

# **Лекция 1: О работе в Data Science и машинном обучении**

Евгений Борисов

# О работе в Data Science

## Автоматические Рекомендеры

прокат фильмов с 1997, 117М подписчиков

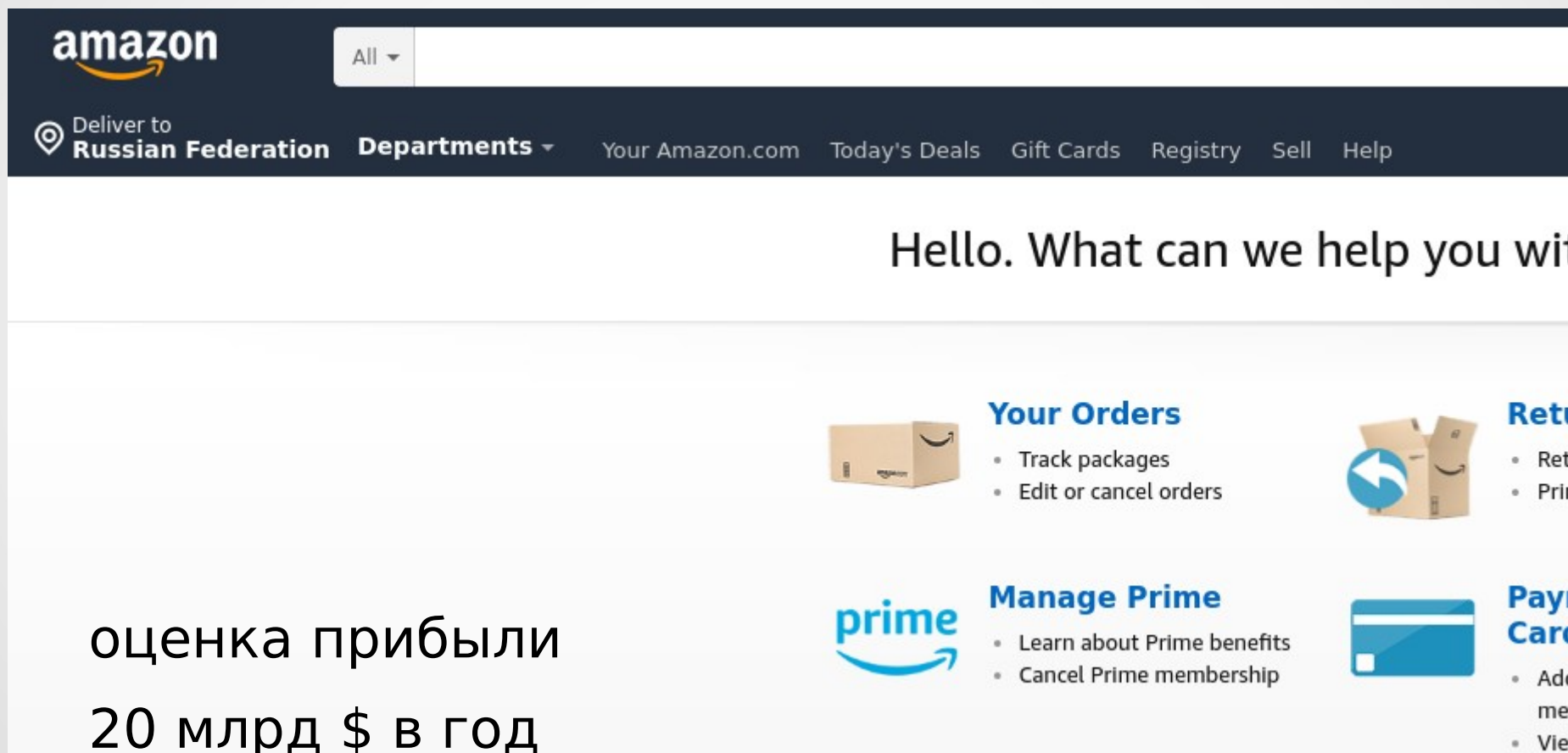


оценка прибыли - 5 млрд \$ в год

2009 Netflix Prize \$1M

# О работе в Data Science

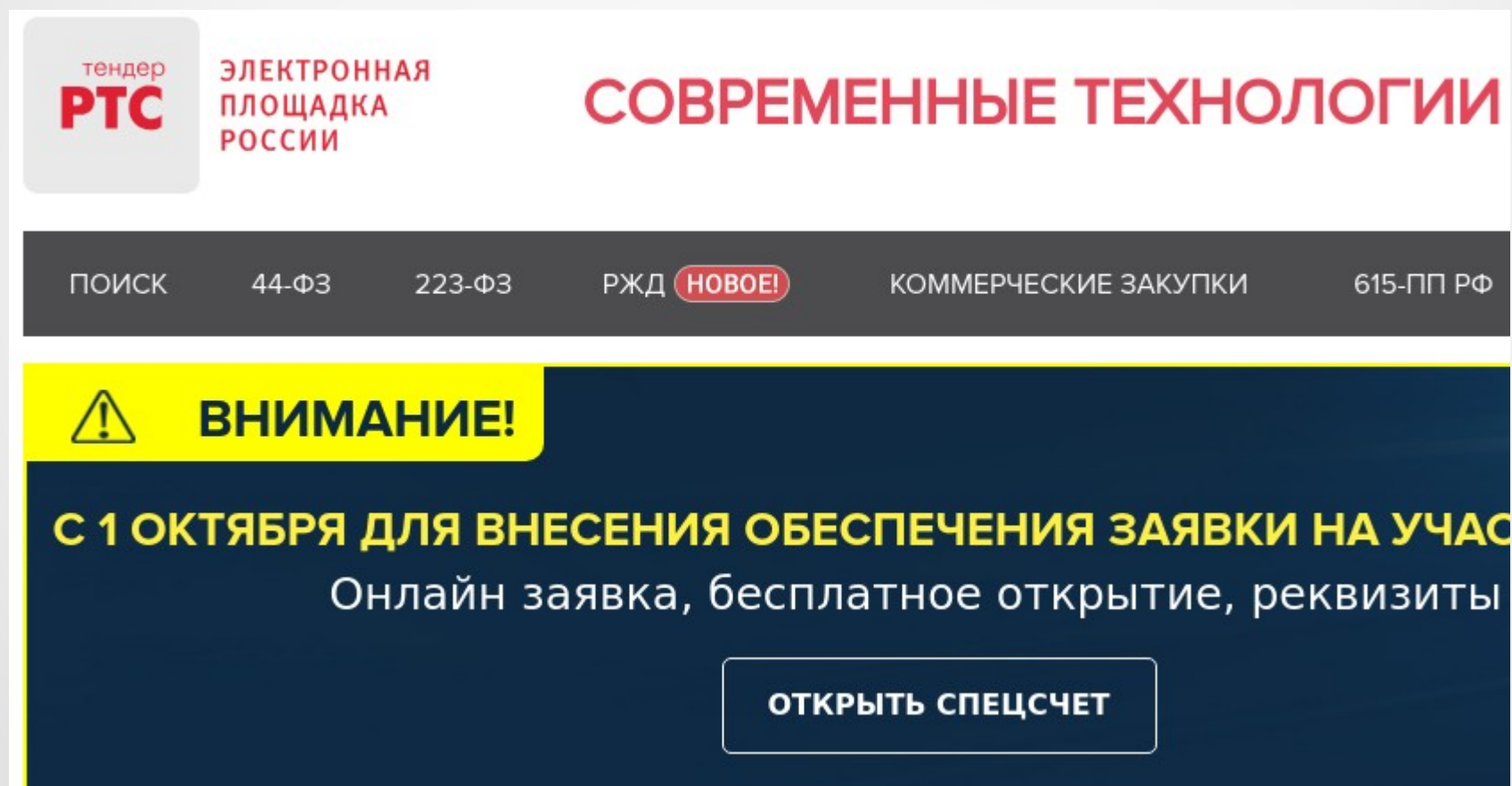
## Автоматические Рекомендеры



# О работе в Data Science

## Автоматические Рекомендеры

до 40% всех госзакупок России



The image shows a screenshot of the Russian Electronic Procurement Platform (РТС) website. The header includes the РТС logo with the text "тендер" (tender) above it, followed by "ЭЛЕКТРОННАЯ ПЛОЩАДКА РОССИИ" (Electronic Platform of Russia) and "СОВРЕМЕННЫЕ ТЕХНОЛОГИИ" (Modern Technologies). Below the header is a navigation bar with links: "ПОИСК" (Search), "44-ФЗ", "223-ФЗ", "РЖД" (Russian Railways) with a "НОВОЕ!" (New!) badge, "КОММЕРЧЕСКИЕ ЗАКУПКИ" (Commercial Purchases), and "615-ПП РФ". A prominent yellow banner with a warning icon and the text "ВНИМАНИЕ!" (Attention!) is displayed. Below the banner, a dark blue section contains the text "С 1 ОКТЯБРЯ ДЛЯ ВНЕСЕНИЯ ОБЕСПЕЧЕНИЯ ЗАЯВКИ НА УЧАСТИЕ" (From October 1 for the submission of a bid security) and "Онлайн заявка, бесплатное открытие, реквизиты" (Online bid, free opening, details). A button labeled "ОТКРЫТЬ СПЕЦСЧЕТ" (Open Special Account) is located at the bottom of this section.

тендер  
**РТС**

ЭЛЕКТРОННАЯ  
ПЛОЩАДКА  
РОССИИ

СОВРЕМЕННЫЕ ТЕХНОЛОГИИ

ПОИСК 44-ФЗ 223-ФЗ РЖД **НОВОЕ!** КОММЕРЧЕСКИЕ ЗАКУПКИ 615-ПП РФ

**ВНИМАНИЕ!**

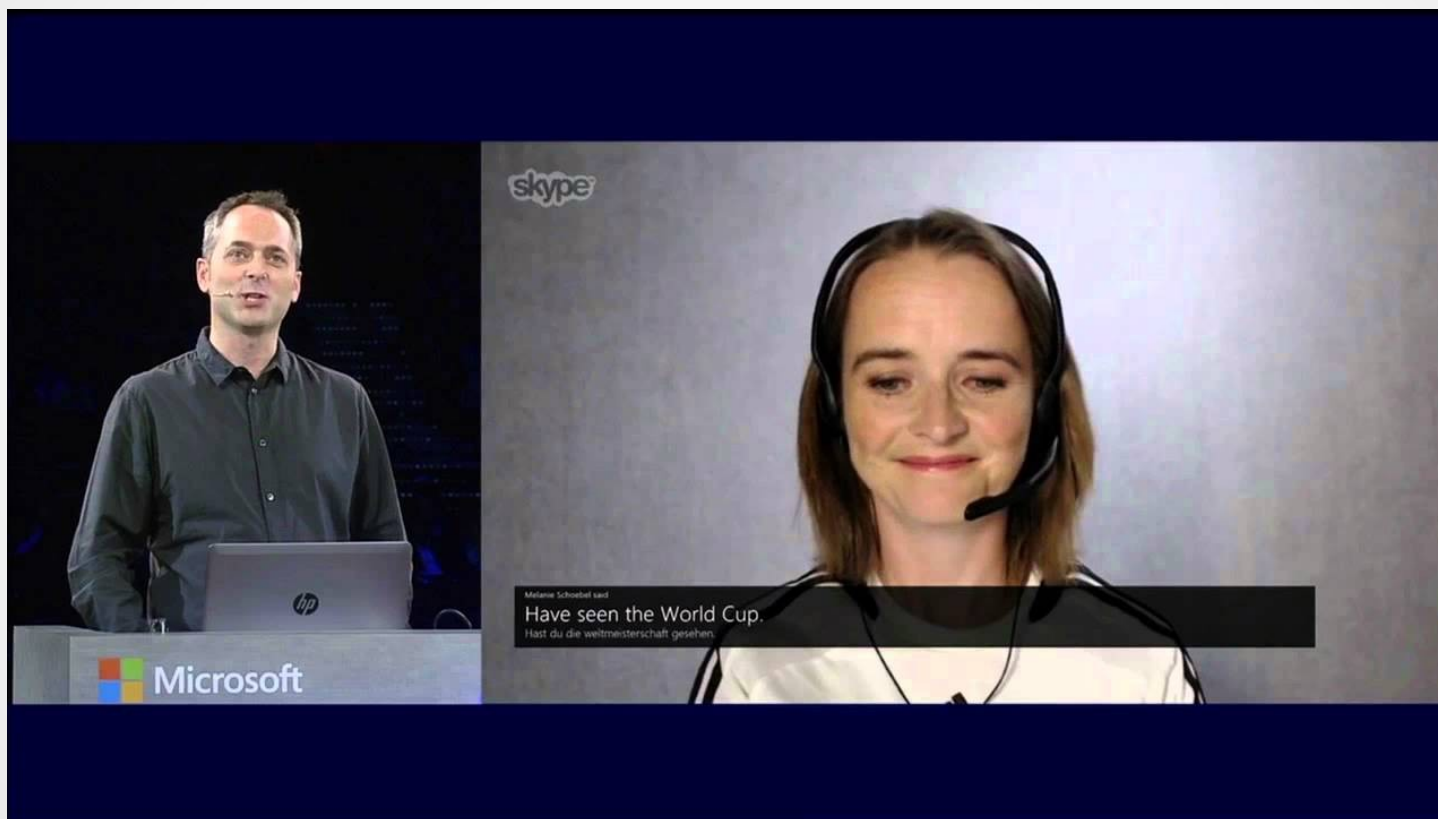
**С 1 ОКТЯБРЯ ДЛЯ ВНЕСЕНИЯ ОБЕСПЕЧЕНИЯ ЗАЯВКИ НА УЧАСТИЕ**  
Онлайн заявка, бесплатное открытие, реквизиты

**ОТКРЫТЬ СПЕЦСЧЕТ**

из них 10% за счёт рекомендера

# О работе в Data Science

## Автоматический Перевод



<https://www.youtube.com/watch?v=C4-qrppl2Nc&t=2m30s>

# О работе в Data Science

## Автоматический Секретарь

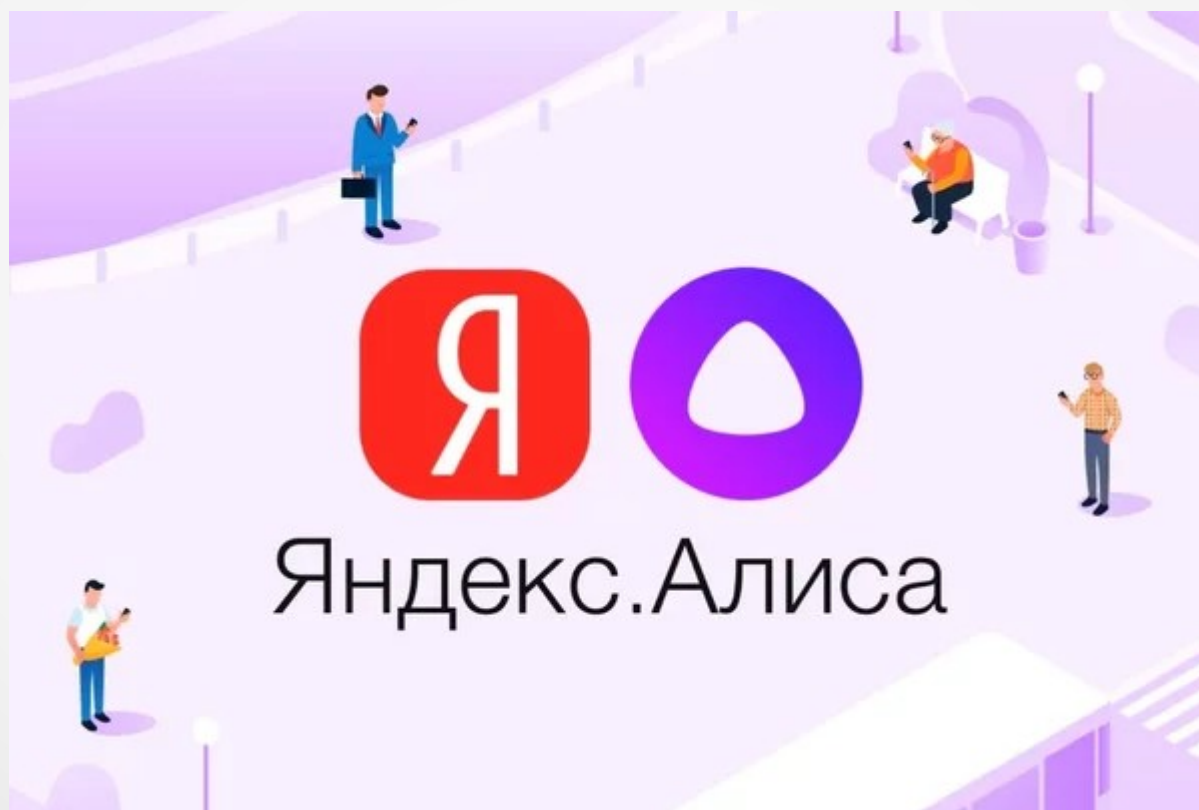


## Siri

Use your voice to send messages, set reminders, search for information, and more.

# 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



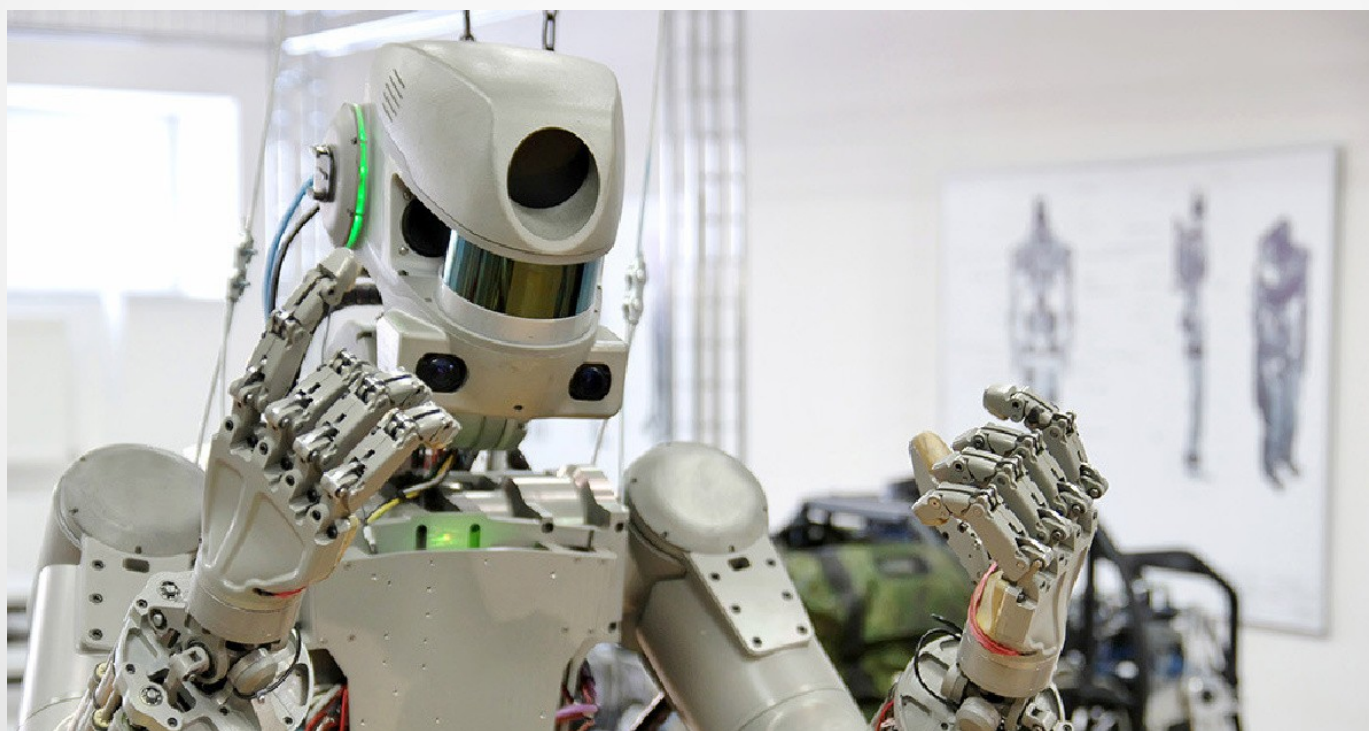
<https://www.youtube.com/watch?v=LikxFZZO2sk>

# 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 ?

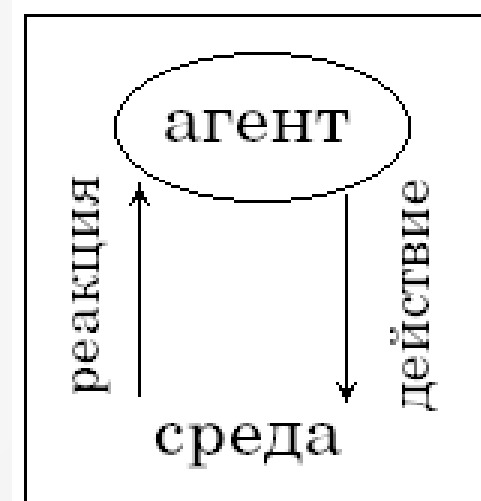
обучение с подкреплением

учебного набора в явном виде нет

собираем историю действий и последствий

пытаемся предсказывать реакцию среды

выбираем оптимальное действие



# ML: с чего все начинается?

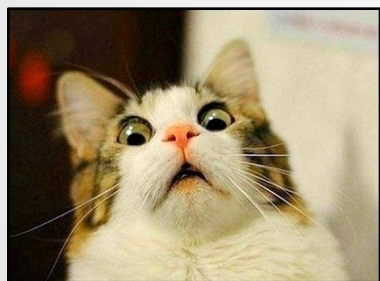
извлечение признаков из объекта  
(feature extracting)

формирование пространства признаков

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

# ML: с чего все начинается?

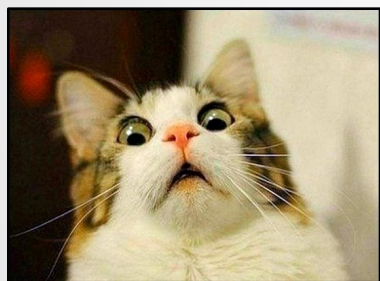
Классификатор: домашние и дикие коты



# ML: с чего все начинается?

Классификатор: домашние и дикие коты

извлекаем признаки  
(цвет, усы, лапы и хвост)



→ [0.14, 12, ..., 345]

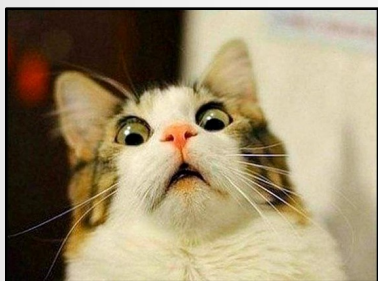


→ [78.0, 20, ..., 177]

# ML: с чего все начинается?

Классификатор: домашние и дикие коты

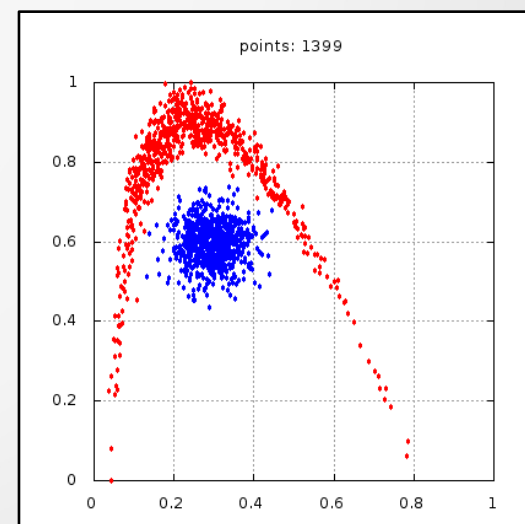
извлекаем признаки  
(цвет, усы, лапы и хвост)



→ [0.14, 12, ..., 345]



→ [78.0, 20, ..., 177]



# ML: обучение

1. формируем учебный набор
2. обучаем модель
3. запускаем модель в работу



# ML: эффект переобучения

эффект переобучения

хорошо обучается =)

плохо работает :(

# ML: борьба с переобучением

формируем 3 набора

учебный / контрольный / тестовый

1. обучаем на учебном  
и проверяем на контрольном
2. итоговый тест на тестовом
3. запускаем модель в работу

# ML: и что дальше?

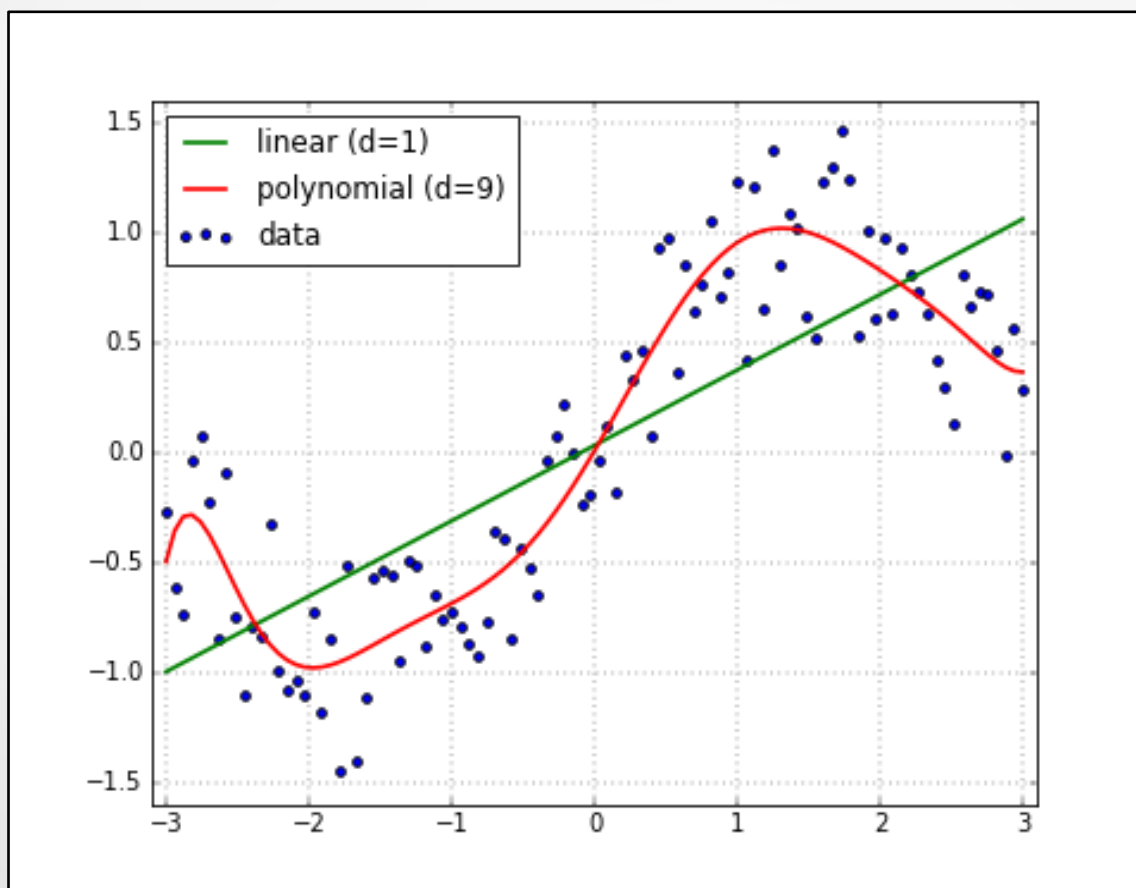
## Задачи:

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

# ML: регрессия

восстановление зависимости по набору точек

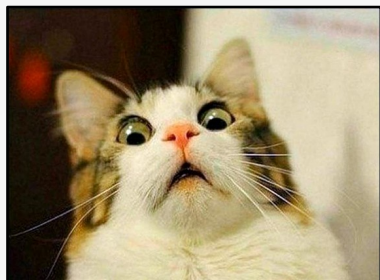
**Оценка недвижимости:** [район, площадь] → цена



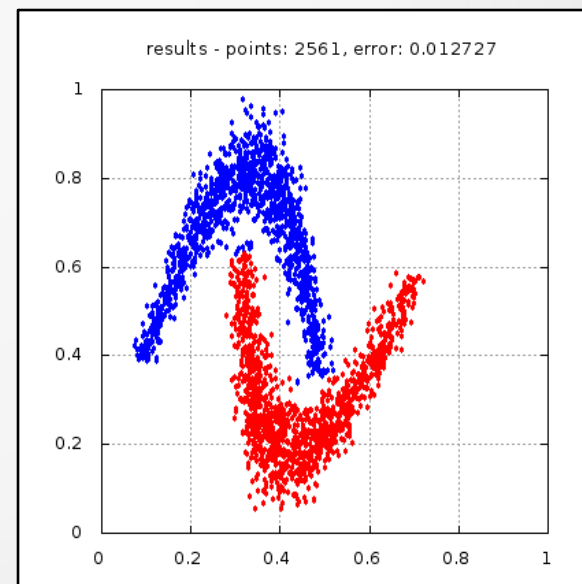
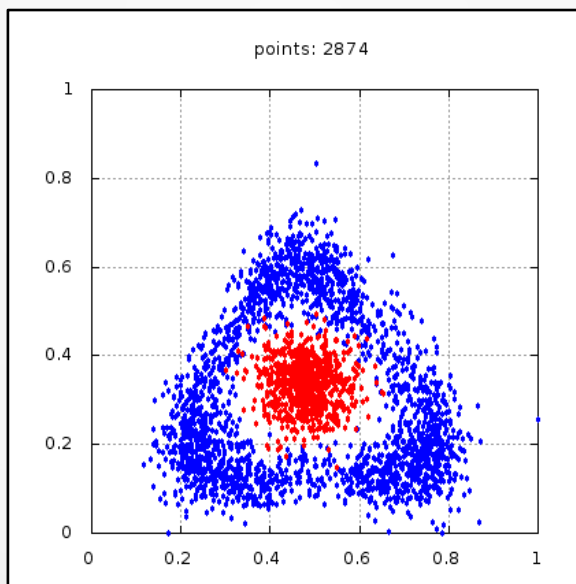
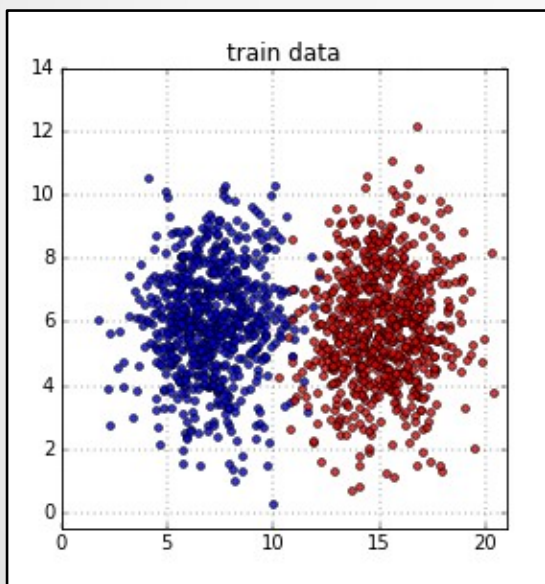
# ML: классификация

разделения объектов на классы

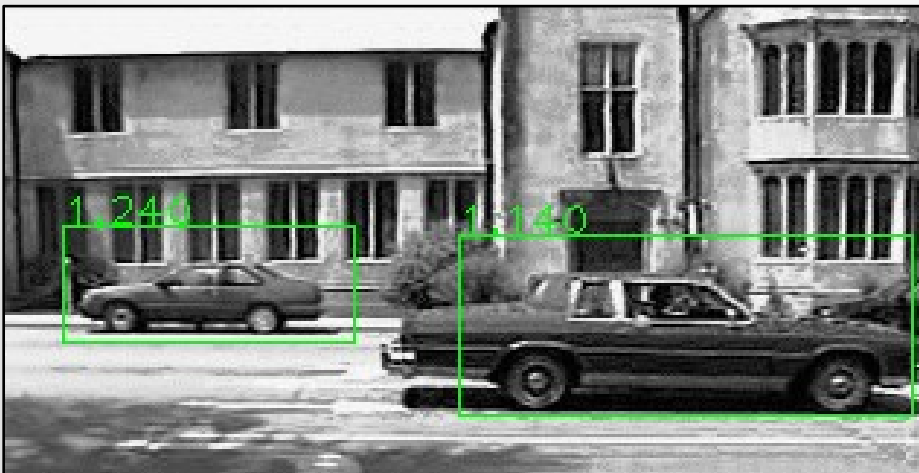
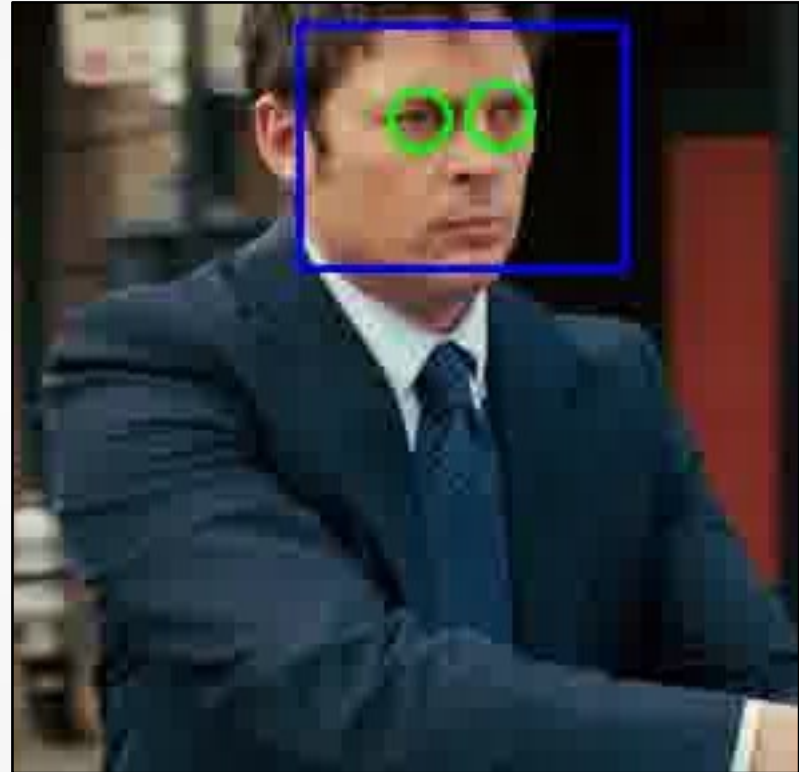
**Детектор котов:**



→ вектор-признак → есть/нет



# ML: Computer vision (CV)

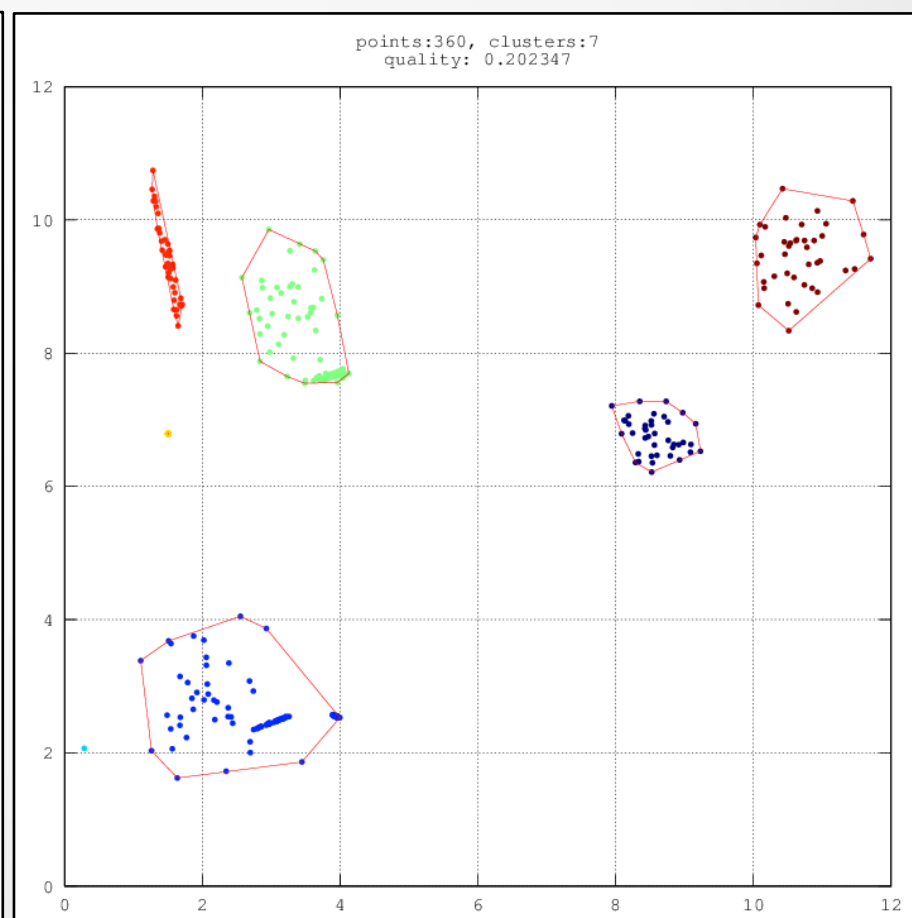
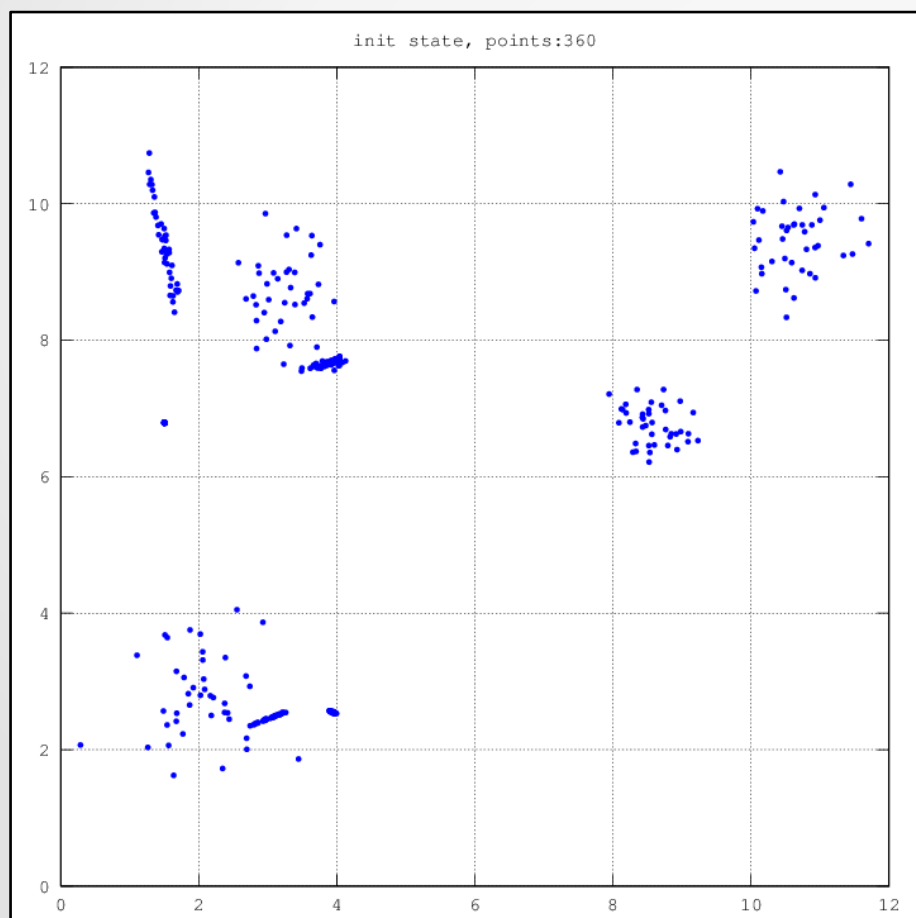




# ML: кластеризация

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 где там обучение?

## Схемы обучения

- с учителем - размеченные данные
- без учителя - данные не размечены
- частичное - данные размечены частично
- с подкреплением - данные заданы неявно

# ML: обучение «с учителем»

**Учебный набор:** [ объект, ответ ]

**Задача:** классификатор, регрессия

объект → вектор-признак → результат

**Обучение:** минимизация ошибки

ошибка = результат — правильный ответ

**Критерий остановки:**

достигнут порог значения ошибки,  
и/или порог количества циклов

# ML: обучение «без учителя»

**Учебный набор:** [ объект ]

**Задача:** кластеризация

объект → вектор-признак → результат

**Обучение:** изменение параметров

**Критерий остановки:**

состояние не изменяется,  
и/или порог количества циклов

# ML: частичное обучение

**Учебный набор:** [ объект, ответ ] + [ объект ]  
частично размечен

**Задача:** классификатор, регрессия, кластеризатор  
объект → вектор-признак → результат

**Обучение:** кластеризация + классификатор



# ML: обучение с подкреплением

## Учебный набор:

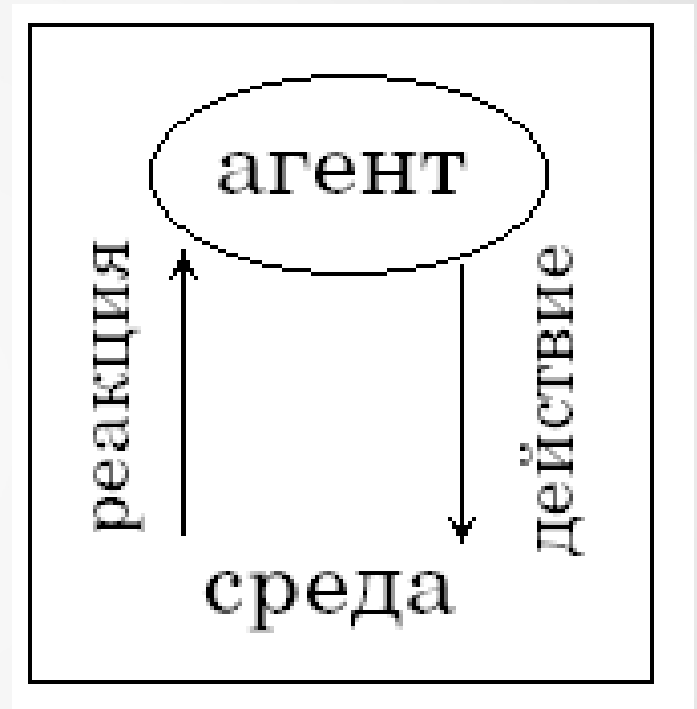
В ЯВНОМ ВИДЕ ОТСУТСТВУЕТ

## Задача: управление

- предсказываем реакцию среды
- выбираем оптимальное действие

## Обучение:

аппроксимация ф-ции оценки действия



## ML: и куда дальше?

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

# О работе в Data Science

## Технические Средства

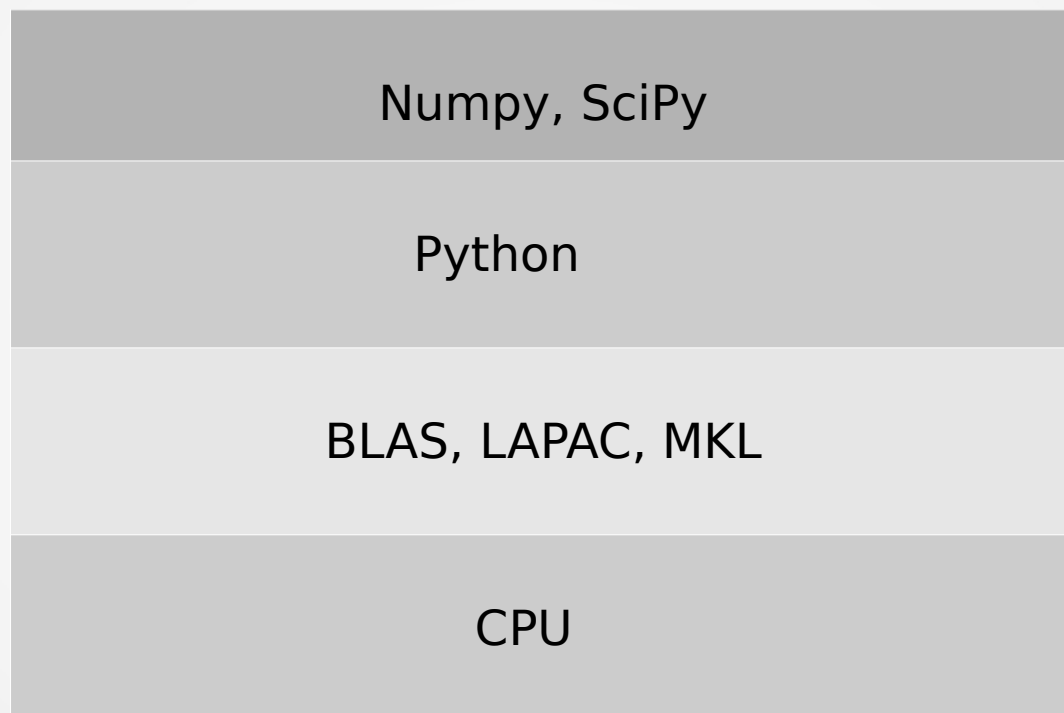
прикладные программные средства

вычислительные библиотеки

программный интерфейс с аппаратурой

аппаратные вычислительные средства

# ML: технические средства



# О работе в Data Science

## Технические Средства

Keras

Theano, TensorFlow

cuBLAS, cuDNN, CNMeM, NCCL

CUDA

NVIDIA GPU

# О работе в Data Science

## Технические Средства



Python



Jupyter

R



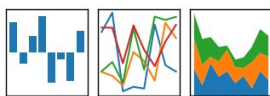
Keras



OpenCV

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



Theano

TensorFlow

theano

scikit-image

Numpy

Matplotlib



Scikit-Learn

NLTK

Pandas



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

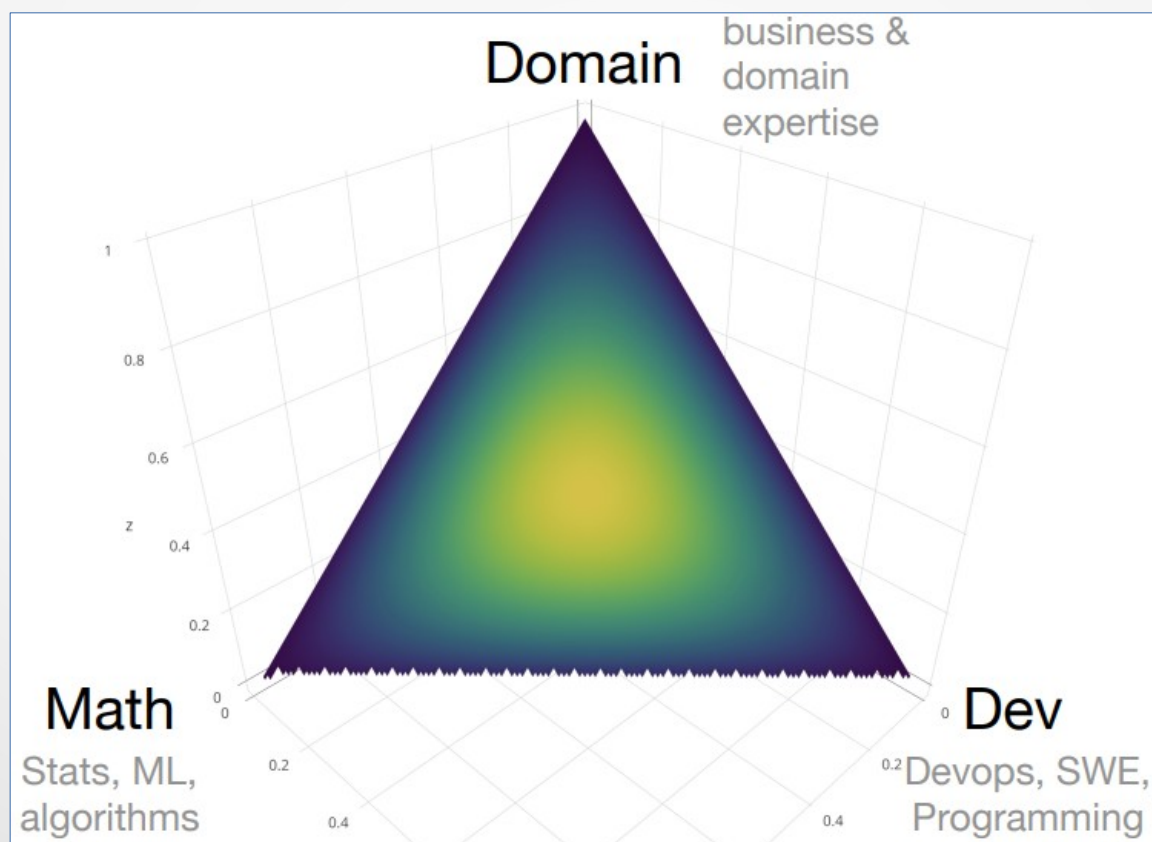
знания по специализации



# О работе в Data Science

## выбор специализации

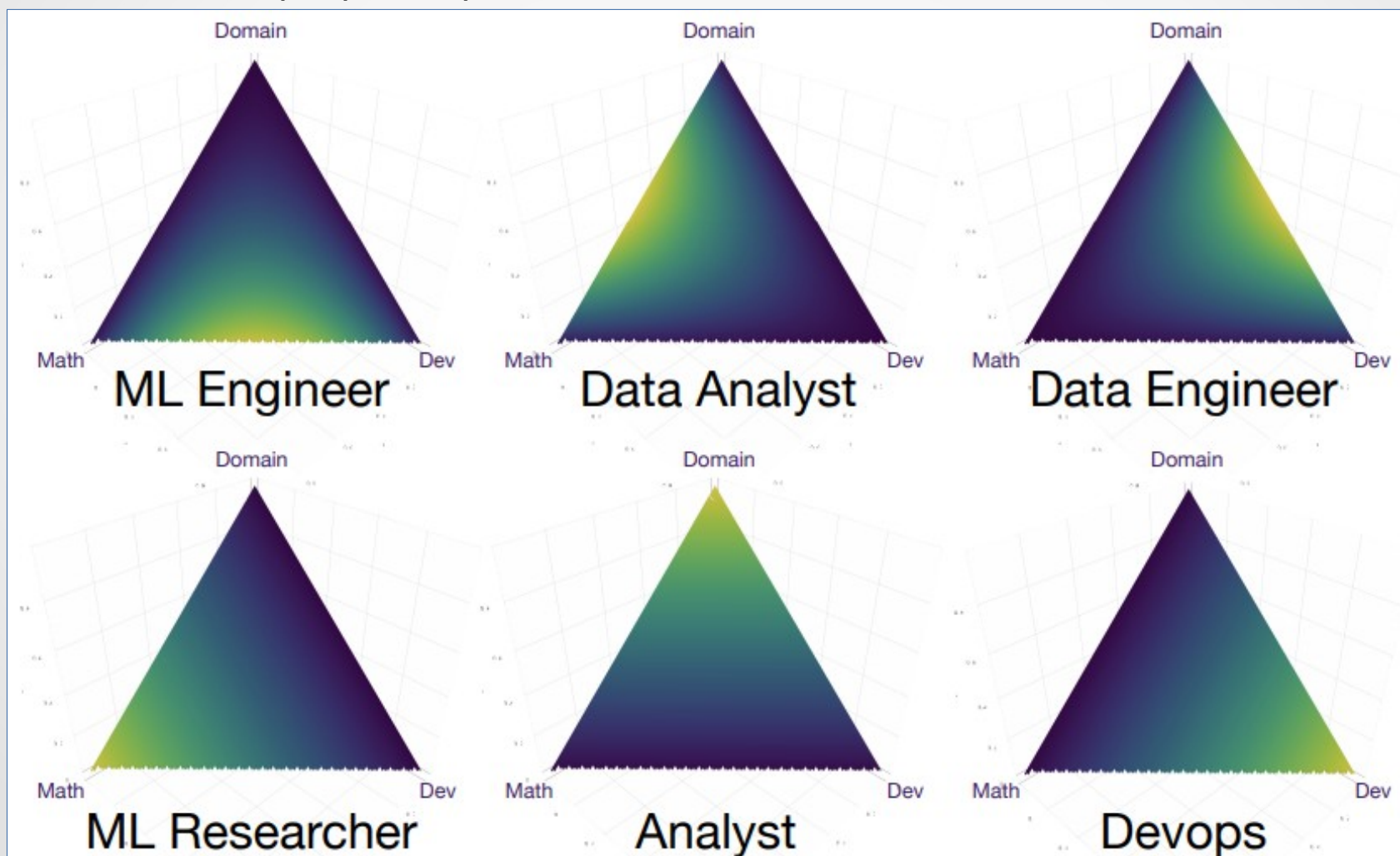
математика / программирование / хозяйственная деятельность



# О работе в Data Science

## выбор специализации

математика / программирование / хозяйственная деятельность



# О работе в Data Science

## Где ещё поучиться DS/ML



ШАД / МШАД Яндекс



Coursera



Kaggle

# ML: что почитать?

- Andrew Ng - Machine Learning
- Константин Воронцов - Машинное обучение
- Евгений Борисов - <http://mechanoid.kiev.ua>
- [http://github.com/mechanoid5/ml\\_lectorium](http://github.com/mechanoid5/ml_lectorium)

# О работе в Data Science

**Вопросы ?**