




Введение в ML

Как устроено машинное обучение?

Политология, 3 курс
Татьяна Рогович

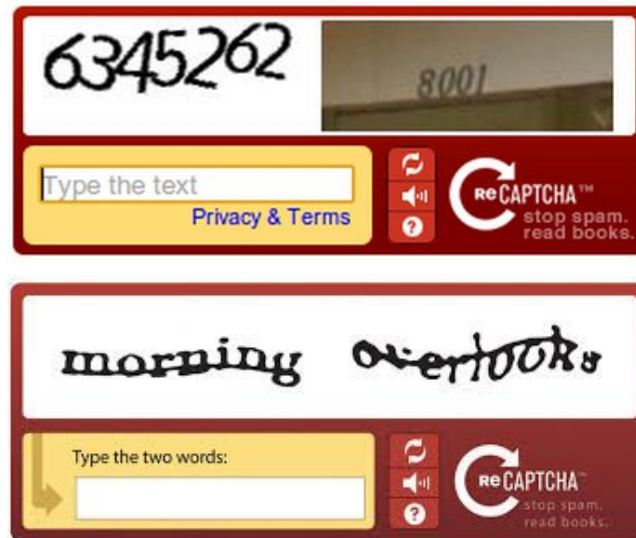
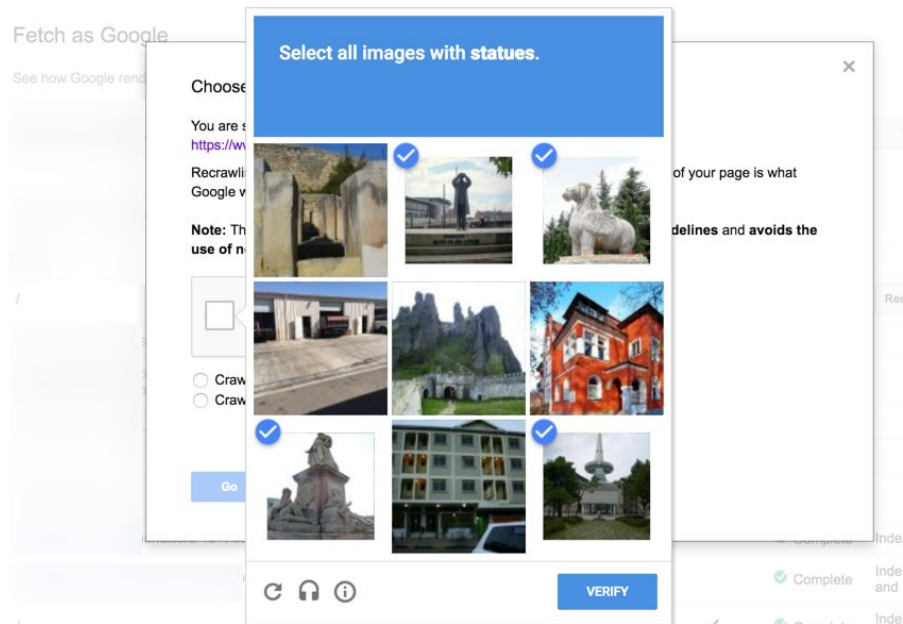


Как машины узнают правильные ответы для данных?

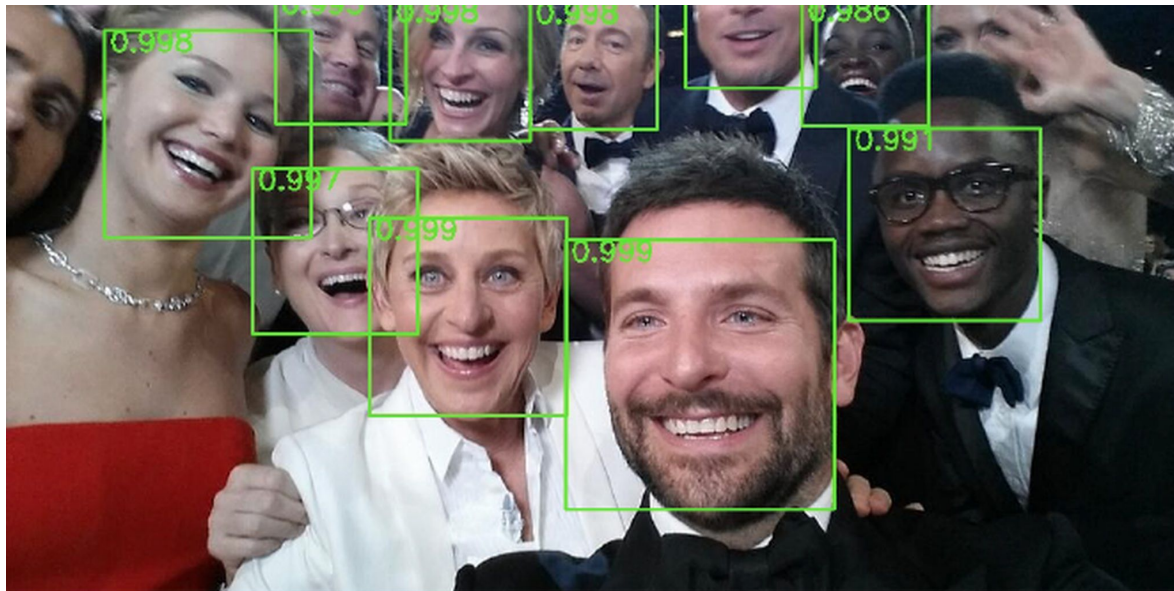
Разметка

- Чтобы компьютер научился выделять объекты, ему нужны примеры с правильными ответами
- Проставление таких ответов — **разметка данных**

Мы сами размечаем данные

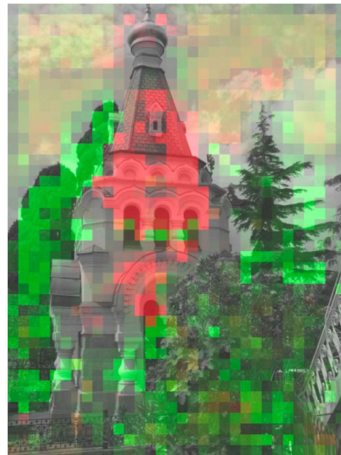
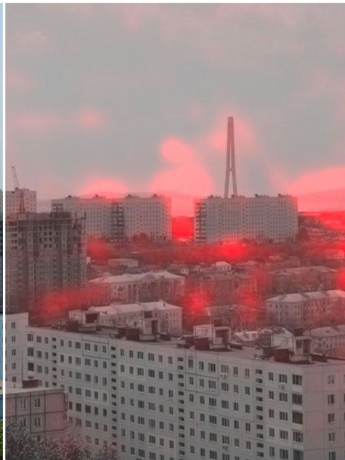


Мы сами размечаем данные



<https://www.dailydot.com/debug/face-detection-algorithm-image-search/>

Кто лучше разбирается в хрущевках?



Как это работает?

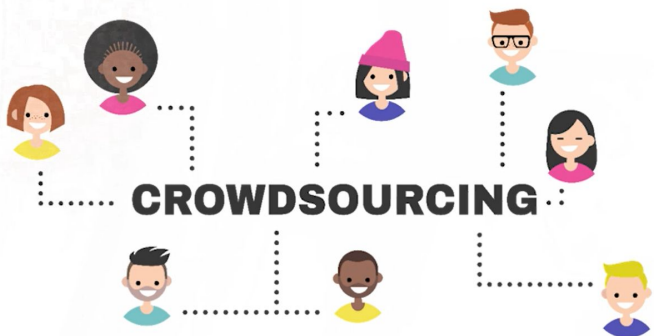
<https://yandex.ru/blog/company/cities-game>

Еще игры с Алисой!

<https://alice.yandex.ru/games>

Что делать, когда вы не Google?

- Разметка силами большого количества людей за небольшую оплату
- Примеры: Amazon Mechanical Turk, Яндекс.Толока



ImageNet



Изображения и разметка объектов на них;



Соревнования — с 2010 года;



Собрано более 10 миллионов размеченных изображений;

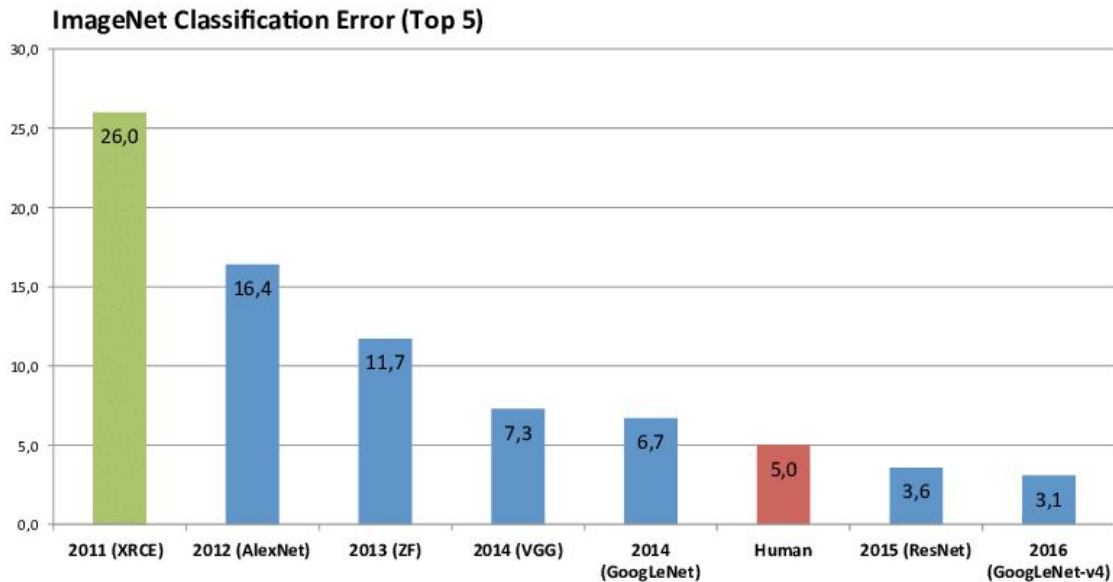


Качество улучшилось с 72% почти до 98%;



Сбор разметки для ImageNet — ключевой шаг в развитии компьютерного зрения.

Люди, кстати, несовершенны





Модель машинного обучения

Некоторая функция, которая предсказывает ответ на основе данных.

Признаки -> Модель -> Предполагаемый ответ



Давайте поиграем в ML

Площадь	Расстояние до метро	Район	Стоимость
50	1	Черёмушки	5.000.000
100	2	Щукино	9.000.000
50	0.5	Мясницкая	20.000.000
100	1	Хамовники	50.000.000



Давайте поиграем в ML

Площадь	Расстояние до метро	Район	Стоимость
50	1	Черёмушки	5.000.000
100	2	Щукино	9.000.000
50	0.5	Мясницкая	20.000.000
100	1	Хамовники	50.000.000

прогноз = 100 000 * площадь



Давайте поиграем в ML

Площадь	Расстояние до метро	Район	Стоимость	Прогноз
50	1	Черёмушки	5.000.000	5.000.000
100	2	Щукино	9.000.000	10.000.000
50	0.5	Мясницкая	20.000.000	5.000.000
100	1	Хамовники	50.000.000	10.000.000

прогноз = 100 000 * площадь



Давайте поиграем в ML

Площадь	Расстояние до метро	Район	Стоимость
50	1	Черёмушки	5.000.000
100	2	Щукино	9.000.000
50	0.5	Мясницкая	20.000.000
100	1	Хамовники	50.000.000

прогноз = 100 000 * площадь - 1 000 000 * расстояние до метро + 1 000 000

Давайте поиграем в ML

Площадь	Расстояние до метро	Район	Стоимость	Прогноз
50	1	Черёмушки	5.000.000	5.000.000
100	2	Щукино	9.000.000	9.000.000
50	0.5	Мясницкая	20.000.000	5.500.000
100	1	Хамовники	50.000.000	10.000.000

прогноз = 100 000 * площадь - 1 000 000 * расстояние до метро + 1 000 000



Давайте поиграем в ML

Площадь	Расстояние до метро	Район	Стоимость
50	1	Черёмушки	5.000.000
100	2	Щукино	9.000.000
50	0.5	Мясницкая	20.000.000
100	1	Хамовники	50.000.000

прогноз = 100 000 * площадь - 1 000 000 * расстояние до метро +
+ 300 000 *(в ЦАО?) * площадь + 1 000 000



Давайте поиграем в ML

Площадь	Расстояние до метро	Район	Стоимость	Прогноз
50	1	Черёмушки	5.000.000	5.000.000
100	2	Щукино	9.000.000	9.000.000
50	0.5	Мясницкая	20.000.000	20.500.000
100	1	Хамовники	50.000.000	40.000.000

прогноз = 100 000 * площадь - 1 000 000 * расстояние до метро +
+ 300 000 *(в ЦАО?) * площадь + 1 000 000



Что же у нас получилось? Линейная модель!

Параметры модели — величины, которые можно подбирать для повышения качества прогнозов

$$\begin{aligned}\text{прогноз} &= \mathbf{100.000} * \text{площадь} \\ &- \mathbf{1.000.000} * (\text{расстояние до метро}) \\ &+ \mathbf{300.000} * (\text{в ЦАО?}) * \text{площадь} \\ &+ \mathbf{1.000.000}\end{aligned}$$

Параметры очень важны для модели

прогноз = **100.000** * площадь

– **1.000.000** * (расстояние до метро)

+ **0** * (в ЦАО?) * площадь

+ **1.000.000**

Площадь	Расстояние до метро	Район	Стоимость	Прогноз
50	1	Черёмушки	5.000.000	5.000.000
100	2	Щукино	9.000.000	9.000.000
50	0.5	Мясницкая	20.000.000	5.500.000
100	1	Хамовники	50.000.000	10.000.000

Изменили параметр стоимости квартиры в ЦАО на менее удачный и модель сразу стала хуже предсказывать



Наша линейная модель в общем виде

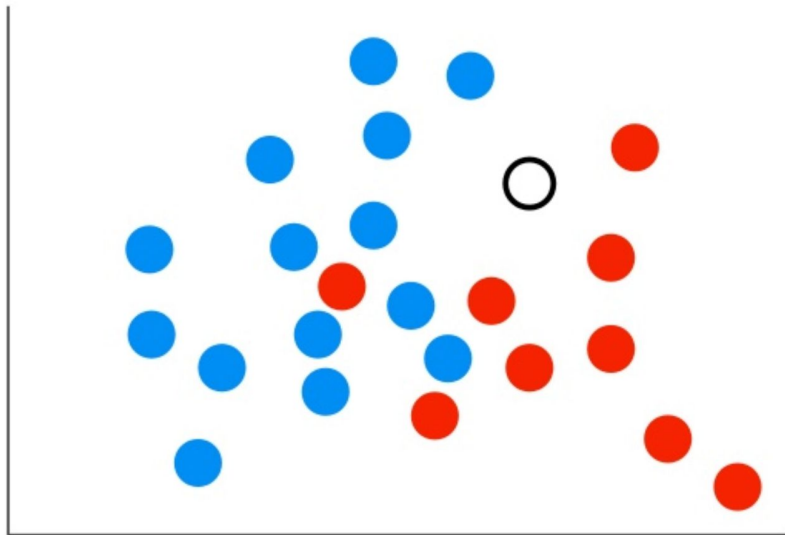
прогноз = **a** * площадь

– **b** * (расстояние до метро)

+ **c** * (в ЦАО?) * площадь

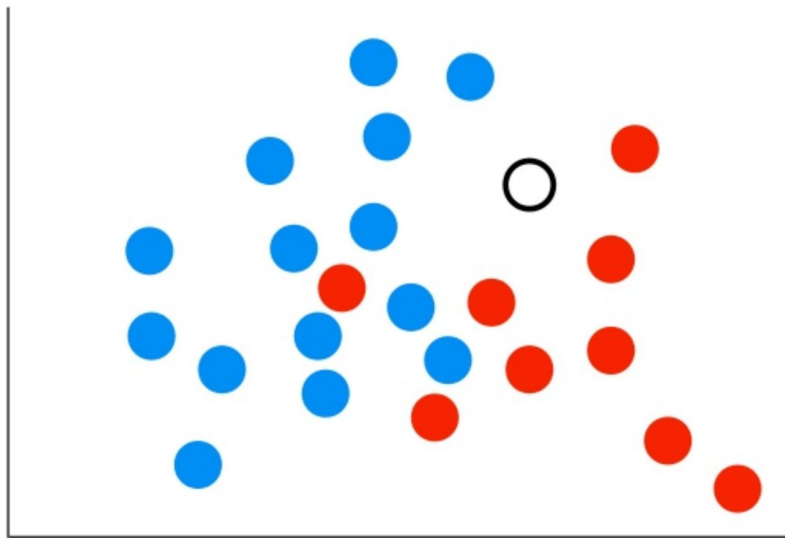
+ **d**

Давайте посмотрим еще на какой-нибудь алгоритм

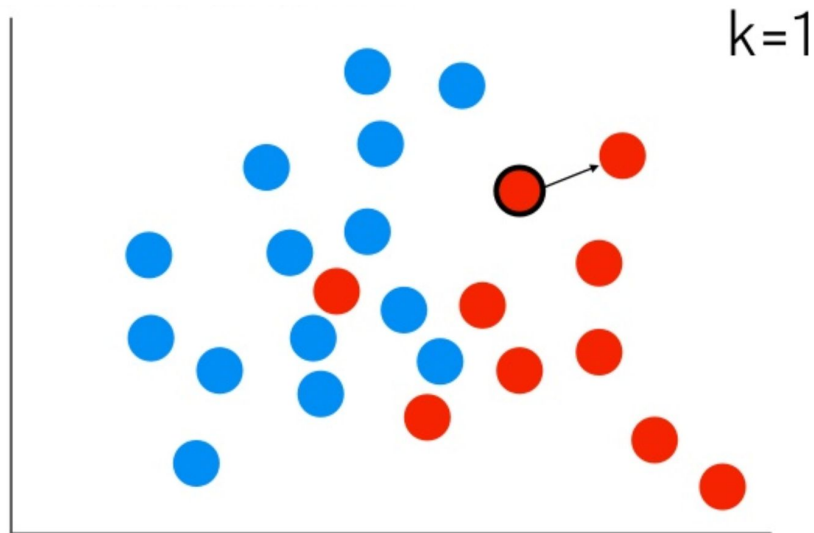


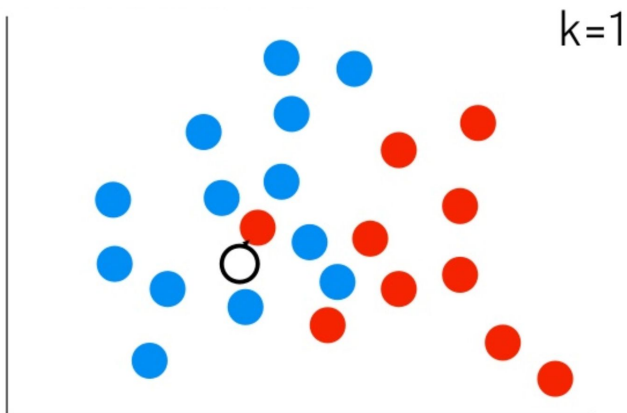
Вам нужно предсказать
цвет белой точки (красный
или синий). Какие есть
идеи?

Давайте покрасим ее
в цвет ближайшего соседа!

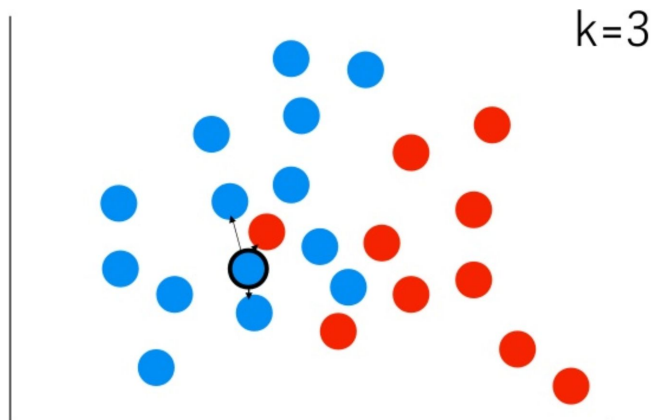
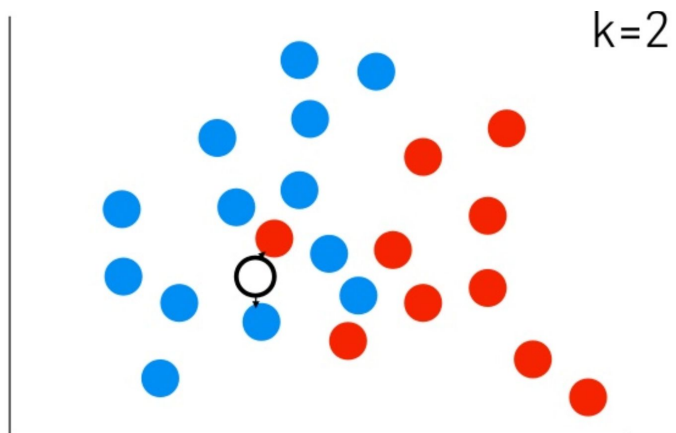


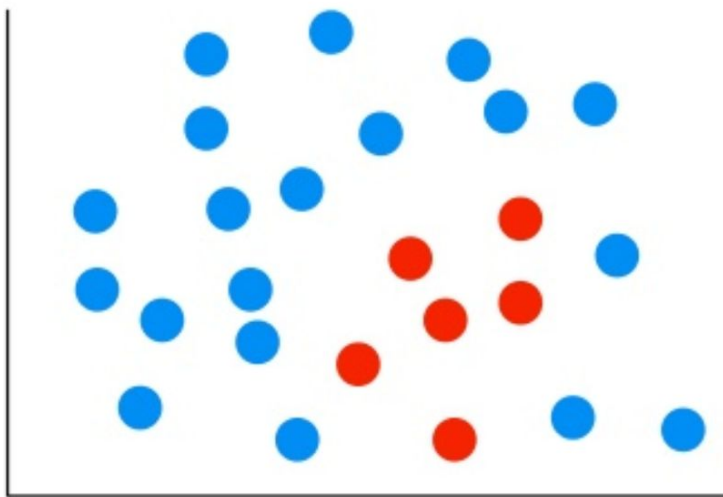
Давайте покрасим ее
в цвет ближайшего соседа!





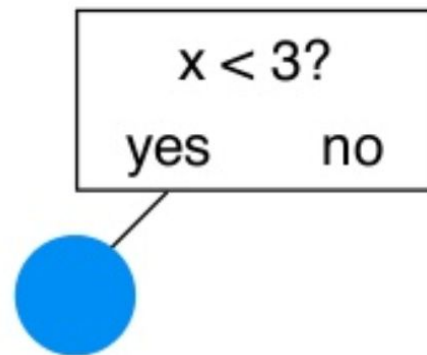
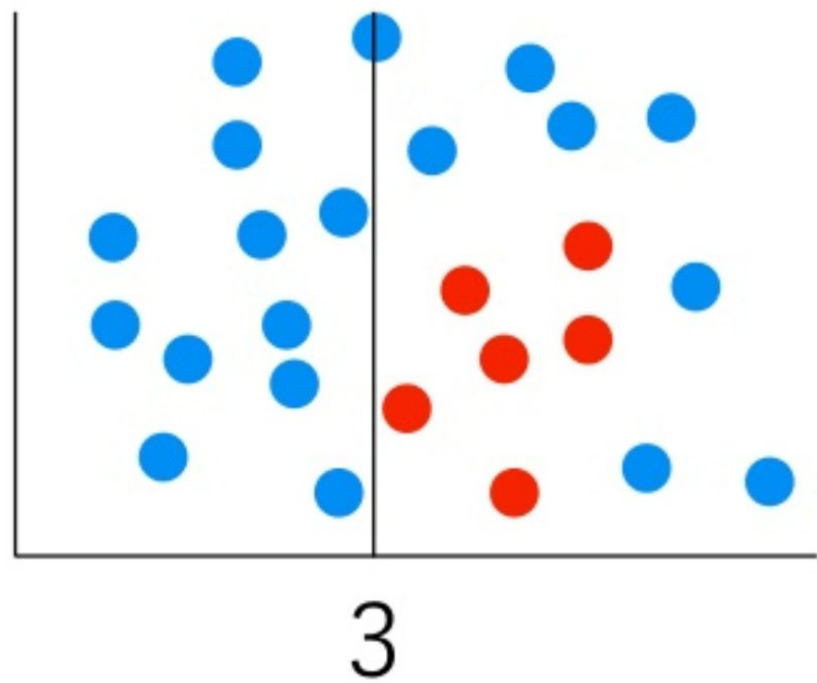
Такая модель будет называться методом ближайших соседей, а параметром, который мы подбираем, будет количество соседей, по которому мы будем определять цвет точки

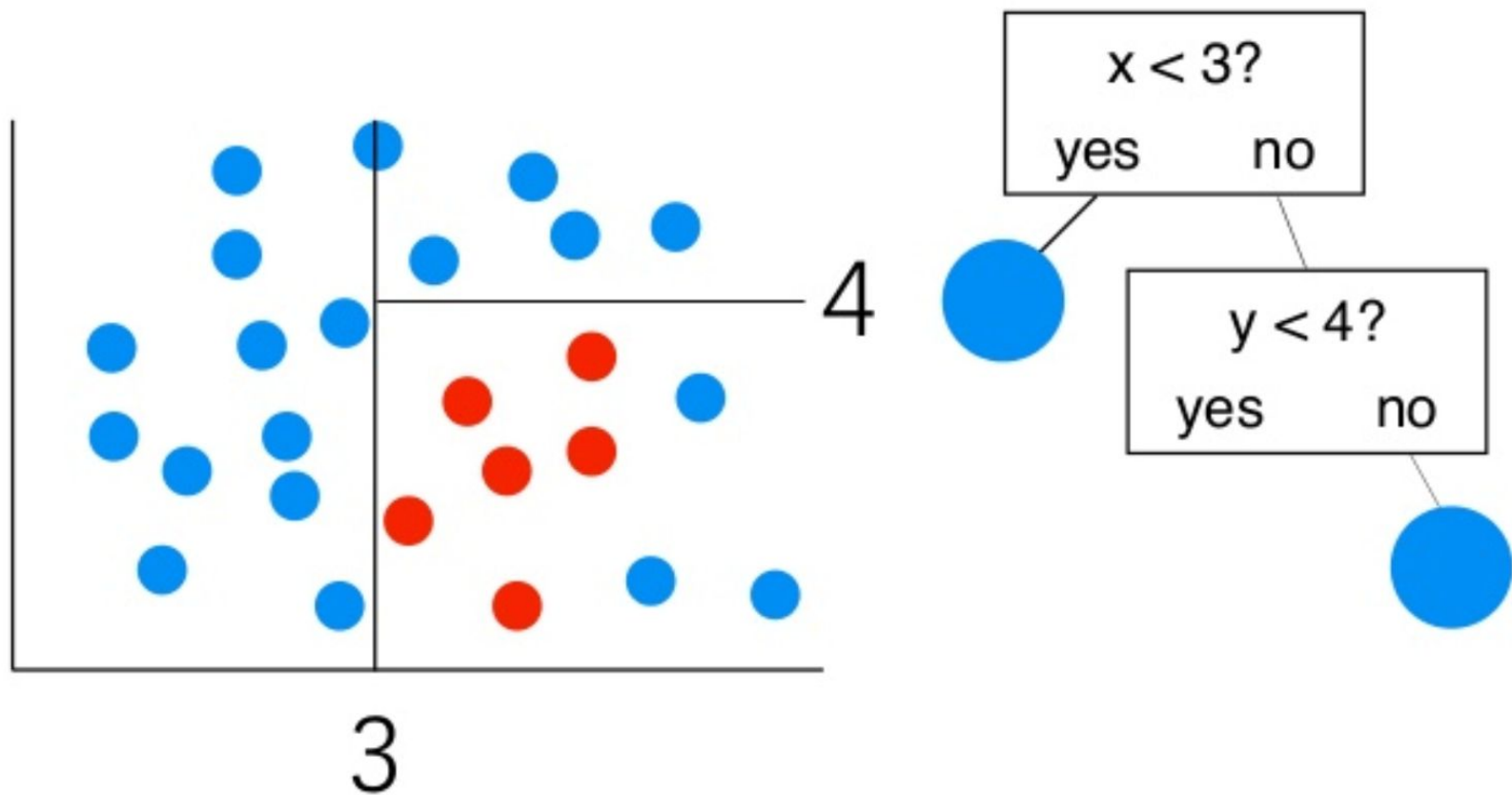


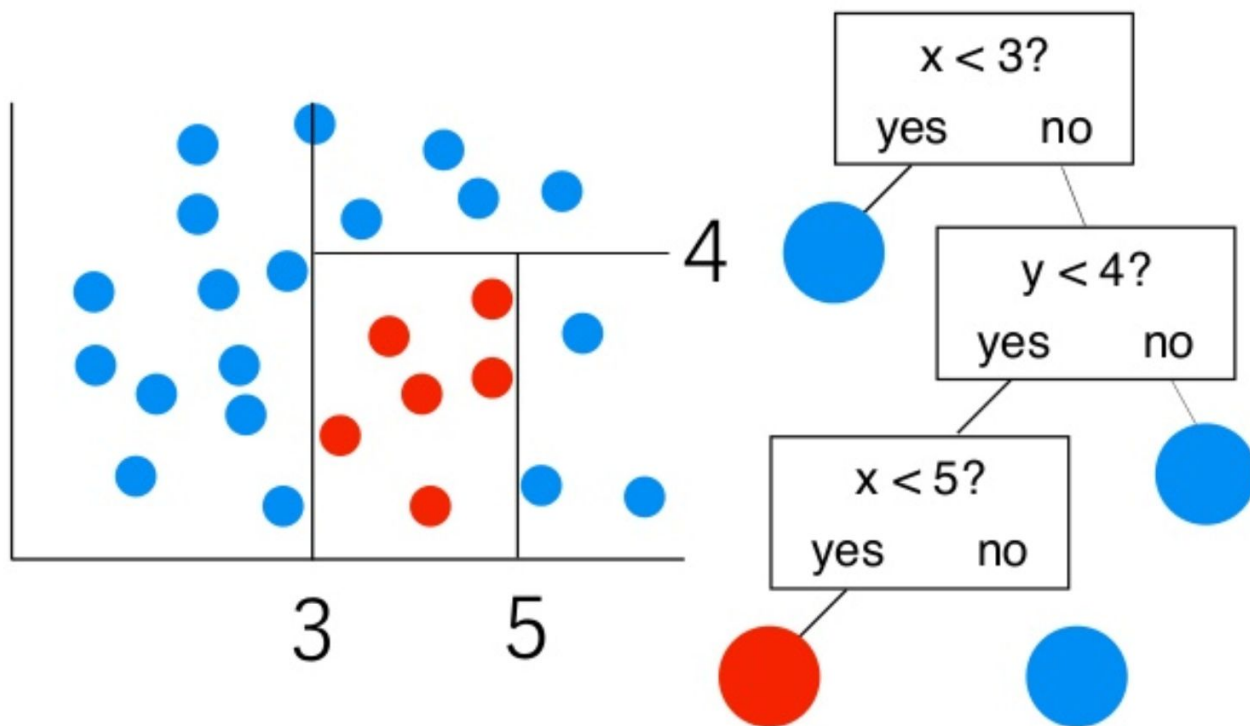


Здесь нам надо не предсказать цвет точки, а нарисовать такие линии, которые лучше всего изолируют красные точки от синих.

Давайте начнем со шкалы X . Где нужно нарисовать линию, чтобы слева и справа от нее состав точек был максимально однородным (насколько это возможно)?







Такая модель называется решающее дерево, а подбираем мы пороги, по которым мы можем максимально однородно разбить наши данные.



Резюме

1. Модель - это некоторая функция, которая предсказывает прогноз на основе признаков.
2. Качество (точность модели) зависит от параметров, которые мы смогли подобрать (насколько хорошо они описывают наши данные)

Как поставить задачу
машинного обучения?



Типы задач в машинном обучении

ML

это название для огромного числа методов и инструментов, которые могут обучаться на данных и делать предсказания.



Типы задач в машинном обучении

- **Supervised learning (обучение с учителем)**

предсказывает или оценивает зависимую переменную, которая обычно либо непрерывная (заработная плата) или категориальная (республиканец/демократ), базируясь на наборе признаков. У нас есть метки для зависимой переменных для обучения или они будут известны спустя какое-то время.

- **Unsupervised learning (обучение без учителя):**

У нас нет меток для зависимой переменной, должны понять структуру данных и правильно ее предсказать (по сути - задачи традиционной статистики)

Задача с разметкой: найти котика на фотографии

Classification



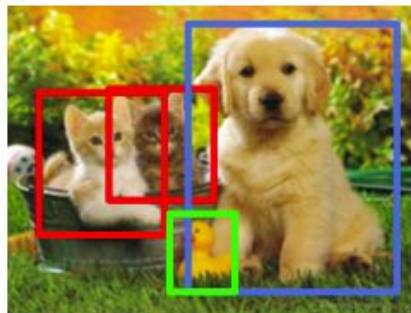
CAT

**Classification
+ Localization**



CAT

Object Detection



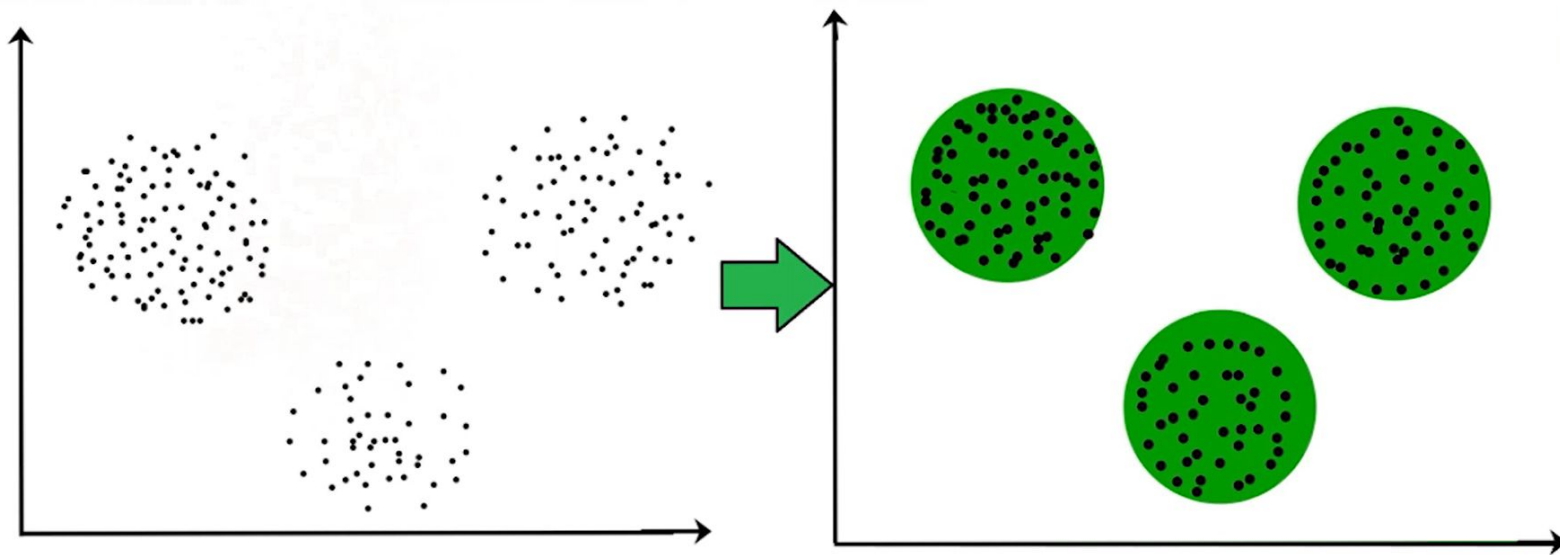
CAT, DOG, DUCK

**Instance
Segmentation**



CAT, DOG, DUCK

Задача без разметки: найти похожих студентов





Терминология ML

Мы хотим построить такую модель, которая будет предсказывать цену (например, чтобы помочь с оценкой квартиры продавцу).

Объектом нашего предсказания будет квартира.

У нас есть одна квартира и ее характеристики (признаки), на основе которых мы будем предсказывать признаки. В статистике мы еще называли объект **наблюдением**.



Терминология ML

Целью нашего предсказания будет: цена квартиры.

В терминологии ML это **ответ, прогноз** или **target**.

В статистике — **целевая** или **зависимая** переменная



Терминология ML

Предсказывать цену (ответ) мы будем с помощью признаков (характеристик) нашего объекта. В статистике мы называли признаки независимыми и контрольными переменными

Что будет признаками наших квартир?

Площадь	Расстояние до метро	Район	Стоимость
50	1	Черёмушки	5.000.000
100	2	Щукино	9.000.000
50	0.5	Мясницкая	20.000.000
100	1	Хамовники	50.000.000

Какие бывают признаки?



Площадь	Расстояние до метро	Район	Стоимость
50	1	Черёмушки	5.000.000
100	2	Щукино	9.000.000
50	0.5	Мясницкая	20.000.000
100	1	Хамовники	50.000.000



Упражняемся

Что будет объектом, целевой переменной и признаком в следующих задачах? Какого типа целевая переменная?

1. Рекомендация музыки
2. Определение котика на фотографии
3. Одобрение кредита



Зачем знать тип целевой переменной?

- **Классификация:** назначить категорию каждому объекту (предсказываем категориальную переменную)
- **Регрессия:** предсказать значение непрерывной переменной для каждого объекта

Тип задачи определяет тип алгоритмов, которые мы можем использовать.

Тип признаков тоже может влиять на выбор (например, есть модели, которые лучше работают с категориальными признаками).



Резюме

1. В задачах обучения с учителем мы можем предсказывать категориальную переменную и решать задачи классификации или количественную в задачах регрессии.
2. Целевую переменную мы предсказываем на основе известных нам объектов и их признаков.

Как мы
выбираем признаки
и параметры?




Скучали по “Титанику”?

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S

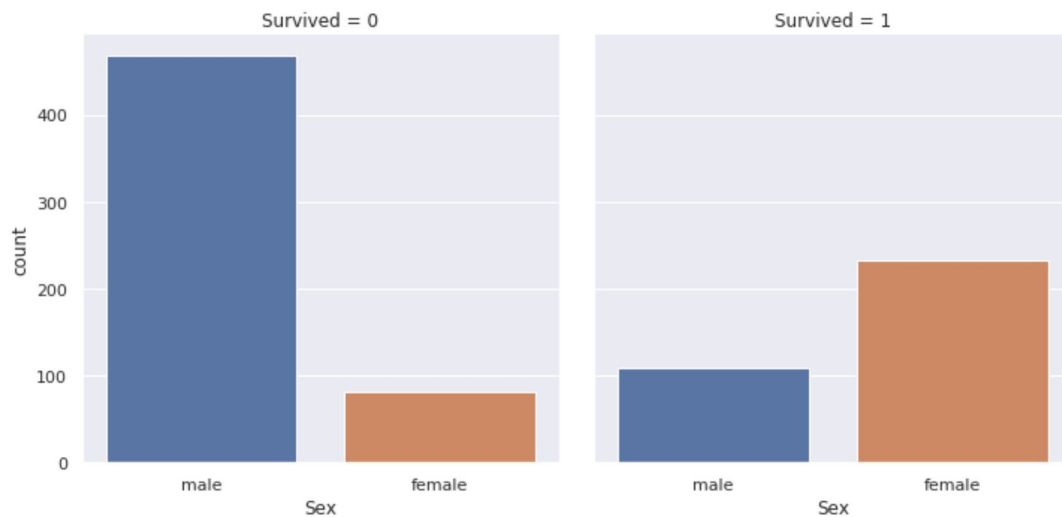


Разведывательный анализ данных (EDA)

- **Exploratory Data Analysis:** процесс, где мы изучаем наши данные (дескриптивные статистики, распределения переменных), пытаемся выявить и решить проблемы, связанные с данными (например, преобразование шкал, работа с пропущенными значениями) и понять, какие независимые переменные могут влиять на нашу целевую.



Какое предположение можем выдвинуть о связи этих двух признаков?





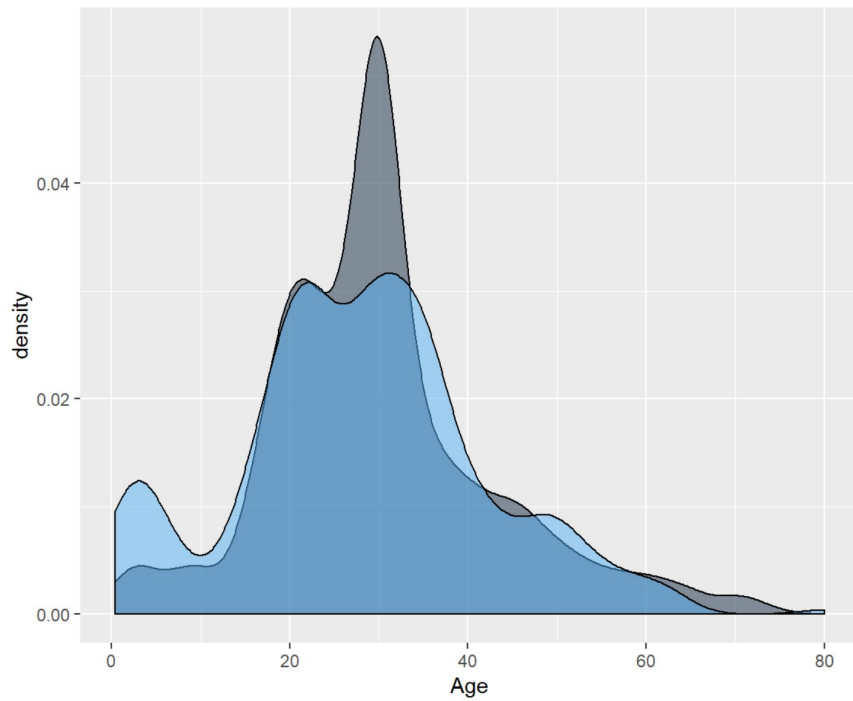
Survived	0	1	All
Pclass			
1	80	136	216
2	97	87	184
3	372	119	491
All	549	342	891

А для этих?



Survived	0	1	All
Pclass			
1	80	136	216
2	97	87	184
3	372	119	491
All	549	342	891

	Pclass	1	2	3	All
Sex	Survived				
female	0	3	6	72	81
	1	91	70	72	233
male	0	77	91	300	468
	1	45	17	47	109
All		216	184	491	891





Формулируем гипотезы

1. Женщины в первом классе почти наверняка выживали.
2. Женщины в третьем классе имели шансы 50/50
3. У мужчин больше шансы погибнуть независимо от класса, но в третьем классе эти шансы растут.
4. У детей до 12 лет шансы выжить выше.

Проверяем гипотезы

Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Survived
3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S	0
1	Cumings, Mrs. John Bradley (Florence Briggs Th...)	female	38.0	1	0	PC 17599	71.2833	C85	C	1
3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S	1
1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S	1
3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S	0
3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q	0
1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S	0
3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S	0
3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S	1
2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C	1
3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.7000	G6	S	1

X: матрица всех значений признаков (независимых переменных)

y: вектор со значениями целевой переменной

X_i - наблюдение (набор значений независимых переменных для одного объекта выборки)



Обучающая и тестовая выборки

- **Training data:** данные, на которых обучаем алгоритм.
Мы даем алгоритму данные, в которых есть значение целевой переменной (**labels, output, target**), чтобы научить его находить такие признаки, которые могут быть связаны со значением целевой переменной.
- **Test data:** данные, значение целевой переменной которых мы знаем, на которых тестируем алгоритм и принимаем решение о его пригодности.

Обычно мы делим наш датасет на два - на одной части тренируемся (обучаемся), а на второй проверяем, чему научились (валидация).

Тестовая выборка

Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	Cumings, Mrs. John Bradley (Florence Briggs Th...)	female	38.0	1	0	PC 17599	71.2833	C85	C
3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S

Survived



Этих данных наш алгоритм раньше не видел и именно на них мы проверяем, хорошо ли мы подобрали параметры.

Y hat -
предсказанное
значение
целевой
переменной



Алгоритм обучения модели

- **Define / Определи.**
- **Fit / Обучи.**
- **Predict / Предскажи.**
- **Evaluate / Оцени.**

Если модель прошла оценивание успешно, то ее можно отправлять в бой (на неведомые данные, для которых значения целевой переменной мы не знаем).

Для некоторых данных в будущем мы можем узнать значение целевой переменной и еще раз проверить качество нашей модели.



Как оценить модель?

Функция потерь - измерение ошибки нашей модели. Выбор конкретной функции зависит от задачи, которую решаем.

Наша цель — минимизировать ее.

Площадь	Расстояние до метро	Район	Стоимость	Прогноз
50	1	Черёмушки	5.000.000	5.000.000
100	2	Щукино	9.000.000	9.000.000
50	0.5	Мясницкая	20.000.000	5.500.000
100	1	Хамовники	50.000.000	10.000.000

Площадь	Расстояние до метро	Район	Стоимость	Прогноз
50	1	Черёмушки	5.000.000	5.000.000
100	2	Щукино	9.000.000	9.000.000
50	0.5	Мясницкая	20.000.000	20.500.000
100	1	Хамовники	50.000.000	40.000.000

Считаем среднюю абсолютную ошибку (MAE)

Площадь	Расстояние до метро	Район	Стоимость	Прогноз	Ошибка
50	1	Черёмушки	5.000.000	5.000.000	0
100	2	Щукино	9.000.000	9.000.000	0
50	0.5	Мясницкая	20.000.000	5.500.000	14.500.000
100	1	Хамовники	50.000.000	10.000.000	40.000.000

Средняя ошибка: $\frac{0+0+14500000+40000000}{4} = 13625000$

Считаем среднюю абсолютную ошибку (MAE)

Площадь	Расстояние до метро	Район	Стоимость	Прогноз	Ошибка
50	1	Черёмушки	5.000.000	5.000.000	0
100	2	Щукино	9.000.000	9.000.000	0
50	0.5	Мясницкая	20.000.000	20.500.000	500.000
100	1	Хамовники	50.000.000	40.000.000	10.000.000

Средняя ошибка: $\frac{0+0+500000+10000000}{4} = 2625000$



Что же такое обучение модели?

Подбор таких параметров,
при которых значение
ошибки минимально!

$$\begin{aligned} \text{прогноз} &= \mathbf{a} * \text{площадь} \\ &- \mathbf{b} * (\text{расстояние до метро}) \\ &+ \mathbf{c} * (\text{в ЦАО?}) * \text{площадь} \\ &+ \mathbf{d} \end{aligned}$$



Резюме

1. Для ML очень важны размеченные данные — данные, для которых значение целевой переменной уже известно.
2. Мы исследуем наши данные, чтобы понять, какие признаки важные.
3. Мы обучаем нашу модель на выбранных признаках и оцениваем ее качество на основе ошибки.
4. После этого мы принимаем решение, насколько наша модель хорошо справляется со своей задачей.