DATAFEST 2023

Не бойтесь выкладывать свои разработки в open-source - даже если кажется, что они не закончены

ДОКЛАДЧИК: САРАФАНОВ МИХАИЛ
SENIOR DATA SCIENTIST, WIREDHUT

Что ждать от доклада

Хочется, чтобы слушатели:

- С шутками и хорошим настроением начали (или продолжили) секцию датафеста посвященную open-source; 😜
- Послушали про то как другие делают кривые open-source продукты и ими все равно пользуются;
- Зарядились желанием делиться собственными разработками

Если чуть серьезней и предметно по структуре:

- Немного про то кто я (докладчик) такой и на примере каких трех маааленьких проектов буду строить рассказ; 😎
- Составляющие почти идеального open-source продукта в вакууме;
- Немного информации про каждый из трех open-source проектов: что делает, какими силами разивался и развивается;
- Что сделано в этих проуктах хорошо, а что ну... плохо;
- Почему считаю, что каждый из них по своему удался

Чувствуйте себя как дома

*тут не будет душных слайдов с теорией, только яркие примеры из личного опыта, и советы, которые (по крайней мере для меня) работают

Прежде чем начать

- Это я, типичный data scientist, которому посчастливилось принять участие в разработке нескольких open-source продуктов

Государственный Гидрологический Институт (2019-2020)
Обработка спутниковых снимков, разработка ML моделей для прогнозирования уровня воды в реках

Университет ИТМО (2020-2022)
Разработка AutoML инструментов. Например, фреймворк FEDOT

Wiredhut Oy (2022-...)
Разработка прогностических моделей перемещения человеков по человековским городам и пригородам, проектирование и реализация систем для обработки данных получаемых с различных сенсоров

python™

1

SimpleSpatialGapfillingProcessor-toolbox

SSGP-toolbox

2

Lines Ranking

Qgis Lines ranking plugin

3

pytsbe

pytsbe

Самая самая главная цель

Показать, что не всегда open-source проекты должны быть большими и качественными, иногда они просто должны быть

© Джейсон Стетхем

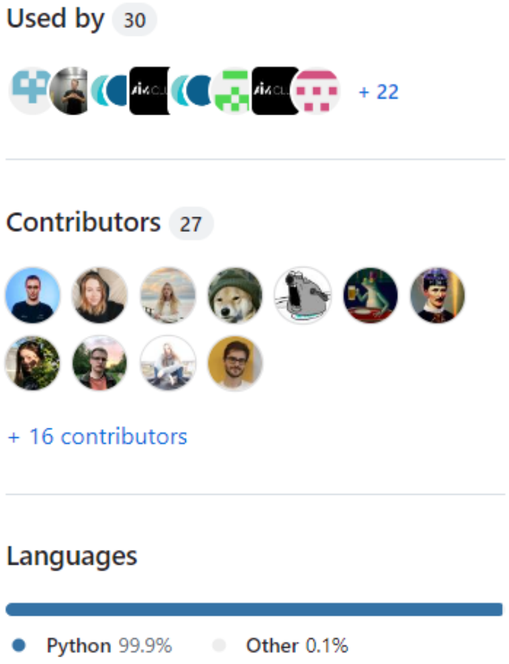
Составляющие крутого open-source

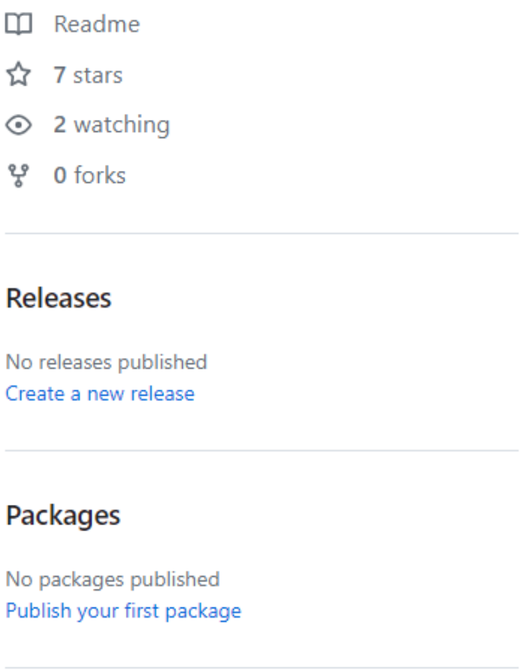
- **Большое сообщество пользователей и контрибьюторов**
- **Архитектура**, которая позволяет быстро расширять функциональность библиотеки. "Красивый" код как таковой
- **Тесты** - достаточное покрытие кода тестами
- **CI/CD** - настроенные автоматические помощники (линтер, автоматический запуск тестов и т.д.)
- **Development guideline** - выстроенный среди maintainer'ов прозрачный процесс утверждения задачи, выполнения и доработки, а также включения изменений в основной код проекта
- **Простой процесс установки** (оба)
- **Простой интерфейс** (насколько это возможно) для запуска open-source модуля
- **Понятная документация** как для пользователей, так и для разработчиков. Доходчивые примеры и пояснения, статьи по теме
- **Сопровождение** - регулярное улучшение библиотеки и поддержка пользователей
- **Много звезд на гитхабе**

Большое сообщество пользователей и контрибьюторов

Свежо, интересно, вызывает доверие

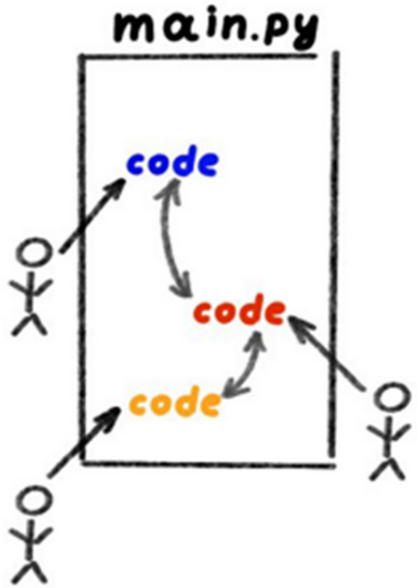
Страшно и ненадежно

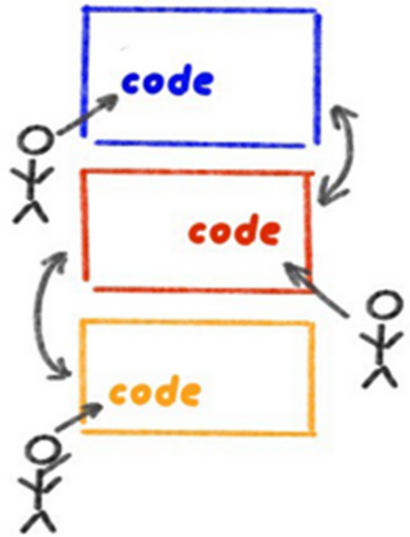


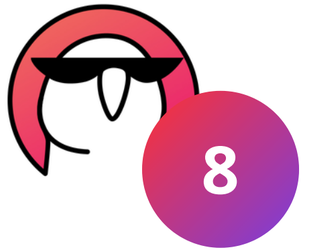


Архитектура

Разнесенные по разным модулям файлы могут изменяться одновременно, что позволяет работать нескольким разработчикам над проектом одновременно. Всегда стоит ожидать, что кто-угодно может захотеть исправить или дополнить какой-угодно модуль, - и эту возможность надо предоставить!

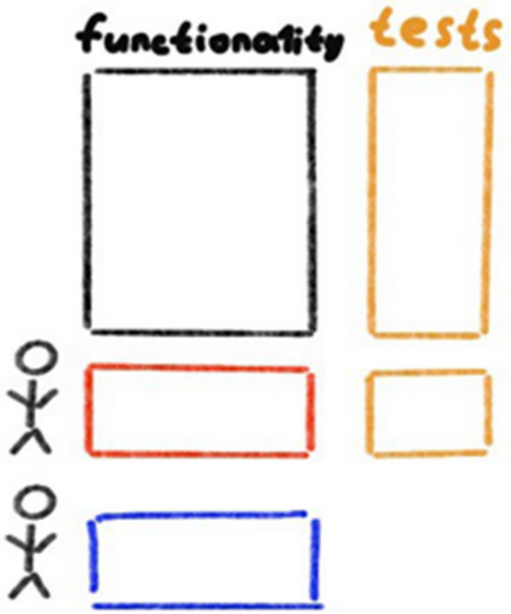


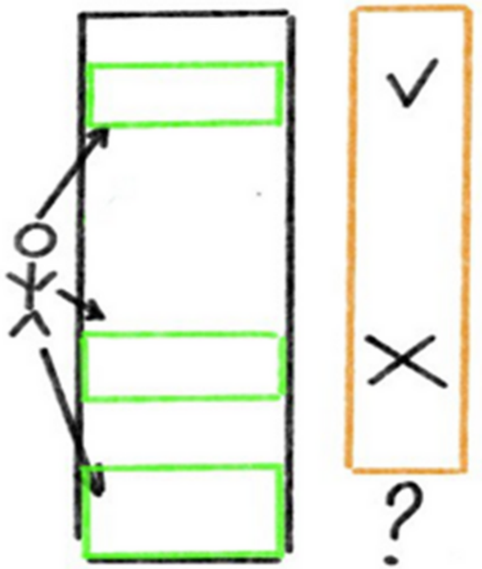




Тесты

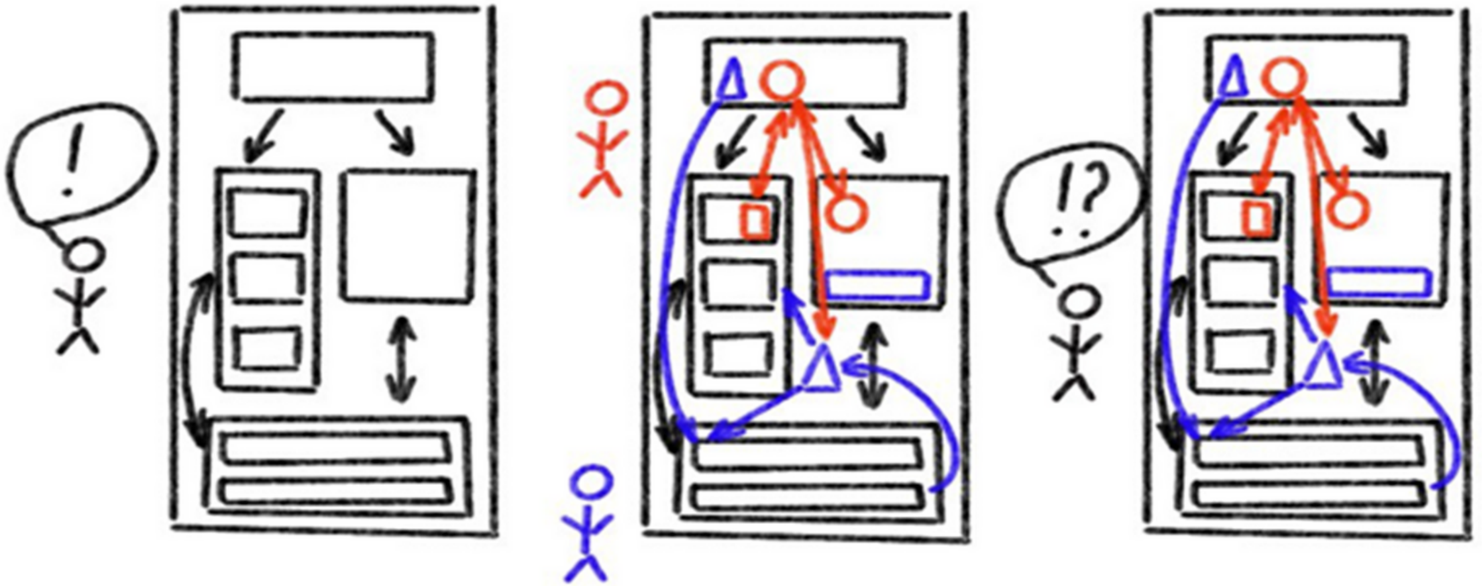
В процессе разработки (особенно когда разработчиков много) неизбежно появляются ошибки. Установить наличие ошибки помогают тесты: если тесты выполняются корректно изменения могут быть интегрированы в master / main ветку. Если хотя бы один не выполняется – контрибьютору следует внести исправления





CI/CD и Development guideline

Что можно автоматизировать и делать без участия людей - лучше автоматизировать.
Для остальных вещей пишется руководство
© Тоже Джейсон Стетхем

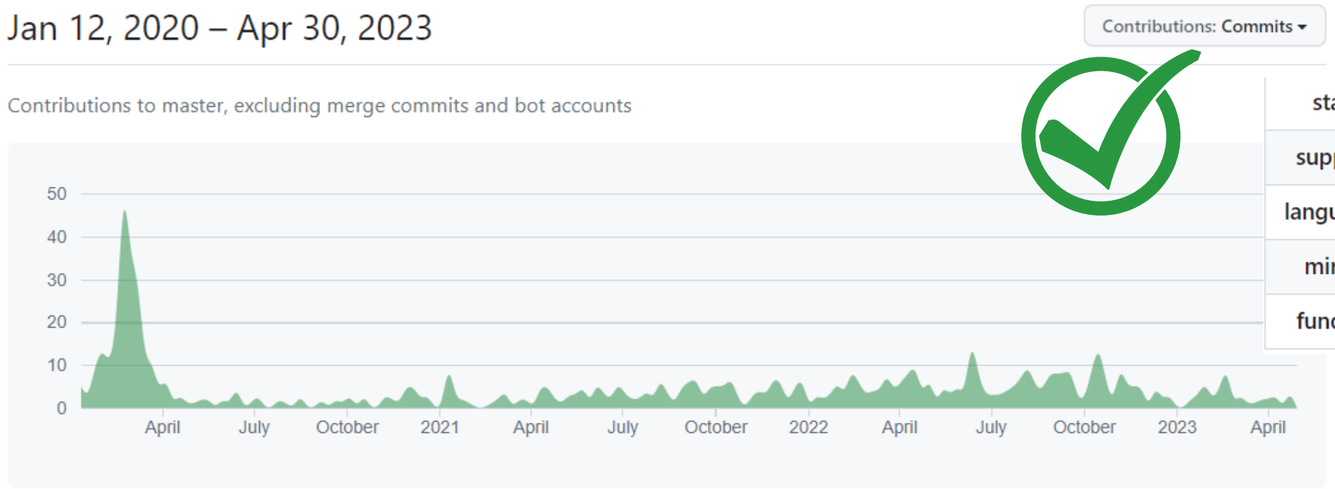


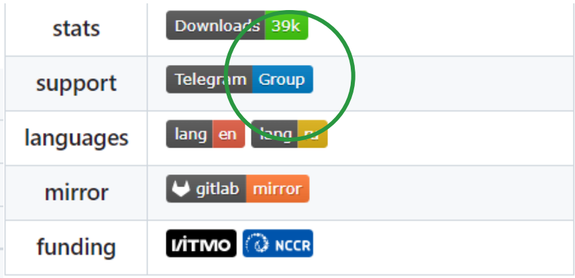
Простота в установке и использовании. Документация

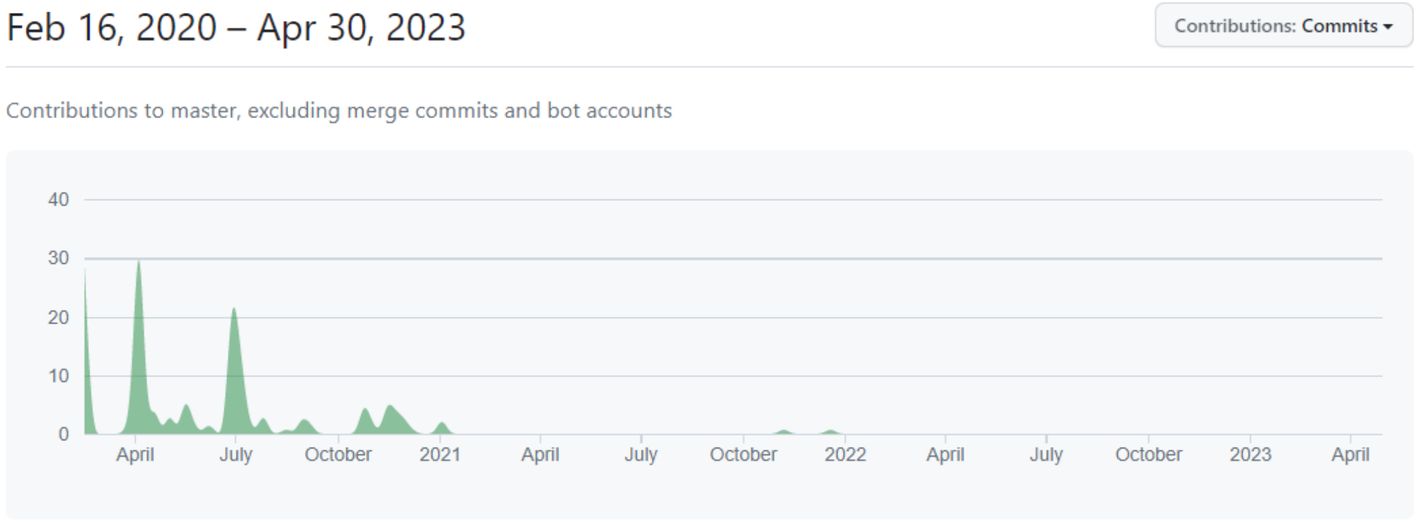
"Интуитивно понятная архитектура"

Кто угодно

Сопровождение







Запомним критерии

Итак, мы будем использовать вышеперечисленные 10 критериев для оценки крутости open-source библиотек.

"Всосали?"

© А так уже говорила моя учительница по математике

Пример "идеального" open-source

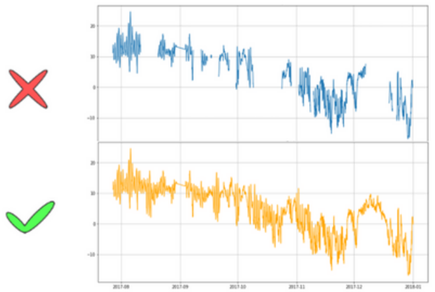
Большое сообщество пользователей и контрибьюторов	✓
Архитектура	✓
Тесты	✓
CI/CD	✓
Development guideline	✓
Простой процесс установки	✓
Простой интерфейс	✓
Понятная документация	✓
Сопровождение	✓
Много звезд на гитхабе	✓

SSGP-toolbox

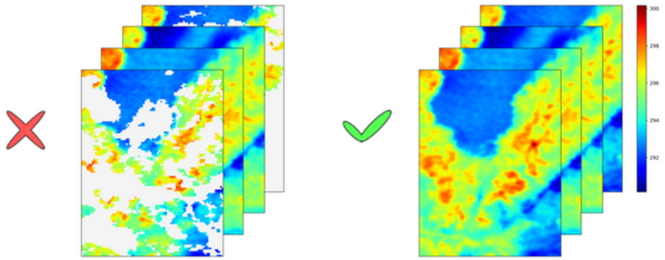
Библиотека для заполнения пропусков в спутниковых снимках при помощи машинного обучения

Проблема

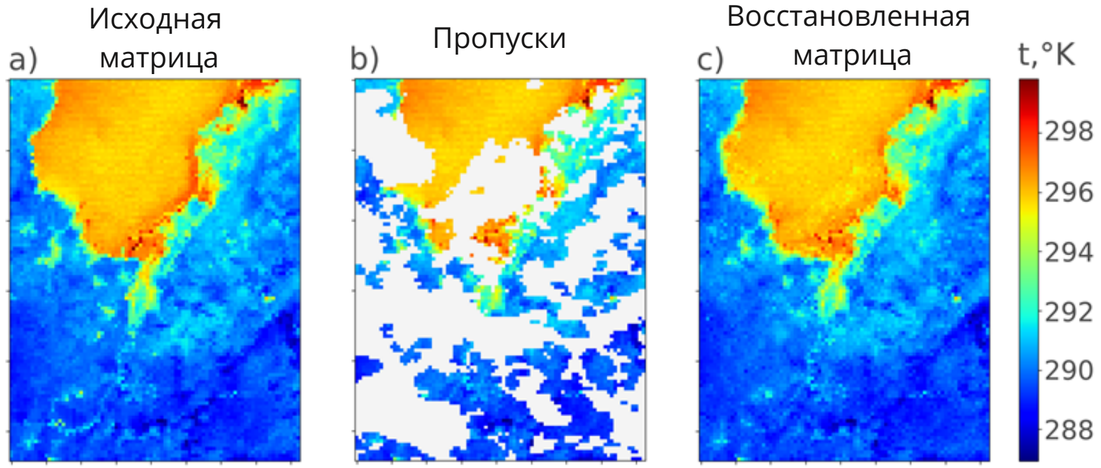
Временные ряды



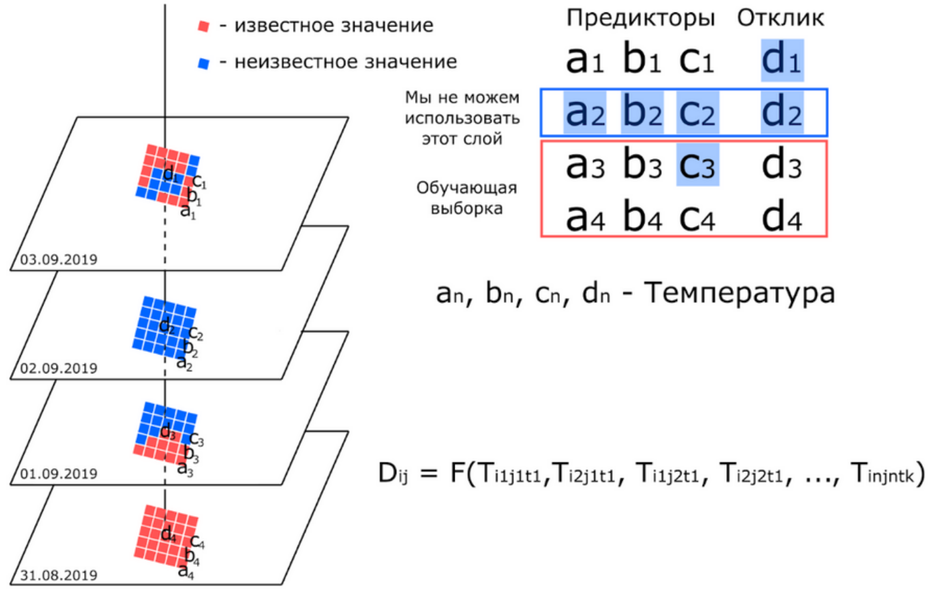
Многомерные массивы



Решение



Заполнение пропущенного значения в пикселе d_1



a_n, b_n, c_n, d_n - Температура

$D_{ij} = F(T_{i1j1t1}, T_{i2j1t1}, T_{i1j2t1}, T_{i2j2t1}, \ldots, T_{injntk})$

Что не так

Большое сообщество пользователей и контрибьюторов	✗
Архитектура	✓ ✗
Тесты	✗
CI/CD	✗
Development guideline	✗
Простой процесс установки	✗
Простой интерфейс	✓ ✗
Понятная документация	✓
Сопровождение	✓ ✗
Много звезд на гитхабе	✓

SSGP-toolbox
когда я его
проектировал

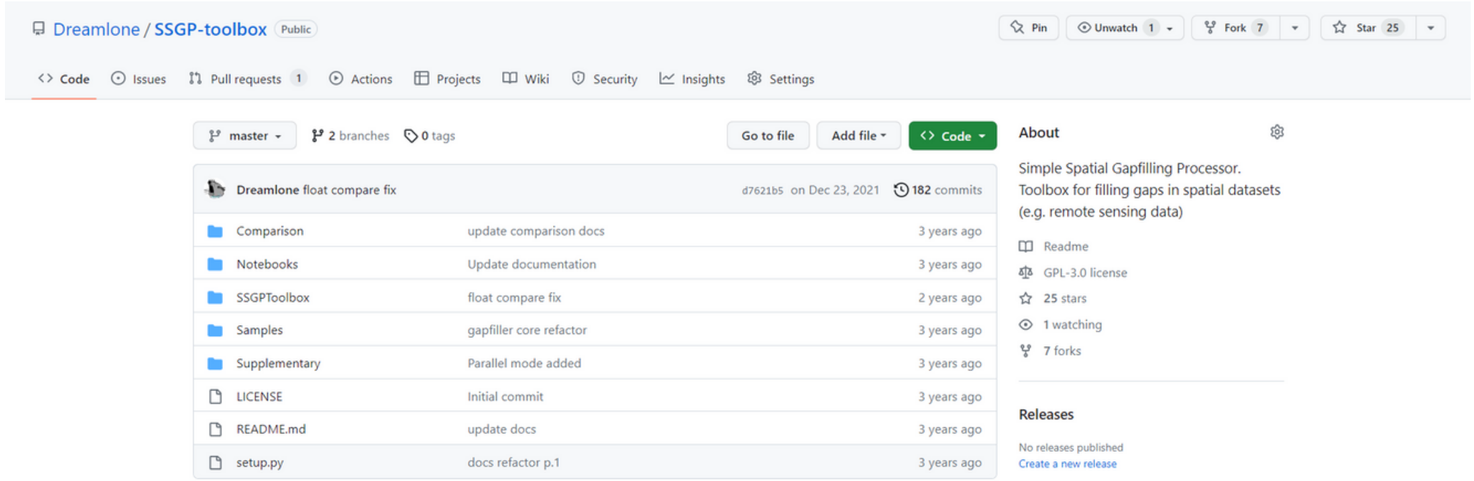
SSGP-toolbox
на момент "выпуска"
в 2020

SSGP-toolbox
когда я смотрю на него
спустя 2 года
(код не изменился за это время)

Документация в jupyter notebooks, никаких билдов на readthedocs и так далее, PEP8? - просто забудьте про стандарты - там их нету, декомпозиции кода почти нет, все держится на if else.

SimpleSpatialGapfillingProcessor-toolbox

Почему это все таки хороший модуль



Mikhail Sarafanov

Edit Delete

A Machine Learning Approach for Remote Sensing Data Gap-Filling with Open-Source Implementation: An Example Regarding Land Surface Temperature, Surface Albedo and NDVI [PDF] from mdpi.com

Authors	Mikhail Sarafanov, Eduard Kazakov, Nikolay O. Nikitin, Anna V. Kalyuzhnaya
Publication date	2020/11/25
Journal	Remote Sensing
Volume	12
Issue	23
Pages	3865
Publisher	MDPI
Description	Satellite remote sensing has now become a unique tool for continuous and predictable monitoring of geosystems at various scales, observing the dynamics of different geophysical parameters of the environment. One of the essential problems with most satellite environmental monitoring methods is their sensitivity to atmospheric conditions, in particular cloud cover, which leads to the loss of a significant part of data, especially at high latitudes, potentially reducing the quality of observation time series until it is useless. In this paper, we present a toolbox for filling gaps in remote sensing time-series data based on machine learning algorithms and spatio-temporal statistics. The first implemented procedure allows us to fill gaps based on spatial relationships between pixels, obtained from historical time-series. Then, the second procedure is dedicated to filling the remaining gaps based on the temporal dynamics of each pixel value. The algorithm was tested and verified on Sentinel-3 SLSTR and Terra MODIS land surface temperature data and under different geographical and seasonal conditions. As a result of validation, it was found that in most cases the error did not exceed 1 °C. The algorithm was also verified for gaps restoration in Terra MODIS derived normalized difference vegetation index and land surface broadband albedo datasets. The software implementation is Python-based and distributed under conditions of GNU GPL 3 license via public repository.
Total citations	Cited by 35

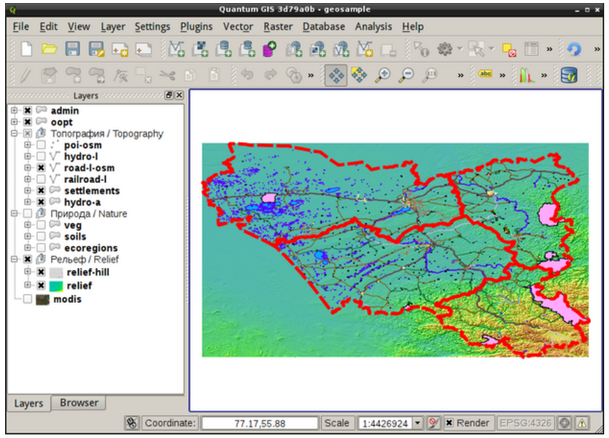
На репозиторий ссылается статья в Q1 журнале. На репозитории аж 25 звезд и 35+ цитирований у статьи

Периодически (раз в месяц) мне на почту пишут пользователи с вопросами. Пользователи - такие же исследователи и инженеры в области дистанционного зондирования Земли - целевая аудитория данного модуля. Им действительно регулярно пользуются

QGIS Lines ranking plugin

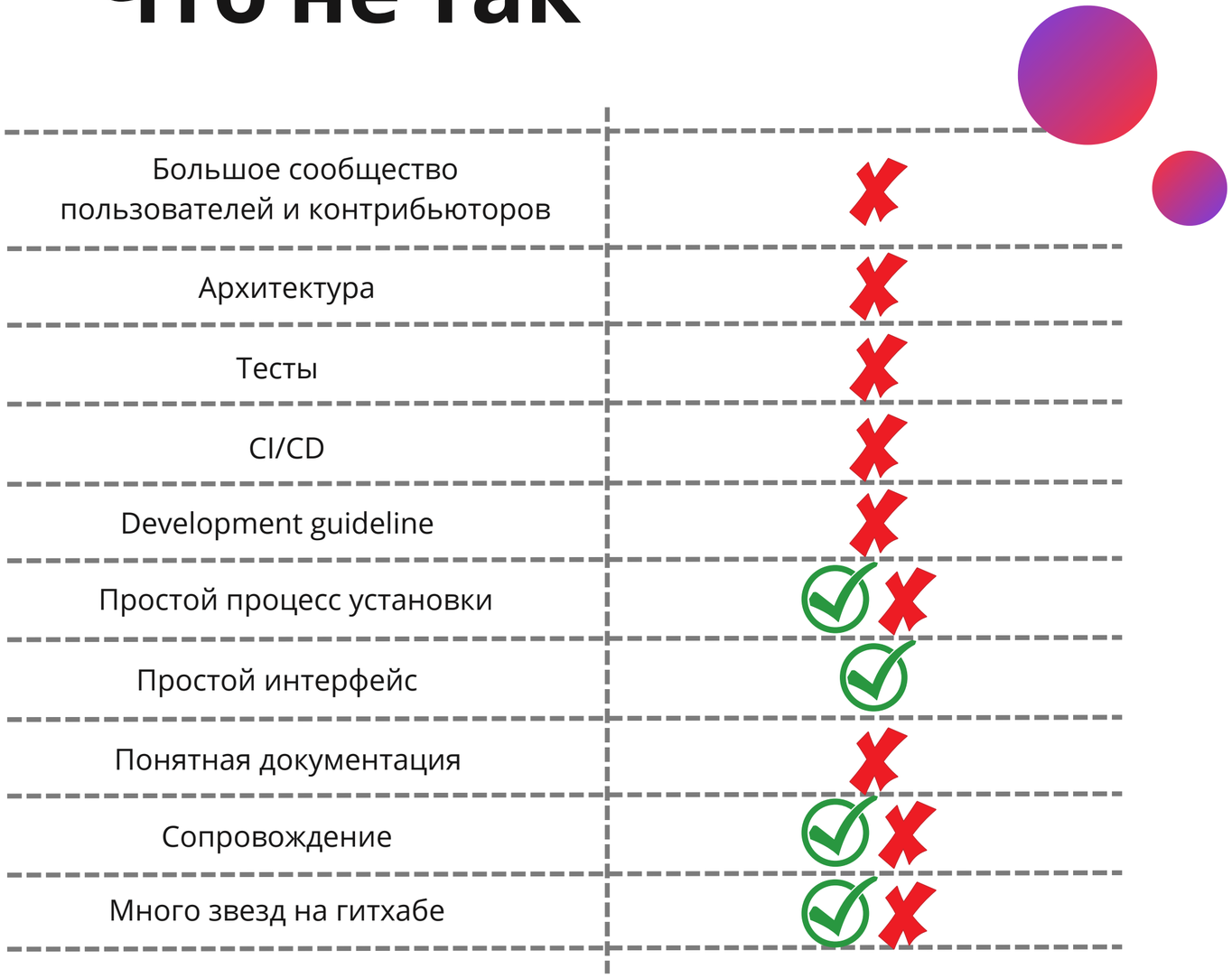
Lines Ranking

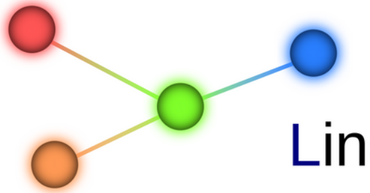
QGIS - open-source приложение для геоинформационного анализа



Lines ranking - позволяет ранжировать сегменты векторного слоя

Что не так





Lines Ranking

И что тогда на нас нашло...

Какие утилиты для версионирования кода вы используете

Код организован по принципу "как пришлось" и распихан на три большие (здоровенные) функции и обвязки для интеграции с QGIS. Тестов нет, ровно как и документации

Почему это все таки хороший модуль

Lines Ranking

QGIS Python Plugins Repository

Download latest **Lines Ranking**

(3) votes

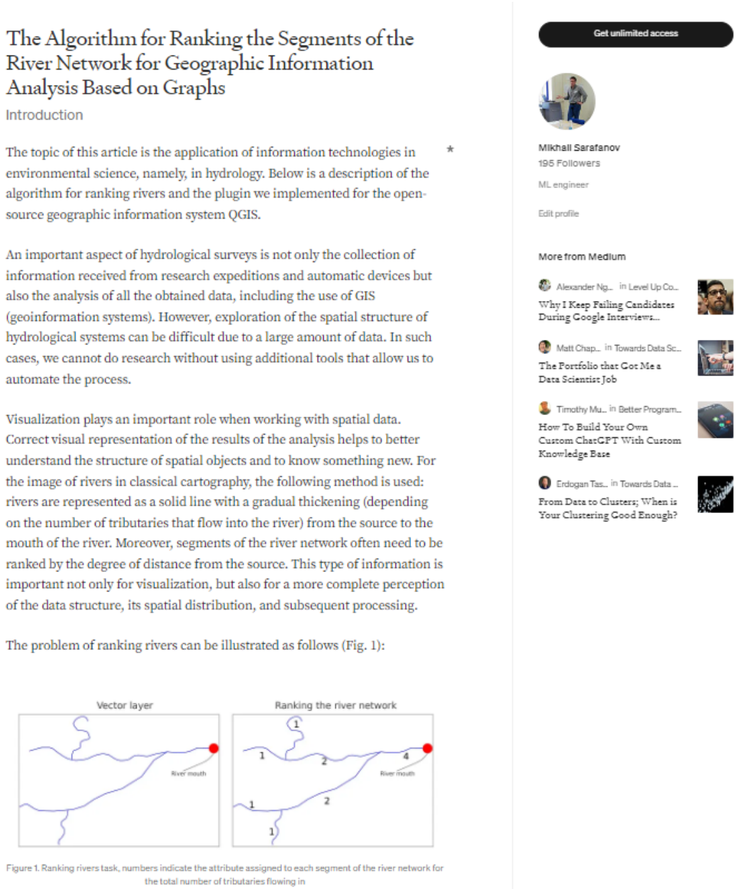
> Plugin for ranking lines features based on the position of starting point

About | Details | Versions

Version	Experimental	Minimum QGIS version	Downloads	Uploaded by	Date
1.0	no	3.14.0	2001	chrislisbon	30 июл. 2020 г., 12:52 GMT+3

Потому что им пользуются, и иногда даже скачивают. Пользователи пишут вопросы на почту.

А ещё мы научились писать плагины для QGIS и статьи на хабр и TowardsDataScience на интересные нам темы и сделали первые шаги в продвижении open-source продуктов



The Algorithm for Ranking the Segments of the River Network for Geographic Information Analysis Based on Graphs

Introduction

The topic of this article is the application of information technologies in environmental science, namely, in hydrology. Below is a description of the algorithm for ranking rivers and the plugin we implemented for the open-source geographic information system QGIS.

An important aspect of hydrological surveys is not only the collection of information received from research expeditions and automatic devices but also the analysis of all the obtained data, including the use of GIS (geoinformation systems). However, exploration of the spatial structure of hydrological systems can be difficult due to a large amount of data. In such cases, we cannot do research without using additional tools that allow us to automate the process.

Visualization plays an important role when working with spatial data. Correct visual representation of the results of the analysis helps to better understand the structure of spatial objects and to know something new. For the image of rivers in classical cartography, the following method is used: rivers are represented as a solid line with a gradual thickening (depending on the number of tributaries that flow into the river) from the source to the mouth of the river. Moreover, segments of the river network often need to be ranked by the degree of distance from the source. This type of information is important not only for visualization, but also for a more complete perception of the data structure, its spatial distribution, and subsequent processing.

The problem of ranking rivers can be illustrated as follows (Fig. 1):

Vector layer | Ranking the river network

River mouth

Figure 1. Ranking rivers task, numbers indicate the attribute assigned to each segment of the river network for the total number of tributaries flowing in

Get unlimited access

Mikhail Sarafanov
195 Followers
ML engineer
Edit profile

More from Medium

Alexander Ng... in Level Up Co...
Why I Keep Failing Candidates During Google Interviews...

Matt Chap... in Towards Data Sc...
The Portfolio that Got Me a Data Scientist Job

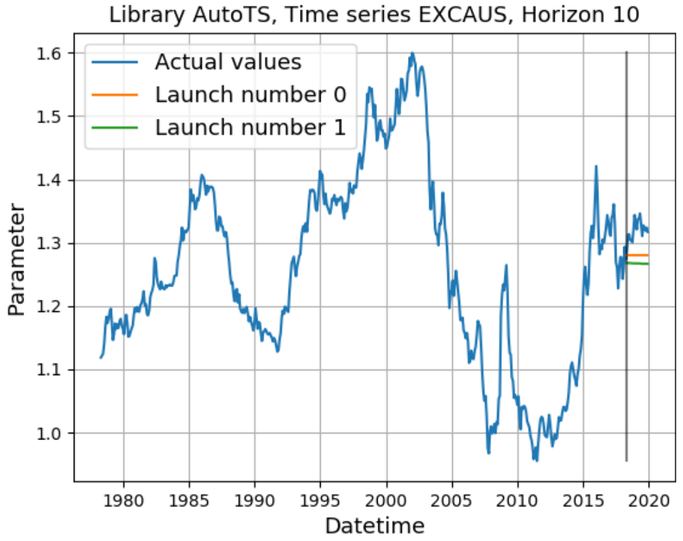
Timothy Mu... in Better Program...
How To Build Your Own Custom ChatGPT With Custom Knowledge Base

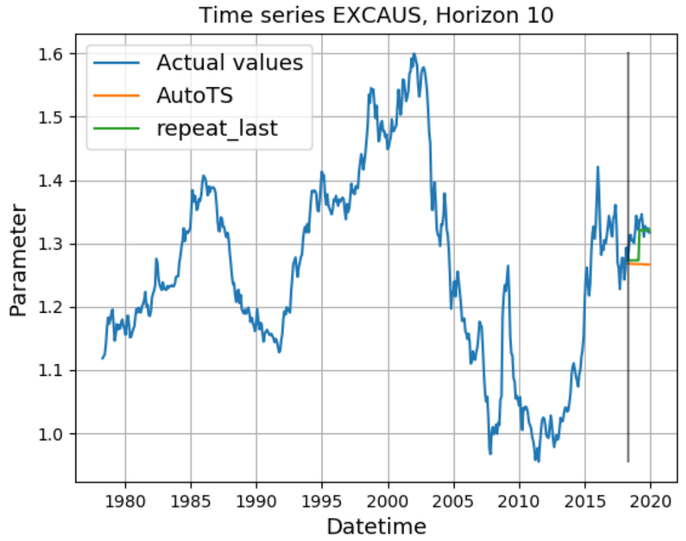
Erdogan Tas... in Towards Data ...
From Data to Clusters; When is Your Clustering Good Enough?

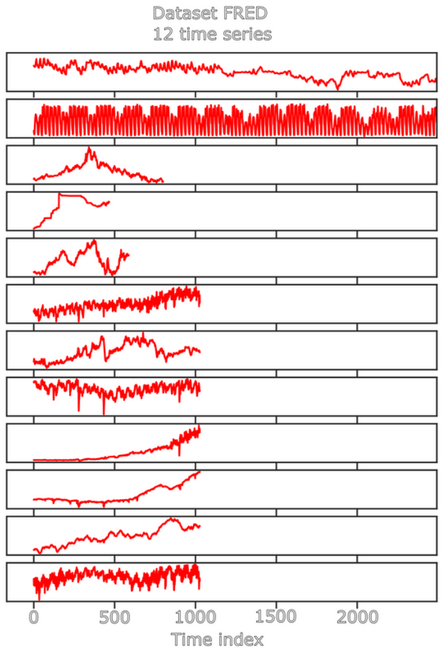
pytsbe

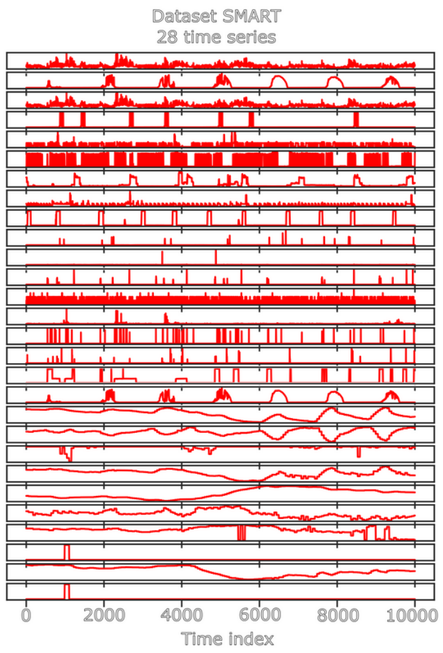
Open-source модуль для бенчмаркинга алгоритмов прогнозирования временных рядов

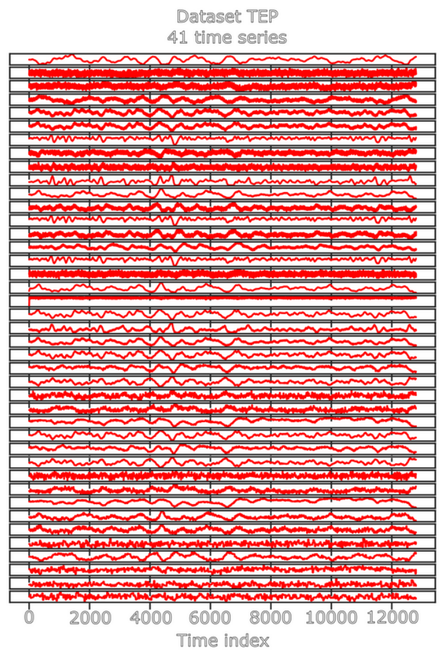
Позволяет запускать несколько алгоритмов прогнозирования временных рядов, измеряет различные метрики и готовит визуализации и сводные таблицы









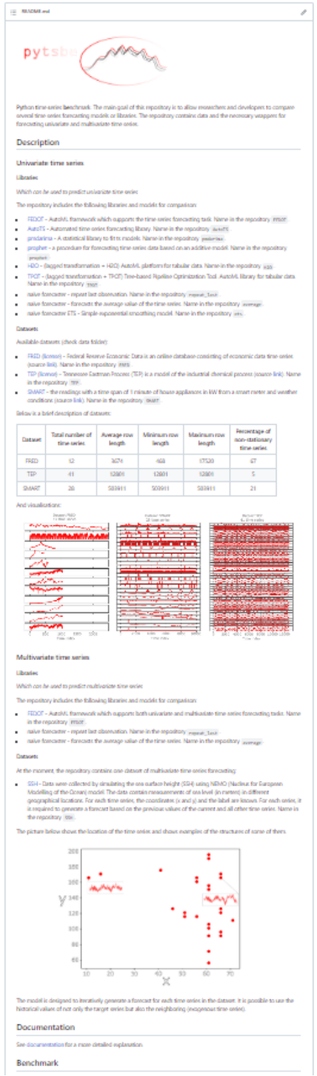


Что не так

Большое сообщество пользователей и контрибьюторов	✗
Архитектура	✓
Тесты	✓
CI/CD	✗
Development guideline	✓
Простой процесс установки	✓ ✗
Простой интерфейс	✓
Понятная документация	✓
Сопровождение	✓ ✗
Много звезд на гитхабе	✓

Документация держится просто в markdown файлах. Модуль так и не "выставлен" в PyPI, половина заявленной функциональности не реализована

Почему это все таки хороший модуль



- Документация (и README в том числе) на самом деле не так плоха
- Руководство для разработчиков "прошло проверку" - без моей помощи разработчики смогли включить дополнительный субмодуль в библиотеку
- Архитектура позволяет быстро и относительно безболезненно расширять функциональность
- Тесты "не хрупкие" и легко модифицируются

- Модуль уже выполнил свое предназначение - позволил провести эксперименты с алгоритмами прогнозирования временных рядов для внутренних исследований лаборатории
- Представляет собой своеобразное "наследие" после моего ухода из команды. Наработки и уже написанные обвязки не пропали, а существуют в виде изолированного модуля и при необходимости с полупинка могут быть запущены бывшими коллегами

Заключение

На этом всё

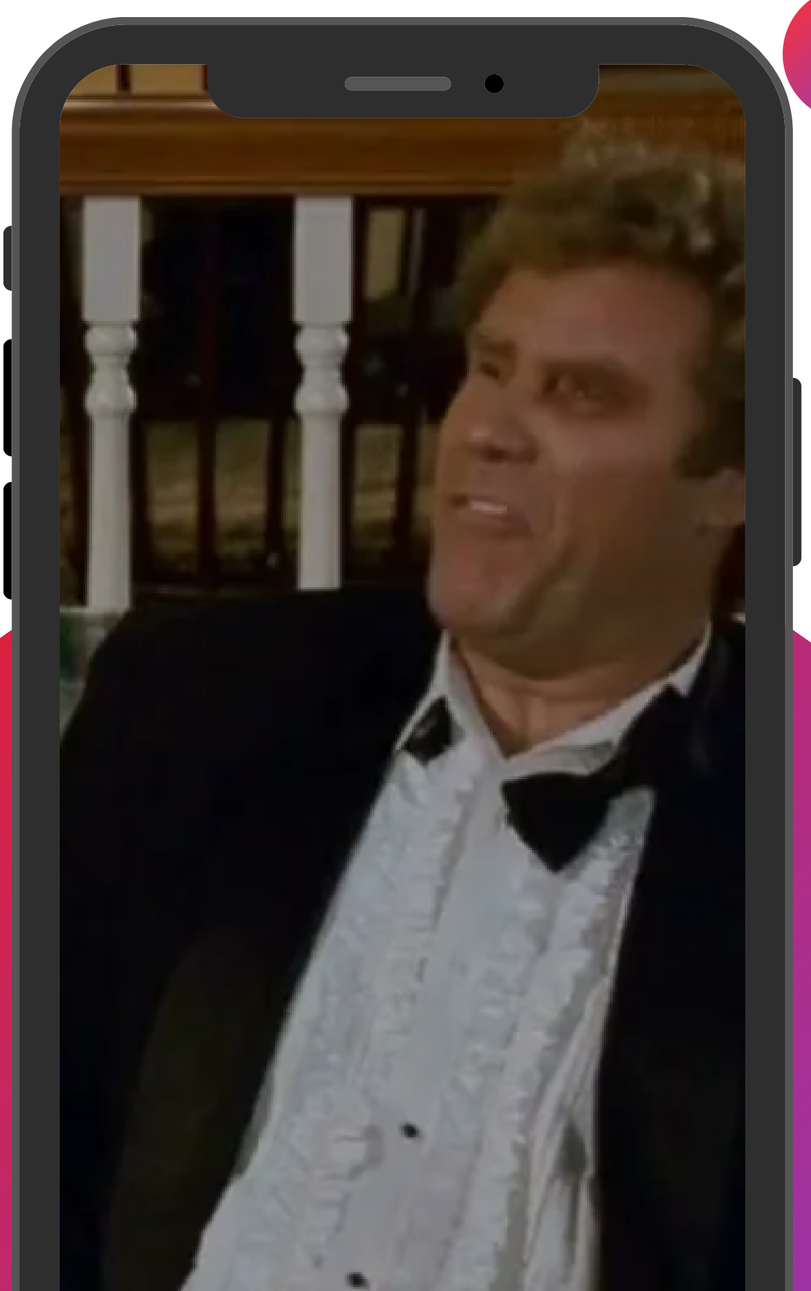
Иногда страшно выкладывать свои разработки в open-source потому что они кажутся незаконченными, слишком маленькими и неказистыми

Кажется, что open-source проекты должны быть большими, с хорошей документацией и активным сообществом

Кажется, что open-source проекты будут отнимать у вас много времени на сопровождение и не смогут должным образом развиваться

Но иногда open-source решениям стоит "быть" просто потому что они сделаны с любовью ❤

*надеюсь, вам понравился рассказ про некоторые из таких



DATAFEST 2023

Спасибо за внимание!

mik.sarafanov@gmail.com