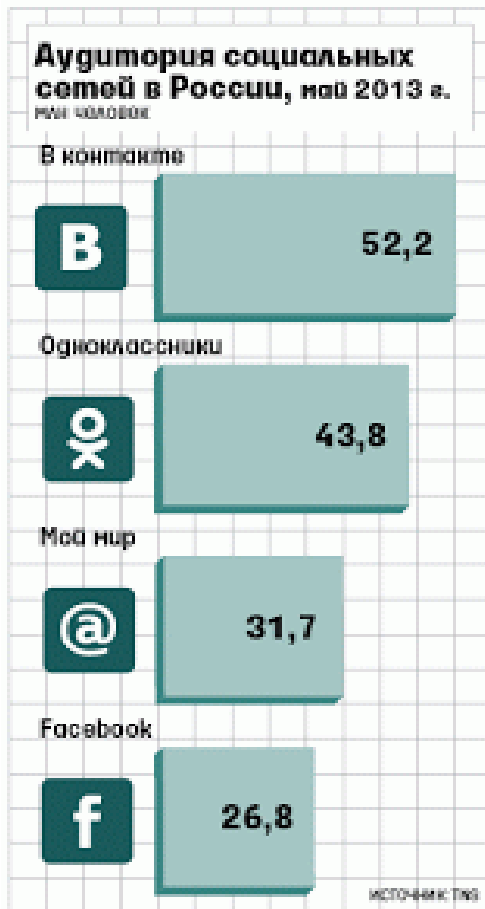


СОЦИАЛЬНЫЕ СЕТИ

Использование данных профилей в оценке кредитных рисков микрофинансовых организаций

*XII Национальная конференция по микрофинансированию и финансовой доступности
24.10.2013 г.*

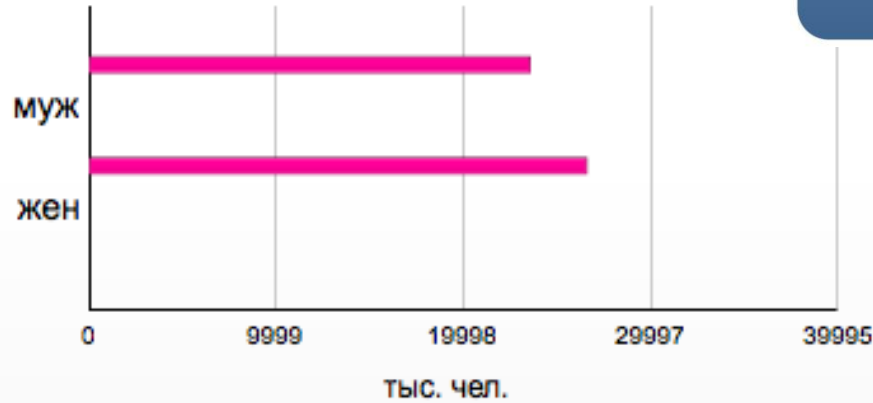
СОЦИАЛЬНЫЕ СЕТИ – НОВЫЙ ИСТОЧНИК ДАННЫХ



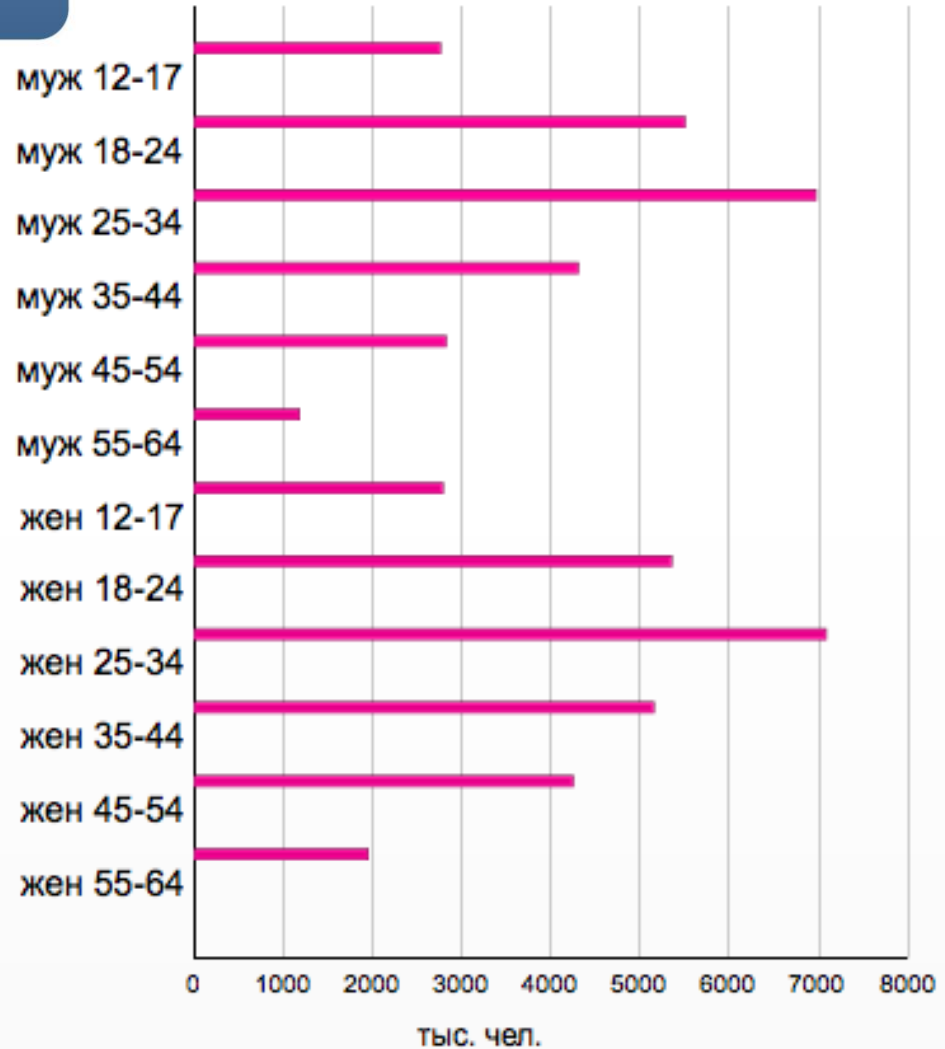
- ◆ Суммарная аудитория ВКонтакте, Одноклассники и Facebook более 50 млн. человек
- ◆ 95% российских пользователей Facebook имеет профиль хотя бы в одной местной социальной сети
- ◆ Количество пользователей, одновременно использующих ВКонтакте и Одноклассники, больше половины аудитории каждой из этих сетей в отдельности
- ◆ Сравним с размером населения страны в 145 млн. человек при уровне проникновения интернета <50%



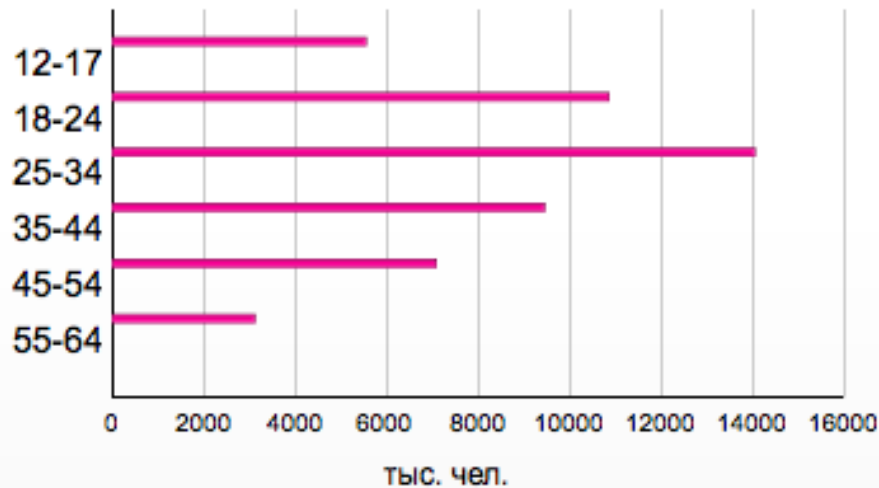
Пол



Пол/Возраст

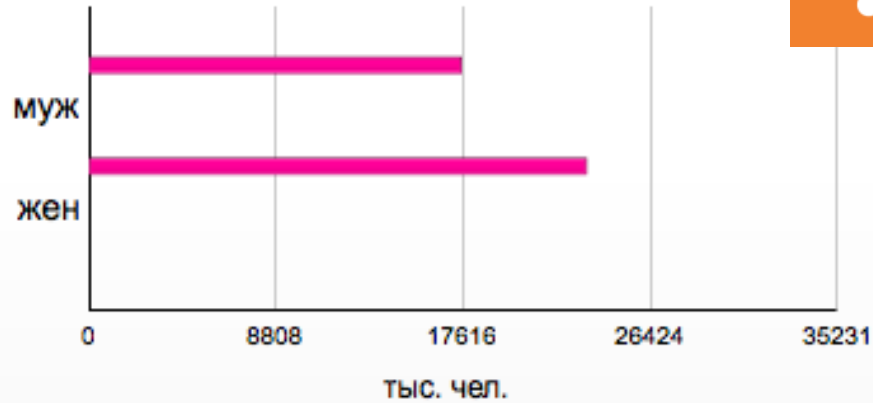


Возраст

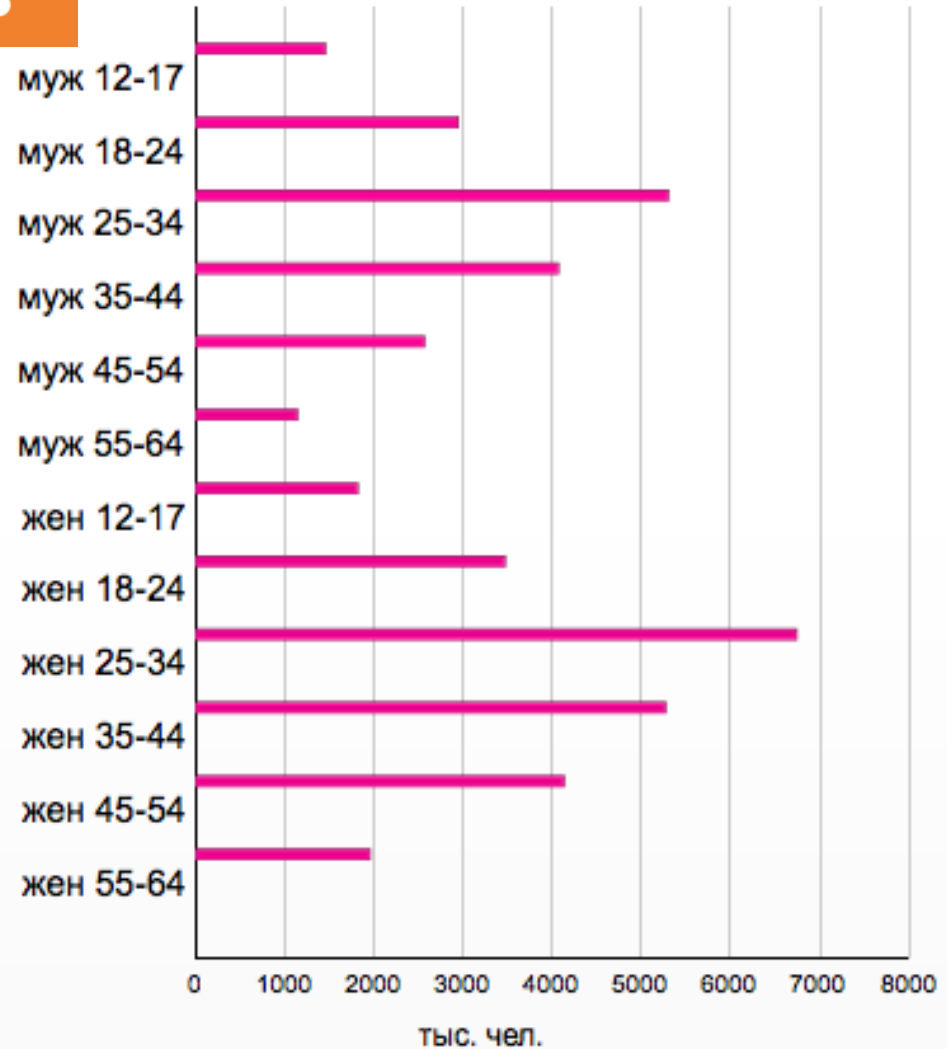




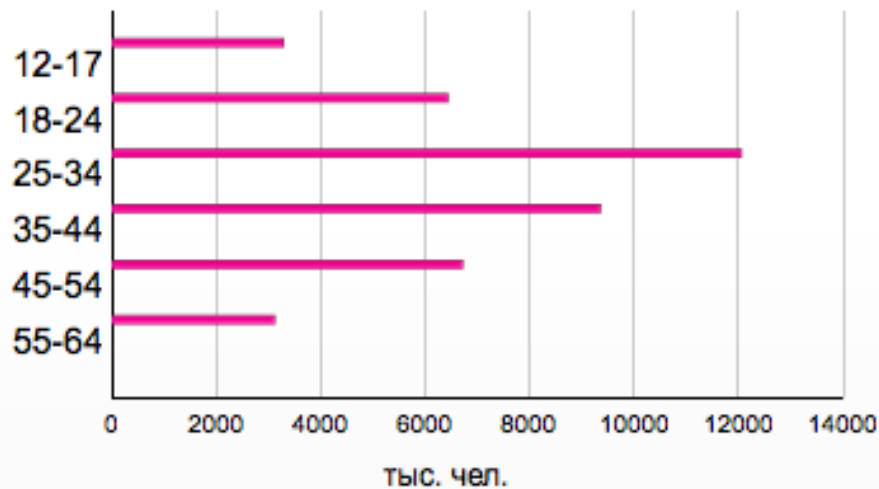
Пол



Пол/Возраст



Возраст



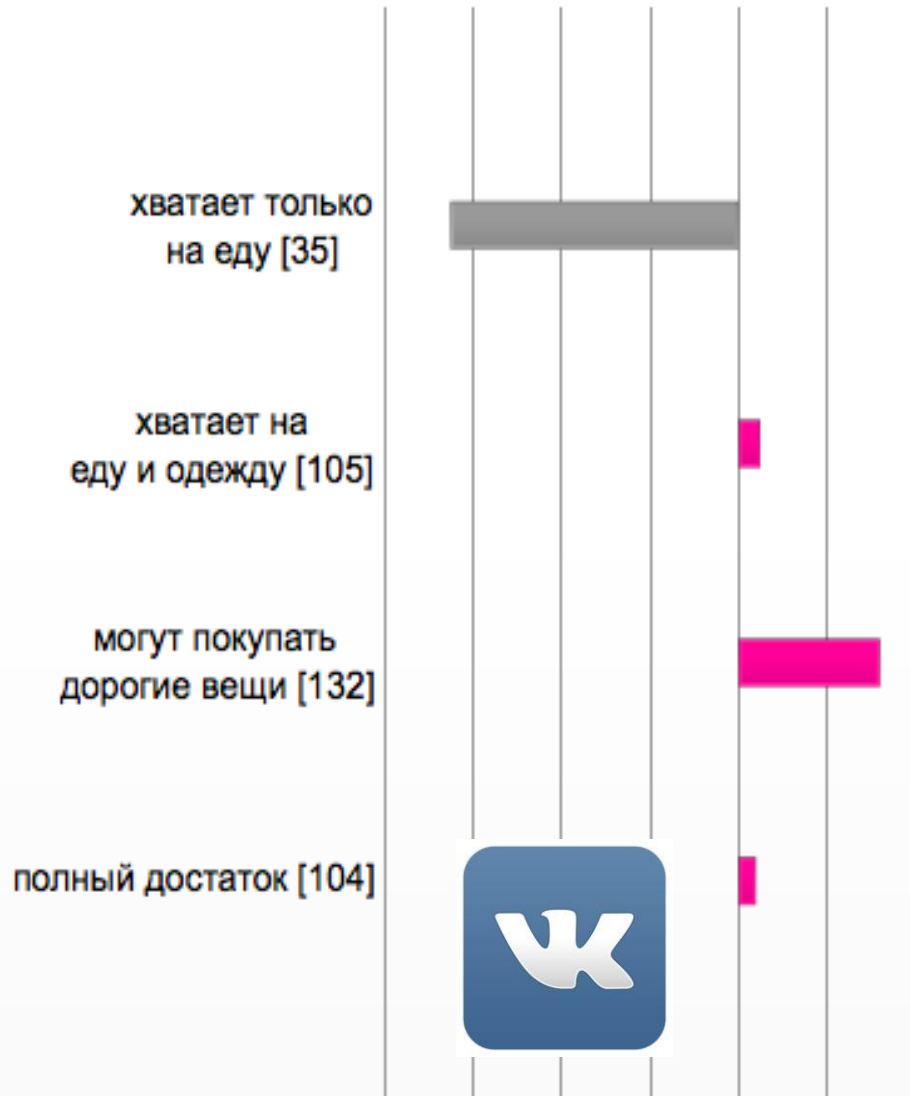
По отношению к населению России 0+ (12-64)



По отношению к населению России 0+ (12-64)



По отношению к населению России 0+ (12-64)



По отношению к населению России 0+ (12-64)



ОБРАБОТКА ДАННЫХ – НАИВНЫЙ ПОДХОД

- ◆ Анализ текстов на частотность по базам стоп-слов
Посты и комментарии сканируются на болезни, наркотики, алкоголизм, криминальный сленг, налоговые и судебные претензии.
- ◆ Анализ социального графа
Количество и «качество» друзей – их интересы, места учебы и работы, оценка социального статуса.
- ◆ Интересы
Сообщества и группы заемщика, тематика постов, тематика «лайков».
- ◆ Отслеживание перемещений между городами / государствами
Анализ гео-координат фотографий, постов, «чекинов» за последние полгода, год, два года.

ОБРАБОТКА ДАННЫХ – BIG DATA SCORING

- ◆ Риск-вектор по профилю содержит тысячи переменных
 - Информация из социальных сетей объединяется с информацией из других источников (поисковые запросы, платежные системы, и др)
 - Данные нормализуются, «оцифровываются»
 - Выделяются ключевые компоненты (основные риск-факторы), в результате формируются гипотезы
 - Data-Scientist строит аналитические модели (классификаторы), проверяет корреляцию нового риск-фактора с платежной дисциплиной, по результатам гипотеза и модель сохраняются или отбрасываются

- ◆ Учитывается платежная история родственников и друзей

- ◆ Необходима аналитическая IT-инфраструктура и экспертиза с
 - SNA, Social Networks Analysis
 - NLP, Natural Language Processing
 - Data Mining, Machine Learning

ДВА ПОДХОДА К ПОЛУЧЕНИЮ ДАННЫХ

СБОР ПУБЛИЧНОЙ ИНФОРМАЦИИ

- Забирается только профиль и «стена»
- Согласие пользователя не нужно, можно собирать данные хоть по всей аудитории социальной сети
- База данных с выгрузками наполняется быстро
- Пользователя нужно искать в полученном индексе по ФИО, полу, возрасту, городу, и т.д. Есть понятие hit-rate << 100%
- Информация публична, 152-ФЗ не применяется

ДОСТУП ПО ПРОТОКОЛУ OAuth через API

- Забирается вся доступная информация (на порядок больше публичной части)
- Нужно явное согласие пользователя на доступ к данным через кнопки соцсетей в онлайн-анкете заемщика
- База данных с выгрузками наполняется медленно
- Если доступ по OAuth получен, то принадлежность профиля уже установлена, hit-rate = 100%
- Выгрузки содержат приватную информацию, 152-ФЗ игнорировать нельзя

ОСНОВНЫЕ ИГРОКИ

СБОР ПУБЛИЧНОЙ
ИНФОРМАЦИИ



ДОСТУП ПО ПРОТОКОЛУ
OAuth через API



Максим Алексеев | SCORR.RU CEO

+7 (916) 501 51 78
malexejev@scorr.ru