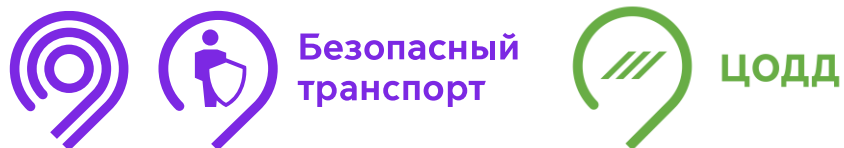


Москва | 2026

Production ML для RTSP-камер: как построить реалтайм платформу видеоаналитики

16 сентября 2026 г.





Александр Мадумаров

Начальник отдела интеллектуальных
систем и машинного обучения
в Инновационном центре
«Безопасный транспорт» при ЦОДД

**Сегодня
мы поговорим
не про ML**



Данные



Потоки

ОД
МЫ ПОГОВОРИМ
ОД



Инженерия



Эксплуатация



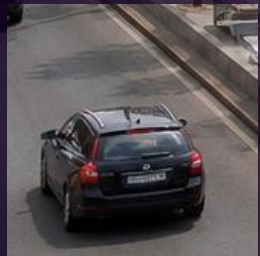
Безопасный
транспорт



ЦОДД

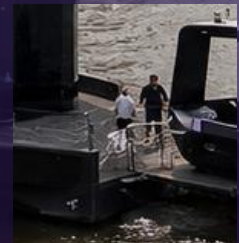
Какие задачи мы решаем

Трекинг и классификация
ТС на дорогах



Детекция СИМ
в парковочных зонах

Трекинг пассажиров
речного транспорта



От ML-модели к Production ML-системе аналитики



RTSP-камера + YOLO —
еще не сервис



**Нужны события
и метрики**

Непрерывное получение
данных

Обработка потока
с приемлемой задержкой

Детекции в треки

Подсчет событий

Сохранение истории

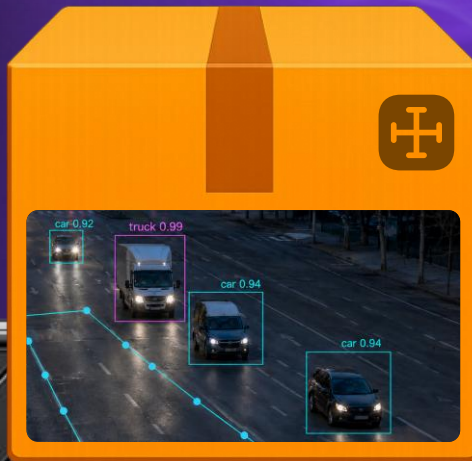
Интеграция с внешними
системами

Настройка бизнес-логики
без изменения кода

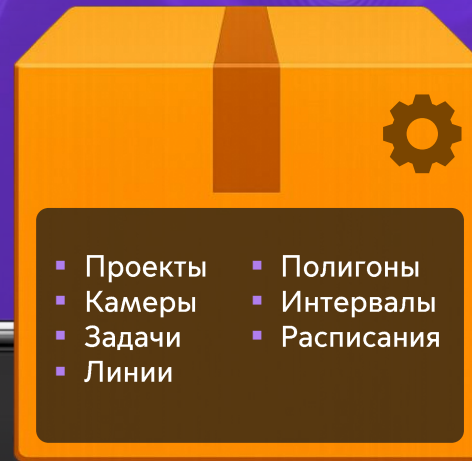


Модель — лишь часть платформы

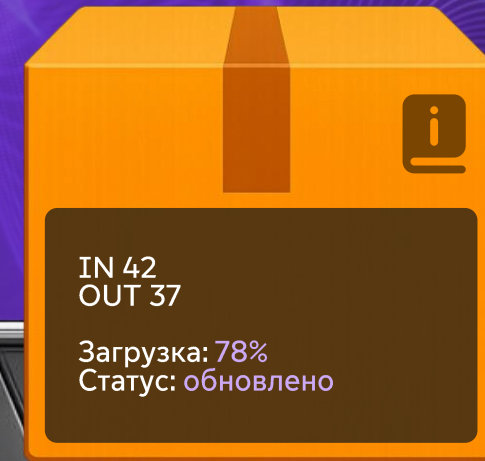
ML-модель



Production-платформа



Бизнес-метрика



Сложности работы с физическими детекторами



Правила работы с физическими детекторами

01

Лимит совместного использования детектора

02

Минимальный интервал между расчетами

03

Дочерний интервал не меньше родительского

04

Разделение real-time и загрузки

05

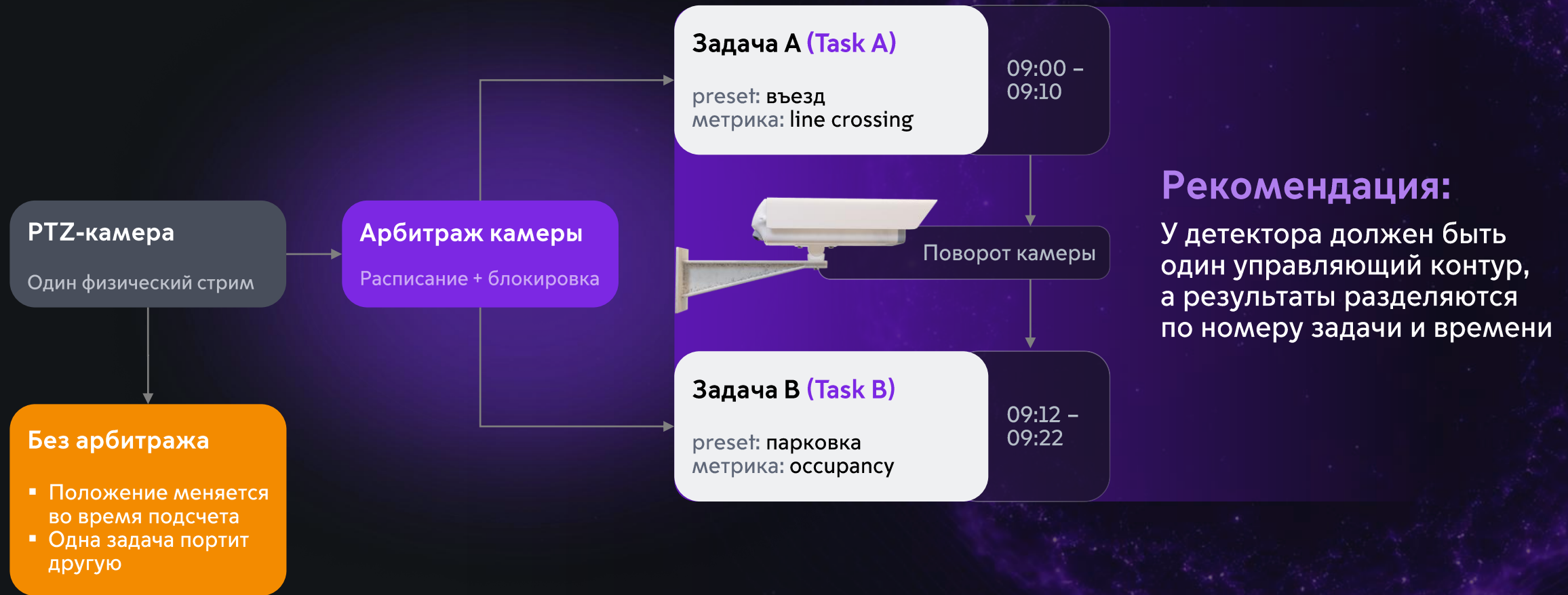
Лимит уникальных ракурсов

06

Разные ракурсы

Один детектор, две задачи: конфликт ракурсов

Детектор не может одновременно
смотреть в разные стороны,
поэтому задачам нужен арбитраж



Сквозная архитектура сервиса



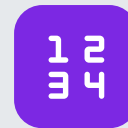
Получение
URL



Обработка
кадров



Отслеживание
объектов



Подсчет
объектов



Сохранение
данных в базе

Обработка видео в реальном времени

Одна камера, один поток, одна метрика

Практический путь от сырого видео к бизнес-метрике

Камера.
Видеопоток



Кадры.
Sampling



Модель
детекции



Трек + линия.
Событие



Ценность появляется после бизнес-логики

Метрика:
 $\text{count_in} / \text{count_out}$

Методы, от которых отказались

Готовое решение и покадровая обработка
не подошли под сценарии проекта

01. Виртуальные детекторы

Ultralytics count

Риск ошибки детектора



Свои зоны + линии

02. Архивное видео

1 кадр

GPU недогружен



Батч 50 кадров

Быстрее для архива

Готовое $\neq$ подходящее под задачу

Как обрабатывается один кадр

01

**Получаем
кадр**

Система
подключается
к видеопотоку
или получает
одиочный кадр

02

**Запускаем
модель**

Передача в модель
детекции
для извлечения
объекта

03

**Обновляем
трекер**

Трекер присваивает
идентификатор
объекту,
отслеживает его
между кадрами

04

**Формируем
аналитическую
запись**

Создается запись
с координатами
объекта, классом
и идентификатором
трекера

05

**Сохраняем
и считаем**

Записи сохраняются
и используются
для подсчета
объектов

06

**Проверяем
визуально**

Кадры с разметкой
сохраняются
для визуальной
проверки

И все-таки немного о моделях

Вычитание фона

Разница кадров

Анализ движения

HOG + SVM

Каскадные классификаторы



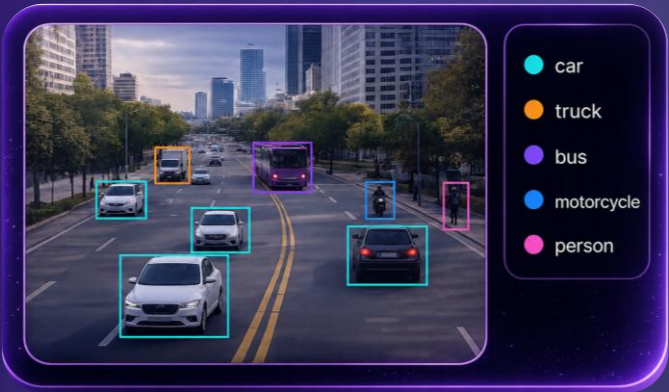
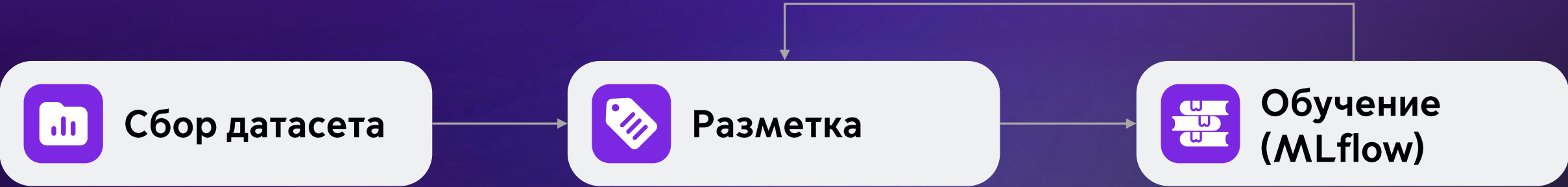
YOLO

Faster R-CNN

DETR /RT-DETR

SAM

Обучение моделей



Сложности разметки

Неоднозначность классификации



Разные типы камер

Обзорные камеры



PTZ камеры



Детекция по точке: когда bbox избыточен

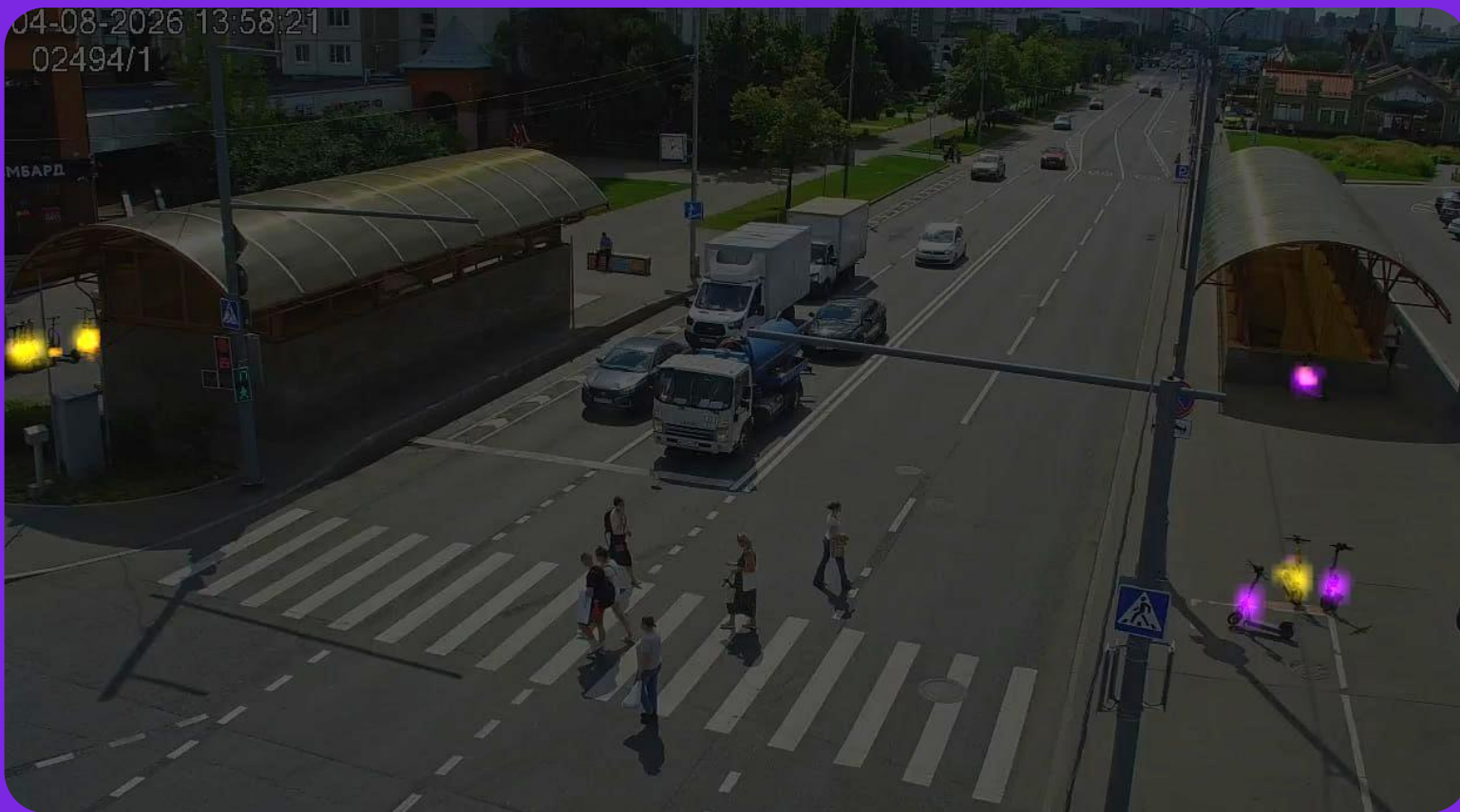


Объекты маленькие или однотипные

Важен подсчет, а не точная рамка

Въех-разметка слишком дорогая

Почему focal loss



● Самокаты Yandex

● Самокаты Whoosh

Модель учится
на редких объектах

Focal loss и плотностное пятно самоката

Градиент на границе
распознанного самоката



Увеличенная область density map

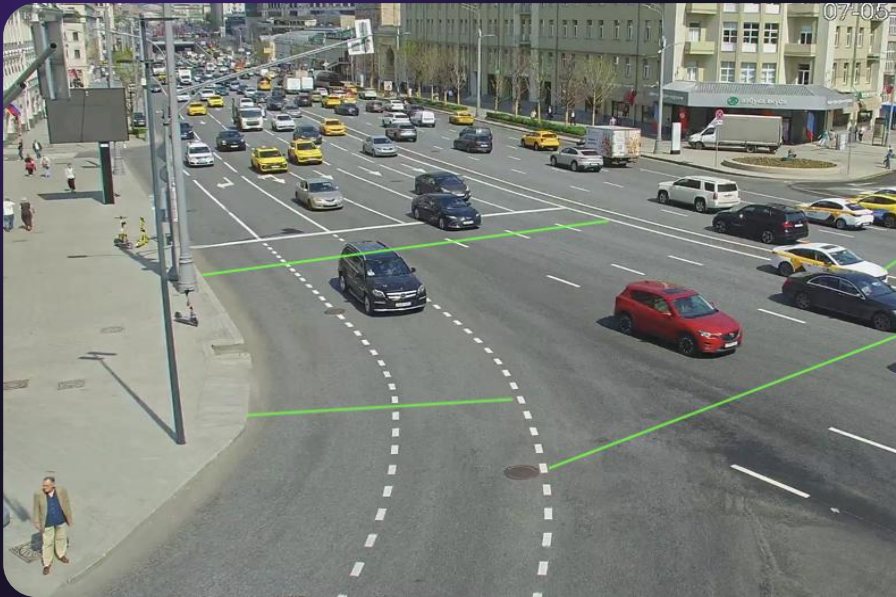


Числа — вклад
ячеек пятна
одного самоката

Один объект
добавляет ровно
1.00 к сумме
яркостей
пикселей карты
плотности

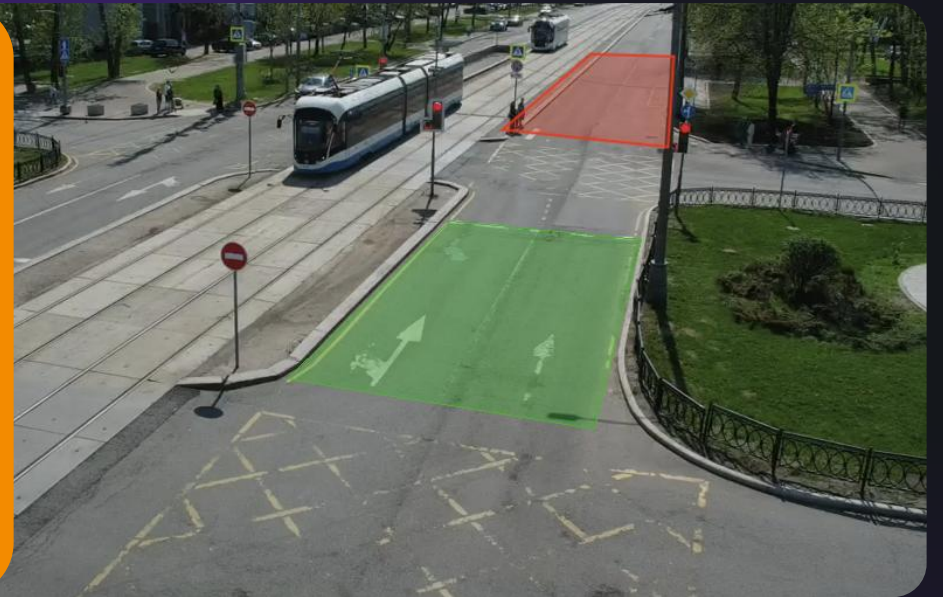
От bbox к треку и событию

Модель отвечает на вопрос
«Что видно на кадре?»



Основная
сложность
production ML
для камер

Продукт отвечает на вопрос
«Какое событие произошло?»



Безопасный
транспорт



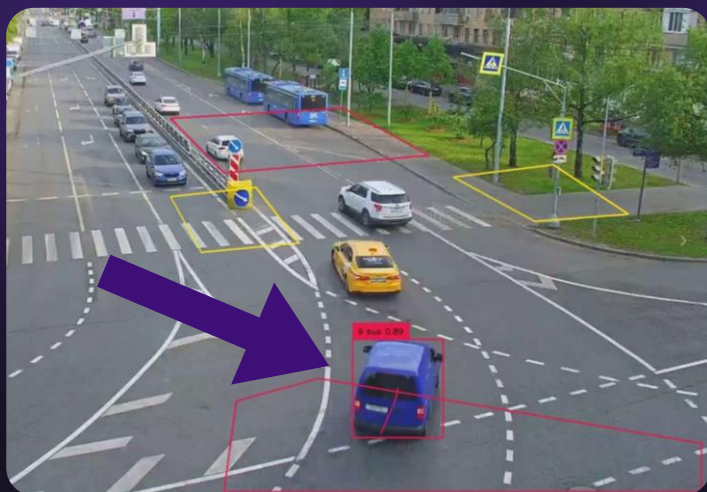
ЦОДД

Вbox — еще не событие

Событие появляется только после анализа движения объекта во времени

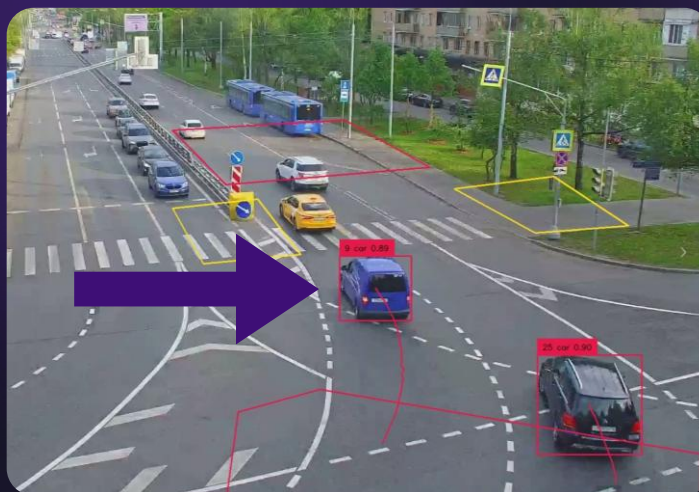
Кадр 1

Объект выезжает из полигона



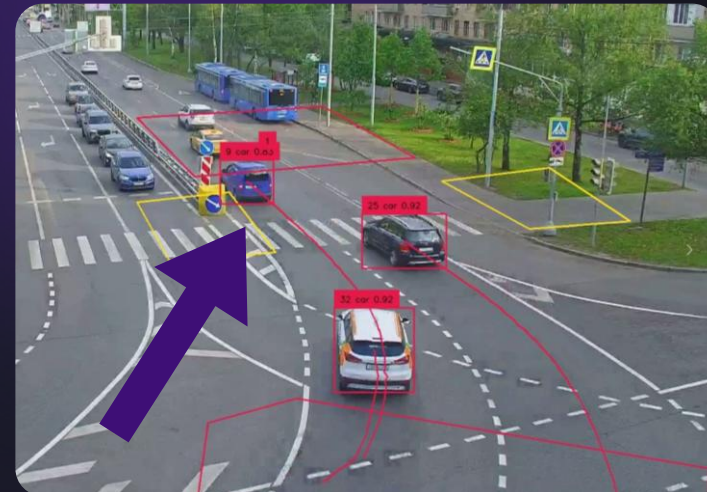
Кадр 5

Объект между полигонами



Кадр 10

Объект заехал в следующий полигон



Событие — перемещение между полигонами

Линии, полигоны и заполняемость

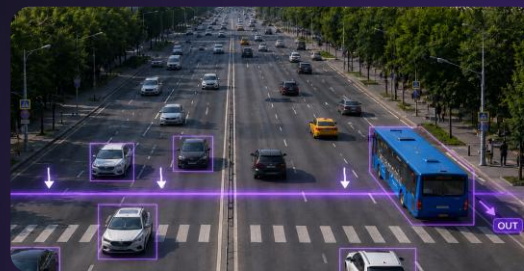
Одна модель детекции —
разные сценарии аналитики

Поток с камеры
RTSP | 25 FPS



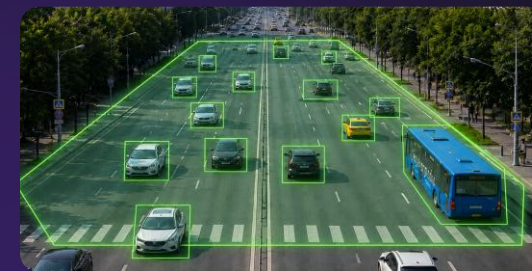
01. Подсчет по линиям

Пересечение объекта
с виртуальной линией



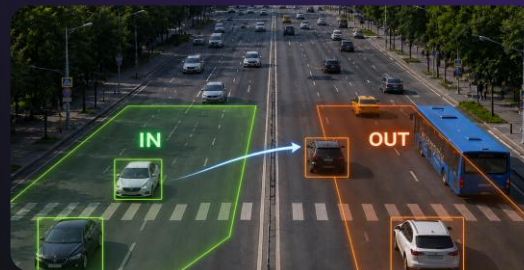
02. Одиночный полигон

Количество объектов
внутри зоны



03. Мультиполигоны

Анализ переходов
между зонами



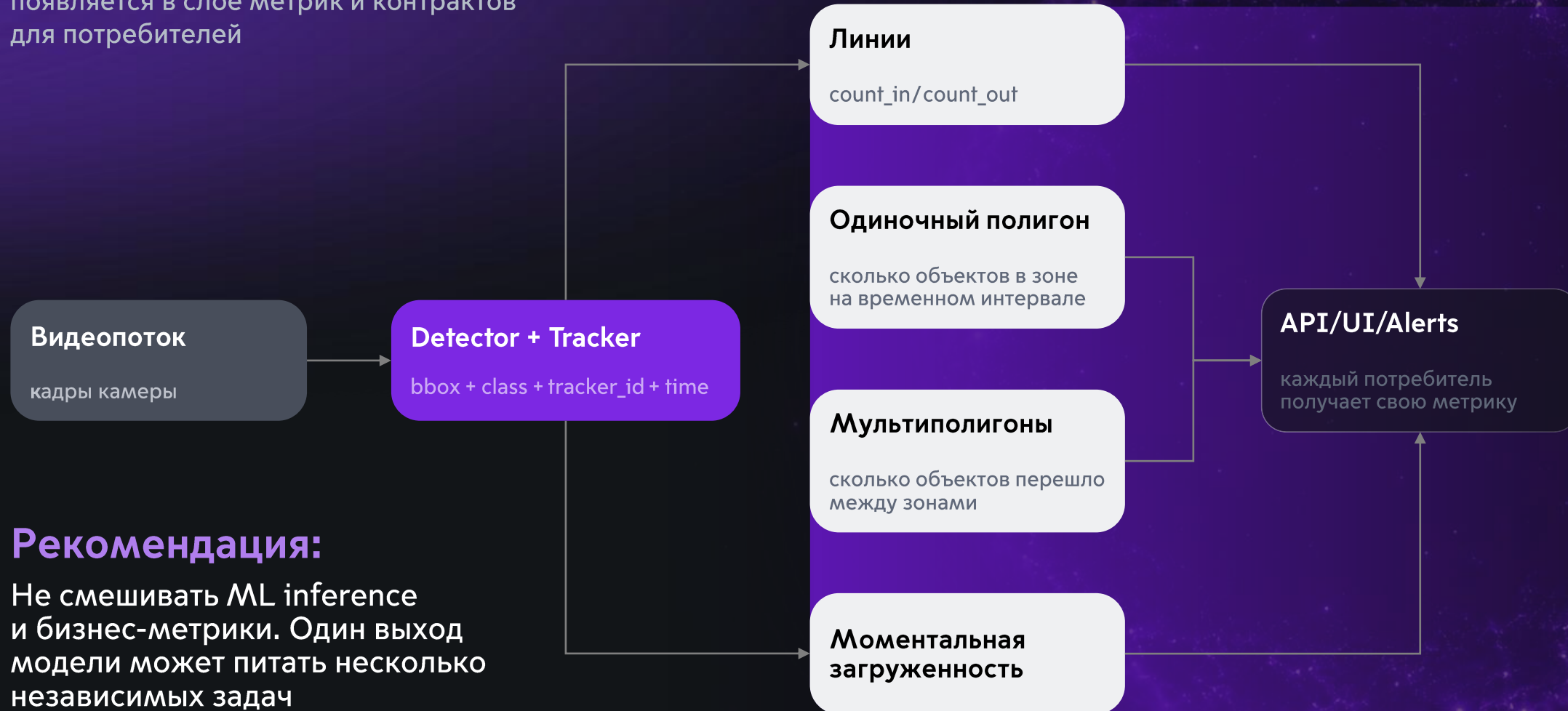
04. Моментальная загруженность

Оценка объектов в зоне
по периодическим снимкам



Одна модель — разные бизнес-метрики

Модель даёт общий сигнал, а ценность появляется в слое метрик и контрактов для потребителей

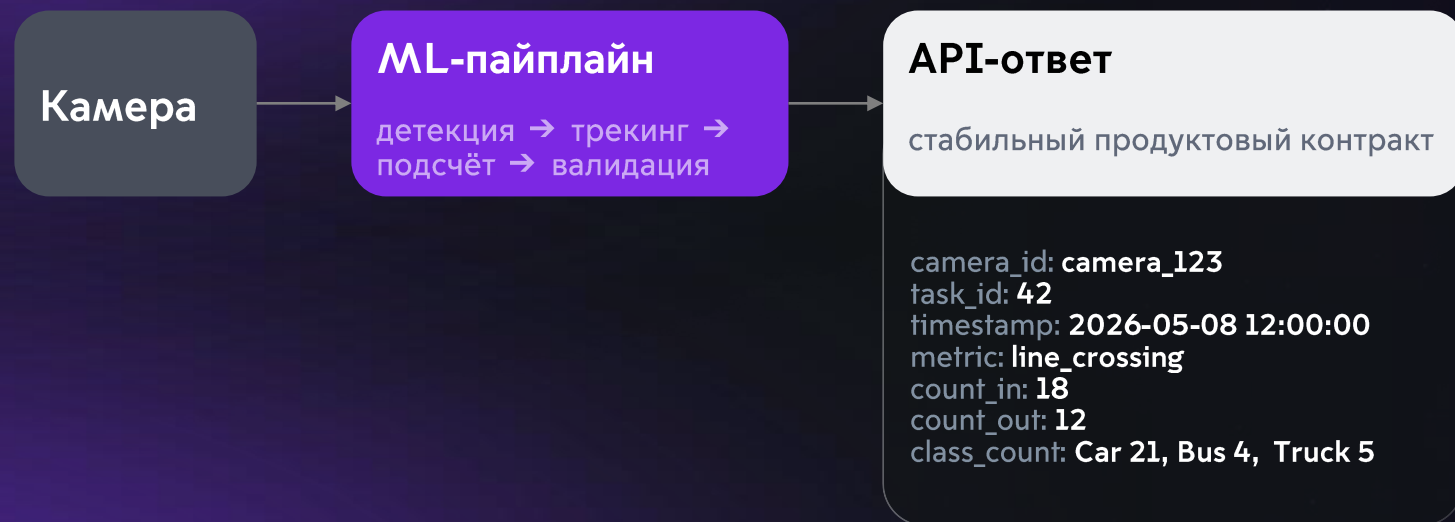


Рекомендация:

Не смешивать ML inference и бизнес-метрики. Один выход модели может питать несколько независимых задач

Что нужно внешней системе

Интеграциям нужны бизнес-события,
а не сырые bbox модели



Отдавать наружу:

- метрику
- время
- источник
- статус
- классы



Не отдавать наружу:

- сырые bbox
- вероятности
- внутренние детали

Производственный деплой



Инфраструктурные требования

Камеры, контейнеры или сервисы приложения, надежное хранилище

Сколько камер?

Сколько GPU / CPU?



Ресурсное планирование

Ресурсы GPU и CPU, декодирование

Как часто писать детекции?

Сколько соединений нужно?



Внешние системы

Взаимодействие с камерами, API и хранилищем

Что при сбое API?

Как восстановить задачи?

**Деплой —
это запуск системы,
а не контейнера
с Python**

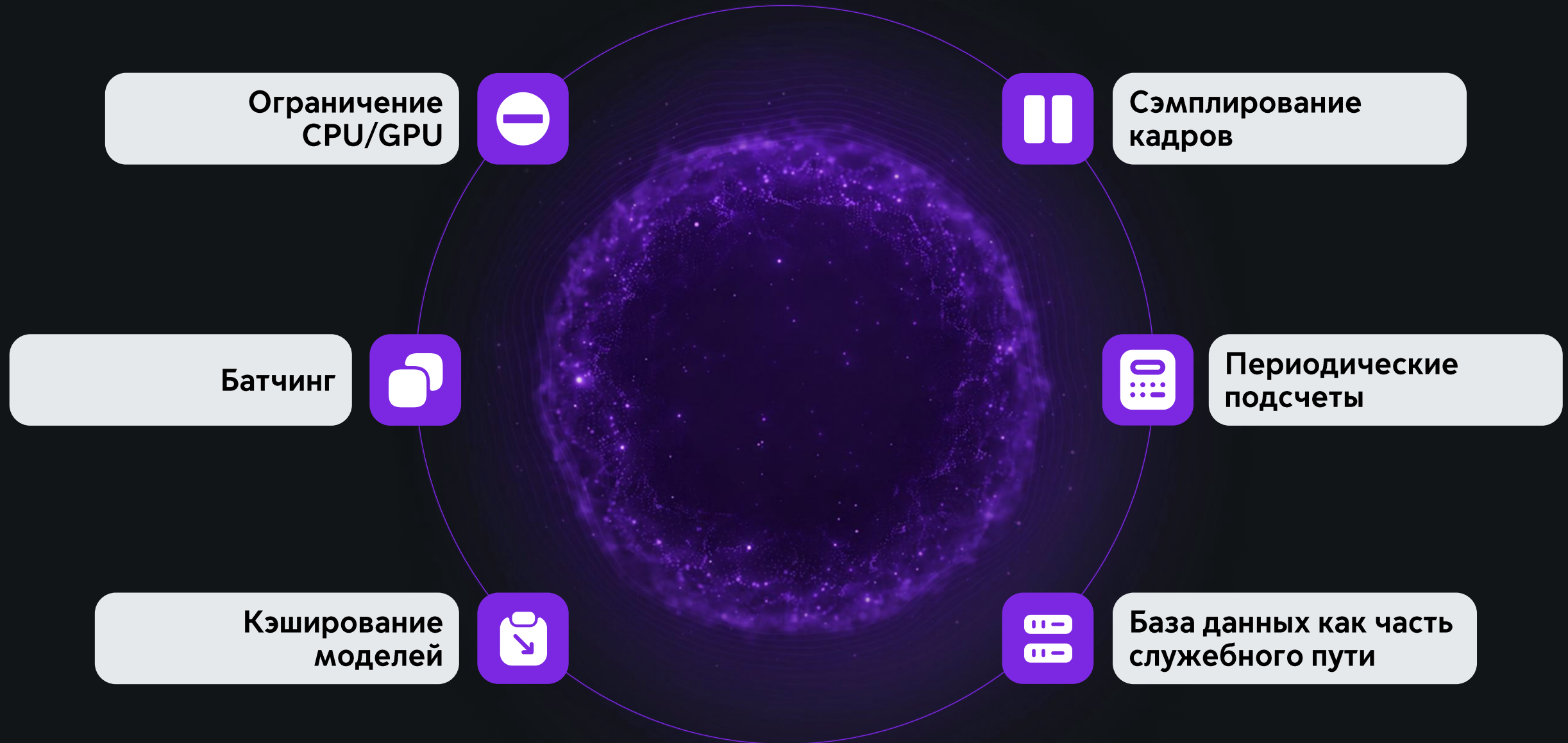


Безопасный
транспорт



ЦОДД

Производительность: FPS, GPU, батчинг, интервалы



Производительность: Тестовый стенд RTX 4060

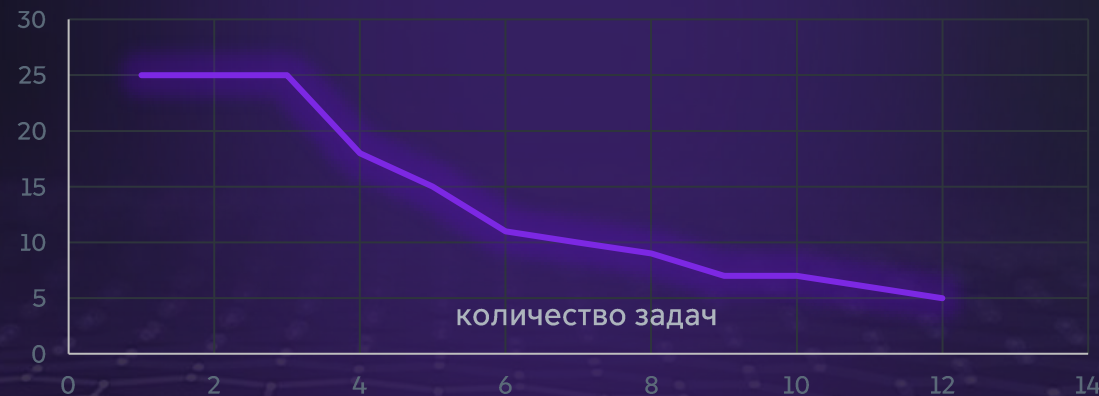
% GPU



VRAM (ГБ)



FPS



FPS против стоимости

Обрабатывать все кадры

250 FPS

 10 камер

х

 25 FPS

Высокая нагрузка,
больше данных, выше лаг

sampling

Подбирать частоту под задачу

80-125 FPS



Каждый 2-й или 3-й кадр

Меньше нагрузка,
метрика достаточно точна

Рекомендация: оптимизировать SLA бизнес-метрики, а не сырой FPS

Нагрузка на БД



1 камера

25 FPS, 10 объектов в кадре

До 250 detection-записей/сек

15 000 записей/мин

900 000 записей/час



х 10 камер

25 FPS, 10 объектов в кадре

До 2 500 detection-записей/сек

150 000 записей/мин

9 000 000 записей/час

Мониторинг — ключ к стабильности Production ML



Состояние задач



Ошибки видеопотоков



TTL ссылок



Метрики фоновых задач



Расписания



Практический сбой: сервис жив, но данных нет



Плохой мониторинг

ОК, контейнер работает

ОК, процесс запущен

ОК, HTTP health

НО новых кадров нет



Хороший мониторинг

Кадры приходят

Детекции и подсчеты обновляются

TTL URL под контролем

Lag не растёт

SLA — про данные, а не только про uptime сервиса



Безопасный
транспорт



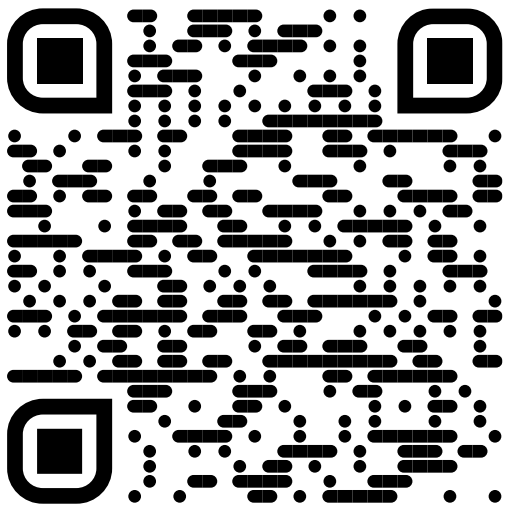
ЦОДД

Production ML — инженерия всей цепочки от камеры до бизнес-системы

ML-модель



Инновационный центр
ictransport.ru



Безопасный
транспорт

Спасибо за внимание!