

# Математический анализ данных и машинное обучение

Лекция 1

Саркисян Вероника

# Коммуникация

Все материалы (презентации лекций, ноутбуки с семинаров, данные, домашние задания) будут выкладываться на [GitHub](#).

Вопросы можно задавать по почте ([impecopeco@gmail.com](mailto:impecopeco@gmail.com)) или в телеграме ([https://t.me/Combo\\_Breaker](https://t.me/Combo_Breaker)).

In case of emergency: 8-926-827-35-45

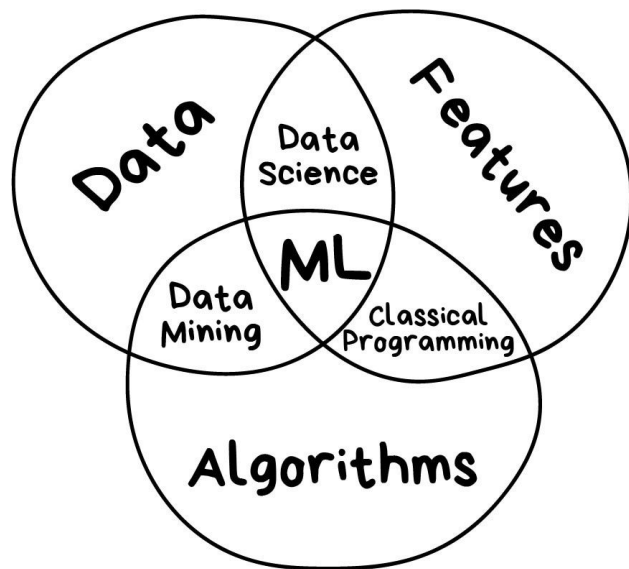
# Структура курса

Модуль 1	Вводная лекция, знакомство с инструментами. Задачи классификации и регрессии. Метрики качества. Kaggle.
Модуль 2	Линейные модели. Деревья, случайные леса. Проблемы в задачах машинного обучения.
Модуль 3	Предобработка данных, категориальные признаки. Ансамблевые методы.
Модуль 4	Презентация проектов

# План на сегодня

9:00 - 10:30	Вводная лекция: что такое машинное обучение?
10:45 - 12:15	Семинар: знакомимся с инструментами
12:15 - 14:00 (?)	Обед
14:00 - 15:30	Лекция: классификация и регрессия
16:00 - 18:00	Семинар: решаем задачи, знакомимся с Kaggle

# Что такое машинное обучение?



# В каких элементах на главной странице Яндекса применяются алгоритмы машинного обучения?

Москва

Настройка

veronica.fkn



Почта новых нет

Написать письмо

Диск

Получить Плюс

Новомосковский



На Хабарова при въезде в Град Московский строят какой то домик и парковку рядом с ним. А та же вид...

Konstantin



SOYUZ.RU



В сиквеле хоррора «Оно» будет один из самых странных моментов книги Стивена Кинга

Сейчас в СМИ в Москве 16 сентября, воскресенье 19:57

СМИ сообщили о двух новых подозреваемых в деле об отравлении Скрипалей

Кандидат от КПРФ стал лидером во втором туре выборов в Приморье

Украина создаст военную базу на Азовском море до конца года

В Петербурге прошел митинг против пенсионной реформы

Ротенберг уверен, что Крымский мост простоят сто лет

USD MOEX 68,05 -0,20 EUR MOEX 79,11 -0,67 НЕФТЬ 78,11 -0,26% ...



Детские товары

Проверенные скидки до 50% на Маркете

Яндекс

Видео Картинки Новости Карты Маркет Переводчик Музыка ТВ онлайн ещё

Найдётся всё. Например, когда цветет амброзия



Найти



Установите Яндекс.Браузер

MTS ТАРИФИЩЕ ДОМА И В ПОЕЗДКАХ ПО РОССИИ: БЕЗЛИМИТНЫЙ ИНТЕРНЕТ И ЗВОНОКИ НА ВСЕХ ОПЕРАТОРОВ ПОСМОТРЕТЬ 5M6

Погода

+14° Ночью +8, утром +11

Пробки

1 До конца дня дороги свободны

Карта Москвы

Метро Такси Расписания

Сообщество соседей

Посещаемое

Маркет — духи и туалетная вода

Авто.ру — пятилетние иномарки

Недвижимость — квартиры на 1 этаже

ТВ онлайн — смотреть МУЗ-ТВ

Деньги — оплата света и воды

Телепрограмма

ТВ онлайн

19:25 Лучше всех! Первый

19:27 Как извести любовницу... ТВ Центр

19:30 Спецы Пятый канал

19:30 Новости культуры с... Культура

Афиша

Хищник премьера

История одного назначения драма

Гоголь. Страшная месть детектив

Великий уравниль 2 боевик



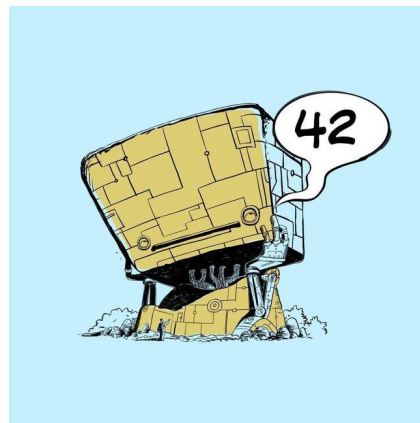
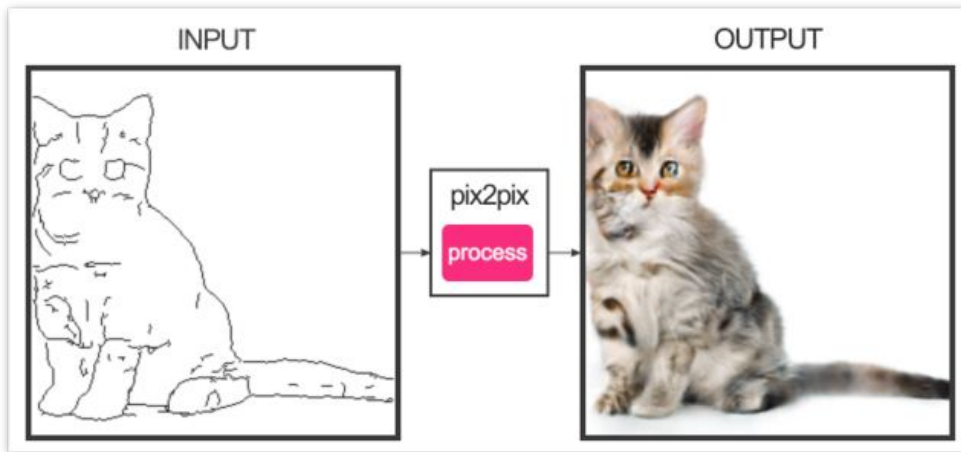
# Что “умеют” модели машинного обучения?

## Умеют:

- Предсказывать
- Извлекать зависимости из данных
- Обобщать
- Генерировать объекты, аналогичные данному

## Не умеют:

- Создавать качественно новые объекты
- Решать новый, неизвестный модели класс задач
- Отвечать на главный вопрос жизни, вселенной и всего такого

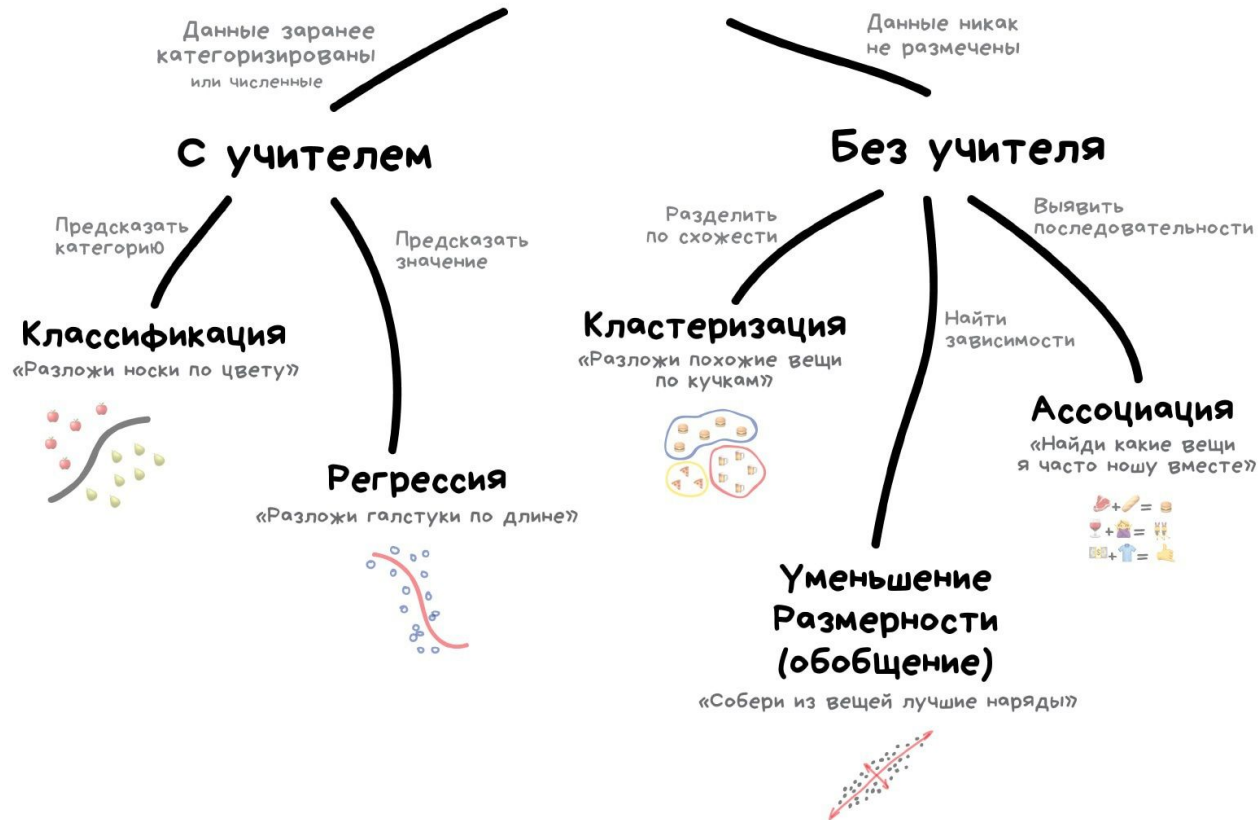


# “Зоопарк” моделей машинного обучения





# Классическое Обучение

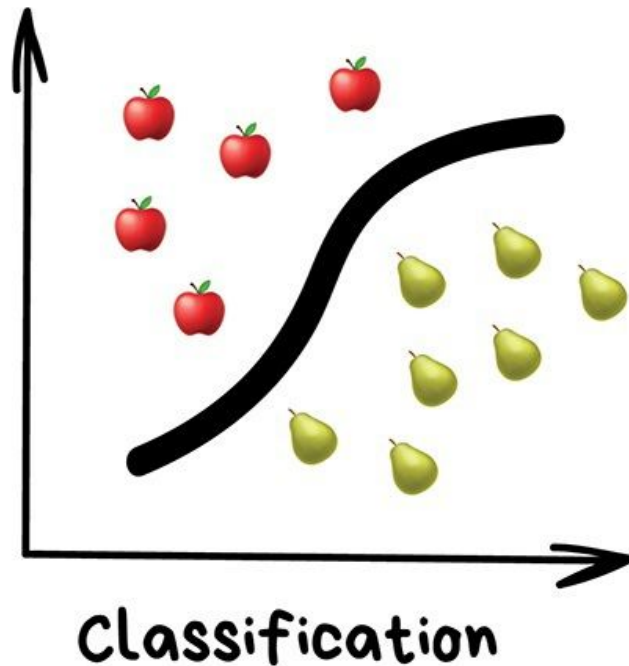


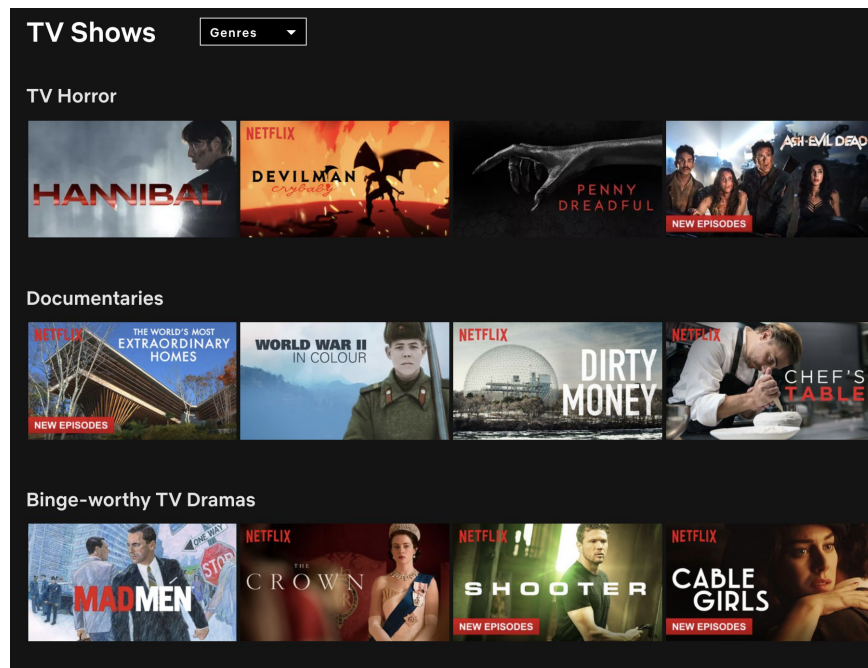
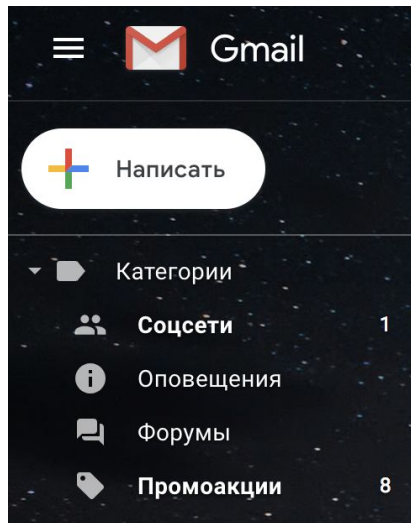
# Классификация

Разделяем объекты на заранее известные классы по набору признаков.

По количеству классов выделяют бинарную классификацию и классификацию с множественными классами (multi-label classification).

- Логистическая регрессия
- Наивный байесовский классификатор
- Деревья решений
- SVM
- KNN



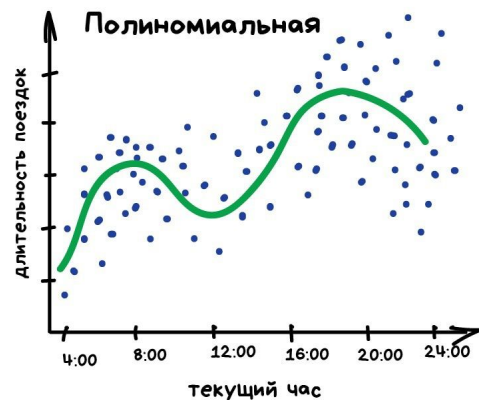
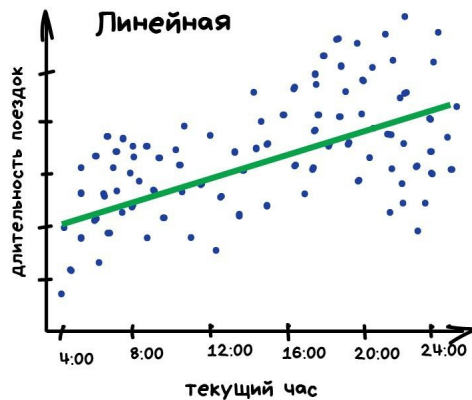


# Регрессия

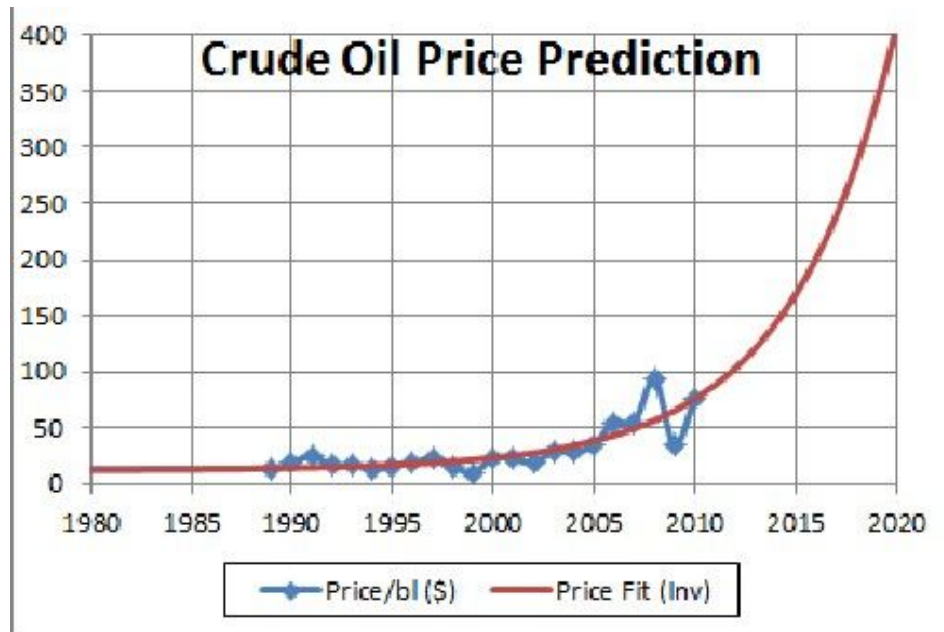
Предсказываем числовую переменную.

- Линейная регрессия
- Полиномиальная регрессия

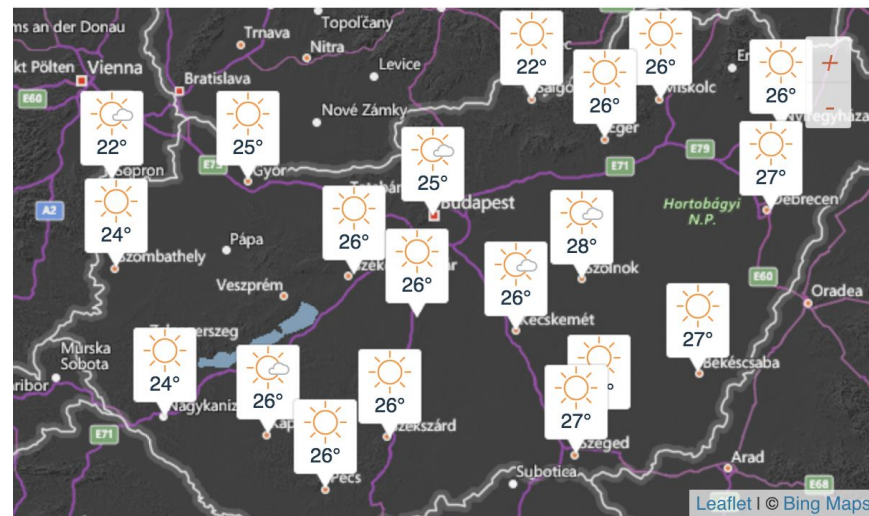
Предсказываем пробки



Регрессия



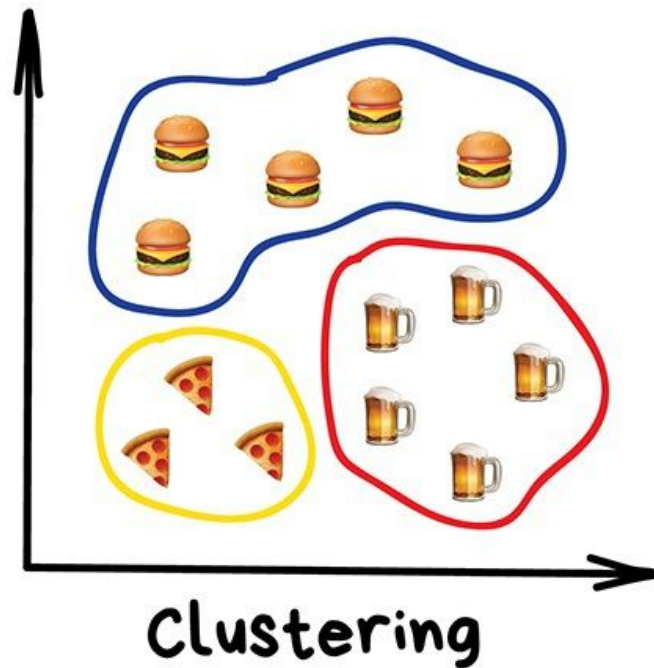
### Венгрия: карта погоды



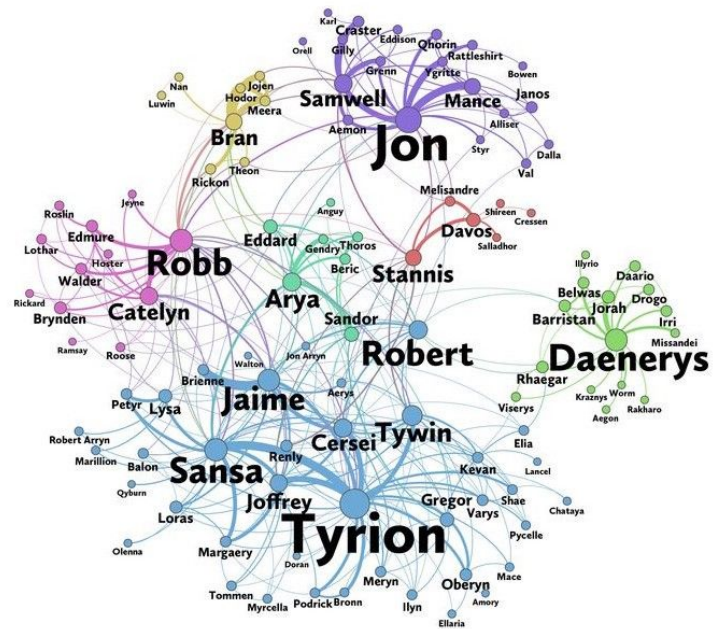
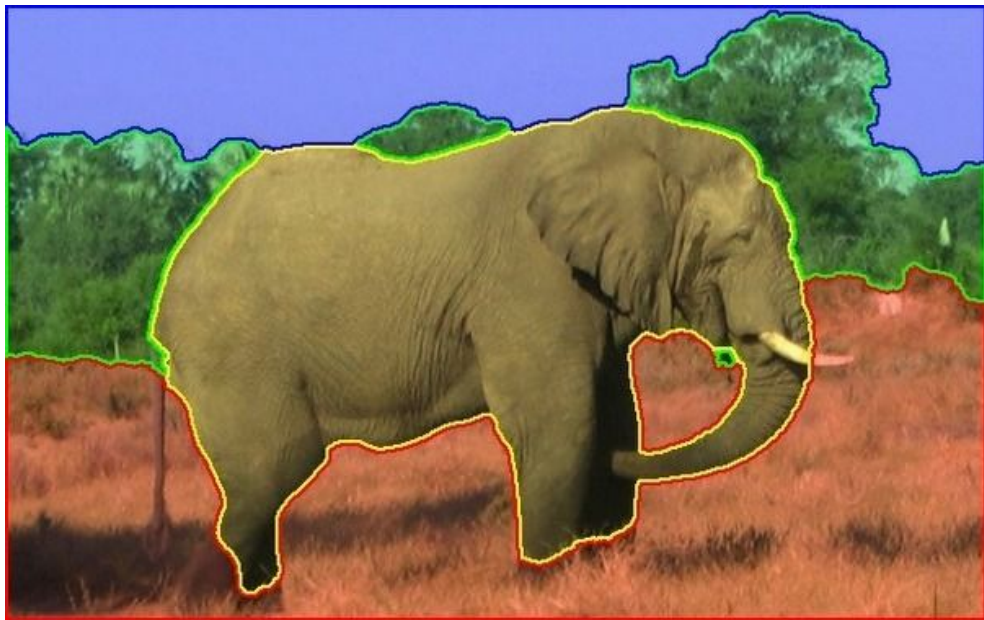
# Кластеризация

Разделяем объекты на классы, но истинных меток объектов не существует (или они нам не известны).

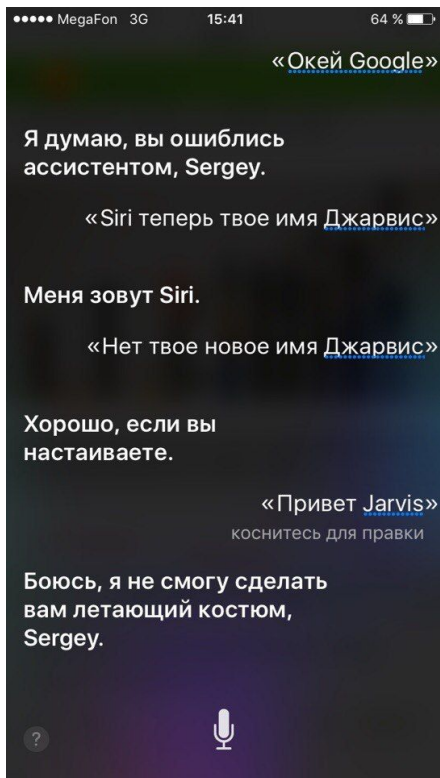
- K-means
- Mean-Shift
- [DBSCAN](#)



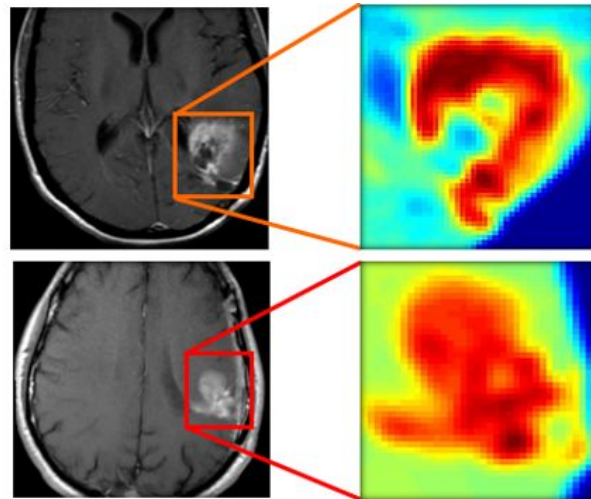




... и многое другое!



DIAGNOSTICS







## Поиск

Ответы на любые вопросы



## Картинки

Изображения всех цветов и размеров



## Видео

Просмотр фильмов, сериалов, телешоу, музыкальных роликов



## Новости

Картина дня, созданная автоматически



## Погода

Прогноз в вашем городе и по всему миру



## Карты

Рекомендации где поесть, куда сходить и чем заняться



## Почта

Электронный ящик без спама и вирусов



## Маркет

Товары, сравнение цен, отзывы покупателей



## Яндекс.Браузер

Простой и безопасный интернет



## Афиша

Развлекательные мероприятия



## Такси

Свободные водители поблизости



## Музыка

Персональные рекомендации



## Деньги

Онлайн-платежи и электронный кошелек



## Диск

Безопасное облако для ваших файлов



## Недвижимость

Объявления о комнатах, квартирах и домах



## Авто.ру

Огромный выбор новых и поддержанных автомобилей



## Авиабилеты

Большой выбор предложений от авиакомпаний и агентств



## Работа

Подбор вакансий с популярных сайтов поиска работы



## Дзен

Публикации на основе ваших интересов



## Коллекции

Ваше избранное в Яндексе: картинки, видео и многое другое



## КиноПоиск

Сервис для выбора фильмов и сериалов

# Постановка задачи машинного обучения

$X$  - множество объектов; каждый объект  $x \in X$  представлен вектором признаков  $(f_1, f_2, \dots, f_n)$

$Y$  - множество допустимых ответов

$L(y, \hat{y})$  - функция ошибки (аргументы - истинные ответы и оцененные); по умолчанию чем больше значение функции, тем больше ошибка алгоритма.

Задача состоит в том, чтобы построить алгоритм  $a: X \rightarrow Y$ , минимизирующий функцию ошибки  $L$ .

# Матрица объекты-признаки

	A	B	C	D	E	F	G	H	I	J	K	L
1	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
3	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
4	3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 31	7.925		S
5	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
6	5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
7	6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
8	7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
9	8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
10	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S
11	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C
12	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S
13	12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
14	13	0	3	Saunders, Mr. William Henry	male	20	0	0	A/5. 2151	8.05		S
15	14	0	3	Andersson, Mr. Anders Johan	male	39	1	5	347082	31.275		S
16	15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14	0	0	350406	7.8542		S
17	16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55	0	0	248706	16		S
18	17	0	3	Rice, Master. Eugene	male	2	4	1	382652	29.125		Q
19	18	1	2	Williams, Mr. Charles Eugene	male		0	0	244373	13		S
20	19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)	female	31	1	0	345763	18		S
21	20	1	3	Masselmani, Mrs. Fatima	female		0	0	2649	7.225		C
22	21	0	2	Fynney, Mr. Joseph J	male	35	0	0	239865	26		S
23	22	1	2	Beesley, Mr. Lawrence	male	34	0	0	248698	13	D56	S
24	23	1	3	McGowan, Miss. Anna "Annie"	female	15	0	0	330923	8.0292		Q
25	24	1	1	Sloper, Mr. William Thompson	male	28	0	0	113788	35.5	A6	S
26	25	0	3	Palsson, Miss. Torborg Danira	female	8	3	1	349909	21.075		S
27	26	1	3	Asplund, Mrs. Carl Oscar (Selma Augusta Emilia Johansson)	female	38	1	5	347077	31.3875		S
28	27	0	3	Emir, Mr. Farred Chehab	male		0	0	2631	7.225		C

**Соотнесите данные задачи с задачами машинного обучения: назовите тип задачи, возможные признаки и модели решения.**

1. Определение возраста (пола) человека по фотографии
2. Кредитный скоринг (оценка кредитоспособности клиента)
3. Распознавание рукописного текста
4. Рекомендации в онлайн-магазине
5. Определение тональности текста
6. Разделение пользователей форума на группы по интересам
7. Фильтрация спама
8. Определение жанра фильма

# Инструменты

# GitHub

[Хороший пошаговый tutorial](#) о том как использовать Git из командной строки.

Combo-Breaker / ML\_course

Watch 0Star 0Fork 0

<> Code

Issues 0

Pull requests 0

Projects 0

Wiki

Insights

Settings

Курс МАДМО в Сбербанке.

Edit

[Manage topics](#)

8 commits

1 branch

0 releases

1 contributor

Branch: master

New pull request

Create new file

Upload files

Find file

Clone or download

Combo-Breaker Wine Dataset

Latest commit df44a58 19 minutes ago

Lecture

Create План

23 minutes ago

Seminar

Wine Dataset

19 minutes ago

README.md

Update README.md

3 days ago

README.md

ML\_course

Математический анализ данных и машинное обучение.

# Python

NumPy - библиотека для удобной работы с векторами, матрицами и т.д.

Pandas - библиотека для работы с датасетами.


Matplotlib - библиотека для визуализации данных.

Sklearn - библиотека с реализацией основных алгоритмов машинного обучения.

[Anaconda](#)

# Jupyter Notebook

## Jupyter Notebook

Jupyter English\_bilstm\_NER Last Checkpoint: 14.12.2017 (autosaved)  Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Code

```
In [1]: import pandas as pd

df = pd.read_csv('ner.csv', encoding = "ISO-8859-1", error_bad_lines=False)

df.head()
```

b'Skipping line 281837: expected 25 fields, saw 34\n'

Out[1]:

	Unnamed: 0	lemma	next-lemma	next-next-lemma	next-next-pos	next-next-shape	next-next-word	next-pos	next-shape	next-word	...	prev-prev-lemma	prev-prev-pos	prev-prev-shape	prev-prev-word
0	0	thousand	of	demonstr	NNS	lowercase	demonstrators	IN	lowercase	of	...	__start2__	__START2__	wildcard	__START
1	1	of	demonstr	have	VBP	lowercase	have	NNS	lowercase	demonstrators	...	__start1__	__START1__	wildcard	__START
2	2	demonstr	have	march	VBN	lowercase	marched	VBP	lowercase	have	...	thousand	NNS	capitalized	Thousai
3	3	have	march	through	IN	lowercase	through	VBN	lowercase	marched	...	of	IN	lowercase	
4	4	march	through	london	NNP	capitalized	London	IN	lowercase	through	...	demonstr	NNS	lowercase	demonstrat

5 rows x 25 columns