СБЕРБАНК
Всегда рядом

Использование массивов данных в банковском бизнесе

Ноябрь 2015 г.

Ценность данных для Банка

Необходимым условием получения ценности Банком от аналитики массивов данных является реализация остальных элементов Бизнес модели: продуктов, удовлетворяющих потребностям клиентов, компетенций необходимых для реализации и использования аналитических сервисов, построение организационной структуры и системы дистрибуции продуктов, заключение партнерств

Доп. доход | Снижение рисков Банка | Сокращение затрат

Модели создания ценности для бизнеса Банка[1]

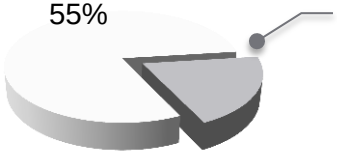
		Описание	Внутренние клиенты
1	**Прибыль от продажи продуктов Банка и партнеров клиентам Банка**	▪ Повышение таргетированности продаж продуктов Банка и партнеров через каналы прямой коммуникации (SMS, e-mail, телемаркетинг, СБОЛ, клиентские менеджеры и др.) за счет дополнительных данных и моделей	РБ, КБ, Спасибо
2	**Удержание и удовлетворенность клиентов**	▪ Выявление клиентов, склонных прекратить отношения с Банком и индивидуальный подбор методов их удержание ▪ Совершенствование продуктов Банка для повышения удовлетворенности клиентов	РБ, КБ
3	**Снижение потерь Банка от мошенничества и социальных дефолтов**	▪ Выявление мошенников и предотвращение мошеннических транзакций за счет дополнительных данных и выявления паттернов поведения мошенников ▪ Предотвращение потерь от невозвращенных кредитов за счет дополнительных данных для анализа платежеспособности заемщиков	ДБ, Блок Риски
4	**Возврат потерь от кредитных дефолтов**	▪ Проведение расследований случаев мошенничеств ▪ Повышение эффективности сбора просроченной задолженности с физических лиц за счет дополнительных контактных данных и таргетированного подбора методов воздействия ▪ Работа с проблемными активами, выявление аффилированных лиц	УРПА, Блок Риски
5	**Сокращение операционных расходов**	▪ Повышение операционной эффективности Банка[2] – Оптимизация подбора сотрудников – Повышение эффективности документооборота – Снижение стоимости хранения и обработки данных – Отчетность	HR, ДУД

1 Инициативы по монетизации используют ту или иную модель создания ценности
2 Данный список не исчерпывающий и по мере проработки новых инициатив будет дополняться

Примеры. HR

СБЕРБАНК
Всегда рядом

Проблема



Порядка 28 тысяч СОЧЛ, МПП, КБП покидают Банк в течении 1 года[1]

12 тысяч (45%) ежегодно нанимаемых СОЧЛ, МПП, КБП увольняются из Банка проработав менее 12 месяцев

0.9 млрд. руб. Ежегодно теряет Сбербанк от ухода сотрудников ВСП профессий СОЧЛ, МПП, КБП, проработавших менее 12 месяцев

Гипотеза

А что если можно будет прогнозировать склонность кандидата к раннему увольнению на этапе подбора?

Склонны к увольнению из Банка ранее 12 месяцев работы

Не склонны к увольнению из Банка ранее 12 месяцев работы

Детали пилота

- Рассмотрены данные 5256 сотрудников Среднерусского ТБ профессий СОЧЛ, МПП, КБП[2]
- Из них 358 уволились ранее 12 месяцев
- По каждому сотруднику рассмотрено более 50 параметров
- Построена модель выявления кандидатов, склонных к увольнению

Результаты

Применяя данные из резюме кандидата на этапе подбора можно выявить до 64% кандидатов склонных к раннему увольнению из Банка. Сокращение потерь[3] Банка при этом составит более 400 млн. руб. в год

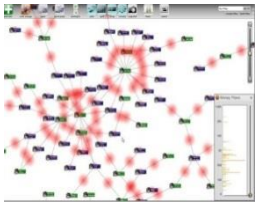
Потенциал

+400 млн. руб. в год

1 Данные взяты на основе цифр за первые 10 месяцев 2014 года (Л. Свириденко предоставил 23306 за первые 10 месяцев)
2 Данные брались за последние 3 года, по тем сотрудникам, по которым они были доступны без технических ошибок
3 Под потерями понимаются как недополученная прибыль в результате ухода сотрудника, так и затраты на обучение и поиск сотрудника

Примеры. УРПА (работа с проблемными активами)

Предпосылки



- Сбор информации по проблемному активу занимает **1 день** и происходит в ручном режиме
- Задача – Разработать инструмент, аналог ПО Palantir, позволяющий ускорить проведение расследований по проблемным активам
- УРПА и ОТИ провели пилот на базе решения IQMen, подтвердивший целесообразность автоматизации

Реализация

Цель

Сократить время на сбор и анализ информации по активу
Улучшить качество анализа

Данные

•Система-источник МДМ РБ (объем 1,1ТБ)
•База Integrum (1ТБ)
•Другие внешние источники…..

Команда

Сотрудники УРПА

Заказчик, эксперт по бизнес-части

Сотрудник дирекции Супермассивов данных

Руководитель проекта

Технические специалисты

Аналитики данных, разработчики, дизайнер интерфейса, архитекторы данных, архитекторы решений

Инструменты

Hadoop, Hive, Spark, Sqoop, Solr, Datomic

Алгоритм

Алгоритмы мэтчинга, обогащения данных, графовая модель связей, полнотекстовый поиск

Результаты

- Создан прототип инструмента для проведения расследований, объединяющий разнородные источники данных и определяющий связи объектов
- Инструмент может масштабироваться для обработки любого объема данных
- Возможности инструмента:
  - оперативный поиск сведений о должниках
  - выявление и визуализация связей между физическими и юридическими лицами

Ожидаемый эффект: сокращение времени расследования с 1-го дня до 30 минут

СБЕРБАНК
Всегда рядом

Определения

Определение супермассивов данных

СБЕРБАНК
Всегда рядом

Супермассивы данных – это методы аналитики больших объемов данных, позволяющие извлечь дополнительную ценность для Банка и характеризующиеся 3 компонентами

1

- **Данные**: **огромные** объемы внутренних и внешних данных **различной** структуры

Таблицы, Текст, Изображение, Голос, Видео

2

- **Аналитика**: выявление скрытых зависимостей и поиск **новых** вопросов и ответов на основе анализа **всего** объема **разнородных** данных

Данные определяют новые вопросы и ответы

3

- **Технологии**: **распределенное** хранение и обработка данных, **самообучающиеся** алгоритмы машинного обучения, принятие решений в режиме **реального времени**

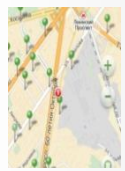
Hadoop, Машинное обучение, Real time,..

Все это в комплексе создает новые огромные возможности для работы с клиентами, повышения качества услуг, сокращения затрат, повышения доходов и создания принципиально новых продуктов и сервисов

Супермассивы данных – это огромные массивы внутренних и внешних данных различной структуры

СБЕРБАНК
Всегда рядом

Данные – информация представленная в цифровом виде. Супермассивы данных характеризуются 3 основными параметрами: Объем, Разнообразие, Скорость (3 V's: Volume, Variety, Velocity). Также, отдельно выделяют категорию "Геолокация"

		"Традиционный подход"	"Супермассивы данных"
Объем (volume)	За прошлый год в мире было создано более **4 млрд** терабайт. Это больше, чем за последние **5 000** лет. При этом к 2020 г. прогнозируется рост совокупного мирового объема данных до **44 млрд** терабайт	**Гигабайты и терабайты** *"Транзакции за один день для выборки клиентов"*	**Сотни терабайт и петабайты** *"Транзакции за целый год для всей клиентской базы"*
Разнообразие (variety)	Развитие интернета, социальных сетей, мобильных устройств и M2M привело к появлению **разнообразия форм**, в которых могут быть представлены данные	**Структурированные данные** *"Классические таблицы"*	**Данные любой структуры** *"Таблицы, текст, изображение, видео, голос"*
Скорость (velocity)	Все новые и новые данные создаются с **огромной скоростью**. Помимо этого данные постоянно меняются. Это **непрерывный процесс**	*"Выгрузка исторических данных за определённый период"*	*"Получение и анализ информации в потоковом режиме"*
Геолокация	**Новое измерение** данных, определяющее **где** совершается событие, **где** находятся объекты. Широкое распространение геолокация получила с ростом количества мобильных телефонов, GPS навигаторов и т.д.		*"Информация о местоположении клиента, объекта с точностью до метра"*

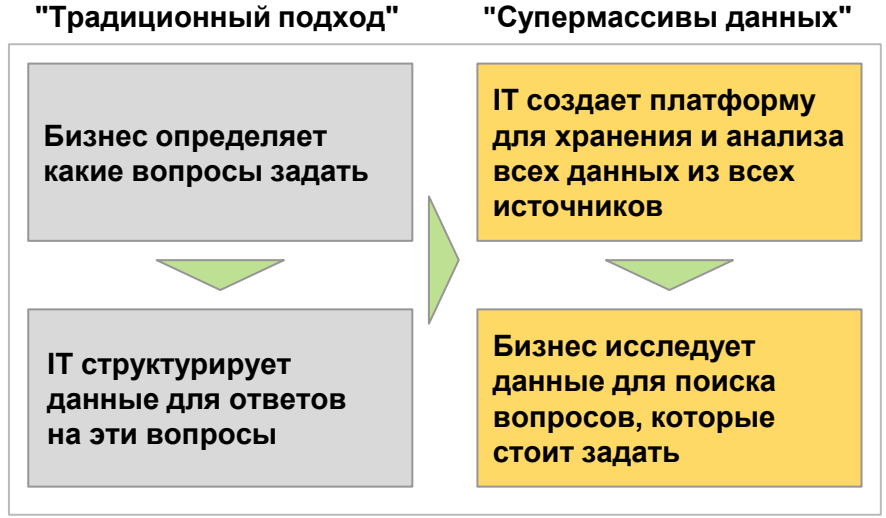
Супермассивы данных – это новые подходы к аналитике: выявление скрытых зависимостей на основе анализа всего объема разнородных данных

СБЕРБАНК
Всегда рядом

Аналитика – совокупность методов математики и статистики, позволяющих извлекать ценность
из данных, определяя взаимосвязи, закономерности и рекомендуя конкретные решения

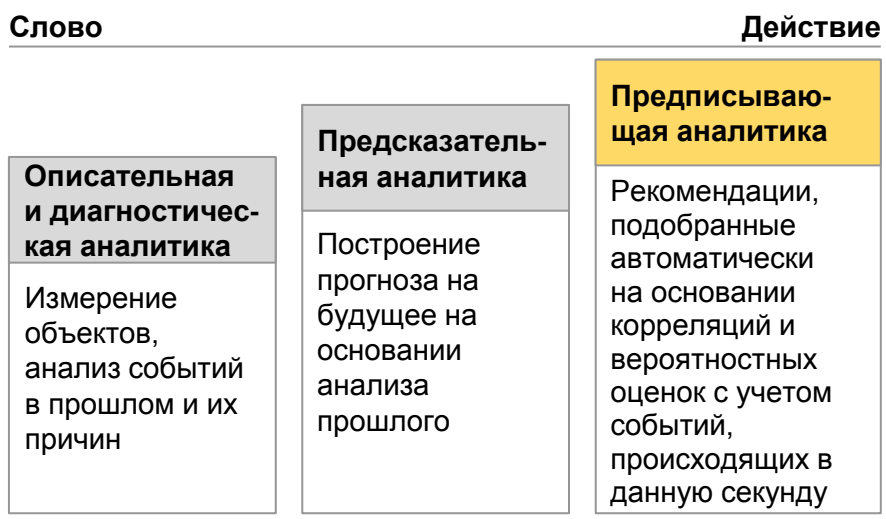
Новая модель проведения анализа

Традиционно недостаток данных и ограничения по возможностям анализа диктовали необходимость **заранее определить гипотезу и проверить ее** на **сильно ограниченном объеме** данных (**выборке**). Сегодня **огромные объемы разнообразных** данных и **мощные** аналитические инструменты позволяют данным "**говорить за себя**" и находить новые вопросы и ответы. Более того, априорное постулирование **причинно-следственных** связей утратило свою значимость и уступило место максимально широкому поиску **зависимостей и закономерностей,** которые могут быть самыми неожиданными



От слов к действиям

Сегодня уже **недостаточно просто описывать** то, что происходило в прошлом, и строить прогнозы на будущее. Настало время **аналитики действий** – рекомендаций что делать, подобранных с помощью **автоматизированных** алгоритмов **машинного обучения**



Супермассивы данных – это новые технологии распределенного хранения данных, машинного обучения, принятия решений в реальном времени

СБЕРБАНК
Всегда рядом

За последние 10 лет появились новые и получили развитие существующие технологии, изменившие мир работы с данными

Технологии хранения и обработки данных

В 2005 г. появились новые **технологии распределенного хранения и обработки** данных **любой** структуры. Ключевая идея – приложение разделяется на большое количество одинаковых элементарных заданий, выполняемых на узлах кластера и естественным образом сводимых в конечный результат. Сейчас решения на базе этих технологий используют все передовые компании, имеющие дело с супермассивами данных (Google, Яндекс, Facebook, ..)

"Традиционный подход"	"Супермассивы данных"
Реляционные базы данных	**Распределенная файловая система (Hadoop)**
▪ Хранение только структурированных таблиц данных ▪ Средняя и высокая стоимость хранения данных	▪ Низкая стоимость хранения данных (в 16 раз ниже) ▪ Возможность хранения данных любой структуры ▪ Легкость масштабирования

Технологии анализа данных

С появлением **Машинного обучения** самообучающие алгоритмы позволили определять скрытые зависимости (корреляции) и закономерности и получать достаточно точные ответы практически **без участия человека**. Также стало возможным **анализировать** данные **любой структуры** и делать это в режиме **реального времени, постоянно** накапливая и **автоматически** включая в анализ все новые и новые данные и повышая **точность** принятия решений

Ручное построение моделей	**Автоматизированное машинное обучение**
Анализ исторических данных	**Анализ в режиме реального времени**
Анализ структурированных таблиц	**Анализ таблиц, текстов, распознавание изображений, видео, голоса**