



# Конкурс Avito.ru:

Задача поиска контактной информации на изображениях и возможные подходы к её решению

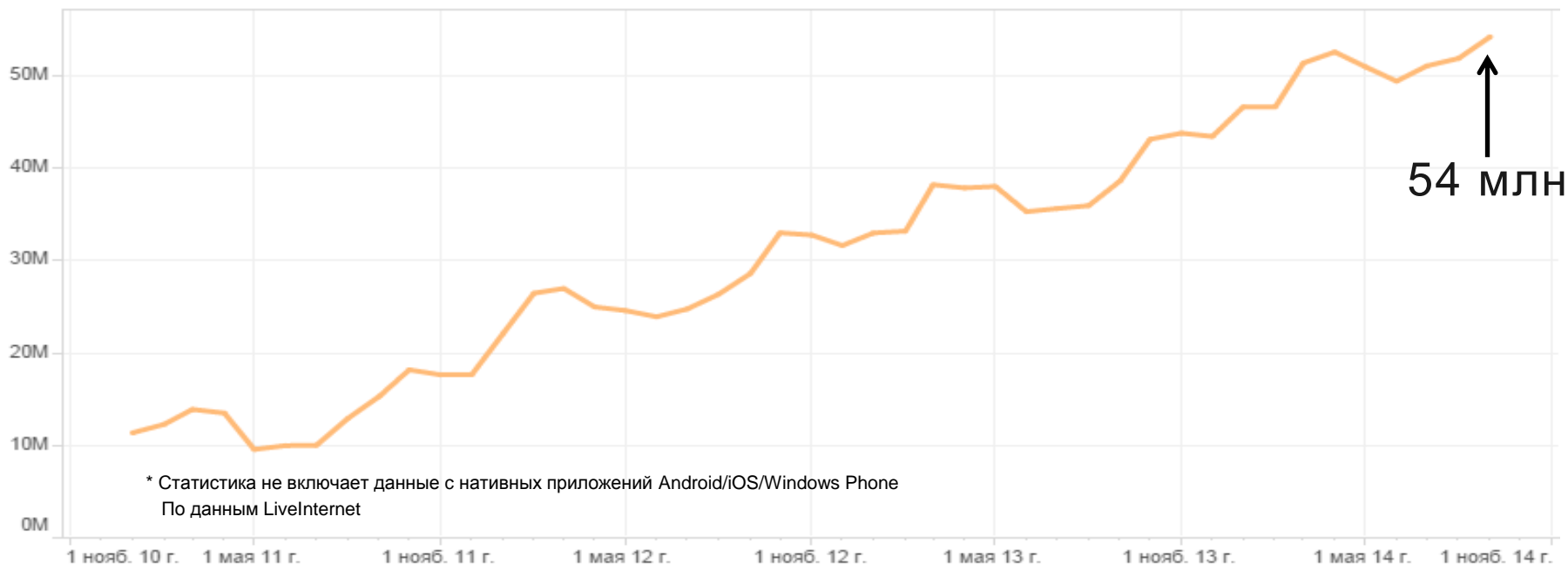
Иван Гуз  
к.ф.-м.н

# Что такое Avito.ru?

- **#1 сайт объявлений в России\***
  - 54млн посетителей / (с моб. версией)
- 150-200млн просмотров в день – **один из крупнейших сайтов объявлений в мире\***
- 25 млн активных объявлений\*\*<sup>\*</sup>: Недвижимость, Транспорт, Работа, Услуги, Товары для дома и т. д.
- Freemium бизнес модель
  - Продвижение объявлений с помощью платных сервисов
  - Реклама
  - Специальные услуги для SMB клиентов

# Ежемесячная аудитория Avito.ru

## Уникальные посетители\* avito.ru



## Новые объявления Gross vs Net



# Почему контакты на фото запрещены?

- Для указания телефонов/email есть специальное поле. Спамеры заполняют их несуществующими или фиктивными значениями чтобы избежать блокировки по черным спискам
- Контакты в тексте объявления мы умеем эффективно находить. Их туда не спрятать
- Кроме телефонов на фото могут указывать прочие контакты (url, email, название компании), по которым также можно найти способ связи

**Подача объявления**

Частное лицо  Компания

Ваше имя

Электронная почта

Я не хочу получать вопросы по объявлению п

Телефон

Город  -- Выбери

Категория

Выберите параметры

Название объявления

Описание объявления

*Fake!* (pointing to email)

*Fake!* (pointing to phone number)

*Fake!* (circled around "Телефон на картинке" in the description)

# Примеры изображений

Есть контакты

Нет контактов

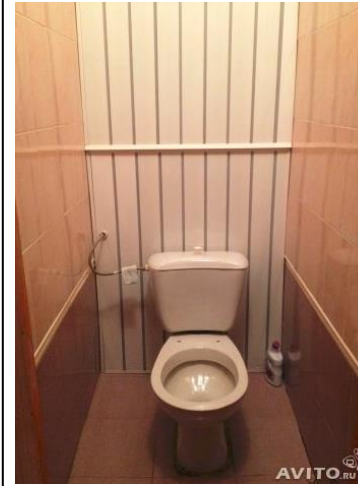
Просто

**НЕБОСКРЁБ** великие Луки СЕТЬ АГЕНТСТВ НЕДВИЖИМОСТИ  
www.neboskrebvl.ru

ул. Дьяконова 10	☎ 5-77-33	+7 911 895-0444
ул. Ставского 22	☎ 7-36-60	+7 911 351-2288
пр. Гагарина 7	☎ 3-39-12	+7 911 387-0888
ул. Щорса 9/3	☎ 6-92-13	+7 911 881-4858

РОСЕЕСТР  
пр. Октябрьский 65 ☎ 3-45-45 +7 911 888-1122

8-951-181-5803  
- Виталий



Сложно



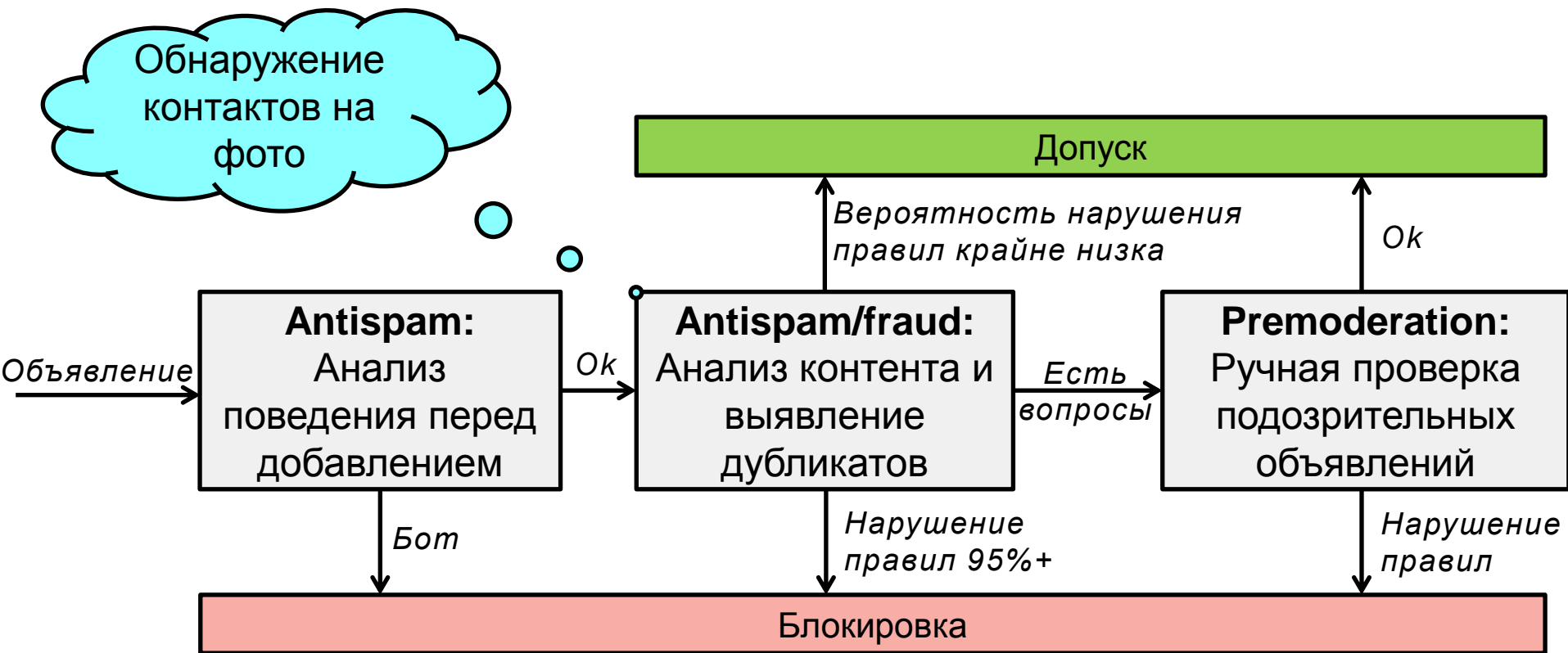
# Где больше всего спама?

Там где стоимость товара самая большая – квартиры



# Процесс модерации:

## 3 уровня защиты: Машины + Машины + Люди



# Критерии качества моделей

## Реальная жизнь:

- Команда модераторов, которая может проверить не более  $K$  объявлений
- 1 объявление  $\rightarrow$  0, 1, 2, ... изображений (фото)



Модель должна ранжировать объявления по вероятности наличия контактов. Чем лучше качество ранжирования в Top  $K$  – тем лучше модель



Доля с контактами среди первых  $i$

$$AP@K \text{ (Average Precision @ } K) = \sum_{i=1}^k P(i) * \begin{cases} 0, \text{ если в } i \text{ нет контактов} \\ \frac{1}{\min(k,m)}, \text{ если в } i \text{ есть контакты} \end{cases}$$

Всего объявлений с контактами

## Конкурс:

- Плоская выборка изображений
- Максимизировать AUC



# Обучающие данные и этапы конкурса

1. За 3 месяца выгрузили все картинки объявлений, заблокированных по причине «контакты на фото»
2. За этот же период выгрузили случайную выборку картинок объявлений, которые были активны достаточно долго и не были заблокированы
3. Отдали 51 122 картинок (п.п 1, 2) на повторную ручную разметку

41 122 изображения – обучающая выборка (разметка дана)

10 000 изображений – тестовая выборка (разметки нет. На ней измеряется качество)

До 14 ноября\* – обучение моделей, отправка (заморозка) моделей

10 000 изображений, аналогичная методика подготовки за более актуальный период

До 19 ноября – скоринг новых картинок с помощью готовых моделей

Определение победителей, проверка воспроизводимости – до 10 декабря

**Важно! Использовать любую информацию кроме изображения бессмысленно!  
В контрольной выборке имена файлов будут случайными.**

# Подходы к решению задачи

- В идеале нужно решать задачу “поиска искусственно нанесенного на картинку текста” => как это делать – не ясно
- Плохо работает: бездумное использование OCR (пробовали Tesseract)
- Лучше работает: бинаризация текста и выделение связных компонент. По этому принципу работает алгоритм SWT (Stroke width transform). Какие компоненты называть текстом/контактами – это эвристика



(a)



(b)



## Отбор победителей, правила и призы

- 1е место (100 000 руб. включая налоги) – максимальный AUC на контрольной выборке среди всех участников
- 2е место (50 000 руб. включая налоги) – максимальный AUC на контрольной выборке среди оставшихся участников, **использующих ПО для некоммерческого использования.**

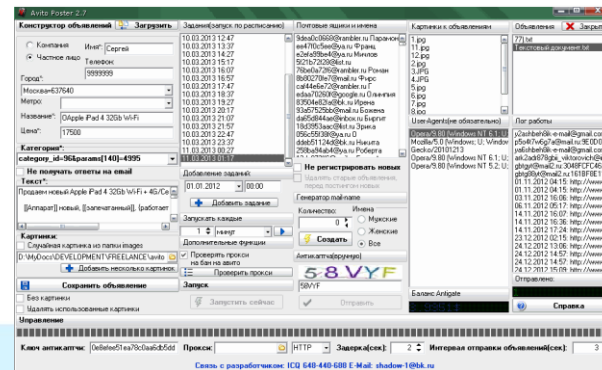
Участники: **только граждане РФ, достигшие 18 лет**

Победителю при необходимости нужно будет показать, как получается соответствующий результат с помощью отправленной модели

# Чего мы хотим добиться?

## Повысить порог входа в спамерский бизнес

	Любители (Физ. лица)	Профессионалы (Юр. лица)
<i>Штат</i>	1 чел	Разработчики, дизайнеры, продавцы, служба поддержки
<i>Маркетинг</i>	Форумы	Реклама, прямые продажи, Email
<i>Объем спама</i>	10-100 в день	100K+ в день
<i>Бизнес модель</i>	C2C	B2B
<i>Обороты</i>	Карманные деньги	MM+
<i>Технологии</i>	Толстый клиент	Облачные платформы
<i>Скорость обновления</i>	Неделя	Несколько раз в день



**Профессиональный сервис публикации объявлений**

Предоставляем Вашему вниманию услуги по привлечению клиентов в сфере недвижимости по средствам массовых публикаций на главной площадке рунета - Avito.

- ✓ Огромная база фото реальных квартир
- ✓ Адаптация к ценам Вашего региона
- ✓ Использование собственного пула сменных номеров для постинга с переадресацией на ваш служебные телефоны

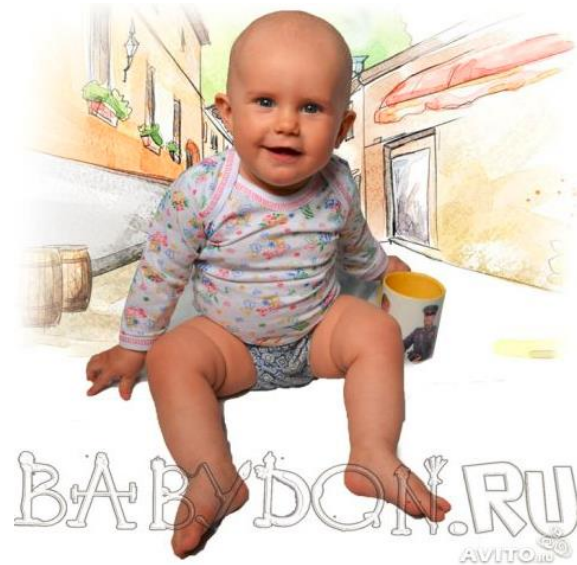
от 6000/мес. **Free test**

**Больше клиентов без трудозатрат с вашей стороны**  
Мы поможем Вам стать заметнее на главном сайте по поиску квартир!

- Больше договоров**  
Странным быть лидером в конкурентной сфере!
- Простота в использовании**  
Мы все делаем сами и не нагружаем Вас работой.
- Вариативность**  
Готовы адаптировать наш сервис под любые Ваши нужды.
- Опыт**  
Ежедневно улучшаем качество постинга.
- Поддержка**  
Всегда готовы помочь 24/7.
- Отчеты**  
Список ссылок на ваши объявления доступен онлайн.

Уникальные методы публикаций!  
2 года работы в этой сфере позволили нам выработать комплекс для публикации, который сводит к минимуму риски от удаления

Спасибо за внимание!



[machinelearning.ru](http://machinelearning.ru)  
[competition.avito.2014@forecsys.ru](mailto:competition.avito.2014@forecsys.ru)