

Модуль 1. Введение в Data Science

План:

1. Организационные моменты. Обзор курса.
2. Экскурс в машинное и глубокое обучение.
3. Стандартные задачи Data Science.
4. Подход к решению задач машинного обучения.
5. Инструментарий Data Science
6. Python / Anaconda / Jupyter Notebook / Google Collab.
Настройка рабочего пространства для последующей работы.

1. Организационные Моменты.

Цели занятия:

- Познакомимся с историей развития машинного обучения
- Узнаем об основных типах задач машинного обучения
- Разберем общий подход к решению задач

2. Экскурс в ML и DL.





From left: Yann LeCun, Geoffrey Hinton and Yoshua Bengio

Все развивается супер быстро

Поэтому рекомендую следить за новостями мира машинного обучения:

- Yann LeCun - <https://twitter.com/ylecun>
- Geoffrey Hinton - <https://twitter.com/geoffreyhinton>
- Ian Goodfellow - https://twitter.com/goodfellow_ian
- Andrej Karpathy -
<https://www.youtube.com/@AndrejKarpathy>
- Yannik Kilcher - <https://www.youtube.com/@YannickKilcher>
- Sam Altman - <https://twitter.com/sama>

Как все начиналось

Как всё начиналось:

1940
- 1950



- Создание первых примитивных компьютеров
- Заявление о направлении “Искусственный Интеллект”
- Алан Тьюринг придумал тест Тьюринга
- Первая программа которая училась играть в шашки
- **Фрэнк Розенблattt создаёт первую модель Персептрана**

Как всё начиналось:

1960 - 1970



- Появление первых алгоритмов машинного обучения (Nearest Neighbors and modifications)
- Первая многослойная нейронная сеть. Метод обратного распространения ошибки
- Разделение Machine Learning и Artificial Intelligence как самостоятельных направлений

Как всё начиналось:

1980-1990



- Первая “Зима ИИ”, стагнация и появления огромного количества систем, основанных на правилах
- Появление рекуррентных нейронных сетей (RNN)
- Появление модели SVM
- Появление бустинга (boosting)
- Конец 90-х бум и появление нейронных сетей для работы со звуком и изображениями

Как всё начиналось:

1990-2000



- Вторая зима, затишье
- Усовершенствование алгоритмов бустинга
- Появление и распространение интернета и компьютеров
- Выход в свет алгоритмов под названием “Машина Больцмана”
- Формирование направления “Обучение с подкреплением (Reinforcement Learning, RL)”

Как всё начиналось:

2000-2010

- 
- С распространением интернета и компьютеров происходит рост данных и больше возможности для обработки
 - Ренессанс нейронных сетей: обработка изображений, видео, звуков, речи, текстов и прочих данных
 - Победа нейросети в игре GO (Google DeepMind Competition)
 - Развитие технологий DeepFake
 - Появление и распространение облачных вычислений

Как всё начиналось:

2010-н.в.



- Усиленное развитие направления Deep Learning
- Появление архитектуры **Transformer (Attention is all you need)**
- Развитие **квантовых** вычислений
- Распространение алгоритмов машинного обучения в IT индустрии и производстве (Copilot, GANs, GPT-3, Bert, ChatGPT etc.)
- Появление self-driving машин, самообучающихся систем, роботов
- Развитие true AI (AGI - artificial general intelligence)
- Распространение больших языковых и мультимодальных

Интересные примеры:

- <https://thispersondoesnotexist.com/>
- <https://talktotransformer.com/>
- <http://www.r2d3.us/visual-intro-to-machine-learning-part-1/>
- <https://www.youtube.com/watch?v=S3F1vZYpH8c>
- <https://www.youtube.com/watch?v=7atZfX85nd4>

Почему этим стоит заниматься

Updated Sep 15, 2022

How much does a Data Scientist make?

Experience

All years of Experience

Industry

All industries

\$126,468 /yr

Total Pay

\$103,825 /yr

Base Pay

\$22,643 /yr

Additional Pay



\$126,468 /yr

\$98K

\$164K

\$79K

\$206K

■ Most Likely Range ■ Possible Range

Total Pay Trajectory

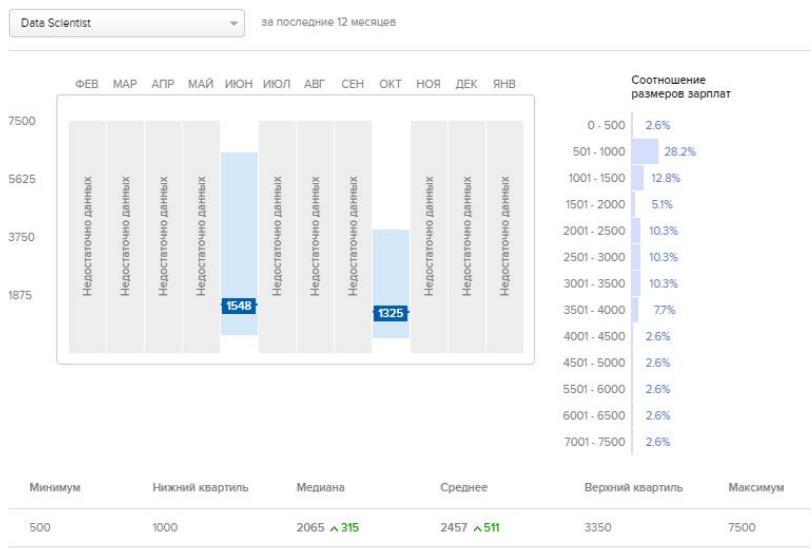
For Data Scientist

- \$126,468 /yr
Data Scientist
- \$165,596 /yr
Senior Data Scientist
- \$163,549 /yr
Lead Data Scientist

[See Full Career Path >](#)

[Download as data table](#)

Почему этим стоит заниматься



▼ Сравнение с данными за предыдущий период

<https://salaries.dev.by/>

hh Помощь

data scientist

Вакансии Резюме Компании

759 вакансий «data scientist»

hh Помощь

Machine learning

Вакансии Резюме Компании

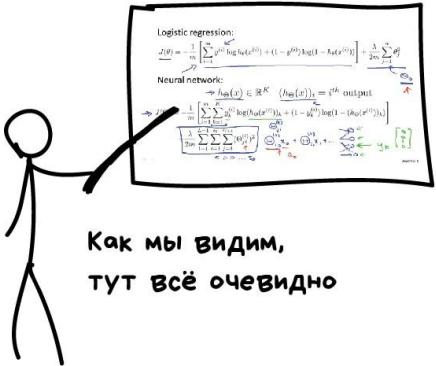
1 426 вакансий «Machine learning»

<https://hh.ru/search/>

Плюсы и минусы

- Мультидисциплинарность
- Востребованность
- Разнообразие задач
- Огромная индустрия
- Самообучение. Очень много самообучения...
- Скорость исследований
- Результативность
- Очень быстрое развитие (необходимо быстро бежать ради того, чтобы просто оставаться на месте)

Плюсы и минусы



Программисты программируют!

Датасенс!

Профессия будущего!

Буквально через пять лет...

Экспоненциально!!!

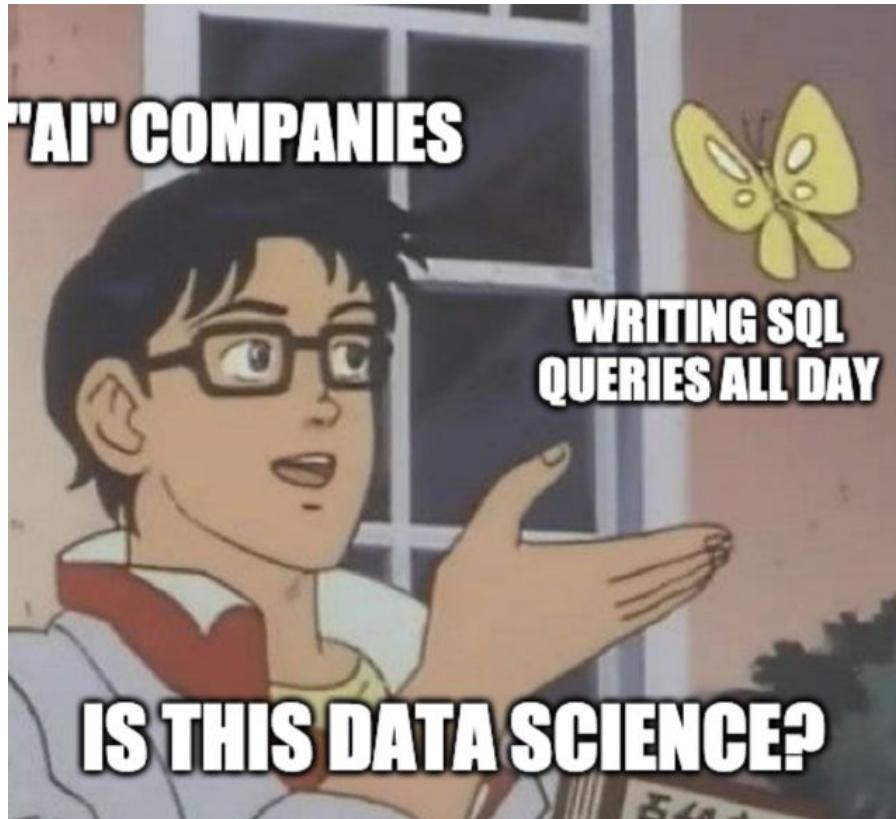
УМНЫЕ РОБОТЫ!

А-А-А-А-А-А-А-А-А-А-ааа!!!!!!

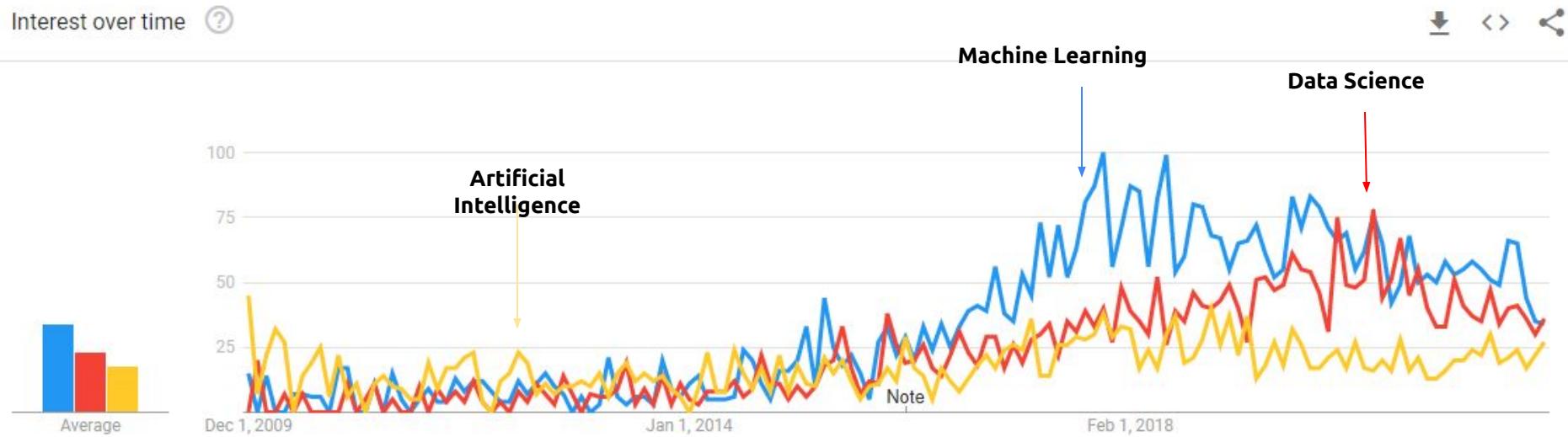


Есть два типа статей про машинное обучение

Плюсы и минусы



Почему всё стало так популярно:

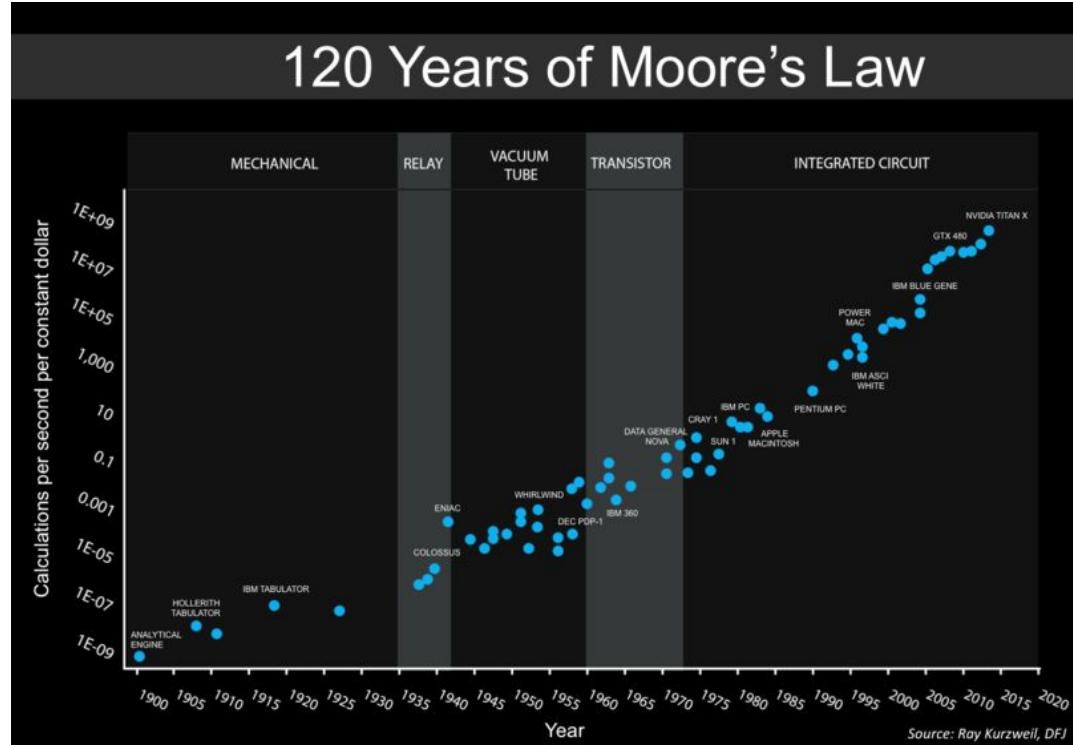


Источник: <https://trends.google.com/trends/>

Скорость вычислений

Закон Мура (1965) - эмпирическое наблюдение, изначально сделанное Гордоном Муром (IBM), согласно которому количество транзисторов, размещаемых на кристалле интеграционной схемы будет удваиваться каждые 24 месяца.

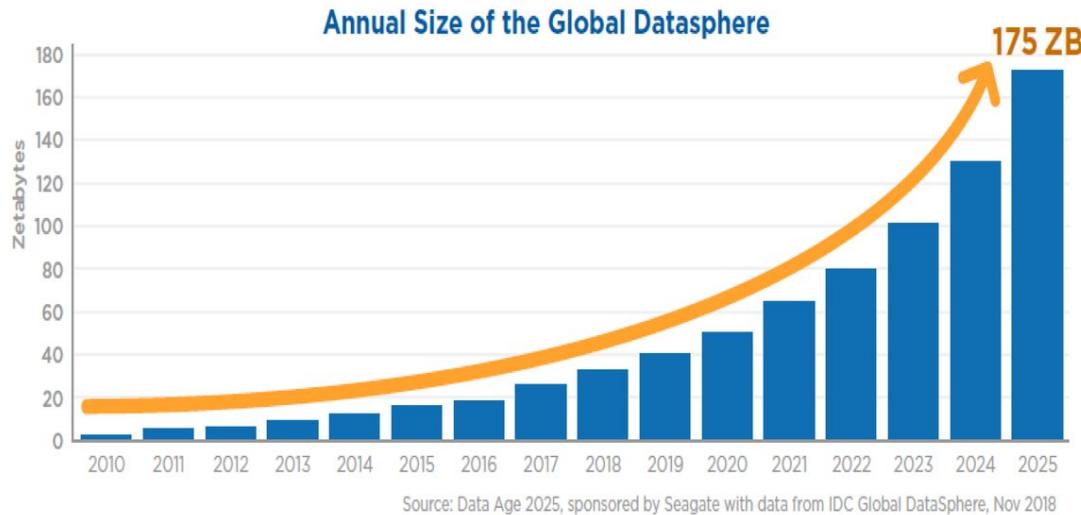
Часто цитируемый интервал 18 месяцев связан с прогнозом Дэвида Хауса (Intel), по мнению которого, производительность процессоров должна удваиваться каждые 18 месяцев из-за сочетания роста количества транзисторов и увеличения тактовых частот процессоров.



Сейчас данный закон опровергли.

Развитие Интернета и накопление данных

Благодаря развитию интернета и всестороннему проникновению компьютеров и мобильных устройств во все сферы жизни человека, происходит увеличение количества данных в сети.



Согласно прогнозу IDC доля глобальной информационной сферы, подвергаемой анализу, к 2025 году вырастет в 50 и более раз, достигнув 5.2 Збайт, а объем данных анализируемых при участии когнитивных систем вырастет в 100 раз, составив 1.3 Збайт. Когнитивные системы позволяют чаще и более гибко анализировать данные во многих отраслях.

Что такое “BIG DATA”

Что такое “BIG DATA”

1880 год - обработка информации и представление данных переписи населения в Европе и США заняло более 8 лет. При этом по прогнозам обработка данных переписи через 10 лет заняла бы еще больше времени, и результаты не были бы готовы даже до проведения последующей переписи (1900). Тогда проблему решила табулирующая машина, изобретенная Германом Холлеритом в 1881 году.

1997 год - термин **Big Data** был впервые введен в 1997 году Майклом Коксом и Дэвидом Эллсвортом на 8-й конференции IEEE по визуализации. Они назвали проблемой больших данных нехватку мощностей и емкости основной памяти, локального и удаленного диска для выполнения обработки и виртуализации. А в 1998 году руководитель-исследователь из SGI Джон Мэши на конференции USENIX использовал термин Big Data в его современном виде.

Большие данные - данные, обработка которых в заданных условиях и ограничениях требует огромного количества вычислительных ресурсов и применения новых технических подходов.



Что такое “BIG DATA”

Переломным моментом стал 2003 год, за который было сгенерировано информации больше, чем за все предыдущее время существования человечества.

Выходит доклад Google File System о вычислительной концепции MapReduce, которая легла в основу Apache Hadoop.

Hadoop стал отдельным полноценным решением для хранения и обработки Больших Данных.

Взаимосвязь Data Mining, Big Data и Data Science

Взаимосвязь Data Mining, Big Data и Data Science

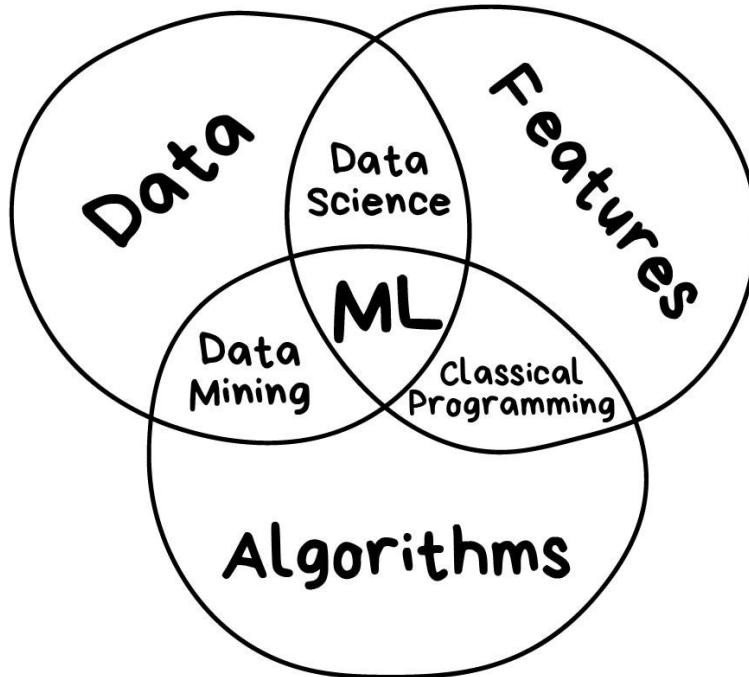
Data Mining (добыча и обработка данных) - деятельность по добыче, обработке и интерпретации данных.

Методы:

- Эксплоративный анализ данных (Exploratory Data Analysis) - поиск полезной информации с использованием инструментов визуализации данных и статистических методов.
- “Классические” алгоритмы машинного обучения - методы классификации, статистические методы (деревья решений, корреляционный и регрессионный анализы, факторный анализ, анализ связей и т.д.)

Одно из важнейших назначений методов data mining состоит в наглядном представлении результатов вычислений (визуализация).

Взаимосвязь Data Mining, Big Data и Data Science



3. Стандартные задачи DS.

Данные

- Чтобы научиться решать задачи ML/DL нужны данные
- Много данных (но есть исключения)
- Данные можно разделить на 2 группы
 - Структурированные
 - Неструктурированные

Когда пытаешься объяснить клиенту как работает твой алгоритм машинного обучения:



Структурированные данные

Структурированные данные

Rows

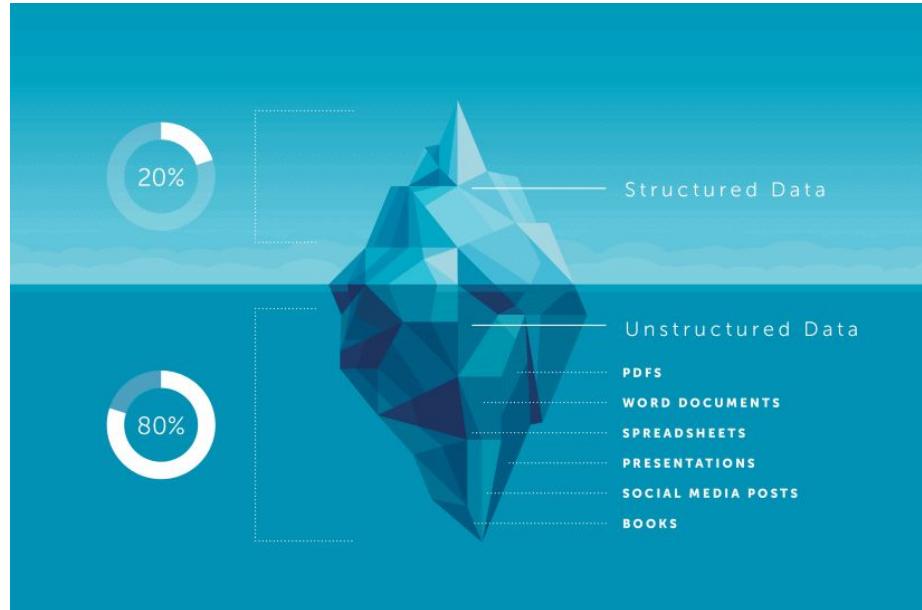
Columns

name	region	sales	expenses
William	East	50000	42000
Emma	North	52000	43000
Sofia	East	90000	50000
Markus	South	34000	44000
Edward	West	42000	38000
Thomas	West	72000	39000
Ethan	South	49000	42000
Olivia	West	55000	60000
Arun	West	67000	39000
Anika	East	65000	44000
Paulo	South	67000	45000

Неструктурированные данные

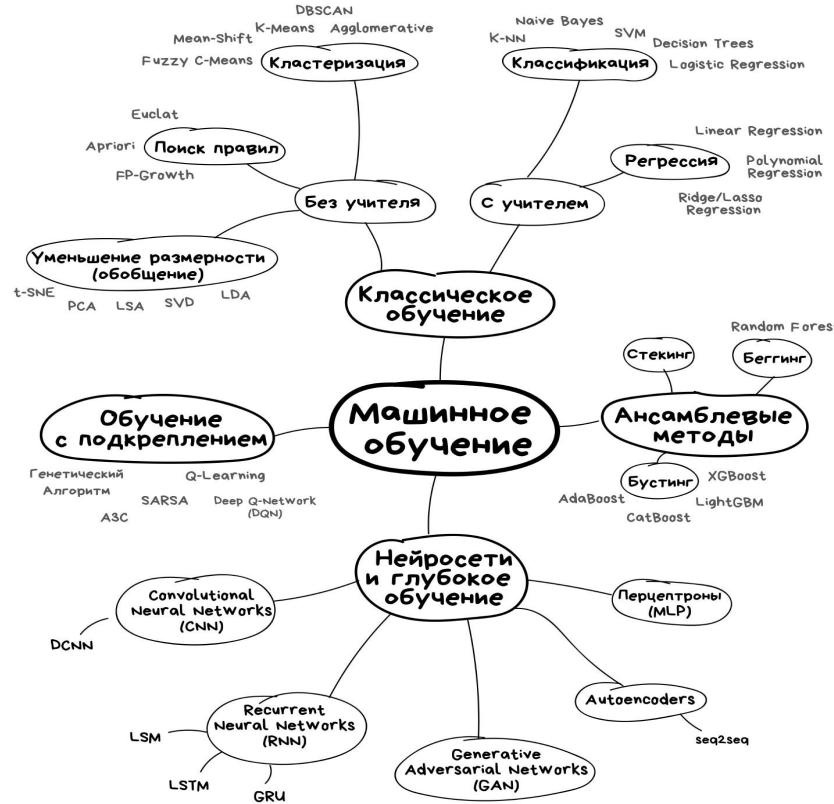
Неструктурированные данные

- Текст
- Картинки
- Видео
- Аудио
- Графовая структура
- Потоковые (streaming) данные
- Метаданные
- Гео-данные



Виды задач машинного обучения

Виды задач машинного обучения



Классическое машинное обучение

Классическое Обучение



Supervised Learning

- У данных есть разметка (Label, Feature Label, Y)
- Существуют наблюдения (правильные и неправильные ответы), которые необходимо предсказать
- Если целевая переменная - категориальная (да/нет, 0/1, 0/1/2/3/4, оплатит/не оплатит и тд) то перед нами стоит задача классификации.
- Если целевая переменная - непрерывная величина (цена ценной бумаги, количество км, уровень загрязнения воздуха и тд) то перед нами стоит задача регрессии.

Supervised Learning

Назовите примеры задач регрессии / классификации из реальной жизни которые вы потенциально встречали?

Unsupervised Learning

- У данных не существует разметки (Unlabeled data)
- Но необходимо извлечь полезную информацию (паттерны, закономерности, “инсайты”)
- Например какие объекты похожи друг на друга и почему? Это задача кластеризации
- Какие объекты не похожи ни на кого из выборки (Задача поиска аномалий / выбросов)
- Изменить число признаков в данных не теряя полезной информации (Задача уменьшения размерности)

Unsupervised Learning

Назовите примеры задач unsupervised learning из реальной жизни которые вы потенциально встречали или можете встретить?

Reinforcement Learning

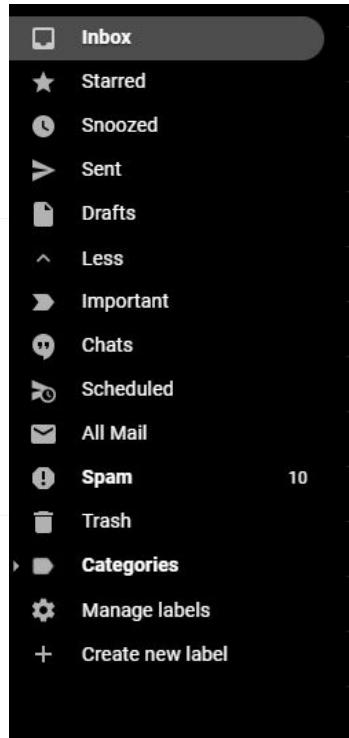
- “True AI”
- Правильных ответов нет, но есть среда, в которой взаимодействуют агенты (то есть симуляция), существует награда за действия и наказание за неправильное действие. Задача разработать оптимальный план поведения в соответствии с поставленной задачей.
- Пример: Dota AI, DeepMind, Searching Spiders etc.

Разминка

Тест на понимание: Назовите где здесь машинное обучение и тип решаемой задачи.

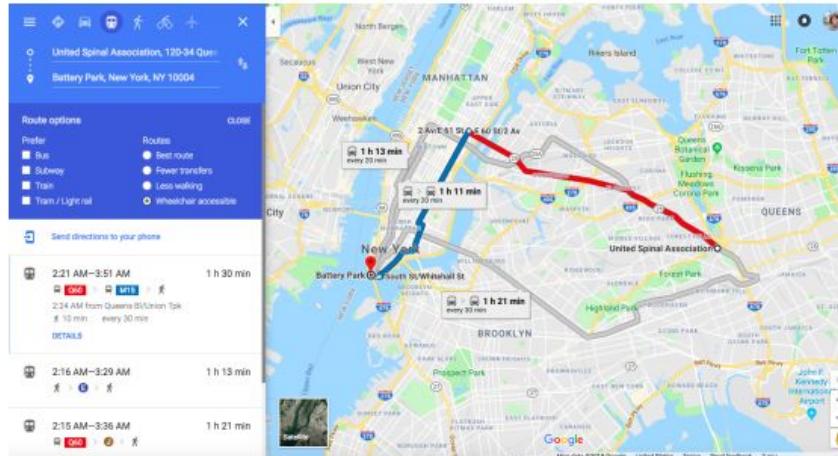
Разминка

Назовите где здесь машинное обучение и тип решаемой задачи.



Разминка

GoogleMaps



Разминка

The image consists of three screenshots from the Uber mobile application, illustrating a surge pricing event.

Screenshot 1: Pickup Location Selection

PICKUP LOCATION
29-49 3rd Street

SET PICKUP LOCATION

1 MIN

Map showing the pickup location at 29-49 3rd Street, San Francisco, with surrounding streets like California St, Mission St, and Powell St.

uberX POOL uberPOOL TAXI ACCESS

THIS RATE EXPIRES IN 2 MIN

Screenshot 2: Surge Pricing Confirmation

Demand is off the charts! Fares have increased to get more Ubers on the road.

2.0x THE NORMAL FARE

\$9 MINIMUM FARE

\$0.52 / MIN \$2.60 / MILE

SAVE UP TO 50%, TRY uberPOOL

NOTIFY ME WHEN SURGE DROPS

I ACCEPT HIGHER FARE

Screenshot 3: Confirmation Keypad

CONFIRMATION

Type (2.0)
to confirm your fare multiple.

MY FARE WILL BE

2 . 0

TIMES THE NORMAL FARE

1	2 ABC	3 DEF
4 GHI	5 JKL	6 MNO
7 PQRS	8 TUV	9 WXYZ
0		X

Разминка



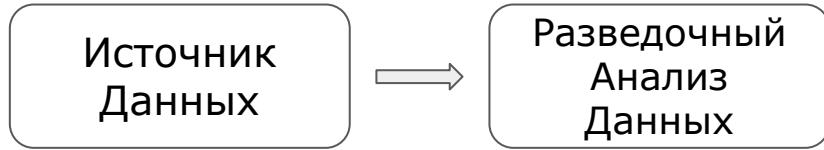
4. Подход к решению задач.

Подход к решению задач

Источник
Данных

- Какие данные
- Источник
- Содержание
- Объем
- Структура

Подход к решению задач



Exploratory Data Analysis (EDA)

- Знакомство с данными
- Подсчет базовых статистик
- Визуализация данных
- Поиск закономерностей и зависимостей
- Потенциальный выбор соответствующего класса моделей

Подход к решению задач



Data Preprocessing:

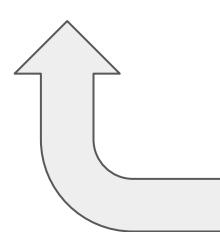
- Перевод данных в подходящий для моделей формат
- Работа с кодированием категориальных признаков
- Поиск выбросов и аномалий
- Очистка данных
- Создание новых признаков
- Сокращение размерности

Подход к решению задач



Modelling / Training :

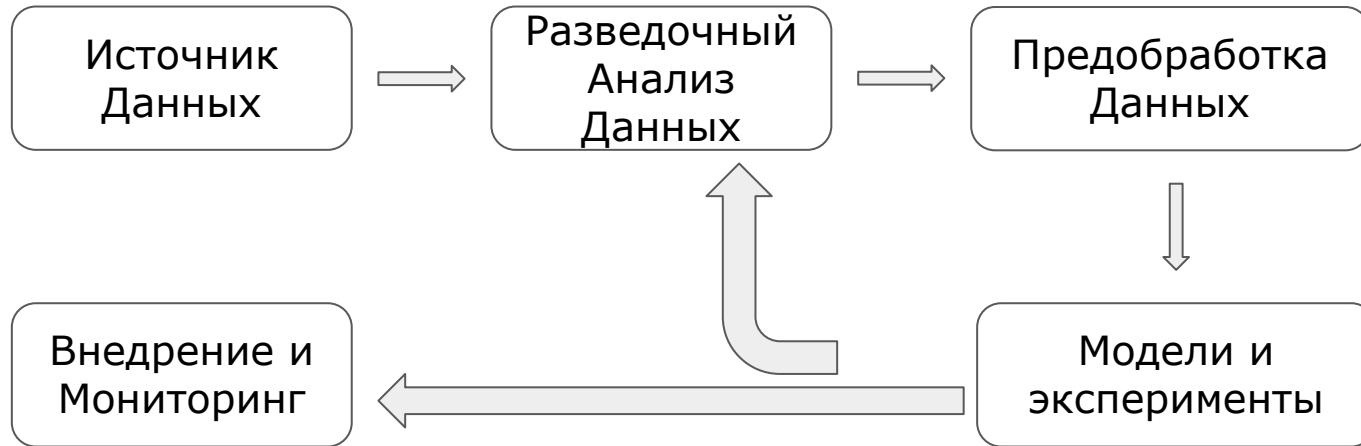
- Построение моделей
- Оценка качества
- Подбор гиперпараметров
- Отбор признаков



Модели и
эксперименты

Повторный цикл если все плохо.

Подход к решению задач

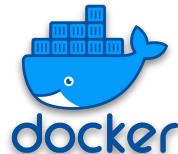


Подход к решению задач

Теперь вы знаете основные пункты методологии работы в отрасли Data Science - **CRISP-DM** (Cross Industry Standard Process for Data Mining).

Теперь вы будете работать только по нему :)

5. С чем будем работать: инструменты



6. Python / Anaconda / Jupyter Notebook / Google Collab



Скачиваем интерпретатор:

<https://www.python.org/downloads/>

6. Python / Anaconda / Jupyter Notebook / Google Collab

Для целей разведывательного и статистического анализа данных будем пользоваться Jupyter Notebook / Google Colab

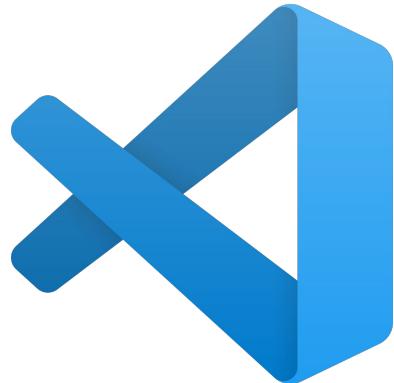


Установка дистрибутивов:

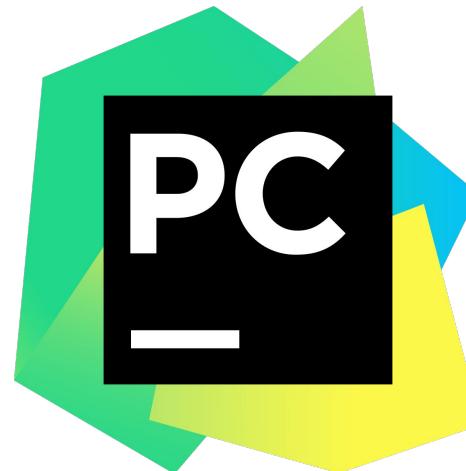
- Anaconda: <https://www.anaconda.com/>
- Google Colab: <https://colab.research.google.com/>

6. Python / Anaconda / Jupyter Notebook / Google Collab

Для реализации моделей и имплементации их в production:



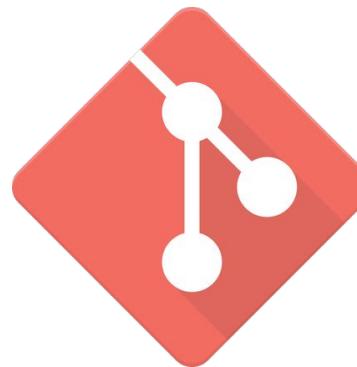
VS Code
<https://code.visualstudio.com/>



PyCharm
<https://www.jetbrains.com/pycharm/download>

6. Python / Anaconda / Jupyter Notebook / Google Collab

Работать будем через систему контроля версий Git



- Заводим аккаунт на GitHub <https://github.com/>
- Скачиваем git bash <https://git-scm.com/downloads>

Выводы

- Машинное обучение - одно из самых востребованных направлений настоящего и ближайшего будущего
- Data Science объединяет множество междисциплинарных направлений, в том числе бизнес в виде domain knowledge
- Междисциплинарность позволяет не привязываться к определенной индустрии, но требует больших доз самообразования и самостоятельного поиска информации на возникающие вопросы

Дополнительная литература



Программирование на

Python

ТОМ I

О'РЕИЛЛЫ®



Mark Lutz

Марк Лутц,
“Программируем на
Python” в 2-х томах



Себастьян Рашка,
“Python и машинное
обучение”



Василий Сабиров,
“Игра в цифры”

Дополнительные материалы

Статистика и математика:

- [Повторение основ статистики. Базовые концепции доступным языком.](#)
- Конспекты избранных лекций курса MIT Statistical Thinking and Data Analysis
 - [Review of Probability](#)
 - [Statistical Distributions](#)
 - [Inferences for Single Samples](#)
 - [Inferences for Multi Samples](#)

Быстрые шпаргалки:

- [Linear Algebra](#)
- [Statistics](#)
- [Probability](#)

Основы Python:

- [Python Data Science Handbook](#)
- [Introduction to NumPy](#)
- [Data Manipulations with Pandas](#)

Q&A