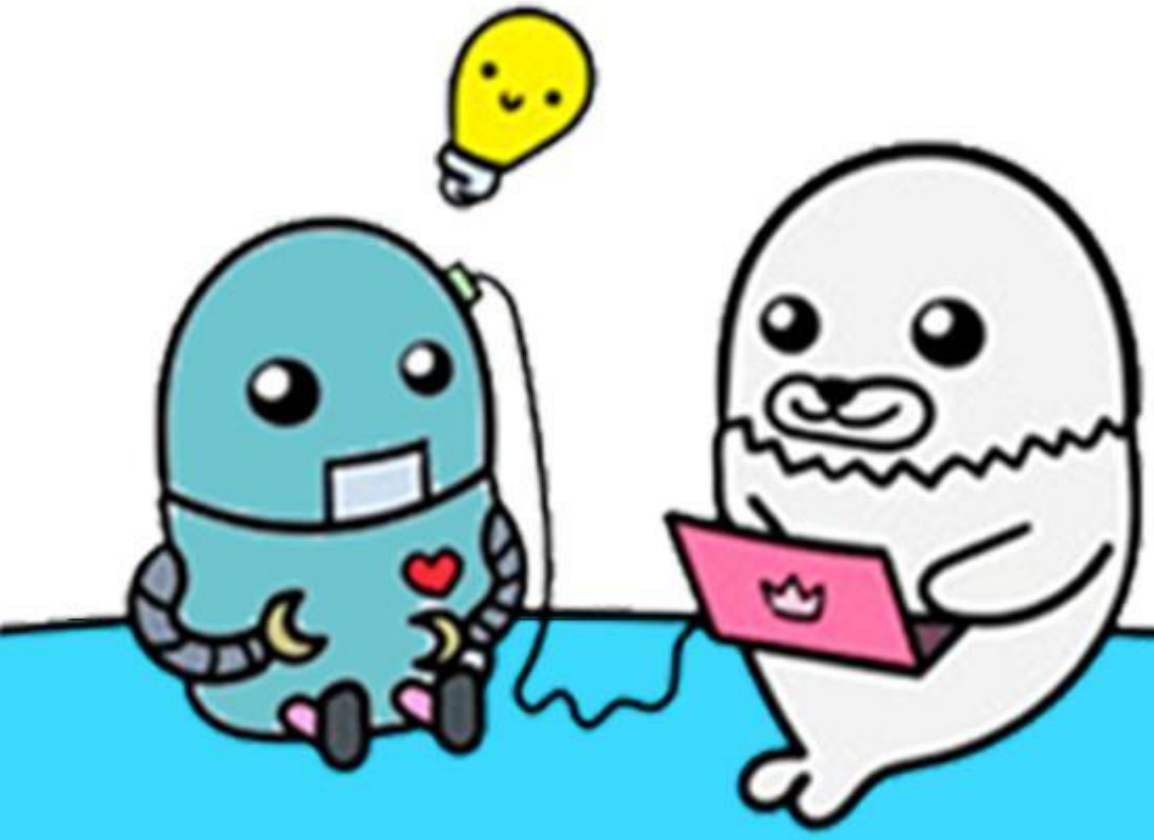
Яндекс Кью

Open Data Science

MLOps и production подход к ML исследованиям



28 марта - 28 мая

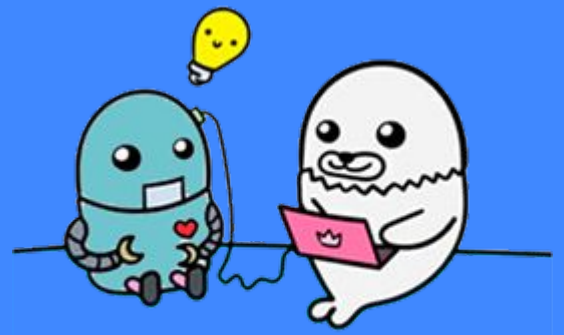
Яндекс Кью

Open Data Science

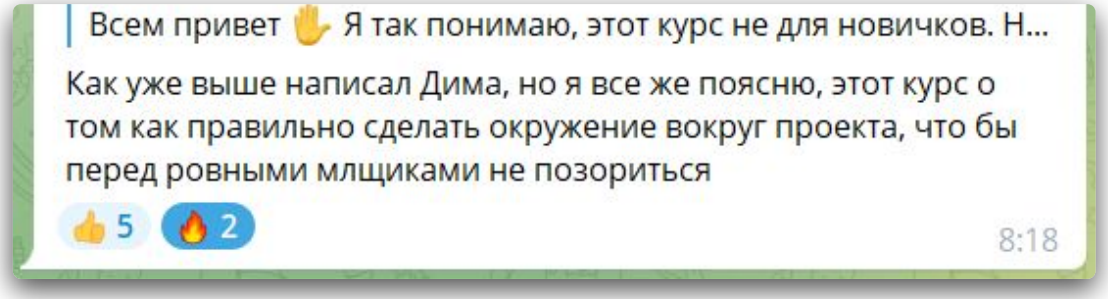
MLOps и production подход к ML исследованиям

Концепция воспроизводимых и масштабируемых исследований в ML

Павел Кикин
Газпромнефть ЦР
Руководитель направления NLP
t.me/pavel_kikin



Для чего этот курс



- Работать в команде над одним проектом
- Превратить модель в сервис
- Корректно передать результаты исследований заказчику или другим разработчикам
- Систематизировать исследования
- Отслеживать и сохранять условия и результаты экспериментов
- Автоматизировать эксперименты
- Повысить качество кода
- Выстроить с нуля процессы МЛ разработки в команде
- **Обеспечить воспроизводимость ваших исследований**

Для чего этот курс

- Написать больше умных слов в резюме
- Казаться умнее на собеседованиях

О курсе

9 занятий теория + практика

Понедельник 18:30

Рейтинг участников

Конференции в Zoom

Онлайн трансляции и запись в Youtube

О курсе

Домашние задания (не оценивается)
Google Drive

Тесты (10 балов)
Страничка курса в ODS

Дополнительные задания (20 баллов)

- *руководство*
- *[illegible]/доклад*

Вопросы и ответы на Кью (5 баллов)

Итоговый проект (50 баллов)

О курсе

Примерные темы для докладов:

- **Опыт реализации проекта Findmybike.ru**
- **Исследование сервисов управления ноутбуками (sagemaker, databricks, datalore, ванильный JH)**
- Amazon S3 и S3-like хранилища
- Docker
- СУБД
- Особенности ООП в МЛ
- Применение Agile в ML командах
- w&b, neptuna, clearml
- Pachyderm
- Dataflow
- Kubernetes and KubeFlow
- TF serving
- Контейнеризация рабочей среды для DS
- Настройка CUDA для Docker

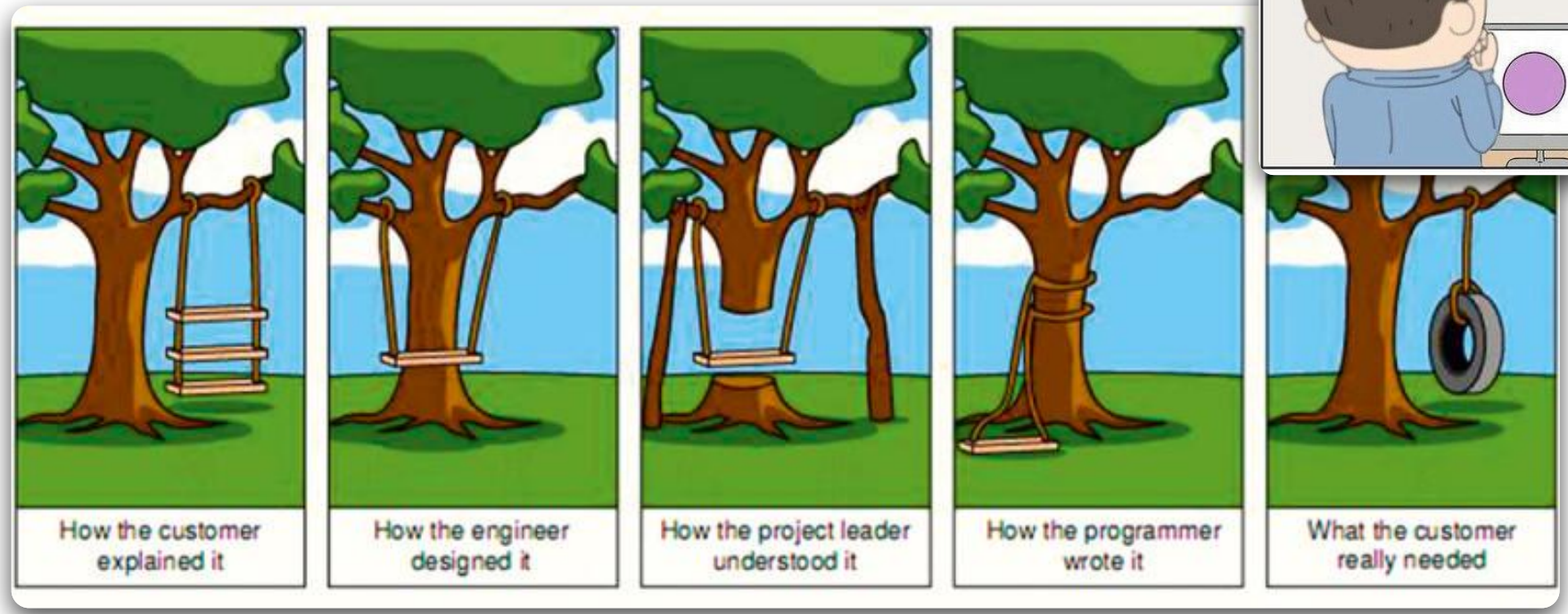
Процесс работы над ML проектом

- Бизнес анализ, инициирование работ
- Обзор и анализ существующих решений – бейзлайнов
- Исследование/собственная разработка
- Подготовка отчета

Бизнес анализ, инициирование работ

Постановка бизнес - задачи

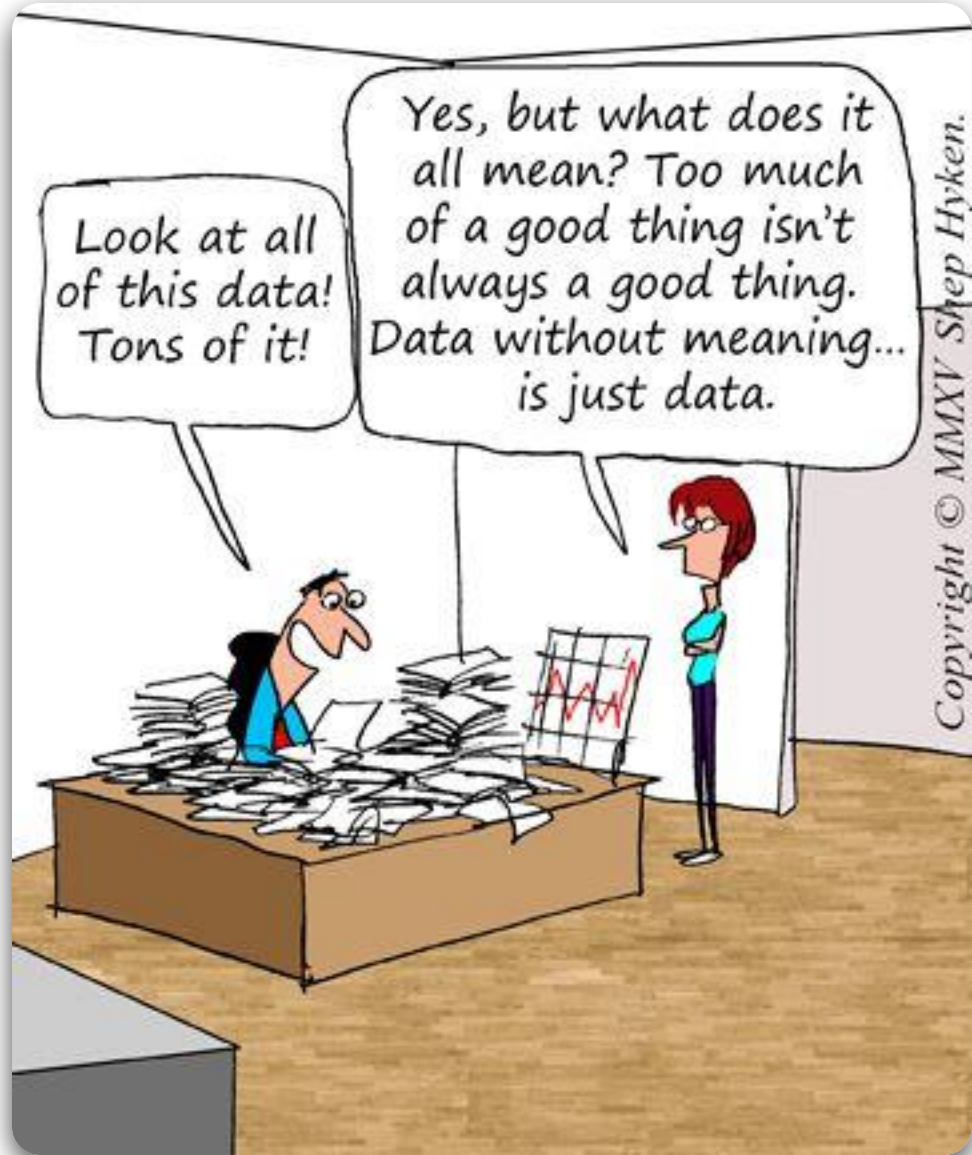
- Понять бизнес-цели (помочь заказчику их сформулировать)
- Очертить возможные подходы и решения
- Обозначить технологические пределы возможностей
- Продумать риски
- Повторно продумать и скорректировать бизнес-цели



Бизнес анализ, инициирование работ

Оценка данных

- Сколько объектов выборки мы имеем?
- Возможно ли ознакомиться со всеми данными до старта проекта?
- Есть ли разметка и какого она качества?
- Есть ли возможность разметить все классы?
- Сколько классов?
- Есть ли дисбаланс?
- Что с выбросами и пропусками?



Бизнес анализ, инициирование работ

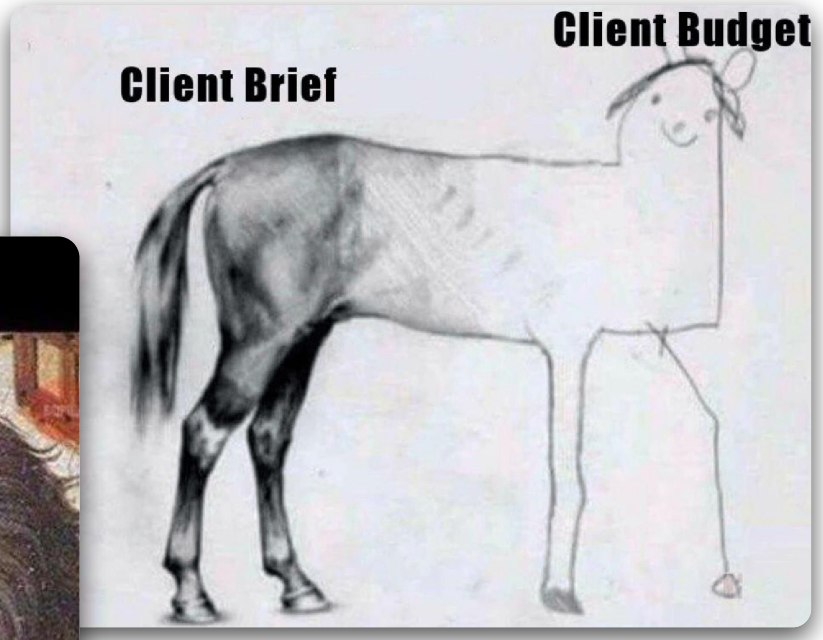
Оценка заказчика и проекта

- Узнайте заказчика поближе.
- Изучите опыт работы других исполнителей с ним.
- Назначьте встречу со всеми заинтересованными сторонами от заказчика, а не только с одним из представителей.

Бизнес анализ, инициирование работ

Оценка заказчика и проекта

- Деньги будут?



Бизнес анализ, инициирование работ

Подбор метрик

- Выберите метрики, которые будут наилучшим образом согласовываться с бизнес-целями
- Объясните их заказчику
- Убедитесь, что заказчик их понял. Еще раз проверьте, что он их **ТОЧНО** понял
- Если метрик много, нужно выбрать набор, который будет использоваться для валидации решения/ий

Бизнес анализ, инициирование работ

Финальный этап

- Четко сформулируйте цели исследования
- Оцените ресурсы с учетом рисков (перезакладывайтесь):
  - Особенности заказчика
  - Неопределенность в требованиях
  - Неопределенность в данных

Бизнес анализ, инициирование работ

Финальный этап

- Сформируйте ТЗ
  - Можно ли по этому ТЗ четко сказать, что проект выполнен в полной мере или нет?
  - Точные метрики
  - Критерии приемки

Обзор, анализ и реализация существующих решений – бейзлайнов

- Формирование критериев оценки:
  - Метрики
  - Производительность
  - Работа под нагрузкой
  - Поддержка
  - Готовность
  - Популярность
- Поиск существующих решений (не беритесь за 1 попавшееся):
  - Популярные сайты для разработчиков
  - Научные источники
  - ODS, TG, Кью, профильные форумы
- Реализация подходящих решений
- Составление валидационного набора данных
- Разработка валидационного пайплайна (бенчмарка)

«*Не изобрети велосипеда*»

Подготовка отчета

- Постановка задачи
- Описание исходных данных
- Обзор и анализ существующих решений
- Методика, технология и критерии оценки существующих решений
- Результаты оценки существующих решений и выводы

Вариант 1 – мы нашли подходящее решение

- Детальное описание выбранного решения и его адаптации под нашу задачу
- Детали предобработки исходных данных

Вариант 2 – пилим своё

- Гипотезы
- Результаты проверки гипотез

- Заключение:
  - Выводы

Концепция воспроизводимости

Воспроизводимость

Мера вероятности того, что, получив один результат эксперимента, вы сможете провести тот же эксперимент с теми же параметрами и получить точно такой же результат. Это способ убедиться, что результаты верны и не являются случайностями.

Виды воспроизводимости

- **Repeatability:** та же команда, **те же условия эксперимента**. Возможность получения заявленных в результатов на тех же входных данных.

- **Reproducibility:** другая команда, **те же условия эксперимента**. Если наблюдение воспроизводимо, оно должно быть выполнено другой командой, повторяющей эксперимент с использованием тех же экспериментальных данных и методов, в тех же рабочих условиях, в том же или другом месте, в нескольких испытаниях.

- **Replicability:** другая команда, **другие условия эксперимента**. Если наблюдение можно воспроизвести, оно должно быть выполнено другой командой, с использованием другой измерительной системы и набора данных, в другом месте, в нескольких испытаниях. Следовательно, это потребует нового сбора данных.

Plesser HE. Reproducibility vs. Replicability: A Brief History of a Confused Terminology. Front Neuroinform. 2018;11:76. Published 2018 Jan 18. doi:10.3389/fninf.2017.00076

Одинаковые условия эксперимента?

В целом в науке:

- Место нахождения
- Измерительные инструменты
- Другое оборудование, использованное в эксперименте
- Наблюдатель
- Гипотеза
- Период времени

Для ML:

- Аппаратное обеспечение
- Программное обеспечение и его версии
- Способ/оборудование для получения данных (датчики климата, спутниковые оборудование и т.д.)
- Время получения данных (особенно важно для часто обновляемых источников, новостные ленты, социальные сети и т.д.)

Кризис воспроизводимости

Крупномасштабные усилия по оценке воспроизводимости научных публикаций дали тревожные результаты. Например, в 2015 году группа исследователей психологии, получившая название «Открытое научное сотрудничество», рассмотрела 100 экспериментов, опубликованных в высокорейтинговых рецензируемых журналах. Из этих 100 исследований только 68 репродукций дали статистически значимые результаты, совпадающие с исходными данными.

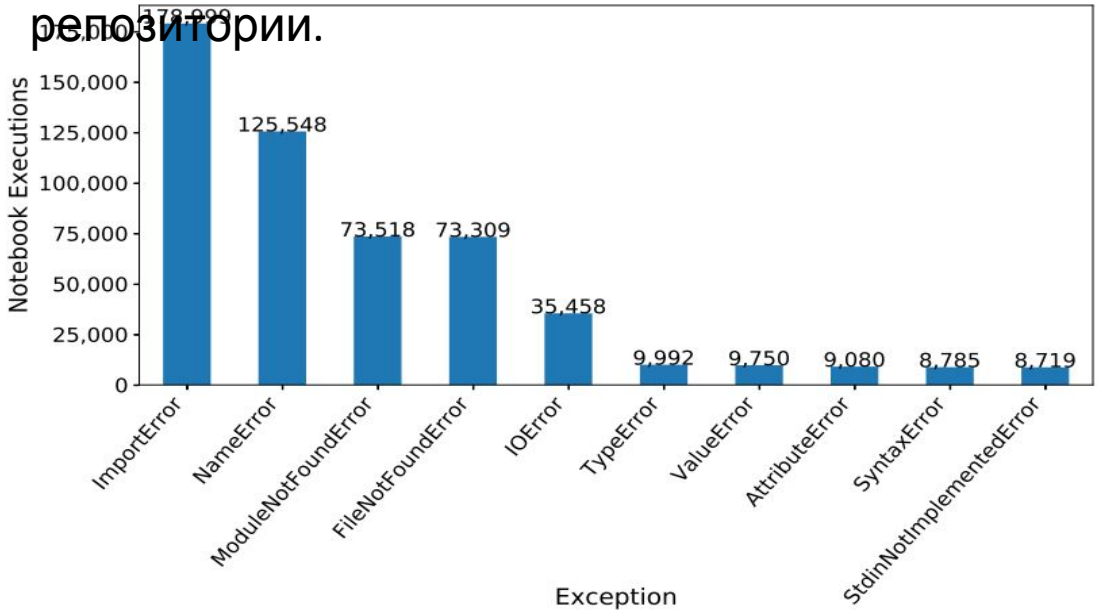
Согласно данным другого анализа, до 85 % всех проведённых в мире исследовательских работ в области биомедицины не привели к значимым результатам.

Признаки кризиса воспроизводимости:

- Фактический недостаток формализации воспроизводимости исследований в опубликованной литературе по многим научным направлениям
- широко распространена невозможность воспроизвести результаты опубликованных исследований;
- высокая распространенность «закрытых» методов исследования, которые завышают количество ложноположительных результатов
- отсутствие данных и алгоритмов анализа в научных публикациях.

Факторы уменьшающие воспроизводимость

- Первое место — проблемы с зависимостями в библиотеках и зависимостями в зависимостях. Часть репозиториев использовали requirements.txt, часть setup.py.
- Второе место — порядок исполнения. Тетрадка сохранена без очистки кода, порядок не сохранен и некоторые переменные объявлены или инициализированы после использования.
- Третье место — нет нужных данных, например, указаны абсолютные пути или данных вообще нет в репозитории.



Pimentel, João Felipe et al. “A Large-Scale Study About Quality and Reproducibility of Jupyter Notebooks.” 2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR) (2019): 507-517.

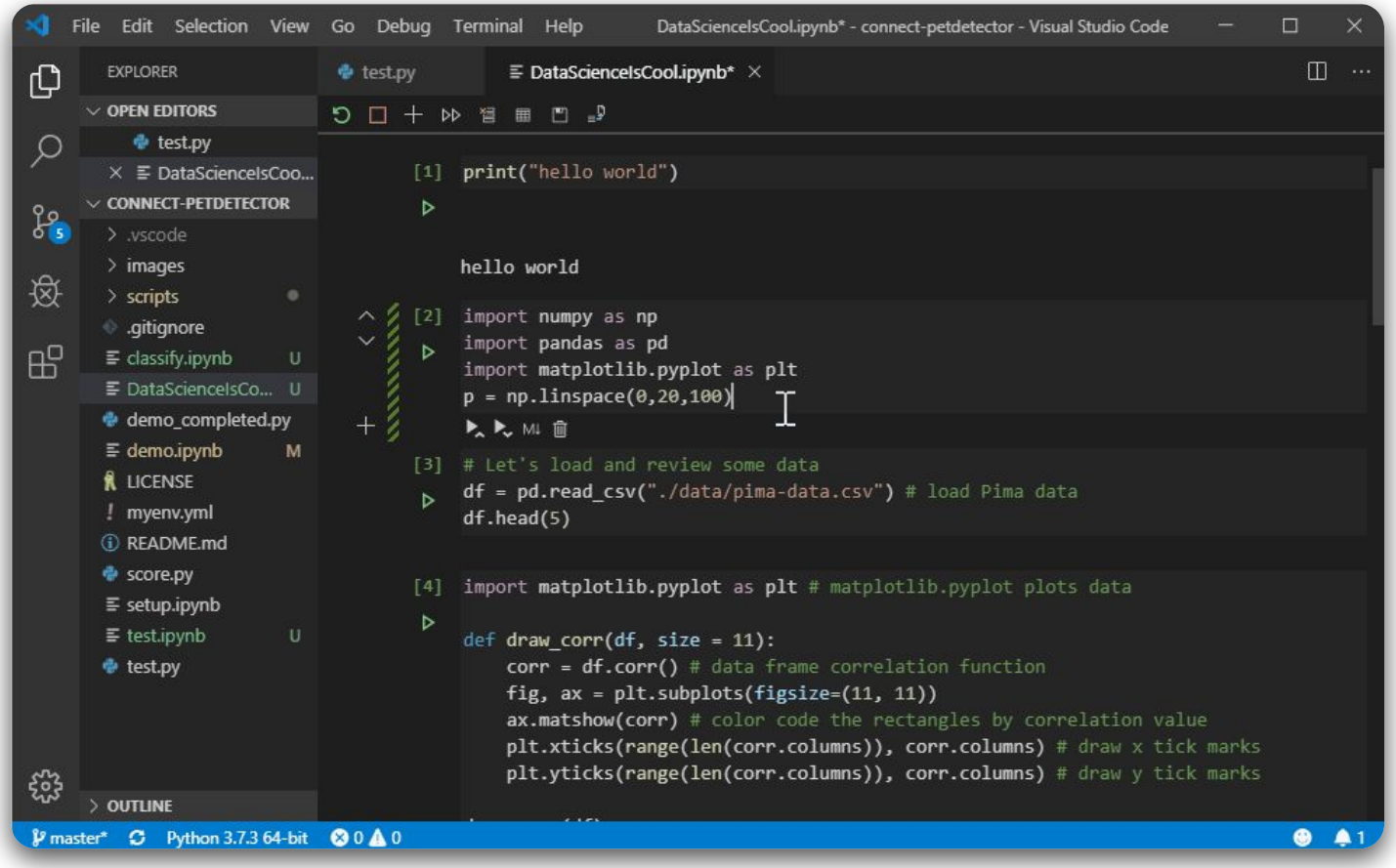
Факторы уменьшающие воспроизводимость

- Неуправляемая случайность в данных или алгоритмах (40%) (Random Seed)
- Исходные данные измененные вручную, а не с помощью скриптов
- Зависимость вывода и результатов от функций времени (13%)
- Различия отображения на графиках (некорректное использование matplotlib в том числе) (52%)
- Недоступны внешние данные (3%)
- Различия в выводе чисел с плавающей запятой (3%)
- Непостоянный порядок обхода словарей и др. контейнеров в python (4%)
- Различия в среде исполнения (27%)

Чего хотят конференции: воспроизводимость экспериментов в data science https://habr.com/ru/post/562262/

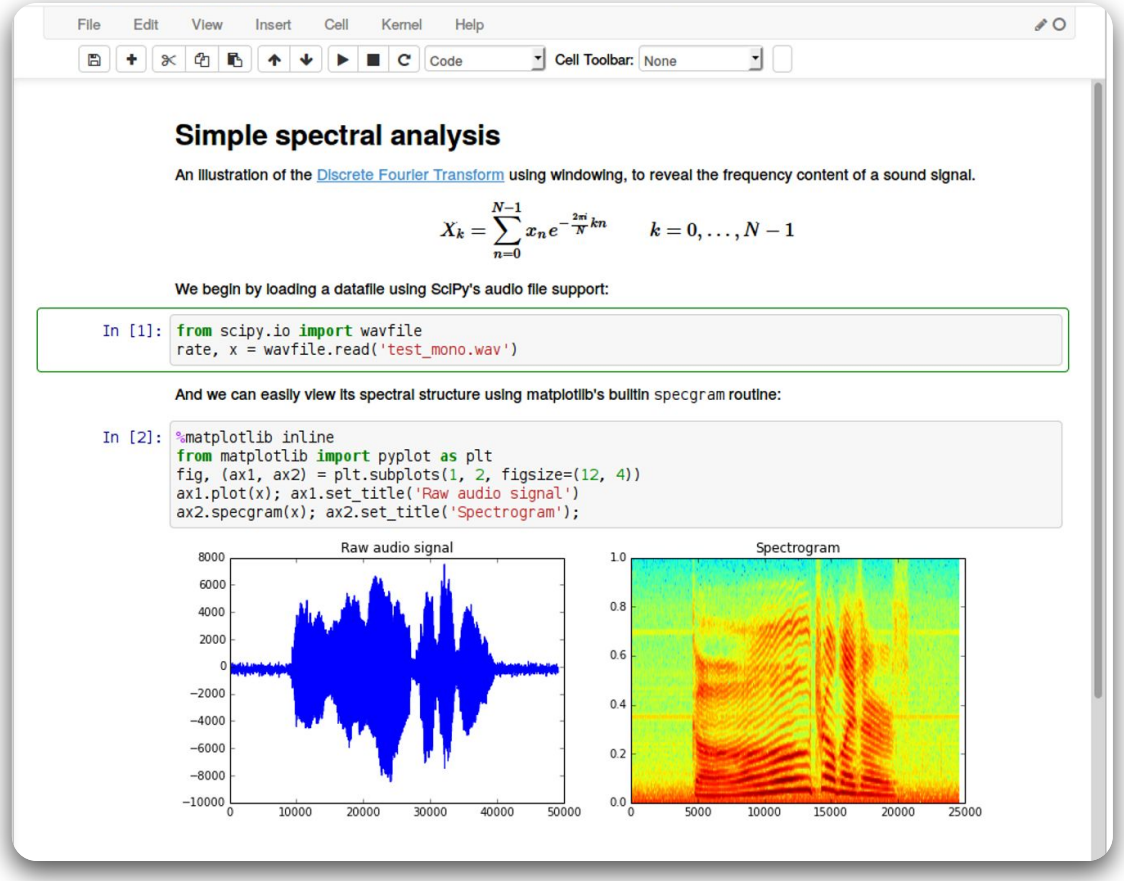
Способы улучшение воспроизводимости

- **Пишите код в виде Python скриптов в полноценных IDE**



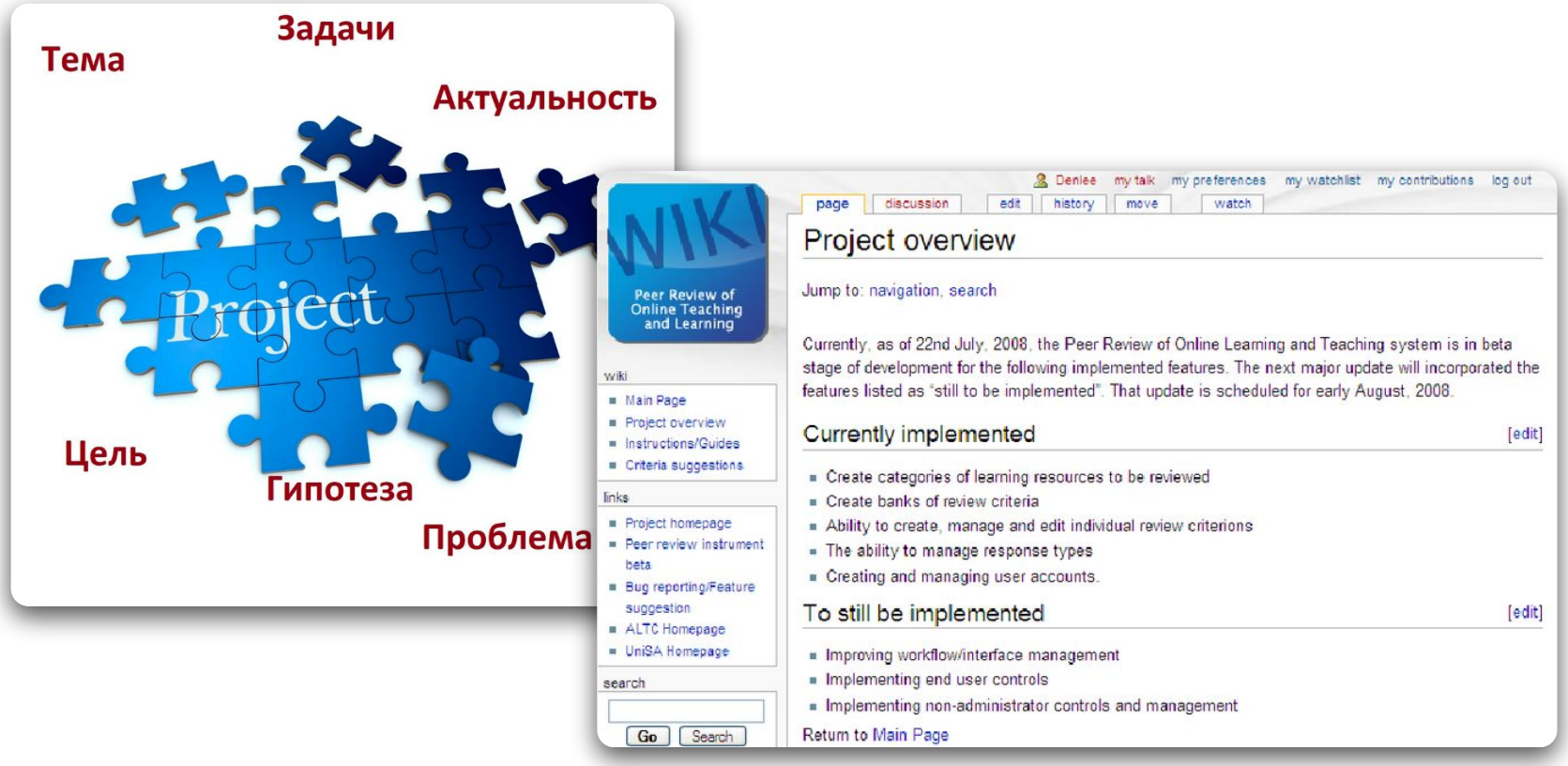
Способы улучшение воспроизводимости

- **Используйте ноутбуки для быстрых экспериментов и их описания**
  - Не должен являться основным артефактом разработки
  - Максимум абстракции (весь основной код во внешних скриптах)
  - Больше описаний
  - Больше графиков



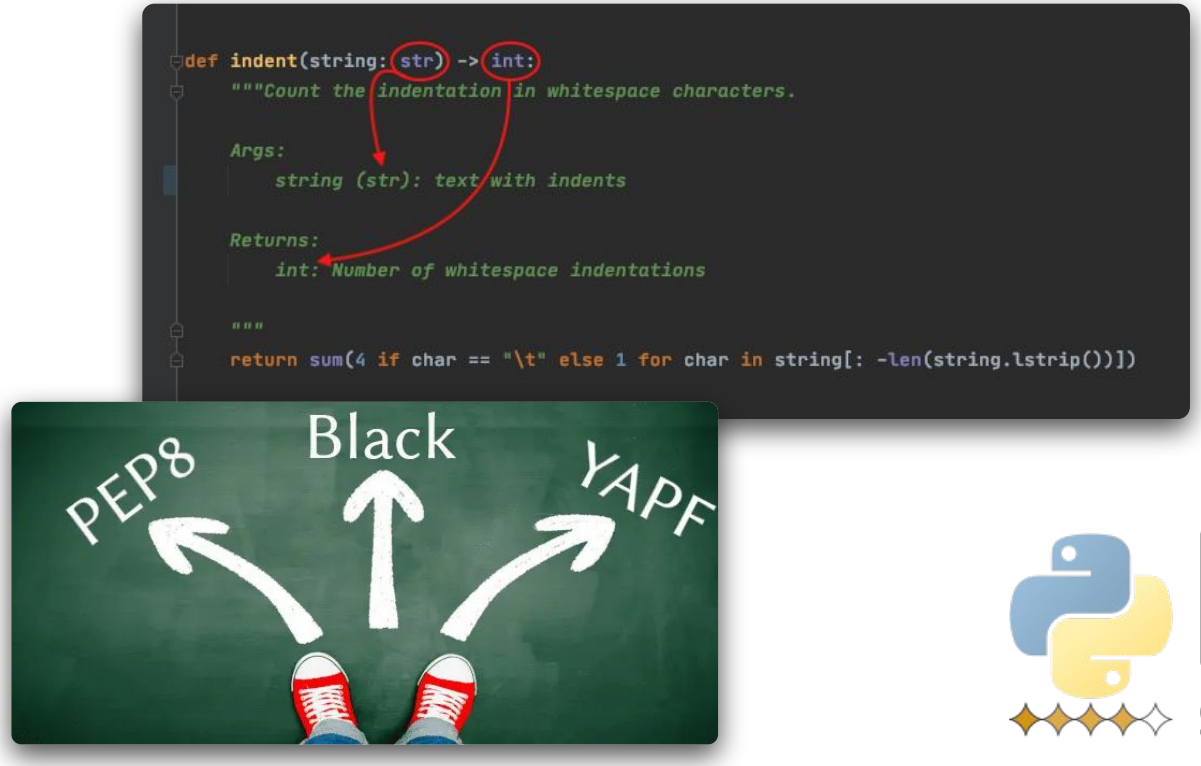
Способы улучшение воспроизводимости

- **Документируйте эксперимент по ходу исследования, а не после его окончания**
- **Ведите вики проекта**



Способы улучшение воспроизводимости

- Соблюдайте Codestyle, пишите комментарии и нотации
- Используйте линтеры и автоформатеры



flake8

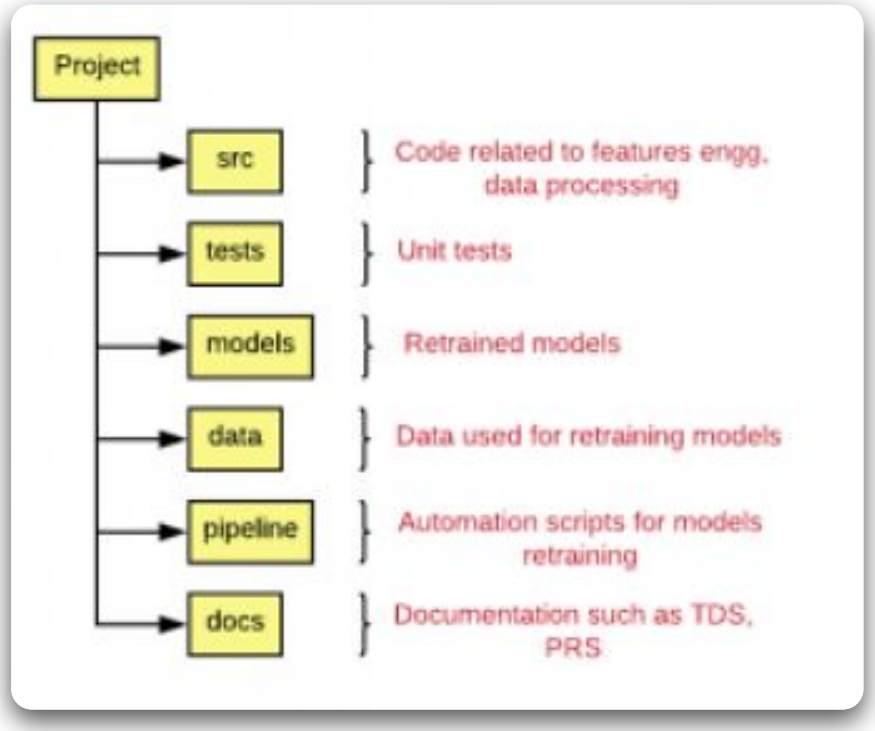
Bandit

Pylint

Star your Python code!

Способы улучшение воспроизводимости

- **Шаблонизируйте DS проекты**



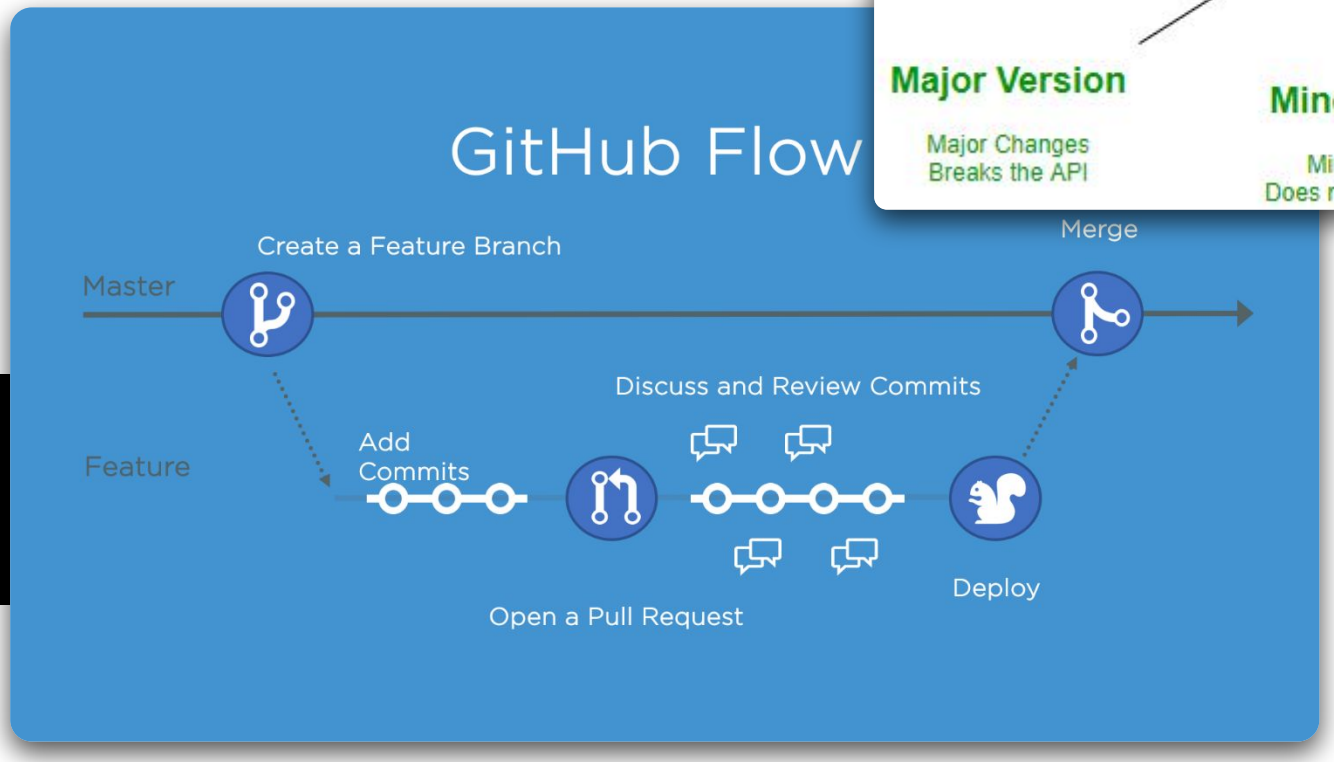
COOKIECUTTER

```
├── LICENSE
├── Makefile           <- Makefile with commands like `make data` or `make train`
├── README.md          <- The top-level README for developers using this project.
├── data
│   ├── external       <- Data from third party sources.
│   ├── interim        <- Intermediate data that has been transformed.
│   ├── processed      <- The final, canonical data sets for modeling.
│   └── raw            <- The original, immutable data dump.
│
├── docs               <- A default Sphinx project; see sphinx-doc.org for details
│
├── models             <- Trained and serialized models, model predictions, or model summaries
│
├── notebooks          <- Jupyter notebooks. Naming convention is a number (for ordering),
│                         the creator's initials, and a short `-` delimited description, e.g.
│                         `1.0-jqp-initial-data-exploration`.
│
├── references         <- Data dictionaries, manuals, and all other explanatory materials.
│
├── reports            <- Generated analysis as HTML, PDF, LaTeX, etc.
│   └── figures        <- Generated graphics and figures to be used in reporting
│
├── requirements.txt   <- The requirements file for reproducing the analysis environment, e.g.
│                         generated with `pip freeze > requirements.txt`
│
├── setup.py           <- makes project pip installable (pip install -e .) so src can be imported
├── src                <- Source code for use in this project.
│   ├── __init__.py    <- Makes src a Python module
│   │
│   ├── data           <- Scripts to download or generate data
│   │   └── make_dataset.py
│   │
│   ├── features       <- Scripts to turn raw data into features for modeling
│   │   └── build_features.py
│   │
│   ├── models         <- Scripts to train models and then use trained models to make
│   │   │                 predictions
│   │   ├── predict_model.py
│   │   └── train_model.py
│   │
│   └── visualization  <- Scripts to create exploratory and results oriented visualizations
│       └── visualize.py
│
└── tox.ini            <- tox file with settings for running tox; see tox.readthedocs.io
```

Способы улучшение воспроизводимости

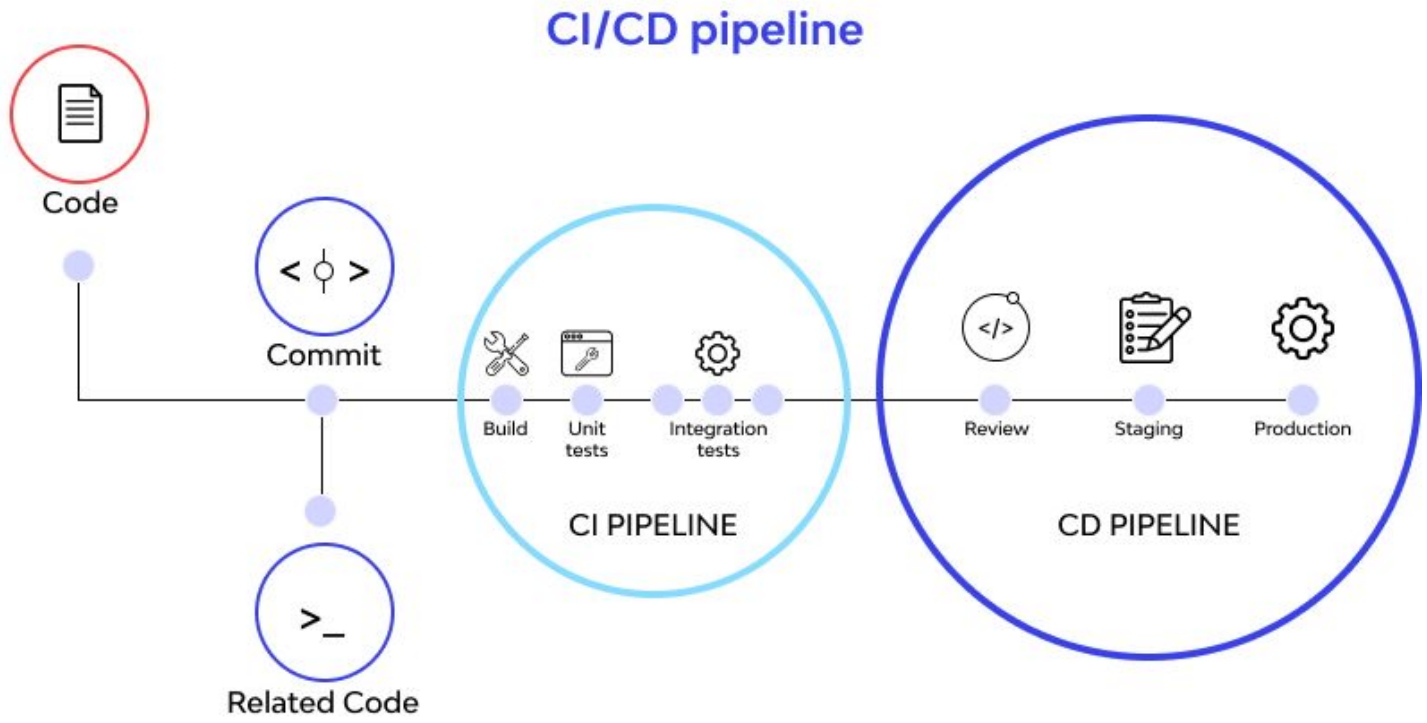
- **Используйте системы контроля версий, соблюдайте версионирование кода, проводите качественное code-review**

GitLab



Способы улучшение воспроизводимости

- **Настройте CI/CD (Continuous Integration, Continuous Delivery)**

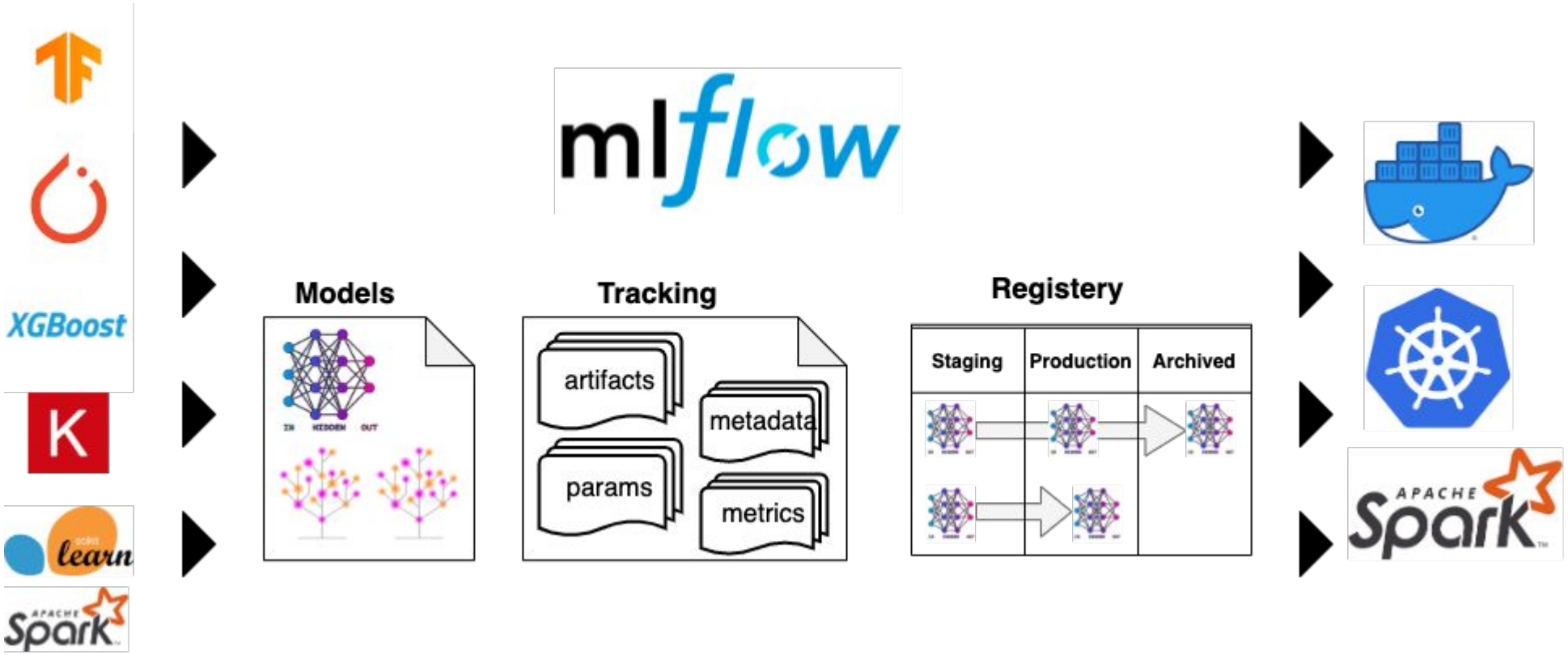


Способы улучшение воспроизводимости

- **Покрывайте код, данные и модели тестами**
  - Юнит тесты
  - Интеграционные тесты
  - Тесты данных
  - Тесты моделей
  - Нагрузочные тесты

Способы улучшение воспроизводимости

- **Применяйте системы трекинга экспериментов (MLFlow)**

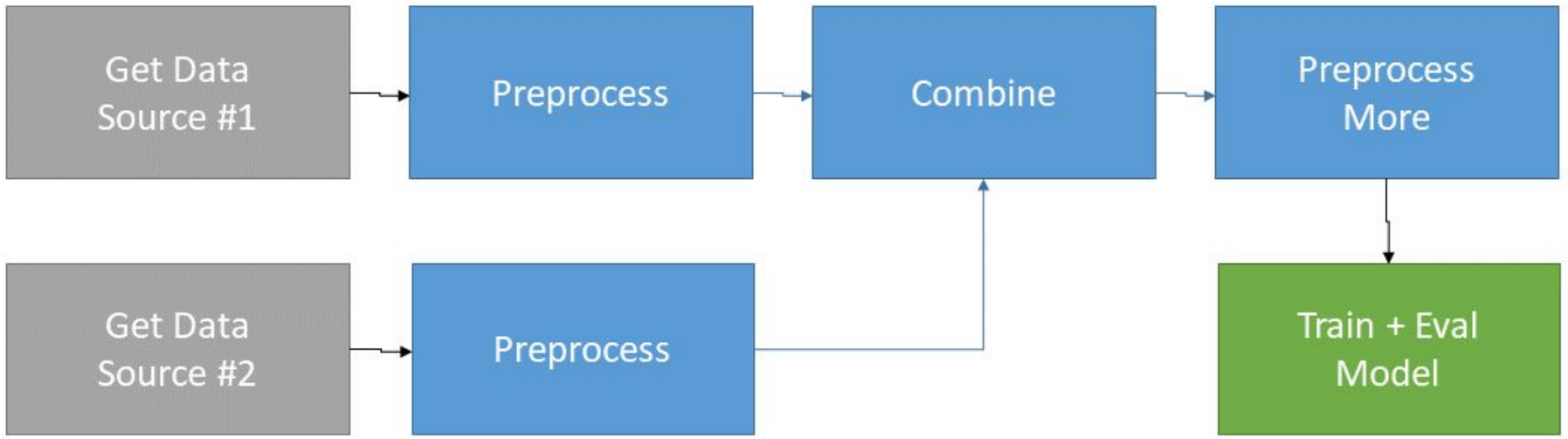


Способы улучшение воспроизводимости

- **Сохраняйте алгоритмы получения для каждого результата (workflow менеджеры с DAG)**

Apache Airflow

snakemake



Способы улучшение воспроизводимости

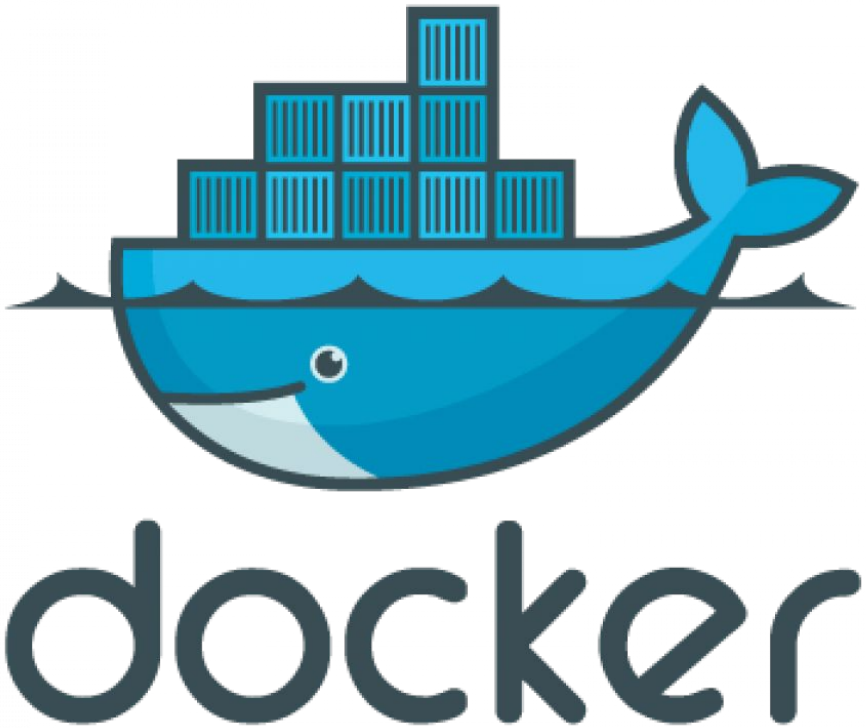
- Оформляйте код, как python пакеты
- Создавайте CLI
- Пишите инструкции по запуску/интерфейсам вашего кода
- Избегайте ручного изменения данных

Способы улучшение воспроизводимости

- **Управляйте зависимостями и сохраняйте конфигурации с точными версиями (conda, pip)**

Способы улучшение воспроизводимости

- **Контейнеризируйте**



Способы улучшение воспроизводимости

- Сохраняйте промежуточные данные (ускорение экспериментов)
- Фиксируйте случайные процессы в алгоритмах оптимизации и изменения данных (Random seed)
- Не работайте в одиночку. Выносите результаты на обсуждение.

Яндекс Кью

Open Data Science

MLOps и production подход к ML исследованиям

Концепция воспроизводимых и масштабируемых исследований в ML

Павел Кикин
Газпромнефть ЦР
Руководитель направления NLP
t.me/pavel_kikin