Управление данными как шаг к применению AI

Антон Агеев
РСХБ-Интех

РСХБ цифра

РоссельхозБанк

Знакомство

Агеев Антон Михайлович

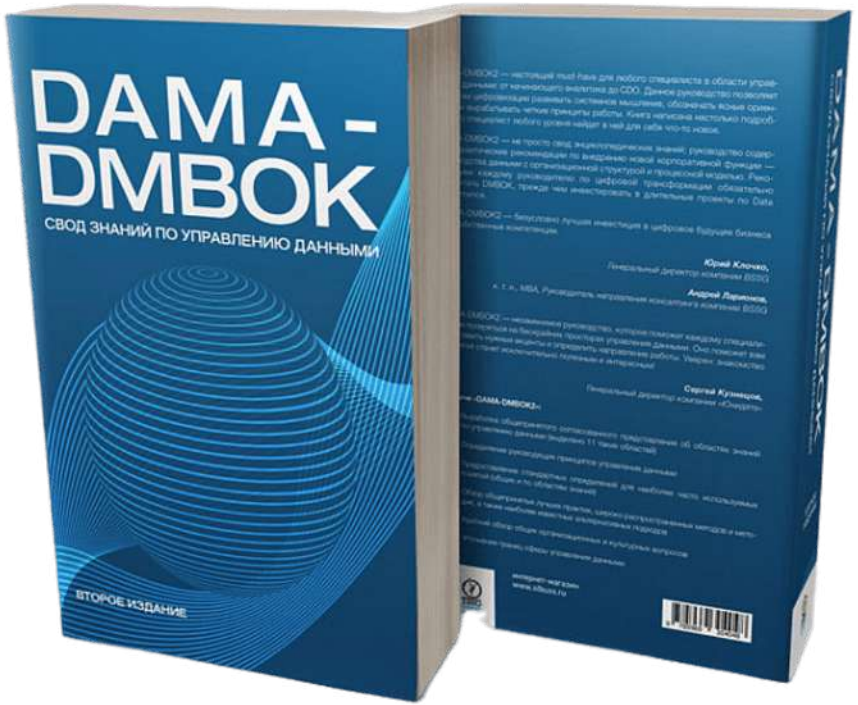
Исполнительный директор РСХБ-Интех
Более четырнадцати лет опыта планирования и воплощения цифровых стратегий крупнейших агрохолдингов России. Практическое решение вызовов четвертой промышленной революции в сельском хозяйстве

Инновационные проекты в агро:

- Беспилотные комбайны
- Машинное зрение и слух
- Цифровые двойники предприятий

DMBoK2 (Data Management Body of Knowledge)

О чем, для чего и как это связано с AI и галлюцинациями?



Часть 1: Введение в управление данными

- виды данных
- источники данных
- требования к управлению данными
- процессы управления данными
- инструменты для работы с данными

Часть 2: Управление данными жизненного цикла

Конкретные шаги и методы для эффективного управления данными в различных сферах деятельности

Часть 3: Управление метаданными

Основные понятия
и принципы управления метаданными:
виды, источники, требования
и процессы.

Методы и инструменты для сбора,
хранения, анализа
и использования метаданных
в различных отраслях.

Часть 4: Управление правами доступа к данным

Вопросы управления правами доступа:

- создание политик доступа
- управление ролями пользователей
- контроль доступа к данным

Методы и инструменты для обеспечения безопасности данных:

- Шифрование
- Аутентификация
- авторизация

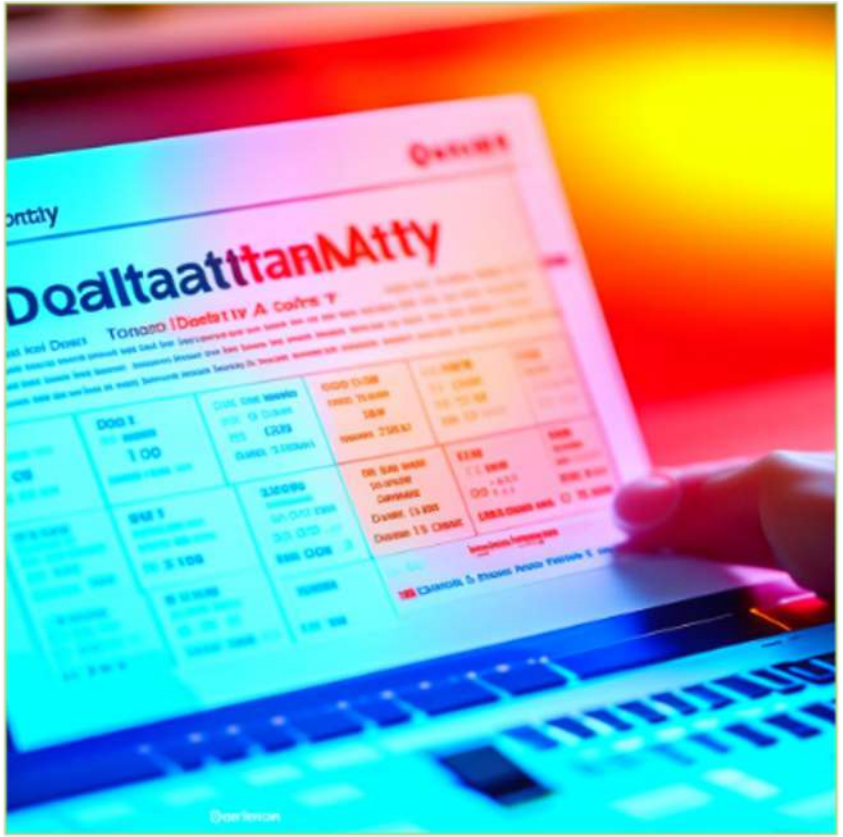
Часть 5: Управление качеством данных

Процесс управления качеством данных:

- Сбор
- Хранение
- Анализ
- Обработка данных с учетом требований качества

Методы оценки качества данных:

- проверка корректности
- валидация и верификация данных



Часть 6: Управление безопасностью данных

Управление безопасностью данных:

- защита от угроз
- контроль доступа
- мониторинг действий пользователей

Методы и технологии для защиты данных от кибератак, вирусов и других угроз

Вопросы аутентификации пользователей, шифрования и авторизации доступа к данным

Методики DAMA

Data Access Management

Управление доступом к данным

- Определение ролей и прав доступа к данным
- Управление правами доступа
- Контроль доступа
- Мониторинг доступа
- Аудит доступа

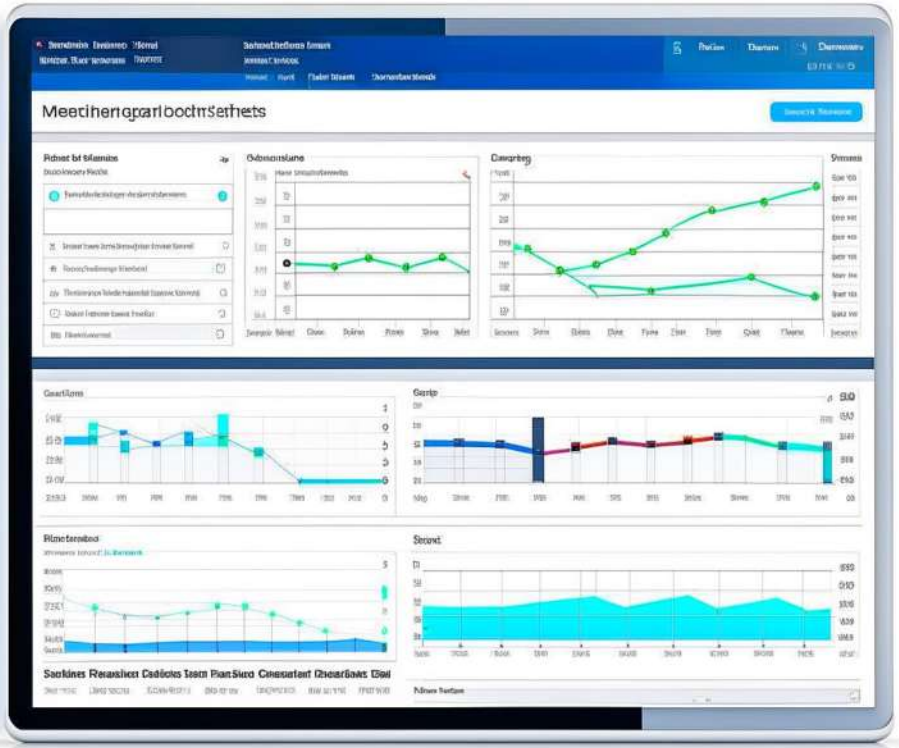
Методики DMIS

Data Monitoring and Information Systems
Мониторинг и анализ данных в реальном времени

Набор методик и инструментов, которые используются для мониторинга данных в реальном времени.

Методики DMIS

Позволяют отслеживать изменения
в данных, выявлять аномалии
и определять тенденции



Методики DMIS

Помогают организациям получать более точную и актуальную информацию о своих данных

Позволяет принимать более обоснованные решения и повышать эффективность своих бизнес-процессов

Методики DMIS

Один из основных методов DMIS — **мониторинг данных с помощью временных рядов**

Временной ряд — это последовательность значений данных, которые измеряются в определенный момент времени

Мониторинг временных рядов позволяет отслеживать изменения в значениях данных, выявлять отклонения от нормального поведения и прогнозировать будущие значения

Методики DMIS

Метод **Data Mining** позволяет извлекать скрытые закономерности и взаимосвязи из больших объемов данных

Data Mining может использоваться для выявления аномалий в данных или для построения моделей прогнозирования

Методики DMIS

В рамках DMIS рассматриваются методы визуализации данных

Они позволяют представить данные в виде графиков, диаграмм и таблиц, что упрощает их понимание и анализ

Методики DSS

Data Storage and Security
Хранение и защита данных

Набор методов и инструментов, которые используются для обеспечения безопасности данных и их хранения:

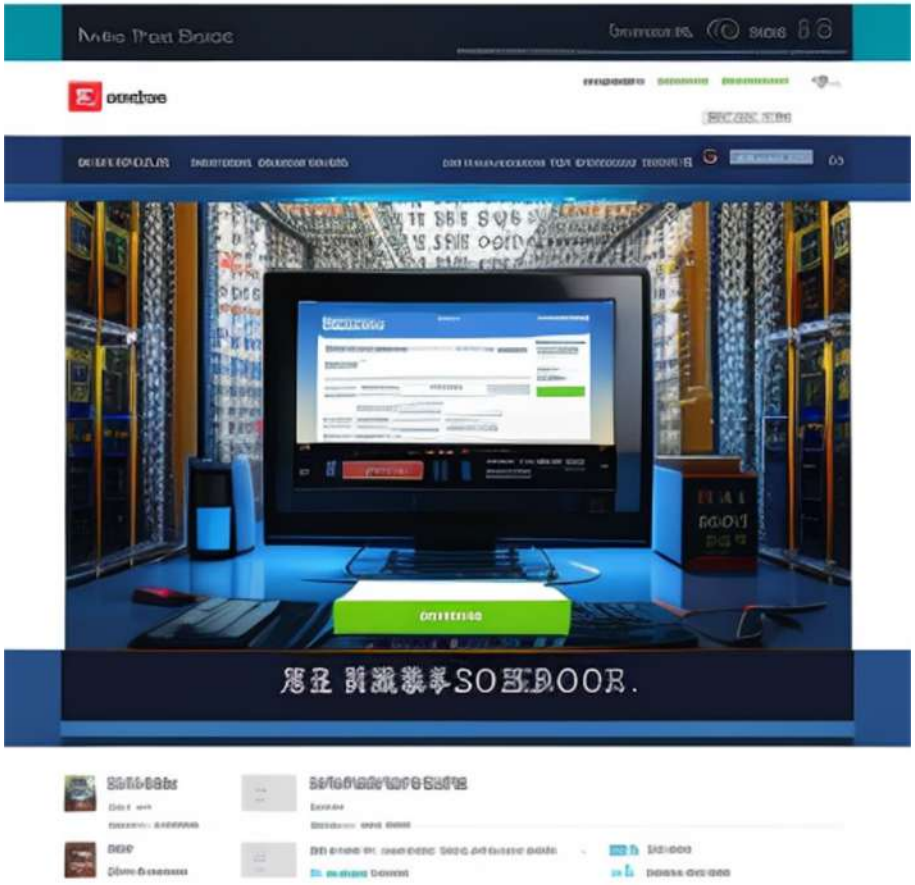
1. Шифрование данных
2. Аутентификация
3. Контроль доступа
4. Резервное копирование
5. Мониторинг и аудит
6. Обучение пользователей
7. Регулярные обновления

Методики DDM

Data Discovery and Management
Обнаружение, сбор, хранение
и анализ данных

Методики DDM

Сбор всех доступных данных
из различных источников:
Базы данных
Файлы
Сайты
Электронная почта и т.д.



Корпоративные источники для майнинга данных:

Финансовые данные Доходы, расходы, активы, обязательства

Данные о клиентах Покупки, предпочтения, демографические данные

Операционные данные Производственные процессы, использование ресурсов, качество продукции

Рыночные данные Конкуренты, тенденции рынка, цены на товары и услуги

Данные об управлении персоналом Информация о сотрудниках, навыки, опыт работы, уровень удовлетворенности

Майнинг данных в производственных процессах

Сбор данных с использованием промышленных контроллеров, датчиков и других устройств, собирающих информацию о процессах в режиме реального времени

Использование специализированных программных решений для мониторинга и анализа производственных процессов

Внедрение систем управления производственными процессами (например, MES-систем), позволяющих собирать, хранить и анализировать данные о производстве

Применение технологий искусственного интеллекта и машинного обучения для анализа данных о производстве и выявления закономерностей и тенденций

Обеспечение доступа к данным о производстве для аналитиков и специалистов по data mining, чтобы они могли проводить исследования и выявлять новые возможности для оптимизации процессов

Методики DDM

Анализ данных для выявления закономерностей, тенденций и корреляций

Для выявления скрытых взаимосвязей между параметрами можно использовать корреляционный, регрессионный или кластерный анализ

Для определения тенденций и закономерностей могут использоваться методы временных рядов или анализа больших данных

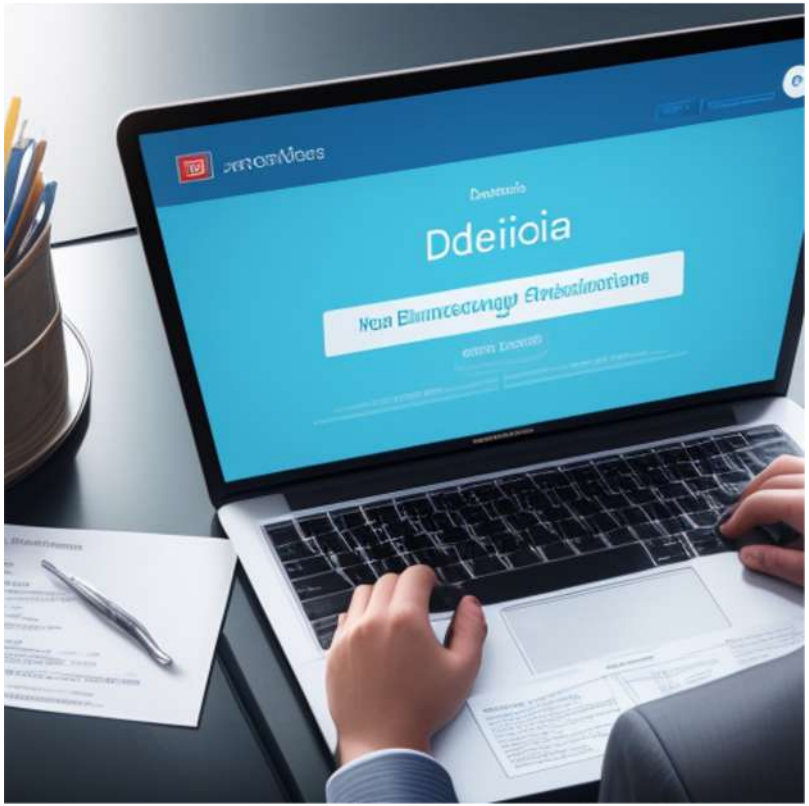
Методики DDM

Визуализация данных представление данных в виде графиков, диаграмм и других визуализаций для более удобного анализа

Методики DDM

Извлечение знаний:
процесс получения полезной информации из данных, которые были извлечены из различных источников

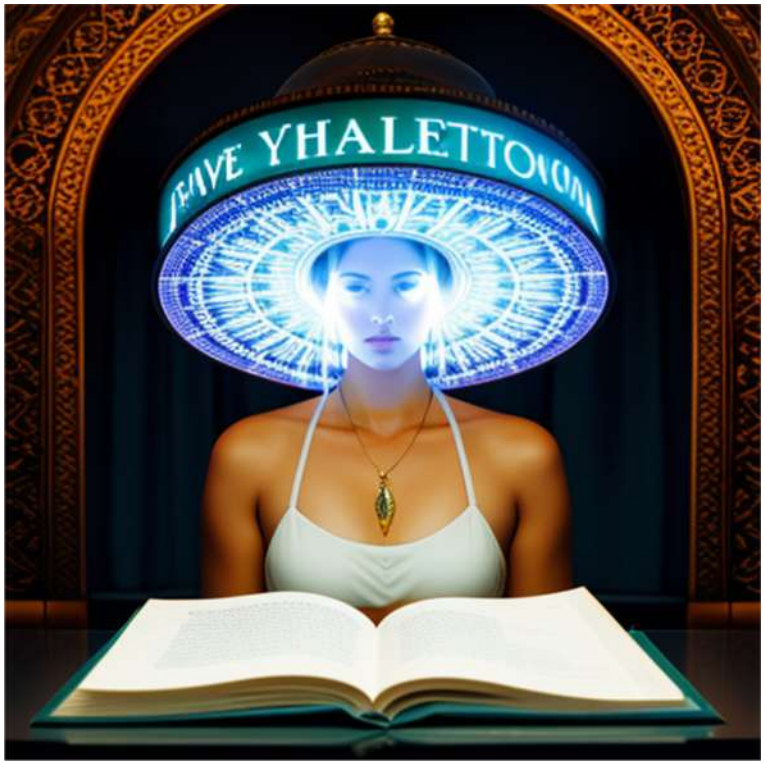
выполняется для принятия решений и оптимизации бизнес-процессов



Извлечение знаний

Существует несколько подходов к извлечению знаний:
1. Машинное обучение
2. Экспертные знания
3. Комбинированный подход
4. Визуализация

Извлечение знаний — непрерывный процесс, который требует постоянного обновления и анализа новых данных



Как внедрить DDM

Регулярно выполнять:

- определение потребностей и целей
- сбор данных
- обработку и очистку данных
- визуализацию данных
- анализ данных
- принятие решений на основе данных

Известные продукты, реализующие DDM

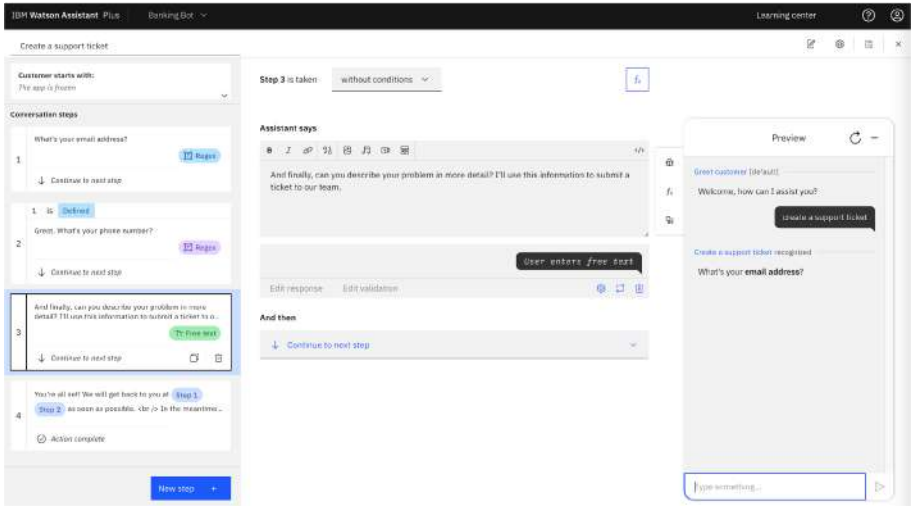
IBM Watson

Система искусственного интеллекта, способная обрабатывать и анализировать большие объемы данных.

Может использоваться для распознавания речи, обработки естественного языка, рекомендательных систем, компьютерного зрения и т.д.

Основные функции IBM Watson:

1) Обработка естественного языка (NLP)
2) Извлечение информации
3) Рекомендации и прогностическая аналитика
4) Распознавание речи и преобразование текста в речь
5) Анализ настроений и мнений
6) Обучение и самообучение
7) Интеграция с другими сервисами и приложениями



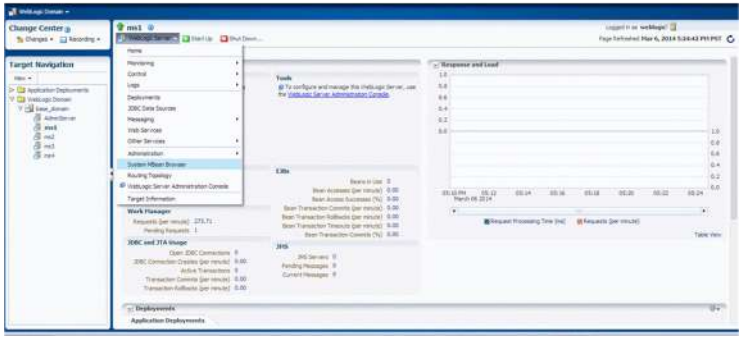
Известные продукты, реализующие DDM

Oracle Coherence

Высокопроизводительное ПО для обеспечения согласованности данных в распределенных системах

Предоставляет решения для кэширования, синхронизации и распределения данных между узлами в кластере

Может использоваться для оптимизации доступа к данным, повышения производительности и масштабируемости приложений, работающих на кластерах или в облачных средах

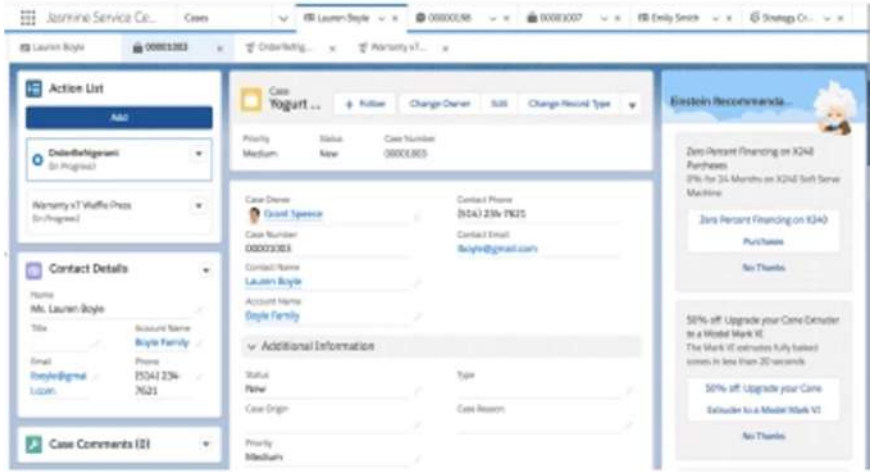


Известные продукты, реализующие DDM

Salesforce Einstein

Набор функций на базе искусственного интеллекта, помогающий компаниям принимать более обоснованные решения, оптимизировать операции и улучшать качество обслуживания клиентов

Включает в себя:
Einstein Discovery
Einstein Bots
Einstein Recommendations
Einstein Vision
Einstein Sales Insights



Data Mining with open source

WEKA - Waikato Environment for Knowledge Analysis

Программная платформа для машинного обучения и интеллектуального анализа данных, разработанная в Университете Вайкато в Новой Зеландии. Создана в 1993 году, с тех пор активно развивается и улучшается

Включает в себя инструменты:

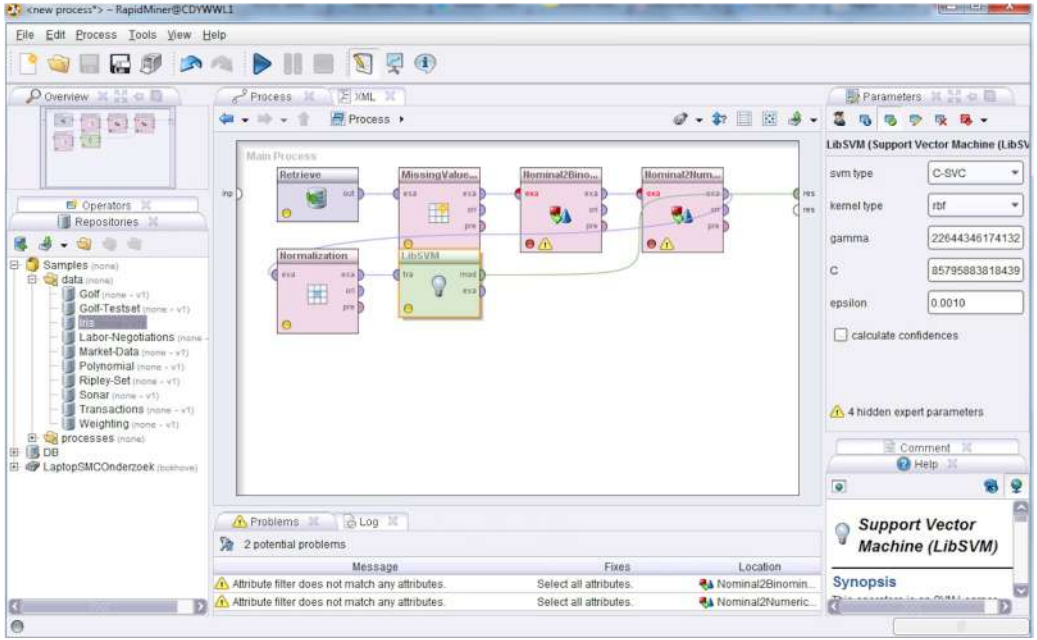
- Проводник
- Экспериментатор
- Поток знаний
- «Верстак»
- Простой интерфейс командной строки

Data Mining with open source

RapidMiner

Представляет собой интегрированную среду для обработки данных в больших информационных массивах, машинного обучения, текстовой аналитики и построения прогностических моделей, а также для решения иных задач

Год выхода: 2006 г.
Платформа: Java Virtual Machine



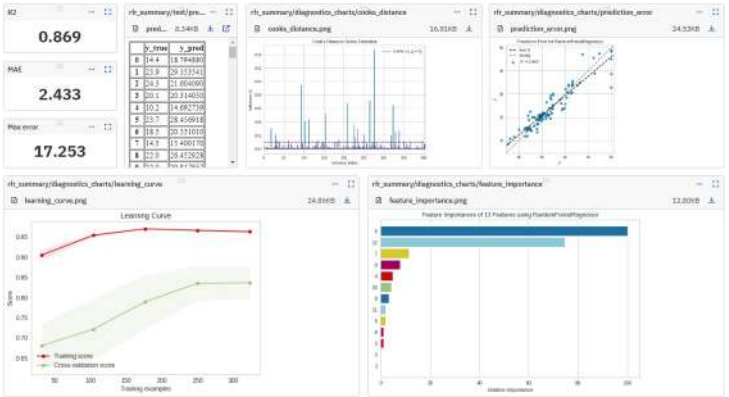
Data Mining with open source

Scikit-Learn

Библиотека Python для машинного обучения. Содержит множество алгоритмов Data Mining

Предоставляет набор инструментов для выполнения задач машинного обучения, таких как кластеризация, регрессия, классификация и снижение размерности

Простой и понятный интерфейс, позволяющий разработчикам легко создавать и тестировать различные модели машинного обучения

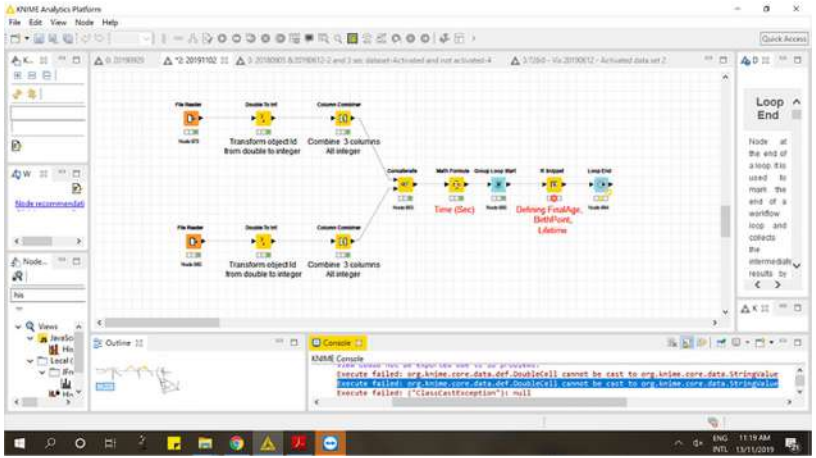


Data Mining with open source

KNIME

Платформа с открытым исходным кодом для обработки и анализа данных, включает в себя различные инструменты для Data Mining

Представляет собой no-code решение для обработки и анализа данных, построения процессов обработки данных, визуализации и машинного обучения



Год выхода: январь 2004 г.
Платформа: Java Virtual Machine

Orange

Среда для машинного обучения и Data Mining

Программная среда с открытым исходным кодом для визуализации, анализа и моделирования данных.
Разработана на Python, доступна для Windows, macOS и Linux

Широкий спектр инструментов для обработки и визуализации данных: графики, диаграммы, таблицы.
Также включает алгоритмы машинного обучения: классификация, регрессия, кластеризация.

Имеет простой и интуитивный интерфейс, позволяющий легко создавать модели машинного обучения без необходимости написания сложного кода.
Может быть интегрирован с другими программами, такими как R и Python, что расширяет его возможности

Версии до 3.0 включают основные компоненты на C ++ с оболочками на Python.
Начиная с версии 3.0, Orange использует для научных вычислений распространенные библиотеки открытым исходным кодом языке Python, такие как numpy, scipy и scikit-learn.
Графический пользовательский интерфейс работает в рамках кроссплатформенной платформы Qt framework

Orange

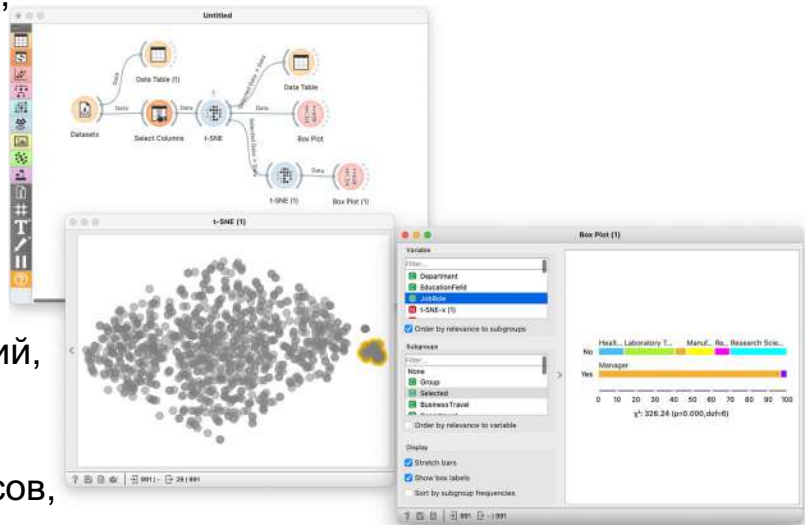
Применение:

Банковское дело и финансы: Анализ кредитных рисков, предсказание дефолтов по кредитам, оценка стоимости кредитов

Медицина: Анализ медицинских данных, диагностика заболеваний и прогнозирование исхода лечения

Маркетинг: Анализ поведения пользователей на сайте, определение наиболее эффективных рекламных кампаний, предсказание продаж

Производство: Оптимизация производственных процессов, предсказание отказа оборудования, улучшение качества продукции



Выводы

- Управление данными — это основополагающий шаг к применению AI, включая GPT модели
- Управление данными недооценено
- Проприетарные продукты не обязательны для начала внедрения AI
- Управление данными — это прекрасная точка роста для молодых специалистов

Спасибо за внимание!

ArchitectureVision

РСХБ цифра

РоссельхозБанк