Большие данные и машинное обучение в HeadHunter: умный поиск, рекомендации, прогнозирование

Александр Сидоров

У каждого свои большие данные

- Дисковые полки в 2002
- Медицинские экспертные системы
- Антивирусы
- Веб-антивирусы
- Веб-поиск и признаки по поведению пользователей
- Classified

Big data – не только про объём данных

Автоматизированная модерация резюме

Задача

- Проверять качество заполнения резюме на самом верху воронки подбора

Проблема

- Значительная часть соискателей плохо и неполно заполняют резюме
- Модерировать вручную – дорого

Решение

- Продолжать модерировать вручную
- Машинное обучение

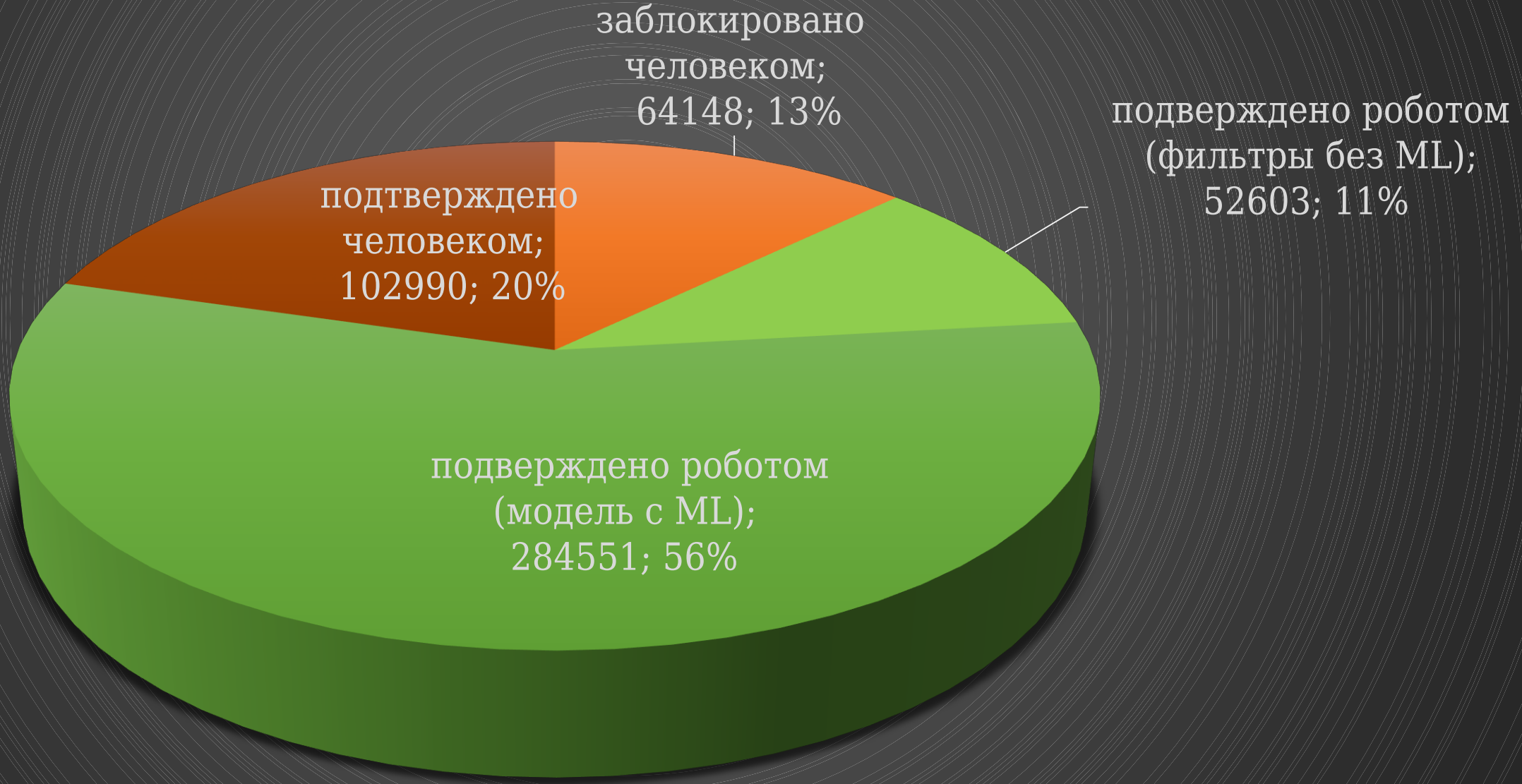
Схема работы автомодерации

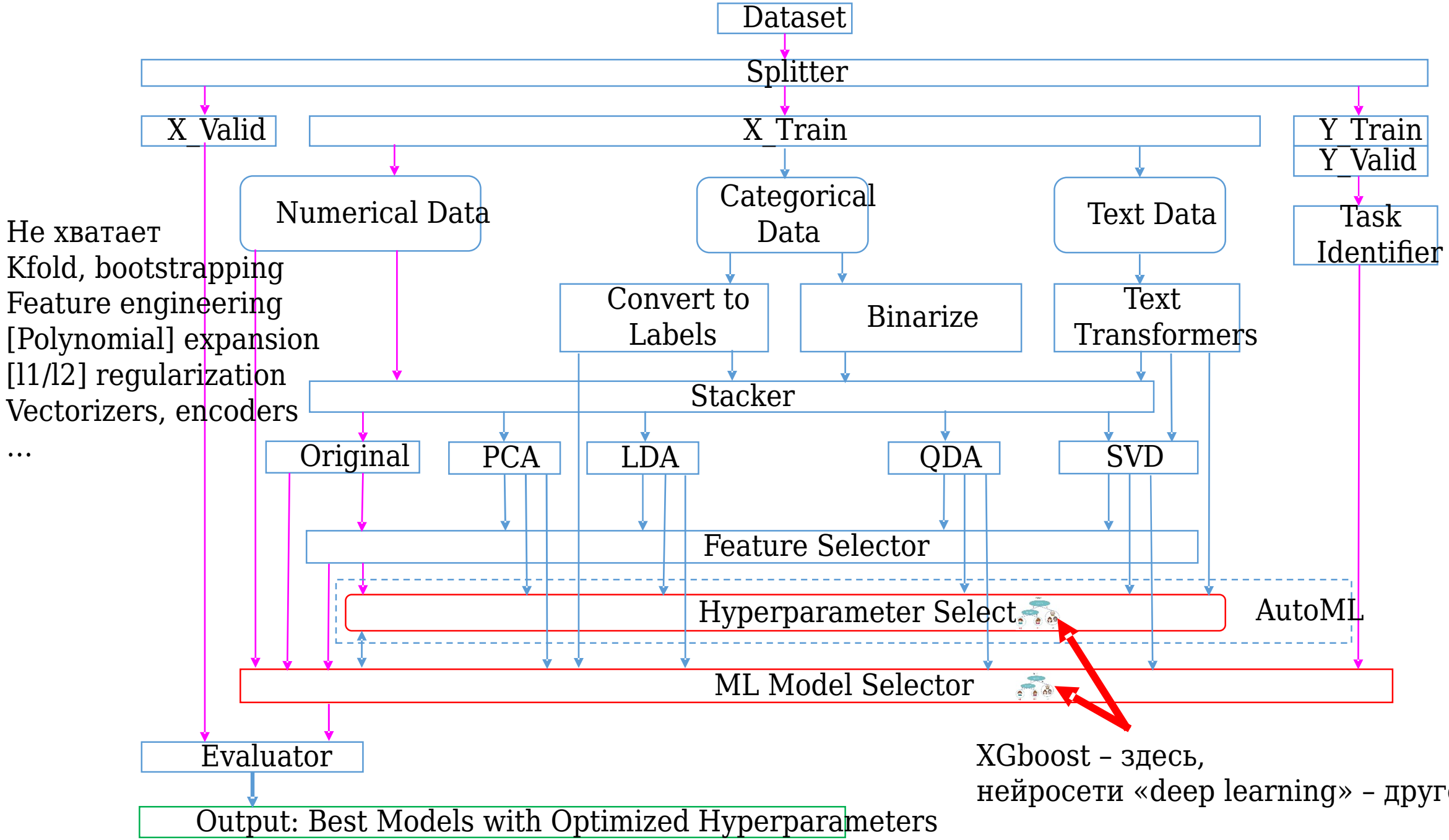
2016: 20k новых резюме в день, 20 модераторов

1. Результаты ручной модерации - исторические данные, 200k рез
2. Научили модель автомодерации, 1k признаков, auc_roc 0.94
3. Новые резюме проверяются автомодератором
4. Качественные резюме автоматически подтверждаются
5. Спорные случаи отправляются людям

2017: >30k новых резюме в день, ↑, 12 модераторов и качество мо
const

Экономия времени модераторов-людей





Не хватает
Kfold, bootstrapping
Feature engineering
[Polynomial] expansion
[l1/l2] regularization
Vectorizers, encoders
…

XGboost – здесь,
нейросети «deep learning» – другой вариа

http://blog.kaggle.com/2016/07/21/approaching-almost-any-machine-learning-problem-abhishel

Кейс 2: Прогнозирование времени работы
KPI, риска ухода для конкретной компани

Задача

- Снизить текучесть персонала, сэкономить ФОТ
- Увеличить производительность и выручку

Решение

- Больше работы HR’ов и руководителей
- Машинное обучение

Функционал

- Дописываем к резюме кандидатов прогнозы по времени работы и KPI

Схема прогнозирования времени работы и

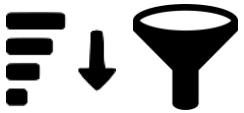
26 000 000 резюме

Отбор по проф. области
и дополнительным критериям

3-5 признаков

Модель, разделяющая по
времени и KPI

Ранжирование по
вероятности,
времени, KPI



≈1000 признаков

10-200k резюме в обучающих выборках

Результаты предпилотных проектов

Компания	Банк 1	Банк 2	Банк 3	FMCG 1
Процент новых сотрудников, которые…				
…работают более 2 года **сейчас**	36,5%	34,0%	40,3%	30,5%
…будут работать более 2 лет **с нашим продуктом**	48,6%	68,2%	75,9%	70,6%
Уменьшение текучести	1.33 раза	2,05 раза	1,88 раз	2,31 раза
Экономия ФОТ на текучести	2,5%	5,1%	4,68%	5,7%

- Сейчас технология рекрутмента у клиентов такая, что они нанимают кандидатов, из которых только **1/3** р
- Мы отбираем и предоставляем резюме кандидатов, из которых до **2/3** проработают больше 2 лет

Кейс 3: Поиск, рекомендации вакансий, реклама вакансий

Задача

- Подобрать по резюме вакансии, интересные соискателю

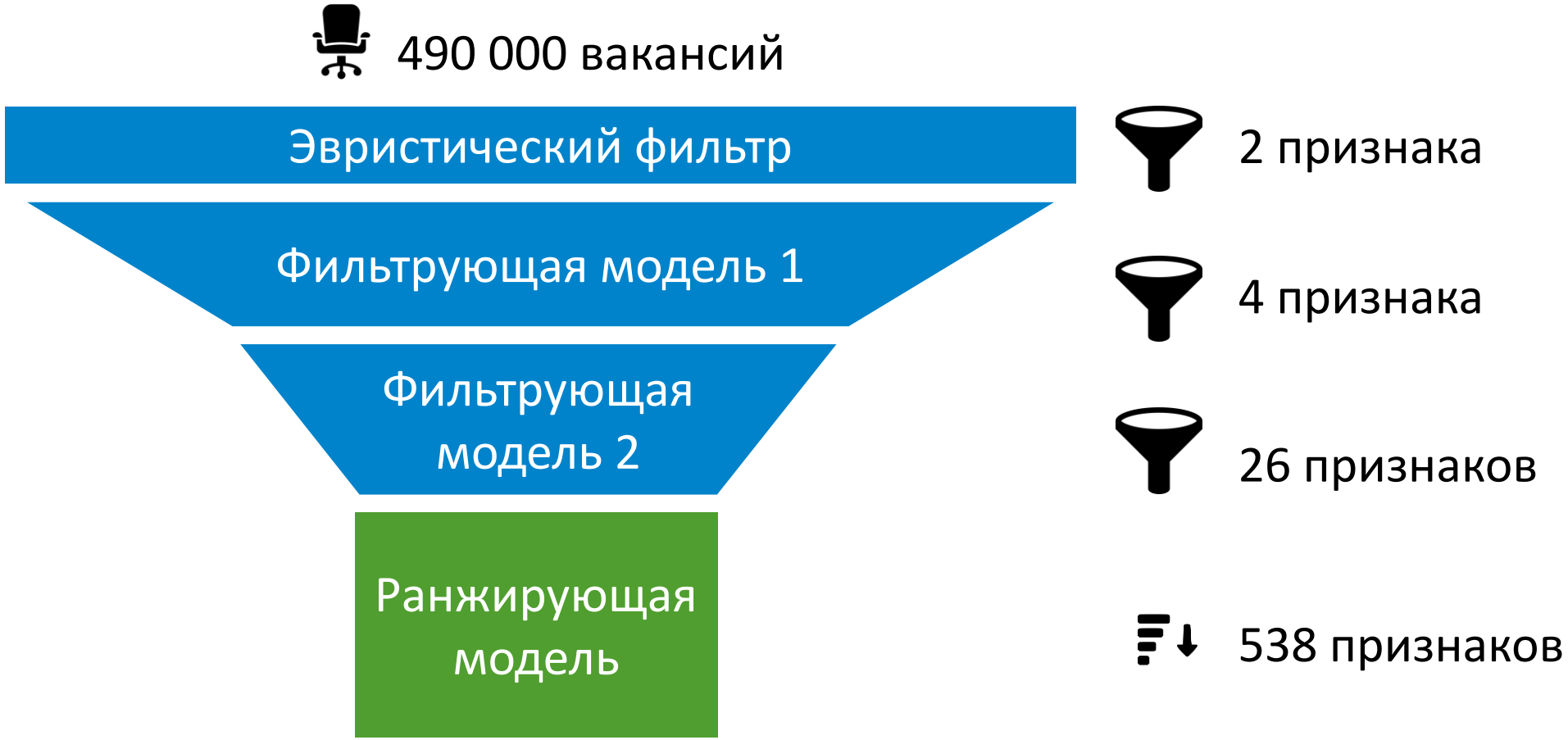
Решение

- Фильтр по параметрам вакансии, ручной таргетинг
- Машинное обучение

Функционал

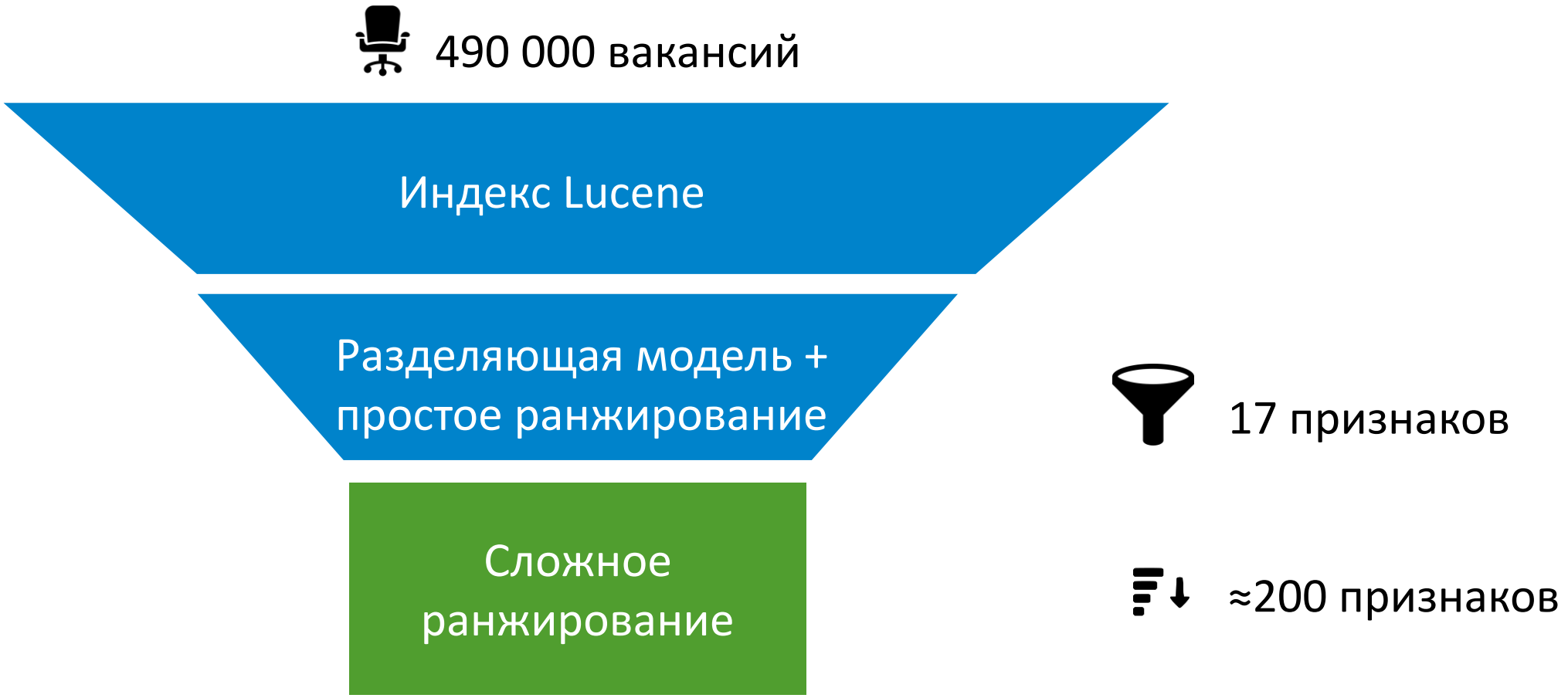
- Главная страница, рекомендуемые вакансии в списке резюме, рассылки с подходящими вакансиями

Схема работы рекомендательной системы



44-500m пар в обучающих множествах, переподбор за ночь

Схема работы поиска



44-500m пар в обучающих множествах, переподбор за ночь

С ML и Big Data связана только часть проек

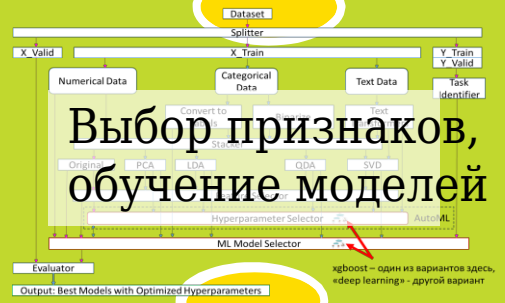
Постановка задачи,
выбор приорите
процессы, команда

Сбор
исторических
данных

Мониторинг
ресурсоёмкости,
контролируемое
поведение при
превышении нагрузки

Оптимизация
использования
ресурсов

Прототипирова



Выбор признаков,
обучение моделей

Автоматизированное
тестирование

Работа с
экспериментальным
кодом

Регулярный пересчё
статических признаков,
повторное обучение
моделей

Расчёт признаков
и работа моделей
во время обработки
поисковых запросов

A/B-тесты,
соблюдение баланса
«польза/ресурсы/
трудозатраты
на оптимизацию»

Синхронное, без останов
обновление
на серверах
в production


соискатель
Frontend, nginx
Интеграционный слой, Java, Python
id соискателя
резюме
API для работы с БД, Java
БД резюме и вакансий
PostgreSQL
резюме
признаки для
резюме
Сервис извлечения статических
признаков, Python
ML pipeline, Python
Расчёт признаков (>200)
Кеш признаков,
PostgreSQL
Индексатор,
Java
признаки
для резюме,
условия поиска
Поиск, Java
Индекс с признаками для вакансий,
Lucene
Расчёт pairwise features, sconfig
Фильтрующие модели, logit
Ранжирующая модель, XGBoost
список
найденных
вакансий
Подбор
моделей,
измерение
offline-
метрик
Выгрузка и
чистка
данных
Хранилище логов
Hadoop, 100 млн. записей
день
A/B-тесты, Python
Tableau
Расчёт online-метрик

Frontend (обычный и мобильный сайт, также API, через который работают мобильные приложени
20 instances

Интеграционный слой: 15 instances

Backend остальных сервисов: 30 instances

Backend поиска и рекомендаций по вакансиям: 30+ instances

Meta (балансировщик, индексатор) 3 instances

Master (обновляет сегменты индексов) 2 instances

Basesearch (Lucene, расчёт pairwise features, модели) 20+ instances

Kardinal (извлечение статических признаков) 7 instances

Data storage: ≈50 instances

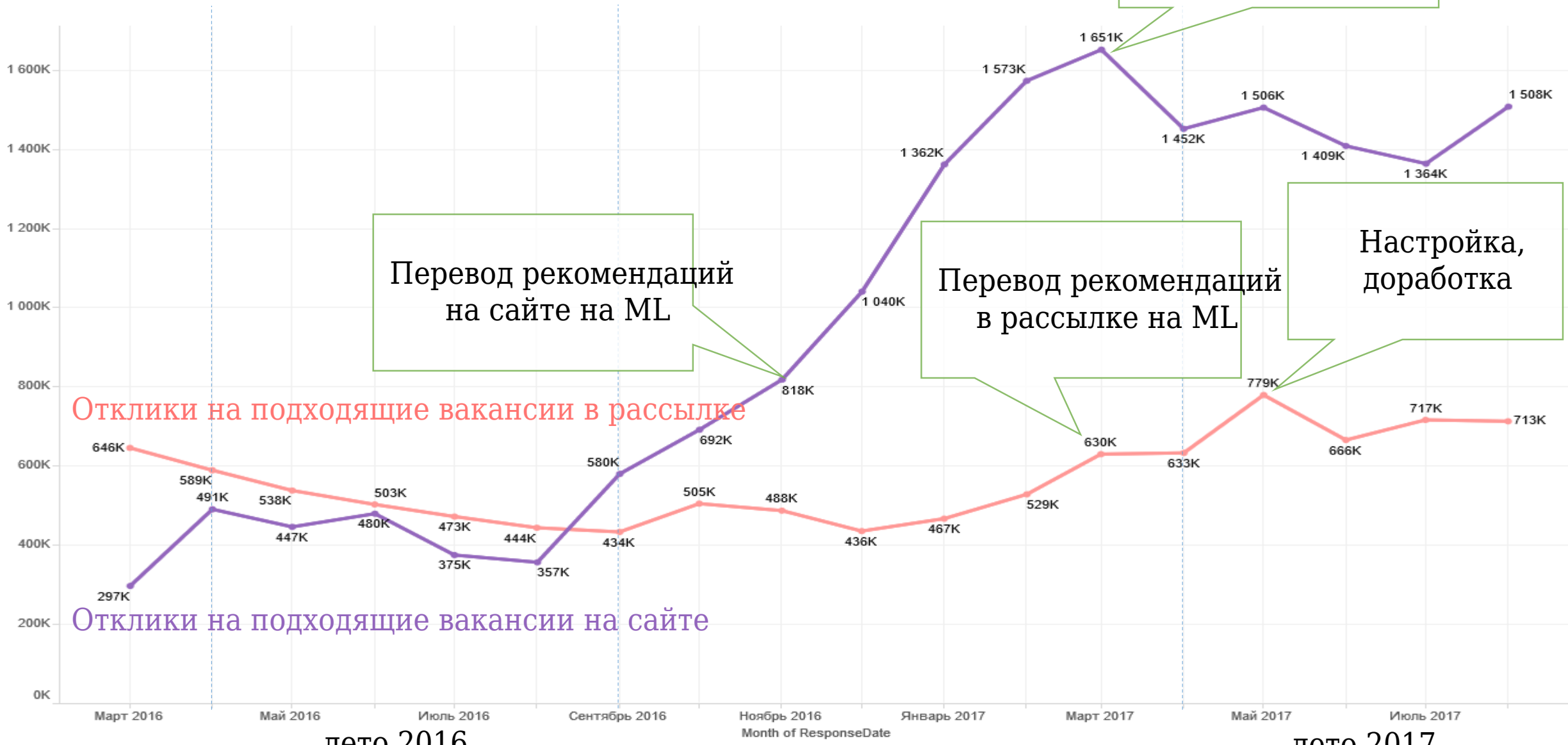
PostgreSQL 7 servers

Hadoop cluster 30+ servers

Cassandra 7 servers

Instance это docker container, vm или сер

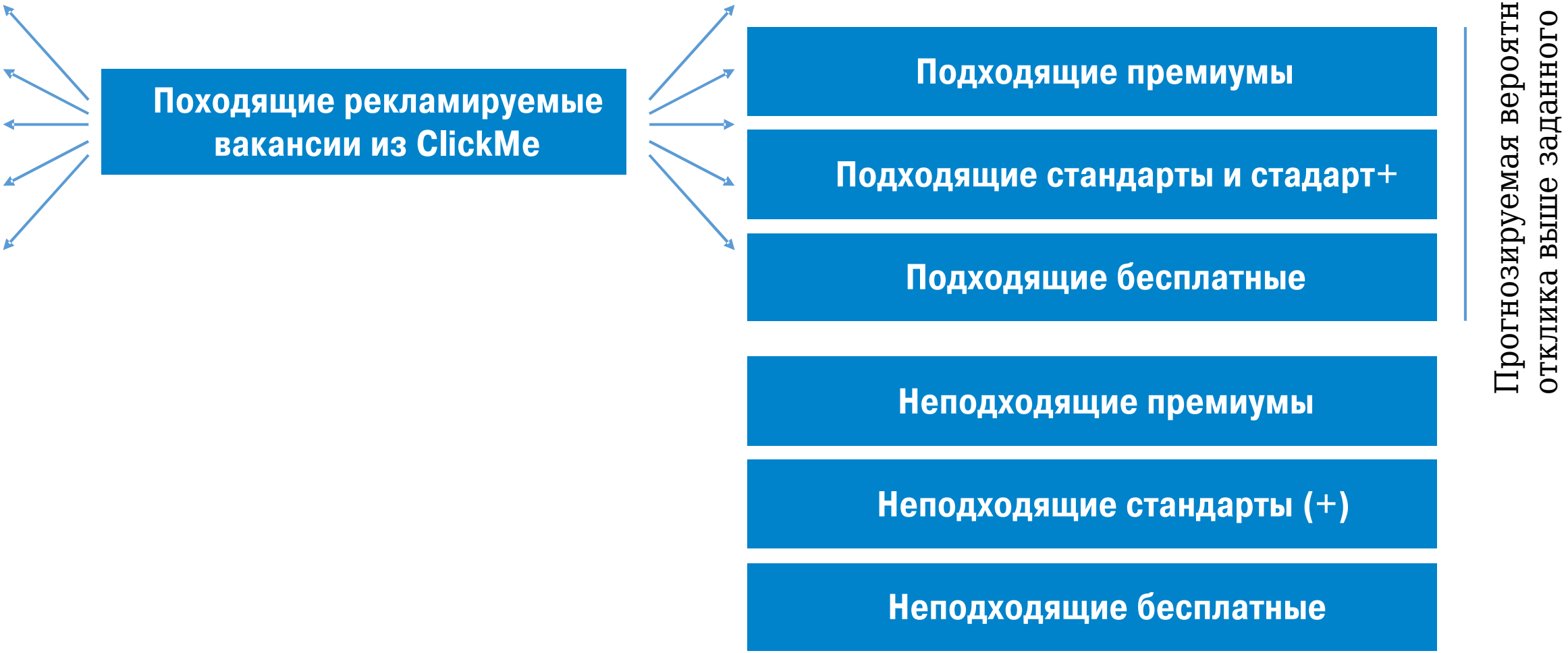
Отклики на вакансии



лето 2016

лето 2017

Умный поиск 1.0 + умная реклама вакансий в С



Воронка найма при обычном поиске, **было**

Воронка найма при умном поиске, **становит**

Воронка при умном поиске и clickme, скоро

Big data – не только про объём данных

Большое спасибо!

Вопросы можно сейчас или потом, по почте