# Непрерывная оптимизация

Лекция 1

# О себе

Мышлянов Алексей Владимирович

Аспирант 3 года кафедры АСУ, НИТУ МИСиС
Аналитик больших данных в МегаФоне (Senior Data Scientist)

Контакты:
+7 999 132 02 16
tg: **@l3lush**
mail: avmysh@gmail.com

# Из чего будет состоять курс

Итого **16 занятий**:

- 16 лекций
- 16 практик

=> будет **5-6 практических/домашних работ**

Изучим:

- MLOps

# Зачем вообще нам нужен ИИ

*Технологии на основе ИИ помогают повысить эффективность и производительность труда за счет автоматизации процессов и задач, которые раньше выполнялись людьми. ИИ также умеет интерпретировать объемы данных, которые не под силу интерпретировать человеку. Это умение может приносить существенные преимущества для бизнеса.*


Если кратко:

1. Автоматизируем производство
2. Повышаем показатели бизнеса

Если совсем кратко:

1. **Деньги**

# ML код – лишь малая часть системы

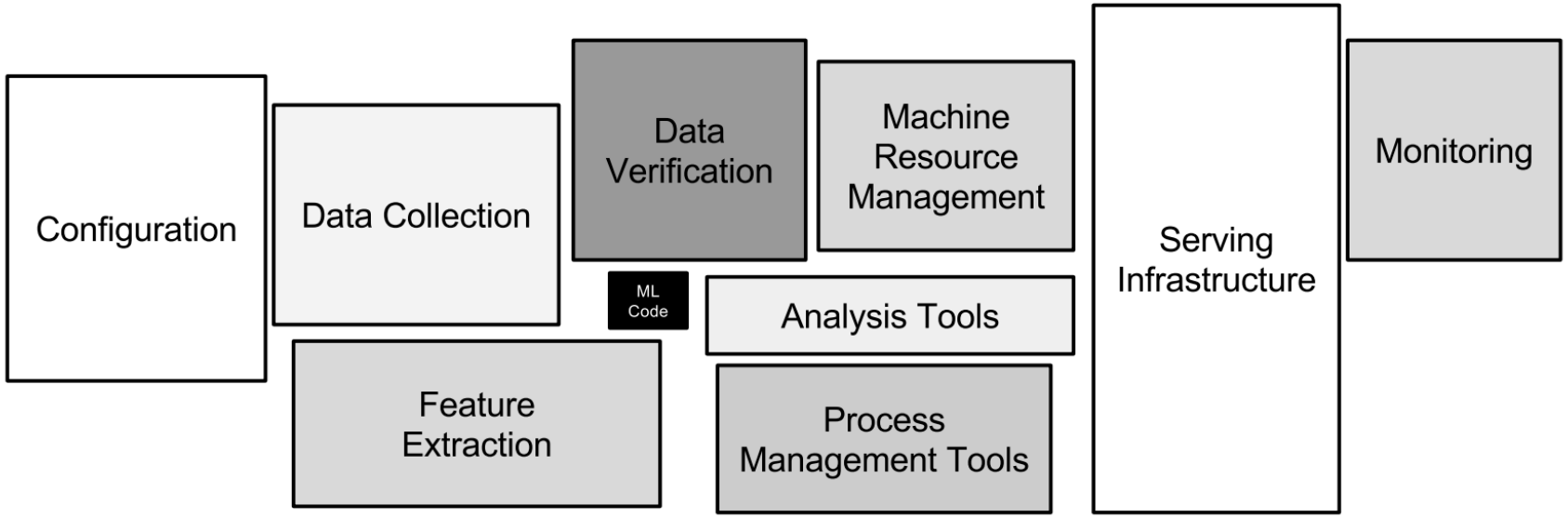



Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

https://proceedings.neurips.cc/paper_files/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf

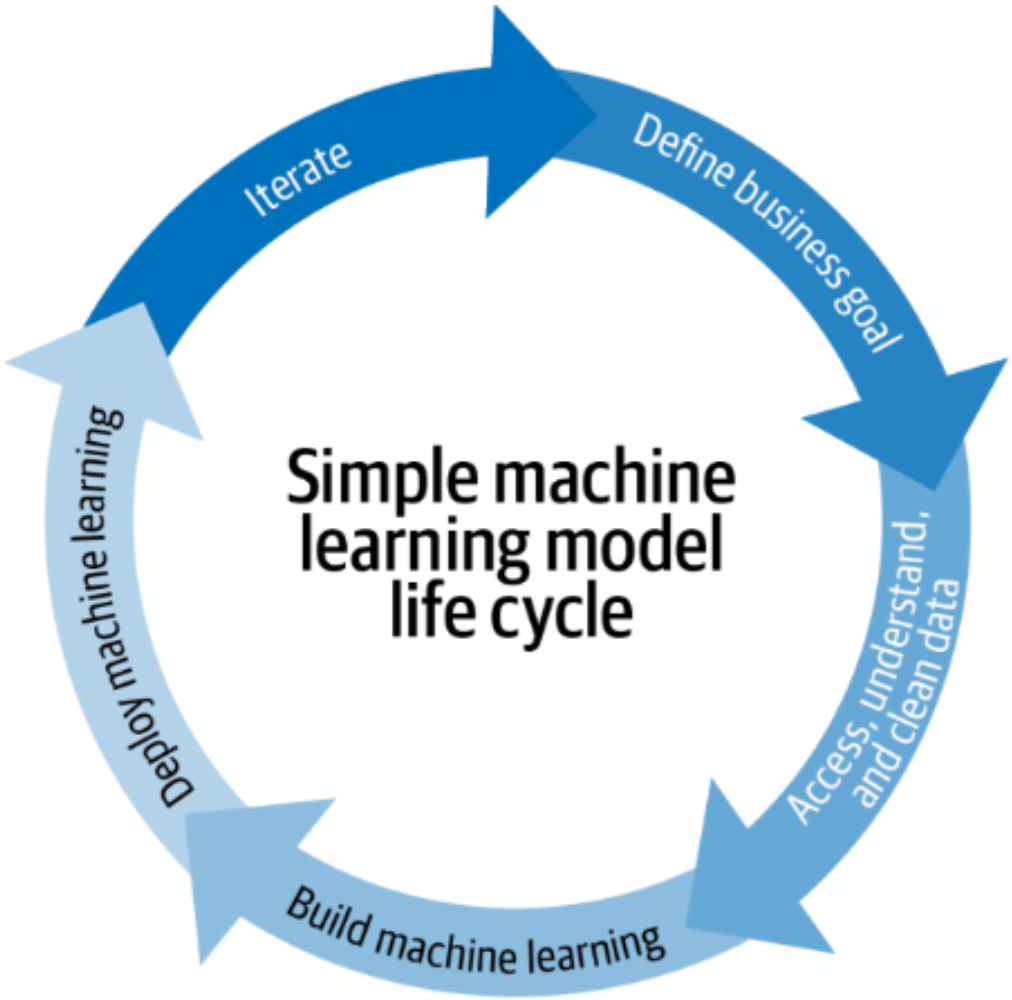
# Что такое MLOps

**MLOps** – расширение DevOps в области ML

**DevOps** - (акроним от англ. development & operations) — методология автоматизации технологических процессов **сборки**, **настройки** и **развёртывания** программного обеспечения.

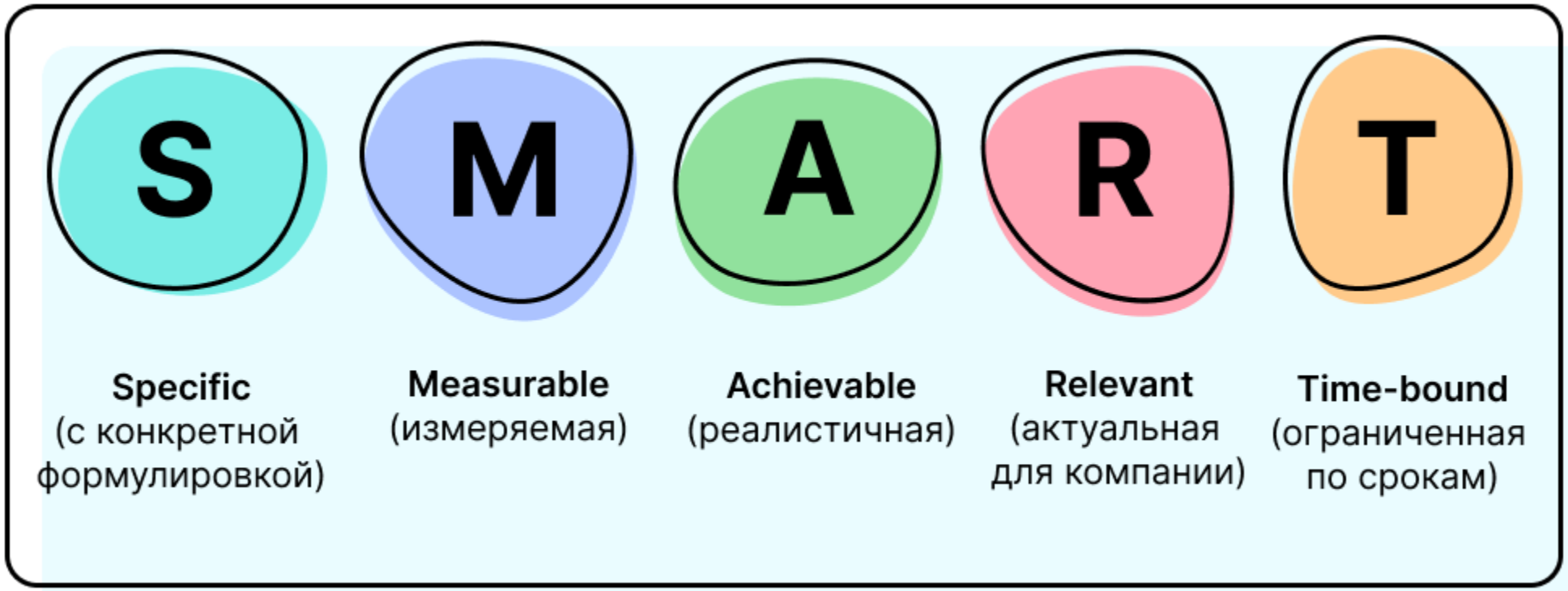
# Этапы жизненного цикла ML модели

1. Постановка цели
2. Работа с данными
3. Тренировка модели
4. Деплой
5. Мониторинг
6. Итерация



# 1. Постановка цели

Например, **увеличить конверсию в звонках с 10% до 15%.**

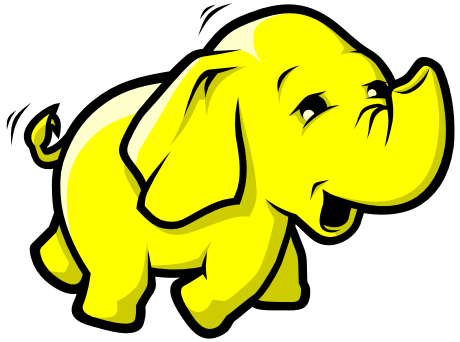


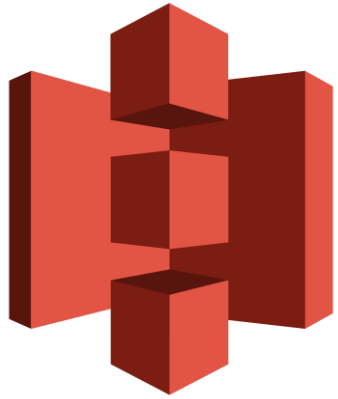
# 2. Работа с данными

Зачем нам данные:

1. Учим на них модели
2. Ищем инсайты
3. Делаем предсказания
4. Оцениваем результаты

Где они могут храниться?



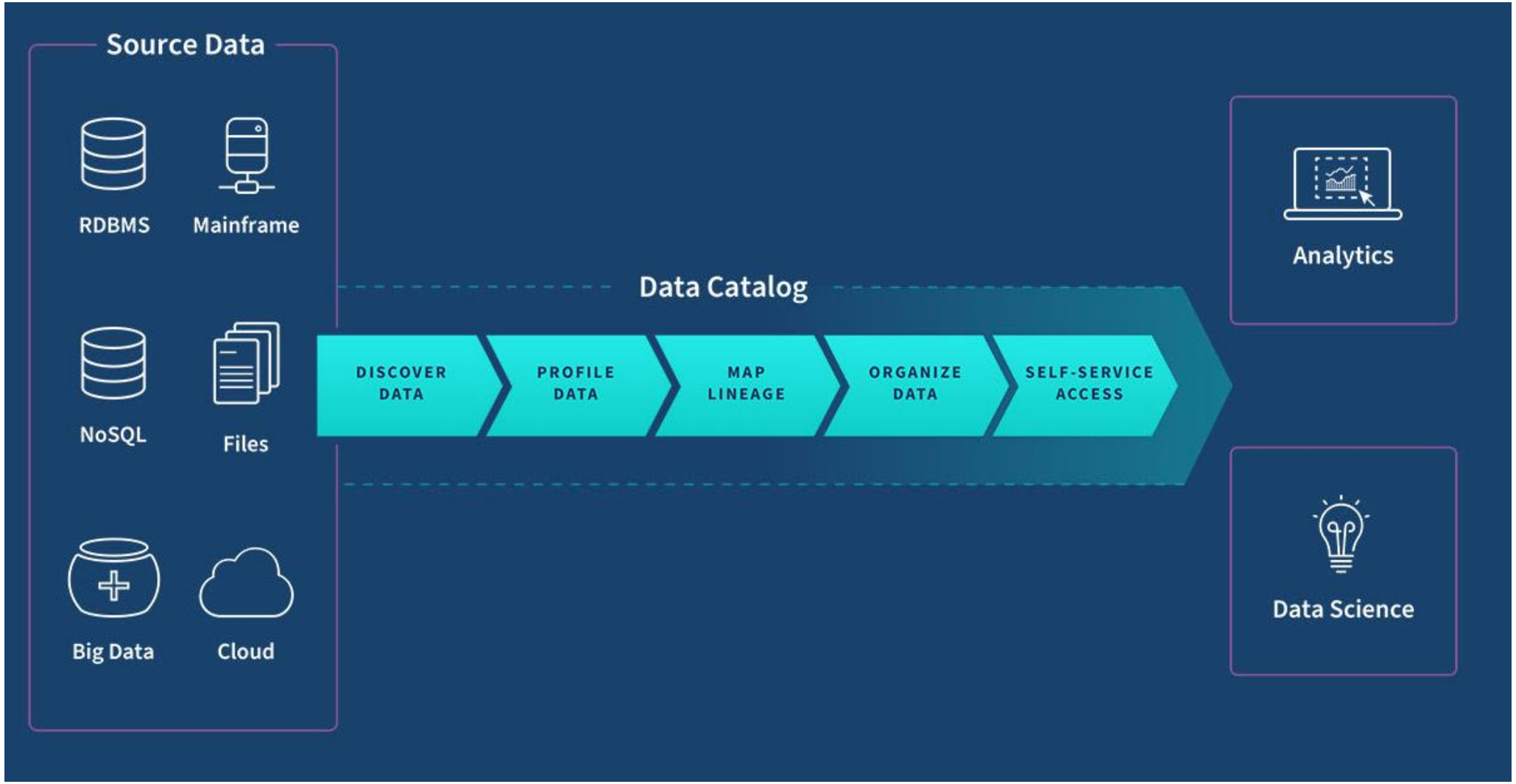



# 2. Работа с данными

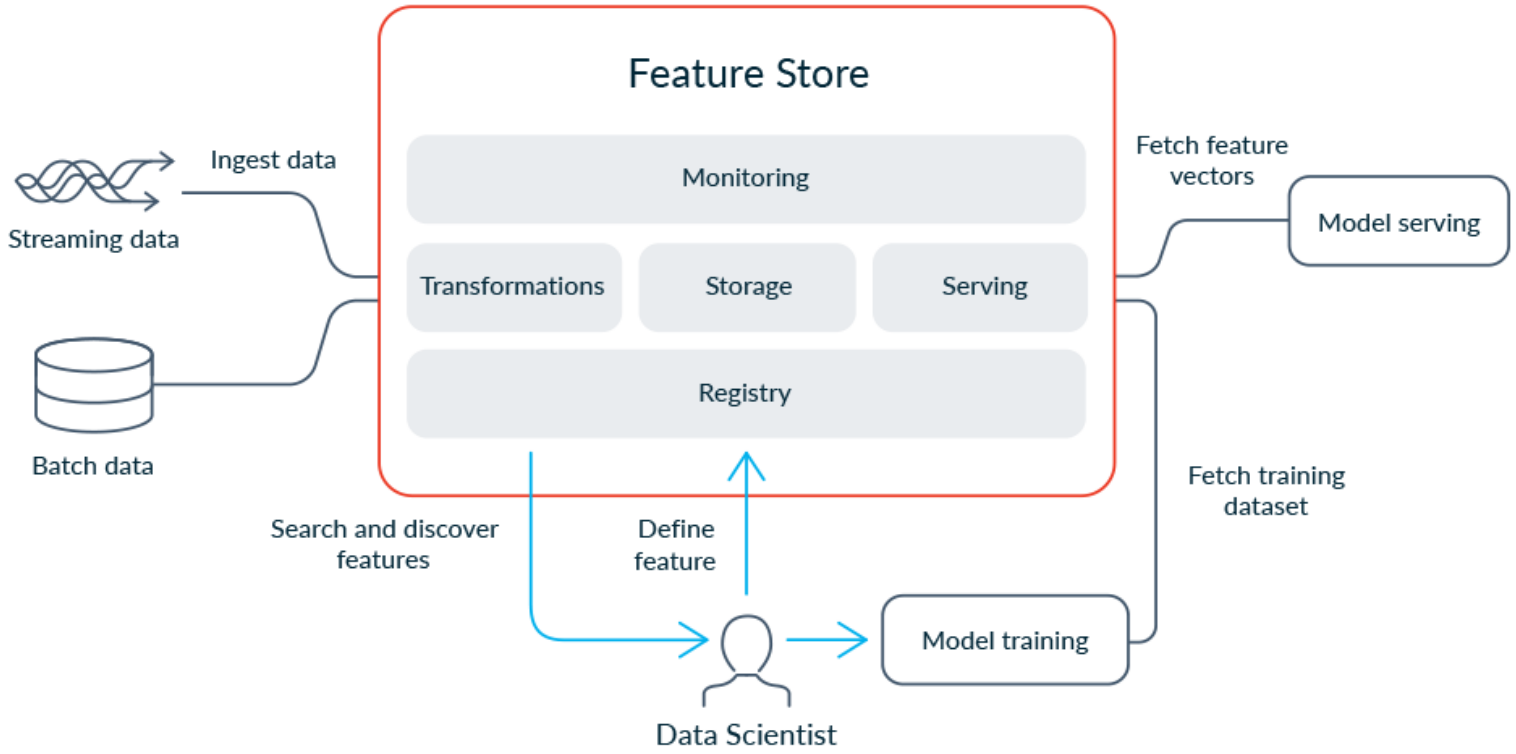
Вопросы к данным:

1. Какие релевантные данные доступны?
2. Достаточно ли точны и надежны данные?
3. Как получить доступ к этим данным?
4. Какие фичи можно будет получить при джоине?
5. Часто ли обновляются данные?
6. Можно ли получить эти фичи в реалтайм?

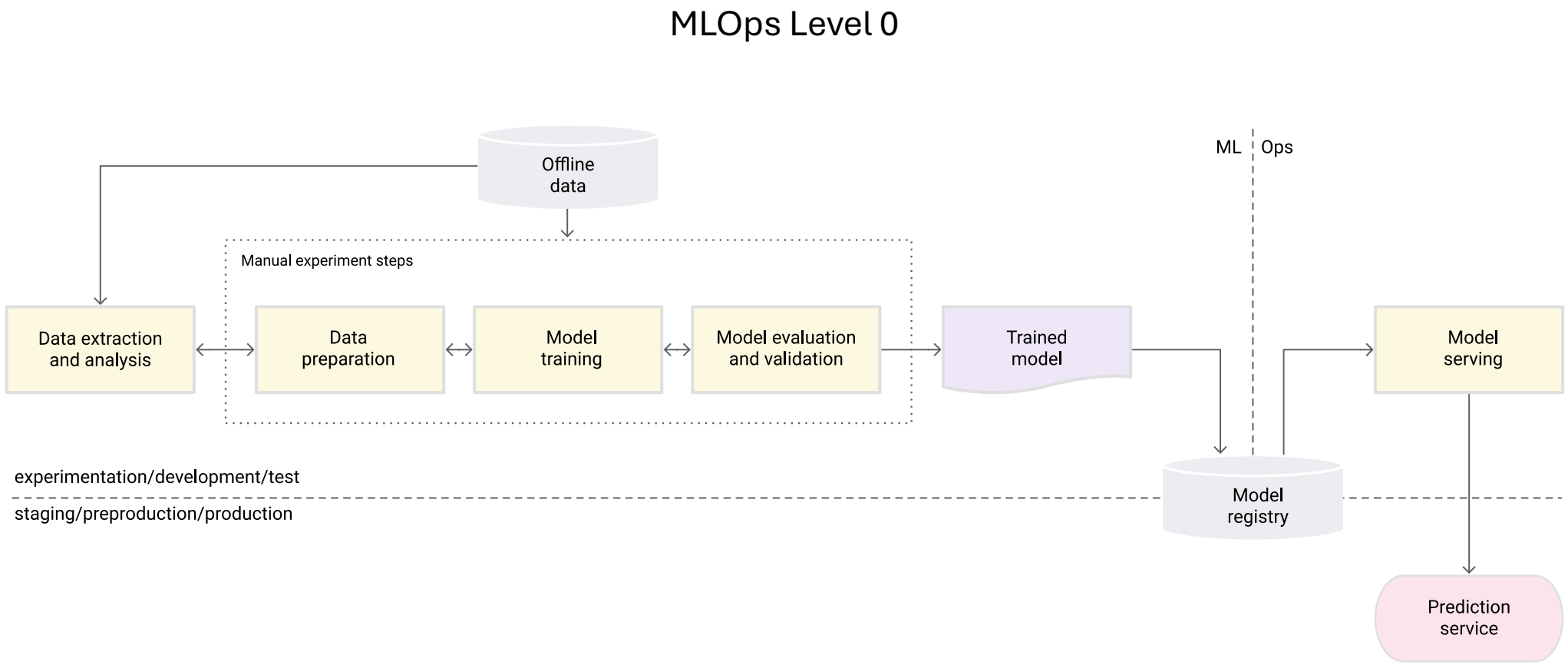
# Data Catalog & Feature Store



# Data Catalog & Feature Store



# 3. Тренировка моделей

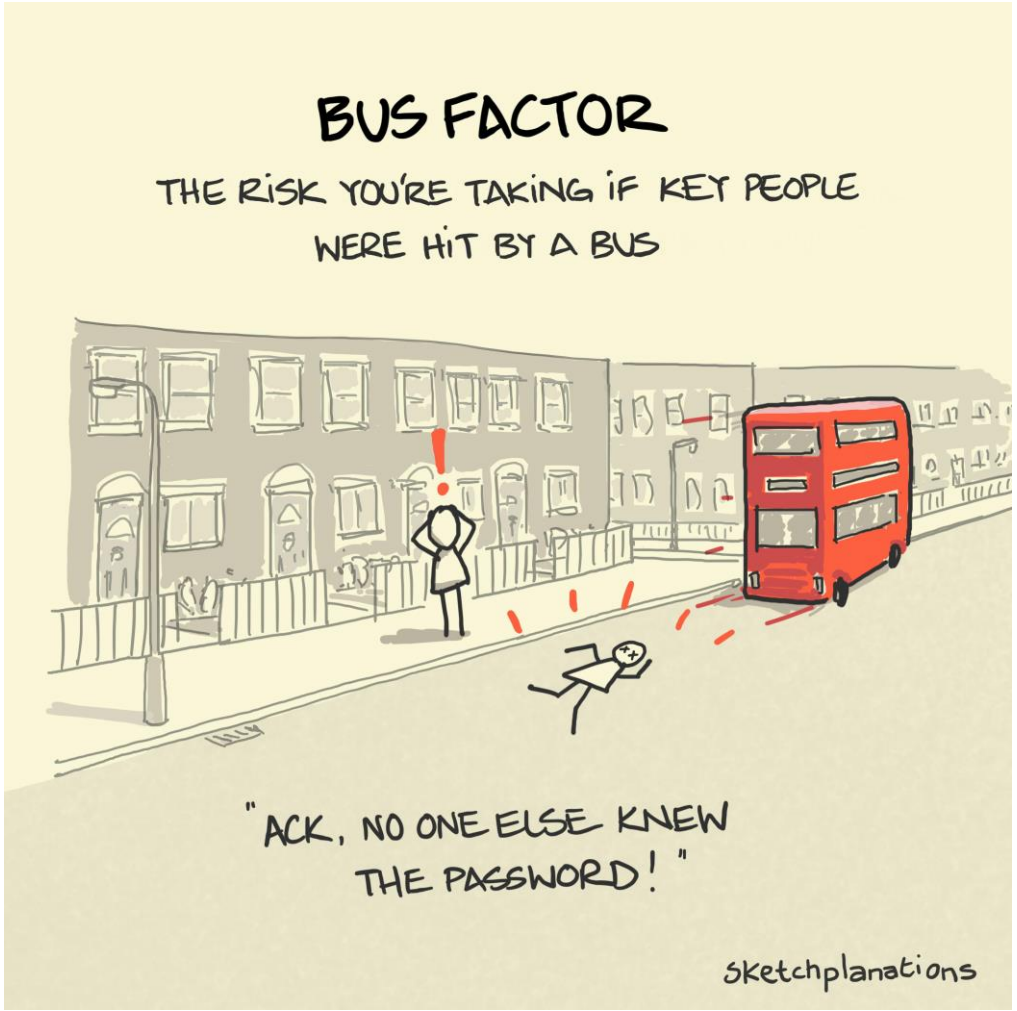



https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning

# 3. Тренировка моделей. Bus factor

Проблемы:

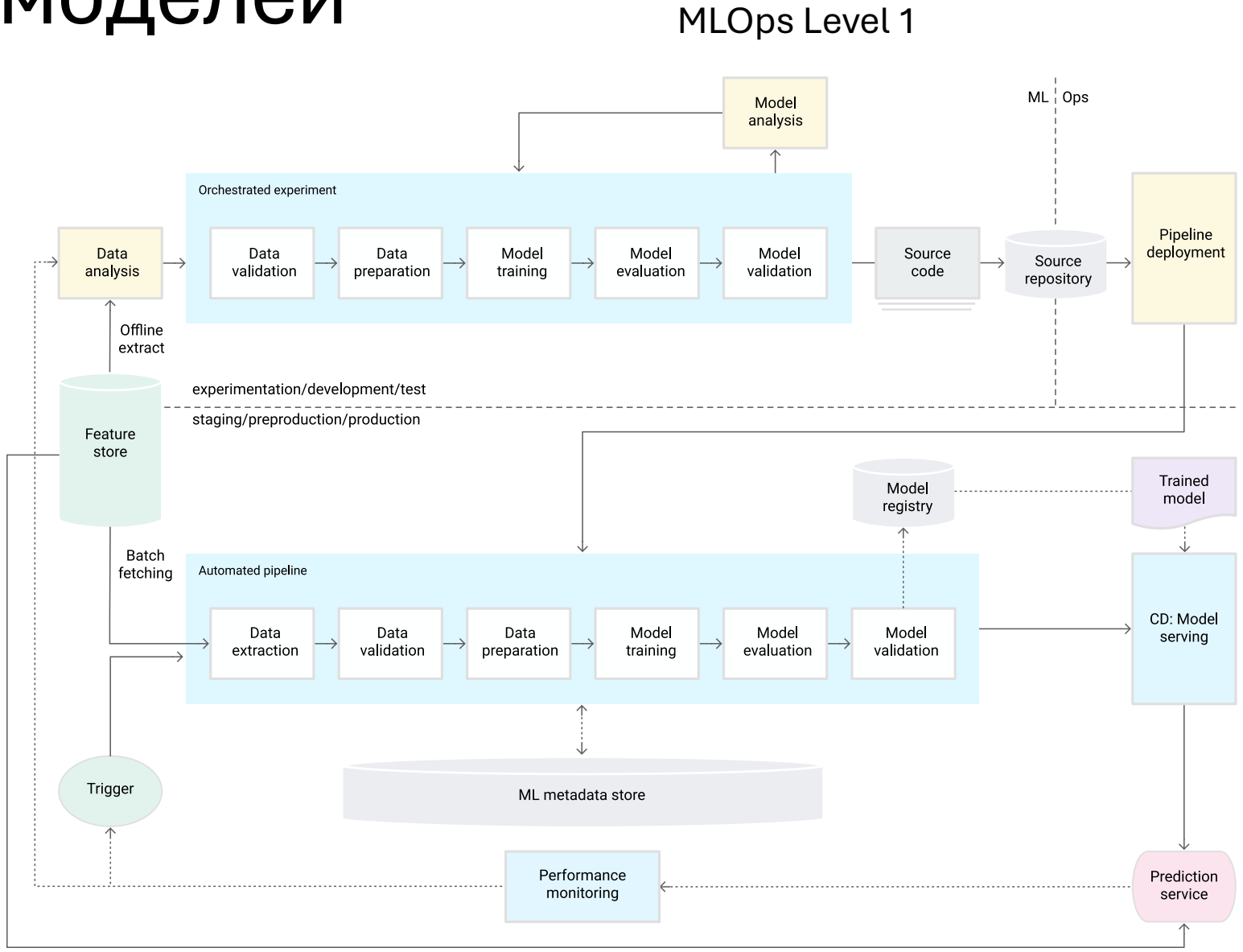
1. Потери данных
2. Потери кода
3. Потери знаний об обучении моделей
4. Как выкатывали в прод?



# 3. Тренировка моделей

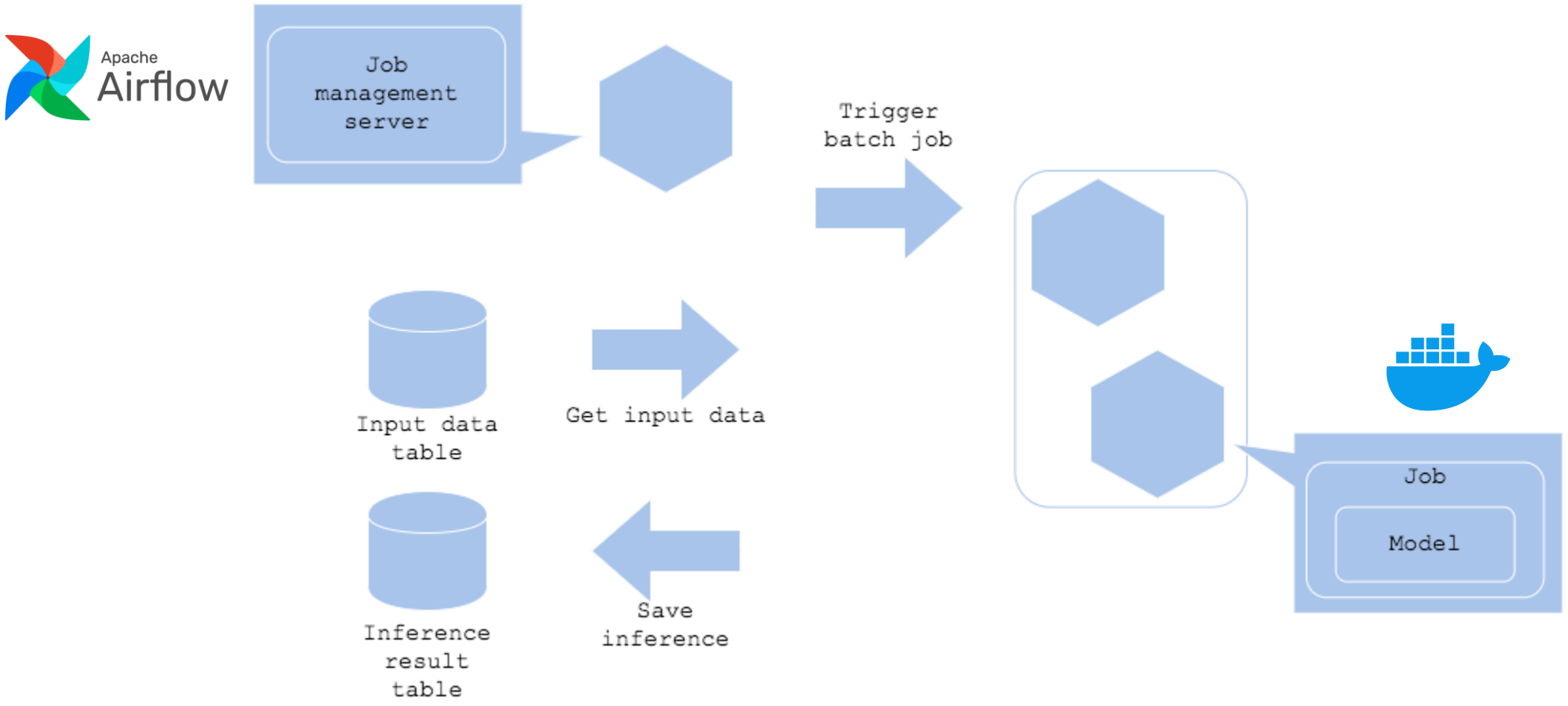
Пишем **переиспользуемый, версионируемый, воспроизводимый** и **тестируемый** код

**Модель** – это **не** цель и **не** финал МЛ процесса

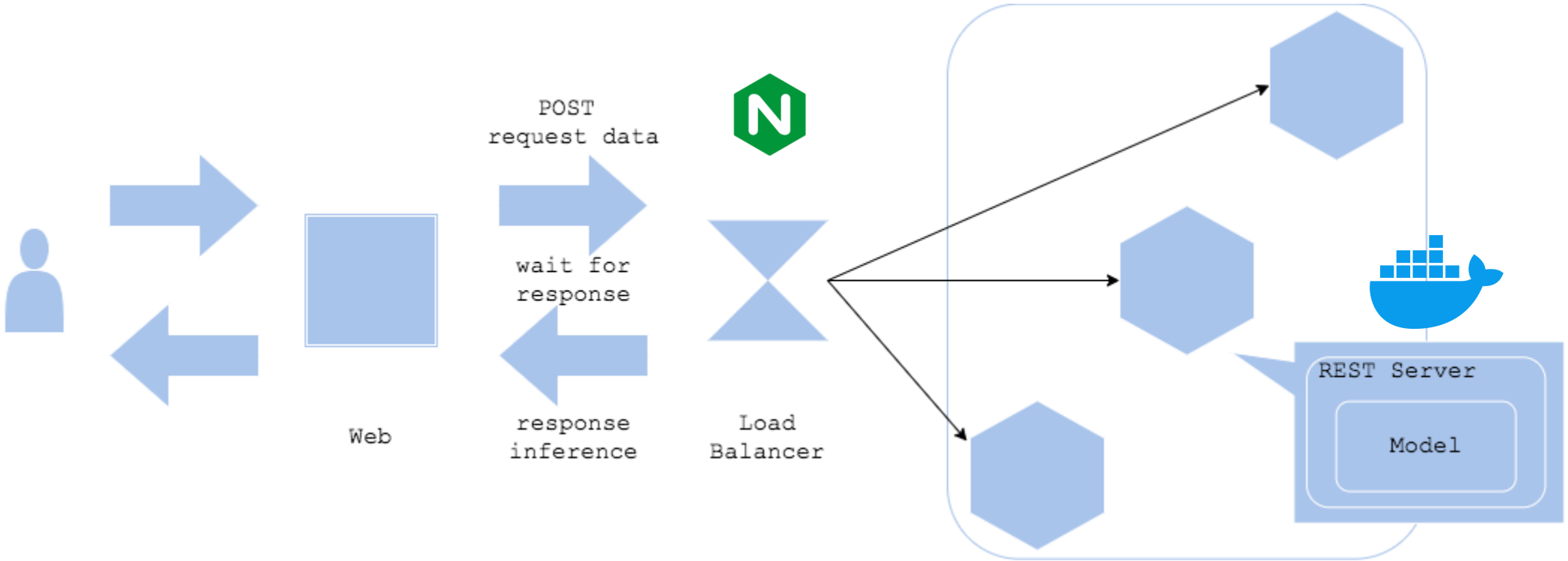


# 4. Деплой

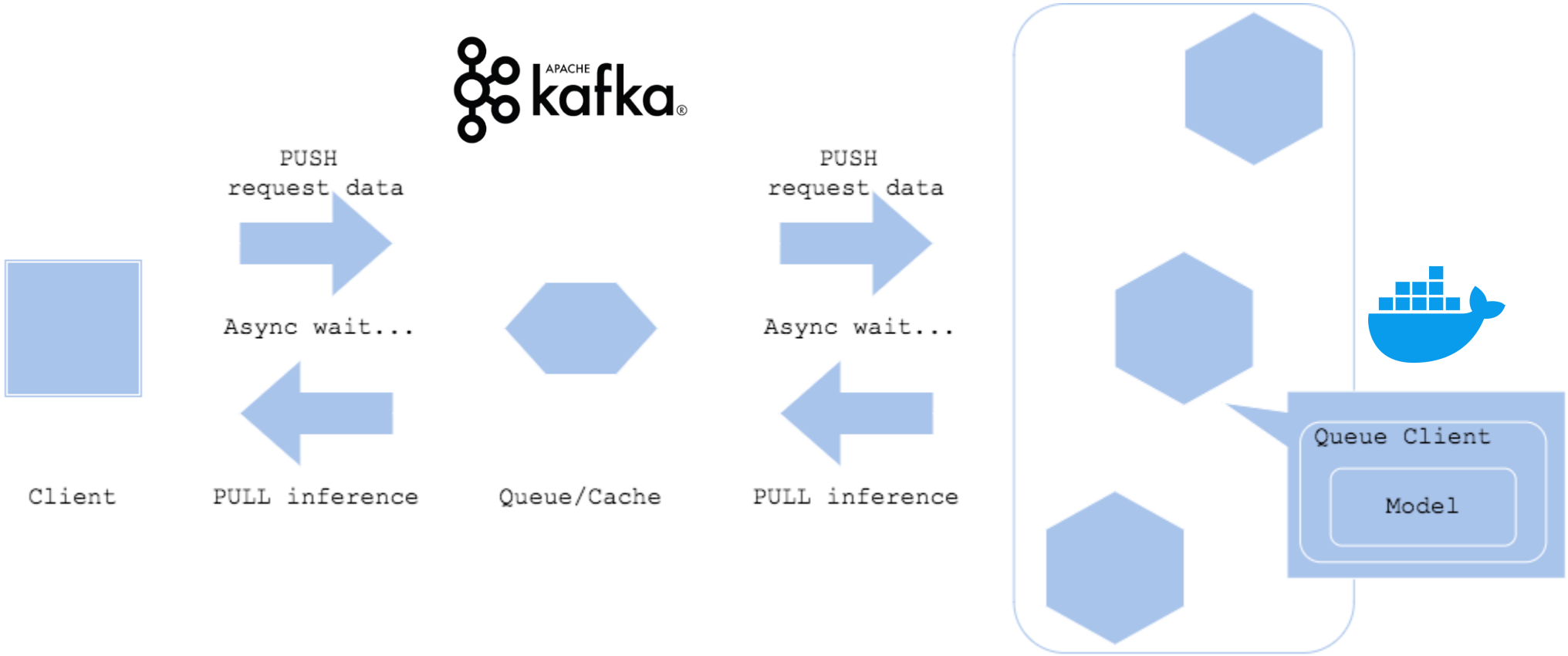
# 4. Деплой. Batch Inference



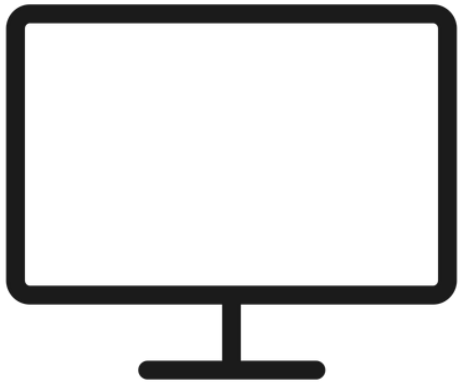
# 4. Деплой. Synchronous pattern (Realtime)

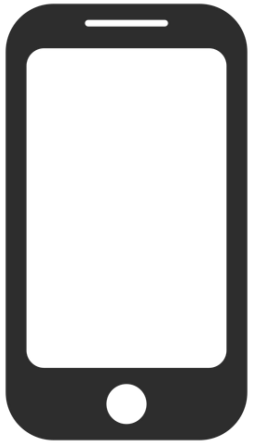


# 4. Деплой. Asynchronous pattern (Near Realtime)



# 4. Деплой. Куда деплоить?

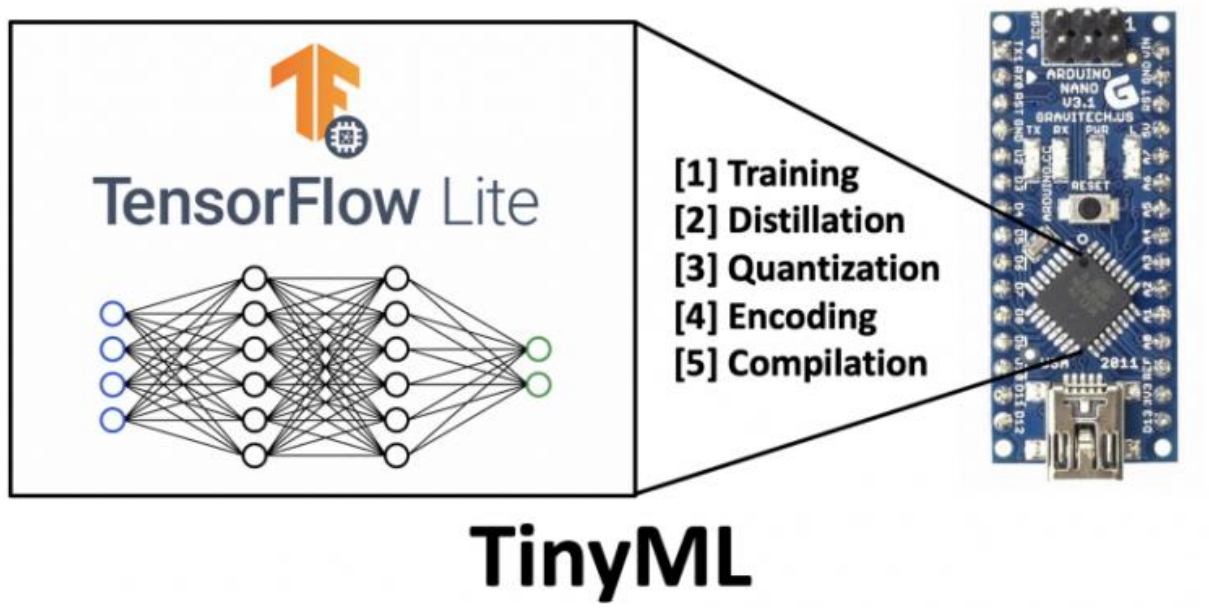





# 4. Деплой. Embedded model

Доставлять модель напрямую на устройство пользователя

+ Нет сетевой задержки
+ Можно запустить на девайсе
- Сложно масштабировать
- Устройства все еще могут быть не такими мощными



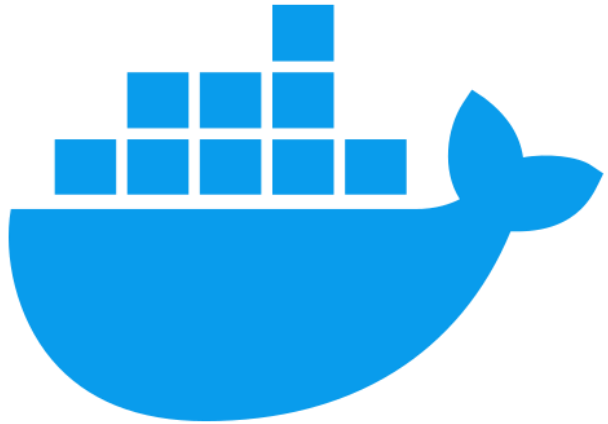
# 4. Деплой. Model as a service

Модель представляет собой
отдельный сервис

+ Можно масштабировать

+ Можно запускать сложные модели

- Есть сетевая задержки

# 4. Деплой. Model as a code

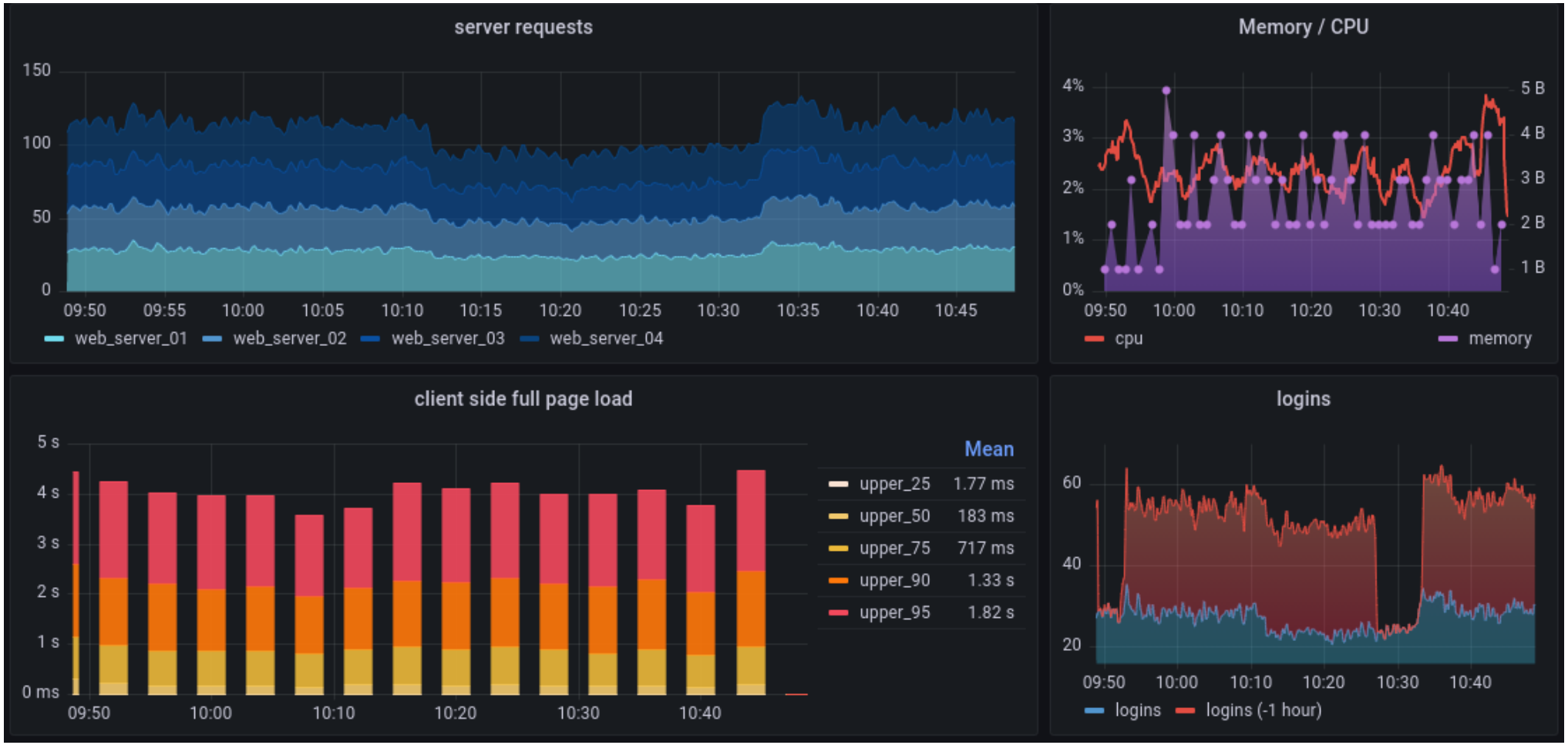
Модель представляет собой код

+ Просто реализовать

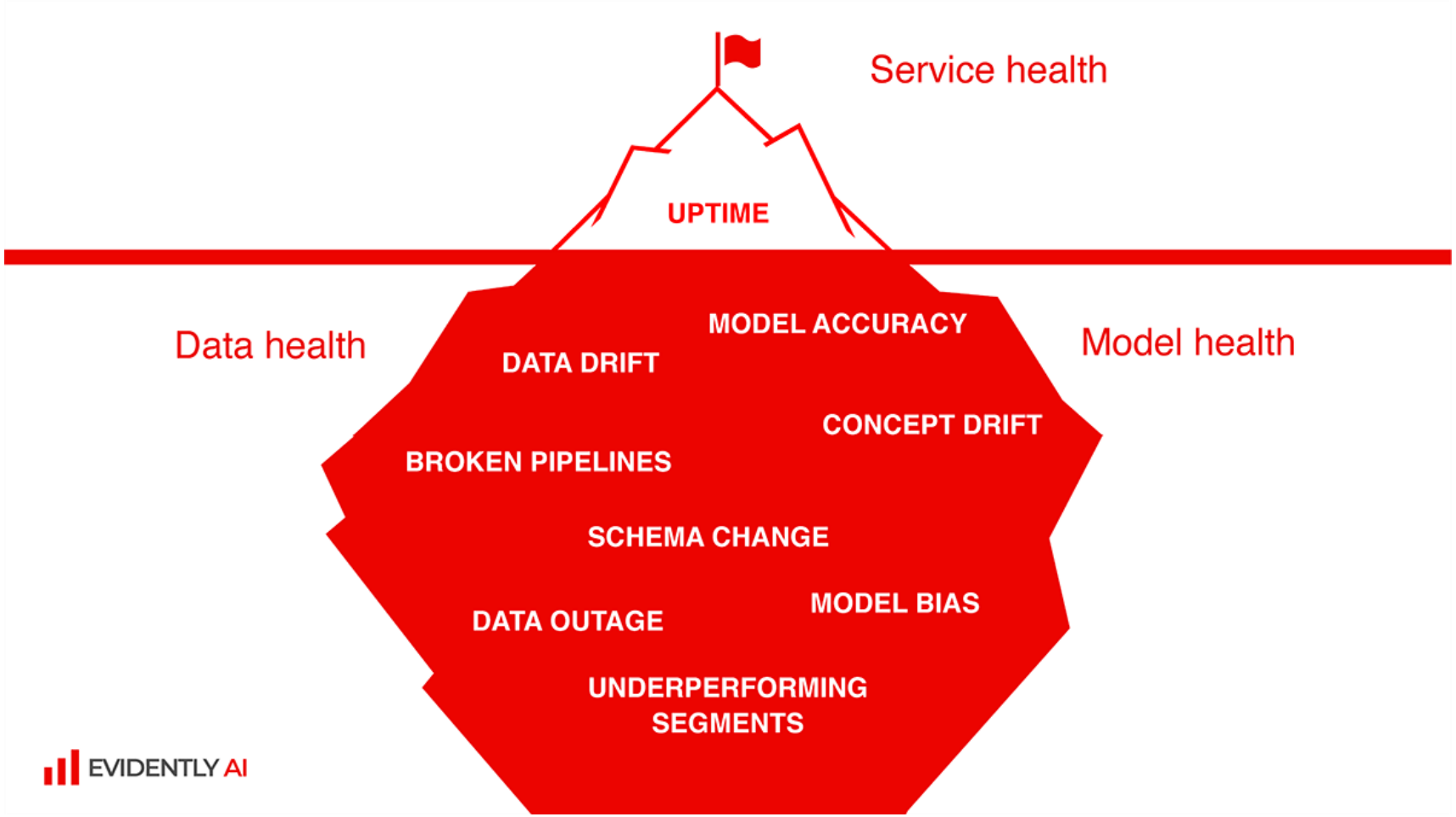
- Невозможно масштабировать
- Работает только с простыми моделями

```sql
SELECT id,
       feature1,
       feature2,
       feature3,
       1 / (1 + EXP(-(
        (SELECT weight FROM model_weights WHERE feature_name = 'feature1') *
        feature1 +
        (SELECT weight FROM model_weights WHERE feature_name = 'feature2') *
        feature2 +
        (SELECT weight FROM model_weights WHERE feature_name = 'feature3') *
        feature3
        ))) AS predicted_probability
  FROM input_data;
```

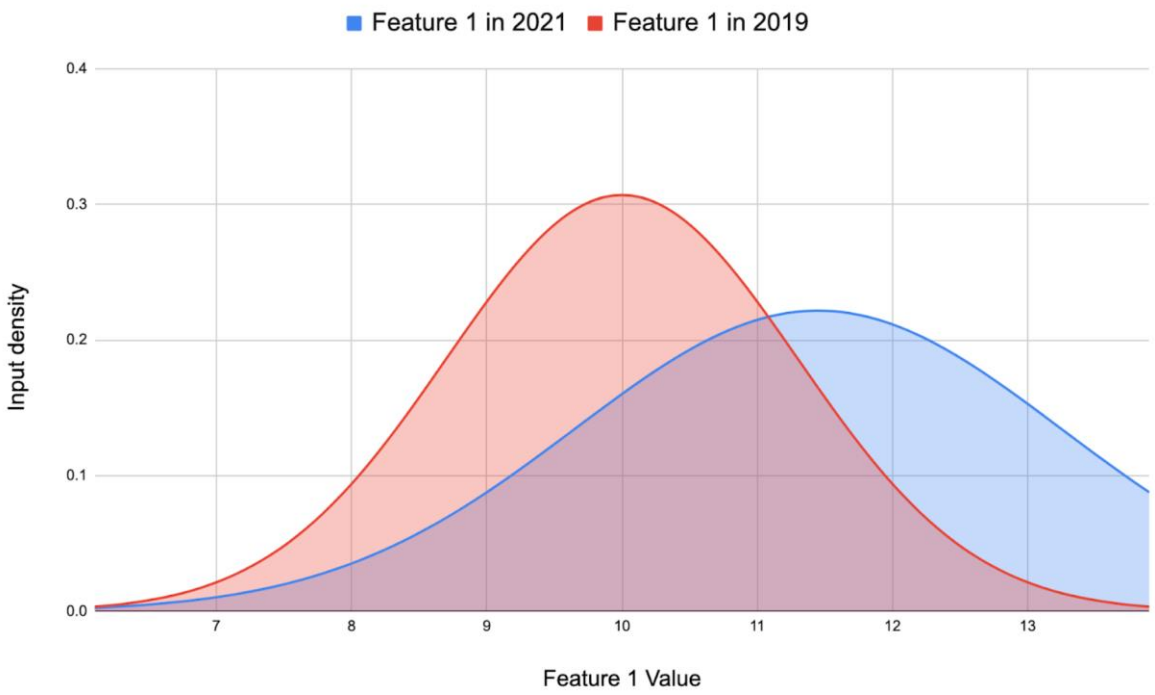
# 5. Мониторинг. Технические метрики

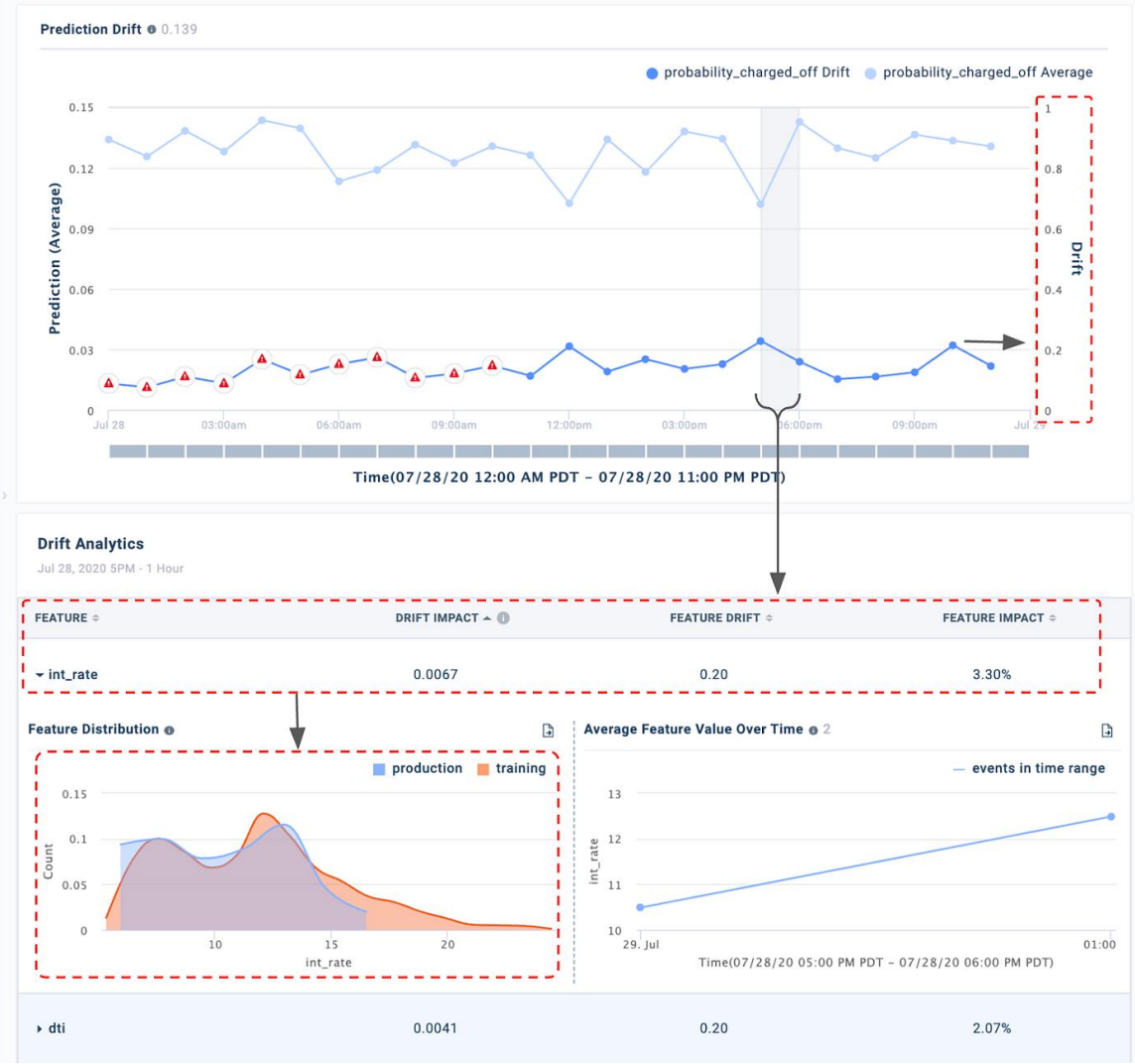


# 5. Мониторинг. Технические + МЛ метрики



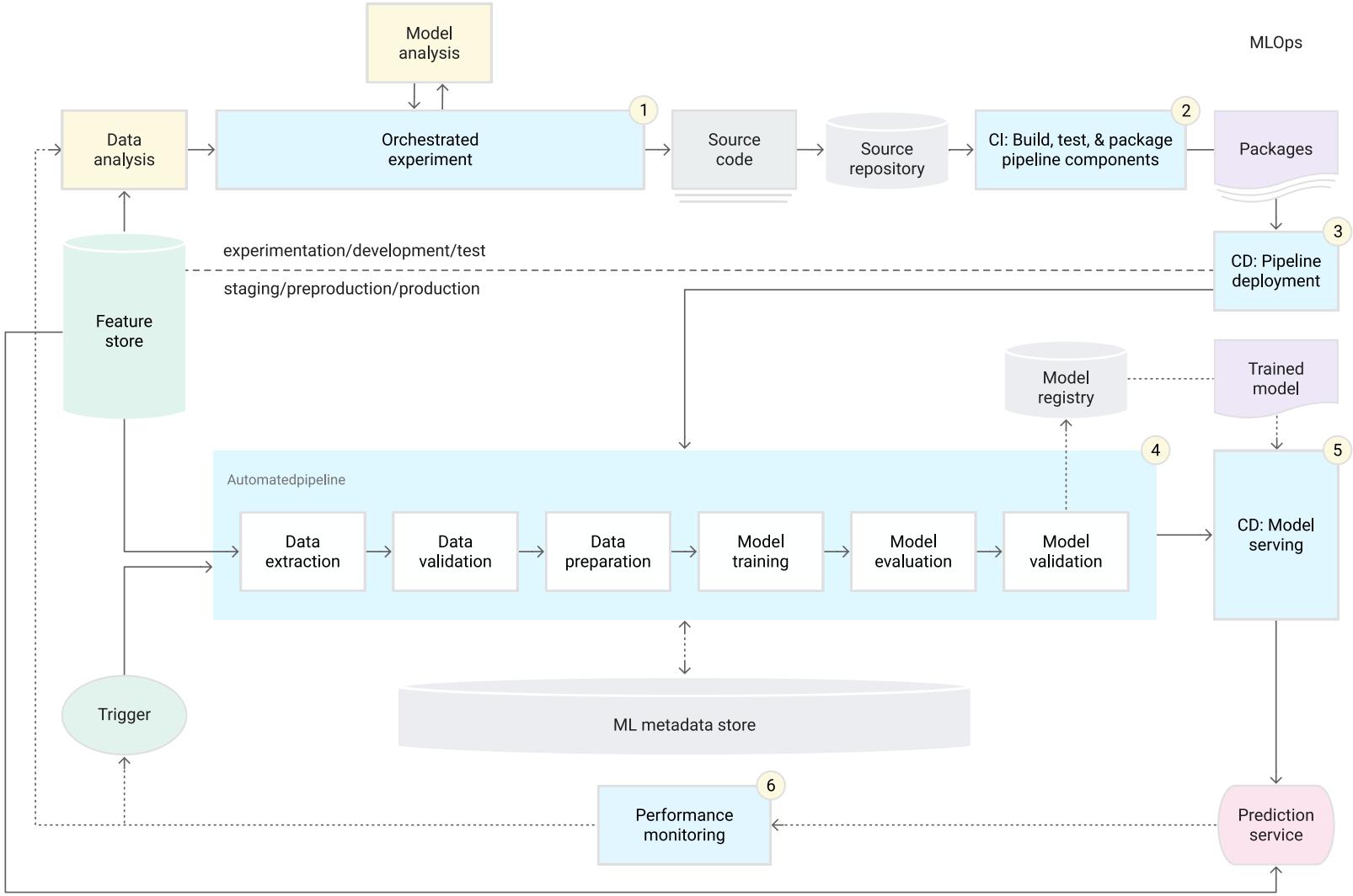
# 5. Мониторинг. МЛ метрики



# 6. Итерация

MLOps Level 2



# 6. Итерация. ML CI/CD