

Машинное обучение

Занятие 2

Машинное обучение

Понятие

Формально: Наука, изучающая способы извлечения **закономерностей** из ограниченного количества примеров.

Неформально: Пусть машина посмотрит на наши данные, найдёт в них закономерности и научится предсказывать для нас **ответ**.

Основные понятия

Объект - то, для чего хотим сделать предсказание

Обычно обозначаем как x (x_1, x_2, \dots, x_n - мы можем нумеровать объекты)

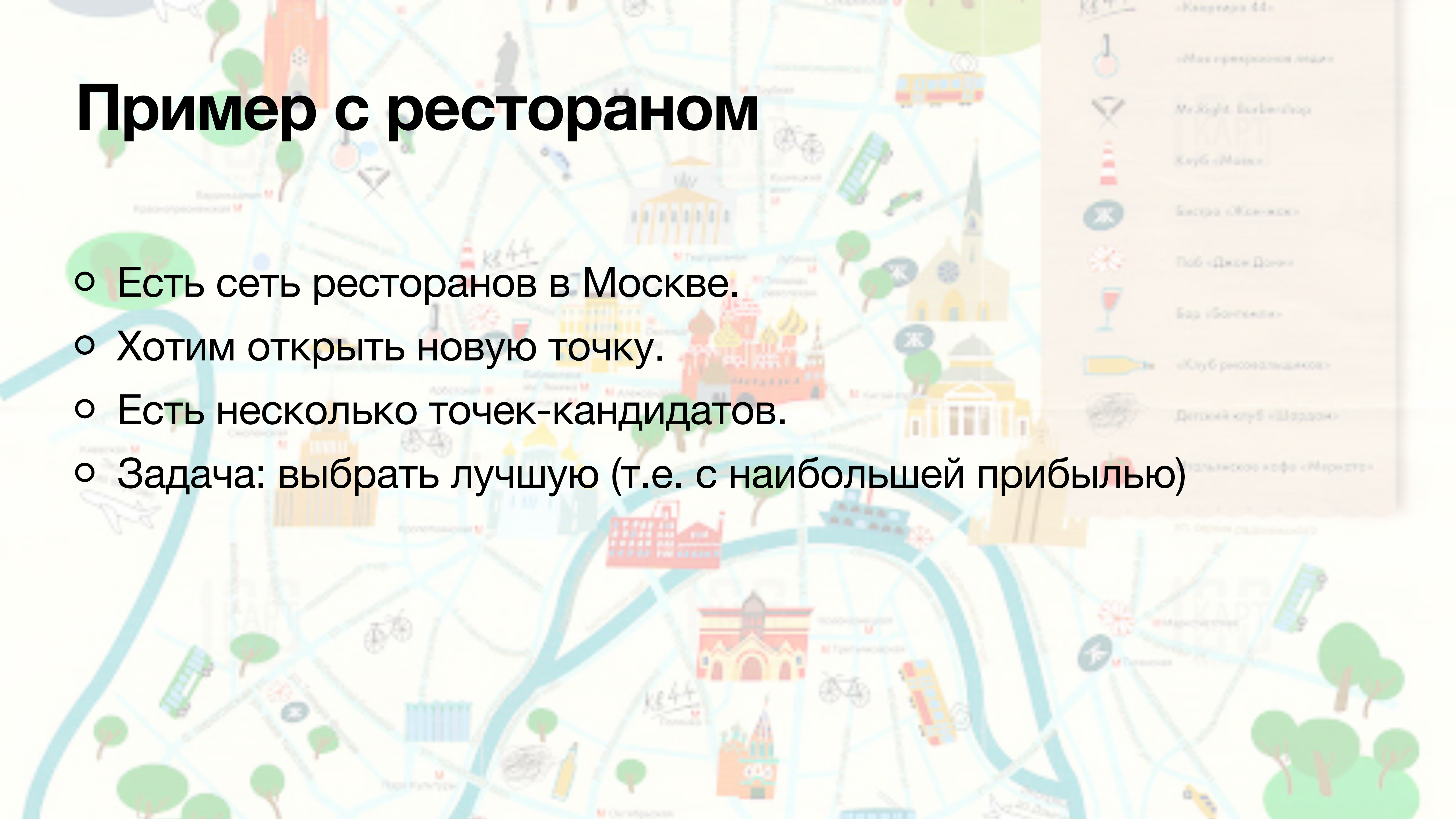
Ответ или целевая переменная - то, что хотим предсказать.
Обозначаем y , тоже нумеруем (y_1, y_2, \dots, y_n)

X - множество всех объектов, пространство объектов

Y - множество ответов

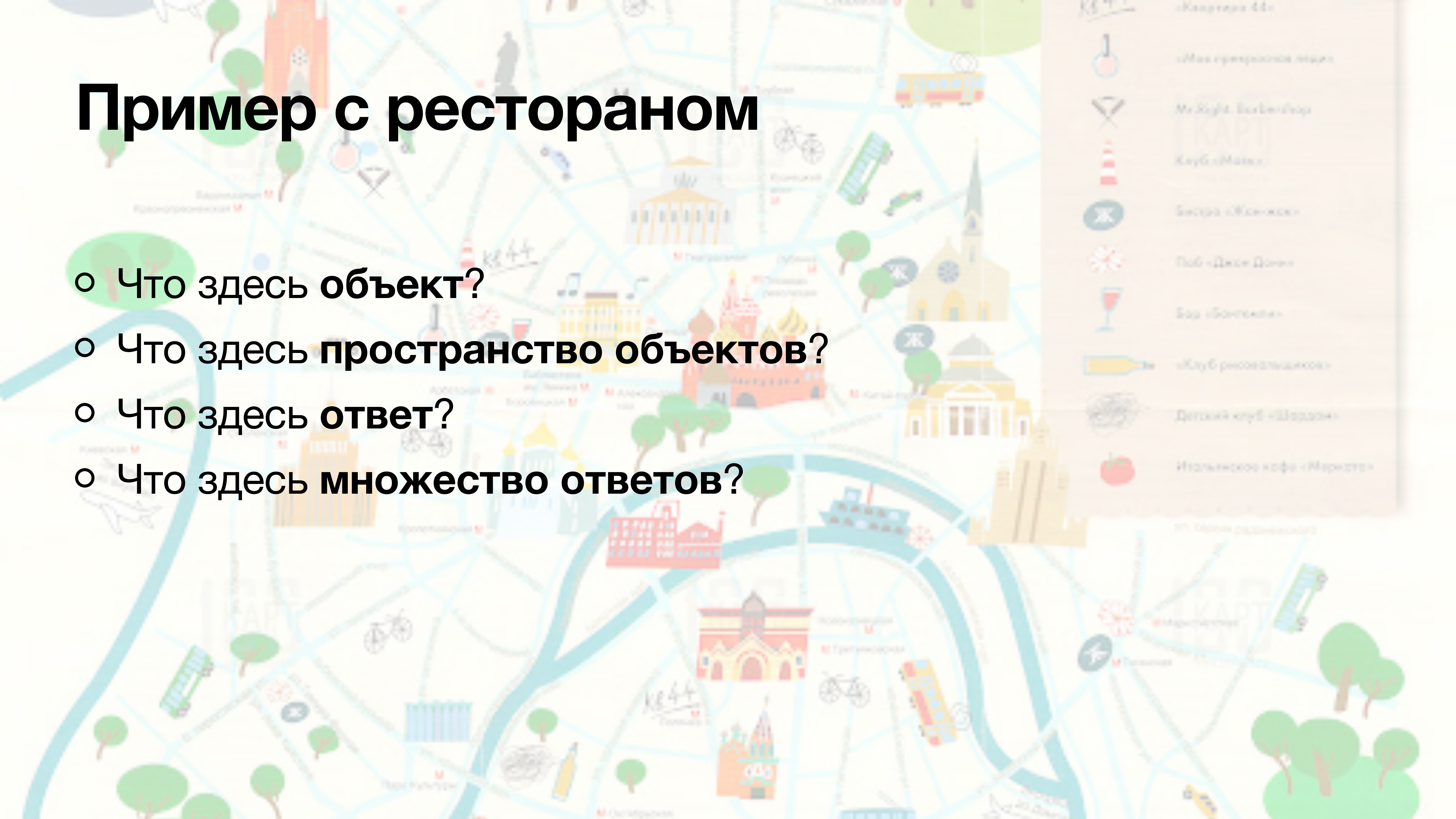
Пример с рестораном

- Есть сеть ресторанов в Москве.
- Хотим открыть новую точку.
- Есть несколько точек-кандидатов.
- Задача: выбрать лучшую (т.е. с наибольшей прибылью)



Пример с рестораном

- Что здесь объект?
- Что здесь пространство объектов?
- Что здесь ответ?
- Что здесь множество ответов?



Основные понятия

Чтобы **предсказать** выручку **новых ресторанов**, мы должны посмотреть на данные по нашим старым ресторанам (мы же владеем целой сетью!)

Обучающая выборка - конечный набор объектов, для которых известны значения целевой переменной. Т.е. это наши старые данные, прецеденты, с известной прибылью.

Основные понятия

Дано:

$\{x_1, \dots, x_n\} \subset X$ – обучающая выборка

$\{y_1, \dots, y_n\}, y_i = y(x_i)$ - известные ответы

Найти:

$a: X \rightarrow Y$ – алгоритм (решающую функцию),
приближающую y на всем множестве X

Объекты

Объекты — некие абстрактные сущности (точки размещения ресторанов), которыми компьютеры не умеют оперировать напрямую

Признаки (= факторы) - наборы характеристик, которыми мы описываем объекты

Пример с квартирами

Объект - квартира

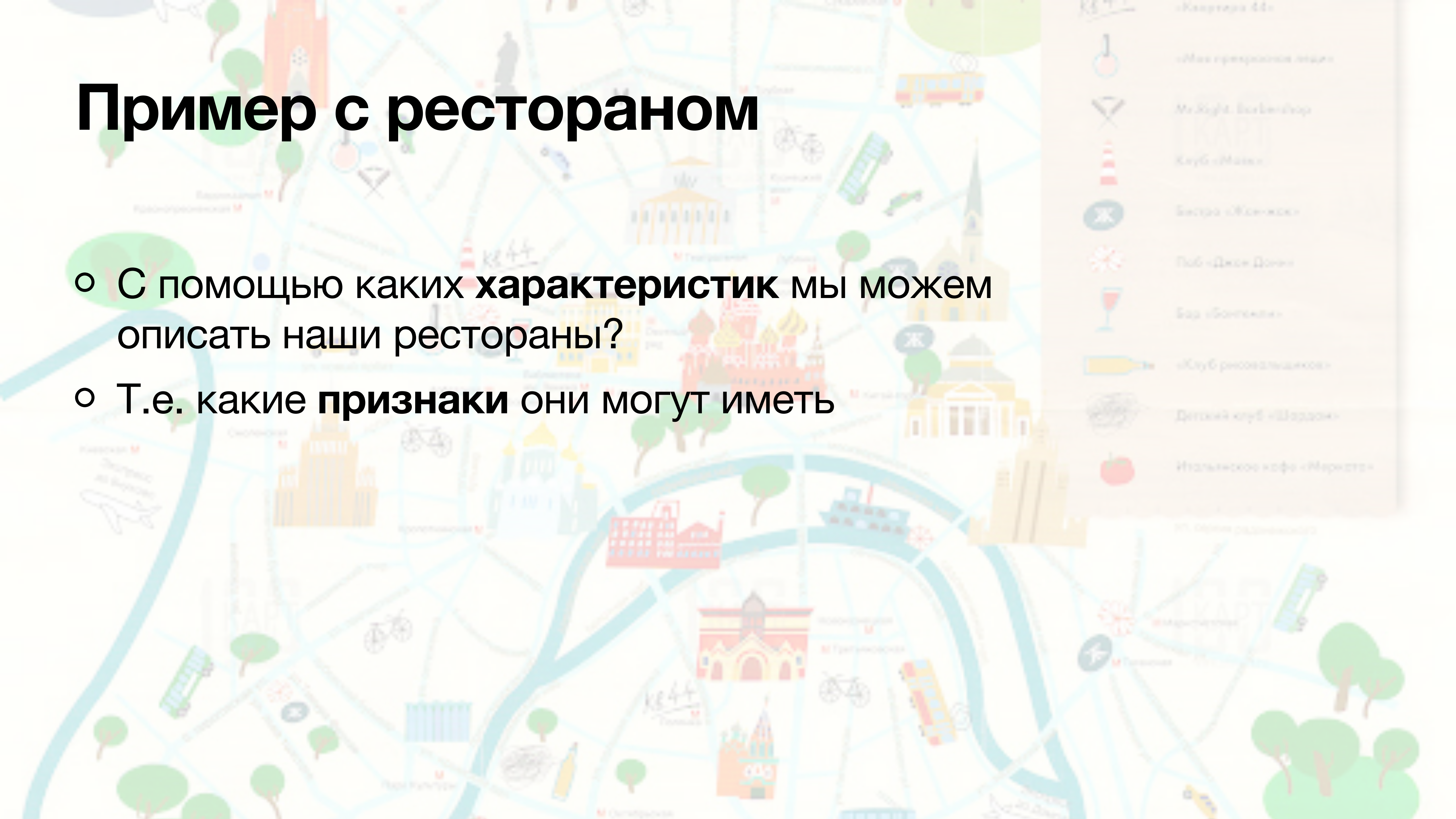
Целевая переменная - цена квартиры

Признаки (что может быть важно):

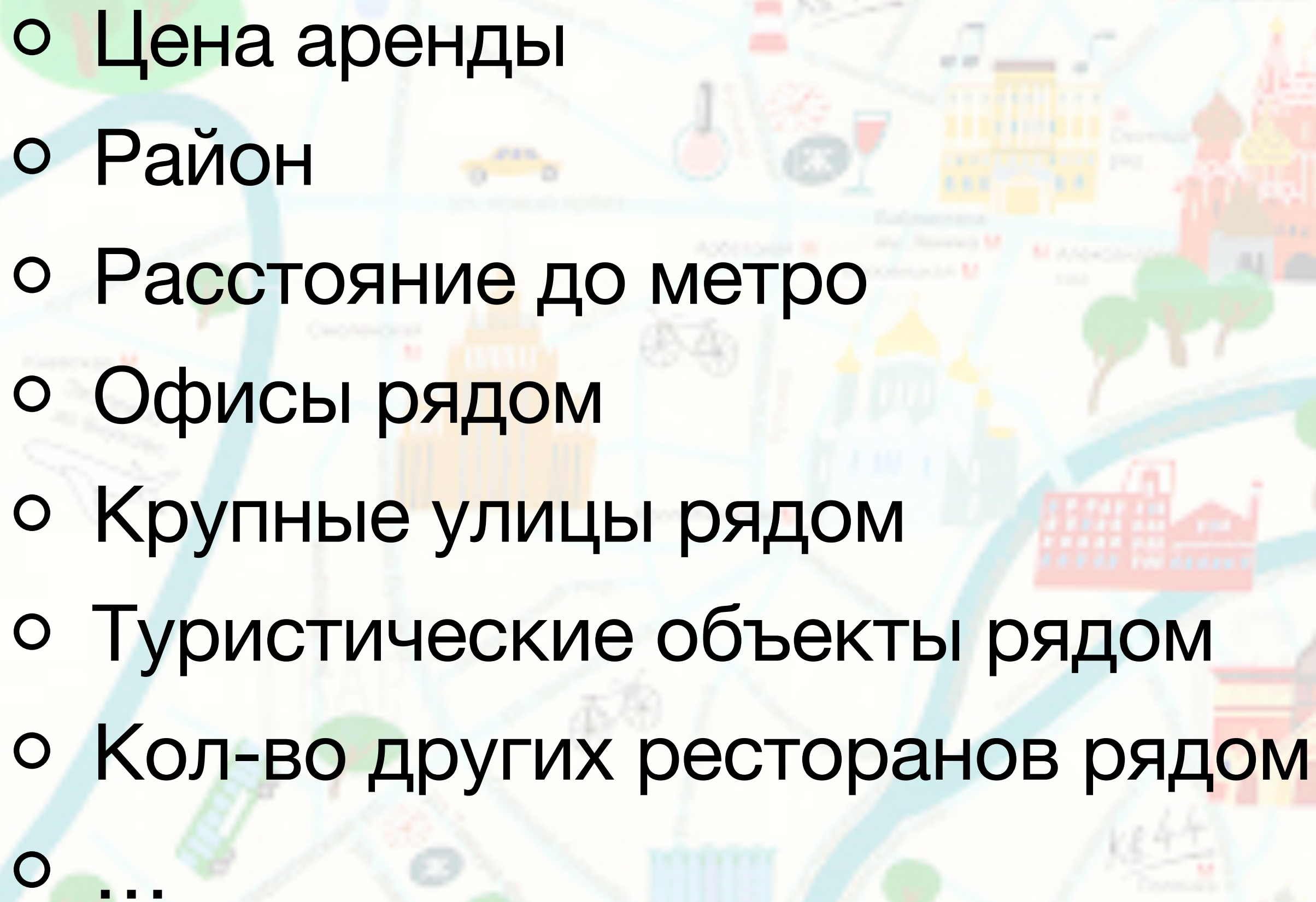
- метраж
- район
- расстояние до метро
- наличие балкона
- Год постройки дома
- ...

Пример с рестораном

- С помощью каких **характеристик** мы можем описать наши рестораны?
- Т.е. какие **признаки** они могут иметь



Пример с рестораном

- 
- Цена аренды
 - Район
 - Расстояние до метро
 - Офисы рядом
 - Крупные улицы рядом
 - Туристические объекты рядом
 - Кол-во других ресторанов рядом
 - ...

Матрица «объект-признак»

Числовая матрица:

	Признак 1	Признак 2	...	Признак K
Объект 1				
Объект 2				
Объект 3				
...				
Объект N				

Матрица признаков

Пример. Титаник

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Матрица признаков

Пример. Титаник

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Виды признаков

- **Бинарные**
- **Числовые**
- **Категориальные** (Принимают значение из множества)
- **Порядковые** (Упорядоченные категориальные по шкале)
- **Признаки со сложной структурой** (изображение)

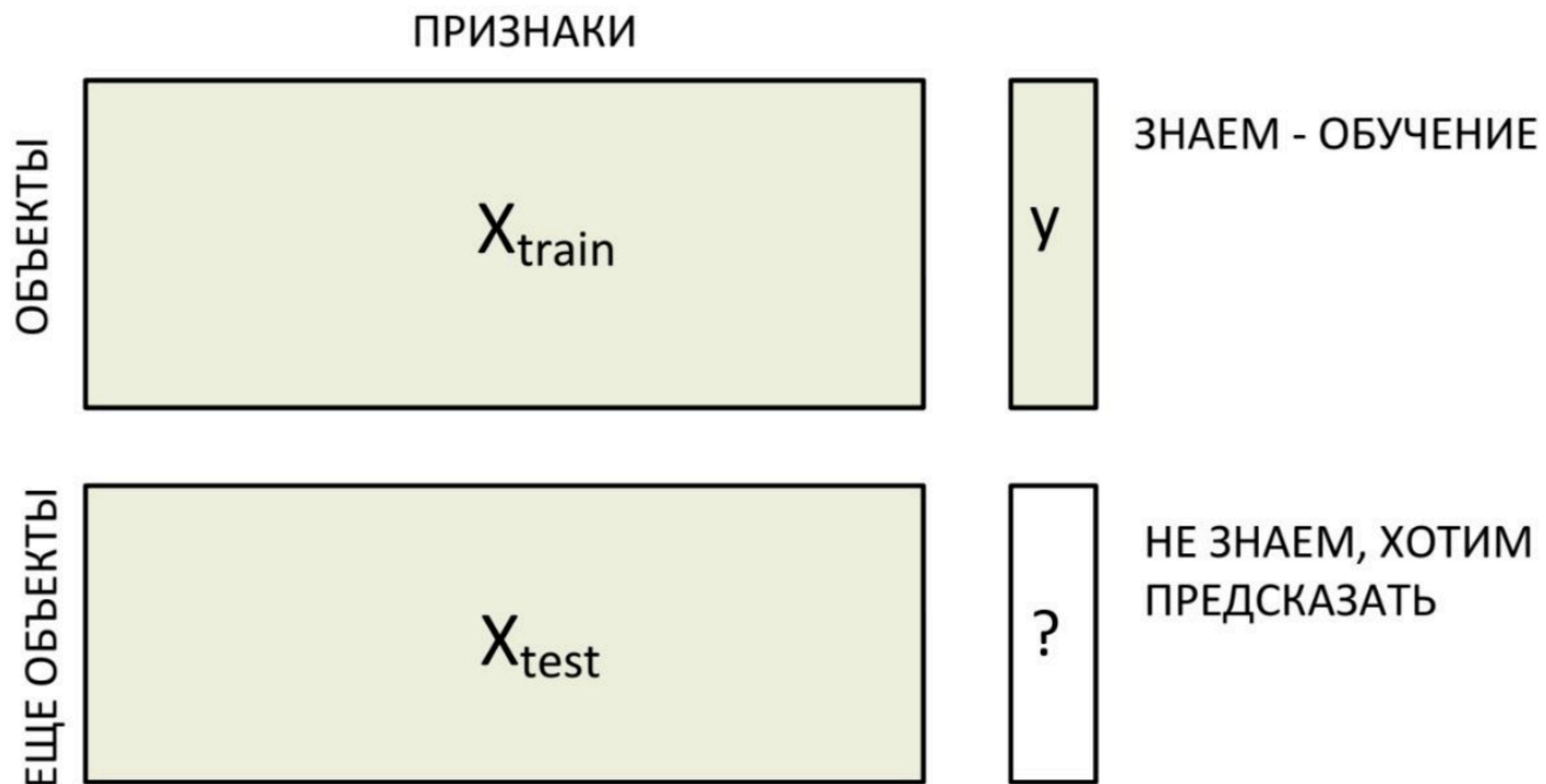
Признаки

- Разные виды признаков
- Машина не умеет работать не с числами, а признаки бывают не числовые
- Работа с ними сильно отличается по сложности

Сложные признаки

- Фотография как признак
- Как сделать числовым?
- Работа со сложными признаками - deep learning

Постановка задач МО



Задачи МО. Основные этапы

В задачах обучения с известными данными (обучение по прецедентам) всегда есть два этапа:

- Этап обучения (training):
по выборке $X=\{x_i, y_i\}$ строим алгоритм a
- Этап применения (testing):
алгоритм a для новых объектов x выдает ответы $a(x)$

Feature engineering

Специалист по анализу данных не является экспертом в предметной области – вся необходимая информация содержится в обучающей выборке.

Эксперты нужны при формировании признаков.

Линейная регрессия

Примеры задач

- Предсказание стоимости недвижимости
- Предсказание прибыли ресторана
- Предсказание поведения временного ряда в будущем (стоимость акций)
- Предсказание зарплаты человека

Линейная регрессия

Подбираем нужную прямую, которая бы хорошо описывала наши данные.

Т.е. наш алгоритм должен подобрать **уравнение прямой**

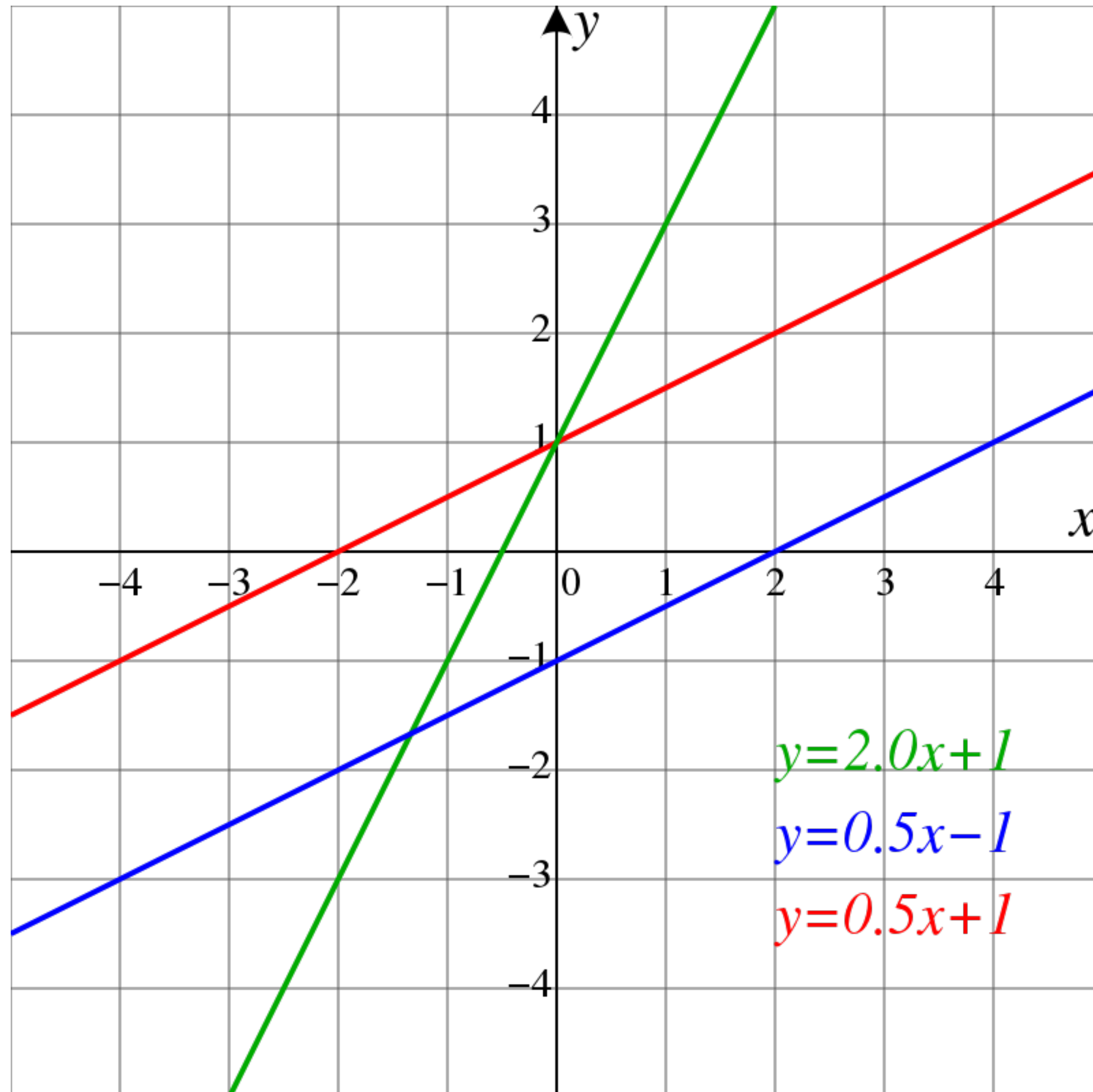
$$y = kx + b, \text{ где}$$

x - аргумент (переменная)

k - угловой коэффициент

b - свободный коэффициент

Уравнение прямой



$y = kx + b$, где

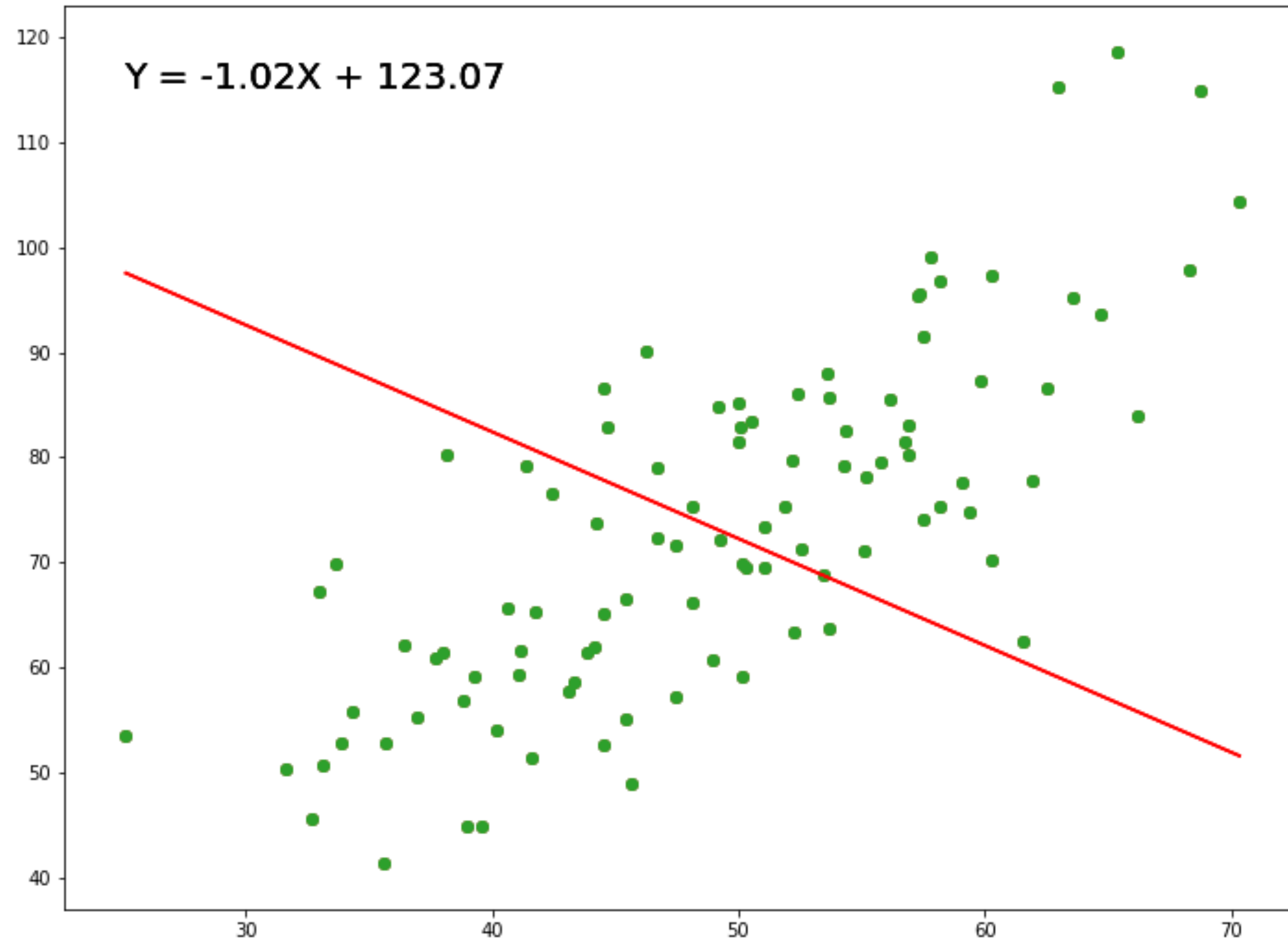
x - аргумент (переменная)

k - угловой коэффициент

b - свободный коэффициент

В МО обычно используют нотацию $a(x) = w_1x + w_0$

Линейная регрессия

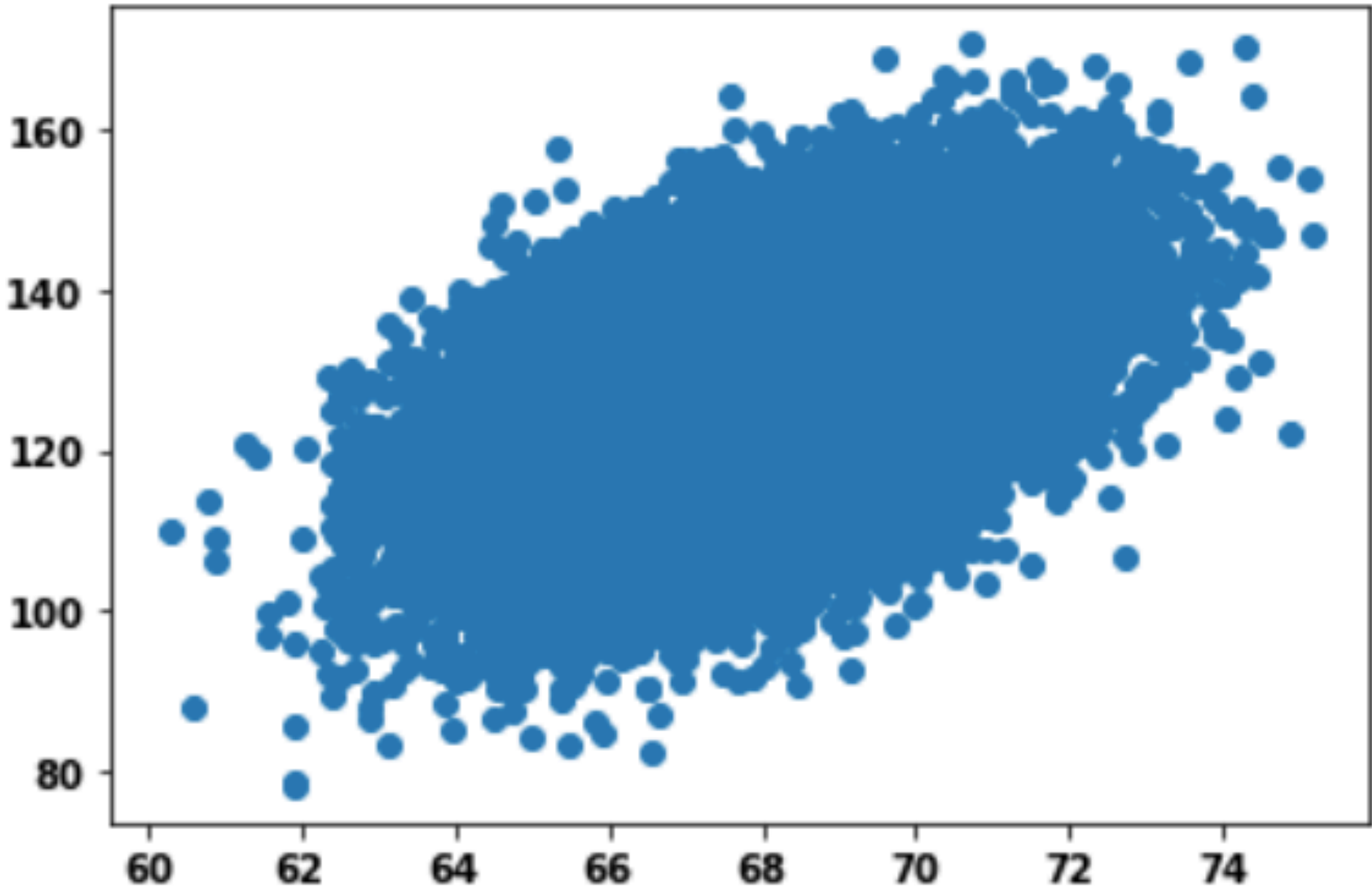


Линейная регрессия

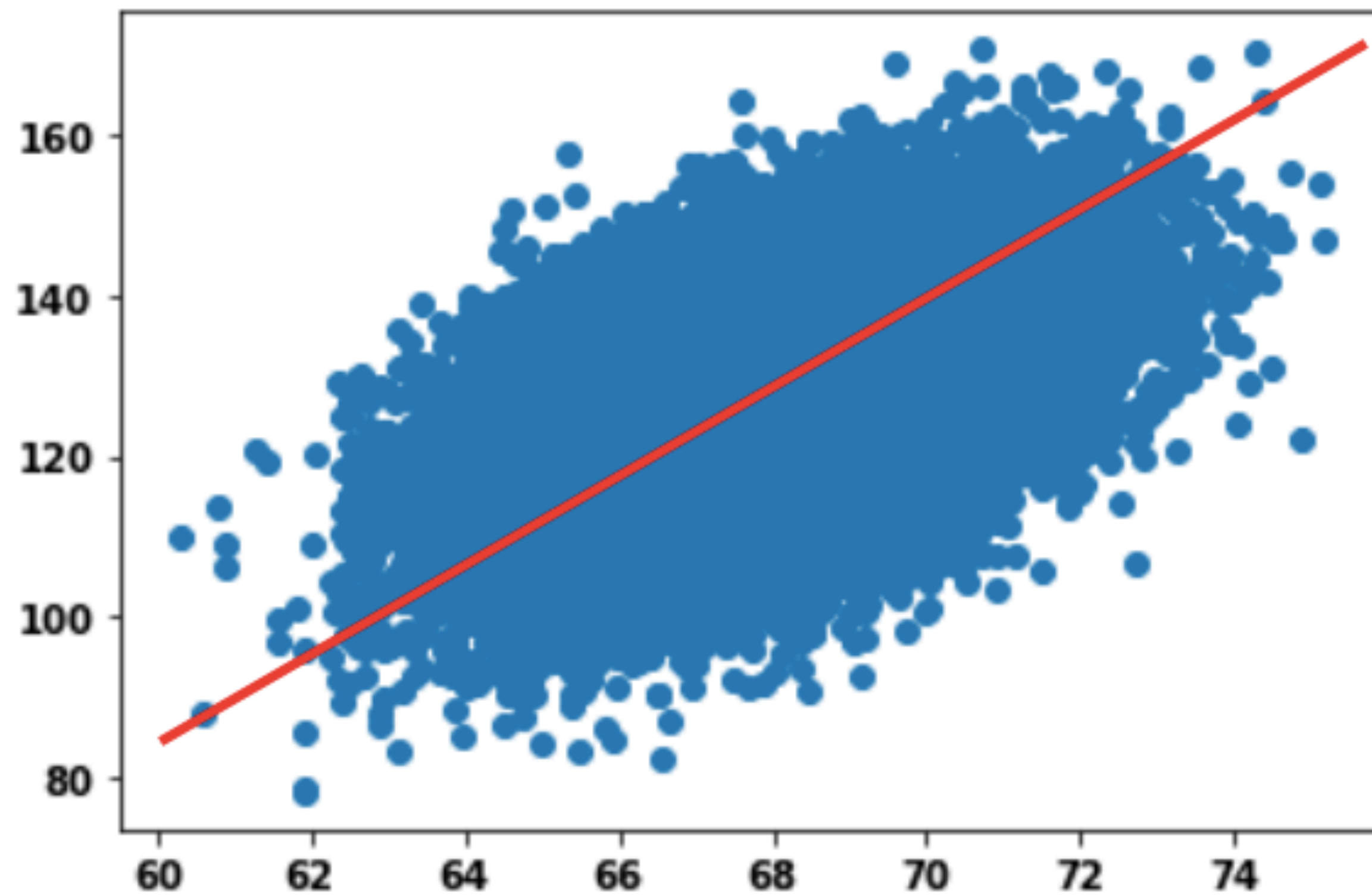
X Y

	Height	Weight
0	65.78331	112.9925
1	71.51521	136.4873
2	69.39874	153.0269
3	68.21660	142.3354
4	67.78781	144.2971
...
24995	69.50215	118.0312
24996	64.54826	120.1932
24997	64.69855	118.2655
24998	67.52918	132.2682
24999	68.87761	124.8740

25000 rows × 2 columns



Линейная регрессия



Не всегда закономерность в данных хорошо описывается прямой... но это уже другой разговор:)

Обучение регрессии

Допустим, мы хотим предсказать вес человека по его росту

	X	Y
	Height	Weight
0	65.78331	112.9925
1	71.51521	136.4873
2	69.39874	153.0269
3	68.21660	142.3354
4	67.78781	144.2971
...
24995	69.50215	118.0312
24996	64.54826	120.1932
24997	64.69855	118.2655
24998	67.52918	132.2682
24999	68.87761	124.8740

25000 rows × 2 columns

Используем линейную модель для предсказания

Она будет выглядеть так:

$$a(x) = w_0 + w_1 x_1,$$

где w_0 и w_1 – параметры модели (веса).

Обучение регрессии

Если бы у нас, кроме роста, был возраст, уравнение бы выглядело так:

$$a(x) = w_0 + w_1 x_1 + w_2 x_2$$

Вес Рост Возраст

(целевая переменная)

Общий вид линейных моделей: $\mathcal{A} = \{a(x) = w_0 + w_1 x_1 + \dots + w_d x_d \mid w_0, w_1, \dots, w_d \in \mathbb{R}\}$

Обучение регрессии

В общем виде, наша задача - построение функции

$$a : X \rightarrow Y$$

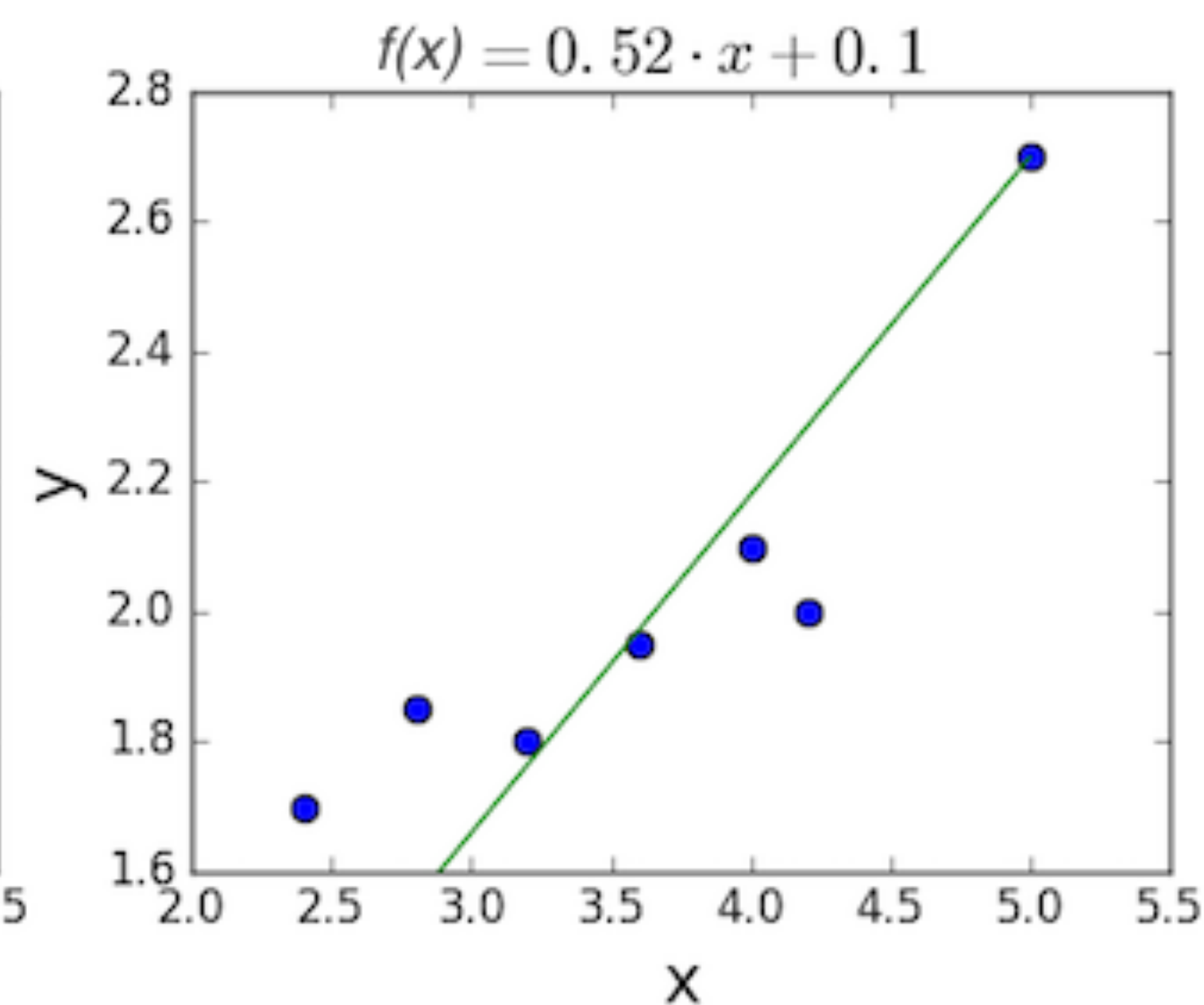
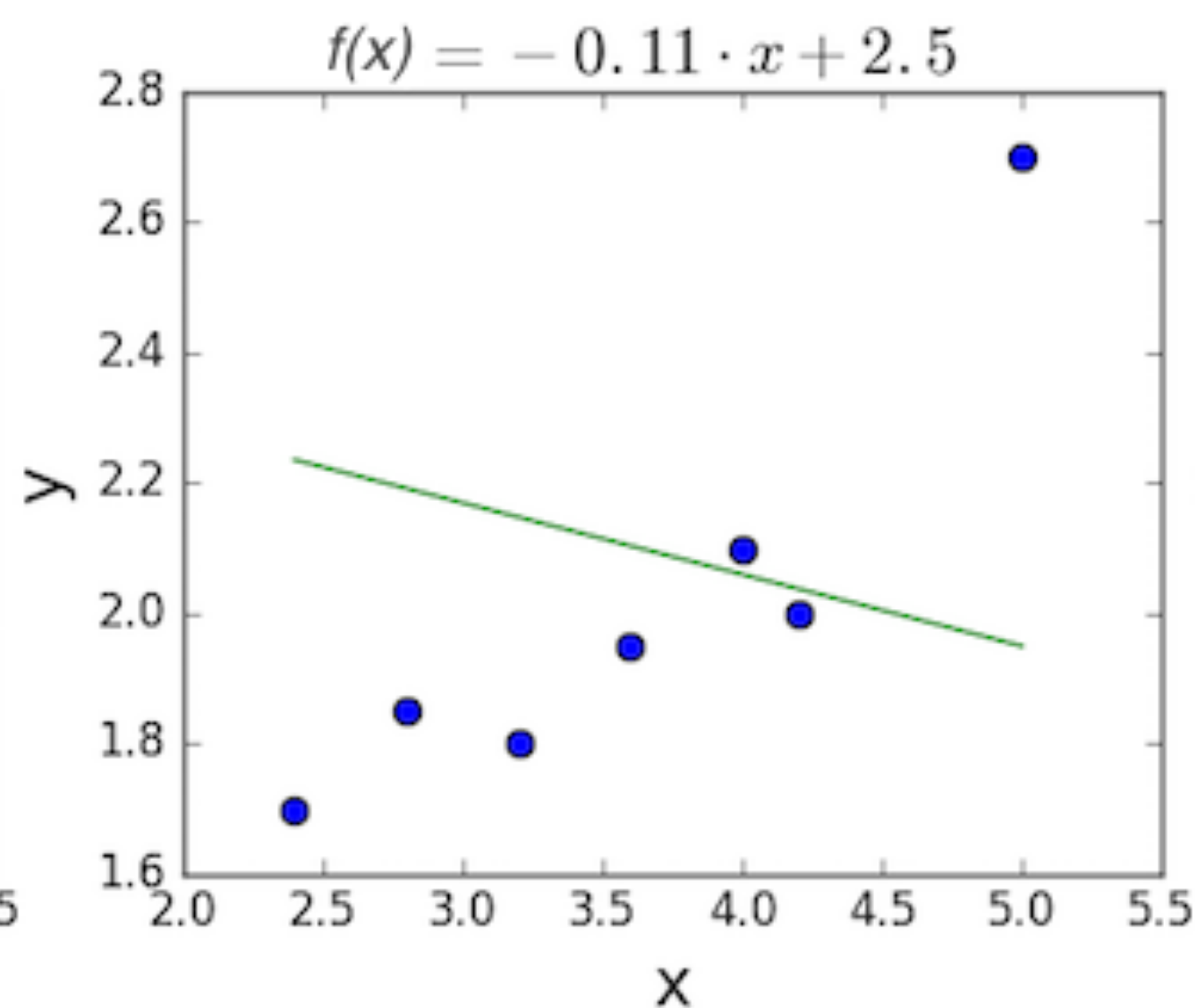
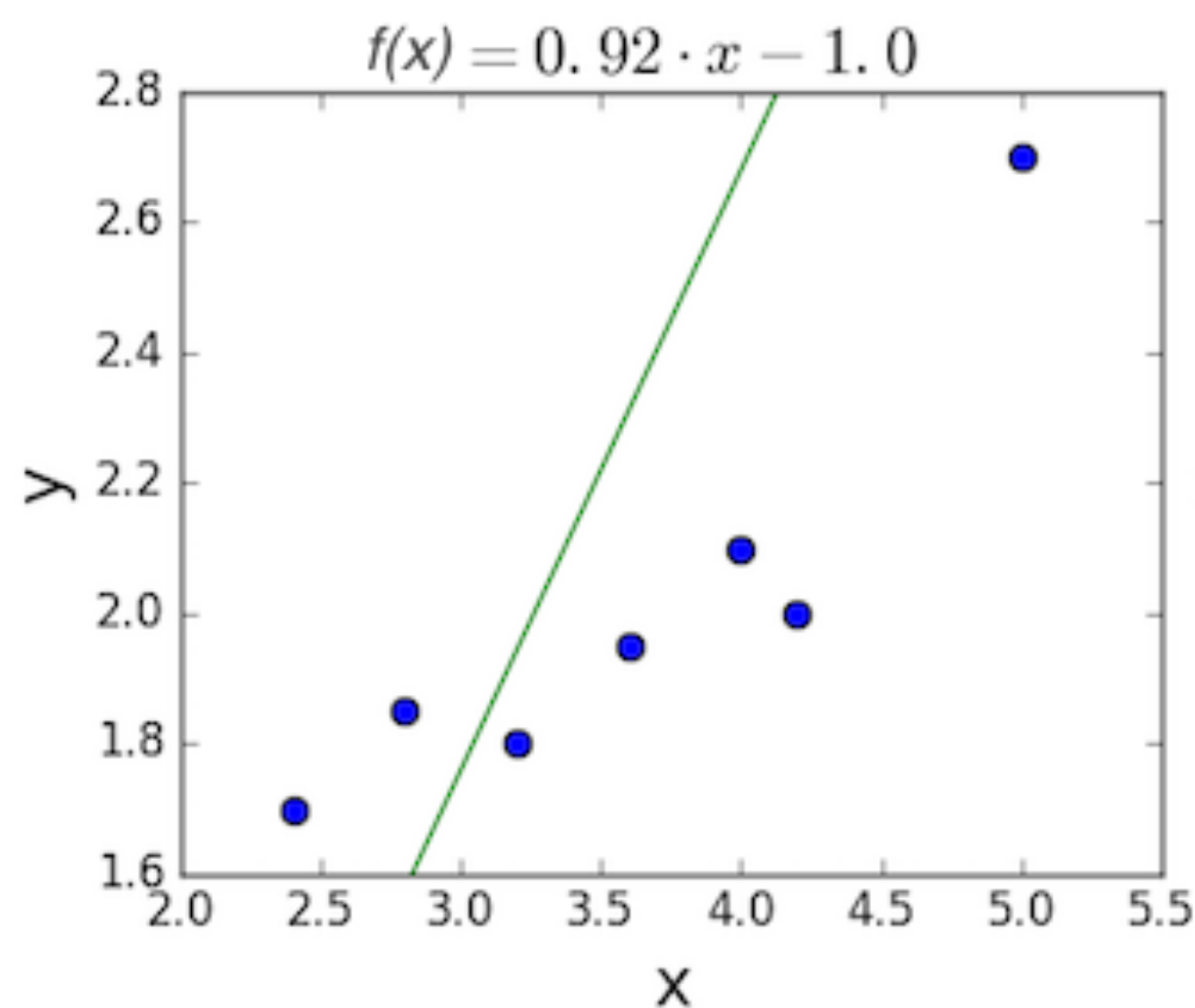
которая для любого нового объекта будет предсказывать ответ (y)

Такую функцию мы называем **алгоритмом** или **моделью**

Как понять, что модель хорошая?

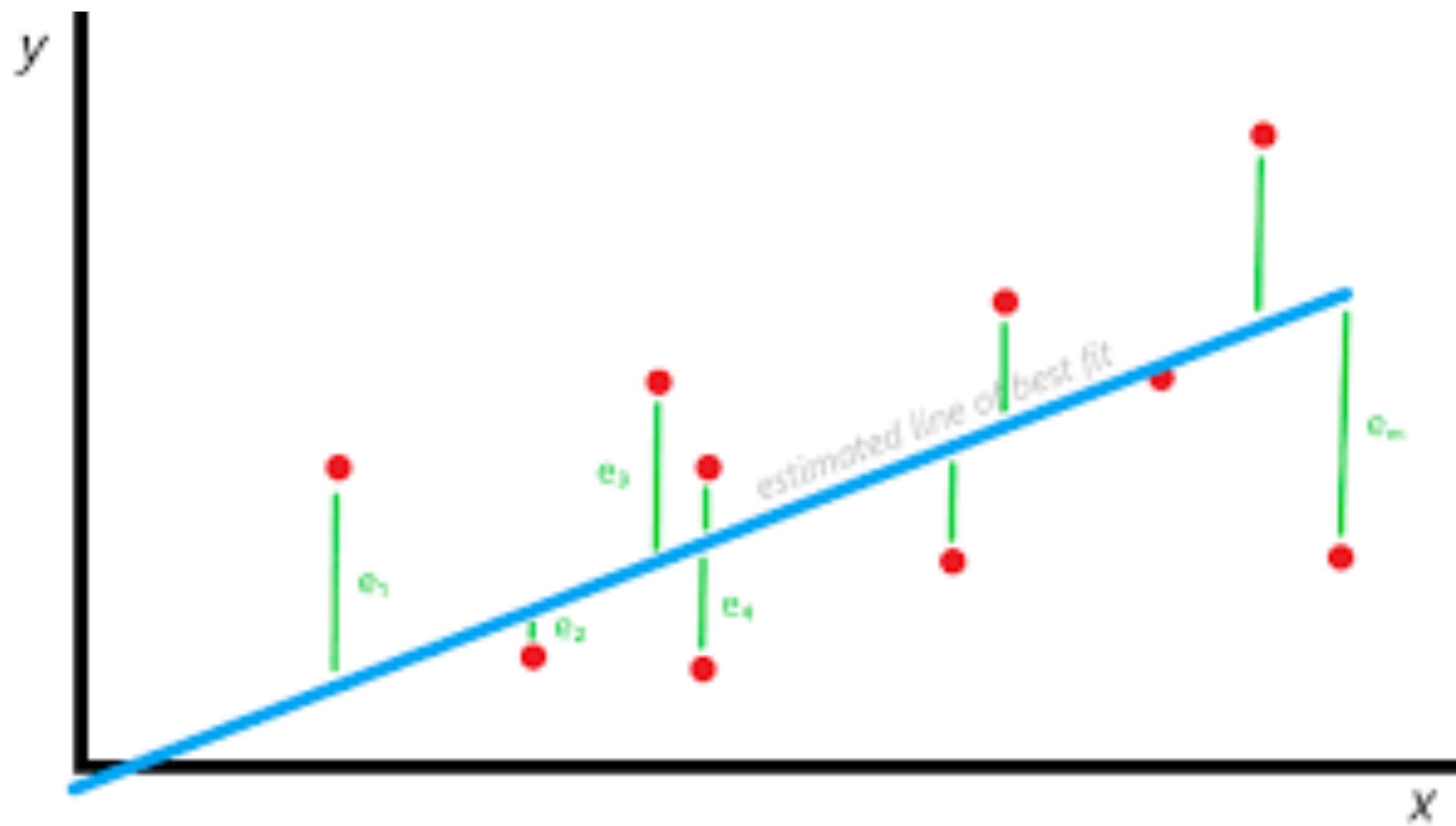
Функционал ошибки

Какую модель (т.е. уравнение прямой) выбрать?



Как понять, что модель хорошая?

Функционал ошибки



Как понять, что модель хорошая?

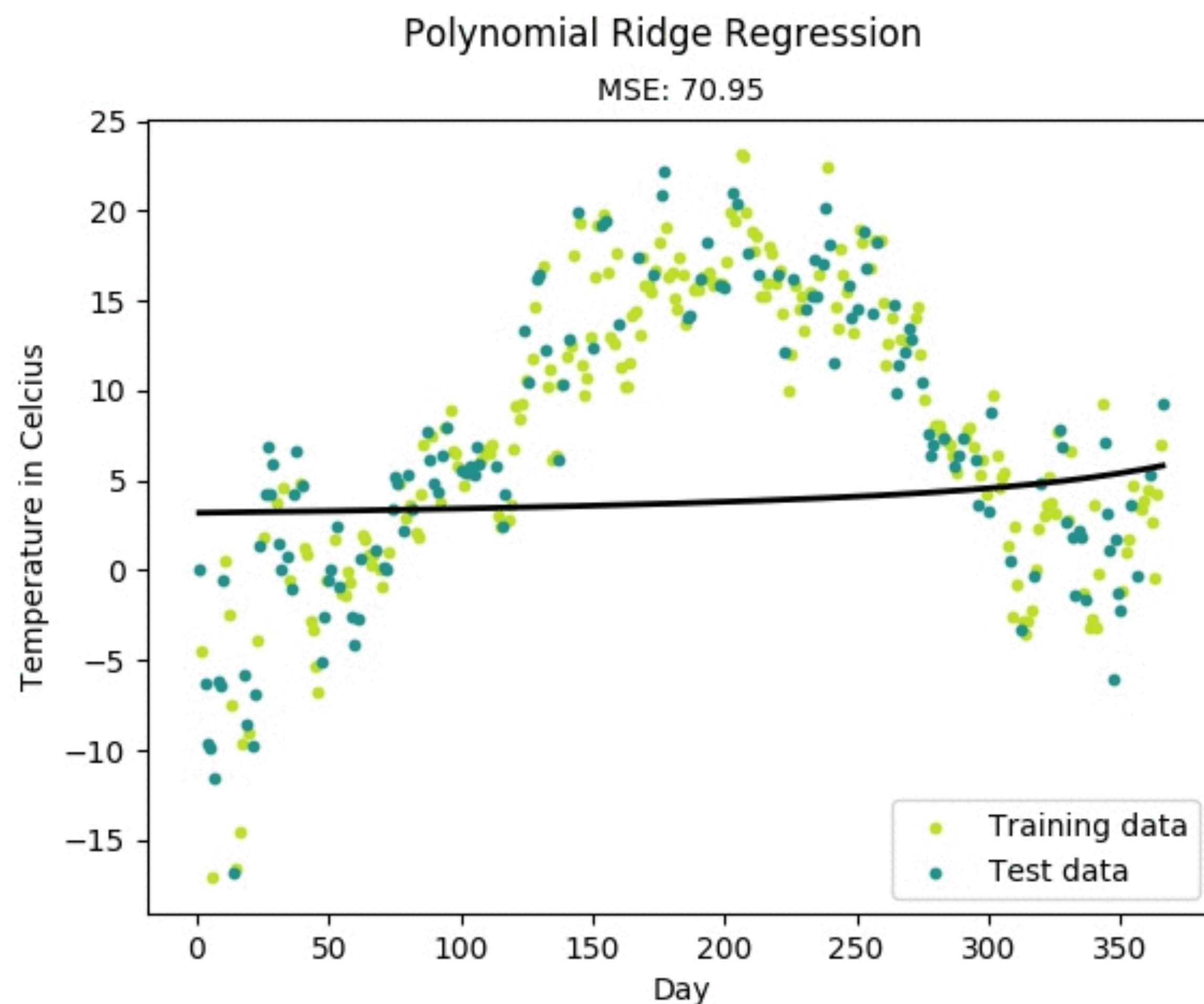
Функционал ошибки

Среднеквадратичная ошибка (mean squared error, MSE)

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

MSE

Чем меньше MSE, тем лучше модель



Регрессия не линейная,
НО СМЫСЛ ТОТ ЖЕ

Функционал ошибки

MSE

Параметры w_0 , w_1 , w_2 подбираются так, чтобы на них достигался минимум функции потерь (на обучающей выборке)

$$Q(a, X) = \frac{1}{l} \sum_{i=1}^l (w_0 + w_1 x_1 + w_2 x_2 - y_i)^2 \rightarrow \min_{w_0, w_1, w_2}$$

При обучении модели мы минимизируем функционал ошибки

Обучение модели

Обучение - процесс поиска оптимального алгоритма
(оптимального набора *весов*)

Обучение модели

Ок, модель обучили (подобрали нужные веса) - но это на **обучающей выборке**. Что дальше?

Нужно применить нашу модель (формулу) к **тестовой выборке** и оценить, как хорошо модель работает на новых данных.

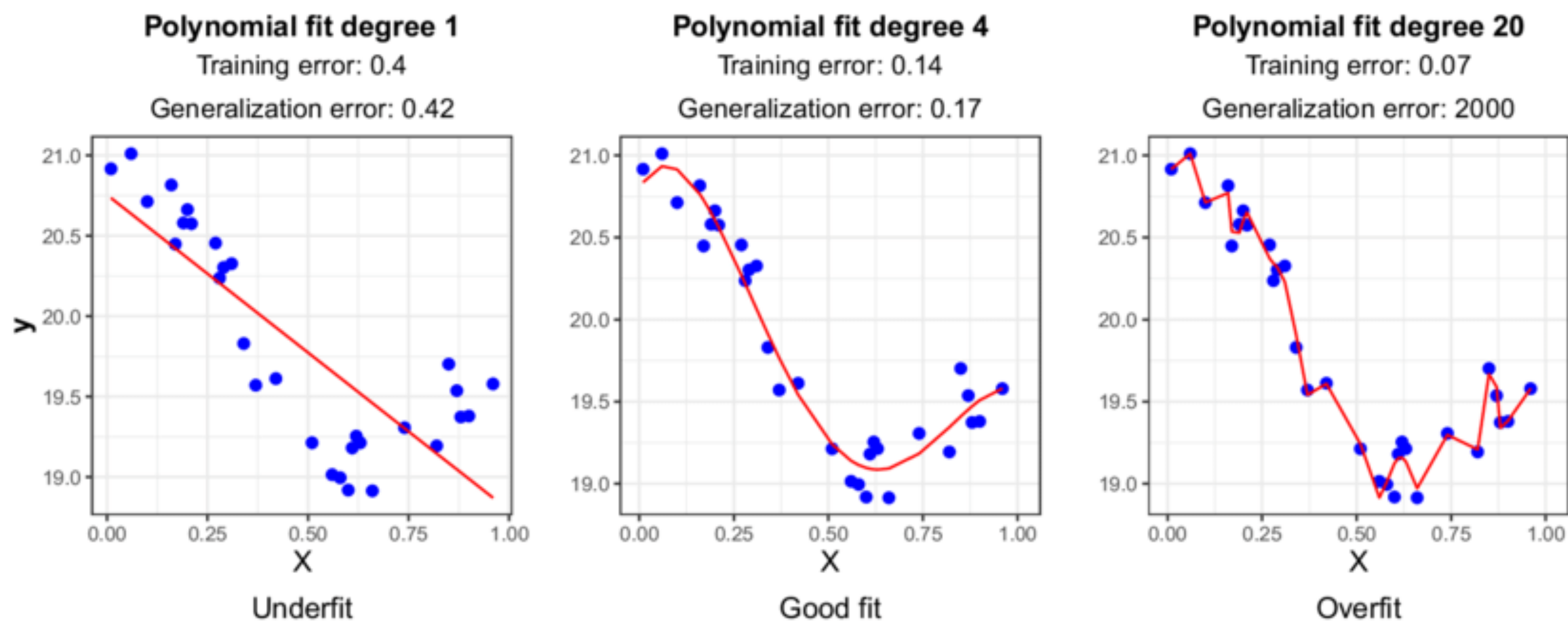
Нужна **метрика качества**.

Часто измерения делают тоже с помощью MSE.

Переобучение и недообучение

Недообучение - модель плохо описывает данные

Переобучение - подгон модели под обучающие данные



Алгоритм решения задач

1. Постановка задачи
2. Выделение признаков
3. Формирование выборки
4. Выбор функции потерь и метрики качества
5. Предобработка данных
6. Построение модели (функционал ошибок, на train-set)
7. Оценивание метрики качества (на новых данных)