

Машинное обучение

Лекция 1

Введение в машинное обучение. KNN.

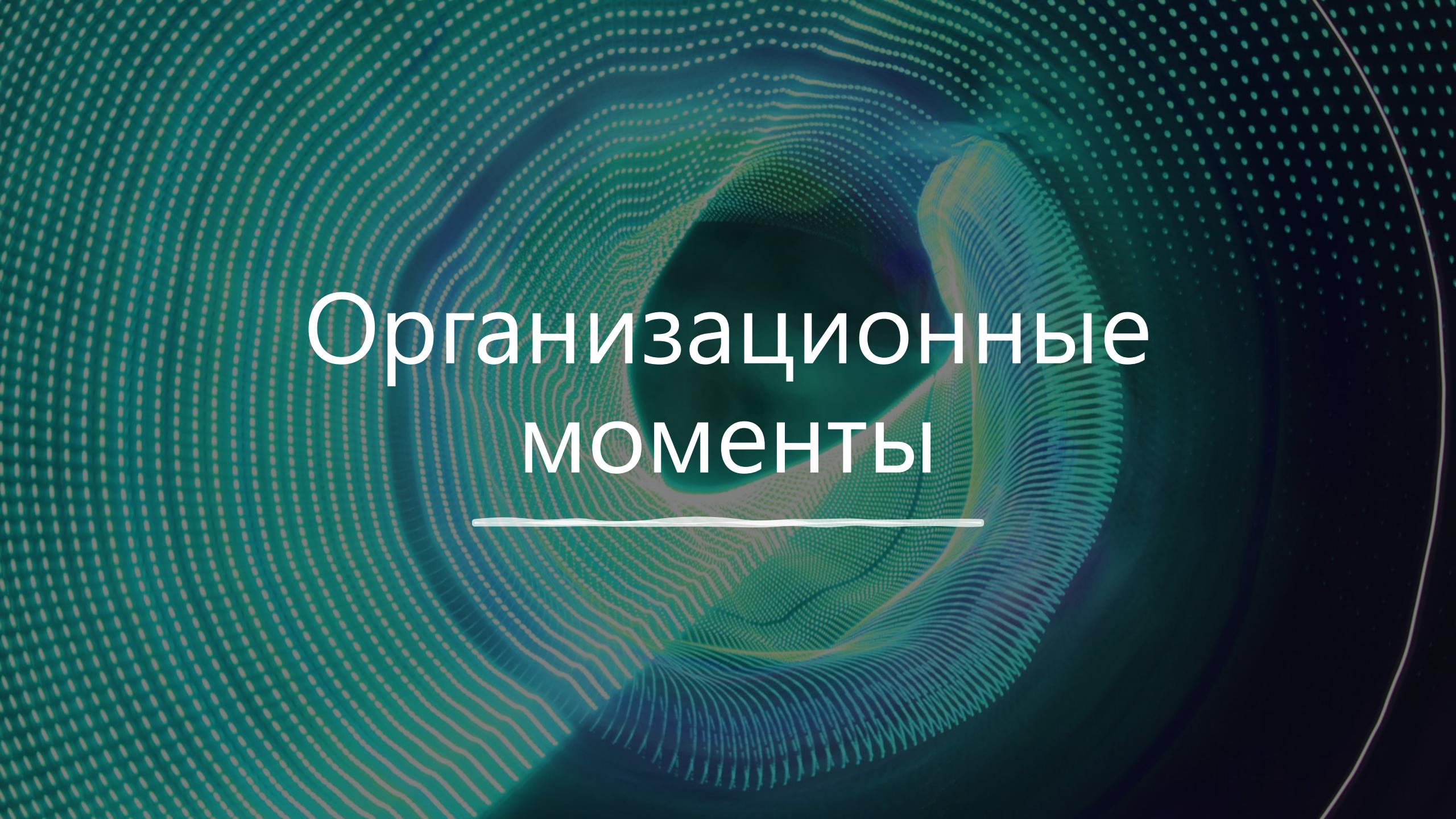
Михаил Гущин

mhushchyn@hse.ru

НИУ ВШЭ, 2023



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ



Организационные моменты

Наша команда

- ▶ Лектор:
 - Михаил Гущин
- ▶ Семинаристы:
 - Владимир Бочарников
 - Сергей Корпачев
- ▶ Ассистенты:
 - Софья Пирогова
 - Артём Станкевич

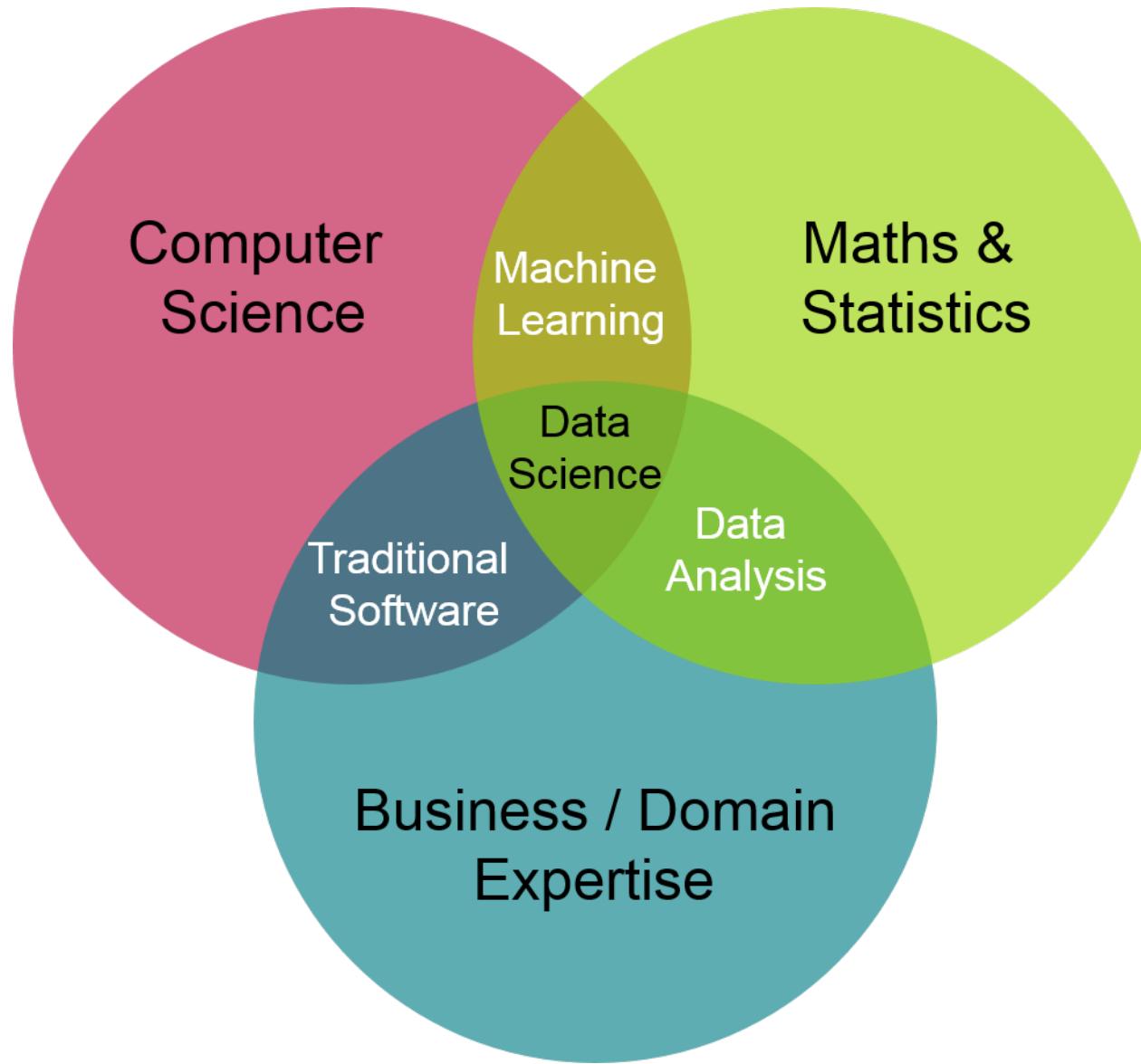
Материалы курса

- ▶ Вики курса:
 - материалы лекций и семинаров
 - ссылки на доп. курсы и книги
- ▶ Чат в ТГ
- ▶ Лекции
 - по вторникам в 16:20 в R205
- ▶ Семинары
 - по средам в 14:40 в R506
 - по субботам в 14:40 в (уточняется)



Оценки

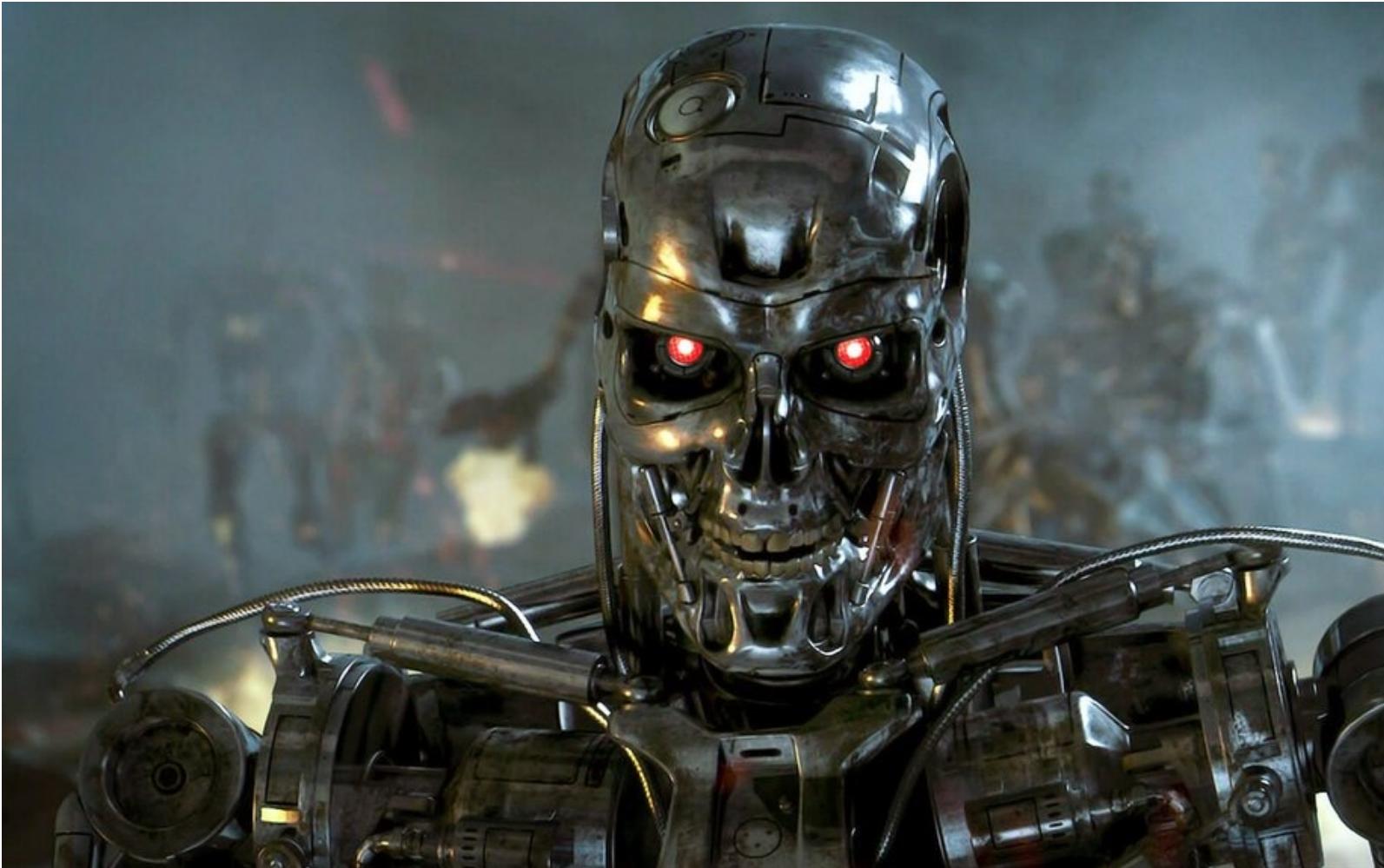
- ▶ Домашние задания
 - Практические задачи на программирование
 - Теоретические задачи
 - Мягкий и жесткий дедлайны
- ▶ Контрольная работа в ноябре
 - Теоретические вопросы + задачи
- ▶ Экзамен в конце курса
 - Теоретические вопросы + задачи
- ▶ Правила выставления оценок:
$$\text{Итоговая} = \text{Округление} (0.5 * \text{ДЗ} + 0.2 * \text{КР} + 0.3 * \text{Э})$$



ИИ в кино



Терминатор



Я, робот



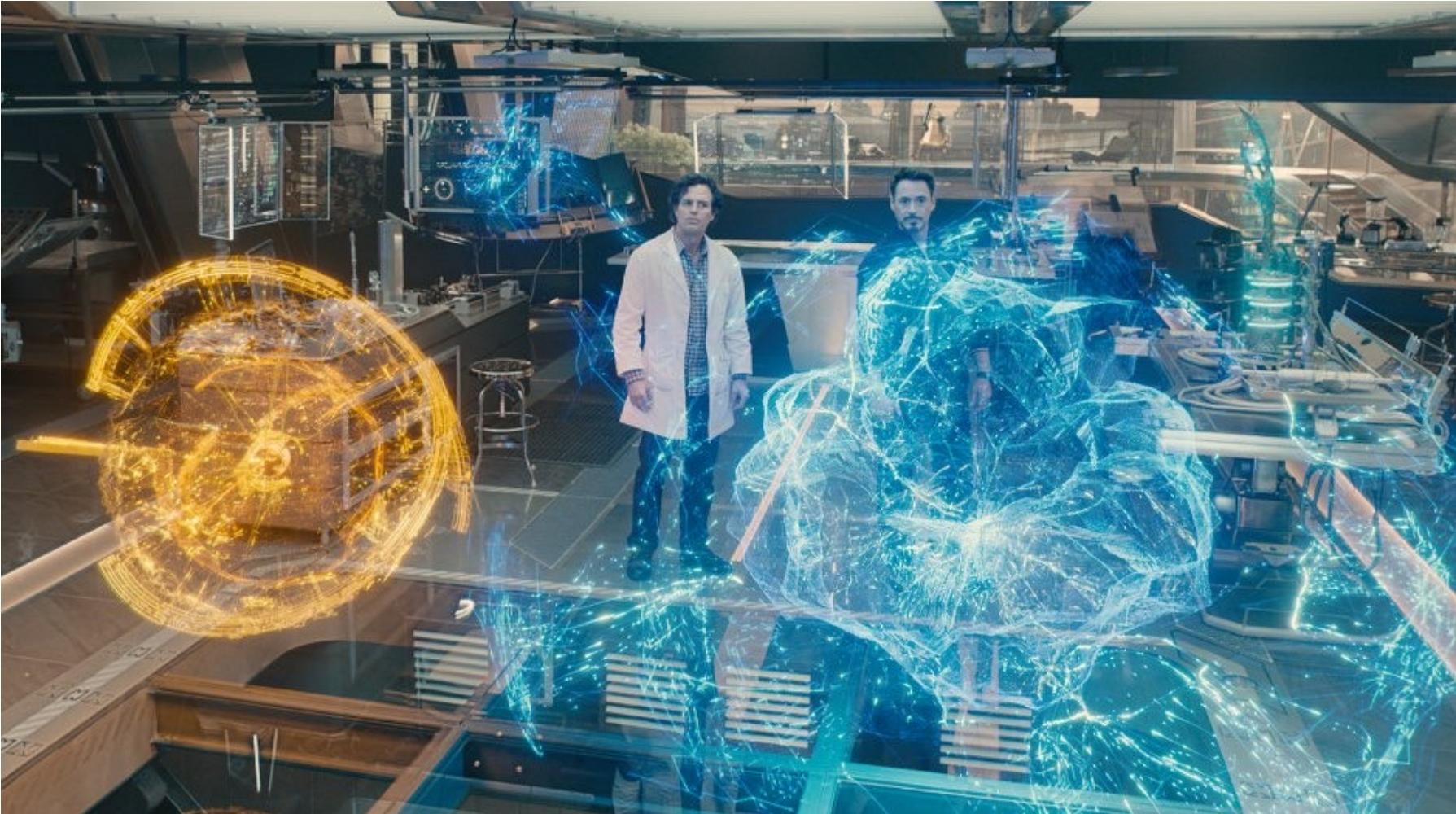
ВАЛЛ-И



Превосходство



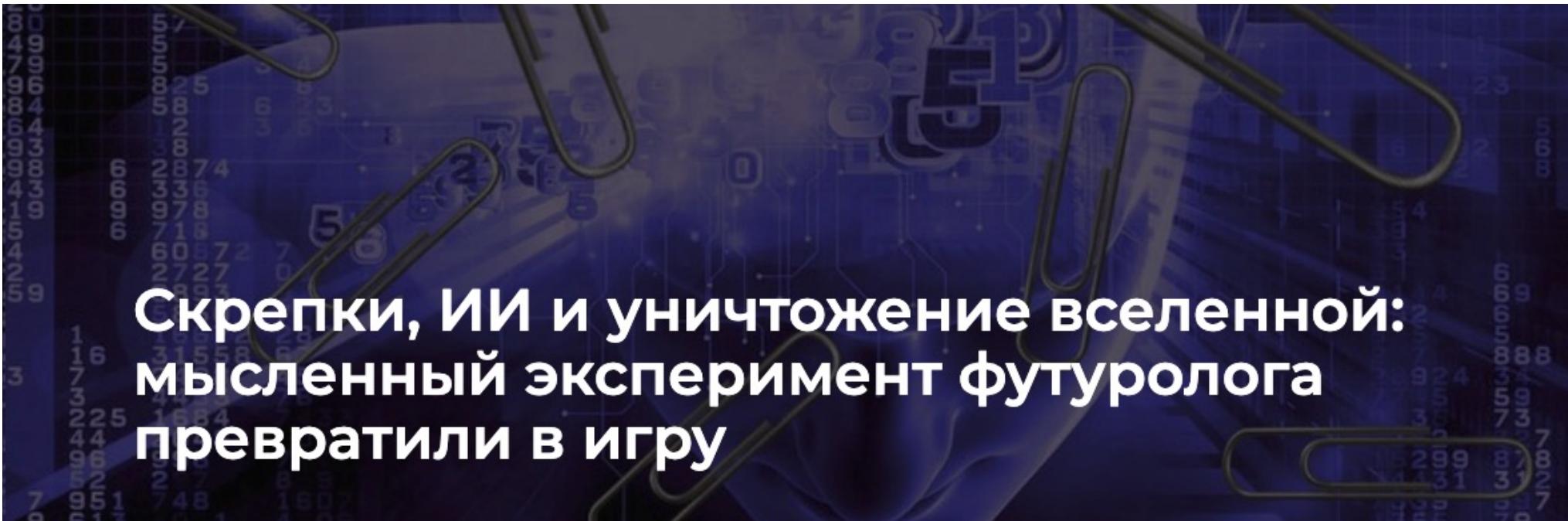
Железный человек / Мстители



Индекс страха



ИИ и скрепки

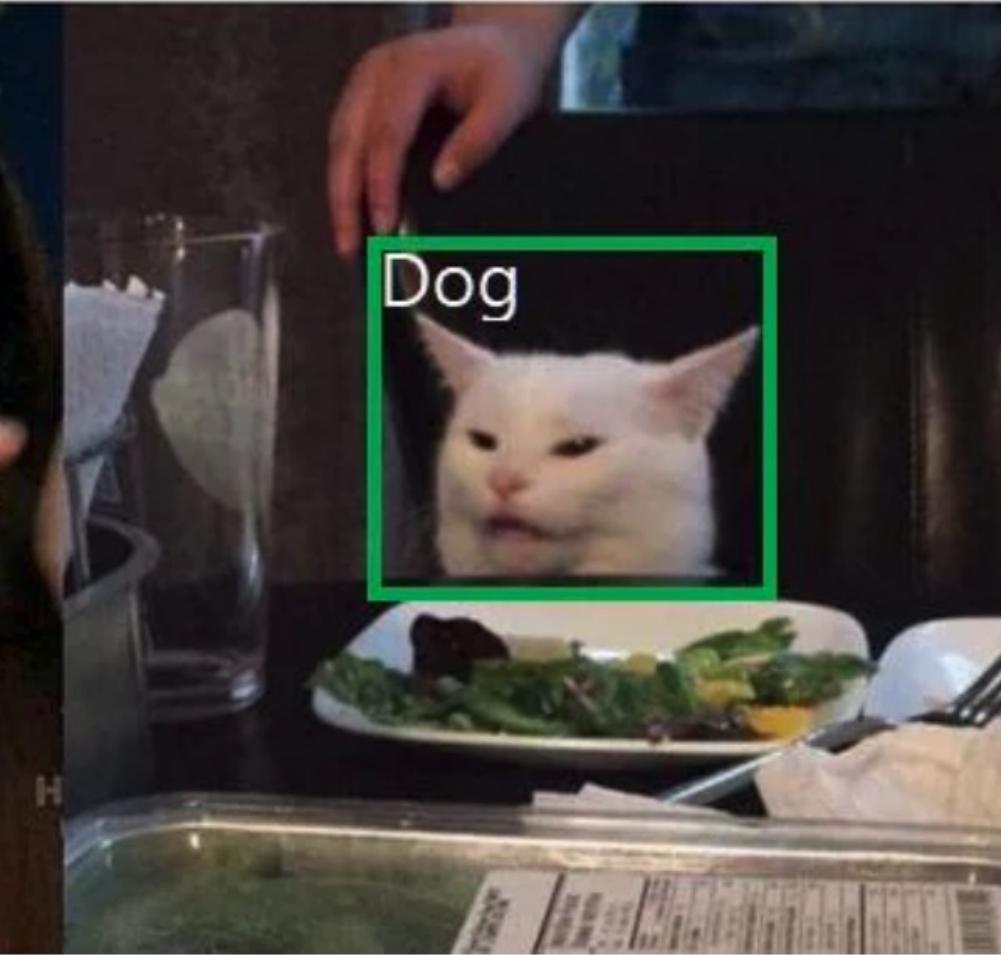


- ▶ Статья: <https://rb.ru/story/universal-paperclips/>
 - ▶ Игра-кликер: <http://www.decisionproblem.com/paperclips/index2.html>

People with no idea
about AI, telling me my
AI will destroy the world



Me wondering why my
neural network is
classifying a cat as a dog..



Реальные приложения



Поиск информации

Google machine learning X | 🔍

All Images Videos News Books More Tools

About 2,080,000,000 results (0.54 seconds)

Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values.

<https://www.techtarget.com/searchenterpriseai/definition> :: [What Is Machine Learning and Why Is It Important? - TechTarget](#)



Unsupervised Learning: Clustering, Dimensionality Reduction, Anomaly Detection, Association Rule Mining, Recommendation Systems.

Supervised Learning: Classification, Regression, Structured Prediction, Unstructured Prediction, Generative Models.

Reinforcement Learning: Reinforcement Learning, Policy Iteration, Value Function Iteration, Actor-Critic Methods.

Feedback About featured snippets · Feedback

https://en.wikipedia.org/wiki/Machine_learning :: [Machine learning - Wikipedia](#)

Machine learning (ML) is a field of inquiry devoted to understanding and building methods that 'learn', that is, methods that leverage data to improve ...

[Machine Learning \(journal\)](#) · [Machine learning control](#) · [Active learning \(machine...](#)

See results about

machine learning Dictionary definition >

People also ask :

What is machine learning with example? ▾

What are the 3 types of learning in machine learning? ▾

What is AI vs machine learning? ▾

Why is machine learning used? ▾

Feedback

Голосовые помощники

The screenshot shows the Yandex.Alice mobile application. At the top center is the Alice logo, a stylized white egg inside a purple circle. Below it is a large white speech bubble containing the text "Привет, я Алиса!". The background is a solid purple color. In the center, the text "Я готова помочь" is displayed in white. Below this, there are four rows of cards, each consisting of an icon and text. The icons are white with a blue outline. The rows are separated by thin horizontal lines.

| | | | | | | | | |
|--|-------------------|---|--|---------------------|---|--|-------------------|---|
| | Определить песню | > | | Узнать, что на фото | > | | Включить сказку | > |
| | Одеться по погоде | > | | Поиграть | > | | Построить маршрут | > |
| | Вызвать такси | > | | Найти нужное место | > | | Купить на Беру | > |

Машинный перевод

Яндекс Браузер

Яндекс.Браузер обновился. Версия 21.8.2

1

The screenshot shows the Yandex Browser homepage. At the top left is the browser logo and name. At the top right are update notifications for 'Яндекс.Браузер обновился. Версия 21.8.2' and like/dislike buttons. The main content area features a video player with a play button. A speech bubble from a character named David says: 'Hi! I'm David and I lead the NLP team at Yandex'. Below the video, there's a large heading 'Закадровый перевод видео с английского' (Subtitles for video from English) and a text block explaining the feature: 'Нейросети Яндекса научились сами переводить и озвучивать видео на английском языке. Пока — не везде, но уже скоро любой ролик на английском можно будет смотреть на русском.' It also includes a link to try it out. To the left of the video player is a small thumbnail of the video frame. Below the video player are social sharing links for Dzen, VKontakte, Twitter, and Telegram. At the bottom, there's a note about DRM protection and a copyright notice.

Яндекс Браузер обновился. Версия 21.8.2

Hi! I'm David and I lead the NLP team at Yandex

Закадровый перевод видео с английского

Нейросети Яндекса научились сами переводить и озвучивать видео на английском языке. Пока — не везде, но уже скоро любой ролик на английском можно будет смотреть на русском.

Сразу попробовать новую функцию можно [по ссылке](#).

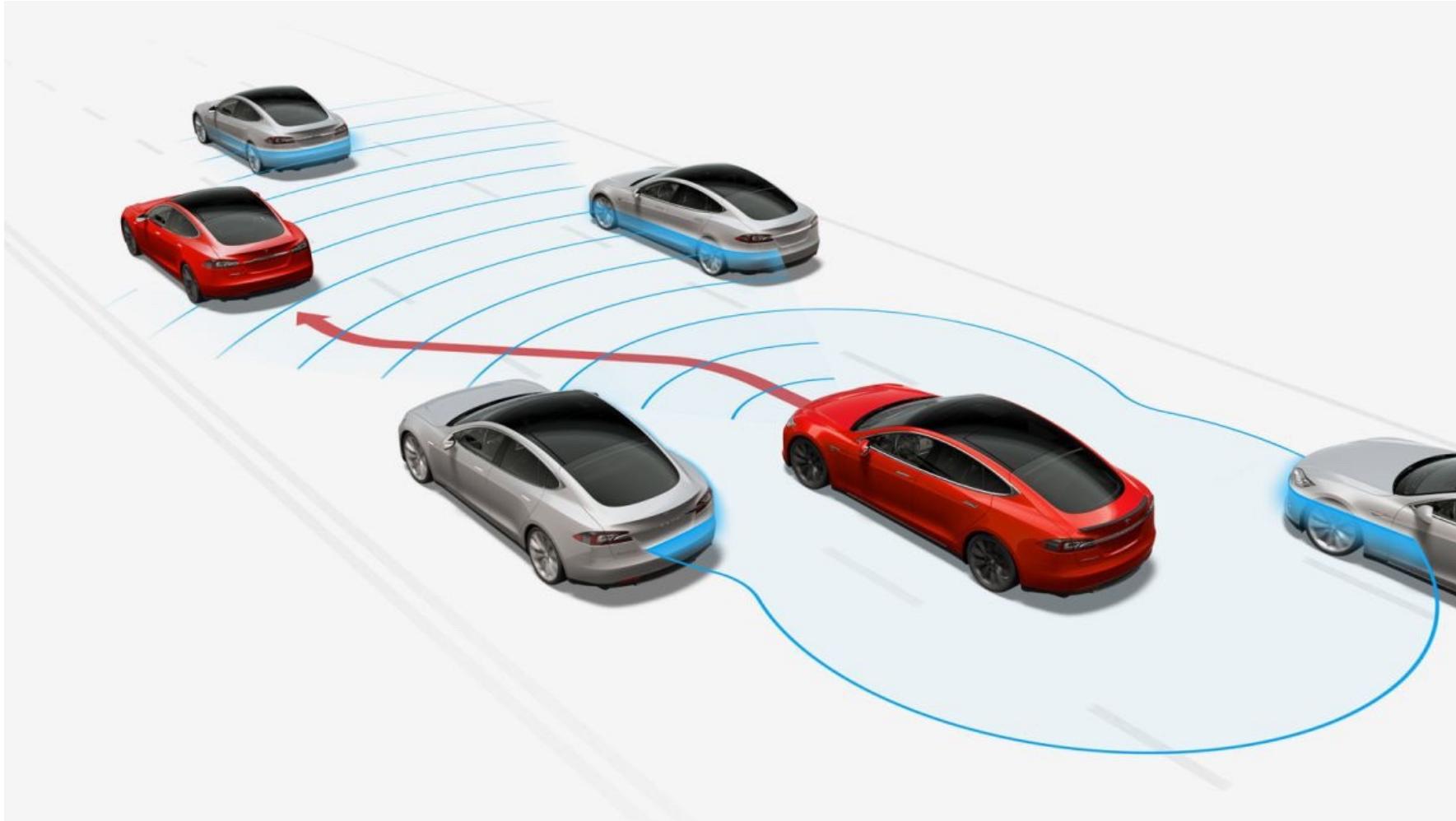
Как включить перевод видео?

Подписывайтесь на новости Яндекс.Браузера:

Дзен ВКонтакте Твиттер Телеграм

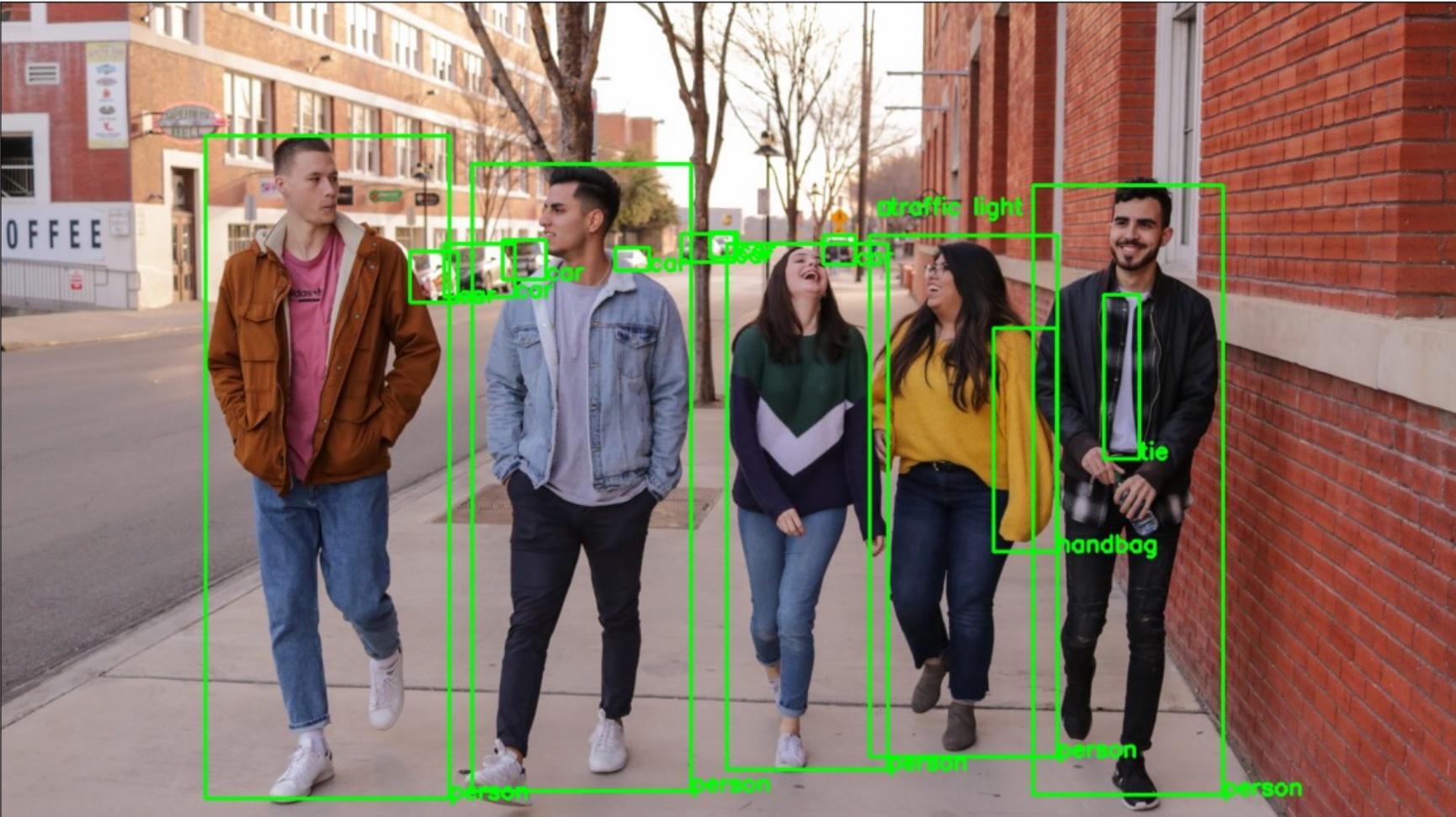
Перевод не доступен для видео, у которых есть технические средства защиты авторских прав (DRM).

Автопилоты



<https://moscowteslaclub.ru>

Анализ видео



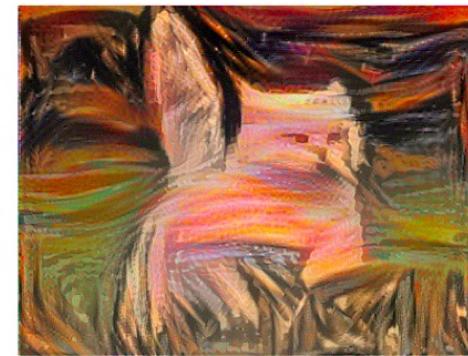
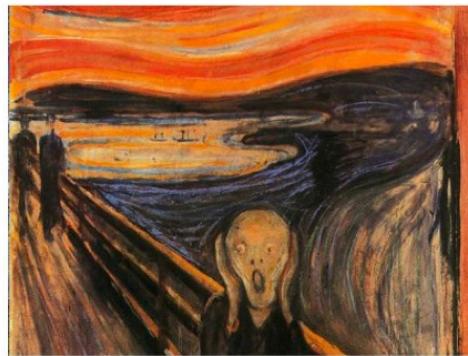
Перенос стиля изображения



+



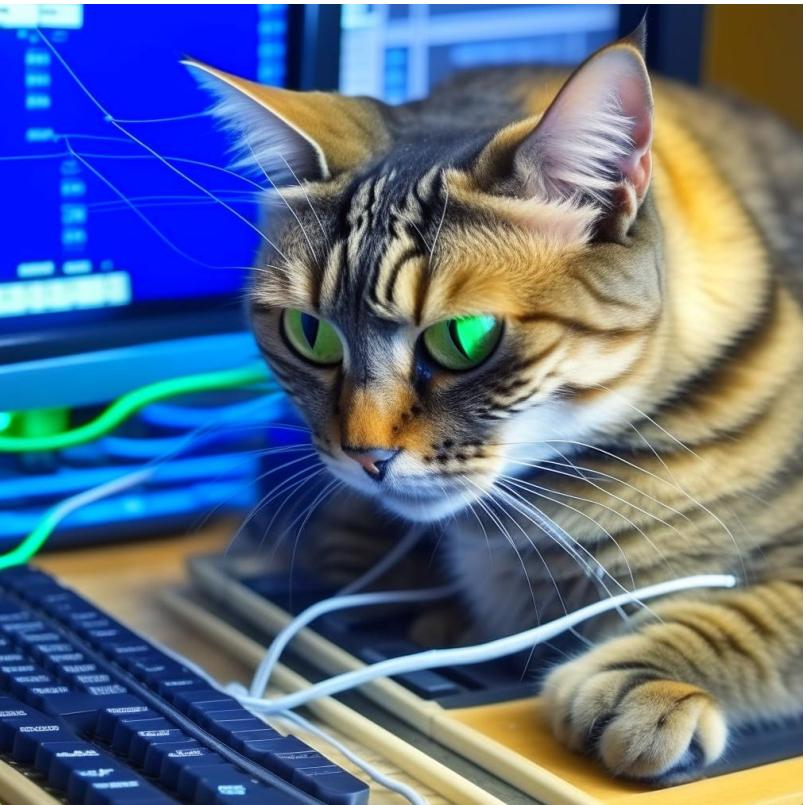
+



+



Генерация изображения по описанию



Кот фиксит баг в обучении
нейронной сети



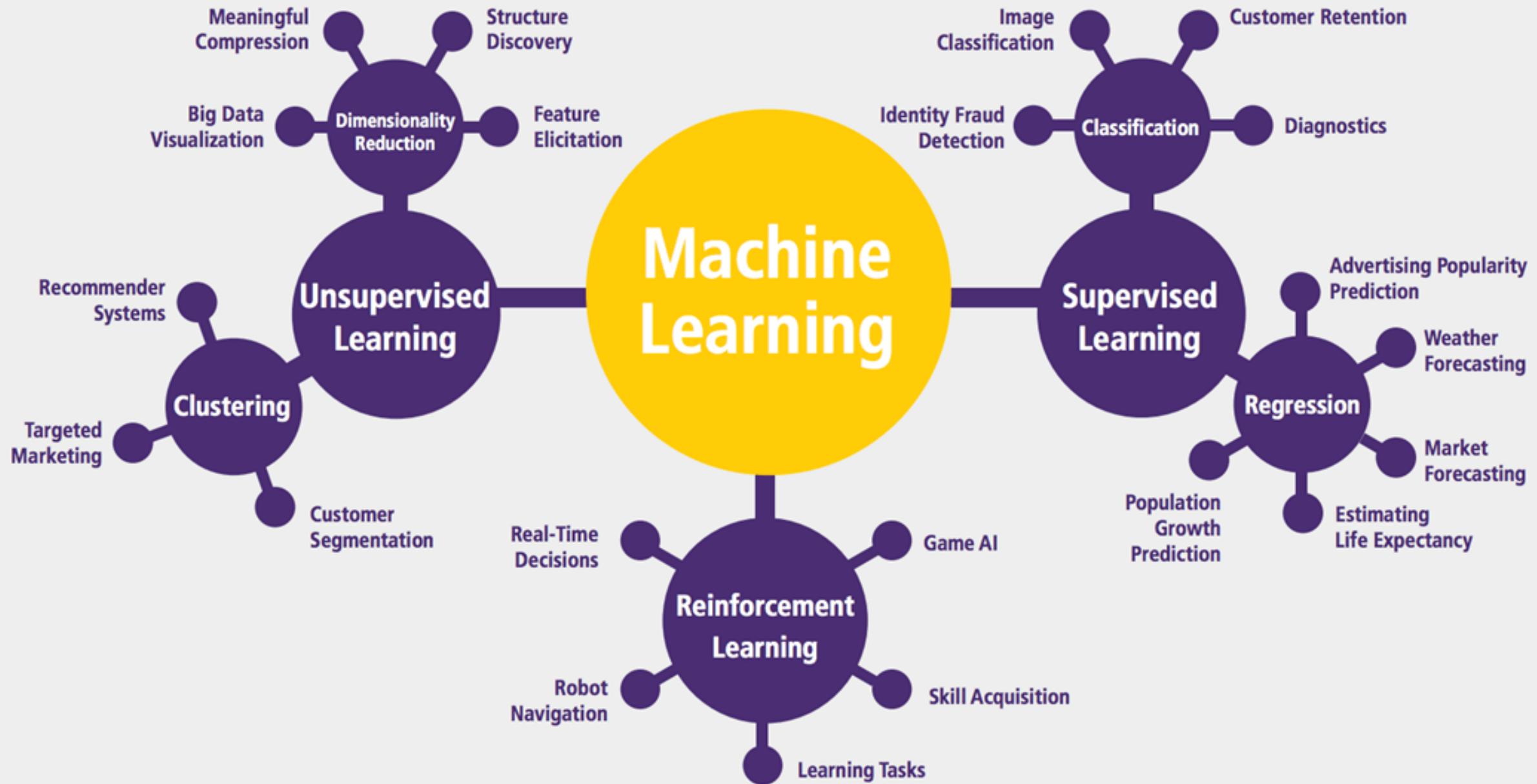
Розовый фламинго стоит на
одной ноге в воде

Ссылка: <https://www.sberbank.com/promo/kandinsky>

Deepfake

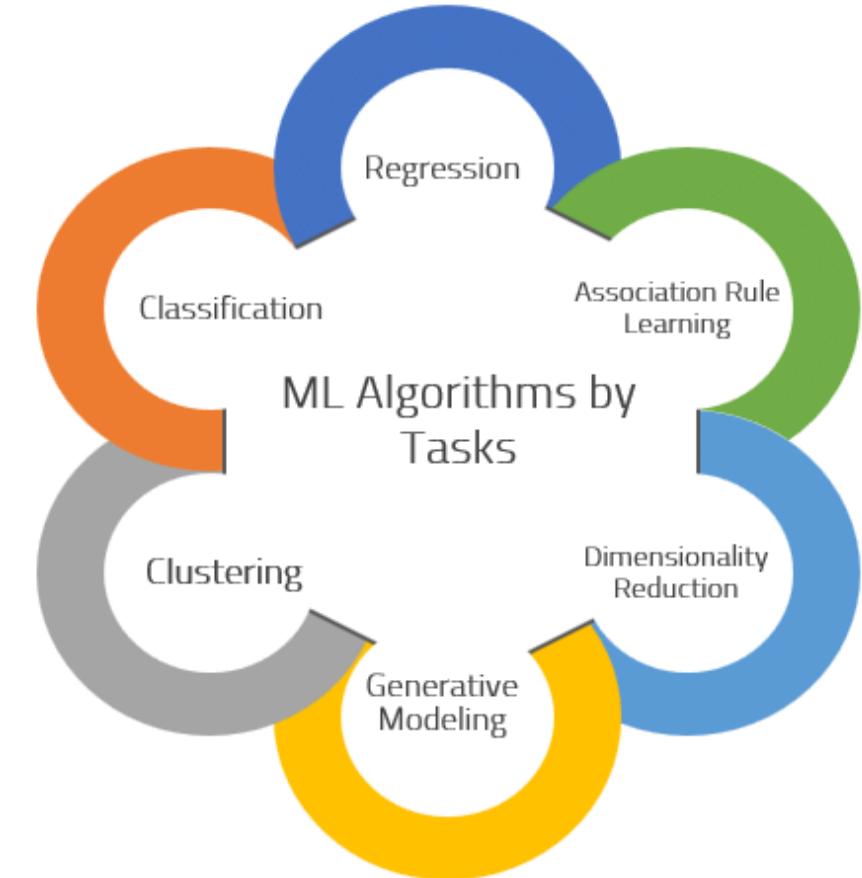


<https://www.youtube.com/c/CtrlShiftFace>



В нашем курсе

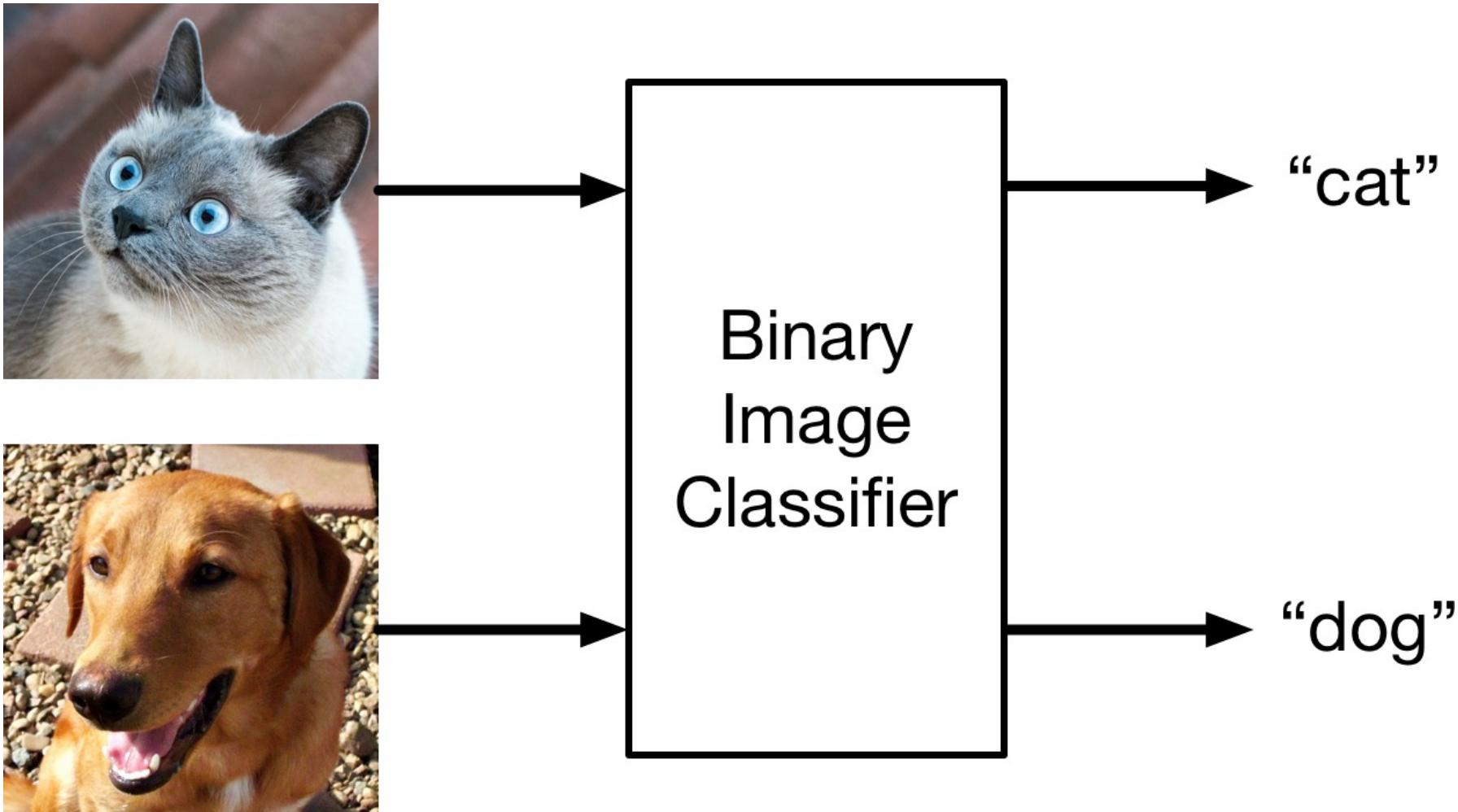
- ▶ Задачи машинного обучения с учителем:
 - Классификация
 - Регрессия
 - Рекомендательные системы
- ▶ Задачи машинного обучения без учителя:
 - Кластеризация
 - Понижение размерности



Классификация



Классификация изображений



<https://www.raywenderlich.com/>

Приложения в банках

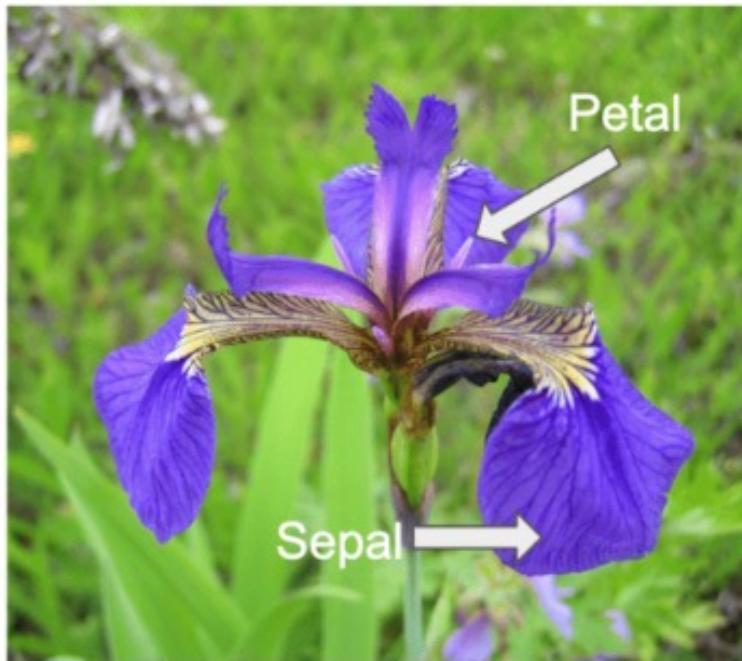
- ▶ Кредитный scoring
- ▶ Прогноз дефолта клиента
- ▶ Обнаружение мошенничества
- ▶ Классификация транзакций



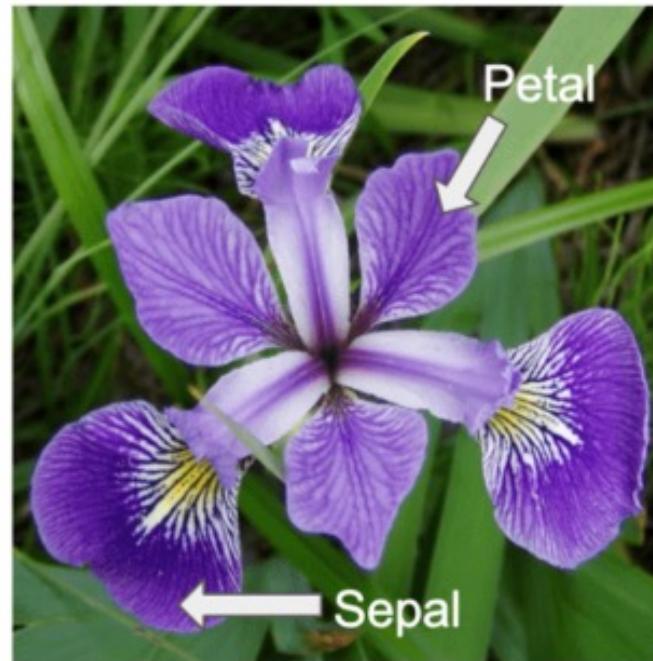
<https://datastart.ru/>

Классификация ирисов

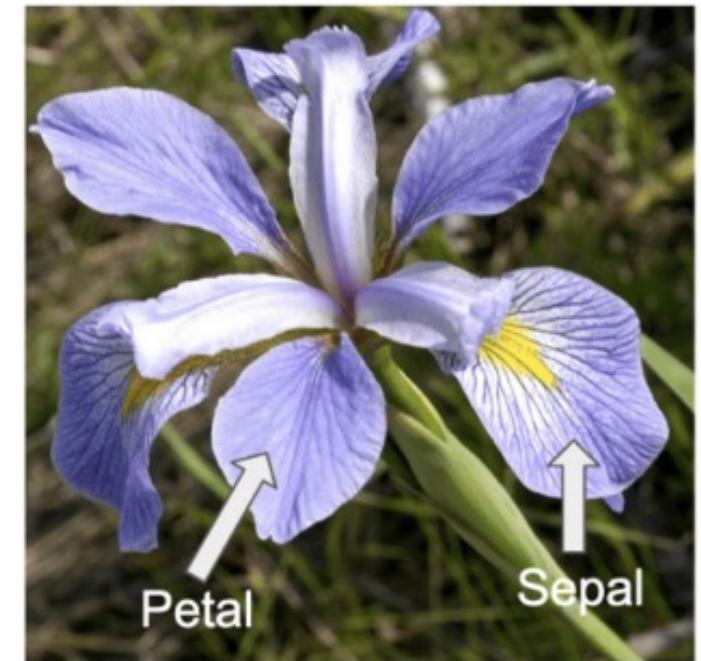
Iris setosa



Iris versicolor



Iris virginica

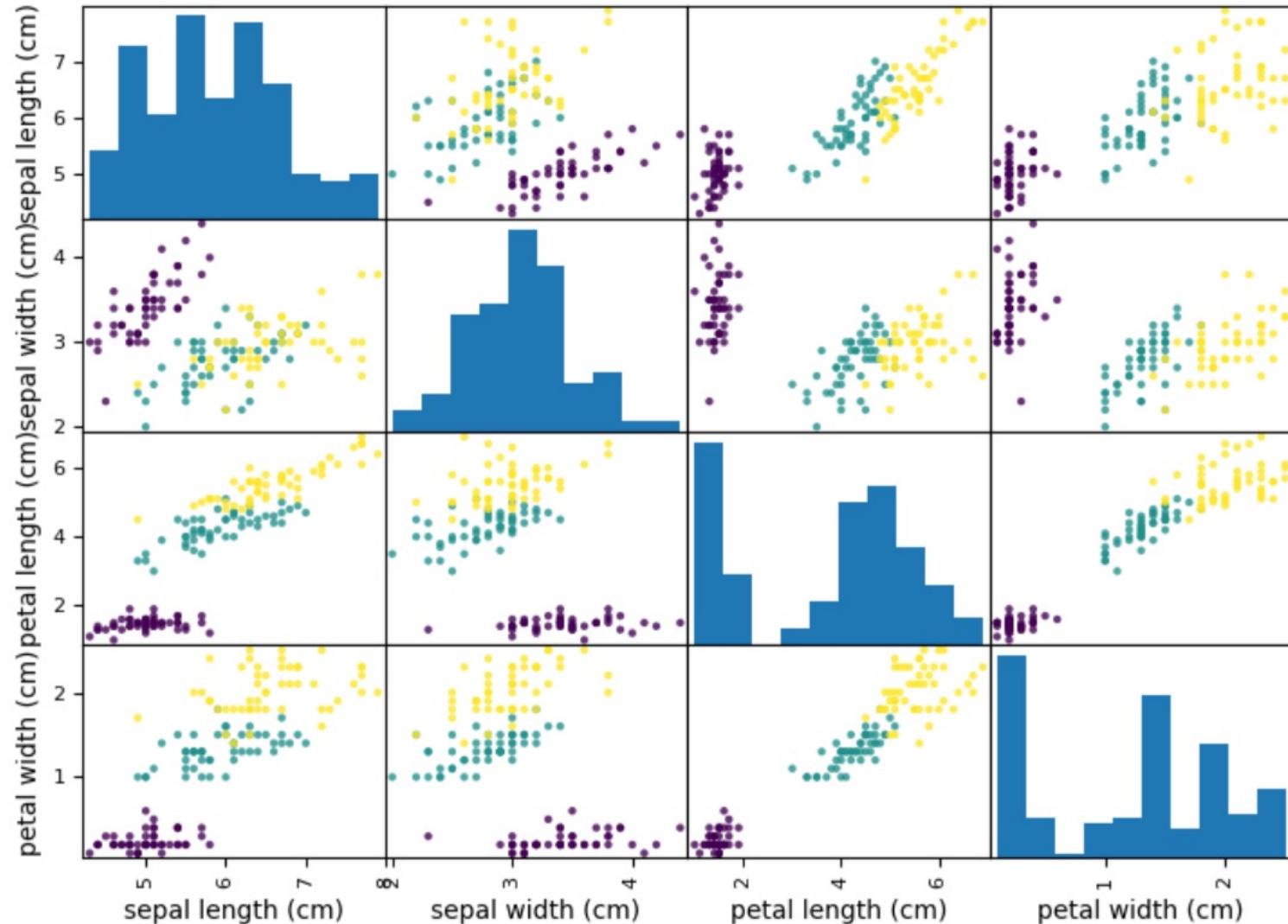


Классификация ирисов

| Id | SepalLengthCm | SepalWidthCm | PetalLengthCm | PetalWidthCm | Species |
|-----------|----------------------|---------------------|----------------------|---------------------|----------------|
| 1 | 5.1 | 3.5 | 1.4 | 0.2 | Iris-setosa |
| 2 | 4.9 | 3 | 1.4 | 0.2 | Iris-setosa |
| 3 | 4.7 | 3.2 | 1.3 | 0.2 | Iris-setosa |
| 4 | 4.6 | 3.1 | 1.5 | 0.2 | Iris-setosa |
| 5 | 5 | 3.6 | 1.4 | 0.2 | Iris-setosa |
| 6 | 5.4 | 3.9 | 1.7 | 0.4 | Iris-setosa |
| 7 | 4.6 | 3.4 | 1.4 | 0.3 | Iris-setosa |
| 8 | 5 | 3.4 | 1.5 | 0.2 | Iris-setosa |
| 9 | 4.4 | 2.9 | 1.4 | 0.2 | Iris-setosa |
| 10 | 4.9 | 3.1 | 1.5 | 0.1 | Iris-setosa |
| 11 | 5.4 | 3.7 | 1.5 | 0.2 | Iris-setosa |
| 12 | 4.8 | 3.4 | 1.6 | 0.2 | Iris-setosa |
| 13 | 4.8 | 3 | 1.4 | 0.1 | Iris-setosa |
| 14 | 4.3 | 3 | 1.1 | 0.1 | Iris-setosa |
| 15 | 5.8 | 4 | 1.2 | 0.2 | Iris-setosa |
| 16 | 5.7 | 4.4 | 1.5 | 0.4 | Iris-setosa |
| 17 | 5.4 | 3.9 | 1.3 | 0.4 | Iris-setosa |
| 18 | 5.1 | 3.5 | 1.4 | 0.3 | Iris-setosa |
| 19 | 5.7 | 3.8 | 1.7 | 0.3 | Iris-setosa |



Классификация ирисов

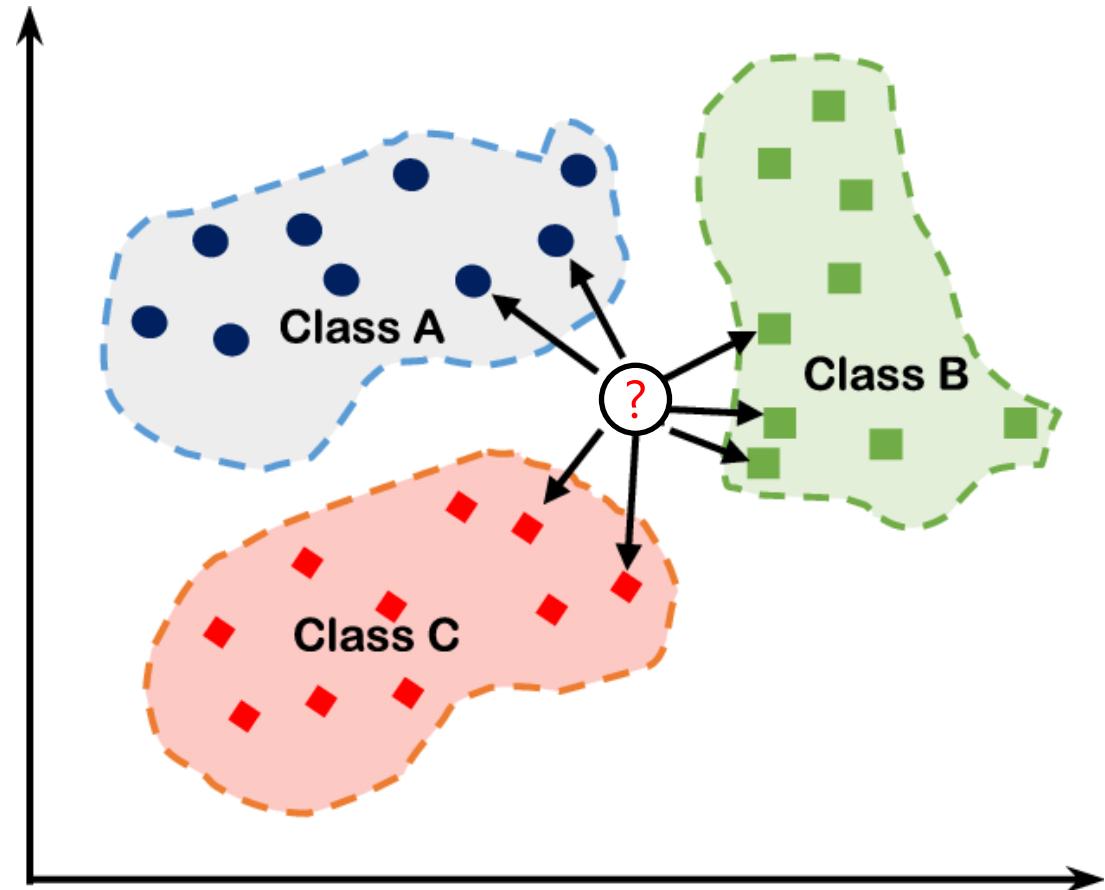


Алгоритм KNN для классификации



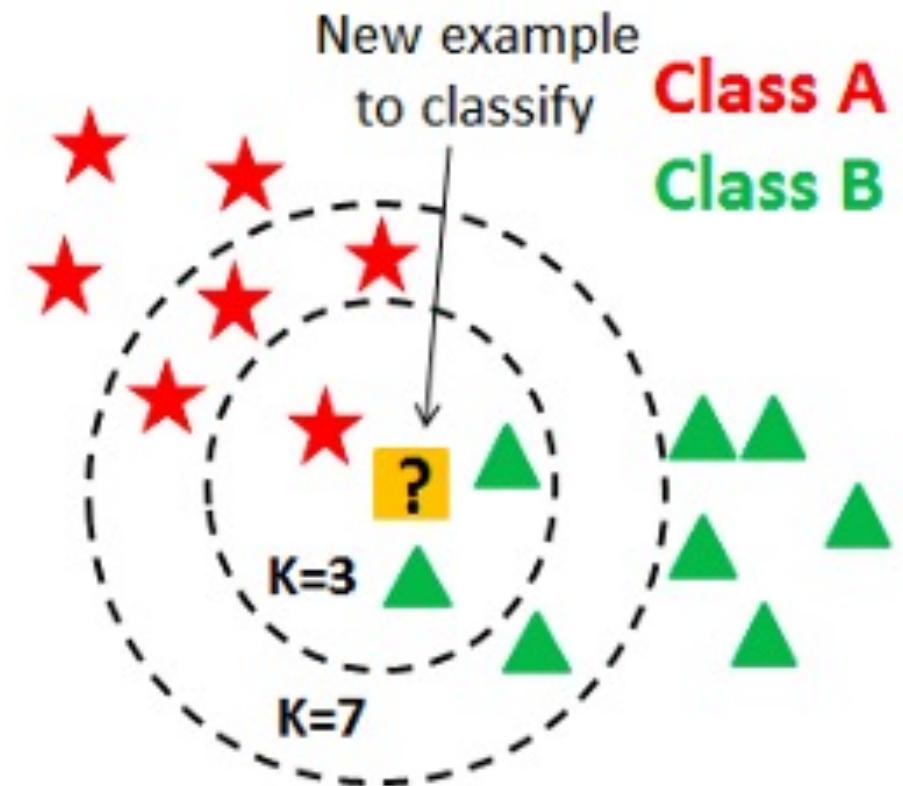
Метрические методы

Метрические методы в машинном обучении – это методы, которые используют **расстояния** между объектами



Задача

- ▶ Есть объекты **нескольких** классов
- ▶ Для каждого **нового** объекта нужно определить его класс
- ▶ Будем **сравнивать** новые объекты с уже известными



Алгоритм K Nearest Neighbors #1

- ▶ Пусть дан набор из n точек: $\{x_i, y_i\}_{i=1}^n$, где
 - x_i – вектор из d признаков объекта;
 - $y_i = \{0, 1, 2, \dots, m\}$ – метка класса объекта.
- ▶ Пусть дана функция расстояния между двумя объектами:

$$\rho(x_i, x_j)$$

- симметричная и неотрицательная
- является мерой (не обязательно)

Алгоритм K Nearest Neighbors #2

- ▶ Запоминаем обучающую выборку $\{x_i, y_i\}_{i=1}^n$.
- ▶ Для каждого нового объекта u сортируем объекты обучающей выборки по расстоянию:

$$\rho(u, x_u^{(1)}) \leq \rho(u, x_u^{(2)}) \leq \dots \leq \rho(u, x_u^{(k)})$$

– $x_u^{(i)}$ – i -й сосед объекта u .

Алгоритм K Nearest Neighbors #3

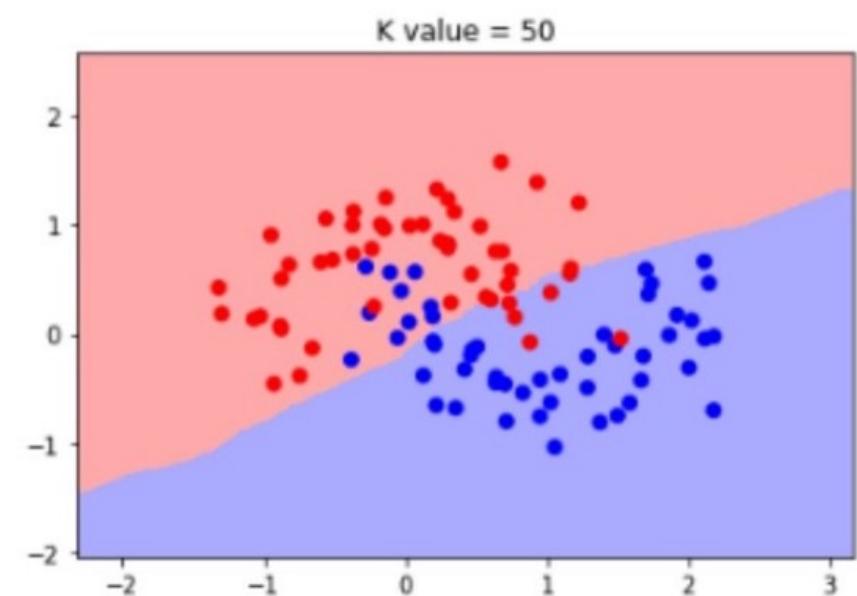
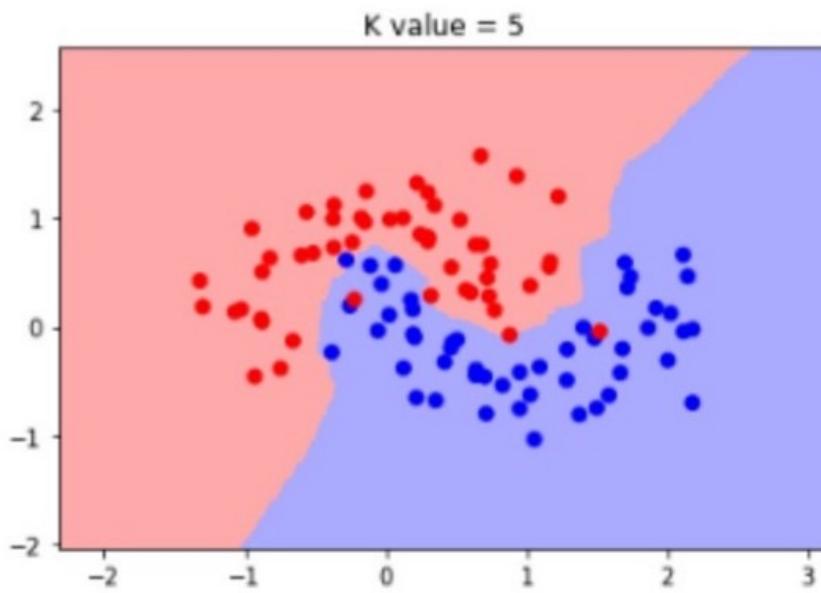
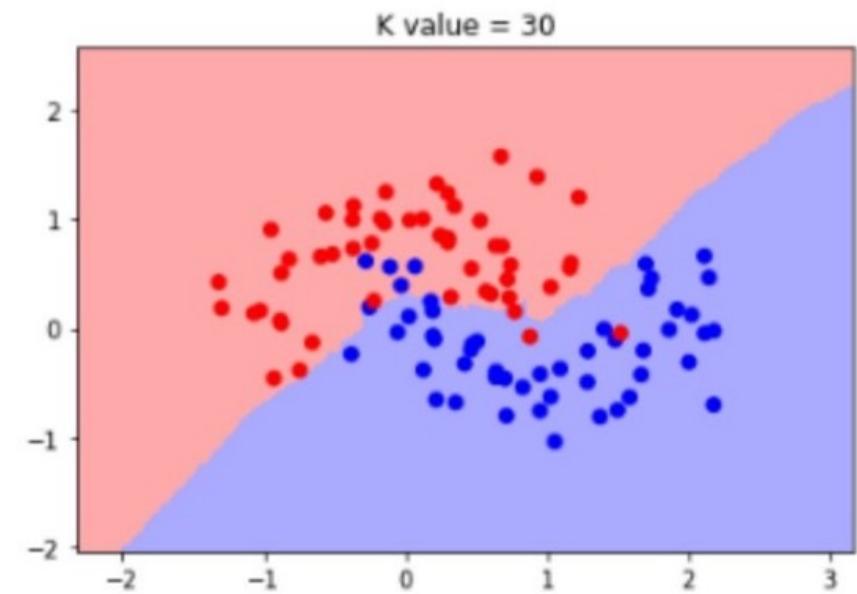
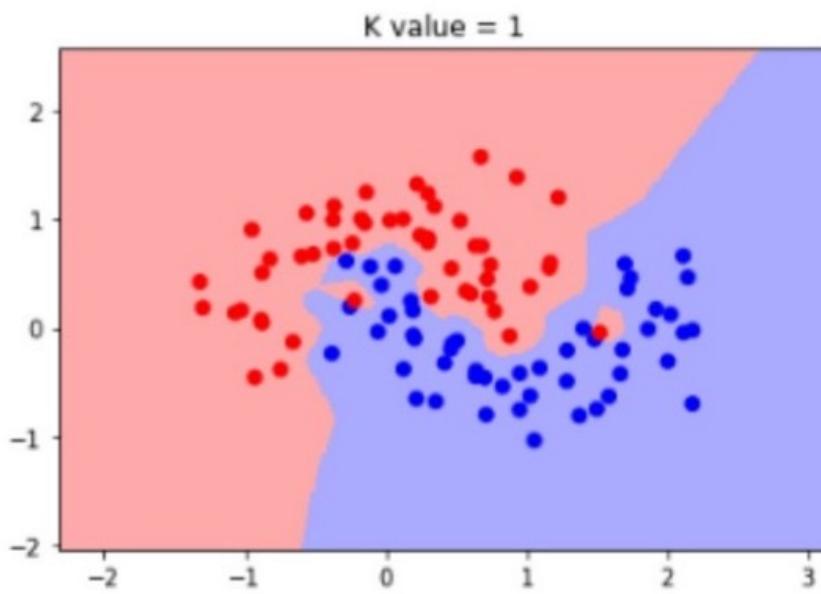
- ▶ Алгоритм относит объект **u** к тому классу, представителей которого окажется больше всего среди **k** его ближайших соседей:

$$\hat{y}(u) = \arg \max_c \sum_{j=1}^k [y_u^{(j)} = c]$$

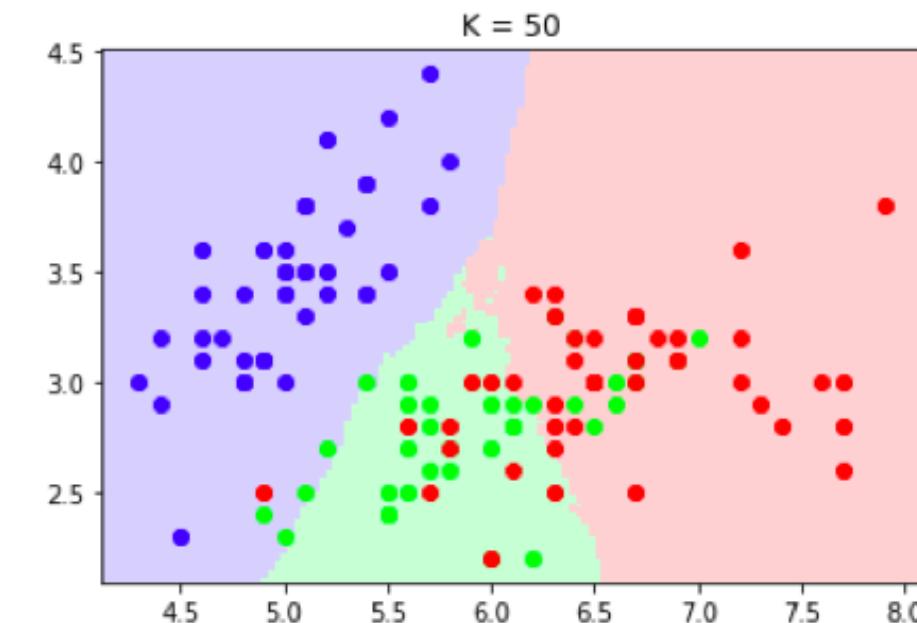
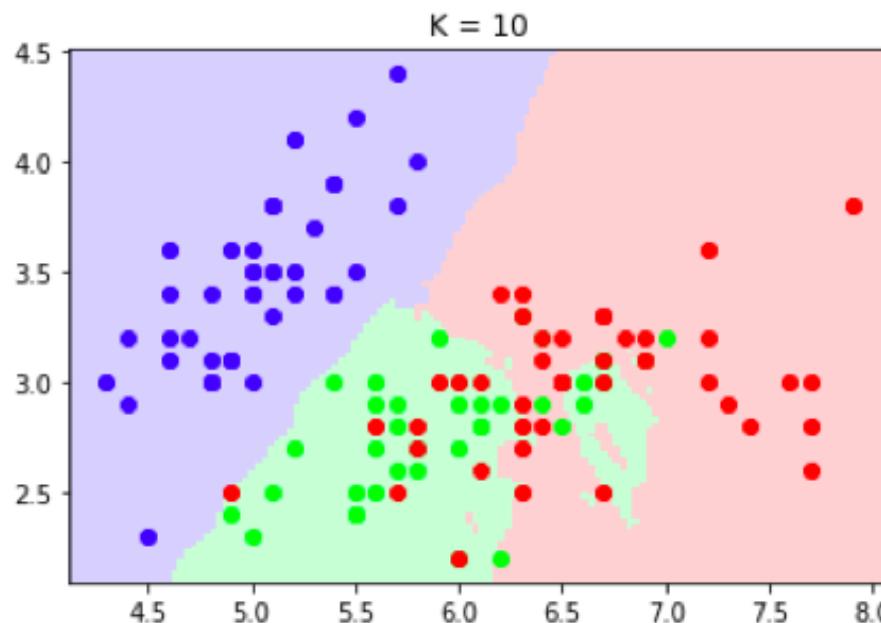
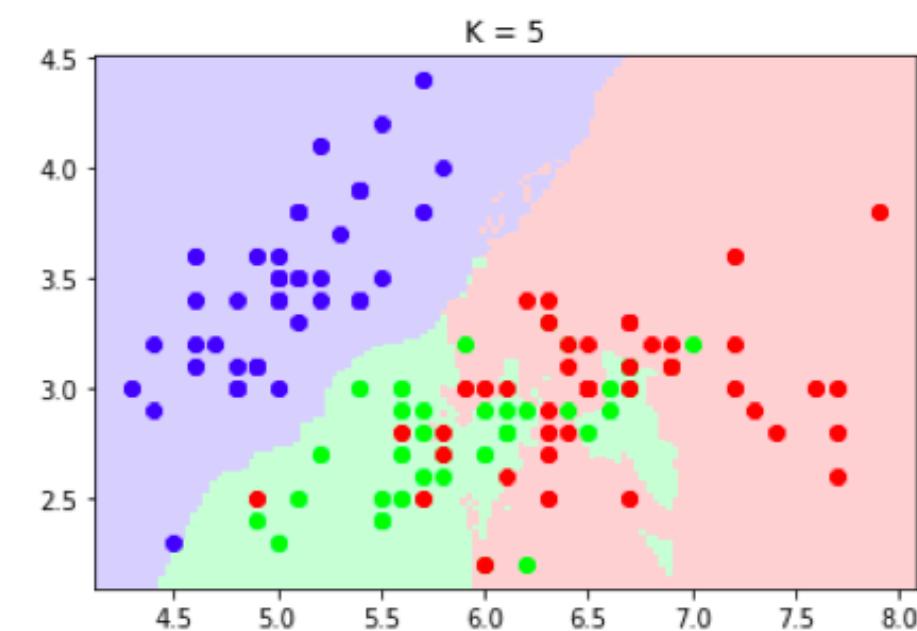
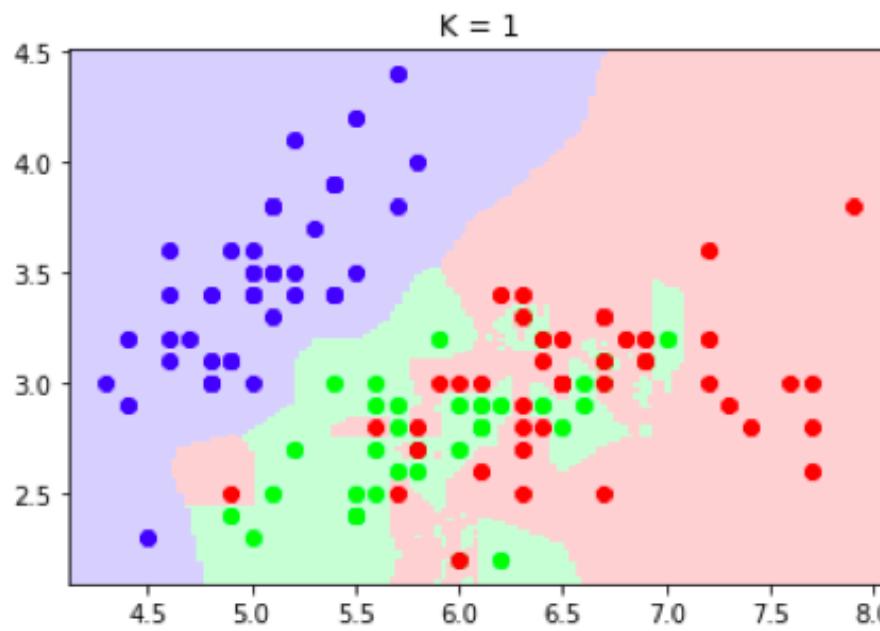
$$\hat{p}_c(u) = \frac{1}{k} \sum_{j=1}^k [y_u^{(j)} = c]$$

- $c = \{0, 1, 2, \dots, m\}$ – метка класса;
- $y_u^{(j)}$ - метка j -го соседа объекта **u**.
- $\hat{y}(u)$ - прогноз метки класса объекта **u**;
- $\hat{p}_c(u)$ - прогноз вероятности класса для объекта **u**;

Пример



Пример



Вопросы

- ▶ Почему все соседи вносят одинаковый вклад в прогноз?
- ▶ Как сделать вклад более близких соседей весомее?

Модификация

Алгоритм относит объект **u** к тому классу, представителей которого окажется больше всего среди **k** его ближайших соседей:

$$\hat{y}(u) = \arg \max_c \sum_{j=1}^k \mathbf{w}_j[y_u^{(j)} = c]$$

$$\hat{p}_c(u) = \frac{1}{\sum_{i=1}^k \mathbf{w}_i} \sum_{j=1}^k \mathbf{w}_j[y_u^{(j)} = c]$$

- \mathbf{w}_j - некоторый вес соседа.

Примеры весов

- ▶ Чем ближе сосед, тем больше вклад:

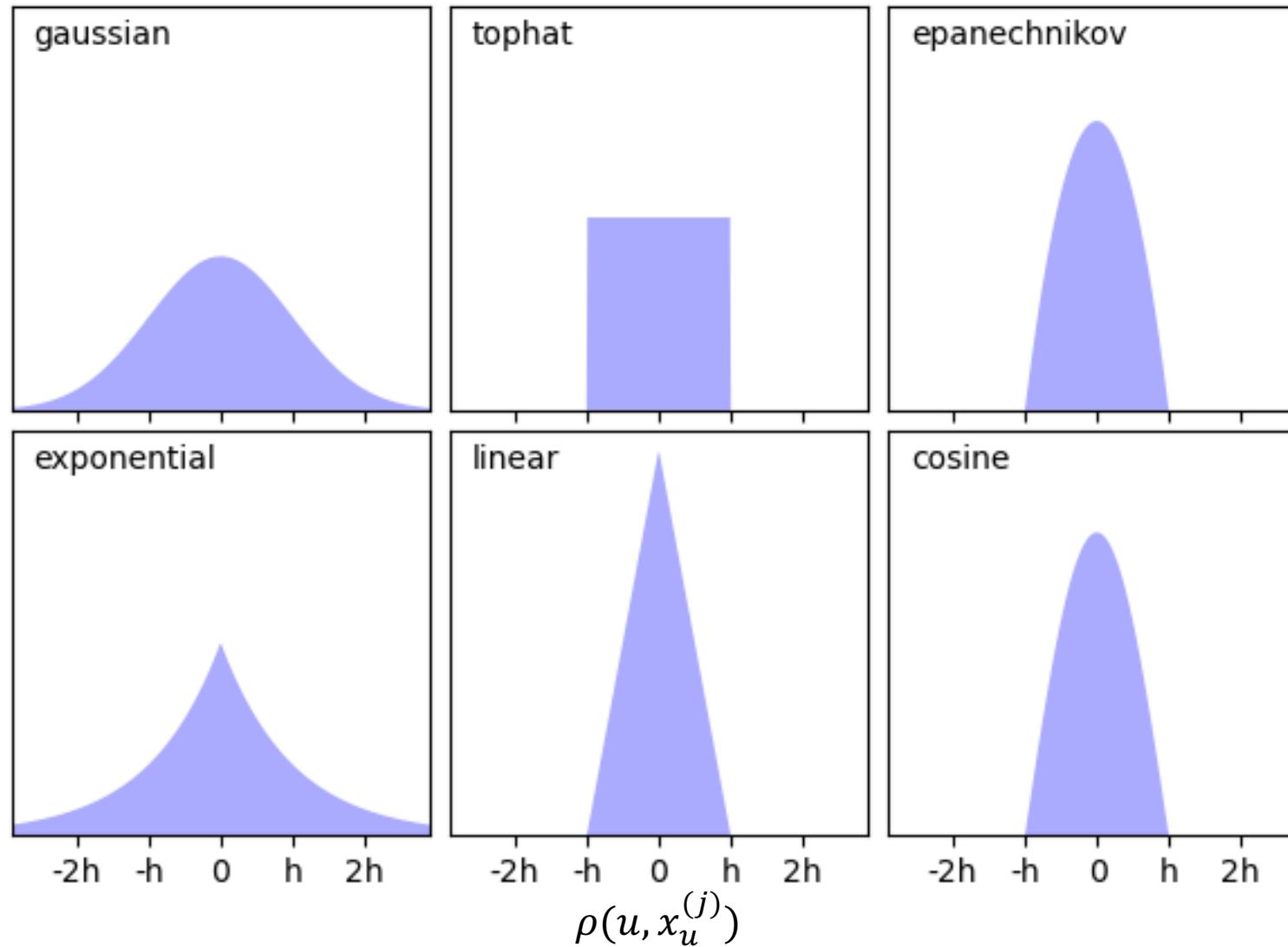
$$w_j = \frac{1}{\rho(u, x_u^{(j)})}$$

- ▶ Метод парзеновского окна:

$$w_j = K \left(\frac{\rho(u, x_u^{(j)})}{h} \right)$$

- h - ширина окна (гиперпараметр)
- K - ядро (некоторая функция близости двух объектов)

Примеры ядер



- Gaussian kernel (`kernel = 'gaussian'`)
$$K(x; h) \propto \exp\left(-\frac{x^2}{2h^2}\right)$$
- Tophat kernel (`kernel = 'tophat'`)
$$K(x; h) \propto 1 \text{ if } x < h$$
- Epanechnikov kernel (`kernel = 'epanechnikov'`)
$$K(x; h) \propto 1 - \frac{x^2}{h^2}$$
- Exponential kernel (`kernel = 'exponential'`)
$$K(x; h) \propto \exp(-x/h)$$
- Linear kernel (`kernel = 'linear'`)
$$K(x; h) \propto 1 - x/h \text{ if } x < h$$
- Cosine kernel (`kernel = 'cosine'`)
$$K(x; h) \propto \cos\left(\frac{\pi x}{2h}\right) \text{ if } x < h$$

Алгоритм KNN для регрессии



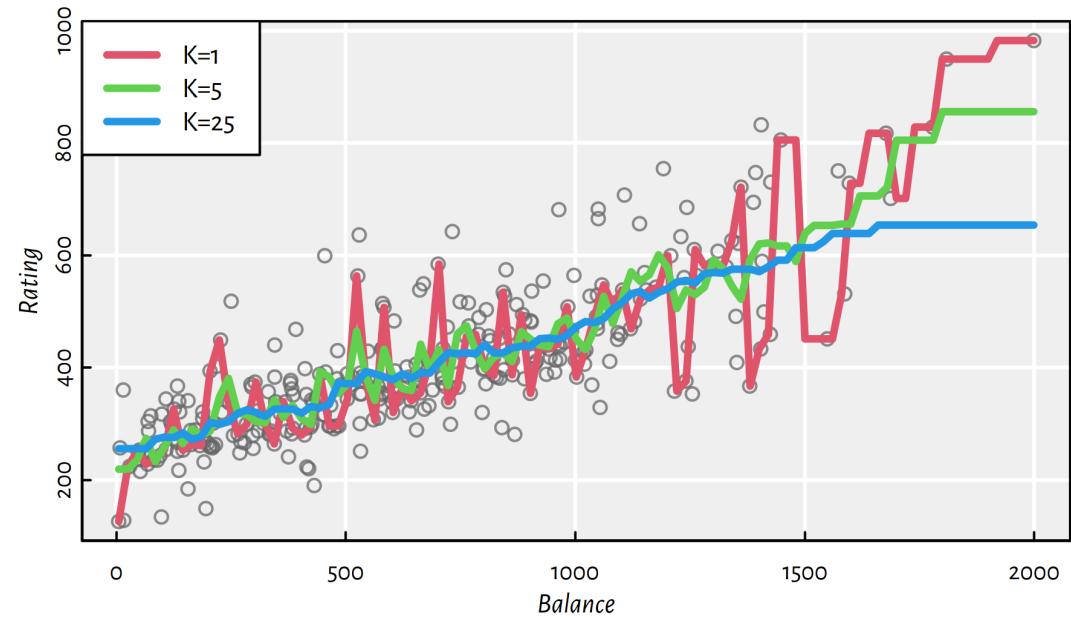
Примеры

- ▶ Прогноз продолжительности жизни
- ▶ Оценка рисков в банках
- ▶ Прогнозирование цены товара
- ▶ Прогнозирование объема продаж



Задача

- ▶ Есть объекты (X)
- ▶ Нужно предсказать некоторую величину (y)
- ▶ Функция, которая описывает зависимость y от X - **модель регрессии**



Алгоритм

Для нового объекта \mathbf{u} находим такое число \mathbf{c} , что:

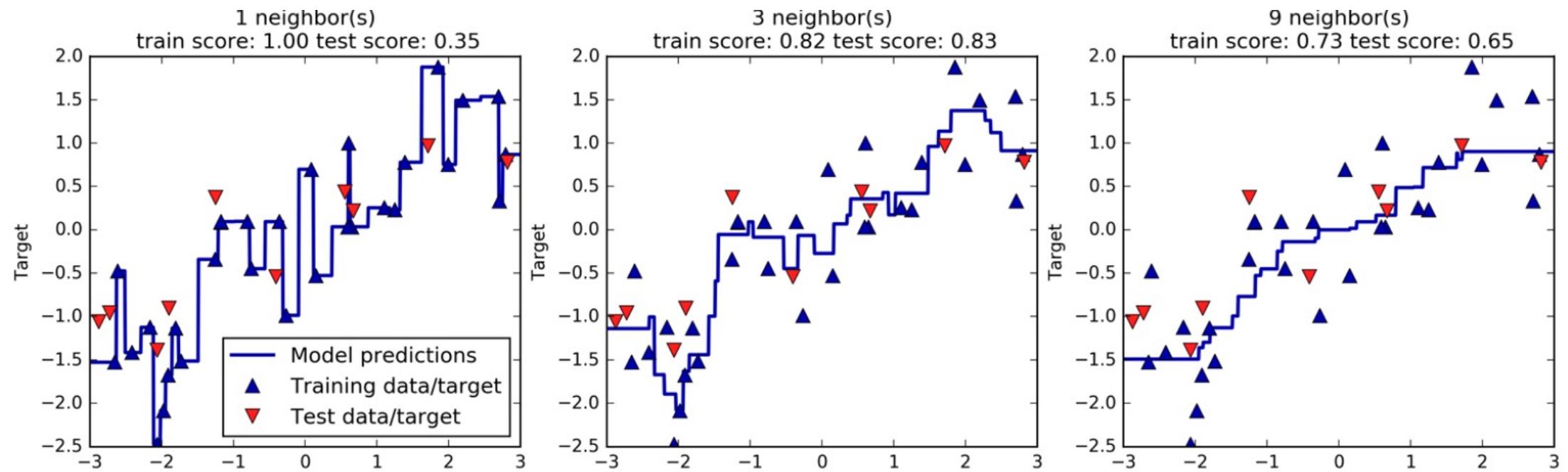
$$\hat{y}(u) = \arg \min_{c \in R} \sum_{j=1}^k \mathbf{w}_j \left(y_u^{(j)} - c \right)^2$$

или в явном виде:

$$\hat{y}(u) = \frac{1}{\sum_{i=1}^k \mathbf{w}_i} \sum_{j=1}^k \mathbf{w}_j y_u^{(j)}$$

- c – 'усредненное' значение целевой переменной по соседям,
- $y_u^{(j)}$ – значение целевой переменной j -го соседа объекта \mathbf{u} ,
- \mathbf{w}_j – некоторый вес соседа,
- $\hat{y}(u)$ – прогноз для объекта \mathbf{u} .

Пример



ФУНКЦИИ РАССТОЯНИЯ

- Метрика Минковского:

$$\rho(a, b) = \left(\sum_{i=1}^d |a_i - b_i|^p \right)^{\frac{1}{p}}$$

- $p = 2$ – Евклидова метрика,
- $p = 1$ – Манхэттенское расстояние,
- $p = \infty$ – метрика Чебышева (наибольшее покоординатное расстояние),

Функции расстояния

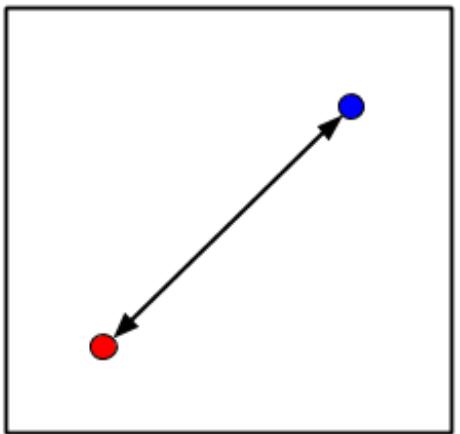
- ▶ Косинусное расстояние:

$$\rho(a, b) = \arccos \frac{a^T b}{\|a\| \|b\|}$$

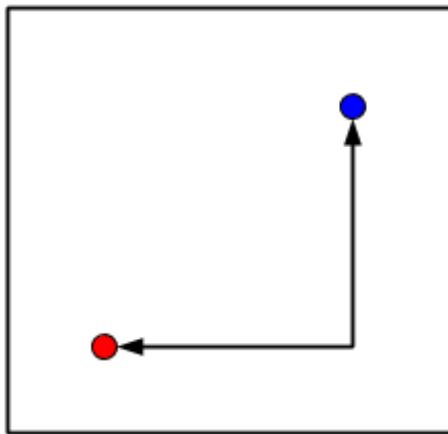
- ▶ Угол между векторами a и b.

ФУНКЦИИ расстояния

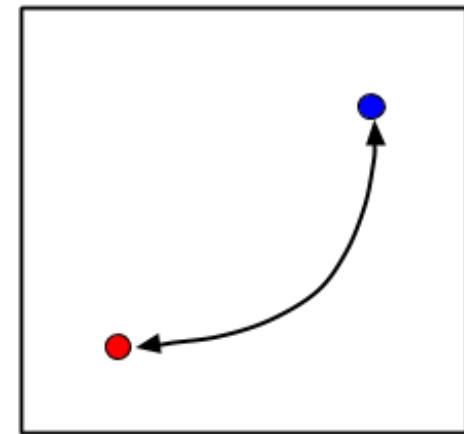
Euclidean



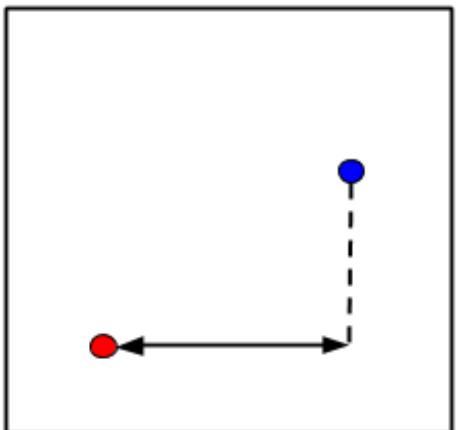
Manhattan



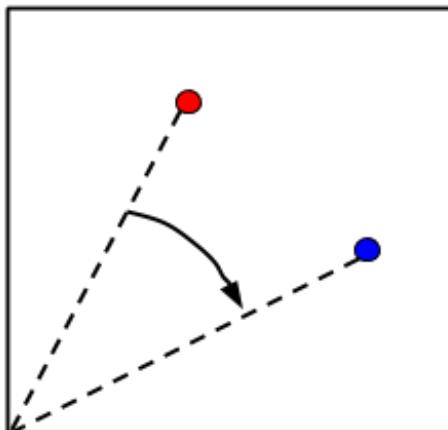
Minkowski



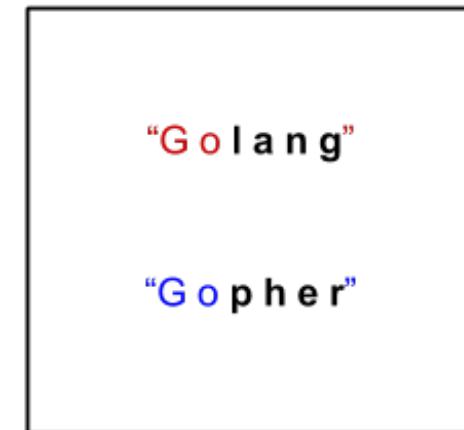
Chebychev



Cosine Similarity



Hamming



Основные проблемы KNN

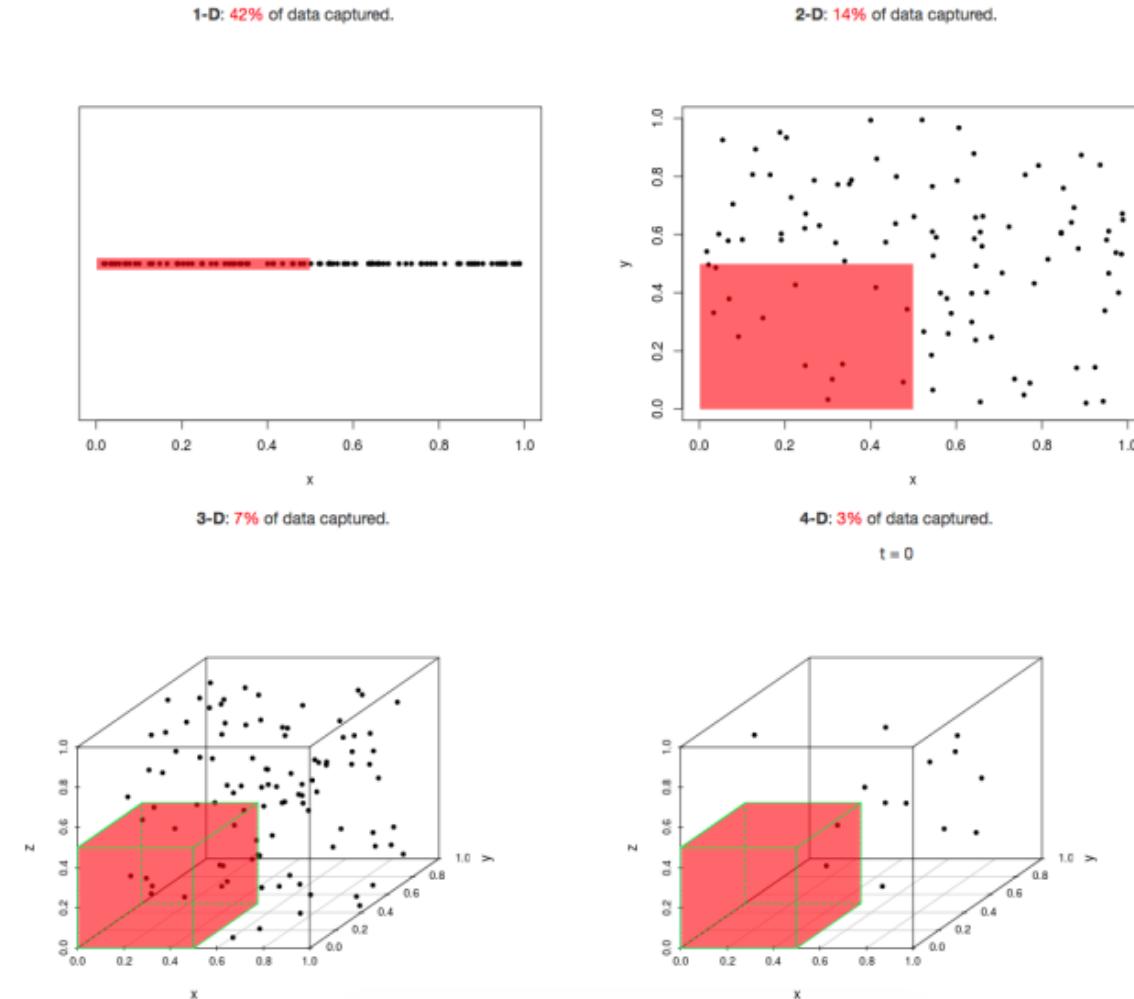
Все проблемы KNN из-за функций расстояния:

$$\rho^2(a, b) = (a_1 - b_1)^2 + (a_2 - b_2)^2 + \dots + (a_d - b_d)^2$$

- ▶ Масштаб признаков сильно влияет на результат
 - Одинаковый масштаб – равный вклад в прогноз
 - Разный масштаб – разный вклад в прогноз ☺
- ▶ Шумовые признаки снижают качество алгоритма
- ▶ Плохо работает на больших размерностях данных
 - Проклятье размерности

Проклятье размерности

- ▶ С ростом размерности D число кубов растет экспоненциально с D
- ▶ Число объектов в каждом кубе падает экспоненциально с D
- ▶ Расстояние между объектами сильно увеличивается
- ▶ Использование KNN может потерять смысл



<https://eranraviv.com/curse-of-dimensionality/>

Заключение



Вопросы

- ▶ Опишите алгоритм k ближайших соседей для задач регрессии и классификации.
- ▶ Каковы проблемы использования метода k ближайших соседей на практике?
- ▶ Запишите формулы для следующий функций расстояния: расстояние Минковского, евклидово расстояние, манхэттэнское расстояние, косинусное расстояние.