

Возможности машинного обучения в бизнесе и промышленности

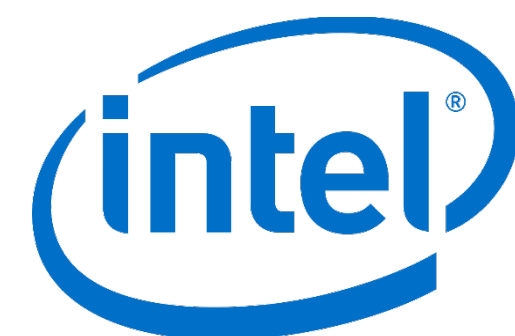
Встреча клуба 4сіо
29 марта 2017

Немного о Yandex Data Factory

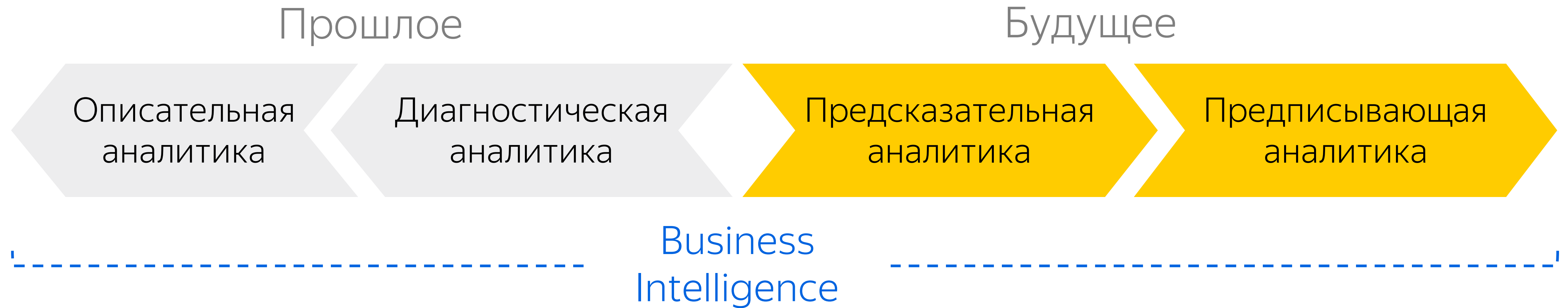


- › Подразделение Яндекса для B2B-рынка
- › Запущено в 2014 году
- › Машинное обучение и технологии работы с большими данными, изначально разработанные для внутреннего использования
- › Международный бизнес (инфраструктура, сотрудники и представительства в России и ЕС)
- › Индивидуальные и типовые решения – машинное обучение на основе анализа данных заказчика
- › Фокусная индустрия – промышленное производство

Среди наших клиентов и партнеров



Фокус нашей экспертизы



Области
экспертизы:

- Поведенческая аналитика
- Временные ряды и аномалии
- Пространственная аналитика
- Распознавание речи, компьютерное зрение

Ожидания и реальность новых технологий

Ожидания

- › Чудеса
- › Очень доходные чудеса
- › Безвредные и даже комфортные чудеса
- › Чудеса, не требующие усилий

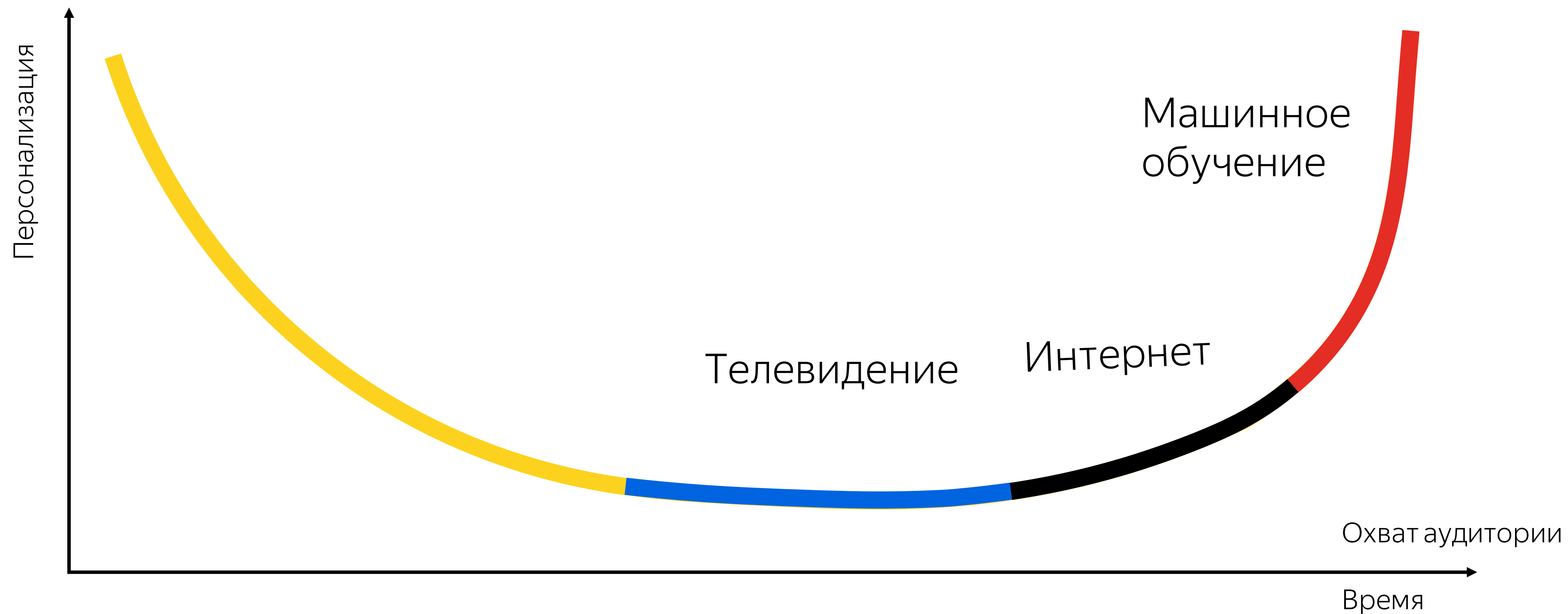
Реальность

- › Что попросил, то и получил (и то не всегда)
- › Нет данных, нет результатов
- › Оптимизация, совершенствование, улучшение
- › Но да, иногда чудеса случаются

Почему сейчас?



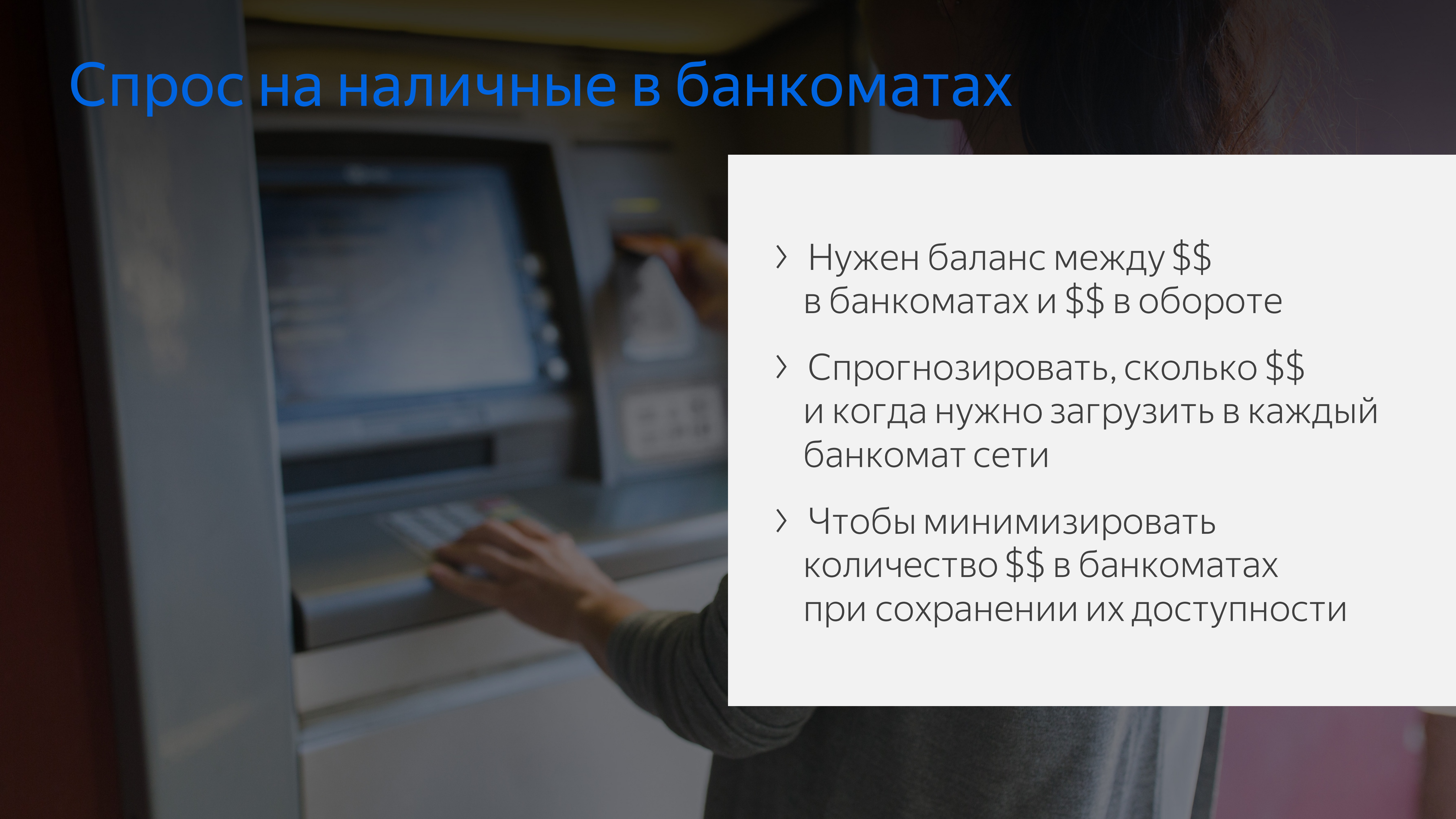
Кривая персонализации



02

Примеры решений

Спрос на наличные в банкоматах

The background of the slide is a blurred photograph of an ATM. A person's hand is visible, inserting a card into the machine's slot. The ATM screen shows some text, but it is out of focus. The overall lighting is dim, and the colors are muted, with a dark blue/purple tint.

- › Нужен баланс между \$\$ в банкоматах и \$\$ в обороте
- › Спрогнозировать, сколько \$\$ и когда нужно загрузить в каждый банкомат сети
- › Чтобы минимизировать количество \$\$ в банкоматах при сохранении их доступности

Прогнозирование спроса

на наличные в банкоматах Райффайзенбанка

Банки и финансы

Использованные данные

Данные 2000 банкоматов:

- › ID, адрес, тип установки
- › Параметры транзакций (сумма, время)

15%

прогнозируемое
снижение издержек
банка

30%

снижение отклонения
прогноза от реального
спроса

Результат

- › Внедрение модели в перспективе позволит банку примерно **на 15% снизить издержки**, связанные пополнением банкоматов и изъятием из оборота лишних средств
- › Модель позволит банку уменьшить отклонение прогноза от реального спроса примерно **на 30%**

Продажи товаров в рамках промоакций

- Точно прогнозировать спрос:
 - › На каждый доступный по промоакции товар
 - › В каждом магазине сети
 - › Заранее
- Чтобы оптимизировать объем поставок в магазины

Прогнозирование продаж

Ритейл

товаров в рамках промоакций в сети «Пятёрочка»

Использованные данные

- › Данные о магазине (ID, регион, филиал и тип магазина)
- › Данные о товаре (ID, категория, величина скидки, стоимость)
- › Статистика продаж за 2 года

Результат

- › В 87% случаев погрешность прогноза составила не более одной упаковки
- › В 61% прогноз оказался абсолютно точным
- › Прогноз успешно построен даже для товаров без истории продаж

87% предсказаний
верны с точностью до
одной упаковки

91 магазин
охват пилота

Оптимизация расхода сырья

- › Сократить расход ферросплавов при производстве стали в кислородно-конвертерном цехе
- › Без потери качества стали

Оптимизация расхода сырья

при производстве стали на ММК

Металлургия

Использованные данные

Исторические данные о более 200 000 плавов за 7 лет, включая:

- › Технические параметры, результаты хим. анализов
- › Требования к хим. составу готовой стали, справочники марок стали и ферросплавов

5%
среднее снижение
расхода ферросплавов

>275 млн руб.
ожидаемая
годовая экономия

Результат

Рекомендательный сервис:

- › Принимает данные по исходному составу, массе и требованиям по содержанию хим. элементов в готовой стали
- › Выдает рекомендации по оптимальному количеству ферросплавов на каждом этапе

Предсказание дефектов листовой стали

- › Предсказывать те слябы, которые приведут к $>10\%$ дефекта в листовой стали
- › Чтобы на раннем этапе производства иметь возможность направить их по другому маршруту обработки и изготовить сталь без дефекта

Предсказание дефектов

Металлургия

листовой стали на ранних этапах производства

Использованные данные

1. Исторические данные о 17 000 ранее изготовленных слабов:
 - › Химический состав
 - › Геометрические параметры
 - › Вес

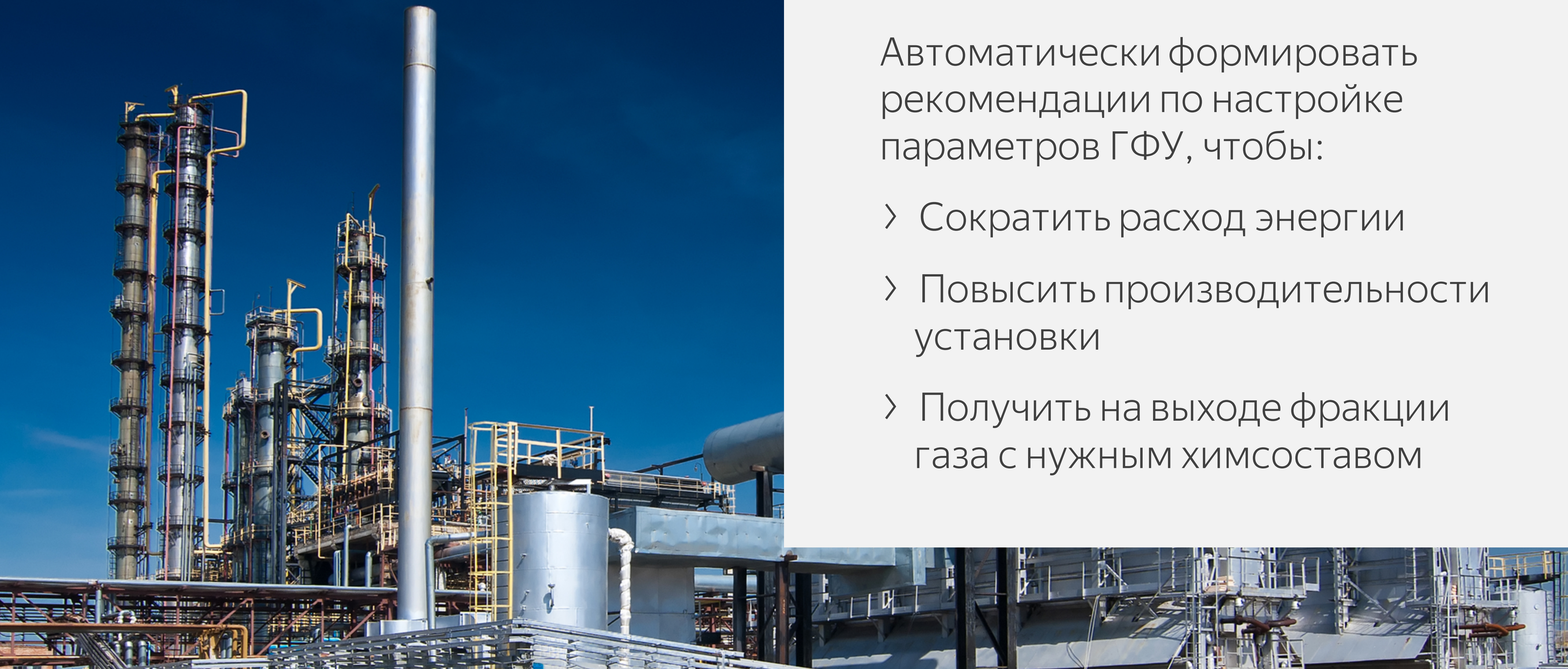
Результат

- › Модель предсказывает слабь, которые после обработки дадут дефект в итоговой стали
- › 48% всех дефективных слабов находятся уже в первых 10% всех слабов, ранжированных моделью

48%

дефектных слабов предсказаны уже в первых 10%

Оптимизация газофракционирования



Автоматически формировать рекомендации по настройке параметров ГФУ, чтобы:

- › Сократить расход энергии
- › Повысить производительности установки
- › Получить на выходе фракции газа с нужным химсоставом

Оптимизация работы

установки газофракционирования

Нефтехим

Использованные данные

- › Параметры входящего потока природного газа (в том числе его химический состав)
- › Внешние параметры (к примеру, температура воздуха)
- › Параметры работы установки
- › Ограничения на химсостав выходящих газов и другие технологические ограничения

Результат

- › Система «подсказчик оператора» для формирования рекомендаций по управлению параметрами установки для поддержания максимальной производительности

Оптимизация процесса цианирования

- › Определять оптимальный объем и концентрацию цианида при извлечении золота из руды
- › Сократить общий расход цианида при сохранении уровня извлечения золота

Оптимизация расхода цианида

Металлургия

при извлечении золота из руды

Использованные данные

Исторические данные о процессе:

- › Хим. состав исходной руды
- › Доступные параметры процесса цианирования (длительность процесса, расход воздуха, характеристики потока пульпы, концентрация раствора цианида и др.)

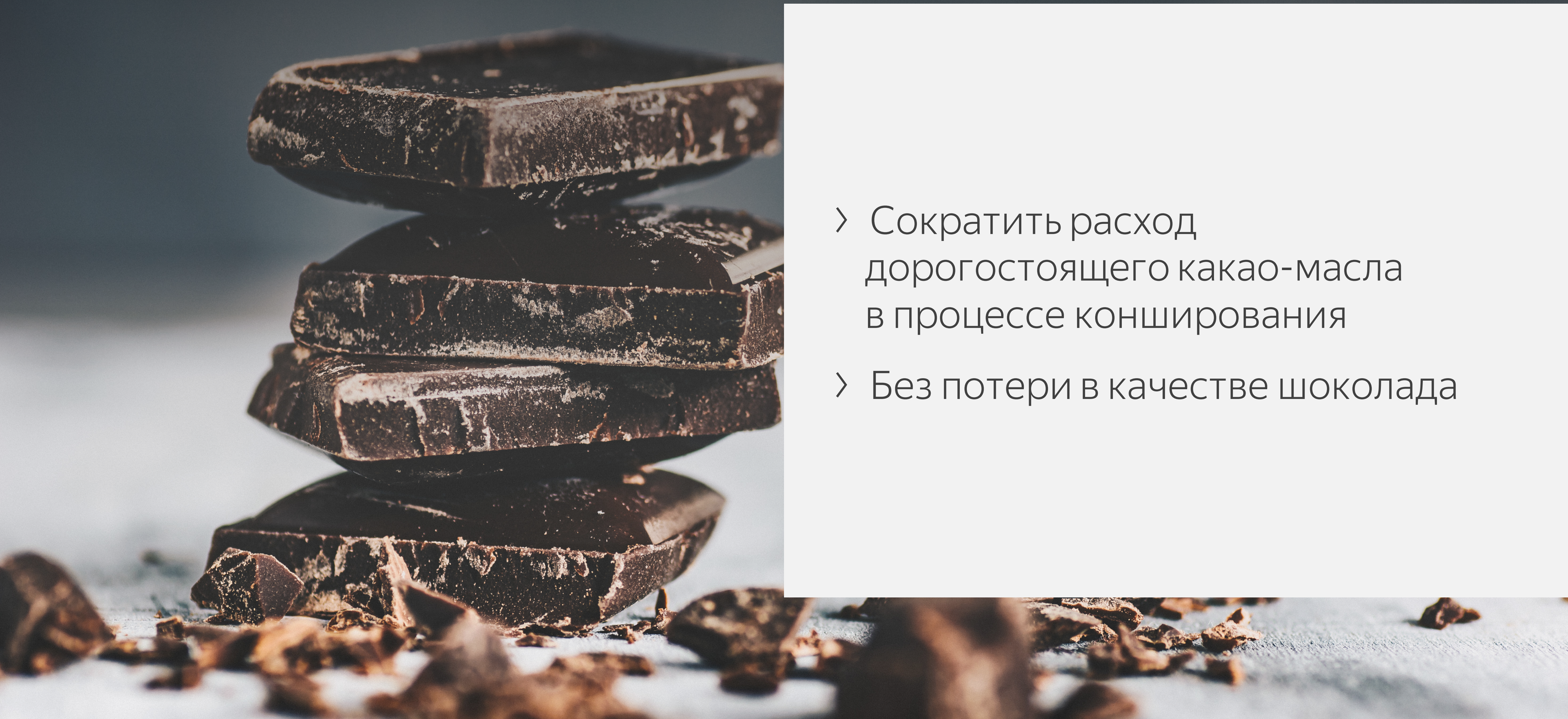
20-40%

себестоимости переработки руды составляют затраты на цианид

Результат

- › Рекомендательная система для определения оптимального объема и концентрации цианида

Оптимизация производства шоколада



- › Сократить расход дорогостоящего какао-масла в процессе конширования
- › Без потери в качестве шоколада

Оптимизация расхода сырья

Пищепром

в процессе конширования шоколада

Используемые данные

- › Состав какао-массы и прочих ингредиентов
- › Параметры процесса конширования (температура, длительность, интенсивность)
- › Требования по содержанию какао-масла в готовом продукте и другие технологические ограничения

Результат

- › Рекомендательная система, оптимизирующая расход какао-масла при соблюдении требований к составу итогового продукта

Рекомендательная система для ритейла

- › Выводить товар с наибольшей вероятностью покупки в блоке рекомендаций на сайте
- › Чтобы повысить конверсионность и выручку

Рекомендательная система

для крупного интернет-магазина электроники

Ритейл

Использованные данные

- › Логи поведения пользователей на сайте (просмотренные товары, содержимое корзины)
- › Ассортимент товаров по 58 регионам присутствия (название, категория, цена, покупки и просмотры, промо)

+1,5%

выручка на сеанс

+3,2%

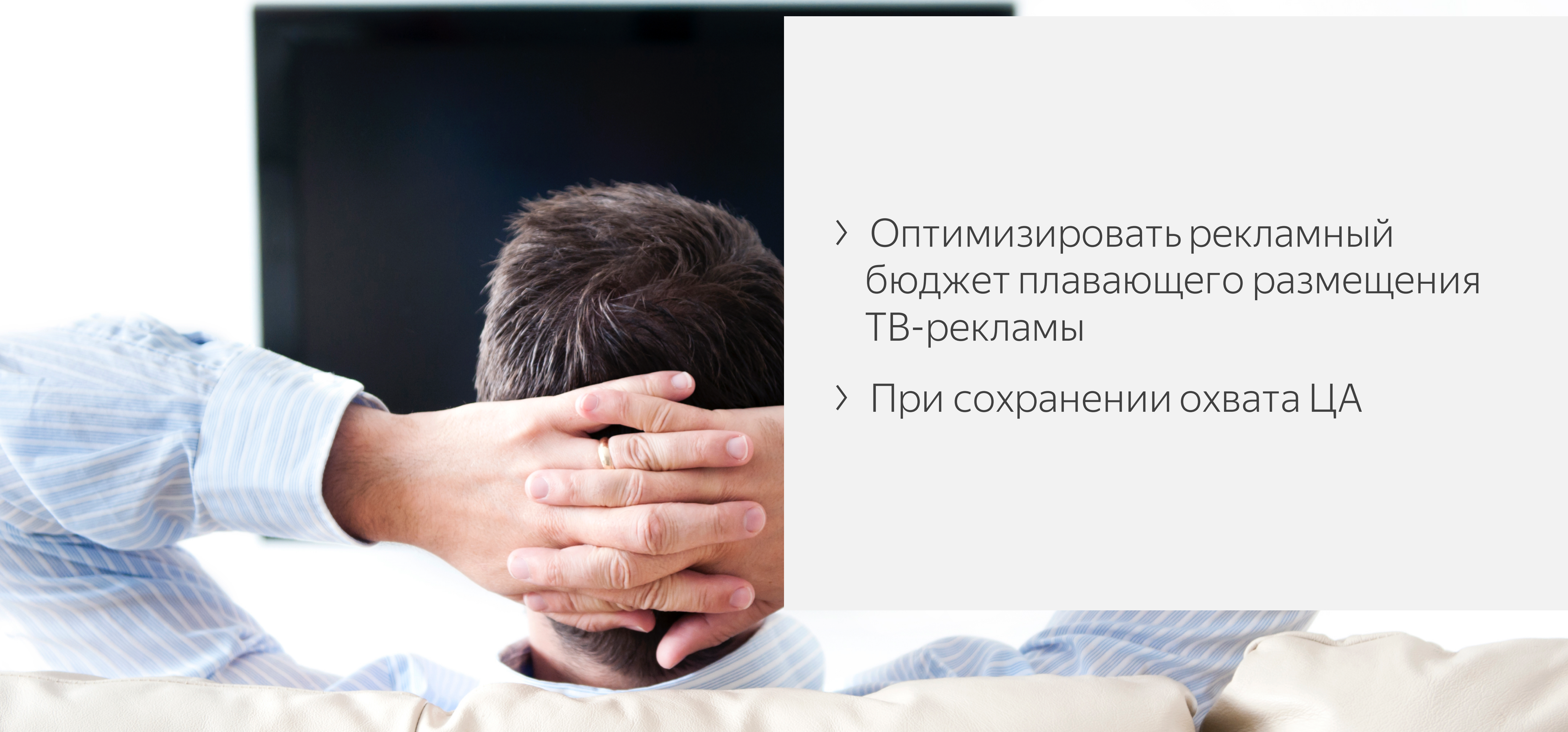
конверсия

Результат

Разработаны две модели:

- › Определяющая вероятность покупки в случае конкретного пользователя
- › Определяющая степень похожести товара/аксессуаров (для кросс-продаж)

Оптимизация рекламного ТВ-бюджета



- › Оптимизировать рекламный бюджет плавающего размещения ТВ-рекламы
- › При сохранении охвата ЦА

Оптимизация рекламного

ТВ-бюджета

Реклама

Задача

- › Спрогнозировать охват РК в группе ЦА по бюджету РК и текущему распределению бюджета по каналам
- › Минимизировать общий бюджет, распределяя его по каналам с плавающим размещением при сохранении охвата

2,9%

оптимизация бюджета
плавающего размещения

2 650 000 руб.

экономия бюджета
по каждой рекламной
кампании (РК) клиента

Использованные данные

- › Время выхода рекламных роликов в эфир
- › Сведения TNS об охвате групп ЦА
- › Сведения о прошедших РК
- › Сетки вещания телекомпаний и стоимость размещения рекламы

Прогнозирование намерения уволиться

- › Предсказать, кто из сотрудников компании собирается уйти в ближайшее время
- › Чтобы сфокусировать усилия на их удержании и сократить затраты на поиск и наем

Прогнозирование увольнений

сотрудников инженерного центра

Промышленность

Использованные данные

- › Статистика по увольнениям для группы из 1255 инженеров за период с 2008 по 2012 годы
- › Информация о сотрудниках (пол, возраст, образование, данные о тренингах, отпусках)

Результат

- › 26 сотрудников из полученного топ-50 действительно уволились
- › Ожидаемая годовая экономия от внедрения решения – до 30 млн руб. (при выделении топ-50 для каждой тысячи позиций инженеров)

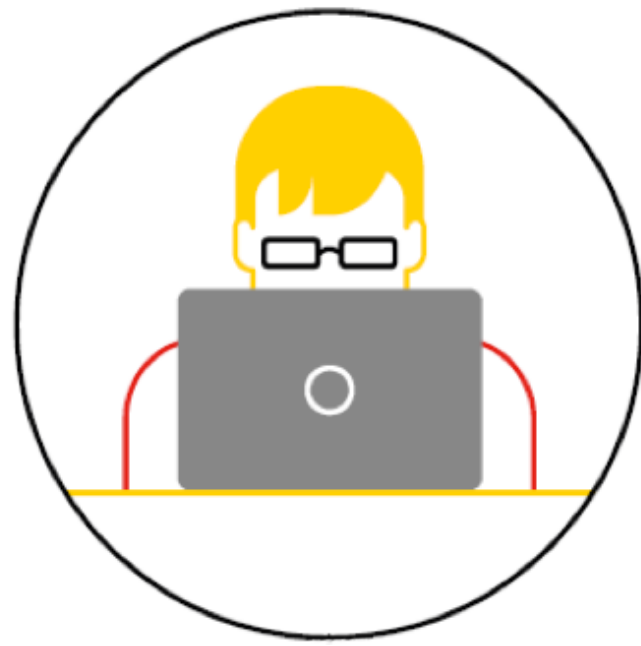
26 из топ-50
действительно
уволились

до 30 млн руб.
ожидаемая
годовая экономия

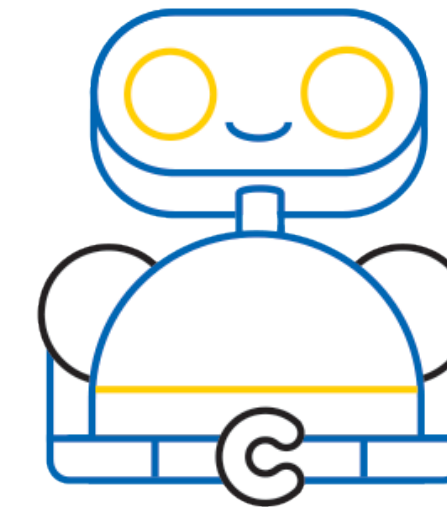
03

Отличия проектов ML от обычных
проектов внедрения нового ПО

Все наоборот

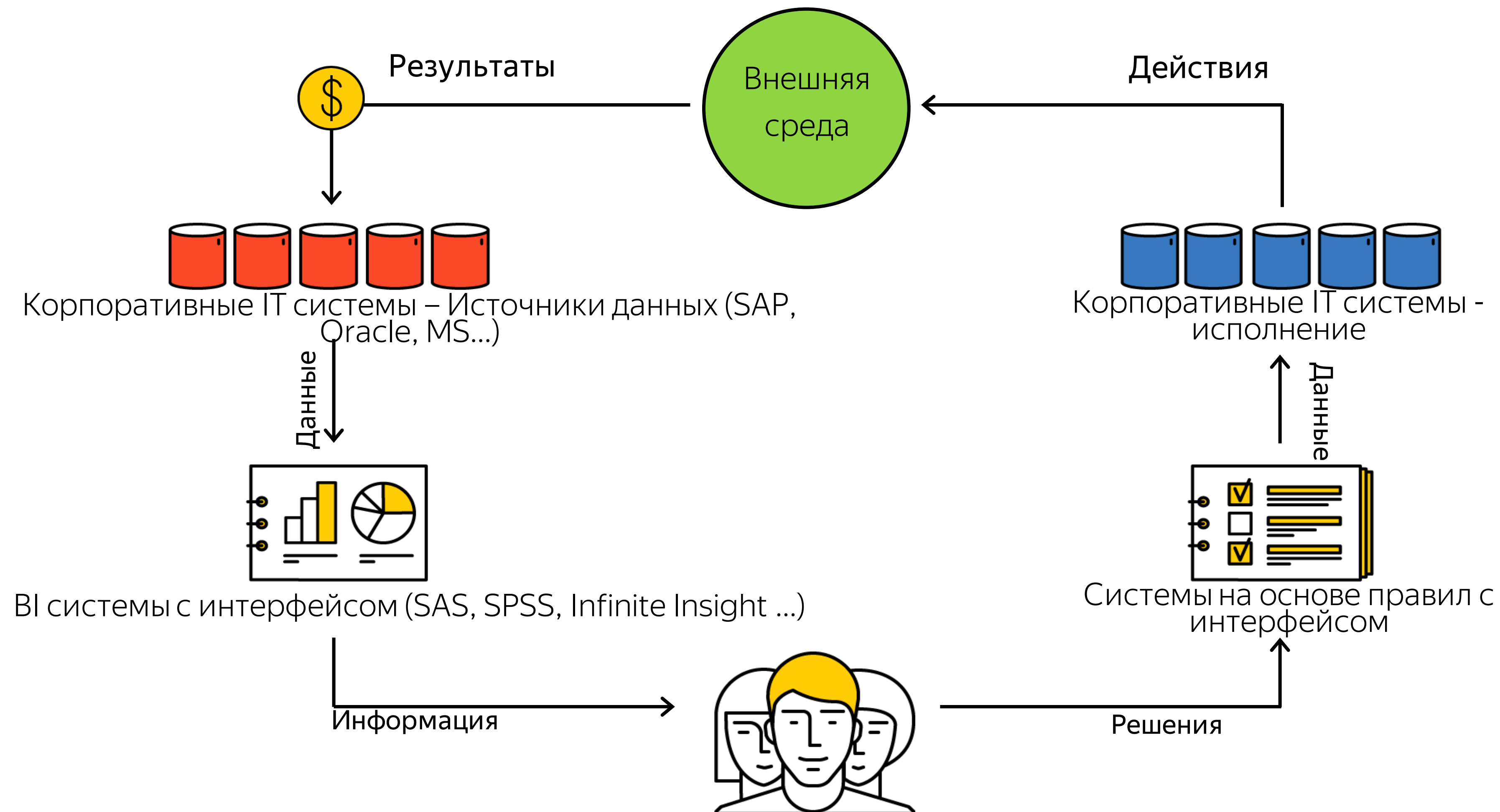


- › Богатый функционал
- › Стремление заменить имеющиеся решения
- › Сложный интерфейс пользователя
- › Сложная архитектура
- › Существенные риски интеграции
- › Внимание на обучение
- › Опытная эксплуатация – сбор мнений и оценок

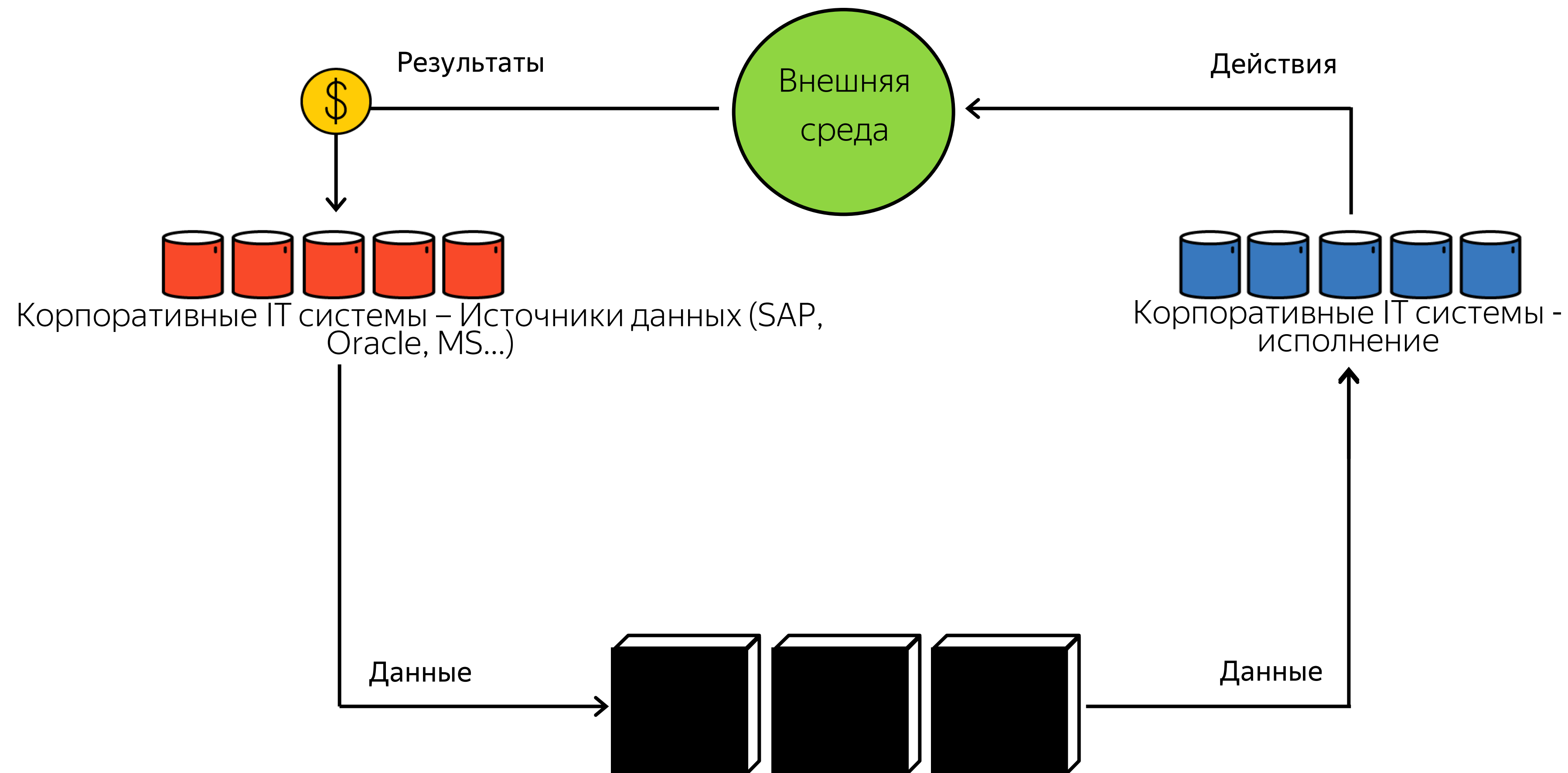


- › Скупой (почти отсутствующий) функционал
- › Встраивание в имеющуюся экосистему, невозможность изолированного внедрения
- › Примитивный или отсутствующий интерфейс
- › Архитектура скрыта внутри решения
- › Тривиальная интеграция
- › Внимание на организационные аспекты
- › Опытная эксплуатация – измерение

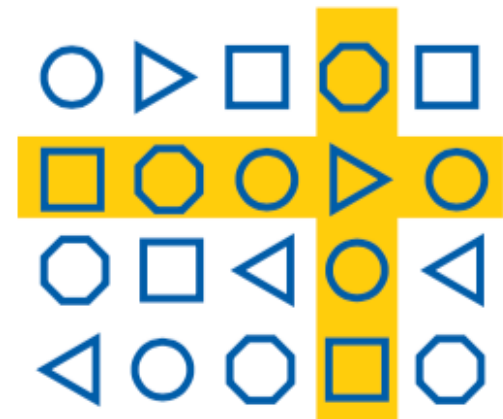
Цикл принятия решений



Цикл принятия решения с ML

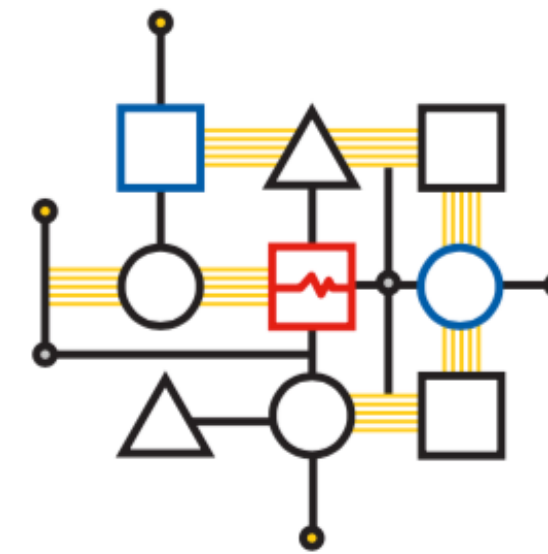


Жизненный цикл решения



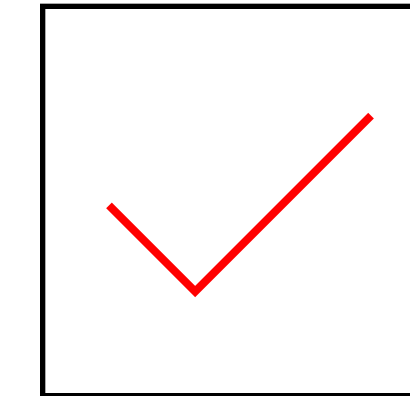
1 Подготовительный этап

- Оценка доступных данных
- Определение требований к работе сервиса, метрик и критериев успеха
- Передача данных (при необходимости – их деперсонализация и обфускация)
- Согласование методики проведения эксперимента



2 Проведение пилота и оценка результатов

- Обучение модели
- Тестирование модели (проведение эксперимента)
- Проверка критериев успеха, оценка эффективности работы модели



3 Промышленная эксплуатация сервиса

- Организация регулярной передачи данных или развертывание модели на мощностях заказчика
- Регулярное A/B тестирование качества работы сервиса
- Поддержка качества работы модели, включая регулярное дообучение

03

Вопросы практического применения

Ключевые вопросы практического применения



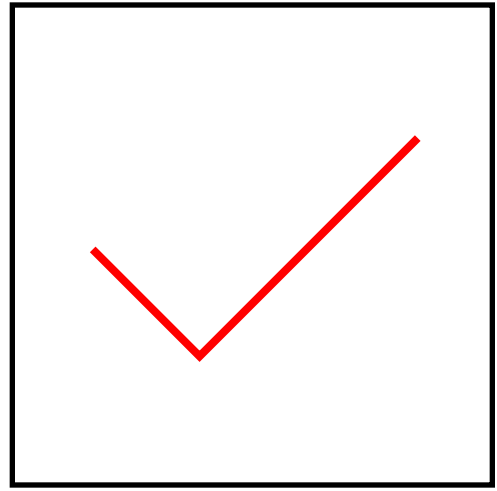
- Постановка задачи: не «узнать и понять», а «предсказать и рекомендовать»
- Выбор метрики
- Экспериментальная проверка, дизайн эксперимента
- Встраивание в бизнес
- Организационные изменения
- Доступ к данным

Основные проблемы



- Баланс между индустриальной экспертизой и Data Science
- Экспериментальный подход в бизнесе
- Проблемы измерений
- Разрушение привычных процессов
- Вопрос ответственности – кто отвечает за рекомендации модели?
- Внутренняя конкуренция
- Внешняя конкуренция

Ответы и пути решения



- Невозможность интерпретировать модель -> измерения, постоянные A/B тесты
- Потеря контроля, зависимость -> конкуренция многих поставщиков
- Потеря экспертизы -> разумная кооперация с внешними поставщиками, разделение инноваций и оптимизации
- Недостаточная гибкость -> «Цифровая демократия» от Walmart

Целевое состояние



- Дружественная к экспериментам архитектура
- Доступность данных
- Набор готовых датасетов для тестирования подрядчиков
- Баланс внешней и внутренней экспертизы



Спасибо за внимание!
А теперь вопросы.

yandexdatafactory.com

[Александр Хайтин](#)

Исполнительный директор
Yandex Data Factory

ahajtin@yandex-team.ru