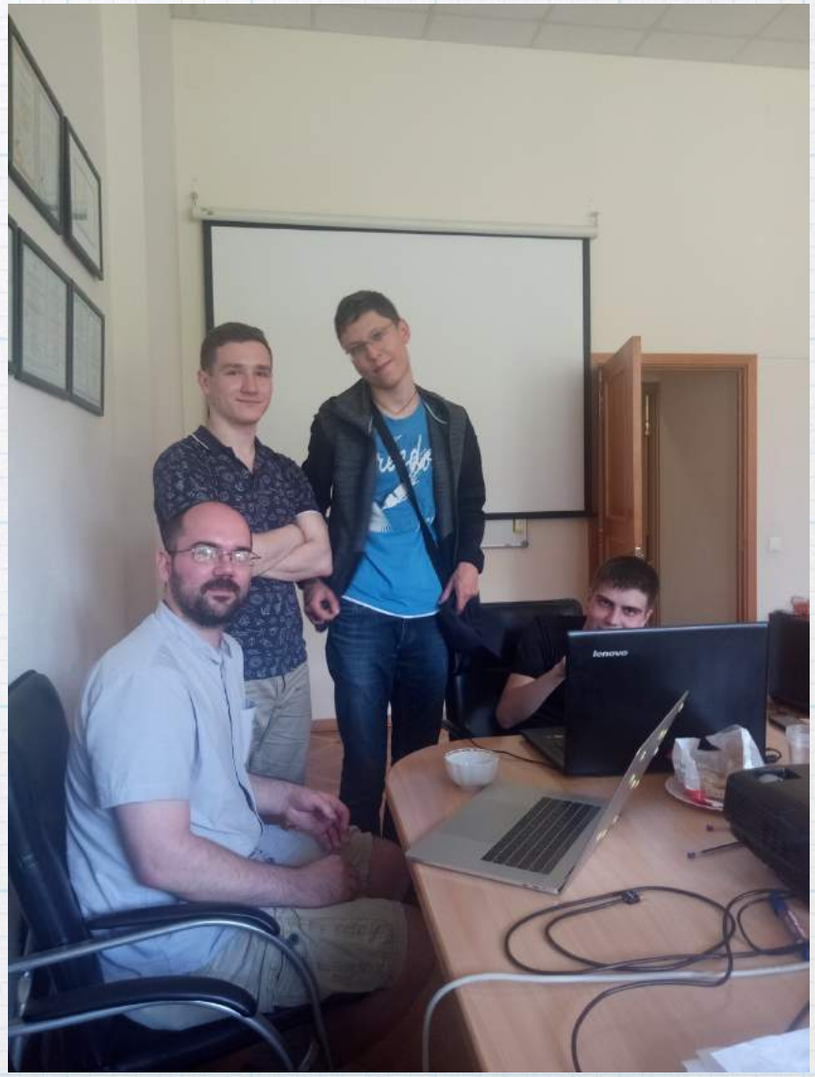
Перспективы развития альтернативных подходов к обработке данных в биоинформатике

* Иван Короткий
vana0pub@gmail.com



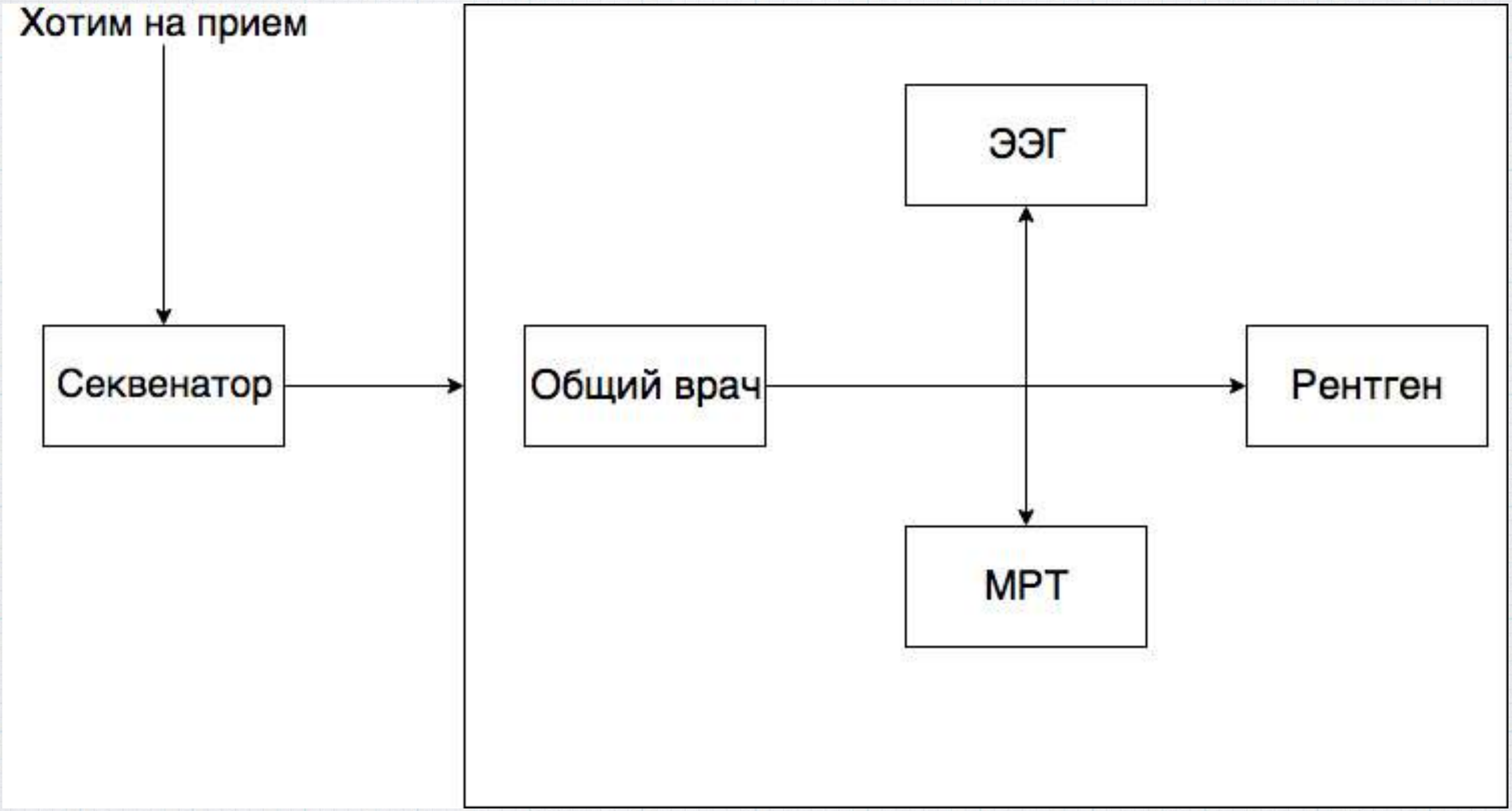
SAINT PETERSBURG
STATE UNIVERSITY
MATHEMATICS AND MECHANICS
DEPARTMENT

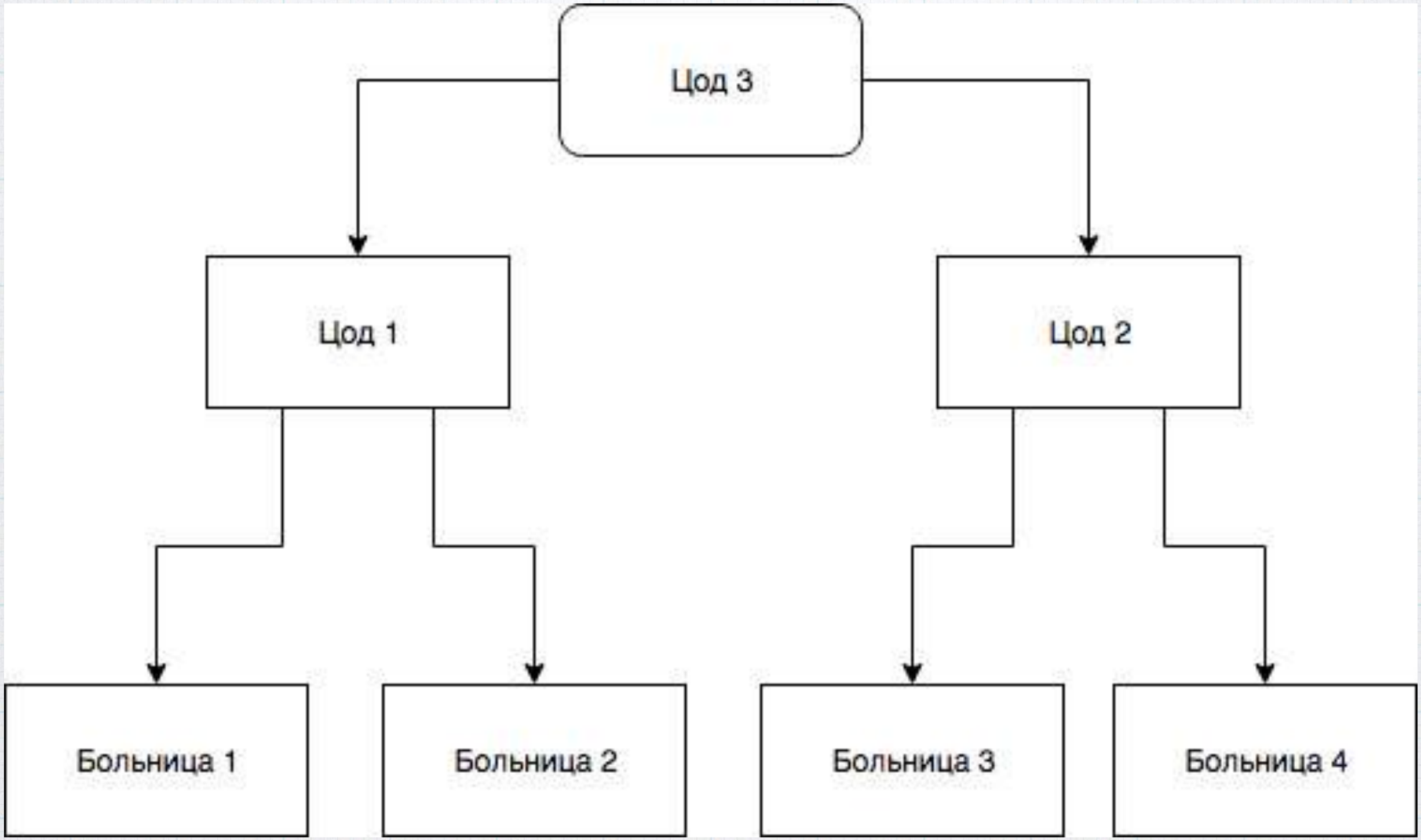
SECR

Что будет в будущем с биоинформатикой?

* 10 лет?
* 30 лет?
* 50 лет?

Обычная больница


Хотим на прием
Секвенатор
Общий врач
ЭЭГ
Рентген
МРТ


Цод 3
Цод 1
Цод 2
Больница 1
Больница 2
Больница 3
Больница 4

Актуально просто изучить предметную область

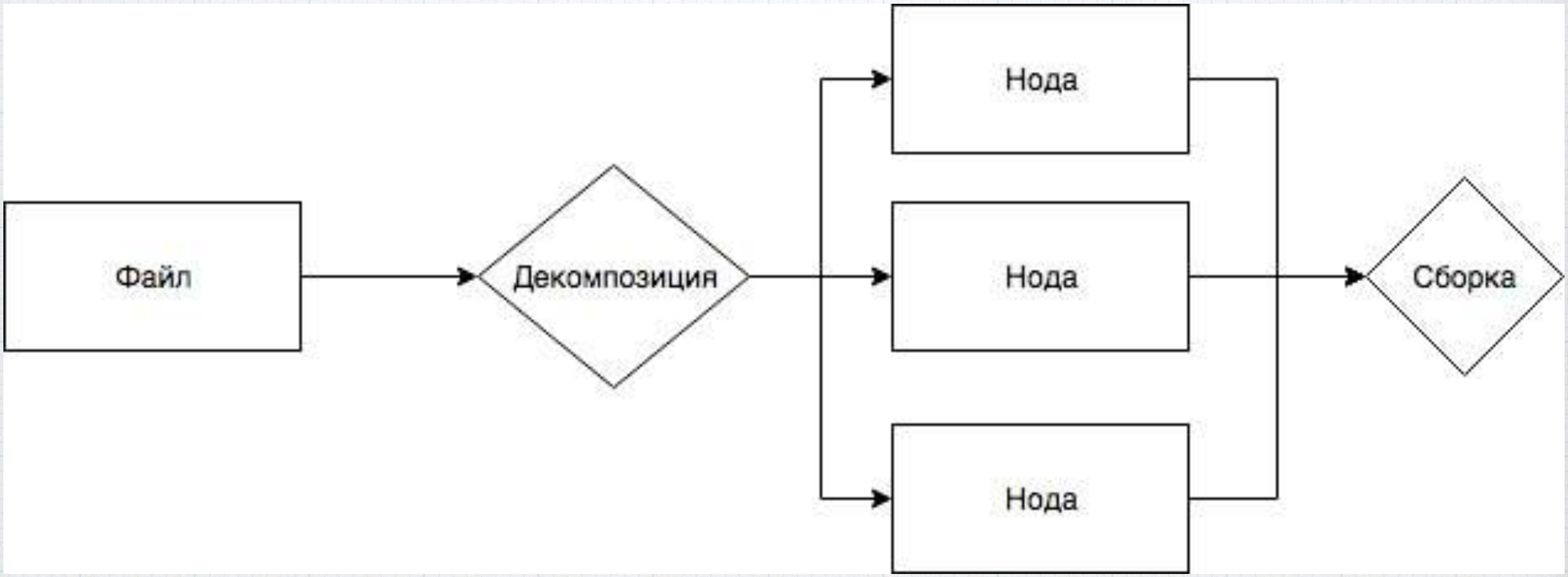
* Чтобы проверить оптимальность конфигурации надо провести эксперимент
* Даже провести эксперимент - уже проблема и это не так тривиально
* Опыт не дает гарантий, что будет эффективно
* Лучший способ это сделать - изучать какую-то проблему вкупе с основной работой.

Введение 1

* Геном - последовательность нуклеотидов размером 3 ГБ за одно его считывание
* В каждом считывании есть ошибки
* Вопрос о избыточности информации в геноме не решен т.е. им занимаются до сих пор (см проект ENCODE)

Ведение 2

* Формат BAM/SAM - стандарт, но при его разработке не задумывались о масштабировании
* Есть региональные базы данных созданные группами стран (ENSEMBL, GENBANK, ENTREZ)
* Основные задачи - выровнять строку на данные из этой базы. Это сложно из-за ошибок секвенатора
* Алгоритмы есть, но ими очень часто не пользуются т.к. методы на основе математической статистики практичнее, хоть и менее точные


Нода
Файл
Декомпозиция
Нода
Сборка
Нода

Наши направления работы

* Декомпозиция BAM/SAM до состояния загрузки в реляционную БД
* Прототип уровня одной больницы с которым можно экспериментировать
* Загрузка в базу данных - очень медленная операция, скорость чтения из неоптимизированной базы эквивалентна с чтением из файлов

Наши направления работы

* Полная замена файлов базами данных возможна, но это стоит памяти (!!на данный момент!!)

* Однако замена перспективна из-за перехода на оптимизатор SQL запросов

* Частичная замена методом выделения промежуточных таблиц с частотными характеристиками более интересна

Возможности, на которые стоит взглянуть

* Применение кластерных решений в процессе обработки учитывая, что данные статичные
* Интеграция с HDFS