Воспроизводимость экспериментов

ИТМО

+

+

Data

Schema

Sampling over Time

Volume

Model

Algorithms

More Training

Experiments

Code

Business Needs

Bug Fixes

Configuration

Воспроизводимость

- Повторяемость измерений(также сходимость результатов измерений, англ. Repeatability)
- Повторяемость исследований (англ. Replicability) (Different team, same experimental setup)
- Воспроизводимость (англ. Reproducibility)

Проведение воспроизводимых исследований

1. Для каждого полученного результата храните алгоритм его получения
2. Избегайте этапов ручного управления данными или процессом
3. Храните точные версии всех использованных внешних инструментов
4. Используйте контроль версий
5. Храните все промежуточные результаты в стандартизированном виде.
6. Для алгоритмов использующих случайность фиксируйте *random_state*.
7. Всегда храните вместе с графиками данные
8. Иерархический подход при генерировании результатов анализа
9. Всегда указывайте вместе текстовые утверждения и результаты исследования
10. Обеспечивайте доступность ваших результатов, данных и исследований

I don't like notebooks

- I don't like notebooks.- Joel Grus
- Clean Code in Jupyter notebooks, using Python

Code Smells .. in ipynb

- Cells can't be executed in order (with runAll and Restart&RunAll)
- Prototype (check ideas) code is mixed with "analysis" code
- Debugging cells
- Copy-paste cells
- Duplicate code (in general)
- Multiple notebooks that re-implement the same function

Управление зависимостями

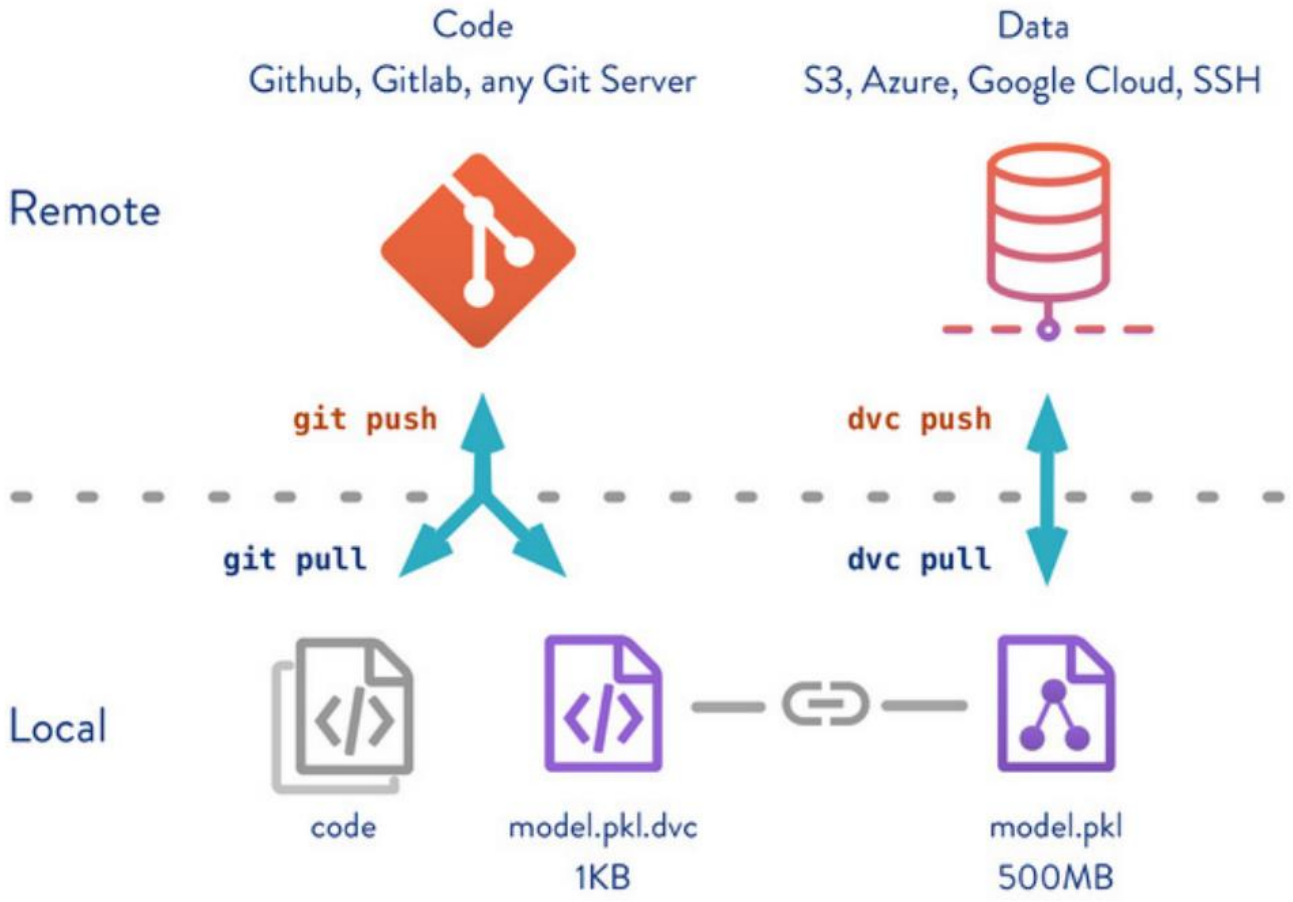
Инженерные практики - код

ИТМО

- Версионирование – Git
- Тестирование – UnitTest, PyTest…
- Качество кода – Pep8, black, isort, flake8, pylint…

Версионирование данных

- DVC



ИТМО

Cookiecutter DS

- Логичная, достаточно стандартизированная, но гибкая структура проекта для выполнения работы по науке о данных и обмена ею.

```
├── 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           <- Make this project pip installable with `pip install -e`
├── 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
```

Пайплайны и трекинг исследований

- Фиксация последовательности исполнения.
- Упрощение воспроизведения результатов.
- Распределение вычислений.
- DVC, AirFlow