

# Как качество данных влияет на внедрение AI



**Евгений Загуменнов**

Руководитель Yandex Crowd Solutions

# 95% AI-проектов не увеличивают производительность\*

Главное препятствие на пути к масштабированию  
AI-систем — отсутствие обучения и регулярной доработки

Большинство систем GenAI не сохраняют  
обратную связь, не подстраиваются  
под контекст и не совершенствуются  
со временем

\*Источник: [The GenAI Divide STATE OF AI IN BUSINESS 2025](#)

# Что выделяет тех, у кого получается



AI-инструменты  
обучаются на всём цикле  
и встраиваются  
в рабочие процессы



Внешние партнёрства имеют  
вдвое больший  
успех по сравнению  
с внутренними разработками



Наибольший ROI  
в операционных  
процессах

# Мы ориентируемся на два подхода

**RLHF + Human-in-the-loop**

**Data-centric AI**

# RLHF + Human-in-the-loop

## Замена:

- ↘ До 90% обучающих данных синтетическими приводит только к незначительному снижению производительности
- ⊗ Последних 10% вызывает серьёзные деградации\*

## Непрерывное улучшение:

- 🎓 Human-in-the-loop создаёт циклы обратной связи, где AI непрерывно учится и улучшается на основе человеческого ввода

\*Источник: [A Little Human Data Goes A Long Way research](#)

# Data-centric AI

Data-Centric AI помогает создавать надёжные, точные и воспроизводимые модели с минимальными усилиями по изменению архитектуры

## Данные — 80% успеха внедрения AI

## «Мусор на входе — мусор на выходе»

Некачественные данные приводят к плохим результатам, несмотря на сложность модели

## Ключевые процессы с данными:

-  Очистка и исправление ошибок
-  Улучшение разметки
-  Стратегическая выборка и аугментация
-  Использование алгоритмов для оценки качества

# Данные нужно «ГОТОВИТЬ»

## Уже много данных

промышленность

сельское хозяйство

медиа ритейл e-com и др.

## Экспертизы и архитектурных решений тоже хватает

## Но сырых данных недостаточно

Чтобы построить эффективную ML-модель, данные нужно подготовить – это позволит выстроить постоянное улучшение модели

# Проблемы подготовки данных



Попытки закрыть всё автоматикой проваливаются



Дообучение требует людей



Отсутствие внутренней экспертизы в разметке данных

# Примеры сценариев

Они все нуждаются в специфичной, часто экспертной подготовке данных

1

Оцифровка  
объектов  
культурного  
наследия  
в музеях

2

Контроль  
содержания животных  
на фермах (болезни,  
питание)

3

Выявление аномалий  
на производственной  
линии, прогнозирование  
поломок оборудования

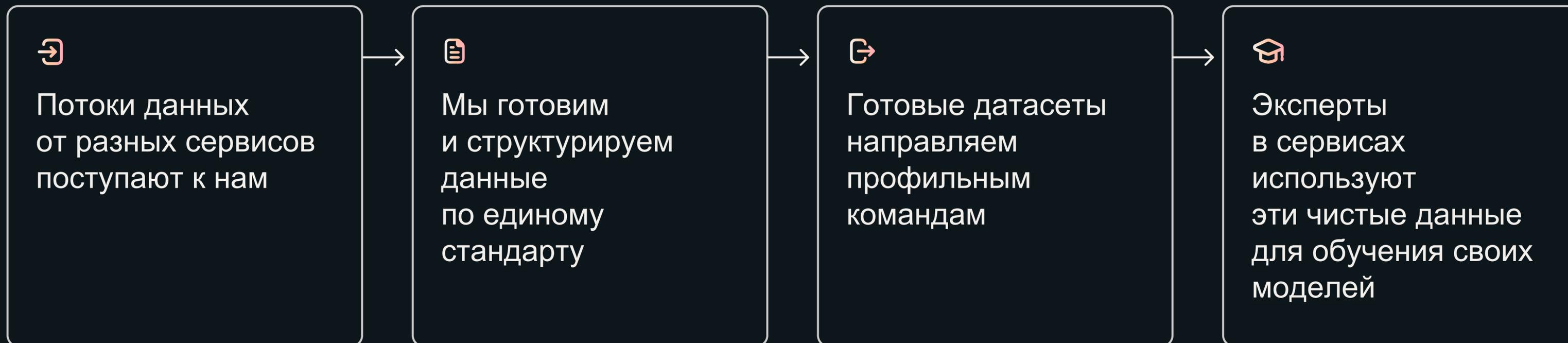
4

Обучение комбайнов  
и сельскохозяйственных  
роботов

5

Конвейерная  
сортировка

# Наш подход к подготовке данных внутри Яндекса



**80+**

продуктов и сервисов Яндекса пользуются нашими решениями

**160K+**

верифицированных и квалифицированных исполнителей

**20 лет**

отраслевого опыта и научных исследований

# Наш подход к подготовке данных вне Яндекса

От

## Проекты «под ключ»

- Услуга
- Высокое качество и экспертиза
- Возможность работы с данными, требующими профильной экспертизы
- Проекты «под ключ»



## Платформа Яндекс Задания

- SaaS
- Эксперименты и быстрый старт
- Широкий пул исполнителей
- Самостоятельная настройка

Yandex Crowd Solutions

# Спасибо!

✈ @ezagumenov

✉ ezagumenov@yandex-team.ru



**Евгений Загуменнов**

Руководитель Yandex Crowd Solutions



Yandex Crowd Solutions