М. Шубин (аспирант ВМК МГУ, кафедра СКИ)
М. Григорьева (ведущий специалист, НИВЦ МГУ)
Н. Попова (доцент ВМК МГУ)

Исследование подходов к анализу популярности данных в области физики высоких энергий на примере эксперимента ATLAS на БАК

Май 2025 г.

Введение

- Работа посвящена разработке подходов направленных на выявление **шаблонов поведения** пользователей по доступу к данным в распределённых системах **мегасайенс** проектов.
- Базовой основой для исследований выбран эксперимент **ATLAS на БАК**.
- Предлагаемые подходы могут использоваться для других аналогичных систем.

Типы задач в эксперименте ATLAS на LHC

- **Централизованная обработка (Central Production)**
  - преобразование данных с детектора / полученных с помощью Монте-Карло к виду, пригодному для анализа
  - централизованное планирование, управление
- **Пользовательский анализ (User Analysis)**
  - пользовательский анализ данных, подготовленных в рамках Central Production
  - **хаотичность** и **непредсказуемость**
  - необходимо сократить время ожидания (может достигать нескольких суток) и улучшить предсказуемость времени выполнения заданий

Особенности распределенной системы WLCG (Worldwide LHC Computing Grid)

- Данные представлены в виде **наборов логически сгруппированных файлов** (**datasets**)
- Каждый датасет имеет несколько **копий (реплик)** в различных центрах грид
- **Управление репликами** датасетов является ключевым фактором, обеспечивающим эффективное функционирование грид-системы. На текущий момент управление осуществляется в гибридном (полуавтоматическом) режиме.

- На вычислительные центры поступают задачи разных типов. Поток задач пользовательского анализа **хаотичный** и **неоднородный**.
- Для эффективного управления распределенными ресурсами важна **предсказуемость загруженности** вычислительных центров.

Актуальность исследования

- Существующие системы мониторинга анализируют общие количественные характеристики работы системы, без учёта накопленной исторической информации.
- Предложенные в ряде работ подходы к прогнозированию не используются на практике, ввиду сложности в реализации и недостаточной точности прогнозов
- В настоящее время стремительно развиваются новые подходы для предиктивного анализа, такие как трансформеры, анализ нерегулярных временных рядов, графический анализ временных рядов и др., которые не применялись ранее к исследуемым данным и могут позволить улучшить точность прогнозов.

Цели и задачи исследования

- **Цель данного исследования** – повышение эффективности работы распределённых вычислительных систем крупномасштабных научных экспериментов при помощи современных методов предиктивного анализа.
- **Задачи:**
  - Обзор и анализ существующих подходов к прогнозированию популярности данных
  - Разработка методов прогнозирования популярности на базе различных алгоритмов машинного обучения
  - Исследование и разработка подхода к кластеризации данных для выделения групп данных со сходным паттерном обращений к ним

Существующие подходы к прогнозированию

- В основном использование статистических методов
- Использование классических нейросетевых моделей на базе многослойного персептрона
- Современные нейросетевые методы, используются как правило непосредственно в анализе самих данных

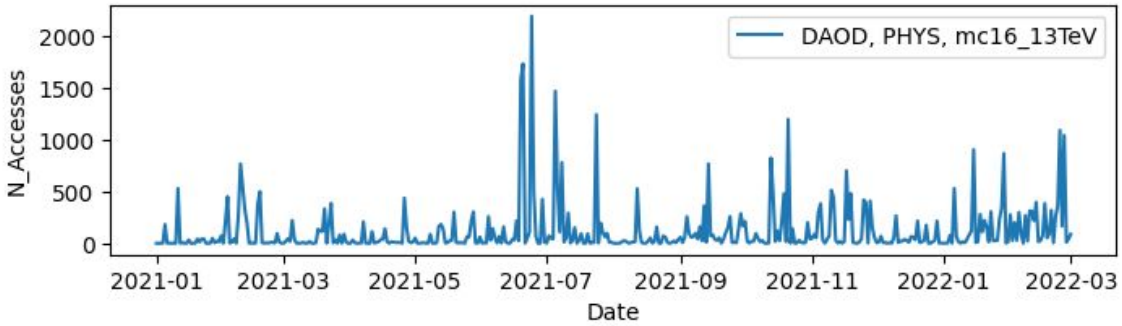
Подходы к прогнозированию популярности данных

Анализ и предсказание обращений может проводится:

- Для **групп датасетов** (датасеты одного формата, проекта). Например: *DAOD, PHYS, mc16_13TeV*
- Для отдельных **датасетов**.
  Например:
  mc16_13TeV:mc16_13TeV.364122.Sherpa_221_NNPDF30NNLO_Zee_MAXHTPTV140_280_BFilter.deriv.DAOD_PHYS.e5299_e5984_s3126_r10201_r10210_p4355_tid23598208_00

Понятие популярности данных

- Под **популярностью датасетов** будем понимать количество обращений к ним за определённый промежуток времени (день, неделя, месяц).
- **Популярность группы датасетов** определим как **суммарное** количество обращений к датасетам данной группы за определённый промежуток времени.
- Предсказываемая величина представляет собой **временной ряд**



Задача прогнозирования популярности **групп датасетов**

Пусть $G = \{D_1, \ldots, D_n\}$ — группа датасетов. $D_i = \{a_{i,1}, a_{i,2}, \ldots, a_{i,m}\}$ — временной ряд i-го датасета. Где $a_{i,t}$ — количество обращений к нему за неделю (день) t; m — длина ряда. Временной ряд группы датасетов G будет строиться как: $S_G = \{\sum_{i=1}^{n} a_{i,1}, \sum_{i=1}^{n} a_{i,2}, \ldots, \sum_{i=1}^{n} a_{i,m}\}$ (поэлементная сумма).

Задача: по заданному временному ряду S_G найти k будущих значений: $\{\sum_i a_{i,m+1}, \ldots, \sum_i a_{i,m+k}\}$ (**задача регрессии**)

- k=1 — **One-step** предсказание
- **Multi-step** предсказание (k > 1)
- **Multi-output** предсказание (k > 1)

Модели и данные для предиктивного анализа на **еженедельных** временных рядах

- One-step и Multi-step предсказания:
  a. Diff(x) + LSTM + Dense layer
- Multi-output предсказание:
  a. Diff(x) + LSTM + MaxPooling + Dense
  b. Facebook Prophet
  c. *Наивный метод предсказания* (выдающий сдвинутые вперёд известные значения)

- Группы датасетов: HIGGD (Higgs physics process), TOPQ (Top Quark), SUSY (Supersymmetry)
- Предобработка данных: логарифмирование Log(x+1)

Точность предсказания моделей на еженедельных временных рядах

Модель	Тип предсказания	Количество недель	HIGGD, %	TOPQ, %	SUSY, %
LSTM	One-step	1	7,21	4,90	11,68
	Multi-step	4	7,70	**4,97**	13,81
		12	7,85	**5,05**	15,03
	Multi-output	4	8,64	6,20	13,05
		12	8,83	6,67	12,78
Prophet	Multi-output	4	8,16	7,23	9,36
		12	7,67	6,83	12,15
Naive	Multi-output	4	9,85	6,77	13,31
		12	11,21	8,03	13,69

- Показатель ошибки: метрика **SMAPE** (Symmetric mean absolute percentage error) на логарифмической шкале
- LSTM и Prophet не показывают существенно лучшей точности по сравнению с Наивным методом
- Популярность группы SUSY предсказывается существенно хуже остальных
- Лучшую точность предсказания нескольких значений вперёд демонстрирует модель LSTM Multi-step для группы TOPQ: для 4, 12 недель ошибка около 5 %
- При переходе **на исходную шкалу** точность существенно **снижается** по причине того, что модель не может справиться с резкими перепадами количества обращений

Анализ **ежедневных** временных рядов: исследование длины обучающей выборки и горизонта прогнозирования

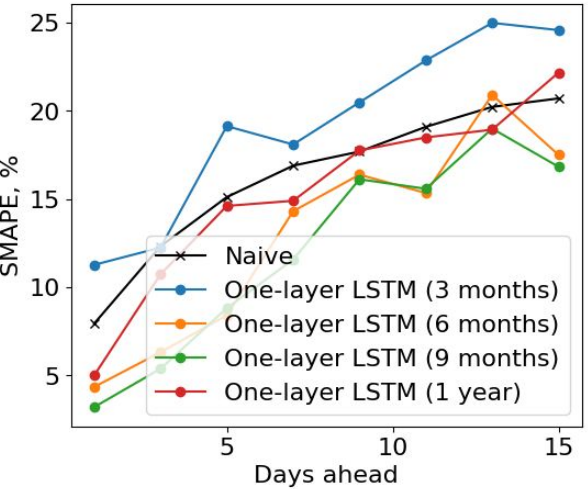
- Группа данных: формат **PHYS**, проект mc16_13TeV
- Длина временного ряда: 2 года 4 месяца (2021–2023 гг.)
- Кросс-валидация: блочная

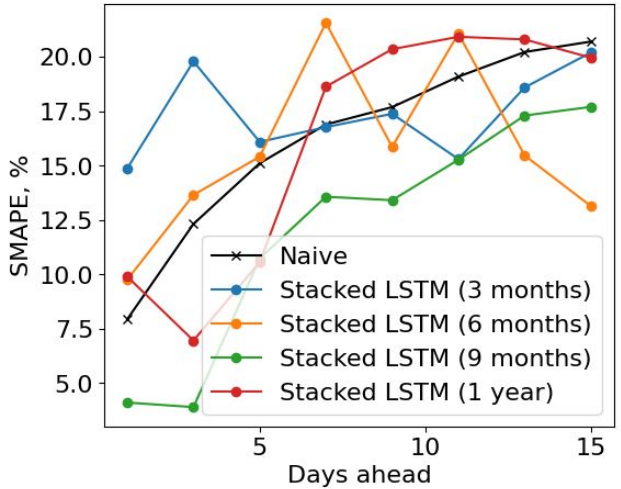
- Multi-output модели:
  1. **One-layer LSTM:** Diff(x) + LSTM + MaxPooling + Dense
  2. **Stacked LSTM:** MinMax + LSTM + LSTM + LSTM + LSTM + LSTM
  3. **Random Forest**

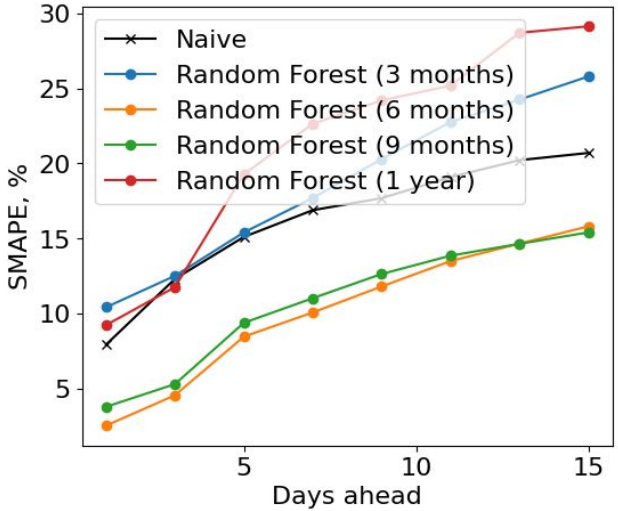
Предобработка: Moving Average, Log(x+1)

- Вариация длины обучающей выборки: 3, 6, 9, 12 месяцев
- Вариация горизонта предсказания: 1 – 15 дней

Зависимость точности прогнозирования от горизонта предсказания

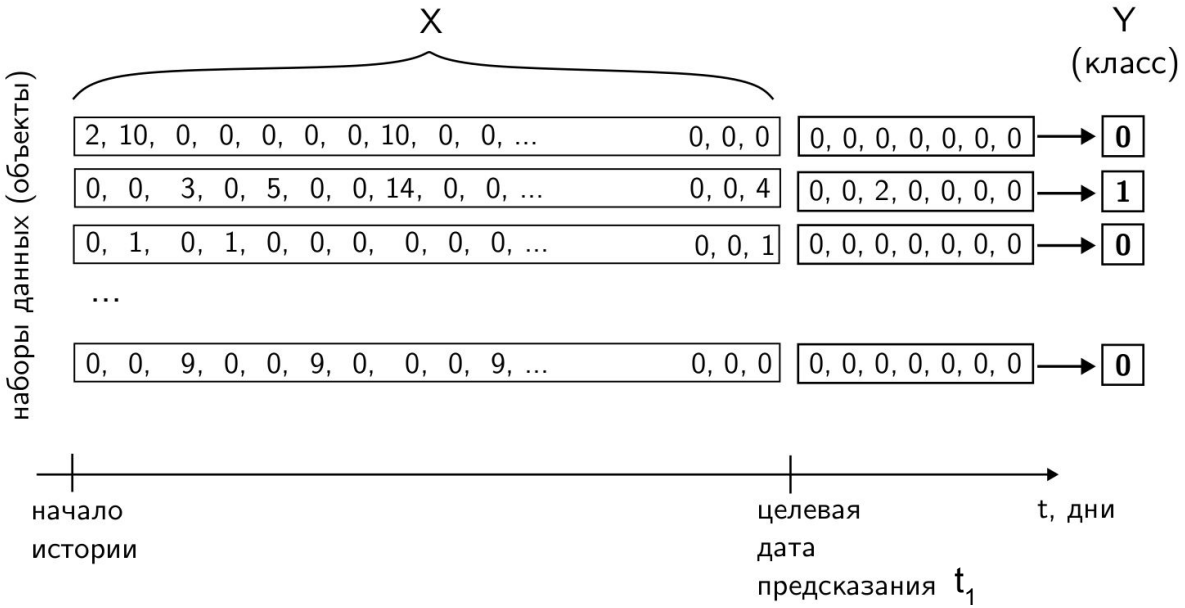






- Устойчивое превосходство по сравнению с наивным методом для выборки длиной 6-9 месяцев
- Лучший результат показала модель на основе Случайного Леса.

Постановка задачи классификации датасетов по популярности



- датасеты, к которым есть обращения в будущую неделю, назовём **популярными (положительный класс)**
- к которым не было обращений — **непопулярными (отрицательный класс)**

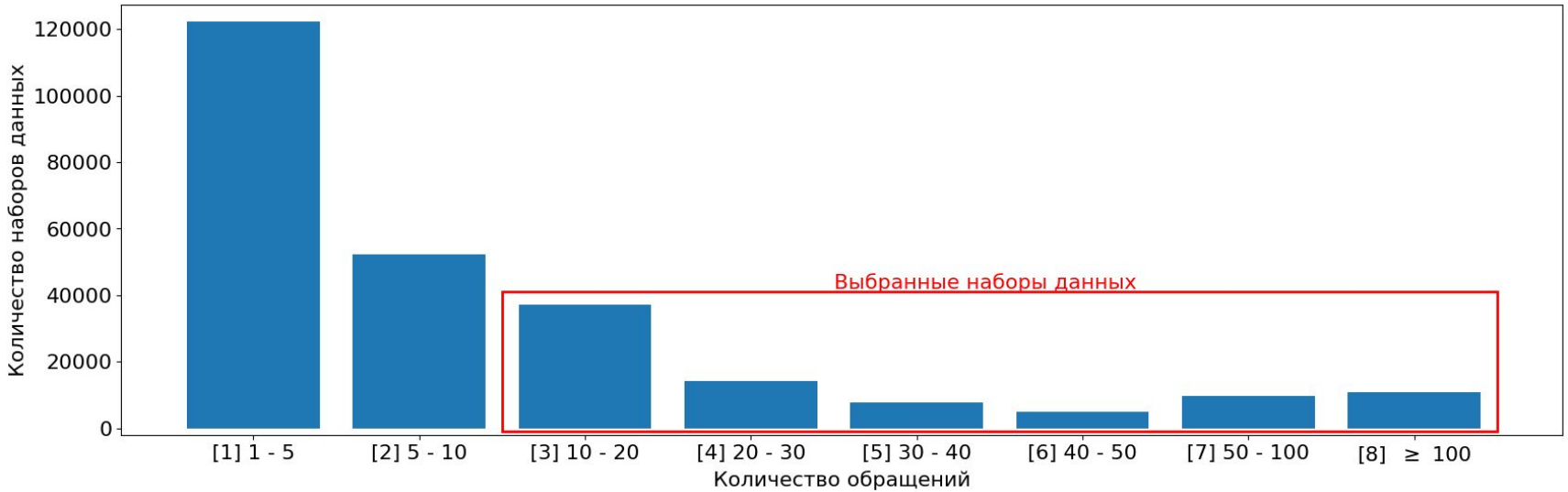
$D_i = \{a_{i,1}, a_{i,2}, \ldots, a_{i,m}\}$ (временной ряд датасета)

- **Дано:** временные ряды всех датасетов $\{D_1, D_2, \ldots, D_N\}$ (обучающая выборка)
- **Найти:** классификатор $\sigma : \{D_i\} \rightarrow \{0, 1\}$, разделяющий датасеты по их популярности на неделю вперёд относительно момента t_1

Предобработка: фильтрация датасетов

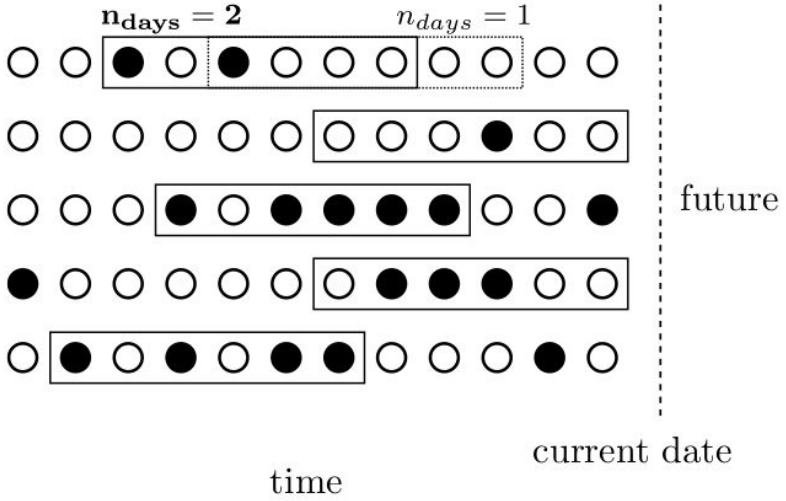
- Оставляем только датасеты с количеством обращений

  $n_{accesses} \geqslant \mathbf{n_{threshold}}$. Порог $\mathbf{n_{threshold}}$ был выбран равным **10**



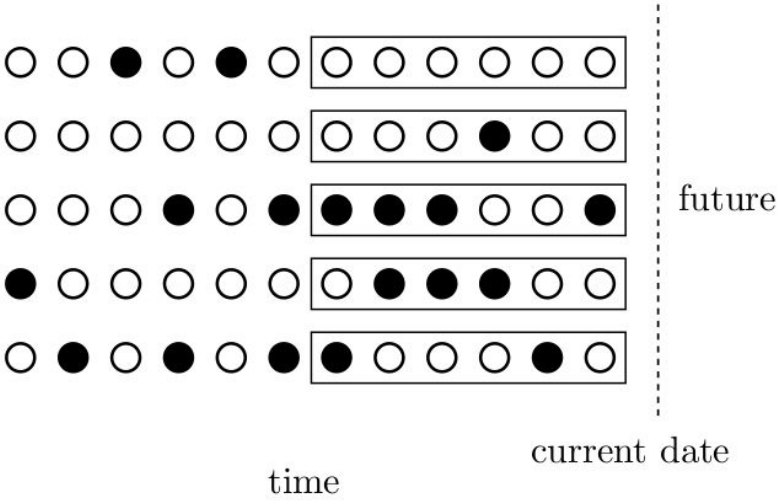
Регуляризация: выделение временных рядов фиксированной длины

Относительные временные ряды



Выбор временного окна независимо для каждого датасета по определённому критерию.
В данном случае: максимизация количества обращений внутри окна.

Абсолютные временные ряды

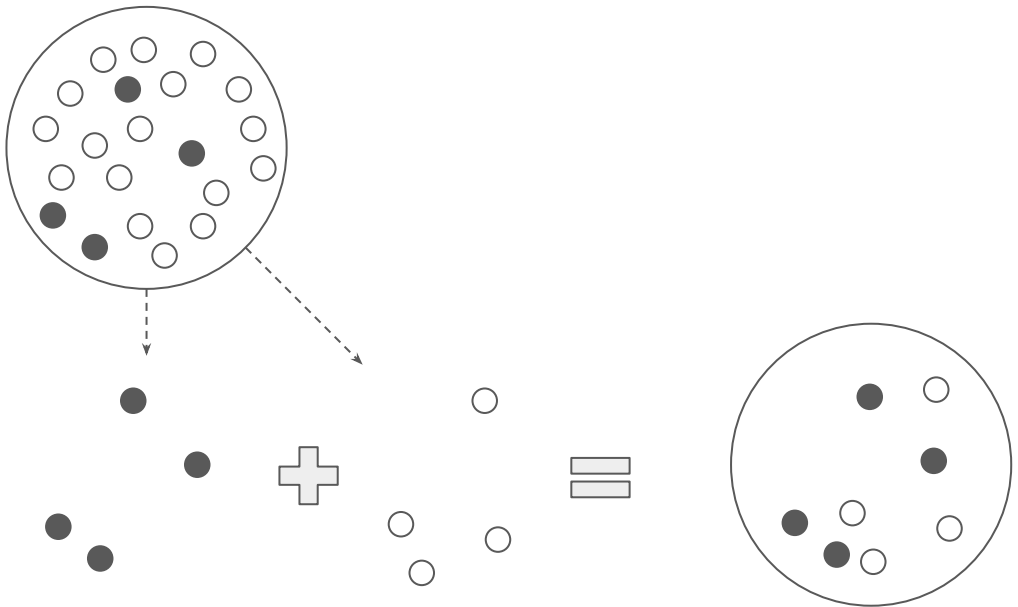


Использование общего интервала для всех датасетов.
Используется самый “поздний” интервал, который отражает актуальную картину.

Балансировка обучающей выборки

- Исходное количество объектов (датасетов): 255 000 (HIGGD,TOPQ,SUSY,PHYS)
- После фильтрации: ~ 30 000 объектов
- Количество объектов положительного класса: около 10 %

⇒ необходимость **балансировки** обучающей выборки



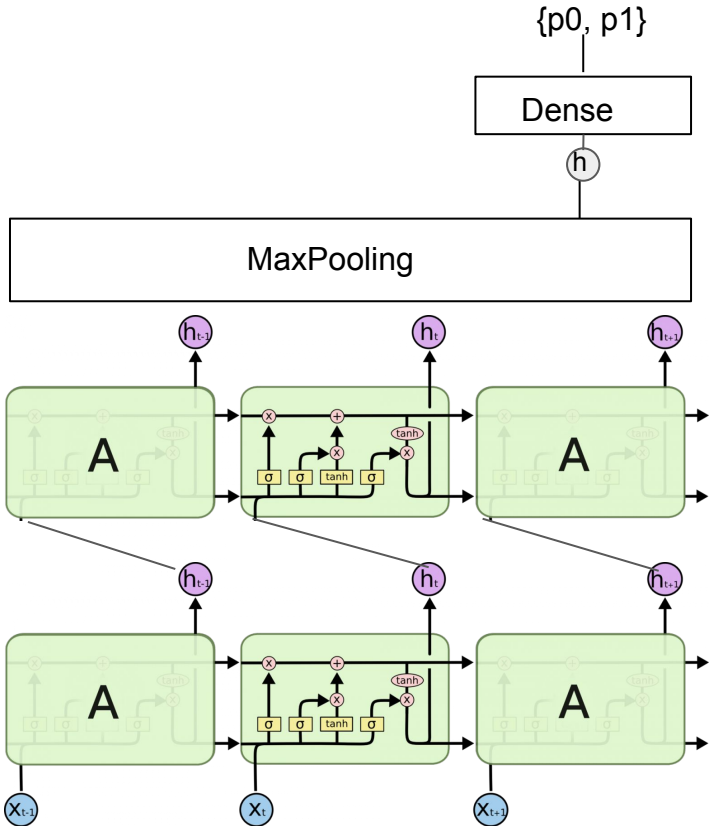
- Выбор всех объектов положительного класса
- Добавление к ним такого же количества случайно выбранных объектов отрицательного класса

Предлагаемый метод классификации датасетов по популярности

- Построение модели:
  a. **Формирование обучающей выборки** на основе текущей истории обращений (разметка классов — согласно обращениям на **последней неделе**)
  b. **Фильтрация** объектов.
  c. **Регуляризация** данных. Выделение временных рядов фиксированной длины (*абсолютных* или *относительных*)
  d. **Балансировка**: дублирование положительных объектов в нужном количестве для достижения равного количества положительных и отрицательных объектов в выборке
  e. **Обучение нейросетевой модели** на основе LSTM на сбалансированной выборке
- Использование:
  a. Проверка условия фильтра для конкретного датасета.
  датасет, не удовлетворяющий заданному условию, классифицируется как непопулярный.
  b. Применение (inference) нейросетевой модели.

Процесс обучения повторяется каждую неделю на основе обновленных данных мониторинга

Предложенная архитектура нейросетевой модели



- **Вход:** временной ряд
- **Выход:** вероятности p0, p1 принадлежности к классам 0 и 1.

Параметры:

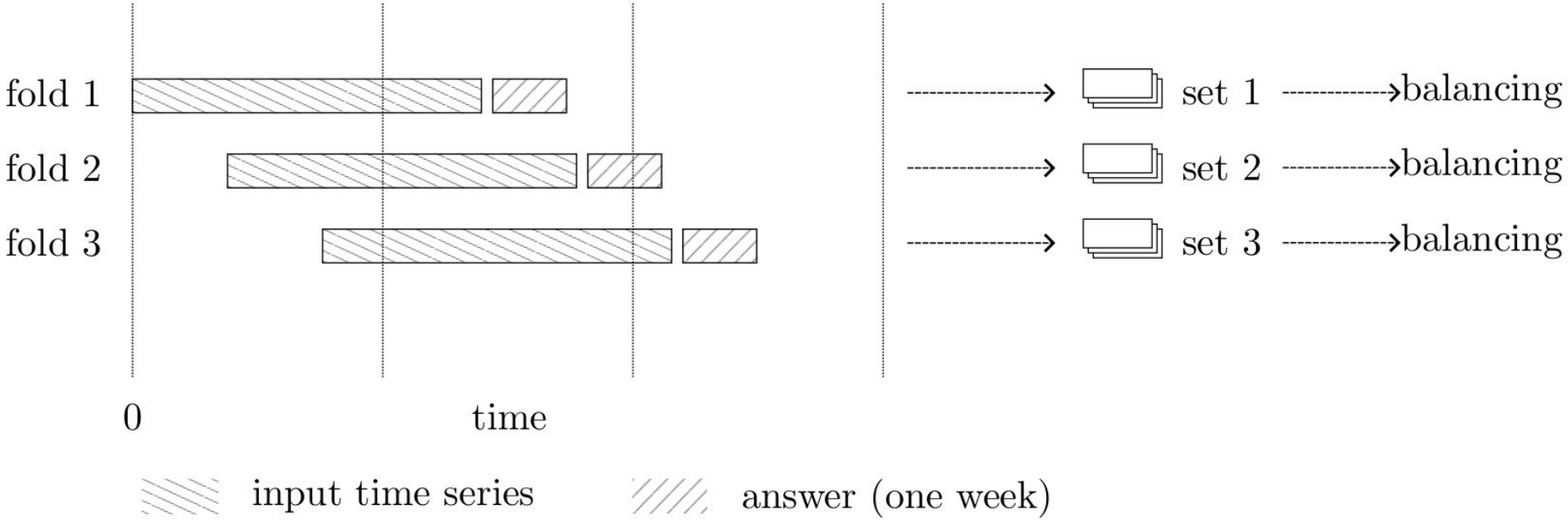
- Количество слоев LSTM
- Размерность скрытого состояния

Рассматриваемые модели бустинга

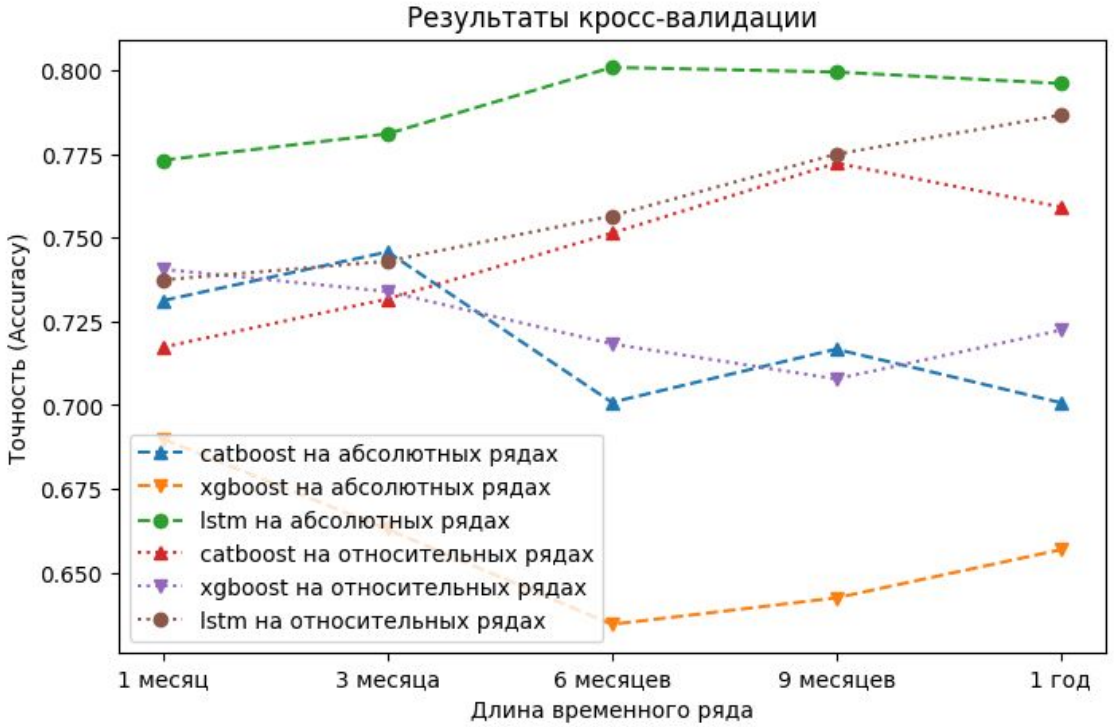
- XGBoost — eXtreme Gradient Boosting.
  - Параметры по умолчанию:
  - количество итераций (деревьев): 100
  - глубина дерева: 3
- CATBoost — unbiased boosting with categorical features
  - Параметры, подобранные эмпирически:
  - количество итераций (деревьев): 200
  - максимальная глубина дерева: 3

Кросс-валидация моделей

- **Блочный** вариант кросс-валидации
- На каждой итерации (fold) кросс-валидации **свой набор объектов** для обучения. Тестирование проводится на наборе следующей итерации (отличающимся сдвигом на неделю).
- **Балансировка** каждого набора проводится независимо



Зависимость точности классификации (Accuracy) от длины обучающего временного ряда



- Наилучший результат достигается при использовании модели LSTM на абсолютных временных рядах
- Увеличение длины временного ряда улучшает точность в половине случаев

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

TP - количество True Positive ответов, FN - количество False Negative ответов

Результаты кросс-валидации для **абсолютных** временных рядов

Model	Timeseries length	f1-score Class "0"	f1-score Class "1"	Accuracy
xgboost	1 year	0.979 / 0.743	0.976 / 0.469	0.978 / 0.657
	9 months	0.974 / 0.736	0.97 / 0.432	0.972 / 0.642
	6 months	0.962 / 0.73	0.956 / 0.421	0.959 / 0.635
	3 months	0.941 / 0.745	0.929 / 0.492	0.936 / 0.663
	1 month	**0.882 / 0.762**	**0.839 / 0.543**	**0.864 / 0.69**
catboost	1 year	0.977 / 0.771	0.973 / 0.551	0.975 / 0.701
	9 months	0.971 / 0.781	0.967 / 0.581	0.969 / 0.717
	6 months	0.959 / 0.769	0.953 / 0.568	0.956 / 0.701
	3 months	**0.938 / 0.795**	**0.926 / 0.659**	**0.933 / 0.746**
	1 month	0.881 / 0.787	0.837 / 0.624	0.863 / 0.731
lstm	1 year	0.965 / 0.83	0.961 / 0.739	0.963 / 0.796
	9 months	0.96 / 0.833	0.955 / 0.741	0.957 / 0.799
	6 months	**0.949 / 0.835**	**0.941 / 0.741**	**0.946 / 0.801**
	3 months	0.933 / 0.819	0.92 / 0.718	0.927 / 0.781
	1 month	0.879 / 0.813	0.835 / 0.707	0.861 / 0.773

- LSTM показывает лучшую точность в сравнении с моделями бустинга
- XGBoost справляется хуже остальных моделей, на объектах положительного класса (популярные датасеты) F1-мера едва превышает 50%.
- Моделям бустинга достаточно временного окна небольшой длины (1-3 месяца) для наилучшего предсказания

Результаты кросс-валидации для **относительных** временных рядов

Model	Timeseries length	f1-score Class “0”	f1-score Class “1”	Accuracy
xgboost	1 year	0.998 / 0.785	0.998 / 0.59	0.998 / 0.722
	9 months	0.998 / 0.776	0.998 / 0.566	0.998 / 0.708
	6 months	0.997 / 0.78	0.997 / 0.598	0.997 / 0.718
	3 months	0.994 / 0.791	0.994 / 0.625	0.994 / 0.734
	1 month	**0.98 / 0.792**	**0.98 / 0.647**	**0.98 / 0.741**
catboost	1 year	0.987 / 0.805	0.987 / 0.674	0.987 / 0.759
	9 months	**0.98 / 0.813**	**0.98 / 0.702**	**0.98 / 0.772**
	6 months	0.966 / 0.799	0.965 / 0.669	0.966 / 0.751
	3 months	0.944 / 0.784	0.941 / 0.637	0.942 / 0.732
	1 month	0.9 / 0.768	0.893 / 0.629	0.897 / 0.717
lstm	**1 year**	**0.97 / 0.822**	**0.97 / 0.726**	**0.97 / 0.787**
	9 months	0.968 / 0.813	0.969 / 0.712	0.969 / 0.775
	6 months	0.965 / 0.8	0.965 / 0.678	0.965 / 0.756
	3 months	0.967 / 0.791	0.967 / 0.658	0.967 / 0.743
	1 month	0.958 / 0.785	0.958 / 0.654	0.958 / 0.737

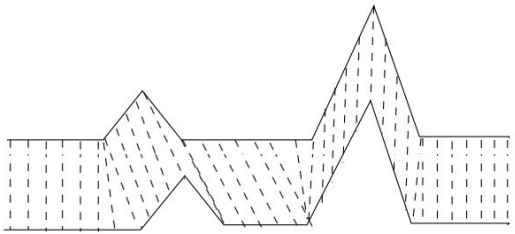
- Модели бустинга показывают лучшие показатели, по сравнению с абсолютными временными рядами
- Однако точность модели LSTM немного снижается при переходе к относительным временным рядам

Задача кластеризации временных рядов датасетов

Дано: $\{D_1, D_2, \ldots, D_N\}$ – временные ряды датасетов, $D_i = \{a_{i,1}, a_{i,2}, \ldots, a_{i,m}\}$.

Найти: разбиение на кластеры: $\mathbf{D_i \rightarrow y_i \in Y}$, где Y - множество меток кластеров, таким образом чтобы объекты (датасеты) внутри каждого кластера были близки относительно метрики ρ, а объекты разных кластеров значительно различались.

В качестве метрики $\rho(D_i, D_j)$ выберем DTW-расстояние (**Dynamic Time Warping**) между временными рядами.



Оценка качества кластеризации

интервал показателя: [-1, 1]
> 0 — кластеризация скорее удачная
< 0 — элемент следовало отнести к соседнему кластеру

Показатель **силует** (**Silhouette score**):
для элемента x ∈ C_I вычисляются:

$$a(x) = \frac{1}{|C_I| - 1} \sum_{y \in C_I, x \neq y} d(x, y)$$

— среднее расстояние от x до других элементов того же кластера

$$b(x) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{y \in C_J} d(x, y)$$

— минимальное расстояние от x до других кластеров

Silhouette score:

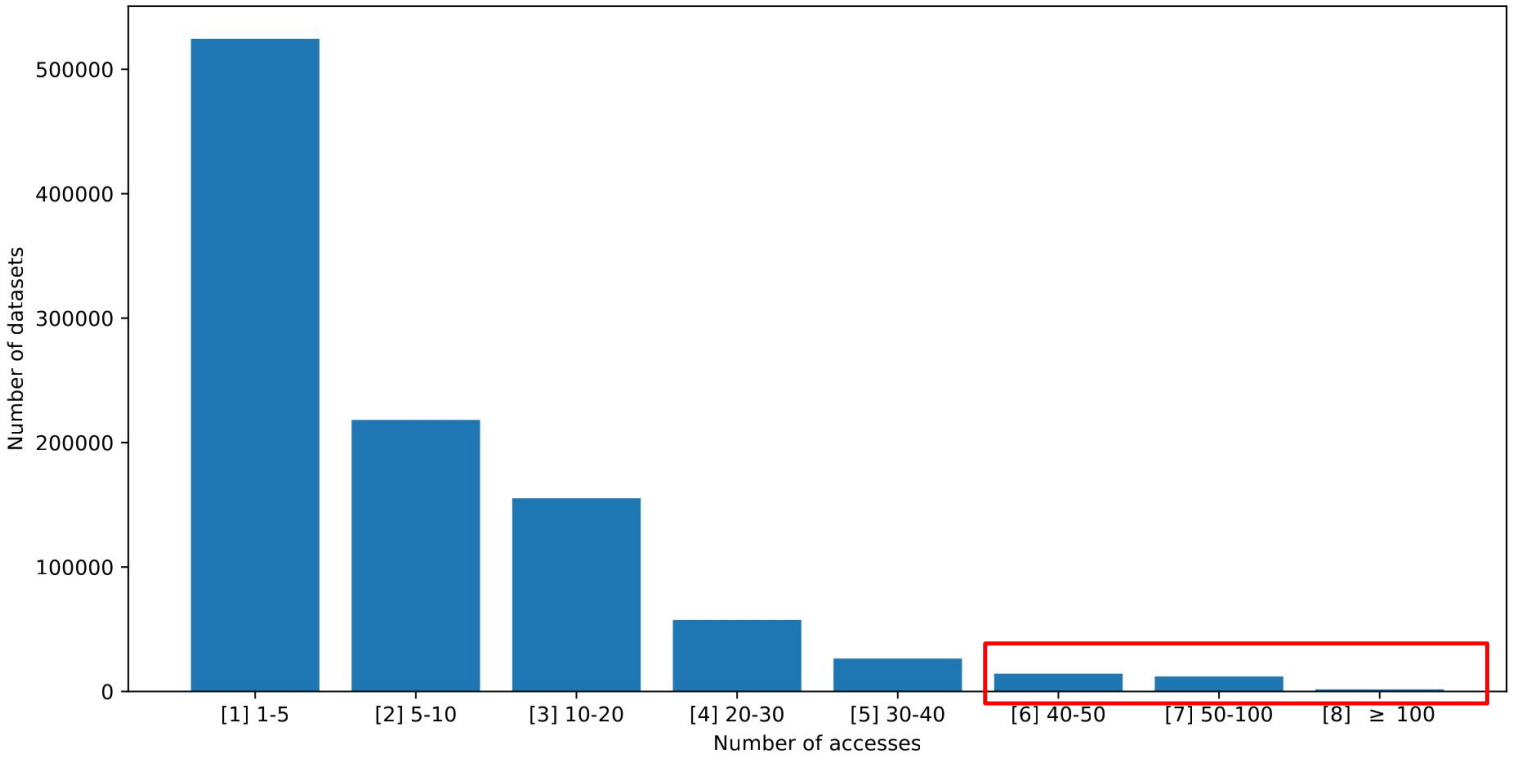
$$s(x) = \frac{b(x) - a(x)}{max\{a(x), b(x)\}} \quad \text{if} \quad |C_I| > 1$$

$$s(x) = 0 \quad \text{if} \quad |C_I| = 1$$

Данные для кластеризации

- Датасеты формата DAOD (Derived Analysis Object Data) экспериментальных данных и данных Монте-Карло.
- Еженедельные временные ряды количества обращений к датасету с начала 2017 года по 2024 год. Длина временного ряда: 385 недель.
- Количество датасетов: 1 009 994

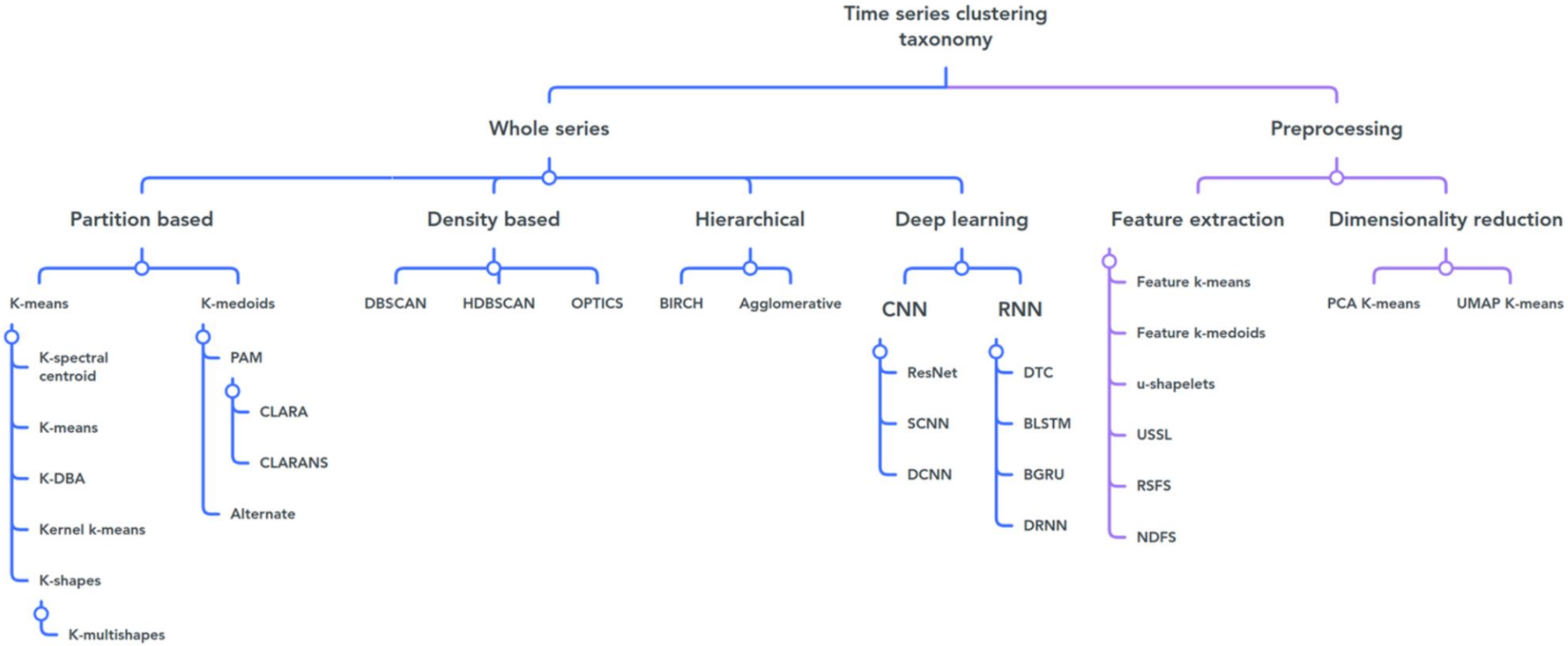
Фильтрация датасетов



Оставляем только датасеты с количеством активных недель > 40 (из 385)

Из 1 009 994 остаются 26 536

Методы кластеризации



Источник: Holder, C., Middlehurst, M., and Bagnall, A. **A review and evaluation of elastic distance functions for time series clustering.** Knowl Inf Syst 66, 765–809 (2024).

Методы кластеризации

- **Плотностные (Density-based)** методы кластеризации не требуют априорного знания о количестве кластеров

- Метод **HDBSCAN** допускает наличие кластеров различной плотности (не требует задания размера окрестности).

Внутреннее представление временных рядов

Рассмотрение временных рядов как арифметических векторов (использование базовых функций расстояния) **нецелесообразно**.

Методы снижения размерности:

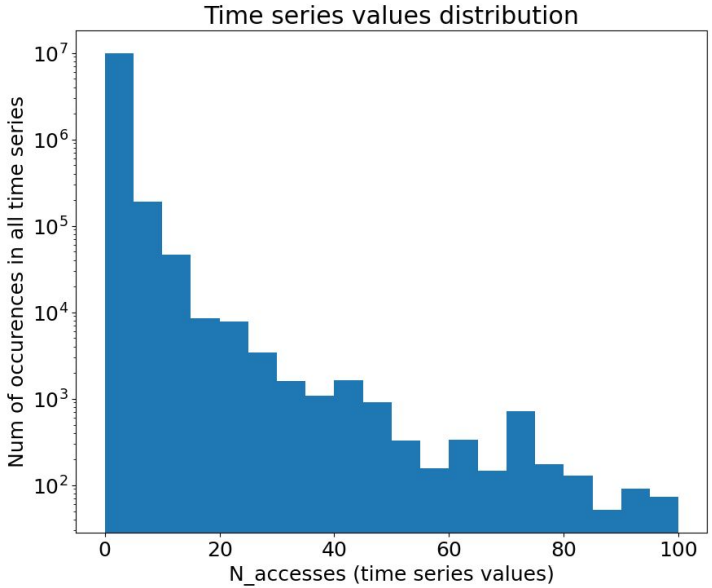
- Метод главных компонент (**PCA**), t-SNE, UMAP

Векторные представления в другом пространстве (**эмбеддинги**)

- Автоэнкодеры (Bidirectional LSTM, MLP)
- Библиотека TS2Vec[1]
- Трансформеры (DistilBERT)

1. Zhihan Yue, Yujing Wang et al. **TS2Vec: Towards Universal Representation of Time Series.** arXiv.2106.10466. 2022.

Предобработка: **квантование** временных рядов



Распределение значений временных рядов

- С ростом натурального числа n “встречаемость” n во временных рядах существенно снижается

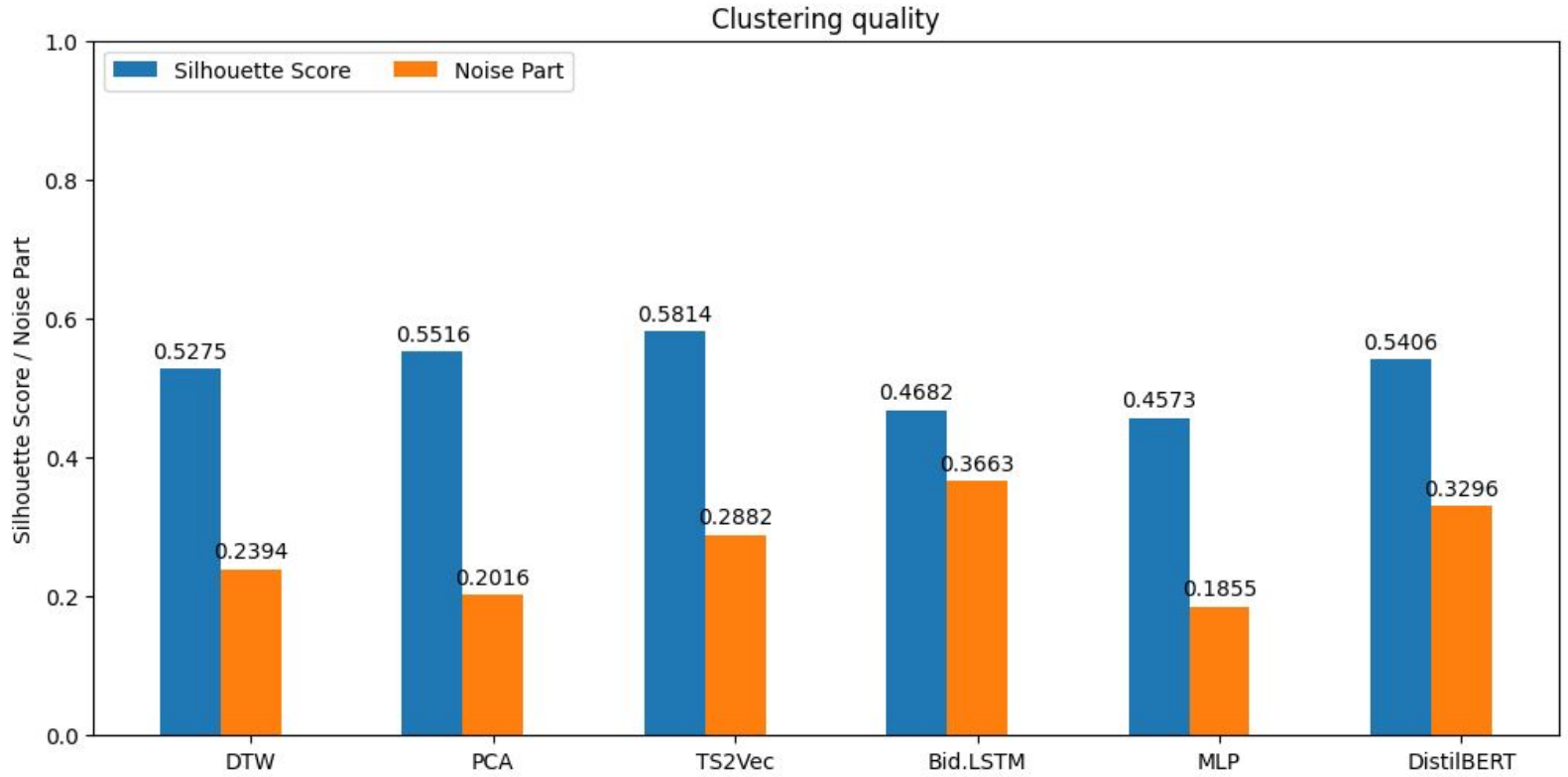
- Для приведения данных к виду, пригодному для подачи на вход моделям машинного обучения, применяется преобразование:

0 -> 0
1 -> 1
[2, 5) -> 2
[5, 10) -> 3
[10, 15) -> 4
[15, 20) -> 5
[20, 25) -> 6
[25, 30) -> 7
[30, 35) -> 8
[35, 40) -> 9
[40, 45) -> 10
[45, 50) -> 11
[50, ∞) -> 12

Результаты кластеризации

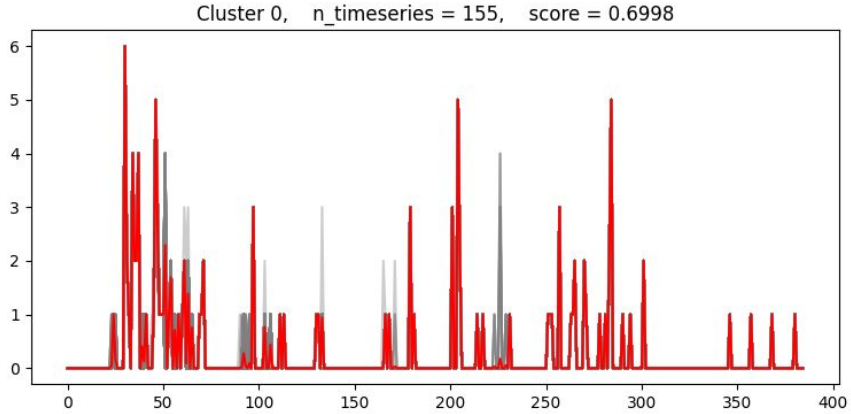
Метод эмбеддинга	Число кластеров	Среднее число в элементов кластере	Доля шумовых элементов	Silhouette score
DTW (исходное пространство)	88	229	23.94%	0.5275
PCA	90	235	20.16%	0.5516
TS2Vec	78	242	28.82%	0.5814
Bidirectional LSTM	66	255	36.63%	0.4682
MLP	81	267	18.55%	0.4573
DistilBERT	69	258	32.96%	0.5406

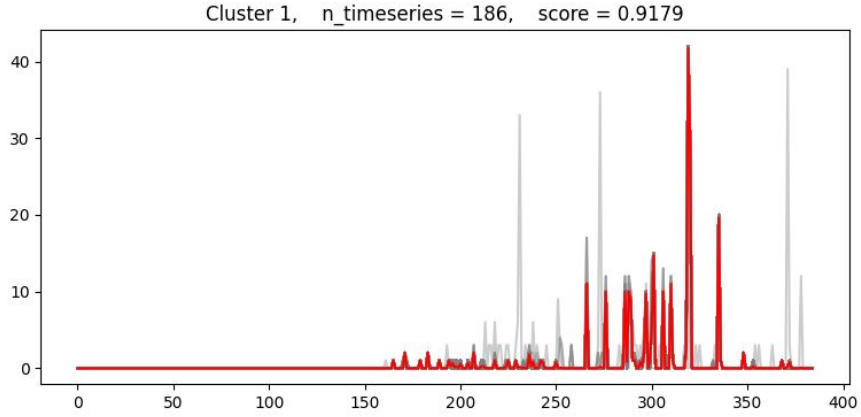
Результаты кластеризации

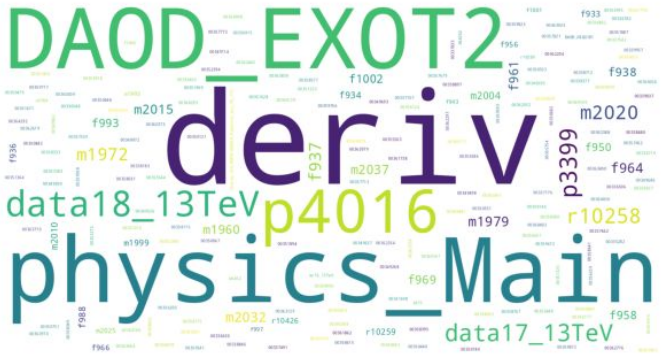


- Наилучшим методом с т.зр. silhouette score показал себя **TS2Vec**, с небольшим отставанием за ним — **PCA**
- Наименьшая доля шумовых элементов выделяется при помощи метода MLP

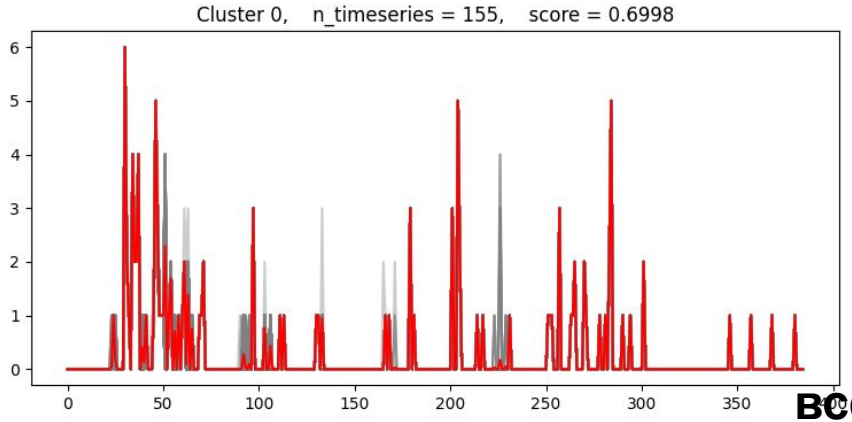
Визуализация отдельных кластеров (TS2Vec)

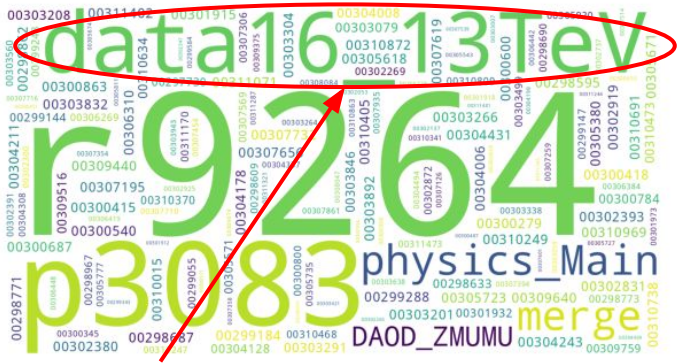




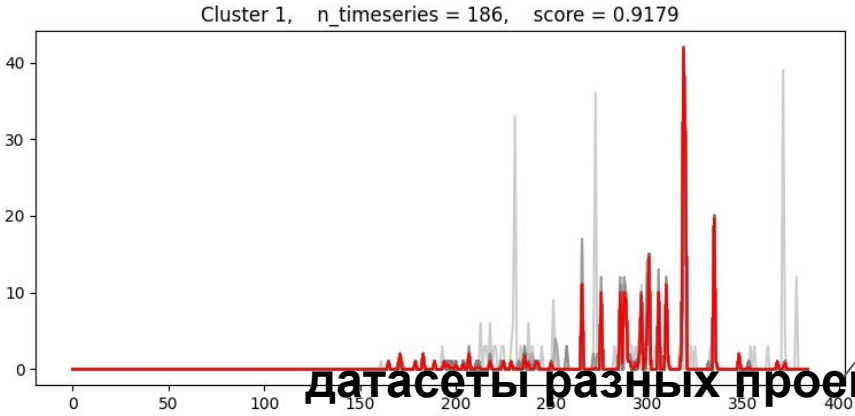


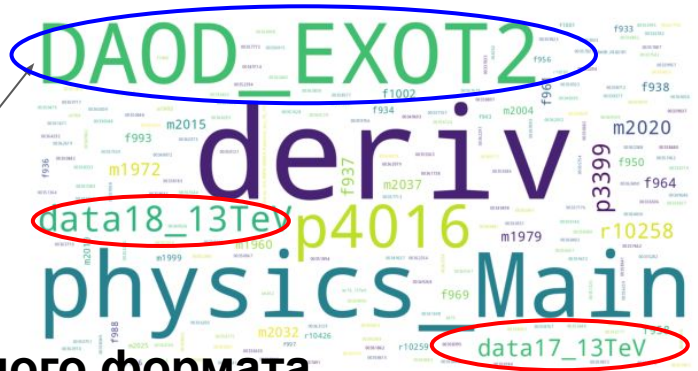
Визуализация отдельных кластеров (TS2Vec)





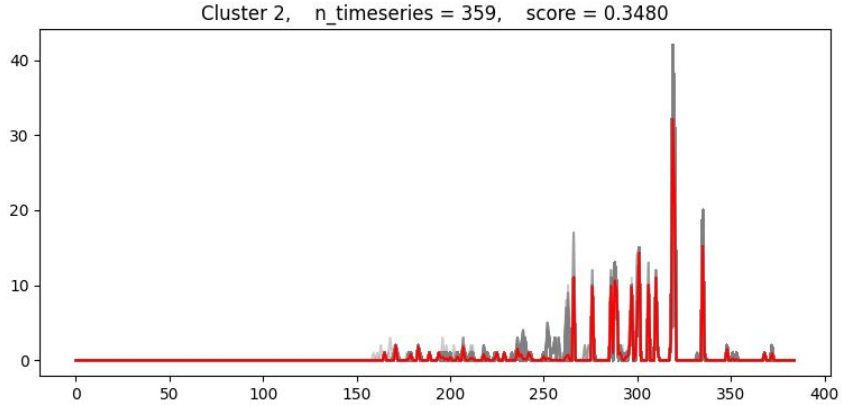
все 155 датасетов из одного проекта

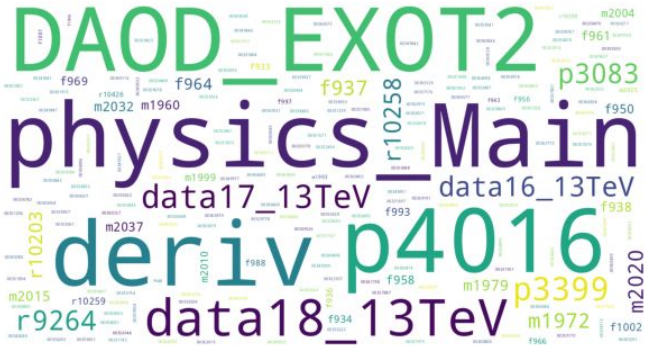


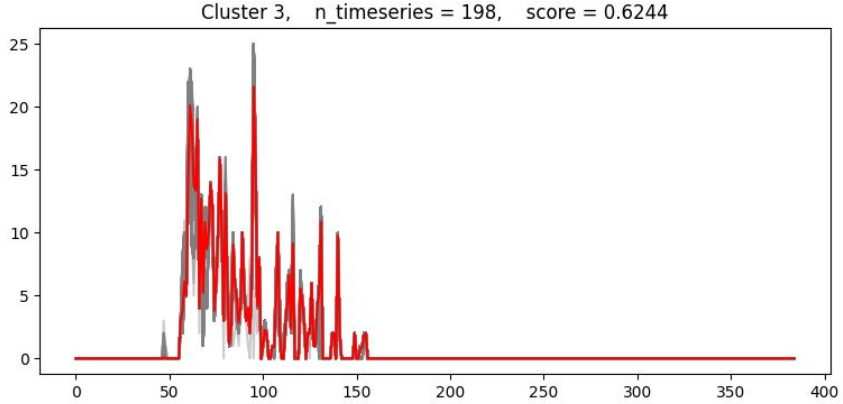


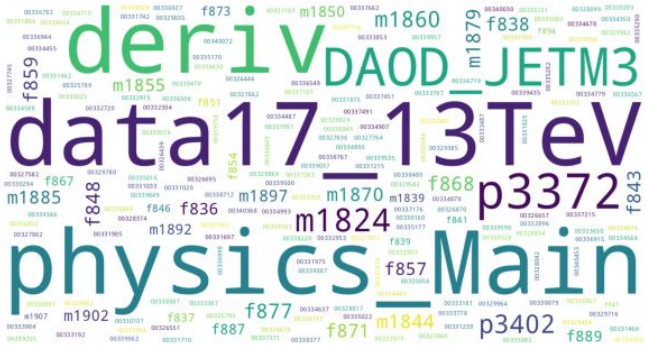
датасеты разных проектов одного формата

Визуализация отдельных кластеров (PCA)

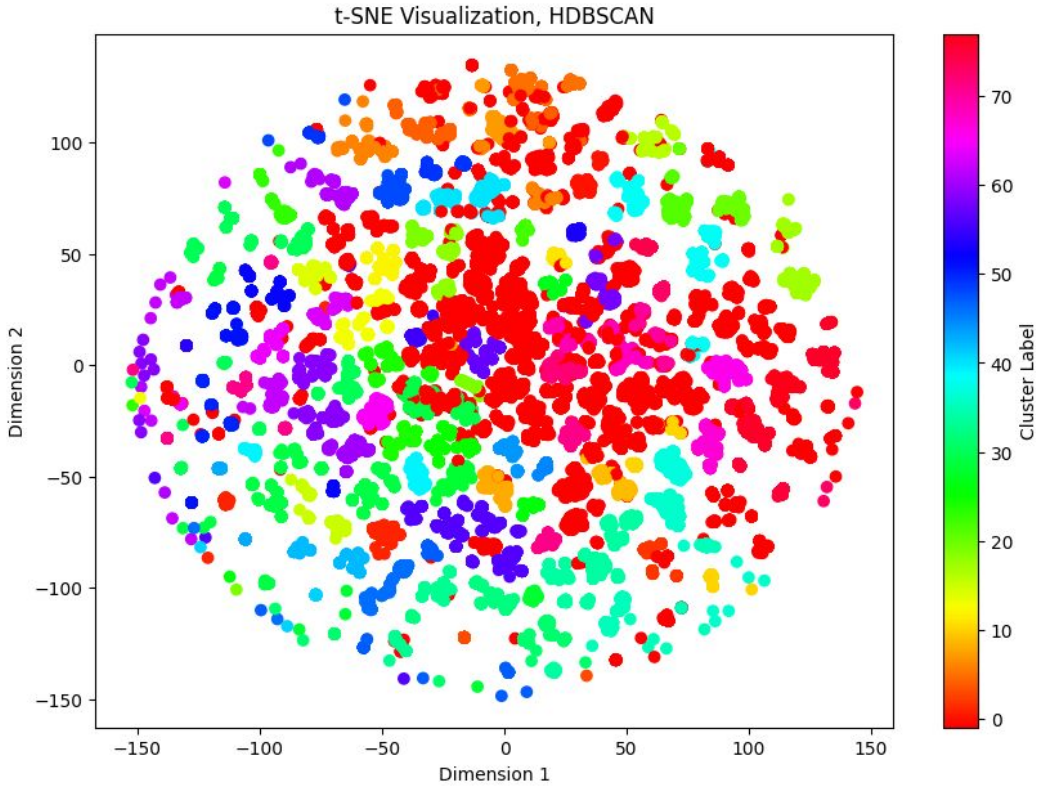








Визуализация в пространстве пониженной размерности (TS2Vec)



Основные результаты

- Исследована применимость методов прогнозирования популярности групп датасетов на основе методов машинного обучения для еженедельных и ежедневных данных (задача регрессии). Для еженедельных данных приемлемый результат показала **нейросетевая модель LSTM**. Для ежедневных - модель на основе **случайного леса**.
- **Предложен метод** классификации датасетов по признаку востребованности в будущем. Метод содержит этапы **фильтрации, регуляризации и балансировки** обучающей выборки. Кросс-валидация показала точность около 80%.
- Исследованы подходы на основе **кластеризации** для нахождения групп датасетов, сходных по характеру обращения к ним. Подходы включают **предобработку, выделение эмбеддингов**. Качество кластеризации достигает 0,5-0,6 по метрике silhouette score. Визуализация кластеров отражает общие характерные поля в названиях датасетов.

Направления дальнейшего развития

- Дальнейшая настройка различных **гипер-параметров** методов во всех трёх задачах.
- Более широкое исследование применимости нейросетевых моделей глубокого обучения (в т.ч. **трансформеров**).
- Реализация и встраивание в **систему мониторинга** (рекомендательную систему) предложенных методов.
- **Предложение стратегии управления репликами** датасетов на основе разработанного метода прогнозирования популярности.
  - **цель:** улучшить сбалансированность загрузки вычислительных ресурсов, снизить время ожидания начала выполнения пользовательских задач, увеличить точность предсказания времени выполнения задач.
- **Оценка эффективности** стратегии управления репликами.

Опубликованные статьи

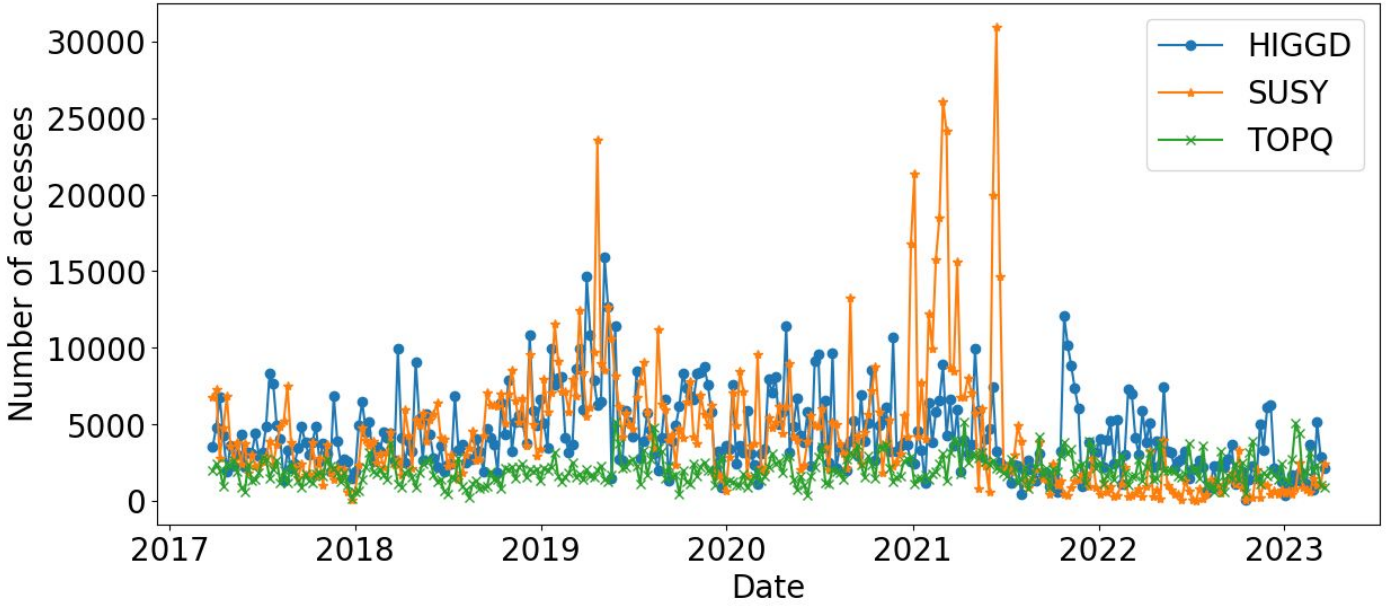
- M.A. Grigorieva, N.N. Popova, D.A. Vartanov, M.V. Shubin, **Exploring Hierarchical Forecasting of Data Popularity in High-Energy Physics Experiments** // Lobachevskii Journal of Mathematics, 2023, Vol. 44, No. 8, pp. 3076-3090.
- M.A. Grigorieva, N.N. Popova, D.A. Vartanov, M.V. Shubin, **Methodology of Data Popularity Forecasting in High-Energy Physics Experiments on Unbalanced and Irregular Time-series Data** // Lobachevskii Journal of Mathematics, 2024, Vol. 45, No. 7, pp. 3072-3084.

Готовится:

- M.A. Grigorieva, N.N. Popova, M.V. Shubin, **Decoding Dataset Access Patterns in High-Energy Physics through Embedding-Driven Density Clustering** // 2025

Дополнительные слайды

Статистический анализ обращений к наборам данных

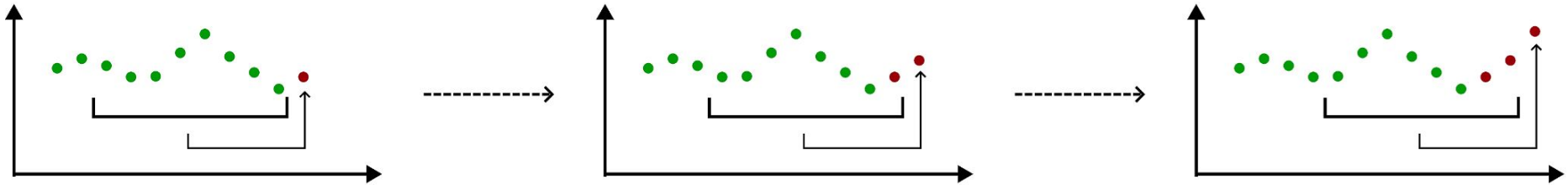


количество обращений к наборам различных групп данных

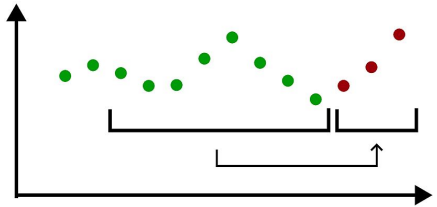
- Необходимость анализа отдельных групп данных независимо

Предсказание нескольких значений в будущем

- **Multi-step** (многошаговое) предсказание с использованием One-step модели



- **Multi-output** предсказание сразу нескольких значений наперёд



Анализ популярности групп датасетов: **выводы**

- В задаче прогнозирования на несколько недель вперёд (еженедельные данные):
  - Модель **LSTM** показала лучший результат для групп HIGGD и TOPQ и худший для SUSY.
- В задаче прогнозирования на ежедневных данных:
  - Оптимальной длиной обучающей выборки является **6-9 месяцев**. Лучший результат показала модель на основе **Случайного Леса**.
- Рассмотренные модели **не позволяют точно** предсказывать количество обращений в будущие недели, но могут быть полезны для прогнозирования увеличения/уменьшения запросов к данным.

Классификация датасетов: **выводы**

- **Предложен метод** классификации датасетов по признаку востребованности в будущем
- Метод содержит этапы **фильтрации, регуляризации и балансировки** обучающей выборки.
- Проведена **кросс-валидация** нескольких моделей классификации. Наилучшим образом себя показала нейросетевая модель на основе LSTM, которая позволяет достичь точности около 80%.

Кластеризация датасетов: **выводы**

- Методы на основе кластеризации позволяют находить группы датасетов со сходным паттерном обращений к ним.
- Снижение размерности PCA и выделение эмбеддингов библиотекой TS2Vec улучшают качество кластеризации по метрике silhouette score, при этом доля шумовых элементов может измениться.
- Нейросетевые методы построения эмбеддингов показали себя хуже в текущей конфигурации.