



Факультет компьютерных наук НИУ ВШЭ

XIX конференция «Свободное
программное обеспечение в
высшей школе»

28 – 30 июня 2024
г. Переславль-Залесский

Проект открытого кода научных исследований ФКН

Гущин Михаил, mhushchyn@hse.ru

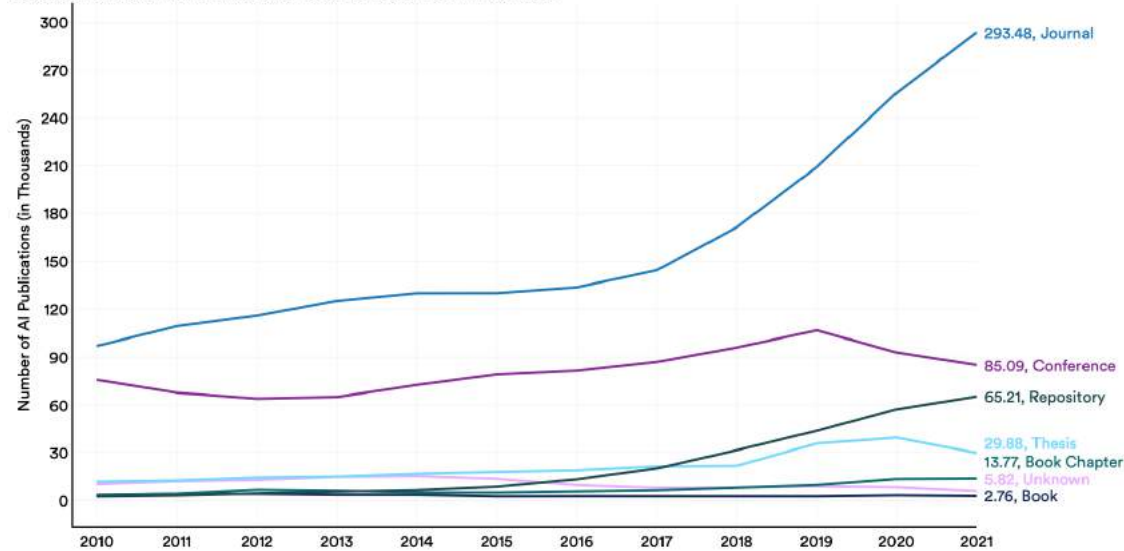
ФКН НИУ ВШЭ



Публикации и открытый код по ИИ

Number of AI Publications by Type, 2010–21

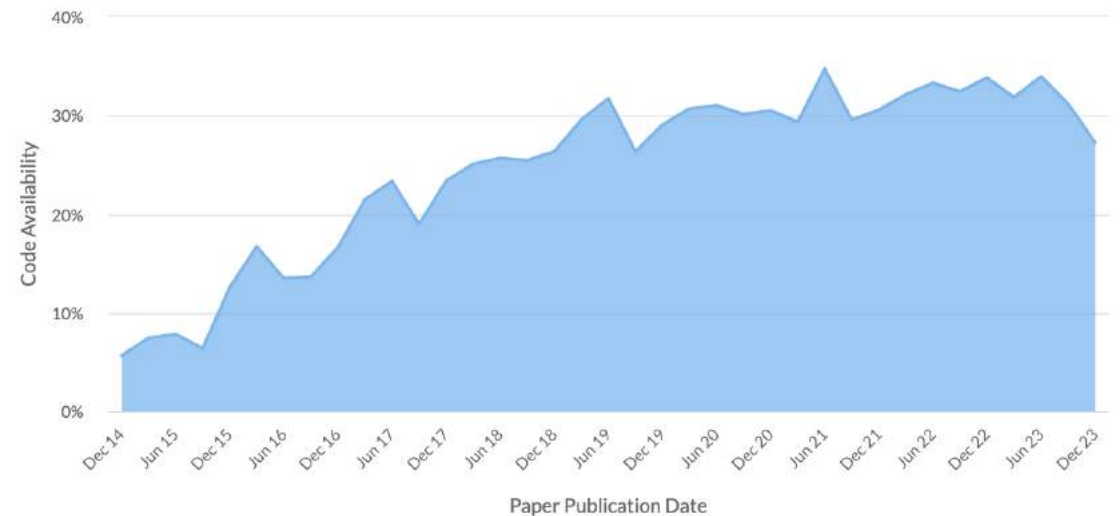
Source: Center for Security and Emerging Technology, 2022 | Chart: 2023 AI Index Report



Более **290 000** публикаций в научных журналах в области ИИ за 2021 год согласно 2023 AI Index Report (Stanford)

Code Availability

Percentage of published papers that have at least one code implementation



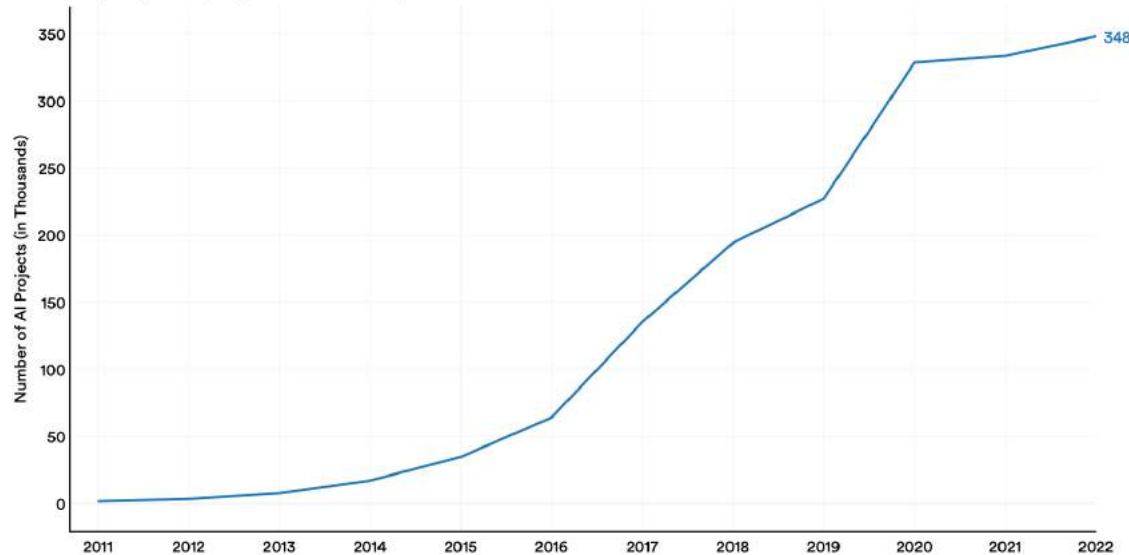
Более **30%** публикаций имеют хотя бы одну открытую реализацию в коде, согласно paperswithcode.com/trends



Проекты по ИИ с открытым кодом на GitHub

Number of GitHub AI Projects, 2011–22

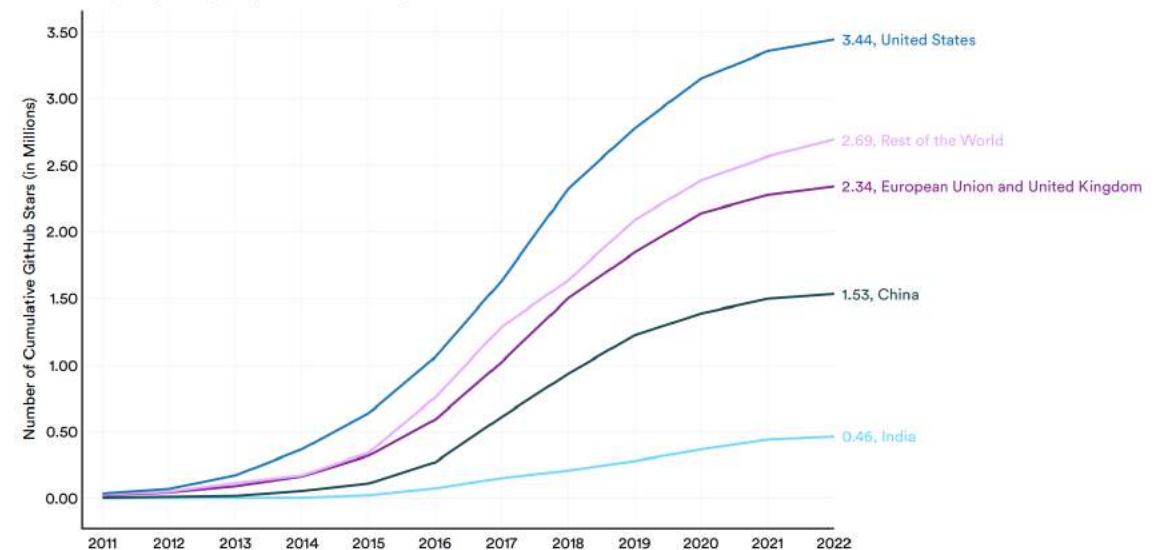
Source: GitHub, 2022; OECD.AI, 2022 | Chart: 2023 AI Index Report



Более **330 000** проектов с открытым кодом в области ИИ за 2021 год согласно 2023 AI Index Report (Stanford) (против **290 000** публикаций в журналах)

Number of GitHub Stars by Geographic Area, 2011–22

Source: GitHub, 2022; OECD.AI, 2022 | Chart: 2023 AI Index Report



Более **8 000 000** звезд в проектах в области ИИ к 2022 году согласно 2023 AI Index Report (Stanford)

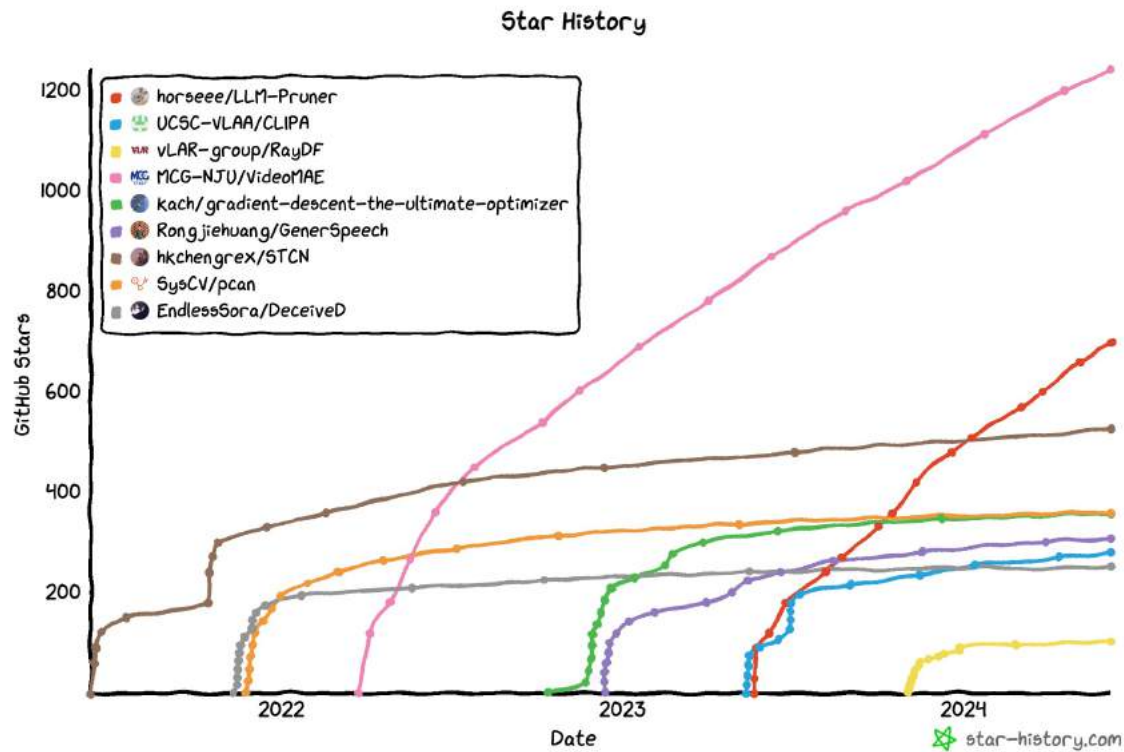


Код статей NeurIPS на GitHub

Год	Статья	Код	# цитирований на Scholar	# форков на GitHub	# звезд на GitHub
2023	LLM-Pruner: On the Structural Pruning of Large Language Models	https://github.com/horseeee/LLM-Pruner	148	74	698
	An Inverse Scaling Law for CLIP Training	https://github.com/UCSC-VLAA/CLIPA	19	10	280
	RayDF: Neural Ray-surface Distance Fields with Multi-view Consistency	https://github.com/vLAR-group/RayDF	2	4	102
2022	VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training	https://github.com/MCG-NJU/VideoMAE	648	123	1241
	Gradient Descent: The Ultimate Optimizer	https://github.com/kach/gradient-descent-the-ultimate-optimizer	34	25	358
	GenerSpeech: Towards Style Transfer for Generalizable Out-Of-Domain Text-to-Speech	https://github.com/Rongjiehuang/GenerSpeech	62	44	307
2021	Rethinking Space-Time Networks with Improved Memory Coverage for Efficient Video Object Segmentation	https://github.com/hkchengrex/STCN	272	71	525
	Prototypical Cross-Attention Networks for Multiple Object Tracking and Segmentation	https://github.com/SysCV/pcan	82	50	360
	Deceive D: Adaptive Pseudo Augmentation for GAN Training with Limited Data	https://github.com/EndlessSora/DeceiveD	99	24	251



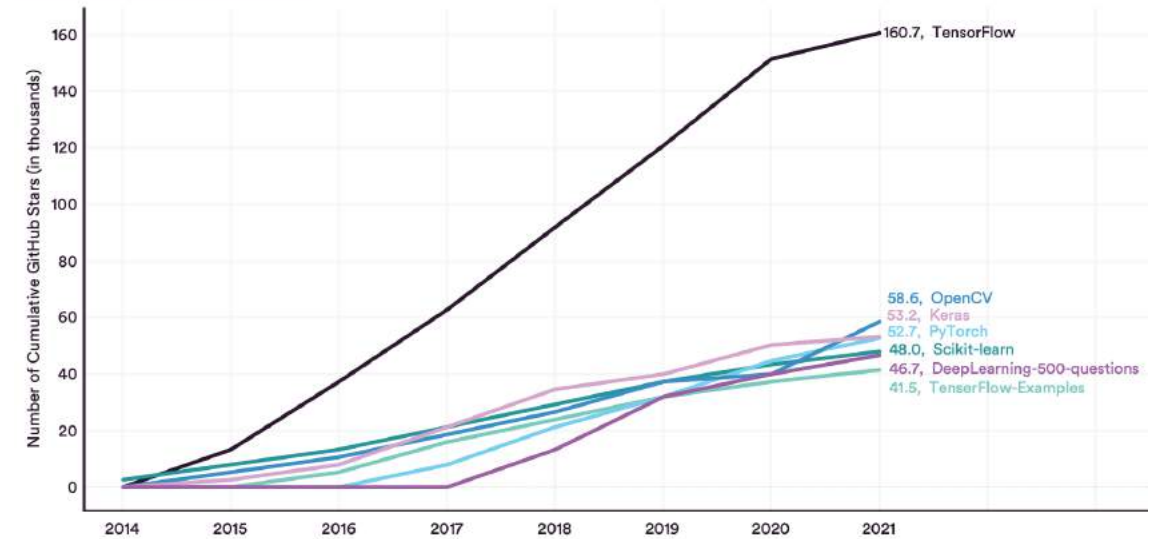
Популярность проектов на GitHub



Открытый код на GitHub создает быстрый **рост**
популярности результатов исследований

NUMBER of GITHUB STARS by AI LIBRARY (OVER 40K STARS), 2014–21

Source: GitHub, 2021 | Chart: 2022 AI Index Report



Библиотеки – самые популярные проекты по ИИ
согласно 2022 AI Index Report (Stanford)



Организации на GitHub (# звезд самого популярного проекта)

Мировые университеты:

- <https://github.com/stanfordmlgroup> (1550*)
- <https://github.com/stanfordnlp> (9343*)
- <https://github.com/stanford-crfm> (1499*)
- <https://github.com/mit-han-lab> (5894*)
- <https://github.com/mit-acl> (821*)
- <https://github.com/MIT-LCP> (2219*)
- <https://github.com/thu-coai> (1639*)
- <https://github.com/AllenInstitute> (316*)
- <https://github.com/flatironinstitute> (581*)

Организации:

- <https://github.com/openai> (56k*)
- <https://github.com/google-research> (32k*)
- <https://github.com/facebook> (219k*)
- <https://github.com/apple> (65k*)

Организации РФ:

- <https://github.com/yandexdataschool> (9570)
- <https://github.com/Yandex> (8197*)

Университеты РФ:

- <https://github.com/SciProgCentre> (637*)
- <https://github.com/aimclub> (618*)
- <https://github.com/SkoltechRobotics> (185*)
- <https://github.com/AIRI-Institute> (339*)
- <https://github.com/cig-skoltech> (94*)

НИУ ВШЭ:

- <https://github.com/bayesgroup> (1046*)
- <https://github.com/HSE-LAMBDA> (100*)



Открытый код ФКН

Цели проекта по открытому коду ФКН:

- Увеличить видимость и узнаваемость кода ФКН
- Увеличить вовлеченность студентов в проектах ФКН
- Привлечь сторонних разработчиков в проекты ФКН

Задачи проекта:

- Объединить код проектов факультета
- Привлекать студентов к проектам ФКН в рамках КР, ВКР и летних практик
- Организация докладов о лучших практиках разработки открытого кода
- Подготовка рекомендаций по разработке и поддержке проектов

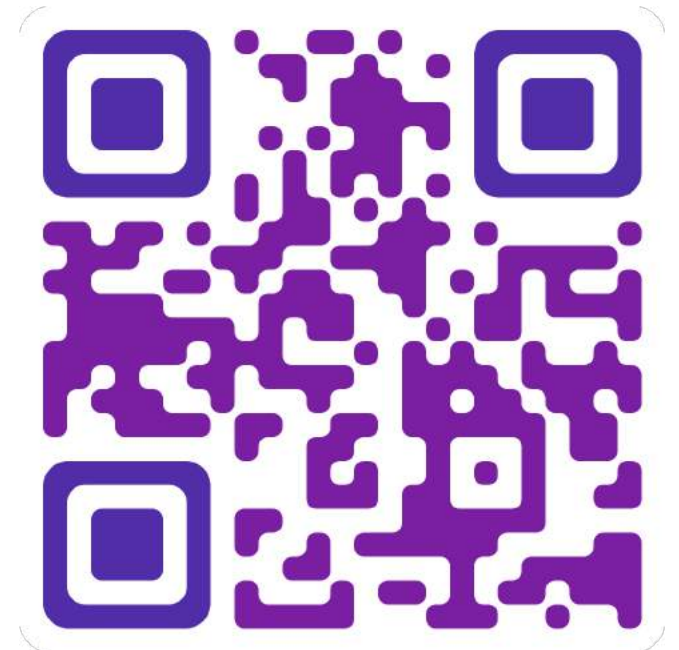


ФКН на GitHub

The screenshot displays the GitHub profile for the organization 'hse-cs'. At the top, there is a navigation bar with tabs for Overview, Repositories (10), Projects, Packages, Teams, People (6), and Settings. The profile header features a cat-themed avatar, the name 'HSE CS', and the bio 'HSE CS open source projects and technologies'. Below this, a section titled 'Popular repositories' lists several projects:

- fulu**: A Python library for supernova light curves approximation. It is a public repository with 23 stars and 2 forks.
- probaforms**: Conditional normalizing flows (NFs), conditional GANs, and conditional variational autoencoders (CVAEs) with a sklearn-like interface. It is a public repository with 23 stars and 6 forks.
- LaNeta**: A Python project with 6 stars and 1 fork.
- LINDA**: Tabular synthetic data generation. It is a public repository with 1 star.
- pytorch_ard**: A Python project with 1 star.
- pro**: A Python project with 1 star.

On the right side of the profile, there is a 'View as: Public' dropdown menu and a section for 'Discussions' with a 'Turn on discussions' link. At the bottom, there is a 'People' section with avatars of team members and an 'Invite someone' button. A search bar at the bottom left allows finding repositories by type, language, or sort order, with a green 'New' button.



<https://github.com/hse-cs>



Наши коллеги 😊

Data Fest в гостях у Альфа-Банка



А как у других?

ИТМО

Организация	Описание	Самые популярные репозитории
ИТМО AIM.CLUB	Объединенный репозиторий AI/ML фреймворков ИТМО	FEDOT , BAMT , FEDOT.Industrial
ВШЭ	Репозитории AI/ML фреймворков ВШЭ	hsemotion , roerich , probaforms
МФТИ, SPC	Подборка проектов МФТИ	DeepPavlov , kmath
Сколтех	Новые официальные форки бывшего репозитория Сколтеха	ttoy , h2tools

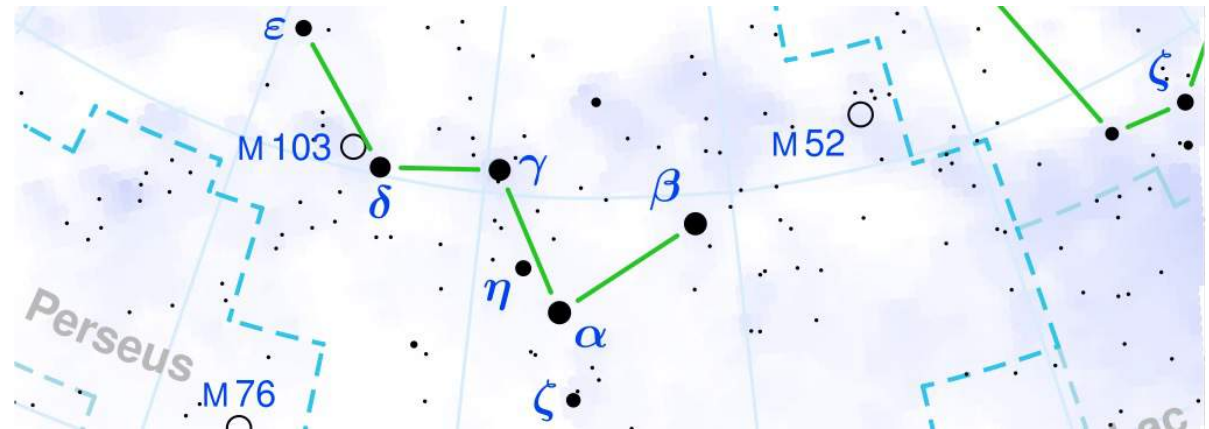
10





Fulu

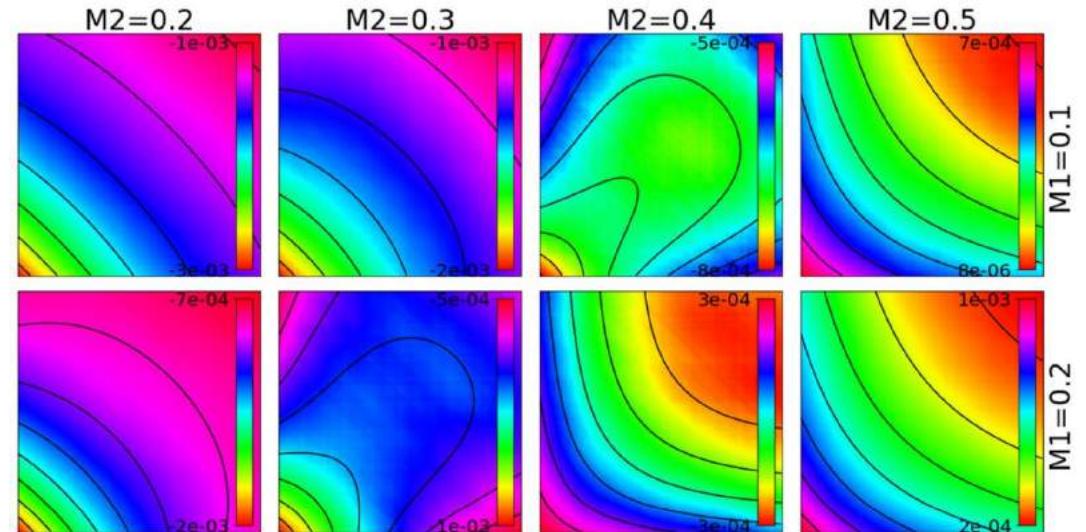
- <https://github.com/hse-cs/fulu>
- Библиотека методов аппроксимации кривых блеска астрономических объектов с использованием нейронных сетей.
- Названа в честь звезды Дзета Кассиопеи в 590 световых годах от нас, которая имеет официальное название Fulu
- Разработана в результате совместного научного исследования 7 организаций из 3 стран





LaNeta

- <https://github.com/hse-cs/LaNeta>
- Библиотека для оценки времен примешивания между двумя популяциями при двух пульсах миграции.
- Позволяет точно исследовать недавнюю (в пределах нескольких десятков поколений) историю примешивания популяций в сложных сценариях, для которых существовавшие ранее методы были неприменимы или неточны.





Заключение

Сайт: <https://cs.hse.ru/opensource>

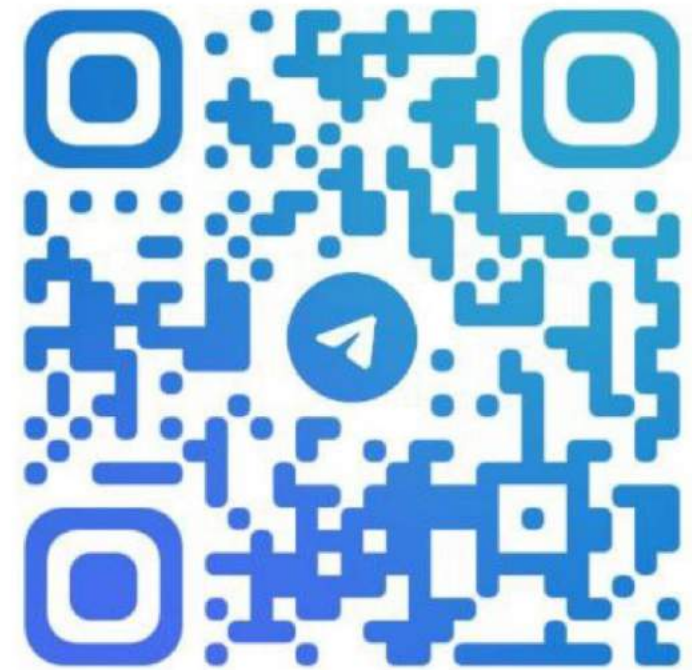
Github: <https://github.com/hse-cs>

Канал: https://t.me/hse_cs_opensource

Контакты:

Гуцин Михаил

mhushchyn@hse.ru, [@mikhail_h91](https://t.me/mikhail_h91)



@HSE_CS_OPENSOURCE





Ruptures

- C. Truong, L. Oudre, N. Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020
 - 988 цитирований
- Код по статье:
<https://github.com/deepcharles/ruptures>
 - 1400+ звезд
 - 11М скачиваний кода

deepcharles / ruptures

Code Issues 12 Pull requests 3 Discussions Actions Projects 1 Security Insights

Public Watch 27 Fork 160 Starred 1.4k

master 3 Branches 18 Tags

File	Commit Message	Time Ago
pre-commit-ci[bot]	chore: pre-commit autoupdate (#319)	2fba882 · last month 582 Commits
.binder	docs: add text segmentation example (#142)	3 years ago
.github	chore: release version 1.1.9 (#316)	2 months ago
docs	docs: fix example notebooks (#287)	last year
images	readme	7 years ago
src/ruptures	Improve speed of BottomUp (#309)	3 months ago
tests	build: add support of cp10 and cp11 wheels when possibl...	8 months ago
.flake8	style: add module imported but unused in flake8 (#191)	3 years ago
.gitignore	build: cleaner build process (#107)	4 years ago
.pre-commit-config.yaml	chore: pre-commit autoupdate (#319)	last month
CHANGELOG.md	docs: update changelog (#205)	3 years ago
CONTRIBUTING.md	build: cleaner build process (#107)	4 years ago
LICENSE	docs: update license in readme (#214)	3 years ago
MANIFEST.in	build: cleaner build process (#107)	4 years ago
README.md	chore: release version 1.1.9 (#316)	2 months ago
mkdocs.yml	docs: Ensemble dimensions (#248)	2 years ago
mkdocs_macros.py	fix broken notebook link in docs (#315)	3 months ago
pyproject.toml	ci: remove coverage from the wheel testing process (#229)	2 years ago
setup.cfg	ci(docs): fix the doc publishing job (#261)	2 years ago
setup.py	style: add module imported but unused in flake8 (#191)	3 years ago

About

ruptures: change point detection in Python

python science signal-processing scientific-computing change-point-detection change-point-detection

Readme

BSD-2-Clause license

Activity

1.4k stars

27 watching

160 forks

Report repository

Releases 12

v1.1.9 Latest on Dec 11, 2023

+ 11 releases

Packages

No packages published

Contributors 20

+ 6 contributors

Deployments 27



NGBoost

- Tony Duan, Anand Avati, Daisy Yi Ding, Khanh K. Thai, Sanjay Basu, Andrew Ng, and Alejandro Schuler. 2020. NGBoost: natural gradient boosting for probabilistic prediction. In Proceedings of the 37th International Conference on Machine Learning (ICML'20), Vol. 119. JMLR.org, Article 252, 2690–2700.
 - 291 цитирование
- Код по статье:
<https://github.com/stanfordmlgroup/ngboost>
 - 1600+ звезд
 - 11к скачиваний кода в месяц
- На странице Stanford ML Group:
<https://github.com/stanfordmlgroup>

stanfordmlgroup / ngboost

Public

Watch 45 Fork 216 Star 1.6k

25 Branches 20 Tags

File/Folder	Commit Message	Time Ago
.github	Add py311, remove py37 (#320)	2 weeks ago
data	Update survival datasets.	5 years ago
docs	Switch to poetry and add robust contributing workflow (#...	4 years ago
examples	Update load_boston in README.md (#340)	3 months ago
figures	Fixing black formatter (#301)	2 years ago
ngboost	Merge pull request #344 from mesenrj/fix/pred-dist-mem...	last week
results	speed-up and robustness changes	5 years ago
scripts	Switch to poetry and add robust contributing workflow (#...	4 years ago
tests	Add py311, remove py37 (#320)	2 weeks ago
.gitignore	Switch to poetry and add robust contributing workflow (#...	4 years ago
.pre-commit-config.yaml	Add py311, remove py37 (#320)	2 weeks ago
LICENSE	Switch to poetry and add robust contributing workflow (#...	4 years ago
MANIFEST.in	Create MANIFEST.in and package license file	4 years ago
Makefile	Fixing black formatter (#301)	2 years ago
README.md	Update load_boston in README.md (#340)	3 months ago
RELEASE_NOTES.md	Typo	4 months ago
pyproject.toml	Add py311, remove py37 (#320)	2 weeks ago
pytest.ini	Wolbra fix deps (#187)	4 years ago
setup.cfg	Switch to poetry and add robust contributing workflow (#...	4 years ago

About

Natural Gradient Boosting for Probabilistic Prediction

python machine-learning uncertainty-estimation gradient-boosting natural-gradients ngboost

Readme

Apache-2.0 license

Activity

Custom properties

1.6k stars

45 watching

216 forks

Report repository

Releases 20

v0.4.2 np.bool fix Latest on Nov 1, 2023

+ 19 releases

Packages

No packages published

Used by 158

Contributors 42



Streaming LLM

- Efficient Streaming Language Models with Attention Sinks, ICLR 2024
 - 48 цитирований
- Код по статье: <https://github.com/mit-han-lab/streaming-llm>
 - 5894 звезды
- На странице MIT HAN LAB: <https://github.com/mit-han-lab>

streaming-llm Public

Watch 60 Fork 349 Star 5.9k

main 1 Branch 0 Tags

Go to file Add file Code

Guangxuan-Xiao Update README.md 9790785 · 5 days ago 33 Commits

assets	add slides	4 months ago
data	upload ppl eval and llama chatbot demo	5 months ago
examples	Move input_ids to model device rather than "cuda"	4 months ago
figures	update readme	5 months ago
streaming_llm	core code	5 months ago
.gitignore	upload ppl eval and llama chatbot demo	5 months ago
LICENSE	Initial commit	5 months ago
README.md	Update README.md	5 days ago
setup.py	core code	5 months ago

README MIT license

Efficient Streaming Language Models with Attention Sinks

[paper] [slides] [video]

(a) Dense Attention (b) Window Attention (c) Sliding Window w/ Re-computation (d) StreamingLLM (ours)

Attention Sink