# Повышаем воспроизводимость исследований в AI/ML с помощью опенсорса

Никитин Николай

лидер движения ITMO.OpenSource,

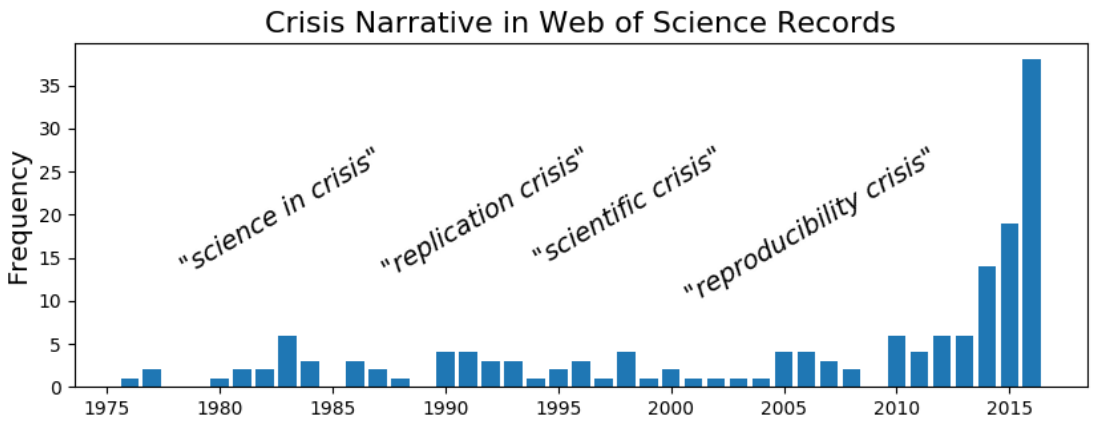
руководитель лаборатории автоматического машинного обучения, ИТМО

# Немного о ИИ, воспроизводимости, и открытом коде

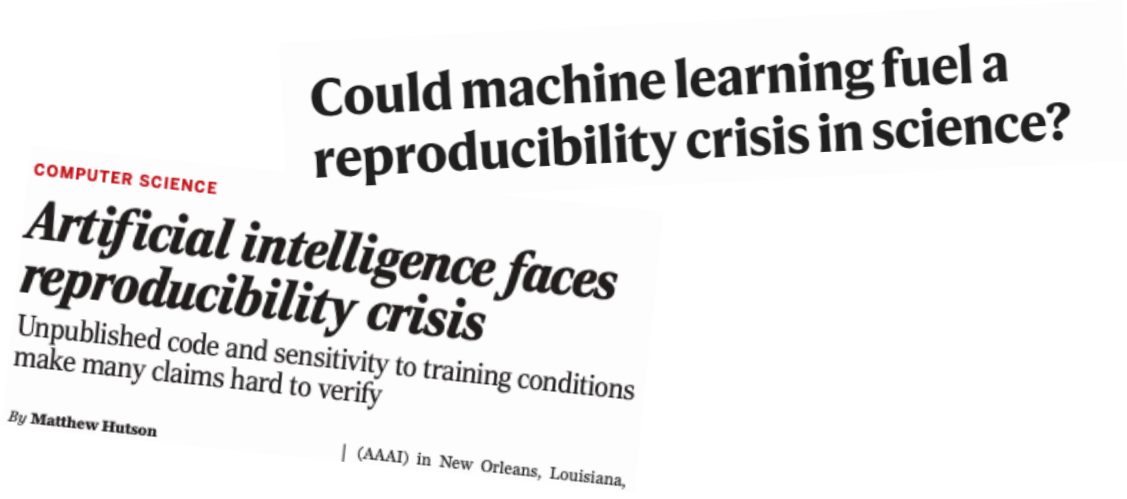
# Кризис воспроизводимости

- **Что хорошего**: темпы прогресса в AI/ML очень высоки.

- **Что плохого**: огромную долю новых результатов очень сложно проверить, воспроизвести, применить на практике. Это сильно снижает их полезность.




«Раньше было лучше»


Could machine learning fuel a reproducibility crisis in science?

COMPUTER SCIENCE

Artificial intelligence faces reproducibility crisis

Unpublished code and sensitivity to training conditions make many claims hard to verify

By Matthew Hutson

(AAAI) in New Orleans, Louisiana,

AI и ML – зона риска

# Что считается результатом?

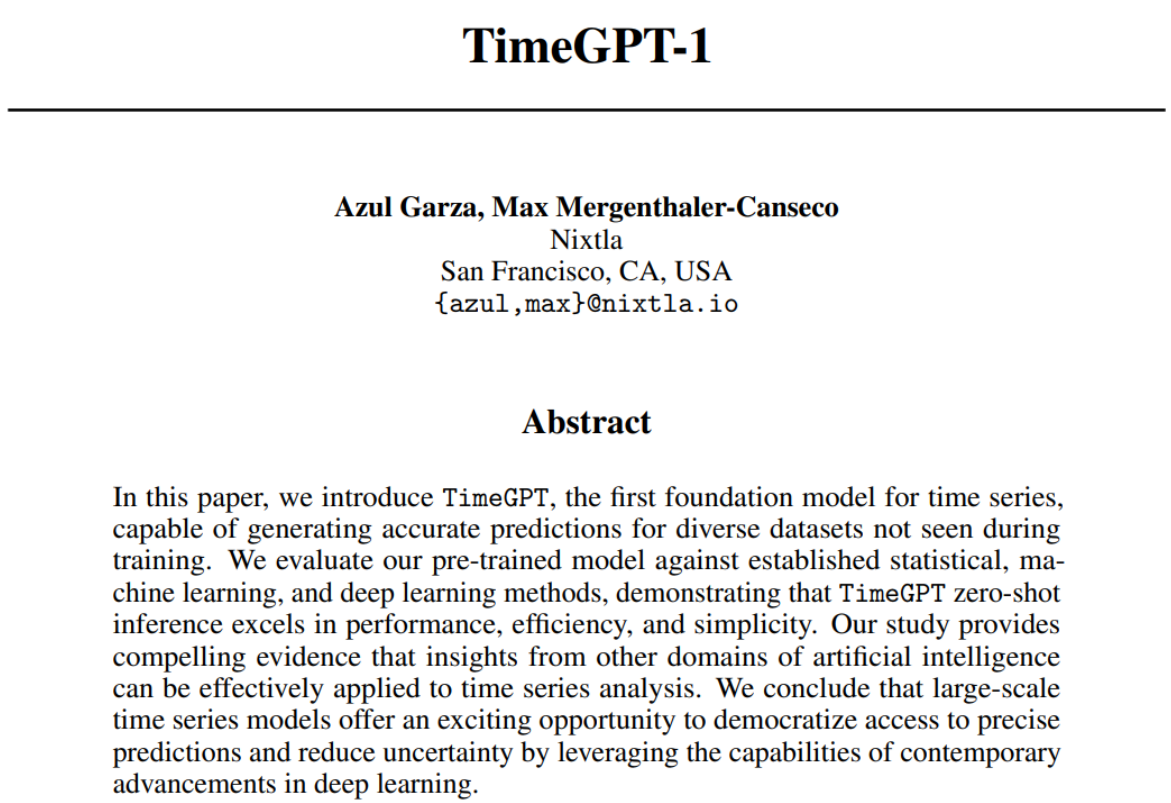
Что обычно хотят от интересного научного результата в области AI/ML:

- Проверить полученные метрики на правдивость;
- Сравнить с аналогами на новых данных;
- Применить для решения своих задач.

# Сложившиеся практики

Как выглядит результат:

- Основной артефакт — статья или препринт.
- В дополнение к ним авторы могут приложить код и данные (а могут и не приложить).



**TimeGPT-1**

**Azul Garza, Max Mergenthaler-Canseco**
Nixtla
San Francisco, CA, USA
{azul,max}@nixtla.io

**Abstract**

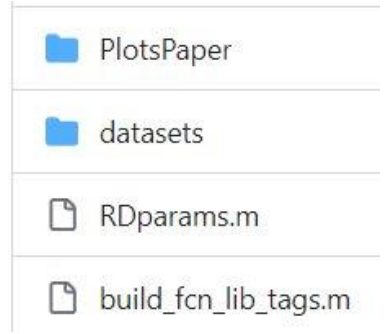
In this paper, we introduce TimeGPT, the first foundation model for time series, capable of generating accurate predictions for diverse datasets not seen during training. We evaluate our pre-trained model against established statistical, machine learning, and deep learning methods, demonstrating that TimeGPT zero-shot inference excels in performance, efficiency, and simplicity. Our study provides compelling evidence that insights from other domains of artificial intelligence can be effectively applied to time series analysis. We conclude that large-scale time series models offer an exciting opportunity to democratize access to precise predictions and reduce uncertainty by leveraging the capabilities of contemporary advancements in deep learning.

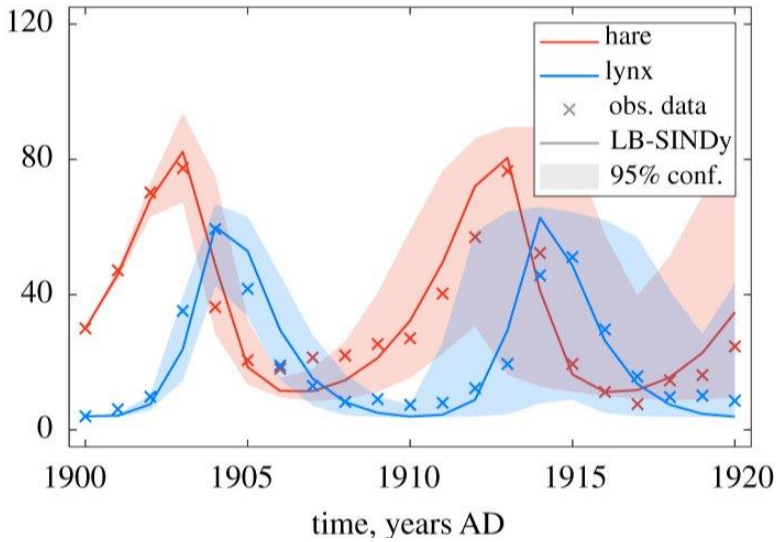
Пример свежей и интересной статьи — без кода, но с доступом к модели через API.

# Реальный пример

Статья: про обучение моделей на основе дифференциальных уравнений по данным.

К ней приложен код — скрипты на языке MATLAB:



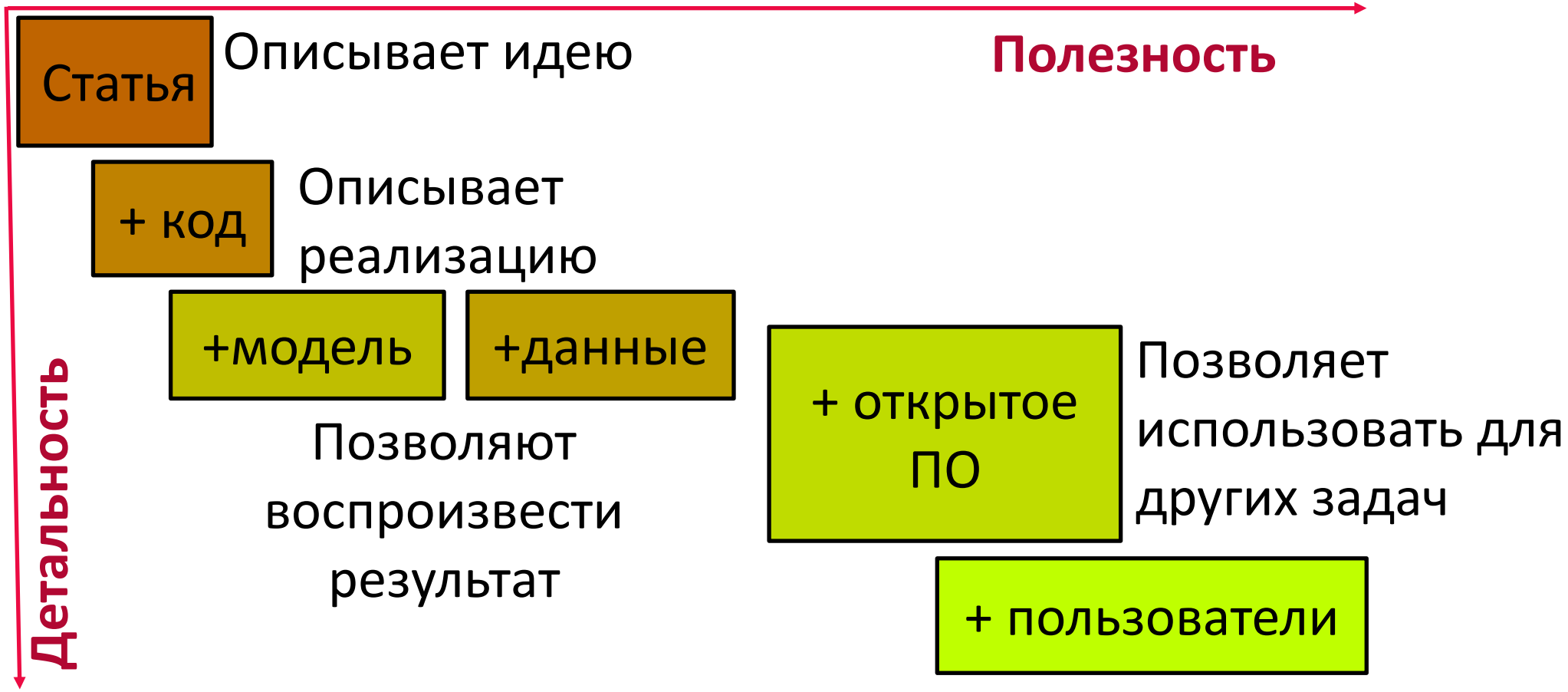

Код позволяет воспроизвести простой эксперимент

| PDE | noise level | form | model error WSINDy \| E-WSINDy | success rate WSINDy \| E-WSINDy |
|---|---|---|---|---|
| inviscid Burgers | 100% | $u_t + 0.5uu_x = 0$ | 2.6% \| **2.5%** | 99% \| **100%** |
| Korteweg de Vries | 100% | $u_t + 0.5uu_x + u_{xxx} = 0$ | 27.5% \| **4.0%** | 93.5% \| **100%** |
| nonlinear Schrödinger | 50% | $iu_t + 0.5u_{xx} + \lvert u\rvert^2 u = 0$ | 13.0% \| **11.3%** | 82.0% \| **100%** |
| Kuramoto–Sivashinsky | 100% | $u_t + 0.5uu_x + u_{xx} + u_{xxxx} = 0$ | 29.5% \| **24.7%** | 87.5% \| **99.5%** |
| reaction–diffusion | 20% | $u_t = 0.1\nabla^2 u + \lambda(A)u - \omega(A)v$<br>$v_t = 0.1\nabla^2 v + \omega(A)u + \lambda(A)v$<br>$A^2 = u^2 + v^2$, $\omega = -\beta A^2$, $\lambda = 1 - A^2$ | 77.7% \| **7.1%** | 0.0% \| **99.5%** |

А для более сложного — кода нет.

Итог: польза от статьи на порядок меньше, чем могла бы быть

# А как хочется?



# Возможен ли идеал?

ИТМО

Что объективно не позволяет повышать воспроизводимость:

- Коммерческие ограничения (NDA, лицензии)
- Объемы данных (например, нужные для обучения LLM)
- Спешка ("Publish or perish") при написании и рецензировании
- Невозможность фиксации всех влияющих факторов
- Разработка ПО — не основная деятельность научных команд

# Как улучшить положение дел?

ИТМО

Что уже есть в мире:

- На хороших конференциях требуют прикладывать к статьям код
- Есть агрегаторы воспроизводимых публикаций https://paperswithcode.com
- Есть специализированные репозитории научных данных
- Ряд топовых журналов имеют Open Source-треки

Чего хотелось бы:

- Внедрения культуры открытого кода, данных и документации в научную среду
- Развития сообщества научного Open Source — как на международном уровне, так и внутри страны.

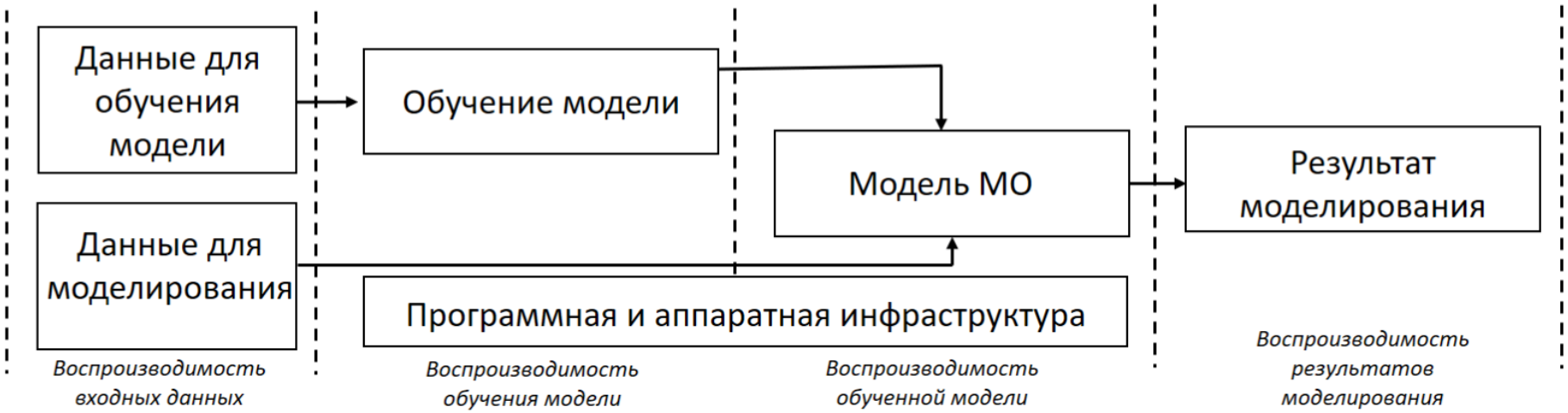
# Как открывать обученные модели?

Что нужно сделать открытым, чтобы воспроизвести модель ИИ:

- входные данные для обучения,
- алгоритм обучения,
- структуру и веса обученной модели,
- описание программной и аппаратной инфраструктуры,
- результаты работы модели на тестовых данных.

<u>В идеале</u> – так:

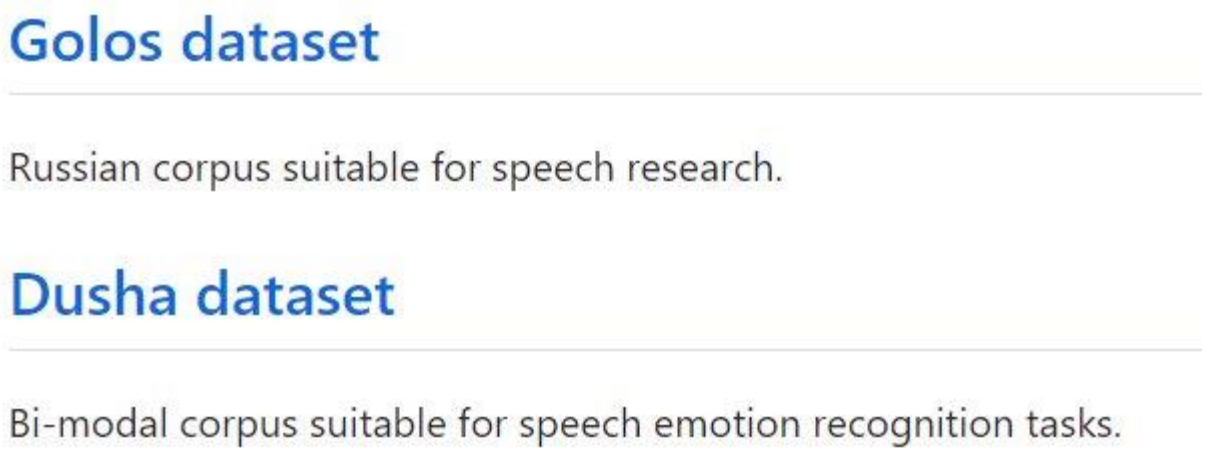
<u>По факту:</u> – есть нюансы



# Как открывать данные?

Вывести набор данных в Open Source — даже полезнее, чем код.

Пример открытых данных для **обучения моделей**:




https://github.com/salute-developers/golos

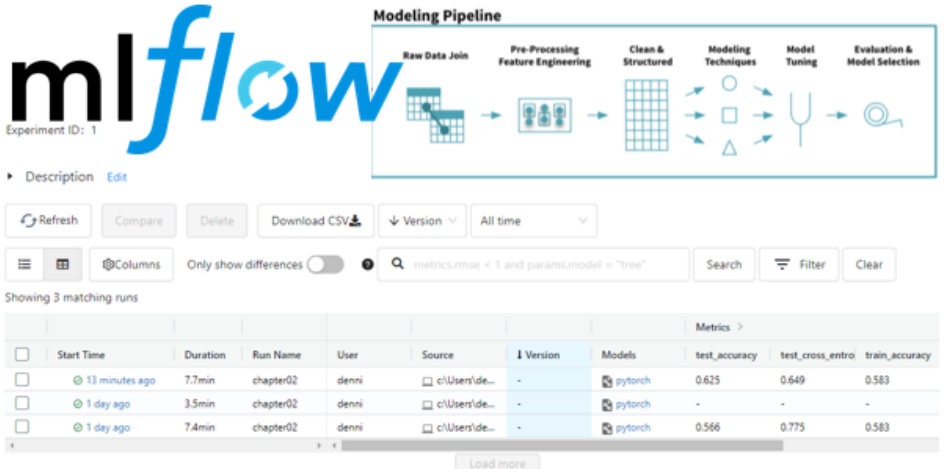
Пример открытых данных для **сравнения моделей** (бенчмарк):


https://github.com/waico/SKAB

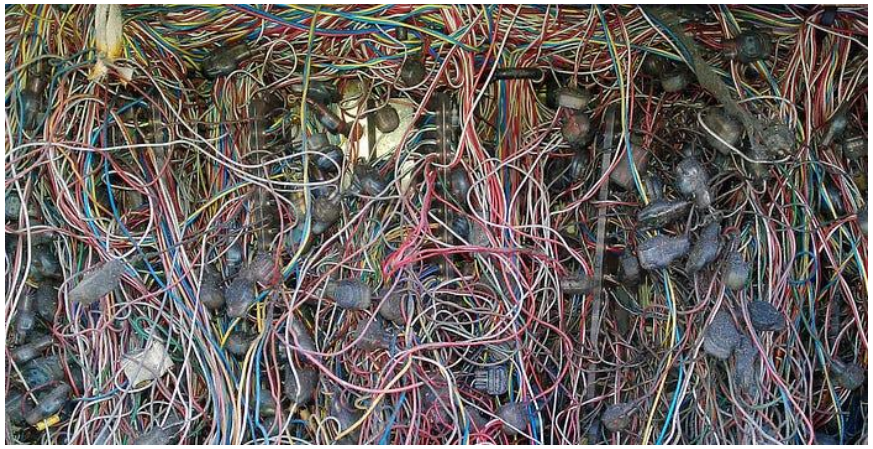
# Как управлять экспериментами?

Что ещё поможет воспроизводимости экспериментов  ИИ:

- Совместимость со стандартными инструментами MLOps — например, mlflow.
- Контейнеризация экспериментов.
- Версионирование моделей и данных
- Детальное логирование.
- Избегание проприетарных инструментов.
- Быть осторожными с использованием Jupyter Notebook (не по назначению).




Иногда workflow выглядит так.



А иногда — так.

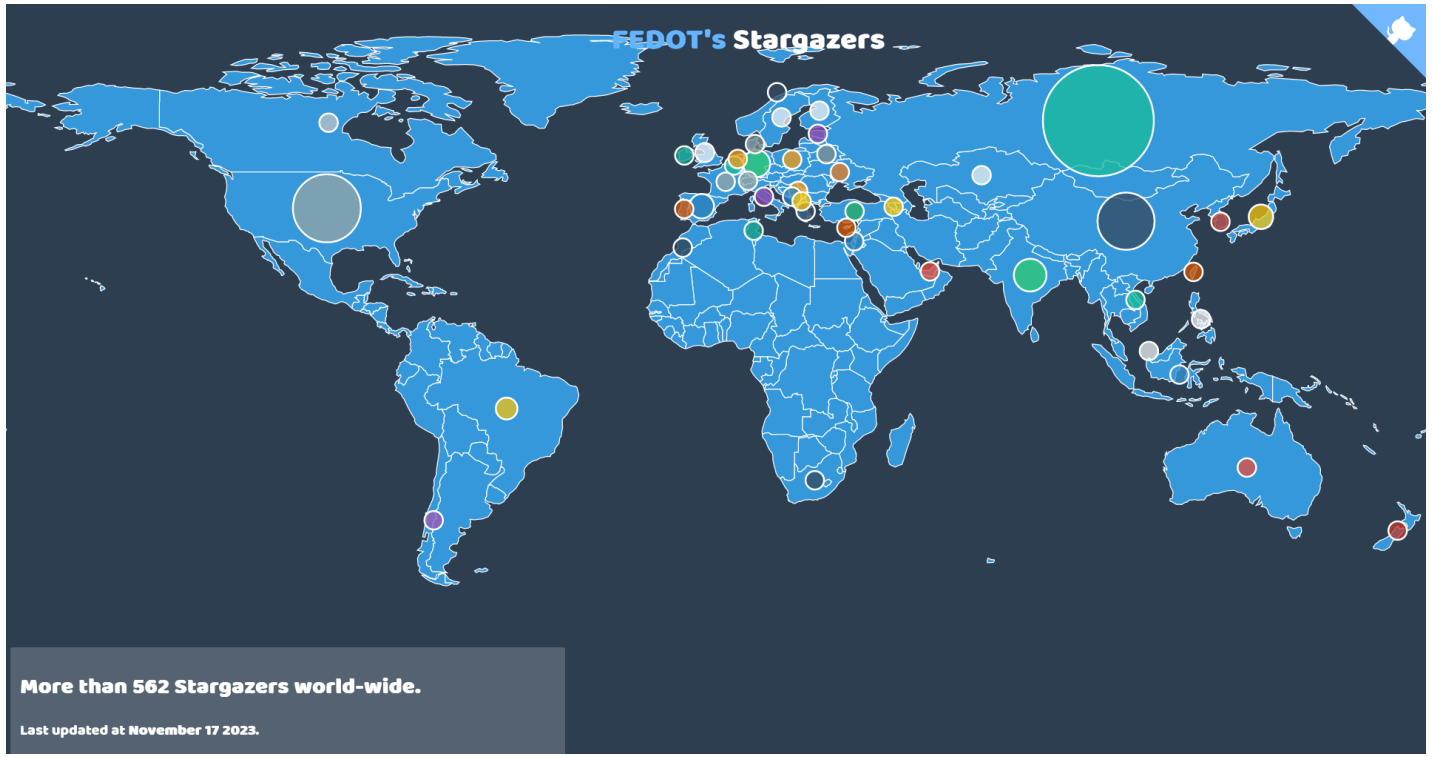
# Популяризация разработок

Публикации:

- Хабр, Towards Data Science;
- Youtube и другие видеохостинги;
- tg-чаты и каналы.

Мероприятия:

- HighLoad++: Open Source-трибуна;
- ODS, AI Journey, etc.;
- Научные публикации и конференции;
- Хакатоны, Kaggle.




Наш опыт: разработанные инструменты используют по всему миру

# Немного об опыте ИТМО

# С чего мы начинали?

ИТМО

В 2020 году мы имели:

Одиночные открытые библиотеки;

Минимум внешних участников;

Отсутствие популяризации решений;

Что сделали:

запустили ряд проектов по созданию Open Source инструментов в области ML (как крупномасштабных, так и студенческих).


Суть

# К чему пришли?

**Наш вызов**: продвигать культуру научного открытого кода

| **Из чего состоит** | **Кто использует** | **Востребованность** | **Зачем?** |
|---|---|---|---|
| 30+ фреймворков и библиотек в области ИИ, готовые к использованию, снабженные тестами и документацией. | Разработанные инструменты используют в десятках стран - в том числе сотрудники NVIDIA, Uber, Google Cloud, JetBrains, Databricks. | 1300+ звезд, 300к скачиваний (ТОП 1 в РФ среди университетов). | • Обучение профессионалов;<br>• Вдохновение тех, кто участвует;<br>• Технологический суверенитет. |

**Репозиторий**: https://github.com/aimclub

# Что нового в экосистеме за 2023?

**Код**:
- >80% существующих библиотек и фреймворков значительно обновлены.
- Суммарное число «звезд» - выросло на 30%.
- Выложено в общий доступ 9 новых библиотеки и фреймворков.
- Развитие взаимной интеграции решений экосистемы.

**Сообщество**:
- Создано сообщество открытого кода ITMO.Opensource (650+ участников), а также одноименный студенческий клуб.
- Проведено 6 очных встреч-митапов для сообщества.
- Подготовлено 15 руководств по созданию открытых проектов.

**Наука**:
Новые методы и алгоритмы в разных областях – от «классического» ML до экзотики.

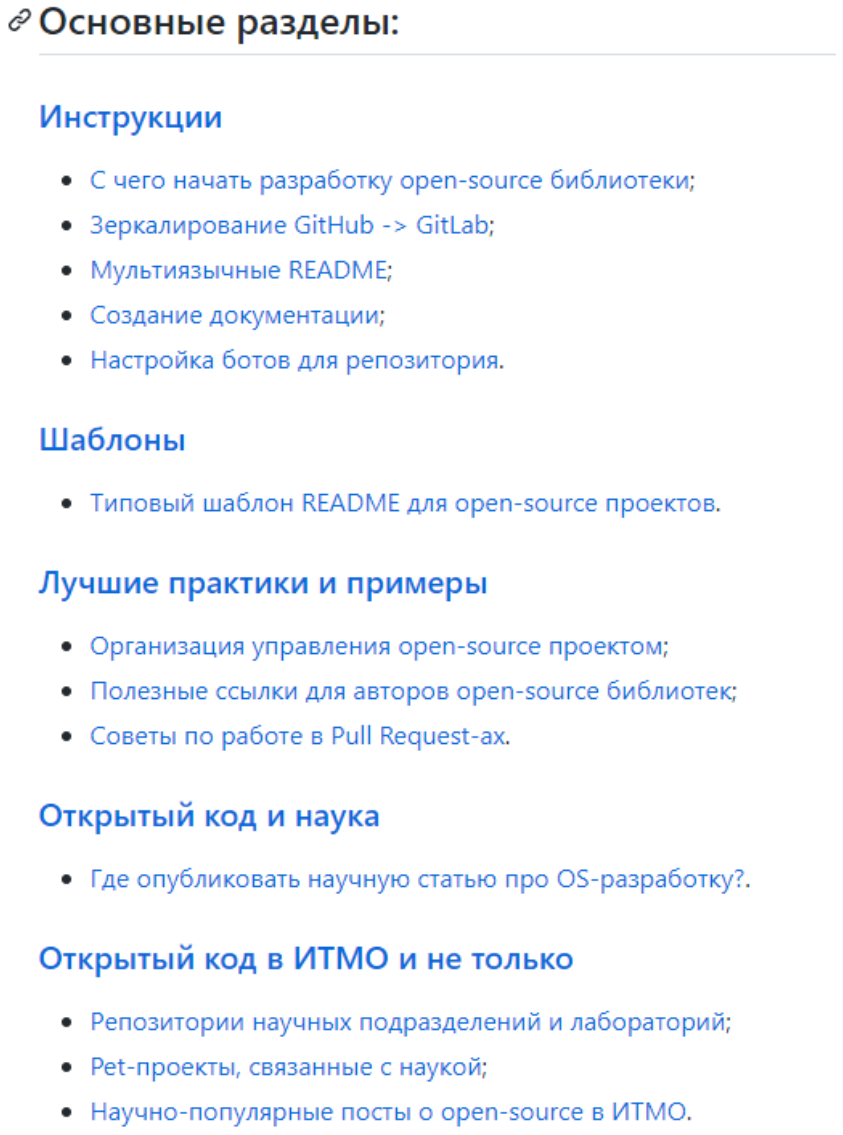
# Как помочь «вкатывающимся»?

ИТМО

У научных команд часто возникают типовые проблемы:

- Как написать понятный readme?
- Как и зачем реализовать автотесты?
- Где и как можно рассказать о своём проекте?

Поэтому мы создали репозиторий с лучшими практиками, руководствами и шаблонами в области Open Source.



Основные разделы:

Инструкции

- С чего начать разработку open-source библиотеки;
- Зеркалирование GitHub -> GitLab;
- Мультиязычные README;
- Создание документации;
- Настройка ботов для репозитория.

Шаблоны

- Типовый шаблон README для open-source проектов.

Лучшие практики и примеры

- Организация управления open-source проектом;
- Полезные ссылки для авторов open-source библиотек;
- Советы по работе в Pull Request-ах.

Открытый код и наука

- Где опубликовать научную статью про OS-разработку?.

Открытый код в ИТМО и не только

- Репозитории научных подразделений и лабораторий;
- Pet-проекты, связанные с наукой;
- Научно-популярные посты о open-source в ИТМО.

# Открытость и воспроизводимость — где только можно

Мероприятия тоже могут проводиться в Open Source-стиле.

Наш пример: для материалов конференции International Young Scientists Conference 2023 создан отдельный репозиторий.

## Deep Learning and Data-Driven Modelling

Section Papers 13 | Papers with Open Code 9

| ID | Title | Links |
|---|---|---|
| 06 | Light-weight ensembling of deep neural models for object recognition in remote sensing data | Code Data |
| 15 | Convolutional Neural Network Graph-based Embedding for Neural Architecture Search | Code |
| 16 | FAMLINN: Representation for Storing Neural Network Architecture | Code |
| 18 | Multimodal prediction of profanity based on speech analysis | Code |
| 24 | Proactive selection of machine learning models for small sample sizes based on PAC-learning theory | Code |
| 31 | CIS Multilingual License Plate Detection and Recognition Based on Convolutional and Transformer Neural Networks | Code Data |
| 32 | Predicting dataset size for neural network fine-tuning with a given quality in object detection task | Code |
| 35 | Forecasting of Sea Ice Concentration using CNN, PDE discovery and Bayesian Networks | Code |
| 45 | mHAR: a novel convolutional recurrent model for recognizing motion-based human activity | Code Data |

https://github.com/itmo-ai/YSC-2023-Papers

# Как поучаствовать?

| **Аспекты культуры Open-Source** | **Пример** |
|---|---|
| Open-Source – это не только про код | Конференции, семинары (ещё один пример от нас - https://github.com/ITMO-NSS-team/scientific-seminars), датасеты – все может быть открытым. |
| Участие в сообществе | На наших мероприятиях мы даем возможность рассказать о своих проектах, найти участников, поделиться трудностями и успехами. |
| Менторство | Помощь более опытных участников – начинающим. Мы реализуем это в рамках AIM. |
| Образование | Форматы «диплом как открытый код», open-source хакатоны, участие в реальных проектах. |

Приглашаем вступать в наше сообщество ITMO.Opensource!

# Спасибо за внимание

ITMO more than a UNIVERSITY

nnikitin@itmo.ru