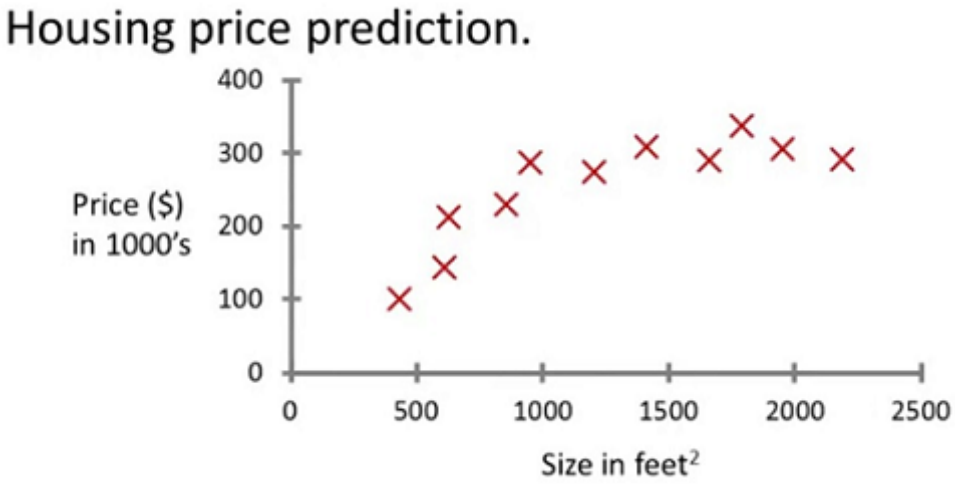
Supervised Learning

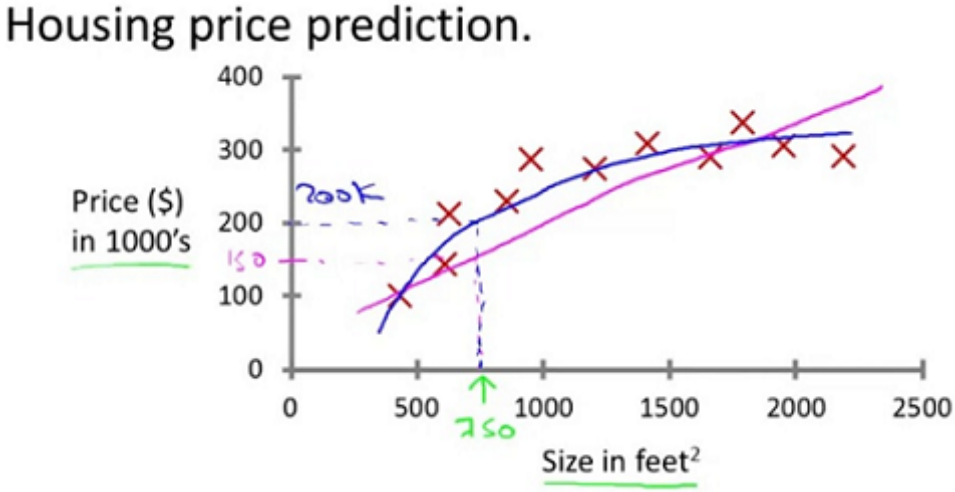
Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labeled training data consisting of a set of training examples.

Example of supervised learning: Let's say we have a friend who owns a house that is say 750 square feet, and they are hoping to sell the house, and they want to know how much they can get for the house. How can the learning algorithm help?



One thing a learning algorithm might be want to do is put a straight line through the data, also fit a straight line to the data. Based on that, it looks like maybe their house can be sold for about $ 150,000.

But maybe this isn't the only learning algorithm we can use, and there might be a better one. For example, instead of fitting a straight line to the data, we might decide that it's better to fit a quadratic function, or a second-order polynomial to this data. If we do that and we make a prediction, then it looks like, maybe they can sell the house for closer to $ 200,000.



Supervised Learning: "right answers" given.

The term Supervised Learning refers to the fact that we gave the algorithm a data set in which the, called, "right answers" were given. That is we gave it a data set of houses in which for every example in this data set, we told it what is the right price. So, what was the actual price that that house sold for, and the task of the algorithm was to just produce more of these right answers such as for this new house that our friend may be trying to sell.

This is also called regression problem. The term regression refers to the fact that we're trying to predict the sort of continuous values attribute. **Regression: Predict continuous valued output (price). No real discrete delineation.**

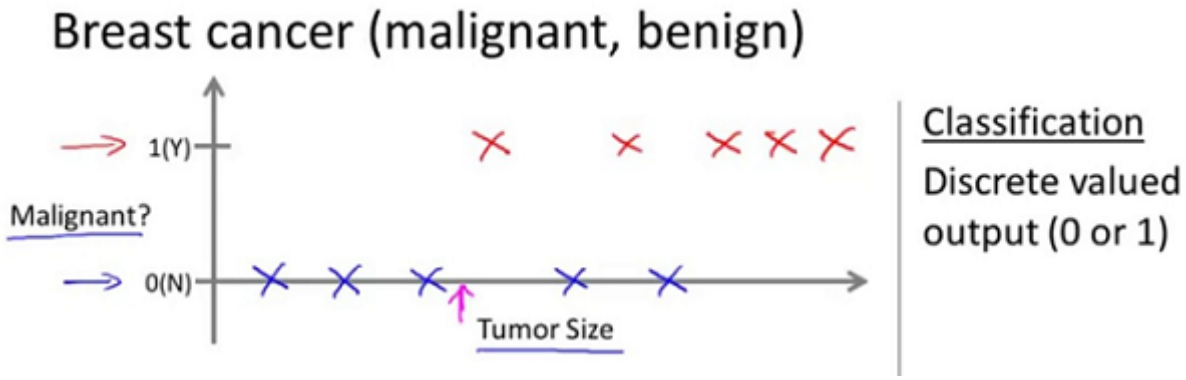
Another example of Supervised Learning:

Let's say we want to look at medical records and try to predict of a breast cancer as malignant or benign. If someone discovers a breast tumor, a lump in their breast, a malignant tumor is a tumor that is harmful and dangerous, and a benign tumor is a tumor that is harmless. Malignang? (Yes/1), (No/0)

Malignant -> 1

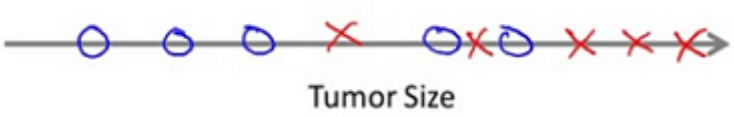
Benign -> 0

This is an example of a **classification problem**. The term classification refers to the fact, we're trying to predict a discrete value output zero or one, malignant or benign. It turns out that in classification problems, sometimes you can have more than two possible values for the output.

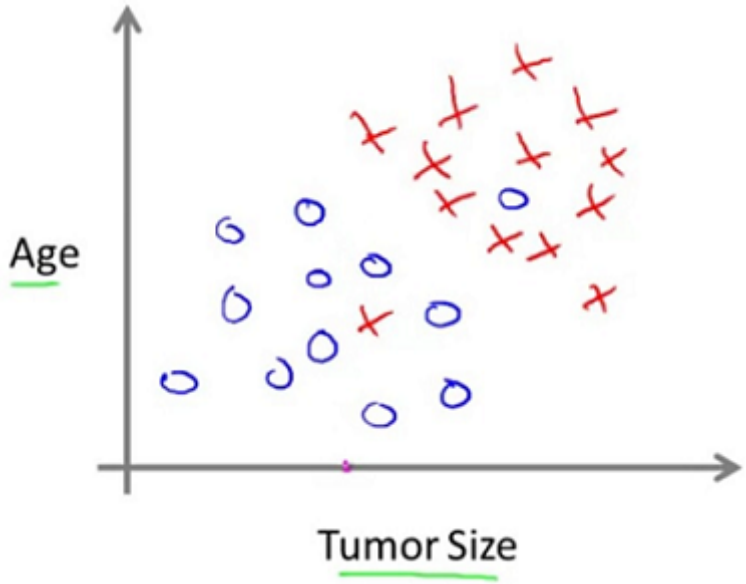


As a concrete example, maybe there are three types of breast cancers. So, we may try to predict a discrete value output zero, one, two, or three, where zero may mean benign, benign tumor, so no cancer, and one may mean type one cancer, and two mean a second type of cancer, and three may mean a third type of cancer. But this will also be a classification problem because this are the discrete value set of output corresponding to we're no cancer, or cancer type one, or cancer type two, or cancer types three.

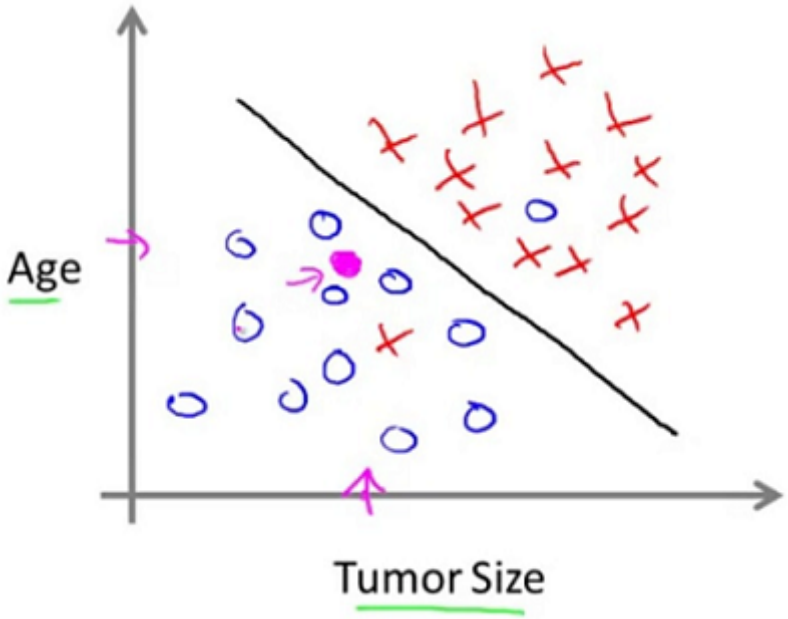
In classification problems, there is another way to plot this data.



In this example, we use only one feature or one attribute, namely the tumor size in order to predict whether a tumor is malignant or benign. In other machine learning problems, when we have more than one feature or more than one attribute. Let's say that instead of just knowing the tumor size, we know both the age of the patients and the tumor size.



So, given a data set, what the learning algorithm might do is fit a straight line to the data to try to separate out the malignant tumors from the benign ones, and so the learning algorithm may decide to put a straight line to separate out the two causes of tumors.



In this example, we had two features namely, the age of the patient and the size of the tumor. In other Machine Learning problems, we will often have more features.

Video Question: You're running a company, and you want to develop learning algorithms to address each of two problems.

Problem 1: You have a large inventory of identical items. You want to predict how many of these items will sell over the next 3 months.

Problem 2: You'd like software to examine individual customer accounts, and for each account decide if it has been hacked/compromised.

Should you treat these as classification or as regression problems?

- Treat both as classification problems.
- Treat problem 1 as a classification problem, problem 2 as a regression problem.

> Treat problem 1 as a regression problem, problem 2 as a classification problem.

- Treat both as regression problems.

Summary of supervised learning

In supervised learning, we are given a data set and already know what our correct output should look like, having the idea that there is a relationship between the input and the output.

Supervised learning problems are categorized into "regression" and "classification" problems. In a regression problem, we are trying to predict results within a continuous output, meaning that we are trying to map input variables to some continuous function. In a classification problem, we are instead trying to predict results in a discrete output. In other words, we are trying to map input variables into discrete categories.

Example 1:

Given data about the size of houses on the real estate market, try to predict their price. Price as a function of size is a continuous output, so this is a regression problem.

We could turn this example into a classification problem by instead making our output about whether the house "sells for more or less than the asking price." Here we are classifying the houses based on price into two discrete categories.

Example 2:

(a) Regression - Given a picture of a person, we have to predict their age on the basis of the given picture

(b) Classification - Given a patient with a tumor, we have to predict whether the tumor is malignant or benign.