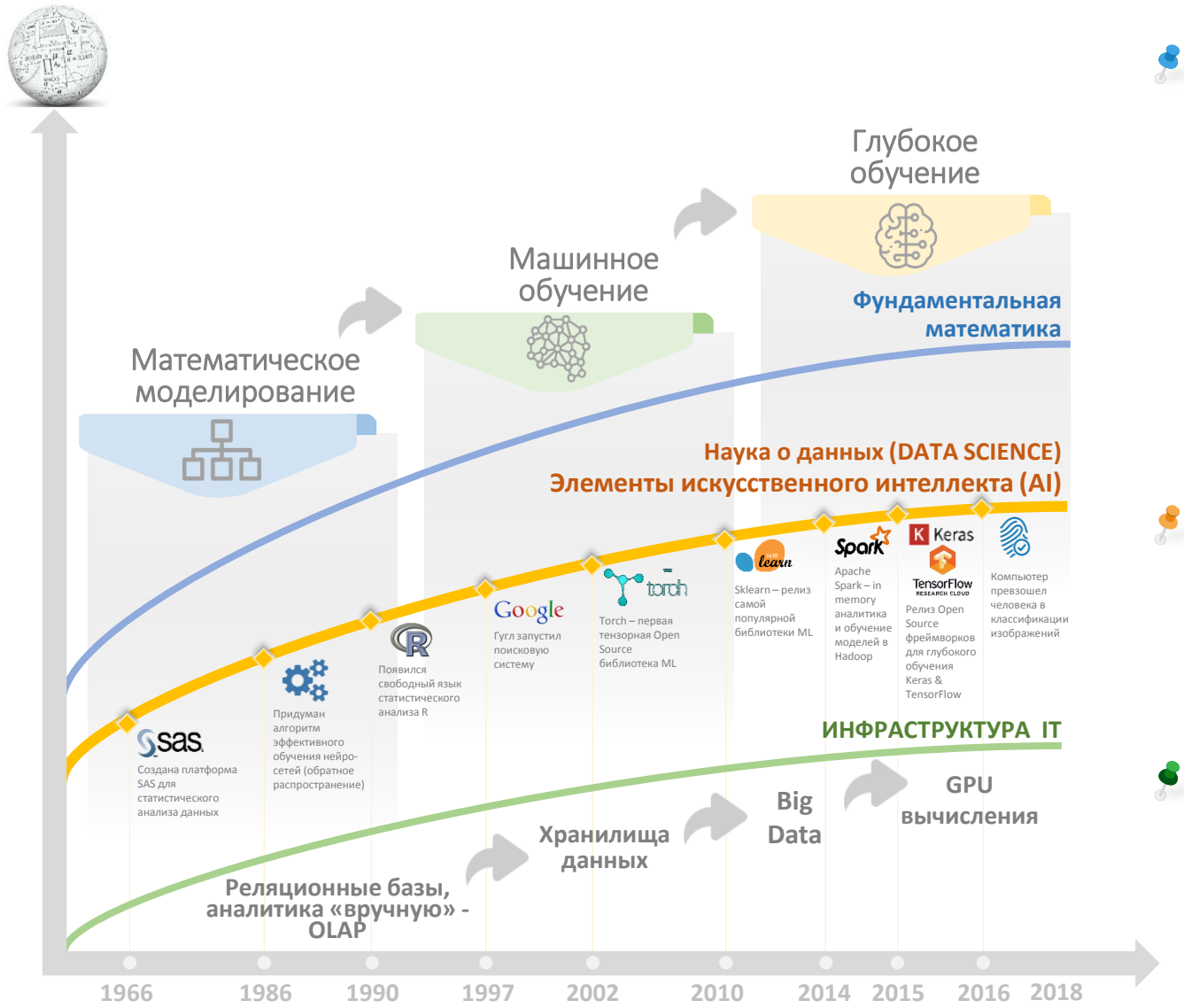
ГАЗПРОМБАНК

ДАННЫЕ КАК ОСНОВА ЦИФРОВОЙ ТРАНСФОРМАЦИИ

ОПРЕДЕЛЕНИЯ

ГАЗПРОМБАНК



Математическое моделирование

Основной инструмент - регрессии, с помощью которых из набора факторов выделяется ограниченный круг сигналов, которые с высокой достоверностью оказывают значимое влияние на целевую переменную. Полученная модель трактуется как математическое представление реальности, т.е. некий идеальный объект, для которого выявлены идеальные «физические» закономерности

Машинное обучение

Класс методов, характерной чертой которых является не прямое решение задачи, а обучение в процессе применения решений большого множества сходных задач.
В отличие от математического моделирования, подход основан не на выявлении узкого круга высокодостоверных закономерностей, а на парадигме того, что множество статистически «слабых» сигналов могут быть объединены в одну статистически «сильную» модель

Глубокое обучение

В отличие от машинного обучения, не предполагается ручного формирования факторов (сигналов), алгоритмы применяются к максимально широкому спектру первичных данных. За счет способности многослойных нейронных сетей к самостоятельному выявлению сигналов из данных при обучении, потеря информации из первоисточника минимизируются. В результате, существенно растет предсказательная способность в сложных задачах с большими данными, в том числе, позволяя превосходить способности человека

Наука о данных (Data Science)

Наука о данных, изучающая проблемы анализа, обработки и представления данных в цифровой форме. Объединяет достижения в области математики с инфраструктурными возможностями ИТ по работе с Большими данными, а также с алгоритмами эффективного распараллеленного хранения и обработки данных в разнородных форматах

Элементы искусственного интеллекта (AI)

Аналогично Data Science, AI основан на синергии математического аппарата и ИТ инфраструктуры данных и вычислений. Принципиальное отлучение от Data Science – применение механизмов Глубокого обучения к Большим данным путем использования специализированных программных и аппаратных решений, таких как Hadoop и GPU вычисления

Хранилище данных

Единый центр сбора и хранения строго структурированных данных в реляционной форме

Большие данные (Big Data)

Структурированные и неструктурированные данные огромных объёмов и значительного многообразия, эффективная обработка которых может быть произведена только путем масштабного распараллеливания вычислений. Решения, обеспечивающие универсальные интерфейсы и программные решения для расспараллеливания вычислений над BigData, называются инструментами работы с Большими данными (например Hadoop)

Эволюция аналитики и организации

ГАЗПРОМБАНК

1.0 Традиционная аналитика

- Преимущественно описательная аналитика и отчетность
- Данные из внутренних источников, относительно небольшие, структурированные
- Разрозненные группы аналитиков
- Аналитика вспомогательный второстепенный инструмент

Аналитика выполняет вспомогательную и необязательную функцию для принятия решений

CDO

2.0 Большие данные

- Сложные, большие, неструктурированные источники данных
- Новые аналитические и вычислительные возможности
- Появление «исследователей данных»
- Продукты и услуги, основанные на данных, - источник прибыли

Создание продуктов и услуг, основанных на данных

CDO + CDS

3.0 Data-driven организация (AI First)

- Целостное сочетание традиционной аналитики и больших данных
- Аналитика как неотъемлемый компонент ведения бизнеса
- Быстрое и гибкое обеспечение решения
- Аналитические инструменты доступны в точке принятия решений
- Аналитика интегрирована в операционные процессы

Аналитика встроена во все операционные и бизнес-процессы

Данные нужны для решения бизнес-задач

ГАЗПРОМБАНК

Сегодня CDO становится ключевым элементом управления компанией, поскольку значимость данных признана в качестве основы цифровой трансформации

Новая информация со старыми алгоритмами дает результаты лучше, чем старая информация с новыми алгоритмами

Данные – важнейший актив

Рост вычислительных мощностей, улучшенная инфраструктура данных, появление новых алгоритмов

Данные без технологий бесполезны

Интеллектуальный анализ данных (Data Mining) включает изучение больших объемов данных и поиска в них закономерностей с применением статистических методов, технологий управления базами данных

Знания из данных дают модели и аналитика

Что принесет нам цифровая информация

ГАЗПРОМБАНК

ДАННЫЕ – ОСНОВА ЦИФРОВОЙ ТРАНСФОРМАЦИИ

Новые модели бизнеса

Новые бизнес-процессы

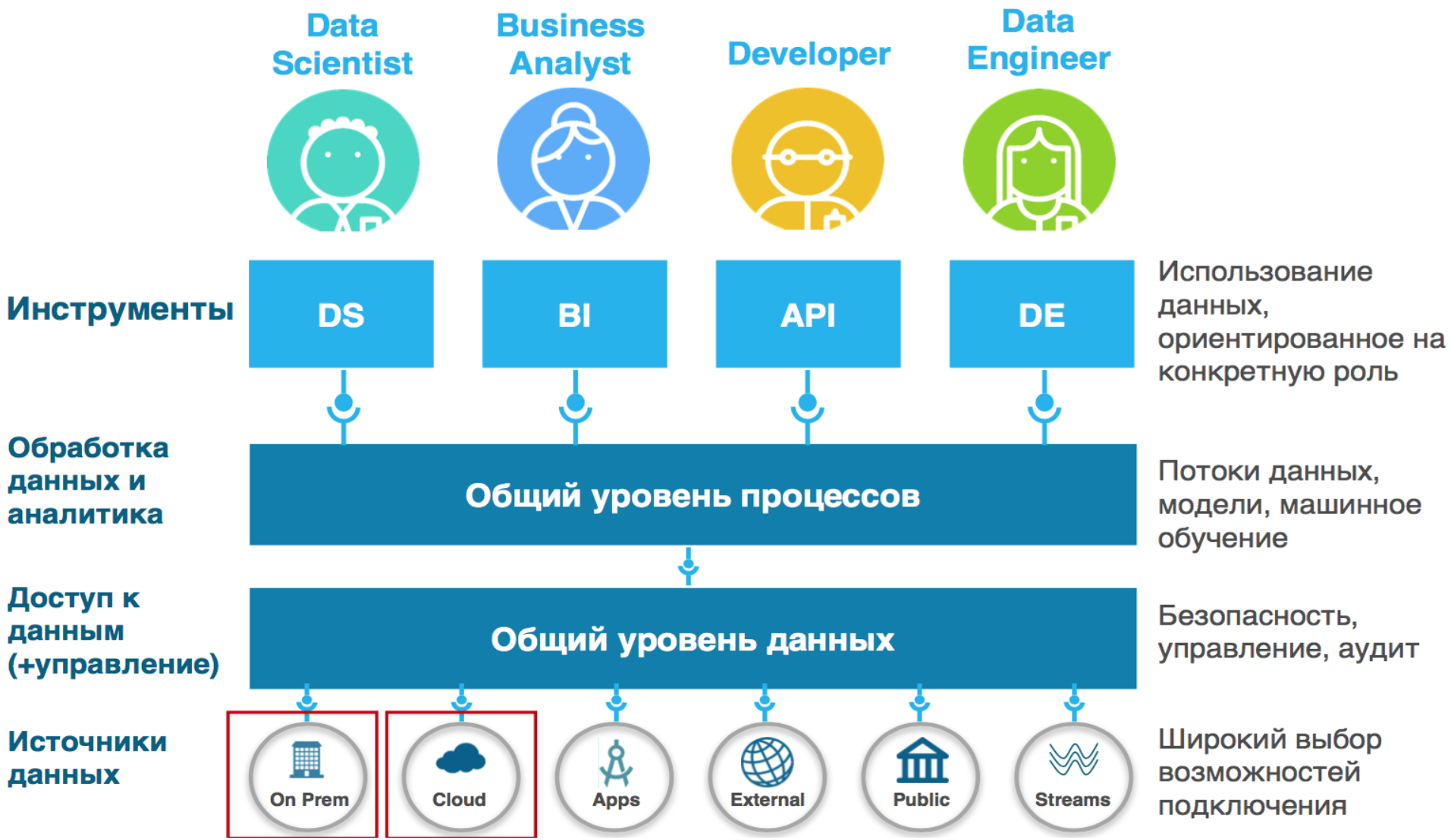
Новые бизнес-роли

Новые приложения

Новая оргструктура

Ключевые роли

ГАЗПРОМБАНК



Технологии работы с данными

ГАЗПРОМБАНК

Принять решения

Машинное обучение и Искусственный интеллект
Machine Learning (ML) and Artificial Intelligence (AI)

Технологии и Инструменты Анализа и Визуализации Данных
Business Intelligence (BI)

Обработать

Сохранить

Реляционные системы управления базами данных (RDBMS)	Massive Parallel Processing (MPP)	Распределенные системы хранения и обработки данных любых форматов	Системы распределенной обработки данных в оперативной памяти	Специализированные системы управления базами данных

Проверить

Технологии и Инструменты Data Governance

Загрузить

Технологии и Инструменты Интеграции и Трансформации Данных
Extract Transform Load (ETL)

Задачи технологий работы с данными

ГАЗПРОМБАНК

Инновации

CDO*

Загрузить
- Получить извне
- Передать между системами
- Собрать в одном месте
- Проверка качества

Сохранить
- Сохранить, предоставить доступ
- Быстро находить нужное

CDS**

Обработать
- Объединить в одной структуре
- Рассчитать показатели
- Обучить модель

Принять решение
- Отчеты, дэшборды, визуализация
- Предписание в операционном процессе

Применение инноваций позволяет предложить варианты развития с учетом новых технологий

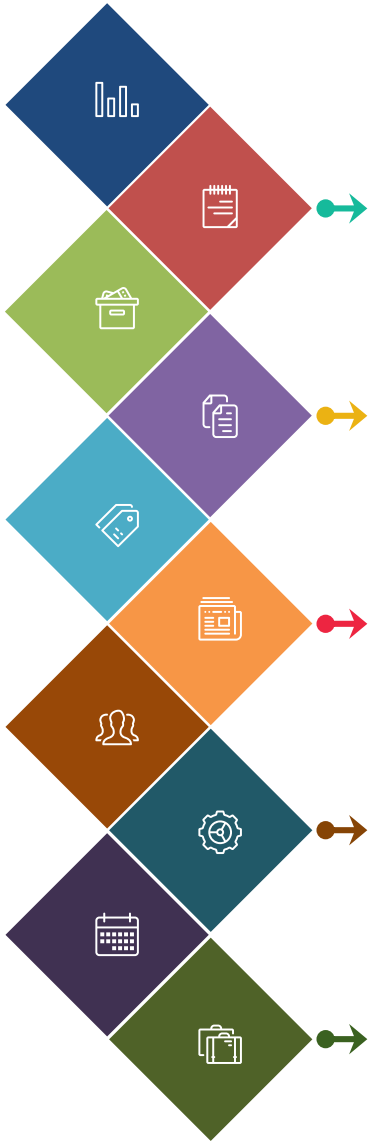
* CDO – Chief Data Officer
** CDS – Chief Data Scientist

Описание целевого функционала основных ролей

Роли в процессах	Цель в процессе (основная функция роли в процессе)
Бизнес заказчик	Извлечение прибыли с помощью моделей на основе данных
CDO Банка	Обеспечение Банка требуемым ассортиментом и качеством внутренних и внешних данных
Инженер данных	Разработка инструментов трансформации и поставки данных
CDS Банка	Обеспечение Банка аналитическими моделями
Исследователь данных Банка	Разработка сложных аналитических моделей Модели как сервис

НЕОБХОДИМЫЕ УСЛОВИЯ

ГАЗПРОМБАНК



- ✓ Комбинация данных, технологий и машинного обучения открывают принципиально новые возможности, но требуют:
  - изменения бизнес-культуры:
    - data driven менеджмент
    - кросс-функциональные команды и сотрудничество
  - включения в компанию специалистов нового типа
  - приобретения и освоения новых средств IT и создания целостной высокотехнологичной платформы организации
- ✓ Начало работ в этом направлении – стратегическое, а не техническое решение
- ✓ Повышение роли искусственного интеллекта должно сопровождаться повышением степени контроля соответствующих модельных рисков

ФУНКЦИОНАЛЬНЫЙ DATA DRIVEN

ГАЗПРОМБАНК

Основной тренд последних двух лет – это переход от общей парадигмы **«Data Driven организации»** (организации, которая в своей деятельности использует культуру изучения данных с помощью Data Science для получения финансового результата) к парадигме функциональных или предметных областей деятельности в Data Driven:

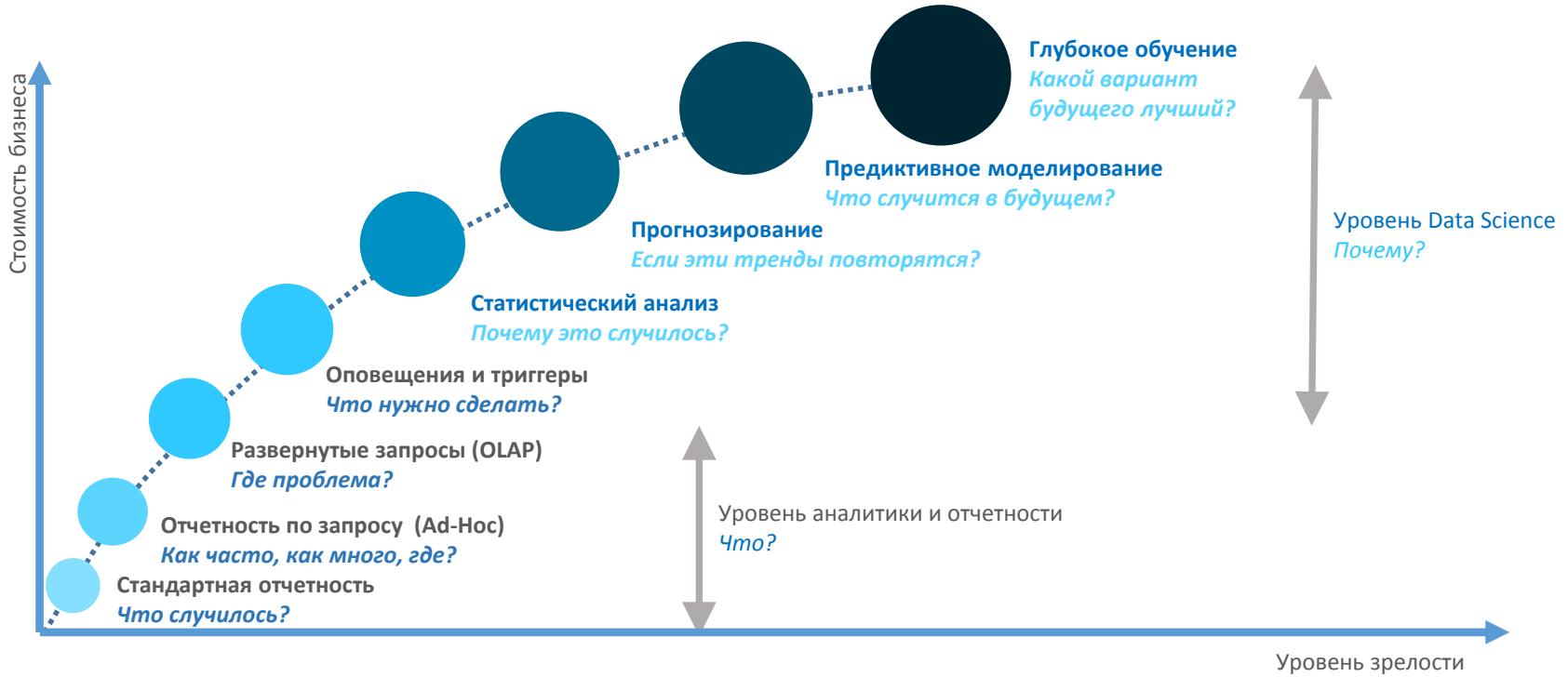
Data Driven Banking – разработка и внедрение банковских продуктов на основании изучения всей совокупности данных о клиентах.

Data Driven Marketing - это построение маркетинговой стратегии на основе анализа полученных данных.

Data Driven Design – это проектирование продукта на основе данных: исследований, тестов, проверки гипотез, машинного обучения, Big Data.

Data Driven Journalism (DDJ) – использование открытых источников данных и обработка их математическими моделями для поиска закономерностей и выявления событий, интересных для новых публикаций.

8 УРОВНЕЙ ЗРЕЛОСТИ ОРГАНИЗАЦИИ НА ПУТИ К DATA DRIVEN



ФИЛОСОФИЯ MODELOPS И MAAS

ГАЗПРОМБАНК

В центре внимания находится все бизнес-процессы по созданию ценности

Обратная связь участникам в виде регулярного мониторинга качества моделей

Корпоративная культура использования моделей в бизнес-процессах

MLOps непрерывно связан с **DevOps** и предполагает внедрение цифровых моделей совместно с IT-инфраструктурой

Финальная цель идеологии **MLOps** – создание инфраструктуры на принципе **MaaS (Model as a Service)** – предоставление сервисной функции по желанию заказчика для встраивания модели в любую точку бизнес-процесса

Цифровая трансформация – быстро и просто не будет

ГАЗПРОМБАНК

Процесс цифровой трансформации требует грамотного подхода:

✓ Оптимизация процессов, создание новой организационной культуры и гибких ИТ-решений.

75% опрошенных Tadviser банков ТОП30 считают, что задачи цифровизации должны быть аккумулированы в рамках единой платформы.

- ✓ **Процесс цифровой трансформации должен развиваться постепенно.**
- ✓ **Вера в технологии – позитивный тренд, но без критичного подхода она может привести к негативным последствиям.**
- ✓ **Сами по себе технологии – не панацея, необходим взвешенный подход, прежде всего – анализ данных**

ЦИФРОВАЯ ТРАНСФОРМАЦИЯ =

Примеры

Data Driven: начало

Самым известным историческим событием в области Data Driven принято считать изобретение печатного станка Гутенбергом в 1440 году – стоимость производства книги вручную на тот момент составляла 20 тыс долларов (в современном эквиваленте). Книги выступали предметом роскоши для благородных семей. Изобретение печатного станка позволило снизить стоимость изготовления книги в 300 раз, до 70 долл за книгу (в современном эквиваленте). За несколько лет печатные станки появились по всей Европе. Производство книг как источника данных в течение первых 100 лет после изобретения печатного станка выросло в 30 раз. Как следствие это бурное развитие финансовых, математических и астрономических наук в 16-17 веках, начало в 1770-х годах промышленной революции, в результате которой цивилизация перешла от состояния практически полного отсутствия научного прогресса к привычным нам быстрым переменам.

Атомная станция – одно из самых ранних технологичных Data driven решений

Атомные станции исторически являются Data Driven организацией. Процессы на АЭС являются сложнейшими объектами автоматизации в мире. Атомная станция – сверхопасное технологическое предприятие, на котором ошибка может привести к фатальным последствиям. Поэтому на атомных станциях в системах безопасности используется многократное дублирование и максимально исключается воздействие человеческого фактора на производственные процессы. Атомную станцию можно рассматривать как один из самых ранних примеров объектов, на которых была проведена цифровизация – 50-е годы. Системы управления и защиты действовали по строгим алгоритмам, которые исключали случайные действия оператора. Об уровне защищенности атомных станций от ошибок и непредвиденных ситуаций можно судить по тому, что за все время произошло всего две аварии, повлекшие серьезные последствия.

2007 Банк Нью-Йорк Меллон Корпорейшн интегрировал чат-боты

Финансовый институт с 233 историей активно занимается внедрением программных ботов и роботов. В мае 2017 года компания выпустила более 220 ботов для обработки запросов, которые часто повторяются, и обычно обрабатываются персоналом. Боты помогали отвечать на запросы данных от внешних аудиторов и при переводе средств, помогали исправить ошибки форматирования и данных в запросах на переводы денежных средств в долларах. Реализация RPA привела к следующим результатам:

* 100-процентная точность валидаций закрытия счетов в пяти ИТ-системах
* 88-процентное уменьшение времени обработки запросов
* ¼ -х секундное роботизированное принятие решения по сделке против 5-10 минут человеком

2002-2010 Amazon – применение цифровизации

Становление монополиста торговли в инете: стоимость акции на конец 2002 год - $12.25, а на конец 2010 - уже $134.52, нв сегодняшний день - $1513.00. В начале 2000 года компания пережила кризис доткомов.

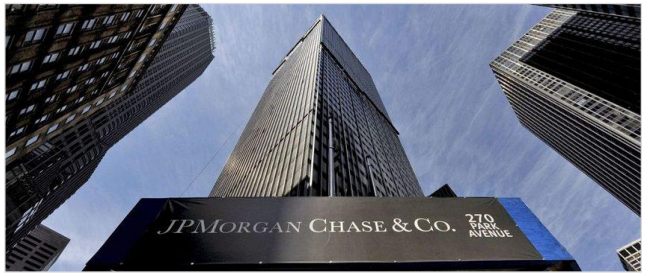
Правильная выбранная стратегия цифровизации, в том числе, таргетированных рекомендаций, представление большей информации для клиентов, автоматизация доступа клиентов к услугам, соглашениям с поставщиками, применение передовых технологий в сортировки и логистики, доставки клиенту мелкорозничных товаров сделали компанию мировым лидером.

Примеры

ГАЗПРОМБАНК

2010 Booking.com. Агрегатор данных по гостиницам

В 2002 году сеть имела убытков $19млн, к 2011 году прибыль составила $1.1 млрд. Покупка этой компании PriceLine Group в 2006 году считается самым лучшим приобретением на рынке цифровых путешествий. Бизнес идея, которая привела к успеху, заключалась в предоставлении клиентам ранее недоступной информации о рынке через цифровые каналы, что привело к географическому перераспределению спроса: росту гостиничного бизнеса там, где стоимость пребывания была недооценена, и снижению цен в районах традиционного пребывания путешественников.



2016 Банк JPMorgan Chase – применение элементов AI

Банк заложил в бюджет 10,8 млрд. долл. США для инвестиций в технологии, причем в передовые технологии - более 5 млрд. долл. Эти инвестиции позволили реализовать несколько инициатив AI, что, в частности, позволило обрабатывать автоматизировано более 12 000 кредитных соглашений в год. Ручной анализ и обработка данных документов раньше занимал 360 000 человеко-часов, теперь выполняется за считанные секунды

2016 Bank of America

Эрика, официально представленная на конференции 2016 Money 20/20 в Лас-Вегасе, - крупнейшее в мире инновационное событие в области платежных и финансовых услуг. Эрика - чат-бот, использующий предиктивную аналитику и модели, основанные на глубоком обучении для предоставления финансовых рекомендаций. Кроме стандартных функций, которыми владеют роботы – отвечать на вопросы клиентов – Эрика способна предложить различные методы улучшения финансовой деятельности благодаря встроенным способностям прогнозирующего анализа - давать советы относительно повышения кредитоспособности и рекомендовать клиентам повышать ежемесячные платежи по непогашенным остаткам на кредитке по снижению процентных отчислений.

Модели и данные как источник прибыли в игровом бизнесе

Caesars Entertainment – один из лидеров игорного бизнеса США, управляет известным казино Caesar’s Palace в Лас Вегасе и еще более 50 игорными заведениями по всему миру. За 17 лет компания накопила впечатляющий объем данных и аналитики по своей программе лояльности Total Rewards, который теперь является одним из самых ценных ее активов и был оценен в 1 млрд. $. Компания использует Hadoop и облачные решения, обрабатывая более 3 млн. записей в час. Анализ данных также используется для сегментации клиентов и повышения стандартов безопасности.

Результат: создан наиболее ценный индивидуальный актив компании стоимостью более 1 млрд. $, достигнут рост прибыльности и стандартов безопасности.

Примеры

1997 Потерянный рынок

В середине 1990-х годов, еще до появления Google, самой продвинутой поисковой машиной была даже не Yahoo, не AltaVista, не Lycos или Hot Wired. Система Open Text, как и Google сегодня, работала максимально быстро, точно, охватывая весь объем информации. В 1995 году менеджеры компании Open Text небезосновательно утверждали, что их система смогла проиндексировать каждое слово из 5 миллионов документов, которые на тот момент и составляли Всемирную сеть. Однако в 1997 году разработчики Open Text посчитали рынок поиска недостаточно перспективным и занялись системами управления корпоративными данными. Ну а через год появилась компания Google, которая показала, что они ошибались на счет перспектив этого рынка. Ситуация могла сложиться иначе, а сам Google мог бы не занимать сейчас лидирующую позицию.

1998 NASA потеряли зонд на Марсе из-за математической ошибки

За несколько лет NASA потратили около 125 миллионов долларов на проект Mars Climate Orbiter. В стоимость вошли научные исследования и разработки, а также запуск зонда в космос.
Зонт был запущен к Марсу в 1998 году и первоначально зонд предназначался для изучения климата на Марсе, а также для сообщений о любых изменениях в атмосфере или на поверхности планеты. Контакт с зондом был потерян вследствие ошибки в расчетах незадолго до того, как он должен был начать свою миссию. В то время как большинство команд, работающих над проектом, использовали стандартные метрические единицы измерения, одна из них использовала дюймы, футы и ярды.
Это и спровоцировало ошибку в передаче координат между командами. В результате Mars Climate Orbiter двигался слишком низко в атмосфере и разрушился при подходе к Марсу.

2008 Фиаско терминала в Хитроу из-за сбоя программы

Незадолго до открытия пятого терминала в аэропорту Хитроу персонал тестировал новейшую систему распознавания для транспортировки больших объемов багажа, поступающего в аэропорт ежедневно. Перед открытием терминала она была тщательно протестирована на 12 000 пробных «единицах» багажа. Все испытания прошли безупречно, но в день открытия терминала оказалась, что система неработоспособна. Вероятно, причиной тому стали непредусмотренные практические ситуации. Например, пассажир мог забыть в сумке какой-то важный предмет, и багаж вручную забирали из транспортной системы. Весь процесс обработки нарушался, и система отказывала. В течение следующих десяти дней около 42 000 мест багажа не были доставлены владельцам, из-за этого пришлось отменить более 500 рейсов.

Примеры

2008 Обвал рынка из-за ошибочного прогноза

Одной из ключевых причин кризиса 2008 года стали слишком оптимистичные оценки обеспечения производных ценных бумаг (и производных бумаг на производные) моделями риск-менеджмента.
В результате, полные потери в мире от кризиса составили $22 трлн согласно оценке Government Accountability Office, что сопоставимо с ВВП США, но многие потери остались неизмеренными или были неизмеримыми.
Потери рынка ценных бумаг $9.1 млрд, последующие расходы на совершенствование законодательства $1.1 млрд

2012 Биржевой робот за 45 минут привел к убыткам в 440 млн долларов

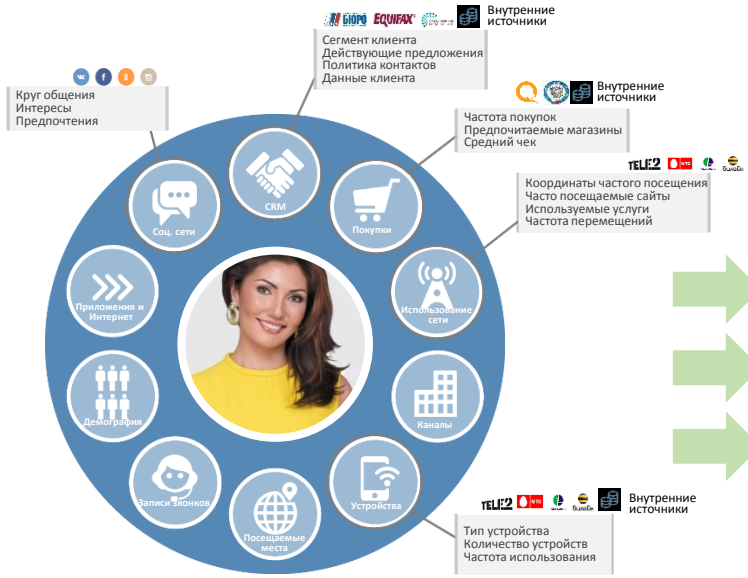
На разного рода биржах работают программные комплексы, «роботы», которые позволяют автоматизировать процесс купли/продажи.
Иногда автоматизация может привести к беде. Так, биржевой «робот» компании Knight Capital Group начал бесконтрольно скупать акции разного рода компаний, включая RadioShack, Ford Motor Company и American Airlines. Торговая вакханалия автомата продолжалась всего 45 минут. После этого компания постаралась как можно быстрее сбыть нежелательные пакеты акций. Подсчитав затраты на покупку акций, а затем выручку от продажи, компания недосчиталась 440 миллионов долларов. Причина таких «удачных» торгов, по мнению, представителей компании — новое программное обеспечение, которое было недавно установлено.
Это в четыре раза больше прибыли, полученной компанией в 2011 году. Убытки настолько велики, что компания испытывает серьезные трудности с дальнейшим ведением бизнеса Руководство компании верило, что автоматизация поможет Knight Capital Group превзойти всех своих конкурентов.
Случившееся может привести к административному ограничению использования биржевых роботов на торгах.

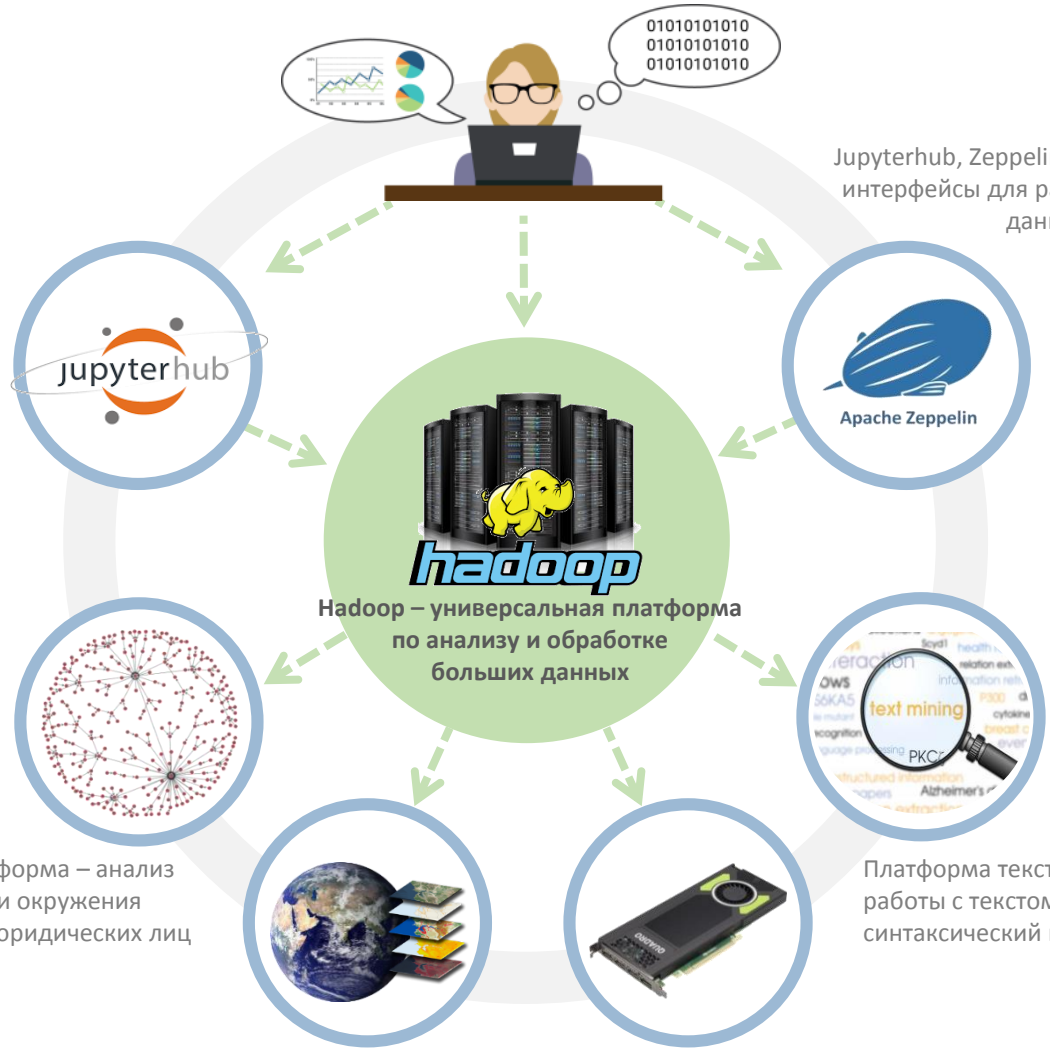
2018 Ошибка моделирования процесса глобального потепления

Сотрудники Института Океанографии Скриппса в октябре 2018 года пришли к выводу, что океаны планеты поглощают на 60 процентов больше тепла, чем считалось на протяжении последних 25 лет. Это означает, что опасность глобального потепления гораздо серьезнее, чем предполагалось ранее. Однако независимый ученый Ник Льюис обнаружил «серьезные (хотя, разумеется, и неумышленные) ошибки в расчетах, лежащих в основе исследования». Для исследования использовалась модель предсказания уровня нагревания океана на основе объемов двуокиси углерода и кислорода в атмосфере. Исследователи признали ошибку, и таким образом было признано, что поглощение тепла океаном соответствует уровню середины 90-х годов, и гипотеза глобального потепления из-за изменений в атмосфере признана не такой критичной, как это было принято в последние годы.

Консолидация данных и аналитических инструментов

ГАЗПРОМБАНК





Jupyterhub, Zeppelin – универсальные интерфейсы для работы аналитиков данных

- Чат-боты
- Event-based Маркетинг
- Выявление мошенничества
- Оптимизация процессов
- Умные советы
- Кредитный скоринг

Графовая платформа – анализ взаимосвязей и окружения физических и юридических лиц

Гео-платформа – универсальный инструмент анализа гео-тегированных данных

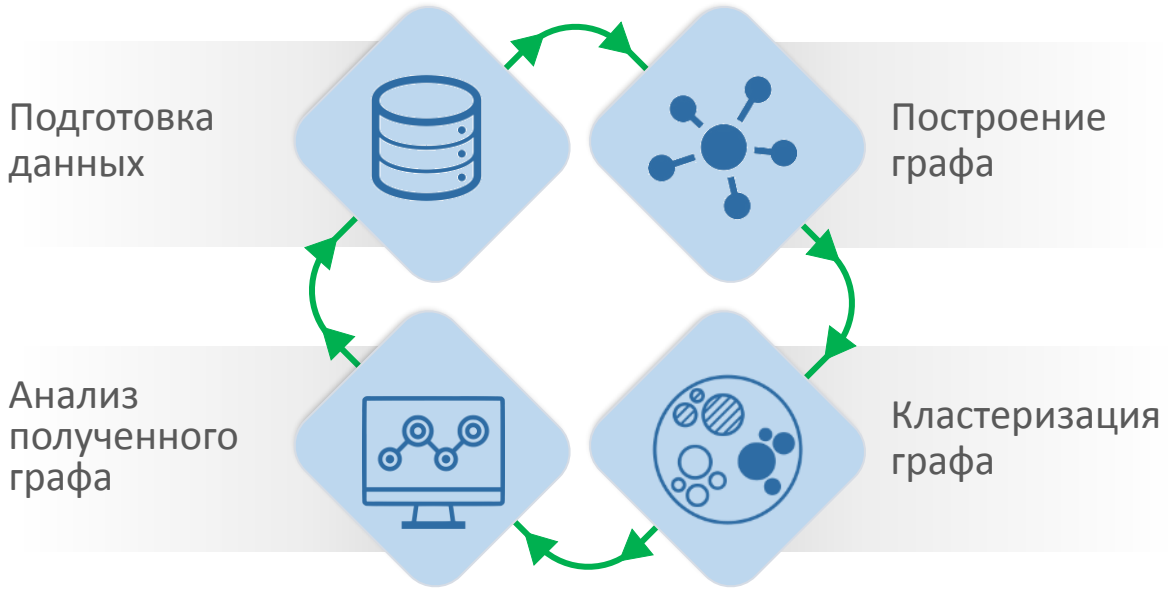
Платформа глубокого обучения – специализированное решение для обучения многослойных нейронных сетей с применением GPU вычислений

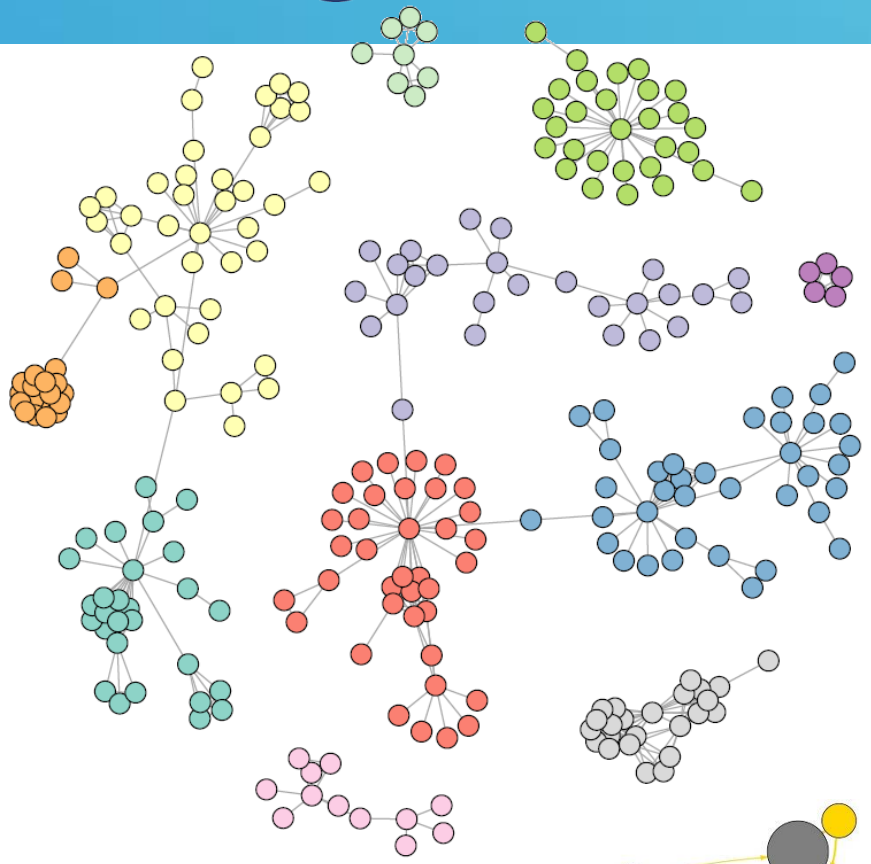
Платформа текстовой аналитики – NLP методы работы с текстом (распознавание, синтаксический и семантический анализ и т.д.)

Примеры ГПБ: ГРАФЫ связей юридических лиц

ГАЗПРОМБАНК

Работа с графами ведется в Python и R с использованием OpenSource решений

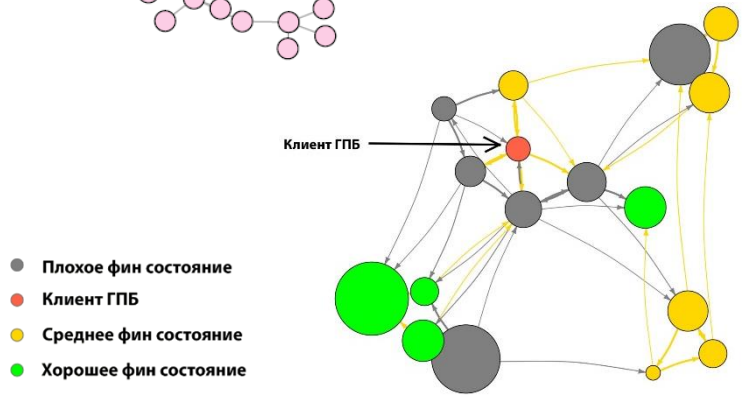




По результатам оценки PD моделью без учета графовой аналитики: компания попадает в «серую зону».

Модель PD с учетом агрегатов, рассчитанных на графе, повышает вероятность дефолта до порогового значения: компания попадает в «черную зону».

Скоринговая модель с учетом метрик, построенных на графе, показывает повышение коэффициента Gini на 11 единиц.



Примеры ГПБ: АВТОМАТИЗАЦИЯ ПРОЦЕССОВ – чат БОТ

ГАЗПРОМБАНК

Парсинг открытых банковских форумов-горячих линий и статей о банковских услугах

~~ох,~~ ~~уж~~ ~~мне~~ ~~эти~~ чат-боты, ~~зачем~~ ~~вообще~~ ~~они~~ ~~в~~ банке?

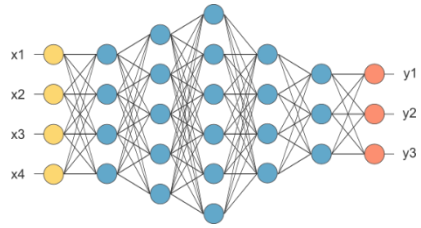
Удаление стоп-слов

чат-боты => чат-бот банке => банк

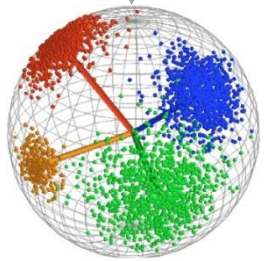
Stemming & Lemmatization

чат-бот => [2 0 … 7 5] банк => [3 17 … 55 1]
чат-бот + банк = [2 0 … 7 5] + [3 17 … 55 1]

Генерация признаков на основе Word2Vec представления слоев



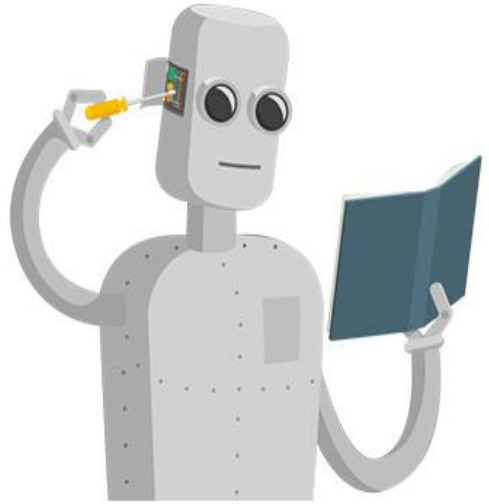
Размеченные поисковые запросы по банковской тематике из открытых источников



Классификация в векторном пространстве

API Мессенджера

Создание Чат-Бота



Примеры ГПБ: ГЕО-платформа

ГАЗПРОМБАНК

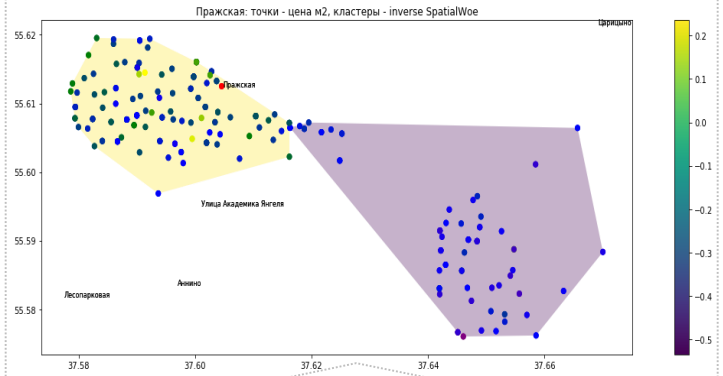
Моделирование с использованием Геоплатформы

Задача: интегрирование гео-данных в «стандартный» процесс анализа данных и применения моделей

- Анализ свойств гео-точек: адресов объектов недвижимости, мест жительства и работы физических лиц, адресов регистрации юридических лиц, точек продаж

Разработан и внедрен принципиально новый инструмент гео-пространственной кластеризации

- Определение оптимального разбиения территории на полигоны, свойства которых имеют высокую предсказательную способность относительно целевой переменной
- Автоматический контроль баланса между недо- и пере-обученностью модели разбиения
- Эффективен для моделирования областей, в отношении которых Банк обладает ограниченным объемом данных и экспертизы



Пример ГПБ: ВЗГЛЯД НА КЛИЕНТА 360°

ГАЗПРОМБАНК

Глубокая интеграция позволит прогнозировать потребности и повысит скорость реакции на изменения

1 Постоянный клиент с большой историей

Стабильный и высокий доход
Низкая кредитная нагрузка
Большая семья
Тратит много времени на дорогу

Бизнес

- Страхование жизни/здоровья
- Подписка на акции застройщиков
- Ипотека на выгодных условиях

2 Перспективный клиент

Стабильная карьера
Низкая кредитная нагрузка
Посещает фитнес и любит путешествия
Тратит много времени на поиск товаров

Оптимум

- Карта фитнесс-клуба
- Карта лояльности авиалиний
- Подписка на распродажи
- Кредит на свадебное путешествие

3 Необязательный клиент

Часто меняет работу
Средняя кредитная нагрузка
Посещает футбол и ходит в бары
Регулярная краткосрочная задолженность

Лайт

- Карта болельщика
- Подписка на расписание матчей
- Chat-напоминания о платежах
- Создание шаблонов для платежей
- Оптимизация кредитной нагрузки

Объединение технологий идентификации и online обработки данных позволяет прогнозировать потребности и мгновенно реагировать на любые события в жизни клиента.

Пример ГПБ: ВЗГЛЯД НА КЛИЕНТА 360°

ГАЗПРОМБАНК

Требуется доработка законодательной и технологической базы для удаленной авторизации новых клиентов

Не существует механизма получения гос. данных о физ. лице по результатам удаленной аутентификации

Ряд ключевых гос. данных не доступен для автоматизированной обработки даже при наличии согласия физ. лица

Биометрическая аутентификация

Согласие на обработку данных

Транзакции

Банк

Геолокация

Окружение

Соц. сети

МТС
МегаФон
Билайн
TELE2

Сотовые операторы

Оптимальное для клиента, персонифицированное предложение

Отсутствие монополии на данные: данные доступны тем, кому доверяет клиент

Централизованный API к социальным данным

Централизованный API к данным по транзакционной информации клиентов

Централизованный API к данным по гео-локации и связи