Data Science

Специалист по Data Science — это человек, который анализирует большие объёмы данных, совершенствует нейросети, создаёт модели машинного обучения и умные алгоритмы. Благодаря этому компании улучшают качество своих продуктов и услуг, оптимизируют расходы и создают прорывные технологии.

Что же такое эти большие данные? Это могут быть тысячи чеков из огромного супермаркета, по которым можно изучить потребительский спрос, сотни рентгеновских снимков, на основе которых компьютер обнаруживает болезни, метеоданные для прогноза погоды, а ещё клиентские базы, видео, аудиофайлы, веб-аналитика и многое другое.

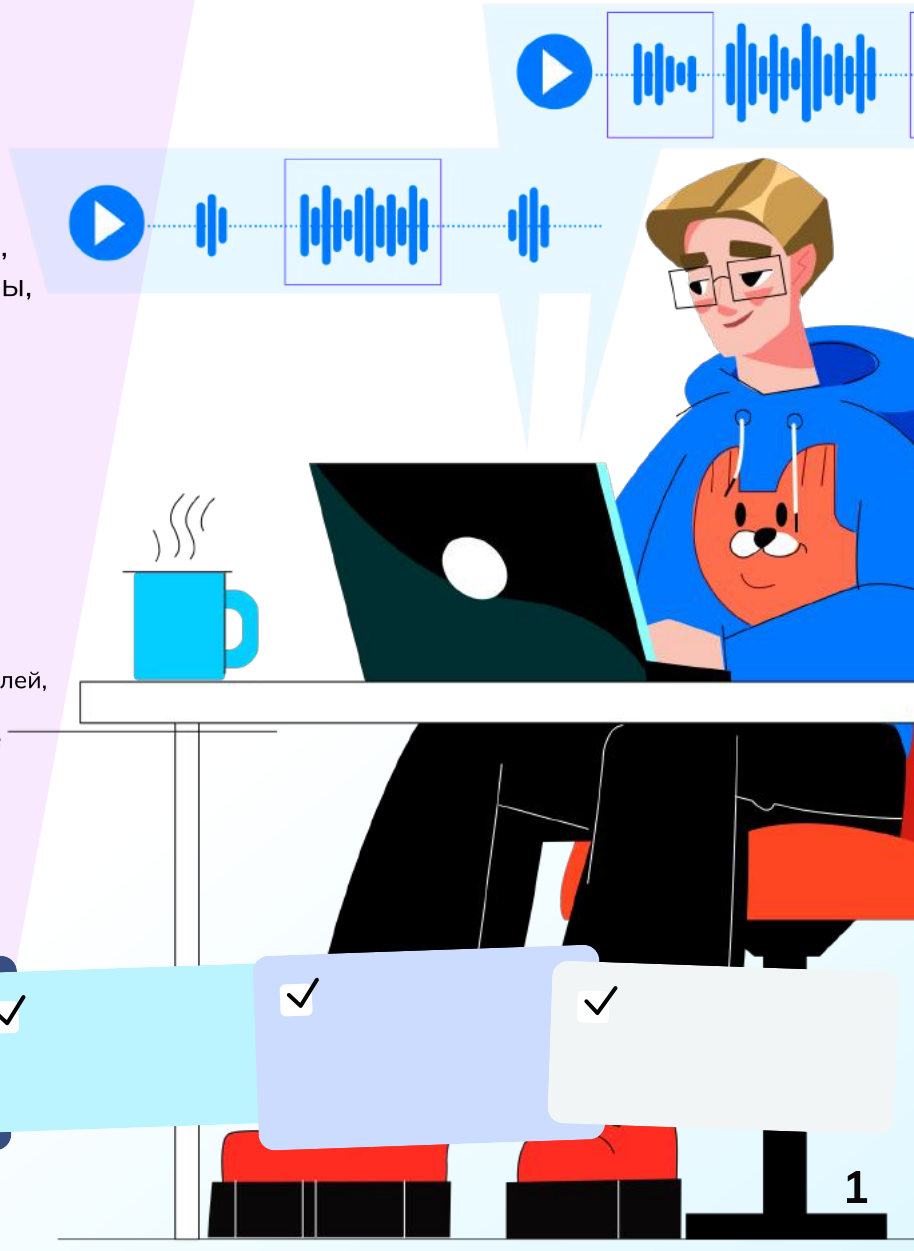
Задача дата-саентиста — так поработать с данными, чтобы на их основании сделать прогнозы или решить конкретную задачу, которая принесёт пользу людям.

Задачи

1. Сбор, подготовка и разметка подходящих для задачи данных
2. Сбор, подготовка и разметка подходящих для задачи данных
3. Анализ данных для выявления закономерностей
4. Создание алгоритмов, способных анализировать огромные объёмы информации
5. Создание моделей, делающих статистические прогнозы

Вам будет интересна эта профессия, если вы...

- ✓ Давно освоили компьютер и даже пробуете себя в простых алгоритмах
- ✓ Быстро разбираетесь в математических формулах
- ✓ Хотите получить современную профессию с достойной зарплатой
- ✓ Усидчивы, аккуратны и внимательны
- ✓
- ✓
- ✓



Задание

Часто для решения сложных задач или даже создания новых IT-продуктов используют искусственный интеллект. Но он появляется не сам по себе: обучением машин занимаются специалисты по Data Science.
Именно такую роль вы сегодня и примерите на себя.

Ваша основная задача: разработать модель машинного обучения для перевода аудиоконтента в текст.

Вы разберётесь, откуда вообще появляются алгоритмы и нейросети, на конкретной задаче обучите модель распознавания слов и проверите, насколько эффективно она работает.

А ещё, несмотря на техническую сложность, вам может пригодиться знание русского языка — ведь надо будет указать алгоритму на его грамматические ошибки.

Вперёд, будущие айтишники!

Этапы

1. Определить цель машинного обучения
2. Собрать и подготовить данные
3. Обучить модель машинного обучения в несколько этапов
4. Проверить работу алгоритма с помощью метрик

Часть 1

Определение задач

Специалистам по Data Science важно сразу чётко определить, какую технологическую задачу им предстоит решать.

Попробуйте определить, что из предложенного будет задачей машинного обучения, а что — нет?

Подсказка: задача машинного обучения — это тип прогноза или вывода о какой-то проблеме на основе данных.

Задачи бизнеса обычно направлены на оптимизацию процессов в компании, увеличение выручки и количества пользователей, создание новых продуктов и функциональных возможностей.

I

Задачи машинного обучения

II

Бизнес-задачи

1. Распознавать данные в аудиодорожке и переводить её в текст.
2. Создать технологию, чтобы люди в любой ситуации могли воспринимать видео- и голосовые сообщения.
3. Увеличивать ежедневное количество пользователей чата на 15%.
4. Дать возможность слабовидящим пользователям проходить «капчу» в виде сгенерированной аудиодорожки.
5. Предсказывать наличие знаков препинания в расшифрованных аудиосообщениях.
6. Поднять оценку показателя «Удобный чат» на 10%.

Часть 1

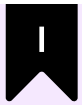
Определение задач

Специалистам по Data Science важно сразу чётко определить, какую технологическую задачу им предстоит решать.

Попробуйте определить, что из предложенного будет задачей машинного обучения, а что — нет?

Подсказка: задача машинного обучения — это тип прогноза или вывода о какой-то проблеме на основе данных.

Задачи бизнеса обычно направлены на оптимизацию процессов в компании, увеличение выручки и количества пользователей, создание новых продуктов и функциональных возможностей.



Задачи машинного обучения

- ✓ Распознавать данные в аудиодорожке и переводить её в текст
- ✓ Предсказывать наличие знаков препинания в расшифрованных аудиосообщениях

Продуктовые задачи

- ✓ Создать технологию, чтобы люди в любой ситуации могли воспринимать видео- и голосовые сообщения
- ✓ Увеличить ежедневное количество пользователей чата на 15%
- ✓ Дать возможность слабовидящим пользователям проходить «капчу» в виде сгенерированной аудиодорожки
- ✓ Поднять оценку показателя «Удобный чат» на 10%

Часть 2

Сбор и подготовка данных

Цель поставлена: необходимо разработать алгоритм для перевода аудиосообщений в текст. Но как его обучить?

Самое важное для машинного обучения — данные. Данных нужно много, они должны быть разнообразны и качественно отобраны.

В среднем человек за одну секунду говорит два слова. Если сообщение из пяти слов длится 20 секунд, будем считать такую дорожку некачественной, не подходящей для обучения модели.

И если на аудиодорожке много пустых мест без «всплесков», она также не подходит для обучения.

Исключите 3 записи, которые не подойдут для машинного обучения.

Запись 1

Пойдем вечером в кино. Там вышел новый классный фильм. 9 слов

Запись 2

Всем привет. Очень нужны ваши сердечки. Кхе. 6 слов

Запись 3

Спасибо большое. Очень рад, что вам понравилась моя статья. 9 слов

Запись 4

А кто в кино собрался? Я не понял. 8 слов

Запись 5

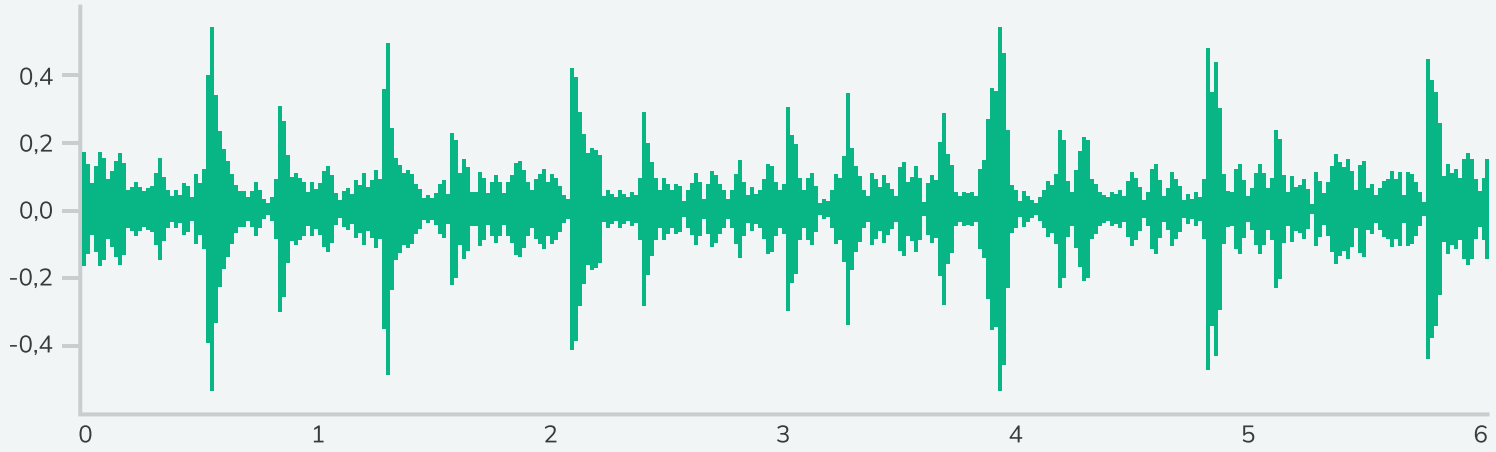
Ого-о-о! 1 слово

Запись 6

Да.... Нет... Угу. 3 слова

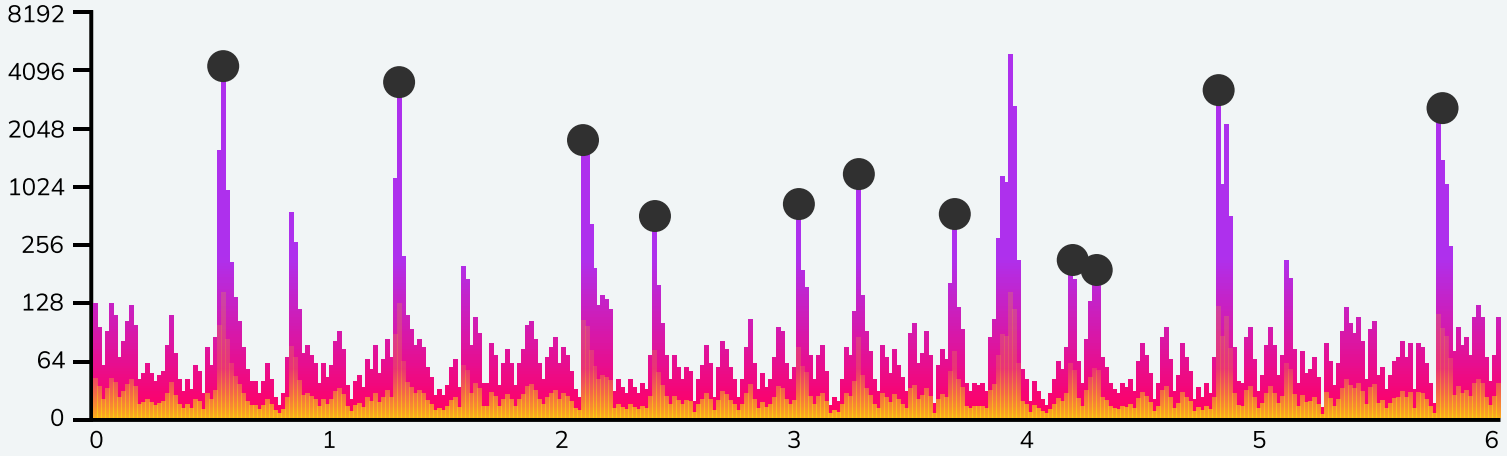
Найдите все пиковые значения на представленной записи и выделите их.

Осциллограмма



Все ли отметки пиковых частот есть на спектрограмме?
Проставьте недостающие, чтобы не потерять никаких данных.

Спектрограмма



Часть 3

Акустическая модель

Основная часть автоматического распознавания речи — работа с акустической моделью. Данные проходят через неё и преобразуются в символы.

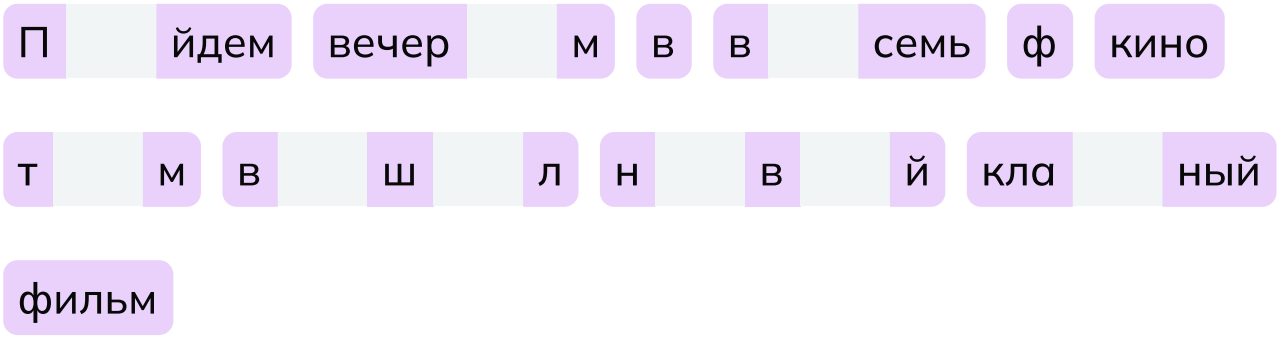
Главная задача модели на этом этапе: определить наиболее вероятный звук в отдельный момент времени.

Установите, какие звуки распознал алгоритм.

Будьте внимательны: нужно отметить именно фонему, то есть «как слышится», а не «как пишется».

Расшифровка звуков:

Определите пропущенные звуки по звуковой дорожке:



Часть 3

Акустическая модель

Основная часть автоматического распознавания речи — работа с акустической моделью. Данные проходят через неё и преобразуются в символы.

Главная задача модели на этом этапе: определить наиболее вероятный звук в отдельный момент времени.

Установите, какие звуки распознал алгоритм.

Будьте внимательны: нужно отметить именно фонему, то есть «как слышится», а не «как пишется».

Расшифровка звуков:

Определите пропущенные звуки по звуковой дорожке:

П а йдем вечер а м в в о семь ф кино

т а м в ы ш э л н о в ы й кла с ный

фильм

Лингвистическая модель

Найдите и исправьте все ошибки в написании. Ниже напишите всю фразу правильно.

Пайдем вечерам в восемь ф кино там вышэл новый класный фильм

Пунктуационная модель

Поставьте галочку там, где должен быть какой-то знак препинания, включая точку.

Пример 1

Прошло ☐ жаркое ☐ лето ☐ и ☐ наступила ☐ осень ☐

Пример 2

Здравствуйте ☐ Иван ☐ Федорович ☐

Пример 3

Летом ☐ ребята ☐ приносили ☐ разные ☐ цветы ☐ лютики ☐ ромашки ☐ васильки ☐ колокольчики ☐

Пример 4

Фу ☐ какая ☐ мерзость ☐

Общее количество знаков 11

1. **Распознать звуки в словах и написать их в пропусках для акустической модели.**
2. **Прописать уже верные буквы по правилам правописания для лингвистической модели.**
3. **Расставить знаки препинания в те ячейки, где они необходимы.**

Пример 1

Х_чу ☐ вам ☐ сказать ☐ что ☐ это ☐ м_я ☐ мечта ☐ играть ☐ на ☐ гитаре ☐ как ☐ профе_ионал ☐

Акустическая модель

Х	а	чу		вам		сказать		что		это		м	а	я		мечта				
играть		на		гитаре		как		профе	с	ионал										

Лингвистическая модель

Х	о	чу		вам		сказать	,	что		это		м	о	я		мечта	—			
играть		на		гитаре		как		профе	сс	ионал	.									

Пример 2

Ф_нетика ☐ это ☐ раздел ☐ науки ☐ о ☐ _зыке ☐ к_торый ☐ изучает ☐ звуки ☐ речи ☐

Акустическая модель

Ф	а	нетика		это		раздел		науки		о		и	зыке		к	а	торый		изучает	
	звуки		речи																	

Лингвистическая и пунктуационная модели:

Ф	о	нетика	—	это		раздел		науки		о		я	зыке	,	к	о	торый		изучает	
	звуки		речи	.																

Пример 3

Как ▢ _сегда ▢ мой с_се_ ▢ пр_спал ▢ и ▢ _п_здал ▢ е_о ▢ не ▢ пустили ▢ в ▢ кла_ ▢

Акустическая модель

Как ▢ ф сегда ▢ мой ▢ с а се т ▢ пр а спал ▢ и ▢ а п а здал ▢

е в о ▢ не ▢ пустили ▢ в ▢ кла с ▢

Лингвистическая и пунктуационная модели:

Как ▢ в сегда , мой ▢ с о се д ▢ пр о спал ▢ и ▢ о п о здал ,

е г о ▢ не ▢ пустили ▢ в ▢ кла сс .

Пример 4

П_года ▢ ре_ко ▢ уху_шилась ▢ и ▢ потемнело ▢ и ▢ сне_ ▢ пош_л ▢ и ▢ ветер ▢ усилился ▢

Акустическая модель

П а года ▢ ре с ко ▢ уху т шилась ▢ и ▢ потемнело ▢ и ▢ сне к

▢ пош о л ▢ и ▢ ветер ▢ усилился ▢

Лингвистическая и пунктуационная модели:

П о года ▢ ре з ко ▢ уху д шилась : и ▢ потемнело , и ▢ сне г

▢ пош ё л , и ▢ ветер ▢ усилился .

Пример 5

Он ▢ х_р_шо ▢ знает ▢ ру_кий ▢ _зык ▢ и ▢ это ▢ его ▢ пр_имущество ▢

Акустическая модель

Он ▢ х а р а шо ▢ знает ▢ ру с кий ▢ и зык ▢ и ▢ это ▢ его ▢ пр и имущество ▢

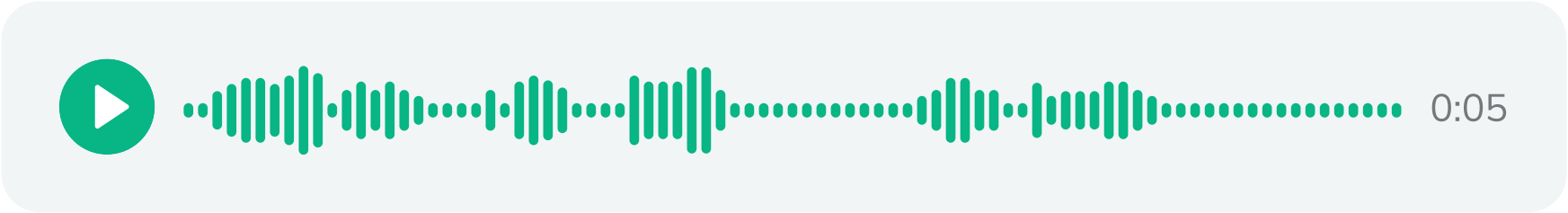
Лингвистическая и пунктуационная модели:

Он ▢ х о р о шо ▢ знает ▢ ру сс кий ▢ я зык , и ▢ это ▢ его ▢ пр е имущество .

Пунктуационная модель

Обведите или заштрихуйте паузы в аудиодорожке.

Дорожка



Классно быть дата саентистом! Что думаешь?

Оценка результата

Разделите фразу на минимальные цепочки, чтобы алгоритму было проще расшифровать их. Обведите или заштрихуйте необходимые участки.

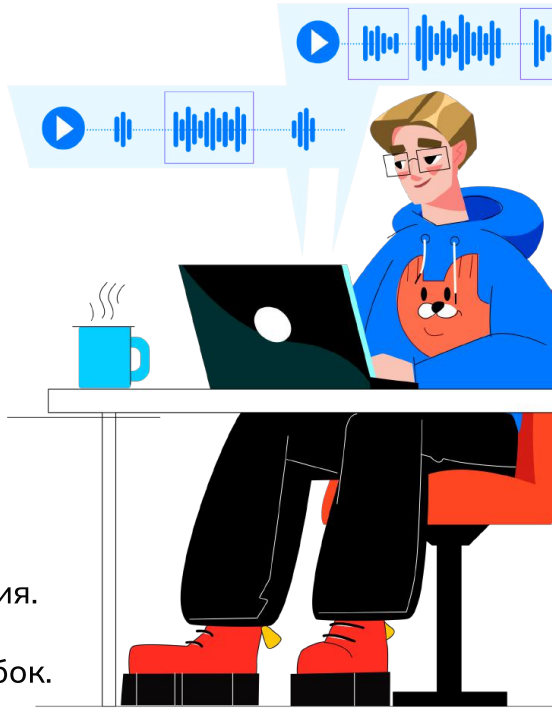
Прости | что | долго | не | отвечала

Отлично, задание выполнено!

Поздравляем! Теперь любое голосовое сообщение или видео можно превратить в разборчивый текст.

Сегодня вы:

- ✓ Познакомились с одной из задач специалиста Data Science.
- ✓ Отличили частоты на осциллограмме от частот на спектрограмме.
- ✓ Поработали с несколькими моделями машинного обучения.
- ✓ Обучили алгоритм и убедились, что он работает без ошибок.



Мне понравилось! Что дальше? Советы от эксперта

Иван Самсонов
Data Science Lead, VK

1. Участвуйте в хакатонах и чемпионатах по программированию! Это даст колоссальный опыт и знакомства.
2. Развивайте насмотренность. То есть наблюдайте, как другие решают какую-то задачу, какую технологию и где применяют. Это очень важный навык.
3. Если вы чувствуете, что хотите в ИТ, но пока не определились, начните делать свой проект. Пробуйте разные роли: дизайнера, разработчика, менеджера и пр.
4. В интернете много лекций и инструкций, много уже написанных программ! Так вы разберётесь во многих нюансах!