

# Исследование открытых данных ФНС<sup>1</sup> России для прогнозирования финансового неблагополучия МСП<sup>2</sup>

- 1. ФНС – федеральная налоговая служба
- 2. МСП – малые и средние предприятия

Рекомендации бизнесу на основе изучения  
финансовых и организационных признаков МСП

ФИО

## Цели проекта:

- собрать данные;
- изучить данные и обучить модель для предсказания;
- дать рекомендации;

Здесь может быть картинка

Нестабильность экономических условий в первую очередь сказывается на малом и среднем бизнесе<sup>1</sup>. Узкий горизонт планирования затрудняет оценку инвестиционной привлекательности и возможности заключения контрактов и договоров с потенциальными подрядчиками или поставщиками.

Поэтому важно собирать и оценивать всю возможную информацию о второй стороне.

И именно здесь – в поиске и работе с информацией - полезен глубокий анализ данных и машинное обучение для выявления неявных закономерностей.

1. Например, статья на эту тему,  
Или ещё статья

## Задачи на пути к целям:

- Какие данные доступны? – Изучить открытые источники.
- В каком виде они доступны? – Изучить структуру данных, собрать, очистить и соединить.
- О чём говорят данные? – Провести исследовательский анализ данных, выбрать цель предсказания, создать новые признаки.
- Что неочевидно в данных? – Обучить модель и изучить вклад признаков
- Что можно получить из всего этого? – Сформулировать рекомендации.

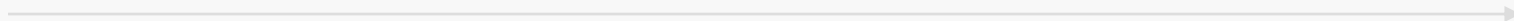
“Куда-нибудь ты обязательно попадешь, — сказал Кот. — Нужно только достаточно долго идти”  
Л. Кэррол

У таймлайна 2 центра тяжести (пока один!):

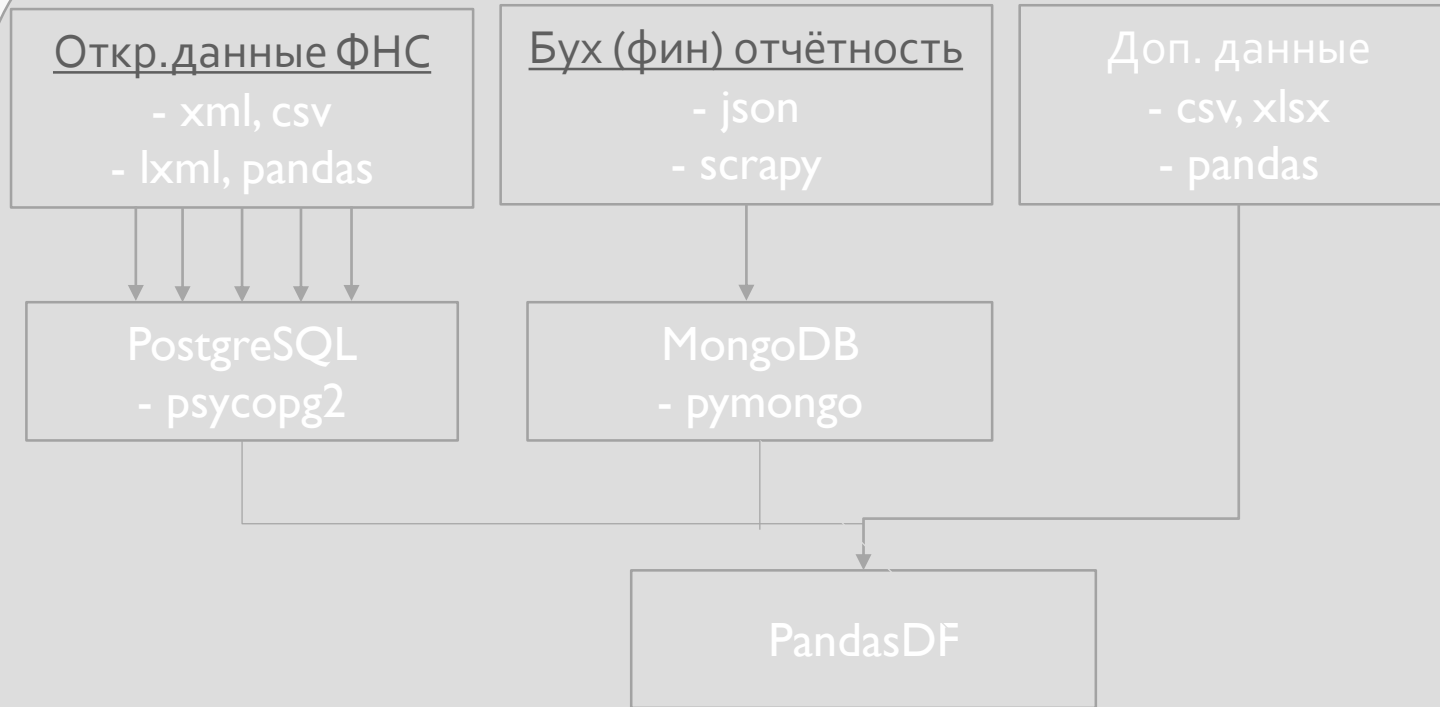
- сбор разнотипных данных из нескольких источников
- понимание данных и их вклада в предсказание модели

(если хватит времени)

- Здесь будет таймлайн с примерными этапами и сроками
- 1 нед поиск источников -> 4 нед сбор и загрузка данных -> ...



# Структура проекта



# 1. Получение данных

«Как закалялась сталь»  
Н. Островский

И здесь может быть картинка

Источник вдохновения – датасет  
Company Bankruptcy Prediction

## Основные источники данных:

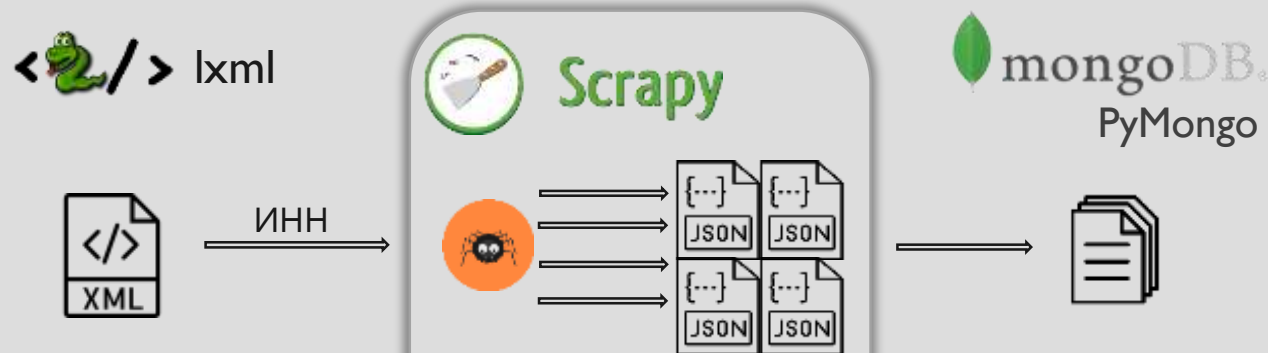
1. <https://bo.nalog.ru/> : государственный информационный ресурс бухгалтерской (финансовой) отчётности.
2. <https://www.nalog.gov.ru/opendata/>: открытые государственные данные ФНС России

Дополнительно: [https://rosstat.gov.ru/vpn\\_popul](https://rosstat.gov.ru/vpn_popul) - итоги Всероссийской переписи населения

# Особенности извлечения данных

«Не спи, не спи, работай,  
Не прерывай труда»  
Б.Пастернак

1. Источник: <https://bo.nalog.ru/>, json



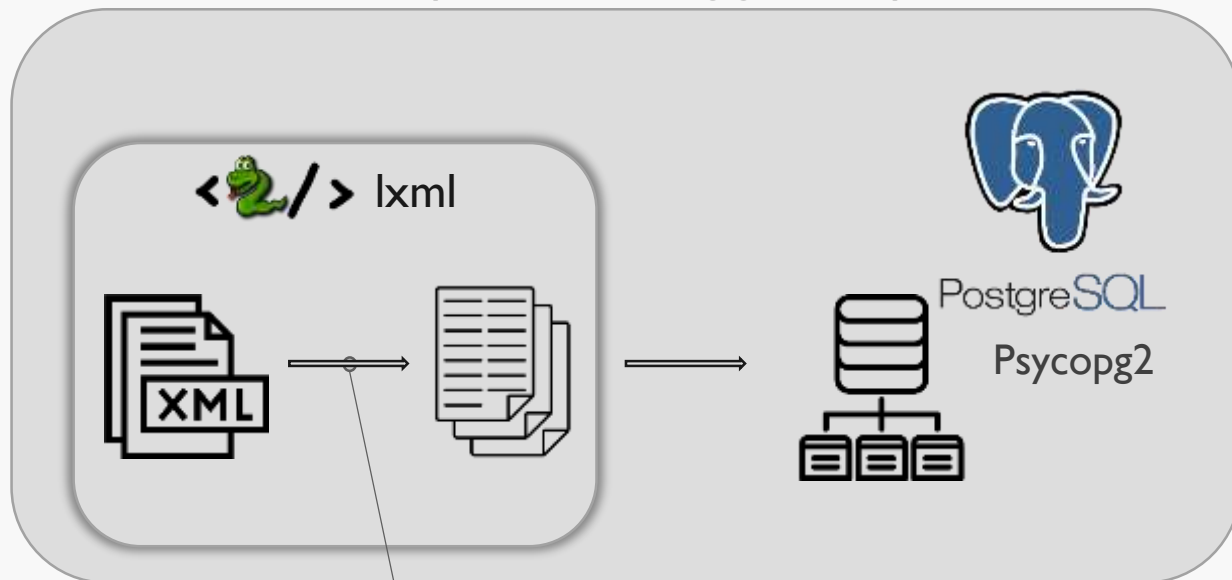
- Страница для каждой организации однозначно определяется номером ИНН. Список ИНН можно получить из реестра МСП ( подробности в следующем слайде).
- Контент формируется динамически: чтобы собрать необходимый минимум информации по одному ИНН – требуется 4 запроса.
- Это наиболее продолжительный этап проекта.

Немного деталей:

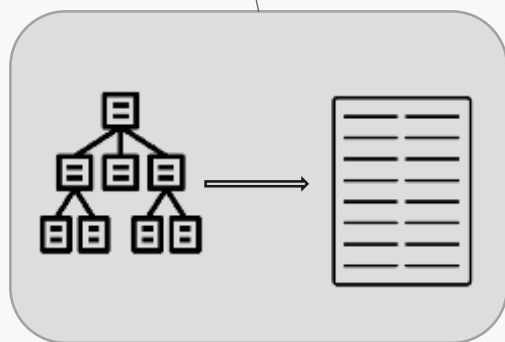
- Всего в реестре около 2.3 млн. уникальных ИНН.
- В итоге получается около 9.2 млн. запросов (на самом деле меньше, так как не для всех ИНН есть отчёты в открытом доступе).
- Паук обрабатывает 50 тыс. ИНН чуть менее чем за 14 часов и получает из них около 40 тыс. документов для mongoDB.
- Можно быстрее, если использовать платный API.
- Повторный проход для обновления или дополнения информации будет быстрее: в базе сохраняются необходимые для запросов id документов.
- Неплоские документы по одному за раз удобно записывать в NoSQL базу данных.

## Особенности извлечения данных

2. Источник - <https://www.nalog.gov.ru/opendata/>, xml



- Данные представляют собой архивы с множеством файлов xml.
- Данные имеют достаточно сложную вложенную структуру.
- Архивы можно сохранить локально, т.е. можно задать реляционную структуру и изменять её, если это необходимо.
- При пофайловой обработке удобно записывать данные пакетами.
- Можно установить связи между таблицами.



Немного деталей:

- Всего получается 4 таблицы:
  - reestr\_msp: данные из единого реестра субъектов МСП
  - support\_msp: данные о полученной субъектами поддержке
  - debt: сведения о налоговых правонарушениях
  - tax\_offense: Сведения о суммах недоимки и задолженности по пеням и штрафам
- Самый большой архив в распакованном состоянии занимает около 20 Гб и включает около 7 тыс. xml файлов.
- Из одного файла можно забрать от 0 до чуть менее 1000 строк.
- Цикл обработки занимает 20-25 минут, конечная таблица занимает около 800 Мб.

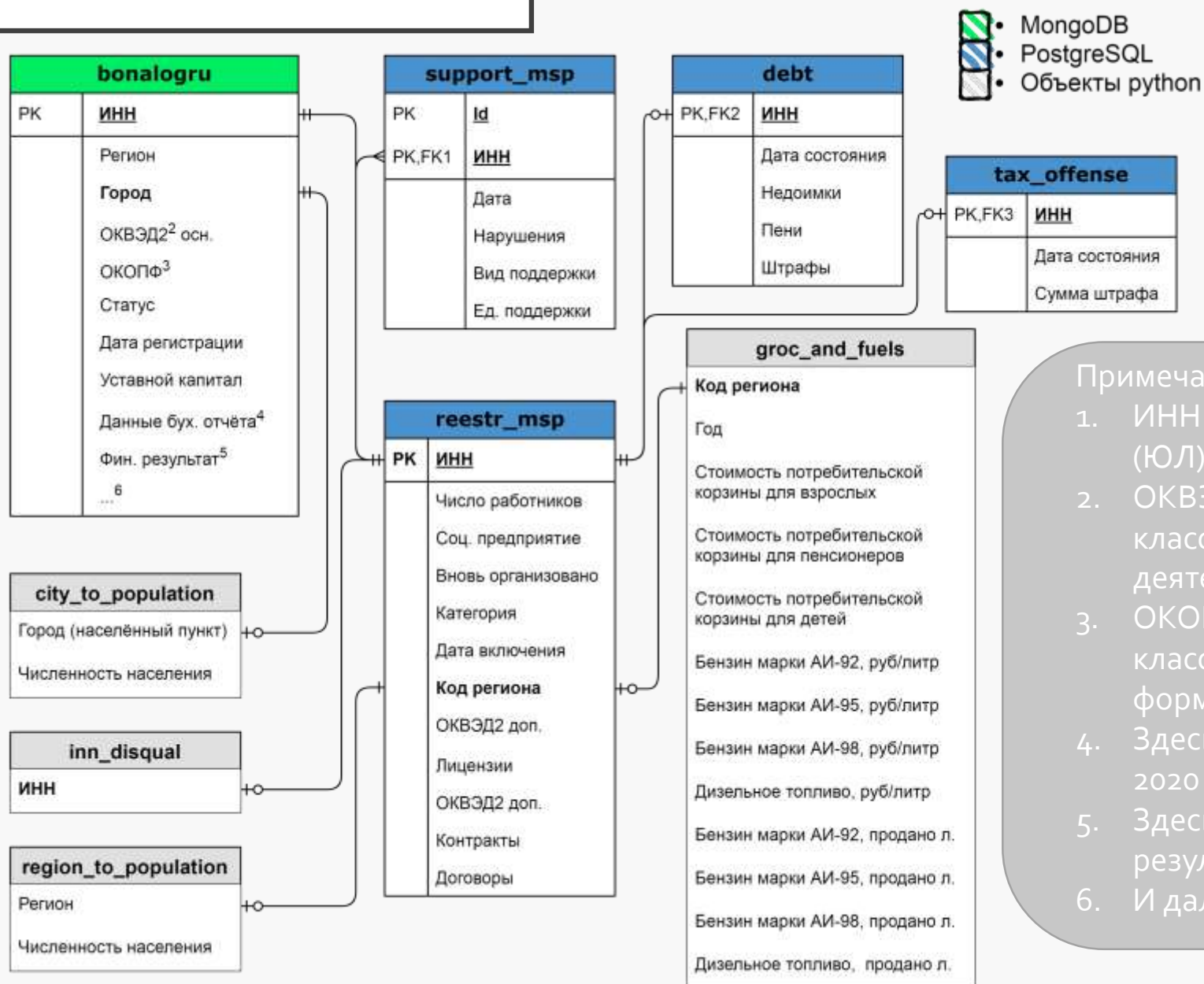
## Особенности извлечения данных

### 3. Дополнительные источники

- <https://www.nalog.gov.ru/opendata/>, csv
- [https://rosstat.gov.ru/vpn\\_popul](https://rosstat.gov.ru/vpn_popul), xlsx



# Структура данных



Здесь м.б. картинка/надпись  
Или расшифровка названий таблиц

## Примечания:

- ИНН здесь и далее – ИНН юридических лиц (ЮЛ)
- ОКВЭД – код по общероссийскому классификатору видов экономической деятельности
- ОКОПФ – код по общероссийскому классификатору организационно-правовых форм собственности
- Здесь данные бухгалтерского отчёта за 2021, 2020 и 2019 гг.
- Здесь данные отчёта о финансовых результатах за 2021 и 2020 гг.
- И далее – в совокупности более 150 полей



