«Иерархическая кластеризация базы данных предварительно посчитанных сигналов для решения параметрической обратной задачи рассеяния света»

Автор – студент Мулюков А.Р.[1,2]
Науч. Рук. – с.н.с., к.ф.-м. н. Юркин М.А. [1,2]
[1] Новосибирский государственный университет,
[2] ИХКГ СО РАН, лаб.ЦиБ

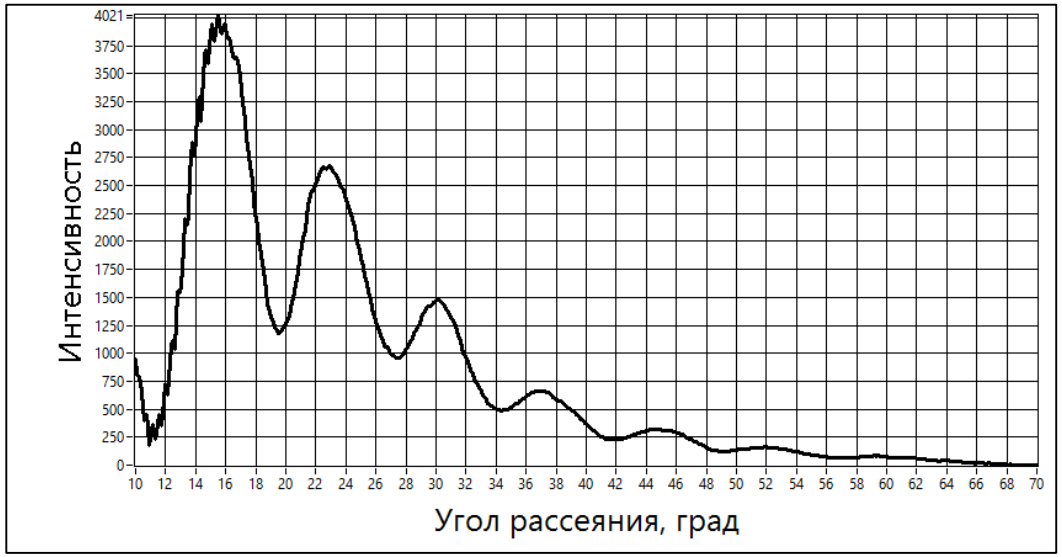
План доклада

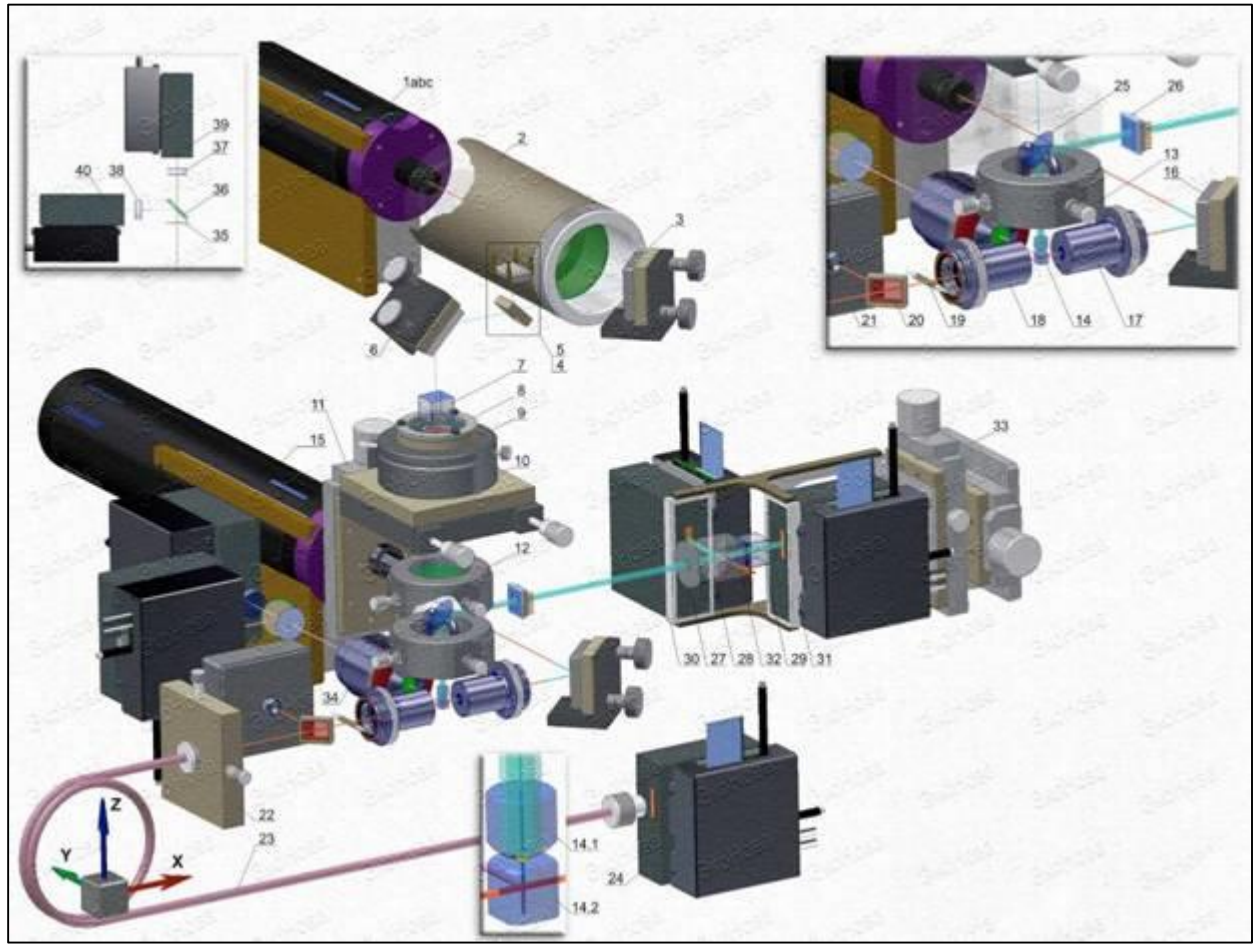
1. Общее описание обратной задачи
2. Новый алгорим
3. Реализация
4. Полученные результаты
5. Заключение

Сканирующий проточный цитометр

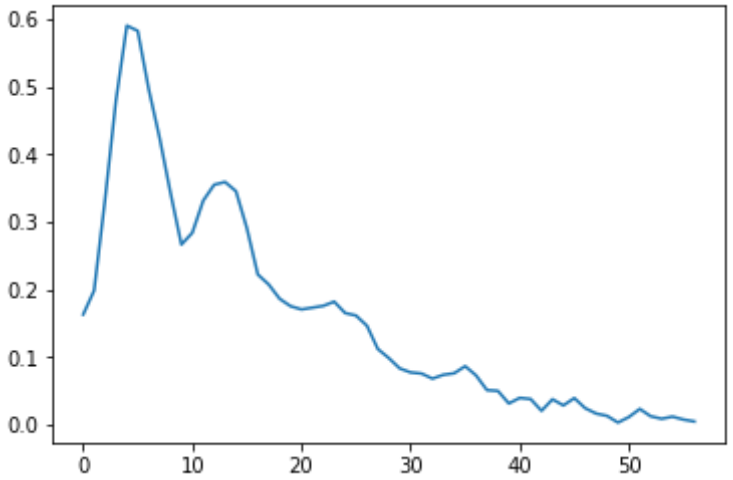
О технологии:

- Позволяет измерять подробный сигнал от каждой клетки (индикатриса светорассеяния)
- Можно определять характеристики клеток (размер, показатель преломления, и т.д.)
- Потенциально – быстрая диагностика различных заболеваний
- Измеряет ~100 клеток в минуту





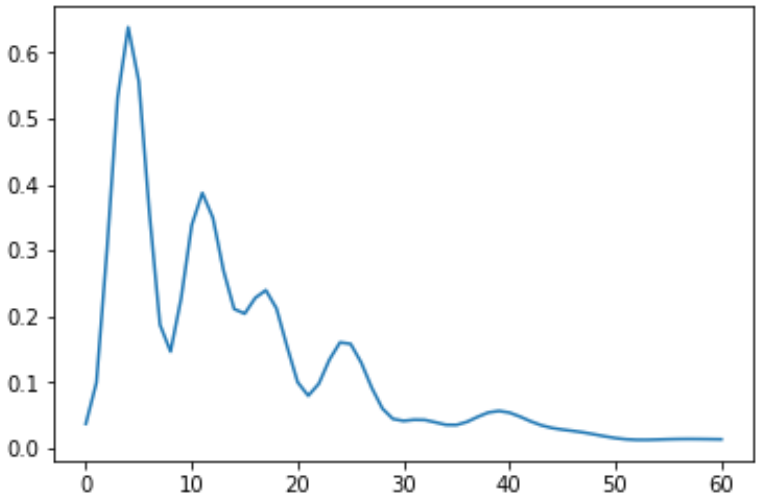
Обратная и прямая задачи



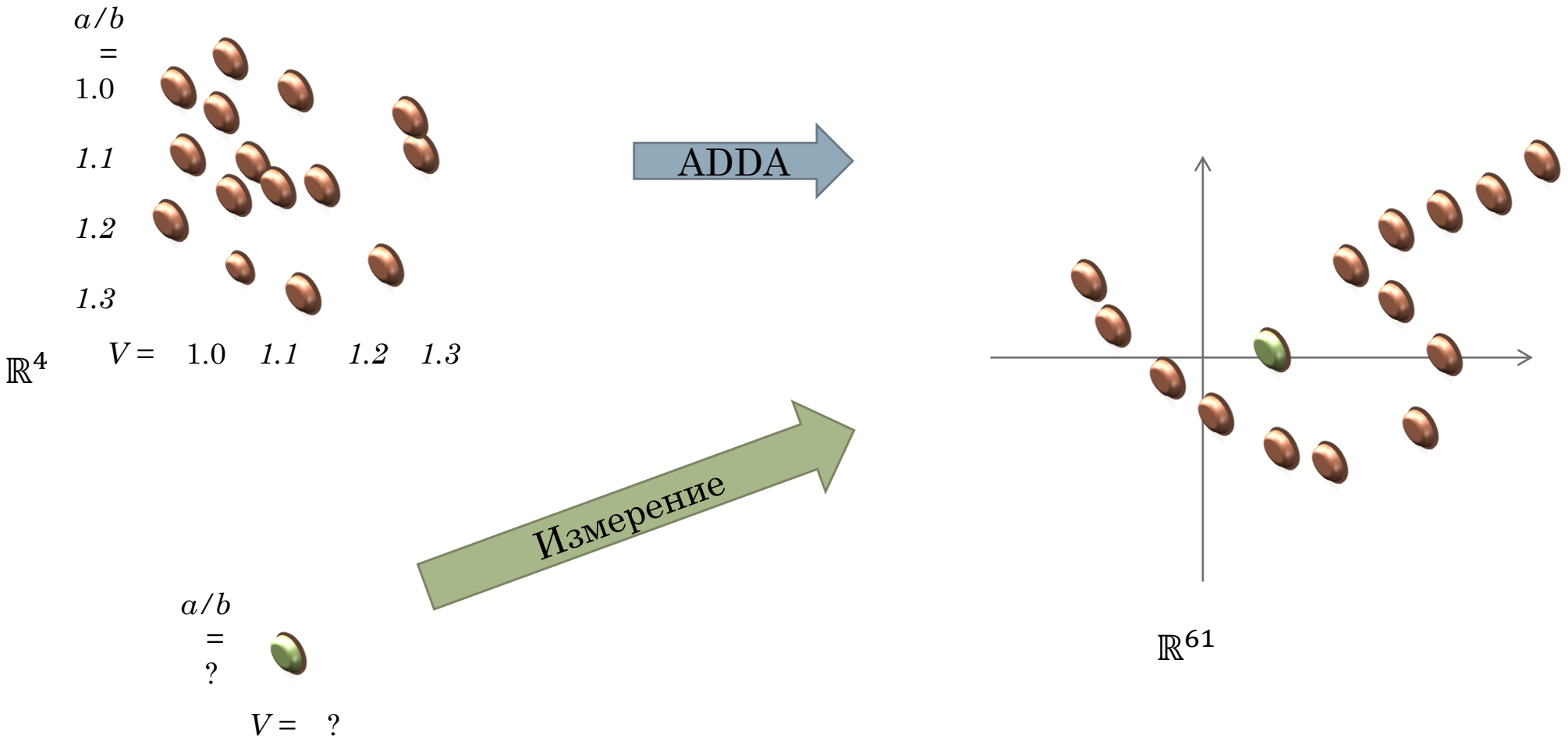
$$V = ?;\ \frac{a}{b} = ?;\ n = ?;\ \beta = ?$$

(ADDA):

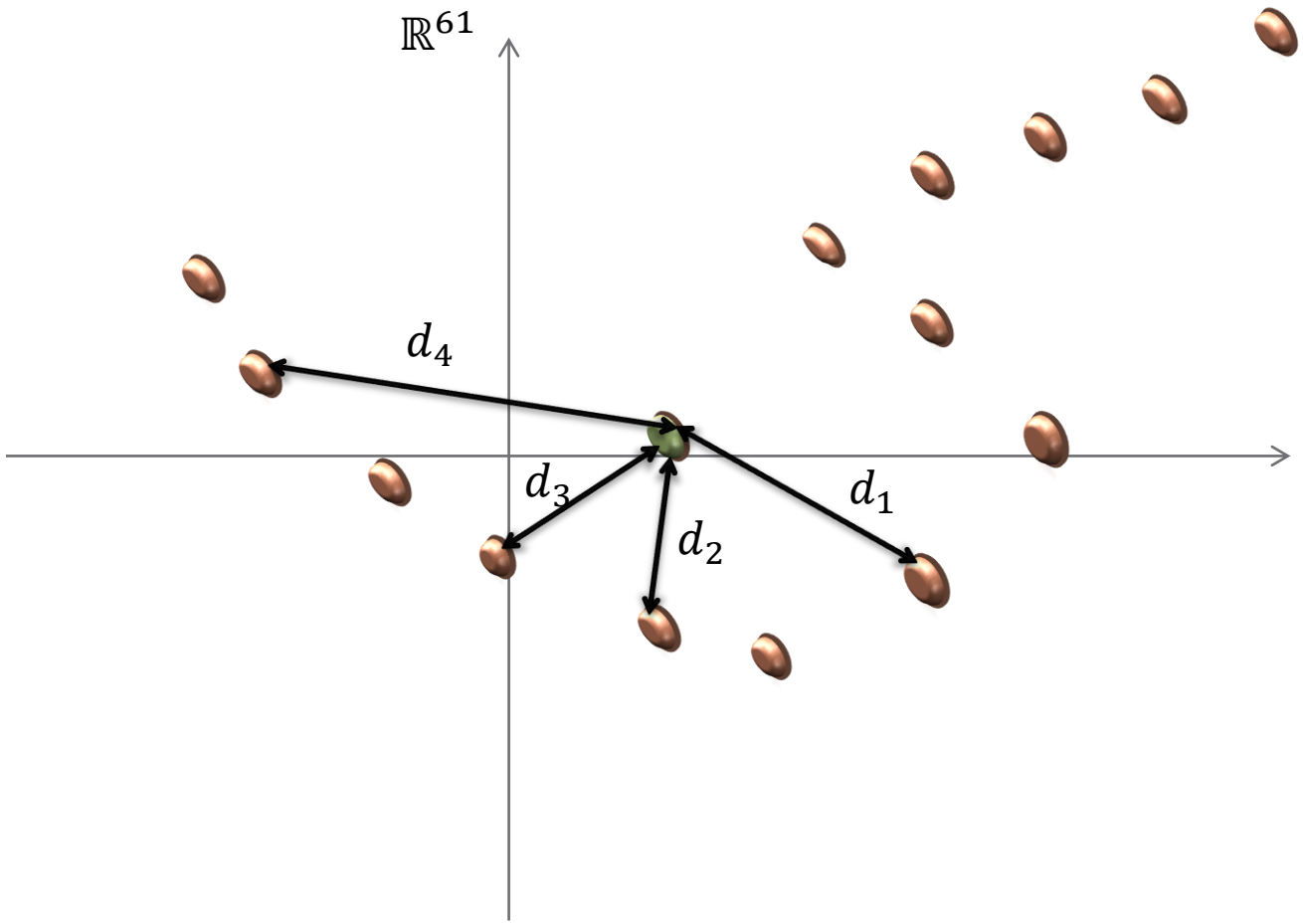
$$V = 1.2;\ \frac{a}{b} = 5{,}4;\ n = 1.47;\ \beta = 24.73$$



От случайной базы данных в $\mathbb{R}^4$ к $\mathbb{R}^{61}$



Поиск ближайшей и оценка точности



- Находим наименьшее расстояние d – находим решение задачи (наиболее вероятные характеристики клетки)
- Все просмотренные расстояния d можно использовать для оценки точности решения

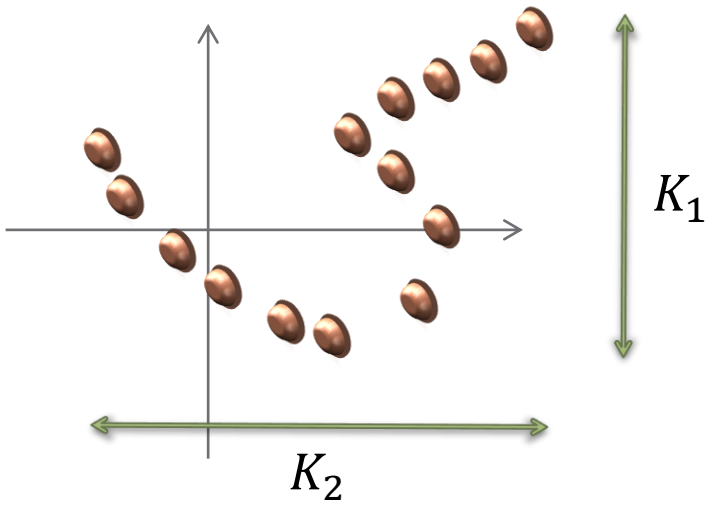
Модификация оригинального метода

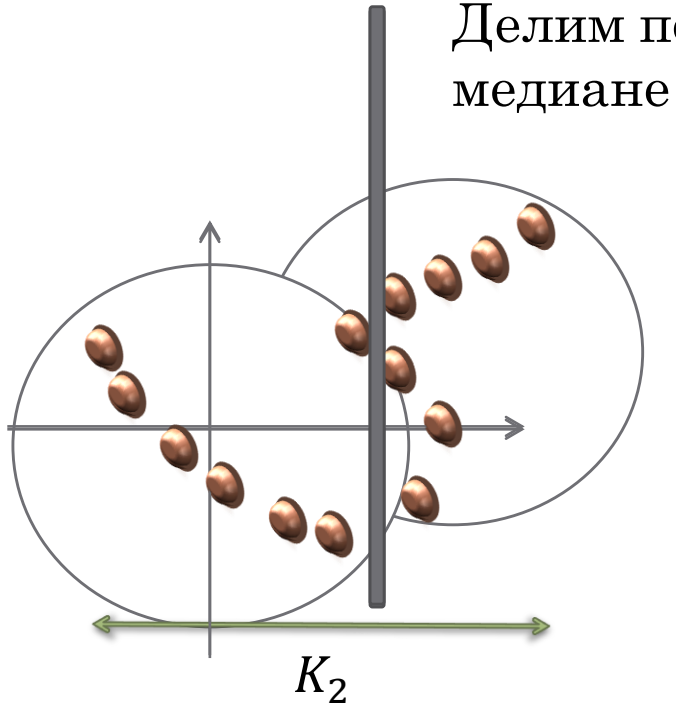
Оригинальный метод:
- Сравниваем каждый теоретический элемент с измеренным
- Находим среди них элемент с наименьшим расстонием l_2
- Используем полученные расстояния для вычисления статистической оценки полученного решения.

Проблема:
- Скорость решения задачи зависит линейно от размера базы данных (а большой размер требуется для точного решения)

Предлагается ускорить перебор теоретических элементов. Для этого:
- Строим иерархическую структуру (дерево) из имеющихся элементов
- Отбрасываем часть точек целыми кластерами на основе одного сравнения
- Предполагаемое ускорение до $\log n$, где n – размер базы данных. При этом ближайший элемент находится точно.

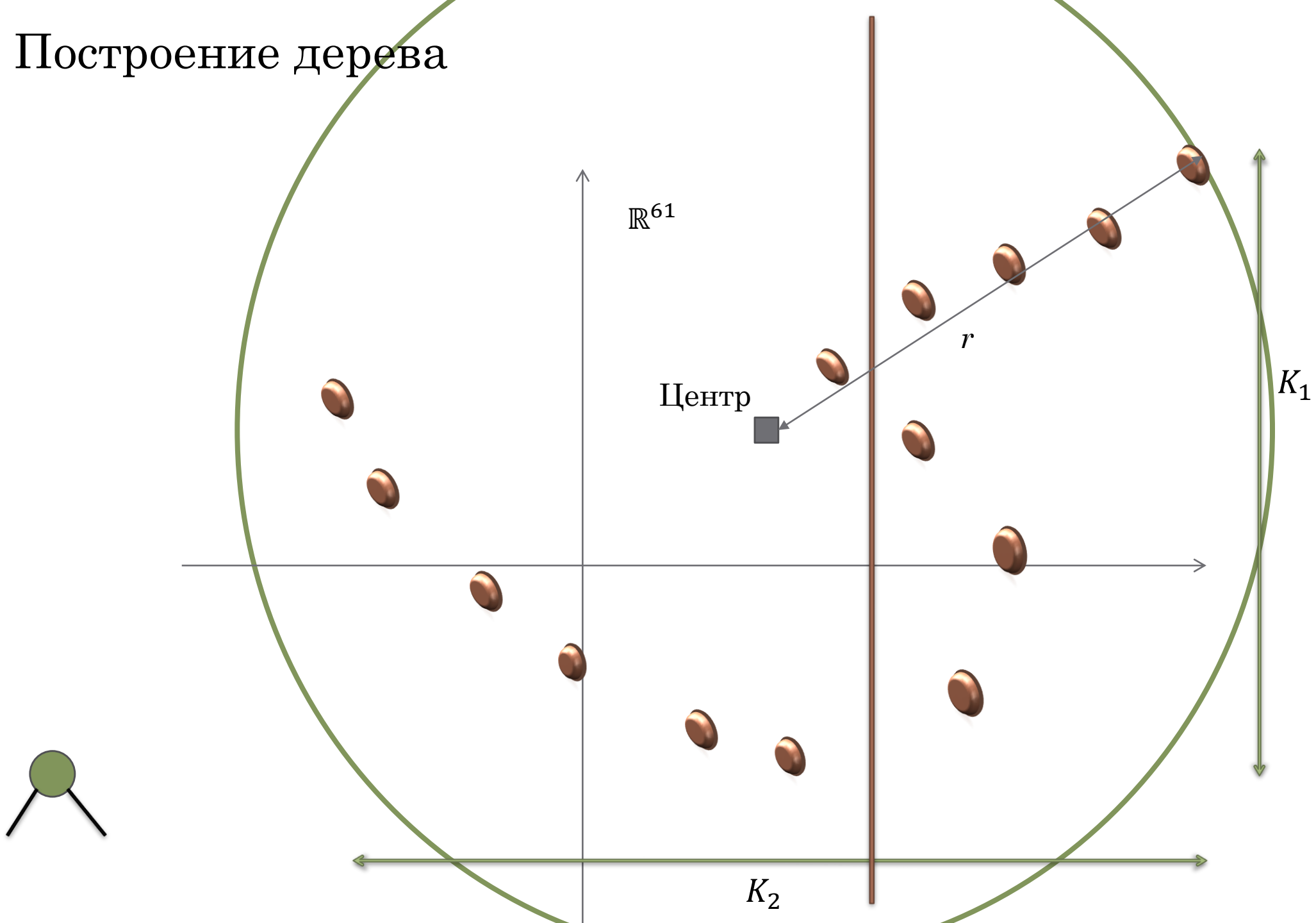
Построение иерархической структуры (бинарное дерево)



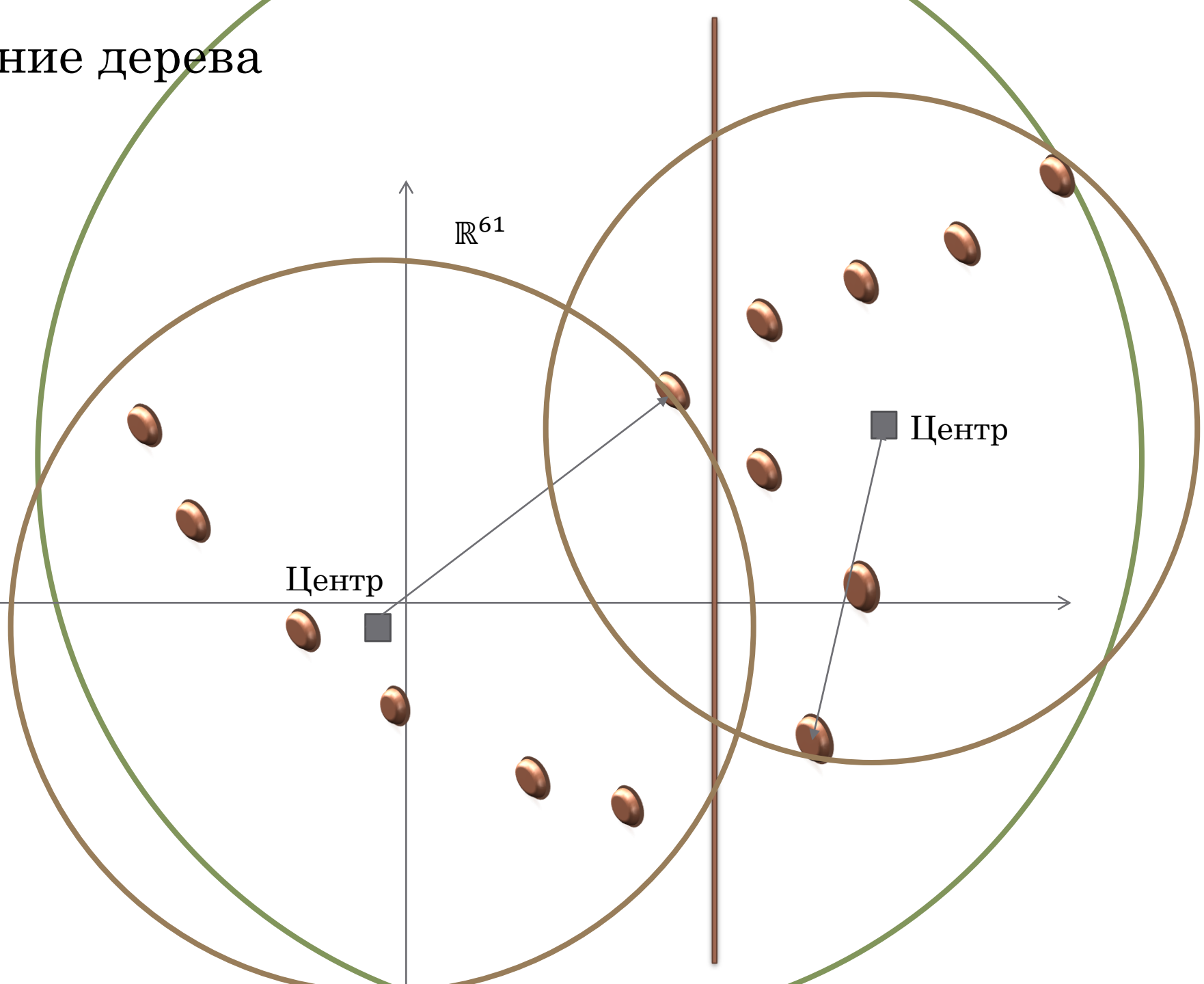


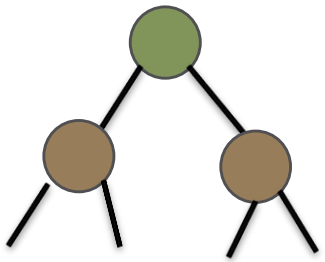
Выбранная координата: $\underset{i}{\operatorname{argmax}} K_i$

Построение дерева

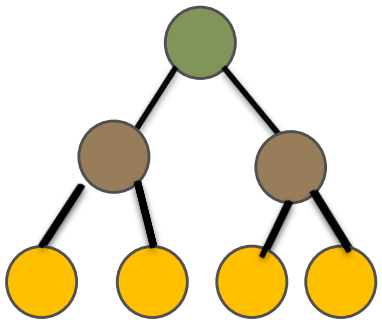


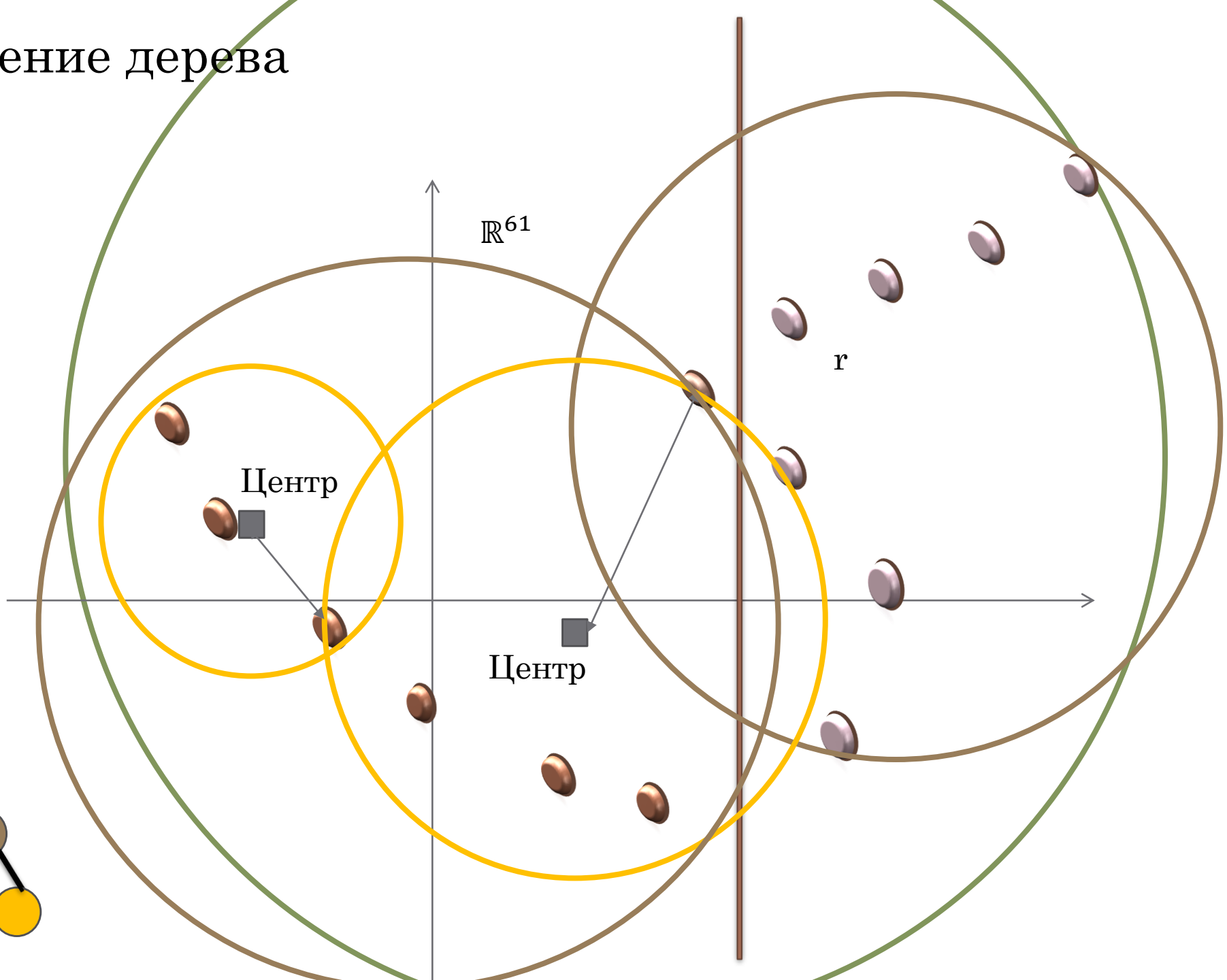
Построение дерева



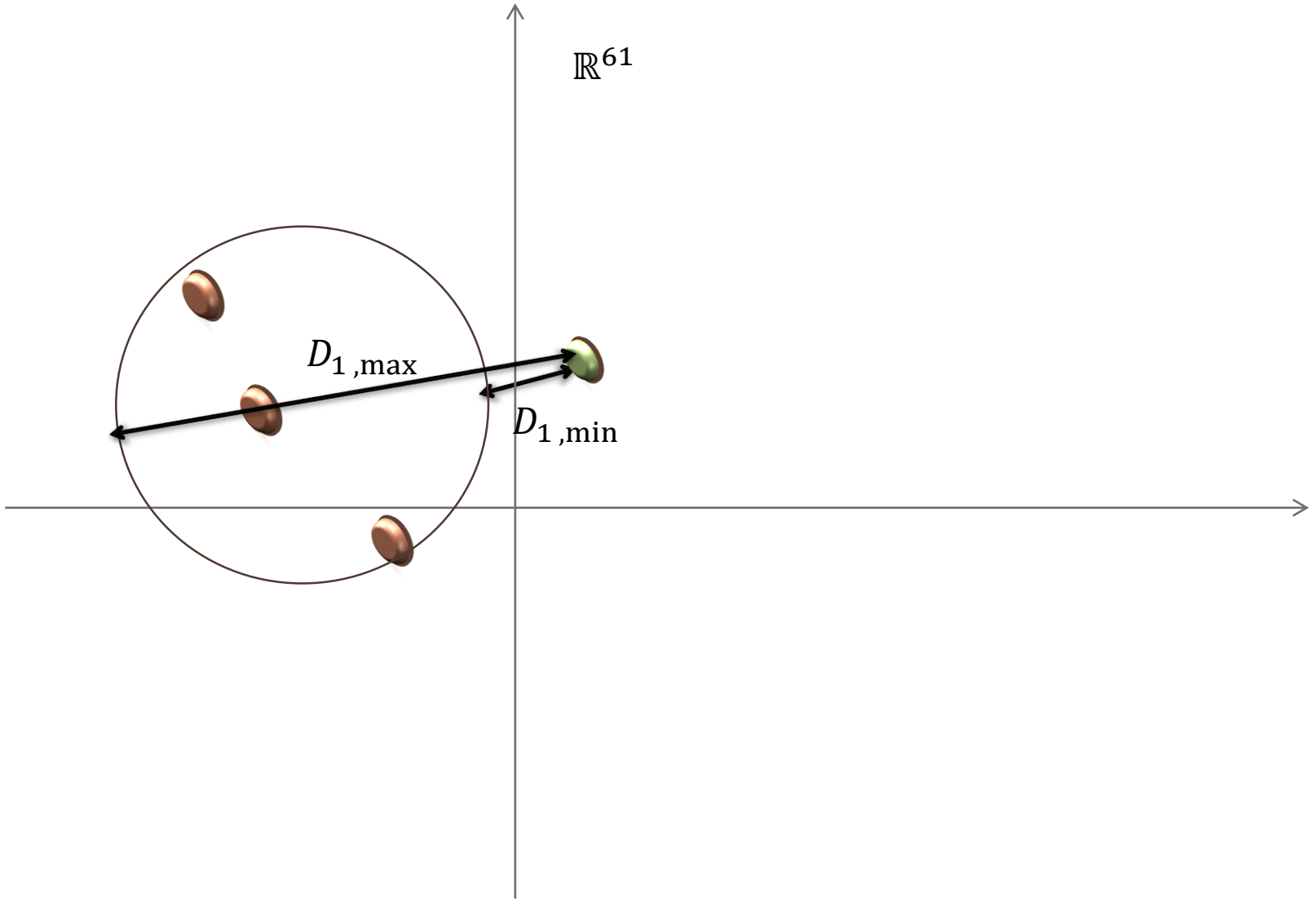


Построение дерева

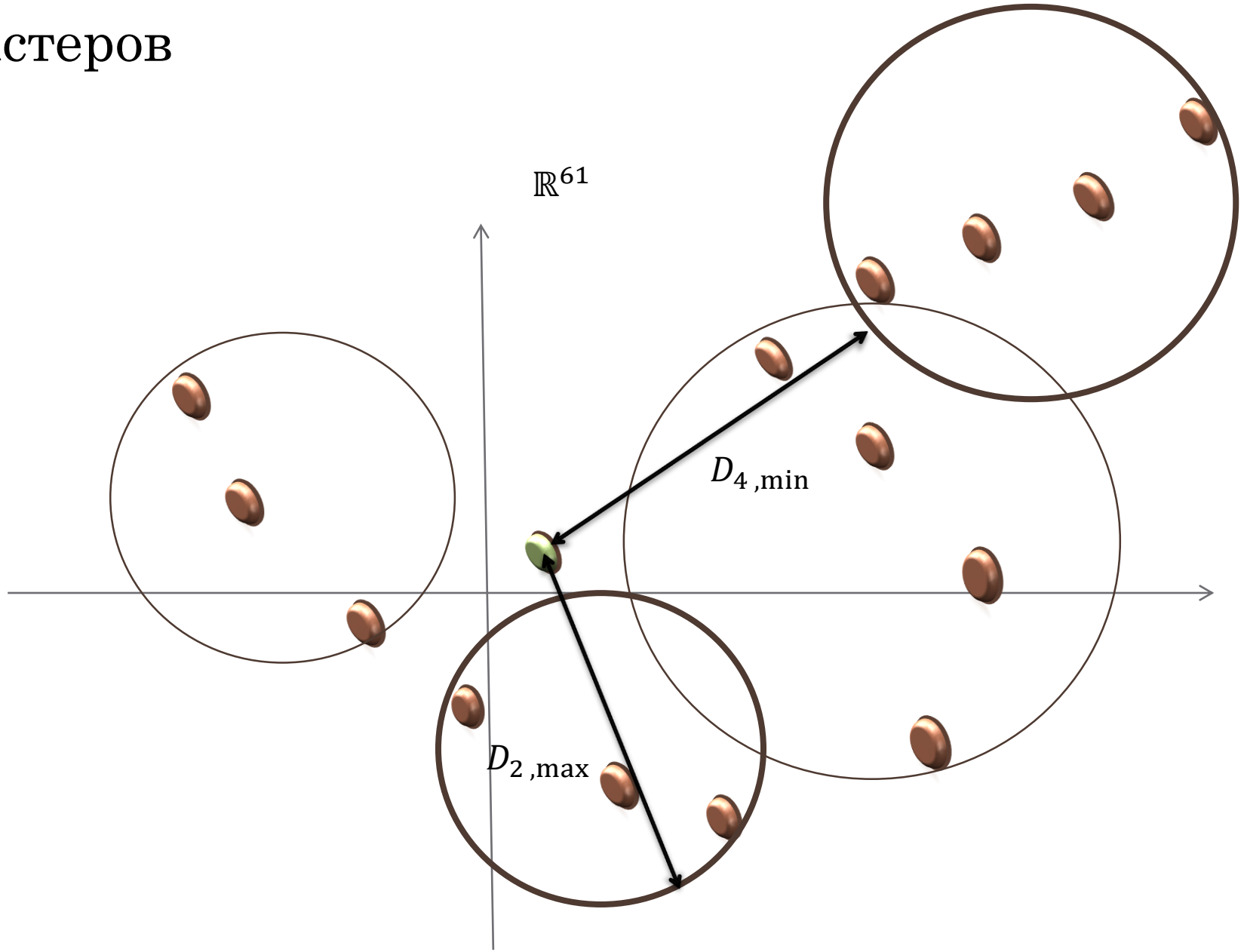


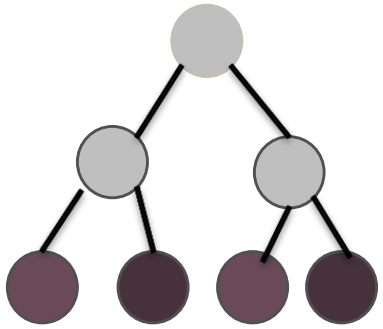


max и min расстояния до кластера

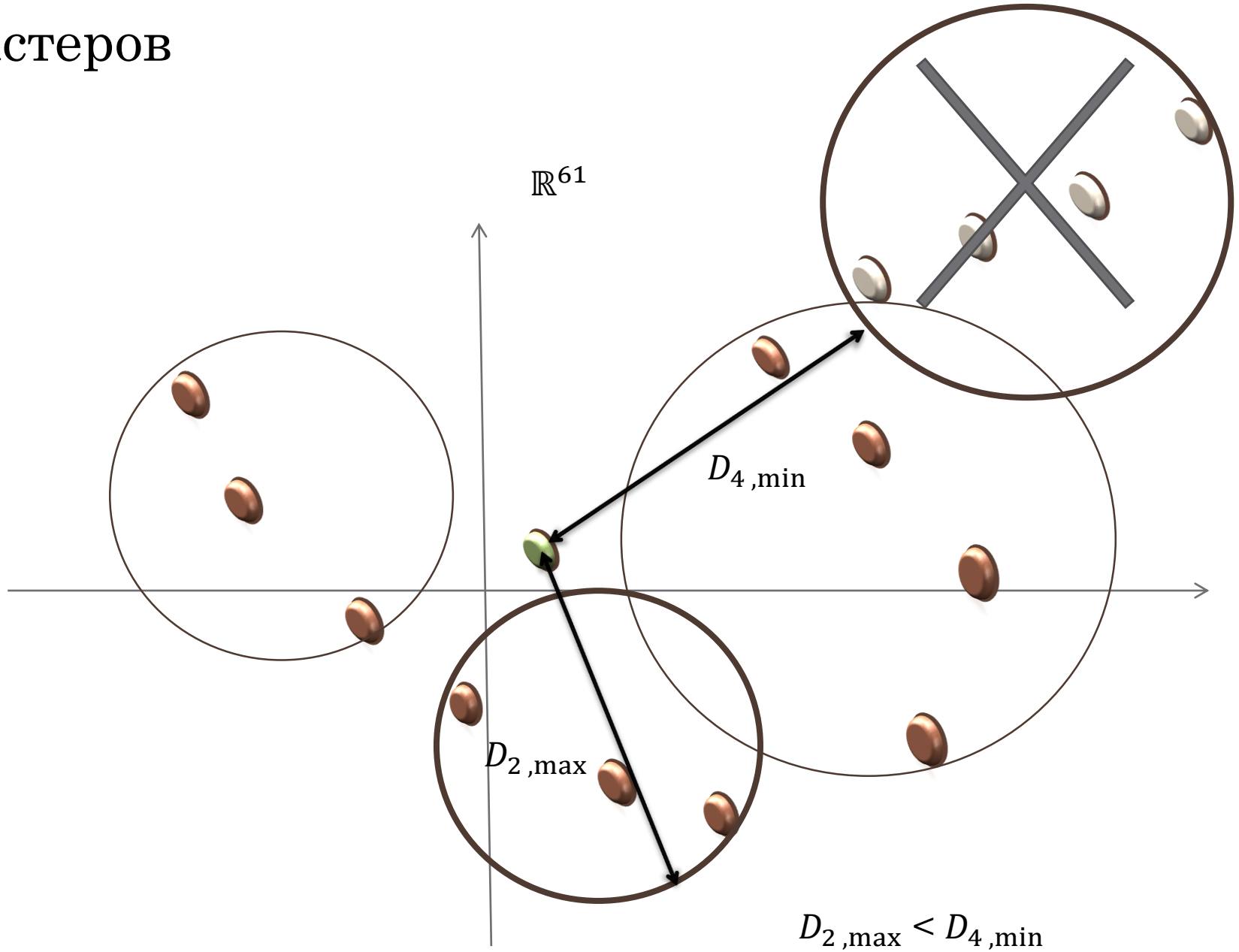


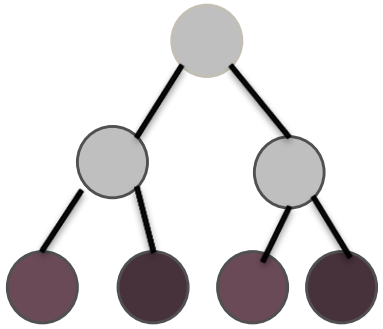
Отброс кластеров





Отброс кластеров





Предварительные тесты

1. Тесты эффективности алгоритма с помощью Python:

- Сравнение количества необходимых сравнений и реальной скорости работы.
- Использовался код открытой библиотеки sklearn

Почему не использовать сразу код из Python?

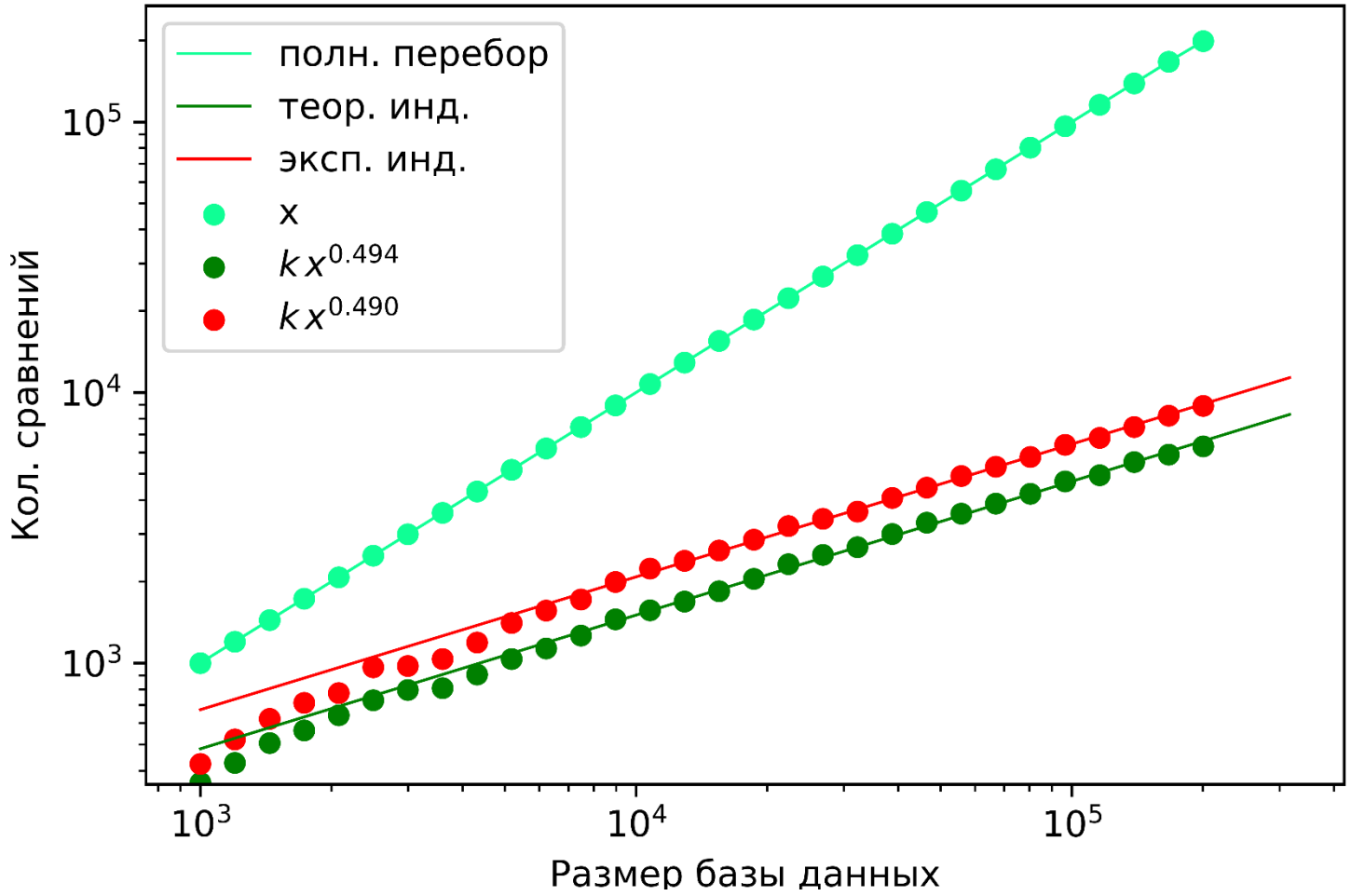
- Он не сохраняет отброшенные расстояния
- Нет возможности гибко настраивать критерии отброса кластеров
- Он не оформлен в формате dll, для работы в LabView (в котором находятся существующие программы для обработки экспериментов)

Реализация: Код для LabView

2. Рабочий код ускорения для LabView

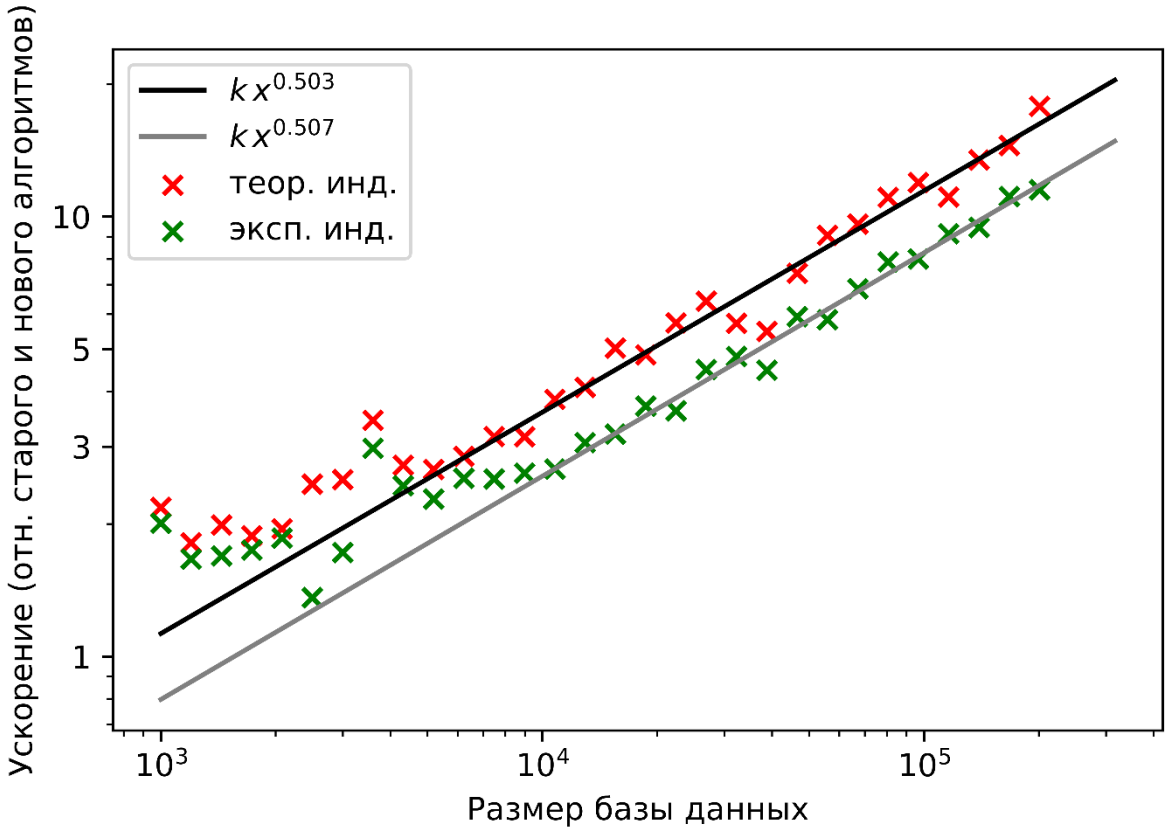
- Код на C++ (в Microsoft Visual Studio)
- Оформлен в виде Dynamic Link Library (DLL)
- Код встраивается как функции в среду LabView
- Отдельная функция для построения дерева и отдельная функция для его обхода
- Параметры для настройки глубины обхода дерева (для более точных статистических оценок)

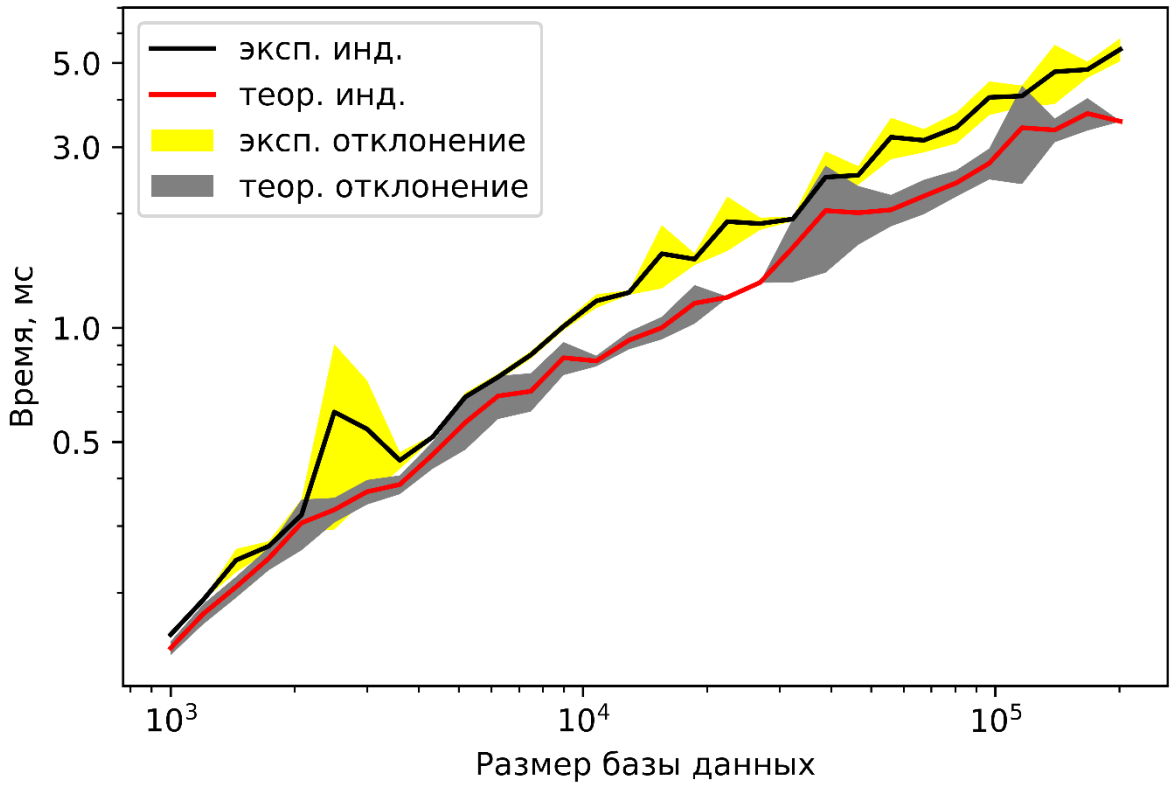
Количество сравнений от размера базы данных



- Полученное количество сравнений имеет ускорение, аппроксимируемое функцией $\sqrt{x}$

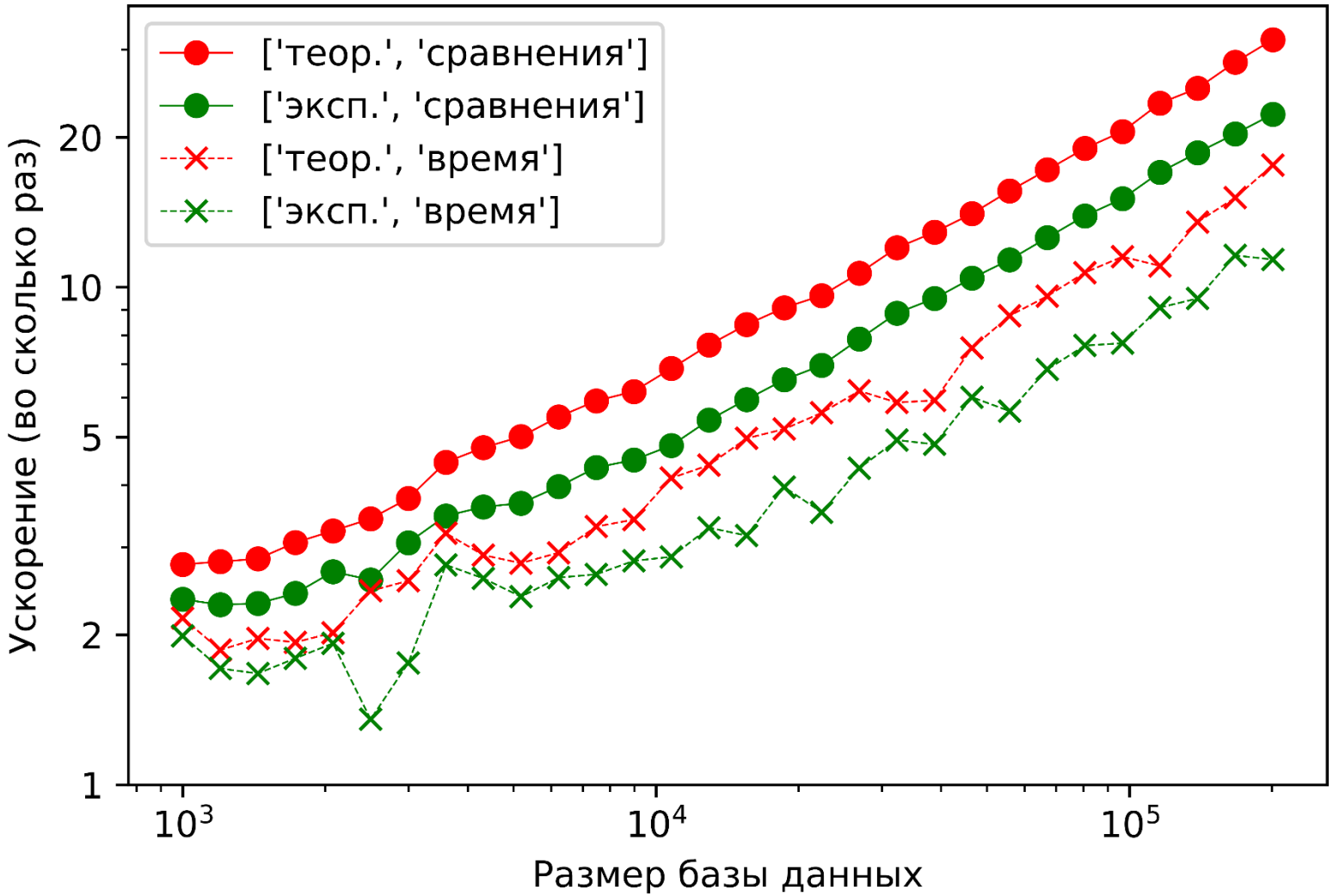
Реальная скорость работы





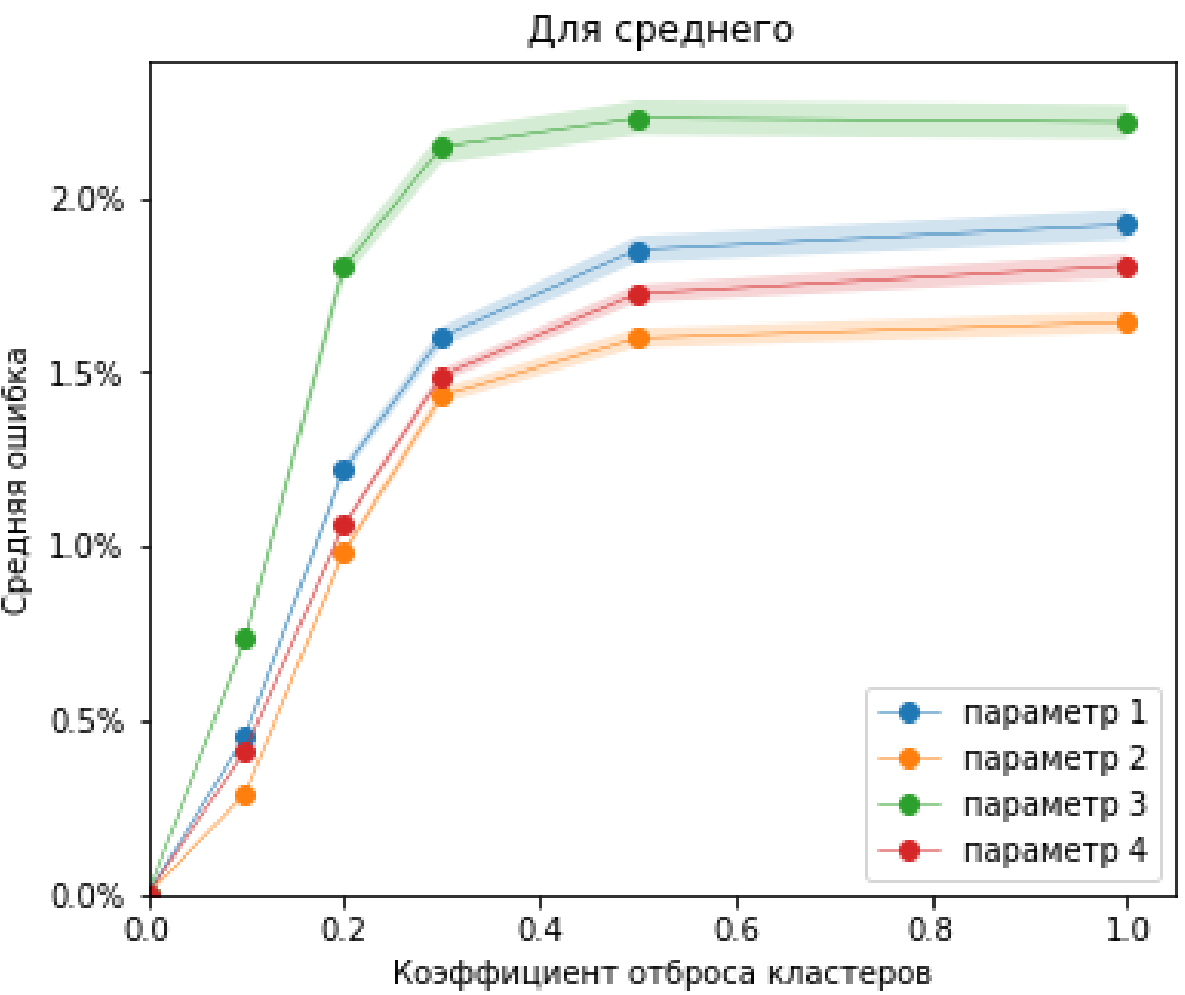
- Несмотря на некую зашумленность, результаты времени работы стабильны

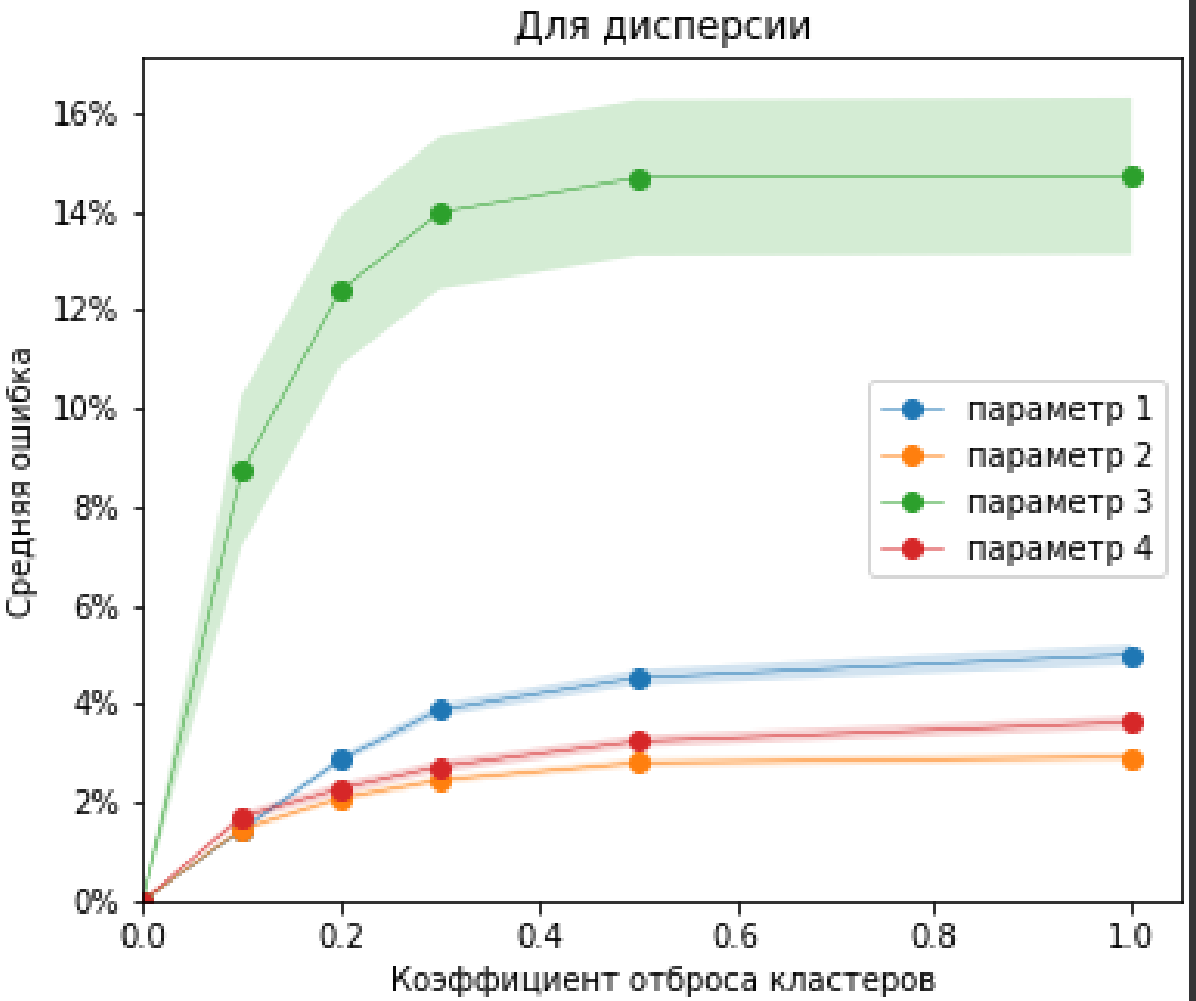
Сравнение ускорения



- Реальное ускорение чуть ниже ожидаемого
- Возможно из-за проблем организации памяти

Оценка уровня ошибки





- Различия статистических оценок для оригинального алгоритма и модифицированного от уровня агресивности отброса кластера k. Используется: $kD_{\max} < R_{\text{new}}$.

Заключение

- Алгоритм помогает получить значительное ускорение работы классического метода решения задачи (примерно на порядок для текущей базы данных)
- Скорость обработки порядка 5 мс → режим онлайн
- Ускорение растет в зависимости от размера базы данных
- Предположительно, на реальное ускорение в значительной степени влияет организация работы с памятью компьютера
- Ближайшая индикатриса совпадает и имеют небольшую погрешность в вычислении статистических оценок полученного решения (до 2.5% для оценки среднего, до 15% для оценки стандартного отклонения)
- Необходимо проработать вопрос выбора типа дерева, провести тестирование на разных базах данных (для разных задач).

«Иерархическая кластеризация базы данных предварительно посчитанных сигналов для решения параметрической обратной задачи рассеяния света»

Автор – студент Мулюков А.Р.[1,2]
Науч. Рук. – с.н.с., к.ф.-м. н. Юркин М.А. [1,2]
[1] *Новосибирский государственный университет,*
[2] *ИХКГ СО РАН, лаб.ЦиБ*

Спасибо за внимание.

Novosibirsk State University
*THE REAL SCIENCE

МНСК