



MLOps

MLFlow и переобучение ML-моделей

• REC

Проверить, идет ли запись

Напишите «+» в чат, если меня
слышно и видно



Тема открытого урока



MLFlow и переобучение моделей

Емельянов Петр



CEO at Bloomtech, R&D Director at Ubic

Эксперт в кибербезе и *privacy-preserving ML*, много знаю про IT-менеджмент, еще больше – про технологии.

- 20 лет опыта в IT;
- 10+ успешных проектов для крупного бизнеса и государства;
- Bloomtech – технологический лидер в области конфиденциальных вычислений;
- преподаватель курса MLOps в ОТУС



<https://www.linkedin.com/in/emelianovpeter/>

Правила вебинара



Активно
участвуем



Задаем вопрос
в чат



Вопросы вижу в чате,
могу ответить не сразу

Маршрут вебинара

1. Знакомство

2. Об ОТУС

3. Воспроизводимость,
переобучение, MLFlow

4. Демо/Практика

5. Команда курса

6. О курсе, программа обучения

7. Бонус: карьерная информация

8. Рефлексия

Расскажите о себе

- Как вас зовут? Откуда вы?
- Ваш опыт работы в IT?
- С какой основной целью вы записались на занятие?



06 ОТУС

О компании



Сфера

ОТУС специализируется на обучении в ИТ.

Наша фишка – продвинутые программы для специалистов с опытом и быстрый запуск курсов по новым технологиям.



Образовательная лицензия

У OTUS есть образовательная лицензия, поэтому вы сможете получить удостоверение о повышении квалификации или диплом о профессиональной переподготовке, а также сделать налоговый вычет.



Направления курсов

Обучение специалистов разных грейдов: junior, middle, senior, lead



- Программирование
- Инфраструктура
- Тестирование
- Аналитика



- Data Science
- Управление
- GameDev
- Информационная безопасность

Мы в цифрах

170+

курсов для junior, middle,
senior специалистов

600+

преподавателей делятся знаниями
и реальными кейсами

7

лет со дня основания
компании

38 000+

выпускников уже
прошли обучение

500 000+

ИТ-специалистов в сообществе, читают
материалы, учатся и общаются
на наших площадках



MLFlow и переобучение моделей

Цели вебинара



После занятия вы сможете

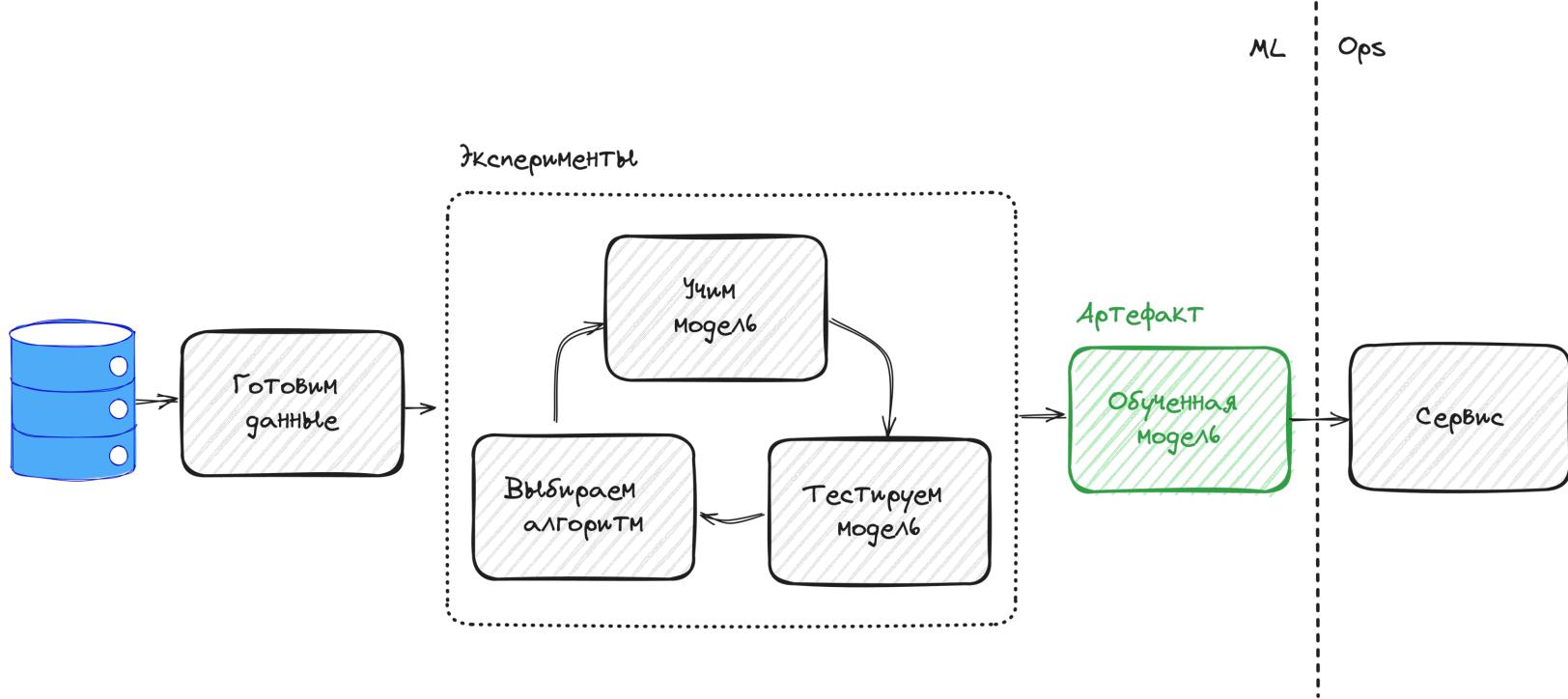
- 1.** Определить сценарии применения MLFlow, разобраться, как этот инструмент помогает в переобучении моделей;

- 2.** Понять, как MLFlow работает в облаке, используя объектное хранилище (S3) для хранения артефактов;

- 3.** Управлять экспериментами, версионировать артефакты моделей;

Воспроизводимость и переобучение

Как мы разрабатываем модели



Что тут не так

1

Много ручного труда

Долго и унизительно, контекст утрачивается очень быстро;

2

Ничего не версионируется

Ни данные, ни результаты, ни модели (веса, иногда даже код);

3

Разрыв между ML и Ops

Разные стеки, трудности перевода;

4

Разрыв в мониторинге

Инфраструктура мониторится (возможно), но если деградировала модель – у нас беда;

5

Локальные эксперименты

Разные окружения, бардак;

6

Разрыв между ML и ML

Нет нормальной коммуникации, обмен результатами в почте/чатиках (бардак);

7

Разрыв в тестировании

Модель и инфраструктура тестируются отдельно;

Кризис воспроизводимости*

70%

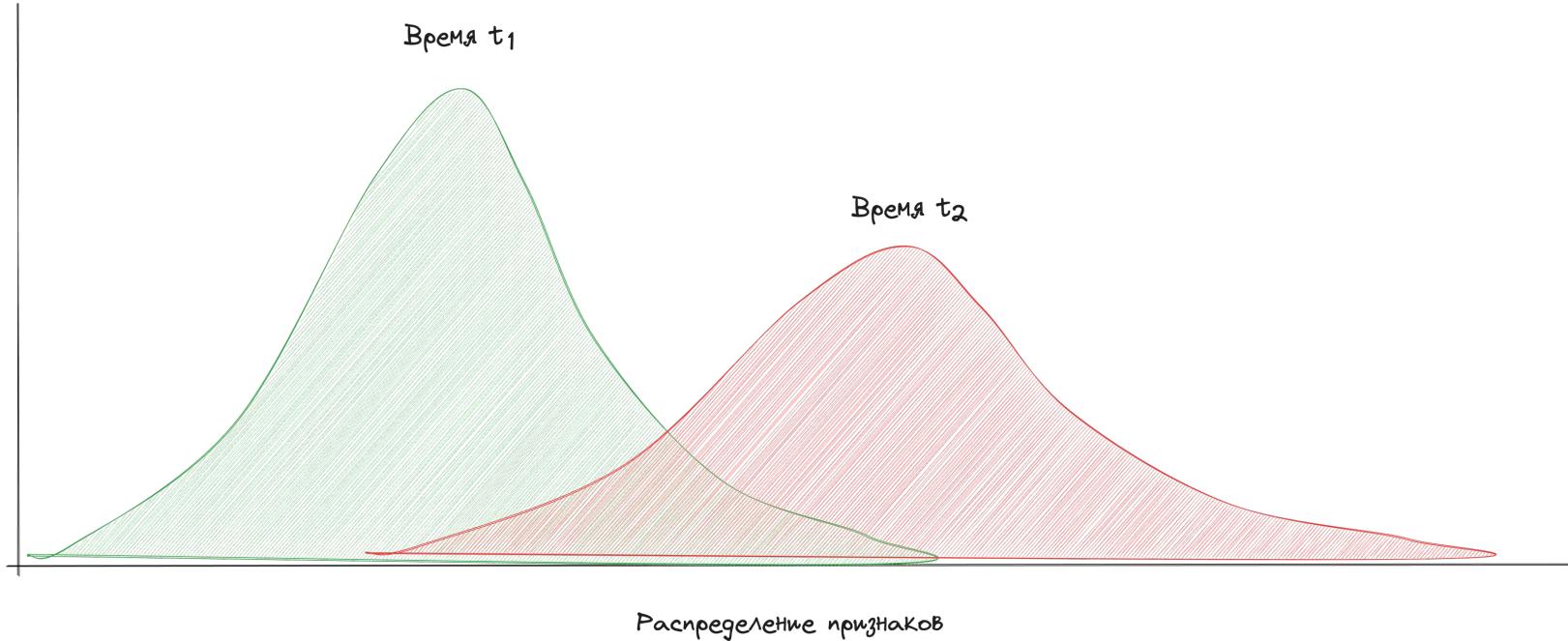
исследователей не смогли воспроизвести результаты других исследователей

>50%

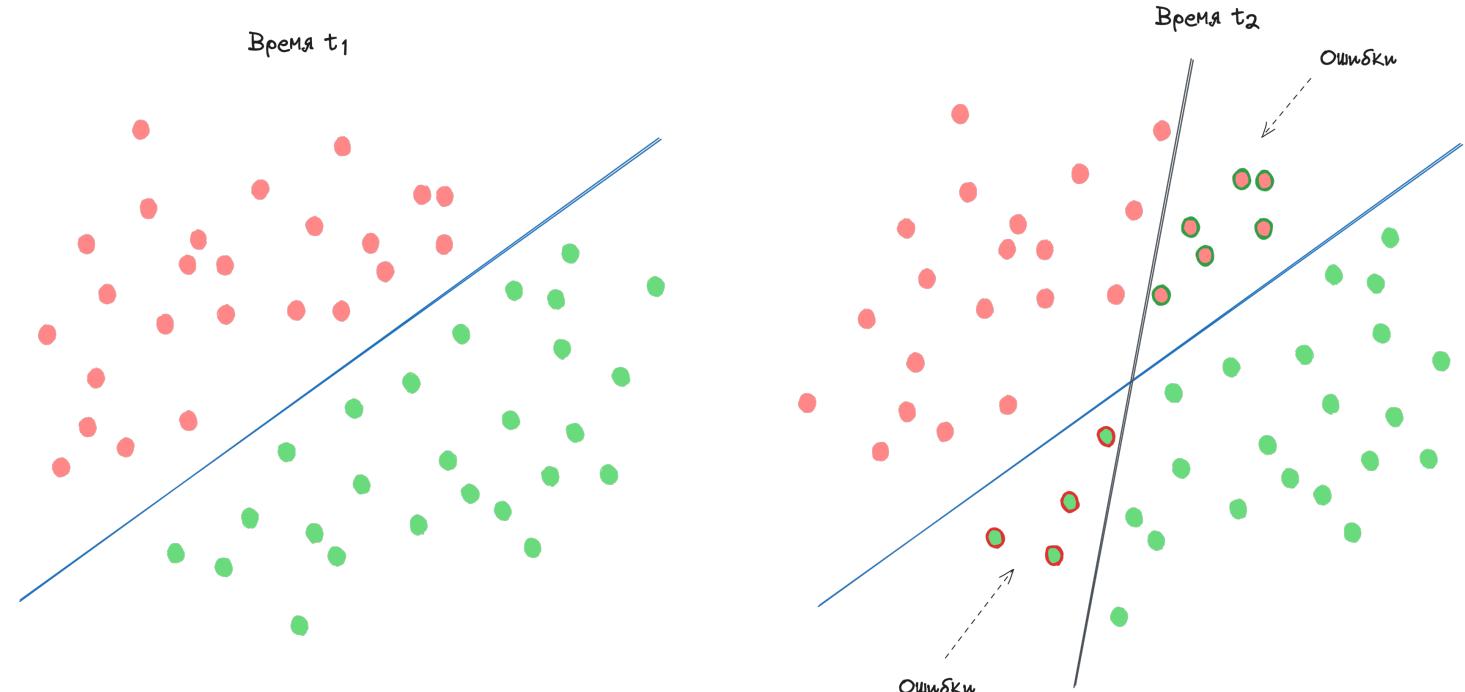
исследователей не смогли воспроизвести результаты собственных исследований

*Nature's survey (2016)

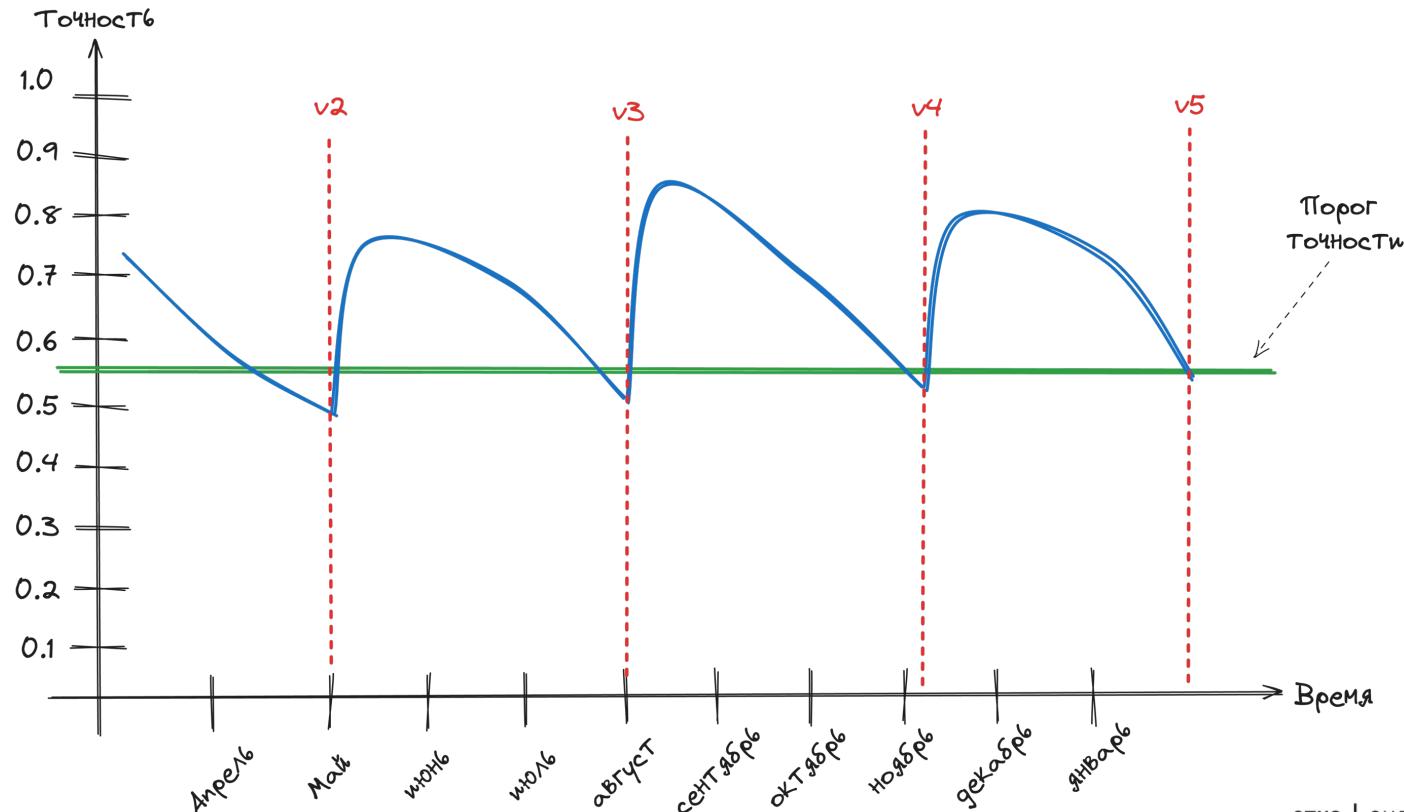
Data Drift



Concept Drift



Регулярное переобучение



Вопросы?



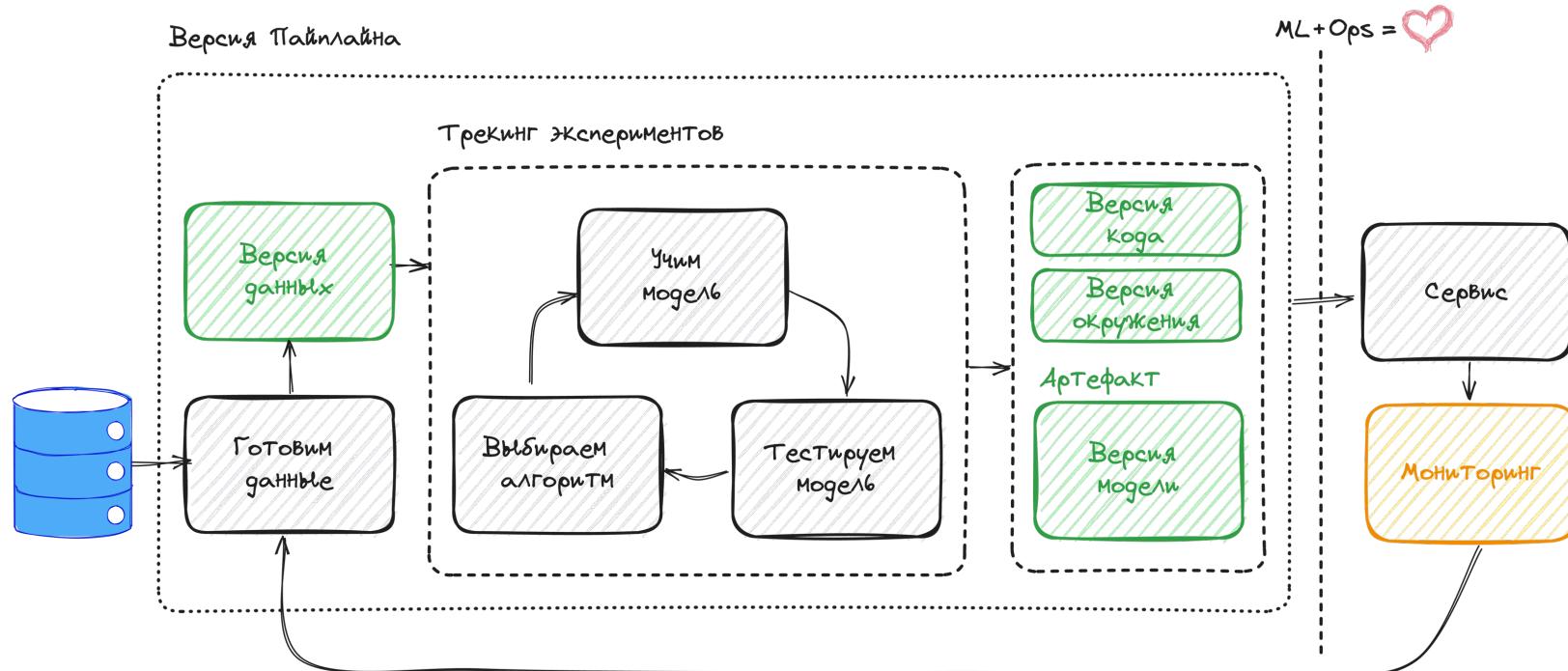
Задаем
вопросы в чат



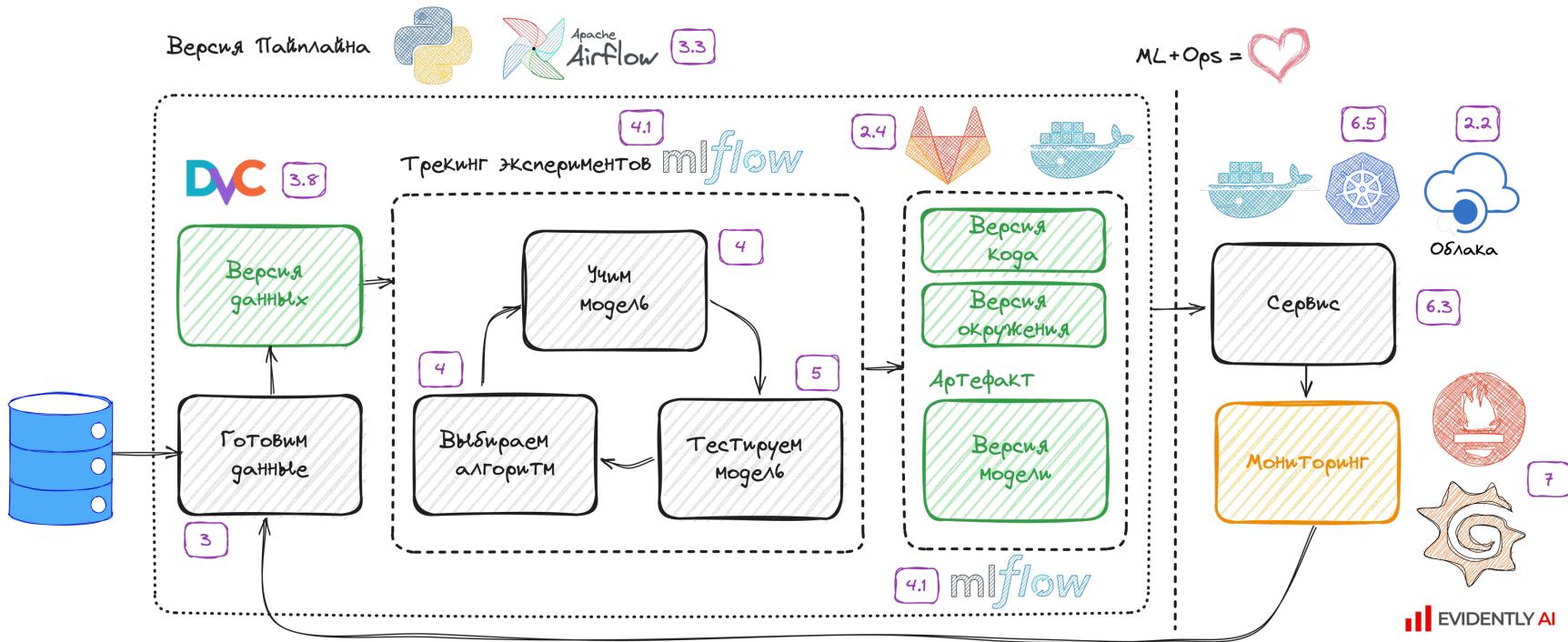
Ставим “-”,
если вопросов нет

Переобучение – задача автоматизации

Как должно быть (примерно)



Инструменты



Вопросы?



Задаем
вопросы в чат



Ставим “-”,
если вопросов нет

MLFlow

Что это/из чего состоит



MLFlow – это open-source платформа, предназначенная для управления жизненным циклом ml-моделей.

mlflow TRACKING

Сохраняйте метаданные о ваших экспериментах: данных, коде, конфигурации и результатах.

mlflow PROJECTS

Сохраняйте DS-код в универсальном формате, который подразумевает воспроизводимость на разных платформах.

mlflow MODEL REGISTRY

Управляйте моделями, их версиями и состоянием в центральном репозитории.

mlflow MODELS

Внедряйте ваши модели в разные среды исполнения.

+MLFlow Recipes

+MLFlow Evaluate

+MLFlow LLM Deployment

+MLFlow Prompt Engineering UI



Сценарии использования



1

Трекинг экспериментов

Храните гиперпараметры, метрики и результаты каждого эксперимента в одном месте; сравнивайте результаты, **улучшайте модели**, сохраняйте их в формате MLFlow Model.



2

Выбор и внедрение модели

Выбирайте лучшую модель, регистрируйте в MLFlow Model Registry, тестируйте, внедряйте, отслеживайте состояние, обучайте и **переобучайтесь**.



3

Мониторинг модели

Вы всегда знаете, какая модель и какое окружение на проде: тестируйте модель в “боевом режиме”, не трогая прод. В случае деградации – **переобучайтесь**.

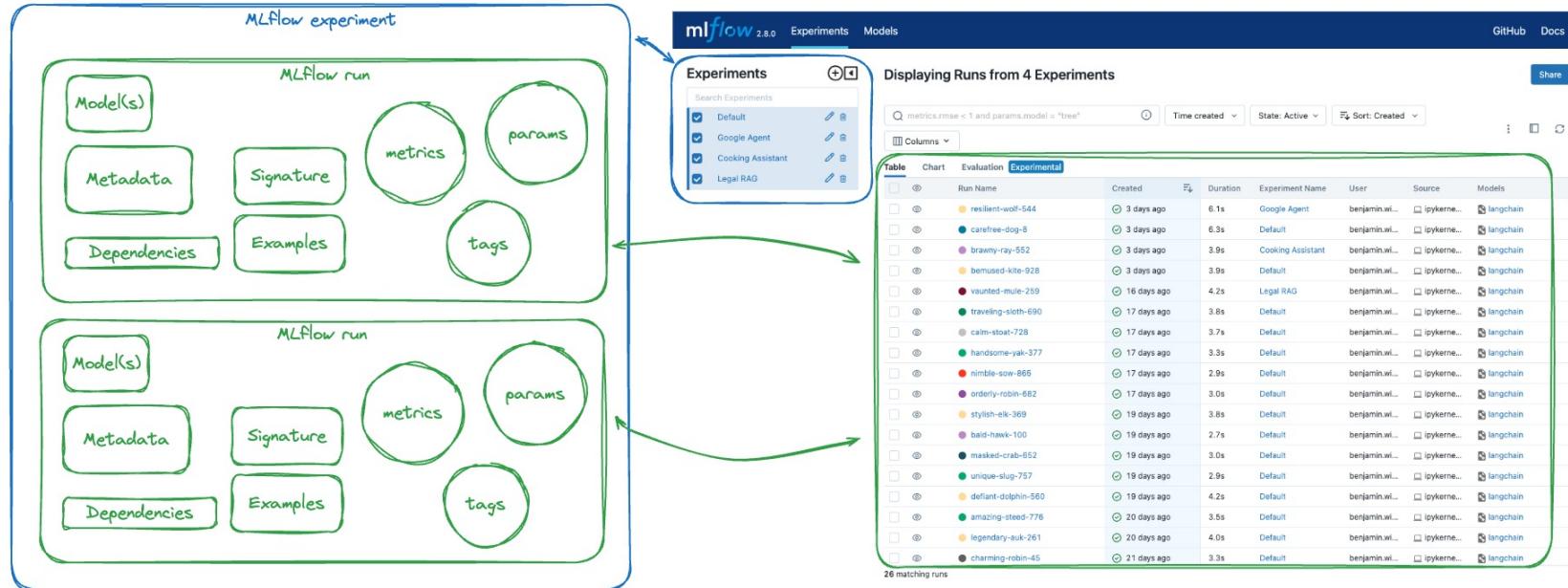


4

Коллаборация

MLFlow – площадка для взаимодействия: ML и ML обмениваются результатами экспериментов, ML и Ops обмениваются знаниями/артефактами.

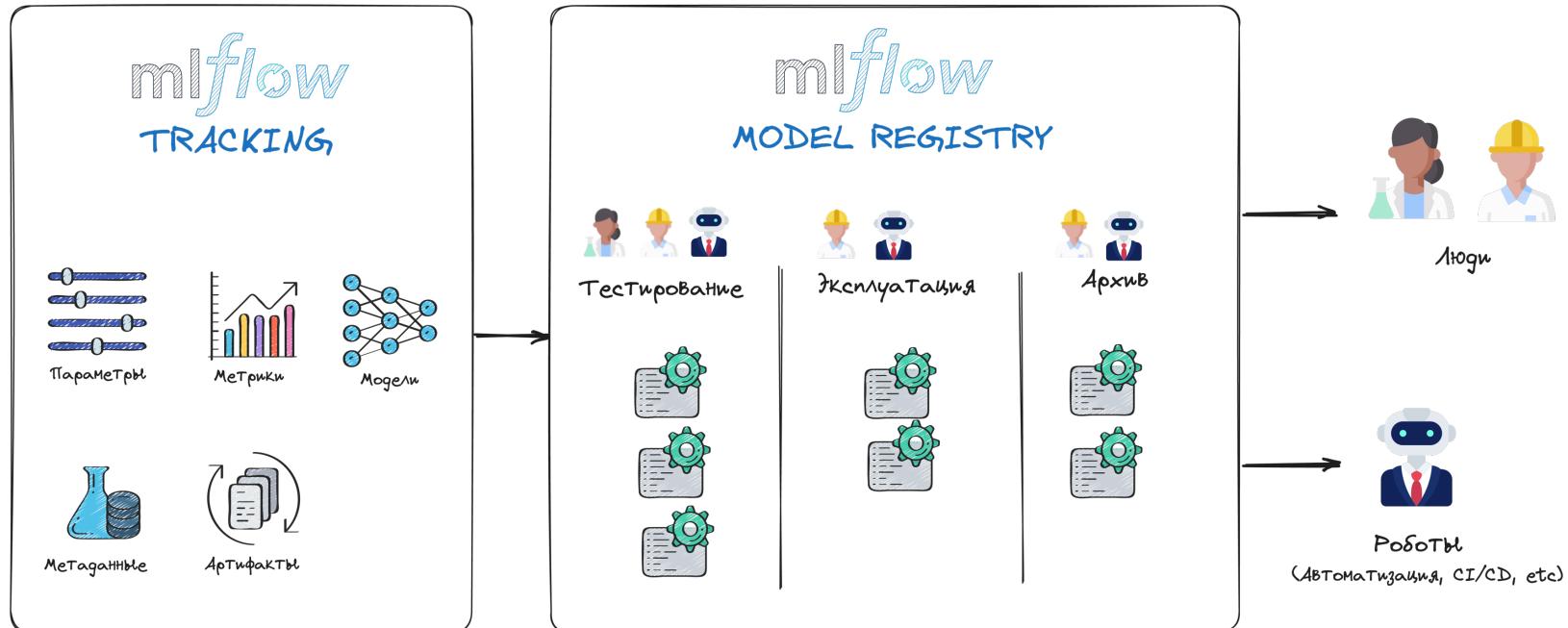
MLFlow Tracking



Источник: [официальная документация](#)



MLFlow Model Registry



Вопросы?



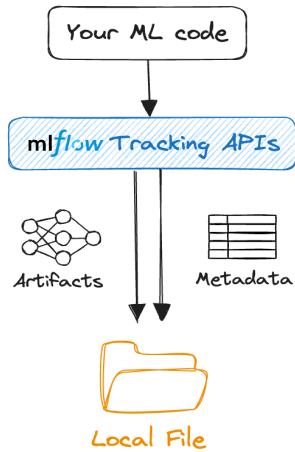
Задаем
вопросы в чат



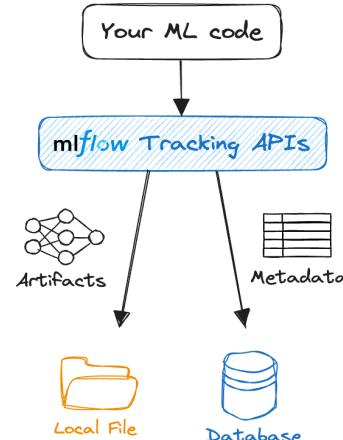
Ставим “-”,
если вопросов нет

HLD

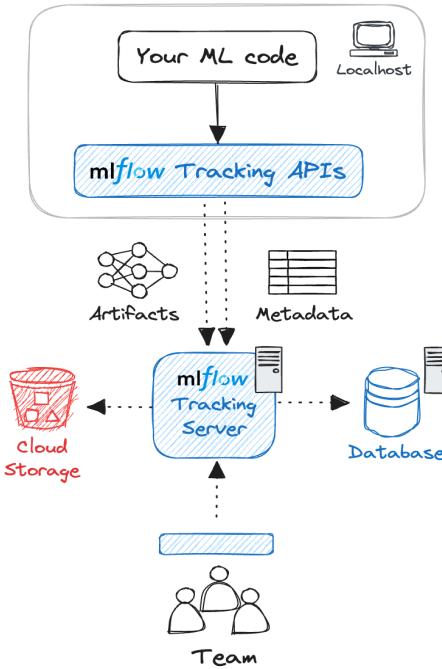
1. Localhost (default)



2. Localhost w/ various data stores



3. Remote Tracking w/ Tracking Server



→ Hall Варнант!

Практика

Ключевые тезисы

1. MLFlow – система управления жизненным циклом ML-проектов. Помогает навести порядок и автоматизировать задачи (в том числе регулярное переобучение ваших моделей).
2. Роботы надежнее людей, все, что можно автоматизировать, нужно автоматизировать.
3. MLOps – это инструменты + навыки + mindset. Причем навыки и mindset важнее инструментов.

Вопросы?



Задаем
вопросы в чат



Ставим “-”,
если вопросов нет

Мы ещё не закончили,
но самое время заполнить
опрос о занятии



Знакомство с командой и программой курса

Процесс обучения



Обучение проходит онлайн по вечерам или в выходные дни.



В процессе можно задать преподавателю вопросы по материалам, уточнять моменты, которые были непонятны на уроке



Все записи занятий и материалы сохраняются в личном кабинете. Вы можете вернуться к ним после окончания обучения.



Время на обучение: от 4 ак. часов на занятия и 4-8 часов на домашнюю работу в неделю



Домашние задания позволяют попрактиковаться. По каждому домашнему заданию преподаватель даёт развернутый фидбек.



Программа обучения на курсах обновляется каждый запуск в зависимости от актуальных запросов в сфере IT-технологий

Карьера и информация



Анализ позиций про MLOps

160+

Вакансий со специализацией
MLOps в апреле 2024 г.

*Источник – hh

Вакансии работодателей

MLOps Engineer

LESTA
GAMES

Какие навыки необходимы:

- Уверенное знание операционных систем семейства Linux, владение bash.
- Уверенное знание и владение технологией контейнеризации Docker.
- Уверенное знание процессов CI/CD и опыт использования систем непрерывной интеграции (Jenkins).
- Знание языка программирования Python.
- Понимание структуры ML-конвейеров и метрик оценки качества ML-моделей.
- Опыт использования систем управления версиями (Git).

Источник – hh



Вакансии работодателей

Ключевой проект - "**Цифровой Билайн**".

Цель проекта - автоматизация процессов управления компанией на основе больших данных. Мы находим цифровые следы, рассчитываем метрики и строим модели, для того, чтобы заместить субъективные факторы и ручной труд. Перед нами большой вызов - более сотни описанных процессов, в которых нужно внедрить data-driven подход.

Наш стек:

- Python + ML libs (pandas, scikit-learn, catboost, pytorch).
- Hadoop (pyspark, hive, hdfs).
- MLFlow, Argo Workflows, Airflow, JupyterHub, K8s.

Источник – hh



Вакансии работодателей

В федеральной розничной сети "ЛЕНТА" открыта вакансия
Исследователя больших данных (MLE)

ЛЕНТА — третий по выручке ритейлер в России.

Сейчас ищем **Middle+**специалиста в команду **GeoML**.

Наш стэк:

Яндекс Облако, PySpark в k8s, Jupyterhub, Airflow, MLFlow, Postgres, Gitlab, Grafana

Мы ждём, что ты знаешь:

- * Python - Писать качественный читаемый код
- * PySpark, Postgres - Хорошее знание SQL и работа с распределенными вычислениями
- * Airflow, Grafana - Постановка задач на расписание и мониторинг основных метрик
- * Git, DVC, MLFlow - Умение работать с системой контроля версий кода и трекать ML-эксперименты
- * Docker - Разработанный ML-сервис должен быть упакован в контейнер для дальнейшего использования
- * Матстат, методы оптимизации, ML-методы - Хорошее понимание работы ML-моделей и способов их применения для оптимизации бизнес задач

Источник – hh

Рефлексия

Цели вебинара



После занятия вы сможете

- 1.** Определить сценарии применения MLFlow, разобраться, как этот инструмент помогает в переобучении моделей;

- 2.** Понять, как MLFlow работает в облаке, используя объектное хранилище (S3) для хранения артефактов;

- 3.** Управлять экспериментами, версионировать артефакты моделей;

Список материалов для изучения

1. [Официальная документация MLFlow](#)
2. [Как развернуть MLFlow в Yandex Cloud](#)
3. [Статья на Habr раз](#)
4. [Статья на Habr два](#)
5. [Machine Learning Engineering with MLflow](#)

О курсе

MLOps



Старт обучения: 28.05.2024



Следующий открытый урок: 14.05.2024

Заполните, пожалуйста, опрос о занятии

**Важно! Пройти опрос могут только залогиненные
пользователи платформы OTUS**



Спасибо за внимание!