

# Machine Learning

Your very first lecture

# Supervised Machine Learning in 1 slide

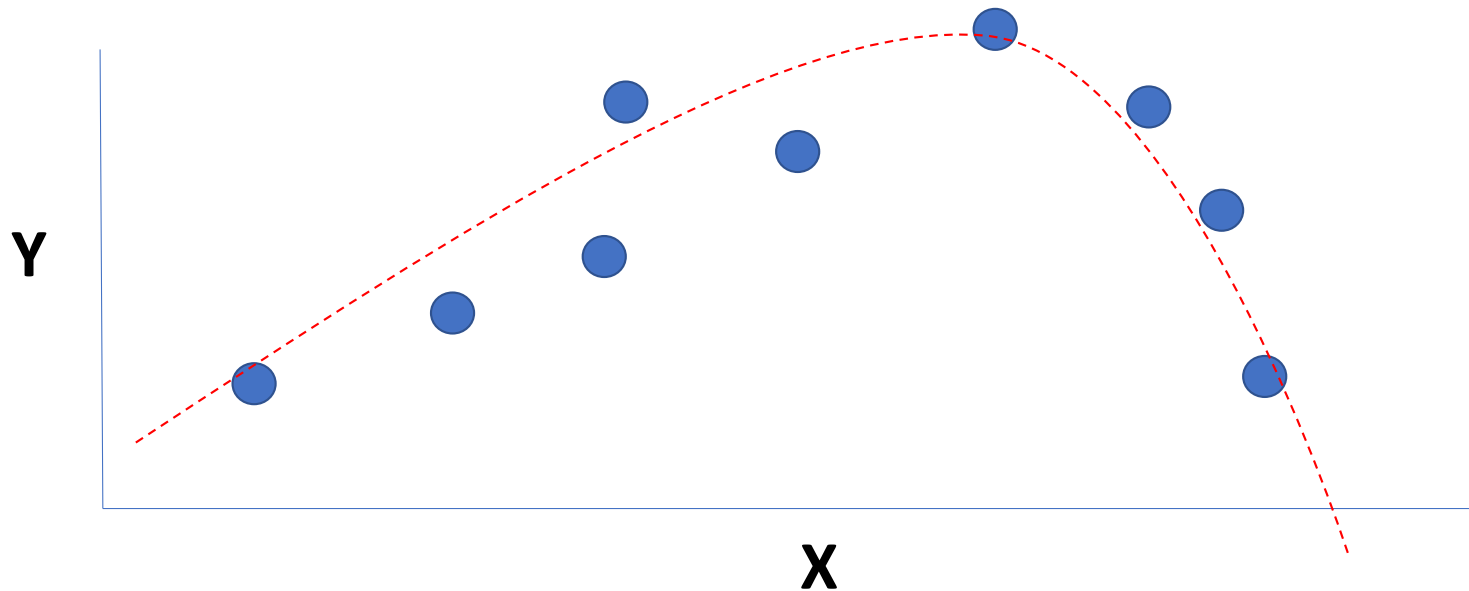
1. Pick data  $\mathbf{D}$ , model  $\mathbf{M}(\mathbf{w})$  and objective function  $\mathbf{J}(\mathbf{D}, \mathbf{w})$
2. Initialize model parameters  $\mathbf{w}$  somehow
3. Measure model performance with the objective function  $\mathbf{J}(\mathbf{D}, \mathbf{w})$
4. Modify parameters  $\mathbf{w}$  somehow, hoping to improve  $\mathbf{J}(\mathbf{D}, \mathbf{w})$
5. Repeat 3 and 4 until you stop improving or run out of time

Pick data  $D$

The data defines a function to learn:  $f(x) = y$

Often, this is from  $\mathbb{R}^d$  to  $\mathbb{R}^d$ .

Learning this function is called **regression**.

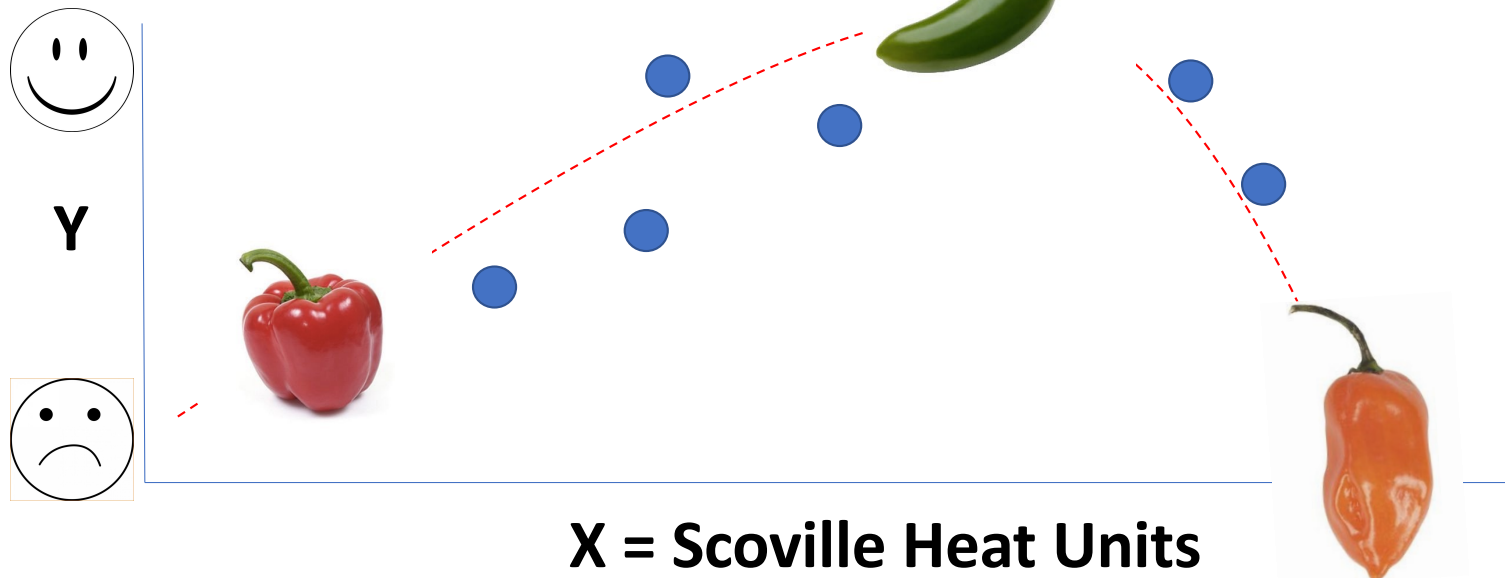


## Pick data D

The data defines a function to learn:  $f(x) = y$

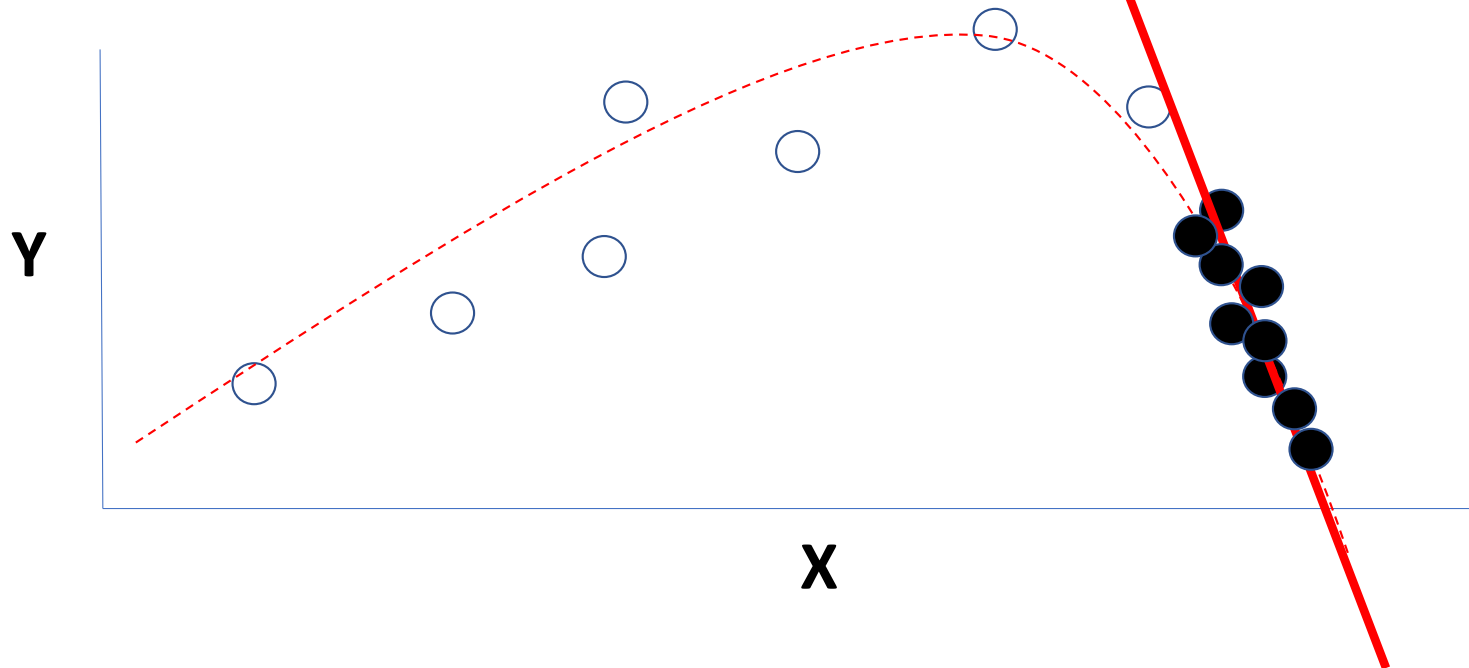
Often, this is from  $\mathbb{R}^d$  to  $\mathbb{R}^d$ .

Learning this function is called **regression**.



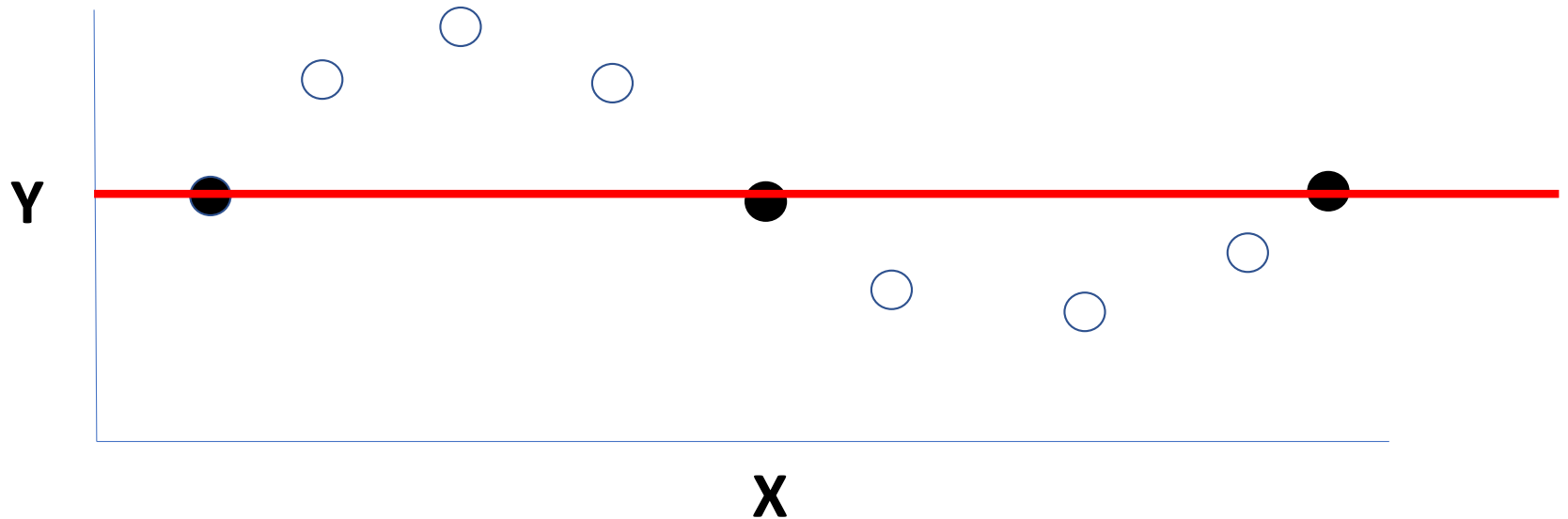
# Pick data D: Is there enough?

- Good coverage of the range of possible values?
- Just because you got lots of data, doesn't mean it covers the space.



# Pick data D: Is there enough?

- Enough density in the space?
- Just because you cover the range, doesn't mean you captured the function.

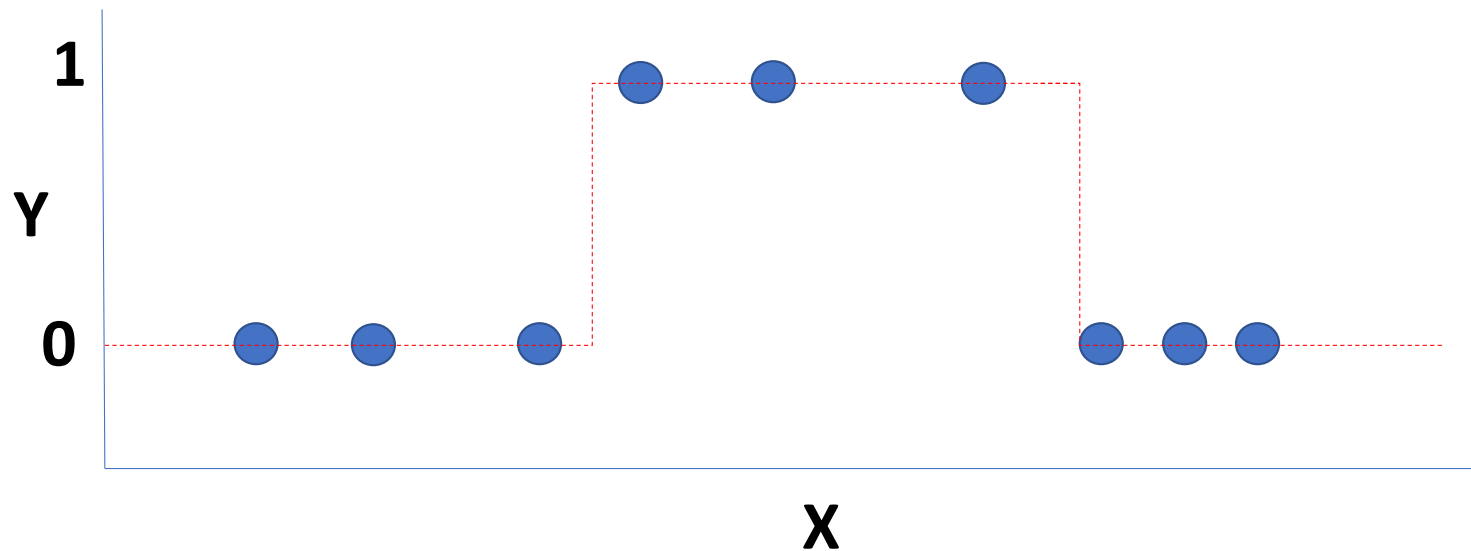


## Pick data D

The data defines a function to learn:  $f(x) = y$

This can also be from  $\mathbb{R}^d$  to a finite set of labels, e.g.  $\{0,1\}$ .

This is **classification**.



# Pick data D

The data defines a function to learn:  $f(x) = y$

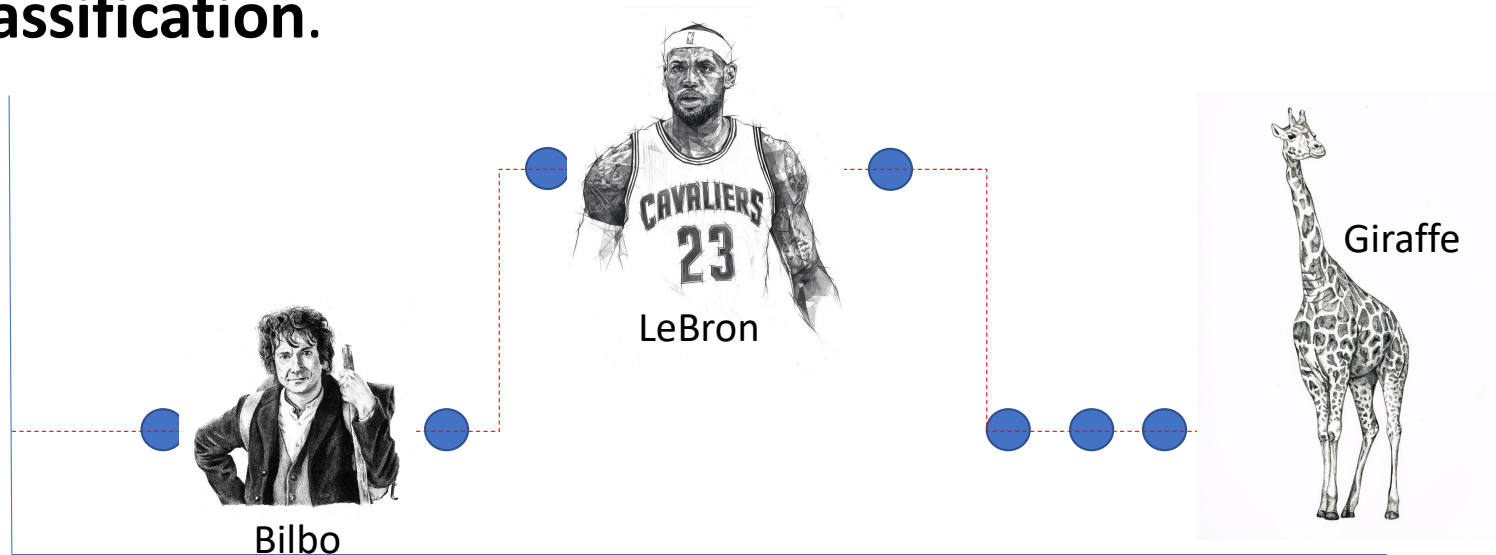
This can also be from  $\mathbb{R}^d$  to a finite set of labels, e.g.  $\{0,1\}$ .

This is **classification**.

In NBA

Y

Not

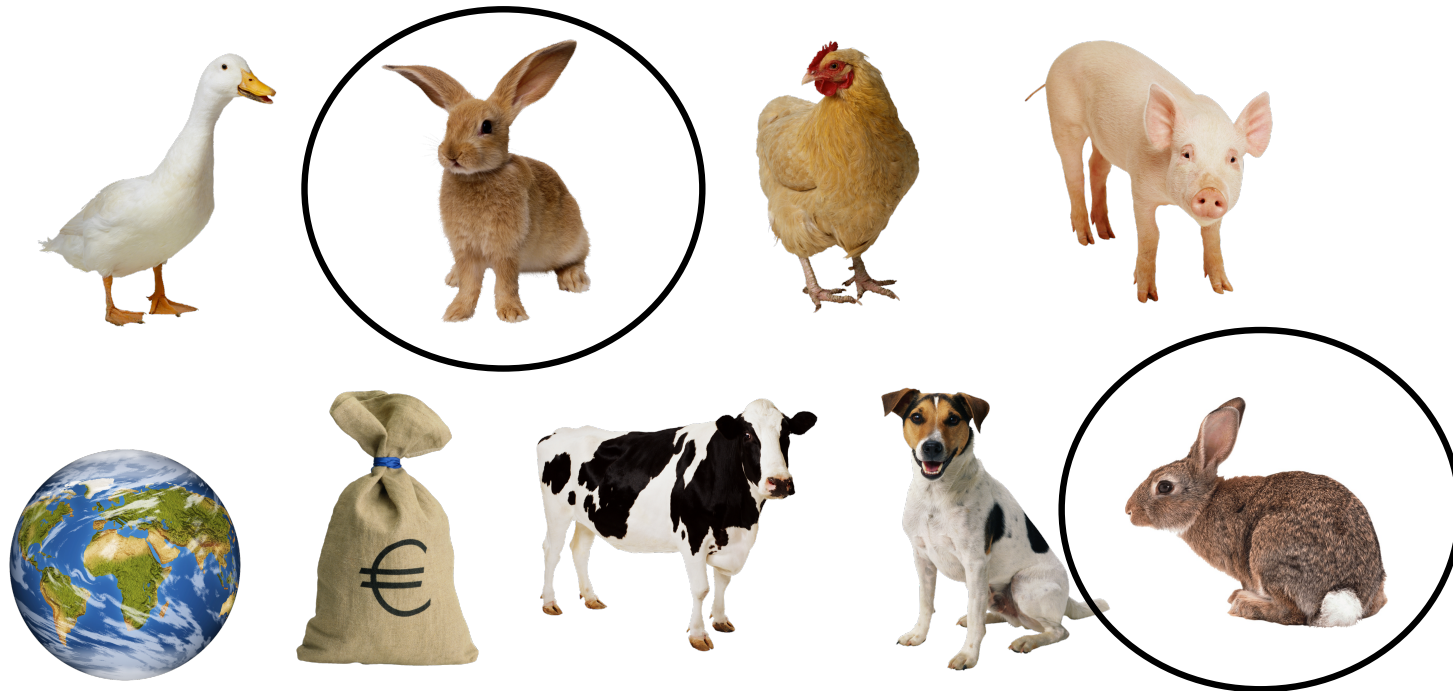


X = Height in CM



# Encoding Matters: Learning "rabbit"

As images, the two rabbits are as distinct from each other as they are from the non-rabbits in our data.



# Encoding Matters: Learning “Rabbit”

We can measure key features to make learning easier, and suppress irrelevant differences. Now both rabbits look identical.



Oops! The dog looks just like a rabbit!



Number of Feet	Fur	Size	Has wings	Warm Blood	$f(x)$
2	No	S	Yes	Yes	0
<b>4</b>	<b>Yes</b>	<b>S</b>	<b>No</b>	<b>Yes</b>	<b>1</b>
2	No	S	Yes	Yes	0
4	No	M	No	Yes	0
0	No	XXL	No	Yes	0
0	No	M	No	No	0
4	Yes	S	No	Yes	0
4	Yes	L	No	Yes	0
<b>4</b>	<b>Yes</b>	<b>S</b>	<b>No</b>	<b>Yes</b>	<b>1</b>

# How many unique instances?

Number of Feet	Fur	Size	Has wings	Warm Blood
Integers 0 to 99	Yes,No	S,M,L,XL,XXL	Yes,No	Yes,No

$100 * 2 * 5 * 2 * 2 = 4000$  instances

Q. How many unique functions to  $\{0,1\}$ ?

A.  $2^{4000}$  unique functions

# The hypothesis space of your model

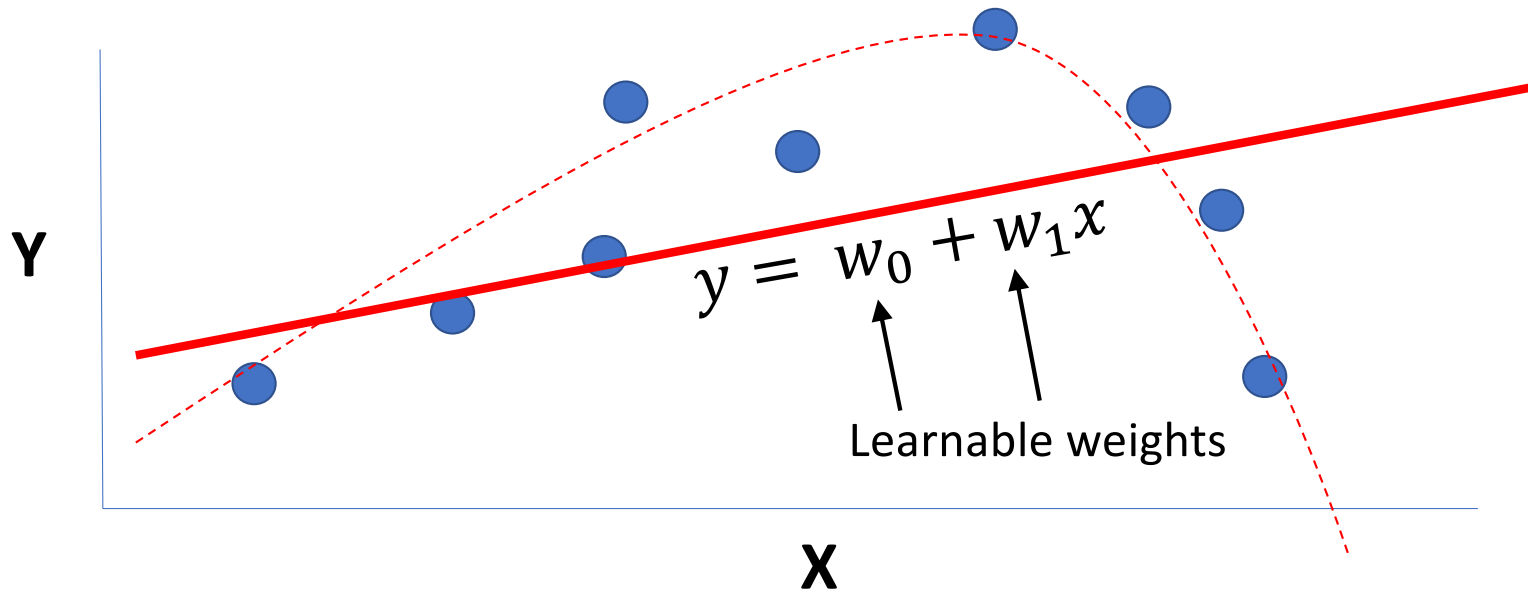
- There are too many concept functions to to try.
- The subset of functions you're willing to actually consider is your hypothesis space.
- The order in which you try the options in your hypothesis space introduces an inductive bias.
- A limited hypothesis space and the search bias are necessary, as the other option (trying every possible function) is impossible.
- So pick the best hypothesis space and search order you can.

# Being unbiased

- The only way to be totally unbiased is to be a “rote learner”
- A rote learner just memorizes its training examples.
- It can't label anything it hasn't seen before because to do so would be to express a bias, somehow.
- This isn't generally practical for real-world use.

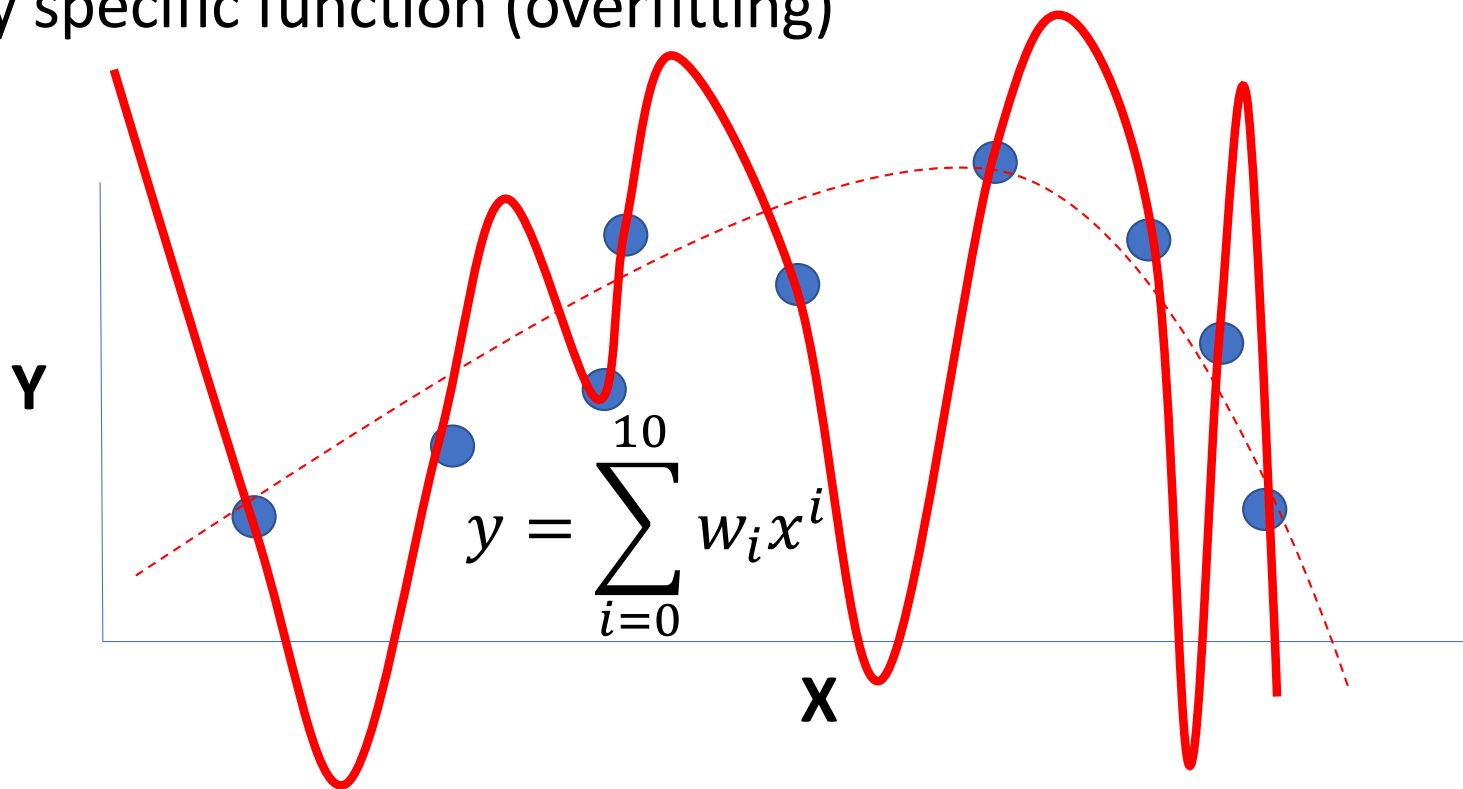
# Fitting & Hypothesis space

If a model's hypothesis space is too small, the true function is probably not in its vocabulary (underfitting)



# Fitting & Hypothesis space

If a model's hypothesis space is too big, it can learn a crazy, overly specific function (overfitting)



# Telling functions apart

- **Definition:** Two functions  $f_1$  and  $f_2$  are *distinguishable*, given the data  $D$ , if they differ in their labeling of at least one of the examples in  $D$ .
- **Definition:** A **set** of hypotheses is distinguishable, given  $D$ , iff ALL pairs of hypotheses in the set are distinguishable given  $D$ .
- Call  $H_D$  a largest set of distinguishable hypotheses, given  $D$ .



# Inductive Learning Hypothesis

- Any hypothesis found to approximate the target function well over the training examples, will also approximate the target function well over the unobserved examples.
- This might not be true. When it isn't the hypothesis does not generalize well.
- In fact, the target concept may not even be in the hypothesis space.
- ...but maybe we can find a hypothesis that is good enough for our purposes

# What kinds of biases are there?

- Choice of data set
  - e.g. Training an image classifier on photos from a foodie website means it won't work well on car photos
- Data representation
  - How you code & represent the data has huge impact
- Hypothesis space
  - e.g. Linear regression only does straight lines and can't fit a curve
- Order in which we select hypotheses to test
- Choice of performance measure (aka loss function, aka objective function)
  - Mean squared error? Maximum Margin? It makes a big difference