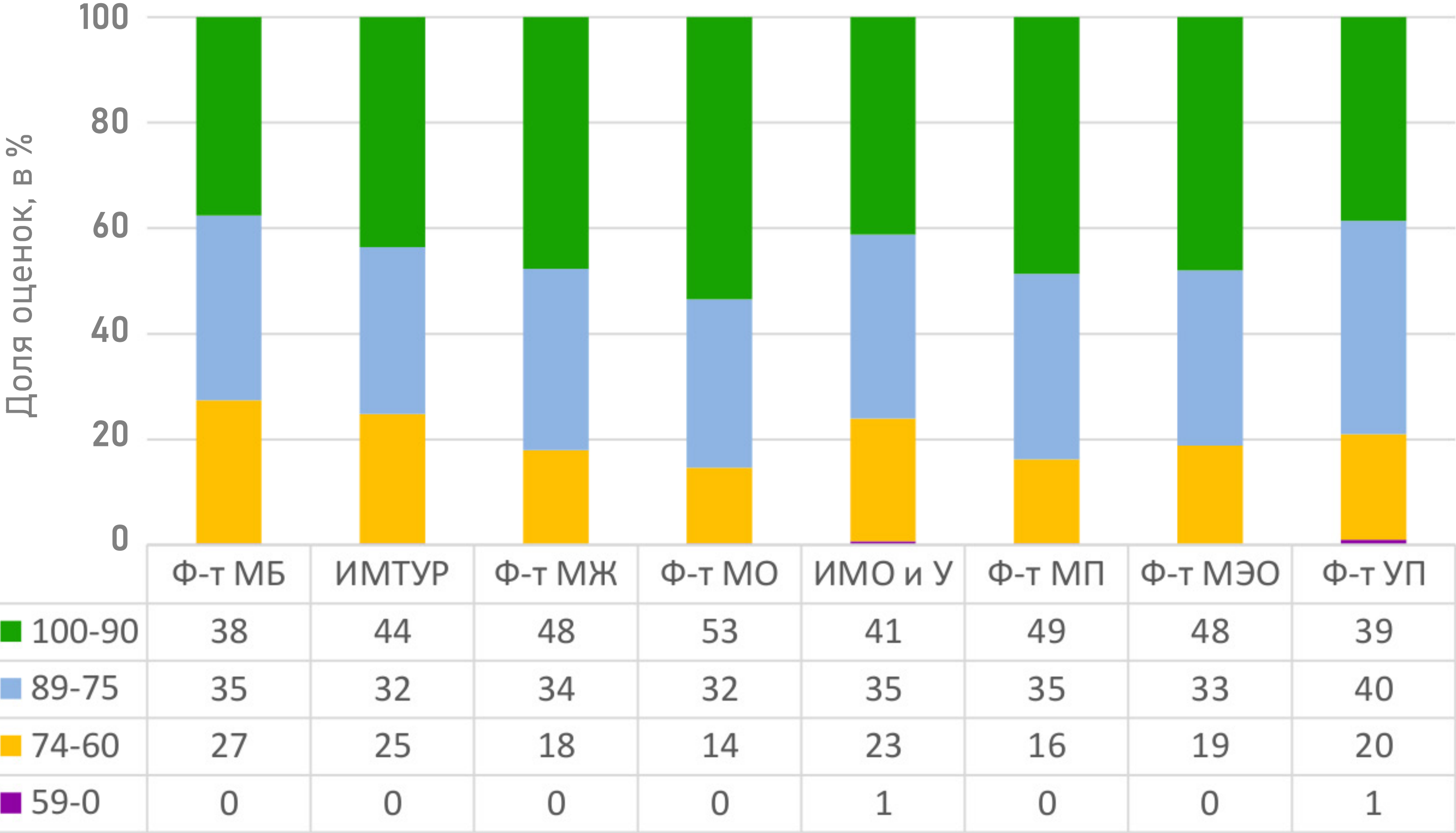


На данной, и на следующих диаграммах приведены иллюстрации работы вычислительного кластера с данными БОД.

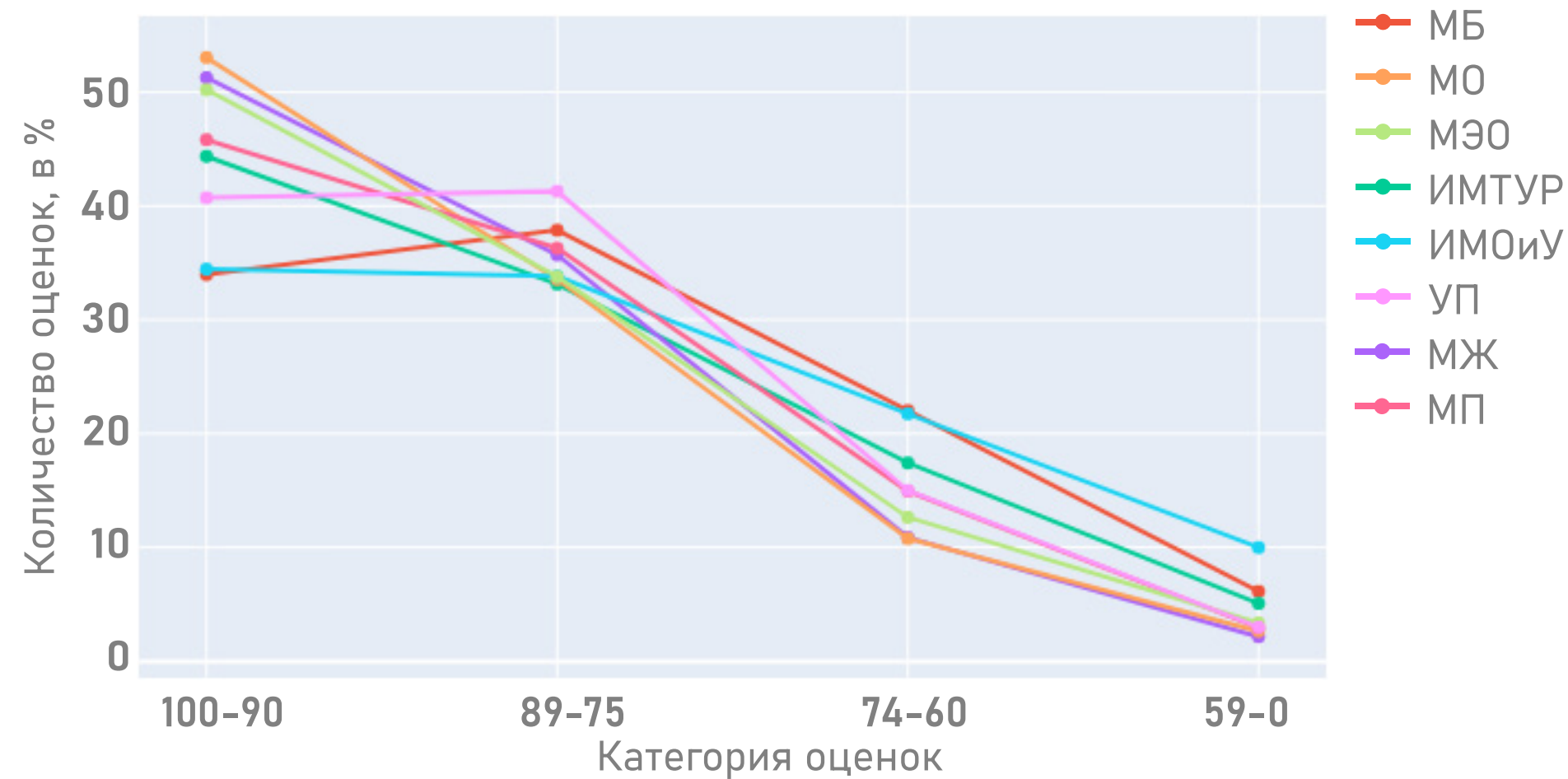
Продолжение на следующих слайдах.



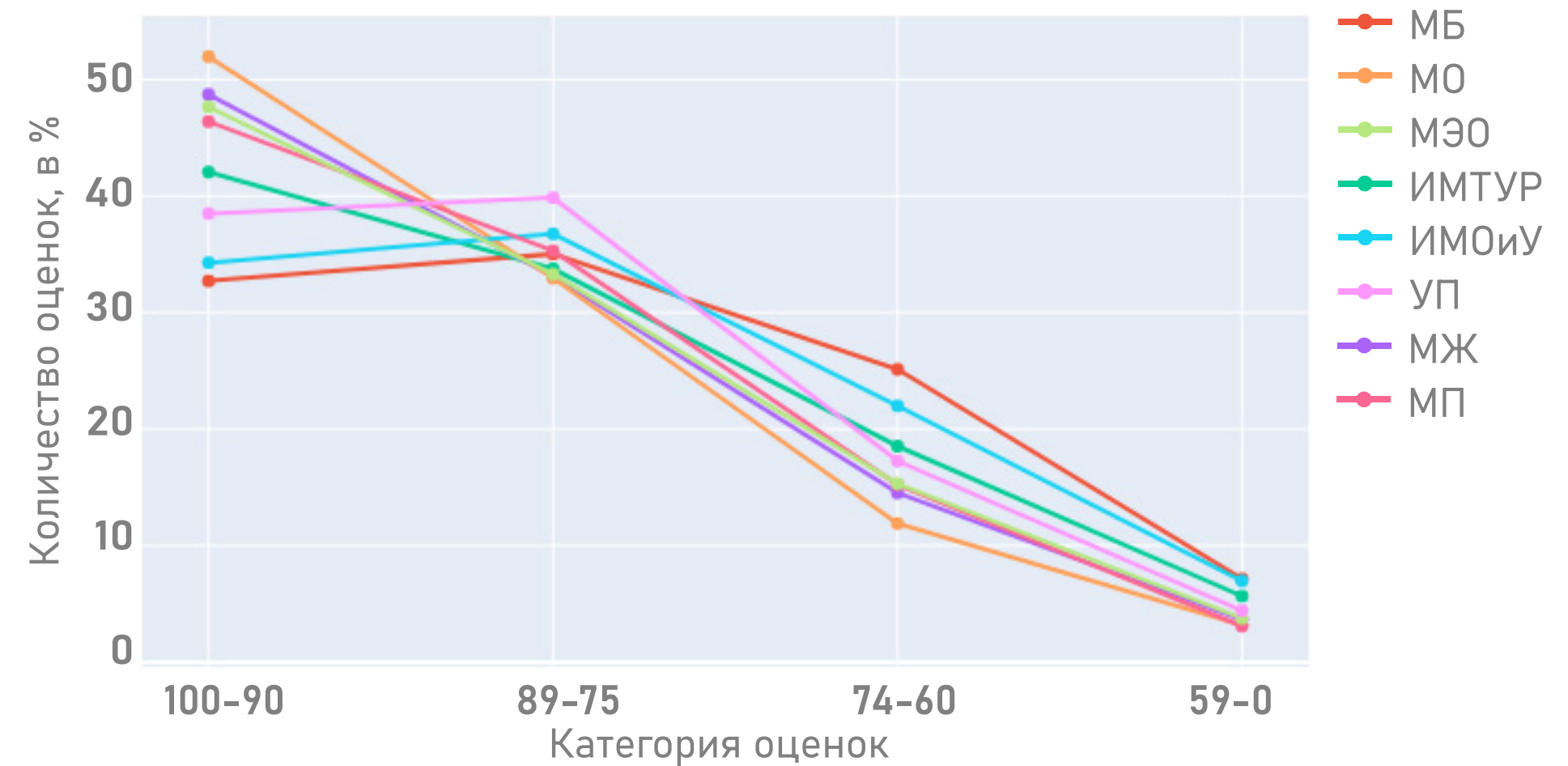
Общая успеваемость на факультетах бакалавриата в 2023–2024 уч. г.



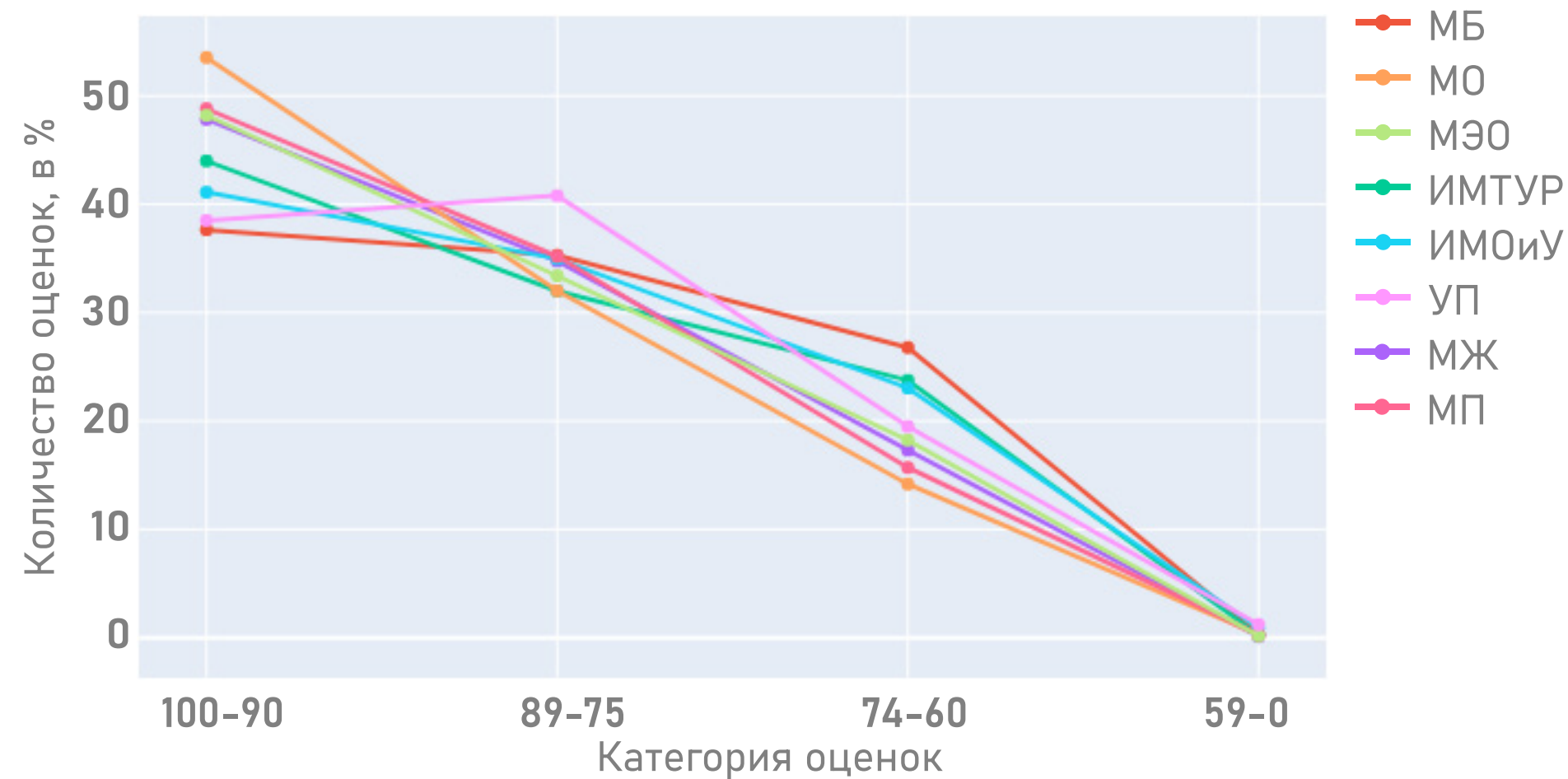
Количество оценок на факультетах бакалавриата за 2019-2020 год.



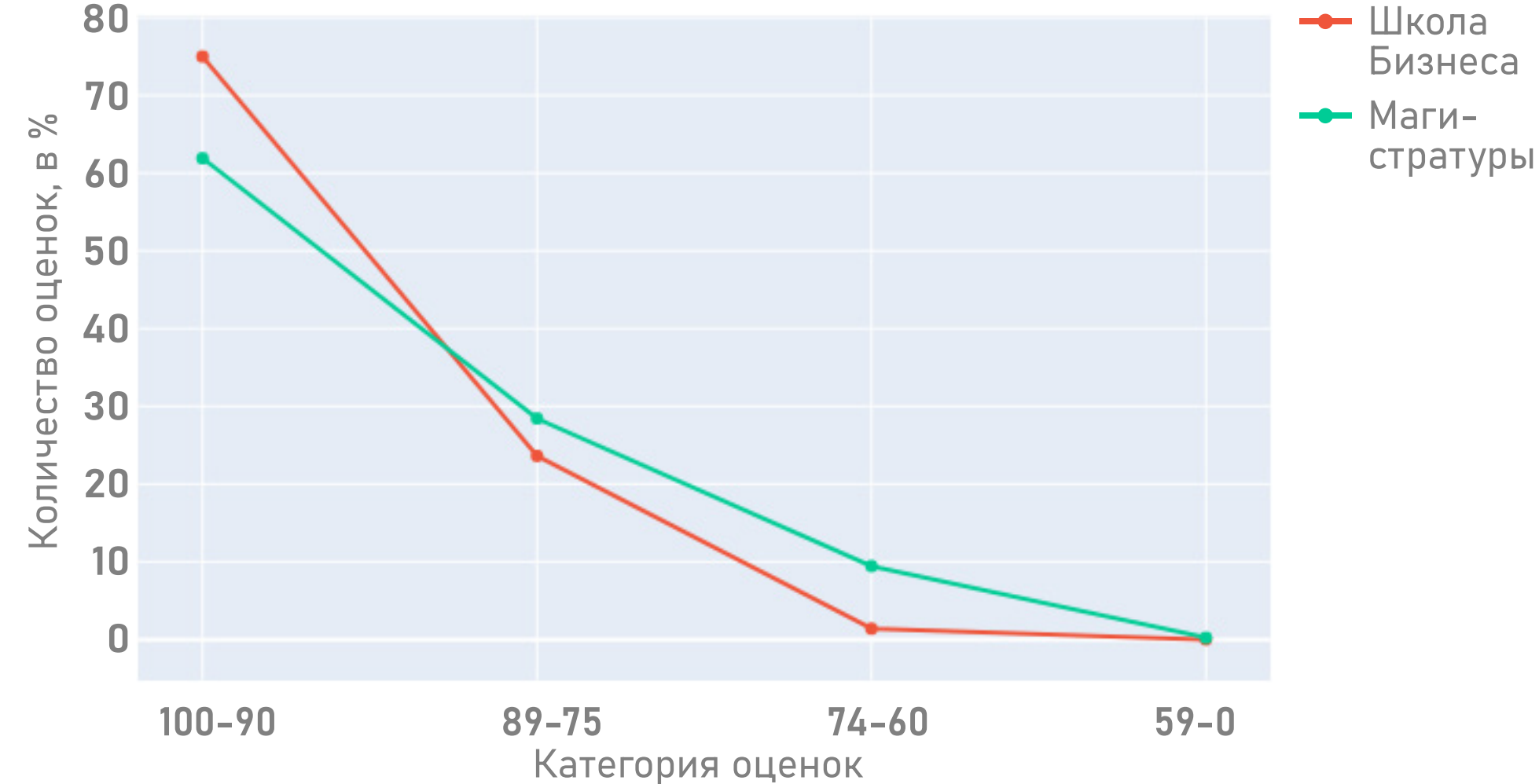
Количество оценок на факультетах бакалавриата за 2021-2022 год.



Количество оценок на факультетах бакалавриата за 2023-2024 год.



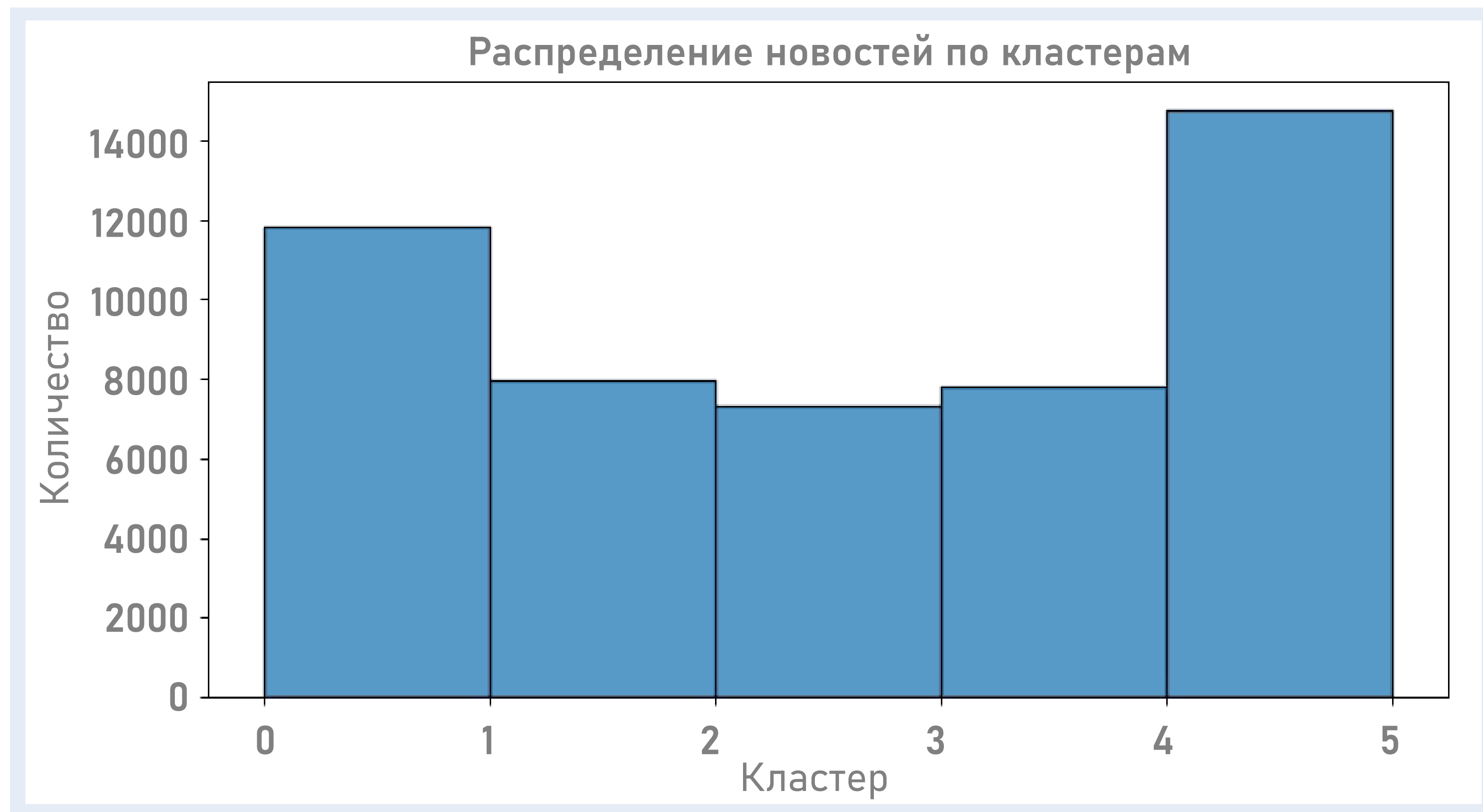
Количество оценок на факультетах магистратуры за 2023-2024 год.



Исследование БОД с помощью Кластера



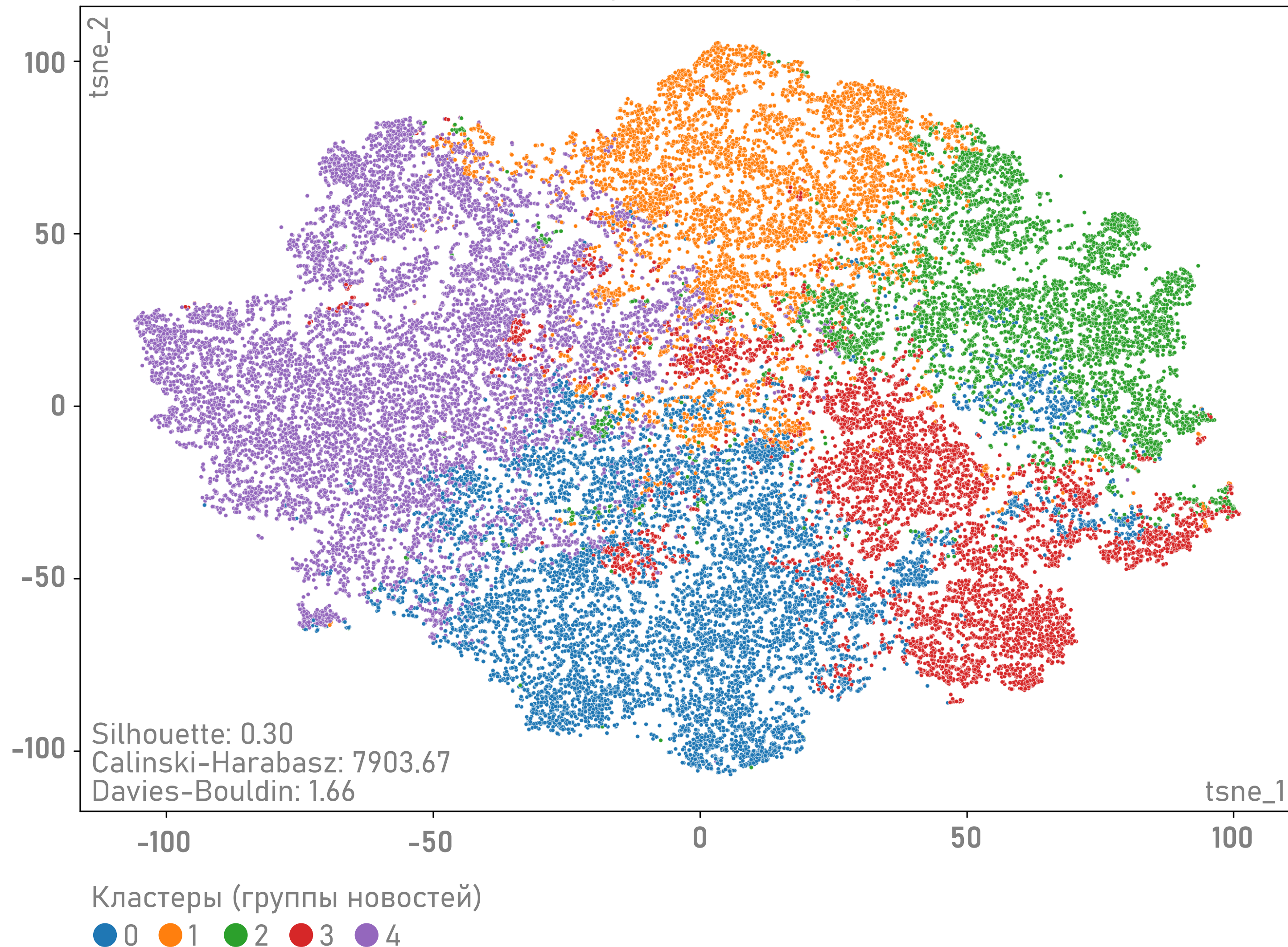
КАФЕДРА
МАТЕМАТИКИ
ЭКОНОМЕТРИКИ и
ИНФОРМАЦИОННЫХ
ТЕХНОЛОГИЙ



На диаграмме сравнивается количество новостей New York Times, рассортированных по пяти группам, обученной на кластере CL120 моделью, которую создал инженер лаборатории ЭВТ им. Чеповского Андрей Слюсарь.

Продолжение
на следующем слайде.

t-SNE визуализация кластеров



На рисунке - двухмерный срез многомерного вектора кластеризации (группировки) базы новостей на кластере CL120

Для повышения качества кластеризации предварительные результаты исследования были загружены в модель o3-mini-high на платформе ChatGPT.

После анализа результатов, моделью ChatGPT были предложены рекомендации, следуя которым в нашу модель были внесены следующие изменения:

- Изменена векторизация текста (применение Word2Vec из библиотеки Spark MLlib);
- Уменьшена размерность векторов с помощью метода главных компонент (PCA);
- Выбран метод кластеризации 'k-means';
- Подобрано оптимальное количество кластеров.

Что позволило повысить коэффициент силуэта с -0.022 до 0.30.



Исследование новостной ленты Нью-Йорк Таймс за период 1987-2007 гг.



КАФЕДРА
МАТЕМАТИКИ
ЭКОНОМЕТРИКИ и
ИНФОРМАЦИОННЫХ
ТЕХНОЛОГИЙ

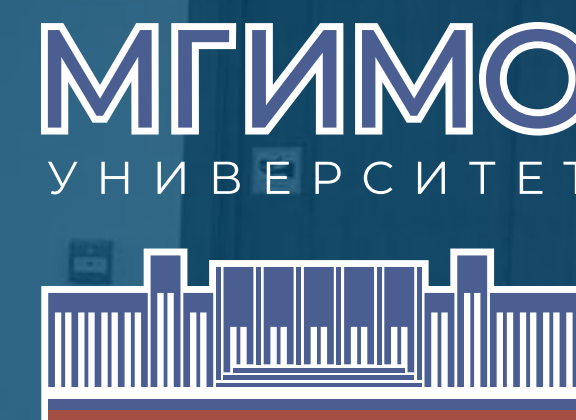
**Данная презентация подготовлена коллективом
Лаборатории ЭВТ им. Павла Чеповского
кафедры МЭИТ факультета МЭО**

**Никита Артамонов,
Александр Немченко, Михаил Звягинцев,
Юрий Акулин, Андрей Слюсарь и другие.**

Контакты:

E-mail: an2k@mail.ru
Телефоны: 16-79, 15-07,
+7 926 932 3001

Кабинеты: 4016, 4011



**КАФЕДРА
МАТЕМАТИКИ
ЭКОНОМЕТРИКИ и
ИНФОРМАЦИОННЫХ
ТЕХНОЛОГИЙ**