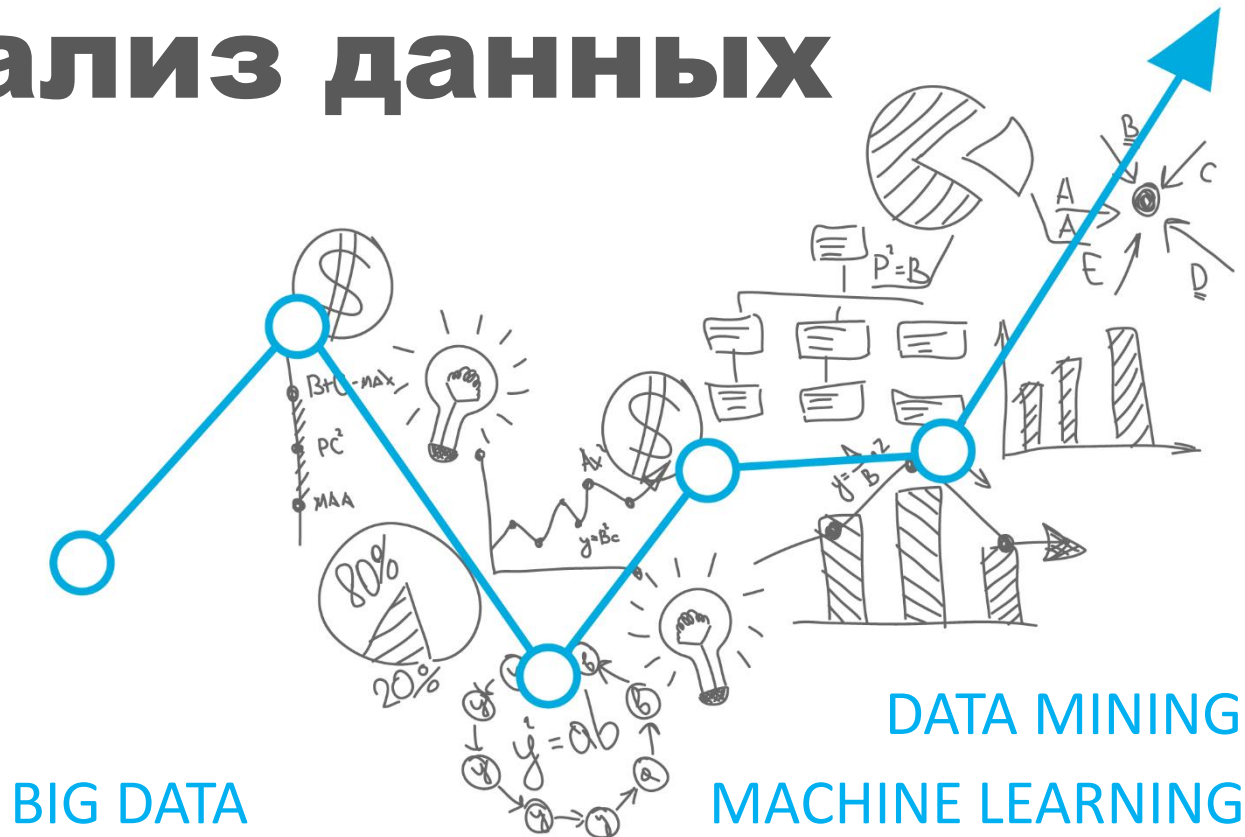


# Интеллектуальный анализ данных



## Лекция 2. Введение в машинное обучение

- Методы предсказания и классификации (линейная и логистическая регрессия, нейронные сети, машины опорных векторов и другие)
- Методы кластеризации (метод k-средних и другие)
- Техники диагностики качества обучения и повышения их точности
- Применение машинного обучения в обработки больших данных
- Практическое использование машинного обучения

- ❑ Информационный портал по машинному обучению

<http://www.machinelearning.ru/>

- ❑ Курс «Machine Learning» на Coursera

<https://www.coursera.org/learn/machine-learning/>

- ❑ Курсы в интернет

<https://proity.ru/analytics/data-science/>

- ❑ Чубукова И.В. Data Mining: учебное пособие

[https://biblioclub.ru/index.php?page=book\\_red&id=233055](https://biblioclub.ru/index.php?page=book_red&id=233055)

Интеллектуальный  
анализ данных

Машинное  
обучение

Базы данных



Информационный  
поиск

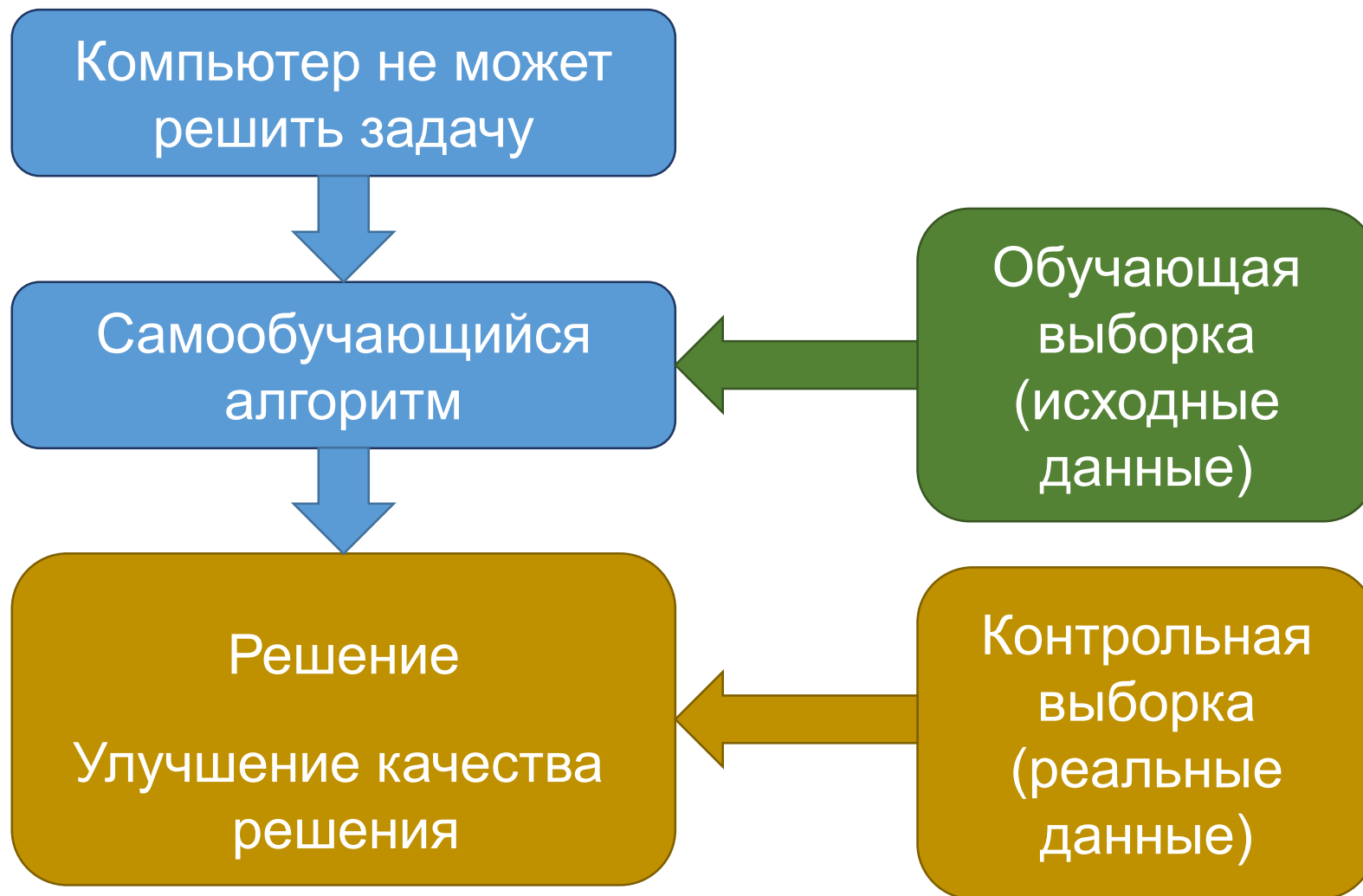


Большие  
данные

добыча данных, раскопка данных, "промывание"  
данных  
извлечение  
информации,  
извлечение  
знаний, анализ шаблонов, информационная  
проходка данных, обнаружение знаний в базах  
данных

Data mining

Big Data



- ❑ Database mining: получение информации из больших объемов данных (переходы по ссылкам веб-страниц, медицинские истории, биоинформатика, ...)
- ❑ Задачи, которые не могут быть хорошо запрограммированы вручную (автономные транспортные средства, распознавание символов, обработка текстов на естественном языке, компьютерное зрение, ...)
- ❑ Самообучаемые программы (системы рекомендаций интернет-магазинов, принятия решений, ...)
- ❑ Искусственный интеллект (имитация работы мозга животных, ...)
- ❑ Множество других областей

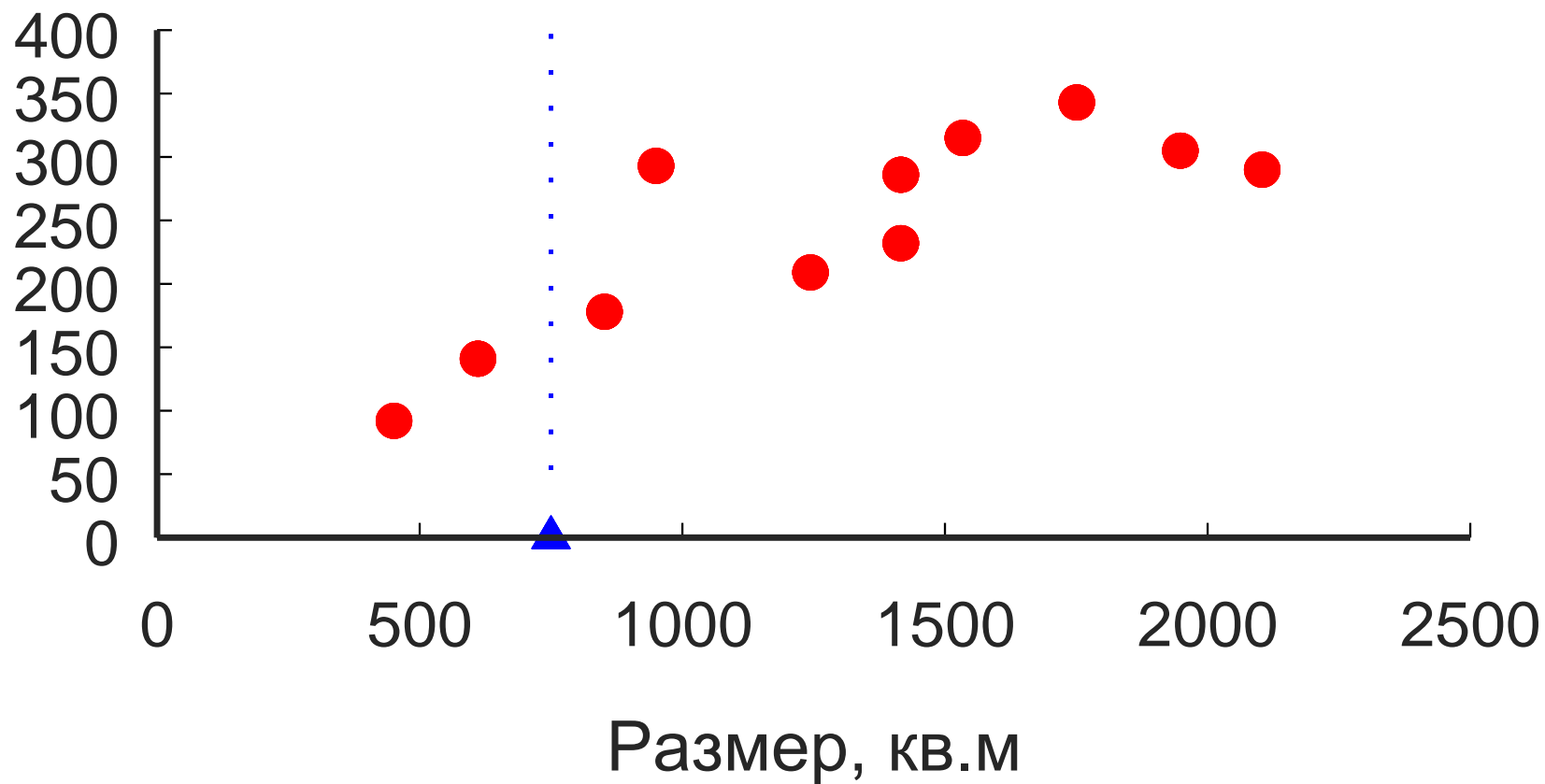
№ объекта	Площадь, м <sup>2</sup>	Цена, тыс. м.к.
1	2104	460
2	1416	232
3	1534	315
4	852	178
5	1948	305
6	950	293
7	611	141
8	1751	343
9	451	102
10	1244	209
11	1416	286



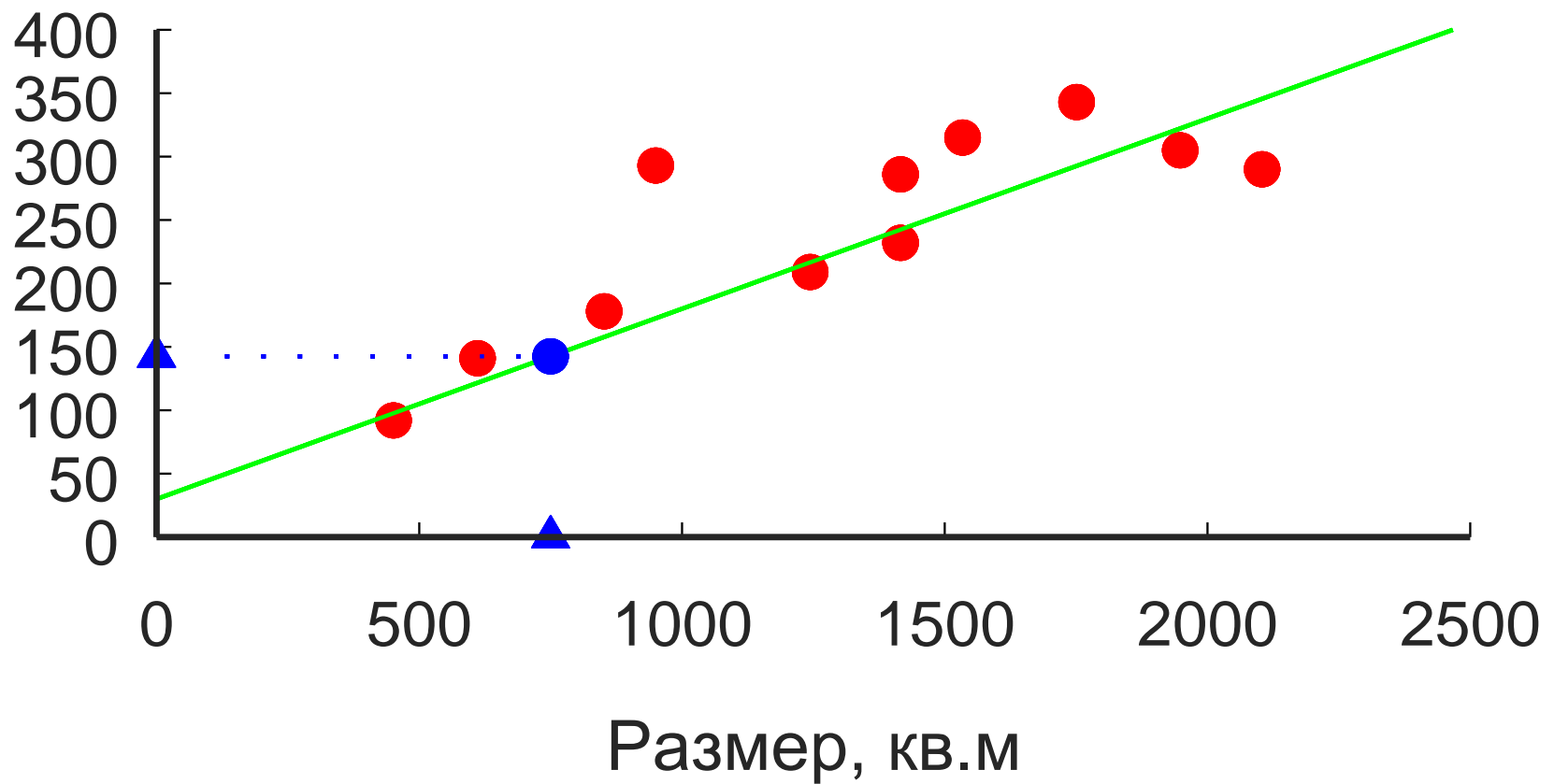
Имеются данные  
о недвижимости  
на Марсе и ее  
стоимости

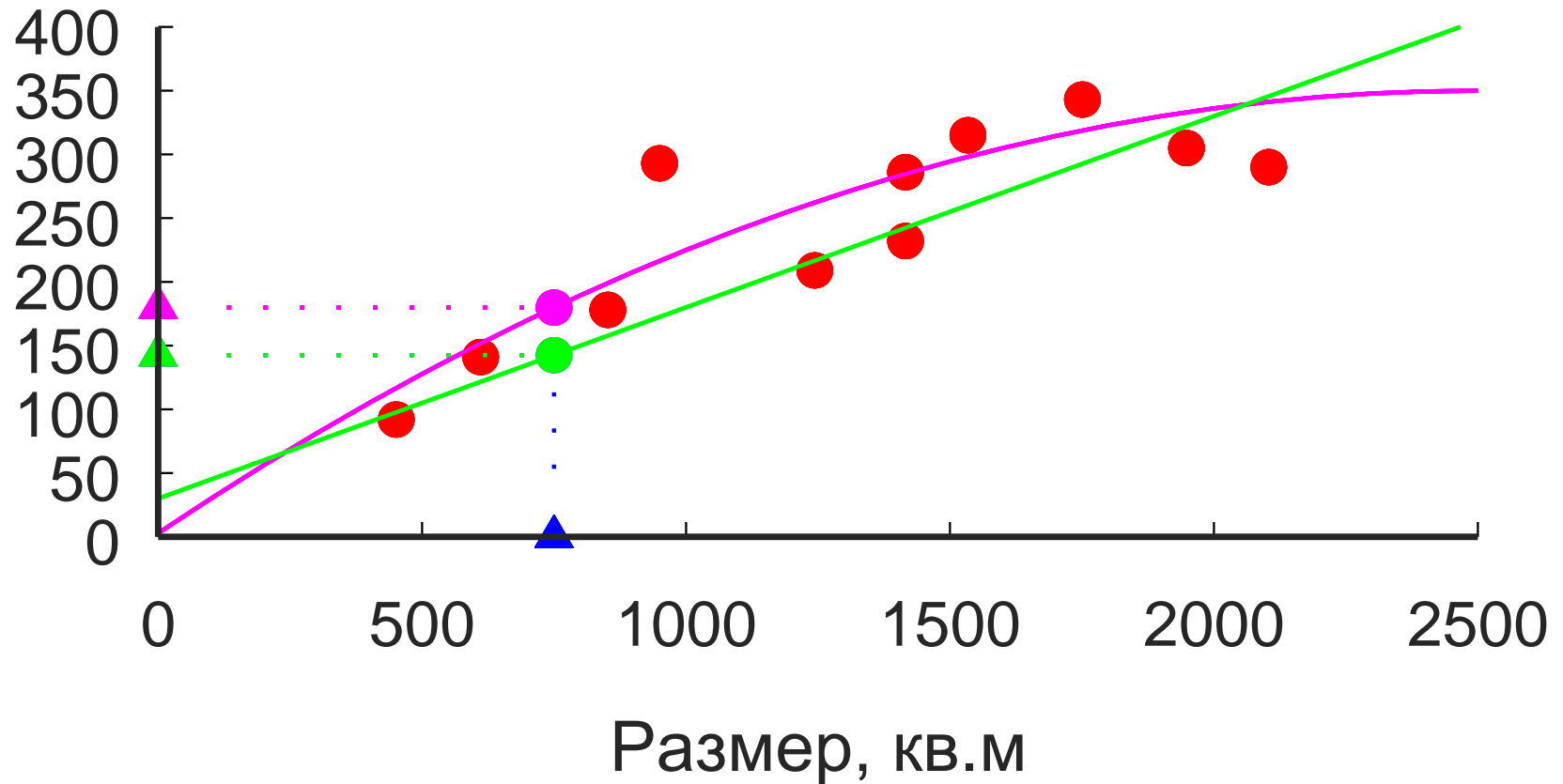
Можно ли на основе  
этих данных  
предсказать, сколько  
будет стоить дом  
площадью **750** м<sup>2</sup> ?







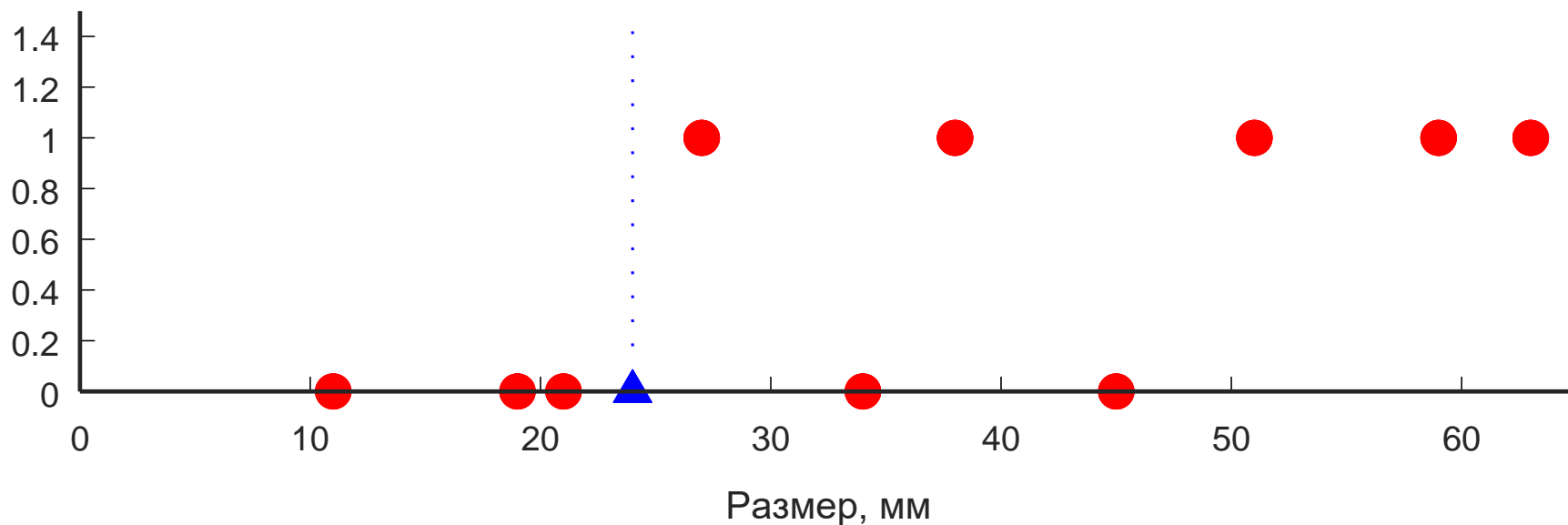




Сведения о пациентах, больных раком

Размер опухоли	11	19	21	27	34	38	45	51	59	63
Злокачественная	0	0	0	1	0	1	0	1	1	1

Если у нас появится пациент с опухолью размером 24, какая она у него будет, скорее доброкачественная или злокачественная?

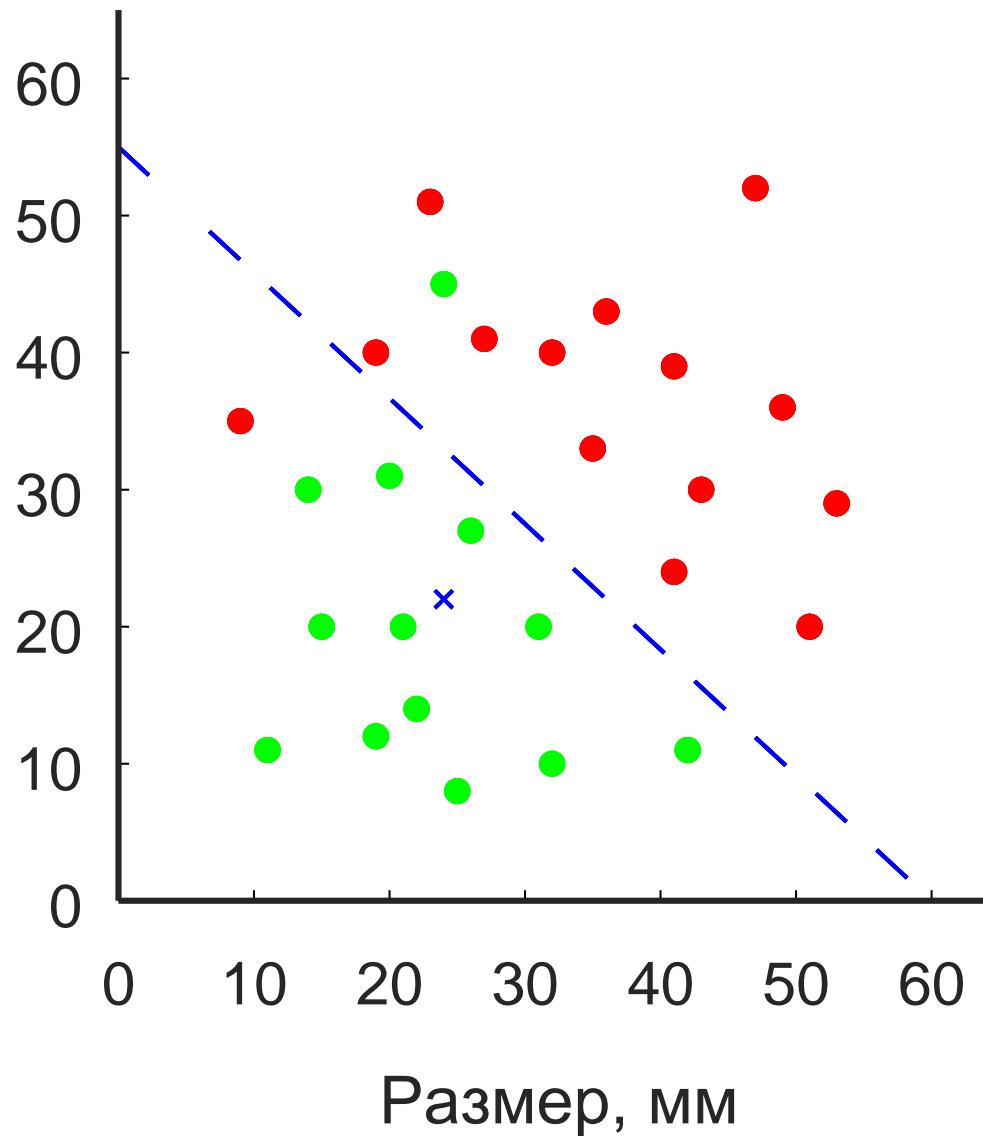


Размер меньше 35 — доброкачественная  
Размер больше или равно 35 — злокачественная

Сведения о пациентах, больных раком

Размер опухоли	15	19	27	24	21	9	51	...	24	47
Возраст	20	12	41	45	20	35	20	...	45	52
Злокачественная	0	0	1	0	0	1	1	...	0	1

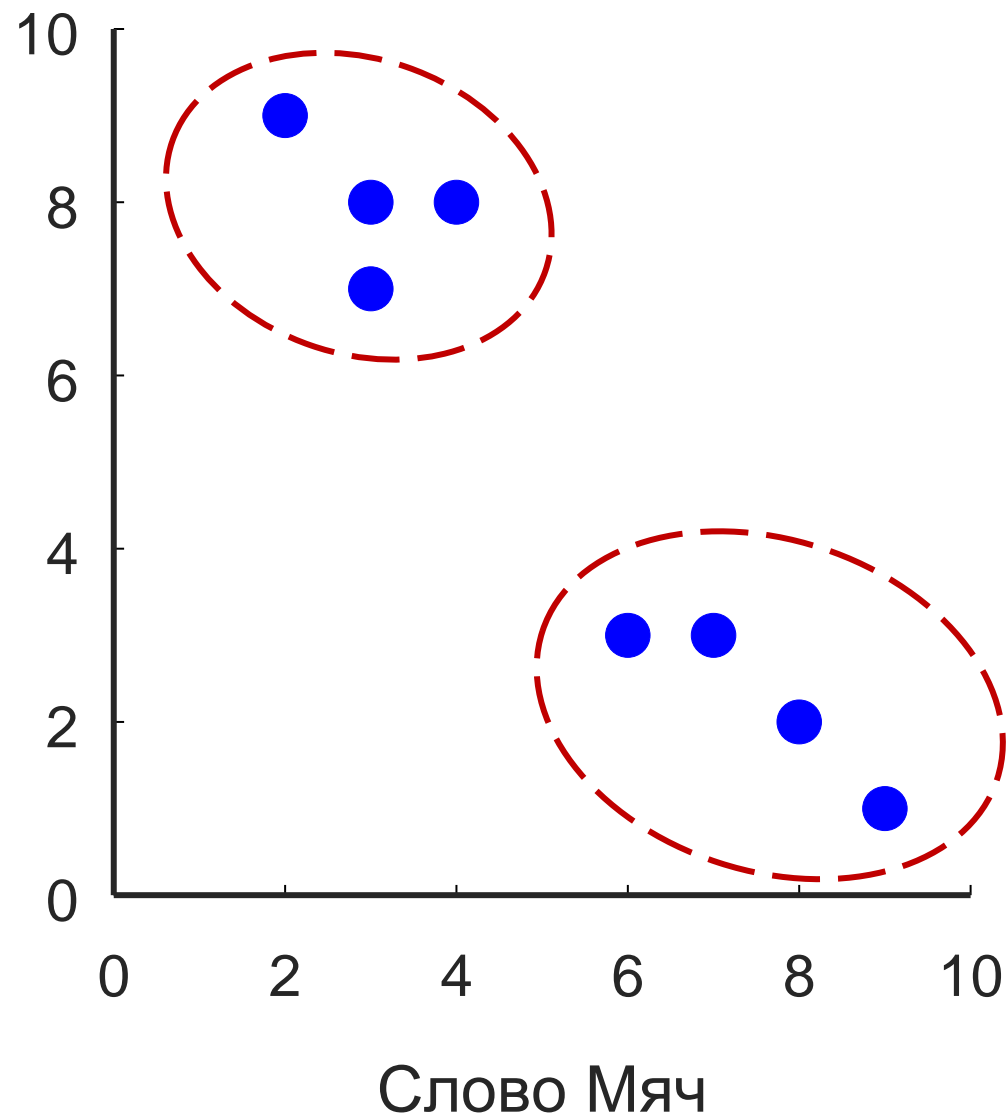
Если у нас появится пациент с опухолью размером 24 и возрастом 22, какая она у него будет, скорее доброкачественная или злокачественная?



Параметры  
линии?

Имеется подборка новостных статей.  
 Задача: группировать их по близким темам.  
 Путь решения: выделить ключевые слова.

Статья	Кол-во слов «Мяч»	Кол-во слов «Реформа»
Нnnnn	6	3
Ееееее	8	2
Ппппп	9	1
Рrrrrrrr	7	3
Иииииии	2	9
Дддд	3	7
Уuuuuu	4	8
Ммм	3	8





**Машинное обучение** – область науки, которая изучает возможность компьютеров обучаться без необходимости их непосредственно программировать [Артур Самюэл, 1959 г.]

Артур Самюэл написал программу игры в шашки, которая обучалась, играя сама против себя.

Говорят, что компьютерная программа **обучается** на основе опыта  $E$  по отношению к некоторому классу задач  $T$  и меры качества  $P$ , если качество решения задач из  $T$ , измеренное на основе  $P$ , улучшается с приобретением опыта  $E$ .

[T.M. Mitchell. Machine Learning. McGraw-Hill, 1997]

Говорят, что компьютерная программа *обучается* на основе опыта  $E$  по отношению к некоторому классу задач  $T$  и меры качества  $P$ , если качество решения задач из  $T$ , измеренное на основе  $P$ , улучшается с приобретением опыта  $E$ .

[T.M. Mitchell. Machine Learning. McGraw-Hill, 1997]

## Пример 1: Игра в шашки

Задача  $T$  — способность выиграть в шашки

Опыт  $E$  — множество сыгранных партий

Качество  $P$  — вероятность выигрыша в следующей партии

## Пример 2: Распознавание рукописных символов

Задача  $T$  — распознавание рукописного текста

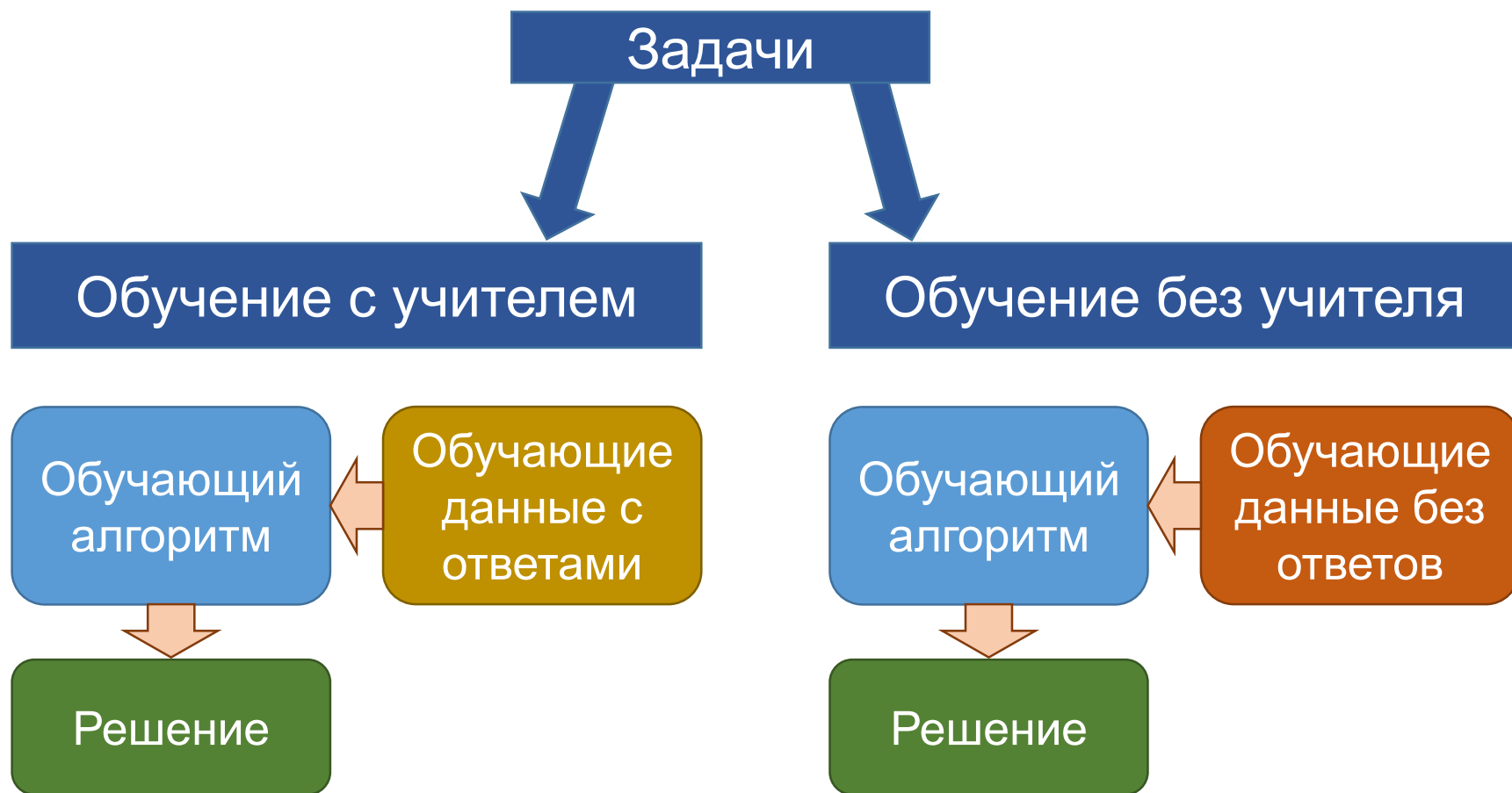
Опыт  $E$  — различные примеры написанных символов

Качество  $P$  — количество правильно распознанных символов

Существует программа, которая изучает, какие электронные письма были отмечены пользователем как спам, а какие как не спам. Основываясь на этих данных, программа определяет в дальнейшем, является ли новое полученное письмо спамом или не является.

Основываясь на определении Митчелла, ответьте, что является задачей  $T$ , опытом  $E$  и критерием  $P$  для этой программы?

- Опыт  $E$*     ☒ база данных меток пользователя
- Задача  $T$*     ☒ классификация писем на спам/не спам
- Критерий  $P$*     ☒ количество писем, правильно классифицированных как спам



Примеры:

- предсказание цен на недвижимость (цена дома)
- медицинская диагностика (злокачественная)

Пример:

- классификация текста

- $x^{(i)}$  — входные переменные в пространстве  $X$  (для примера с недвижимостью  $x^{(i)}$  это площадь,  $X \in \mathbb{R}$ )
- $y^{(i)}$  — выходные переменные в пространстве  $Y$  (в примере с недвижимостью  $y^{(i)}$  это стоимость,  $Y \in \mathbb{R}$ )
- пара  $(x^{(i)}, y^{(i)})$  называется обучающим примером
- $(x^{(i)}, y^{(i)}), i=1, \dots, t$  называется обучающей выборкой

*Задачей обучения с учителем* для является получение такой функции  $h: X \rightarrow Y$ , чтобы  $h(x)$  являлось «хорошим» предсказанием для значения  $y$ .

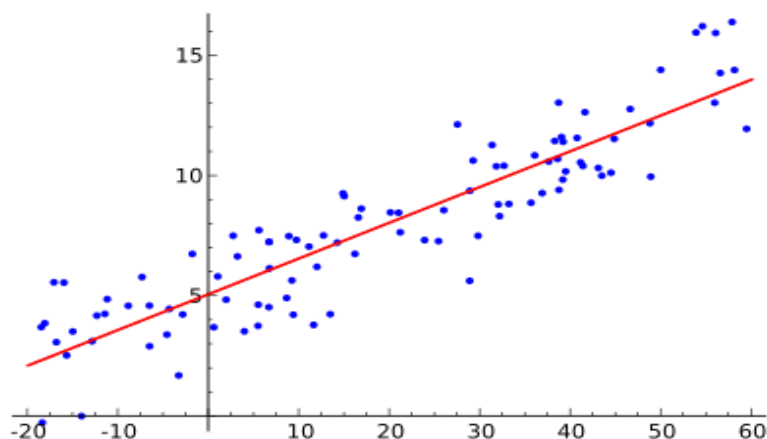
Функцию  $h(x)$  называют *гипотезой*.

В зависимости от типа данных  $y$

Задачи регрессии

$$y \in \mathbb{R}$$

Действительные числа



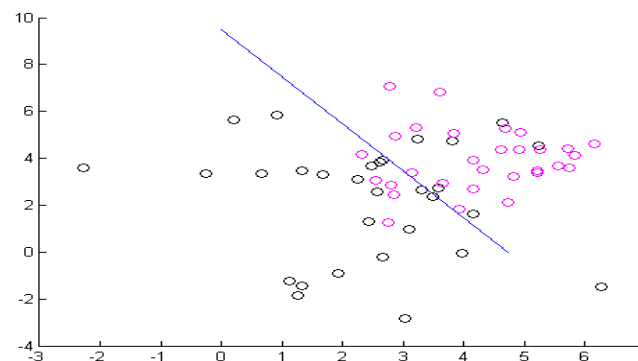
Пример:

- предсказание цен на дома  
( $y$  — цена дома)

Задачи классификации

$$y \in \{y_1, y_2, \dots, y_N\}, N \geq 2$$






Значение из конечного множества  
(идентификаторы классов)








Пример:

- медицинская диагностика  
( $y$  — классы: злокачественная или нет)

**Какие из следующих задач относятся к обучению с учителем, а какие без учителя?**

- С учит.**  имеется база e-mail писем, для которых указано, которые из них являются спамом. Задача научить алгоритм определять спам.
- Без уч.**  имеется база новостных заметок, которые необходимо разделить на группы по близким темам.
- С учит.**  дана база медицинских данных пациентов, больных диабетом. Необходимо научиться определять, имеет ли пациент диабет.
- Без уч.**  дана база данных о потребителях (список товаров, которые потребитель покупал), необходимо автоматически определить наиболее популярные сегменты товаров и их покупателей
- Без уч.**  имеется база данных переходов посетителей веб-сайтов (с одного сайта на другой), необходимо сгруппировать сайты на группы, внутри которых выполняется большая часть переходов.

## К какому классу задач (регрессии или классификации) относятся следующие проблемы?

- Регрес.**  У вас есть склад товаров. Вам нужно решить, сколько товаров будет продано в следующем месяце.
- Классиф.**  Вам нужно написать программу, которая оценит учетные записи почтовых ящиков пользователей и решит, какие из них могут быть взломаны в будущем.
- Классиф.**  Имеются сведения о заемщиках банка. Необходимо определить, кому из заемщиков стоит выдавать кредит, а кому нет.
- Регрес.**  Основываясь на курсе акций различных компаний в прошлом, необходимо предсказать их стоимость на завтра, через неделю, месяц и т.д.
- Классиф.**  Получая изображение с камеры в аудитории, необходимо определить, кто из студентов присутствует на лекции.

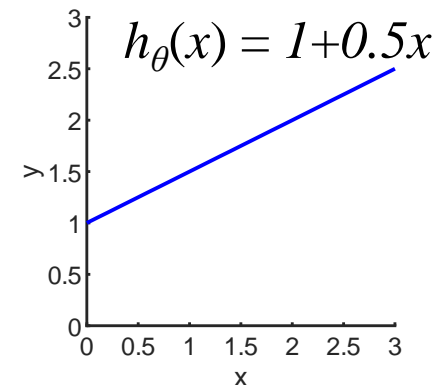
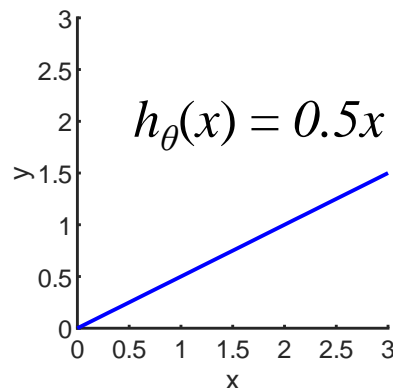
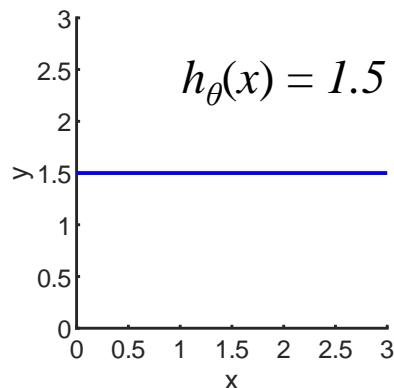


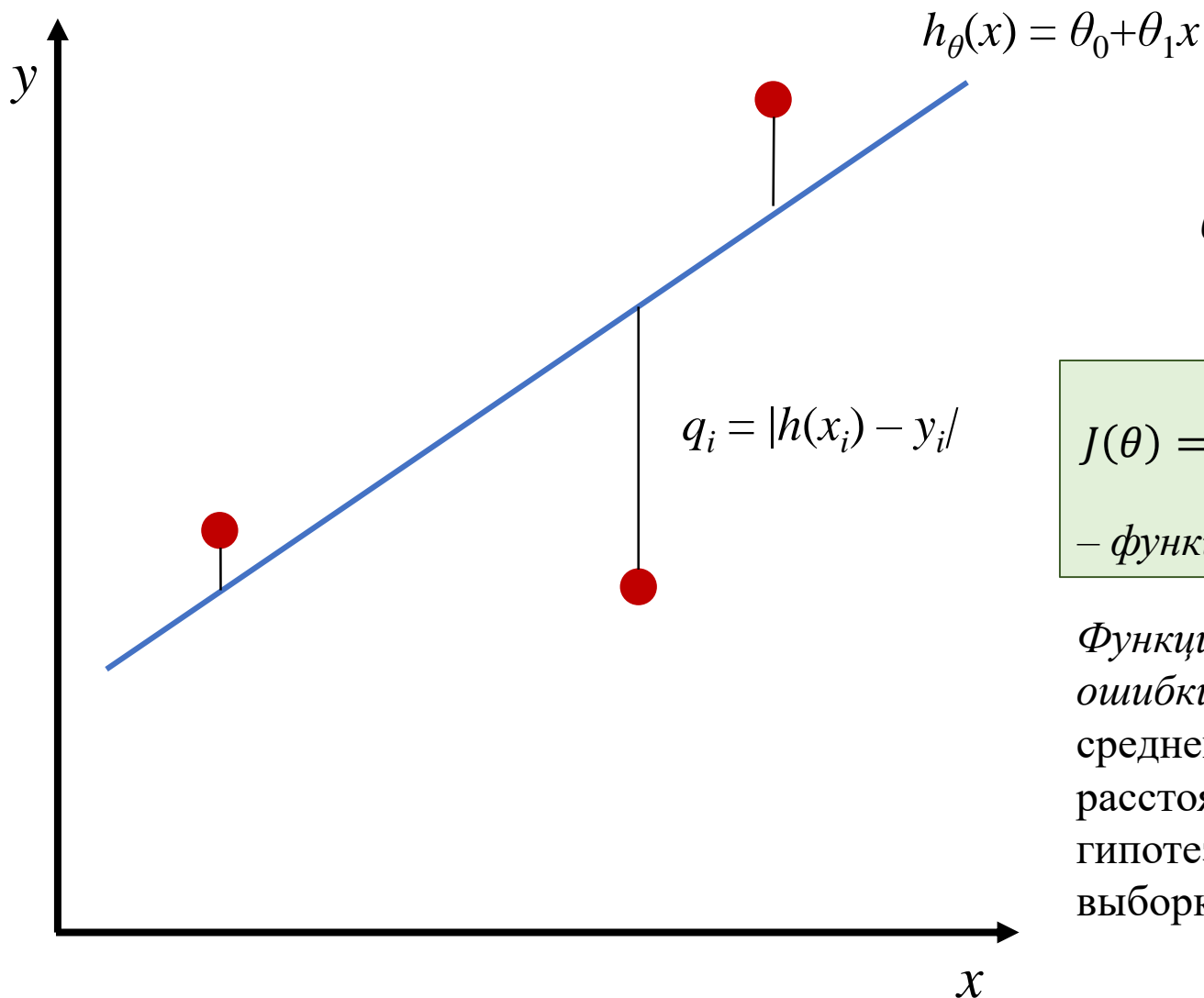
*Задачей обучения с учителем* для является получение такой функции  $h:X \rightarrow Y$ , чтобы  $h(x)$  являлось «хорошим» предсказанием для значения  $y$ .

Что значит быть «хорошим» предсказанием?

Площадь, м <sup>2</sup>	Цена, тыс. м.к.
2104	460
1416	232
1534	315
852	178
...	...

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$



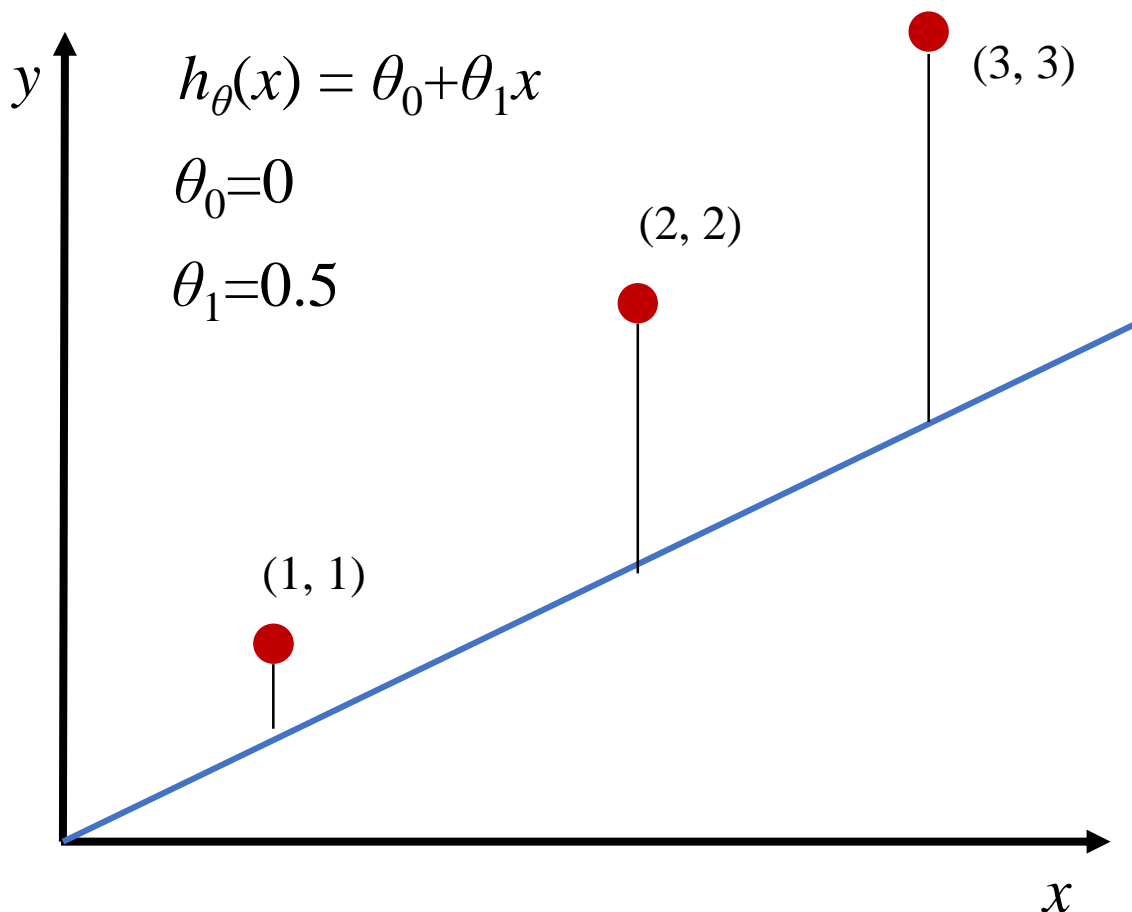


$$Q = \sum_{i=1}^m |h(x_i) - y_i|$$

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x_i) - y_i)^2$$

– функция стоимости

Функция квадратичной ошибки, равная половине от среднего значения квадрата расстояния между значениями гипотезы и обучающей выборки



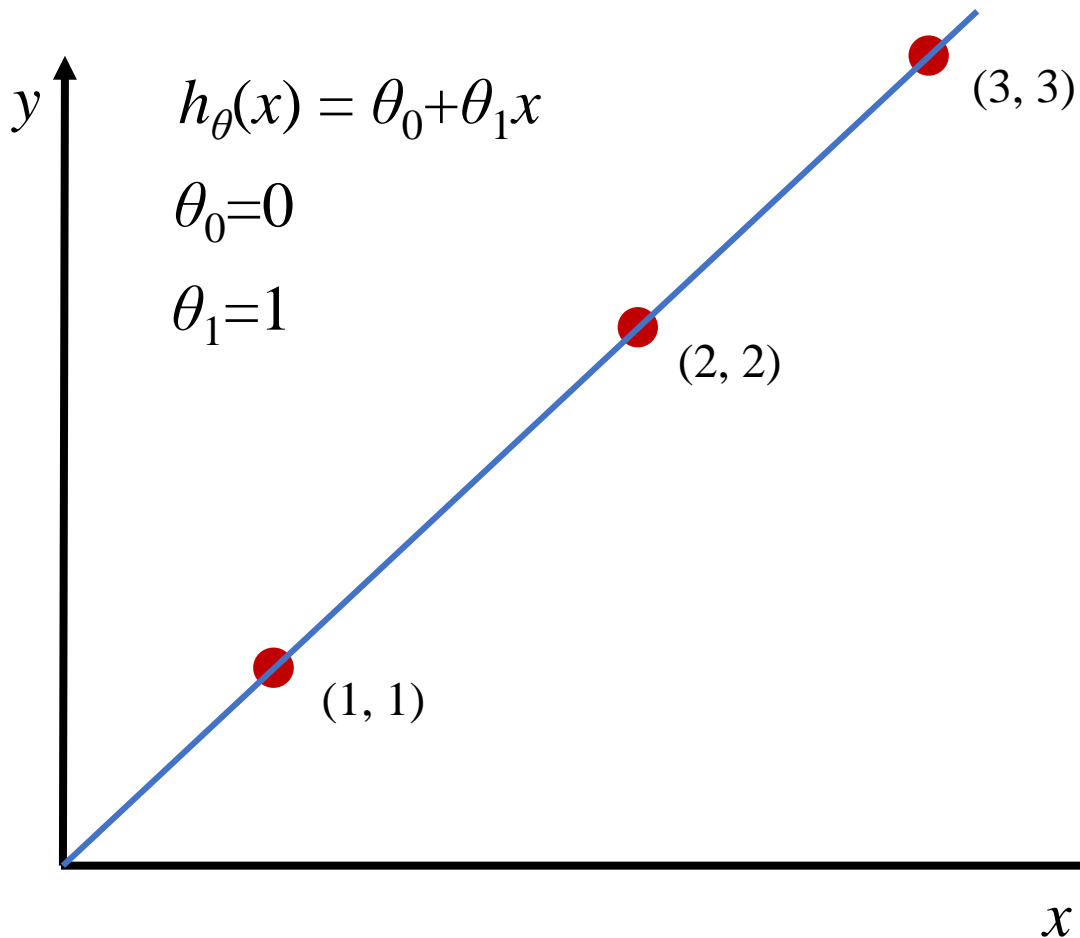
$$h_{\theta}(1) = 0.5$$

$$h_{\theta}(2) = 1$$

$$h_{\theta}(3) = 1.5$$

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x_i) - y_i)^2$$

$$J(\theta_0, \theta_1) = \frac{1}{2 \cdot 3} [(0.5 - 1)^2 + (1 - 2)^2 + (1.5 - 3)^2] \approx 0.58$$



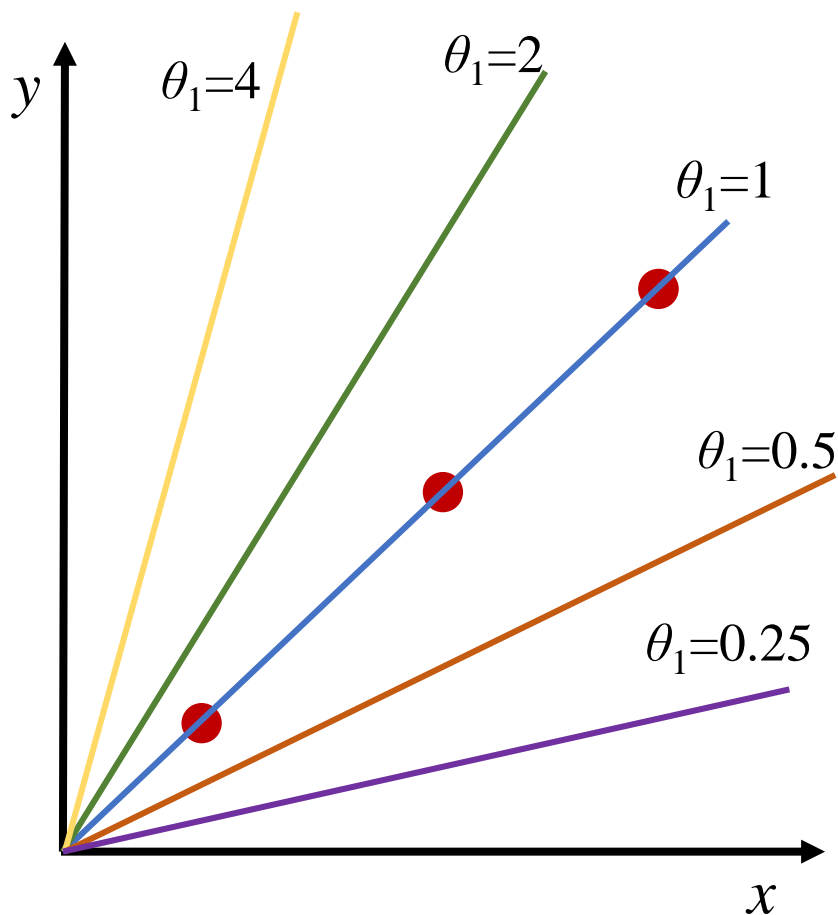
$$h_{\theta}(1) = 1$$

$$h_{\theta}(2) = 2$$

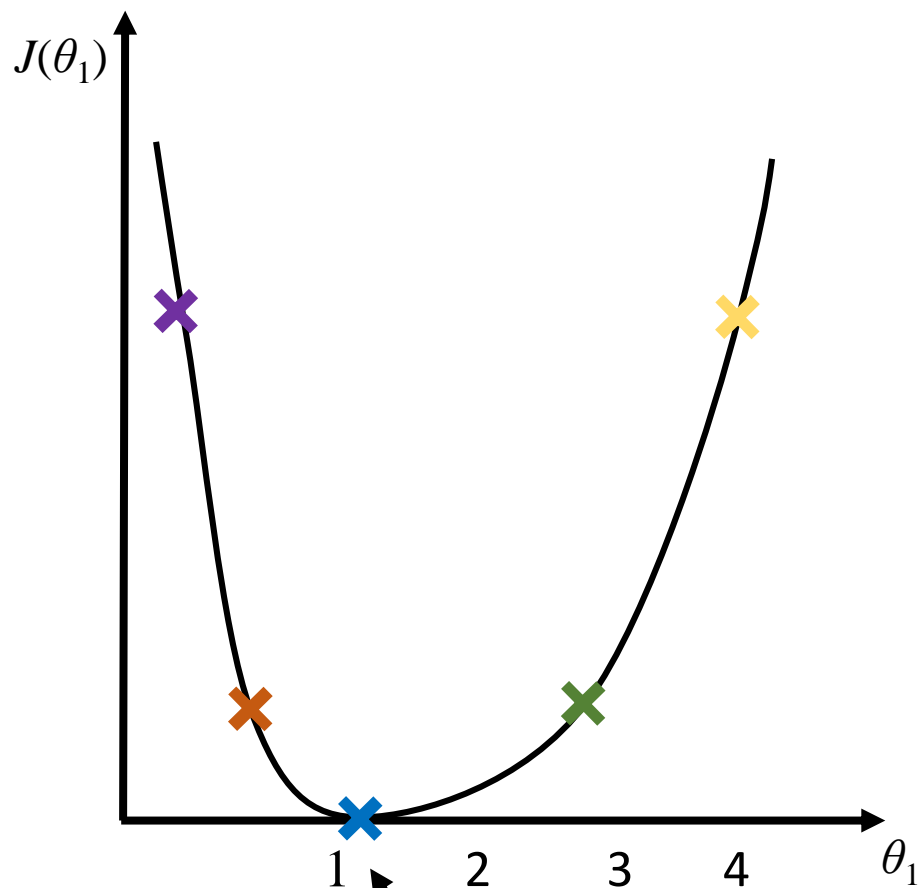
$$h_{\theta}(3) = 3$$

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x_i) - y_i)^2$$

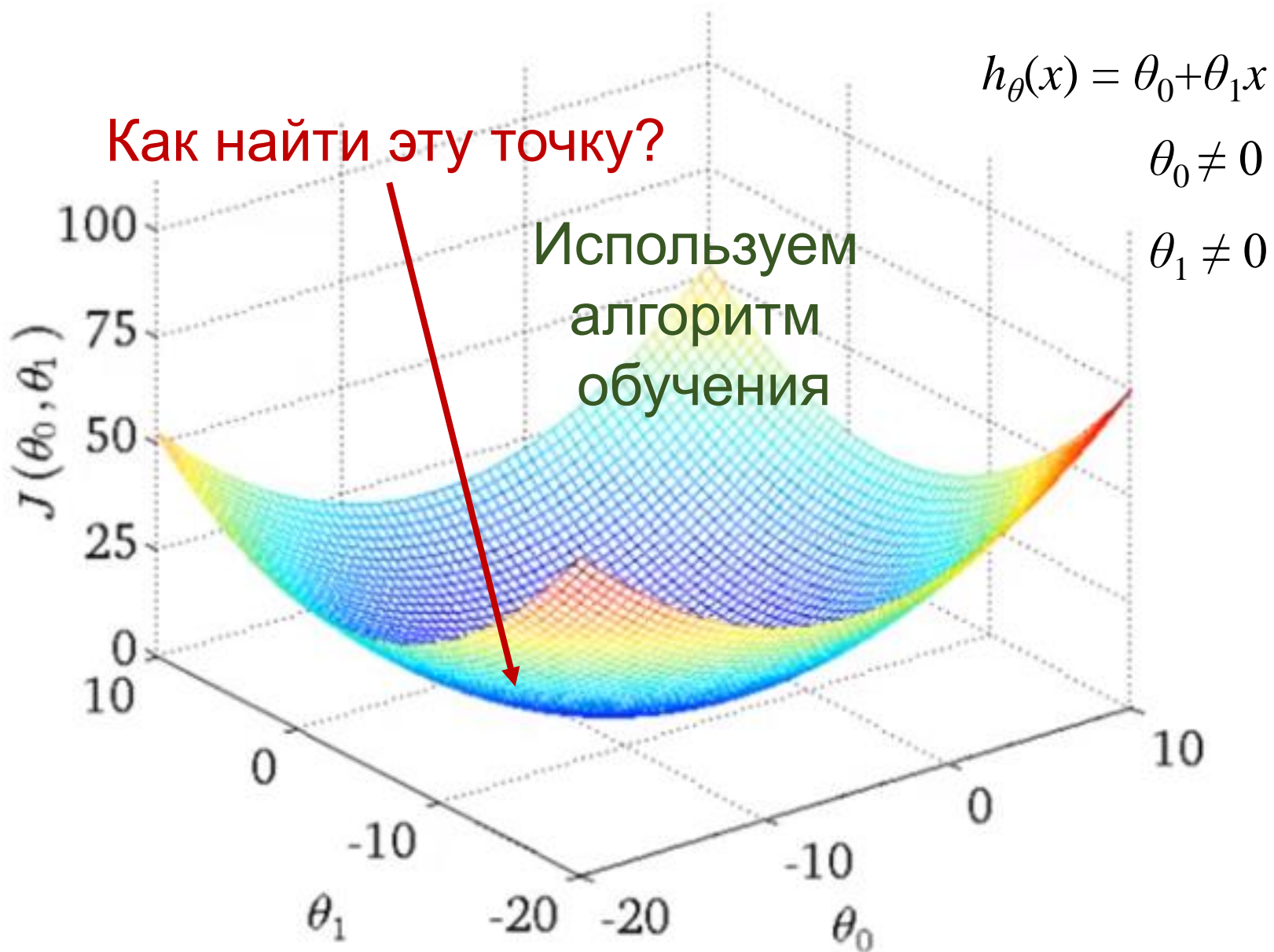
$$J(\theta_0, \theta_1) = \frac{1}{2 \cdot 3} [(1 - 1)^2 + (2 - 2)^2 + (3 - 3)^2] = 0$$

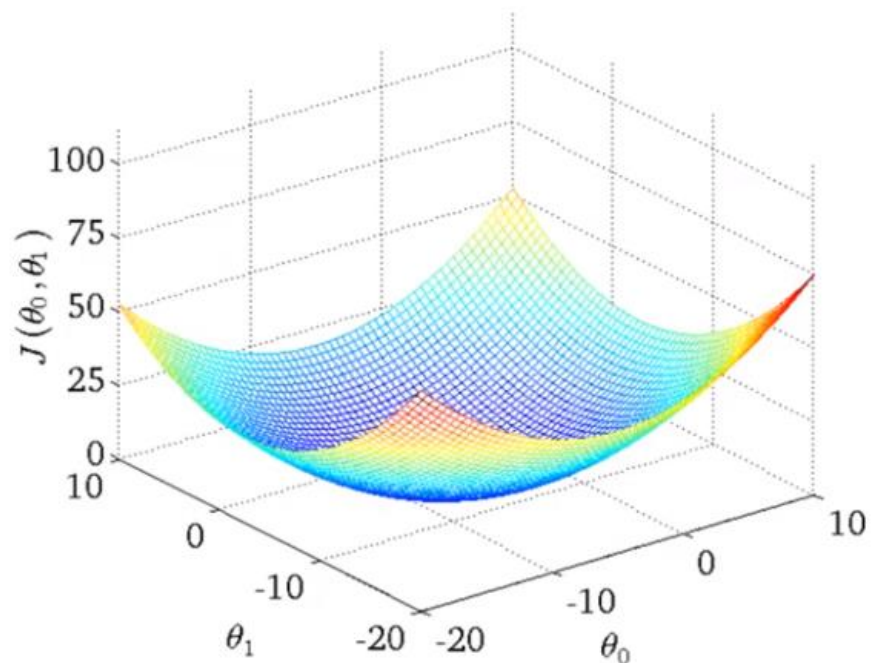


$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x_i) - y_i)^2$$



глобальный минимум



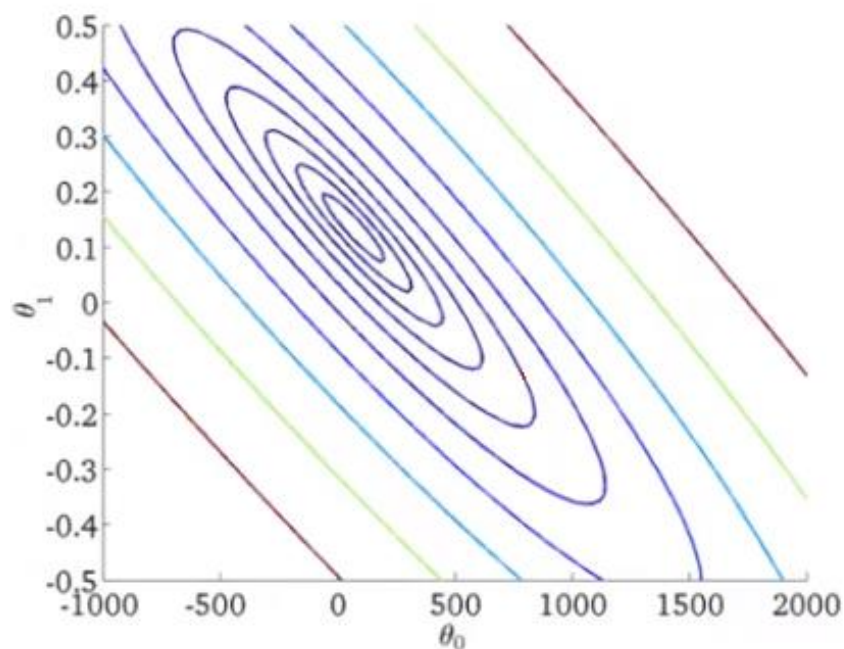


3D график

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x_i) - y_i)^2$$

Эквивалентный  
контурный график  
(линии уровня)

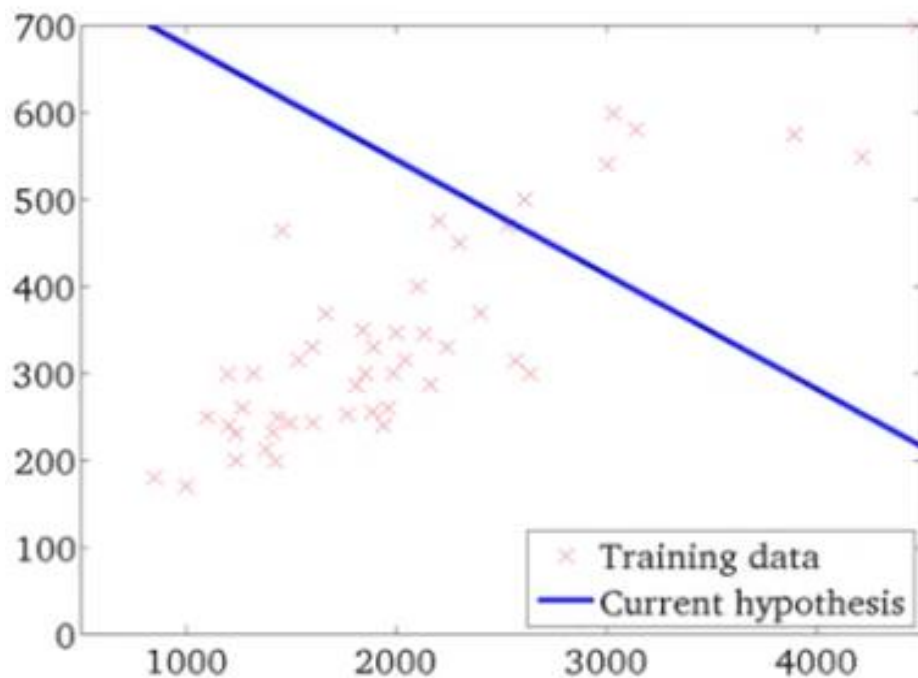
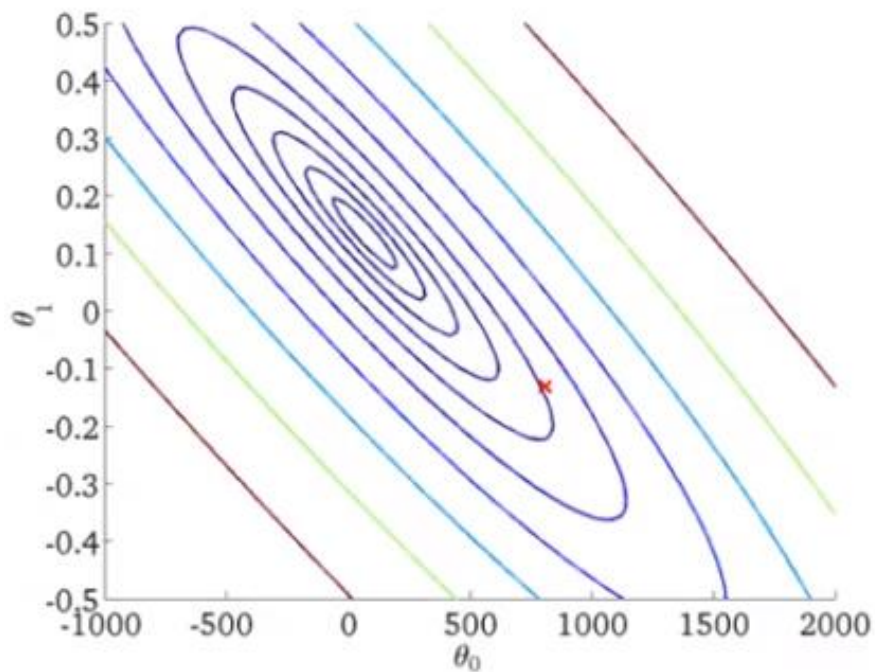


$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x_i) - y_i)^2$$

$$\theta_0 = 800$$

$$\theta_1 = -0.15$$



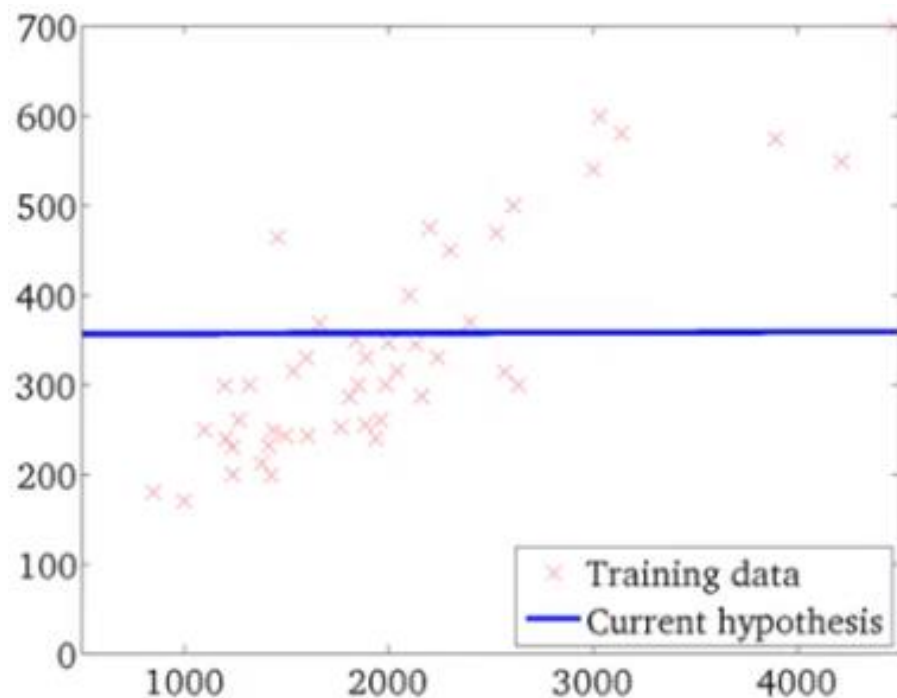
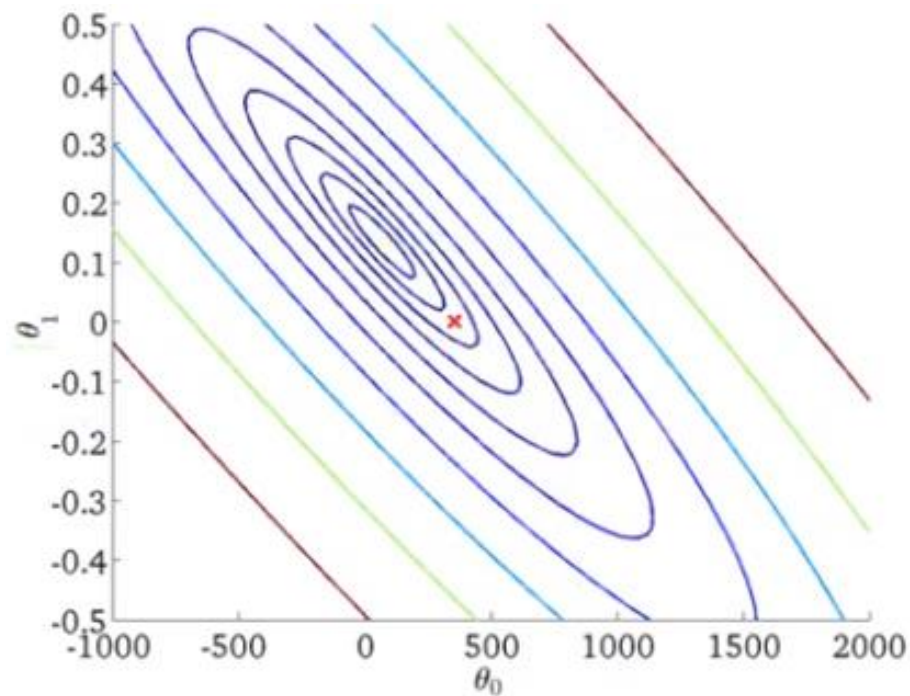


$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x_i) - y_i)^2$$

$$\theta_0 = 360$$

$$\theta_1 = 0$$

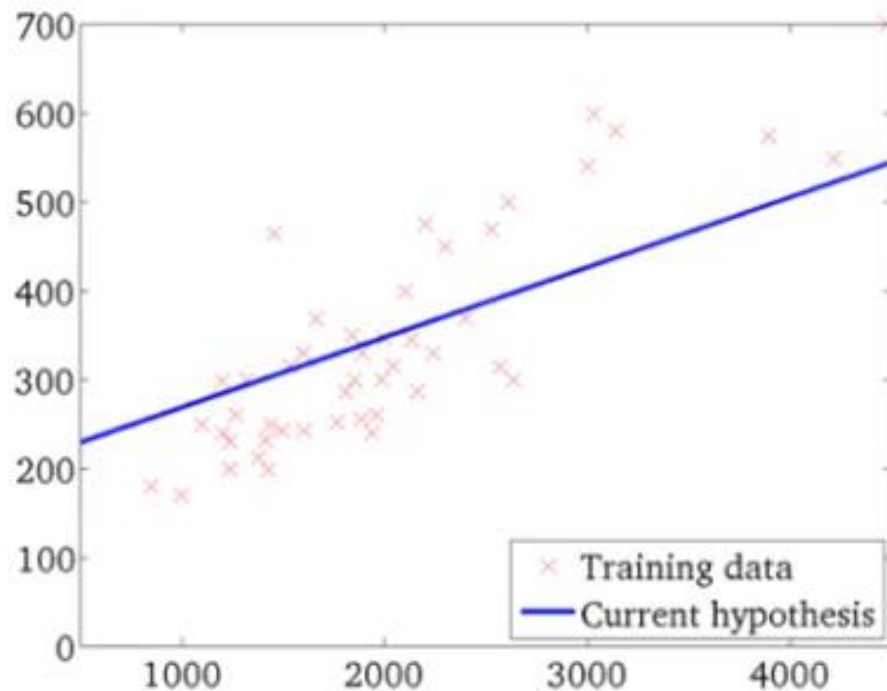
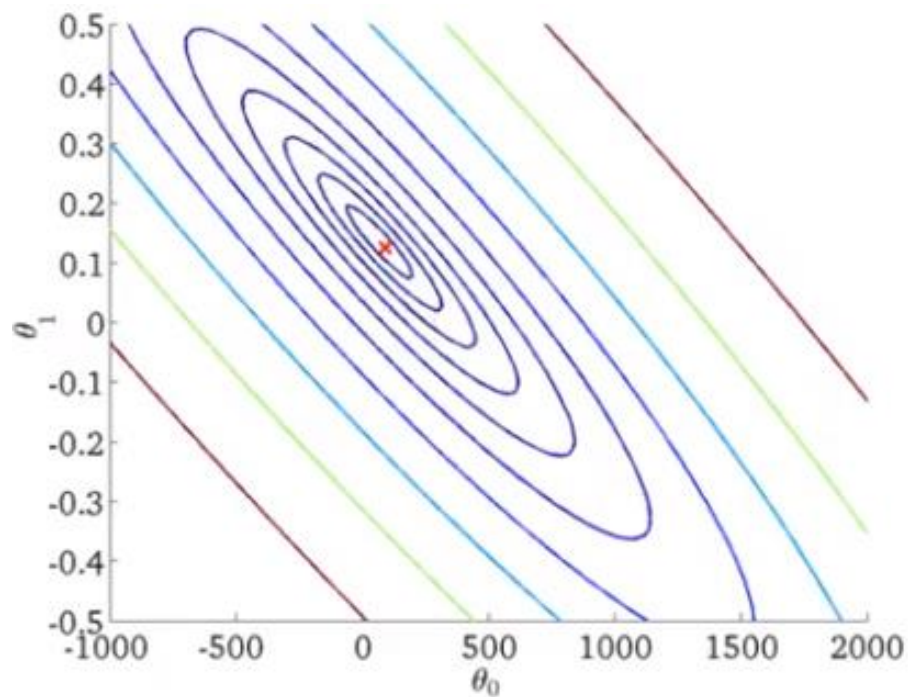


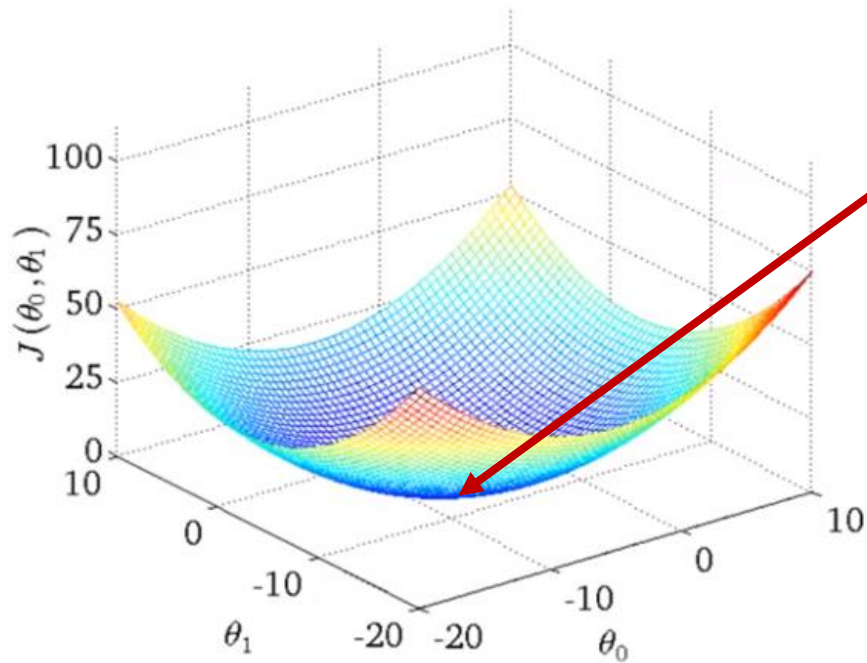
$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m ((h_{\theta}(x_i) - y_i)^2)$$

$$\theta_0 = 250$$

$$\theta_1 = 0.12$$





Как найти эту точку?

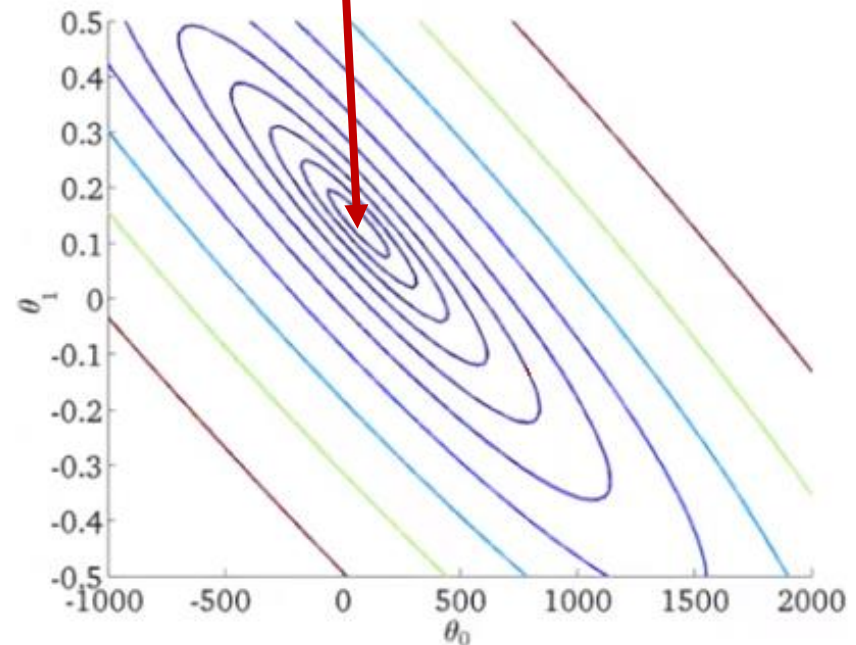
Используем  
алгоритм  
обучения

Самый простой алгоритм обучения

Алгоритм градиентного спуска  
(Gradient descent)

$$\min_{\theta_0, \dots, \theta_n} J(\theta_0, \theta_1, \dots, \theta_n)$$

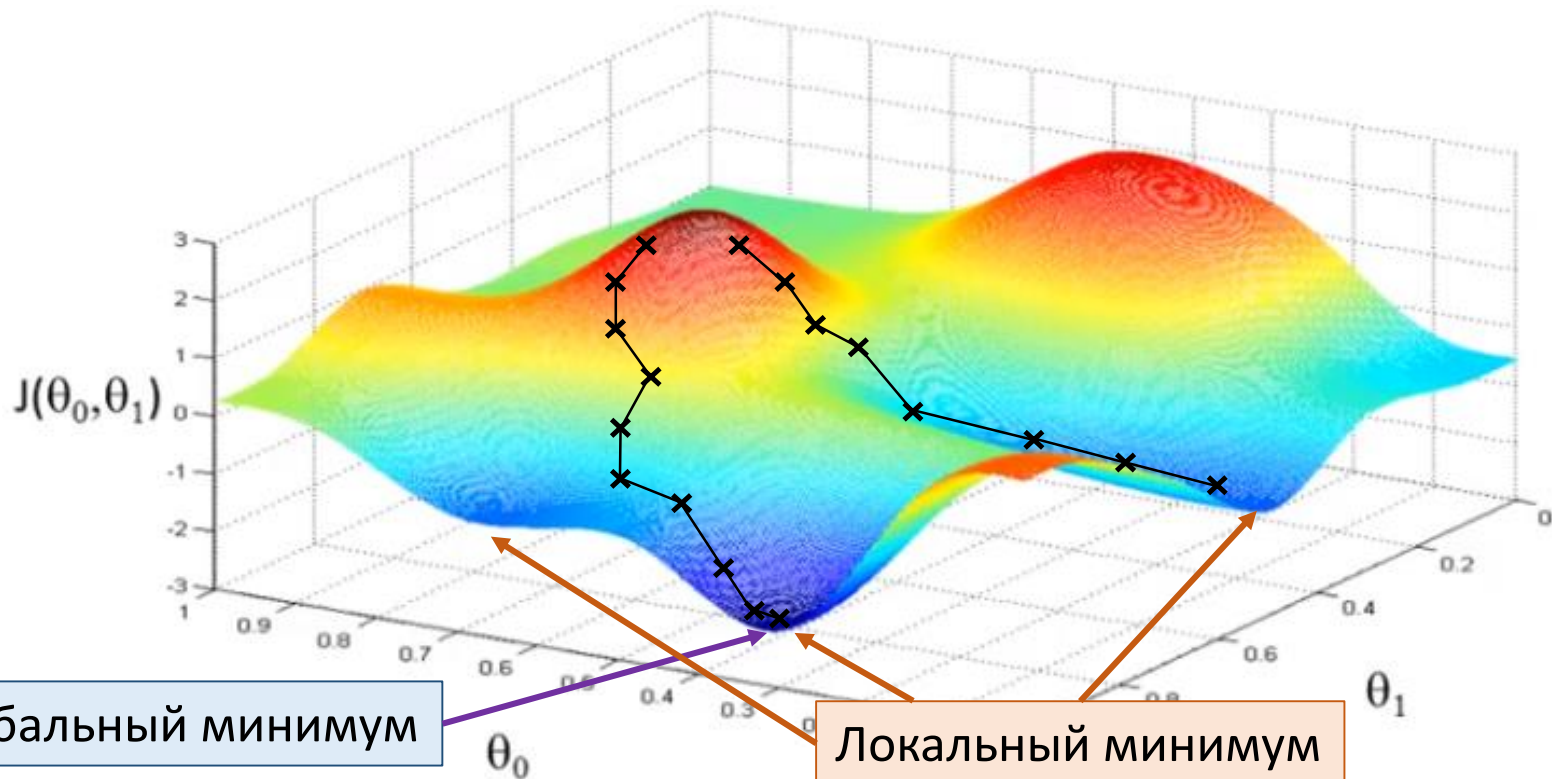
(математическая задача оптимизации)



## Общий принцип алгоритма:

1. Выбираем начальные значения  $\theta_0, \theta_1$
2. Изменим  $\theta_0, \theta_1$  на некоторое значение так, чтобы  $J(\theta_0^1, \theta_1^1)$  уменьшилась
3. Повторяем шаг 2 до тех пор, пока уменьшение  $J(\theta_0, \theta_1)$  не станет достаточно малым

Подобно катящемуся с горы мячику



**Алгоритм градиентного спуска.**

**Вход:**  $J(\theta_0, \dots, \theta_n)$  – функция,  $\theta_0^0, \dots, \theta_n^0$  – начальные значения переменных,  $\alpha$  – темп обучения,  $\varepsilon$  – критерий остановки.

**Выход:**  $\theta_0^*, \dots, \theta_n^*$  – найденные значения для локального минимума.

**Действия:**

для  $j = 0, \dots, n$  :

$$\theta_j := \theta_j^0$$

повторять {

для  $j = 0, \dots, n$ :

$$d\theta_j := \frac{\partial}{\partial \theta_j} J(\theta_0, \dots, \theta_n)$$

для  $j = 0, \dots, n$ :

$$\theta_j := \theta_j - \alpha d\theta_j$$

} пока  $d\theta_j > \varepsilon, j = 0, \dots, n$

для  $j = 0, \dots, n$  :

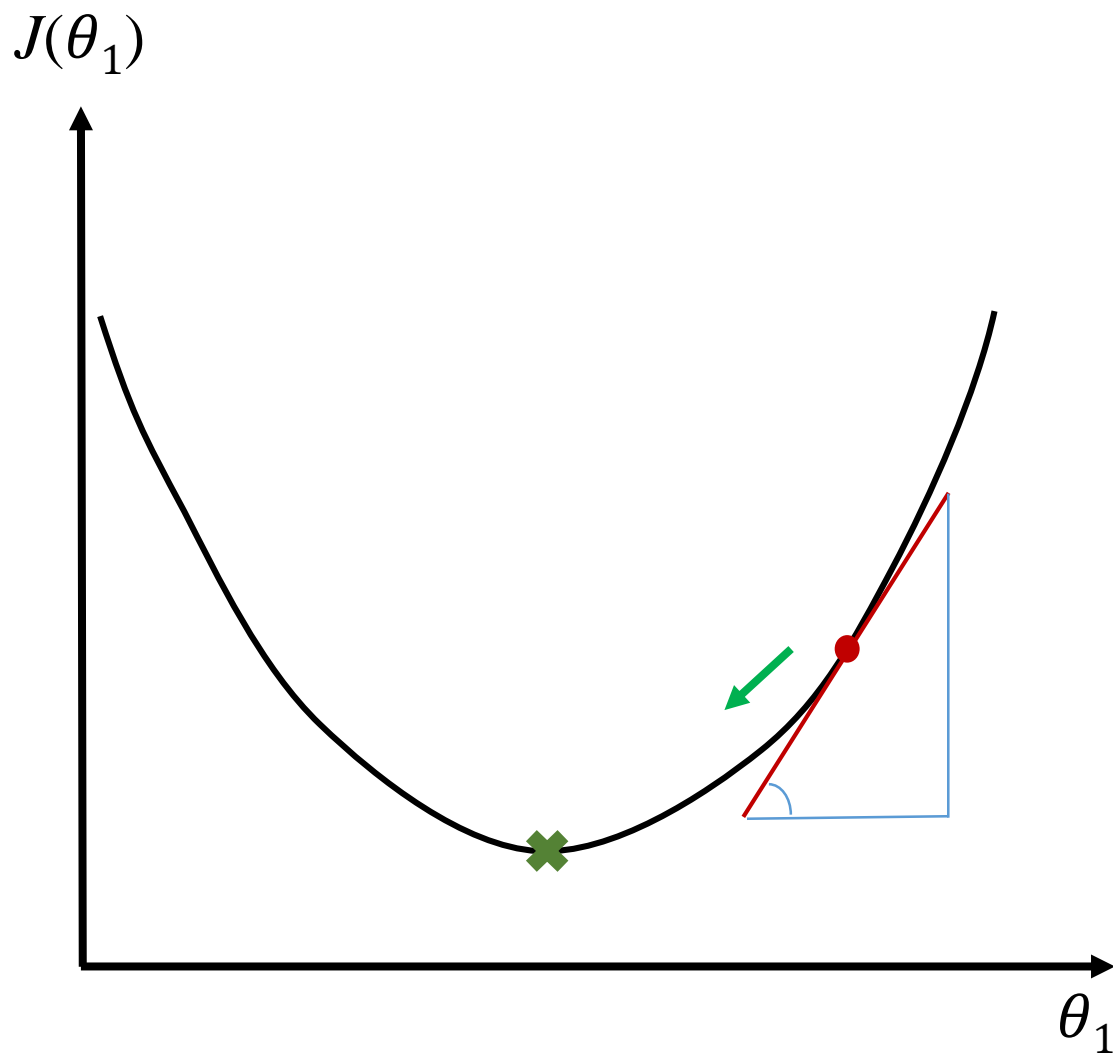
$$\theta_j^* := \theta_j$$

**Конец алгоритма.**

Частная производная

Градиент

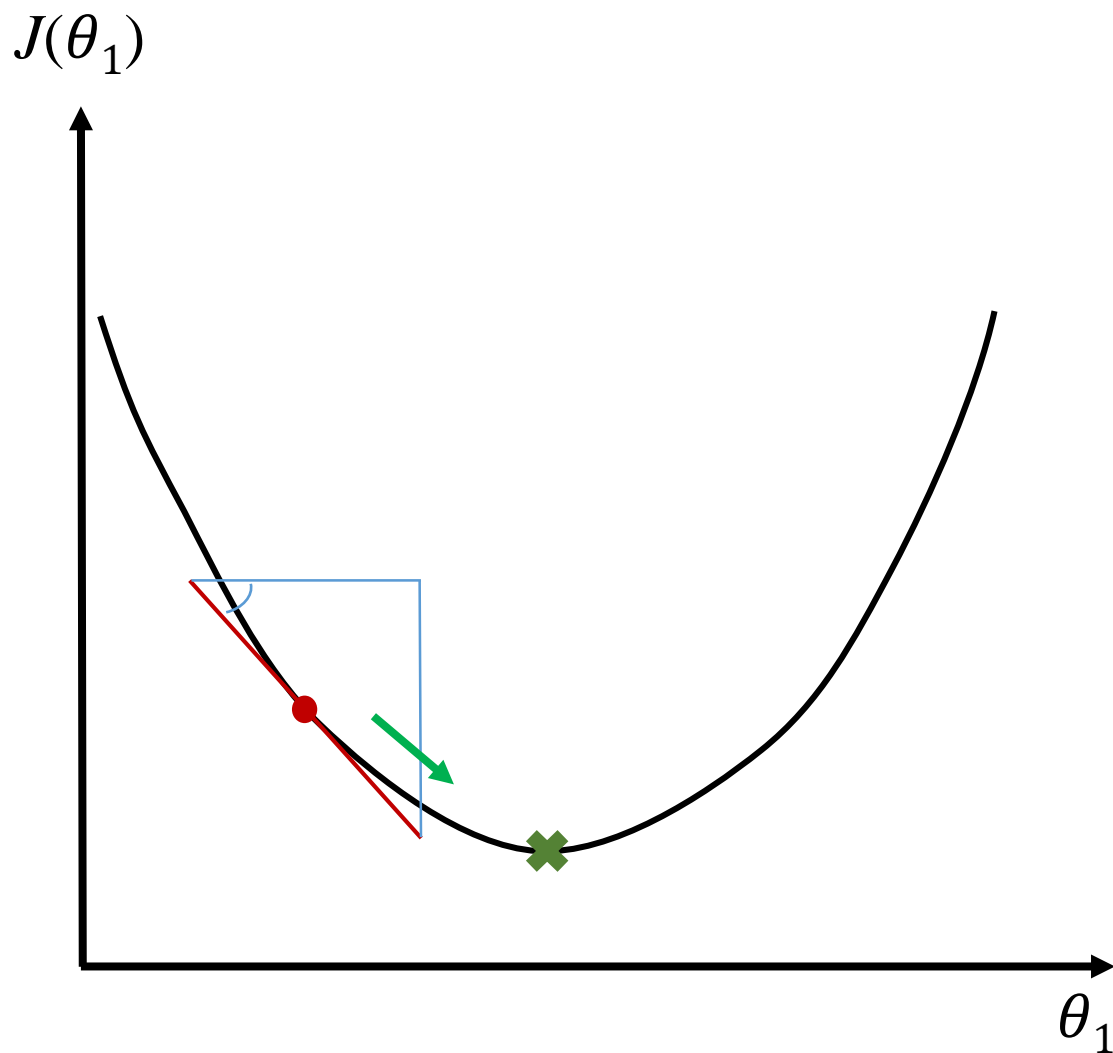
$$\nabla J(\theta) = \left[ \frac{\partial}{\partial \theta_0} J(\theta), \dots, \frac{\partial}{\partial \theta_n} J(\theta) \right]$$



$$\theta_1 = \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_1)$$

$$\frac{\partial}{\partial \theta_1} J(\theta_1) > 0$$

Точка сдвинется влево



$$\theta_1 = \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_1)$$

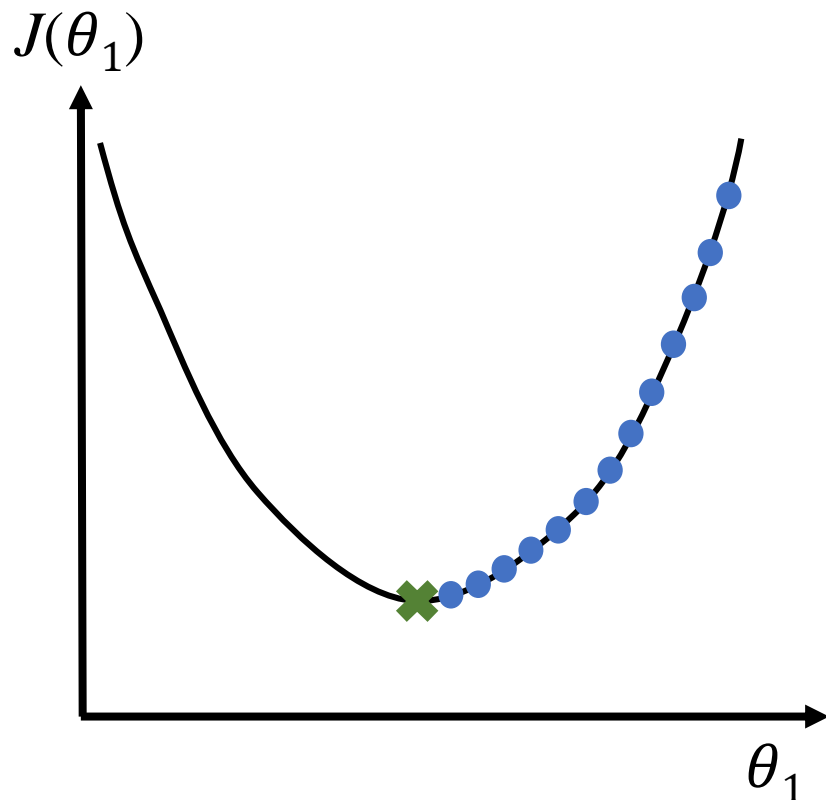
$$\frac{\partial}{\partial \theta_1} J(\theta_1) < 0$$

Точка сдвинется вправо

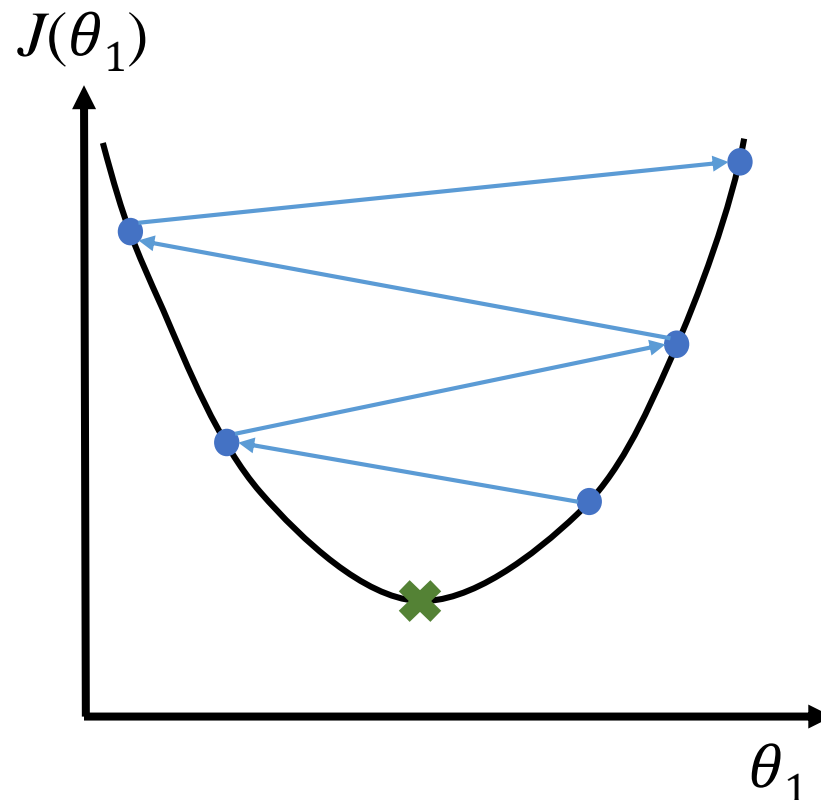
$$\theta_1 = \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_1)$$

$\alpha$  – темп обучения

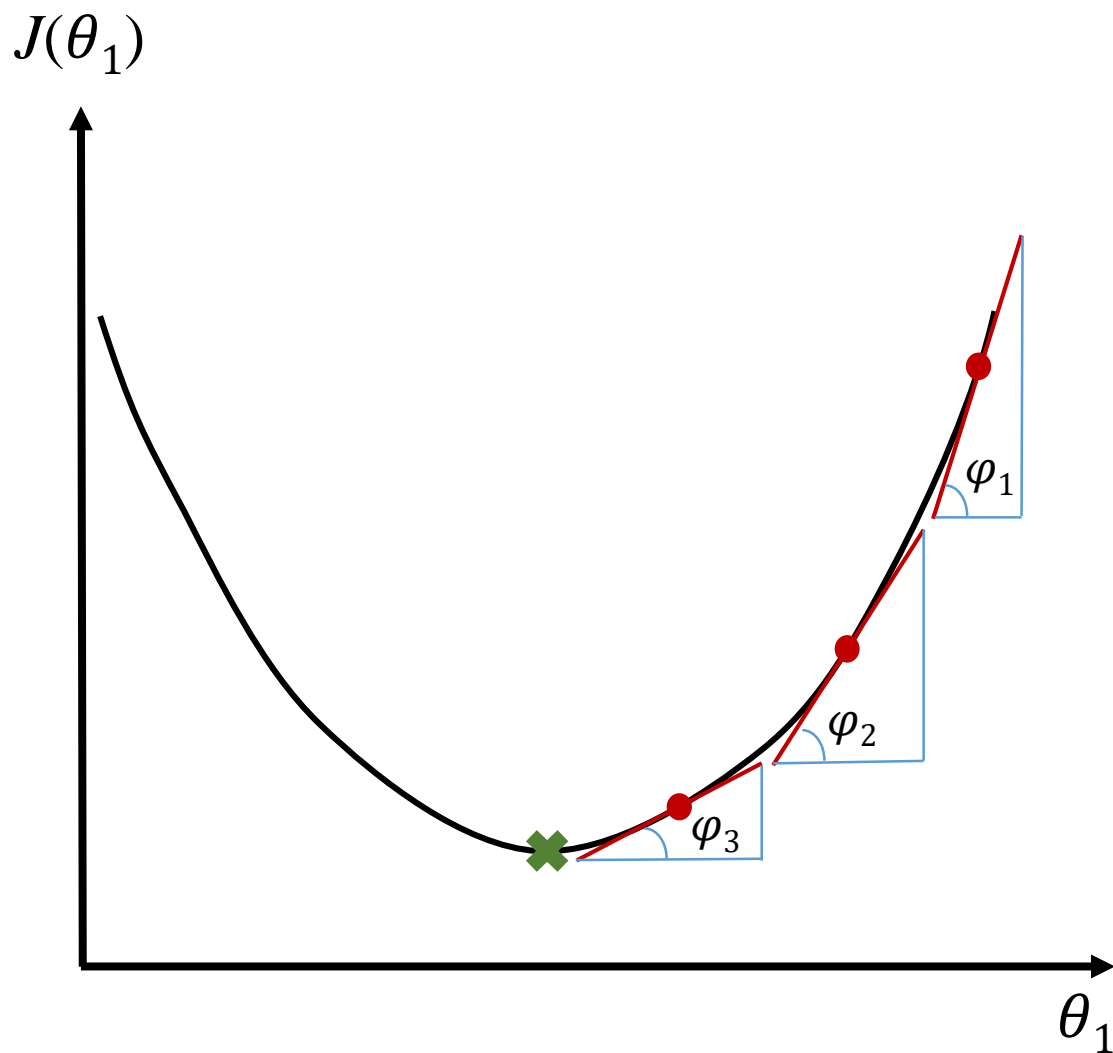
Малое значение  $\alpha$  – низкая скорость сходимости (дольше работает)



Большое значение  $\alpha$  – алгоритм «проскочит» минимум (вообще не сойдется)







$$\theta_1 = \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_1)$$

$$\tan \varphi_1 > \tan \varphi_2 > \tan \varphi_3$$

При приближении к точке минимума, алгоритм сам снижает величину шага. Изменять  $\alpha$  в процессе работы не нужно!

Задача предсказания цен на недвижимость

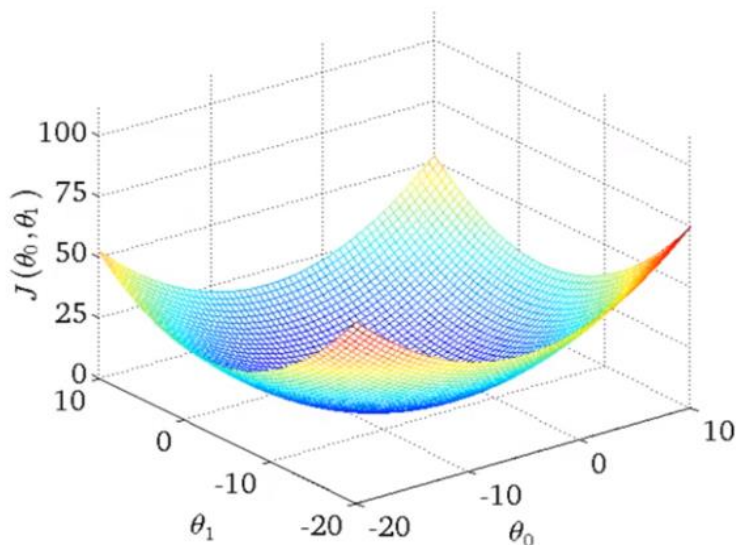
Площадь, м <sup>2</sup>	Цена, тыс. м.к.
2104	460
1416	232
1534	315
852	178
...	...

Гипотеза

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

Функция стоимости

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x_i) - y_i)^2$$



Используем алгоритм  
градиентного спуска, чтобы  
найти параметры  
 $\theta_0, \theta_1$

повторять {

для  $j = 0, \dots, n$ :

$$d\theta_j := \frac{\partial}{\partial \theta_j} J(\theta_0, \dots, \theta_n)$$

для  $j = 0, \dots, n$ :

$$\theta_j := \theta_j - \alpha d\theta_j$$

} пока  $d\theta_j > \varepsilon, j = 0, \dots, n$

Вычислить частные  
производные по  $\theta_0, \theta_1$  для:

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

$$\begin{aligned} \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) &= \\ &= \frac{\partial}{\partial \theta_j} \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 = \\ &= \frac{\partial}{\partial \theta_j} \frac{1}{2m} \sum_{i=1}^m (\theta_0 + \theta_1 x^{(i)} - y^{(i)})^2, \\ j &= 0, 1 \end{aligned}$$

$$\frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^m (\theta_0 + \theta_1 x^{(i)} - y^{(i)})$$

$$\frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^m (\theta_0 + \theta_1 x^{(i)} - y^{(i)}) x^{(i)}$$

**Алгоритм градиентного спуска для линейной регрессии.**

**Вход:**  $J(\theta_0, \theta_1)$  – функция,  $\theta_0^0, \theta_1^0$  – начальные значения переменных,  $\alpha$  – темп обучения,  $\varepsilon$  – критерий остановки.

**Выход:**  $\theta_0^*, \theta_1^*$  – найденные значения для локального минимума.

**Действия:**

$$\theta_0 := \theta_0^0$$

$$\theta_1 := \theta_1^0$$

повторять {

$$d\theta_0 := \frac{1}{m} \sum_{i=1}^m (\theta_0 + \theta_1 x^{(i)} - y^{(i)})$$

$$d\theta_1 := \frac{1}{m} \sum_{i=1}^m (\theta_0 + \theta_1 x^{(i)} - y^{(i)}) x^{(i)}$$

$$\theta_0 := \theta_0 - \alpha d\theta_0$$

$$\theta_1 := \theta_1 - \alpha d\theta_1$$

} пока  $d\theta_0 > \varepsilon, d\theta_1 > \varepsilon$

$$\theta_0^* := \theta_0$$

$$\theta_1^* := \theta_1$$

**Конец алгоритма.**

Сначала вычисляем  
градиент,  
потом делаем шаг

Площадь, м <sup>2</sup>	Цена, тыс. м.к.
2104	460
1416	232
1534	315
852	178
...	...

$$h(x) = \theta_0 + \theta_1 x$$

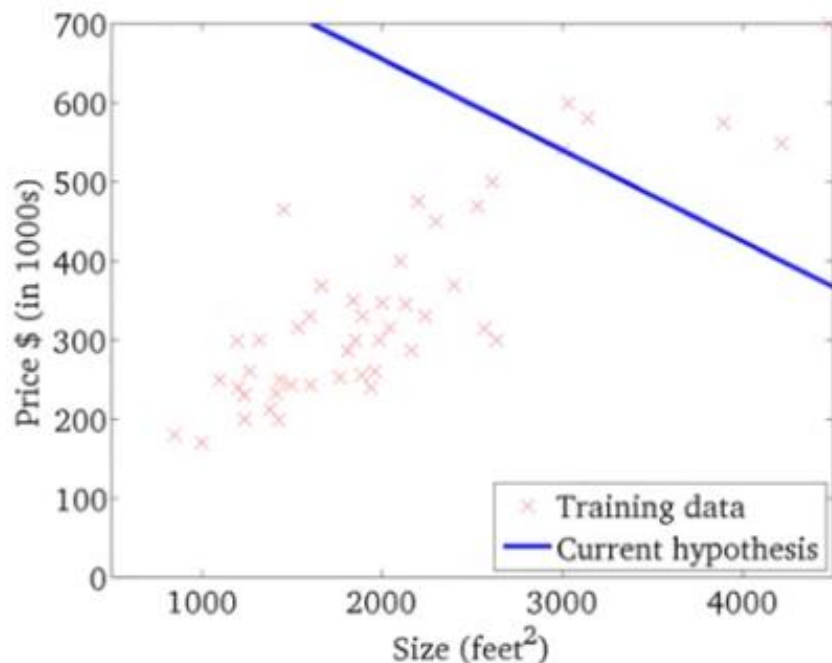
Начальная точка:

$$\theta_0 = 900$$

$$\theta_1 = -0.1$$

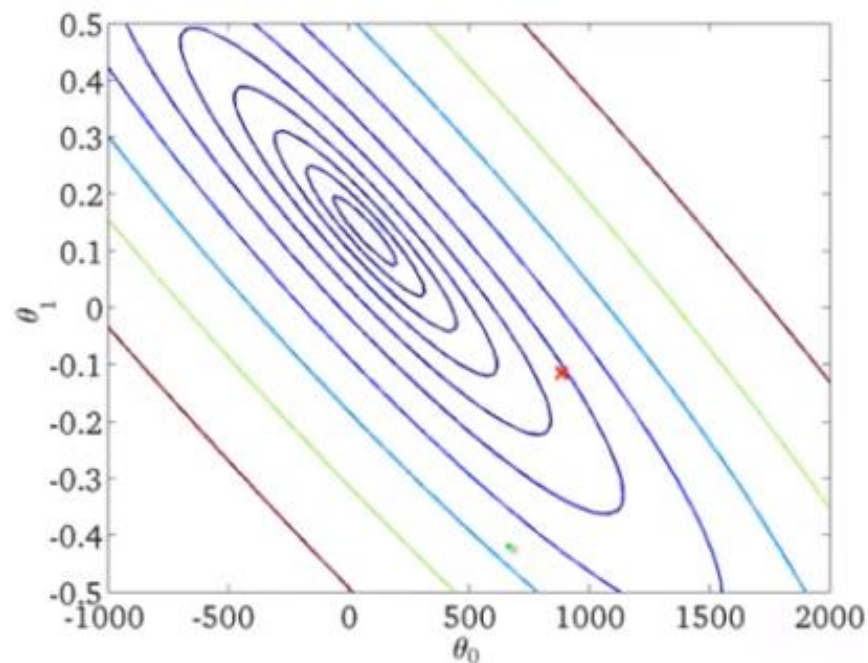
$$h_{\theta}(x)$$

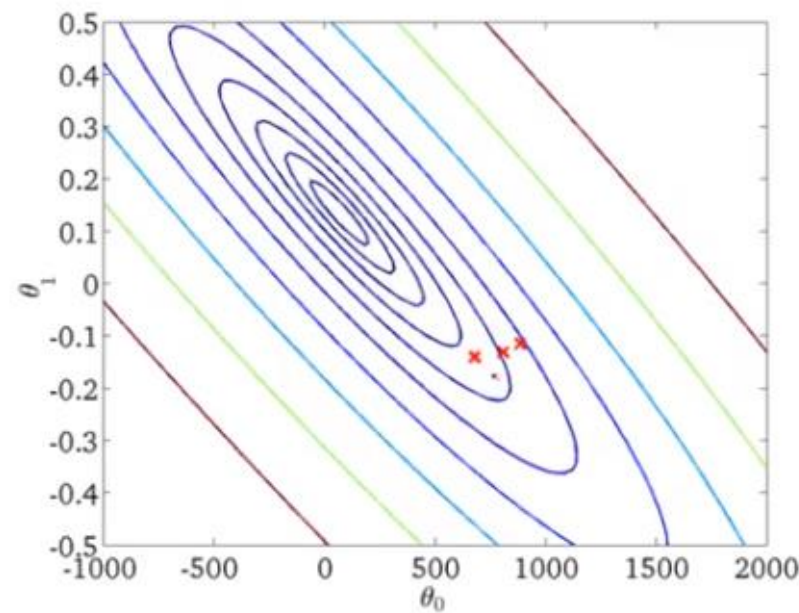
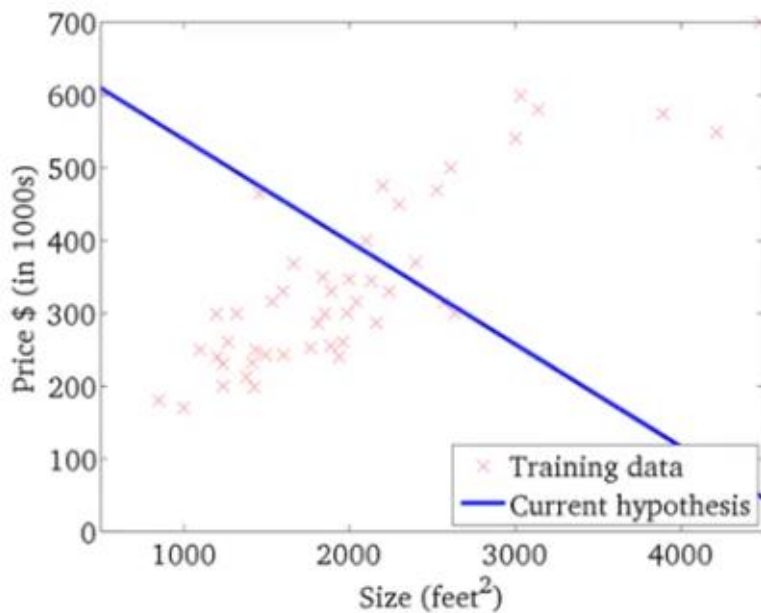
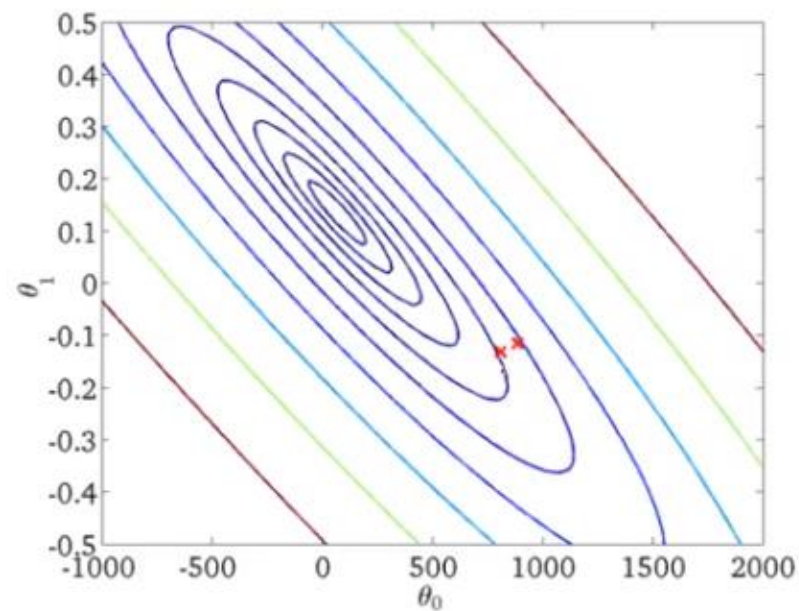
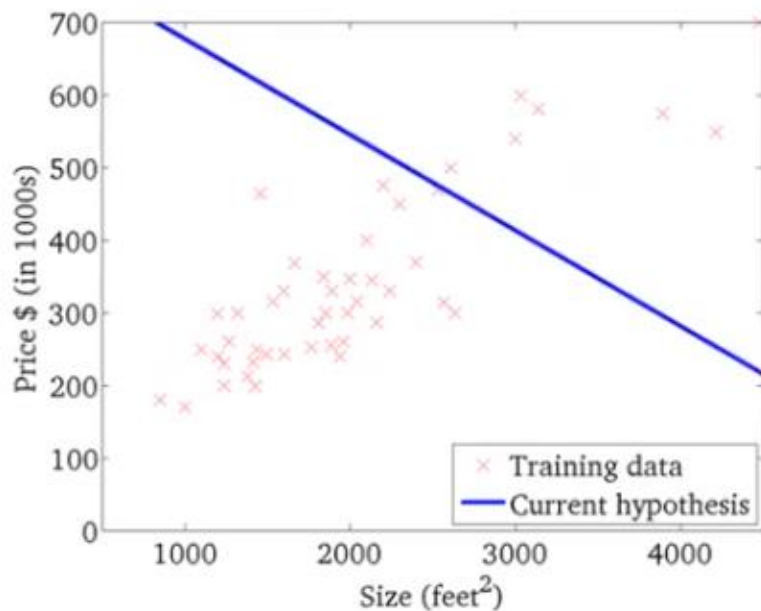
(for fixed  $\theta_0, \theta_1$ , this is a function of  $x$ )



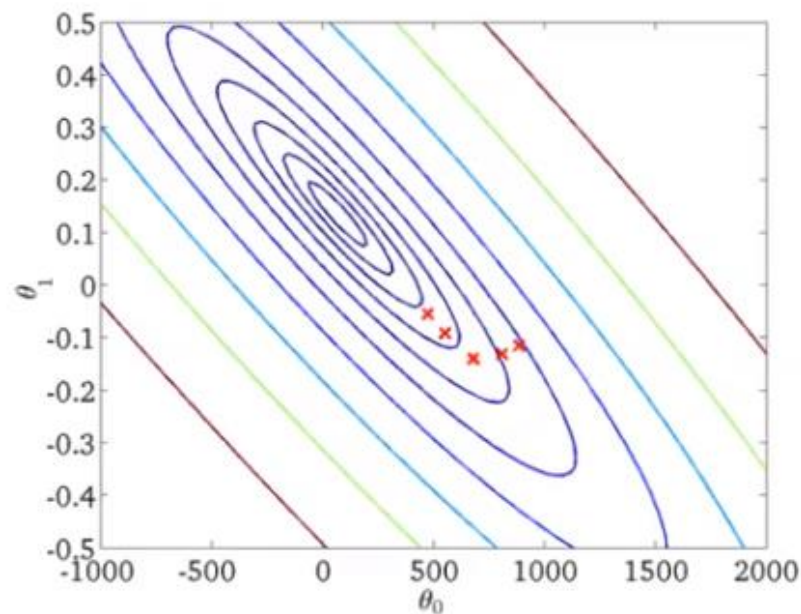
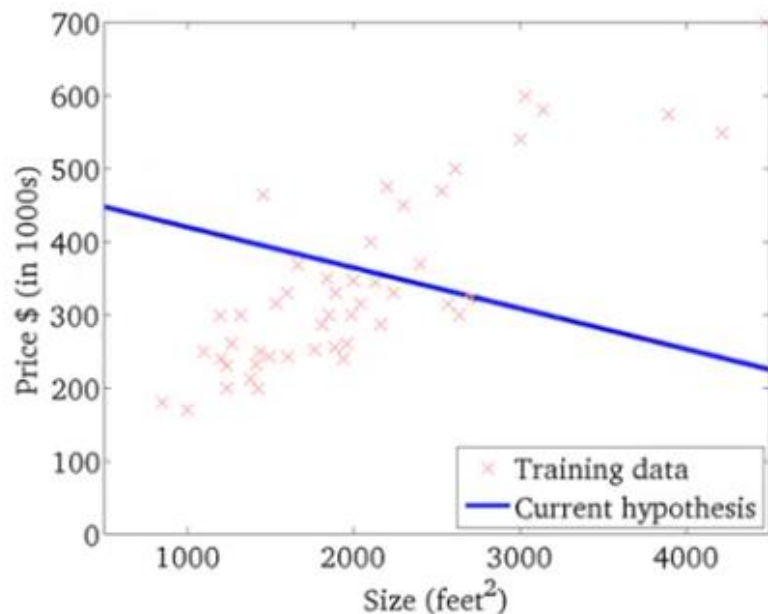
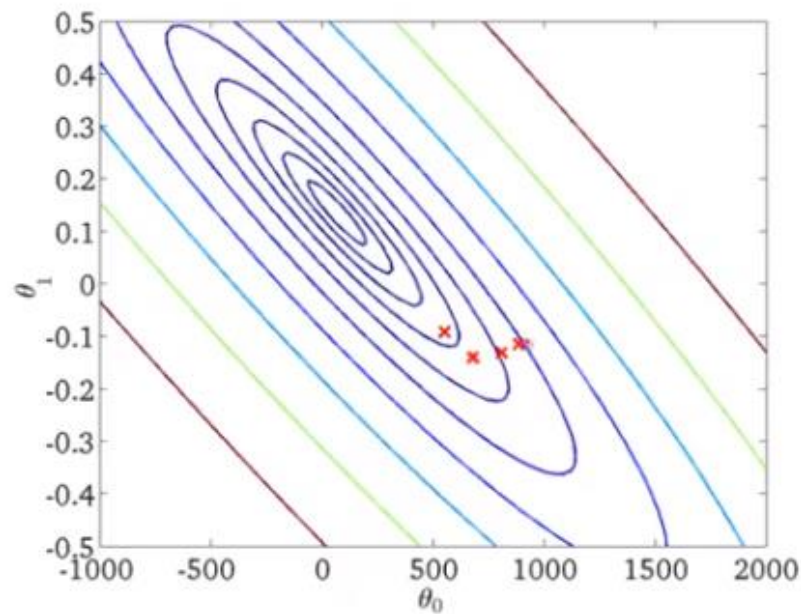
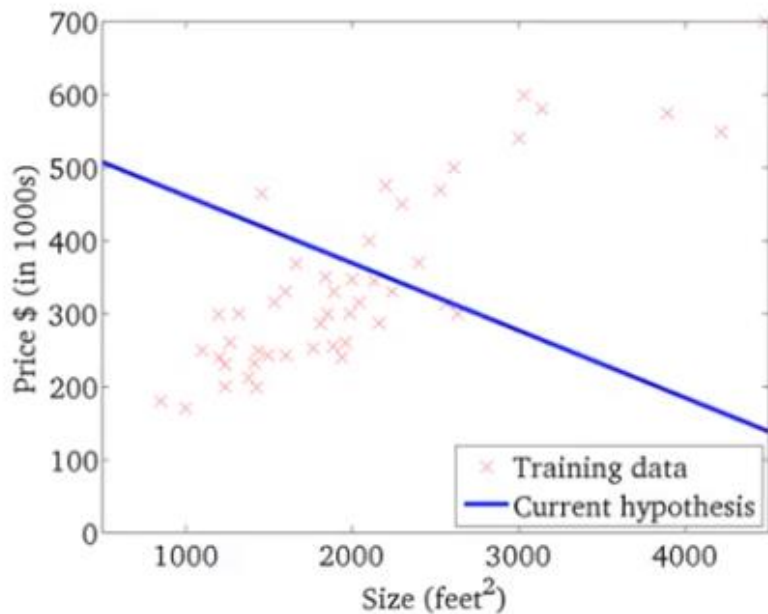
$$J(\theta_0, \theta_1)$$

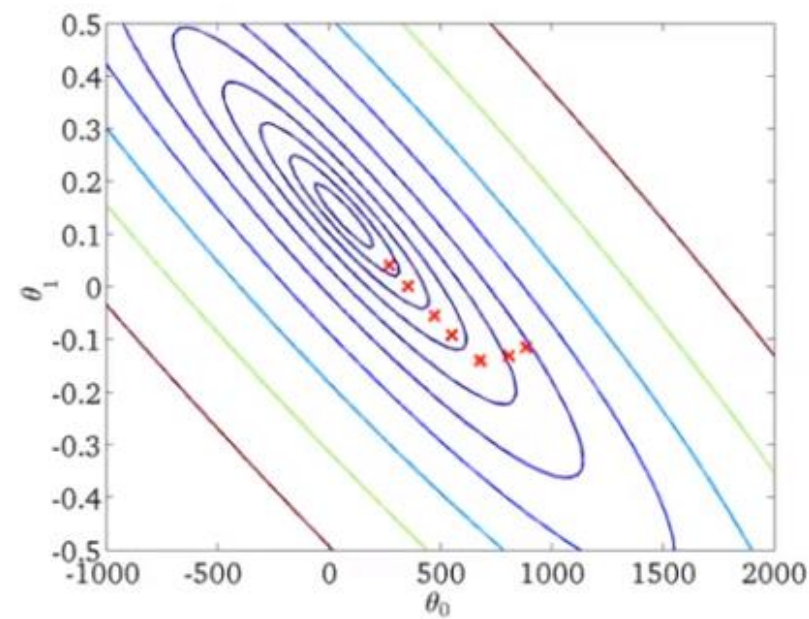
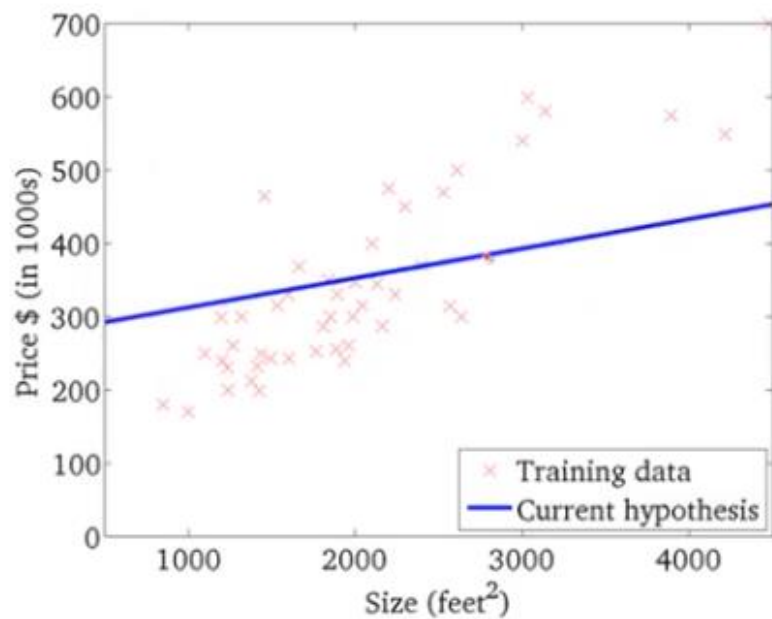
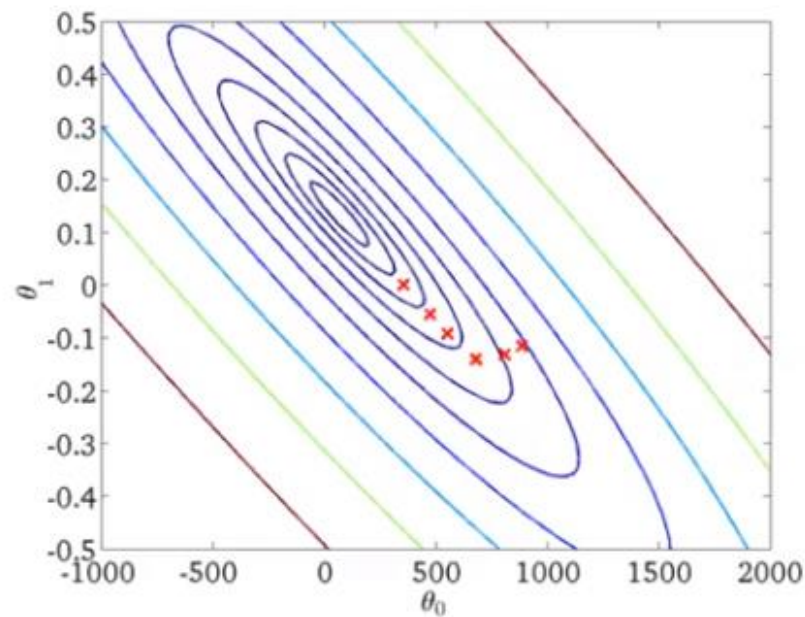
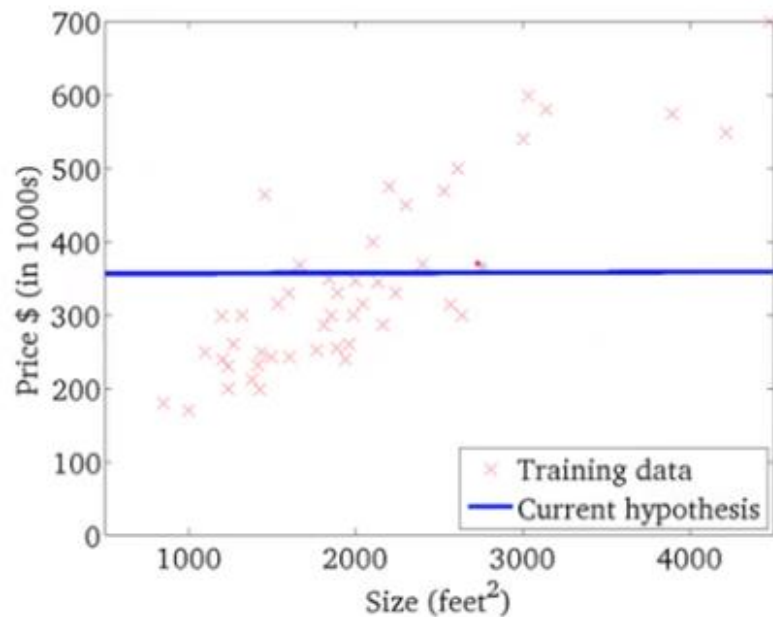
(function of the parameters  $\theta_0, \theta_1$ )



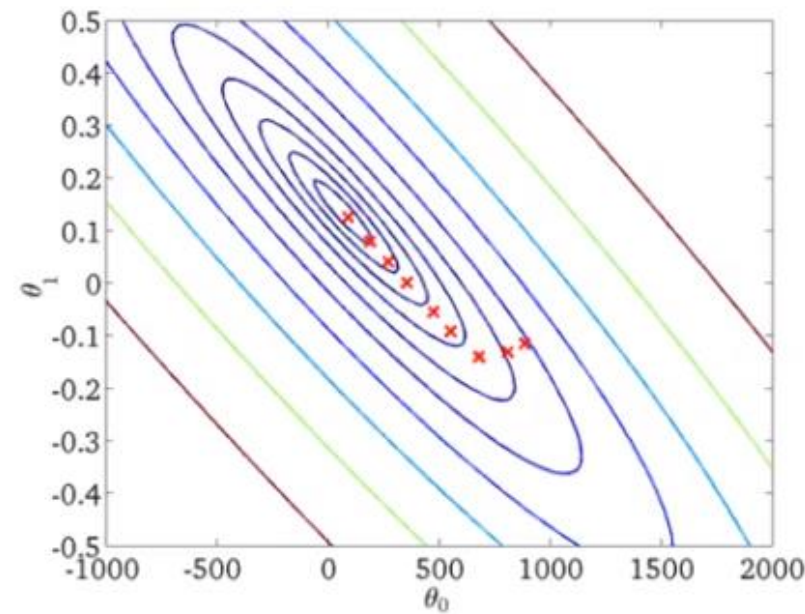
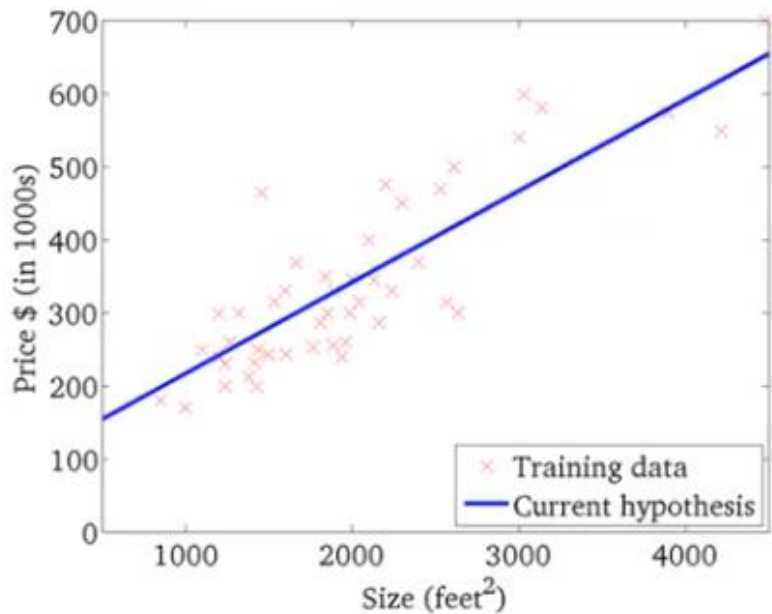
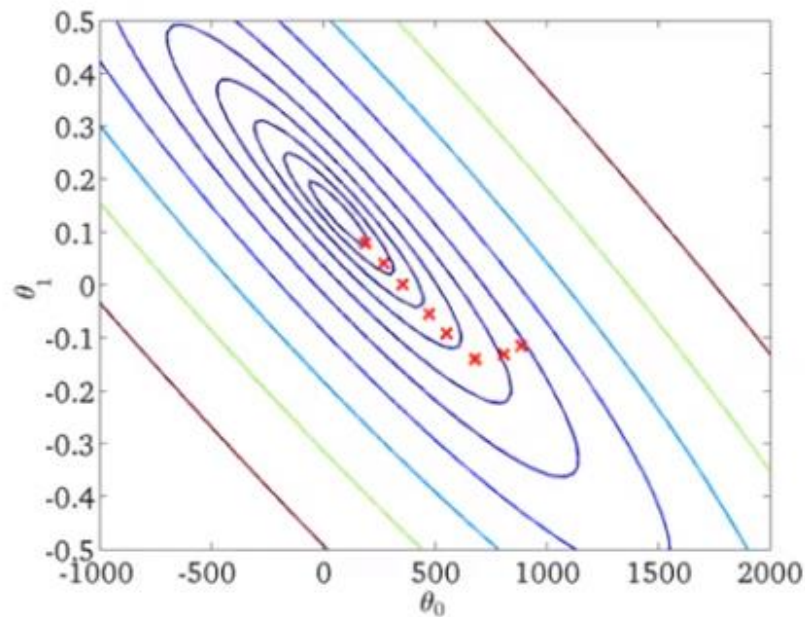
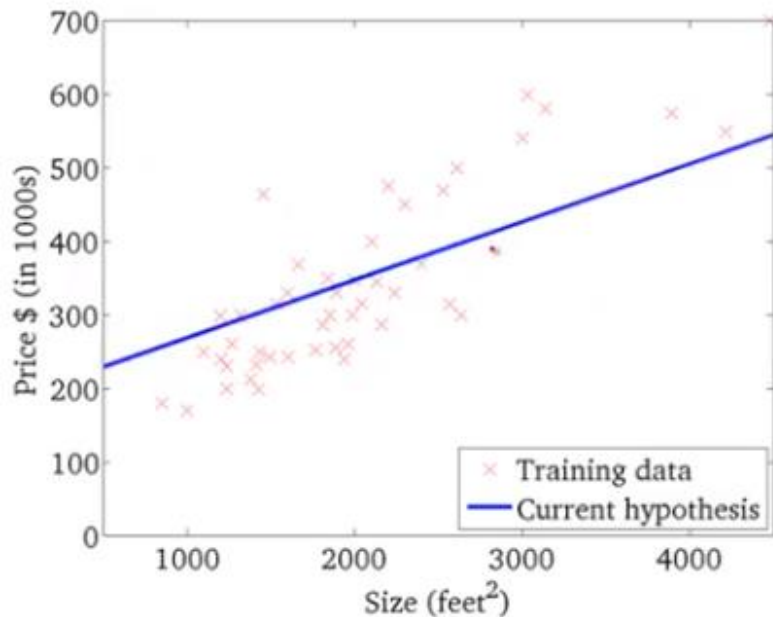








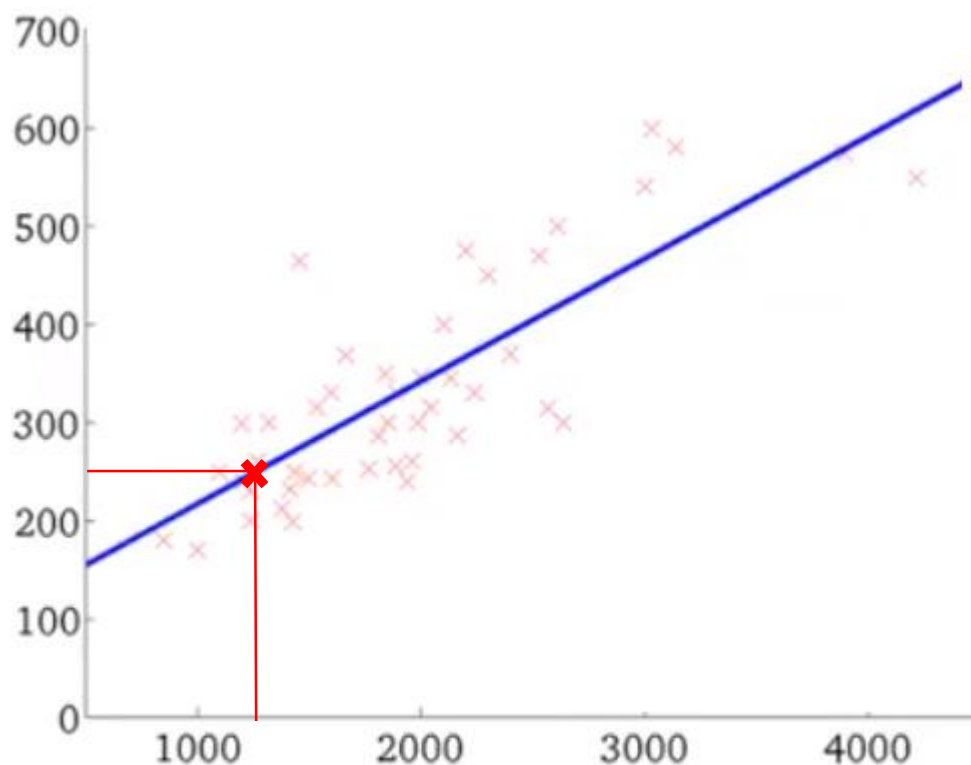




Решение задачи предсказания стоимости недвижимости при помощи линейно регрессии алгоритмом градиентного спуска:

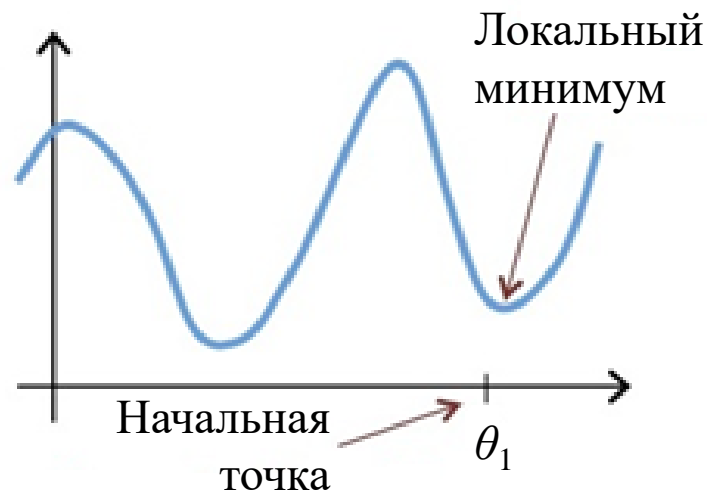
$$h(x)=89.3+0.12x$$

Площадь, м <sup>2</sup>	Цена, тыс. м.к.
2104	460
1416	232
1534	315
852	178
...	...



Для дома площадью  
1250 м<sup>2</sup>  
Предсказанное значение  
стоимости будет  
240 тыс. м. к.

Предположим, что начальная точка  $\theta_1$  находится в локальном минимуме функции  $J(\theta_1)$ , как показано на рисунке:



Чему будет равно значение  $\theta_1$  после выполнения одного шага алгоритма?

- ☒ не изменится
- ☐ изменится в случайном направлении
- ☐ изменится в направлении глобального минимума функции
- ☐ уменьшится
- ☐ увеличится