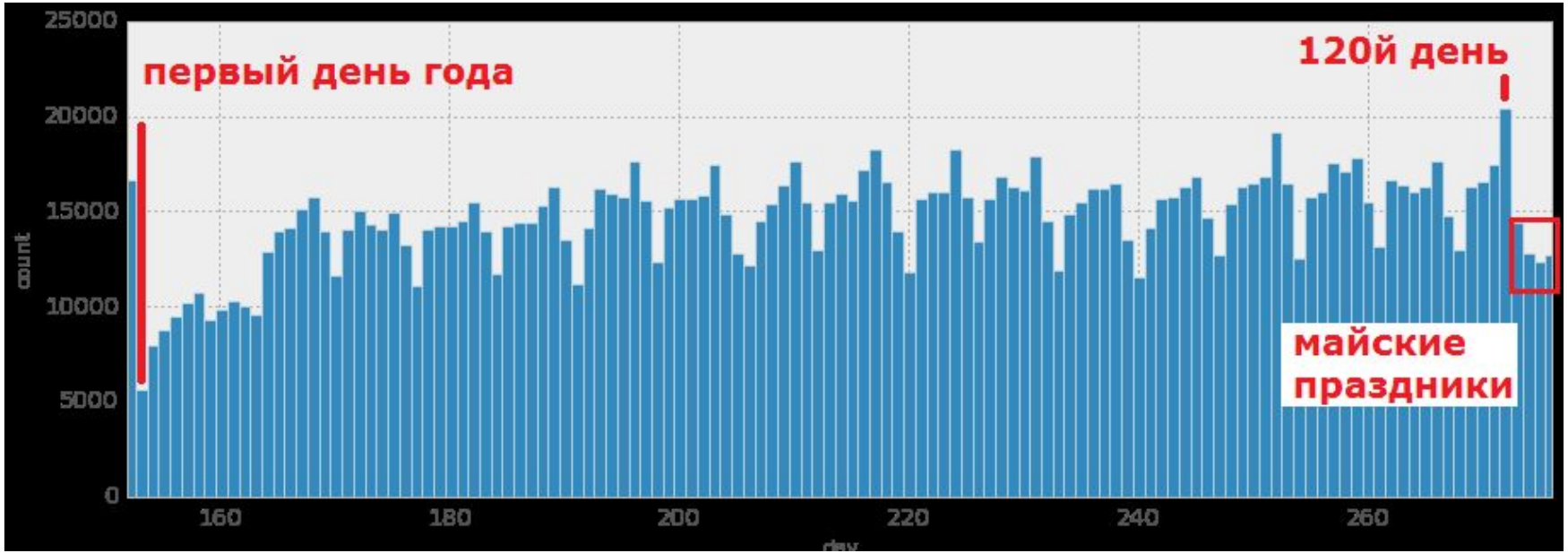
Sberbank Data Science Journey

Хасянов Расул, МФТИ, 4 курс

Данные какого года?

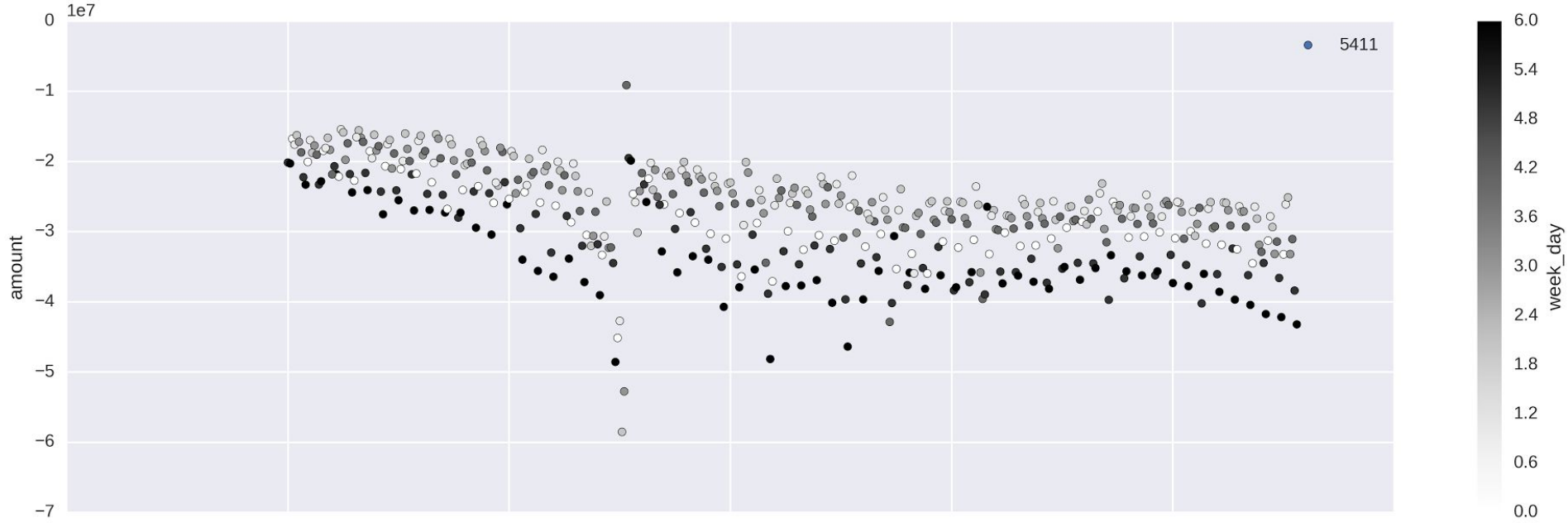
Август 2014 - Октябрь 2015



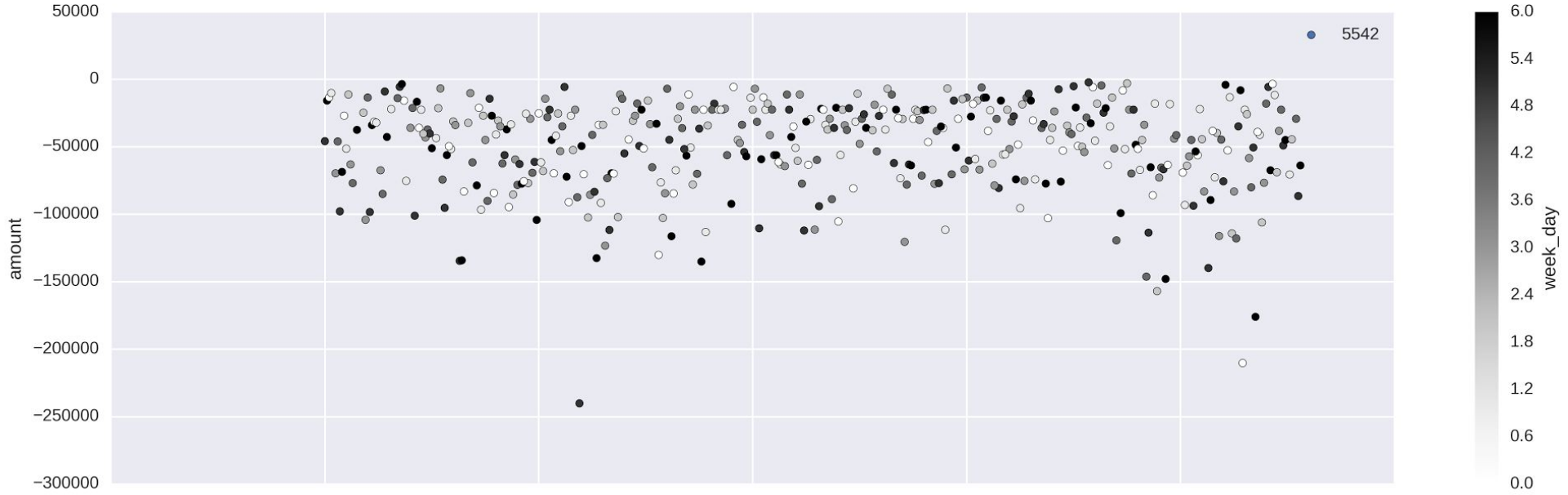
https://contest.sdsj.ru/#/posts/show/7?_k=vstf65

Временные ряды

Гипотеза: в среднем данные повторяются каждые 7 дней.

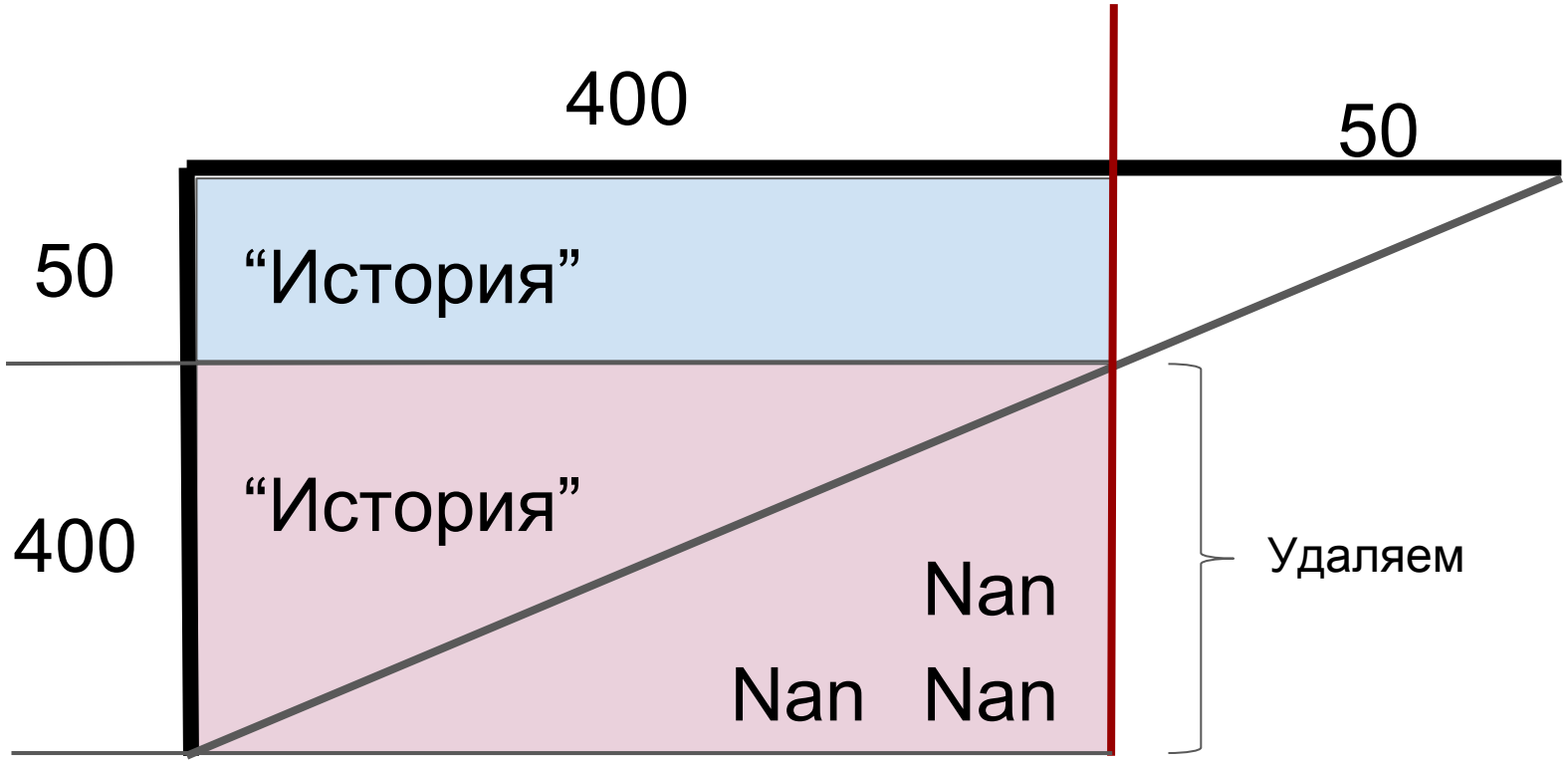


Вообще говоря,
не всегда



Получаем больше признаков

- Всего дней: **450**
- Признаки: **400** дней



Эксперимент

1. Тест: последний месяц из train
2. Модель: Ridge-регрессия
3. Меняющиеся параметры:
    - Длина истории - основной параметр
    - Объем выборки

Сколько брать дней в качестве признаков?

- Признаков **больше**, выборка **меньше**: тренируемся предсказывать меньше дней, но на основе хорошей истории
- График - 1.5635; Итог - 1.5616

Вывод

- Чем больше признаков, тем лучше!
- Основная задача: получить больше признаков
  - Заполнить средними значениями
  - Взять значения из “будущего”
- Обучаемся, пока число признаков << размер выборки
  - Признаков ~250, выборка ~8500 => ~0.03
  - На обучении получилось ~0.015

Итоговое решение

- Lasso с регуляризацией 0.0001
- 52 временных признака
- Обучение на основе 93 дней. Остальные 364 дня - в признаках

Итоговый результат: **1.5616**