# Обработка корпоративной информации на естественном языке

при помощи онтологий и инструментов машинного обучения

**Сергей Горшков, Роман Шебалов, Константин Кондратьев**

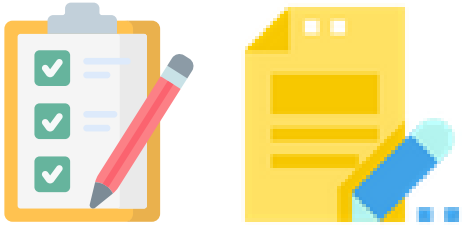
info@trinidata.ru
trinidata.ru

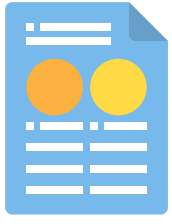
ТриниДата

АрхиГраф.Мир | АрхиГраф.MDM | АрхиГраф.СУЗ

УПРАВЛЕНИЕ ДАННЫМИ 2020

# О чем поговорим?





Проблемы, которым более 30 лет:

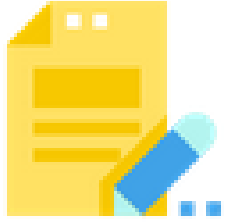
1. Корпоративные документы содержат больший объем информации, чем хранилища структурированных данных.

2. Главный инструмент работы с ними – полнотекстовый поиск «как в Интернете».

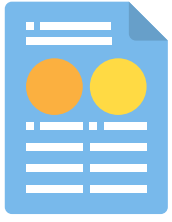
3. Большая часть информации в документах лежит «мертвым грузом»: накопленный опыт не используется, анализ и оптимизация деятельности не проводится, владение информацией не имеет коммерческого смысла.

# Что есть в документах? Как это использовать?

Результаты проверок, инспекций, освидетельствований
=> обобщить, чтобы избежать повторения замечаний



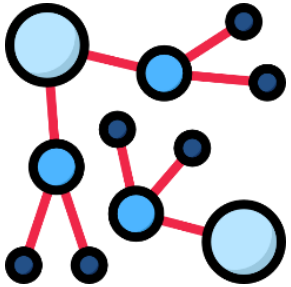
Описания ситуаций, конфликтов, разногласий
=> найти и устранить их причины



Характеристики объектов и событий
=> найти аналогичные объекты или события,
чтобы быстрее и эффективнее принимать решения

Все это – опыт, накопленный за время работы организации.
Чтобы его использовать, нужно из неструктурированной (текстовой) формы перевести его в структурированную.

# Как структурировать произвольную информацию?



Традиционно в ИТ-системах структурируют информацию с помощью таблиц. Но содержимое миллионов текстов нельзя превратить в таблицы.

В нашем сознании такая информация хранится с помощью сети концептов (идей).

Для ее машинно-читаемого представления используются онтологии, как способ представления концептов, и графовые СУБД как средство хранения.

Онтология:

1. Содержит структуру наших знаний (концепты) и конкретные факты, представленные с их помощью.
2. Связывает понятийные структуры (концепты) с лексическими.
3. Содержит фактическую информацию, к которой можно делать структурированные запросы.

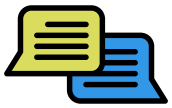
TBox – модель предметной области

Лексическая модель

ABox – факты об объектах и событиях

УПРАВЛЕНИЕ ДАННЫМИ 2020

# Как работает преобразование текста в факты?




заводов

завод

… технология внедрена на заводе «Альфа», где …

2) Из документов извлекается чистый текст

3) Выполняется лемматизация

4) Извлекаются именованные сущности (NER, Named Entity Recognition), а также понятия предметной области

1) Источником являются тексты документов, сообщения, запросы пользователей

6) Сущности и понятия, выделенные в тексте, соотносятся с онтологией, в которой содержится описание предметной области, в т.ч. с использованием правил логического вывода

… технология внедрена на заводе, где …

5) Определяется структура фразы (POS tagging), каждое слово текста аннотируется в соответствии с лексической онтологией

7) Строится граф фактов, описывающих текст (то, о чем в нем говорится) и/или непосредственно изложенных в тексте

# От текста к фактам

«По итогам экспертизы выявлены нарушения требований ГОСТ в части следующих положений:

- использование технических материалов, не соответствующих требованиям по устойчивости к воздействию неблагоприятных факторов окружающей среды с учетом географического района РФ и климатической зоны расположения объекта проектирования (Ангарский район, Иркутская область, Россия);
- применение гидроизоляционных материалов марок Г-ИГ, С-РМ, что не соответствует назначению и планируемым условиям эксплуатации объекта проектирования;
- применение кровельных материалов марок С-РК, С-РЧ, Г-ИК, что не соответствует назначению и планируемым условиям эксплуатации объекта проектирования.»

(Экспертное заключение)

## Семантическая модель:


Результат: автоматическое создание фактов,
которые могут использоваться для проверки параметров новых проектов

# Разрешение смысловых неоднозначностей

"В настоящее время наблюдается разрушение элементов кровли, а именно разрывы, растрескивания и деформация кровельных листов (полотен), что вызывает коррозию несущих элементов конструкции кровли строений, проникновение влаги и осадков внутрь. Мы полагаем, что это стало следствием нарушений, допущенных специалистами «ООО Гамма» при подготовке проектной и рабочей документации. Просим подготовить мотивированное обоснование использования кровельных и изоляционных материалов указанных в документации марок. "

(Обращение организации-заказчика)

Смысл слова «Лист»:
строительный материал

"Наши специалисты приступили к проверке указанных вами фактов. Со своей стороны хотим напомнить, что работы по проекту были проведены в полном соответствии с действующими нормативами, сооружения были сданы в эксплуатацию без замечаний и приняты госэкспертизой. Что подтверждается соответствующими документами.

Прилагаем документацию по приемке на 40 листах.

(Ответ организации-подрядчика)

Смысл слова «Лист»:
лист бумаги/документа

# Разрешение смысловых неоднозначностей

Обычный поиск по тексту не поможет различить смыслы
Пользователь утонет в поисковой выдаче


Лексическое вхождение: «Лист»

Лексический смысл: строительный материал

Лексическое поле: полотно, палета, моток

Лексическое вхождение: «Лист»

Лексический смысл: лист документа

Лексическое поле: бумага, папка, печать

Результат: автоматизированное отсеивание контекстов,
в которых используются слова с ненужными пользователю смыслами

# Разрешение смысловых неоднозначностей

"В период с 01.02.2021 по 26.02.2021 были проведены мероприятия по устранению замечаний заказчика «ООО Альфа» по проекту «Титул 2.76.2021 Комплекс ангаров». Замечания были устранены. 26.02.2021 было проведено совещание с представителями «ООО Альфа». Дополнительных замечаний от «ООО Альфа» не поступало.

(Отчет об устранении недостатков по проекту)

Смысл слова «Ангар»:
сооружение

"Трубопроводное оборудование станции аэрации подвержено усиленной коррозии. Просим провести консультацию для организации проектных работ по реконструкции станции с применением изоляционных и кровельных материалов. Планируемая реконструкция является сложным инженерно-техническим мероприятием, особенно в связи с тем, что станция располагается в непосредственной близости от реки Ангара. "

(Обращение потенциального заказчика)

Смысл слова «Ангара»:
река

# Разрешение смысловых неоднозначностей

Как различить смыслы и избавиться от ненужной информации?

Лексическое вхождение: «Ангар»

Лексический смысл: сооружение

Лексическое поле: склад, хранилище, постройка, строение

Лексическое вхождение: «река Ангара»

Лексический смысл: -

Лексическое поле: -

Каталог объектов: Река

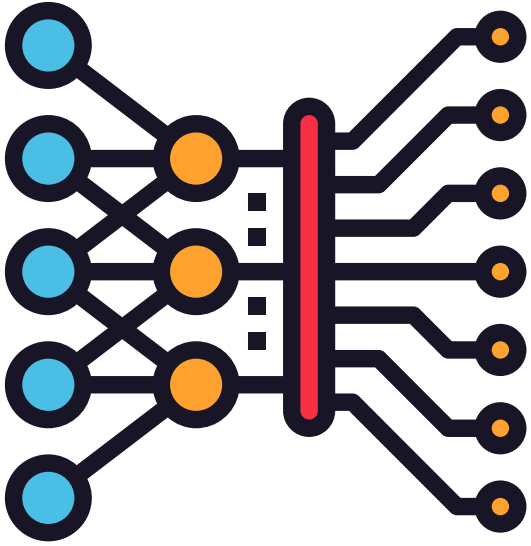
Экземпляр: река Ангара

# Как выбрать инструмент для решения задачи?

Логические правила

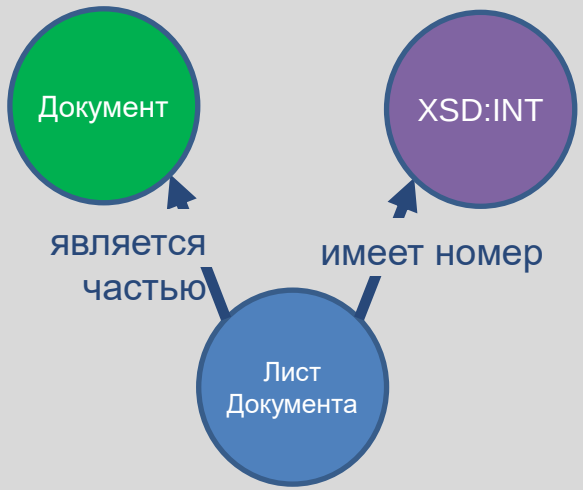
VS

Нейросети



# Как это работает?

## Основные идеи




Онтология предметной области

Ошибка на первом листе документа

Упоминания сущностей в тексте

```
документ_ххх    тип              Документ .
лист_ххх        тип              Лист ;
                имеетНомер       1 ;
                являетсяЧастью   документ_ххх .
```

Выделение объектов

# Как это работает?

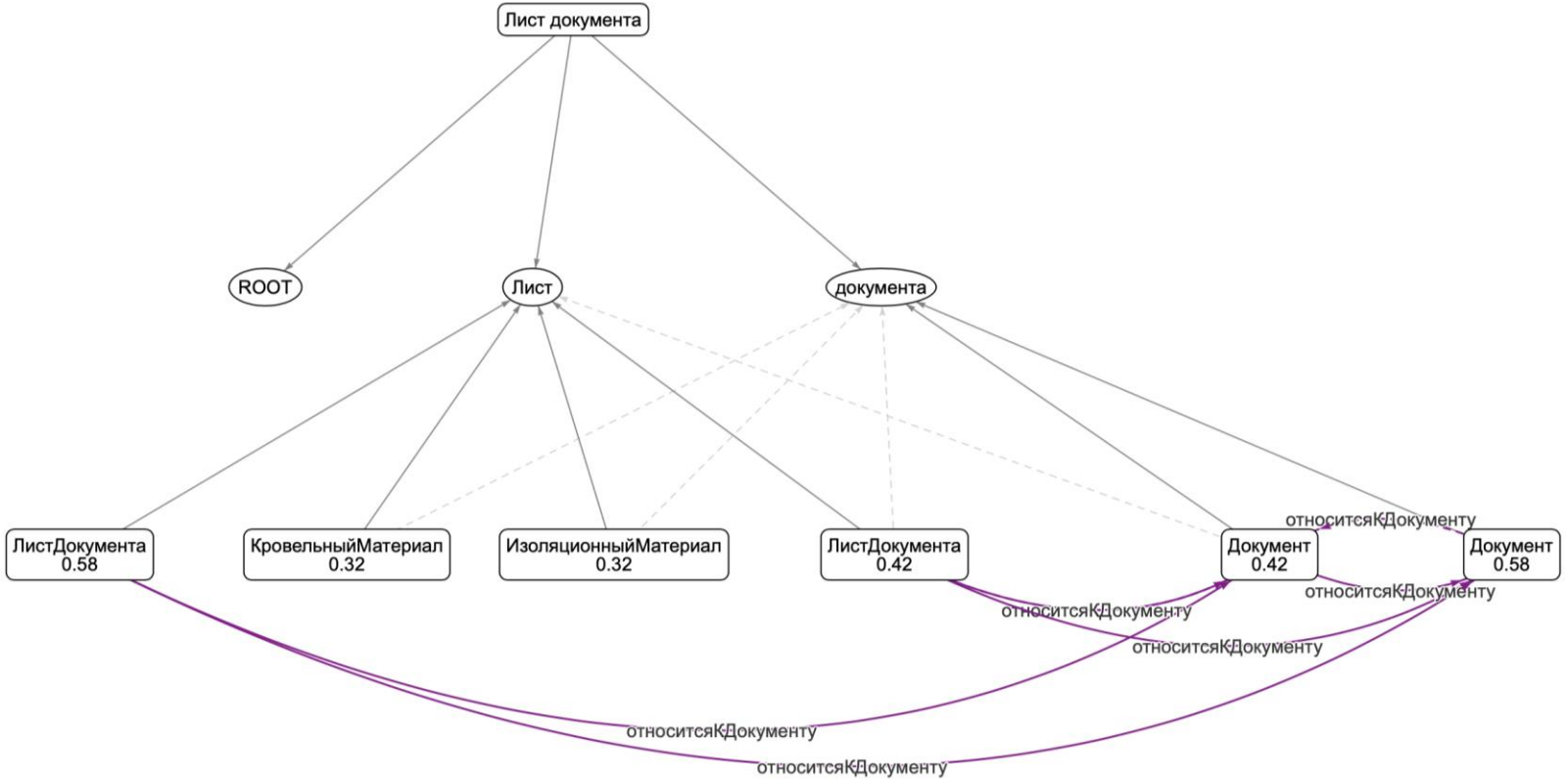
## Pipeline (шаги обработки текста):

- Разбить текст на предложения.
- Синтаксический анализ, лемматизация, POS-тэги.
- Выделение возможных Упоминаний сущностей онтологии в тексте.
  - Правила, словари, нейросетевые модели (NER).
  - Неоднозначность. Через LexicalSense, LexicalField.
  - Неоднозначность. С учетом других сущностей, найденных в тексте.
  - Термины, «разбитые» другими словами.
- Добавление предикатов (семантических связей) между найденными Сущностями.
  - По существующим связям в онтологии.
  - Неоднозначность. Через LexicalSense, LexicalField.
- Выделение объектов. Определение, относятся ли упоминания одинаковых Сущностей к одному и тому же Объекту.
- Составление SPARQL-запроса или записи о фактах.

# Как это работает?

## Выделение возможных Упоминаний сущностей онтологии в тексте.

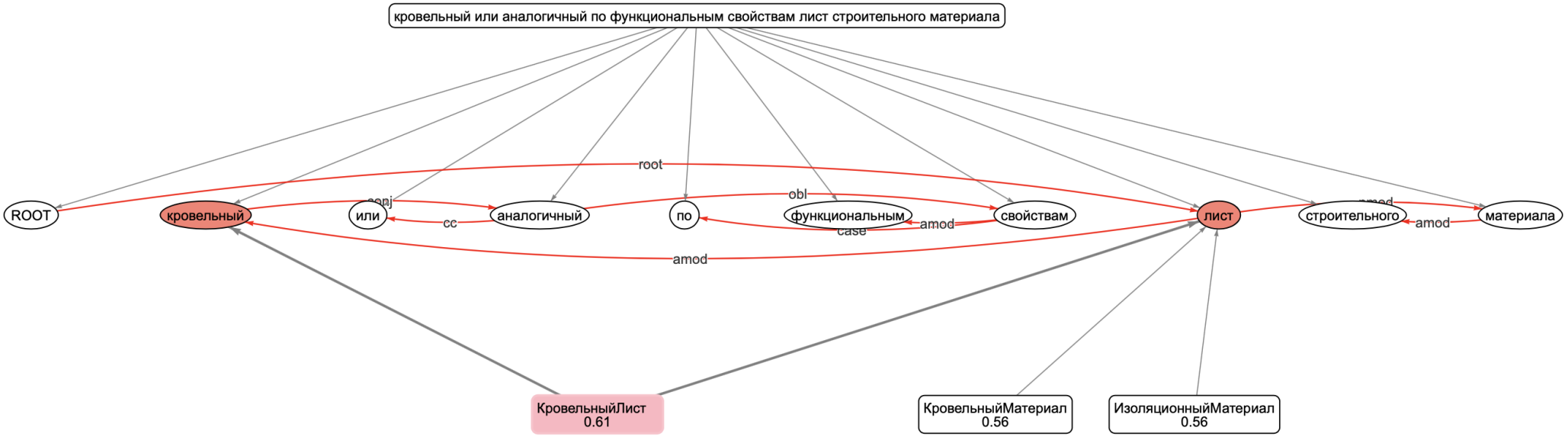
- Неоднозначность. С учетом других сущностей, найденных в тексте.



# Как это работает?

## Выделение возможных Упоминаний сущностей онтологии в тексте.

- Термины, «разбитые» другими словами.



# Как это работает?

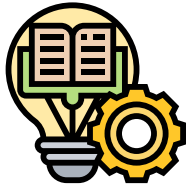
## Составление SPARQL-запроса или записи о фактах.

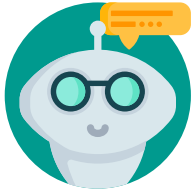
```
@prefix base: <http://trinidata.ru/fire/>
@prefix mdm: <http://trinidata.ru/archigraph-mdm/>
@prefix owl: <http://www.w3.org/2002/07/owl#>
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT *
WHERE {
        ?gm_d8fjtn7k rdf:type base:ГидроизоляционныйМатериал .
        ?mgm_leh75mcy rdf:type base:МаркаГидроизоляционногоМатериала .
        ?mgm_h5jd7fnh rdf:type base:МаркаГидроизоляционногоМатериала .
        ?po_h6jd8nmn rdf:type base:ПромышленныйОбъект .
        ?gm_d8fjtn7k base:имеетМаркуГидроизоляционногоМатериала ?mgm_leh75mcy, ?mgm_h5jd7fnh ;
                     base:используетсяНаОбъекте ?po_h6jd8nmn .
        }
```

Применение гидроизоляционных материалов марок Г-ИГ , С-РМ , что не соответствует назначению и планируемым условиям эксплуатации объекта проектирования.

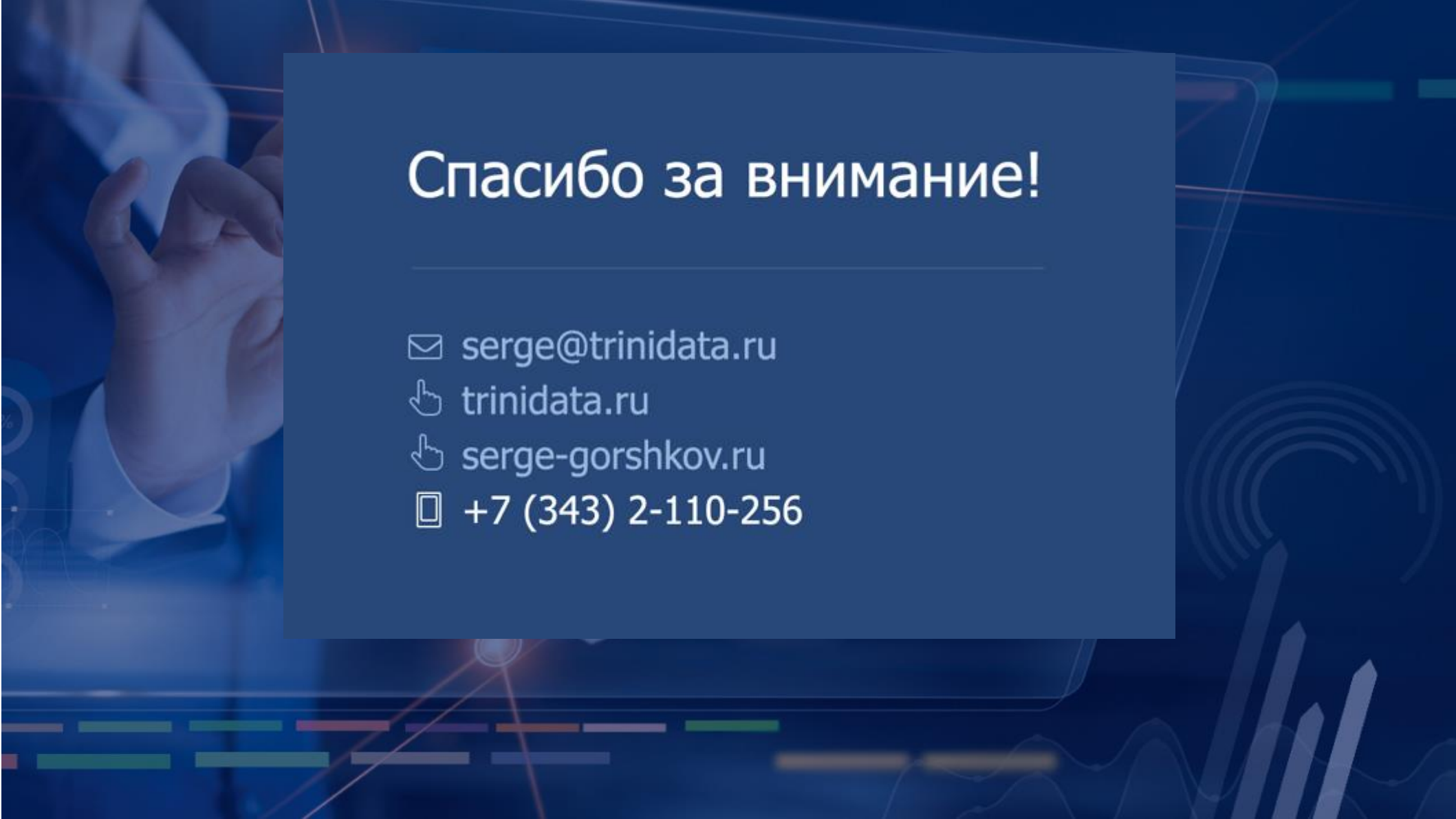
# Что в результате?





Совместное применение онтологий и технологий машинного обучения открывает путь к созданию:

1. Поисковых систем, которые точно отвечают на поставленные вопросы на обычном языке.
2. Систем управления знаниями, которые делают доступными огромные массивы текстовой информации.
3. Диалоговых систем, которые действительно «понимают» собеседника, а не генерируют что-то созвучное.


Спасибо за внимание!
serge@trinidata.ru
trinidata.ru
serge-gorshkov.ru
+7 (343) 2-110-256