О работе в Data Science и машинном обучении

Евгений Борисов

О работе в Data Science

Автоматические Рекомендеры

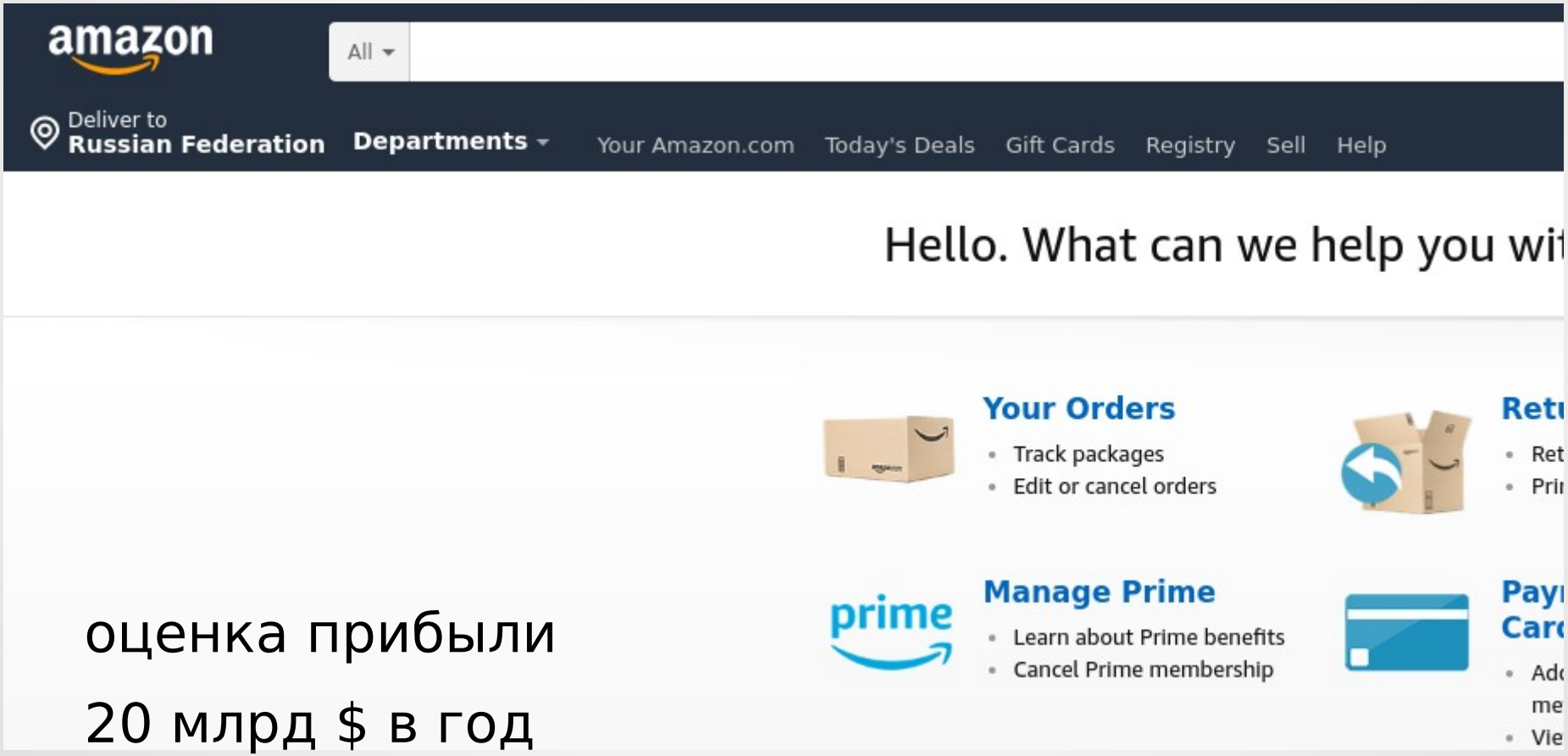
прокат фильмов с 1997, 117M подписчиков

оценка прибыли - 5 млрд $ в год

2009 Netflix Prize $1M

О работе в Data Science

Автоматические Рекомендеры



оценка прибыли

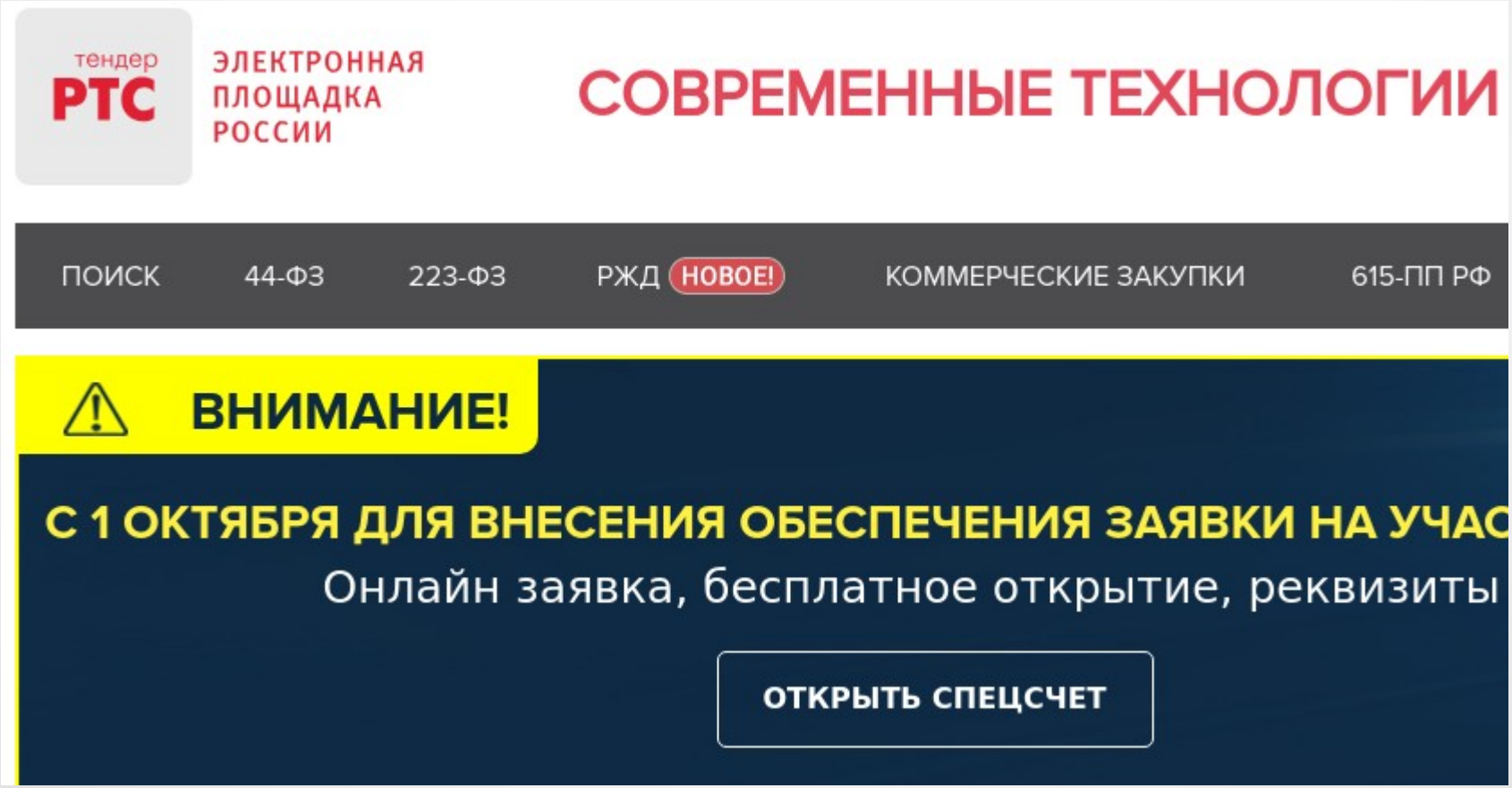
20 млрд $ в год

30% за счёт рекомендера

О работе в Data Science

Автоматические Рекомендеры

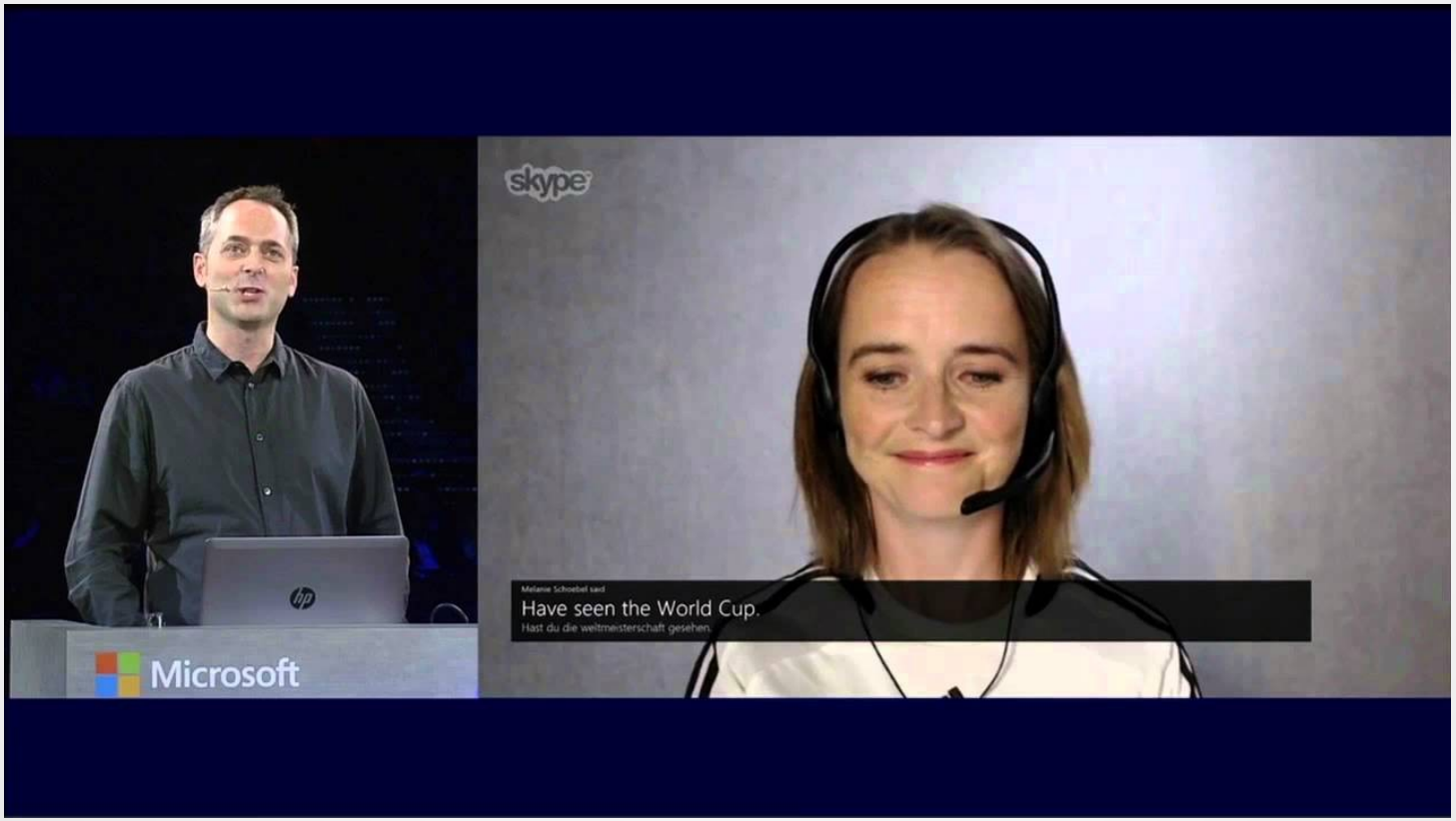
до 40% всех госзакупок России



из них 10% за счёт рекомендера

О работе в Data Science

Автоматический Перевод



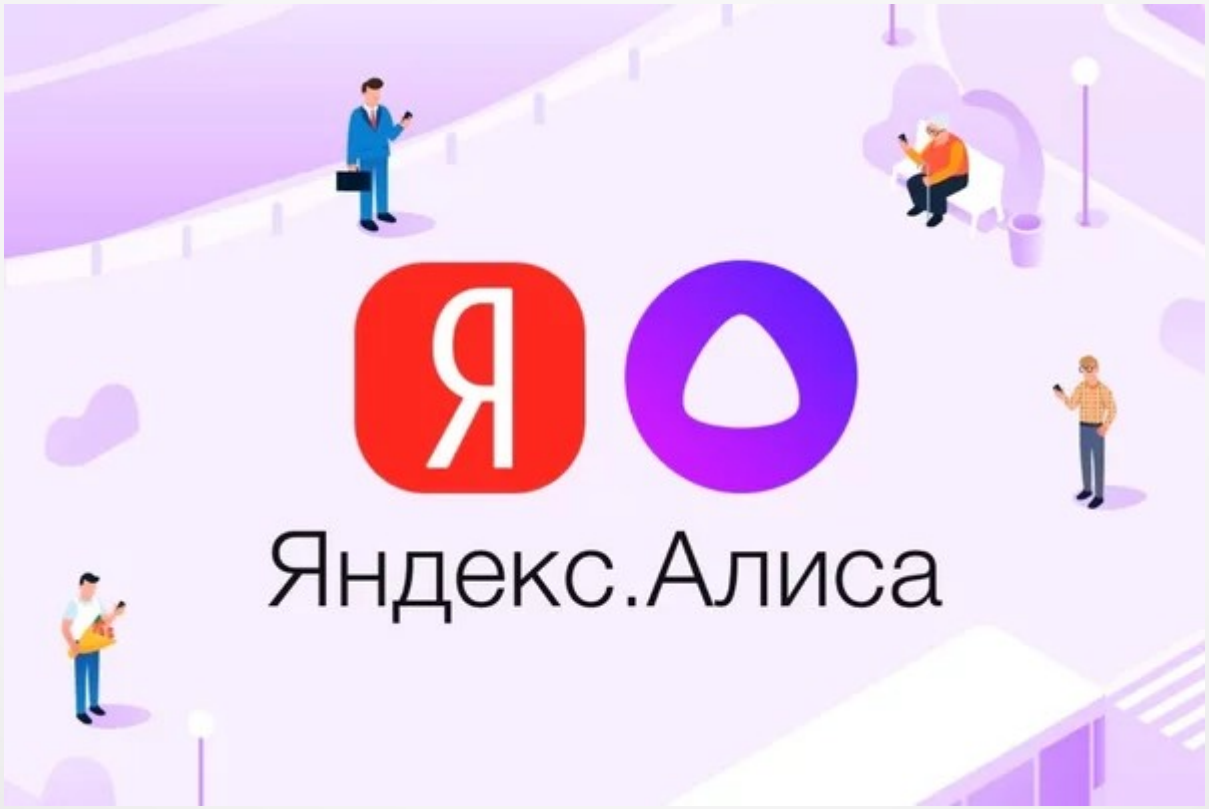
https://www.youtube.com/watch?v=C4-qrppI2Nc&t=2m30s

О работе в Data Science

Автоматический Секретарь

О работе в Data Science

Автоматический Секретарь



О работе в Data Science

Беспилотный Автомобиль

https://www.youtube.com/watch?v=Bx08yRsR9ow

О работе в Data Science

Автономные Роботы

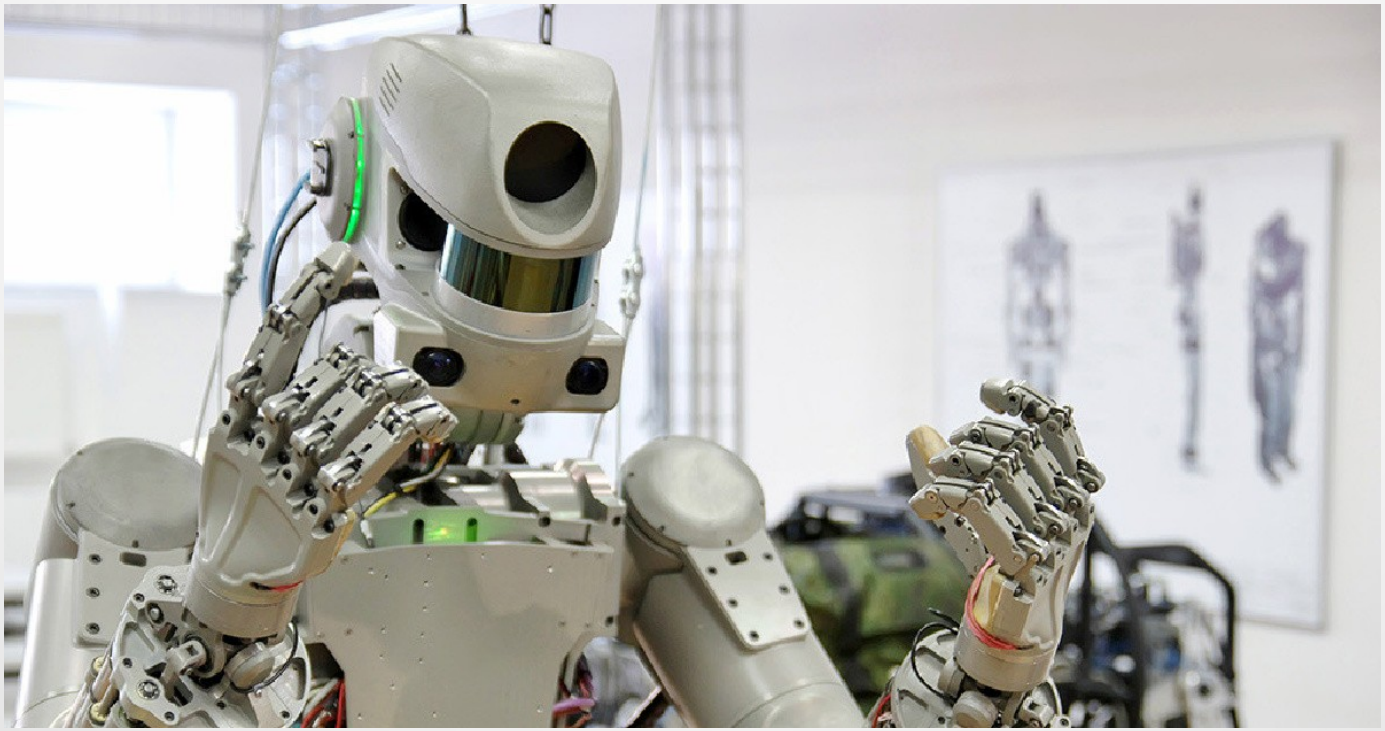
https://www.youtube.com/watch?v=LikxFZZO2sk

О работе в Data Science

Автономные Роботы

Фёдор (FEDOR — Final Experimental Demonstration Object Research)

НПО "Андроидная техника"



О работе в Data Science

Военные Дроны

О работе в Data Science

Data Science

Computer Vision / Natural Languages Processing / Data Analysis / Speech Recognition

Области применения ML

обработка изображений (CV)

обработка текстов (NLP)

обработка звуков (SR)

анализ соц.сетей (DA, SNA)

автоматическое управление (Robotics)

торгово-экономические модели (DA, Econometrics)

О работе в Data Science

Как это работает ?

формируем учебный набор

обучаем модель

запускаем модель в работу

О работе в Data Science

Как это работает ?

формируем учебный набор

обучаем модель

запускаем модель в работу

на самом деле всё немного сложнее :)

О работе в Data Science

...а чтобы сам учился ?

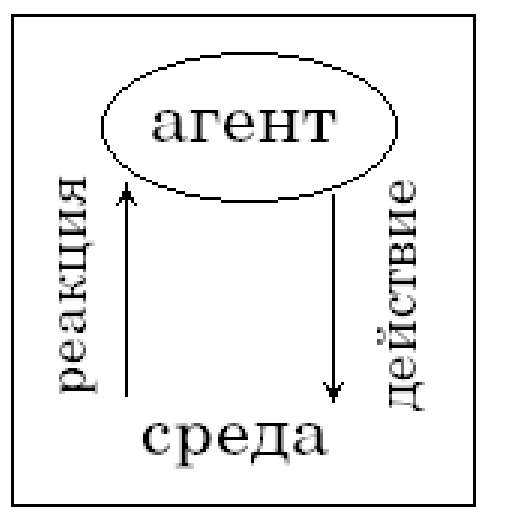
обучение с подкреплением

учебного набора в явном виде нет

собираем историю действий и последствий

пытаемся предсказывать реакцию среды

выбираем оптимальное действие



ML: с чего все начинается?

извлечение признаков из объекта
(feature extracting)

формирование пространства признаков

объект -> [FE] -> признаки -> [ML] -> результат

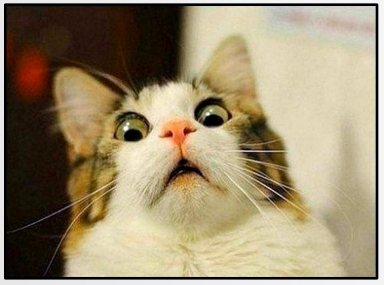
ML: с чего все начинается?

Классификатор: домашние и дикие коты

ML: с чего все начинается?

Классификатор: домашние и дикие коты

извлекаем признаки
(цвет,усы,лапы и хвост)



→ [0.14, 12, ..., 345]

→ [78.0, 20, ..., 177]

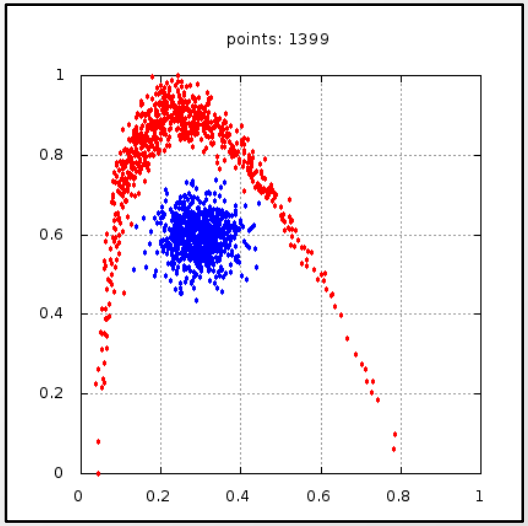
ML: с чего все начинается?

Классификатор: домашние и дикие коты

извлекаем признаки
(цвет,усы,лапы и хвост)

→ [0.14, 12, ..., 345]

→ [78.0, 20, ..., 177]



ML: и что дальше?

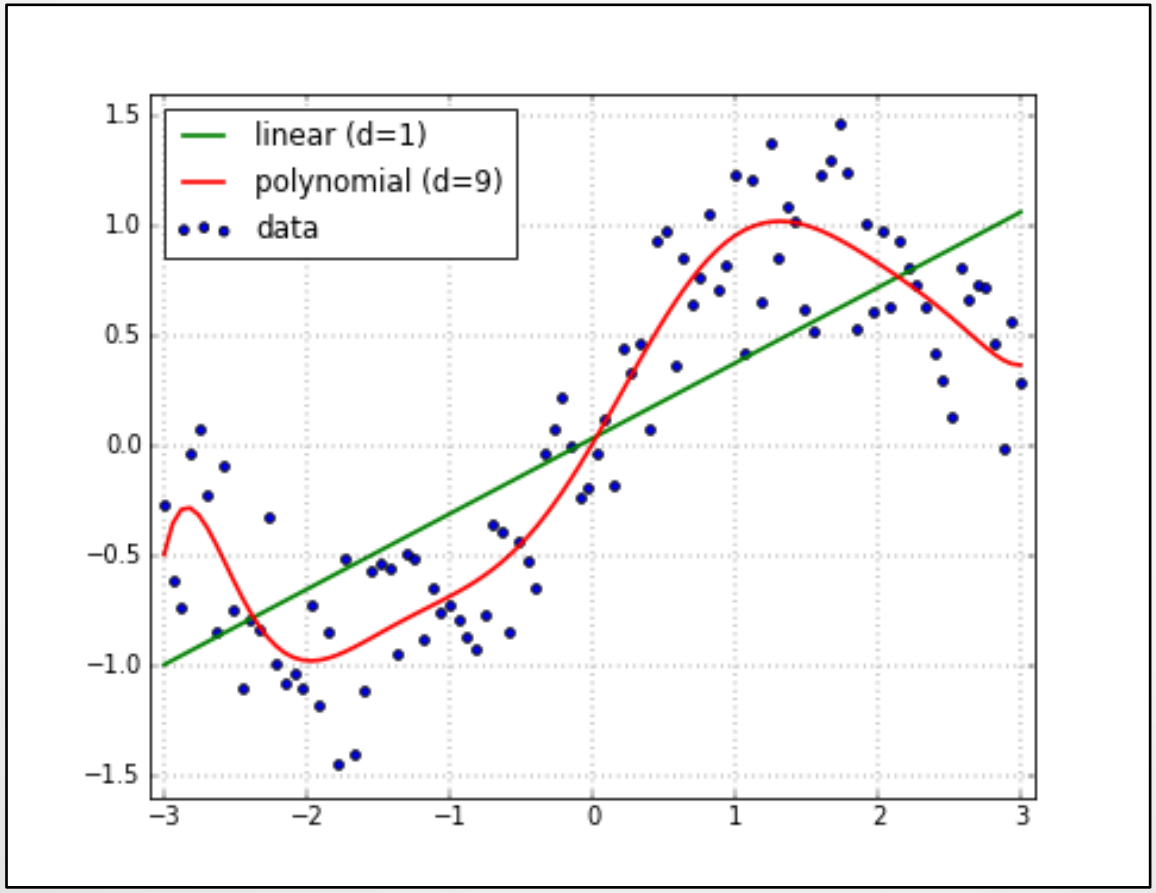
Задачи:

- Регрессия - восстановление зависимости
- Классификация - разделение на части
- Кластеризация - формирование групп

ML: регрессия

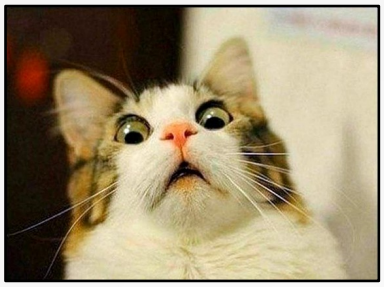
восстановление зависимости по набору точек

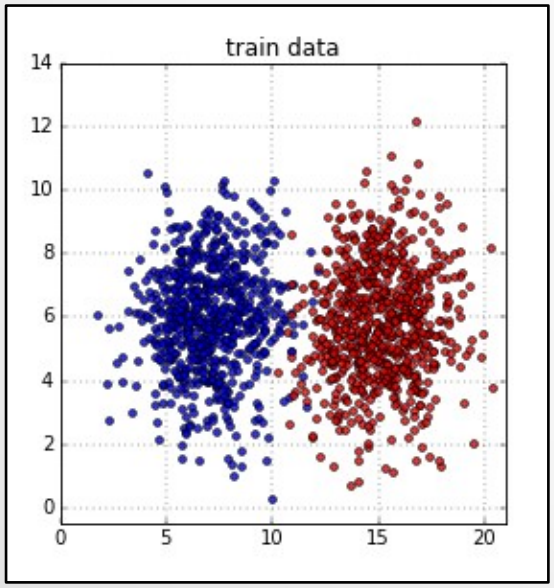
Оценка недвижимости: [район,площадь] → цена

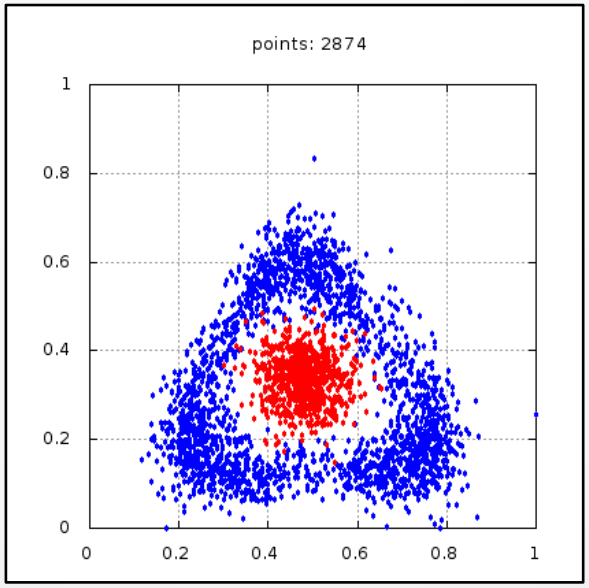


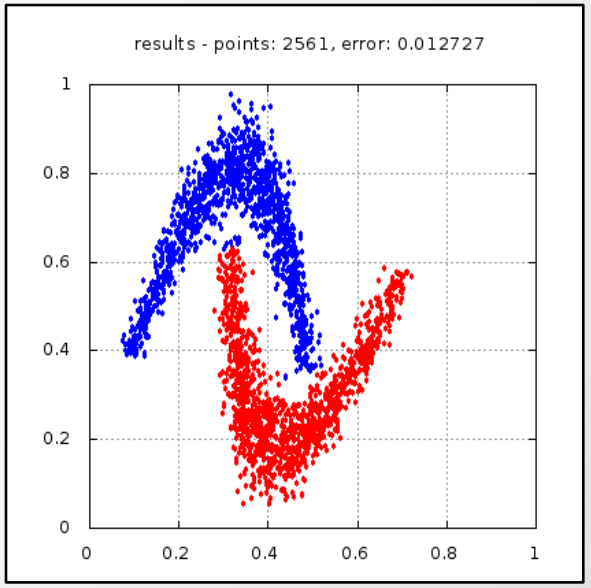
ML: классификация

разделения объектов на классы

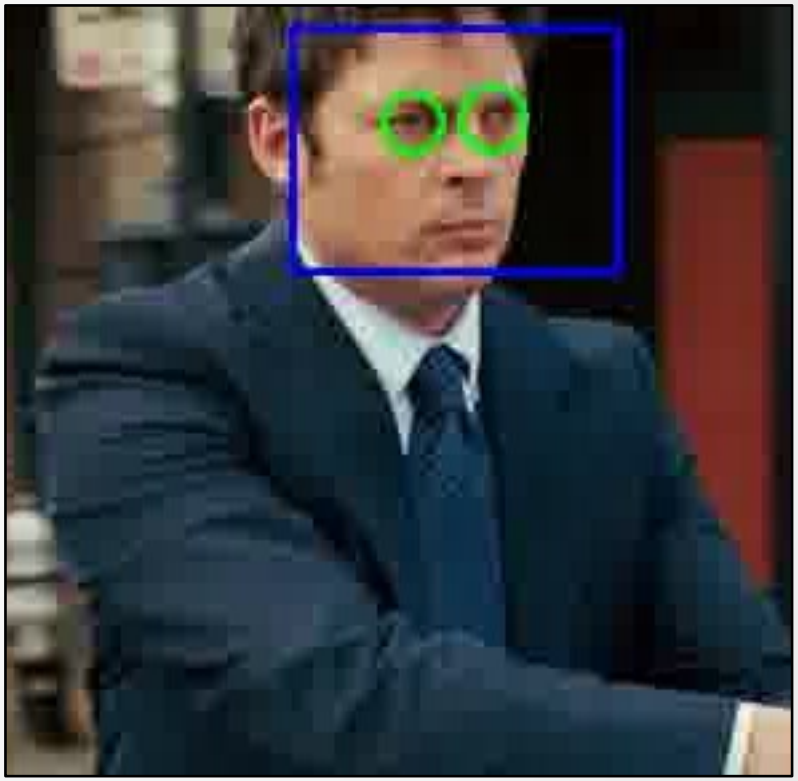
Детектор котов:  → вектор-признак → есть/нет

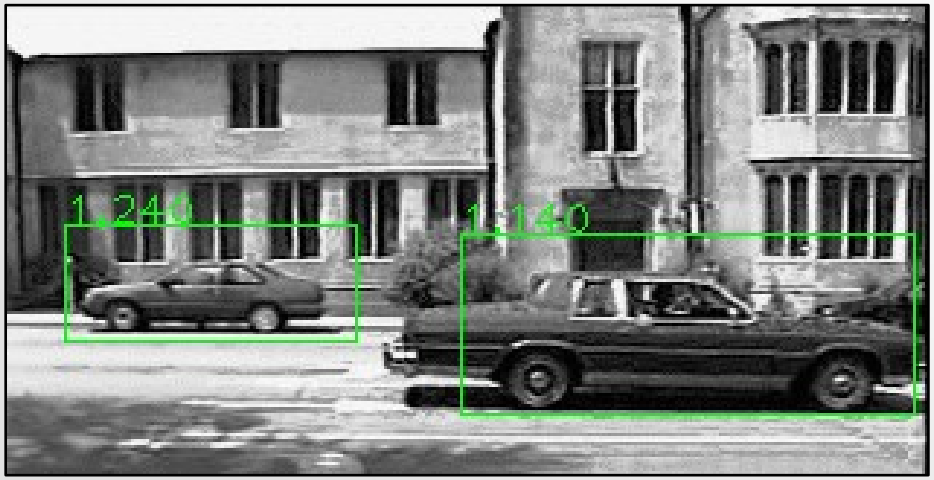






ML: Computer vision (CV)

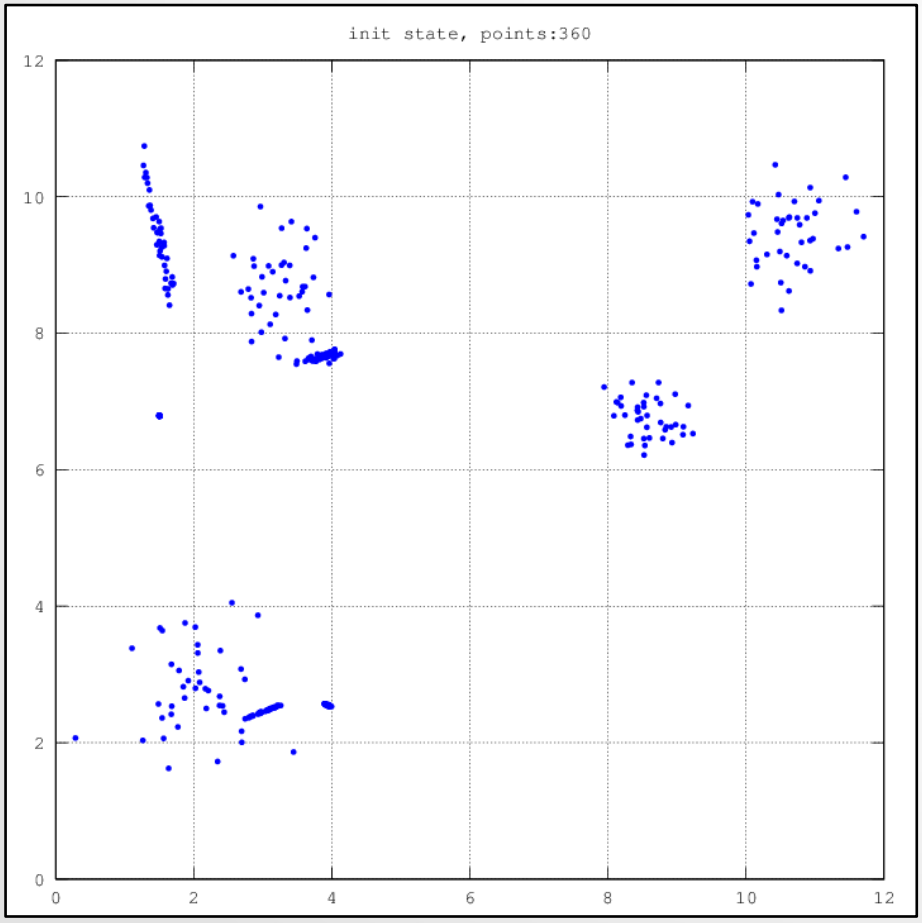


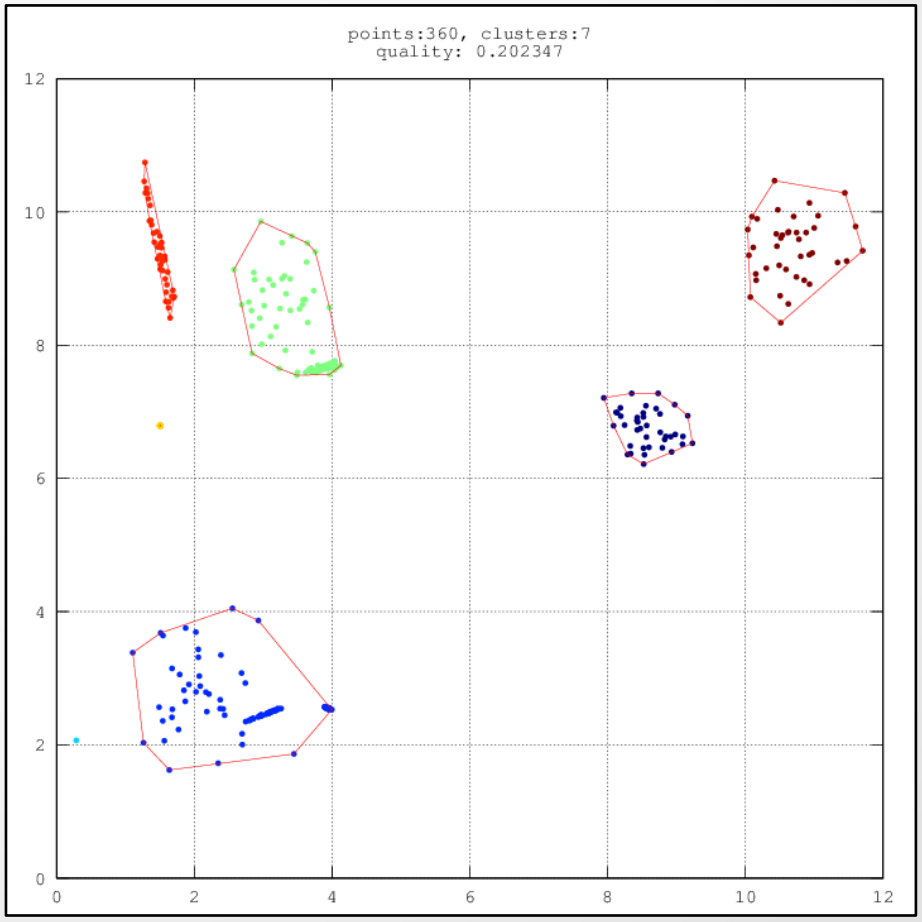


ML: кластеризация

объединение схожих объектов в группы

Поиск похожих текстов: текст → признаки → группа





ML: Natural Language Processing (NLP)

Поиск похожих текстов

Около 18 тысяч человек покинули подконтрольные боевикам районы Алеппо. За минувшие сутки из подконтрольных боевикам районов сирийского города Алеппо было выведено около 17,971 тысячи жителей, в их числе 7,542 тысячи детей. Об этом в субботу, 10 декабря, сообщает ТАСС со ссылкой на российский Центр примирения враждующих сторон в Арабской Республике.

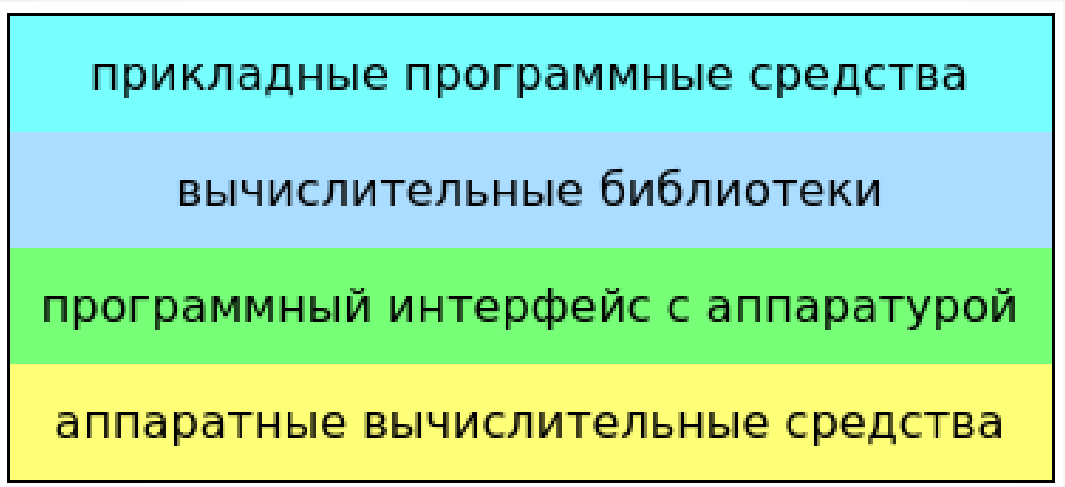
Битва за Алеппо: повстанцы просят дать им вывезти раненых Сирийские повстанцы просят о пятидневном перемирии, чтобы эвакуировать раненых из районов в восточной части Алеппо, после того как они вывели все свои отряды из исторического центра — Старого города.

ML: и куда дальше?

- Статистические: naiveBayes, EM
- Логические: *decision tree*
- Метрические: *k-neighbors, k-means*
- Линейные: *SGD*
- Композиции: *AdaBoost*
- *Deep Learning*

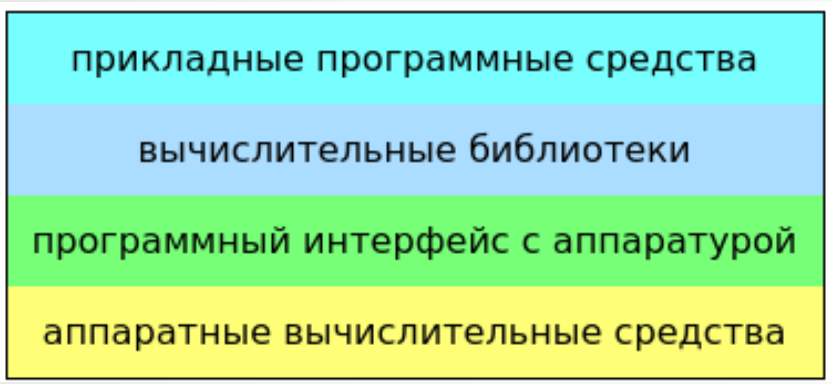
ML: технические средства

общее описание стека технологий



ML: технические средства

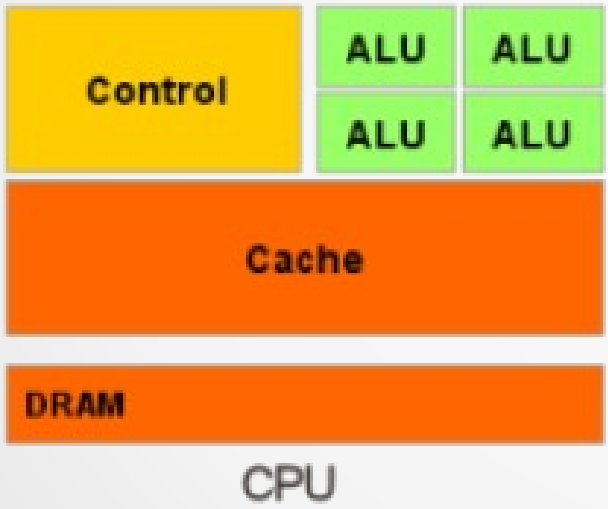
общее описание стека технологий

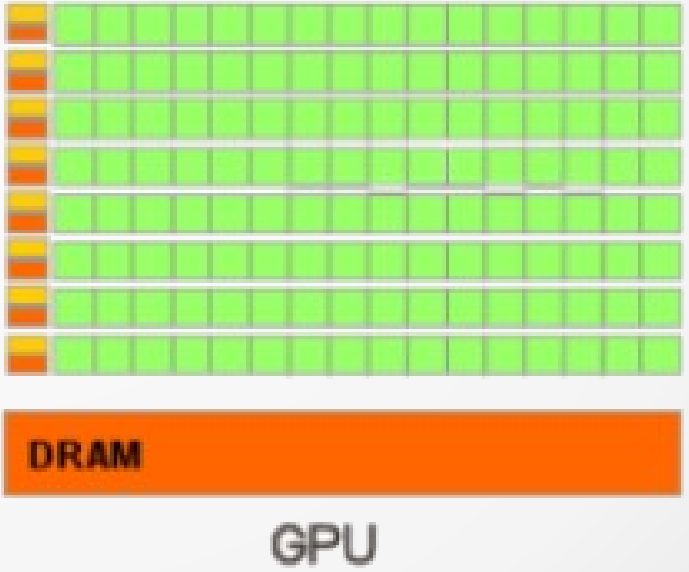


Numpy, SciPy
Python
BLAS, LAPAC, MKL
CPU

ML: технические средства

GP-GPU General-Purpose Graphics Processing Units

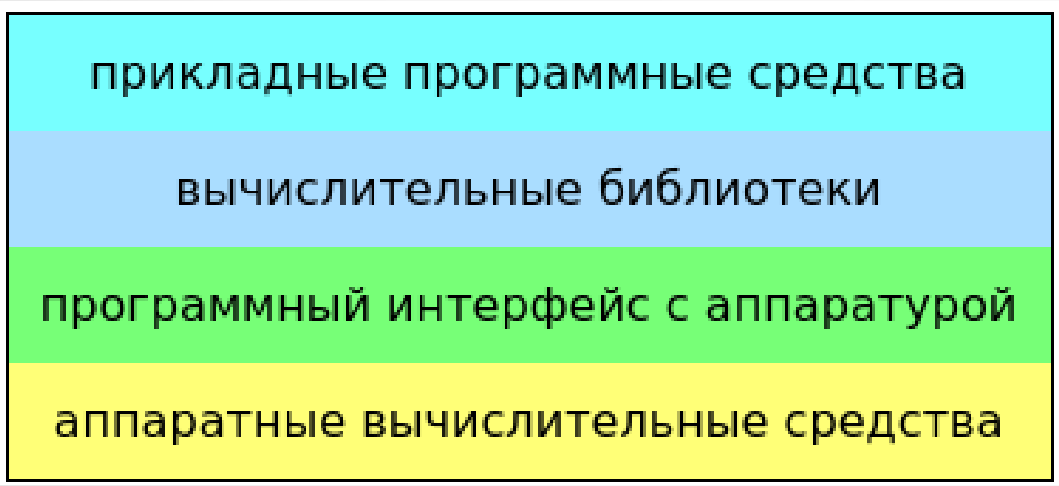


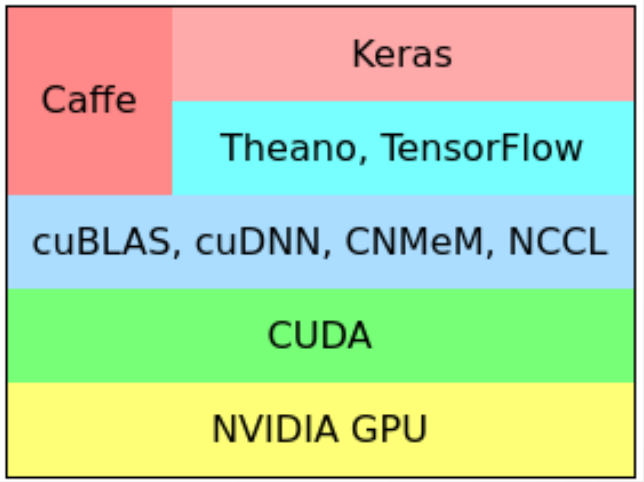


CUDA / OpenCL

ML: технические средства

описание стека технологий





О работе в Data Science

Технические Средства

Python

Jupyter

R

Numpy

Matplotlib

Scikit-Learn

NLTK

Pandas

Keras

Theano

TensorFlow

OpenCV

scikit-image

GPU

CUDA

OpenCL

Spark

О работе в Data Science

Что нужно чтобы стать data scientist'ом ?

мат.анализ

алгебра

теория вероятностей и мат.статистика

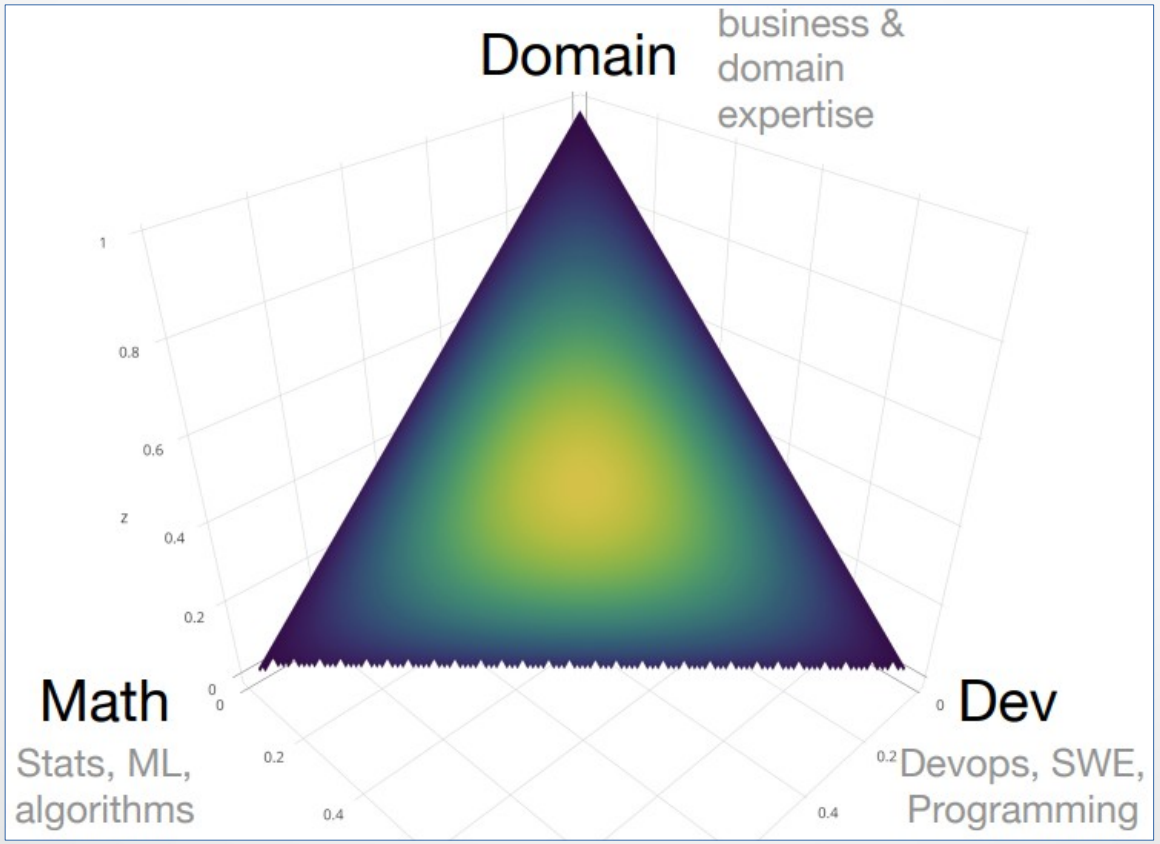
программирование с уклоном в HPC

знания по специализации

О работе в Data Science

выбор специализации

математика / программирование / хозяйственная деятельность

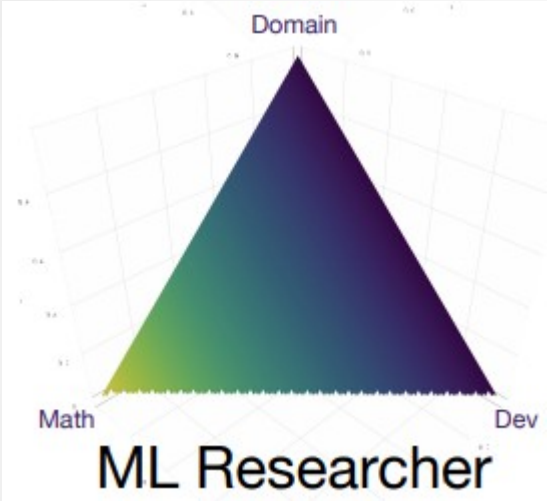


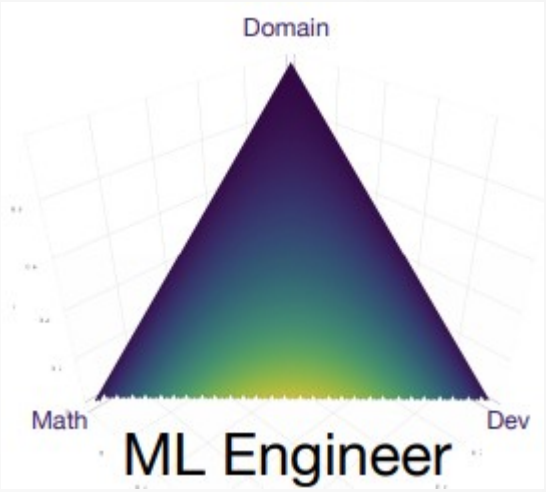
Алексей Натекин Чем отличаются data analyst, data engineer и data scientist // ODS 2018

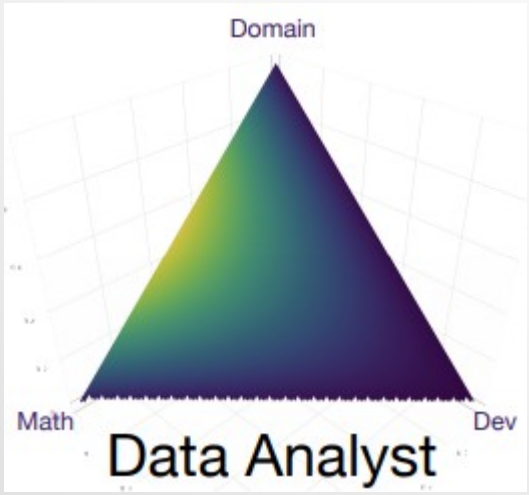
О работе в Data Science

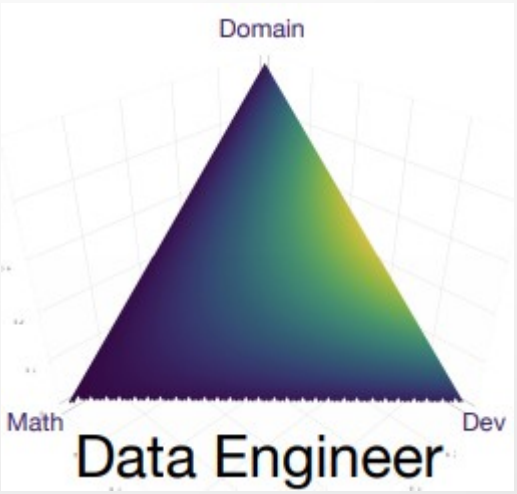
выбор специализации

математика / программирование / хозяйственная деятельность









Алексей Натекин Чем отличаются data analyst, data engineer и data scientist // ODS 2018

О работе в Data Science

Где ещё поучиться DS/ML

ШАД / МШАД Яндекс

Coursera

Kaggle

ML: что почитать?

- Константин Воронцов - Машинное обучение
- Антон Конушин - Компьютерное зрение
- Радослав Нейчев - Машинное обучение, ФПМИ, 2020
- Andrew Ng - Machine Learning
- Евгений Борисов - http://mechanoid.su
- http://github.com/mechanoid5/ml_lectorium

О работе в Data Science

https://habr.com/ru/post/440602/

О работе в Data Science

Вопросы ?