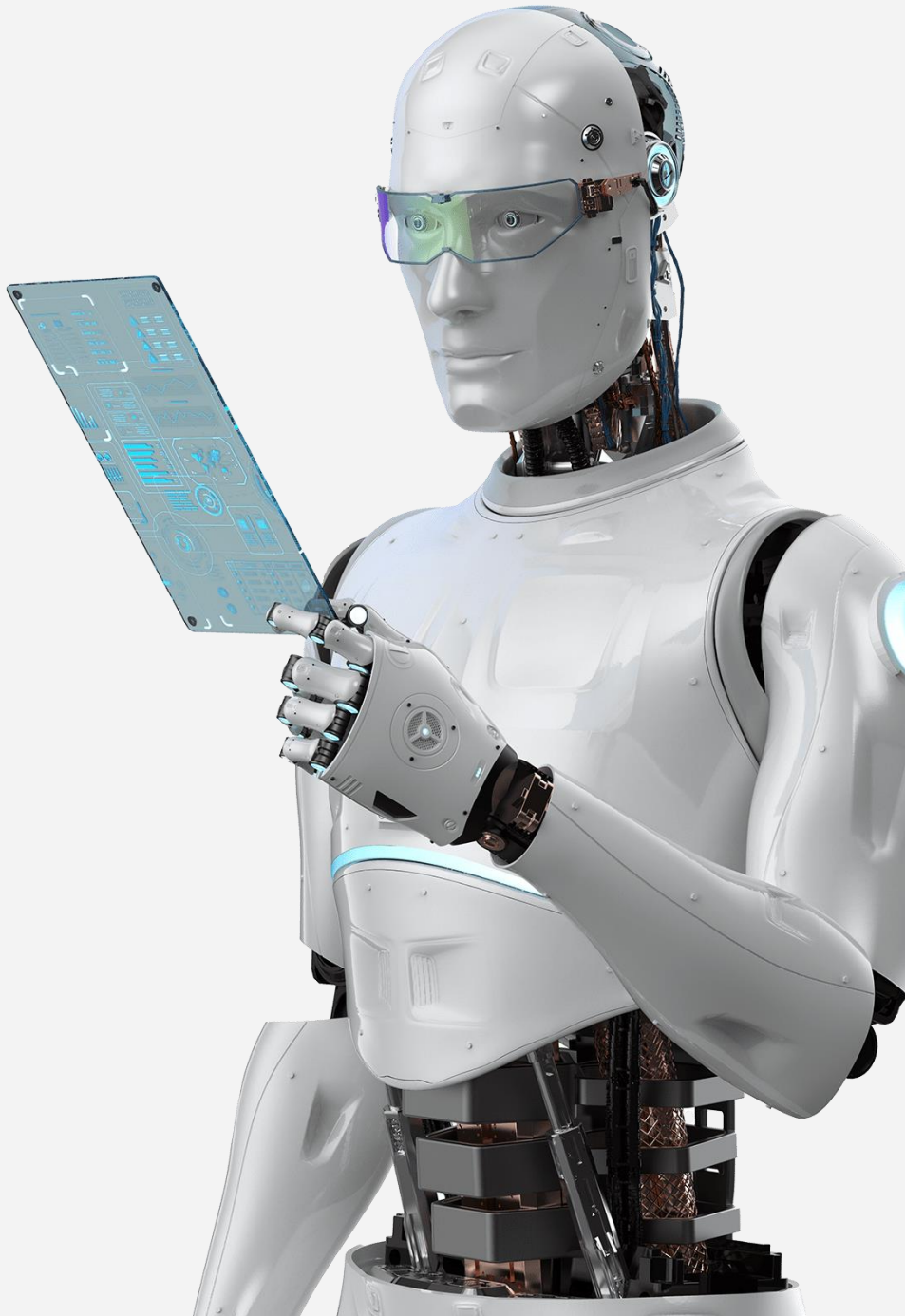
ЦПОД ТО «СОЗВЕЗДИЕ»

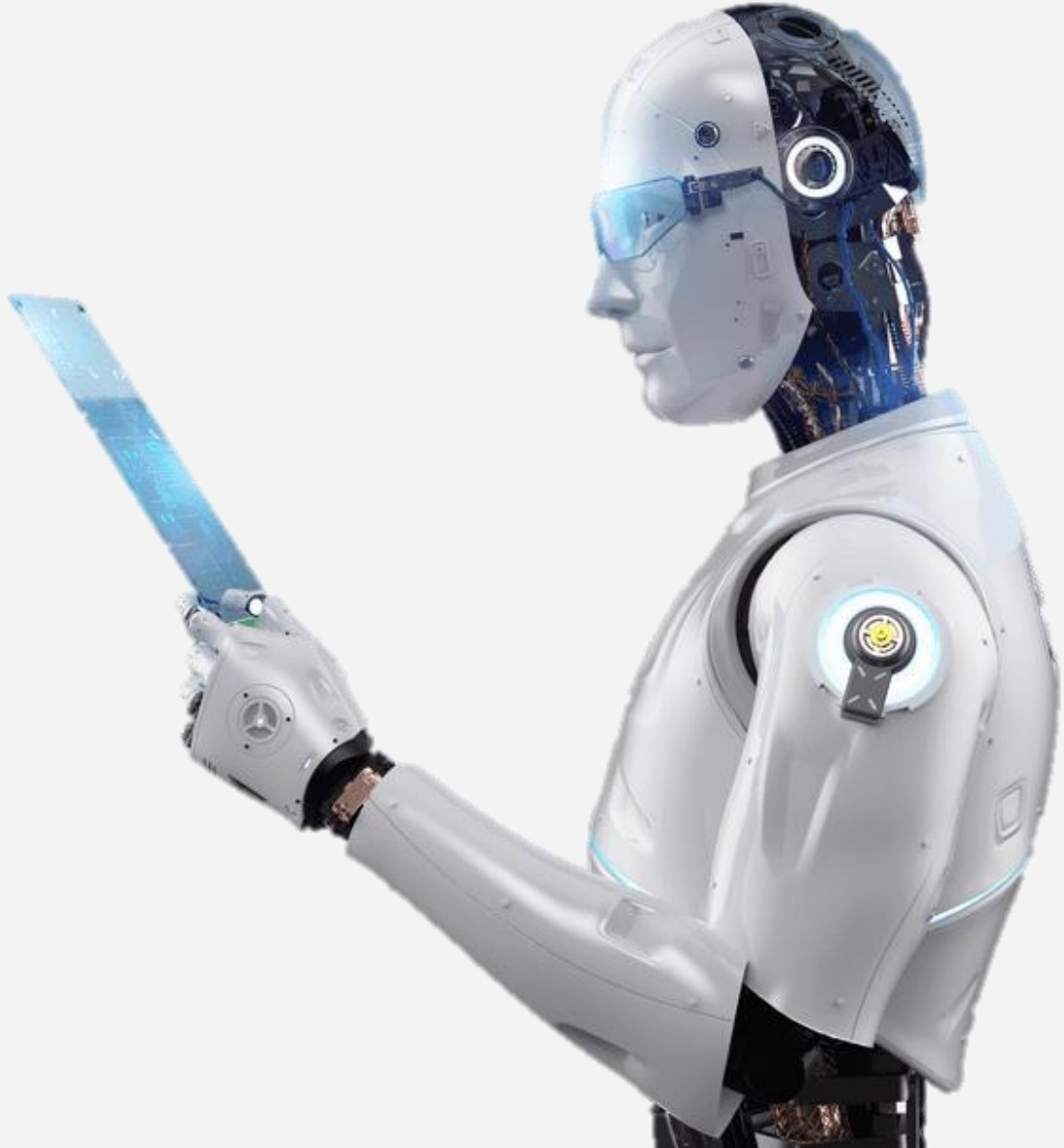
Исследование массива медицинских данных для создания предсказательного сервиса



Подготовили Гончаров Владислав, Григорьев Илья, Кончаков Павел

Типы медицинских рекомендательных систем :

- mediktor
- Symptomate
- Helzy



ПРЕИМУЩЕСТВА И НЕДОСТАТКИ MEDIKTOR :

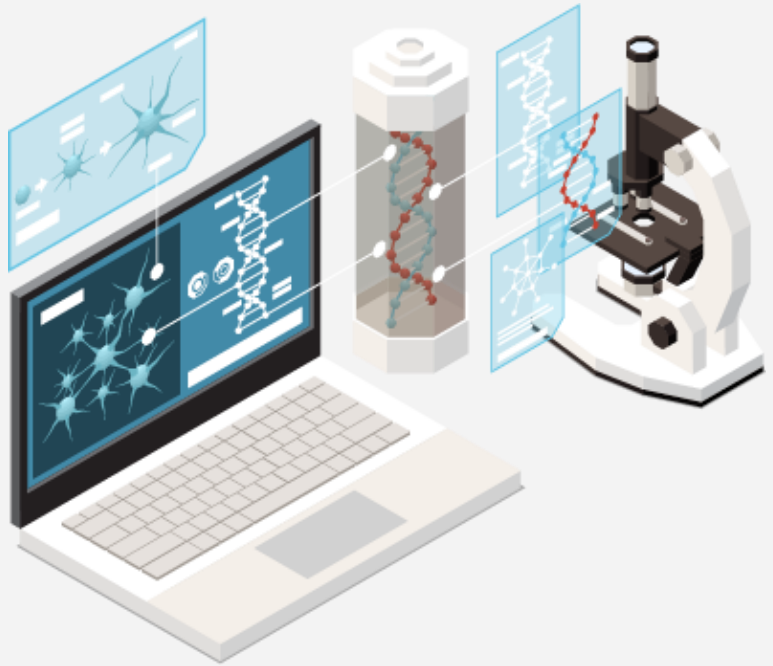
- Сопровождение голосом
- Вывод нескольких возможных диагнозов
- Вариативное анкетирование

- Нет поддержки русского языка
- Необходимо знать название симптома

Преимущества и недостатки Symptomate :

- Открытый код
- Подробное анкетирование

- Очень частая капча
- Необходимо знать название симптома



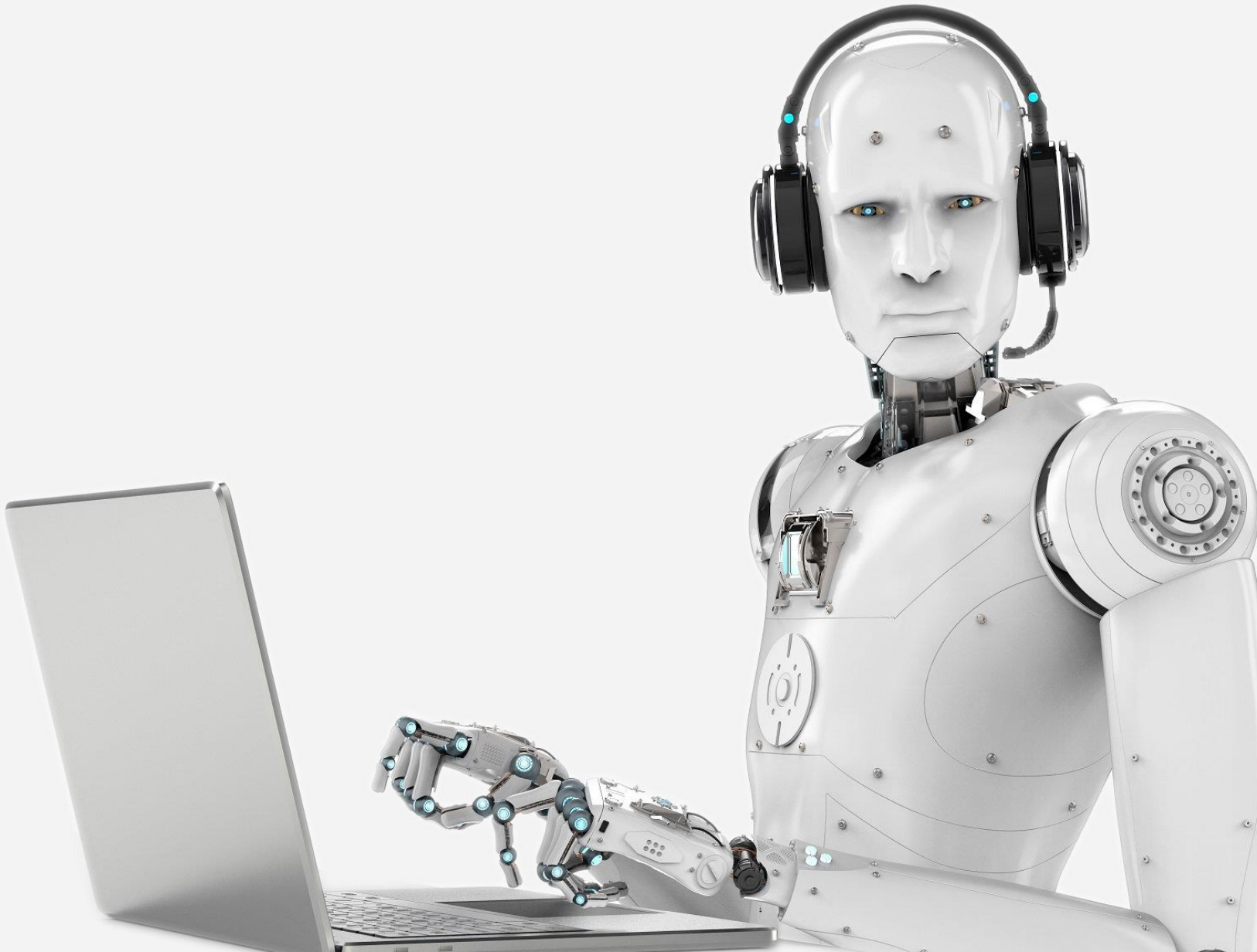
Преимущества и недостатки Helzy :

- Подробное анкетирование
- Вывод нескольких возможных диагнозов
- Вольное описание симптомов

- При описании симптомов часто неправильно их распознаёт
- Иногда вопросы не соответствуют симптомам

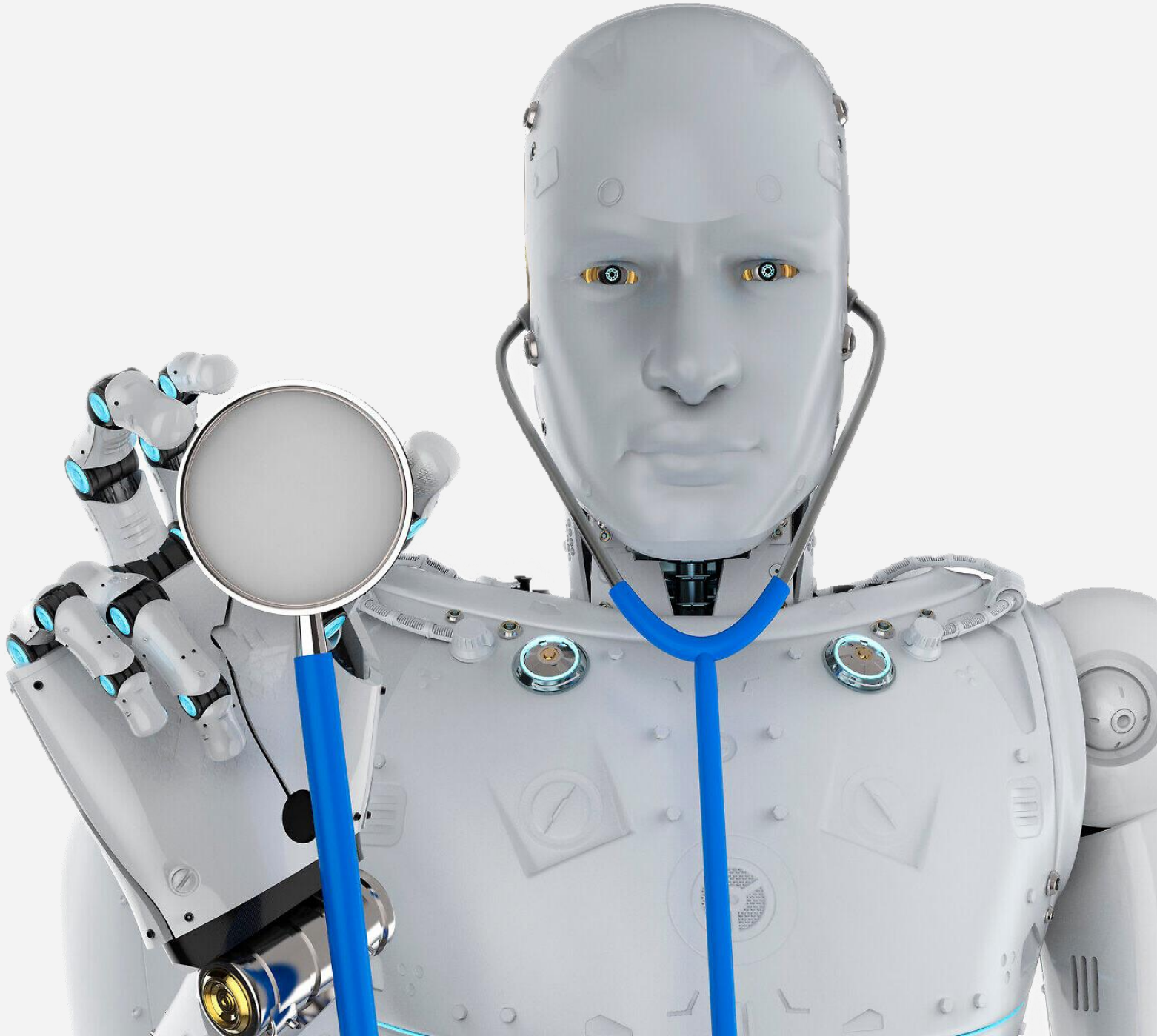
НАША ИДЕЯ :

При обращении за **медицинской** помощью, достаточно будет просто позвонить и **озвучить** жалобы на здоровье. ИИ примет звонок, **проанализирует** жалобы, **определит** диагноз и запишет к необходимому специалисту



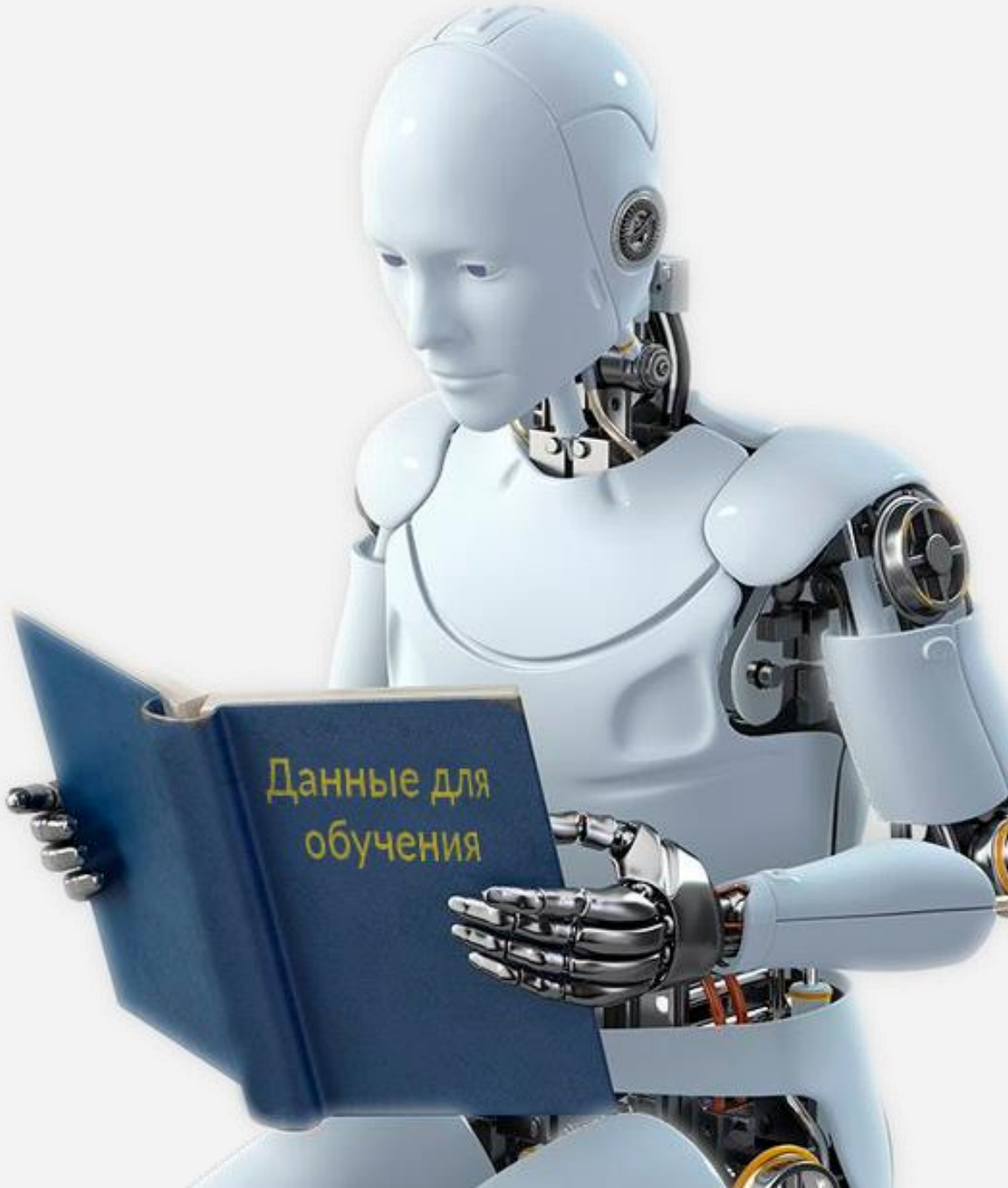
ЦЕЛИ ПРОЕКТА :

- **Упростить** процесс записи к врачу
- **Избавить** пациентов от необходимости самостоятельно **выбирать** специализацию врача
- **Помочь** пациентам, освободив их от необходимости **описывать** свои **симптомы** другому человеку



Задачи проекта :

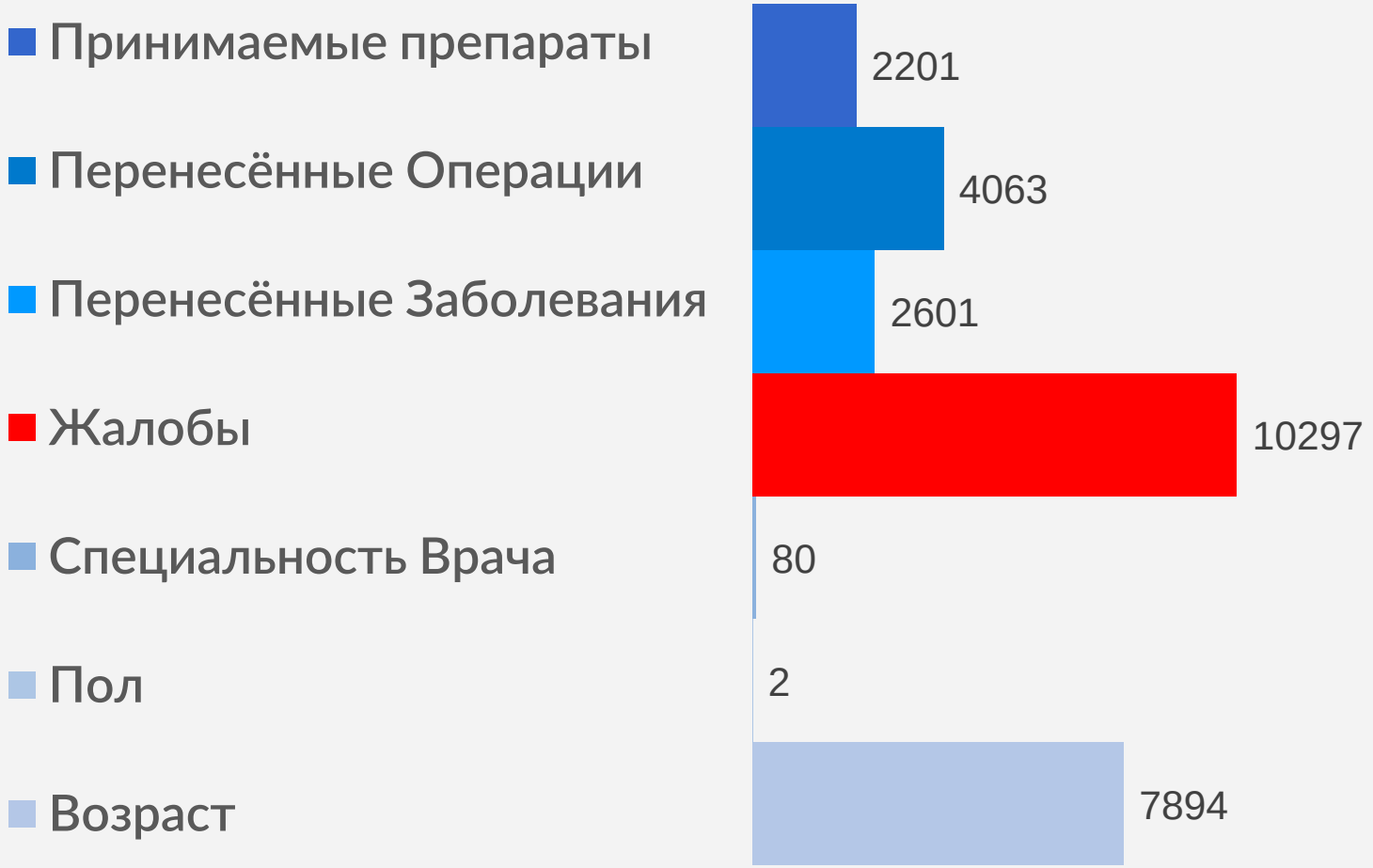
- Провести **анализ** данных
- Представить **данные** в удобном для **обучения ИИ** виде
- Подобрать **модель** для **нейронной** сети
- Провести **обучение**
- Добиться **максимальной** точности



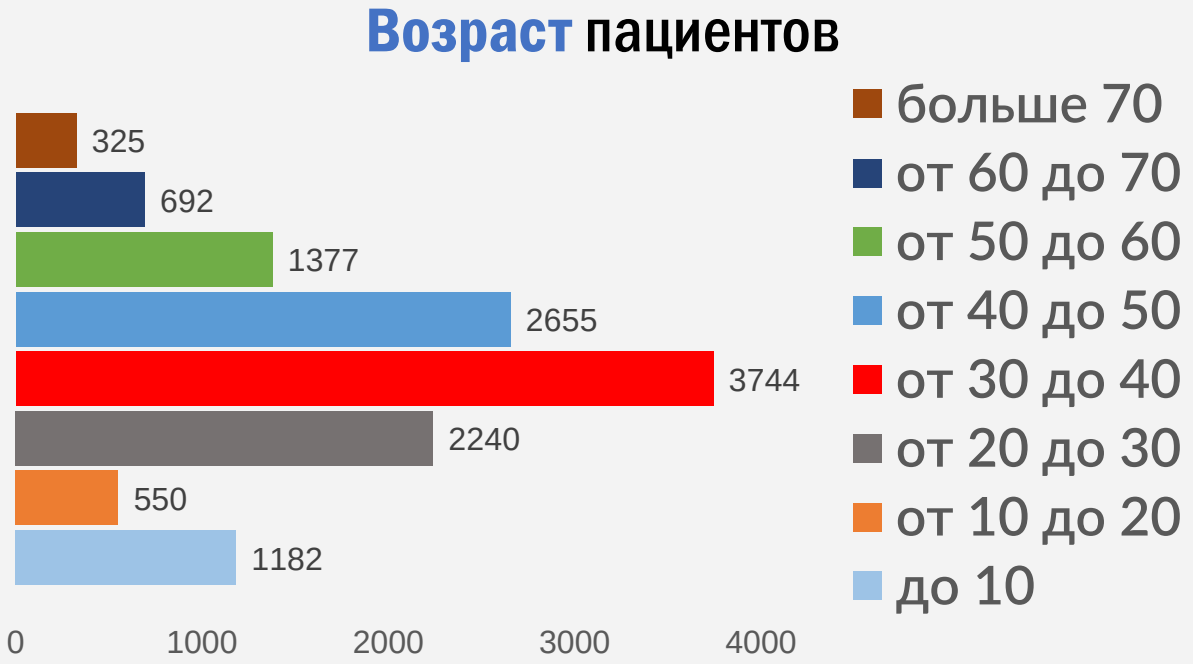
Исследование медицинских данных :

- Уникальных записей **12765**
- Самая популярная жалоба означает **отрицание жалоб**
- В последних трёх столбцах в общей сложности отсутствуют данные по **9322** записям
- **2/3** записей содержат информацию об отсутствии данных.

Количество **информации** в базе данных



Исследование медицинских данных :



Чаще всего к врачам обращались в возрасте около от 30 до 40 лет.

Женщины посещают врачей чаще, чем мужчины в 1,5 раза.

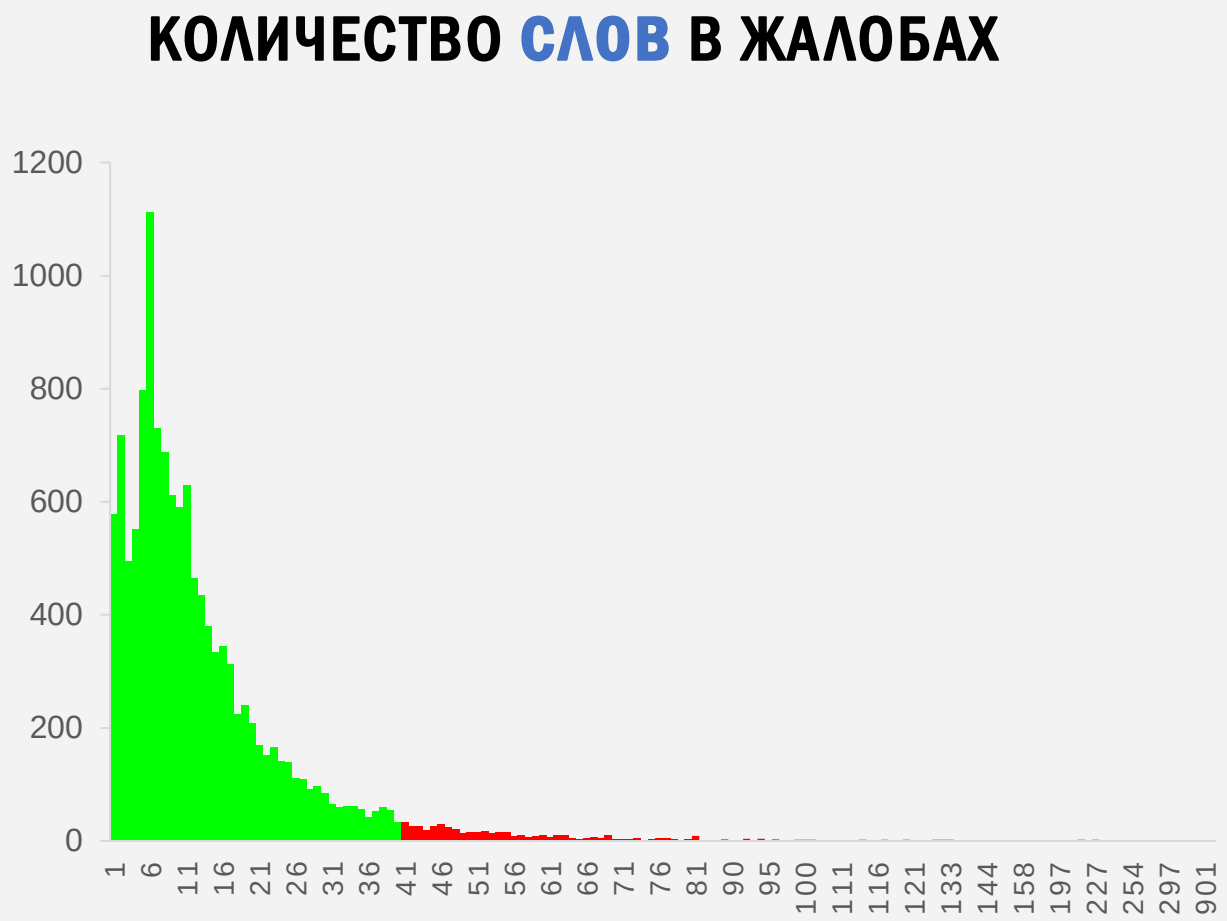
ИССЛЕДОВАНИЕ МЕДИЦИНСКИХ ДАННЫХ :

КОЛИЧЕСТВО ВРАЧЕЙ В БАЗЕ ДАННЫХ

Этот график показывает, что база распределена **неравномерно**, поэтому для **увеличения** точности модели стоит выделить несколько самых **востребованных** направления, а остальные **объединить** в категорию другие.

ИССЛЕДОВАНИЕ МЕДИЦИНСКИХ ДАННЫХ :

- Самая длинная жалоба состоит из **1162** слов
- Большинство жалоб не больше **40** слов



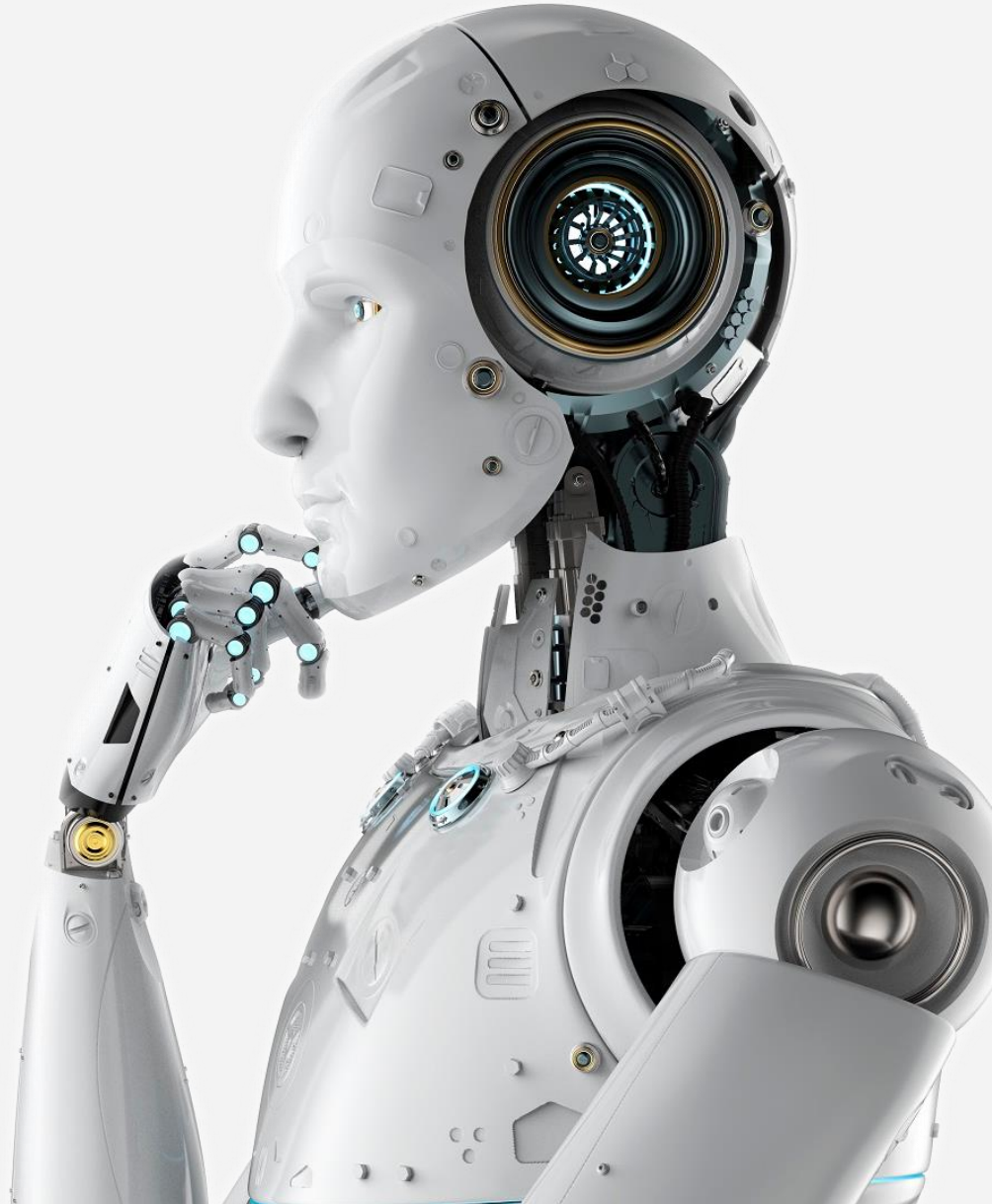
ПРОЦЕСС ОБРАБОТКИ МАССИВА ДАННЫХ :

1. **Уберём** лишние столбцы (**PatientKey, MedicalRecordDate, MedicalRecordKey**)

2. **Удаляем пустые** значения

3. **Группируем схожие** значения

4. Описываем **признаки**



ОПИСАНИЕ ПРИЗНАКОВ :

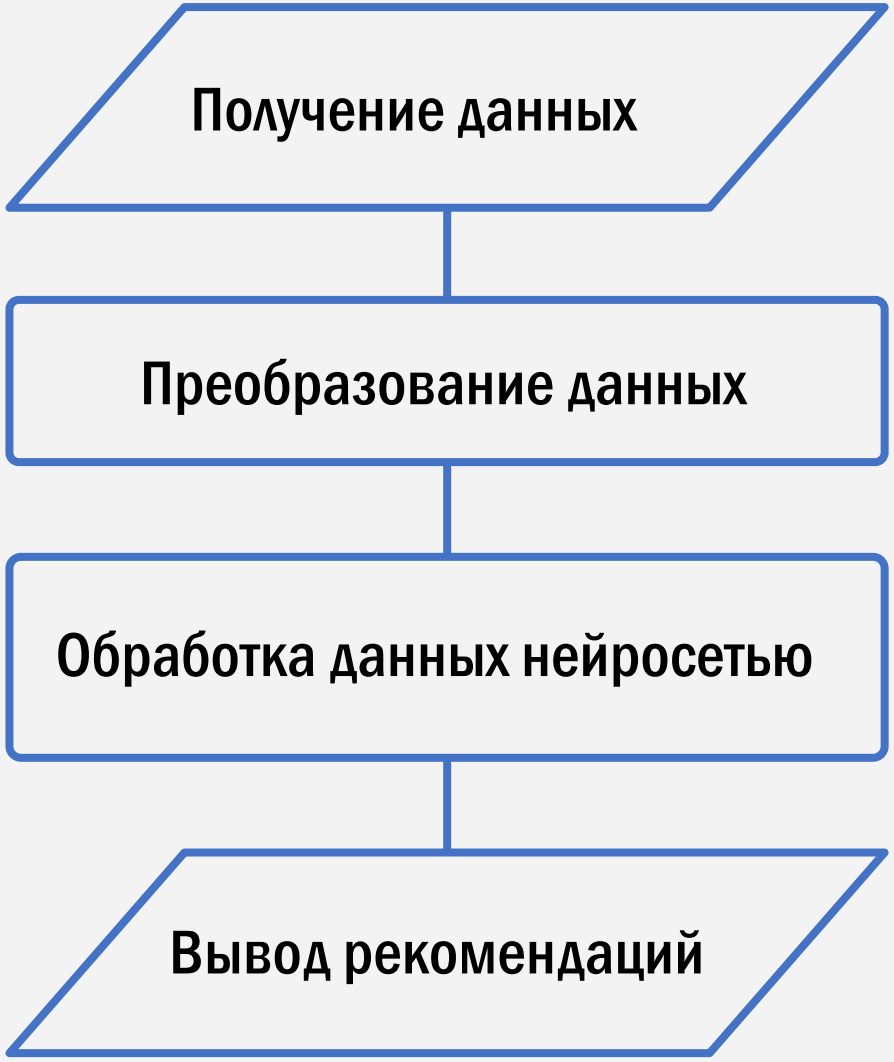
Признак	Пример
Жалоба. Представлена в виде **индексов** частотного списка	на заложенность в левом ухе, заложенность носа [2(на),32(заложенность), 3(в), 79(левом), 142(ухе), 32(заложенность), 27(носа)]
Пол. Оставим в том же виде, что представлен в базе	Мужской - 0 Женский - 1

ОПИСАНИЕ ПРИЗНАКОВ :

Признак	Пример
Возраст. Приведен к **нормированному** значению в интервале от **0** до **1** (делим на 365 и 90) .	1500 дней → 0,456621
Специальность Врача. Представим в виде **One Hot Encoding**	[0 0 0 0 **1** 0 0 0 0 0 0 0 0 0] **Индекс 1** соответствует **индексу** Врача

Описание алгоритма рекомендательной системы :

- Пока что **входные данные** представлены в виде **текста (временное решение)**.
  Финальная версия получает на вход — **аудио файл**.
- **Преобразование** данных к необходимому виду, для **отправки** в нейронную сеть
- Получение **предсказания** на основе входных данных
- Пока что, **выходные данные** представлены в виде **текста (временное решение)**.
  Финальная версия также синтезирует **аудио файл**.



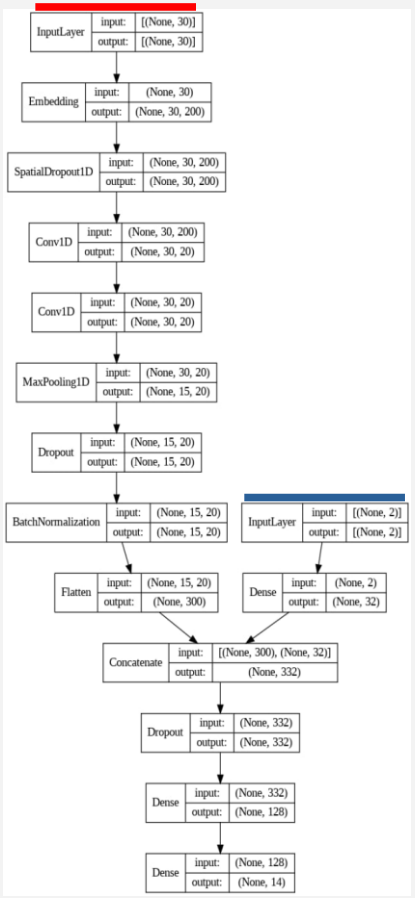
Характеристики модели :

- Точность около 80%
- 2257806 параметров
- Оптимизатор rmsprop
- Функция потерь – Кросс-энтропия

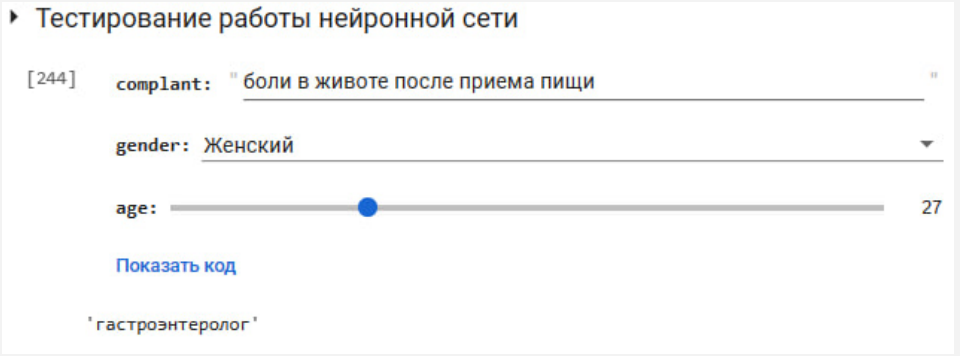
ПРОТОТИП ИСКУССТВЕННОГО ИНТЕЛЛЕКТА:

Нейронная сеть имеет два различных **входа** данных:

На **1-ый вход** поступают жалобы в формате индексов, а затем, в процессе обучения, с помощью эмбеддинга (англ. «embedding» — вложение) располагаются в многомерном пространстве таким образом, что что близкие по смыслу слова кодируются схожими векторами. Для вычисления эмбеддингов используется гипотеза локальности: слова, которые встречаются в одинаковом окружении, имеют близкие значения.



На **2-ой вход** поступают данные о возрасте и поле пациента. Эти данные обрабатываются полносвязными слоями



Результаты обработки **объединяются** и формируется результат