

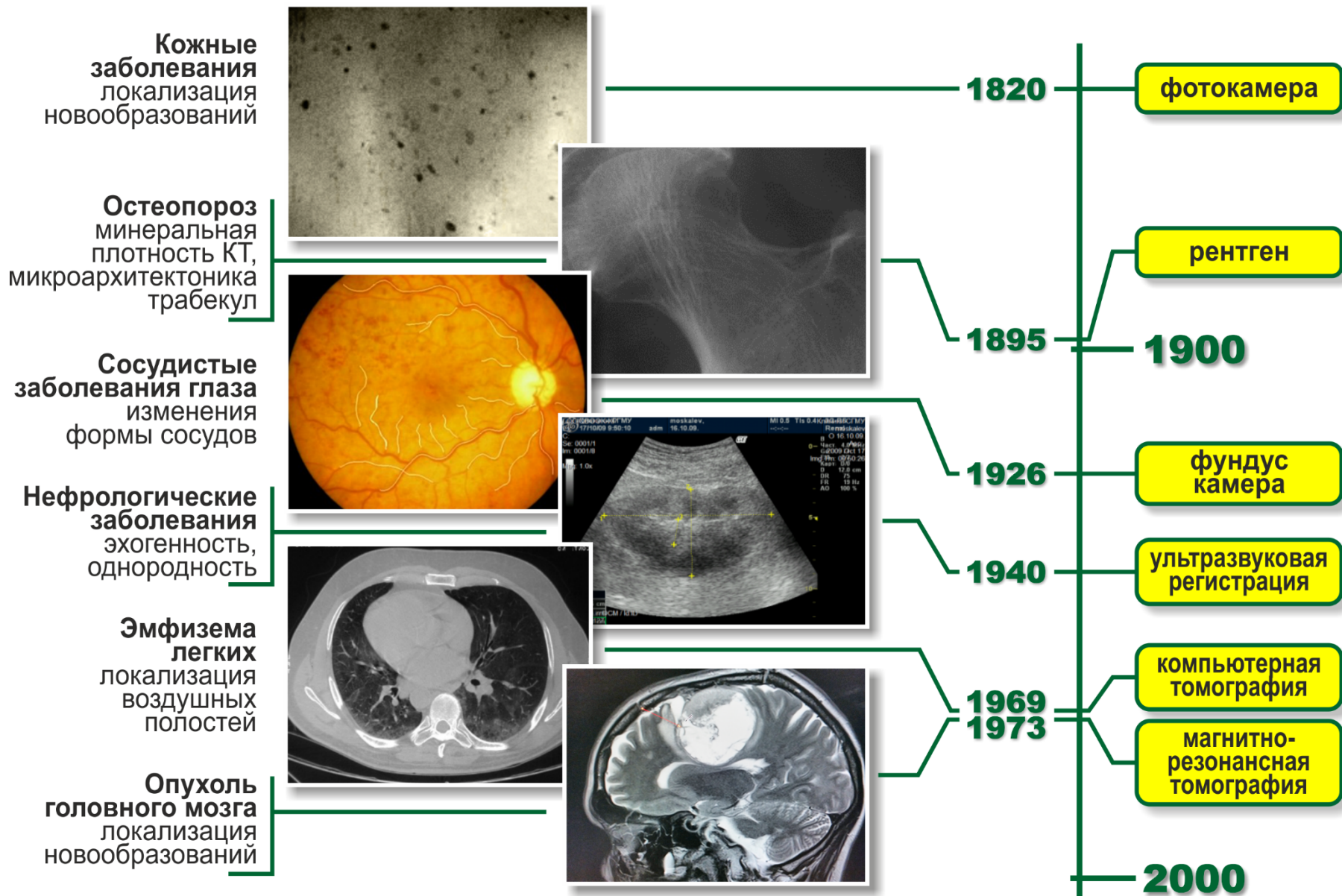


Особенности использования технологий BIG DATA в задачах медицинской диагностики

Авторы доклада:

д.т.н. Ильясова Наталья Юрьевна,
д.т.н. Куприянов Александр Викторович,
аспирант Парингер Рустам Александрович.

Медицинские изображения как носители диагностической информации



Роль, влияние и проблемы Big Data в здравоохранении

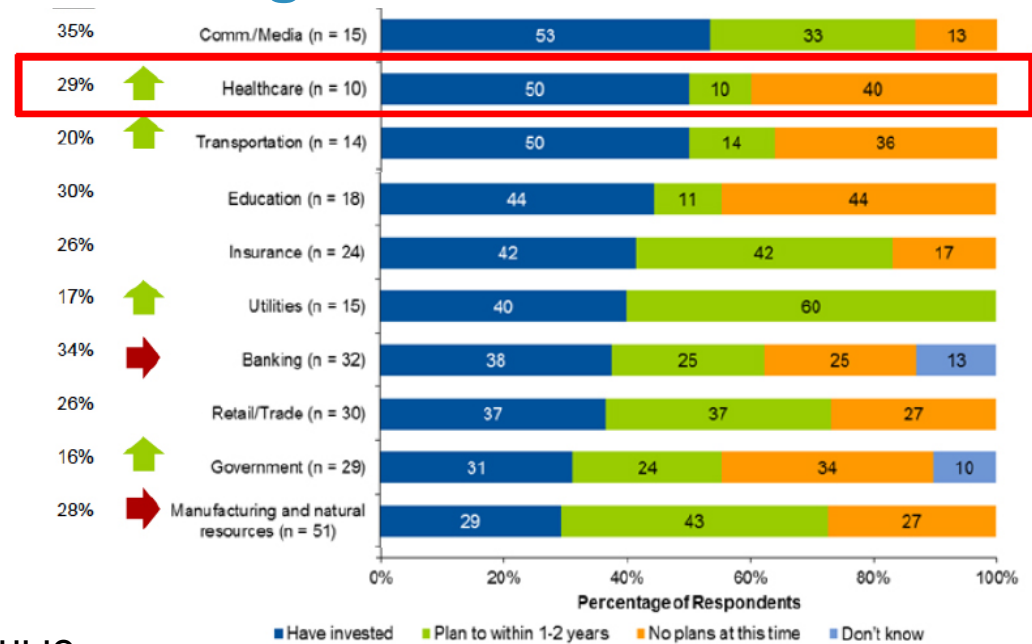
Big Data: Why?

- Более быстрая идентификация рискованных пациентов
- Более эффективные вмешательства
- Принятие лучших решений
- Близкий мониторинг
- Сопоставление значимых клинических данных с результатами Big Data

Big Data: How?

- Массивное хранилище данных
- Все более и более подробные данные для каждого человека — включая геномные данные, сотовую связь, данные о состоянии окружающей среды, историю записей, клинические испытания.
- Масштабные вычисления.
Распределённое и параллельное вычисление на аппаратных средствах.
- Мощная аналитика. Способность обработки большие объёмы данных в пакете или в потоковой передаче.

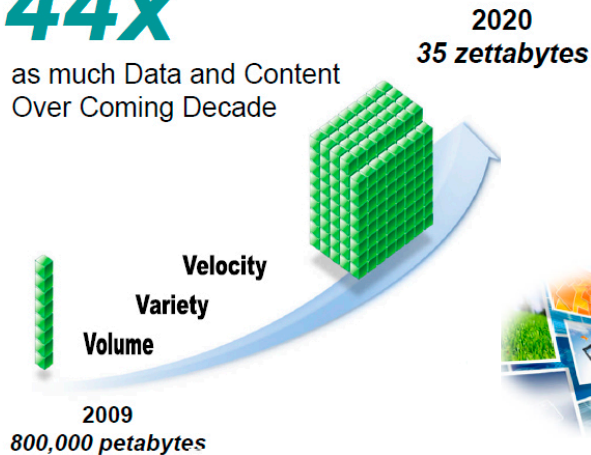
Big Data: инвестиции



Большие данные в медицине: объём и темпы роста

44x

as much Data and Content
Over Coming Decade



90%

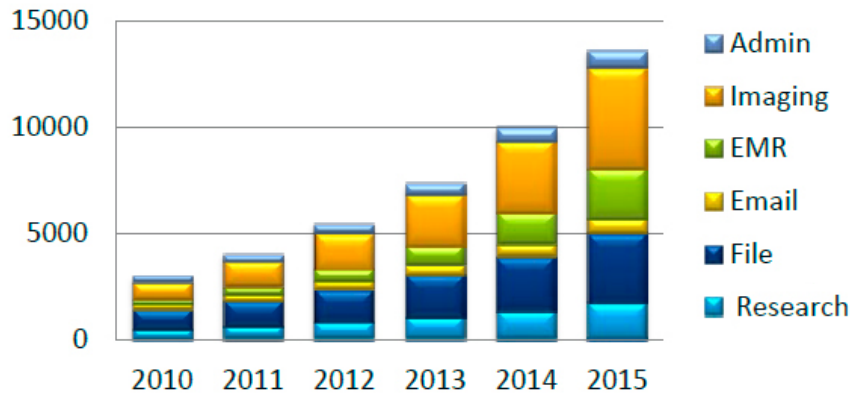
Of world's data
is unstructured



STRUCTURED
DATA

90%
UNSTRUCTURED
DATA

Рост объёма данных (ПБ)

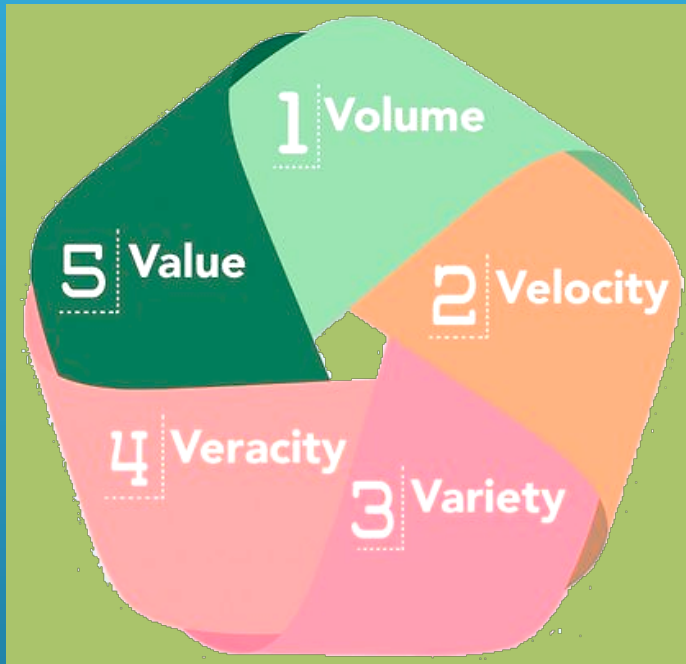


- **90% данных – неструктурированные**
Как правило данные здравоохранения представлены в различных департаментах в различных форматах, из разных источников и из различных клинических систем (80% ЭМК и медицинских изображений (СТ/ПЕТ/MRI))
- **44 кратное увеличение объёма данных в течении предстоящего десятилетия**
(один экзабайт к 2020 году)
- **Большую часть роста данных составляют неструктурированные данные**
(медицинские изображения, видео, текст, речь)
- **Информация ориентирована на индивидуальный подход к пациенту в режиме реального времени** (в дополнение к медицинской информации необходимо знание контекста: оборудование, окружающая среда, демографическая ситуация, социальный и поведенческий профили (5 Терабайт на пациента).

Объём данных по прогнозам достигнет 35 зеттабайт 2020 году с 44-кратным увеличением с 2009 года

«Большие данные» – Big Data

КАК ХАРАКТЕРИСТИКА ДАННЫХ:



КАК ХАРАКТЕРИСТИКА ТЕХНОЛОГИИ:

Технологии **Big Data** – это технологии обработки информации, которые применяются тогда, *когда традиционные технологии* обработки *не применимы* для решения стоящих задач

Большие данные объединяют техники и технологии, которые *извлекают смысл* из данных на экстремальном пределе практичности. [Forrester]

Bigger Data – лучшее решение



В биомедицине существует факт: чем больше размер выборки, тем более точные получаются оценки



Однако размеры большой выборки с плохими по качеству данными могут опасно вводить в заблуждение



IBM Watson Analytics



В здравоохранении точность и надёжность - оба одинаково важны



Анализ данных здравоохранения

Данные проходят 3 этапа перед их использованием для конструктивной аналитики



Инструментарий BIG DATA

• МАШИННОЕ ОБУЧЕНИЕ

Это совокупность подходов, алгоритмов и инструментов, позволяющих делать модели, способные обучаться на основе примеров (индуктивное обучение). Находится на стыке математической статистики, методов оптимизации и дискретной математики, однако имеет собственную специфику, связанную с проблемами вычислительной эффективности и переобучения

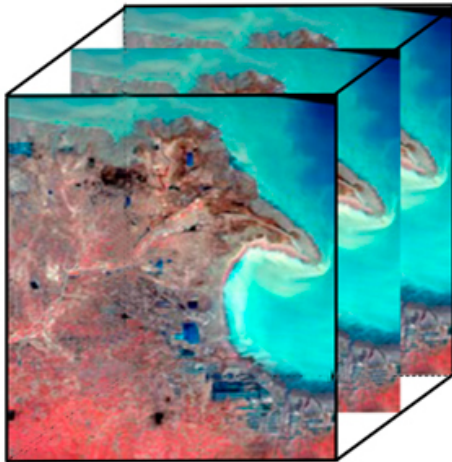
• ИЗВЛЕЧЕНИЕ ИНФОРМАЦИИ

Это задача автоматического построения (извлечения) структурированных данных из неструктурированных или слабоструктурированных машиночитаемых документов. Например, обработка текста на естественном языке

• ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ (Data Mining)

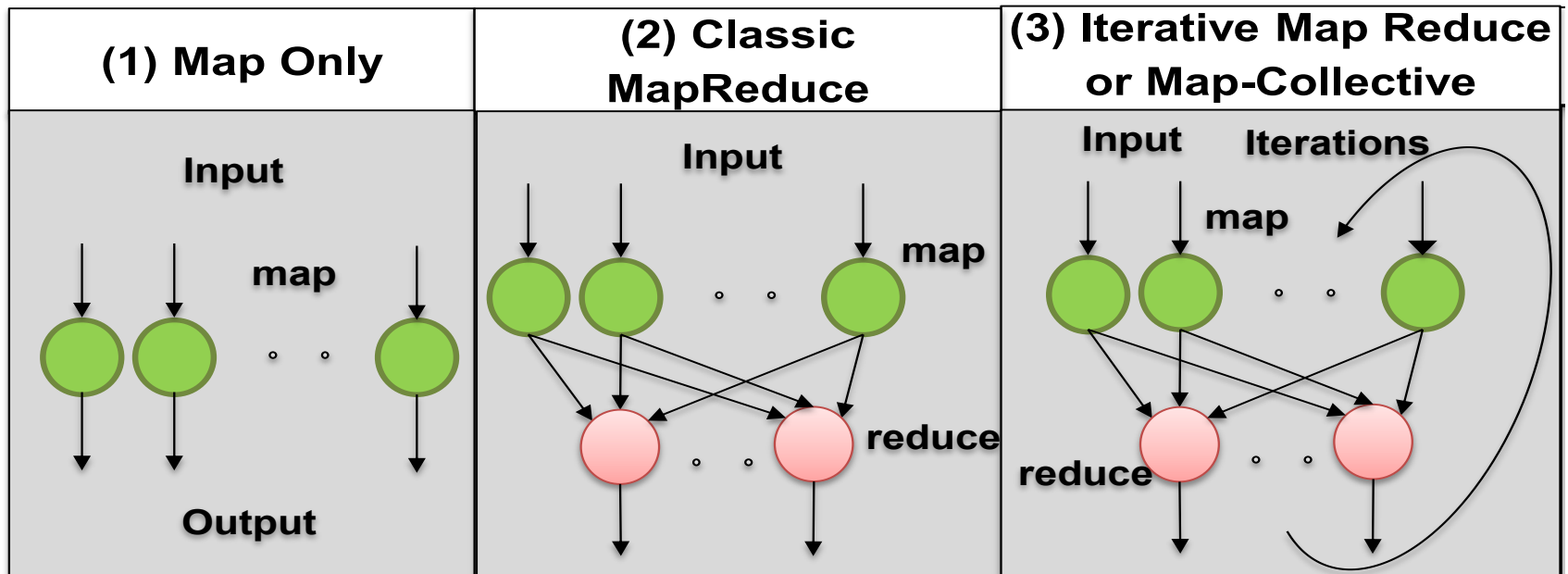
Это совокупность методов обнаружения в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний. Основу методов Data Mining составляют методы классификации, моделирования и прогнозирования, основанные на применении деревьев решений, искусственных нейронных сетей, генетических алгоритмов, эволюционного программирования, ассоциативной памяти, нечёткой логики.

Big Data: технологии Hadoop



Hadoop - информационная среда распределённой обработки и хранения неструктурированных данных, набор программных средств, которые:

- 1) реализуют обработку с использованием методологии MapReduce
- 2) реализуют распределённое отказоустойчивое хранение данных с использованием HDFS

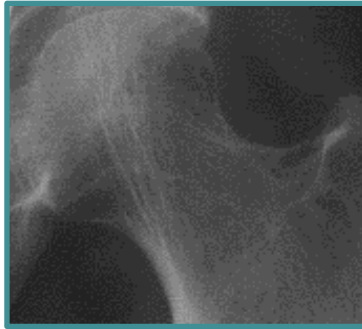


Big Data в здравоохранении: исследования

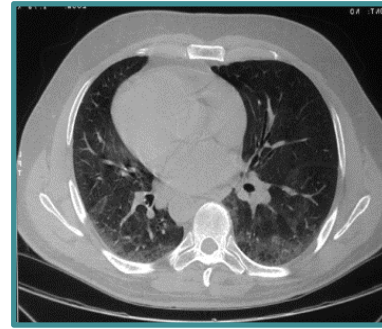
Обработка биомедицинских сигналов



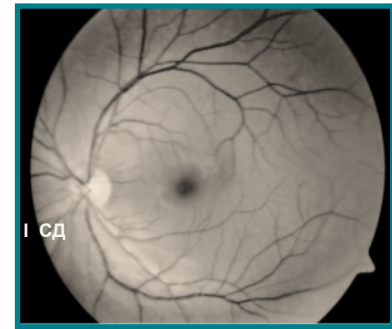
УЗИ почек.
Нефрологические исследования.



Рентген кости.
Остеопороз.
Распознавание на основе текстурного анализа

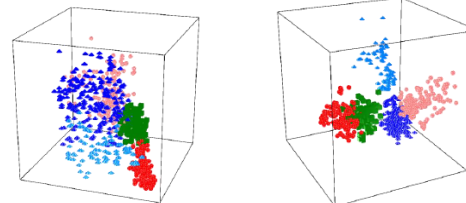
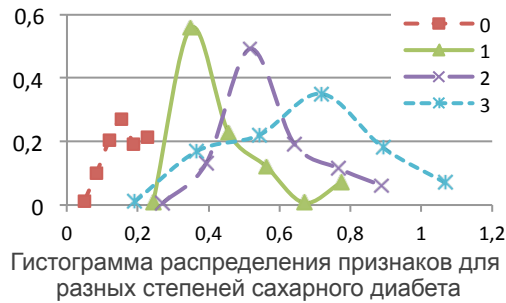


КТ лёгких.
Эмфизема лёгких.



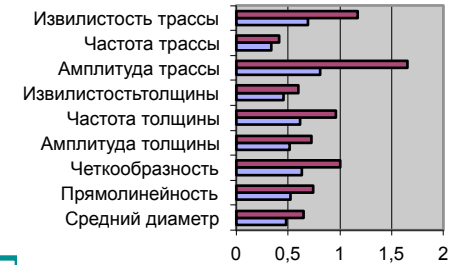
Диагностика глазного дна.
Оценивание геометрических параметров сосудов

Анализ данных

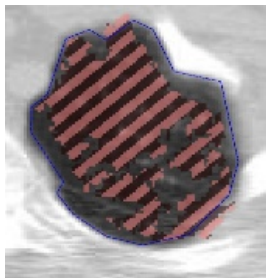


Выборки в пространстве признаков до и после дискриминантного анализа

Артериолы 2 порядка:



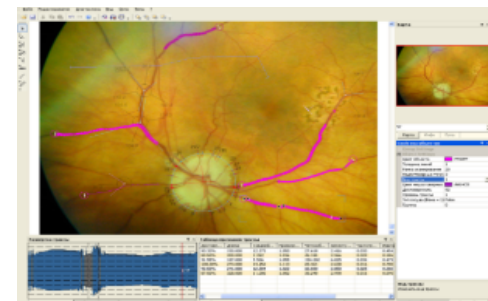
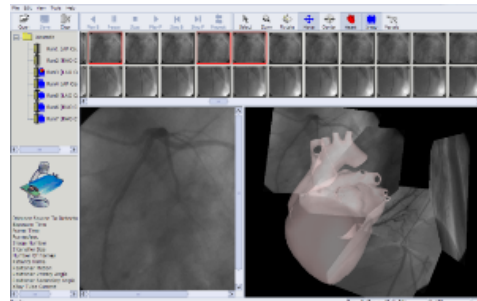
Визуализация и результаты



Выделение областей интереса



Восстановление и визуализация коронарных сосудов



Оценивание признаков сосудов

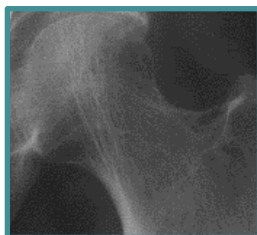
Технологии больших данных в интеллектуальном анализе медицинских диагностических изображений

Анализ биомедицинских изображений

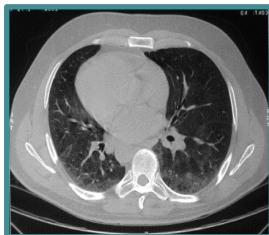


УЗИ почек.
Нефрологические исследования.

Распознавание на основе текстурного анализа



Рентген кости.
Остеопороз.



КТ лёгких.
Эмфизема лёгких.

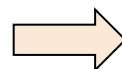


Диагностика глазного дна.
Оценивание геометрических параметров сосудов

Диагностика сахарного диабета по изображениям глазного дна

		J_1	J_2	Повышение критерия разделимости	Ошибка класси- фикации
Артериолы 1 порядка	до	1,98166	3,52161	32%	0,105
	после	2,62686	5,06405		0,024
Артериолы 2 порядка	до	2,26688	4,06699	14,1%	0,073
	после	2,58673	5,28026		0,018
Венулы 1 порядка	до	2,05964	3,60372	25,9%	0,061
	после	2,59395	4,86507		0,020
Венулы 2 порядка	до	2,36223	4,01792	15,8%	0,143
	после	2,61406	4,86384		0,018

Использование технологий Big Data в разработанных системах медицинской диагностики позволило за счёт привлечения разнородных источников диагностической информации **усовершенствовать обучающую выборку** и повысить достоверность постановки диагноза.



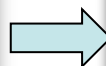
Результаты:

- Разработаны **системы интегральных показателей** патологических изменений, позволившие увеличить спектр диагностической информации.
- Разработаны **новые полиномиальные признаки**, допускающие согласование с текстурными свойствами рентгеновских диагностических изображений, что позволило повысить достоверность распознавания остеопороза с 0,9 до 0,95.
- Предложен **метод автоматического выделения области интереса** в компьютерной томографии (КТ) лёгких, позволивший повысить достоверность распознавания эмфиземы лёгких с 0,94 до 0,97.

Набор данных	Повышение разделимости классов	Достовер- ность
Рентген костей	8%	0,95
УЗИ почек	23%	0,87
КТ лёгких	17%	0,97
Кровеносные сосуды	21%	0,94

Информационные технологии интеллектуального анализа биомедицинских изображений

Информационная технология формирования пространства эффективных признаков

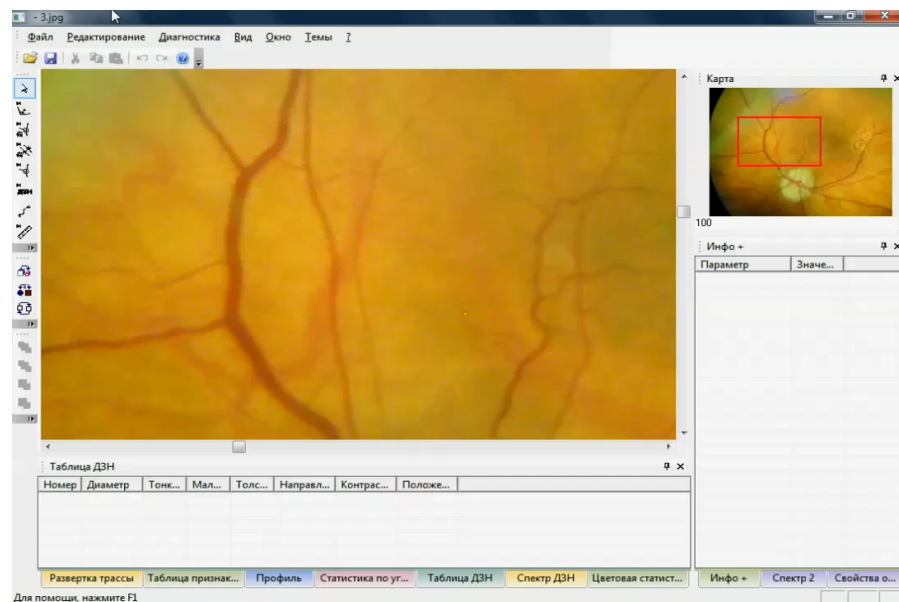


Компьютерная система ранней диагностики глазных заболеваний (сахарный диабет)

- Предварительная обработка изображений.
- Количественная оценка диагностических признаков, основных типов изменений;
- Объективный контроль динамики изменения размеров патологических участков на последовательных изображениях;
- Оценка степени патологии.

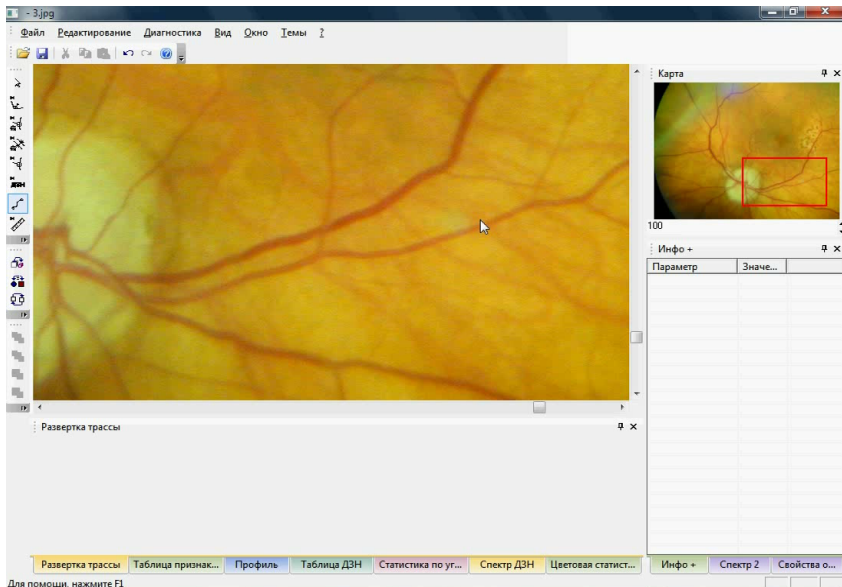
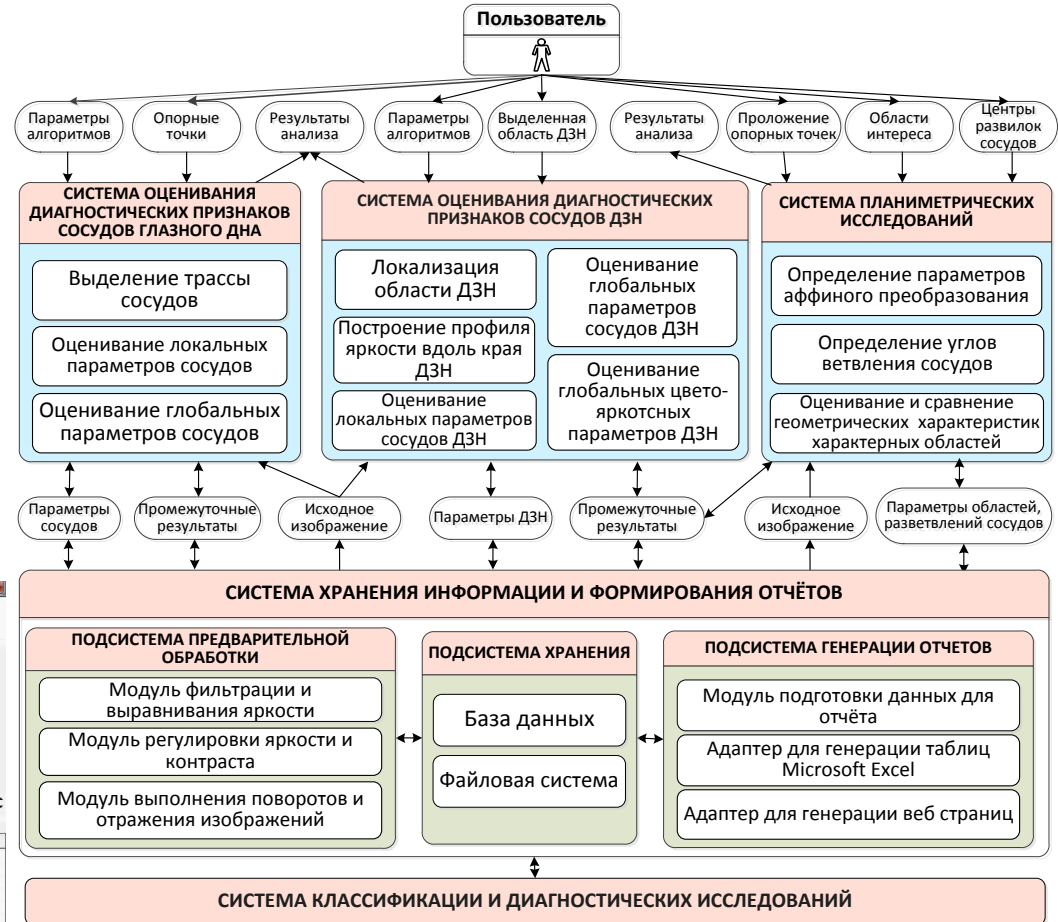
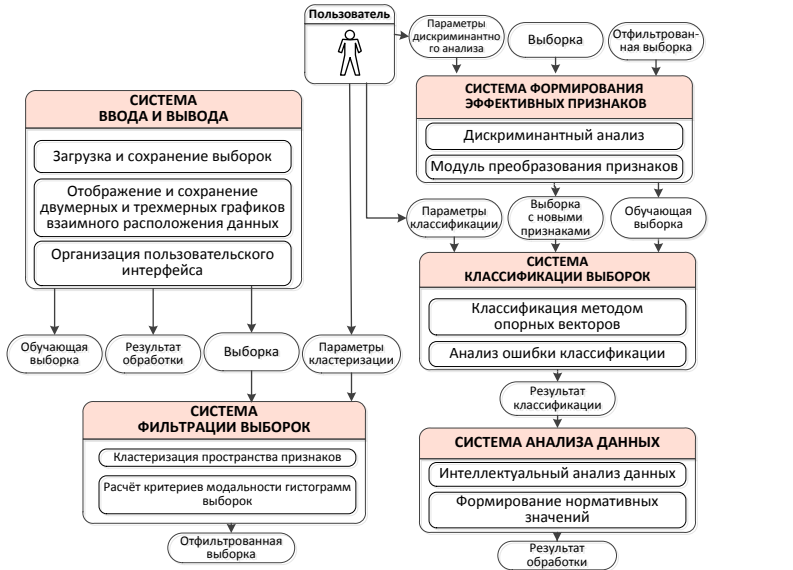
BIG DATA:

- + Усовершенствование обучающей выборки;
- + Уменьшение ошибки классификации;
- + Повышение достоверности постановки диагноза;
- + Уменьшение время диагностики;
- + Телемедицина, удалённый мониторинг.



Подсистема планиметрических измерений на изображениях глазного дна

Компьютерная система анализа сосудов глазного дна «OPHTHALMOFFICE»



Архитектура программного комплекса
«OphthalmOffice»

Анализ информативности и построение эффективных диагностических признаков анализа биомедицинских изображений

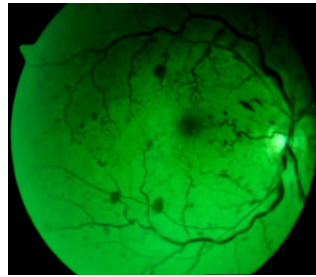
Дискриминантный анализ на основе максимизации критериев разделимости:

$$J_1 = \text{tr}((\mathbf{W} + \mathbf{B})^{-1} \mathbf{B}); \quad J_2 = \ln |\mathbf{W}^{-1} (\mathbf{W} + \mathbf{B})|$$

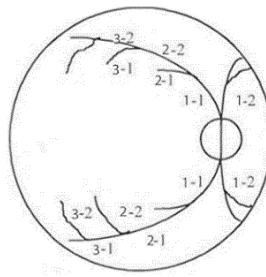
\mathbf{W} — матрица рассеяния внутри классов \mathbf{B} — матрица рассеяния между классами



норма

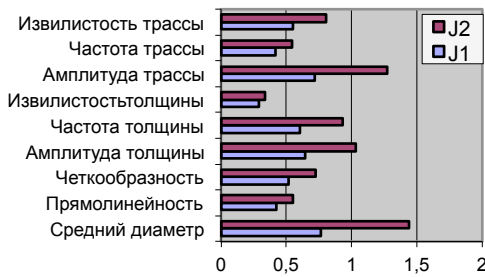


патология (4 СД)

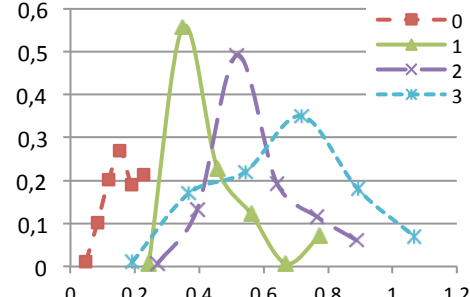
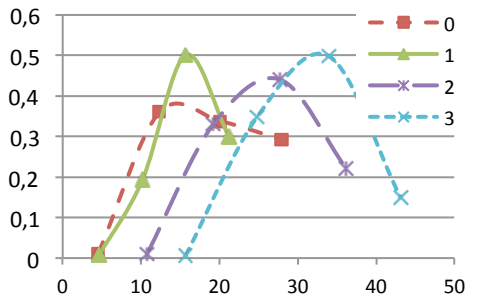
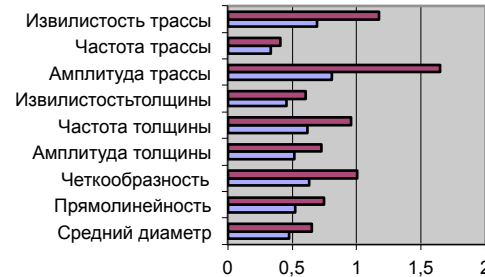


измерения

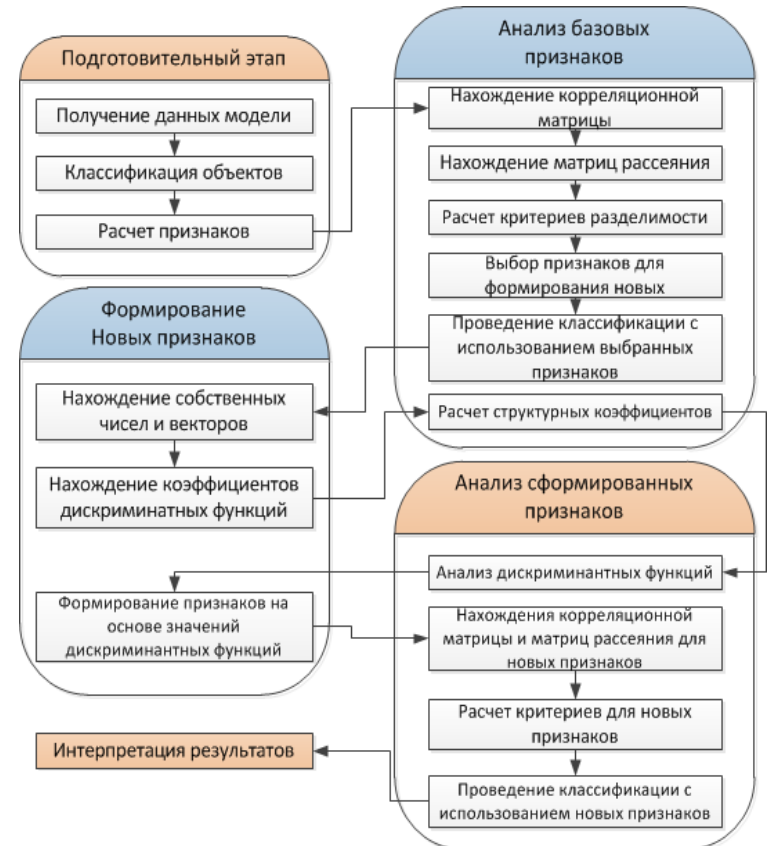
Артериолы 1 порядка:



Артериолы 2 порядка:



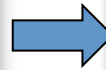
Гистограмма распределения признаков для разных степеней сахарного диабета



Алгоритм анализа информативности и построения эффективных диагностических признаков

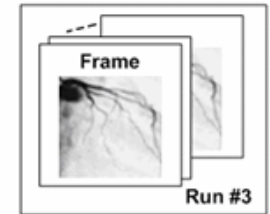
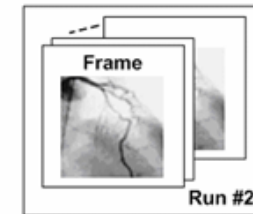
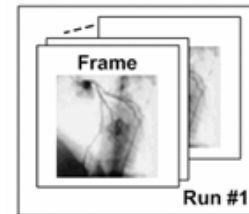
Информационные технологии интеллектуального анализа биомедицинских изображений

Информационная технология восстановления пространственной структуры сосудов сердца и оценивания их геометрических признаков



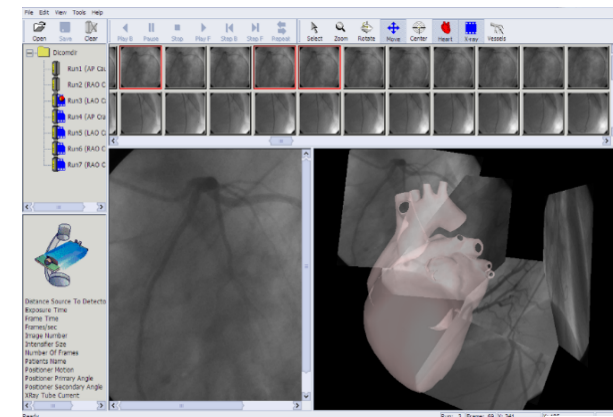
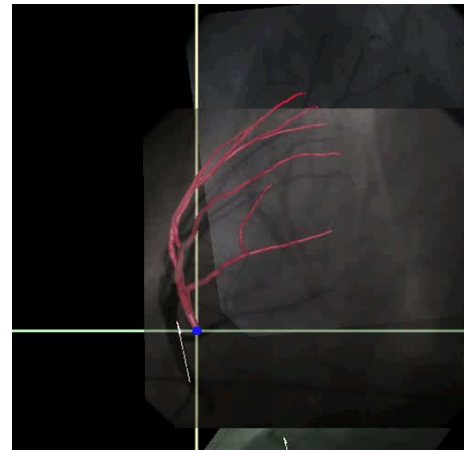
Компьютерная система визуализации, диагностирования стенозированных сосудов

- Моделирование геометрических параметров съемки, компенсация геометрических искажений и синхронизация проекций;
- Численное моделирование и оценивание геометрических параметров пространственной структуры коронарных сосудов;
- Визуализация пространственной структуры коронарных сосудов.



BIG DATA:

- + Визуализация данных для клиницистов;
- + Онлайн-обучающие курсы диагностики и экспертизы сложных патологий;
- + Удалённый мониторинг.



Лаборатория обработки данных сверхбольшого объёма

Институт систем обработки изображений РАН,
Самарский государственный аэрокосмический университет имени академика С.П. Королёва
(национальный исследовательский университет)

Специализированный программно-аппаратный комплекс хранения/обработки Big Data

- Программно-аппаратный комплекс (подсистема) хранения и аналитического анализа структурированных данных **IBM Puredata for Analytics (Netezza)**;
- Комплекс из **5** серверов **IBM System X** для подсистемы хранения и аналитической обработки неструктурированных данных **IBM InfoSphere BigInsights**.

Задачи лаборатории:

- Разработка моделей, методов и технологий сбора, хранения, обработки, визуализации, методов формирования знаний на основе интеллектуального анализа неструктурированной информации
- Обучение специалистов в области создания обучаемых экспертных систем поддержки принятия корпоративных решений, основанных на методах и платформах технологии Big Data.

