

FAQ по общим вопросам специализации “Машинное обучение и анализ данных” от Яндекса и МФТИ на Coursera

Специализация:

<https://www.coursera.org/specializations/machine-learning-data-analysis>

telegram chat специализации: <https://t.me/datasciencecourse>

Составитель FAQ: <http://t.me/rogozinushka>

P.S. Этот FAQ создан не авторами курса, так что все прочитанное вами принимайте на свой страх и риск, так как это субъективное мнение. Если есть вопросы или есть что поправить, пишите в [телеграм](#), я не кусаюсь. И на вопросы отвечаю с удовольствием. Критику приветствую.

О документе: Мне иногда пишут в личку по каким-нибудь общим вопросам: какие зарплаты, что на собеседах надо знать, как тебе специализация, какие скиллы нужны для работы, где учить нейронки, нужно ли знать матан и т.п. Постепенно стал замечать, что даю иногда одни и те же ответы. Так и возникла идея о гугл доке, в котором бы я собрал все подобные вопросы с ответами в одну кучку.

Этот документ больше предназначен для начинающих дата сайентистов и тех, кто думает о том, чтобы начать свой путь в машинном обучении. Но для этого нужно знать матан и Питон) Их можно параллельно изучать на первом курсе. Зачем? [Вот ответ](#).

Думаю, вам пригодится этот материал. Приятного чтения!



Оглавление

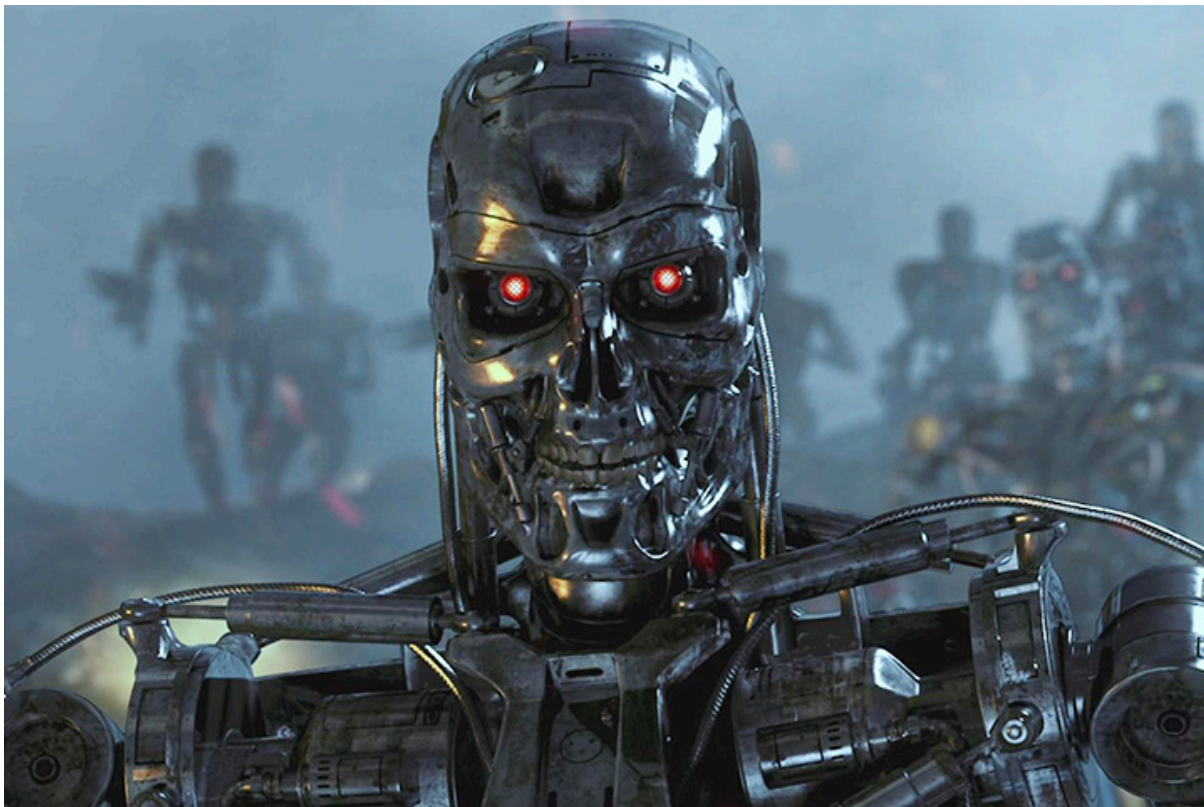
Да кто этот такой ваш дата сайнс	4
Искусственный интеллект захватит мир. Или нет?	4
Дата сайнс или машинное обучение?	6
О пороге вхождения в дата сайнс, или снимите розовые очки	6
Зачем надо знать Python для дата сайнса в целом	6
Зачем надо знать матан для дата сайнса в целом	7
Самый важный навык для работы (и вообще в жизни)	8
Open Data Science - крупнейшее русскоязычное сообщество дата сайнсистов	9
Вопросы по специализации	10
Не получается оплатить курс / курсы пропали с Курсеры	10
График обучения	10
Какие знания может дать специализация	10
Мнение о каждом курсе специализации	10
Нейронкам будут обучать?	11
Почему в видеоматериалах специализации используют Python 2, а в интернете чаще всего говорят про Python 3	11
Как решить самую сложную задачу специализации - сходство строк с кошками	11
Что делать с устаревшими заданиями, или танцы с бубнами	11
Где искать помощи	12
Что такое материальная помощь	12
Не успеваю вовремя сделать домашние и тесты: меня накажут?	13
Как записаться на следующий курс, если написано "Свободных мест нет"?	13
Дополнительные навыки	14
Почему не стоит учиться на geekbrains, skillfactory, skillbox и подобных	14
Хорошие курсы по ML	14
Курсы по нейронным сетям для новичков	15
Где учить Python	16
Надо ли знать SQL	16
Где учить SQL	16
Где учить матан	16
Где учить статистику	17
Где учить Git	18
Где учить алгоритмы (и зачем они нужны дата сайнсистам)	18
Какие хард скиллы ещё были бы полезны	19
Список литературы по машинному обучению и нейронным сетям	19
Как найти работу и что для неё надо	21
После специализации меня устроят на работу?	21
Почему с джунами на рынке сейчас не ок?	21
Короткий список советов	21
Хватит ли специализации для собеседования?	22
Какие вопросы могут быть на собеседовании?	22

Как проходить собеседование	22
Где найти стажировку или работу	23
Какие бывают стажировки	24
А что если я не смогу ответить на вопрос на собеседовании?	24
Как говорить о денежном вопросе на собеседовании	24
Советы по тестовым заданиям	25
Примеры тестовых заданий	25
Мысли о работе	26
Какая профессия в мире дата сайнс тебе подходит больше	26
Почему не стоит быть дата ремесленником	27
Зачем знать матан и теорию моделей, если на работе им не пользуются	28
Что представляет собой дата сайнс на работе	29
Сколько платят дата сайнтистам	30
Сколько платят аналитикам данных	31
Сколько платят на N-ой должности	31

1. Да кто этот такой ваш дата сайнс

Искусственный интеллект захватит мир. Или нет?

В новостных источниках можно часто услышать “Инженеры в X компании сделали Y фичу с помощью искусственного интеллекта”, где под фичей может быть беспилотный автомобиль, собачки из Boston Dynamics, и т.п. Так что же скрывается под вашим ИИ? Почему он еще не захватил наш мир и не поработил человечество, как показывают в фильмах???



Что же такое ИИ простыми словами? ИИ - это свойство компьютерных систем выполнять задачи, которые может сделать обычный человек. К примеру, задачей может быть

- классификация животных на изображении: коты это или собаки
- распознавание знакомых лиц среди лиц в толпе
- распознавание текста на изображении
- создание сценариев, рассказов и т.п.
- и т.п.

То есть некоторая задача, которую сложно описать условиями if-else, так как эта функция творческая, не очень определенная. И тут на помощь приходит искусственный интеллект, который может “творчески” найти решение той или иной задачи. ~~Творчески уничтожить человечество.~~

ИИ может делать разнообразные вещи, но не все сразу. Тут мы и подходим к тому, что задачи бывают как простые, так и сложные. Потому ИИ разделяют на два типа:

1. Слабый искусственный интеллект
2. Сильный искусственный интеллект

Слабый искусственный интеллект решает конкретную узкую задачу. К примеру, это может быть:

- распознавание голоса в текст
- отнесение каждой новости к определенной категории (политика, спорт, развлечения и т.п.),
- определение похожих товаров в магазинах, чтобы показать, определение оттока клиентов (если он потенциально откажется от услуг, его будет дешевле удержать, чем искать нового клиента),
- предсказание ипотечной ставки в зависимости от характеристик клиента и т.п.

Как раз в новостях и фигурирует слабый интеллект, который распространен в IT-компаниях. Машине нужен контроль, так как слабый ИИ не признает своих ошибок, поэтому роль дата сайентистов так и важна. Слабый ИИ банально не понимает, что он ошибается. А так как ИИ сейчас выполняет только одну конкретную функцию, которой его обучили люди, то захватить он нас не сможет. ~~Если только человек его этому не научит.~~



А какой же тогда ваш сильный ИИ. Смотрели “Терминатора 2”? Да кто его не смотрел) Так вот, Арнольд Шварценеггер в роли Т-800 как раз хороший пример. Иными словами, компьютер может мыслить, принимать решения в условиях неопределенности, обучаться, делать осмысленные решения и многое что ещё. Кроме этого он может выполнять разнообразные творческие функции:

- спасти Сару Коннор из психлечебницы
- защитить Джона от Т-1000
- стрелять так, чтобы не убивать полицейских и т.п.

Эти функции заранее нигде не описаны, Т-800 старался сам найти решение всех этих проблем. К примеру, как настоящий человек он может обучаться. Чего стоит сцена, когда Джон учит улыбаться Т-800.



В данный момент небольшое количество компаний занимается сильным ИИ, так как это очень сложная область, требующая огромных вычислительных ресурсов и сложных алгоритмов. Поэтому в данный момент чаще всего ограничиваются слабым искусственным интеллектом, который хорошо может решать конкретные узкие задачи. Это проще и дешевле.

Дата сайнс или машинное обучение?

Если про дата сайнс, ИИ, машинное обучение вы слышите впервые, то для вас есть хорошая статья на человеческом языке, а не на языке машин - https://vas3k.ru/blog/machine_learning/

[Data science](#) - общее понятие, которое включает в себя все знания, связанные с анализом данных и извлечения из них какой-то полезной информации и предсказанием каких-то выводов.

[Анализ данных](#) - методы обработки данных для получения из них полезной информации. Иными словами, просто копаемся в данных, пытаемся найти вручную закономерности, не прибегая к машинному обучению.

[Machine learning](#) - модели, которые можно обучить предсказывать неизвестные данные на основе известных данных. Сюда входит классический ML (линейные модели, случайный лес, градиентный бустинг и др) и глубокие нейронные сети.

Держите хорошую статью, которая вполне доступно говорит, с чем едят дата сайнс - <https://ain.ua/special/data-science/>

О пороге вхождения в дата сайнс, или снимите розовые очки

Увы, но не получится стать высокооплачиваемым специалистом с зарплатой 300к/сек, изучив всего лишь один курс. Красивые вывески онлайн курсов лгут: как и для любой работы, необходимо обладать большим объемом знаний. Для дата сайнса надо ботать и усиленно) Одним курсом не обойдешься.

Мир дата сайнса начал формироваться в России относительно недавно, из-за чего только сейчас начали появляться [достойные курсы](#) на русском языке. Но всех их объединяет одно: порог вхождения. С техническим образованием будет полегче, но если нет оно, то попотеть придется знатно, так как требуются знания во многих областях: [матан](#), [программирование](#) и [желание](#) искать [дополнительные материалы](#). И тут всё зависит от вас: потянете ли вы этот объем знаний и будет ли вам интересно это изучать или нет.

Из-за нехватки знаний иногда задания и кажутся сложными, но это только на первый взгляд. Сейчас смотрю на некоторые задания и диву даюсь, что потратил на их решение столько времени. Со временем набираются опыт и знания, так что дерзайте)

Если есть желание погрузиться в дата сайнс вместе с умением разбираться в теме не только с материалами из курса, то перед вами все двери открыты.

Зачем надо знать Python для дата сайнса в целом

Ответ прост: так исторически сложилось) Делать дата сайнс можно на Java, C#, R и других языках, но стандартом в последнее время служит Python. Почему? Этот язык достаточно прост для изучения: легкий синтаксис и популярные библиотеки позволяют прямо из коробки написать красивый и понятный код. Из-за этого порог вхождения у этого языка прост и менее болезнен по сравнению с C++, к примеру.



**Python
developer
learning C++**



**C++
Developer
learning
Python**

Дата сайентисты, если можно так выразиться, не чистые программисты: они больше аналитики и математики. Поэтому и требований к написания кода у них меньше: простой синтаксис и следующая за ним высокая продуктивность при написании кода, уже готовые встроенные функции в типах, крутые библиотеки, низкий порог вхождения, динамическая типизация и многое другое. Python скрывает низкоуровневую логику, за счет чего он более дружелюбен к новичкам.

Подвох есть лишь в производительности. За красивую обертку, синтаксический сахар и интерпретатор мы расплачиваемся не такой быстрой работой кода. Поэтому некоторые Python библиотеки вызывают внутри себя функции из C/C++ модулей, которые работают быстрее, чем те же самые функции на Python. Самая популярная

библиотека по работе с векторами и матрицами numpy внутри себя содержит Сишный код)

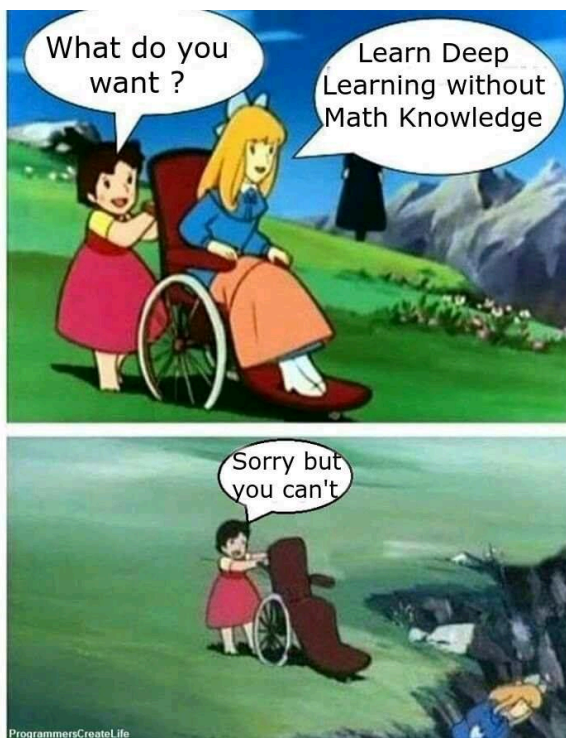
Но это не так страшно, так как плюсов Python-а для дата сайентистов больше: дружелюбное комьюнити, простой синтаксис для написания кода, готовые библиотеки для решения многих задач.

Если у вас программированием на Python есть проблемы, то [соответствующий ответ](#) вам в помощь.

Зачем надо знать матан для дата сайнса в целом

Да, надо. Без него на курсе будет туговато. Очень. Матан является основой как классического машинного обучения, так и нейронных сетей.

Хоть какая-то база по нему должна быть. Поэтому стоит его изучить отдельно или параллельно с прохождением первого курса. Просто так не отвернуться от него, даже если хочется.



Если хочется матан пройти понятнее и доступнее, то [соответствующий ответ](#) в помощь.

Самый важный навык для работы (и вообще в жизни)

Дата сайнс не дается так просто: много чего надо знать и понимать. Корочки курсов по ML и смежным темам хороши, чтобы на linkedin их приложить, чтобы профиль стал более интересным, но не более. Если учить ради них, на собеседованиях и на работе это вскроется. Так что поговорим о том, зачем вообще надо учиться и как.

Сами сертификаты не особо стоят без навыков: все же вы учите материал для себя. Это даже лучше, так как будет формироваться навык искать информацию вовне специализации на сторонних источниках. После чего это войдет в привычку и человек будет понимать, что сделать, если он что-то недопонял и где это искать.

И главным фактором успеха является самостоятельный поиск тех или иных материалов для закрепления и углубления материала. То бишь правильно “юзать гугл”) На самом деле даже не сам гугл важен, как само желание разобраться самостоятельно в той или иной проблеме.

К примеру, на 1-2 курсах специализации рассказали о спрямляющем пространстве в теории: мол, можно увеличить размерность исходного пространства, применить на нем линейную модель, чтобы в исходном пространстве предсказанная линия была нелинейной. Ты такой думаешь, да как так может быть, хочу проверить. Делаешь ноутбук, в котором создаешь новые фичи, пробуешь линейную модель на новых полиномиальных фичах запустить (читаешь справку, как это сделать), отображаешь истинную и предсказанную линии на графике и удивляешься, что это работает.

Зачем это надо? Потому что за тебя никто делать работу не будет и разжевывать материал тоже. Надо уметь и самому разбираться в тех или иных вещах. А если не получается, просить помощи в чате специализации. Большинство ребят помогают от доброты душевной и не надо их за это винить: и в основном это одни и те же лица)

К примеру, если человек попросил помощи в задании, а ему скинули теорию с формулой, по которой это можно сделать, а он взял и применил сам и разобрался в ней - это хороший подход, респект и мое уважение. Если же человек начал возмущаться, что ему не дали готовый ответ - это плохой подход, который показывает, что человек хочет всё готовенькое, а думать сам не шибко хочет.

Open Data Science - крупнейшее русскоязычное сообщество дата сайентистов

Советую вступить в [ODS](#) - Open Data Science. Это крупнейшее русскоязычное сообщество дата сайентистов. Контингент весьма разнообразный: от джунов до сеньоров, от дата сайентистов до дата инженеров, от России до стран СНГ. Очень дружелюбное и опытное сообщество.



Open Data Science

Международное сообщество, объединяющее специалистов, исследователей и инженеров, связанных с Data Science

- 🔥 Вместе делаем крутые проекты, мероприятия и курсы
- 👉 Делимся опытом, прокачивая себя и друг друга
- 📈 Продвигаем открытый Data Science и развиваем область

Несколько полезных каналов для новичков в ODS:

- #jobs - список вакансий
- #career - канал про карьеру. Выкладывают интересные статьи, общаются на вечные темы и т.п.
- #welcome - новички и опытные знакомятся друг с другом
- в #edu_courses советуют курсы, либо спрашивают, какие можно изучить
- в #ods_resume_mastering советуют, как делать нормальные резюме
- и многое другое

2. Вопросы по специализации

Не получается оплатить курс / курсы пропали с Курсеры

Информация актуальна на 05.03.2022.

Вслед за платформой онлайн-обучения edX из России ушла аналогичная платформа – Coursera:

- с нее удалены курсы всех российских образовательных учреждений;
- россияне больше не смогут оплатить курсы на платформе

Что делать:

- изучать DS и смежные навыки на Степике, более подробно можно найти в разделе [Дополнительные навыки](#)
- зайти на Курсеру под ВПН и попытаться оплатить / подать мат помощь
- записаться на курс с Университета 2035

График обучения

О графике обучения есть хороший FAQ от создателей курса. Для интересующихся прошу пройти по ссылке - <https://datasciencecourse.ru/faq>

Какие знания может дать специализация

В этой специализации дают неплохую базу классического ML и смежных тем: матан, Питон, работа с numpy, pandas, sklearn, обработка данных, обучение на размеченных данных, обучение без учителя, статистика, собственный проект в портфолио.

В начале поведают две фундаментальные вещи, без которых не обойтись: [матан](#) и [Python](#). Чуть позже расскажут о том, как модели машинного обучения работают под капотом, какие им надо давать данные и как в них искать закономерности. Познакомят со статистикой и то, как можно проверять гипотезы на данных с помощью статистики и котиков 🐱. А на последнем курсе дадут сделать свой проект в портфолио из 4 предложенных. Он позволит на практике вспомнить всё то, что было на последних 5 курсах - и это круто!

В целом специализация даст базу, при наличии которой можно начать путешествие в мире данных и изучать с пониманием более сложные вещи уже самому.

Мнение о каждом курсе специализации

Если пройти первые три курса специализации, то можно смело идти на собеседование на джуна или стажера дата сайентиста. На большую часть вопросов по машинному обучению ответить сможете. О том, какие вопросы будут ещё на собеседовании, читайте ниже в категории [“Собеседование”](#).

На 4 курсе изучается проверка гипотез, статистика, АВ тесты. Для дата сайентистов он может быть полезен, так как в некоторых вакансиях эти знания приветствуются. Да и для общего развития тоже не помешает)

5 курс - самый противоречивый. Если теория про временные ряды и текстовую обработку полезна и доступна, то по компьютерному зрению проседает:

рассказывают все в общих чертах, явно недостаточно для такой обширной темы. У этого курса самая низкая оценка из всех курсов специализации.

6 курс. Лакомка. Курс представляет собой четыре практических проекта на выбор в портфолио. Начинайте с анализа текста или оттока: они самые понятные и доступные. Остальные два проекта требуют намного большего погружения в теорию и свободное время.

Нейронкам будут обучать?

Специализация посвящена классическому ML: обработка данных, линейные модели, случайный лес и т.п. На нейронные сети идет слабый упор, за их изучением лучше идти на [соответствующие курсы](#) по нейронным сетям.

Почему в видеоматериалах специализации используют Python 2, а в интернете чаще всего говорят про Python 3

Python 3 более актуальный, так что лучше качать именно его. В специализации Python 2 используется, так как специализация немного старенькая) Отличий между версиями не шибко много, так что поправить скрипт на втором Питоне в третий и наоборот не составит труда. Вот, к примеру, статья с различиями этих версий - <https://pythonworld.ru/osnovy/python2-vs-python3-razlichiya-sintaksisa.html>

Как решить самую сложную задачу специализации - сходство строк с кошками

Задача про кошачьи строчки из первого курса, по моему мнению, самая сложная из всех. Неподготовленному не особо понятно, что от тебя хотят в этой задаче: банально опыта не хватает. Эта задача похожа на некоторый фильтр: если ты его преодолел, то значит, что есть порох в пороховницах и дальше материал ты сможешь понять. Сейчас эта задача кажется детским садом: решить ее можно в два счета. Далее задания будут более-менее понятны, но попадались и совсем [устаревшие](#), которые без танцев с бубном не решить.

Если у вас есть проблемы с пониманием этой задачи и путями решения оной, то советую изучить эти материалы из курса “Практикум по математике и Python” - <https://stepik.org/lesson/48784/step/1?unit=30893>

Что делать с устаревшими заданиями, или танцы с бубнами

В курсах есть моменты, которые ставят в ступор. К примеру, устаревшее задание по нейронным сетям, задача с bigartm и другие. Благо таких моментов мало и это не возведено за правило. На форуме курсов есть темы, в которых помогают разобраться в подобных заданиях)

В целом вся специализация сделана достойно и поражаешься труду людей, вложивших в это столько времени и усилий.

Где искать помощи

Не понятно, как решить задачу с первого раза? Попытки безуспешны? Когда кажется, что сил больше нет, на помощь приходят они:

- форум курса на Coursera. Большинство проблем там уже описаны)
- google и stackoverflow. Умение находить быстро справку и распространенные ответы на вопросы - полезный навык)
- telegram chat - <https://t.me/datasciencecourse> . Активный чатик, где можно поговорить за жизнь или попросить помощи в своем задании
- <https://stepik.org/course/3356/promo> - Практикум по математике и Python на Stepik. Курс был создан по мотивам первого курса специализации, так что не удивляйтесь, если увидите знакомые задачи
- В крупнейшем сообществе русскоязычных датасайентистов [ODS](#) можно попробовать поискать советов в направлении, куда двигаться: какие пет-проекты можно сделать, какие фреймворки сейчас актуальны, какие курсы можно поизучать и т.п. И вообще слак ODS очень полезная вещь, так как в одном месте содержит всю русскоязычную тусовку датасайентистов, так что обязательно регистрируйтесь) К примеру, в #welcome новички и опытные знакомятся друг с другом, а #edu_courses советуют курсы, либо спрашивают, какие можно изучить, а в #ods_resume_mastering советуют, как делать нормальные резюме. И многое другое!

Что такое материальная помощь

Обзор > Наука о данных > Машинное обучение

Этот курс входит в специализацию "Специализация Глубокое обучение"

Нейронные сети и глубокое обучение

★★★★★ 4.9 Оценки: 109 115 • Рецензии: 21 751

Andrew Ng и еще 2 преподавателя [ЛУЧШИЕ ПРЕПОДАВАТЕЛИ](#)

Участвовать бесплатно
Начинается 30 сент. г.

[Доступна финансовая помощь](#)

997 226 уже зарегистрированы

Материальная помощь - это бесплатный доступ на 6 месяцев к курсу специализации, если с финансами проблемы. Для каждого курса (их шесть) подаётся свое заявление на материальную помощь. Заявление подается на английском или русском языке на странице курса: через две недели вам дают ответ. Если курс не получилось пройти за шесть месяцев, можно еще раз подать заявление.

Как правило, всем одобряют мат помощь: главное преодолеть лимит требуемых слов.

Самым удобным способом прохождения специализации является следующий сценарий:

1. Подаем заявление на один курс, начинаем проходить его
2. Как курс будет подходить к концу, подаем заявление на следующий
3. Завершаем этот курс и заодно получаем доступ к следующему: то есть пока курс проходим до конца, наше заявление на следующий уже рассматривают

Не успеваю вовремя сделать домашние и тесты: меня накажут?

Если несколько дедлайнов было пропущено, вскоре появится кнопка "Сбросить сроки" на странице с заданиями. Выполненные задания не пропадут, ваши сроки сбросят и перенесут на более дальние числа. Никаких санкций не будет.

Как записаться на следующий курс, если написано "Свободных мест нет"?

Если вы записались на курс от университета, то при попытке записаться на следующий курс специализации может быть предупреждение "Свободных мест нет".

Обучение на размеченных данных

★★★★★ 4.8 2 487 оценок • 332 оценки

Сохранить на потом

Спонсор: Yandex

Об этом курсе

Обучение на размеченных данных или обучение с учителем – это наиболее распространенный класс задач машинного обучения. К нему относятся те задачи, где нужно научиться предсказывать некоторую величину для любого объекта, имея конечное число примеров. Это может быть предсказание уровня пробок на участке дороги, определение возраста пользователя по его действиям в интернете, предсказание цены. по

[Показать все](#)

от партнера



Свободных мест нет



Сертификат, ссылками на который можно делиться с другими людьми

Получите сертификат по завершении



100% онлайн

Начните сейчас и учитесь по собственному графику.

Чтобы появилась возможность записаться на курс, нажмите на стрелочку рядом со словом Coursera и переведите свой аккаунт в личный. Это даст вам возможность записаться на курс: подать на [матпомощь](#) или оплатить его.

coursera



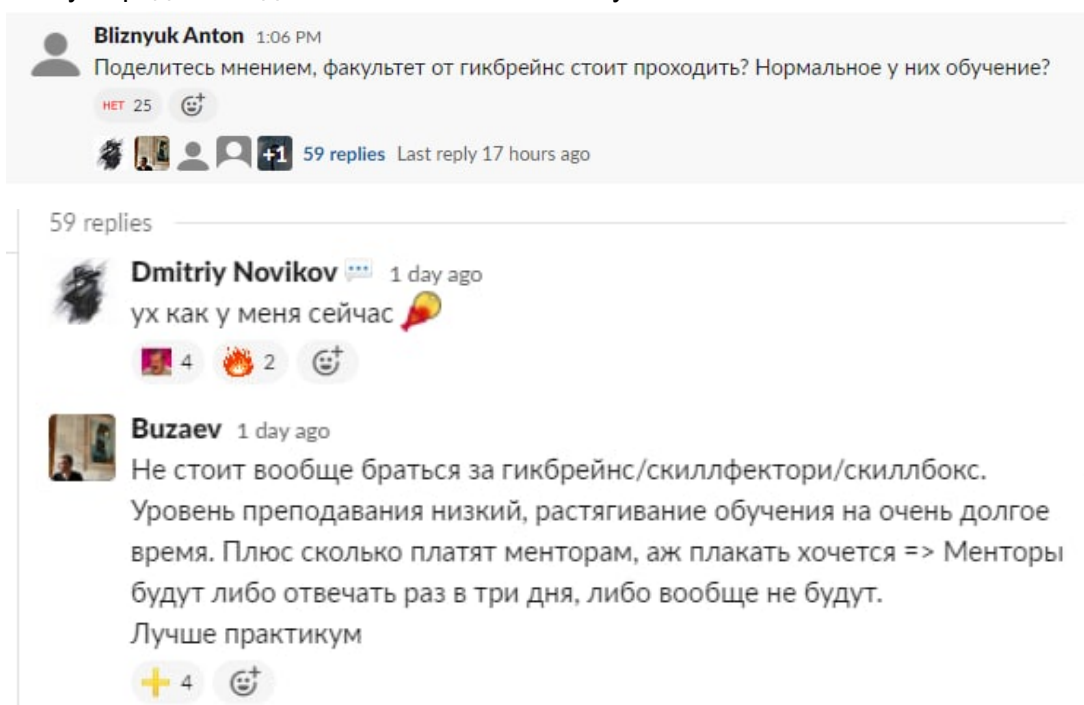
3. Дополнительные навыки

Почему не стоит учиться на geekbrains, skillfactory, skillbox и подобных

Дата сайнс сейчас хайпует: о нем трубят из всех щелей. Мол, пошли, товарищ без профильного образования, к нам учиться в X-ую онлайн школу, после чего будешь зарабатывать 300к/сек без регистрации и смс как в меме про Карлсона. Только для начала заплати за обучение)

Реальность оказывается коварнее. У большинства подобных курсов одна цель - срубить как можно больше бабла с нерадивых студентов. У них нет особой цели обучить, из-за чего больше всего страдают сами студенты.

Недавно в чате ODS в Slack-е задали вопрос о результативности подобных курсов, от чего у порядочных дата сайнстов бомбануло.



Где же тогда учиться? В [этом ответе](#) составил небольшой список неплохих курсов. Дерзайте)

Хорошие курсы по ML

Специализация от Яндекса и МФТИ вполне хорошая, но и другие курсы не отстают. В основном, курсы по ML рассказывают про одно и то же, разница лишь в углублении в материал. Вот несколько хороших курсов для начинающих (они бесплатные, а на Coursera можно [матпомощь](#) подать):

- Курс “Введение в Data Science и машинное обучение” от Института биоинформатики на Stepik - <https://stepik.org/course/4852/promo>
- Курс “Машинное обучение” от ОмГТУ на Stepik - <https://stepik.org/course/8057/promo>

- Open Machine Learning Course - открытый курс машинного обучения. Хороший курс, который на человеческом языке объясняет базовые вещи. Вот, к примеру, переводы на хабре - <https://habr.com/ru/company/ods/blog/322626/> , а вот гитхаб с ноутбуками с курса - <https://github.com/Yorko/mlcourse.ai>
- Яндекс Практикум хорош, вы не догадаетесь чем... практикой) URL - <https://praktikum.yandex.ru/data-scientist>
- Курсы от Эндрю Ына по Machine Learning на Coursera (на англ!) - <https://www.coursera.org/learn/machine-learning>
- Блог Александра Дьяконова "Анализ малых данных" - <https://alexanderdyakonov.wordpress.com/>

Курсы по нейронным сетям для новичков

Понятные курсы по нейронным сетям для новичков:

- курс "Нейронные сети и компьютерное зрение" от Samsung на Stepik. URL - <https://stepik.org/course/50352>
- курс "Нейронные сети и обработка текста" от Samsung на Stepik. URL - <https://stepik.org/course/54098>
- Специализация "Deep Learning" от Эндрю Ына на Coursera (на англ). URL - <https://www.coursera.org/specializations/deep-learning>

Далее можно и более сложные проходить.



ЭНДРЮ ЫН



Где учить Python

Есть несколько хороших курсов для изучения Python на Stepik:

- <https://stepik.org/course/58852/promo> - "Поколение Python": курс для начинающих. Курс позиционируется для школьников (не смотрите, что он для школьников, для взрослых он тоже может подойти). Простой и понятный материал, большое количество практики
- <https://stepik.org/course/67> - Программирование на Python. Хороший курс для новичков. Этот курс будет немного скучнее, чем первый. Я бы сказал он более академический
- <https://stepik.org/course/3356/promo> - Практикум по математике и Python на Stepik. Курс был создан по мотивам первого курса специализации, так что не удивляйтесь, если увидите знакомые задачи
- <https://www.coursera.org/specializations/python#courses> - Python для всех на Coursera (англ)

Кроме этого может быть полезным самостоятельный поиск в гугле разных мелочей Python-а: какие типы данных можно поставить в качестве ключа в словарь, что такое изменяемый и неизменяемый тип данных, чем отличается list от tuple, зачем надо знать о деерсору и многое другое.

По Python рекомендую прочитать хороший список вопросов по этому языку с собеседований - <https://m.habr.com/ru/company/mailru/blog/506824>

Надо ли знать SQL

SQL должен знать любой уважающий программист.

Для дата сайентистов он особенно ценен теорией: как объединять таблицы, как писать условия и группировать и т.п. Те же самые операции мы проводим с помощью pandas)

Плюс, если надо данные тащить из БД, то SQL будет кстати.

Где учить SQL

Есть несколько достойных ресурсов, где можно набить руку на sql-запросах:

- <https://stepik.org/course/63054/promo> - интерактивный тренажер по SQL. Мне нравится этот тренажёр на степике: сразу на задачах усвоите материал. В этом тренажере теории немного, что круто, зато самого полезного - практики - предостаточно.
- <https://sql-ex.ru/> - хоть и старый сайт, но ещё пользующийся популярностью для обучения SQL
- <https://sqlbolt.com/> - интерактивный тренажер SQL на английском языке

Где учить матан

Достаточно стандартного курса по матану в универе. Если матан не ваша стезя или вы его немного подзабыли, то помогут следующие источники:

- <http://mathprofi.ru/> - а именно раздел с первым курсом (мастхев, в университете помогал мне и всем моим одногруппникам)

- <https://stepik.org/course/3356/promo> - Практикум по математике и Python на Stepik. Курс был создан по мотивам первого курса специализации, так что не удивляйтесь, если увидите знакомые задачи

- <https://mml-book.github.io/> - Mathematics for Machine Learning (на англ.)
- <http://vmls-book.stanford.edu/> - Introduction to Applied Linear Algebra – Vectors, Matrices, and Least Squares (на англ.)

Лучше учить матан параллельно со специализацией, чтобы на практике видеть те или иные его закономерности. Если бы меня так матану в универе учили бы, я его любил бы - а так нам говорили лишь теорию без применения её на практике. Это специализация как раз поможет матан применить на практике. Больше осознанности видишь в этом предмете)

Где учить статистику



Самая простая и понятная книга, которая объясняет статистику для не технарей (и технарям она тоже зайдет) - это манга “Занимательная статистика”. Очень наглядное введение в статистику. Каждую тему разбирают на каком-то реальном примере, что круто. Плюс можно посмотреть на нарисованных тян и кунов)

Если вы со статистикой на ВЫ, то могу порекомендовать несколько хороших книжек, которые в доступной форме могут рассказать о тех или иных аспектах статистики человеческим языком. Они будут хороши и для повторения статистики, если вы её когда-то учили в вузе или самостоятельно:



- Манга “Занимательная статистика”. Для повторения материала подойдёт) И для тех, кто ни разу не занимался статистикой

- Савельев Владимир. Статистика и коты. То же хорошая книжка для не технарей, но посложнее, чем манга, так как в этой книге описываются более сложная статистика. Небольшой минус - вместо формул нарисованы коты (серьезно), и это может запутать, если переходить к формулам, которые страшные порой. Если посмотреть формулу нормального распределения в котах и её реальную альтернативу, то они будут идентичны)

- Чарльз Уилан. Голая статистика. Нет формул, но вместо них много умных

мыслей, как статистика в нашей жизни нас постоянно обманывает: как нас обманывает

средняя зарплата по стране, почему Голливуд неправильно считает самые кассовые фильмы и т.п.

- Дарелл Хафф. Как лгать при помощи статистики

По курсам:

- Статистику неплохо рассказывают на 4-м курсе специализации Машинное обучение от Яндекса и МФТИ
- Ещё советуют курс Statistics with R на Coursera - <https://www.coursera.org/specializations/statistics>
- Курсы Основы статистики. Часть [1](#), [2](#), [3](#) на Степике по статистике от Института биоинформатики странные: в этой [статье](#) объясняется почему. С каждым курсом в отзывах встречаются оценки все ниже и ниже. Если хотите вкратце узнать, что с ними не так, держите выжимку в виде анекдота.

Диалог с одним студентом:

- А что тебе в курсе не понравилось?
- Ну там много ошибок, например...
- Но понятный курс же
- Ээээ, понятным языком объяснены неверные вещи
- Но понятно же!

Где учить Git

Хватит несколько первых уроков курса “Основы Git” на Степике. Остальное наберется с опытом. Ссылка на курс - <https://stepik.org/course/3145/promo>

Где учить алгоритмы (и зачем они нужны дата сайентистам)

У некоторых компаний есть фильтр для приема сотрудников под названием “Алгоритмическая секция”. Особенно это любят в Яндексе. Суть заключается в решении задачи на программирование в онлайн режиме в текстовом блокноте:

- <https://www.youtube.com/watch?v=0yxjWwoZtLw> - Как решать алгоритмические секции: помощь разработчикам, собеседующимся в Яндекс. Часть 1
- <https://www.youtube.com/watch?v=zU-LndSG5RE> - Как решать алгоритмические секции: помощь разработчикам, собеседующимся в Яндекс. Часть 2

Немного странно, так как знание алгоритмов, которые требуются, не особо нужны в дата сайнсе, имхо. Но если вы хотите попасть в компанию мечты, где это требуется, держите список годных материалов и курсов, на которых можно набить руку:

- <https://leetcode.com> - мастхев, если вы желаете пройти алгоритмическую секцию. Проходите задачи уровня easy-medium: такого же уровня будут и задачи на собеседовании
- https://t.me/coding_interviews - разбор задач с Leetcode на русском языке

- <https://stepik.org/course/217> - Алгоритмы: теория и практика. Методы от Computer Science Center (CS центр) на Stepik
- <https://stepik.org/course/1547> - Алгоритмы: теория и практика. Структуры данных от Computer Science Center (CS центр) на Stepik

Какие хард скиллы ещё были бы полезны

Из хард скиллов было бы неплохо знать:

- практика на kaggle, то есть свои решения, которые можно показать будущему работодателю

- python и работу с numpy, pandas, sklearn
- регулярные выражения
- умение работать с командной строкой linux

На работе эти знания пригодятся)

Также можете зарегистрироваться в крупнейшем сообществе русскоязычных датасайентистов [ODS](#). После в канале career задать вопросы про то, куда расти: какие пет-проекты можно сделать, какие фреймворки сейчас актуальны, какие курсы можно поизучать и т.п.

Список литературы по машинному обучению и нейронным сетям

Если вас заинтересовала какая-то тема (статистика, Питон и т.п.), то смело ищите книги в Интернете по этой теме. Не забывайте посмотреть отзывы: не все книги полезны и мнения других читателей помогут отсеять водянистые.

Анализ данных:

- Уэс Маккини: Python и анализ данных. Книга для начинающих. Смотрите на год издания: в нашем мире все быстро устаревают

Machine Learning:

- Кристофер Бишоп. Распознавание образов и машинное обучение (Pattern Recognition and Machine Learning). Считается библией Machine Learning-a. Может показаться немного скучноватой, так как содержит одну теорию. **Для чтения нужно хорошо знать матан и тервер!**
- Рашка Себастьян, Мирджалили Вахид. Python и машинное обучение. Машинное и глубокое обучение с использованием Python, scikit-learn и TensorFlow

Deep Learning:

- С. Николенко, А. Кадури, Е. Архангельская Глубокое обучение. Погружение в мир нейронных сетей
- Я. Гудфеллоу, И. Беджио, А. Курвилль. Глубокие сети
- Рашка Себастьян, Мирджалили Вахид. Python и машинное обучение. Машинное и глубокое обучение с использованием Python, scikit-learn и TensorFlow

Эти книги на английском и их рекомендует прочесть разработчики из Яндекса:

- С. М. Bishop, Neural Networks for Pattern Recognition,

- David MacKay, Information Theory, Pattern Recognition and Neural Networks,
- Trevor Hastie, Robert Tibshirani, Jerome Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction

4. Как найти работу и что для неё надо

После специализации меня устроят на работу?

100-процентной гарантии никто вам не даст. Все зависит от ваших навыков, понимания материала, напора дальше развиваться в дата сайнсе и желания найти работу. Такой вещи, как найти работу, и предназначен этот раздел)

Те, кто прошел специализацию, могут подать заявку на программу трудоустройства - <https://datasciencecourse.ru/work/> . Это не гарантия трудоустройства, ваше резюме лишь рассматривают в приоритетном порядке Яндекс и компании-партнеры. Увы, сейчас ссылка сломана (на 25.11.2020).

Процитирую текст из FAQ создателей курса (<https://datasciencecourse.ru/faq>):

- Чтобы принять участие в программе трудоустройства выпускников, вам необходимо завершить программу обучения и подтвердить ее успешное окончание сертификатом. Резюме наших выпускников регулярно просматривает Яндекс и другие наши партнеры в приоритетном режиме. Но, надо сказать, что большинство учащихся трудоустраивается самостоятельно еще в период обучения.

Почему с джунами на рынке сейчас не ок?

Из-за [разнообразных курсов](#) на рынке появились так называемые вайтишники, которые за несколько месяцев обучились дата сайнсу, но глубоко его не понимают, а хотят получать много денег и сразу. Из-за этого у компаний к джунам стали требования выше, чем в прошлые года, так как общий уровень новичков упал. К примеру, теперь компании стали требовать техническое образование и примеры проектов, в которых работал соискатель.

- Про техническое образование. Наверно, это условие стало для компаний значимым: что-то в духе первичного фильтра. Имею в виду, что сейчас из каждого утюга рекламируют трехмесячные курсы "никакого математического и технического бекграунда не нужно, вы сразу начнете получать 300к в сек". В итоге на рынок могут выходить соискатели с кашей в голове(

- Про проекты. Поэтому сейчас работодатель просит гитхаб, пет-проекты и другие доказательства ваших умений. Таким образом они отсеивают тех, кто хочет войти в IT, но не шибко-то и хочет развиваться и стараться.

Короткий список советов

- по возможности иметь пет проекты и свой гитхаб, чтобы выделиться среди других кандидатов. Для компании нанимать джуна рискованно, так как он может уйти из-за нагрузки или будет долго окупаться. Поэтому если вы себя покажите молодцом по сравнению с другими, у вас будет больше шанс устроиться на работу
- знать матан, питон и тот ML, на который вы идёте (классика, NLP, CV)
- полезно будет посмотреть требуемые знания на желаемых должностях и их прокачать

- у эйчаров на первом интервью поинтересоваться, о чем спрашивать будут и к чему подготовиться. Они не звери, чтобы вас валить
- написать всем крупным компаниям о том, что хотите идти именно к ним. Можно на LinkedIn, ODS, на летних стажировках, других источниках подать заявку
- спросить про возможность удалённой стажировки, если вы не в Мск и Питере

Хватит ли специализации для собеседования?

2-3 курсов достаточно, чтобы на собеседовании не тупить и показать, что некоторая база по машинному обучению у тебя есть. Остальные курсы больше про проверку гипотез с помощью статистики и закрепление материала. В целом полученных знаний хватит, чтобы говорить с коллегами на одном языке, постепенно погружаясь в то, как происходит анализ данных в крупной компании.

На данный момент знания одной теории иногда бывает недостаточно, поэтому ищите пет-проекты, решайте соревнования kaggle, чтобы у вас было больше преимуществ перед другими кандидатами на работу.

Как говорится, не одними знаниями по ML едины, смотрите, что еще требуют в ответе ниже.

Какие вопросы могут быть на собеседовании?

Меня спрашивали про опыт в разработке, что узнал в специализации от Яндекса, про тему магистерской диссертации, про метрики качества классификации и то, от чего зависит количество данных для хорошего качества. Иногда говорят, что спрашивают, чем отличаются L1 от L2 регуляризатора и почему первый обнуляет признаки. Иными словами, задают стандартные вопросы, которые позволяют понять, что ты базу ML понимаешь. Это основная часть вопросов, которая задается соискателю, поэтому на подготовку к ним надо приложить как можно больше усилий.

Для подготовки можете посмотреть статьи вопросами из собеседований из чата для начинающих дата сайентистов - https://t.me/start_ds

Могут также проверять знания по [матану](#) и [Python](#). Их знать надо. По Python рекомендую прочитать хороший список вопросов с собеседований - <https://m.habr.com/ru/company/mailru/blog/506824>

Кроме этого могут проверить и другие [смежные знания](#): статистику, SQL, git, linux и другое. Понимать их желательно, но не обязательно. Если вы их знаете, то это будет плюсом в вашу карму, если нет, то не страшно. Хотя всё зависит от компании, куда вы собираетесь пойти работать, поэтому лучше уточнить список необходимых компетенций в тексте вакансии.

Как проходить собеседование

Несколько полезных источников:

- <https://yandex.ru/jobs/ya-interview/#mldev> - хорошая статья, что обычно требуется от кандидата на собеседовании. Алгоритмы, в основном, требуют в Яндексе, в некоторых других компаниях на них нет такого упора.
- https://www.youtube.com/watch?v=C8v3UduKTW&feature=emb_logo - большое видео о стажировке в Яндексе. Похожие стажировки есть и в других компаниях)

- <https://www.youtube.com/watch?v=3vPNReMUzBE> - IT собеседование. Как понять, что компания Вам не подходит? Видео больше предназначено для уже прошаренных дата сайентистов (мидлов и сеньоров), но и для начинающих есть хорошие мысли

Где найти стажировку или работу

Следует гуглить компании, где ты хочешь работать и которые связаны с твоими предпочтениями: размер компании, похожая предметная область (медицина, банки, продажи и т.п.), как-то коррелирует с обучением в универе, хороший ДМС или зарплата и т.п.

Составив такой список, можно уже начать писать эйчарам и откликаться на резюме. Есть несколько способов найти работу для начинающего специалиста. Здесь я укажу общие способы получения работы:

- Выложить резюме на hh.ru. Но для этого надо уметь писать резюме. Можно перечислить основные вещи, которые должны быть в резюме: свои контактные данные, фото, где учился, где работал, какие навыки есть (можно поискать ключевые слова в похожих вакансиях), какие проекты делал и всё на странице A4. Краткость - сестра таланта. Можно поискать в интернете, как грамотно составить резюме
- Зарегистрироваться в крупнейшем сообществе русскоязычных датасайентистов [ODS](#). После в канале _jobs поискать вакансии или в каналах welcome и career рассказать о себе, чтобы тебя направили на путь истинный
- Зарегистрироваться в соц сети для работников <https://www.linkedin.com/> (заходите через VPN, в России сайт заблокировали), заполнить профиль, набрать в контакты дата сайентистов и рекрутеров вашего города, заодно попробовав писать рекрутерам напрямую, есть ли у них вакансии.
- Искать вакансии на профильных ресурсах: [hh.ru](#), <https://career.habr.com/vacancies?type=suitable>, <https://geekjob.ru>, <https://t.me/foranalysts>, <https://github.com/solarrust/junior-friendly> и прочих
- Иногда бывают специальные странички на сайтах компаниях, где можно скинуть свое резюме со всеми навыками, а эйчар по ключевым словам сам подберет, есть ли такие вакансии или нет. К примеру, <https://kontur.ru/career/vacancies/send>, <https://tochka.com/hr/> и другие
- Иногда вакансии не выкладывают на сайте, но эйчар знает о них, можно у него их выпросить)
- Узнать об открытии стажировок в крупных компаниях: Тинькофф, Яндекс, Сбербанк и т.п. Некоторые компании круглый год проводят стажировки: к примеру, Яндекс. Через успешную стажировку вполне можно получить оффер
- Получить работу через успешное прохождение школ какого-то направления при компаниях: к примеру, <https://fintech.tinkoff.ru/study/start/>
- Попасть в дата сайнс через [внутреннюю стажировку](#), о которой будет ниже. Вкратце звучит так: устраиваетесь на похожую работу (программист, аналитик данных) в компанию мечты, набираетесь опыта, после чего пробуете постучаться в “соседнюю дверь”, то есть пойти стажироваться дата сайентистом. Плюсы такого подхода описаны ниже

Какие бывают стажировки

Есть два вида стажировки:

- обычная стажировка. Когда молодого специалиста берут без опыта или кардинально изменившим свою специализацию. Перед этим он должен пройти отбор: тестовые задания, собеседования, код на листочке и т.п. С одной стороны, такой фильтр позволяет избежать потенциально незнающих людей, с другой стороны, этим могут злоупотребить
- внутренняя стажировка. Ты уже работаешь внутри компании и хочешь попробовать себя в новой команде. Ты знаешь процессы, у тебя уже есть опыт в IT: можно сказать, что ты стучишься в соседнюю дверь. Перейти в другую команду легче, так как тебя принимают за своего, могут даже какие-то поблажки делать. Даже если крупные компании говорят о том, что у них нет каких-то особых предвзятых отношений к людям извне и внутри, все равно это неявным образом видно.

По моему мнению, стажировка внутри компании проще)

А что если я не смогу ответить на вопрос на собеседовании?

Тут момент в том, сколько из заданных вопросов ты не знаешь. Интервьюеры не звери же, чтобы вас валить)

Если же не получается объяснить грамотно ответить на вопрос, можно попробовать своими словами, даже на пальцах: интервьюеру станет понятно, что вы понимаете, о чем говорите)

Как говорить о денежном вопросе на собеседовании

Говорить о размере зарплаты в начале собеседования - дурной тон. Это показывает, что вам важнее всего зарплата, а не задачи. Мол, если вам работа не интересна, то вы её качественно выполнять не будете. Можно сказать, что для эйчара это будет звоночек)

Так как тогда говорить о зарплате? Не спорю, деньги тоже важны, но всё же я предпочитаю, чтобы кроме хорошей зарплаты были и условия хорошие: коллектив, плюшки от компании, интересные задачи. Если вам работа будет не нравится, то будет ли смысл говорить о зарплате? В самом начале можно обсудить часть из этих вопросов, а вопрос о зарплате оставить на конец.

Для начинающих говорить о зарплате в начале собеседования не очень хорошо, а мидлы и сеньоры это уже могут: если хорошего специалиста не будет устраивать зарплата, то зачем тратить время на интервью?

Можно поэкспериментировать в этом [калькуляторе](#) и самим, чтобы узнать среднюю зарплату IT-шника и зарплату должности, на какую вы хотите попасть, в вашем городе) Этой зарплатой вы можете оперировать в разговоре на собеседовании, так как она подкреплена фактами. Вы же её не из воздуха взяли)

Советы по тестовым заданиям

Обычно в тестовом задании для джуна дата сайентиста требуется взять размеченные данные и найти к ним подход: как их правильно обработать и как найти для них хорошую модель. Либо ему дают задачки на парсинг сайтов, алгоритмы для проверки опыта в программировании. Но в этом ответе поговорим только о машинном обучении.

Базовые советы по работе с тестовыми заданиями:

- EDA, то есть показать ноутбуки, в которых вы экспериментировали с данными и моделями (их может быть несколько). В этих ноутбуках не забывайте показывать качество, которое получалось, и делать свои выводы. Даже если качество вышло не очень, укажите это и попробуйте подумать, почему так случилось. Это покажет, что вы старались и подходили к проблеме с разных сторон
- красивое форматирование по рер8 и краткие объяснения подходов в markdown ячеек тоже приветствуется
- гуглить библиотеки и подходы для вашей задачи. Это покажет ваше напористость, что вы не только свои знания используете, но можете и в теме постараться разобраться
- предметную область изучить, построить на основе этого гипотезы, как работать с данными и с моделями
- показать конечный ноутбук или ru скрипт, который дает самое лучшее качество
- ноутбуки пронумеровать, чтобы было понятно, в каком направлении проверяющему смотреть ваши задания
- кросс-валидация с грид серчем - хороший подход для проверки качества модели

Примеры тестовых заданий

<https://github.com/slgero/testovoe> - примеры тестовых заданий в разные компании. Посмотрите и решите парочку, узнайте свои силы

5. Мысли о работе

Какая профессия в мире дата сайнс тебе подходит больше

Все зависит от бэкграунда и желания развиваться, я считаю:

- Под бэкграундом я подразумеваю ваш опыт работы, даже не IT-шный. Это может быть опыт сисадмином, программистом, продавцом и т.п.
- Желание развиваться. Даже если опыта нет никакого, но есть невероятный энтузиазм копаться в новом, то это даже в плюс.

Сочетание этих двух характеристик и поможет вам определить, куда вас больше тянет. Может быть ситуация, что опыт в IT есть, поэтому изучить ML будет проще, подпинывать себя получится реже. А может быть ситуация, что опыта нет никакого, прежняя работа достала и вы в пьяном угаре с бешеным желанием развиваться начинаете изучать дата сайнс, чтобы получить новую работу. Всё зависит только от вас)

Для удобства составил табличку, которая **утрировано** показывает, кто из вас может получиться в зависимости от вашего опыта. Плюсы в табличке нужны для сравнения должностей: на какой из должностей на что идёт упор по сравнению с соседними.

P.S. Но это не отменяет, что вы должны становиться тем, кто указан в таблице. Вовсе нет. Если вам нравится изучать дата сайнс, но у вас опыт только в управлении, то продолжайте. Самое же главное заниматься тем, что нравится лично вам)

P.P.S. Должностей, без которых дата сайнсистам тяжело и которые вносят большой вклад, больше: это и DevOps-ы, и аналитики, и инженер-программисты, и многие другие. В табличку я не стал их писать, так как они выходят за рамки этой специализации.

	Machine Learning	Анализ данных	Программирование	Опыт в управлении	Софт скиллы
Дата сайентист	+++	++	+		+
Дата инженер		+	+++		+
Аналитик данных	+	+++	+		+
Тимлид дата сайентистов	++	++	+	+	++
Менеджер разработки	в общих чертах понимать надо			+++	+++

Немного поясню все должности, упомянутые выше:

- У дата-сайентиста задача состоит в том, чтобы из имеющихся данных (которые ещё надо проанализировать и обработать) построить некоторую модель, которая будет давать предсказания, полезные для бизнеса. Для этого ему надо обязательно знать про машинное обучение, анализ данных и на среднем уровне знать программирование.

- Аналитик данных же отвечает за анализ данных: проверка гипотез, АБ тестирование и т.п., чтобы оценить риски при внедрении той или иной фичи в продакшн. Из-за этого аналитик данных лучше работает с анализом и визуализацией данных, чем дата сайентист.

- Дата-инженер обеспечивает надежное хранение и получение информации из разных источников, чтобы дата-сайентисты и аналитики данных имели к ним доступ. Поэтому дата-инженер очень хорош в программирование, разработке систем хранения.

- Тимлид - это человек, который на 50% инженер, а на другие 50% руководитель, который отвечает за формирование команды и назначение задач на коллег. На него накладываются управленческие задачи и теперь уделять все время задачам он и не может, поэтому инженерный уровень со временем у него немного падает. К примеру, он не берет крупных и горящих инженерных задач, так как его в любой момент может вызвать руководство на ковер, иначе задача будет гореть и все будут страдать, когда этой задачей мог бы заняться мидл или сеньор.

- Менеджер разработки отвечает за общение с заказчиками, стратегическое развитие проекта, бюджетирование, построение рабочего процесса в команде и т.п. Он не инженер, он чистый руководитель.

Несколько хороших статей, которые поясняют, чем отличаются эти должности:

- <https://rb.ru/opinion/data-scientist/> - Кто такой дата-сайентист и как им стать
- <https://habr.com/ru/company/yoomoney/blog/504930/> - Дата-инженеры в бизнесе: кто они и чем занимаются?
- <https://academy.yandex.ru/posts/chem-zanimayutsya-analitiki-dannykh-i-kak-nachat-rabotat-v-etoy-oblasti> - Чем занимаются аналитики данных и как начать работать в этой области?
- <https://medium.com/pochemumuzachemkak/pochemu-data-sayntist-i-data-inzhenер-эт-о-не-одно-и-то-же-90c78fbc94a4> - ПОЧЕМУ ДАТА САЙНТИСТ И ДАТА ИНЖЕНЕР — ЭТО НЕ ОДНО И ТО ЖЕ?
- <https://rb.ru/story/data-scientist-or-data-analyst/> - Дата-сайентист или дата-аналитик: какая из этих профессий вам больше подходит?
- https://m.habr.com/ru/company/habr_career/blog/508398/ - Три дороги для программиста: эксперт, руководитель, основатель

Почему не стоит быть дата ремесленником

Типичная работа дата сайентиста - использование готовых библиотек. Из-за этого могут и попадать ремесленники, которые делают задачу копипастом гайд серча из stackoverflow, а шаг вправо или влево для улучшения качества не могут сделать из-за недостатка знаний. Всё же дата сайентист хотя бы на базовом уровне должен знать матан и то, что под капотом моделей находится: [и вот ответ почему](#). Иначе такому специалисту грош цена.

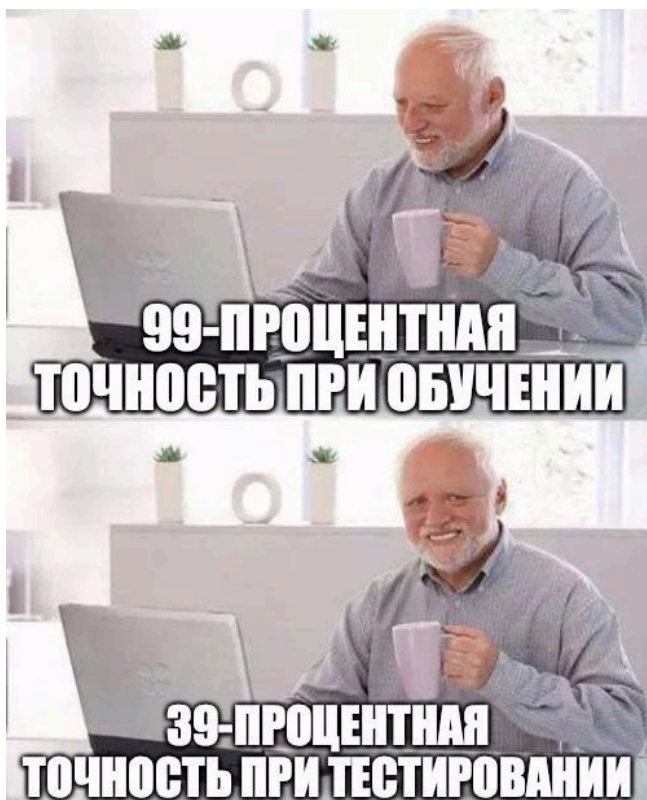
КОГДА ДАННЫЕ ИМЕЮТ ТАБЛИЧНЫЙ ВИД



Зачем знать матан и теорию моделей, если на работе им не пользуются

В специализации мы изучали чуть ли не то, как свой градиентный бустинг или логистическую регрессию замутить. На работе же больше готовыми библиотеками будешь пользоваться, либо пилить свои. Эти знания хороши как бэкграунд, чтобы понимать, что под капотом тех или иных моделей.

Допустим, есть горе-самоучка, который за месяц “освоил” дата сайнс. Он может запилить грид серч, взять оттуда лучшую модель по качеству и всё. А вот шаг влево, шаг вправо: он теряется и не знает что делать.



А если есть бэкграунд, то есть некоторая ясность, почему не работает и что можно поправить. К примеру, модель работает не так, как нужно, а ты знаешь, как она работает, какие данные требует и т.д. Благодаря этим знаниям ты можешь ещё раз проанализировать и исправить данные, нужные гиперпараметры модели выставить и тому подобное. И это круто, так как без этого бэкграунда дата ремесленник бы сломался и не смог ничего сделать.

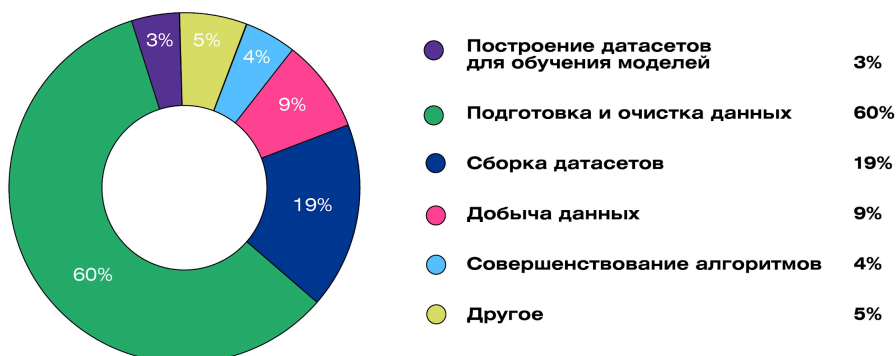
Что представляет собой дата сайнс на работе

Это тебе не простой анализ данных в небольшом ноутбуке, который можно потом залить в гитхаб.

На работе все иначе: скачиваешь сырые данные, погружаешься в предметную область, чтобы разобраться в том, как из всего этого непотребства в данных сделать конфетку, то есть хорошие данные. После экспериментируешь с моделями, признаками и выбираешь лучшую по теоретическим и практическим соображениям. В конце пишешь красивенький пайплайн, который позволит на проде без особых проблем получить нужные предсказания на сырых данных.

Примерное распределение времени дата сайентистов можете посмотреть на картинке ниже, взятой из <https://habr.com/ru/company/skillfactory/blog/529112/>

На что дата-сайентисты тратят больше всего времени



В общих чертах дата сайнс можно описать так, но у каждой компании своя специфика: у кого-то больше упор на нейронные сети, у кого-то на обработку данных и т.п. И поэтому круговая диаграмма выше может приобрести иные проценты.



Сколько платят дата сайентистам

Зарплаты можно посмотреть тут:

- [калькулятор зарплат Хабр Карьеры](#)
- вилки вакансий из [слага ODS](#)

Сколько платят на N-ой должности

Можно поэкспериментировать в этом [калькуляторе](#) и самим, чтобы узнать среднюю зарплату IT-шника и зарплату должности, на какую вы хотите попасть, в вашем городе)