DIALOGUE 2020

# ПРОЕКТИРОВАНИЕ И РАЗРАБОТКА ВОПРОСНО-ОТВЕТНОЙ СИСТЕМЫ НА ОСНОВАНИИ ГРАФОВ ЗНАНИЙ

Сомов Олег, МФТИ

# ЦЕЛЬ И МОТИВАЦИЯ ИССЛЕДОВАНИЯ

- **Цель работы -** разработать инструмент для быстрого создания доменных вопросно ответных систем и на его основании реализовать вопросно-ответную систему для двух доменов – фильмов и музыки.
- **Мотивация исследования –** улучшение методологии построения мульти-доменных вопросно-ответных систем.
  - Большинство вопросно-ответных систем плохо масштабируемы и не позволяют добавить новый домен(область знаний) быстро и без ущерба для качества.
  - Используется большой объем неструктурированных данных - обработка требует высоких вычислительных мощностей
  - Сложно контролировать  корректность текстовых данных и модифицировать/обновлять их.

# ОСНОВНЫЕ ТРЕБОВАНИЯ И ОГРАНИЧЕНИЯ ВОПРОСНО-ОТВЕТНОЙ СИСТЕМЫ

**Требования**:

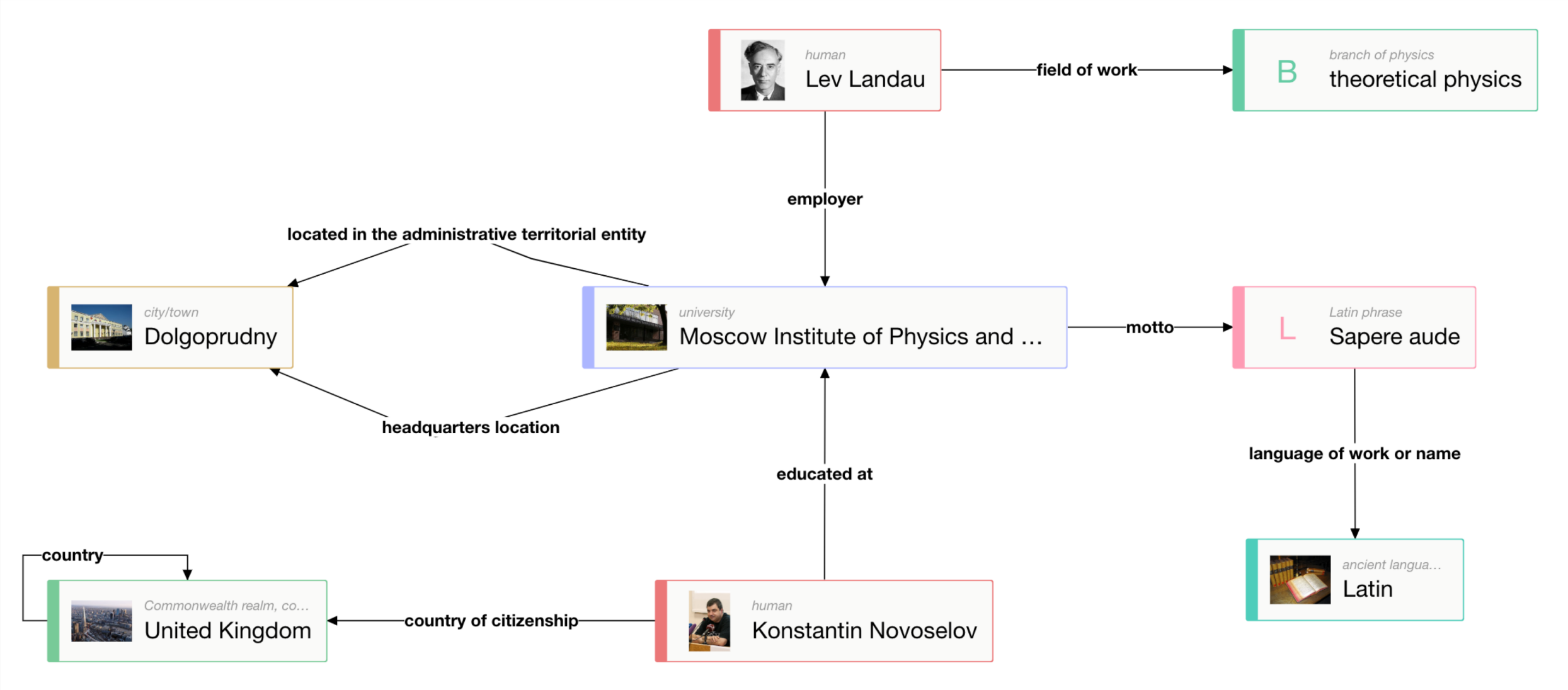
- Используемые данные должны содержать большинство знаний домена и должны быть корректны
- На вопросы, одинаковы по смыслу, но разные по формулировке должен быть дан **1** ответ
- Возможность быстрого добавления новых доменов к вопросно-ответной системе
- Система должна обладать высоким ответным качеством – метрика точность > **50%**
- Система должна быстро отвечать на входные вопросы **- 200 – 300** мс

**Ограничения**:

- Данные доступны и структурированы в формате RDF
- Система не подразумевает уточняющих вопросов и слежения за состоянием диалога
- Есть возможность использовать доменную краудсорс разметку

# ТЕРМИНОЛОГИЯ И ПОНЯТИЯ

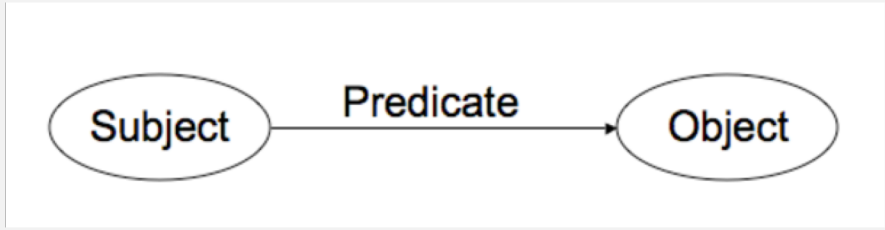
# ГРАФ ЗНАНИЙ


human
Lev Landau
field of work
B
branch of physics
theoretical physics
employer
located in the administrative territorial entity
city/town
Dolgoprudny
university
Moscow Institute of Physics and …
motto
L
Latin phrase
Sapere aude
headquarters location
language of work or name
educated at
country
Commonwealth realm, co…
United Kingdom
country of citizenship
human
Konstantin Novoselov
ancient langua…
Latin

# ПРЕДСТАВЛЕНИЕ ГРАФА ЗНАНИЙ

**RDF модель – формат хранения графа**

- Каждый узел имеет только 1 значение – ID либо значение
- Каждый узел связан как минимум с 1 другим узлом
- Онтология выражается триплетом – **субъект – предикат – объект**
- Используется для графов знаний
- Язык запросов – **SPARQL**




Пушкин – место рождения - Москва

```
@prefix ex: <http://example.com/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

ex:msg1 rdf:type ex:msg;
        ex:to "bob";
        ex:from "alice" ;
        ex:subj "alice to bob 1";
        .
```

Turtle data

# ПРИМЕР SPARQL ЗАПРОСА

**Какая самая высокая гора в Испании?**

```
Select ?mountain ?mountainHeight
Where
{
 ?mountain  wdt:P31/wdt:P279* wd:Q8502;
            wdt:P17 wd:Q29;
              wdt:P2044 ?mountainHeight.
}
ORDER BY DESC(?mountainHeight)
LIMIT 1
```

## Обозначения

**P31, P279** – определение объекта
**P17** – свойство “Локация”
**P2044** – свойство “Высота”

**wdt** – предикат
**wd** – обьект

**Q8502** – гора
**Q29** – Испания

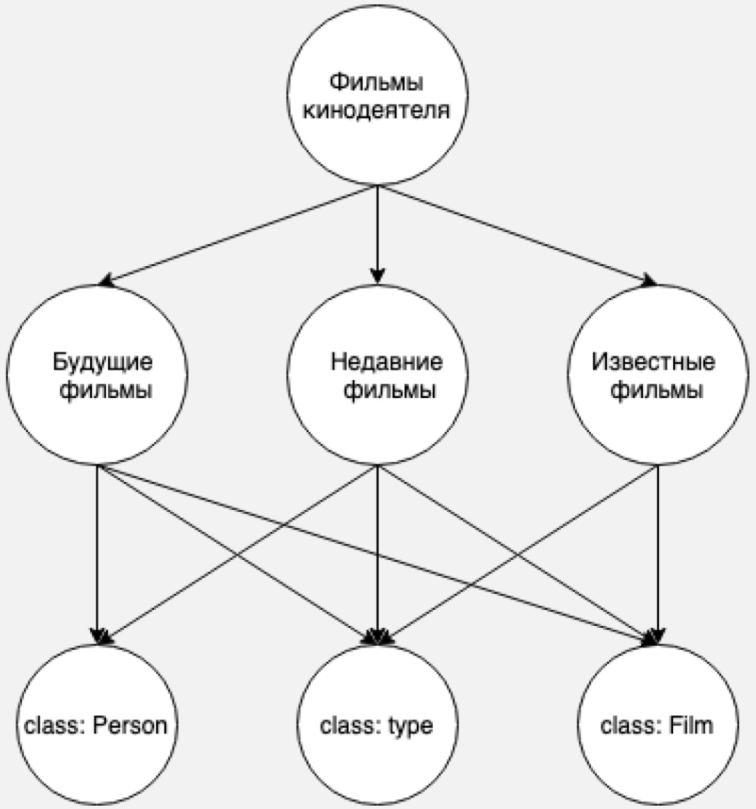
# СОЗДАНИЕ ГРАФА ЗНАНИЙ

# СОЗДАНИЕ ДОМЕННЫХ ГРАФОВ ЗНАНИЙ

1. На платформе разметки собирается информация типов вопроса к выбранному домену знаний
2. На основании данных выявляются необходимые типы сущностей и предикаты, которые должны быть в графе
3. Вопросы разбиваются на категории вопросов:
   - Информация о фильме, альбоме, исполнителе и т.д
   - Фильмы с двумя/тремя исполнителями
   - Популярные фильмы/альбомы в определенное время
4. Каждая категория вопросов – это некоторый подграф в графе знаний WikiData. Для каждого подграфа создается набор SPARQL запросов, которые извлекают всю информацию из WikiData.
5. При необходимости в других данных, используются другие источники. Соединение с сущностями из базового графа создается через новые предикаты.

# СТРУКТУРА ПОДГРАФА

- Каждый подграф имеет внутри себя набор сущностей, каждая из которых агрегирует по определенному признаку сущности графа  внутри себя
- Такая структура позволяет превратить задачу генерации SQL запроса в задачу классификации
- Собрать подграф можно с помощью 3-5 SPARQL запросов.
- Граф знаний – множество таких подграфов.



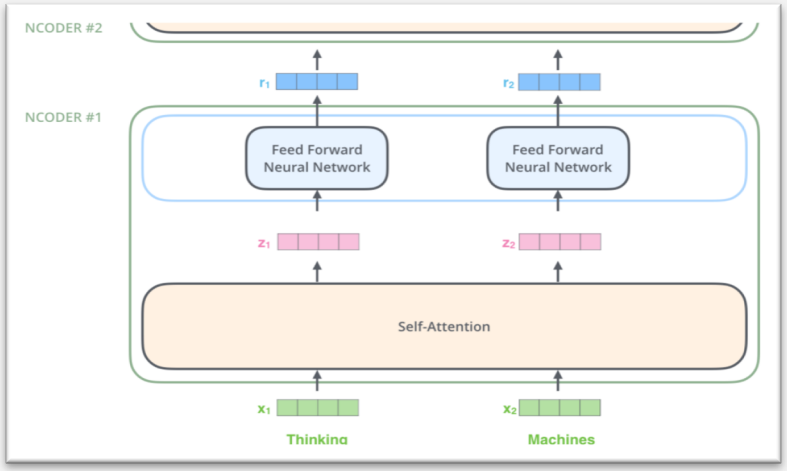

Пример подграфа “Фильмы кинодеятеля”

# СОЗДАНИЕ ВОПРОСНО-ОТВЕТНОЙ СИСТЕМЫ

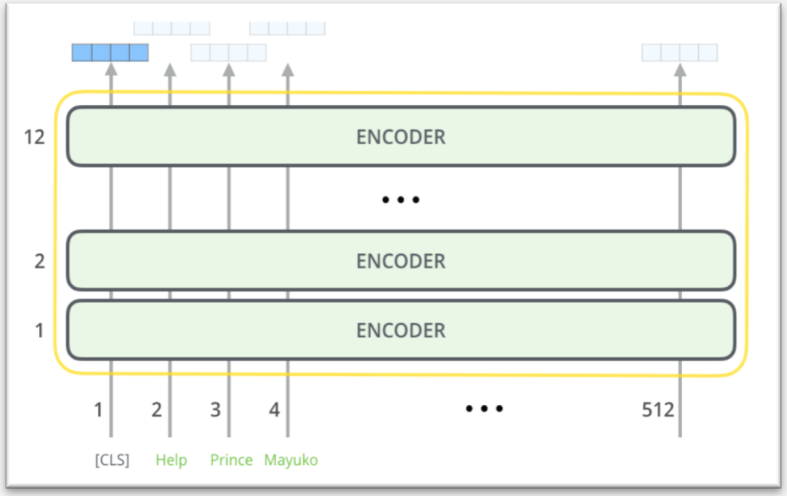
# BIDIRECTIONAL ENCODER REPRESENTATION FROM TRANSFORMERS

**BERT** – языковая модель, позволяющая преобразовать входное предложение в вектор, кодирующий семантику предложения.

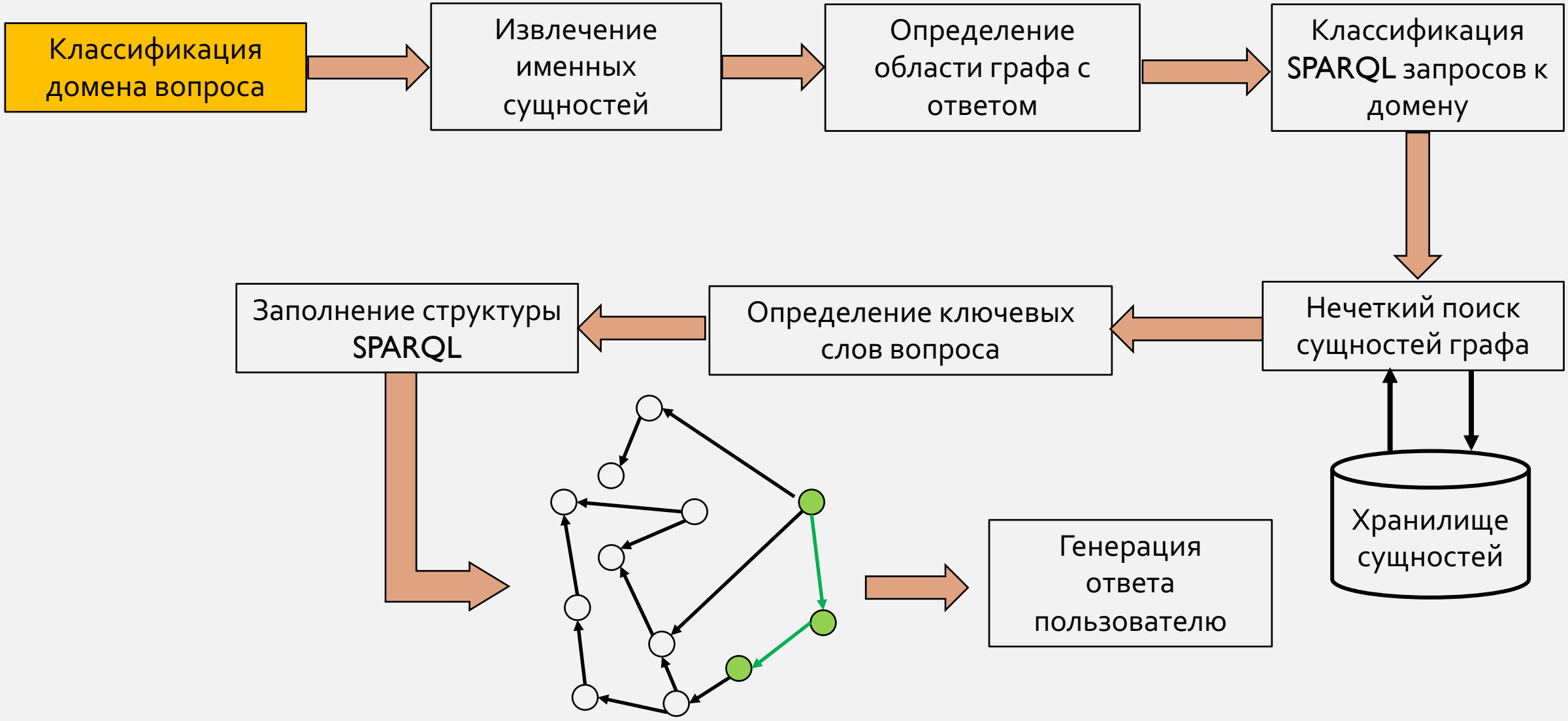
- **Архитектура сети**
  - Архитектура – стек трансформеров
  - Каждый блок(трансформер) – механизм внимания(attention) и полносвязная нейронная сеть
- **Тренировка сети за счет суммарной logloss ошибки на двух задачах:**
  - Предсказание некоторого слова в предложении
    - Я пошел в MASK и купил MASK
  - Определение логичности следующего предложения(Fine-tune)
    - Я пошел в магазин. Там я купил молоко. – True
    - Я купил собаку. В Москве холодает. – False
- **Преимущества модели**
  - Учет длительных зависимостей в предложении и контекста слов
  - Быстрое обучение на TPU(Google Collab)




Transformer

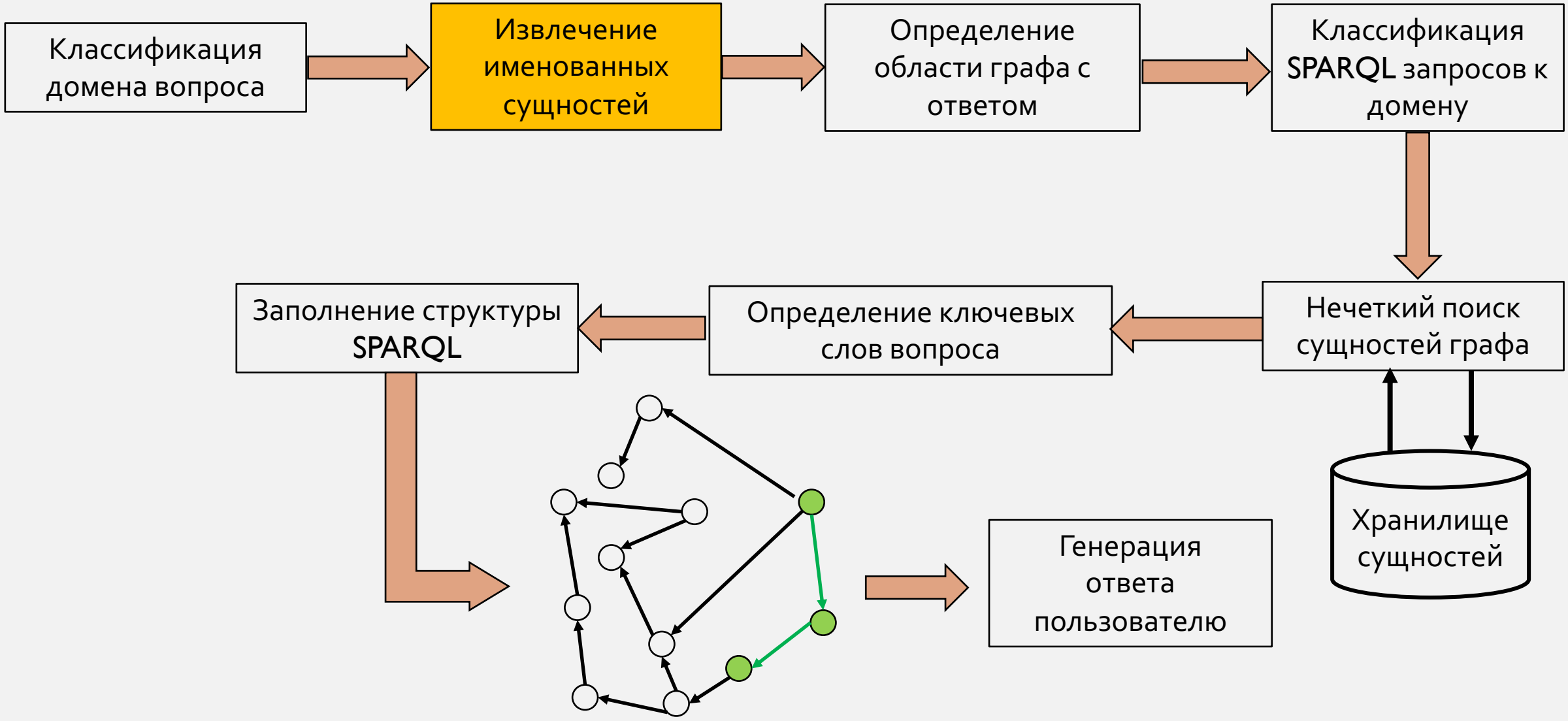



Encoder Stack

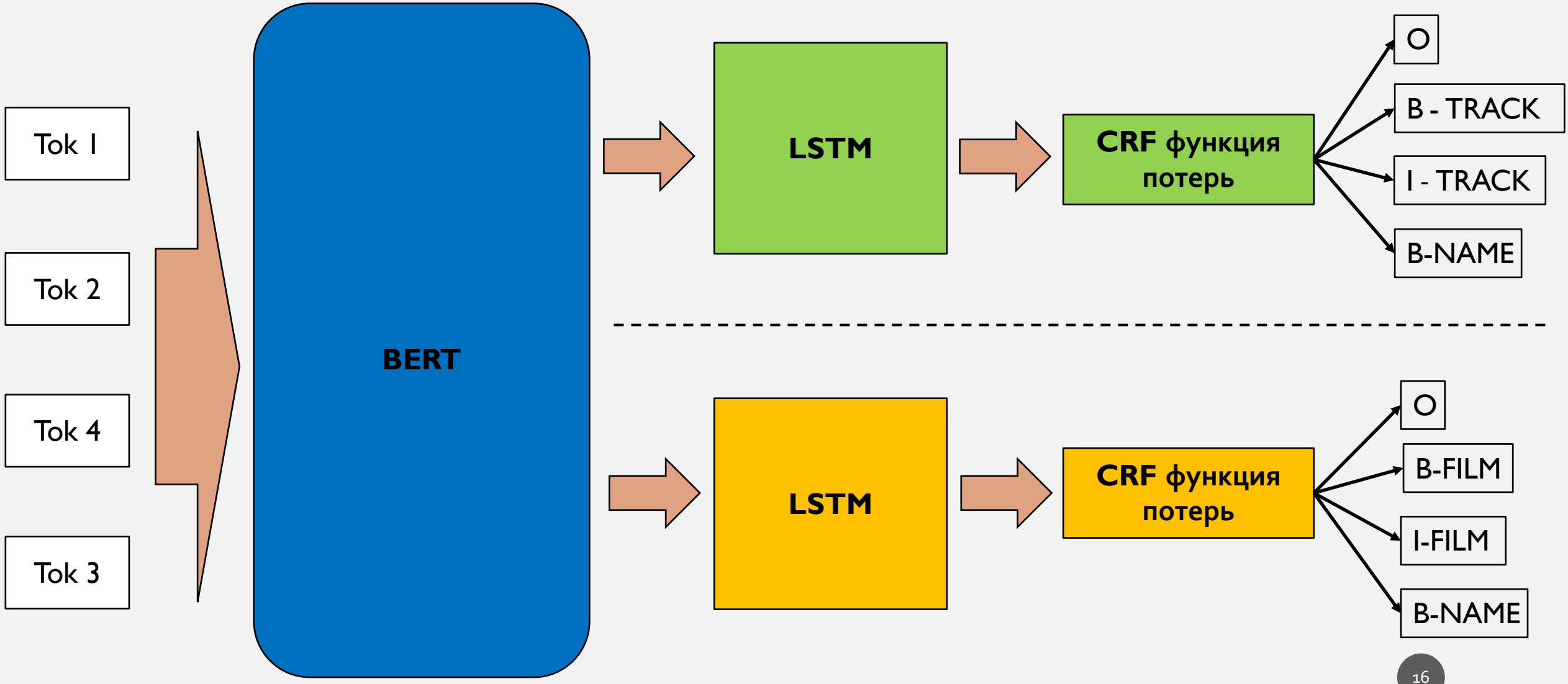

Классификация домена вопроса
Извлечение именных сущностей
Определение области графа с ответом
Классификация SPARQL запросов к домену
Нечеткий поиск сущностей графа
Определение ключевых слов вопроса
Заполнение структуры SPARQL
Хранилище сущностей
Генерация ответа пользователю

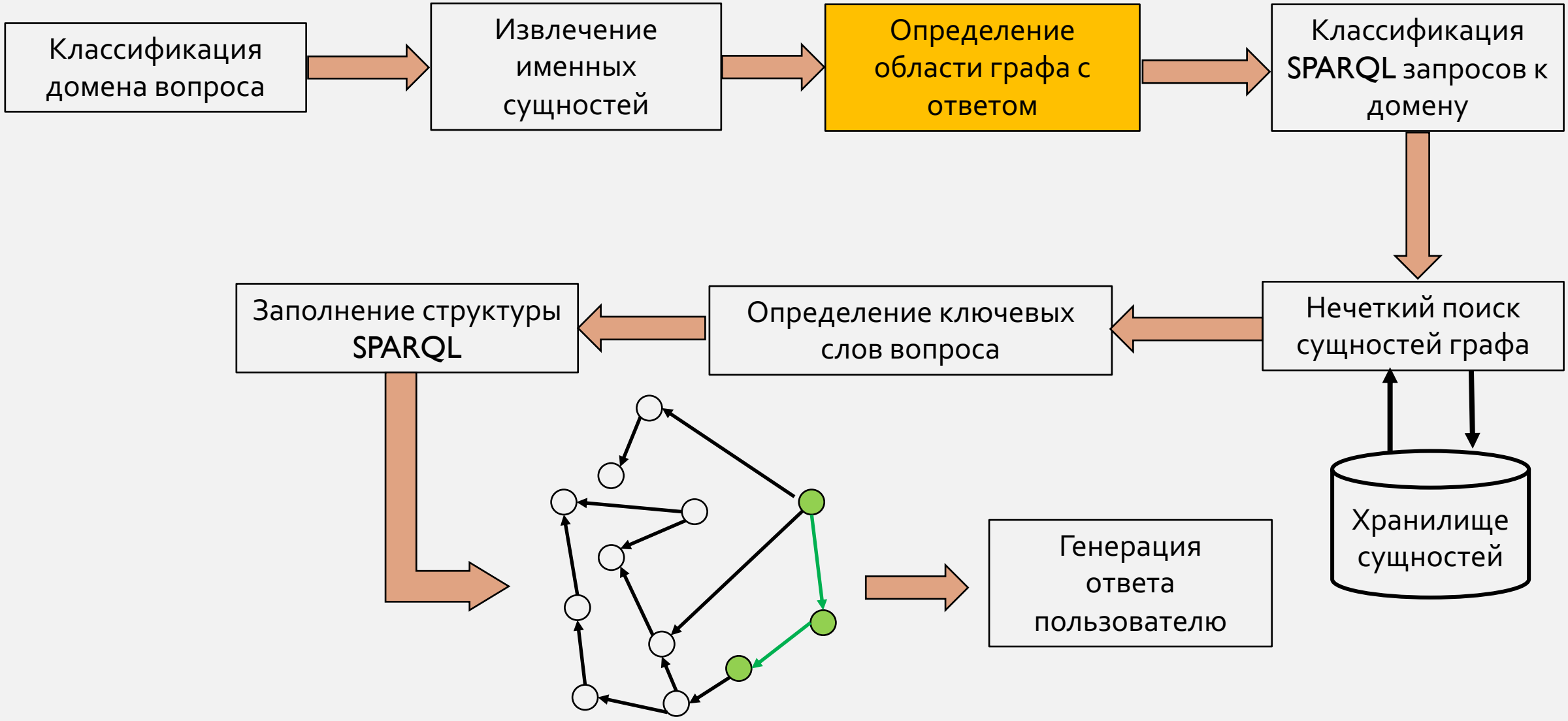
# КЛАССИФИКАЦИЯ ДОМЕНА

- Классификация домена позволяет определить к какому домену относится данный вопрос
- Знание домена помогает с определить именованные сущности в вопросе.
- Реализация модели классификации – 2ух слойная нейронная сеть на эмбеддингами BERT с логистической функцией ошибки.


Классификация домена вопроса
Извлечение именованных сущностей
Определение области графа с ответом
Классификация SPARQL запросов к домену
Нечеткий поиск сущностей графа
Определение ключевых слов вопроса
Заполнение структуры SPARQL
Хранилище сущностей
Генерация ответа пользователю

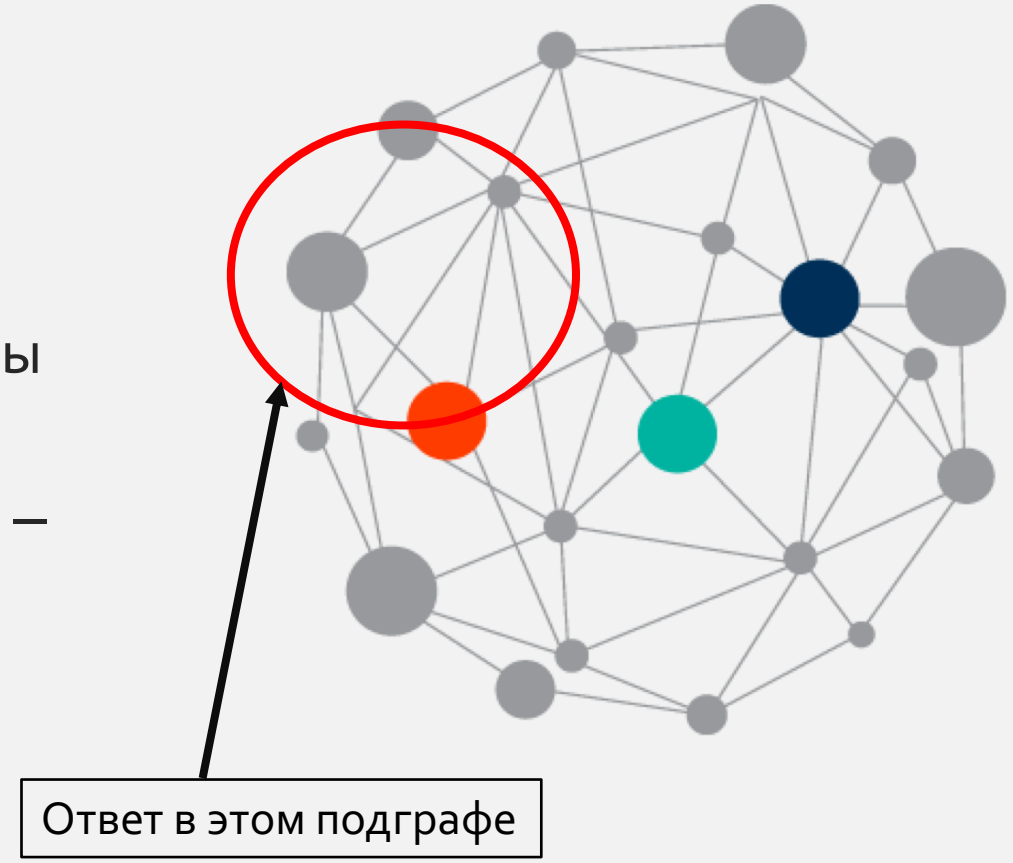
# МОДЕЛЬ ИЗВЛЕЧЕНИЯ ИМЕННОВАНЫХ СУЩНОСТЕЙ

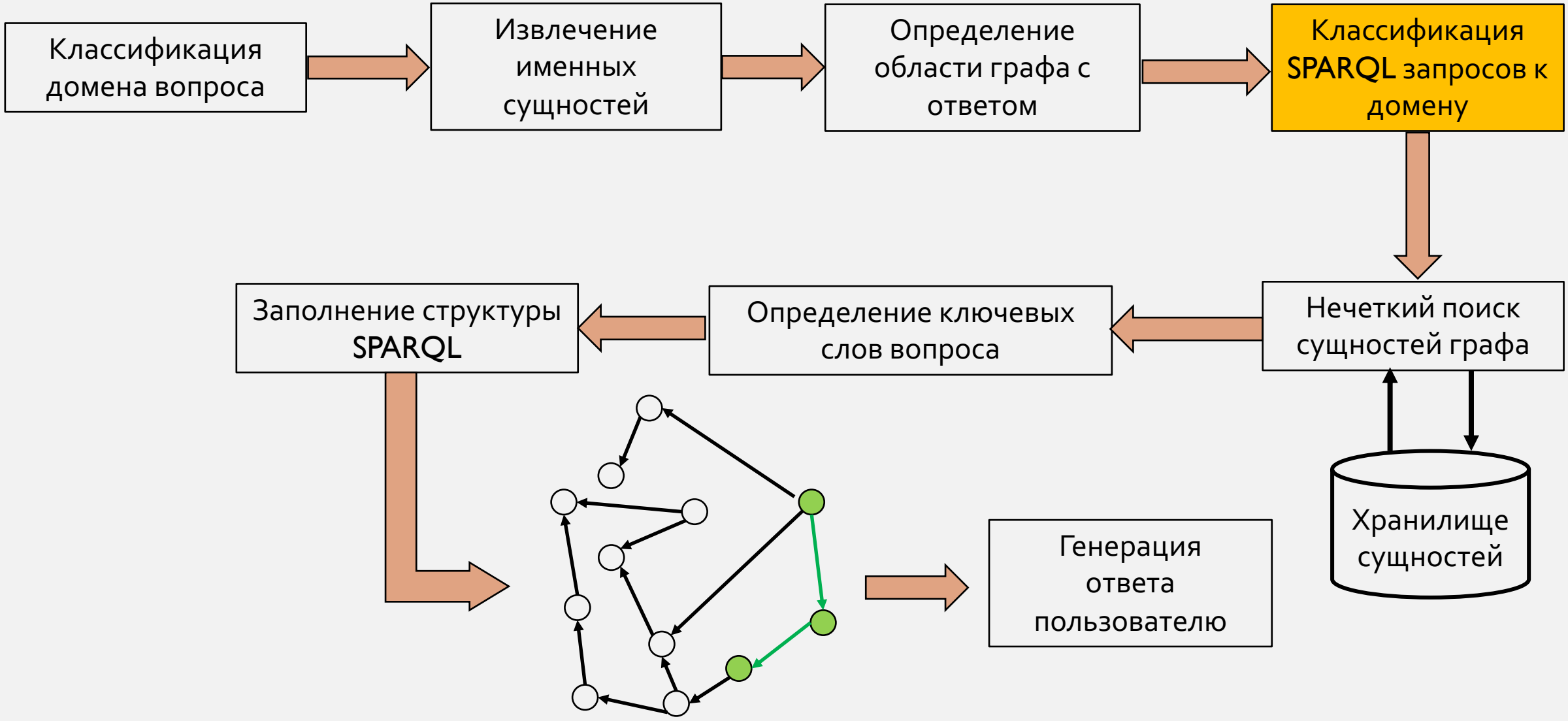



Классификация домена вопроса
Извлечение именных сущностей
Определение области графа с ответом
Классификация SPARQL запросов к домену
Нечеткий поиск сущностей графа
Определение ключевых слов вопроса
Заполнение структуры SPARQL
Хранилище сущностей
Генерация ответа пользователю

# КЛАССИФИКАЦИЯ ПОДГРАФА

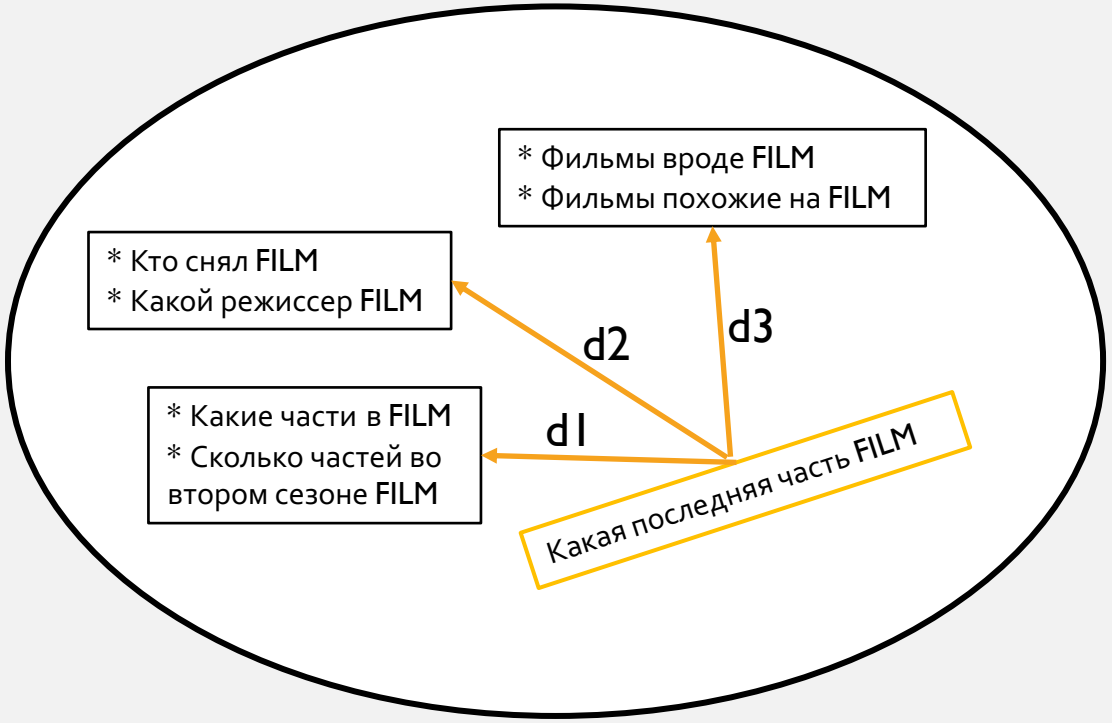
- Определяем в каком подграфе будем искать ответ на вопрос
- Разметка датасета:
  - Кто сыграл в фильме FILM – информация о фильме
  - В каком фильме недавно снялся NAME –фильмы актера
  - В каких фильмах вместе играли NAME и NAME – информация об альбоме
- Подход к быстрому добавление новых областей графа – DSSM модель с triplet loss для формирования вектора + KNN классификация.
- Не нужно переобучать модель для добавления нового подграфа.




Классификация домена вопроса
Извлечение именных сущностей
Определение области графа с ответом
Классификация SPARQL запросов к домену
Нечеткий поиск сущностей графа
Определение ключевых слов вопроса
Заполнение структуры SPARQL
Хранилище сущностей
Генерация ответа пользователю

# ОПРЕДЕЛЕНИЕ SPARQL ШАБЛОНА

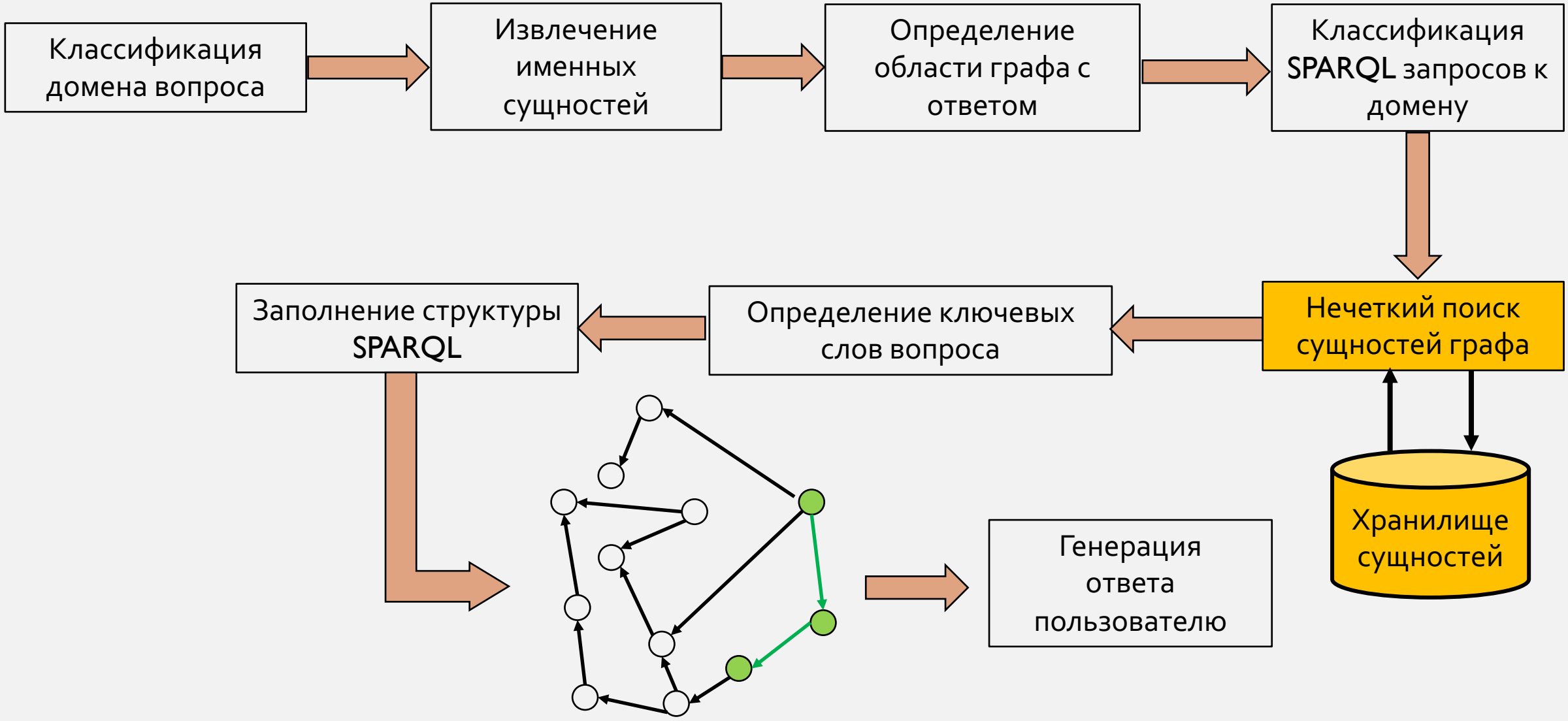
- С помощью KNN выделен подходящий базовый шаблон вопроса, то есть путь в графе




Подграф “Информация о фильме”

```
SELECT DISTINCT ?answer  WHERE
{
    ?x dr:isPartOfSeries FILM ;
                    rdfs:label ?answer .
}
```

Путь в графе “Части фильма”


Классификация домена вопроса
Извлечение именных сущностей
Определение области графа с ответом
Классификация SPARQL запросов к домену
Нечеткий поиск сущностей графа
Определение ключевых слов вопроса
Заполнение структуры SPARQL
Хранилище сущностей
Генерация ответа пользователю

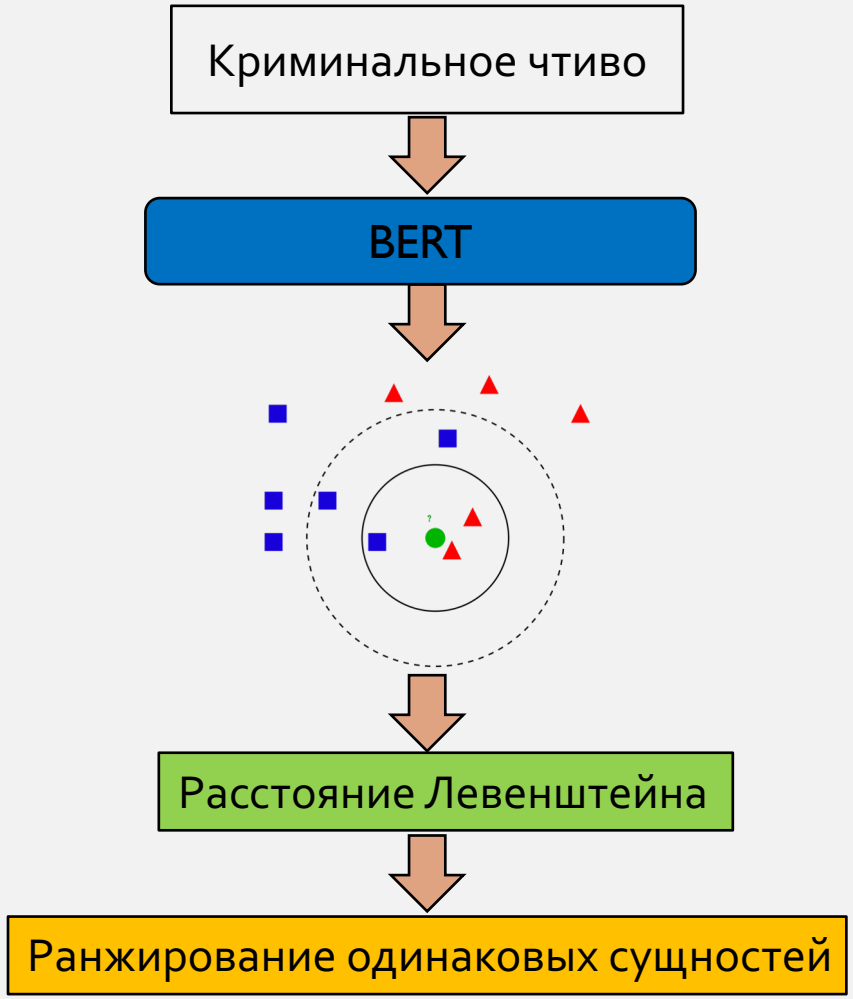
# НЕЧЕТКИЙ ПОИСК СУЩНОСТЕЙ

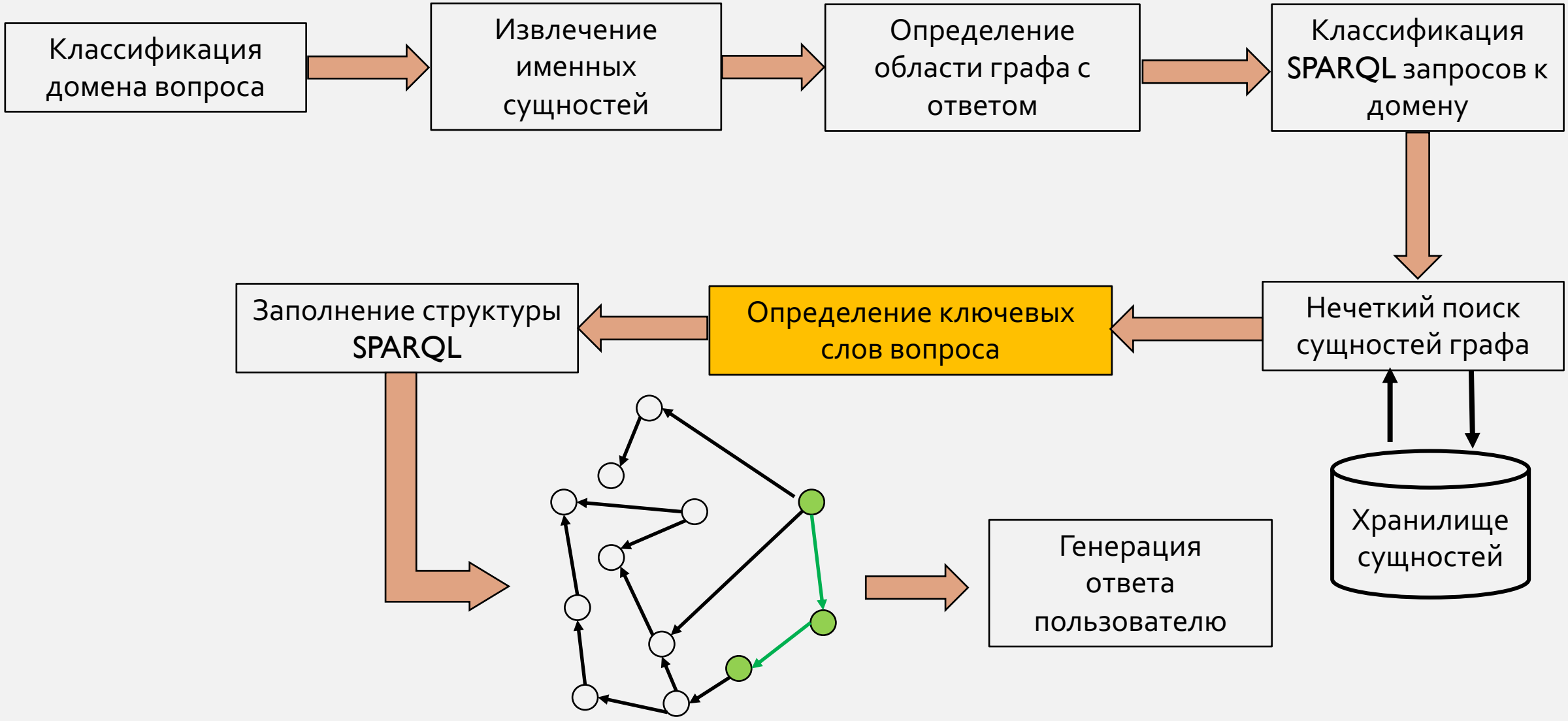
- Для каждого класса именной сущности есть коллекция этих сущностей в базе. На основании разметки NER производится поиск нужного ID .

Алгоритм поиска ID сущности в базе знаний:

1. Векторизация с помощью BERT модели
2. Поиск ближайшего соседа с помощью метода ближайших соседей в $R^n$
3. Ранжирование ближайших соседей объекта с помощью алгоритма сравнения строк – алгоритм Левенштейна
4. Ранжирование сущностей одинаковых сущностей за счет эвристик(Титаник, 1953 vs Титаник, 1997)
   - Популярность сущности
   - Год выпуска

- Возможность поиска ID связанных сущностей вопроса:
  - *Кто сыграл Анри Фортена в фильме Отверженные?*




Классификация домена вопроса
Извлечение именных сущностей
Определение области графа с ответом
Классификация SPARQL запросов к домену
Нечеткий поиск сущностей графа
Определение ключевых слов вопроса
Заполнение структуры SPARQL
Хранилище сущностей
Генерация ответа пользователю

# ОПРЕДЕЛЕНИЕ КЛЮЧЕВЫХ СЛОВ И ЗАПОЛНЕНИЕ ШАБЛОНА SPARQL

- Выделить ключевые слова, которые влияют на структуру запроса
  - Первый фильм Квентина Тарантино – первый – ORDER BY ASC(?date)
  - Сколько фильмов – COUNT(?x)
  - Кто был режиссером – wdt:P57
  - Сколько лет – (current date - date of birth)
- Заполнить соответствующий обработчику шаблон SPARQL предикатами и сущностями

Сколько фильмов снял Квентин Тарантино?

```
Select COUNT(?x) WHERE {
   ?x wdt:P57 wd:Q3772
}
```

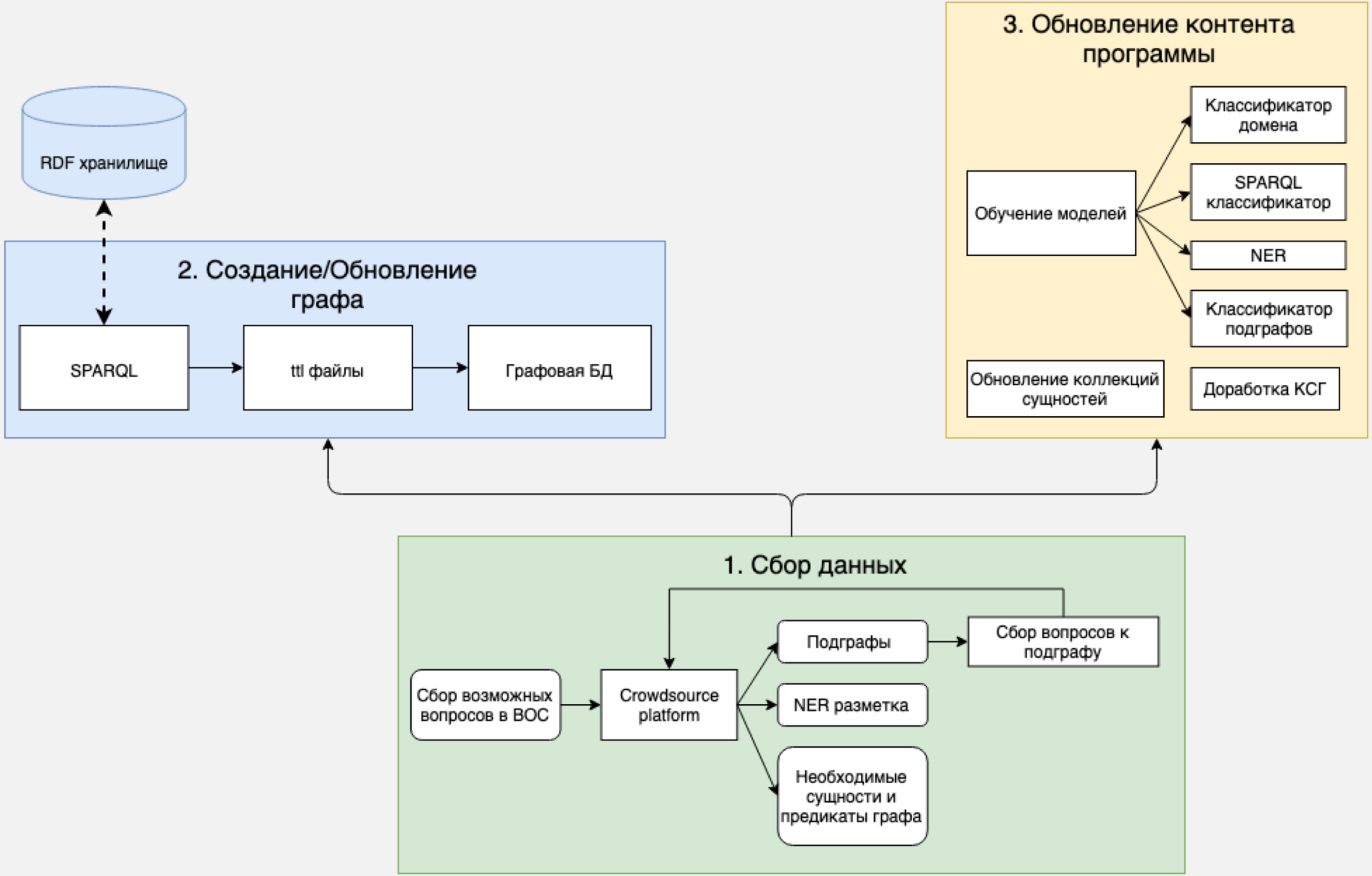
Последний трек The Beatles

```
SELECT DISTINCT ?xLabel
 WHERE {
 ?composition_id_1 wdt:P175 wd:Q1299 .
 ?x dr:hasPublicationDate ?date ;
 rdfs:label ?xLabel .
 }
 ORDER BY DESC(?date)
 LIMIT 1
```

# МЕТРИКИ СИСТЕМЫ

| | Модель | Macro F1 score | Accuracy | Датасет |
|---|---|---|---|---|
| Классификация домена | BERT + 2 Dense + logloss | 0.96 | 0.97 | Набор вопросов к домену |
| NER модель | BERT +LSTM + CRF | 0.73 | 0.98 | Разметка NER(дисбаланс классов) |
| Entity Linking | BERT + KNN + fuzz + ranging | - | 0.95 | Набор сущностей с опечатками, без опечаток и с измененным порядком слов |
| Классификация подграфа | BERT + DSSM + triplet loss | 0.91 | 0.93 | Маскированные парафразы к каждой области графа |
| Классификация шаблона | BERT + KNN | 0.97 | 0.98 | Маскированные парафразы к шаблону SPARQL |
| Ответы на вопросы по двум доменам | | - | 0.59 | Контрольный датасет вопросов |

# МЕТОДОЛОГИЯ ДОБАВЛЕНИЯ НОВОГО ДОМЕНА



# ИТОГИ

- Разработан и реализован универсальный инструмент для создания вопросно-ответных систем для различных доменов знаний
- С помощью инструмента создания ВОС реализованы вопросно-ответные системы для домена музыки и фильмов
- Представлена методология по добавлению нового домена знаний в вопросно-ответную систему
- Использован алгоритм машинного обучения Metric Learning для быстрого добавления новых подграфов без переобучения нейронной сети