

Олег Седелев
Корпоративно-инвестиционный блок (КИБ).
ПАО Сбербанк
декабрь 2019

NLP и Auto ML решения на гетерогенном инфраструктурном стеке

teradata.

1. Сбербанк и Корпоративно-инвестиционный блок (КИБ)
2. Данные в Сбербанке
3. Решения ML/DL в КИБ Сбербанка
4. Гетерогенная инфраструктура данных
5. NLP (векторное представление и контекстно-ориентированный подход)
6. Data Science инструменты и Auto ML

ДКК 360 Корпоративно-инвестиционного блока (КИБ) Сбербанка

Наша цель

Внедрение машинного обучения для улучшения клиентского опыта наших корпоративных клиентов.

Наша миссия

Достичь нового уровня жизнеспособности бизнеса, обеспечивая возможность конкуренции с международными финансовыми организациями, становясь лучшим Банком для бизнеса.

В разработке стратегии работы с корпоративными клиентами КИБ уделяет особое внимание

технологиям Искусственного интеллекта

Знаете ли Вы, что...



~ **10** банков

звонят клиенту после регистрации нового бизнеса в тот же день?



3,2 года

в среднем живет бизнес корпоративного клиента?



до **40%**

рабочего времени клиентский менеджер и продуктовые эксперты банка тратят на поиск и обработку информации о клиентах?

Собираем цифровые следы

Традиционные



Профиль



Платежи



Гос. данные



Продукты

Новые



Общение



Clickstream



Экосистема



Новости

Задача

Получать максимальную ценность и выгоду от технологий ML - NER, NLP, AutoML. Сделать это быстро для получения прибыли сейчас



Вопросы

- Перевести инфраструктуру данных на гомогенный кластер Hadoop или нет?
- Как внедрять новые технологии и подходы ML/DL в существующей инфраструктуре?
- Какие самые эффективные шаги трансформации инфраструктуры для бизнес-кейсов КИБ?
- Какие новые ML бизнес-кейсы могут быть разработаны и какой у них потенциал?

Как применяем

Таргетирование продаж

Таргетирование на структурированных и неструктурированных данных

1

Кредитный скоринг и кредитный мониторинг

Выдача кредитов онлайн до 2 млрд. рублей

2

Ценообразование

Индивидуальное ценообразование

3

Клиентский опыт

Аналитика обращений и обратной связи, роботы-ассистенты, комплаенс-процедуры

4

Планирование

Оценка рынка и потенциала клиентов

5

Экосистема

Рыночная аналитика для внешних клиентов

6

Гетерогенная инфраструктура данных



Почему гетерогенная:

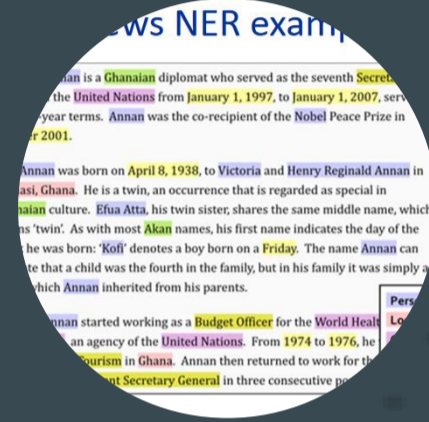
1. Нет необходимости менять всю инфраструктуру: нереализуемо технически с учетом большого количества legacy-систем
2. Можно использовать преимущества различных систем и архитектур: Teradata надежнее и быстрее, в Hadoop – более дешевое хранение данных

ДАННЫЕ



- Транскрибация диалогов
- Определения эмоций на основе аудио
- Emotion detection based on audio

ЗНАНИЯ



- Клиентский контекст
- Auto CSI
- NER: конкуренты (банки), продукты
- spaCy и Tomita parser для NER
- Кластеризация и библиотеки NER

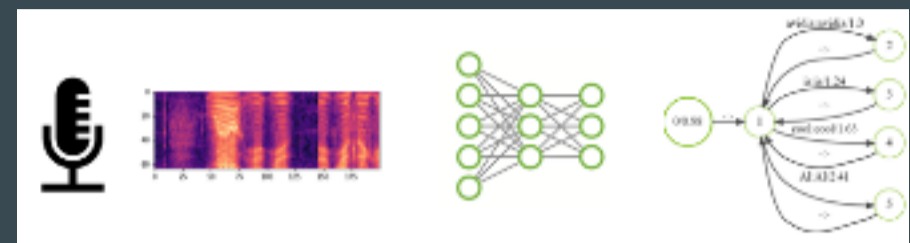
ПРИЛОЖЕНИЯ



- Клиентский опыт: определения и кластеризация причин жалоб
- Анализ эффективности клиентского менеджера (насколько эффективность коррелирует с отклонением скрипта диалога)
- Для cross/up-sales : определение планов клиента: развитие бизнеса, покупка активов, найм сотрудников
- Для cross/up-sales : модель приобретения нового клиента – рост AUC + 8% (embeddings based на AllenNLP)
- Аналитика для владельцев продуктов о ключевых проблемах в продукте
- Ценообразование: идентификация планов перехода в другой банк из-за цены
- Данные для обучения моделей

NLP для таргетирования продаж – рост AUC + 8%

Speech recognition + Named-Entity Recognition



Канал 0: а потом у вас у меня как бы сейчас я вам расскажу у меня
 Канал 1: угу. угу
 Канал 0: значит счёт **центроинвест** потому что там самой маленькие ст
 авки
 Канал 1: угу
 Канал 0: платёжке онлайн я делаю сейчас все там перешли на онлайн
 Канал 1: угу
 Канал 0: **эквайринга**, ваш сберовский и я значит с **эквайринга** с вашег
 о **эквайринга** мне на счёт на **центроинвест** перекидывает я работаю как
 бы вот так вот вот мне предлагали когда я подключал переподключала
 Канал 1: угу
 Канал 0: к вашим у меня было раньше до этого **эквайринга**, **центроинве
 ст**
 Канал 1: угу
 Канал 0: выгодные условия мне предложили сбербанковские выгодные ус
 ловия я, **эквайринга** а вот э
 Канал 1: угу
 Канал 0: отчётного кассового яркового вашего не очень выгодно
 Канал 1: угу
 Канал 0: если я перейду и приехать с банка на расчётный счёт это ч

зать по онлайн, значит э сколько сейчас онлайн у вас платёжка стои
 т
 Канал 1: все платежи внутри сбербанка бесплатно пополнения три внеш
 них платежа ежемесячно вы сможете сделать так же без комиссии четвё
 ртого платежа комиссия за платёж составит сто рублей в внешние друг
 ие банки
 Канал 0: триста рублей **центрэвест** тридцать рублей онлайн тридцать
 рублей **центроинвест** и также внутри банка бесплатно вот и всё дальше
 можно как бы вообще понимаете
 Канал 1: угу. угу
 Канал 0: тридцать и сто разница весомые согласитесь
 Канал 1: да согласусь с вами андрей сейчас заведения расчётного счё
 та в **центроинвест** оплачиваете
 Канал 0: оплачиваю шестьсот рублей по-моему там подорожало шестьсот
 рублей по-моему
 Канал 1: угу, а плюс ещё э интернет э банк оплачивается ежемесячно
 какую-то сумму нет
 Канал 0: нет нет нет ничего
 Канал 1: ну а что касается платежей сколько ориентировочно формируе
 тся

ИСТОЧНИКИ



Комментарии
клиентского
менеджера



Звонки клиентов

ML ТЕХНОЛОГИИ

Классификация комментариев
клиентского менеджера

Определение потребностей
клиента
из комментариев

Преобразование речи в текст

Определение потребностей
клиента
и контекста из текста

Определение эмоций
клиента



РЕЗУЛЬТАТЫ

Определение оптимальной
даты следующего контакта

Бизнес-потребности клиента

Определение лучшей даты
следующего контакта

Информация о конкурентах,
где обслуживается клиент

DL Embeddings

Бизнес-потребности клиента



Переобучение системы
рекомендаций

ML360 портал

Embeddings

Ai Lab

Проприетарные embeddings
на базе транзакций
корпоративных клиентов

Open Source

DeepPavlov (ELMO),
BERT

Витрины

Более 2200 фич, готовых к
использованию

Фичи на базе клиентских
транзакций, данных CRM,
транскрипции данных колл-
центра

Еще больше фич в
разработке


Поиск по метаданным

Поиск Уведомления FAQ

Поиск по ключевым словам, например, "кредиты"

Фильтры: Тип данных: [День-ИНН](#), [Месяц-ИНН](#), [Квартал-ИНН](#) Статус вывода в пром: [sandbox](#), [pre-production](#), [production](#)

Введите ваш запрос

 [mon auto gibdd](#) new!

триггеры указывающие на наличие у клиента тех или иных признаков, касающихся ТС

Тип: **Витрина фичей**

Записей в витрине: ~ **18 000 000**

Интервал данных: 2016-01-31 – 2019-08-31

Фичей в витрине: **11**

Последнее обновление: 2019-10-04

[avaya_pds](#)

Информация по результатам обзвона клиентов

Тип: **Аналитическая витрина**

[elmo-wiki](#)

Эмбединги, обученные на статьях русскоязычной Википедии - ELMo (Embeddings from Language Models)

Тип: **Эмбединг**

tensorflow hub module_spec

на ЛД (ПРОМ) их можно скачать через: `hadoop fs -get /user/team/team_ml360/elmo-wiki.tar.gz ~/`



Быстрое прототипирование

```
from core.DL import init_spark_context
n_nodes = 10
dynamicAllocation_enabled = 'false'
sc = init_spark_context(spark_config=[ ('spark.driver.memory', '10g'),
                                       ('spark.executor.memory', '10g'),
                                       ('spark.driver.maxResultSize', '5g'),
                                       ('spark.port.maxRetries', '150'),
                                       ('spark.executor.cores', 4),
                                       ('spark.executor.instances', n_nodes),
                                       ('spark.default.parallelism', 1000),
                                       ('spark.sql.shuffle.partitions', 1000),
                                       ('spark.yarn.executor.memoryOverhead', '5g'),
                                       ('spark.dynamicAllocation.enabled', dynamicAllocation_enabled),
                                       ('spark.dynamicAllocation.minExecutors', n_nodes),
                                       ('spark.kryoserializer.buffer.max', '1g'),
                                       ('spark.yarn.queue', 'root.g_dl_u_corp.AutoML360'),
                                       ('spark.kryoserializer.buffer.max', '1g'),
                                       ('spark.blacklist.enabled', 'true'),
                                       ('spark.blacklist.timeout', '3h'),
                                       ('spark.blacklist.task.maxTaskAttemptsPerNode',
                                       ('spark.yarn.queue', 'root.g_dl_u_corp.AutoML360'),
                                       ('spark.ext.h2o.nthreads', -1),
                                       ('spark.ext.h2o.cluster.size', n_nodes)])

#
```

```
PandasDF = SparkDF.toPandas()
```

```
PandasDF.head()
```

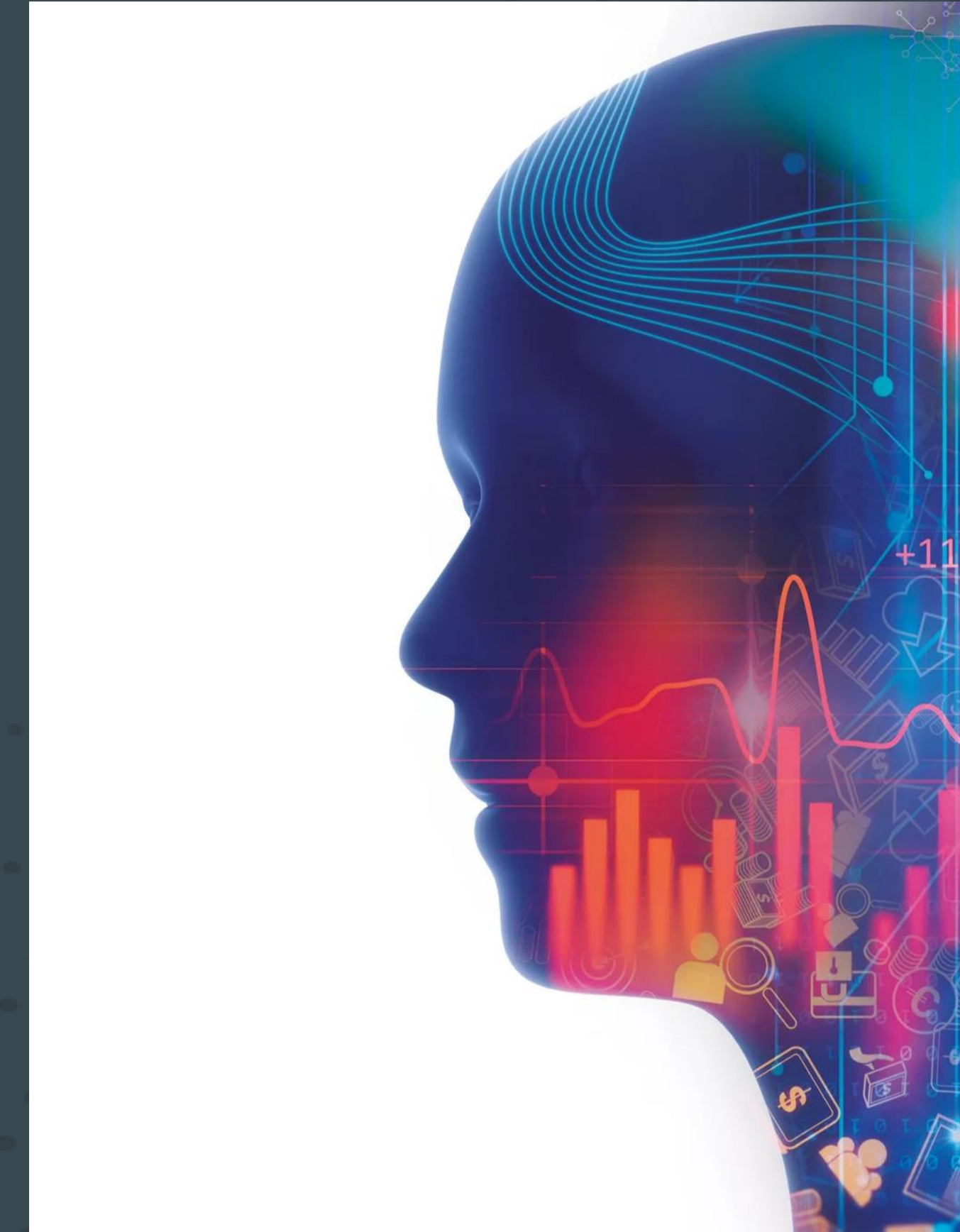
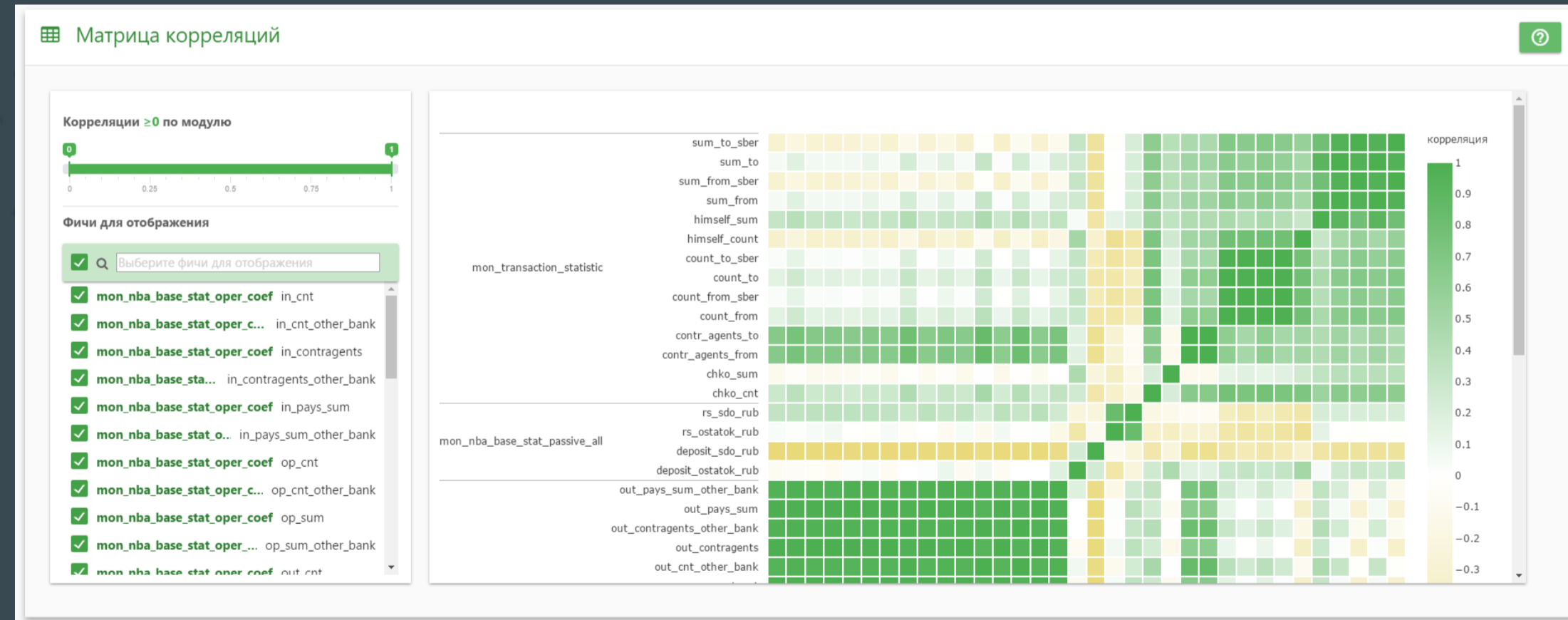
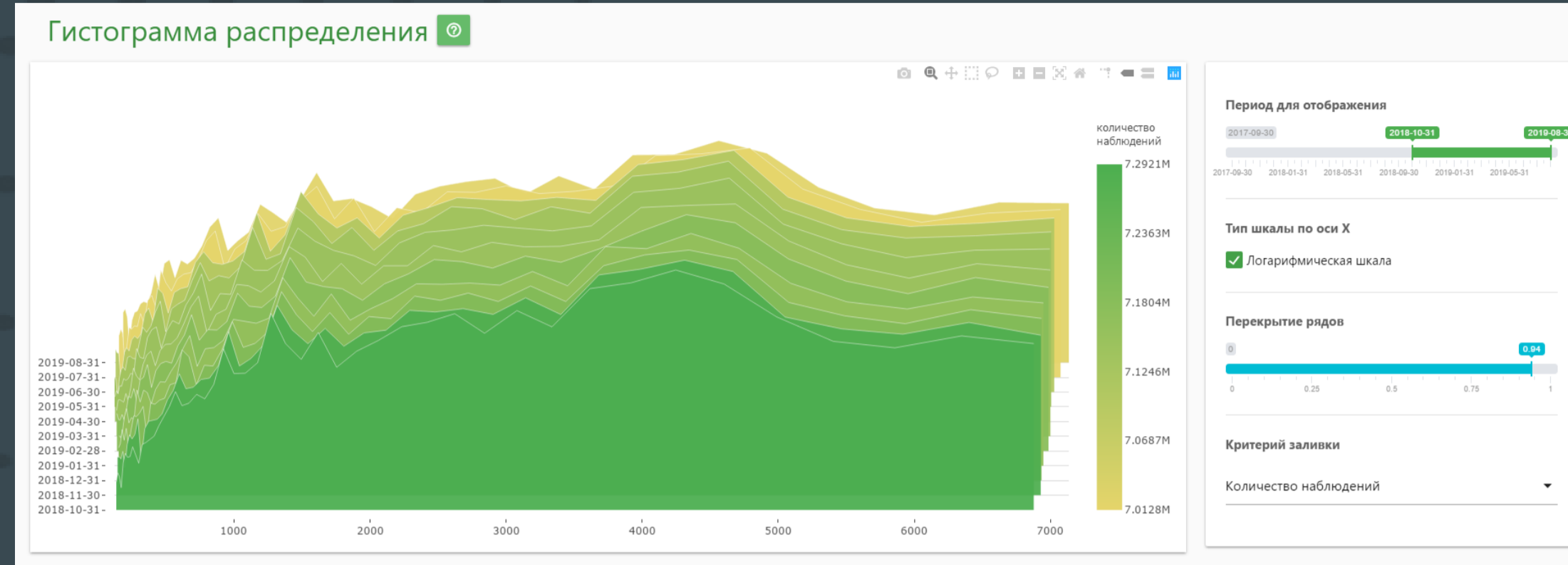
	PassengerId	Survived	Age	Fare	Embarked_LE	Sex_LE	Cabin_LE	Parch_LE	SibSp_LE	Pclass_LE	...	MAX_Fare_	MEAN_Age_	ROOT_Fa					
0	308	1	17.0	108.9000	1	0	75	0	1	0	...	512.3292	31.325843	10.435516	29.0	33.0	62.0	4.699571	53
1	428	1	19.0	26.0000	3	0	0	0	0	1	...	512.3292	31.325843	5.099020	29.0	33.0	62.0	3.295837	53
2	557	1	48.0	39.6000	1	0	3	0	1	0	...	512.3292	31.325843	6.292853	29.0	33.0	62.0	3.703768	53
3	143	1	24.0	15.8500	3	0	0	0	1	2	...	512.3292	31.325843	3.981206	29.0	33.0	62.0	2.824351	53
4	609	1	22.0	41.5792	1	0	0	2	1	1	...	512.3292	31.325843	6.448194	29.0	33.0	62.0	3.751366	53

Автогенерация кода

```
# Импорт библиотеки по получению данных из различных таблиц
from etl360.common import generic
# Задания параметров для получения сводной таблицы (заполняются автоматически с портала)
select_params = dict(
    sqlContext = hc,
    dict_param = {'t_team_ml360.titanic_features': {'columns': []},
}
)
data_pipe = generic(**select_params)
SparkDL = data_pipe["dest"]
SparkDF = SparkDL.get()
SparkDF.show(3)
```

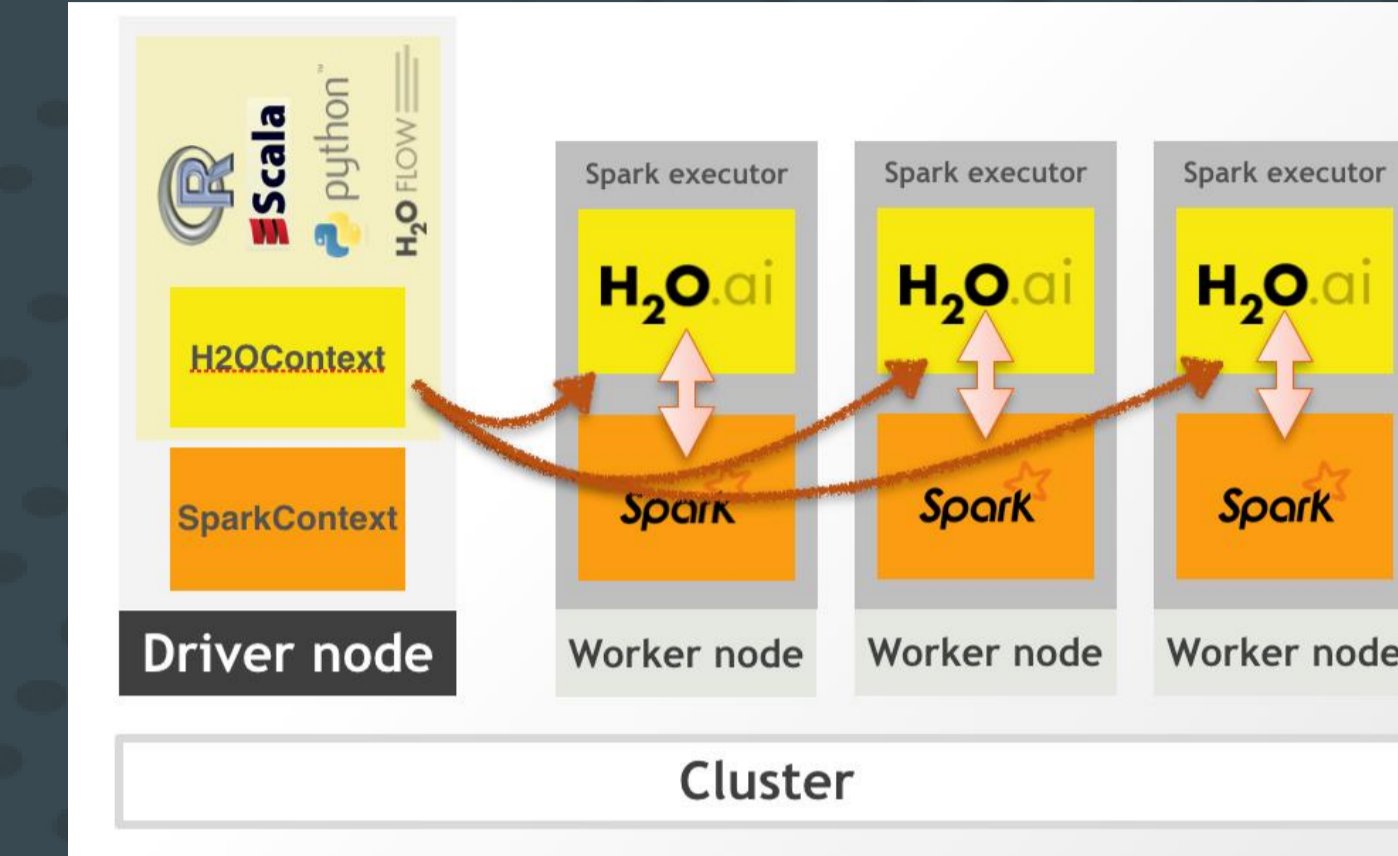
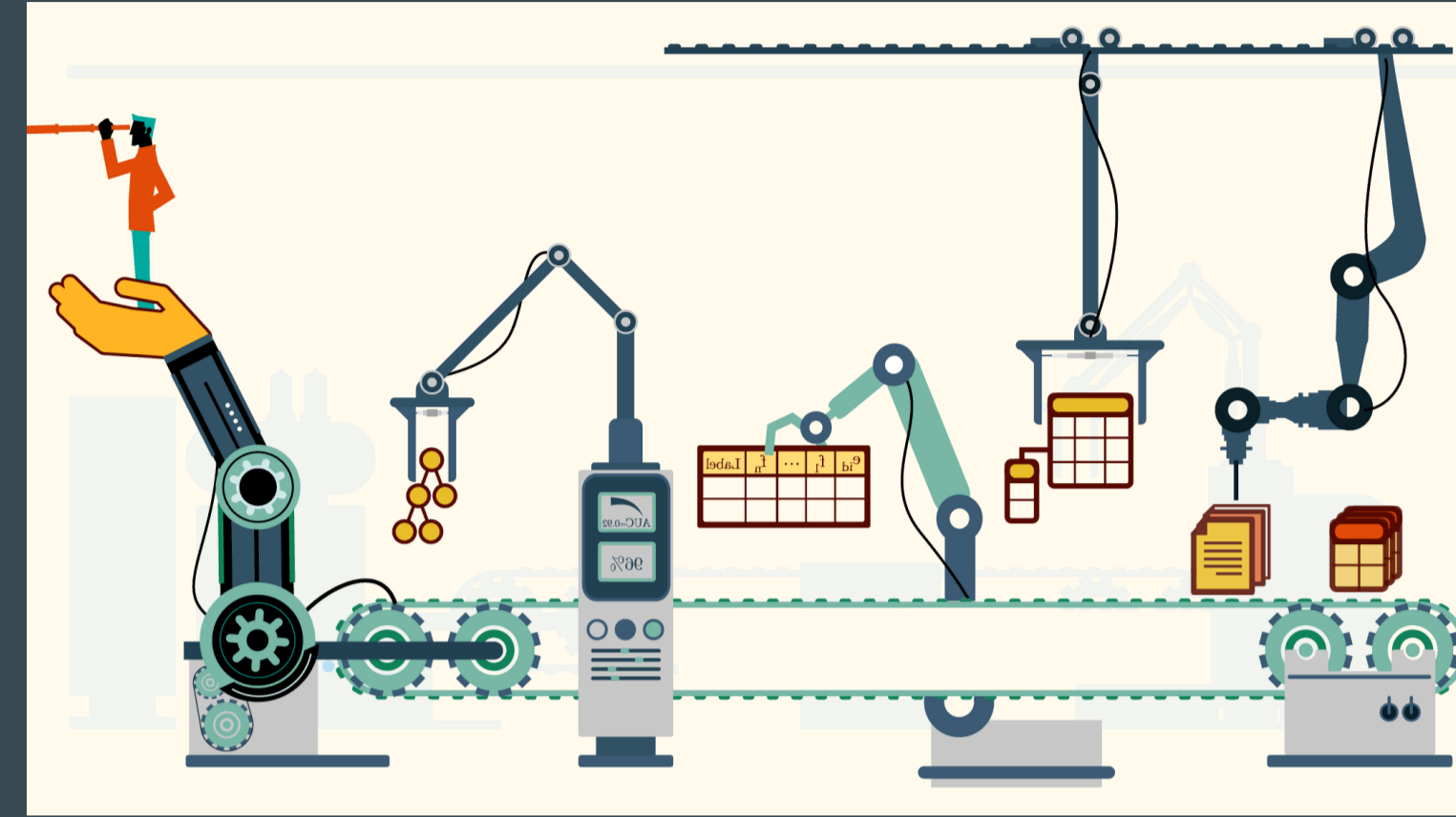


Качество данных



H2O B ML360

```
from core.DL import init_spark_context
from models.AutoML import init_h2o_context, AutoML360
n_nodes = 10
dynamicAllocationEnabled = 'false'
sc = init_spark_context(
    spark_config=[
        ('spark.driver.memory', '10g'),
        ('spark.executor.memory', '10g'),
        ('spark.driver.maxResultSize', '5g'),
        ('spark.port.maxRetries', '150'),
        ('spark.executor.cores', 4),
        ('spark.executor.instances', n_nodes),
        ('spark.default.parallelism', 1000),
        ('spark.sql.shuffle.partitions', 1000),
        ('spark.yarn.executor.memoryOverhead', '5g'),
        ('spark.dynamicAllocation.enabled',
         dynamicAllocationEnabled),
        ('spark.dynamicAllocation.minExecutors', n_nodes),
        ('spark.kryoSerializer.buffer.max', '1g'),
        ('spark.yarn.queue', 'root.g_dl_u_corp.AutoML360'),
        ('spark.kryoSerializer.buffer.max', '1g'),
        ('spark.blacklist.enabled', 'true'),
        ('spark.blacklist.timeout', '3h'),
        ('spark.blacklist.task.maxTaskAttemptsPerNode', 2),
        # ('spark.yarn.queue', 'root.g_dl_u_corp.AutoML360'),
        ('spark.ext.h2o.nthreads', -1),
        ('spark.ext.h2o.cluster.size', n_nodes)
    ]
)
h2o = init_h2o_context(sc, n_nodes)
```



ML портал – сервис, обеспечивающий нашим командам функцию поиска в данных и ETL для постройки и размещения моделей ИИ

The screenshot shows the ML portal interface with a search bar and filters. The search results are as follows:

Item Name	Description	Type	Additional Info
mon auto gibdd new!	триггеры указывающие на наличие у клиента тех или иных признаков, касающихся ТС Записей в витрине: ~ 18 000 000 Фичей в витрине: 11	Витрина фичей	Интервал данных: 2016-01-31 – 2019-08-31 Последнее обновление: 2019-10-04
avaya_pds	Информация по результатам обзвона клиентов	Аналитическая витрина	
elmo-wiki	Эмбединги, обученные на статьях русскоязычной Википедии - ELMo (Embeddings from Language Models) tensorflow hub module_spec на ЛД (ПРОМ) их можно скачать через: <code>hadoop fs -get /user/team/team_ml360/elmo-wiki.tar.gz ~/</code>	Эмбединг	

2200 атрибутов, готовых к использованию в ML

Скачивание одним кликом в Jupyter notebook

Автвалидация

DevOps моделей

Мониторинг

УСКОРЕНИЕ

3-6
мес.
до ПРОМ



1 день
для стандартных
ML моделей

Организация

Исследователи и инженеры данных (около 170)

в составе Agile - команд, разрабатывающих продукты



Команды CDS / CDO / Data protection



Инструменты разработки / внедрения

- Библиотека моделей
- Витрины модельных фичей
- Инструменты мониторинга моделей
- DevOps моделей и витрин данных

Time-to-market внедрения моделей – от 2 недель

РЕЗУЛЬТАТЫ

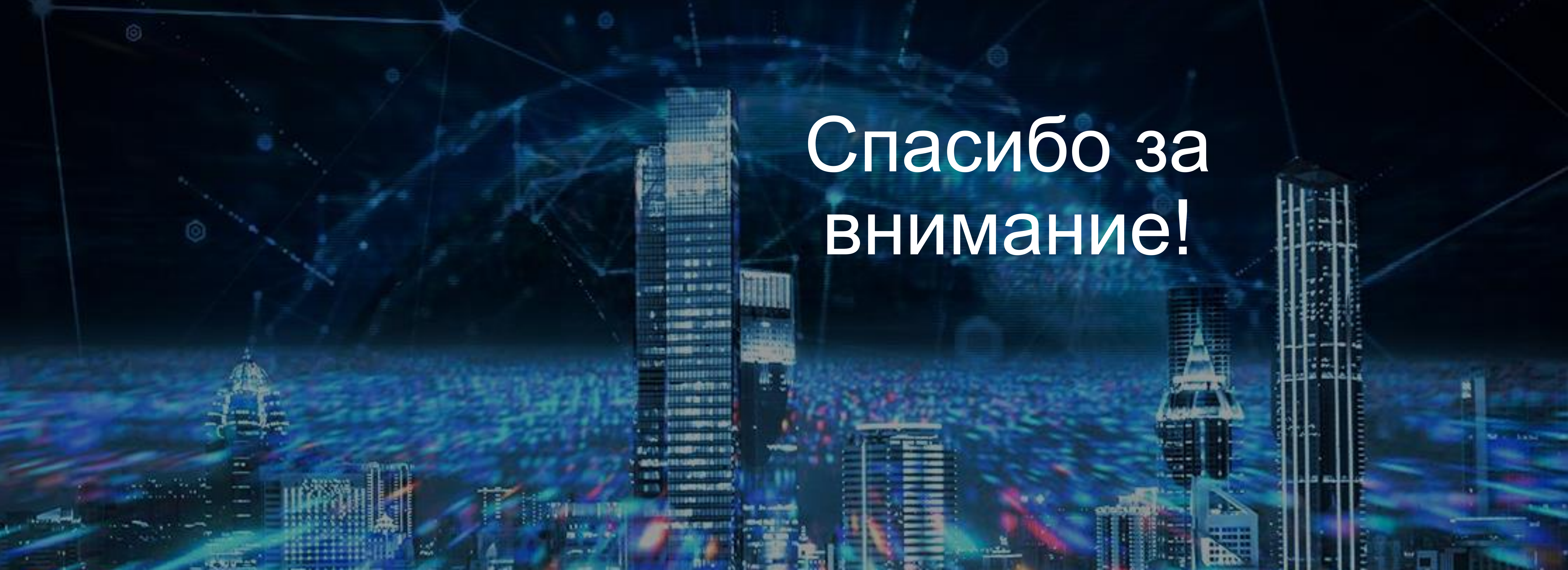
Для наших целей
гетерогенная
инфраструктура
данных дешевле и
более эффективно,
чем использование
одной системы

Использование
существующей
гетерогенной
инфраструктуры дало
2-3 года форы – без
необходимости
создания новой
гомогенной «с нуля»

Данные –
ключевой
элемент
инструментария
Data Scientist
(включая Auto ML)

Большой потенциал
неструктурированных
голосовых данных для
ML моделей (до 10%
AUC)

результат



Спасибо за
внимание!