

# Распределенная система статистической обработки данных: модели, алгоритмы и реализация программного и пользовательского интерфейсов

Выполнила: Можина Алина, 4 курс ФИТ НГУ  
Руководитель: Городничев Максим Александрович

# Актуальность

Основные трудности использования существующих инструментов для биологов и врачей-исследователей:

- Со стороны навыков пользователя:
  - Необходимость свободного владения аппаратом математической статистики
  - Наличие навыков программирования (в т.ч. низкоуровневого)
- Со стороны возможностей инструмента:
  - Отсутствие интерпретации вычисленных результатов
  - Отсутствие возможностей работы с интенсивными вычислениями и большими объемами данных

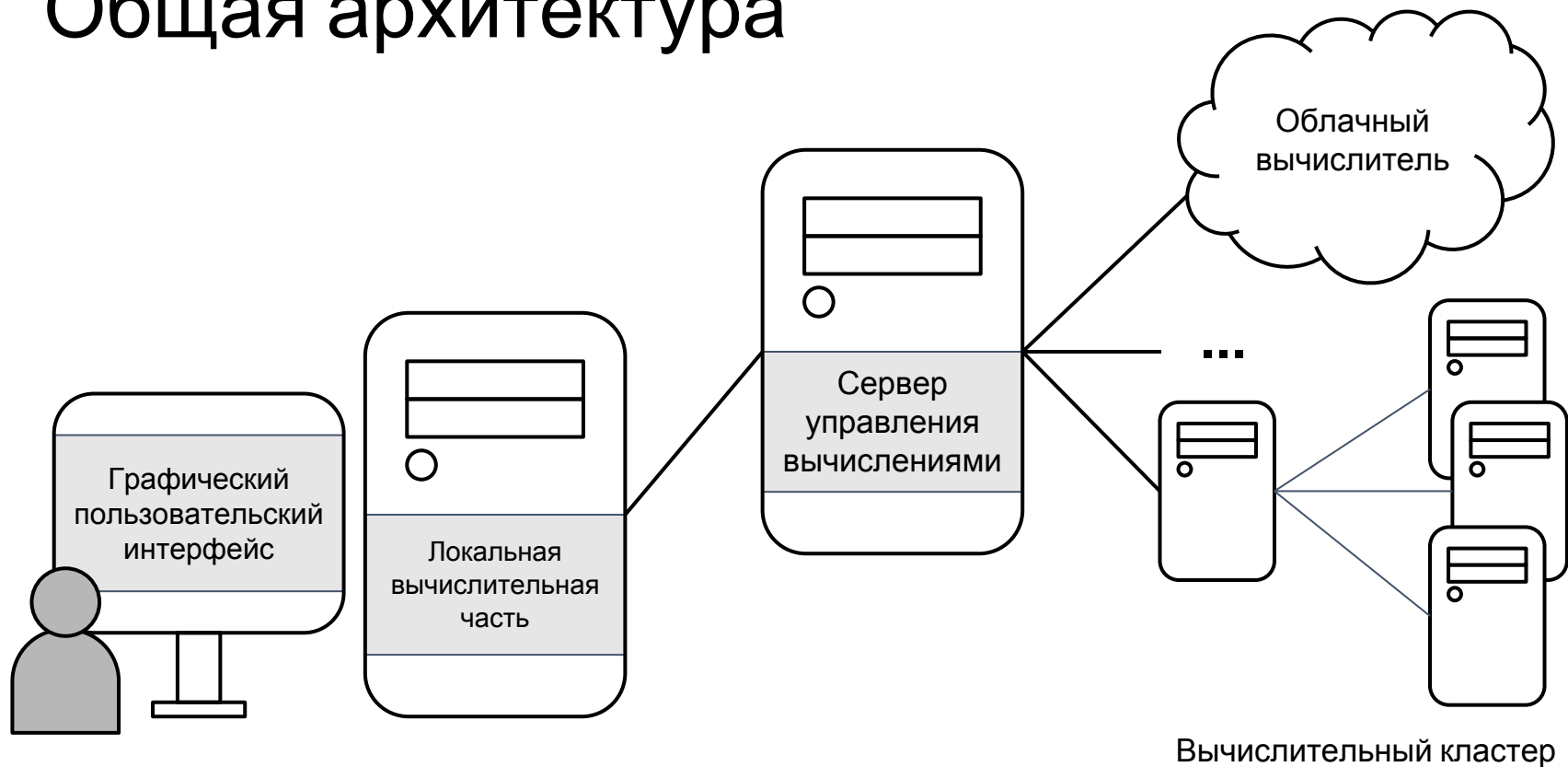
# Проект Statium

Система статистического анализа данных в биологии и медицине.

Ориентирована на обеспечение работы с большими объемами данных для медиков.



# Общая архитектура



# Цель работы

Спроектировать и реализовать клиент-серверное приложение с графическим пользовательским интерфейсом (GUI) для организации локальных и удаленных вычислений в распределенной системе статистической обработки биомедицинских данных Statium.

# Задачи

- Выделить критерии адаптированности GUI для медиков и провести анализ существующих интерфейсов
- Сформулировать требования к приложению
- Спроектировать GUI, модель удаленных вычислений, определить коммуникационные протоколы
- Реализовать рабочий прототип приложения

# Анализ требований к GUI: существующие решения

	Пояснения к методам	Интерпретация результатов	Простота интерфейса	Подробные руководства
Microsoft Office Excel	Отсутствуют	Отсутствует	Отсутствует, но компенсируется привычностью	Присутствуют, русскоязычные
Dell STATISTICA	Отсутствуют	Минимальная	Отсутствует	Присутствуют, русскоязычные, профильные
AnalystSoft BioStat	Краткие	Отсутствует	Отсутствует, но компенсируется привычностью	Отсутствуют
IBM SPSS Statistics	Имеется помощь в выборе метода, непрофильная	Отсутствует	Отсутствует	Присутствуют, русскоязычные

# Требования к GUI

Необходимо реализовать графический пользовательский интерфейс,

а) позволяющий управлять как локальными, так и удаленными вычислениями,

б) удовлетворяющий критериям адаптированности для медиков.



# Модель работы пользователя

Данные – таблицы объект-признак с количественными и качественными значениями.

Основная последовательность действий:

1. Выбор класса вычислительной задачи
2. Указание таблиц с исходными данными
3. Заполнение формы с параметрами задачи (зависят от класса задачи)
4. Выбор вычислителя
5. Запуск задачи
6. Получение таблиц результатов с пояснениями

# Проектирование GUI

Были сформулированы следующие принципы построения интерфейса:

- Вычислительные методы должны сопровождаться описанием принципов работы, условий применения и входных данных
- Пользователь должен получать текстовые пояснения к вычисленным результатам, а также визуализацию в виде диаграмм
- В активном диалоге должны присутствовать только элементы управления, необходимые для выполнения текущего шага сценария

# Плохо: результат вычислений в GUI IBM SPSS

## T-критерий

[Наборданных1]

Статистика группы

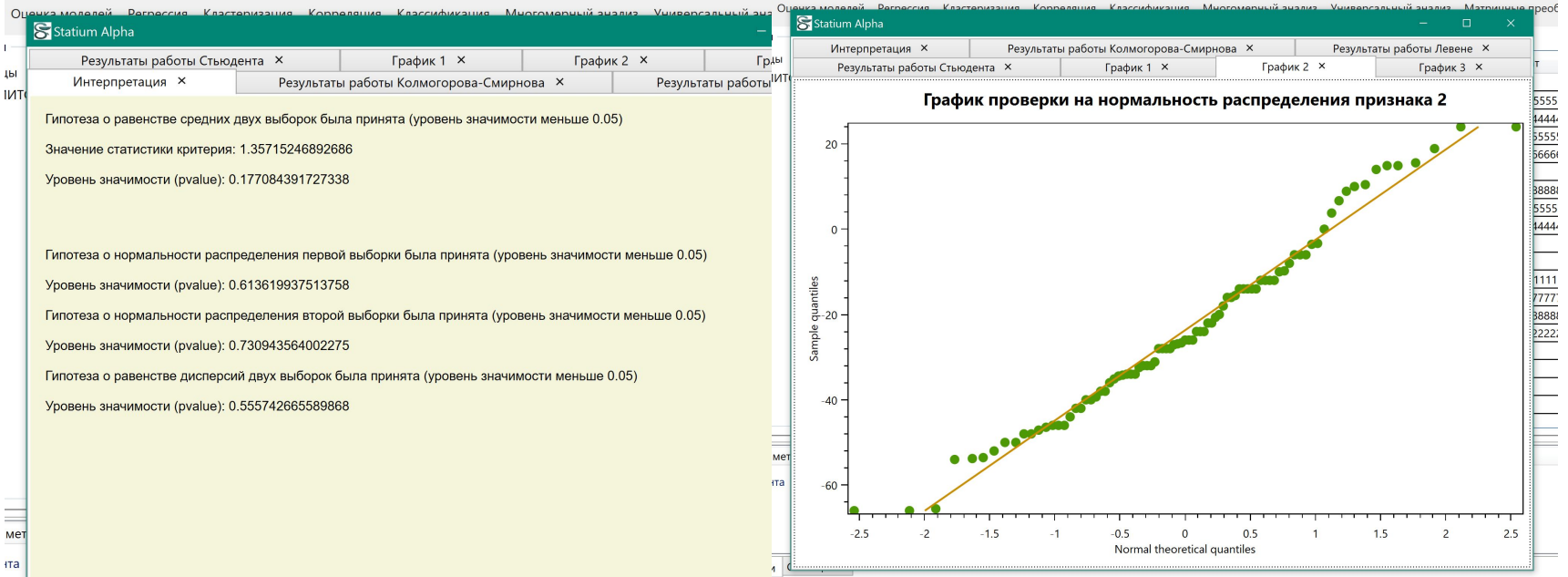
	Имплант	N	Среднее	Среднекв. отклонение	Среднекв. ошибка среднего
Результат	,0	87	-23,686	21,3166	2,2854
	1,0	45	-18,479	20,0423	2,9877

Критерий для независимых выборок

		Критерий равенства дисперсий Ливиня		t-критерий для равенства средних						
		F	Значимость	t	ст.св.	Знач. (двухсторонняя)	Средняя разность	Среднеквадратичная ошибка разности	95% доверительный интервал для разности	
									Нижняя	Верхняя
Результат	Предполагаются равные дисперсии	,349	,556	-1,357	130	,177	-5,2068	3,8366	-12,7970	2,3834
	Не предполагаются равные дисперсии			-1,384	94,075	,170	-5,2068	3,7616	-12,6754	2,2618

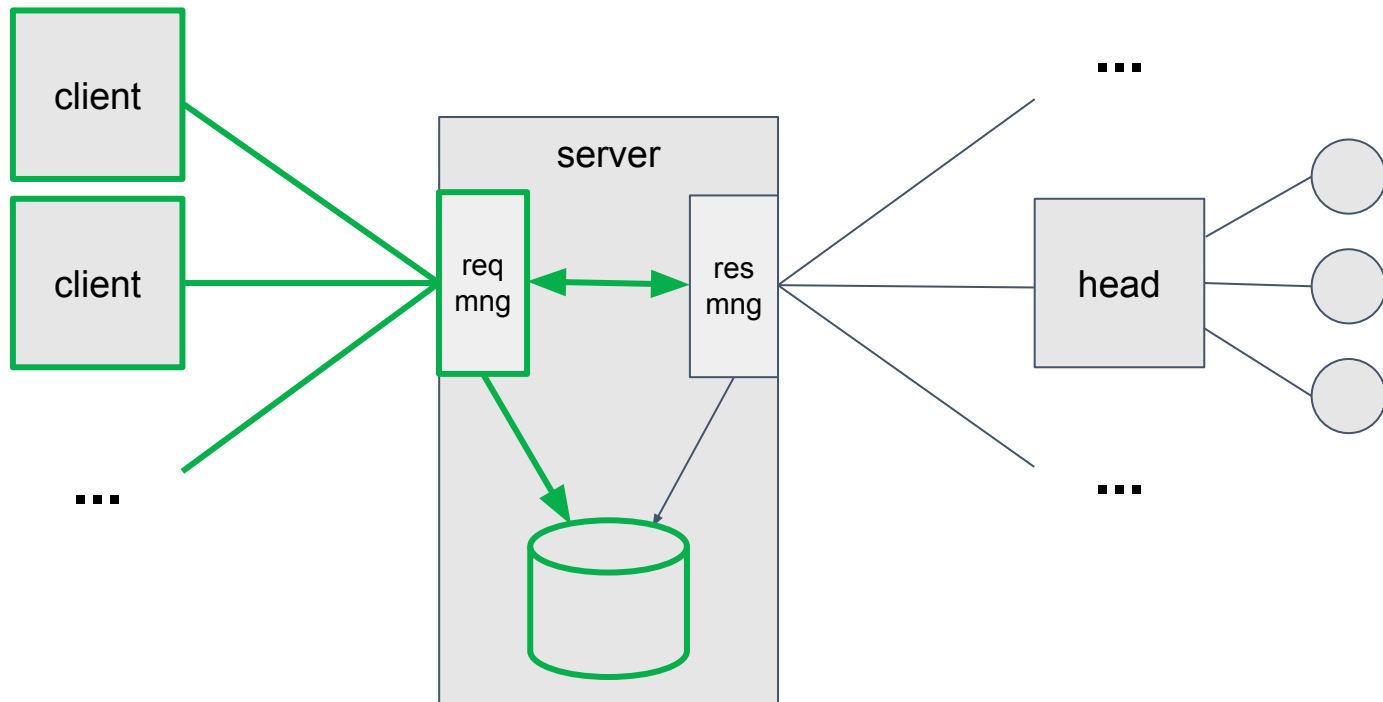
Пример результата вычислений в IBM SPSS 23

# Хорошо: результат вычислений в GUI Statium



Пример вычислений в интерактивном режиме: получение результатов с пояснениями

# Удаленные вычисления



# Удаленные вычисления

Удаленный вычислительный ресурс = вычислительный кластер.

Необходимо скрыть от пользователя сложность коммуникации и постановки задач на кластер.

Два сценария доступа:

1. Пользователь не имеет учетных записей на кластерах. Вычисления происходят в рамках штатного доступа системы к вычислительным центрам.
2. Пользователь имеет учетные записи на кластерах. Вычисления происходят на выбранном пользователем кластере под одной из его учетных записей.

Выбор между локальными и удаленными вычислениями делает пользователь с помощью одного переключателя.

# Требования к серверному приложению

Необходимо реализовать серверное приложение,

а) поддерживающее систему пользователей с возможностью привязывания к профилю учетных записей на кластерах,

б) позволяющее формулировать вычислительную задачу для кластера и забирать результат расчётов,

в) позволяющее делиться входными данными и результатами между пользователями,

г) расширяемое на новые типы вычислителей помимо кластеров.

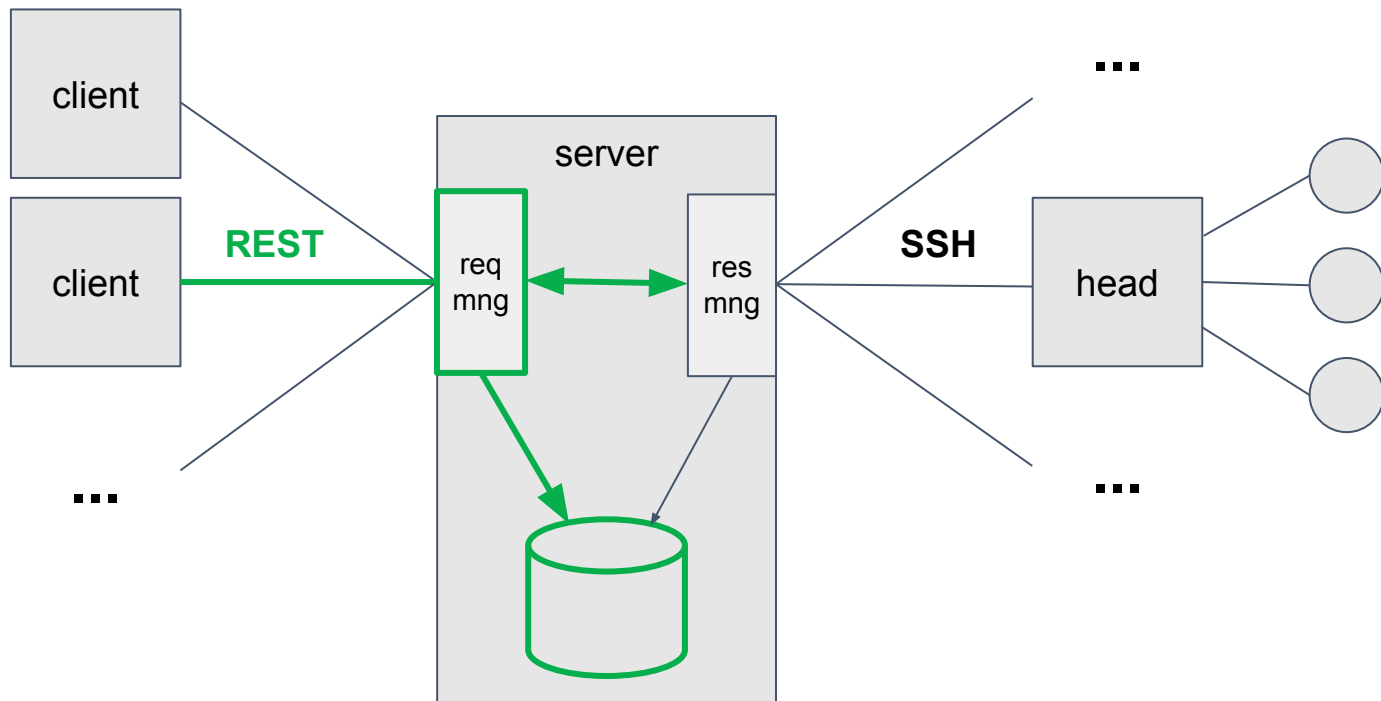
# Проектирование серверного приложения

Ресурсы, которыми оперирует сервер управления вычислениями:

- пользователи (users)
- круги пользователей (circles)
- задачи (jobs)
- классы задач (methods)
- наборы классов задач (method\_profiles)
- таблицы (grids)
- кластеры (clusters)
- учетные записи на кластере (cluster\_profiles)
- ...



# Удаленные вычисления



# Использованные технологии

Модуль	Технологии	Причины
GUI	C#, фреймворк WPF	<ol style="list-style-type: none"><li>1. Удобство разработки – в WPF большое количество стандартных элементов интерфейса</li><li>2. Простота интеграции с менеджером задач Statium – также написан на C#</li></ol>
Серверное приложение (reqmng)	C#, ASP.NET Web API	<ol style="list-style-type: none"><li>1. Развитые средства работы с БД – LINQ, EntityFramework</li><li>2. Скорость разработки – минимум начальной конфигурации с возможностью детального изменения</li></ol>

# Результаты

На текущий момент:

- Построена модель организации удаленных вычислений
- Спроектирован и реализован программный интерфейс (API) сервера управления вычислениями
- Спроектирован и реализован GUI, позволяющий запускать задачи локально и удаленно
- Проведено тестирование GUI с участием студентов МедФ НГУ и преподавателей НГМУ

# Планы

- Добавить в GUI встроенное руководство пользователя (раздел «Помощь»)
- Ввести метрики потребленных вычислительных ресурсов
- Добавить поддержку работы с большими объемами данных
- Провести тестирование работы системы на кластерах НГУ и ССКЦ СО РАН

# Достижения

- Диплом I степени на МНСК-2016 в секции «Распределенная обработка информации, корпоративные информационные системы»
- Внедрение системы в НГМУ при обучении студентов основам математической статистики и в работе Отдела статистической обработки данных

**Спасибо за внимание**

# Термин

Неспециалист – пользователь, не являющийся специалистом в математической статистике и программировании; в рамках данной работы – биолог или врач-исследователь.

# Задачи

## По GUI:

- Выделить критерии адаптированности GUI для неспециалистов и провести анализ существующих интерфейсов
- Сформулировать требования к GUI
- Спроектировать GUI
- Реализовать рабочий прототип GUI

## По серверному приложению:

- Сформулировать требования к серверному приложению
- Спроектировать модель распределенных вычислений, определить коммуникационные протоколы
- Реализовать рабочий прототип серверного приложения



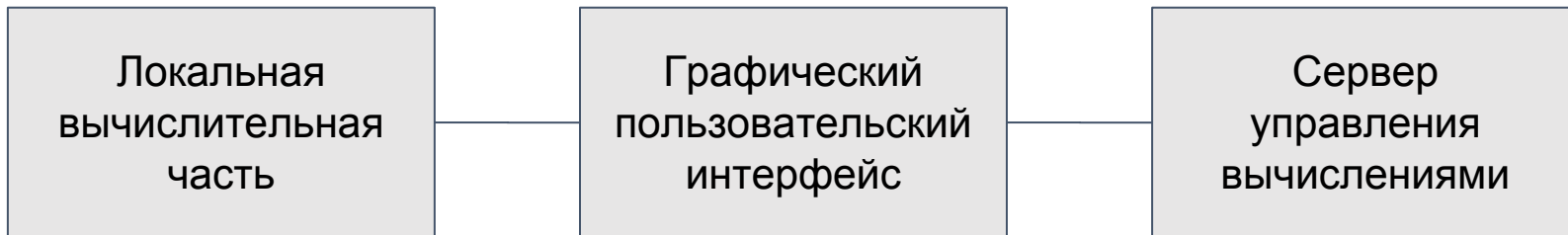
# Актуальность

Имеется необходимость в новом инструменте, который должен:

а) обеспечивать статистическую обработку данных как локально, так и на удаленных высокопроизводительных вычислительных ресурсах,

б) быть понятным для медиков,

в) позволять использовать данные из различных удаленных источников.



# Предложения\замечания

- сформулировать одну цель — клиент-серверное приложение
- убрать слайд с "плохо: SPSS" ?
- нарисовать трехчастную схему с GUI, локальным вычислителем и сервером
- возможно, убрать требования из слайда про актуальность (многократное повторение)
- раньше показать схему распределенных вычислений
- добавить модель пользователя
- добавить про переключение счета локально\кластер одной кнопкой
- говорить, что данные обратно с кластера через меня тоже идут – устно на слайде 18
- возможно, совместить требования в один список
- **избавиться от термина неспециалист**
- сделать крупнее номера слайдов

# Актуальность

Имеется необходимость в новом инструменте, который должен:

а) быть понятным для медиков,

б) обеспечивать статистическую обработку данных как локально, так и на удаленных высокопроизводительных вычислительных ресурсах,

в) позволять использовать данные из различных удаленных источников.

# Требования к GUI

Необходимо реализовать графический пользовательский интерфейс,

а) позволяющий инициировать вычисления как локально, так и удаленно,

б) удовлетворяющий критериям адаптированности для медиков.