



ОНЛАЙН-ОБРАЗОВАНИЕ

# Онлайн-образование





# Меня хорошо видно && слышно?

Ставьте ++, если все хорошо  
Напишите в чат, если есть проблемы



Проверить, идет ли запись!





# «Обнаружение сдвигов в данных»

---

Миленькин Александр



Senior Data Scientist  
X5 Retail Group  
Otus Slack: AleMil



# Преподаватель



Миленькин Александр

- 6 лет в IT
- **Сейчас:** ML Engineer в **Redmadrobot**
- **Ранее:** Старший менеджер по работе с большими данными в **X5 Retail Group**
- **Еще ранее:** Старший аналитик в **Асна**, data scientist в **Gero**, биоинформатик в **Insilico Medicine**.
- Выпускник и преподаватель **МФТИ** (Физтех)
- Победитель нескольких хакатонов по анализу данных.
- **Kaggle Expert**



Insilico  
Medicine

асна



GERO



RED  
mad  
ROBOT

# Правила вебинара



Активно участвуем. Реагируем в чате



Задаем вопросы в чат или голосом. Лучше голосом)



Off-topic обсуждаем в Slack #канал группы или #general



Вопросы вижу в чате, могу ответить не сразу

# Маршрут вебинара

Введение в сдвиги



Теория



Практика поиска сдвигов



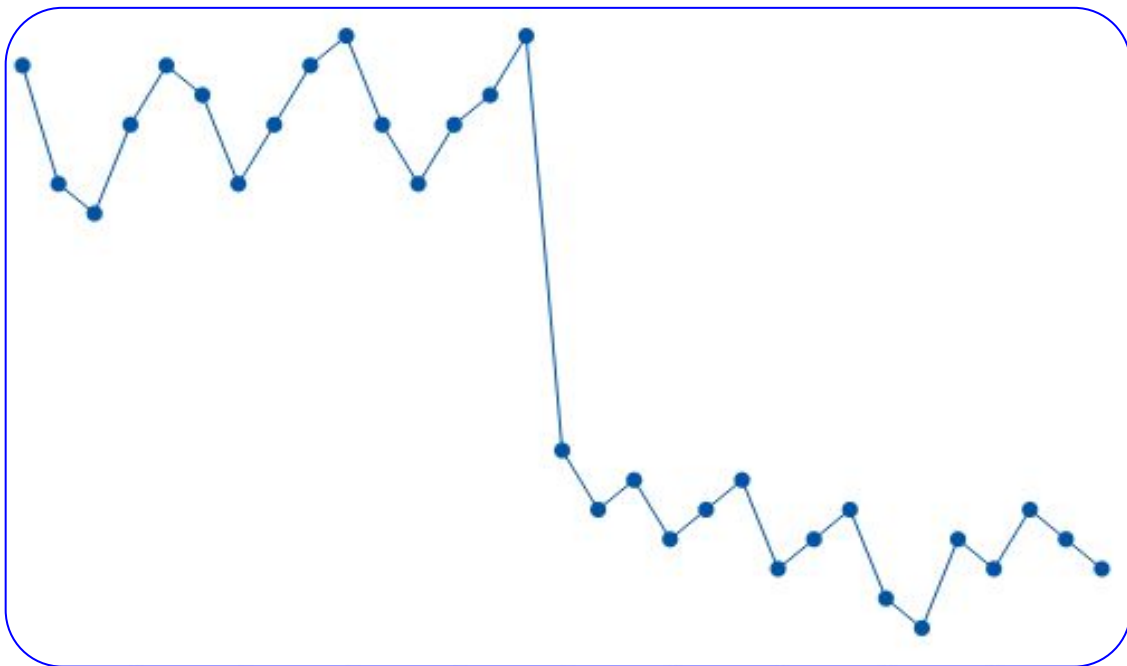
Рефлексия

# Цели занятия

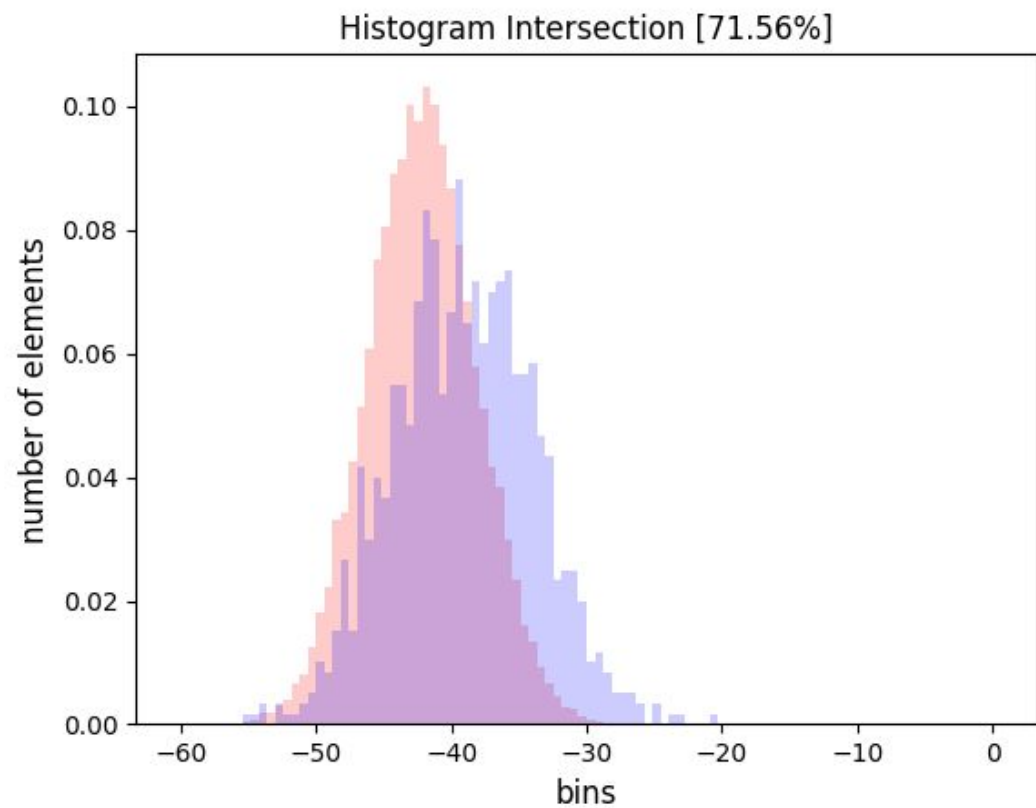
- разобраться, что такое сдвиги и откуда они берутся
- ознакомиться с основными способами их поиска
- научиться искать сдвиги на практике



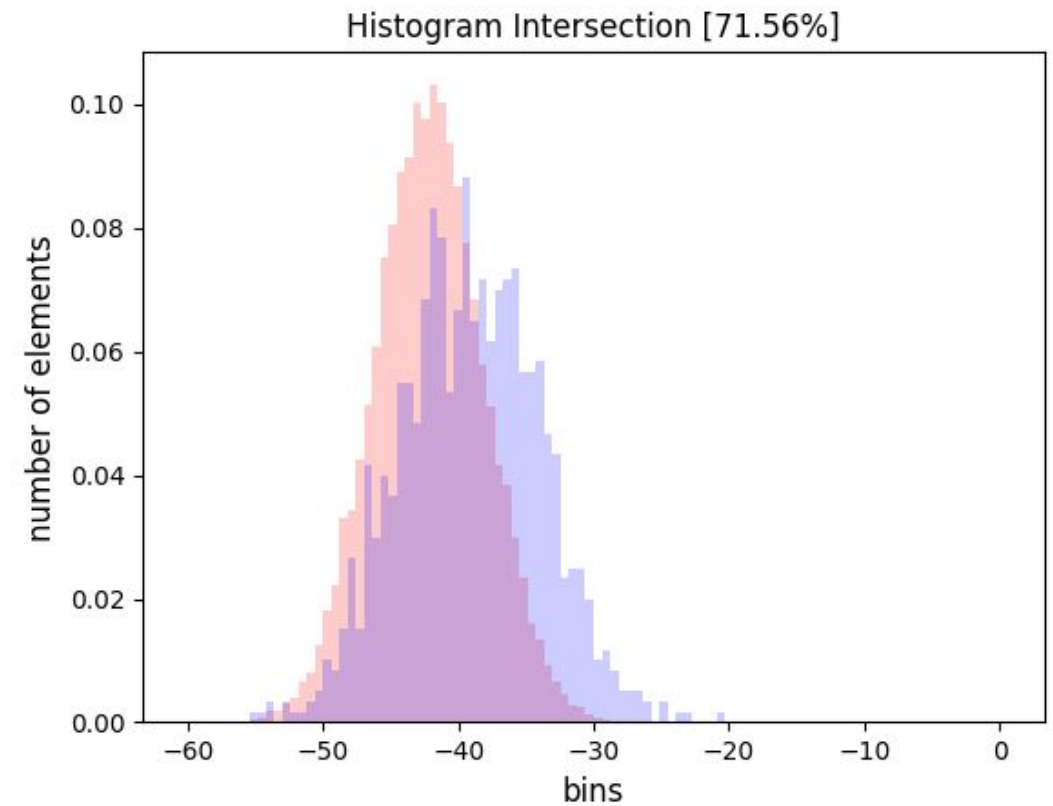
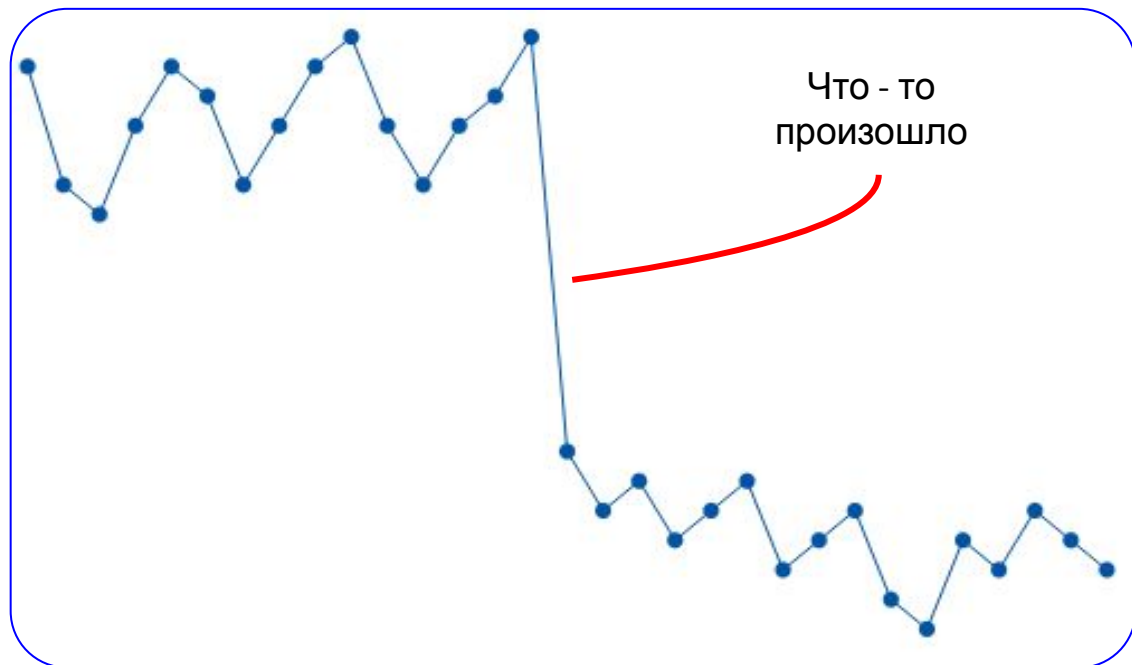
# Сдвиг / он же шифт (shift)



**Причины:** События или просто устаревания данных



# Сдвиг / он же шифт (shift)





# Почему плохо?

- ML плохо обобщаются на данных с разными распределениями
- Снижаются метрики точности. Надо переобучать модели.
- Валидация перестает отражать истинную ситуацию

# Как распознать сдвиг в данных?

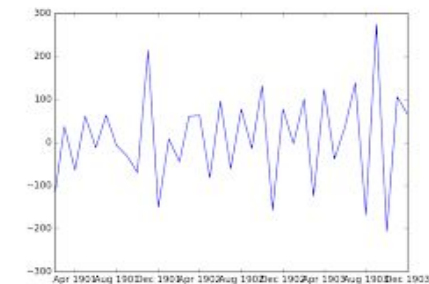
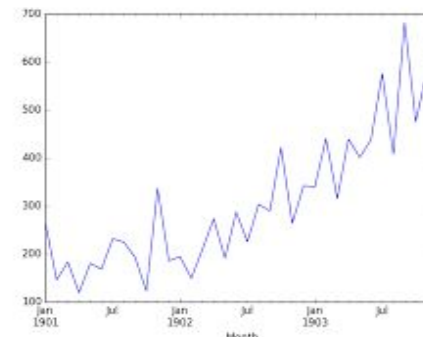
- Визуально (подходит только, если сдвиг произошел в одно фиче)
  - Строим временной ряд
  - Гистограмма / Бокс - плот
- Статистические критерии (t-test / Манна-Уитни / ДИ)
- Adversarial validation (Пусть ищут ML модели)

\*population stability index



# Как бороться с шифтами

- **Мониторить!** (алерты в пайплайне / tests)
- Переобучать модели
- Удаляем shift'ные параметры
- Дифференцирование временного ряда
- Нормализация VS Стандартизация
- Аугментация данных
- Костыли (выровнять разницу)



# Еще пример

## Scented candles and COVID-19

Before COVID-19:

- "No scent" review and low rating  
-> likely the product is bad

After COVID-19:

- Much more "No scent" reviews
- "No scent" review and low rating  
-> doesn't mean the product is bad

The review contents didn't change! Very hard to infer problem from data alone.



**До ковида данные содержали такую зависимость:** если в отзыве пишут, что свечка без запаха, то свечка плохая.

**После ковида зависимость изменилась:** отзывы про отсутствие запаха не означают, что свечка плохая.

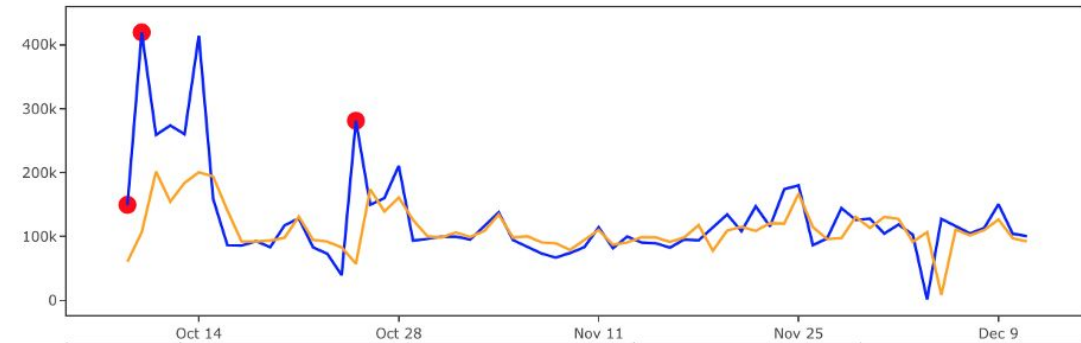
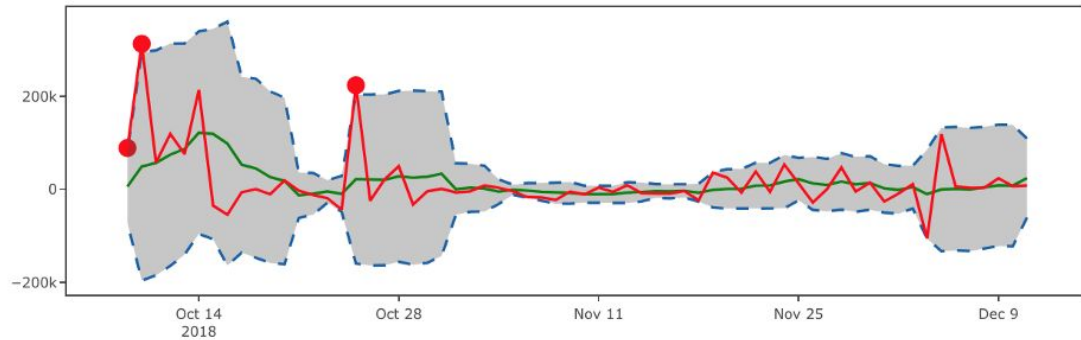
При этом с точки зрения "данных" ничего не меняется. Отзывы состоят из тех же слов, распределения фичей почти не меняются. Одна из тех проблем, которую очень сложно заметить смотря только на данные.

Мониторьте свои модели!



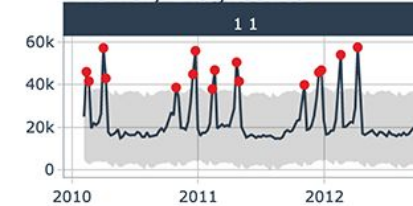
# Еще про тулы

metric\_name



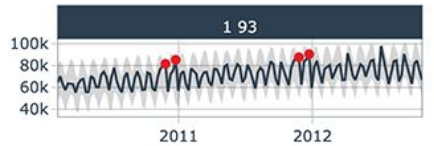
Date	Actual Values	Predicted	% Difference	Severity (0-3)
2018-11-23	113732	120900.074	-6.333%	0
2018-11-22	147069	108739.168	26.062%	0
2018-11-21	108126	114978.294	-6.337%	0
2018-11-20	134514	109228.065	18.798%	0
2018-11-19	113630	77791.011	31.54%	2
2018-11-18	94102	117629.725	-25.002%	1
2018-11-17	95410	98640.56	-3.386%	0

## Anomaly Diagnostics



Anomaly

• Yes



# Дополнительно

- Еще статья про shift'ы:

<https://towardsdatascience.com/understanding-dataset-shift-f2a5a262a766>

Пример борьбы с shift в train и test для улучшения валидации

<https://dyakonov.org/tag/adversarial-validation/>



# Рефлексия




С какими основными мыслями и инсайтами уходите с вебинара



Каких целей вебинара не удалось достичь



The background of the entire image is an aerial photograph of a city, likely New York City, showing numerous skyscrapers and buildings. The image is overlaid with a semi-transparent blue layer. In the center of this layer, there is a network of white lines connecting various points, creating a geometric pattern. The text is written in white, bold, sans-serif font, centered within this blue area.

Заполните, пожалуйста,  
опрос о занятии по ссылке в чате



# Спасибо за внимание!

## Приходите на следующие вебинары

---



Миленькин Александр

Senior Data Scientist  
X5 Retail Group  
Otus Slack: AleMil