Устойчивость моделей машинного обучения

14.03.2024 @ Сколтех

Дмитрий Колодезев, ООО «Промсофт»

Про меня

- Дмитрий Колодезев
- Директор ООО Промсофт
- Reliable ML
- ML System Design Document
- Курс ML System Design
- Datafest
- Kolodezev.ru

О чем мы тут

- fit — predict — profit!
- Иногда сразу плохо работает
- Иногда сначала хорошо
- Потом как обычно
- Иногда не понятно, работает ли вообще

data scientist performing the fit-predict-profit process with joy and ease

Почему оно сломалось

- Могла сразу работать плохо или не работать вовсе
- Проблемы с данными
- Программные, аппаратные, организационные отказы
- Проблема с моделью
  - нестабильность, недоопределенность, низкое качество
- Атаки - см Атлас
- Неправильное использование — model card, datasheets
- Изменение бизнес-процессов

Сдвиг данных и концепций

- $P(Y=y \mid X=x)$
- Меняется распределение X. Data Drift
  Модель работает так же, но метрики другие
- Меняется зависимость Y от X. Concept Drift
  Модель работает хуже, метрики — как повезет
- Меняется распределение Y. Target Drift
  Обычно следствие чего-то выше

Откуда сдвиг

- Мир меняется
  - Ковид, СВО, санкции
  - Новые рынки
  - Новые источники клиентов
  - Новые продукты
- Доступные нам данные меняются
  - БКИ что-то у себя скорректирует
  - Форму ввода на сайте подкорректируют
  - Датчики на станке заменят

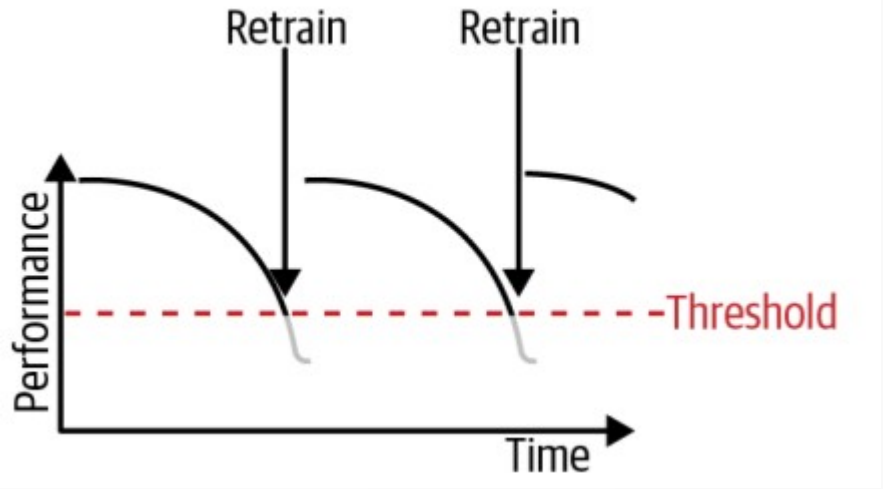
Оказывается, его надо кормить

- Ожидания:
  тратим деньги
  получаем систему
  система приносит деньги
- Реальность:
  тратим деньги
  получаем систему
  система требует денег
  на поддержание качества

data scientist in a suit looking inside his empty leather wallet, white bills with red numbers flying around him in a chaotic manner, blue background

Переобучаем!

- Собираем метрики
- Сравниваем с порогом
- Учим на новых данных
- Проблемы:
  - Какой порог?
  - Станет ли лучше?
  - Что со старыми данными?
  - Где брать разметку?
  - А оно точно будет постепенно ухудшаться?



Какой порог выбрать

- Это инвестиционное решение
- Стоимость дообучения vs вероятный ущерб
- Бизнес всегда попросит «самое лучшее»
- Можно оценить ущерб от падения качества
- Иногда трудно измерить
- Проще переобучать модель, как только надежно детектировали падение качества
- Переобучайте так часто, как можете

Как обучать чаще

- Малый поток данных
  - Синтетика и аугментация
- Медленное вызревание меток
  - Конструируем обратную связь, слабые и прокси метки
- Громоздкая подготовка данных
  - Сохранять предикты и признаки
- Трудно сравнивать модели
  - Интерливинг вместо A/B тестов, откат модели

Будет ли новая модель лучше

- На свежих данных новая модель всегда лучше
- Как будет дальше — вопрос
- Возможно, временная аномалия / шум
- Нужен надежный способ тестирования в проде:
  - Теневой деплой
  - Канареечный деплой
  - АБ тесты
  - Интерливинг
  - Многорукие бандиты

Что со старыми данными

- Если распределение изменилось,
  зачем мы учимся на старых данных?
- Курирование датасета:
  - Обрезаем по горизонту (последний год?)
  - Отбираем наиболее характерные точки (Core set)
  - Сэмплируем пропорционально возрасту данных
  - Сэмплируем пропорционально схожести распределений

Где брать разметку

- Иногда разметка есть сразу:
  - Пользователь кликнул по рекламе
- Иногда разметка задерживается:
  - Пользователь не вернул кредит
- Иногда разметки нет:
  - Долгосрочная прибыльность инвестиций
- Иногда не поймешь:
  - Ошибки распознавания речи/интента

Где брать разметку

- Прокси-метрики
  - Скоррелированные с бизнес-метрикой
  - Легкодоступные
- Слабые метки
  - «Положил в корзину» вместо «купил», с весом
- Где взять
  - Искать в данных
  - Спрашивать бизнес

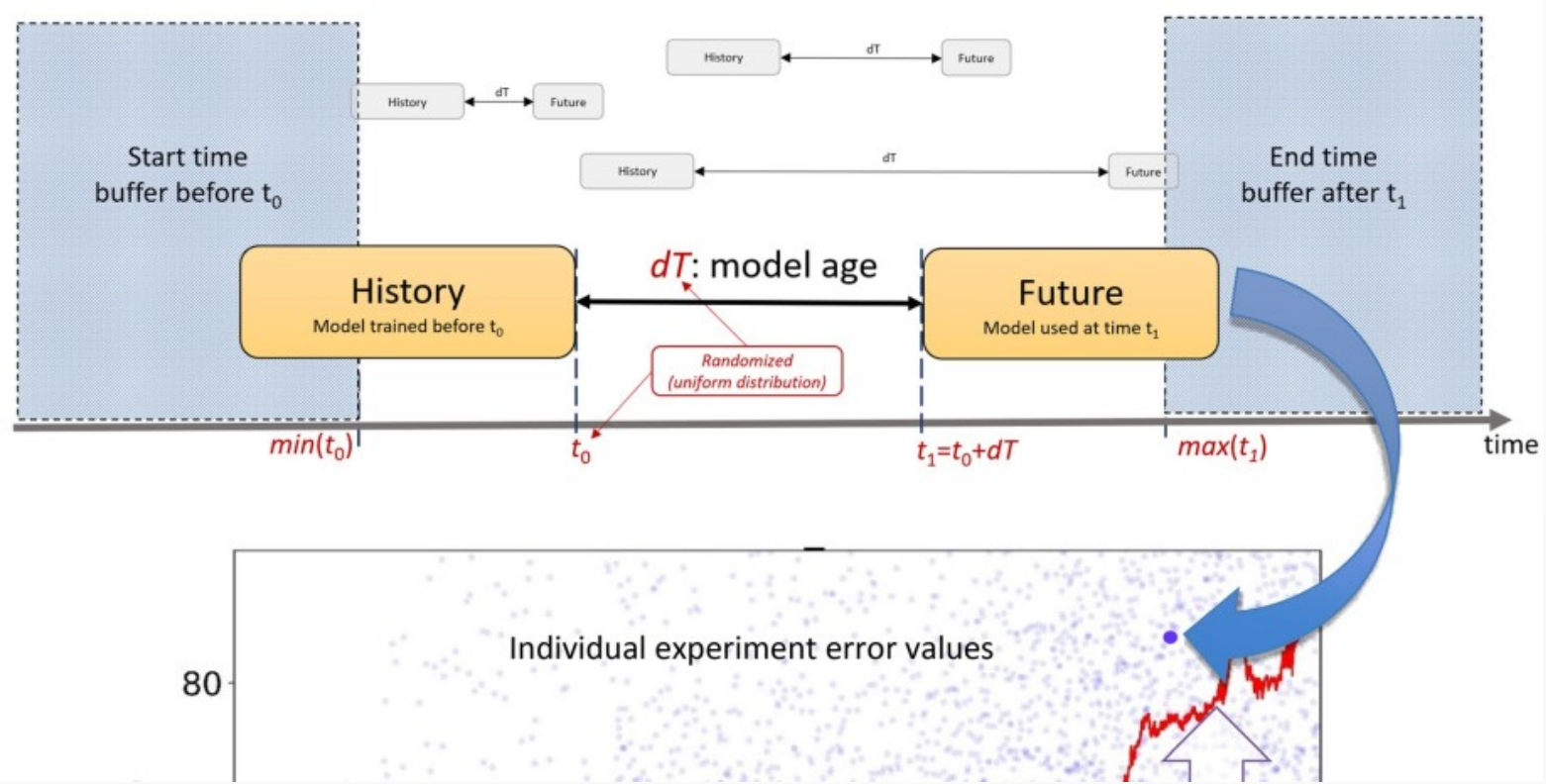
Про постепенное устаревание

Temporal quality degradation in AI models

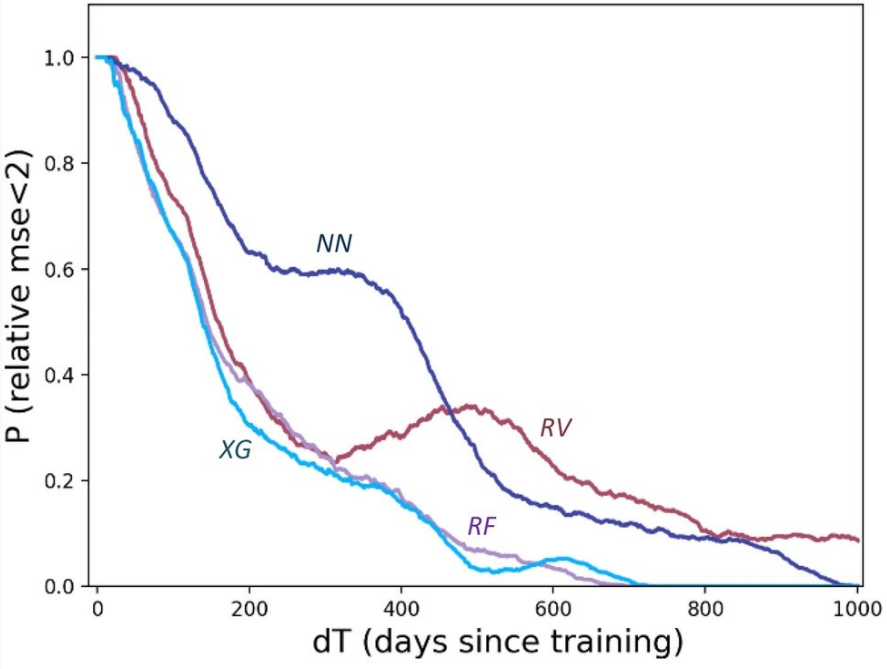
Daniel Vela[1], Andrew Sharp[4], Richard Zhang[2], Trang Nguyen[4], An Hoang[3] & Oleg S. Pianykh[4]✉

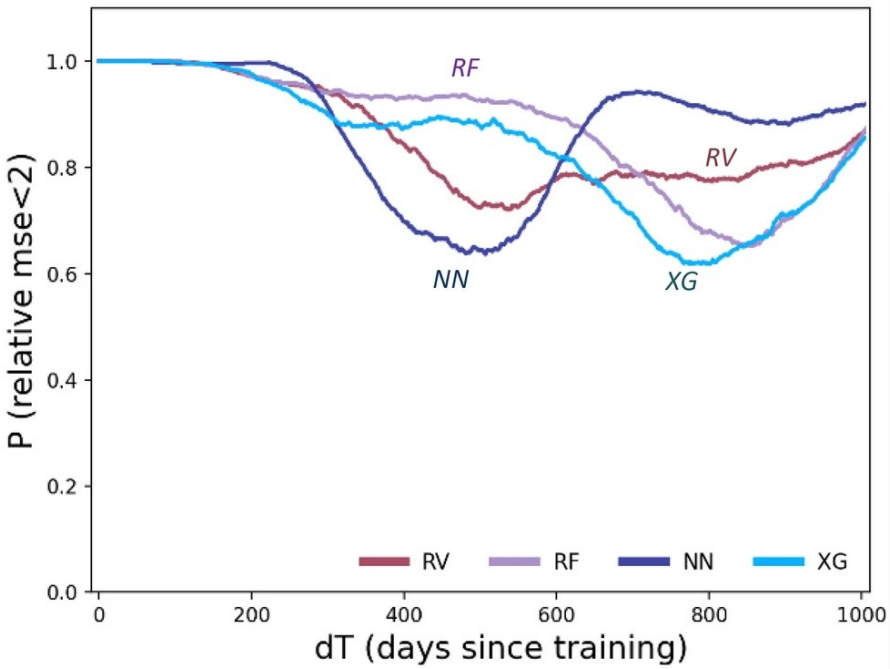
As AI models continue to advance into many real-life applications, their ability to maintain reliable quality over time becomes increasingly important. The principal challenge in this task stems from the very nature of current machine learning models, dependent on the data as it was at the time of training. In this study, we present the first analysis of AI "aging": the complex, multifaceted phenomenon of AI model quality degradation as more time passes since the last model training cycle. Using datasets from four different industries (healthcare operations, transportation, finance, and weather) and four standard machine learning models, we identify and describe the main temporal degradation patterns. We also demonstrate the principal differences between temporal model degradation and related concepts that have been explored previously, such as data concept drift and continuous learning. Finally, we indicate potential causes of temporal degradation, and suggest approaches to detecting aging and reducing its impact.

Моделируем устаревание



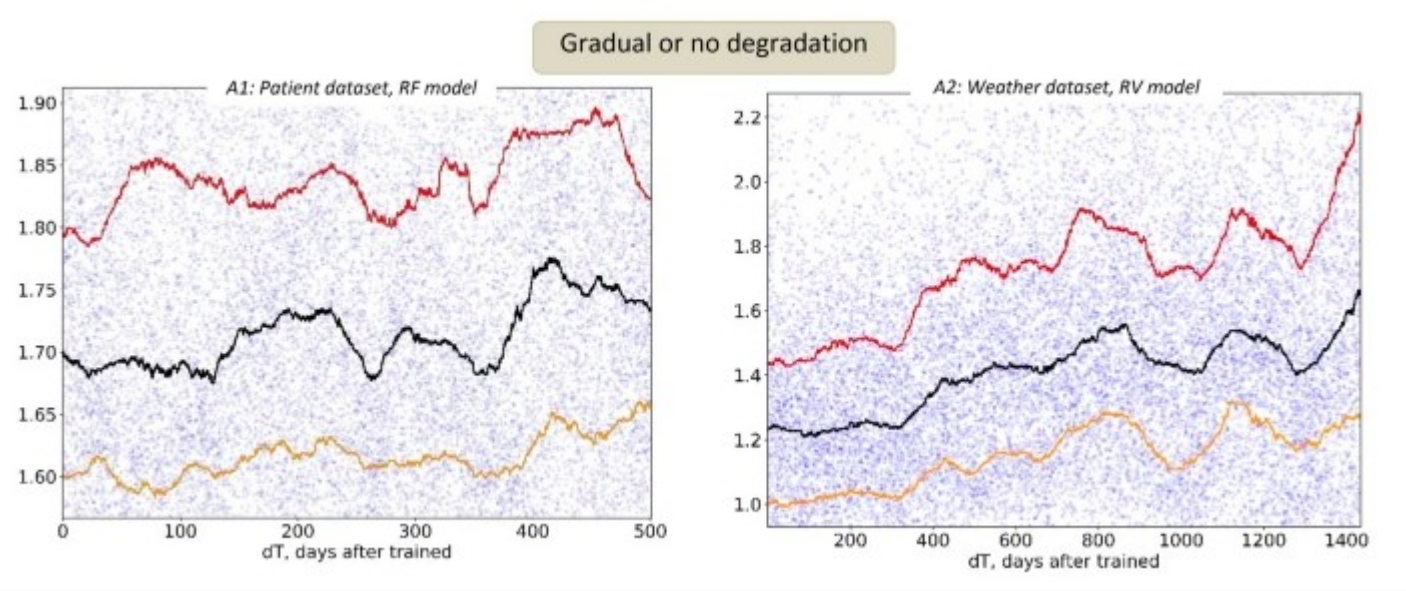
Деградируем





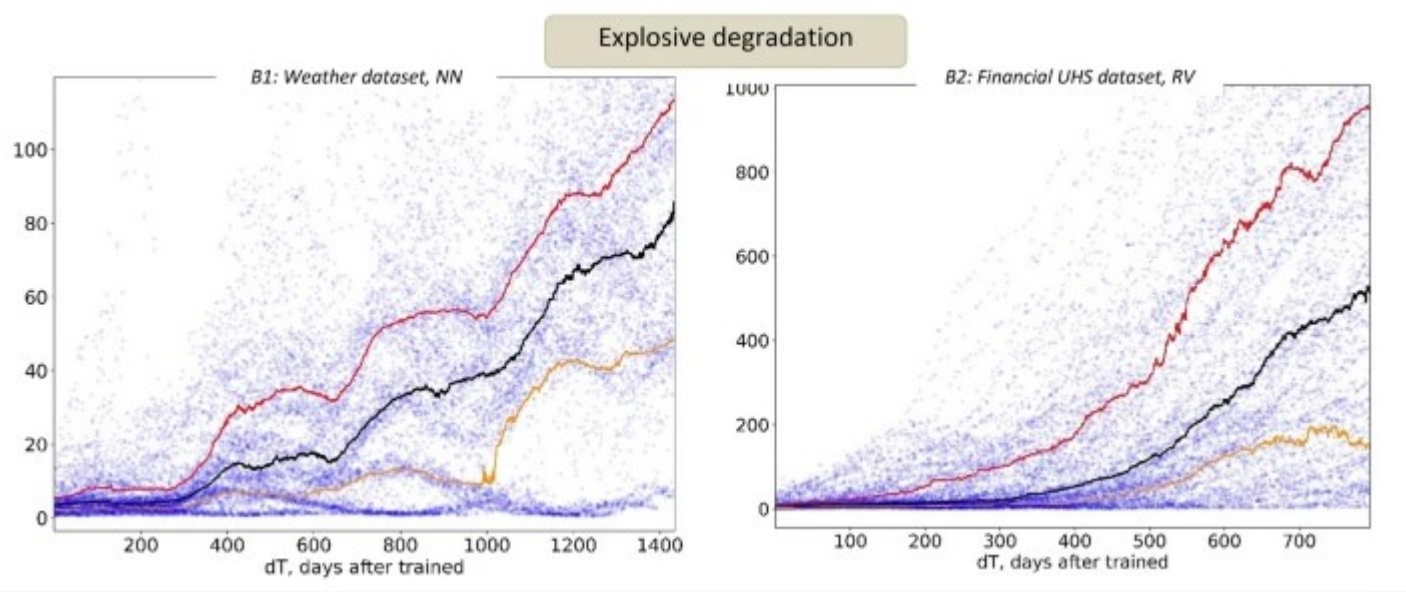
https://www.nature.com/articles/s41598-022-15245-z

Предсказуемая деградация



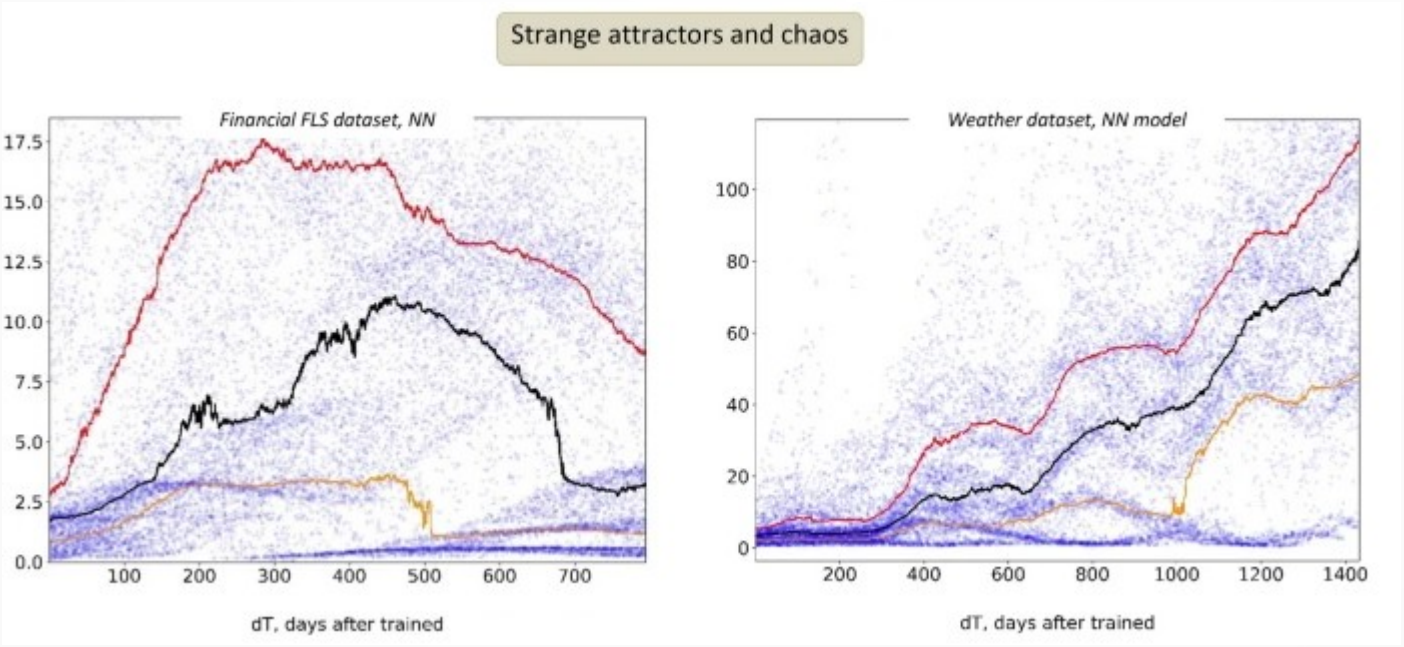
https://www.nature.com/articles/s41598-022-15245-z

Взрывная деградация



https://www.nature.com/articles/s41598-022-15245-z

Черт-те что



https://www.nature.com/articles/s41598-022-15245-z

Выводы

- Устаревание не обязано быть линейным
- Разметка не всегда есть
- Хотелось бы знать, что модель ушла не туда
  - Trust Score
  - Model Performance Predictor
  - Trust Score для LLM
- Без меток мы поймаем только Data Drift

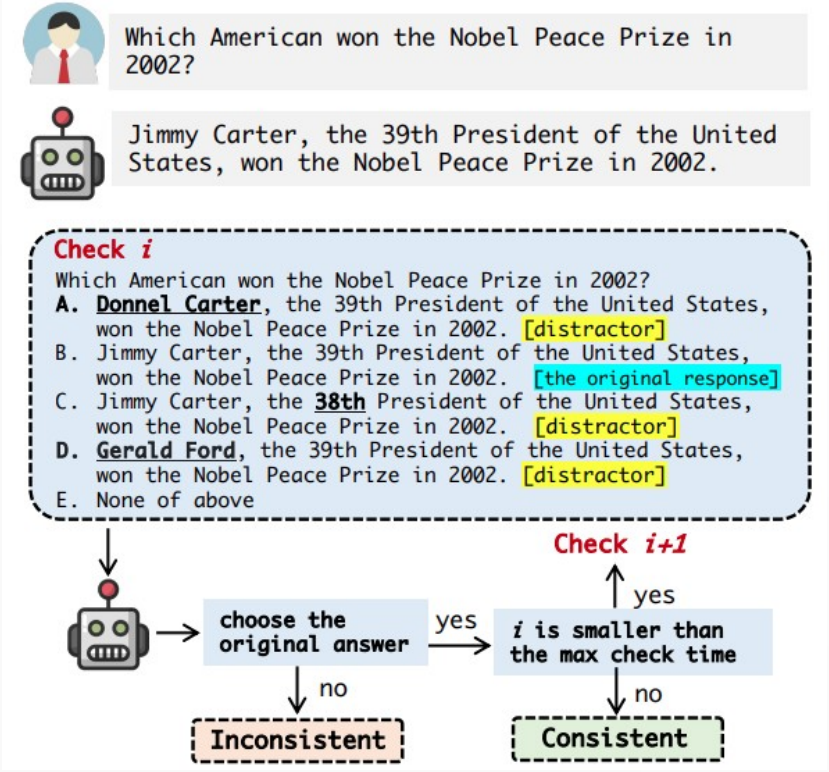
Trust Score

- Смотрим, насколько предикт поддержан обучающими данными
- Если у нас были похожие точки — доверяем
- Если не было — не доверяем ;-)
- Мониторим Trust Score

Model Performance Predictor

- Три набора данных:
  - Train учим модель1
  - Validation учим модель2 предсказывать ошибки модели1
  - Test проверям, что вышло
- На проде — модель 1 предсказывает
- На проде — модель 2 предсказывает качество
- Мониторим распределение предикта модели 2

Trust Score для LLM



https://arxiv.org/pdf/2402.12545.pdf

Как ловить Concept Drift

- RCD Reverse Concept Drift
- Shapley Values
- Множество подходов

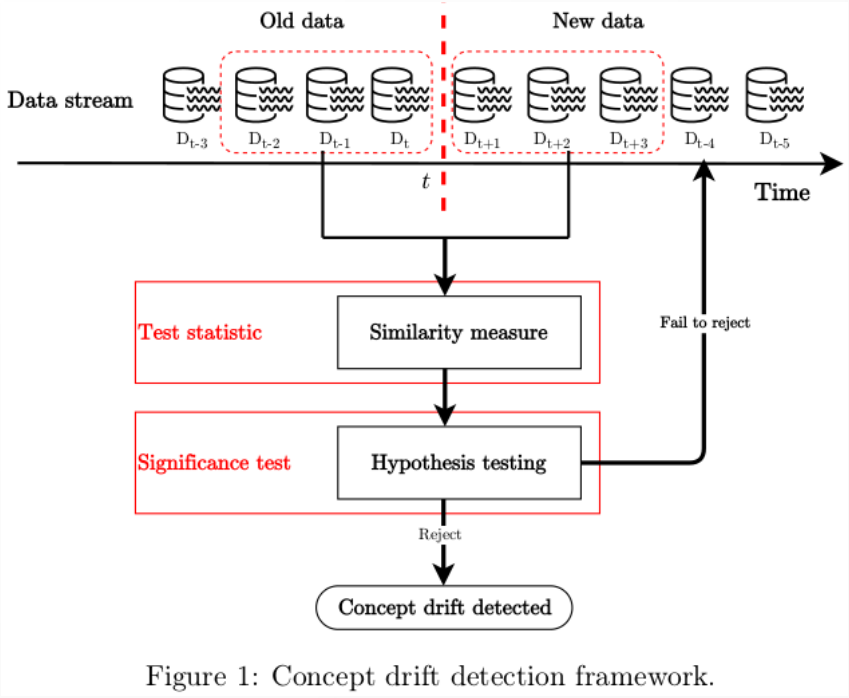
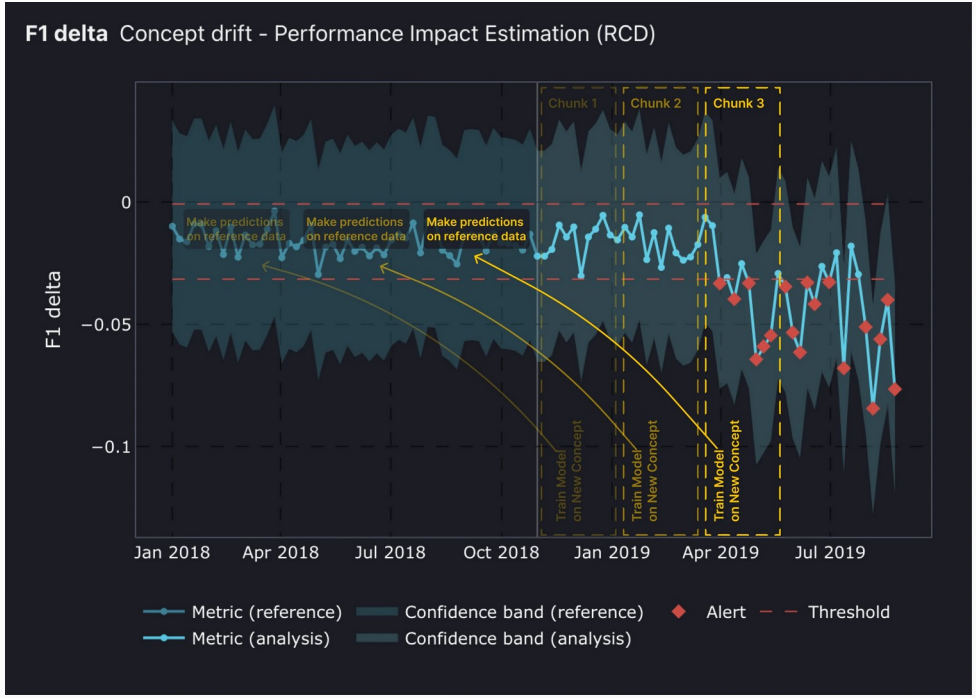


Figure 1: Concept drift detection framework.

Reverse Concept Drift

- Учим на новых
- Предсказываем старые
- Измеряем разницу
- Не метки, а скор
- См NannyML



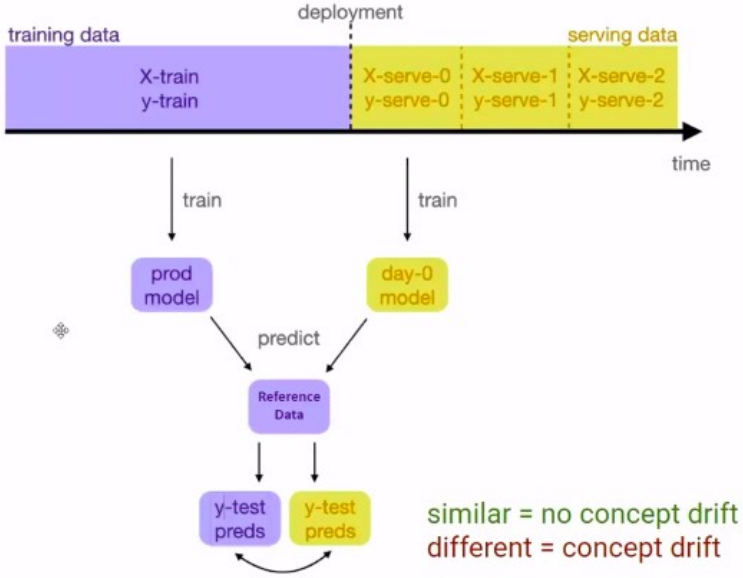
https://docs.nannyml.com/cloud/monitoring-deep-dives/reverse-concept-drift-rcd

https://www.reddit.com/r/MachineLearning/comments/1b2vsbb/r_reversed_concept_drift_rcd_and_algorithm_for/

Reverse Concept Drift

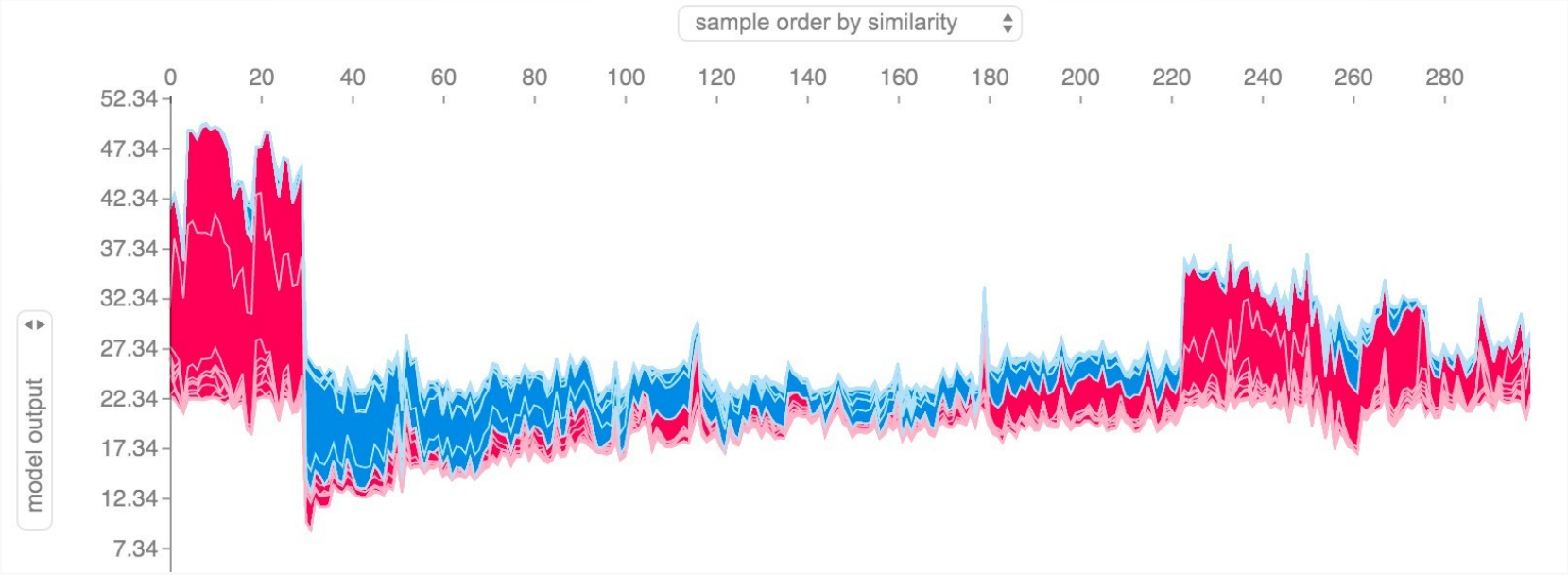
Reversed Concept Drift (RCD)

Intuition



https://www.youtube.com/watch?v=f7EzyecgSqQ

Shapley Values drift



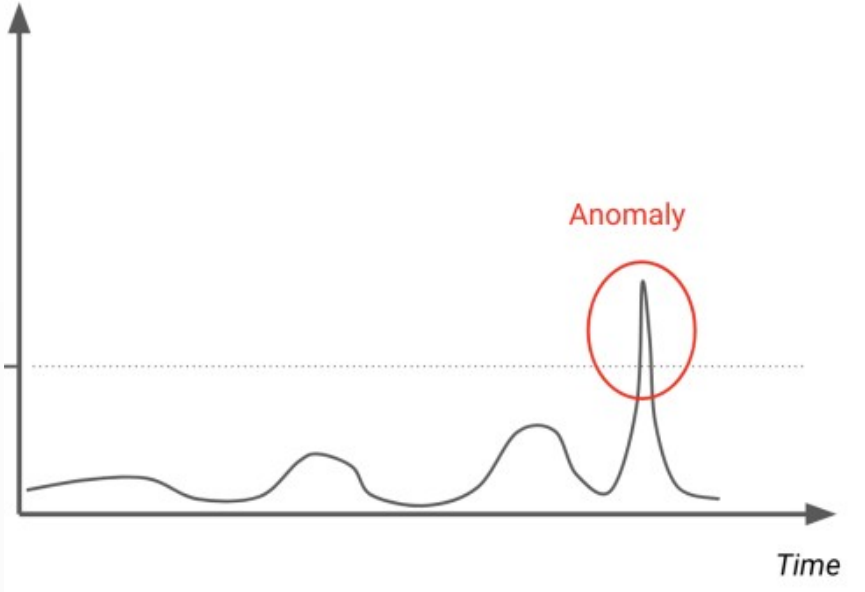
Из исходной статьи про Shap. Есть похожее

Evidently AI

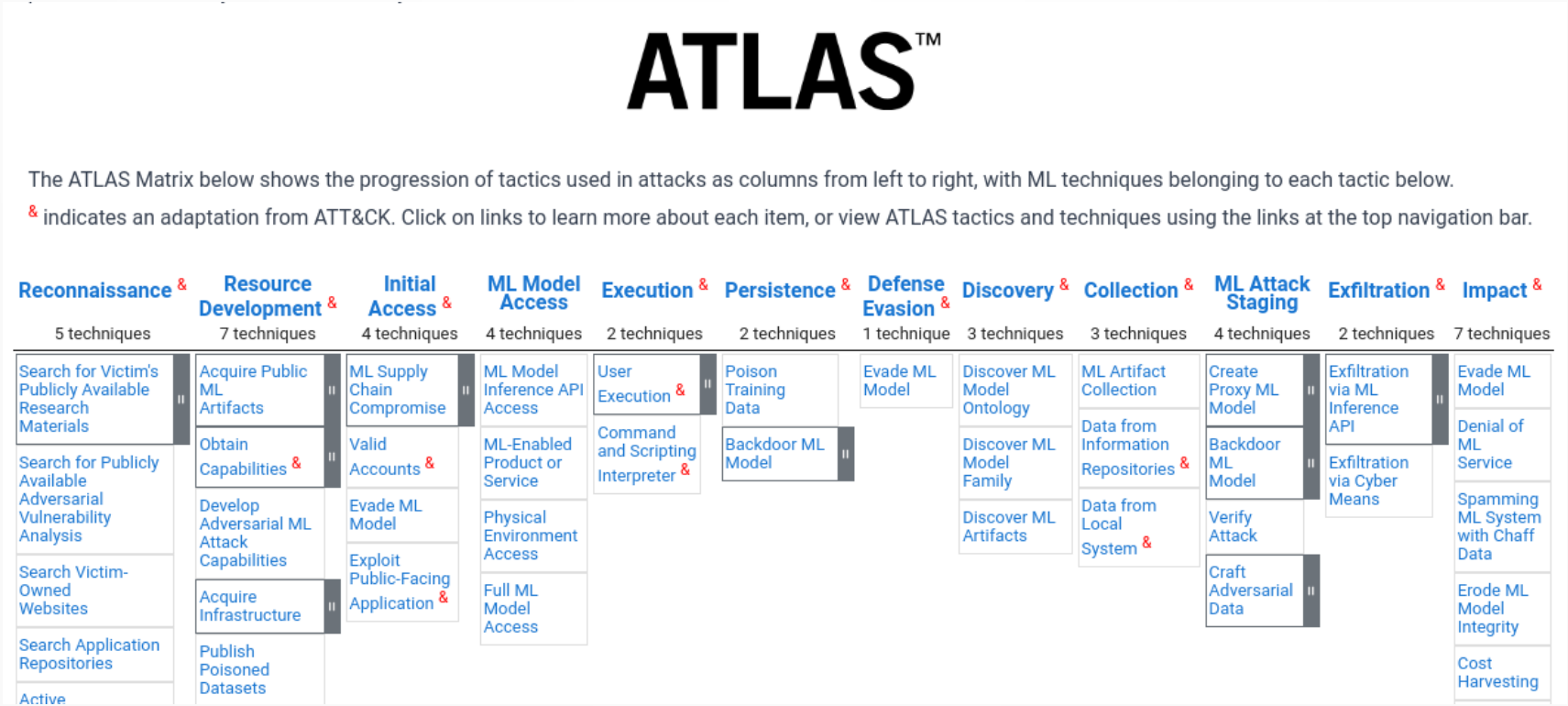
https://www.evidentlyai.com/blog/data-drift-detection-large-datasets

Мониторить предсказания

- Сдвиг данных
- Сдвиг концепции
- Сбой
- Атака
- Что мониторить:
  - Средний скор
  - Распределение скора

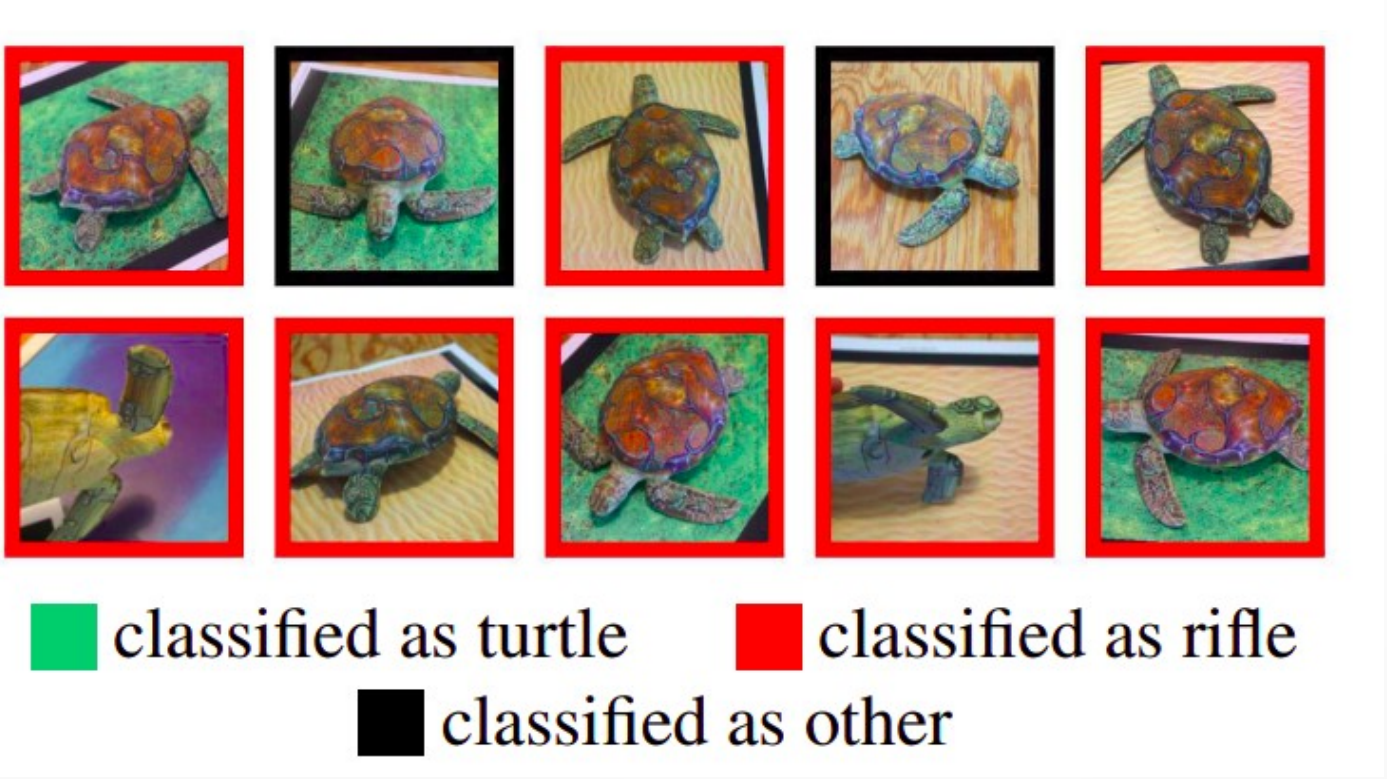


Атаки



https://atlas.mitre.org/

Нас атакуют



https://arxiv.org/abs/1707.07397

Вояки пока держатся

KEY FINDINGS

- Adversarial attacks designed to hide objects pose less risk to U.S. Department of Defense applications than academic research implies.
- In the real world, such adversarial attacks are difficult to design and deploy because of high knowledge requirements and infeasible attack vectors—there are often less expensive, more practical, and more effective nonadversarial techniques.
- Fusing data and predictions across sensor modalities, signal-sampling rates, and image resolution can further mitigate the risk of adversarial attacks.

RAND_RRA866-1

data scientist in a modern suit holding a sword piercing through complex structure made out of different diagram elements

Банки пока держатся

Data Fusion Contest 2023

Ежегодное соревнование по машинному обучению Data Fusion Contest. В 2023 году это турнир по Adversarial ML между командами атакующих и защищающих ML модели на транзакционных данных.

21 18 17 14 11 9 4 4 3 2 2 1 +

Ежегодное соревнование Data Fusion Contest 2023 продолжается! Регистрация открыта для участников до 2 апреля!

Вас ждёт уникальное соревнование по атакам и защите моделей машинного обучения в турнирном формате:

🗡️ В задаче **Атака** участники будут создавать атаки на нейросеть, обученную на данных транзакций.

🛡️ В задаче **Защита** — наоборот, учиться защищать свои модели от заранее оговоренного вида атак.

🏆 Призеров определят **Турниры** — лучшие команды обеих задач столкнутся друг с другом за призовой фонд в **2 000 000 рублей**!

https://ods.ai/tracks/data-fusion-2023-competitions

Поисковые машины - нет

накрутка поисковых подсказок

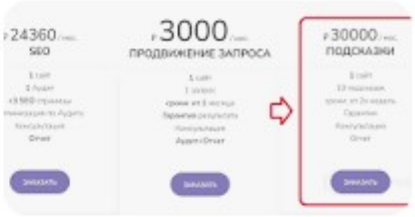
Все | Видео | Картинки | Новости | Книги | Ещё | Инструменты

Результатов: примерно 2 430 (0,38 сек.)

Накрутка поисковых подсказок относится к «серым» способам продвижения.

...

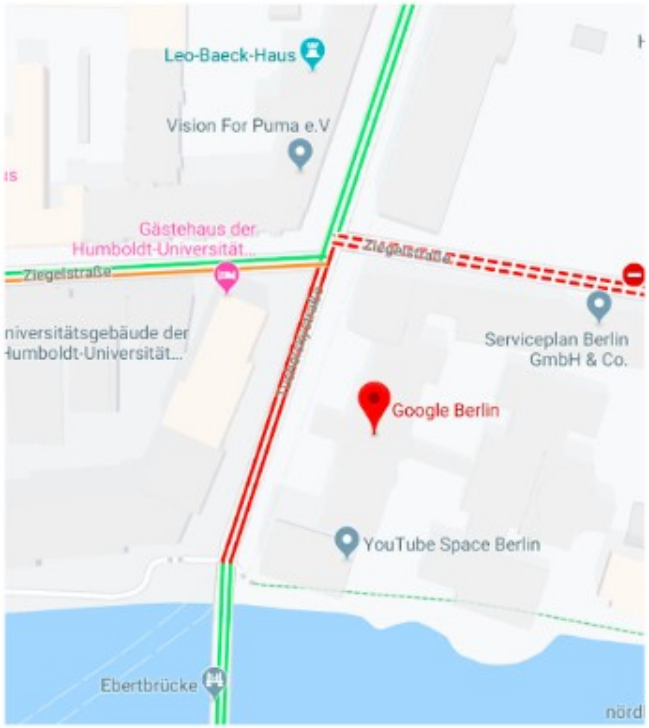
Эксперты, которые плотно работают с накруткой поисковых подсказок, отмечают следующие санкции за некачественную накрутку:



1. Чистка списка **подсказок**. ...
2. Временный или постоянный фильтр на вывод **подсказки** или бренда на определенном ключевом слове.

Ещё • 24 сент. 2022 г.

Против тачки нет приема



https://www.simonweckert.com/googlemapshacks.html

Что почитать

- Supervisory Guidance On Model Risk Management
- Machine Learning for High-Risk Applications
- Graceful Degradation and Related Fields
- Блог nannyML
- Блог Evidently AI
- Introduction to streaming for data scientists
- Переобучению быть или не быть
- Пару слов о дрейфе данных
- Курс ML System Design

Вопросы

Слайды тут

dkolodezev

dmitry_kolodezev

Promsoft

Reliable ML

https://kolodezev.ru/download/sustainability2024.pdf