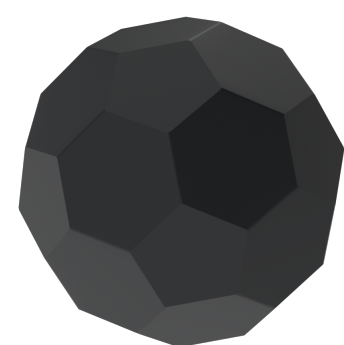


КАК BIG DATA ПОМОГАЕТ SPORTS.RU ПОНИМАТЬ СВОИХ ПОЛЬЗОВАТЕЛЕЙ И СОЗДАВАТЬ НОВЫЕ СЕРВИСЫ

Владислав Денисов



SPORTS.RU

Спортс – digital-издательство для неравнодушных к спорту

MAU

20 млн

уникальных
пользователей

DAU

1,5 млн

уникальных
пользователей

SOCIAL MEDIA

8 млн

уникальных
пользователей

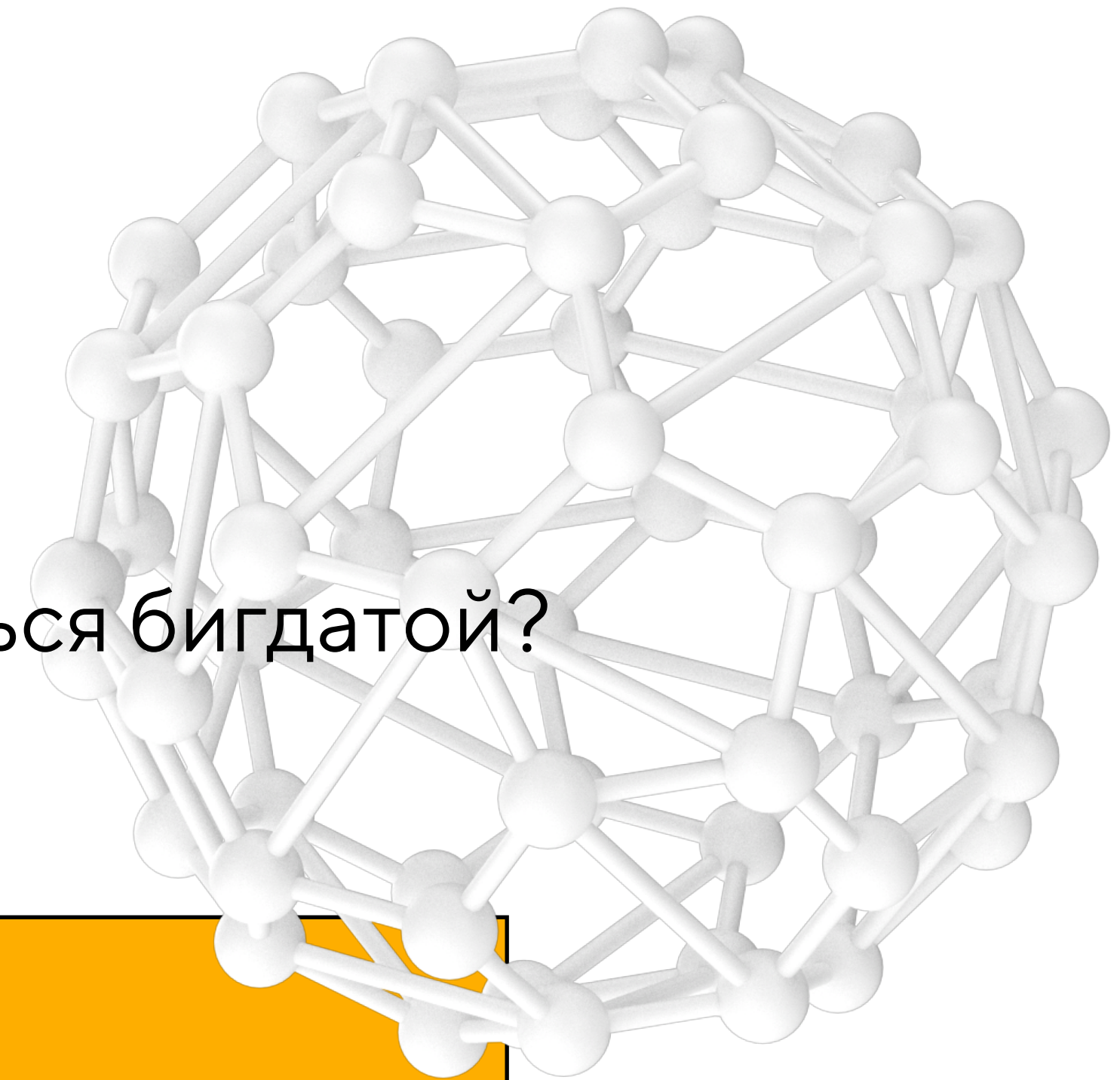
У нас огромное число UGC-контента, мы снимаем фильмы, создаем игры и много чего еще



здесь подробнее про нас

ЧТО ТАКОЕ БИГДАТА?

- Какой объем данных по-настоящему big в 2023-м?
- Был ли он другим 10 лет назад?
- А если нет map-reduce / spark / hadoop, можно заниматься бигдатой?



BUZZWORD?

ДААННЫЕ В SPORTS.RU

Ключевые для бизнес-логики

Спортивная статистика
(~ 3к событий на матч)

Новости и материалы (~ 3к в месяц)

Активность пользователей:
комментарии (до 80к в сутки),
рейтингование (до 250к в сутки)
и подписки

Картинки
и другие медиа-файлы

ДААННЫЕ В SPORTS.RU

Аналитические

Кликстрим: ежедневно +50 ГБ
сжатых данных о взаимодействии
пользователей с сайтами,
приложениями и другими сервисами

Экспорты из разных платформ
(реклама, CRM, технический
мониторинг и т.д.) для обогащения
отчетов

КОМАНДА

- 4 человека
- Нет разделения по ролям (веб / мобайл / продукт / маркетинг / бизнес / еще что-то)
- Есть условное разделение по зонам ответственности (описание, инжиниринг, n/r, исследования, спорт)
- Нанимали людей даже без знания SQL, главное – математика
- Аналитик – не калькулятор и не мастер по презентациям: все отчеты автоматизированы

АНАЛИТИКА

Базы данных



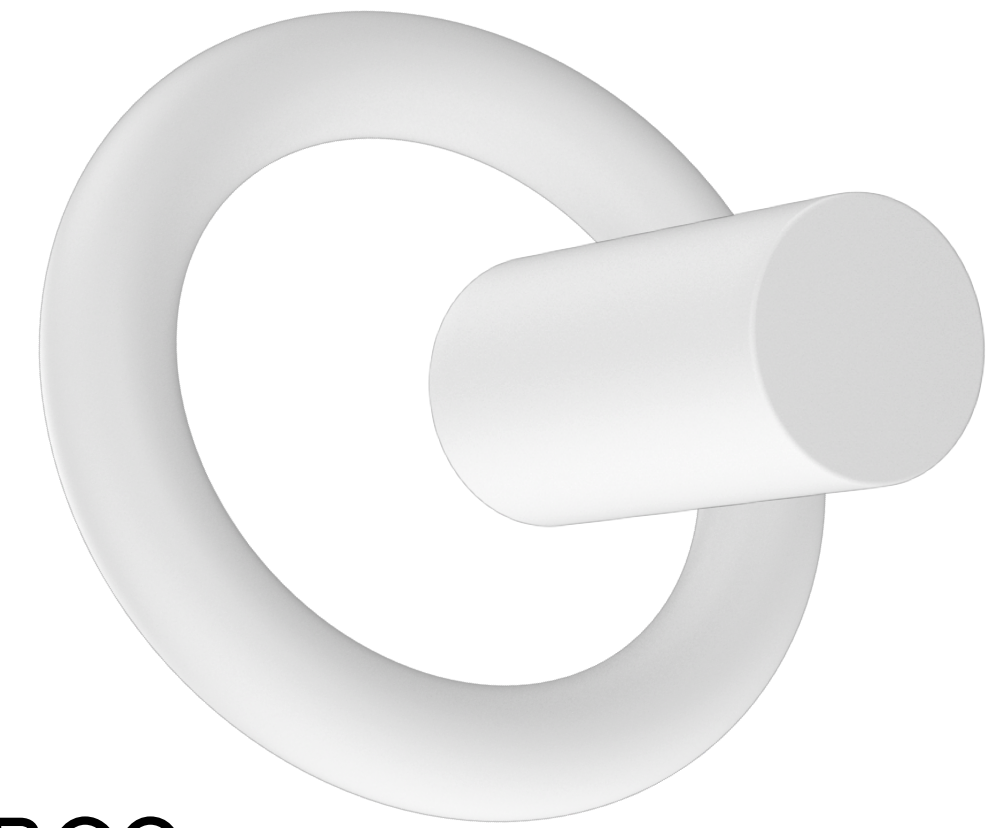
АНАЛИТИКА

Свой счётчик. Зачем, когда есть GA/ЯМ?

- Возможность построения кастомных метрик и отчетов
- Мощная детализация (фильтры по конкретному пользователю или последовательности событий)
- Единая структура для веба и приложений
- Обходим адблок – видим больше пользователей (+15% юзеров)
- Соединение с метаданными для красивых дашбордов
- Открыт в open-source, готовы консультировать по внедрению / использованию

АНАЛИТИКА

Визуализация



- Минусы своего счетчика – на выходе просто таблица в БД, все представления данных нужно собирать с нуля
- BI – Redash (даже контрибьютим в развитие) – огромный набор визуализаций и возможность собирать дашборды на Python
- Если нужна интерактивность – пишем приложение на Streamlit (Python)



main

oracle

sync

режим выгрузки

product

adblock

robot

sports

All

All

исключить просмотры маркетинга

platform

mobile web

geo

Мск и МО

Прогноз

Чем больше период, тем точнее прогноз.

existing start

existing end

2021-07-25

2021-08-01

Advanced

Построить

Динамика

Выгрузить в Google Sheets

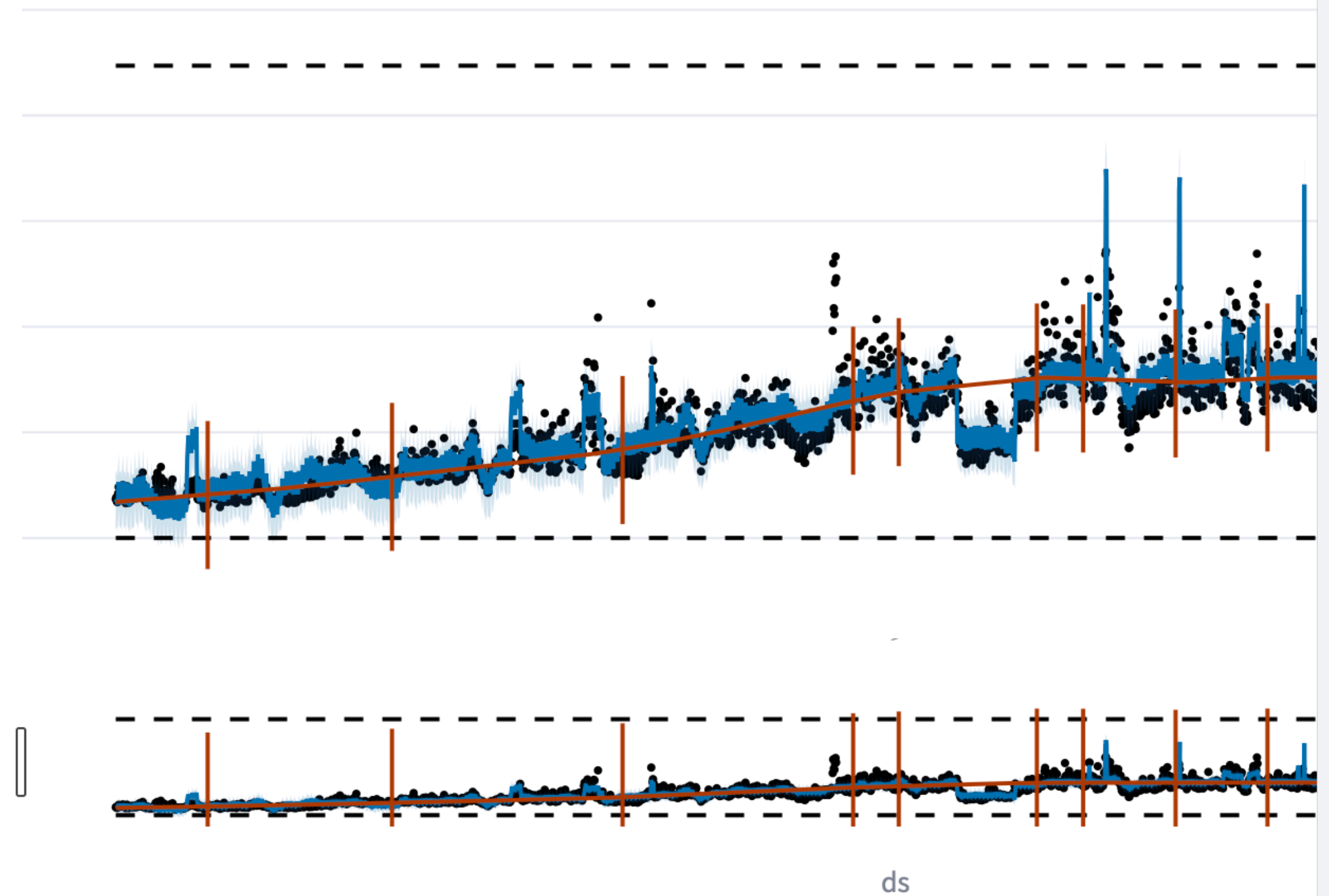
1w

1m

6m

1y

all



Аномалии

Данные с сервера <https://docs.google.com/spreadsheets/d/1t>

АНАЛИТИКА

Конкретнее про бигдату и детализацию

- Суммарно около 1000 строк в таблице с описанием событий
- Отслеживаем просмотр страницы и явные и неявные действия пользователей на них, каждое действие – 1 строка в базе, а 1 просмотр страницы генерирует около 100 событий
- Примеры вопросов, на которые умеем отвечать: какая дочитываемость у контента? а только у тех, кто заходит 5 дней подряд и чаще всего читает про Спартак? сколько денег нам приносит пользователь из ВК? какое пересечение у пользователей между вебom и приложениями? что мешает пользователям в процессе регистрации?



АНАЛИТИКА

Data Discovery

- Знания по структуре БД и событий не должны лежать в людях – выносим в общедоступные ресурсы и стараемся поддерживать актуальность
- Документация ивентов – в Google Sheets
- Документация БД – в DataHub



Google
Sheets



DataHub

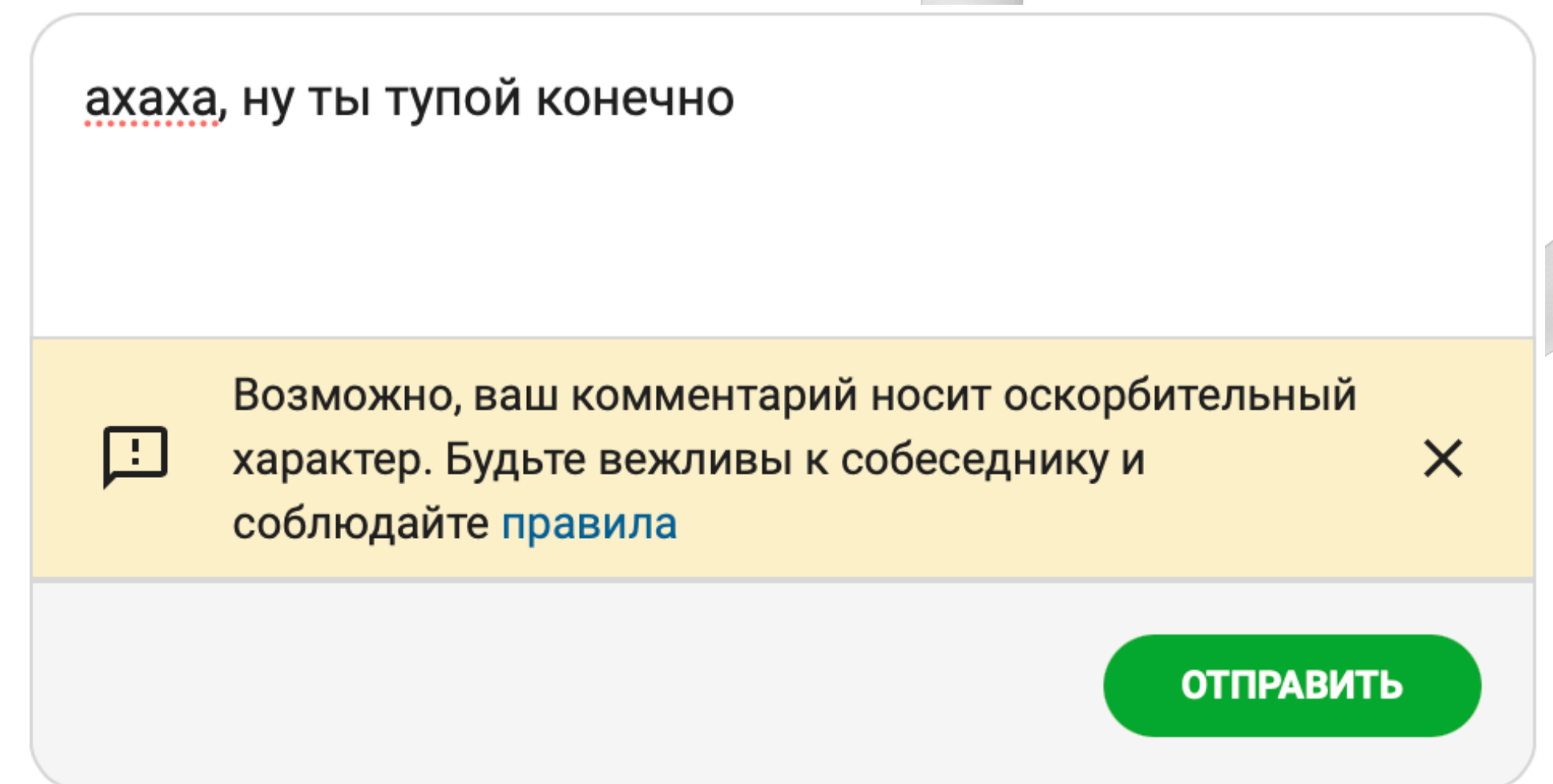
СЕРВИСЫ НА ОСНОВЕ ДАННЫХ

- Аналитики в Sports.ru почти не занимаются ресерчем в ML: часто используем лучшие практики в уже готовых библиотеках
- Команда аналитики сама пишет backend для таких сервисов
- Нейросети – не панацея, если классические алгоритмы справляются не хуже, чаще «классика» работает значительно быстрее, что важно для продакшна
- Есть задачи, которые приходят от бизнеса, но значимая доля сервисов рождается из желания попробовать новый инструмент

СЕРВИСЫ НА ОСНОВЕ ДАННЫХ

Антихейт / антиспам в комментариях

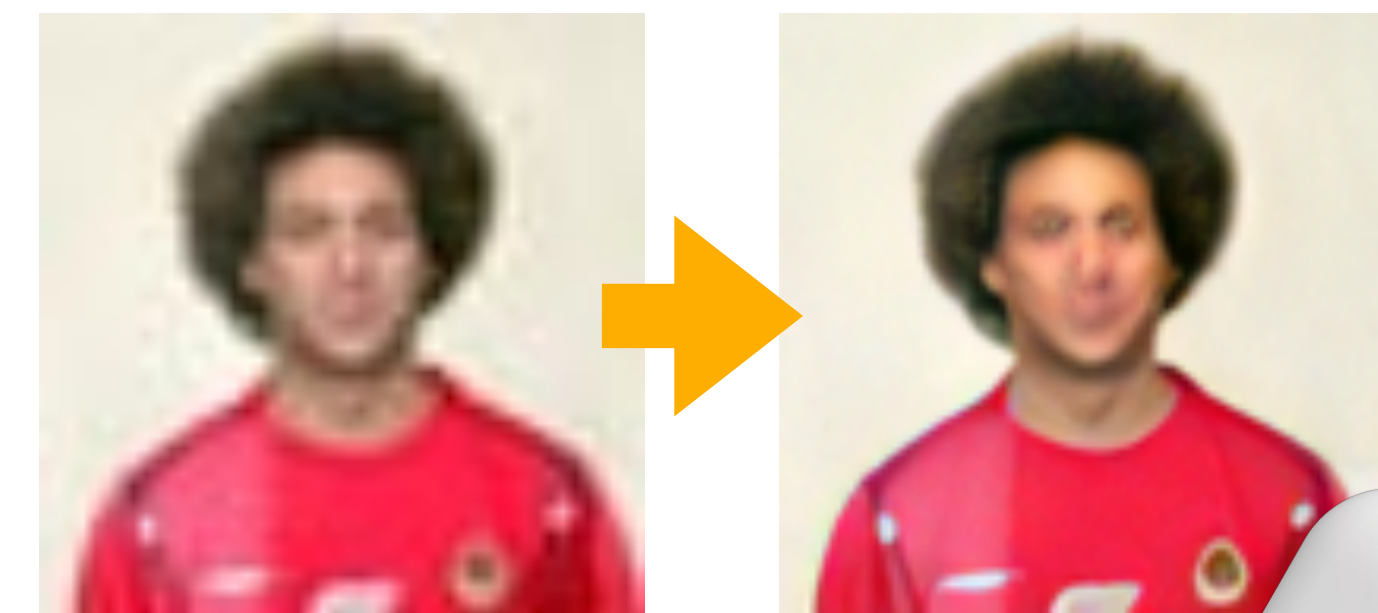
- Следующий слой после примитивного фильтра по запрещенным словам
- Сервис – модель классификации на основе BERT, использовали готовую, дообучив на своих данных
- Главный челлендж – использование в продакшне с минимальным временем ответа – решили через tritonserver от NVIDIA



СЕРВИСЫ НА ОСНОВЕ ДАННЫХ

Super resolution для старых изображений

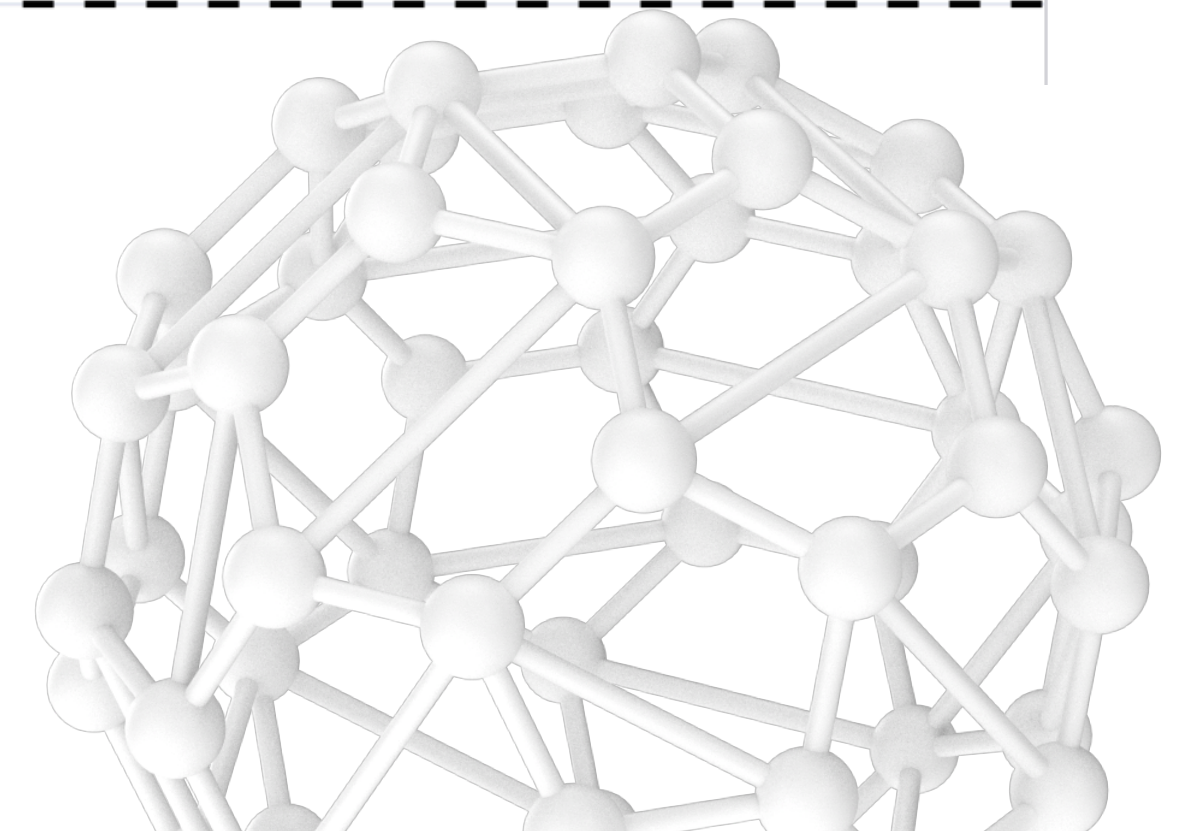
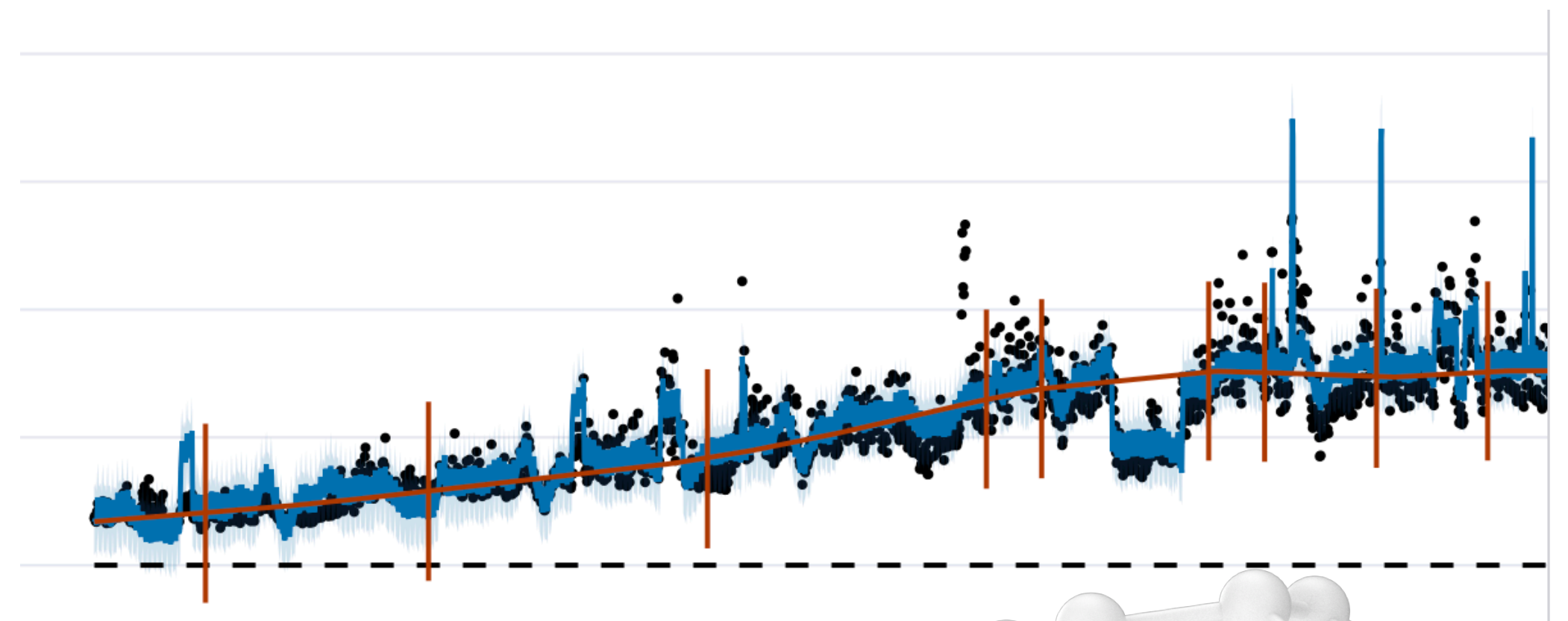
- Загруженные пользователями аватарки слишком маленькие для обновленного дизайна – нужно увеличить в 2-4 раза без потери качества
- Использовали готовую архитектуру, но обучали на своих данных – готовые модели работают с большим разрешением, у нас 48рх



СЕРВИСЫ НА ОСНОВЕ ДАННЫХ

Прогноз трафика

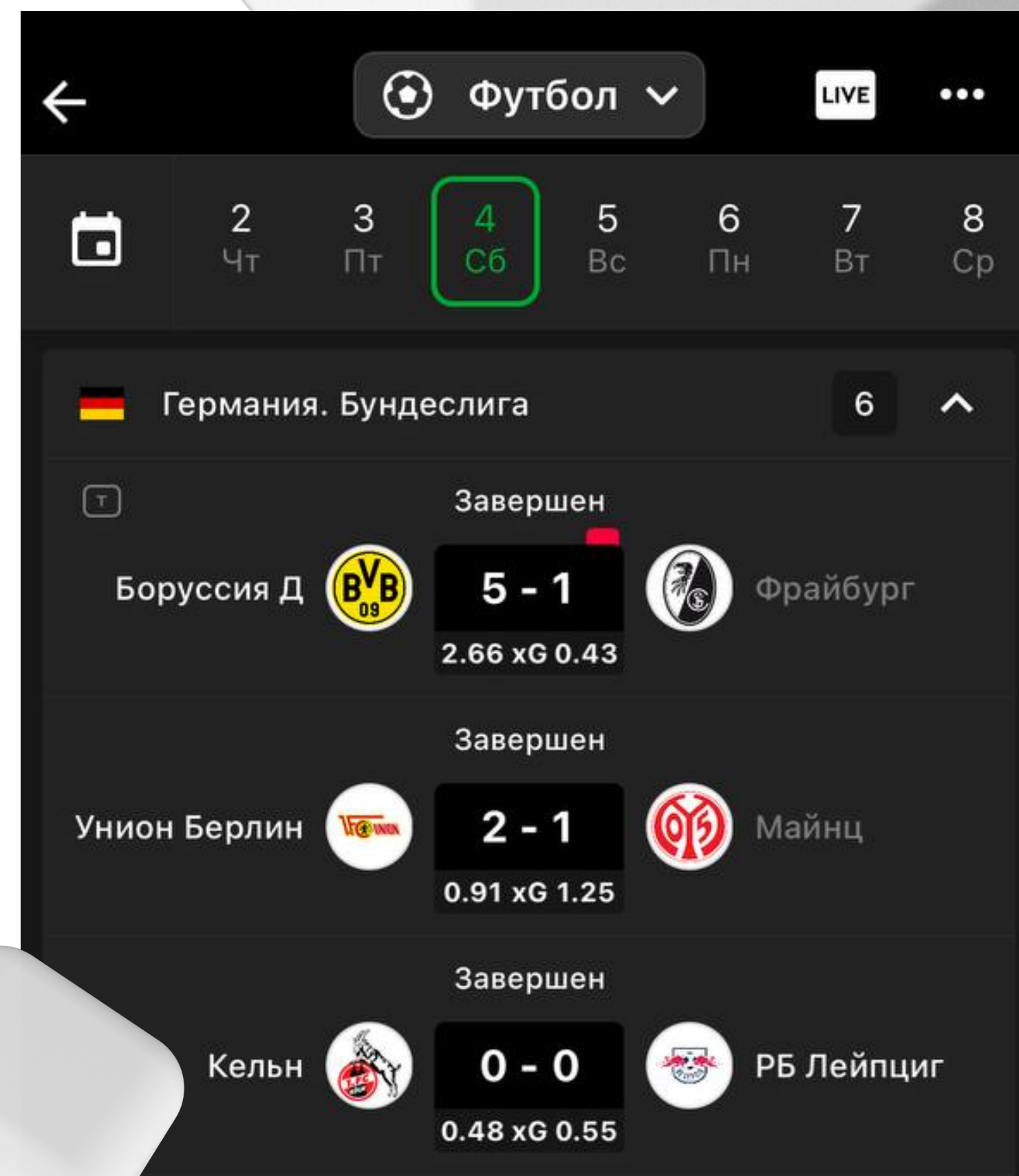
- У нас рекламная модель бизнеса – выручка коррелирует с просмотрами страниц
- Делаем краткосрочный (месяц) и долгосрочный (год) прогнозы
- Сервис прогноза – инструмент для отдела продаж – делаем так, чтобы рекламы было ровно столько, сколько инвентаря доступно



СЕРВИСЫ НА ОСНОВЕ ДАННЫХ

Expected goals в футболе

- xG – показатель расширенной футбольной статистики, который показывает остроту моментов
- Обучили модель классификации на основе бустинга, которая предсказывает вероятность гола по 100+ параметрам
- Топ-5 лиги + РПЛ имеют данный показатель, который обновляется прямо во время матча



The screenshot shows a mobile application interface for football. At the top, there is a navigation bar with a back arrow, a 'Футбол' (Football) dropdown menu, a 'LIVE' indicator, and a menu icon. Below this is a calendar view with days 2 (Чт), 3 (Пт), 4 (Сб), 5 (Вс), 6 (Пн), 7 (Вт), and 8 (Ср). The date '4 Сб' is highlighted with a green box. The main content area displays the 'Германия. Бундеслига' (Germany. Bundesliga) league. Three matches are listed, each with the status 'Завершен' (Completed), the score, and the xG values for both teams.

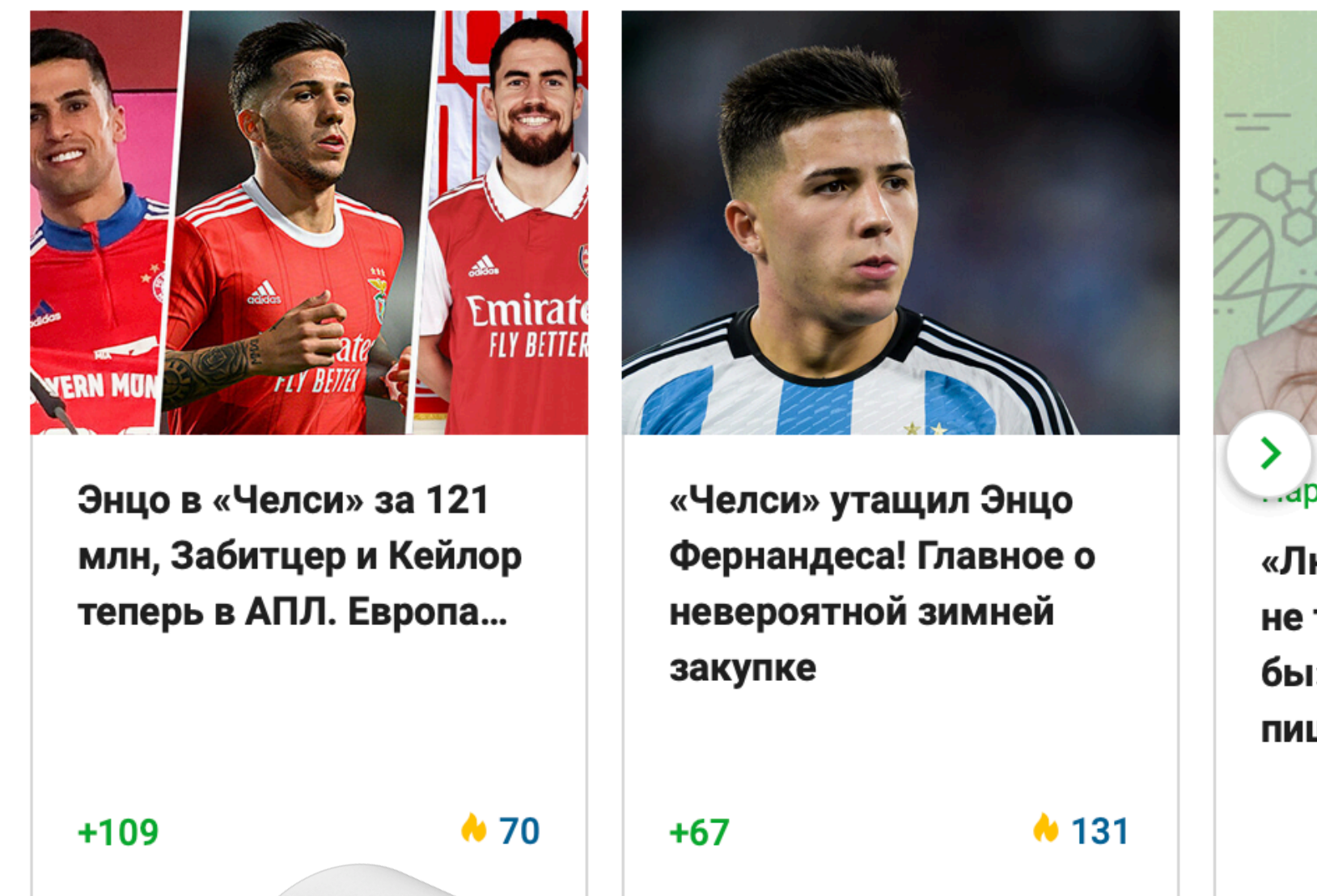
Match	Score	Team 1 xG	Team 2 xG
Боруссия Д vs Фрайбург	5 - 1	2.66	0.43
Унион Берлин vs Майнц	2 - 1	0.91	1.25
Кельн vs РБ Лейпциг	0 - 0	0.48	0.55

СЕРВИСЫ НА ОСНОВЕ ДАННЫХ

Кросспроектные рекомендации

- Делаем свою систему рекомендаций с 2017-го года, сейчас подключаем сторонние ресурсы (пока бесплатный SaaS)
- Работаем качественнее «обменок» (CTR + глубина), но без кликбейта
- Множество алгоритмов: многорукие бандиты, «классика», item2item, next-item – подстраиваемся под конкретную задачу

Рекомендуем



The screenshot displays a recommendation interface with two main article cards. The first card features three Arsenal players in red kits and a headline about Enzo Fernandez's transfer to Chelsea for 121 million. The second card features a close-up of Enzo Fernandez and a headline about Chelsea's winter transfer strategy. Both cards show engagement metrics: the first has +109 and 70 likes, while the second has +67 and 131 likes. A third card is partially visible on the right, showing a green background and a headline starting with '«Лю... не т... бы»: пиц...'. A grey 3D cylinder is positioned at the bottom of the interface.

Article Title	Engagement (+)	Engagement (🔥)
Энцо в «Челси» за 121 млн, Забитцер и Кейлор теперь в АПЛ. Европа...	+109	70
«Челси» утащил Энцо Фернандеса! Главное о невероятной зимней закупке	+67	131

ВСЕ!

Остались вопросы? Хотите потестировать
рекомендалку или счетчик аналитики? Пишите!

TELEGRAM: @VLADENISOV

MAIL: DENISOV@SPORTS.RU

